Noor Ahmad Shaik
Khalid Rehman Hakeem
Babajan Banaganapalli · Ramu Elango
Editors

Essentials of Bioinformatics, Volume I

Understanding Bioinformatics: Genes to Proteins

Springer

Essentials of Bioinformatics, Volume I

Noor Ahmad Shaik • Khalid Rehman Hakeem
Babajan Banaganapalli • Ramu Elango
Editors

Essentials of Bioinformatics, Volume I

Understanding Bioinformatics: Genes to Proteins

Springer

Editors
Noor Ahmad Shaik
Department of Genetic Medicine
Faculty of Medicine
King Abdulaziz University
Jeddah, Saudi Arabia

Babajan Banaganapalli
Princess Al-Jawhara Center of Excellence
in Research of Hereditary Disorders
Department of Genetic Medicine
Faculty of Medicine
King Abdulaziz University
Jeddah, Saudi Arabia

Khalid Rehman Hakeem
Department of Biological Sciences
King Abdulaziz University
Jeddah, Saudi Arabia

Ramu Elango
Princess Al-Jawhara Center of Excellence
in Research of Hereditary Disorders
Department of Genetic Medicine
Faculty of Medicine
King Abdulaziz University
Jeddah, Saudi Arabia

ISBN 978-3-030-02633-2 ISBN 978-3-030-02634-9 (eBook)
https://doi.org/10.1007/978-3-030-02634-9

Library of Congress Control Number: 2019930435

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

This book is dedicated to King Abdulaziz University, Jeddah, Saudi Arabia.

Foreword

Bioinformatics is an interdisciplinary branch of modern biology which utilizes the advanced computational methods to generate, extract, analyze, and interpret the multidimensional biological data. The rapid technological improvements made, especially in molecular and genetic data generation and collection, an easier task for scientists, but left its handling an extremely complex affair. Thus, the logical gap existing in between high-throughput data generation and analysis cannot be fully bridged without the utilization of computational methods, and this is where bioinformatics comes into the scene. All steps of data collection and analysis have their particular computational tools, and beginners trying to experience bioinformatics are often inevitably lost in the huge amount of ambiguously defined terms and concepts.

The reason for choosing essentials of bioinformatics as the main theme of this book series is highly justifiable. For example, the genome and protein biologists regularly deal with large amounts of lightly annotated data, especially which is produced from large-scale sequencing projects related to animal, microbial, plants, or human studies. The fundamental understanding about important computational methods becomes an unescapable need for them to perform first-hand analysis and interpretation of complex data without relying on external sources.

The chapters in this book mainly discusses about introduction to bioinformatics field, databases, structural and functional bioinformatics, computer-aided drug discovery, sequencing, and metabolomics data analysis. All the chapters provide an overview of the currently available online tools, and many of them are illustrated with examples. The methods include both basic and advanced methodologies, which help the reader become familiar with the manifold approaches that characterize this varied and interdisciplinary field.

This book is designed to allow anyone, regardless of prior experience, to delve deep into this data-driven field, and it is a learner's guide to bioinformatics. This book helps the reader to acquaint basic skills to analyze sequencing data produced by instruments, perform advanced data analytics such as sequence alignment, and variation recording or gene expression analysis. Students and researchers in protein research, bioinformatics, biophysics, computational biology, molecular modeling, and drug design will find this easy-to-understand book a ready reference for staying current and productive in this changing, interdisciplinary field. The editors of this book chose a well-defined target group given the fact that bioinformatics is a fast-evolving field and made a good summary of diverse gene to protein level analysis methods available in present-day bioinformatics field.

It is with these thoughts that I recommend this well-written book to the reader.

Kaiser Jamil
School of Life Sciences and Centre for Biotechnology and Bioinformatics
Jawaharlal Nehru Institute of Advanced Studies
Hyderabad, India

Preface

Bioinformatics is a relatively young subject, compared to numerous well-established branches of biomedical sciences. The recent revolutionary technological developments taken place in interdisciplinary fields like chemistry, biology, engineering, sequencing, and powerful computing methods have greatly contributed to the rapid emergence of bioinformatics as a major scientific discipline. With the exponential accumulation of sequencing data generated from different biological organisms and more forthcoming, accessing and interpreting the highly complex genomic information have become central research themes of current-day molecular biology and genetics. However, mining this huge genetic data generated by whole genome or exome or RNA sequencing methods, often coming giga- to terabytes in digital size, demands for advanced computational methods. It has therefore become the necessity of not just the biologists and statisticians but also computer experts to develop different bioinformatics softwares for tackling the unprecedented challenges in the next-generation genomics.

In the recent decades, we all have witnessed the release of several new easy-to-use and effective bioinformatics tools into the public domain by numerous multidisciplinary academic groups on a regular basis. Since one single textbook cannot cover all bioinformatics tools and databases available, ordinary bench scientists are feeling difficulty in keeping themselves updated with the latest developments in this field. Hence, to support this important task, the current textbook discusses the diverse range of cutting-edge bioinformatics tools in a clear and concise manner.

The current book introduces the reader to essential bioinformatics concepts and to different databases and software programs applicable in gene to protein level analysis, highlights their theoretical basis and practical applications, and also presents the simplified working approaches using graphs, figures, and screenshots. In some instances, step-by-step working procedures are provided as a practical example to solve particular problems. However, this book does not aim to serve as a manual for each software program discussed inside, rather it only briefs the purpose of each tool, its source, and working options, in most instances. Majority of the computational methods discussed in this book is available online, is free to use, and does not require special skills or previous working experience. They are rather straightforward to use and only demand simple inputs like nucleotide or amino acid sequence and protein structures to analyze and return the output files.

Since majority of the chapters in this book are prepared by scientists, who utilize different bioinformatics tools in their day-to-day research activities, we believe that this book will mainly help young biologists keen to learn new skills in bioinformatics. Our authors have taken care in simple presentation of chapter contents, easier enough for any practicing molecular biologist to independently employ bioinformatics methods in their regular research tasks without consulting computational experts. The reader of this book is expected to be familiar with fundamental concepts in biochemistry, molecular biology, and genetics. Our authors have consulted numerous original articles and online reading materials to provide the most updated information in preparing their chapters. Although we have taken a care to cover major bioinformatics tools, but in case if any important tool is missed out, then it is only due to space limitation but not a bias against any particular software program.

The chapters in this book mainly covers the introduction to bioinformatics, introduction to biological databases, sequence bioinformatics, structural bioinformatics, functional bioinformatics, computer-aided drug discovery methods, and some special concepts like in silico PCR and molecular modeling. A total of 17 chapters are included in this book, and most of them are relatively independent from each other. All the chapters in each section are arranged in a logical manner where one chapter acts as follow-ups to the next one. Since this book is basically meant for practicing molecular biologists, few selected molecular or mathematical formulas, which are prerequisites for understanding the corresponding concepts, are used. A general discussion about computational program is often described along with its web links. The conclusion part is provided at the bottom of each chapter to refresh the understanding of readers.

We sincerely thank the Princess Al-Jawhara Center of Excellence in Research of Hereditary Disorders (PACER-HD) and Department of Genetic Medicine, Faculty of Medicine and Department of Biology, Faculty of Science at King Abdulaziz University (KAU) for providing us the opportunity to teach and train the bioinformatics course, because of which we have been able to bring up this book. We thank Prof. Jumana Y. Al-Aama, director of the PACER-HD, KAU, for her excellent moral and administrative support in letting us involved in teaching and training the bioinformatics to young minds in biology and medicine. We also thank Dr. Musharraf Jelani, Dr. Nuha Al-Rayes, Dr. Sheriff Edris, and Dr. Khalda Nasser our colleagues in the PACER-HD for their wonderful friendship. We would also like to thank the chairman of the Department of Biological Sciences, Prof. Khalid M. AlGhamdi, and the head of Plant Sciences section, Dr. Hesham F. Alharby, for providing us the valuable suggestions and encouragement to complete this task. Last but not the least, we would like to acknowledge the support of Springer Nature publishing house for accepting our book proposal, their regular follow-up, and the final publication of this book.

Jeddah, Saudi Arabia

Noor Ahmad Shaik
Khalid Rehman Hakeem
Babajan Banaganapalli
Ramu Elango

Contents

About the Editors

Noor Ahmad Shaik is an academician and researcher working in the field of Human Molecular Genetics. Over the last 15 years, he is working in the field of molecular diagnostics of different monogenic and complex disorders. With the help of high throughout molecular methods like genetic sequencing, gene expression, and metabolomics, his research team is currently trying to discover the novel causal genes/biomarkers for rare hereditary disorders. He has published 63 research publications including 50 original research articles and 13 book chapters in the fields of human genetics and bioinformatics. He has been a recipient of several research grants from different national and international funding agencies. He is currently rendering his editorial services to world-renowned journals like Frontiers in Pediatrics and Frontiers in Genetics.

Khalid Rehman Hakeem is associate professor at King Abdulaziz University, Jeddah, Saudi Arabia. He has completed his PhD (Botany) from Jamia Hamdard, New Delhi, India, in 2011. Dr. Hakeem has worked as postdoctorate fellow in 2012 and fellow researcher (associate professor) from 2013 to 2016 at Universiti Putra Malaysia, Selangor, Malaysia. His speciality is plant ecophysiology, biotechnology and molecular biology, plant-microbe-soil interactions, and environmental sciences and so far has edited and authored more than 25 books with Springer International, Academic Press (Elsevier), CRC Press, etc. He has also to his credit more than 120 research publications in peer-reviewed international journals, including 42 book chapters in edited volumes with international publishers.

Babajan Banaganapalli works as bioinformatics research faculty at King Abdulaziz University. He initiated and is successfully running the interdisciplinary bioinformatics program from 2014 to till date in King Abdulaziz University. He has more than 12 years of research experience in bioinformatics. His research interest spreads across genomics, proteomics, and drug discovery for complex diseases. He published more than 40 journal articles, conference papers, and book chapters. He has also served in numerous conference program committees and organized several bioinformatics workshops and trainings programs and acts as editor and reviewer for various international genetics/bioinformatics journals. Recently, he was honored as young scientist for his outstanding research in bioinformatics by Venus International Research Foundation, India.

Ramu Elango is a well-experienced molecular geneticist and computational biologist with extensive experience at MIT, Cambridge, USA, and GlaxoSmithKline R&D, UK, after completing his PhD in Human Genetics at All India Institute of Medical Sciences, New Delhi, India. At GlaxoSmithKline, he contributed extensively in many disease areas of interest in identifying novel causal genes and tractable drug targets. Dr. Ramu Elango presently heads the Research & Laboratories at the Princess Al-Jawhara Center of Excellence in Research of Hereditary Disorders, King Abdulaziz University. His research focus is on genetics and genomics of complex and polygenic diseases. His team exploits freely available large-scale genetic and genomic data with bioinformatics tools to identify the risk factors or candidate causal genes for many complex diseases.

Chapter 1
Introduction to Bioinformatics

Babajan Banaganapalli and Noor Ahmad Shaik

Contents

1.1 Introduction

Bioinformatics is a fairly recent advancement in the field of biological research, and its contribution towards the cutting-edge medical research is phenomenal. Developing skills in bioinformatics has become essential to all biologists due to its interdisciplinary nature and involvement of cutting-edge technology through which it has effectively provided an insight into the inherent blueprint of molecular cell systems. The subject of bioinformatics is an amalgamation of various concepts derived from domains such as computer science, mathematics, molecular biology, genetics, and statistics and other domains. Bioinformatics can be defined as matrimony between biology and computer science, which routinely solves complex

B. Banaganapalli (✉)
Princess Al-Jawhara Center of Excellence in Research of Hereditary Disorders,
Department of Genetic Medicine, Faculty of Medicine, King Abdulaziz University,
Jeddah, Saudi Arabia
e-mail: bbabajan@kau.edu.sa

N. A. Shaik
Department of Genetic Medicine, Faculty of Medicine, King Abdulaziz University,
Jeddah, Saudi Arabia
e-mail: nshaik@kau.edu.sa

N. A. Shaik et al. (eds.), *Essentials of Bioinformatics, Volume I*,
https://doi.org/10.1007/978-3-030-02634-9_1

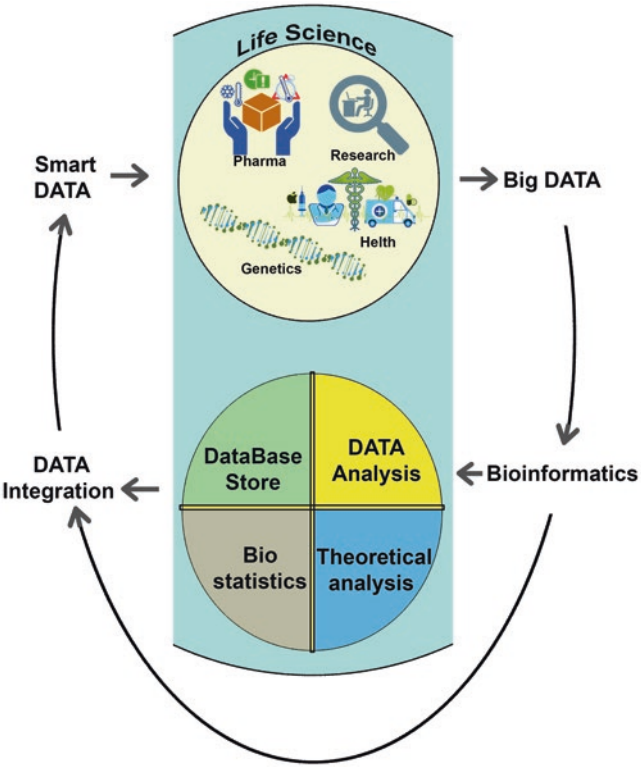


Fig. 1.1 Bioinformatics and biological research

biological problems through computer science, mathematics, and statistics (Can 2014). Bioinformatics endeavors to achieve modeling of complex biological processes at the cellular level and gain insights into the disease mechanisms from the analysis of large volumes of data that is generated. Figure 1.1 given below illustrates the contribution of bioinformatics toward different biological and medical research domains.

Ever since the completion of the human genome project in 2003, biological data generation has increased explosively and the domain of bioinformatics is playing a very key role in drawing critical inference from them (Greene and Troyanskaya 2011). During the Human Genome Project, comprehensive analysis of Human DNA was performed from multiple perspectives. The project started with the sequencing of all the base pairs, discovery of genes (more than 20,000) and mapping of those genes onto the chromosomes leading to further development of linkage maps. The most critical achievement of the Human Genome Project is the complete understanding of the dynamics and modalities of the human gene transcripts, their presentation and localization in the genome and their fundamental molecular functions. The human genome project generated huge amounts of biological data and its significance toward knowledge generation depended on the efficient data mining. The interdisciplinary domain of bioinformatics has played a key role in the data mining process, and in the recent years, a wide range of sophisticated bioinformatics methods, techniques, algorithms and tools which immensely contributed toward extracting the knowledge from major biological data repositories are developed. All of these developments have potentially contributed towards the progression of biological research in general but also facilitated the progress of new scientific domains such as molecular medicine, targeted gene therapy, and in silico drug design approaches to name a few. State-of-the-art bioinformatic programs used in high-throughput sequence analysis allows researchers to accurately monitor and identify the minute genomic alterations on genes, and these kinds of endeavors result in the generation of massive amounts of biological data that requires highly efficient analysis (Goldfeder et al. 2011; Yang et al. 2009). This once again highlights the advantage of bioinformatics over traditional molecular techniques where one can study only a single gene at one time (Jorge et al. 2012; Blekherman et al. 2011; Kihara et al. 2007).

From a biological data analysis perspective, bioinformatics is a robust set of data mining paradigms that help convert textual data into human perceivable knowledge. In every domain of scientific research, computing technologies have virtually become indispensable and the situation is no different in case of bioinformatics (Akalin 2006; Bork 1997; Brzeski 2002). Biological sequencing machines generate tremendous volumes of data, and it has become critically important to archive, organize, and process such data using powerful computational methods efficiently to extract the knowledge. Similarly, computational simulation of complex molecular processes can be an asset as it allows researchers to gain rapid insights in silico without the need for time consuming experimentations (Akalin 2006).

Some of the key contributions of the discipline of bioinformatics include (Akalin 2006):

- Conceptualization, design, and development of biological relational databases for archiving, organizing, and retrieving biological data.
- Development of cutting-edge computer algorithms to model, visualize, mine and compare biological data.
- Highly intuitive and user-friendly curation of biological data that will help biological researchers who lack IT knowledge to derive useful representation of information.

Bioinformatic tools and techniques have become virtually ubiquitous in modern biological research. From a molecular biology perspective involving microarrays and sequencing experiments, bioinformatics techniques can be used for efficient analysis of raw transcript signals and sequence data. In the field of genetics and genomics, bioinformatics can assist in the analysis of sequencing data, annotation of genomic landmarks, and identification of genetic mutations that could have a direct correlation with disease pathology (Batzoglou and Schwartz 2014; Yalcin et al. 2016; Zharikova and Mironov 2016). Furthermore, bioinformatics data mining tools can also be highly effective in the analysis and extraction of knowledge from biological literature so that sophisticated gene ontologies could be conceptualized for future query and analysis. Bioinformatics tools and algorithms also find application in the field of proteomics such as the analysis of protein expression and its regulation. Bioinformatics can contribute toward the analysis of biological pathways and molecular network analysis from a systems biology approach (Hou et al. 2016). In the field of structural biology, bioinformatics tools can help in the in silico simulation and modeling of genomic and proteomic data that can help researchers better understand the key molecular interactions. A key domain where bioinformatics has immensely contributed is, in the field of biomedicine. It will not be wrong to comment that every human disease is somehow connected to a genetic event. In this context, the complete draft of the human genome has greatly helped in the mapping the disease-associated genes and elucidation of their molecular function. As a result it has become possible for researchers to gain comprehensive insights into pathogenesis at the cellular level, thus creating grounds for the development of effective therapeutic interventions. Through the use of the state-of-the-art bioinformatics and computational tools, it has now become possible to simulate, identify, and establish potential drug targets that will have much greater efficacy against diseases with minimal side effects. With the advent of different bioinformatics paradigms, it has now become possible to carry out analysis of an individual's genetic profile leading to the innovative concept of personalized medicine. In domains such as agriculture, bioinformatics tools can be used to alter the genomic structure so that there is an increase in the resistance of crops toward different plant pathogens and insects (Bolger et al. 2014; Edwards and Batley 2010).

1.2 The Role of Bioinformatics in Gene Expression Data Analysis

A molecular biology experimental paradigm that is extensively used to decipher the inherent characteristics and functionality of an entire cohort of gene transcripts on a genome is microarray. A major strength of the microarray is its ability to characterize the complete genes using a single microarray chip or plate (Konishi et al. 2016; Li 2016). Microarrays contain a solid substrate where protein and DNA are printed as microscopic spots. DNA microarray chip may contain short single-stranded oligonucleotides or large double-stranded DNA strands. These spots are called as probes, and they hybridize with the cDNA samples from the control and the test subjects whose identity is unknown (Konishi et al. 2016; Li 2016). The microarray process starts with the extraction and purification of RNA from the control and test samples and their transformation into cDNA using reverse transcriptase enzyme. To differentiate the test and control samples, the cDNA from control samples is labelled with green Cy3 fluorescent dye while the test sample cDNA is labelled with red Cy5 fluorescent dye. The role of the fluorescent dye is to help the researchers accurately estimate the expression profile of the genes. After the experiment is complete, microarray chips/plates are scanned using a scanning device, and the data is collected for analysis. The probes (both test and control) offers a complete representation of the genes that have completed the process of transcription. Certain gene transcripts may show hybridization with both the test and the control samples and appear as yellow spots on the array (Frye and Jin 2016; Non and Thayer 2015; Soejima 2009). At this juncture, the role of

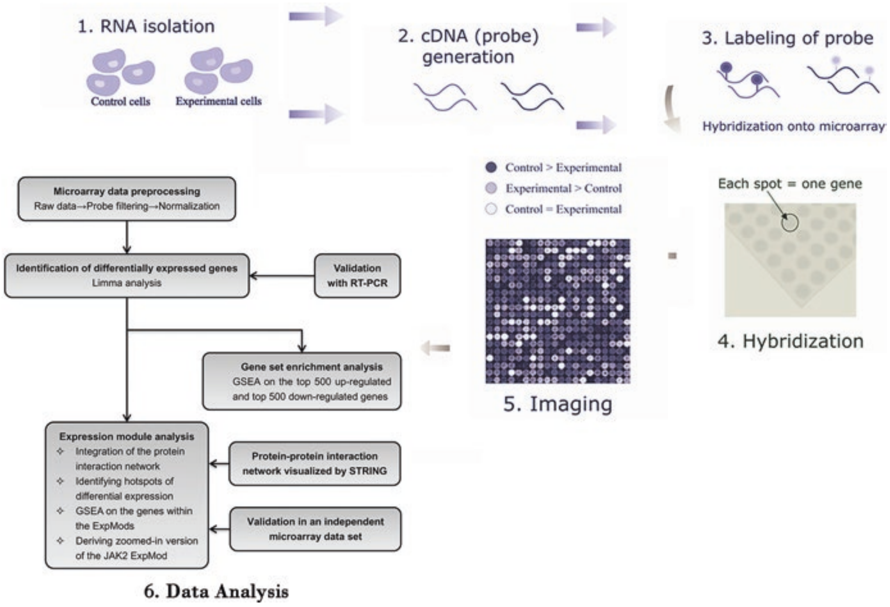

1. RNA isolation
Control cells
Experimental cells
2. cDNA (probe) generation
3. Labeling of probe
Hybridization onto microarray
Each spot = one gene
4. Hybridization
● Control > Experimental
● Experimental > Control
○ Control = Experimental
5. Imaging
Microarray data preprocessing
Raw data→Probe filtering→Normalization
Identification of differentially expressed genes
Limma analysis
Validation with RT-PCR
Gene set enrichment analysis
GSEA on the top 500 up-regulated and top 500 down-regulated genes
Expression module analysis
Integration of the protein interaction network
Identifying hotspots of differential expression
GSEA on the genes within the ExpMods
Deriving zoomed-in version of the JAK2 ExpMod
Protein-protein interaction network visualized by STRING
Validation in an independent microarray data set
6. Data Analysis

Fig. 1.2 The microarray workflow

bioinformatics has become critical with regard to analyzing and deciphering the microarray data to extract the desired information. Bioinformatics data mining tools and work pipelines such as cluster analysis, and heatmap show high efficacy in microarray data analysis (Bodrossy and Sessitsch 2004; Loy and Bodrossy 2006). A typical microarray experiment workflow is presented in Fig. 1.2 (Macgregor and Squire 2002).

1.3 The Role of Bioinformatics in Gene/Genome Mapping

High-throughput sequencing technologies have allowed us in generating whole genome sequencing data at a phenomenal rate. Today a wide range of user-friendly genome browsers are available on the web which can help researchers analyze their gene or genome landmark of interest with the click of a button (Mychaleckyj 2007). Many such genome browsers offer cutting-edge genome mapping tools but it needs to be noted that de novo genome mapping is still a necessity for many researchers. The technique of gene or genome mapping involved the characterization of key landmarks on the genome. The fundamental aim of genome mapping is to create a comprehensive map of the genome of interest and identify landmarks such as gene regulatory sites, short-sequence repeats, single nucleotide polymorphism, or even the discovery of completely new gene transcripts (Waage et al. 2018; Brown 2002). Researchers across the globe are carrying out more and more sequencing experiments with each passing day, and the fundamental question that they face concerns

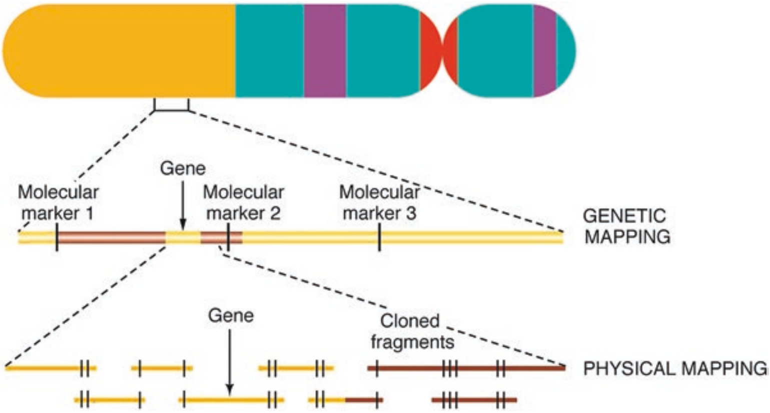


Fig. 1.3 Illustration of large-scale mapping of a gene by genetic mapping and physical mapping

the need for genome maps. These maps allow identification of the key genetic features that could have major therapeutic significance (Brown 2002).

In silico gene mapping (Schadt 2006), uses the publicly available genomic databases to discover and map genes on the genome. In silico gene mapping techniques using bioinformatic applications have been very successful in identifying or mapping QTLs (quantitative trait loci) that help understand the molecular pathways associated with the pathogenesis of different polygenic diseases (Burgess-Herbert et al. 2008). The excellent and informative BAC (bacterial artificial chromosome) clone contig map for Sus scrofa is achieved with help of in silico mapping approaches. In silico allows efficient and fast mapping of genes, which may specifically influence resistance or susceptibility of a particular animal towards a specific disease. Moreover the availability of excellent computational resources, public genomic and phenotypic databases like UniGene and GenBank, and tools like BLAST considerably accelerates the mapping process, when done in silico as compared to the conventional methodologies that are more labor intensive, costly, and time consuming (Schadt 2006). The simple workflow of the gene mapping technique is illustrated in Fig. 1.3.

1.4 Role of Bioinformatics in Sequence Alignment and Similarity Search

Deciphering the similarity between DNA, RNA, or protein sequences is a critical step toward gaining insights into different biological processes and disease pathways. Sequence similarity can help researchers identify and characterize candidate genes or single nucleotide polymorphisms that can predispose an individual toward major pathological conditions such as cancer and autoimmune diseases. Sequence similarities can be determined by carrying out pairwise or multiple sequence alignments that can help in the identification of similar patterns and characteristics. In essence, sequence alignment is the key when it comes to the identification of functional and evolutionary relationships between different organisms. During the process of sequence alignment, the sequences are arranged one over the other and a match is determined. If two nucleotide entries in the test sequences show a mismatch, then a gap in the form of a special character such as "-" is introduced. Mismatches can also have key molecular, biological, and evolutionary significance (Taylor and Triggle 2007). A gap in a sequence could be due to an insertion of a new base or the deletion of an existing base. Sequence alignments in general can be either global or local (Taylor and Triggle 2007). An alignment is considered global when the entire length of the two query sequences are matched.

On the other hand, a local alignment is about achieving alignments of specific regions in the matching sequences (Chandramouli and Qian 2009). From a bioinformatics perspective, the paradigm of sequence similarity is an assent when it comes to a carrying out a wide variety of molecular analysis. Different bioinformatics sequence alignment algorithms are available today, and all of them implement probabilistic models with regard to presence of either matches or mismatches (Hickey and Blanchette 2011). The probability of sequence matches are obtained from curated repositories that archives sequence match incidents. When seen from an evolutionary perspective, the databases which contains multiple sequence alignments that are properly vetted by researchers and curators are considered to be accurate (Eddy 2009).

With regard to the bioinformatics algorithms that are commonly used for sequence similarity searches, BLAST is undoubtedly the most commonly used software platform. The BLAST algorithm is based on a heuristic approach that tries to create an alignment without gaps and with the objective of creating an optimal local similarity score. The USP of the BLAST algorithm does not lie on its speed of execution but rather on the assignment of a value called the p-value that determines the quality of the alignment (Altschul et al. 1990). The BLAST algorithm carries out comparison of nucleotide as well as amino acid sequence pairs to look for local similarities (Altschul et al. 1990; Babajan et al. 2011). The BLAST algorithm not just looks for similarities in nucleotide and amino acid sequences but also helps in the detection of key sequence features such as motifs and genes that code proteins responsible for disease phenotypes. There are many servers available on the web that allows BLAST similarity searches but the most popular one used for nucleotide and protein applications is NCBI BLAST (Pertsemlidis and Fondon 2001). The basic BLAST workflow is presented in Fig. 1.4.

As illustrated in Fig. 1.4, the sequence similarity process during BLAST proceeds with the processing of the input sequences followed by the construction of a reference lookup table to be used for sequence matching. Once the table is ready, the actual sequence comparison starts between the input sequence and the lookup table created earlier. If hits are detected, the algorithm tries to extend the alignment to the maximum limit, and a score is assigned. The alignments that have been assigned an acceptable score are retained and they undergo further analysis with the objective of detecting other types of genomic landmarks such as insertions or

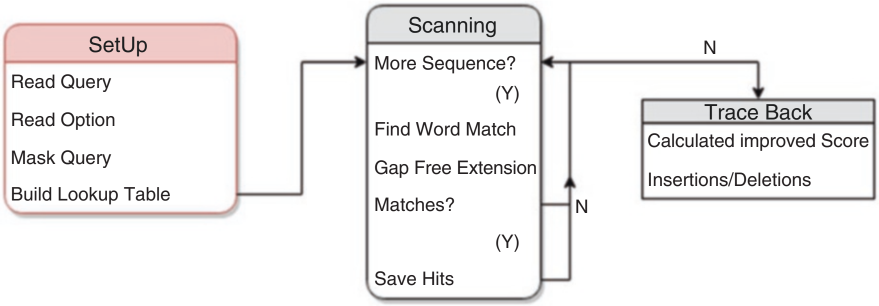


Fig. 1.4 BLAST working principle

deletions (Camacho et al. 2009). With regard to protein sequences, BLAST sequence alignment is of key utility to look for phylogenetically and functionally significant conserved regions. The presence of such conserved regions could very well indicate a homology between different protein groups or families that were otherwise thought to be phylogenetically distant (Varon and Wheeler 2012). It may be possible that the similarity score between two protein sequences is not that significant but the presence of conserved regions between the two could have tremendous biological and therapeutic significance.

Furthermore, protein sequence alignment algorithms such as BLAST can also help in detecting conserved structures, and this can help in the in silico prediction and simulation of protein structures that were previously unknown (Varon and Wheeler 2012). Different BLAST distributions that are available through the NCBI portal include the following (Koltes et al. 2009; Neumann et al. 2014; Madden et al. 1996):

- BLASTn: BLAST algorithm for nucleotide sequence alignments.
- BLASTp: BLAST algorithm for amino acid sequence alignments.
- BLASTx: A BLAST algorithm that uses a nucleotide sequence as the input to compare with a protein database.
- TBLASN: It search translated nucleotide databases using a protein sequences. The TBLASx algorithm translates the nucleotide database during the alignment process.

While pairwise alignments are resource intensive to carry out, alignment of multiple sequences can be even more computationally intensive. The dynamic programming algorithm implemented by BLAST can be extended for multiple alignments as well, but the time of execution becomes large and virtually impractical. To deal with this issue, a heuristic approach is adopted where all the possible pairwise alignments are created first and then assembled in a stepwise manner. The approach involves alignment of clusters of aligned sequences to create a final alignment. Some of the popular bioinformatics programs used in multiple sequence alignment are T-COFFEE, ClustalW, and MUSCLE (Thompson et al. 2002; Di Tommaso et al. 2011; Edgar 2004).

1.5 Contribution of Bioinformatics toward Modern Cancer Research

The burden of cancer on the global healthcare sector is increasing at a tremendous rate, and according to a study by Fitzmaurice et al. (Global Burden of Disease Cancer et al. 2018), in a span of 10 years between 2005 and 2015, the global cancer incidence rate has gone up by 33%. The interdisciplinary domain of bioinformatics can immensely contribute toward cancer research through the development of in silico tools, which can provide insights into molecular cancer pathways and help in the better understanding of the pathogenesis and cancer progression dynamics. The cancer bioinformatics discipline utilizes the information derived from clinical informatics and medical informatics fields to understand, conceptualize, and develop effective diagnostic and therapeutic interventions against different forms of human cancers. Researchers have carried out numerous studies which have focused on the development of highly effective semantic models that brings together different types of genomic and transcriptomic data gathered from cancer samples and their integration with gene ontology data for the construction of cancer networks (Holford et al. 2012). Efforts such as these facilitate the evaluation and better understanding of the molecular response toward anticancer interventions that are currently in practice. There are state-of-the-art bioinformatics utilities such as miRTrail that were developed with the objective of gaining key insights into the modalities of interaction taking place between genes and miRNA molecules. Such insights can contribute immensely towards a better understanding of the underlying malignant processes. According to Laczny et al. (Laczny et al. 2012), the miRTrail utility can be used to integrate information collected from more than 1000 miRNA samples, 20,000 genes, and more than 250,0000 molecular interactions so that it becomes possible for researchers to comprehend the regulatory mechanisms that take place during the development of different types of cancer. Another major benefit of cancer bioinformatics is its ability to identify potential cancer biomarkers that could be associated with different cancer phenotypes. Information on cancer biomarkers can be highly useful in malignancy characterization, early diagnosis, and development of more effective interventions.

Within the domain of cancer bioinformatics, analysis of the expression profiles of gene transcripts that are associated with different cancer phenotypes is a well-established approach (Subramanian et al. 2004). Cancer genes that have got connection with different types of malignancies have unique expression profiles compared to their normal genes. The identification and characterization of these genes using bioinformatics tools can help in discovering new cancer markers or signatures with tremendous therapeutic and diagnostic significance. Characterization of cancer markers can contribute towards the identification of population groups that are under higher risk of developing cancer and can even help in predicting the outcome of treatment interventions (Jones et al. 2005). A thorough bioinformatics analysis of the cancer genes is also very important because it can help in the development of new anticancer interventions with molecular entities such as siRNAs. It is interesting to note that the diverse range of bioinformatics tools can be used in studying molecular basis of cancer depending on the inheritance mode of the disease, pathogenic mechanisms, and metastatic characteristics. Within the discipline of bioinformatics, the sub-domain of clinical bioinformatics is highly effective when it comes of early diagnosis and effective intervention because it integrates multiple research paradigms such as clinical informatics, omics technology, and medical informatics (F. Wang et al. 2011). In the post-genomic era, diagnosis and treatment of different forms of cancer is no longer limited to the clinical setting alone. Effective cancer management in the present era is a combined outcome following the integration of multiple disciplines such as computational tools, software utilities, and biological databases that help in the identification and characterization of gene markers, quantitative trait loci, and candidate genes that have direct association with different forms of malignancies. Given the regular developments taking place in the domain of cancer bioinformatics, it will not be wrong to assume that it will play a much more significant role in the very near future towards establishing valid relationships between potential cancer candidate genes and cancer phenotypes. Accurate identification of such cancer phenotypes can be highly instrumental in achieving early diagnosis that would eventually contribute towards the early start of treatment and better prognosis thresholds (Wang et al. 2018).

Furthermore, valid characterization of cancer phenotypes can also facilitate continuous monitoring of the treatment outcomes (Y. Wang et al. 2018). Efficient use of bioinformatic resources for the identification of cancer-associated markers can help healthcare providers make the most informed decision with regard to selecting the best intervention from a range of available options such as surgery, radiation therapy, and chemotherapy (Katoh and Katoh 2006). When a biomarker associated with non-metastatic tumors is detected, then it is advisable to go for surgical intervention, and if biomarker associated with highly metastatic cancer is are detected, then combinations of radiation and chemotherapy are more advisable (Katoh and Katoh 2006). Bioinformatics analysis of the transcriptome has allowed researchers to identify biomarkers associated with malignancies of the lung, uterus, and esophagus (C. Kihara et al. 2001).

Pathogenic genetic mutations in the human genome can predispose an individual to cancer and even affect the cancer intervention paradigms in a negative manner (Mount and Pandey 2005). One of the most common forms of variation that is present in genomes is single nucleotide polymorphism, commonly referred to as SNP, and researchers have established that these polymorphisms could have serious implications with regard to cancer pathogenesis and progression (Zienolddiny and Skaug 2012). The role of cancer bioinformatics in this context is established once again in the sense that bioinformatics analysis of cancer-related SNPs is emerging as an effective intervention strategy.

When there is a single-base substitution in the genome, the event is referred to as single nucleotide polymorphism, and this could have immense medical and phylogenetic significance. The occurrence of SNPs is rather common, and most of them are harmless without any pathological significance. However some SNPs can result in serious medical conditions such as sickle cell anemia putting immense pressure on the global healthcare sector (Botstein and Risch 2003). After the release of complete draft of the human genome, numerous medical research studies have identified more than ten million single nucleotide polymorphisms that are uniformly distributed across the entire length of the human genome (Carlson et al. 2003). Single nucleotide polymorphisms are a direct result of point mutational events; their presence in the genome is very diverse from one human population to the another (Erichsen and Chanock 2004). Genome-wide association studies have been able to define the dynamics of the association between SNPs and cancer, and it has been established that SNPs can either make an individual or a population of individuals susceptible to a particular form of cancer or even influence the outcome of cancer treatment (Erichsen and Chanock 2004).

A good example is the MPO gene carrying an SNP (−463G>A) that can predispose Caucasians to different forms of lung malignancies (Cascorbi et al. 2000). Bioinformatics tools can also help in the identification of pharmacogenetic markers through the identification and analysis of SNPs, characterization of the gene expression profiles, and delineation of drug metabolism pathways (Wilkinson 2005). Identification and characterization of pharmacogenetic markers can be extremely helpful in predicting if there will be any drug induced side effects in the patient and also in estimating the optimal drug dose (Need et al. 2005; Banaganapalli, Mulakayala, D, et al. 2013a; Banaganapalli, Mulakayala, Pulaganti, et al. 2013b). Following the establishment of the role of SNPs in cancer pathogenesis and cancer therapeutic outcomes, a comprehensive SNP database called dbSNP was developed by NIH. This is an excellent bioinformatics resource that helps in the identification and characterization of not just cancer-related SNPs but also other forms of genomic variants such as insertions and deletions, microsatellites, and short tandem repeats (Sayers et al. 2012).

1.6 The Domain of Structural Bioinformatics

The unprecedented advancements in high-throughput molecular methods, like next generation sequencing technology, microarray based gene expression assays and mass spectrophotometric identification of metabolites, have made the developments in structural biology inevitable. Furthermore, the elucidation of molecular structures or theoretical analysis of proteins has become a central step in understanding the complex biological mechanisms and also in drug discovery processes (Gutmanas et al. 2013). In this context, the bioinformatics component of structural biology, known as structural bioinformatics, which deals with both prediction and analysis of 3 dimensional structures of proteins, DNA and RNA rapidly became a potential alternate option compared to labor intensive traditional laboratory methods. Structural bioinformatics utilizes the information from both experimentally proved molecular structures and computationally predicted models to shed light into the structural organization of bio molecules in terms of molecular folding, motifs, interacting residues, binding affinity and structure-function relationships (Samish et al. 2015; Al-Abbasi et al. 2018). Computational methods not only provides useful predictions about the molecular structures and functions, but they can also provide the in depth analysis of structures solved through experimental approaches.A wide range of bioinformatics tools useful in amino acid sequence analysis, modeling, and structural visualization have been developed in the recent years. For example, the protein data bank (PDB) is the largest model repository which holds the structures of more than 130000 proteins, in addition to 3200 nucleic acids and 7200 nucleic acids-protein structure complexes. The vast molecular structural information available in PDB have not only allowed us in predicting ligand binding sites and active sites (statistical methods) in the target proteins, but also to predict other homologous proteins and to elucidate the structure and functional relationships (Banaganapalli et al. 2017; Vaser et al. 2016; Shaik et al., 2018). Another widely used computational server used in visualizing the 3D structures is modeller. The Modeller utilizes both sequence and structural data as input in FASTA and PDB formats and executes database similarity searches using BLAST and PSI-BLAST methods. Modeller allows efficient and rapid detection of structural homology and template-based modeling of query protein sequences (Altschul et al. 1990). The homology modeling workflow is presented in Fig. 1.5 (Webb and Sali 2017; Shaik et al. 2014; Banaganapalli et al. 2016). The protein structure information when integrated with genomic and metabolomics data, provides multifold information in better understanding the pathways and cellular pathways at the molecular level (Chicurel 2002).

1.7 Bioinformatics Processing of Big Data

With the rapid advancement of high-throughput sequencing paradigms, the rate of generation of biological data has increased at a phenomenal rate. Huge volumes of molecular data is being exponentially generated from sequencing experiments on different clinical conditions. Following the generation of such huge volumes of data, a pertinent need for efficient paradigms that can store and process the data in a highly time-sensitive manner became evident. It quickly became apparent to the researchers that the old statistical data analysis protocols are incapable of handling such data volumes and more sophisticated data analysis paradigms are required (Greene et al. 2014; Chen and Gao 2016). Requirements such as these led to the conceptualization and development of robust bioinformatics techniques that are highly efficient in the identification, evaluation, and characterization of inherent data patterns. Both supervised and unsupervised machine learning protocols are being implemented for rapid and efficient processing of huge data volumes curated

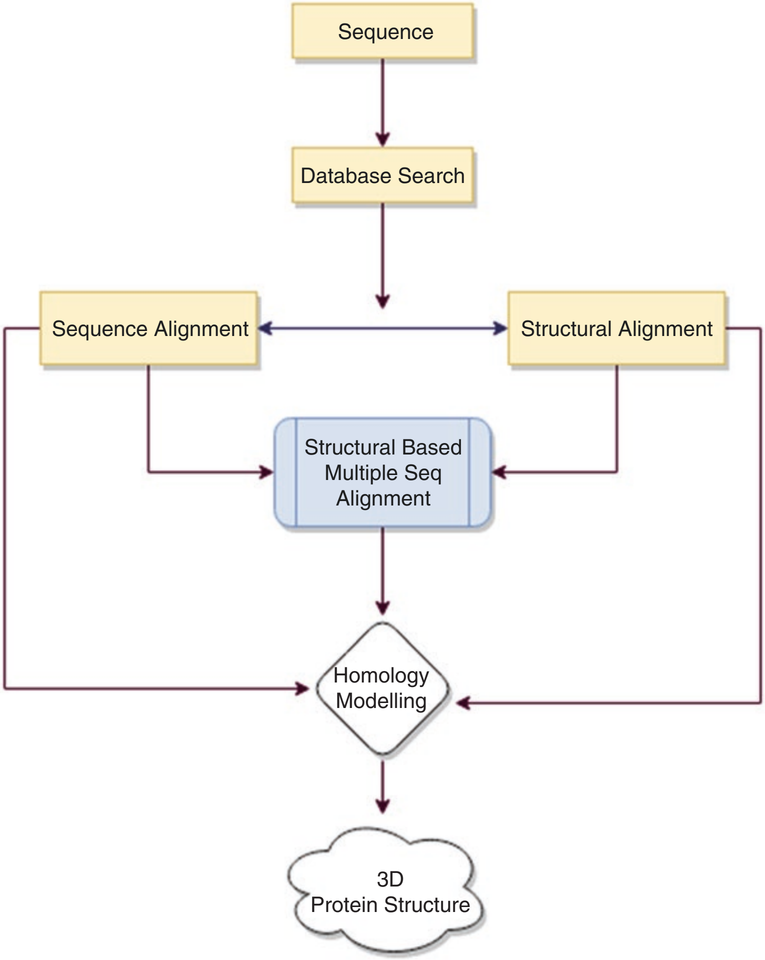


Fig. 1.5 Homology modeling workflow

and stored by repositories such as the Cancer Genome Atlas (Hinkson et al. 2017). State-of-the-art bioinformatics web utilities such as PILGRM (Platform for Interactive Learning by Genomics Results Mining) facilitates rapid and accurate identification of candidate genes that may be implicated in different disease pathogenesis (Greene and Troyanskaya 2011). Big data can be an excellent asset for researchers to discover and identify key findings that could have major ramifications for human and animal health, but there are major challenges associated with efficient data storage, management, and analysis. It is imperative that biologists and bioinformatics researchers collaborate to conceptualize and develop paradigms that can efficiently overcome these impending challenges (Chen and Gao 2016; Greene et al. 2014; Puhler 2017).

1.8 Conclusion

The interdisciplinary domains of bioinformatics have revolutionized the way inferences are drawn from biological data sets. Diverse bioinformatics paradigms for efficient biological data analysis and knowledge interpretation have not only shed light on complex biological processes but have also led to the identification of previously unknown disease markers, contributing immensely toward the development of effective therapeutic interventions. The rapid pace of development taking place in the field of bioinformatics can further bring revolutionary changes in the field of biological data management, archival, and processing.

References

Akalin PK (2006) Introduction to bioinformatics. Mol Nutr Food Res 50(7):610–619. https://doi.org/10.1002/mnfr.200500273

Al-Abbasi FA, Mohammed K, Sadath S, Banaganapalli B, Nasser K, Shaik NA (2018) Computational protein phenotype characterization of IL10RA mutations causative to early onset inflammatory bowel disease (IBD). Front Genet 9:146. https://doi.org/10.3389/fgene.2018.00146

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Babajan, B., Chaitanya, M., Rajsekhar, C., Gowsia, D., Madhusudhana, P., Naveen, M., . . . Anuradha, C. M. (2011). Comprehensive structural and functional characterization of Mycobacterium tuberculosis UDP-NAG enolpyruvyl transferase (Mtb-MurA) and prediction of its accurate binding affinities with inhibitors. Interdiscip Sci, 3(3), 204–216. doi:https://doi.org/10.1007/s12539-011-0100-y

Banaganapalli B, Mohammed K, Khan IA, Al-Aama JY, Elango R, Shaik NA (2016) A computational protein phenotype prediction approach to analyze the deleterious mutations of human MED12 gene. J Cell Biochem 117(9):2023–2035. https://doi.org/10.1002/jcb.25499

Banaganapalli, B., Mulakayala, C., D, G., Mulakayala, N., Pulaganti, M., Shaik, N. A., . . . Chitta, S. K. (2013a). Synthesis and biological activity of new resveratrol derivative and molecular docking: dynamics studies on NFkB. Appl Biochem Biotechnol, 171(7), 1639–1657. doi:https://doi.org/10.1007/s12010-013-0448-z

Banaganapalli, B., Mulakayala, C., Pulaganti, M., Mulakayala, N., Anuradha, C. M., Suresh Kumar, C., . . . Gudla, D. (2013b). Experimental and computational studies on newly synthesized resveratrol derivative: a new method for cancer chemoprevention and therapeutics? OMICS, 17(11), 568–583. doi:https://doi.org/10.1089/omi.2013.0014

Banaganapalli, B., Rashidi, O., Saadah, O. I., Wang, J., Khan, I. A., Al-Aama, J. Y., . . . Elango, R. (2017). Comprehensive computational analysis of GWAS loci identifies CCR2 as a candidate gene for celiac disease pathogenesis. J Cell Biochem, 118(8), 2193–2207. doi:https://doi.org/10.1002/jcb.25864

Batzoglou S, Schwartz R (2014) Computational biology and bioinformatics. Bioinformatics 30(12):i1–i2. https://doi.org/10.1093/bioinformatics/btu304

Blekherman G, Laubenbacher R, Cortes DF, Mendes P, Torti FM, Akman S et al (2011) Bioinformatics tools for cancer metabolomics. Metabolomics 7(3):329–343. https://doi.org/10.1007/s11306-010-0270-3

Bodrossy L, Sessitsch A (2004) Oligonucleotide microarrays in microbial diagnostics. Curr Opin Microbiol 7(3):245–254. https://doi.org/10.1016/j.mib.2004.04.005

Bolger ME, Weisshaar B, Scholz U, Stein N, Usadel B, Mayer KF (2014) Plant genome sequencing - applications for crop improvement. Curr Opin Biotechnol 26:31–37. https://doi.org/10.1016/j.copbio.2013.08.019

Bork P (1997) Bioinformatics and molecular medicine--introduction and call for papers. J Mol Med (Berl) 75(1):3–4

Botstein D, Risch N (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat Genet 33(Suppl):228–237. https://doi.org/10.1038/ng1090

Brown TA (2002) Genomes (Second Edition), Bios Scientific Publishers Ltd, Oxford; ISBN 1-85996-201-7

Brzeski H (2002) An introduction to bioinformatics. Methods Mol Biol 187:193–208. https://doi.org/10.1385/1-59259-273-2:193

Burgess-Herbert SL, Cox A, Tsaih SW, Paigen B (2008) Practical applications of the bioinformatics toolbox for narrowing quantitative trait loci. Genetics 180(4):2227–2235. https://doi.org/10.1534/genetics.108.090175

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. BMC Bioinformatics 10:421. https://doi.org/10.1186/1471-2105-10-421

Can T (2014) Introduction to bioinformatics. Methods Mol Biol 1107:51–71. https://doi.org/10.1007/978-1-62703-748-8_4

Carlson CS, Eberle MA, Rieder MJ, Smith JD, Kruglyak L, Nickerson DA (2003) Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. Nat Genet 33(4):518–521. https://doi.org/10.1038/ng1128

Cascorbi I, Henning S, Brockmoller J, Gephart J, Meisel C, Muller JM et al (2000) Substantially reduced risk of cancer of the aerodigestive tract in subjects with variant--463A of the myeloperoxidase gene. Cancer Res 60(3):644–649

Chandramouli K, Qian PY (2009) Proteomics: challenges, techniques and possibilities to overcome biological sample complexity. Hum Genomics Proteomics 2009:1. https://doi.org/10.4061/2009/239204

Chen XW, Gao JX (2016) Big Data Bioinformatics. *Methods* 111:1–2. https://doi.org/10.1016/j.ymeth.2016.11.017

Chicurel M (2002) Genome analysis at your fingertips. Nature 419:751. https://doi.org/10.1038/419751b

Di Tommaso P, Moretti S, Xenarios I, Orobitg M, Montanyola A, Chang JM et al (2011) T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. Nucleic Acids Res 39(Web Server issue):W13–W17. https://doi.org/10.1093/nar/gkr245

Eddy SR (2009) A new generation of homology search tools based on probabilistic inference. Genome Inform 23(1):205–211

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32(5):1792–1797. https://doi.org/10.1093/nar/gkh340

Edwards D, Batley J (2010) Plant genome sequencing: applications for crop improvement. Plant Biotechnol J 8(1):2–9. https://doi.org/10.1111/j.1467-7652.2009.00459.x

Erichsen HC, Chanock SJ (2004) SNPs in cancer research and treatment. Br J Cancer 90(4):747–751. https://doi.org/10.1038/sj.bjc.6601574

Frye SV, Jin J (2016) Novel therapeutics targeting epigenetics: new molecules, new methods. ACS Med Chem Lett 7(2):123. https://doi.org/10.1021/acsmedchemlett.6b00037

Global Burden of Disease Cancer Collabration, Fitzmaurice C, Akinyemiju TF, Al Lami FH, Alam T, Alizadeh-Navaei R et al (2018) Global, regional, and National Cancer Incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 29 Cancer groups, 1990 to 2016: a systematic analysis for the global burden of disease study. JAMA Oncol. https://doi.org/10.1001/jamaoncol.2018.2706

Goldfeder RL, Parker SC, Ajay SS, Ozel Abaan H, Margulies EH (2011) A bioinformatics approach for determining sample identity from different lanes of high-throughput sequencing data. PLoS One 6(8):e23683. https://doi.org/10.1371/journal.pone.0023683

Greene CS, Tan J, Ung M, Moore JH, Cheng C (2014) Big data bioinformatics. J Cell Physiol 229(12):1896–1900. https://doi.org/10.1002/jcp.24662

Greene CS, Troyanskaya OG (2011) PILGRM: an interactive data-driven discovery platform for expert biologists. Nucleic Acids Res 39(Web Server issue):W368–W374. https://doi.org/10.1093/nar/gkr440

Gutmanas A, Oldfield TJ, Patwardhan A, Sen S, Velankar S, Kleywegt GJ (2013) The role of structural bioinformatics resources in the era of integrative structural biology. Acta Crystallogr D Biol Crystallogr 69.(Pt 5:710–721. https://doi.org/10.1107/S0907444913001157

Hickey G, Blanchette M (2011) A probabilistic model for sequence alignment with context-sensitive indels. J Comput Biol 18(11):1449–1464. https://doi.org/10.1089/cmb.2011.0157

Hinkson IV, Davidsen TM, Klemm JD, Kerlavage AR, Kibbe WA (2017) A comprehensive infrastructure for big data in Cancer research: accelerating Cancer research and precision medicine. Front Cell Dev Biol 5:83. https://doi.org/10.3389/fcell.2017.00083

Holford ME, McCusker JP, Cheung KH, Krauthammer M (2012) A semantic web framework to integrate cancer omics data with biological knowledge. BMC Bioinformatics 13(*Suppl 1*):S10. https://doi.org/10.1186/1471-2105-13-S1-S10

Hou J, Acharya L, Zhu D, Cheng J (2016) An overview of bioinformatics methods for modeling biological pathways in yeast. Brief Funct Genomics 15(2):95–108. https://doi.org/10.1093/bfgp/elv040

Jones J, Otu H, Spentzos D, Kolia S, Inan M, Beecken WD et al (2005) Gene signatures of progression and metastasis in renal cell cancer. Clin Cancer Res 11(16):5730–5739. https://doi.org/10.1158/1078-0432.CCR-04-2225

Jorge NA, Ferreira CG, Passetti F (2012) Bioinformatics of Cancer ncRNA in high throughput sequencing: present state and challenges. Front Genet 3:287. https://doi.org/10.3389/fgene.2012.00287

Katoh M, Katoh M (2006) Bioinformatics for cancer management in the post-genome era. Technol Cancer Res Treat 5(2):169–175. https://doi.org/10.1177/153303460600500208

Kihara C, Tsunoda T, Tanaka T, Yamana H, Furukawa Y, Ono K et al (2001) Prediction of sensitivity of esophageal tumors to adjuvant chemotherapy by cDNA microarray analysis of gene-expression profiles. Cancer Res 61(17):6474–6479

Kihara D, Yang YD, Hawkins T (2007) Bioinformatics resources for cancer research with an emphasis on gene function and structure prediction tools. Cancer Inform 2:25–35

Koltes JE, Hu ZL, Fritz E, Reecy JM (2009) BEAP: the BLAST extension and alignment program-a tool for contig construction and analysis of preliminary genome sequence. BMC Res Notes 2:11. https://doi.org/10.1186/1756-0500-2-11

Konishi H, Ichikawa D, Arita T, Otsuji E (2016) Microarray technology and its applications for detecting plasma microRNA biomarkers in digestive tract cancers. Methods Mol Biol 1368:99–109. https://doi.org/10.1007/978-1-4939-3136-1_8

Laczny C, Leidinger P, Haas J, Ludwig N, Backes C, Gerasch A et al (2012) miRTrail--a comprehensive webserver for analyzing gene and miRNA patterns to enhance the understanding of regulatory mechanisms in diseases. BMC Bioinformatics 13:36. https://doi.org/10.1186/1471-2105-13-36

Li PC (2016) Overview of microarray technology. Methods Mol Biol 1368:3–4. https://doi.org/10.1007/978-1-4939-3136-1_1

Loy A, Bodrossy L (2006) Highly parallel microbial diagnostics using oligonucleotide microarrays. Clin Chim Acta 363(1–2):106–119. https://doi.org/10.1016/j.cccn.2005.05.041

Macgregor PF, Squire JA (2002) Application of microarrays to the analysis of gene expression in cancer. Clin Chem 48(8):1170–1177

Madden TL, Tatusov RL, Zhang J (1996) Applications of network BLAST server. Methods Enzymol 266:131–141

Mount DW, Pandey R (2005) Using bioinformatics and genome analysis for new therapeutic interventions. Mol Cancer Ther 4(10):1636–1643. https://doi.org/10.1158/1535-7163.MCT-05-0150

Mychaleckyj JC (2007) Genome mapping statistics and bioinformatics. Methods Mol Biol 404:461–488. https://doi.org/10.1007/978-1-59745-530-5_22

Need AC, Motulsky AG, Goldstein DB (2005) Priorities and standards in pharmacogenetic research. Nat Genet 37(7):671–681. https://doi.org/10.1038/ng1593

Neumann RS, Kumar S, Haverkamp TH, Shalchian-Tabrizi K (2014) BLASTGrabber: a bioinformatic tool for visualization, analysis and sequence selection of massive BLAST data. BMC Bioinformatics 15:128. https://doi.org/10.1186/1471-2105-15-128

Non AL, Thayer ZM (2015) Epigenetics for anthropologists: an introduction to methods. Am J Hum Biol 27(3):295–303. https://doi.org/10.1002/ajhb.22679

Pertsemlidis A, Fondon JW 3rd (2001) Having a BLAST with bioinformatics (and avoiding BLASTphemy). Genome Biol *2*(10):REVIEWS2002

Puhler A (2017) Bioinformatics solutions for big data analysis in life sciences presented by the German network for bioinformatics infrastructure. J Biotechnol 261:1. https://doi.org/10.1016/j.jbiotec.2017.08.025

Samish I, Bourne PE, Najmanovich RJ (2015) Achievements and challenges in structural bioinformatics and computational biophysics. Bioinformatics 31(1):146–150. https://doi.org/10.1093/bioinformatics/btu769

Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K et al (2012) Database resources of the National Center for biotechnology information. Nucleic Acids Res 40(Database issue):D13–D25. https://doi.org/10.1093/nar/gkr1184

Schadt EE (2006) Novel integrative genomics strategies to identify genes for complex traits. Anim Genet 37(Suppl 1):18–23. https://doi.org/10.1111/j.1365-2052.2006.01473.x

Shaik NA, Awan ZA, Verma PK, Elango R, Banaganapalli B (2018) Protein phenotype diagnosis of autosomal dominant calmodulin mutations causing irregular heart rhythms. J Cell Biochem. https://doi.org/10.1002/jcb.26834

Shaik NA, Kaleemuddin M, Banaganapalli B, Khan F, Shaik NS, Ajabnoor G et al (2014) Structural and functional characterization of pathogenic non- synonymous genetic mutations of human insulin-degrading enzyme by in silico methods. CNS Neurol Disord Drug Targets 13(3):517–532

Soejima H (2009) Epigenetics-related diseases and analytic methods. Rinsho Byori 57(8):769–778

Subramanian S, West RB, Corless CL, Ou W, Rubin BP, Chu KM et al (2004) Gastrointestinal stromal tumors (GISTs) with KIT and PDGFRA mutations have distinct gene expression profiles. Oncogene 23(47):7780–7790. https://doi.org/10.1038/sj.onc.1208056

Taylor JB, Triggle DJ (2007) Comprehensive medicinal chemistry II. Amsterdam; London: Elsevier

Thompson JD, Gibson TJ, Higgins DG (2002) Multiple sequence alignment using ClustalW and ClustalX. Curr Protoc Bioinformatics., Chapter 2, Unit 2.3. https://doi.org/10.1002/0471250953.bi0203s00

Varon A, Wheeler WC (2012) The tree alignment problem. BMC Bioinformatics 13:293. https://doi.org/10.1186/1471-2105-13-293

Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC (2016) SIFT missense predictions for genomes. Nat Protoc 11(1):1–9. https://doi.org/10.1038/nprot.2015.123

Waage J, Standl M, Curtin JA, Jessen LE, Thorsen J, Tian C et al (2018) Genome-wide association and HLA fine-mapping studies identify risk loci and genetic pathways underlying allergic rhinitis. Nat Genet 50(8):1072–1080. https://doi.org/10.1038/s41588-018-0157-1

Wang F, Kong J, Cooper L, Pan T, Kurc T, Chen W et al (2011) A data model and database for high-resolution pathology analytical image informatics. J Pathol Inform 2:32. https://doi.org/10.4103/2153-3539.83192

Wang Y, Zhang Y, Huang Q, Li C (2018) Integrated bioinformatics analysis reveals key candidate genes and pathways in breast cancer. Mol Med Rep 17(6):8091–8100. https://doi.org/10.3892/mmr.2018.8895

Webb B, Sali A (2017) Protein structure modeling with MODELLER. Methods Mol Biol 1654:39–54. https://doi.org/10.1007/978-1-4939-7231-9_4

Wilkinson GR (2005) Drug metabolism and variability among patients in drug response. N Engl J Med 352(21):2211–2221. https://doi.org/10.1056/NEJMra032424

Yalcin D, Hakguder ZM, Otu HH (2016) Bioinformatics approaches to single-cell analysis in developmental biology. Mol Hum Reprod 22(3):182–192. https://doi.org/10.1093/molehr/gav050

Yang MQ, Athey BD, Arabnia HR, Sung AH, Liu Q, Yang JY et al (2009) High-throughput next-generation sequencing technologies foster new cutting-edge computing techniques in bioinformatics. BMC Genomics 10(Suppl 1):I1. https://doi.org/10.1186/1471-2164-10-S1-I1

Zharikova AA, Mironov AA (2016) piRNAs: biology and bioinformatics. Mol Biol (Mosk) 50(1):80–88. https://doi.org/10.7868/S0026898416010225

Zienolddiny S, Skaug V (2012) Single nucleotide polymorphisms as susceptibility, prognostic, and therapeutic markers of nonsmall cell lung cancer. Lung Cancer (Auckl) 3:1–14. https://doi.org/10.2147/LCTT.S13256

Chapter 2
Introduction to Biological Databases

Noor Ahmad Shaik, Ramu Elango, Muhummadh Khan, and Babajan Banaganapalli

Contents

2.1 Introduction to Databases

Database is a computerized resource where data is structured in a way that makes it easy to add, access, and update it. The main purpose of databases is to enable easy handling and retrieval of information through multiple search features

N. A. Shaik
Department of Genetic Medicine, Faculty of Medicine,
King Abdulaziz University, Jeddah, Saudi Arabia
e-mail: nshaik@kau.edu.sa

R. Elango · B. Banaganapalli (✉)
Princess Al-Jawhara Center of Excellence in Research of Hereditary Disorders, Department of Genetic Medicine, Faculty of Medicine, King Abdulaziz University, Jeddah, Saudi Arabia
e-mail: relango@kau.edu.sa; bbabajan@kau.edu.sa

M. Khan
Genomics and Biotechnology Section, Biology Department, Faculty of Science,
King Abdulaziz University, Jeddah, Saudi Arabia
e-mail: mkmkhan@kau.edu.sa

N. A. Shaik et al. (eds.), *Essentials of Bioinformatics, Volume I*,
https://doi.org/10.1007/978-3-030-02634-9_2

(Garcia-Molina et al. 2002). Database is organized in a way that each data entry represents a record. A record contains multiple items of data; therefore it consists of a number of fields. For example, names, phone numbers, addresses, etc. are all attached to one particular record (Garcia-Molina et al. 2002). To search the database for a specific record, the user can use information given in a single field to retreive the whole record (Jagarlapudi and Kishan 2009). In data science, this is called as making a query.

Biological databases have tools for higher level information processing. Their objective is not only to store information but also to discover. They often have the ability to detect connections among new entries and previously archived data to avoid overlapping of the data (Birney and Clamp 2004). They also can perform computational operations on the stored data, for example, detect sequence homology or certain motifs. This facilitates a comprehensive holistic approach to biological data.

2.2 Types of Databases

When databases originated, they were in the form of plain text file with multiple entries separated by vertical or horizontal delimiter, e.g., a vertical bar or any other suitable character, much like a large table where raw data is stored. It didn’t support further tasks like finding similarities or repetitions or adopting keywords. For a computer to search for a certain piece of information, it had to read through all the regions of the text (Özsu and Valduriez 1991). This is a time-consuming process that requires a heavy-duty processor and intense-performance memory. For small databases, it shouldn’t be a problem, but biological ones are often very large with a huge amount of data in different fields. Computers are often crashed during the search processes if they exceed the capacity of their operating systems. Thus, to facilitate accessing and searching through the data, many sophisticated software programs have been designed and installed. They are programmed to find connections among the raw information entered into the database. These software programs are called database management systems with the function of storing, monitoring, and sorting all types of information. They structure the raw data into sets according to different combinations and connections they can find, making the search processes more organized and effective. Hence, they can make reports of the search and conclude more information from the raw data entries (Gibas 2001). Databases management systems are classified into two large groups: relational databases and object-oriented databases. These have their own management systems.

2.2.1 Relational Databases

Relational databases categorize the data into columns/entities with different values/ instances (Codd 1998). These entities have something in common called as attribute. The attribute represents information about the values of the entity

Table 2.1 Relational databases data

Student no.	Student name	School no.	School name
1.	Jaan	1.	X
2.	Rayyan	2.	Y
3.	Sarah	3.	Z

Table 2.2 Relational databases data

School no.	School name
1.	X
2.	Y
3.	Z

Table 2.3 Relational databases data

Student no.	Student name	School no.
1.	Jaan	1.
2.	Rayyan	2.
3.	Sarah	3.

(Ramakrishnan and Gehrke 2003). Let's take an example to make things easier. A survey is conducted to list all the students in a given state and the schools they study in. The table representing the state will have a column for the schools' ordinal number and another for its name, likewise a column for the students' number and another for their names (Table 2.1). So, if you look for a certain student, the computer has to read through the whole table. In relational database, the following will happen: the big table will be divided into two. One is for the school name and number (Table 2.2). The other will have the student name, number, and the number of the school he visits (Table 2.3). If he or she is a student at school number 3 the computer will automatically look up the second table and get the name of the corresponding school which is in this case Z.

Relational databases segregate the flat file which nothing but a plain text file into smaller ones according to their relation facilitating a quick and simple query. Retrieving information from these databases is more efficient, and new data can be added with no modifications in the old ones. However, generation of relational databases is a complicated process that takes a lot of planning and designing. The specialized programming language used for their creation is called Structured Query Language (SQL). Once they are generated, searching through them will be an easy task. New data categories can be added to the database without having to change the whole system. Also, different tables can be connected to each other logically by relations. The connection could be one-to-one, meaning each table connected separately to the previous or the next in succession, or many-to-many where multiple tables are connected together. The more the specialization, better the performance of the database. Creation of relational databases is tiresome. Once finished, searching and getting a final report of whatever search objects is easily possible.

2.2.2 *Object-Oriented Database*

Sometimes one of the features making relational databases inadequate is that they are isolated entities of abundant information. It may be difficult to integrate sophisticated hierarchical pieces of information into the database or connect them together (Garvey and Jackson 1989). Therefore, another type of databases emerged to overcome this problem. Object-oriented databases set an "object" as the unit of combination or description of the data. The notion "object" also implies the behavior of this data set and computational operations done on it. In other words, object is a single word to describe the concept behind a set of data and the processes these data undergo. Objects of a database are linked together by a group of pointers that indicate the relationships among these objects. This hierarchical system allows for easy accessing of information without the need of some index to understand the links. If we consider the previous example, there would be three objects: student objects, school object, and state object. Their interrelations are pointers which are represented by arrows in the Fig. 2.1. A pointer may refer to the state that the student lives in and to the school he or she attends (Fig. 2.1).

Object-oriented databases usually use programming languages like C++ and Java since these languages show more flexibility and inheritance. Objects can be searched and accessed easily. New objects can be added dynamically at any time without the demand to change previously added tables. Related objects are organized into classes which facilitates homology identification. Also, each object has a predictable behavior, so new data can be extracted from the database just

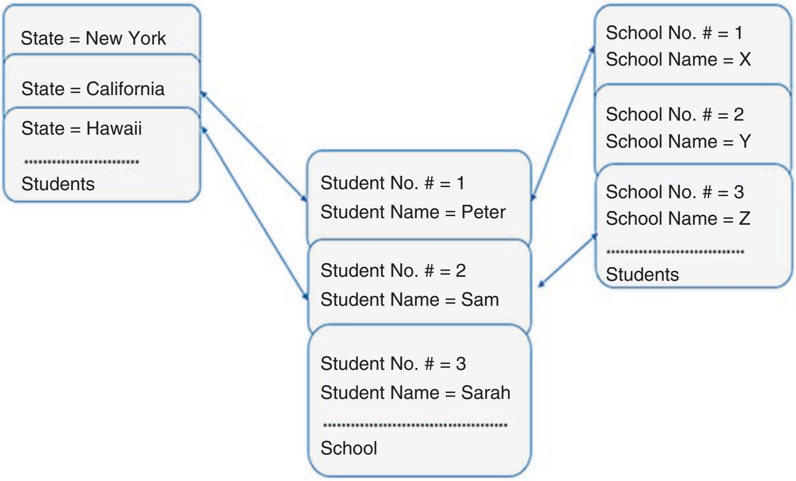


Fig. 2.1 Example of hierarchical system of object-oriented database

by observing the mathematical behavior of the entries. The new inputs have the capability to inherit data attributes from previously added objects. Object-oriented databases are better at modeling to real world. But on the other hand, they may lack the precise mathematical structure of relational databases. Pointers and relations between the objects could be misinterpreted or wrongly displayed. For that reason, object-relational databases now have integrated characteristics of both types databases.

2.3 Introduction to Biological Databases

Biological databases make use of the three aforementioned database types: plain flat text, object-oriented, and relational databases. Combination of their features brings out the best results (Bourne 2005). Despite many limitations and restrictions the flat text is used only in biological databases and in small databases (Stein 2002), they resort to the plain text table format keeping in mind that the researchers understand the operation and output in all cases (Baxevanis 2009, 2011).

2.3.1 Classification of Biological Databases

According to the information added to the database, they are classified into three main categories: primary databases, secondary databases, and composite or specialized databases. Primary databases serve as computational archives containing only raw data, e.g., nucleic acids and protein sequences. Examples include GenBank and Protein Data Bank (Mullan 2003).

Secondary biological databases use the original data in primary databases to derive new data sets by using specialized software programs or by manual annotation of information. Secondary data include translated protein sequences and active site residues. Examples of secondary databases are InterPro, a database for protein families, PIR (Protein Information Resources), Swiss-Prot for protein structure, and Ensembl that specializes in studying variation, classification, and function of the genome sequences.

A composite database may combine more than one primary database so that instead of searching each one separately, the user can search related databases together for quick results. The NCBI is the best example for this type. Specialized databases are dedicated to a specific research field such as Ribosomal Database Project which stores only rRNA gene sequence. Table 2.4 contains a list of frequently used databases (Stein 2013).

Table 2.4 Most popular biological databases

Database	Short description	URL
GenBank	National Center for biotechnology information primary database	https://www.ncbi.nlm.nih.gov/
EMBL	Molecular biology primary database	http://www.embl.org/
DNA data Bank of Japan	DNA sequence primary databank	http://www.ddbj.nig.ac.jp/
OMIM	Online Mendelian inheritance in man – Human genome secondary database	http://www.omim.org/
FlyBase	Specialized organism database	http://www.flybase.org/
Swiss-Prot	Protein sequence database	http://www.flybase.org/
InterPro	Protein-based secondary database	http://www.ebi.ac.uk/interpro/
Reactome	A specialized database for human reactions and metabolic pathways	http://www.ebi.ac.uk/interpro/

2.3.2 *Primary Database*

Three of the most outstanding nucleic acid sequence databases created by scientists and researchers around the world are open access i.e. they are freely available on the Internet (Bishop 1999). Anyone can access these three databases which are the European Molecular Biology Laboratory (EMBL) (1990), DNA Data Bank of Japan (DDBJ) (Mashima et al. 2017), and the GenBank database (Benson et al. 2018). These databases contain data which has been freely deposited by the researchers from around the world but this data lacks proper annotation. Most international journals nowadays require the authors to annotate and submit their protein and/or DNA sequence to any of the three databases before accepting their article for publication. Publishers in this way can validate all sequences or structures submitted against those already available in the database for redundancy or other purposes. These three databases collaborate with each other exchanging new data nearly every day. So, a researcher only needs to access one of them to avail the latest information. The three databases together make the International Nucleotide Sequence Database Collaboration. Despite having the same data, they have different designs and display formats. The protein data bank (PDB) (Rose et al. 2013) stores the three-dimensional structure of protein and nucleic acids. Although it uses a flat file format, for the sake of convenience, its interface supports simple manipulation of the 3D structures.

2.3.3 *Secondary Databases*

Primary databases often lack annotation and multiple other features. Thus, information processing tools are of much importance as they turn the plain raw data into more sophisticated knowledge. Secondary databases have this capability alongside their ability to analyze information and draw conclusions. Secondary databases vary in how advanced they process the data extracted from the primary ones. Some secondary databases have only a limited degree of processing and translate the sequence data from stored DNA. Others have higher levels of analyzing and handling information. They provide info about annotation, family rank, function, and structure. For example, UniProt/TrEMBL database stores information regarding various aspects of translated nucleic acid sequences and functional macromolecules. Data are extracted from published literature and added to the database for curation and annotation. This is done manually ensuring a high level of accuracy and up-to-date information. Any protein in the database is annotated with its function, structure, catalytic sites, domain, metabolic pathway, disease association, and any other relative information (Zou et al. 2015; D'Eustachio 2013; Toby 2012; Banaganapalli et al. 2016).

2.3.4 Specialized Databases

Specialized databases are large-scale collaborations between members of the scientific community devoted to a specific disease or a certain organism. The abundance of data and lack of organization were the main reasons for their appearance. Information in these specialized databases is usually curated by authors or experts in the field (Toby 2012), which indicates that they may have advanced computational capabilities than secondary databases. This databases may also allow interactive platforms where the users can add notes or comments. An example of this is the Parkinson disease map project by department of life sciences and biological systems at Luxembourg University. Taxonomic-specific databases are also specialized ones. Examples include FlyBase, TAIR, and AceDB (Zhang et al. 2016).

2.3.5 Interconnection between Biological Databases

The user usually needs to search in various databases to get what he or she needs. So, it is better to incorporate them together through composite ones. The difficulty lies in the design and format differences; flat file, relational, or object-oriented databases. A specification language called Common Object Request Broker Architecture (CORBA) has emerged to enable database programs to collaborate regardless of their database structure. Another system called eXtensible Markup Language (XML) also specializes in combining databases. Any data entry or record is divided into small components labelled by specific tags. Bioinformatics researchers have also developed a new method for data exchange called the distributed annotation system. This protocol allows one computer to connect to multiple servers at a time and extract data concerning different sequence annotations. It then integrates it into a final search report (Baxevanis 2011).

2.4 Retrieval from Databases

The most important function of any database is to provide an easy access to the stored information. Therefore, there are systems specially designed for the retrieval process. Entrez and Sequence Retrieval Systems (SRS) are examples of these systems. They provide advanced search options through an user-friendly interface and Boolean operators to search through multiple databases. Boolean operators help define the relation between the variables in the search box to give a refined output, e.g., and/or/not and so on. It also supports the keyword search. Obviously "and" gives the order to bring results containing all mentioned variables, "or" indicates the case when either one of the variable is relevant, whereas "not" means to exclude search results with this piece of information. Using these advanced options facilitates the search process to a great degree and makes it very specific (Sreenivasaiah and Kim 2010; Shaik et al. 2018).

2.5 Conclusion

Biological databases are digital libraries where data such as DNA and protein sequences are stored. With the progress of Life sciences, it became important to store biological data in an organized way to facilitate the search process. Biological databases make it easy to access, analyze, and do sophisticated calculations on the data. This helps researchers to draw new conclusions by linking different knowledge aspects together. Biological databases are classified, according to the program management system, into flat file, relational, and object-oriented databases. Raw data is stored in the primary databases secondary databases store information generated by further processing of the raw data, and specialized databases offer information about a specific aspect of interest.

References

Banaganapalli B, Mohammed K, Khan IA, Al-Aama JY, Elango R, Shaik NA (2016) A computational protein phenotype prediction approach to analyze the deleterious mutations of human MED12 gene. J Cell Biochem 117(9):2023–2035. https://doi.org/10.1002/jcb.25499

Baxevanis AD (2009) The importance of biological databases in biological discovery. Curr Protoc Bioinformatics 27(1):1.1.1–1.1.6. https://doi.org/10.1002/0471250953.bi0101s27

Baxevanis AD (2011) The importance of biological databases in biological discovery. Curr Protoc Bioinformatics Chapter 1, Unit 1 1. https://doi.org/10.1002/0471250953.bi0101s34

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Ostell J, Pruitt KD et al (2018) GenBank. Nucleic Acids Res 46(D1):D41–D47. https://doi.org/10.1093/nar/gkx1094

Birney E, Clamp M (2004) Biological database design and implementation. Brief Bioinform 5(1):31–38

Bishop, M.J. (1999). Genetics databases

Bourne P (2005) Will a biological database be different from a biological journal? PLoS Comput Biol 1(3):e34. https://doi.org/10.1371/journal.pcbi.0010034

Codd EF (1998) A relational model of data for large shared data banks. 1970. MD Comput 15(3):162–166

D'Eustachio P (2013) Pathway databases: making chemical and biological sense of the genomic data flood. Chem Biol 20(5):629–635. https://doi.org/10.1016/j.chembiol.2013.03.018

European Molecular Biology Laboratory (1990) New nucleotide sequence data on the EMBL File Server. Nucleic Acids Res 18(17):5329–5341

Garcia-Molina H, Ullman JD, Widom J (2002) Database systems: the complete book. Prentice Hall, Harlow

Garvey MA, Jackson MS (1989) Introduction to object-oriented databases. Inf Softw Technol 31(10):521–528. https://doi.org/10.1016/0950-5849(89)90173-0

Gibas C et al. (2001) Developing bioinformatics computer skills. Shroff Publishers & Distributors, Mumbai, India. http://shop.oreilly.com/product/9781565926646.do

Jagarlapudi SA, Kishan KV (2009) Database systems for knowledge-based discovery. Methods Mol Biol 575:159–172. https://doi.org/10.1007/978-1-60761-274-2_6

Mashima J, Kodama Y, Fujisawa T, Katayama T, Okuda Y, Kaminuma E et al (2017) DNA Data Bank of Japan. Nucleic Acids Res 45(D1):D25–D31. https://doi.org/10.1093/nar/gkw1001

Mullan LJ (2003) Biological sequence databases. Brief Bioinform 4(1):75–77

Özsu MT, Valduriez P (1991) Principles of distributed database systems. Prentice Hall, Englewood Cliffs

Ramakrishnan R, Gehrke J (2003) Database management systems. McGraw-Hill Education, Boston

Rose PW, Bi C, Bluhm WF, Christie CH, Dimitropoulos D, Dutta S et al (2013) The RCSB protein data Bank: new resources for research and education. Nucleic Acids Res 41.(Database issue:D475–D482. https://doi.org/10.1093/nar/gks1200

Shaik NA, Awan ZA, Verma PK, Elango R, Banaganapalli B (2018) Protein phenotype diagnosis of autosomal dominant calmodulin mutations causing irregular heart rhythms. J Cell Biochem. https://doi.org/10.1002/jcb.26834

Sreenivasaiah PK, Kim DH (2010) Current trends and new challenges of databases and web applications for systems driven biological research. Front Physiol 1:147. https://doi.org/10.3389/fphys.2010.00147

Stein L (2002) Creating databases for biological information: an introduction. Curr Protoc Bioinformatics. Chapter 9, Unit 9 1. https://doi.org/10.1002/0471250953.bi0901s00

Stein L (2013) Creating databases for biological information: an introduction. Curr Protoc Bioinformatics Chapter 9, Unit9 1. https://doi.org/10.1002/0471250953.bi0901s42

Toby I (2012) Biological databases as research tools in the post-genomic era. Aviat Space Environ Med 83(4):452–453

Zhang Q, Ding N, Zhang L, Zhao X, Yang Y, Qu H et al (2016) Biological databases for hematology research. Genomics Proteomics Bioinformatics 14(6):333–337. https://doi.org/10.1016/j.gpb.2016.10.004

Zou D, Ma L, Yu J, Zhang Z (2015) Biological databases for human research. Genomics Proteomics Bioinformatics 13(1):55–63. https://doi.org/10.1016/j.gpb.2015.01.006

Chapter 3
Sequence Databases

Vivek Kumar Chaturvedi, Divya Mishra, Aprajita Tiwari, V. P. Snijesh, Noor Ahmad Shaik, and M. P. Singh

Contents

V. K. Chaturvedi · A. Tiwari · M. P. Singh (✉)
Centre of Biotechnology, University of Allahabad, Allahabad, India

D. Mishra
Centre of Bioinformatics, University of Allahabad, Allahabad, India

V. P. Snijesh
Innov4Sight Health and Biomedical System Private Limited, Bangalore, India

N. A. Shaik
Department of Genetic Medicine, Faculty of Medicine, King Abdulaziz University,
Jeddah, Saudi Arabia
e-mail: nshaik@kau.edu.sa

N. A. Shaik et al. (eds.), *Essentials of Bioinformatics, Volume I*,
https://doi.org/10.1007/978-3-030-02634-9_3

3.1 Introduction

Bioinformatics involves the use of information technology to collect, store, retrieve, and analyze the enormous amount of biological data that are available in the form of sequences and structures of proteins, nucleic acids, and other biomolecules (Toomula et al. 2011). Biological databases are mainly classified into sequence and structure databases. The first database was created soon after the sequencing of insulin protein in 1956. Insulin was the first protein to be sequenced; it contains 51 amino acid residues (Altschul et al. 1990). Over the past few decades, there has been a high demand for powerful computational methods which can improve the analysis of exponentially increasing biological information, finally giving rise to a new era of "bioinformatics." Development in the field of molecular biology and high-throughput sequencing approaches has resulted in the dramatic increase in genomic and proteomic data such as sequences and their corresponding molecular structures. Submission of such facts into public information has led to the development of several biological databases which can be accessed for querying and retrieving of stored information through the research community. During the mid-1960s, the first nucleic acid sequence of yeast tRNA was found out. Around this period 3D structure of the protein was explored, and the well-known Protein Data Bank (PDB) was developed as the primary protein structure database with approximately ten initial entries (Ragunath et al. 2009).

The ultimate goal of designing a database is to collect the data in the suitable form which may be easily accessed through researches (Toomula et al. 2011). In this chapter, we are cataloguing the various biological databases and also provide a short review of the classification of databases according to their data types.

3.2 Sequence Data Generation

The sequencing technique has played an essential role in analyzing the biological data of organisms. The initial pioneering work in the field of sequencing was done by Frederick Sanger, as well as by Maxam and Gilbert. Their initial sequencing methods have greatly helped in the development and validation of current-day advanced sequencing technologies (Heather and chain 2016). Over the years, the technological advances made on sequencing, molecular biology and automation increased the technological efficiency of sequencing and allowed the analysis of multiple DNA sample sequencing at a single run. As a result, researchers moved from expensive and time-consuming in vitro and in vivo *methods* to quick and reliable in silico analysis as a first-line option in biomedical research (Ansorge 2009). Recent decades have witnessed the continuous decoding, publishing, and hosting of genomes from multiple organisms by sequence repositories.

3.2.1 The First Generation of Sequencing

The first generation of sequencing is dominant over the other three decades in which the Sanger technique was widely used. The Sanger technique, as well as Maxam and Gilbert technique, was categorized as the first generation of sequencing. The Sanger technique relies on DNA synthesis in vitro, coupled with chain termination. The first genome elucidated through Sanger technique was bacteriophage φX174 (genomic size is 5374 base pairs). However, Sanger technique had some drawbacks like difficulty in handling complex genomic species, and it is still an expensive and time-consuming approach. Maxam and Gilbert sequencing method, which is based on chemical degradation steps, is another widely used first-generation sequencing method. However, this method is considered to be a risky sequencing approach as it uses the toxic chemical to sequence the data.

3.2.2 The Second Generation of Sequencing

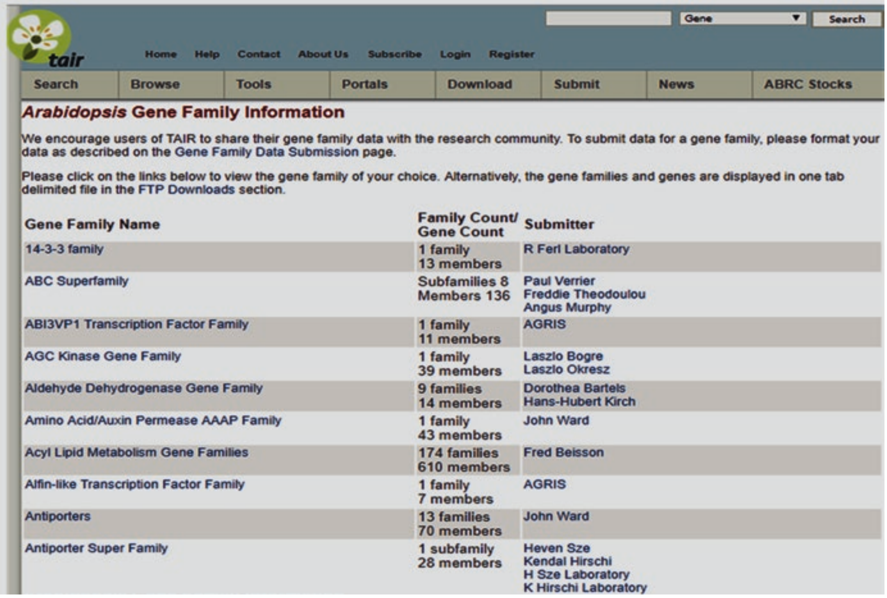
The second generation sequencing has some specific advantages like the ability to generate millions of parallel short reads at a time. This technique is less expensive as well as less time-consuming than the first generation of sequencing, and sequence output is generated without the involvement of the electrophoresis method. Two approaches are widely used in short read sequencing. The first approach is based on sequencing via ligation method, and another approach is sequencing via synthesis (Ansorge 2009).

3.2.3 The Third Generation of Sequencing

The second generation of sequencing methods are not efficient to solve the very complex repetitive area of the genome. Also, aligning the samples to the reference genome based on short reads makes the task more complex and difficult one. To solve these issues, researchers have developed the new generation of sequencing called the third generation of sequencing. This technique is less costly as well as less time-consuming compared to the second generation of sequencing. This approach can generate long reads of sequences at a time.

3.3 Classes of Biological Databases

Biological databases were broadly classified into three major categories as sequence, structure, and functional databases. The sequences of proteins and the nucleic acids are stored in sequence databases, and their solved structure of transcripts, as well as proteins, are stored in structural databases. Based on their contents, biological databases can be divided into three classes, i.e., primary, secondary, and specialized databases. Primary databases incorporate authentic biological records (Hughes 2001). They give information of raw sequence or structural information submitted using the scientific community. Common examples of primary databases involve GenBank and Protein Data Bank (PDB) (Berman et al. 2000). Secondary databases comprise computationally processed or manually curated information, derived from original data from primary databases. Translated protein sequence databases containing functional annotation are an example of a secondary database (John et al. 2011). Likewise, specialized databases contain special feature-based information that includes model organisms' databases, pathways, as well as disease-related information of human being (Neelameghan 1997) (Fig. 3.1).



Arabidopsis Gene Family Information

We encourage users of TAIR to share their gene family data with the research community. To submit data for a gene family, please format your data as described on the Gene Family Data Submission page.

Please click on the links below to view the gene family of your choice. Alternatively, the gene families and genes are displayed in one tab delimited file in the FTP Downloads section.

Gene Family Name	Family Count/ Gene Count	Submitter
14-3-3 family	1 family 13 members	R Ferl Laboratory
ABC Superfamily	Subfamilies 8 Members 136	Paul Verrier Freddie Theodoulou Angus Murphy
ABI3VP1 Transcription Factor Family	1 family 11 members	AGRIS
AGC Kinase Gene Family	1 family 39 members	Laszlo Bogre Laszlo Okresz
Aldehyde Dehydrogenase Gene Family	9 families 14 members	Dorothea Bartels Hans-Hubert Kirch
Amino Acid/Auxin Permease AAAP Family	1 family 43 members	John Ward
Acyl Lipid Metabolism Gene Families	174 families 610 members	Fred Beisson
Alfin-like Transcription Factor Family	1 family 7 members	AGRIS
Antiporters	13 families 70 members	John Ward
Antiporter Super Family	1 subfamily 28 members	Heven Sze Kendal Hirschi H Sze Laboratory K Hirschi Laboratory

Fig. 3.1 The Arabidopsis Information Resource (TAIR), a specialized database for *Arabidopsis thaliana* species

3.4 Types of Sequence Databases

All published genome sequence is available over the public repositories, as it is a requirement of every scientific journal that any published DNA or RNA or protein records need to be deposited in a public database. The three huge databases which stores and dispense the sequence information are: the NCBI databases (www.ncbi.nlm.nih.gov), the European Molecular Biology Laboratory (EMBL) database (www.ebi.ac.united kingdom/ensemble), and the DNA Data Bank of Japan (DDBJ) databases. These databases gather all publically available DNA, RNA, and protein information and make it freely available (Whitfield et al. 2006). They exchange their data regularly, so, basically, these databases contain the same type of information. An accession number recognizes collections in NCBI sequence databases (or EMBL/DDBJ). This is a unique number that is simply associated with one sequence. As well as the sequence itself, for every collection, the NCBI databases (or EMBL/DDBJ databases) also store a few extra annotation data, consisting of the name of the species it comes from, references to the publication describing that sequence, and so on (Pearson 1994). There are several kinds of sequence databases as described below.

3.4.1 Nucleotide Sequence Databases

3.4.1.1 EMBL/DDBJ/GenBank

The EMBL Nucleotide Sequence Database (also known as EMBL-Bank) is the primary nucleotide sequence resource maintained by the European Bioinformatics Institute (EBI), situated in the United Kingdom. The DNA and RNA sequences are submitted directly from individual researchers, genome sequencing projects, and patent applications. The EMBL Nucleotide Sequence Database (www.ebi.ac.uk/embl/) is the European member of the tri-partied International Nucleotide Sequence Database Collaboration DDBJ/EMBL/GenBank. The EBI provides bioinformatics tools for database searching, sequence and homology searching, multiple sequence alignments, etc. The EBI provides a comprehensive set of sequence similarity algorithms that is easily accessible by the EMBL-EBI web site represented in Fig. 3.3 (www.ebi.ac.uk/Tools/).

The main data sources are large-scale genome sequencing centers, individual scientists, and the European Patent Office (EPO). DNA Data Bank of Japan (DDBJ) initiated the DNA data bank activities in 1986 at the National Institute of Genetics (NIG). DDBJ has been working as one of the prominent international DNA databases along with EBI (European Bioinformatics Institute) in Europe and the National Center for Biotechnology Information (NCBI) in the USA. Direct submissions to

EMBL-Bank are set off by daily data exchange with collaborating databases DDBJ (Japan) (Tateno et al. 2000) and GenBank (USA) (Benson et al. 2000). The database is created in a global joint effort with GenBank (USA) and the DDBJ. The DDBJ collects DNA sequence data mainly from Japanese and researchers from all over the world. Many other tools have been developed at DDBJ from data retrievals and their analysis. A Web-based tool of DDBJ is SAKURA used for nucleotide sequence submission, annotation, and information about the submitter. Each of the three groups gathers a bit of the aggregate succession information announced around the world, and all new and refreshed database sections are traded between the groups once a day. The EMBL Nucleotide Succession Database is a piece of the Protein and Nucleotide Database (PANDA) group.

GenBank is the most comprehensive and annotated collection of publicly available DNA sequence and a part of the International Nucleotide Sequence Database Collaboration (INSDC), which consists of DDBJ, EMBL, and GenBank at NCBI (Fig. 3.2) (Dennis et al. 2000). The NCBI was established in 1988 as a subsidiary of the National Library of Medicine (NLM) at the National Institutes of Health (NIH), USA. The EMBL database along with GenBank and DDBJ plays the pivotal role in the acquisition, storage, and distribution of human genome sequence data. The coding sequence (CDS) features in EMBL entries mark the translation of protein-coding regions which are automatically added to the TrEMBL protein database (Farrell et al. 2014). In consequence, Swiss-Prot curators subsequently generate the Swiss-Prot database using these entries. DDBJ is the only authorized DNA data bank in Japan for the collection of DNA sequences by the researchers worldwide and to issue the internationally accepted accession number in databases. In addition, DDBJ has developed and provided data retrieval and analysis tool (Fig. 3.3).

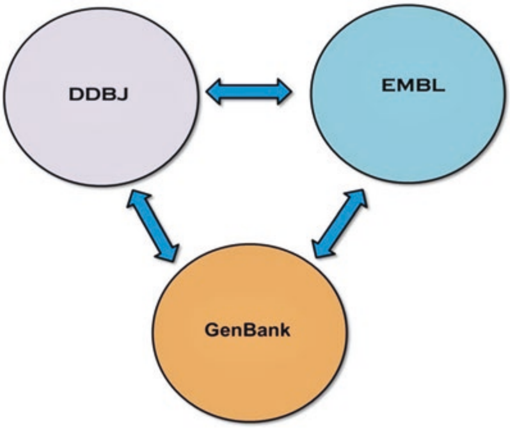
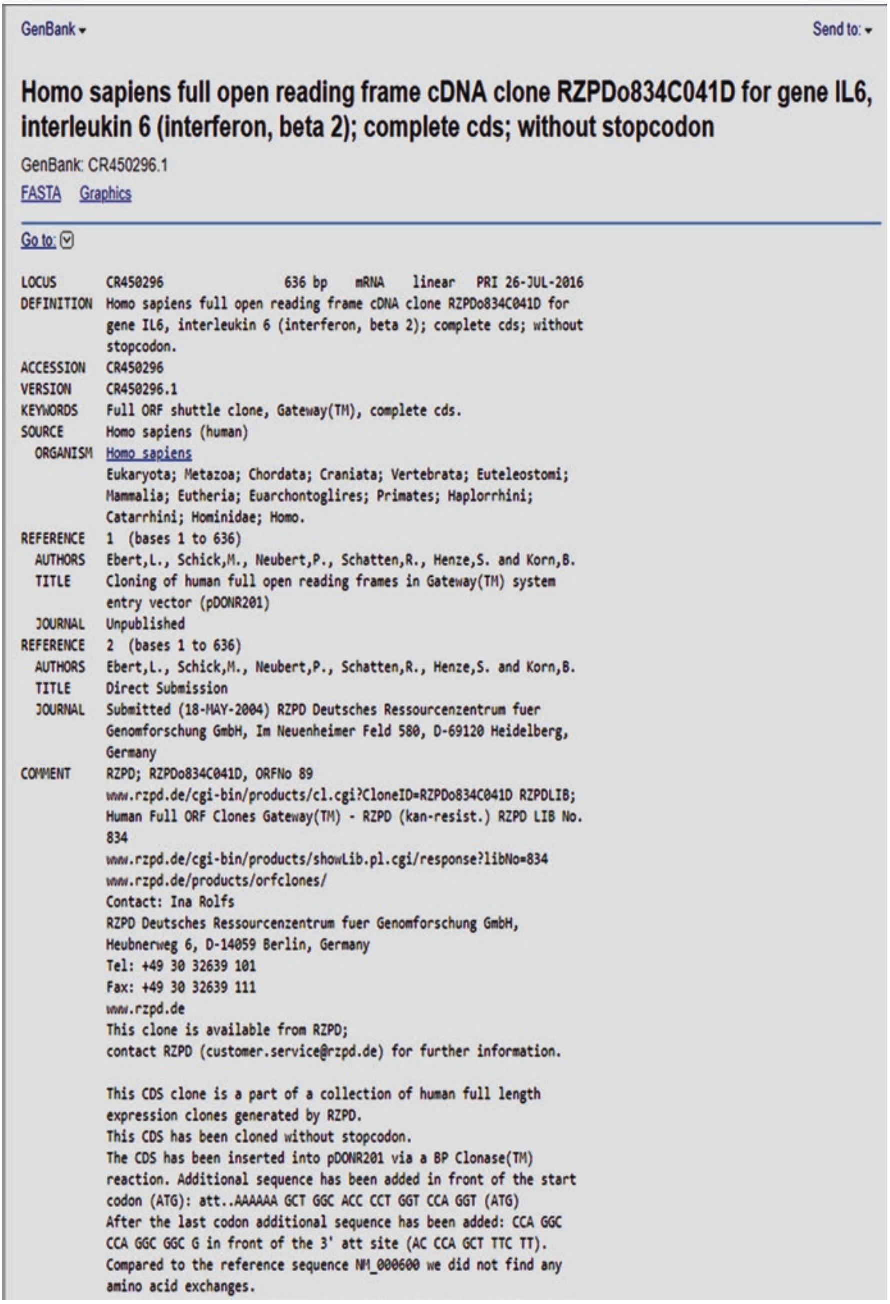


Fig. 3.2 Information stored at GenBank, EMBL, and DDBJ shared with each other



Homo sapiens full open reading frame cDNA clone RZPDo834C041D for gene IL6, interleukin 6 (interferon, beta 2); complete cds; without stopcodon

GenBank: CR450296.1

```
LOCUS       CR450296                 636 bp    mRNA    linear   PRI 26-JUL-2016
DEFINITION  Homo sapiens full open reading frame cDNA clone RZPDo834C041D for
            gene IL6, interleukin 6 (interferon, beta 2); complete cds; without
            stopcodon.
ACCESSION   CR450296
VERSION     CR450296.1
KEYWORDS    Full ORF shuttle clone, Gateway(TM), complete cds.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 636)
  AUTHORS   Ebert,L., Schick,M., Neubert,P., Schatten,R., Henze,S. and Korn,B.
  TITLE     Cloning of human full open reading frames in Gateway(TM) system
            entry vector (pDONR201)
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 636)
  AUTHORS   Ebert,L., Schick,M., Neubert,P., Schatten,R., Henze,S. and Korn,B.
  TITLE     Direct Submission
  JOURNAL   Submitted (18-MAY-2004) RZPD Deutsches Ressourcenzentrum fuer
            Genomforschung GmbH, Im Neuenheimer Feld 580, D-69120 Heidelberg,
            Germany
COMMENT     RZPD; RZPDo834C041D, ORFNo 89
            www.rzpd.de/cgi-bin/products/cl.cgi?CloneID=RZPDo834C041D RZPDLIB;
            Human Full ORF Clones Gateway(TM) - RZPD (kan-resist.) RZPD LIB No.
            834
            www.rzpd.de/cgi-bin/products/showLib.pl.cgi/response?libNo=834
            www.rzpd.de/products/orfclones/
            Contact: Ina Rolfs
            RZPD Deutsches Ressourcenzentrum fuer Genomforschung GmbH,
            Heubnerweg 6, D-14059 Berlin, Germany
            Tel: +49 30 32639 101
            Fax: +49 30 32639 111
            www.rzpd.de
            This clone is available from RZPD;
            contact RZPD (customer.service@rzpd.de) for further information.

            This CDS clone is a part of a collection of human full length
            expression clones generated by RZPD.
            This CDS has been cloned without stopcodon.
            The CDS has been inserted into pDONR201 via a BP Clonase(TM)
            reaction. Additional sequence has been added in front of the start
            codon (ATG): att..AAAAAA GCT GGC ACC CCT GGT CCA GGT (ATG)
            After the last codon additional sequence has been added: CCA GGC
            CCA GGC GGC G in front of the 3' att site (AC CCA GCT TTC TT).
            Compared to the reference sequence NM_000600 we did not find any
            amino acid exchanges.
```

Fig. 3.3 Overview of GenBank database

3.4.1.2 RefSeq

The RefSeq is a commonly used database in genomic and proteomic research. The database entries are incorporated into NCBI's repositories involving the nucleotide and protein. Records in the RefSeq database can be easily searched either by the keyword "RefSeq" or by their specific accession prefix (Nosek et al. 2015). The key characteristic of the RefSeq dataset is the combination of computation, collaboration, and curation of submitted reports by NCBI. The RefSeq group works together with many expert groups including official nomenclature authorities such as HUGO Gene Nomenclature Committee (HGNC) and Zebrafish Information Network (ZFIN), UniProtKB, and miRBase (Gray et al. 2015; Ruzicka et al. 2015; UniProtC 2015). The curators of RefSeq improve the quality of the database via review of QA test results, involvement in the selection of certain inputs for genome annotation processing, sequence analysis, taxonomic analysis, as well as functional review.

The RefSeq sequence records are generated by various methods depending on the sequence class and organism. NCBI's prokaryotic genome annotation pipeline (ncbi.nlm.nih.gov/books/NBK174280/) is used for archaeal and bacterial genome annotation, while collaboration and manual curation sustain a small number of reference bacterial genomes. The channel for a subset of eukaryotes including fungi, protozoa, and nematodes involves propagating annotation that has been submitted with standardization format to a RefSeq copy of the submitted genome assembly to the International Nucleotide Sequence Database Collaboration (INSDC). RefSeq sequence data are retrieved by using NCBI's nucleotide and protein databases, BLAST databases, NCBI's programmatic interface, or File Transfer Protocol.

3.4.1.3 Ensembl

Ensembl database supports various publicly available vertebrate genome assemblies by providing great-quality genomic resources. Curwen and co-workers introduced the Ensembl gene annotation system in 2004 (Curwen et al. 2004). Ensembl database was designed to annotate the species with high-quality draft genome assemblies, where same-species protein sequences and full-length cDNA sequences existed as input for identifying numerous protein-coding genes. Ensembl database was designed to annotate high-quality draft genome assemblies of different species (Fig. 3.4). The Ensembl gene annotation method is divided into four main stages: genome preparation, protein-coding model building, filtering, and gene set finalization (Curwen et al. 2004).

Ensembl does not produce the genome assemblies. However, it provides annotation on genome assemblies that have been deposited into a member database of the International Nucleotide Sequence Database Collaboration (INSDC) such as GenBank (Benson et al. 2014), ENA (Cochrane et al. 2013), and DDBJ (Kosuge et al. 2014). Upon getting an assembly from one of the INSDC databases, further it is loaded into a database and prepares it for sequence alignment through running the

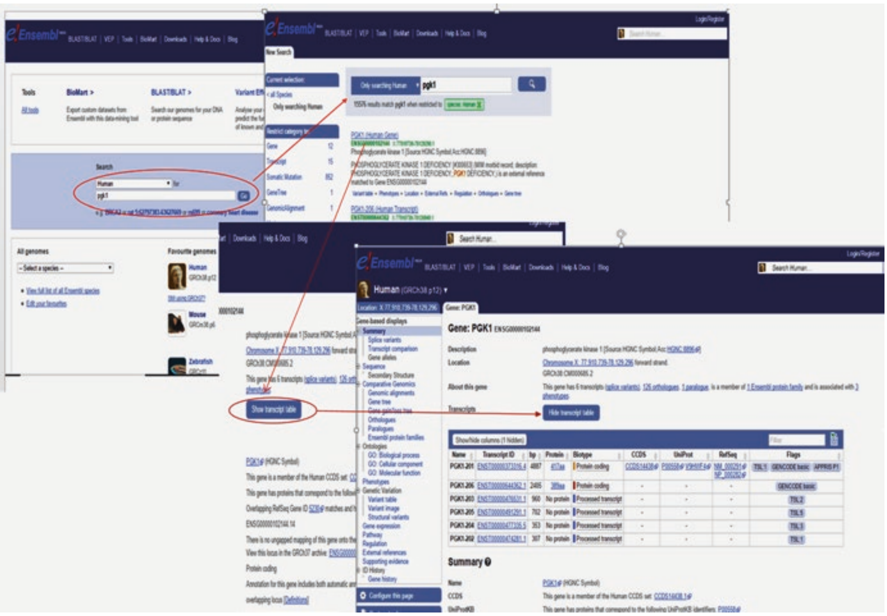


Fig. 3.4 Annotation of the gene phosphoglycerate kinase 1 (PGK 1) using the Ensembl database

repeat masking and raw compute examination. Retrieving information about the protein phosphoglycerate kinase 1 (PGK1) is depicted in Fig. 3.4.

For vertebrate genome assemblies, assembly loading involves introducing a list of the contig (component), scaffold, and chromosome accession into an Ensembl core database schema (Stabenau 2004). DNA sequences for all the contigs are the first pileup in the database. After that they load mappings between each coordinate system, using the AGP ("A Golden Path") files provided with the assembly. All annotation processes of the gene are run across the top-level coordinate system (Potter et al. 2004). RepeatMasker (Smit et al. 2013), Dust (Morgulis et al. 2006), and Tandem Repeat Finder (Benson 1999) (TRF) are used for disguising repetitive genomic sequence. Repbase repeat libraries (Jurka et al. 2005) are useful for RepeatMasker. Several RepeatMasker analyses are run for each of different chosen Repbase libraries and one for the custom RepeatModeler library generated in-house. Raw computes (Potter et al. 2004) is a term used for the selection of primary annotation analyses that are run across the genome assembly instantly after repeat masking. The protein model building stage involves the alignment of protein, cDNA, EST, and RNA-seq sequences to the genome assembly. This phase-specific method usually relies on the availability of input data at the time of annotation. Then the input datasets are selected according to attribution, with same-species data which is preferable over data from other species.

3.4.2 Protein Sequence Database

Different types of protein sequence databases ranging from simple to complex sequence databases exists (Finn et al. 2006). It is the collection of sequence data extracted from many sources, i.e., the annotated coding region of translations in GenBank, third-party annotation, RefSeq, etc. Commonly used sequence databases are given in Table 3.1.

3.4.2.1 TrEMBL

TrEMBL database contains computer-based entries that are derived from the translation of all coding sequences present in the DDBJ/EMBL/GenBank nucleotide sequence database not included in Swiss-Prot. To make sure the completeness, TrEMBL contains several protein sequences mined from the reported literature or submitted directly by the users (Apweiler et al. 2004). This database allows rapid access to protein sequence data. The data is made, and if a match is found, a set of secondary patterns computed with the eMotif algorithm is used to check the significance. It is a combination of two resources, Swiss-Protein + TrEMBL at the EBI, and is nominally redundant. It can be accessed at the SRS (Sequence Retrieval System) on the EBI web server (Emmert et al. 1994).

3.4.2.2 GenPept

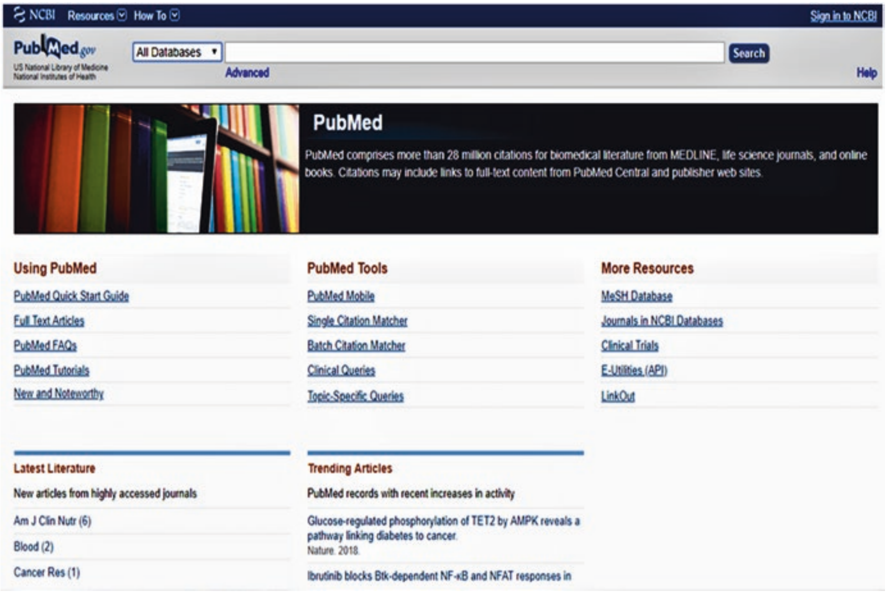
The GenPept database is developed by the National Center for Biotechnology Information (NCBI) (Apweiler et al. 2004). GenPept is a database of coding sequence features with a translation qualifier (Whitfield et al. 2006). This format is text-based and derived from the parent GenBank format. It comprises approximately 135,440,924 numbers of sequences which hold around 126,551,501,141 numbers of bases (Bagchi 2012). GenBank database allots a unique GenBank identifier or GenBank accession number to each submitted sequence.

3.4.2.3 Entrez Protein

Entrez is a WWW-based data retrieval tool developed by the NCBI, which can be used to search for information in 11 integrated NCBI databases, including GenBank and its subsidiaries, OMIM, and the literature database MEDLINE, through PubMed. Entrez is the common front end to all the databases maintained by the NCBI and is an extremely easy system to use (Whitfield et al. 2006). The Entrez main page, as with all NCBI pages, is usually quickly downloadable and does not have any specific requirements for web browsers (Fig. 3.5).

Table 3.1 Commonly used sequence databases and their descriptions

Category	Name	Link	Description
DNA	AFND	allelefrequencies.net	Allele Frequency Net Database
	dbSNP	ncbi.nlm.nih.gov/snp	Database of single nucleotide polymorphisms
	DEG	essentialgene.org	Database of essential genes
	EGA	ebi.ac.uk/ega	European Genome-phenome Archive
	Ensembl	ensembl.org	Ensembl genome browser
	EUGene	eugenes.org	Genomic information for eukaryotic organisms
	GeneCards	genecards.org	Integrated database of human genes
	JASPAR	jaspar.genereg.net	Transcription factor binding profile database
	JGA	trace.ddbj.nig.ac.jp/jga	Japanese Genotype-phenotype Archive
	MITOMAP	mitomap.org	Human mitochondrial genome database
	RefSeq	ncbi.nlm.nih.gov/refseq	NCBI Reference Sequence Database
	PolymiRTS	compbio.uthsc.edu/miRSNP	Polymorphism in miRNAs and their target sites
	1000 Genomes	1000genomes.org	A deep catalog of human genetic variation
Protein	EKPD	ekpd.biocuckoo.org	Eukaryotic Kinase and Phosphatase Database
	HPRD	hprd.org	Human Protein Reference Database
	InterPro	ebi.ac.uk/interpro	Protein sequence analysis and classification
	ModBase	salilab.org/modbase	Database of comparative protein structure models
	PDB	rcsb.org/pdb	Protein Data Bank for 3D structures of biological macromolecules
	PDBe	ebi.ac.uk/pdbe	Protein Data Bank in Europe
	Pfam	pfam.xfam.org	Database of conserved protein families and domains
	PIR	pir.georgetown.edu	Protein Information Resource
	SysPTM	lifecenter.sgst.cn/SysPTM	Posttranslational modifications
	UniProt	uniprot.org	Universal protein resource
	UUCD	uucd.biocuckoo.org	Ubiquitin and Ubiquitin-like Conjugation Database
	TreeFam	treefam.org	Database of phylogenetic trees of animal species
	CATH	cath.biochem.ucl.ac.uk	Protein structure classification
	CPLM	cplm.biocuckoo.org	Compendium of Protein Lysine Modifications
	DIP	dip.doe-mbi.ucla.edu	Database of Interacting Proteins



PubMed

PubMed comprises more than 28 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher web sites.

Using PubMed
- PubMed Quick Start Guide
- Full Text Articles
- PubMed FAQs
- PubMed Tutorials
- New and Noteworthy

PubMed Tools
- PubMed Mobile
- Single Citation Matcher
- Batch Citation Matcher
- Clinical Queries
- Topic-Specific Queries

More Resources
- MeSH Database
- Journals in NCBI Databases
- Clinical Trials
- E-Utilities (API)
- LinkOut

Latest Literature

New articles from highly accessed journals
- Am J Clin Nutr (6)
- Blood (2)
- Cancer Res (1)

Trending Articles

PubMed records with recent increases in activity
- Glucose-regulated phosphorylation of TET2 by AMPK reveals a pathway linking diabetes to cancer.
  Nature. 2018.
- Ibrutinib blocks Btk-dependent NF-κB and NFAT responses in

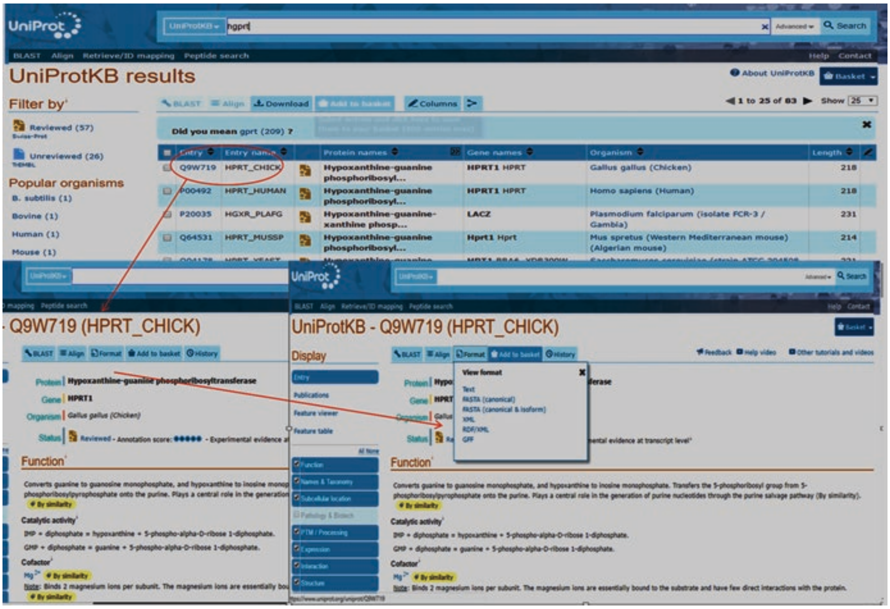
Fig. 3.5 Web home page of NCBI

3.4.2.4 UniProt

The National Institutes of Health (NIH) awarded a grant to combine the three protein sequence databases, Swiss-Prot, TrEMBL, and PIR-PSD databases, into a single resource, i.e., UniProt (Apweiler et al. 2004). It has many components: UniProt Knowledge Base (UniProtKB). It is a central part of UniProt Consortium's activities. It is a curated protein database which comprises two sections known as UniProtKB/Swiss-Prot (Boeckmann et al. 2003) and UniProtKB/TrEMBL (Whitfield et al. 2006). Data retrieval using UniProt is shown in Fig. 3.6.

3.5 Sequence Submission

For researchers to enter their sequence data, GenBank implements the World Wide Web sequence submission tool called BankIt, and a stand-alone program is called Sequin. Both the software are easy to handle that facilitate the researcher to enter as well as submit the annotated information to GenBank. By the worldwide collaboration of DDBJ and EMBL with GenBank database, the daily submission is forwarded to the respective databases.



UniProtKB results

Did you mean gprt (209) ?

Entry	Entry name		Protein names	Gene names	Organism	Length
Q9W719	HPRT_CHICK		Hypoxanthine-guanine phosphoribosyl...	HPRT1 HPRT	Gallus gallus (Chicken)	218
P00492	HPRT_HUMAN		Hypoxanthine-guanine phosphoribosyl...	HPRT1 HPRT	Homo sapiens (Human)	218
P20035	HGXR_PLAFG		Hypoxanthine-guanine-xanthine phosp...	LACZ	Plasmodium falciparum (isolate FCR-3 / Gambia)	231
Q64531	HPRT_MUSSP		Hypoxanthine-guanine phosphoribosyl...	Hprt1 Hprt	Mus spretus (Western Mediterranean mouse) (Algerian mouse)	214

UniProtKB - Q9W719 (HPRT_CHICK)

Protein | Hypoxanthine-guanine phosphoribosyltransferase
Gene | HPRT1
Organism | Gallus gallus (Chicken)
Status | Reviewed - Annotation score: ●●●●● - Experimental evidence at transcript level

Function

Converts guanine to guanosine monophosphate, and hypoxanthine to inosine monophosphate. Transfers the 5-phosphoribosyl group from 5-phosphoribosylpyrophosphate onto the purine. Plays a central role in the generation of purine nucleotides through the purine salvage pathway (By similarity).
By similarity

Catalytic activity

IMP + diphosphate = hypoxanthine + 5-phospho-alpha-D-ribose 1-diphosphate.
GMP + diphosphate = guanine + 5-phospho-alpha-D-ribose 1-diphosphate.

Cofactor

Mg^{2+} By similarity
Note: Binds 2 magnesium ions per subunit. The magnesium ions are essentially bound to the substrate and have few direct interactions with the protein.

Fig. 3.6 Exploring details of the protein hypoxanthine-guanine phosphoribosyltransferase (HGPRT) of *Gallus gallus* species using UniProt database

3.5.1 Sequin

Sequin has an interactive web interface as well as a graphical screen-based program. This stand-alone tool is designed to simplify the process of sequence submission as well as gives the handling capability to the increased amount of information to accommodate the long reads, complex sequence data, as well as authentic error analysis. The process of submission of the nucleic acid sequence is given in Fig. 3.7.

3.5.2 BankIt

The BankIt is the simplest sequence and descriptive data submitting tool in which the data is directly submitted to GenBank through the international collaboration web interface and instantly forwarded to DDBJ and EMBL databases.

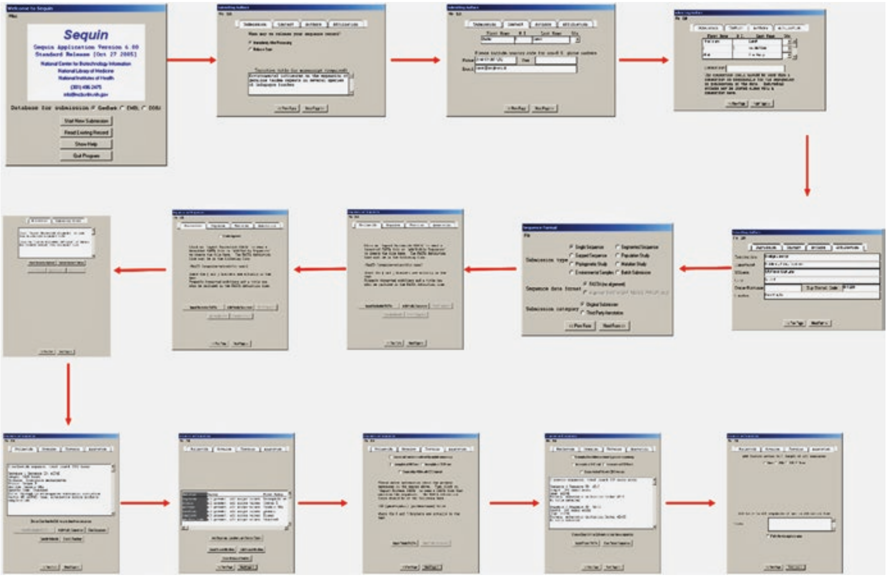

Welcome to Sequin
Sequin
Sequin Application Version 6.00
Standard Release [Oct 27 2005]
National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
Database for submission ● GenBank ○ EMBL ○ DDBJ
Start New Submission
Read Existing Record
Show Help
Quit Program

Fig. 3.7 Steps involved in the submission of nucleic acid sequence using Sequin

3.5.3 *Webin*

It is the European Bioinformatics Institute submitting program which guides users via a sequence checklist and their forms to allow the interactive as well as descriptive submission information. All the information required to create a databases access could be amassed during this process, i.e.:

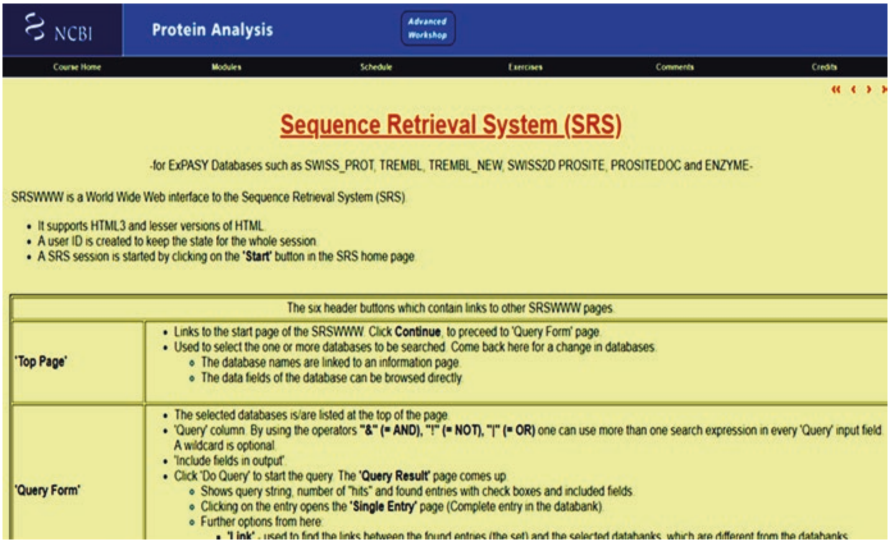
1. Submitter data
2. Launch date information
3. Sequence statistics, description, and source information
4. Reference quotation information

This program is used to enter the data as in single as well as multiple entries.

3.6 Retrieval

3.6.1 *SRS (Sequence Retrieval System)*

The Sequence Retrieval System, developed in EBI, is mainly used in the text searching data from multiple biological databases (Fig. 3.8). It gives the link to their appropriate biological information for entries that match the exploration criteria. This system is a highly recommended retrieval system for use.



Sequence Retrieval System (SRS)

-for ExPASY Databases such as SWISS_PROT, TREMBL, TREMBL_NEW, SWISS2D PROSITE, PROSITEDOC and ENZYME-

SRSWWW is a World Wide Web interface to the Sequence Retrieval System (SRS)

- It supports HTML3 and lesser versions of HTML
- A user ID is created to keep the state for the whole session
- A SRS session is started by clicking on the **'Start'** button in the SRS home page

The six header buttons which contain links to other SRSWWW pages	
'Top Page'	• Links to the start page of the SRSWWW. Click **Continue**, to preceed to 'Query Form' page • Used to select the one or more databases to be searched. Come back here for a change in databases ∘ The database names are linked to an information page ∘ The data fields of the database can be browsed directly
'Query Form'	• The selected databases is/are listed at the top of the page • 'Query' column. By using the operators **"&" (= AND), "!" (= NOT), "|" (= OR)** one can use more than one search expression in every 'Query' input field. A wildcard is optional • 'Include fields in output' • Click 'Do Query' to start the query. The **'Query Result'** page comes up ∘ Shows query string, number of "hits" and found entries with check boxes and included fields ∘ Clicking on the entry opens the **'Single Entry'** page (Complete entry in the databank) ∘ Further options from here ▪ **'Link'** - used to find the links between the found entries (the set) and the selected databanks, which are different from the databanks

Fig. 3.8 The home page of Sequence Retrieval System (SRS)

3.6.2 *Entrez*

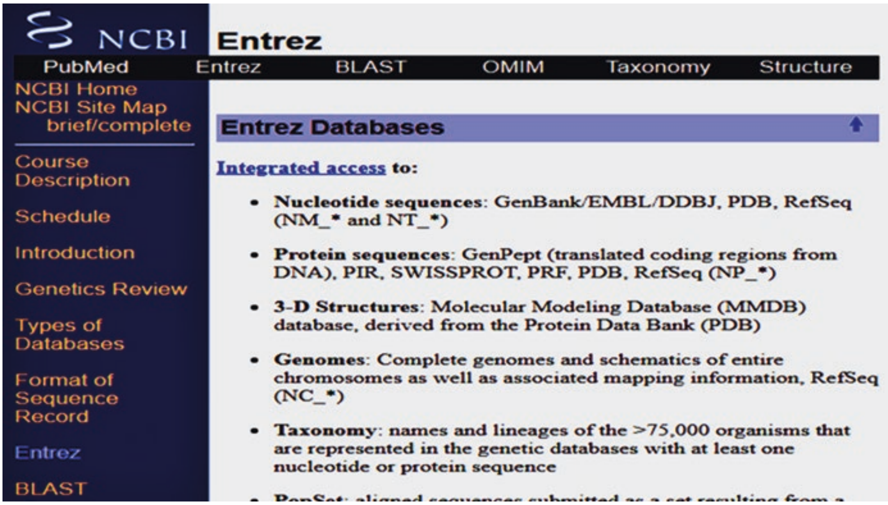
NCBI developed the Entrez retrieval system. The Entrez is an organized search engine that provides the users to retrieve many NIH biomedical information science databases at NCBI. The home page of the Entrez database is shown in Fig. 3.9.

3.6.3 *DBGET*

The DBGET retrieval system is developed in the University of Tokyo. This system provides the multiple databases of molecular biology database entry at a time. The home page of DBGET is shown in Fig. 3.10.

3.7 Conclusion

The phenomenal production of genome and proteome data underscores the necessity to develop and maintain biological databases. This huge-scale data provides an opportunity for data scientist to retrieve sequence as well as structure information of the data extracted from the diverse group of organisms. The raw data coming from the massive numbers of biological studies can provide critical insights if it is properly coded, stored, analyzed, and interpreted with the help of sequence

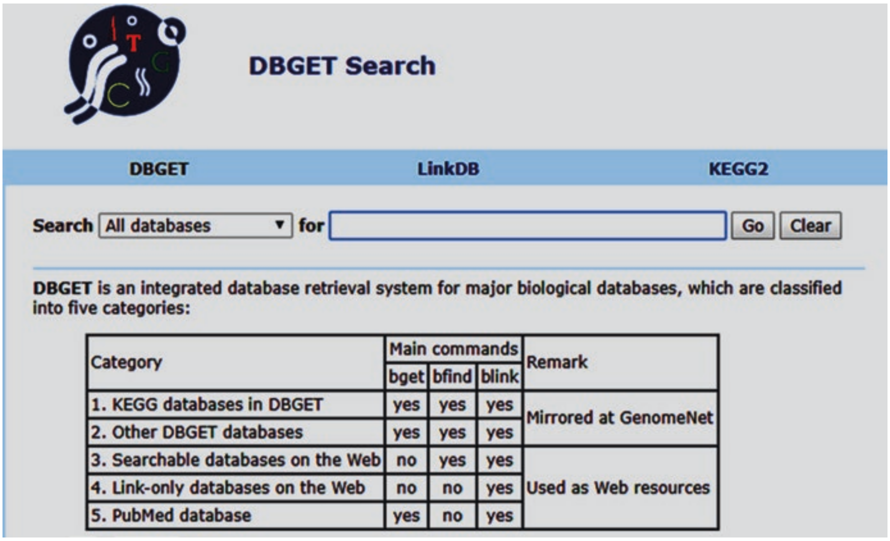


Entrez Databases

Integrated access to:

- **Nucleotide sequences**: GenBank/EMBL/DDBJ, PDB, RefSeq (NM_* and NT_*)
- **Protein sequences**: GenPept (translated coding regions from DNA), PIR, SWISSPROT, PRF, PDB, RefSeq (NP_*)
- **3-D Structures**: Molecular Modeling Database (MMDB) database, derived from the Protein Data Bank (PDB)
- **Genomes**: Complete genomes and schematics of entire chromosomes as well as associated mapping information, RefSeq (NC_*)
- **Taxonomy**: names and lineages of the >75,000 organisms that are represented in the genetic databases with at least one nucleotide or protein sequence
- **PopSet**: aligned sequences submitted as a set resulting from a

Fig. 3.9 Overview of NCBI Entrez databases



DBGET is an integrated database retrieval system for major biological databases, which are classified into five categories:

Category	Main commands			Remark
	bget	bfind	blink	
1. KEGG databases in DBGET	yes	yes	yes	Mirrored at GenomeNet
2. Other DBGET databases	yes	yes	yes	
3. Searchable databases on the Web	no	yes	yes	Used as Web resources
4. Link-only databases on the Web	no	no	yes	
5. PubMed database	yes	no	yes	

Fig. 3.10 DBGET retrieval system

databases. The sequence databases offer two kinds of fundamental benefits including the deposition of newly resolved sequences and their comparison with previously deposited sequence repositories. In the current chapter, we provided a brief overview about different types of knowledge bases like primary, secondary, and composite biological databases along with some specialized databases which host information about RNA molecules, protein-protein interactions, metabolic pathways, and so on.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Ansorge WJ (2009) Next-generation-sequencing techniques. J New Biotechnol 25(4):195–203

Apweiler R, Bairoch A, Wu CH (2004) Protein sequence databases. Curr Opin Chem Biol 8:76–80

Bagchi A (2012) A brief overview of a few popular and important protein databases. Comput Mol Biosci 2:115–120

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27:573–580

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL (2000) GenBank. Nucleic Acids Res 28(1):15–18

Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2014) GenBank. Nucleic Acids Res 42:32–37

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein databank. Nucleic Acids Res 28(1):235–242

Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL. Nucleic Acids Res 31:365–370

Cochrane G, Alako B, Amid C, Bower L, Cerdeño-Tárraga A, Cleland I, Gibson R, Goodgame N, Jang M, Kay S, Leinonen R, Lin X, Lopez R, McWilliam H, Oisel A, Pakseresht N, Pallreddy S, Park Y, Plaister S, Radhakrishnan R, Rivière S, Rossello M, Senf A, Silvester N, Smirnov D, Ten Hoopen P, Toribio A, Vaughan D, Zalunin V (2013) Facing growth in the European nucleotide archive. Nucleic Acids Res 41:30–35

Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, Searle SM, Clamp M (2004) The Ensembl automatic gene annotation system. Genome Res 14:942–950

Dennis AB, llene KM, David JL, James O, Barbara AR, David LW (2000) GenBank. Nucleic Acids Res 34:16–20

Emmert DB, Stoehr PJ, Stoesser G, Cameron GN (1994) The European Bioinformatics Institute (EBI) databases. Nucleic Acids Res 22:3445–3449

Farrell CM, O'Leary NA, Harte RA, Loveland JE, Wilming LG, Wallin C, Diekhans M, Barrell D, Searle SM, Aken B et al (2014) Current status and new features of the Consensus Coding Sequence database. Nucleic Acids Res 42:865–872

Finn RD, Mistry J, Schuster BB, Griffiths JS, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer EL, Bateman A (2006) Pfam: clans, webtools and services. Nucleic Acids Res 34:247–251

Gray KA, Yates B, Seal RL, Wright MW, Bruford EA (2015) Genenames.org: the HGNC resources in 2015. Nucleic Acids Res 43:1079–1085

Heather JM, Chain B (2016) The sequence of sequencers: the history of sequencing DNA. Genomics 107:1–8

Hughes EA (2001) Sequence databases and the internet. Methods Mol Biol 167:215–223

John GSM, Chellan R, Satoru T (2011) Understanding tools and techniques in protein structure prediction. In: System and computational biology. InTech London UK, 185–212

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J (2005) Repbase update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110:462–467

Kosuge T, Mashima J, Kodama Y et al (2014) DDBJ progress report: a new submission system for leading to a correct annotation. Nucleic Acids Res 42:44–49

Morgulis A, Gertz EM, Schäffer AA, Agarwala R (2006) A fast and symmetric DUST implementation to mask low-complexity DNA sequences. J Comput Biol 13:1028–1040

Neelameghan A (1997) S.R. Ranganathan's general theory of knowledge classification in designing, indexing, and retrieving from specialized databases. Inform J 34:25–42

Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Breckler SJ, Buck S, Chambers CD, Chin G, Christensen G (2015) Promoting an open research culture. Science 348:1422–1425

Pearson WR (1994) Using the FASTA program to search protein and DNA sequence databases. Methods Mol Biol 24:307–331

Potter SC, Clarke L, Curwen V et al (2004) The Ensembl analysis pipeline. Genome Res 14:934–941

Ragunath PK, Venkatesan P, Ravimohan R (2009) New curriculum design model for bioinformatics postgraduate program using systems biology approach. J Comput Sci Syst Biol 2:300–305

Ruzicka L, Bradford YM, Frazer K, Howe DG, Paddock H, Ramachandran S, Singer A, Toro S, Van Slyke CE, Eagle AE, Fashena D, Kalita P, Knight J, Mani P, Martin R, Moxon SA, Pich C, Schaper K, Shao X, Westerfield M (2015) ZFIN, the zebrafish model organism database: updates and new directions. Genesis 53:498–509

Smit AFA, Hubley R, Green P (2013) RepeatMasker Open-4.0.2013–2015. http://www.repeatmasker.org

Stabenau A (2004) The Ensembl core software libraries. Genome Res 14:929–933

Tateno Y, Miyazaki S, Ota M, Sugawara H, Gojobori T (2000) DNA Data Bank of Japan (DDBJ) in collaboration with mass sequencing teams. Nucleic Acids Res 28:24–26

Toomula N, Kumar A, Kumar S, Bheemidi VS (2011) Biological databases-integration of life science data. J Comput Sci Syst Biol 4(5):087–092

UniProtC (2015) UniProt: a hub for protein information. Nucleic Acids Res 43:204–212

Whitfield EJ, Pruess M, Apweiler R (2006) Bioinformatics database infrastructure for biotechnology research. J Biotechnol 124:629–639

Chapter 4
Biological 3D Structural Databases

Yasser Gaber, Boshra Rashad, and Eman Fathy

Contents

Y. Gaber (✉)
Department of Pharmaceutics and Pharmaceutical Technology, Faculty of Pharmacy,
Mutah University, Al-Karak, Jordan

Microbiology Department, Faculty of Pharmacy, Beni-Suef University, Beni-Suef, Egypt
e-mail: Yasser.Gaber@mutah.edu.jo; Yasser.Gaber@pharm.bsu.edu.eg

B. Rashad
Biotechnology and Life Sciences Department, Faculty of Postgraduate Studies for Advanced Sciences (PSAS), Beni-Suef University, Beni-Suef, Egypt
e-mail: boshramohamed@psas.bsu.edu.eg

E. Fathy
Biotechnology and Life Sciences Department, Faculty of Postgraduate Studies for Advanced Sciences (PSAS), Beni-Suef University, Beni-Suef, Egypt

Microbiology Department, Directorate of Health Affairs at Ministry of Health,
Beni-Suef, Egypt
e-mail: emanfathy@psas.bsu.edu.eg

N. A. Shaik et al. (eds.), *Essentials of Bioinformatics, Volume I*,
https://doi.org/10.1007/978-3-030-02634-9_4

4.1 Introduction

Structural databases are storage platforms that are devoted to the three-dimensional (3D) structural information of macromolecules. The 3D structure determination of biomacromolecules is essential for understanding phenomena such as the mechanisms of disease development that can aid in the design of new drugs. Also, 3D structures of biomacromolecules help to find the structure-function relationship. For instance, a point mutation in an enzyme can lead to a serious disease; this is exemplified by the glucose-6-phosphate dehydrogenase mutant enzyme that has lower ability to bind NADP+cofactor, thus resulting in the hemolytic anemia syndrome (Wang et al. 2008). The availability of 3D structural information of macromolecules will unveil the mysterious protein-protein interaction. Also, the conserved amino acid analysis using 3D structural features of proteins facilitates understanding the structure activity relationships. Proteins are polymers of amino acid sequence; it is amazing that only 20 different amino acids account for all the diversities of proteins, which are mainly arranged into primary, secondary, tertiary, and quaternary structural forms. The primary structure refers to the linear attachment of amino acids that make up the polypeptide chain. Secondary structure denotes repeated and regular folding patterns of the main chain sequence either an alpha helix or beta sheets connected via coils, turns, or loops. Tertiary structure is the characteristic three-dimensional shape resulted from the secondary structure elements found in the protein. Quaternary structure refers to two or more protein subunits that are linked to each other via non-covalent interaction.

The start of original structure biology dates back to the 1950s, when DNA double helix, hemoglobin, and myoglobin structures were determined. In the following years, scientists paid great attention to the evaluation and study of protein structure in terms of the relation between protein sequence, structure, and function. In 1971, structure biologists held an important meeting to discuss the allowance of the public accessibility to structural data; as a result, the Brookhaven National Laboratory hosted the Protein Data Bank (Berman et al. 2012). The structural databases aim at keeping the information about the structures of each biomacromolecule, annotate its properties, and facilitate to the users finding relevant information and related structures. Table 4.1 lists the main 3D structural databases, tools, and servers that are essential for biologists, bioinformaticians, and even the public interested in structure biology. There are several structural databases that are available free of charge for public use and are responsible for archiving and organizing the 3D structural information of biological macromolecules and proteins such as RCSB PDB, PDBe, PDBj, BMRB, SCOP, and CATH. The 3D structural information can be seen as a primary source of data that requires effort for extraction and interpretation of the useful information. Therefore, other several types of databases and web servers are developed to add further levels of information such as comparison to other structures or focus on certain property, for example, the membrane protein databases (Bagchi 2012).

Table 4.1 List of important structural biological databases and related web resources for structure analysis

Database	Use/description	Link	References
1. Primary structural data centers and other browsers			
PDBj	Protein Data Bank Japan archives macromolecular structures and provides integrated tools	https://pdbj.org/	Kinjo et al. (2016)
BMRD	Biological Magnetic Resonance Data Bank (NMR), a repository for data from NMR spectroscopy on proteins, peptides, nucleic acids, and other biomolecules	http://www.bmrb.wisc.edu/	Markley et al. (2008)
PDBe	Protein Data Bank in Europe (PDBe) archives biological macromolecular structures	http://www.ebi.ac.uk/pdbe/	Velankar et al. (2010) and Velankar et al. (2015)
RCSB PDB	Research Collaboratory for Structural Bioinformatics Protein Data Bank archives information about the 3D shapes of proteins, nucleic acids, and complex assemblies	https://www.rcsb.org/	Berman et al. (2000)
PDBsum	Pictorial analysis of macromolecular structures	www.ebi.ac.uk/pdbsum	Laskowski (2007) and Laskowski et al. (2018)
2. Structure classification databases			
CATH	Domain classification of structures	http://www.cathdb.info/	Knudsen and Wiuf (2010)
SCOP	SCOP2, structural and evolutionary classification	http://scop2.mrc-lmb.cam.ac.uk/	Lo Conte et al. (2000)
3. Nucleic acid databases			
NDB	Nucleic acid database	http://ndbserver.rutgers.edu/	Coimbatore Narayanan et al. (2013)
RNA FRABASE	3D structure of RNA fragments	http://rnafrabase.cs.put.poznan.pl/	Popenda et al. (2010)
NPIDB	3D structures of nucleic acid-protein complexes	http://npidb.belozersky.msu.ru/	Zanegina et al. (2015)
4. Membrane protein database			
MemProtMD	MemProtMD, database of membrane protein	http://sbcb.bioch.ox.ac.uk/memprotmd/	Stansfeld et al. (2015)
5. Ligands and binding sites and metalloproteins			
PeptiSite	Is a comprehensive and reliable database of biologically and structurally characterized peptide-binding sites that can be identified experimentally from co-crystal structures in the Protein Data Bank	http://peptisite.ucsd.edu/	Acharya et al. (2014)

(continued)

Table 4.1 (continued)

Database	Use/description	Link	References
ComSin	Database of protein structures inbound (complex) and unbound (single) states in relation to their intrinsic disorder	http://antares.protres.ru/comsin/	Lobanov et al. (2009)
MetalPDB	MetalPDB collects and allows easy access to the knowledge on metal sites in biological macromolecules	http://metalweb.cerm.unifi.it/	Putignano et al. (2017)
Pocketome	The Pocketome is an encyclopedia of conformational ensembles of druggable binding sites that can be identified experimentally from co-crystal structures in the wwPDB	http://www.pocketome.org/	An et al. (2005)
MIPS	A database of all the metal-containing proteins available in the Protein Data Bank	http://dicsoft2.physics.iisc.ernet.in/cgi-bin/mips/query.pl	Mewes et al. (2002)
6. Structure comparison servers			
DALI	The Dali server is a service used for comparing protein 3D structures	http://ekhidna2.biocenter.helsinki.fi/dali/	Holm and Rosenström (2010)
VAST+	**V**ector **A**lignment **S**earch **T**ool, web-based tool for comparing 3D structure against all structures in the Molecular Modelling Database (MMDB), NCBI	https://structure.ncbi.nlm.nih.gov/Structure/VAST/vastsearch.html	Madej et al. (2013)
CE	A method for comparing and aligning protein structures	http://source.rcsb.org/ceHome.jsp	Shindyalov and Bourne (1998)
7. Other databases			
PTM-SD	Posttranslational modification database	http://www.dsimb.inserm.fr/dsimb_tools/PTM-SD/	Craveur et al. (2014)
PED3	Protein Ensemble Database The database of conformational ensembles describing flexible proteins	http://pedb.vib.be/	Varadi and Tompa (2015)
GFDB	Glycan Fragment Database (GFDB), identifying PDB structures with biologically relevant carbohydrate moieties and classifying PDB glycan structures based on their primary sequence and glycosidic linkage	http://www.glycanstructure.org/	Jo and Im (2012)
ChEBI	**C**hemical **E**ntities of **B**iological **I**nterest (**ChEBI**), a database focused on "small" chemical compounds	https://www.ebi.ac.uk/chebi/	Hastings et al. (2015)
ChEMBL	ChEMBL is a database of bioactive drug-like small molecules	https://www.ebi.ac.uk/chembl/	Gaulton et al. (2016)

Table 4.2 Experimental methods used for determination of macromolecule 3D structures

	X-Ray crystallography	Nuclear magnetic resonance	Cryo-EM
Experimental steps	1. X-rays are scattered by electrons in the atoms of crystal. 2. Then recorded on a detector, e.g., CCDS. 3. Phase estimation and calculation of electron density map. 4. Fit primary sequence to electron density map (model). 5. Model refinement. 6. Deposition in PDB	1. Molecules absorb radiofrequency radiation held in a strong magnetic field. 2. Resonance frequency detection influenced by chemical environment. 3. Collection of conformational interatomic distance constraints. 4. Calculation of the 3D structure. 5. Deposition in PDB	1. Sample is vitrified at liquid nitrogen temp. 2. High-energy electron beam passes through it under high vacuum. 3. Image is produced when transmitted electrons are projected to a detector 4. Structure determination
Specimen	Crystals	Solution	Vitrified solution[a]
Protein size	Wide range	Below 40–50 KDa	>150 KDa
Contribution[b]	>89% of PDB entries	> 9% of PDB entries	>1% of PDB entries
Resolution	Higher resolution	High resolution	Significantly low >3.5 Å
Advantages	Well-developed Accurate, easy for model building	Provide dynamic information	Easy sample preparation Samples in its native environment
Disadvantages	Crystallization step Slow process	High purity sample is required Less precise than X-ray Intensive computational simulations	Cost Mainly for large molecules and assemblies

[a] A vitrified solution is the solidification of a liquid into a noncrystalline or amorphous solid known as glass

The determination of the 3D structure for biological macromolecules is done by four fundamental techniques arranged in terms of familiarity and contribution as follows: X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, cryo-electron microscopy (Cryo-EM), and neutron diffraction. Table 4.2 summarizes the experimental steps adopted in the first three techniques and shows main advantages and disadvantages of these techniques. Although these techniques are viable and inestimable, they cannot build an atomic structure model from scratch without former knowledge of the proteins' chemical and physical properties and the proteins' primary sequences.

4.1.1 X-Ray Crystallography

X-ray protein crystallography is a branch of science that plays a vital role in many aspects including the determination of the 3D structure of proteins. Proteins' 3D structure determination enables us to perceive the relationships between the structure and the function of these molecules and characterizes drug targets such as G-coupled protein receptors (Rosenbaum et al. 2009), 3D structures of enzymes, DNA structure, and others. In 1895, Wilhelm Roentgen eternalized his name by discovering a new unknown type of rays that has a shorter wavelength than the UV rays; he named it X-rays. In 1912 Max von Laue demonstrated that X-rays can be diffracted upon interacting with a crystalline material. The following year, Bragg, the father, and his son, could solve a very challenging step in using X-rays for structure determination that was known as the phase problem; they succeeded in paving the way to use X-ray diffraction to know the 3D structure of a crystalline material. According to the current status, X-ray protein crystallography can be summarized in two main successive steps:

4.1.2 Crystal Formation

The X-ray crystallography experiment is based on shooting a protein crystal with X-rays. The process of getting crystals can be a cumbersome task since it is somehow a trial-and-error rather than systematic experiment. The process starts with obtaining a protein sample in high concentration. This step is done nowadays using different techniques of recombinant DNA technology. It is noteworthy to mention that the advancement in DNA synthesis has facilitated the process of gene cloning and expression. Advancement in genetics has not only facilitated the synthesis of genetic sequences at very reasonable cost but also assisted in controlling the gene expression by manipulating the molecular regulatory elements in the host cells (e.g., *Escherichia coli*, *Pichia*, or mammalian cells). Aided by different DNA techniques, the gene of interest can be overexpressed in suitable expression host to yield the target protein in a very good yield. Taking advantage of DNA recombinant technology, it is possible to add tags to the overexpressed proteins that will help in the purification steps (e.g., multi-histidine residues to the overexpressed protein to aid in metal affinity chromatography or SUMO tag that helps in an expression of the protein in a good yield) (Gaber et al. 2016). The overexpressed protein will undergo a process of purification until it is obtained in a high purity as judged by SDS-PAGE analysis. Afterward, a concentrated protein solution will be subjected to a crystal formation experiment. In such experiment, the concentrated protein solution will be exposed to different buffer solutions with different additives such as ethylene glycol; the process is run in a miniaturized setting that allows testing hundreds of crystallization conditions in a short time and in an automated manner. The appearance of crystals in any of the tested conditions will be considered a positive hit that will lead to picking this specific condition and pursuing with the condition to reach a big crystal size of the protein. It is worth mentioning that membrane proteins are among the very difficult protein types to be crystallized. The difficulty comes from different reasons such as flexibility issues, instability, usage of detergent for extraction from cell membrane, purification, crystallization, data collection, and structure solution (Carpenter et al. 2008; Wlodawer et al. 2008).

4.1.3 Structure Determination

A special facility named synchrotron is used in the process of X-ray shooting. These facilities are located mostly in Europe, the USA, Japan, and Australia, for example, in Grenoble, France, and Lund, Sweden. The synchrotron is big laboratories that accelerate electrons to generate X-rays. The crystals obtained from the crystallization process are kept frozen in liquid nitrogen to protect them from destruction upon exposure to the high-energy rays. Special types of detectors collect the diffraction patterns obtained from the process of crystal exposure to the X-ray. These detectors have witnessed continuous development in order to facilitate the data collection process. The obtained data are then subjected to what is known data reduction in order to reduce the number of data obtained. Eventually, the data obtained will lead to what is known as electron density map which can be described as an in silico representation of a 3D shape of the protein revealed from the X-ray shooting experiment. The electron density map can be figured numerically by Fourier transformation (Wlodawer et al. 2008). The following step is to fit the protein primary amino acid sequence into the obtained electron density map providing the preliminary 3D model; this was a challenging task; however a plethora of programs are created to alleviate this issue; the most common one is COOT (Emsley et al. 2010). COOT is a widely used molecular graphics program for model building and biological molecule validation. It unveils atomic models and electron density maps and permits the manipulations of built models. Moreover, COOT supplies access to numerous validation and refinement tools. Validation of the preliminary model is vital before depositing final structure model into the PDB as a misinterpretation of data is liable. Many programs can help with this issue like PROCHECK. In addition, attempts to re-evaluate structures after deposition into PDB have been spotted; PDB-REDO is a good example of such efforts which can re-refine formerly deposited structures (Joosten et al. 2010).

4.2 Macromolecular Structural Databases

4.2.1 Protein Data Bank wwPDB

The Worldwide Protein Data Bank abbreviated as wwPDB (www.wwpdb.org) is the central organization that takes the responsibility to maintain and archive the 3D structural information of biomacromolecules. wwPDB stores 141,150 records of 3D structures (updated April 2018).

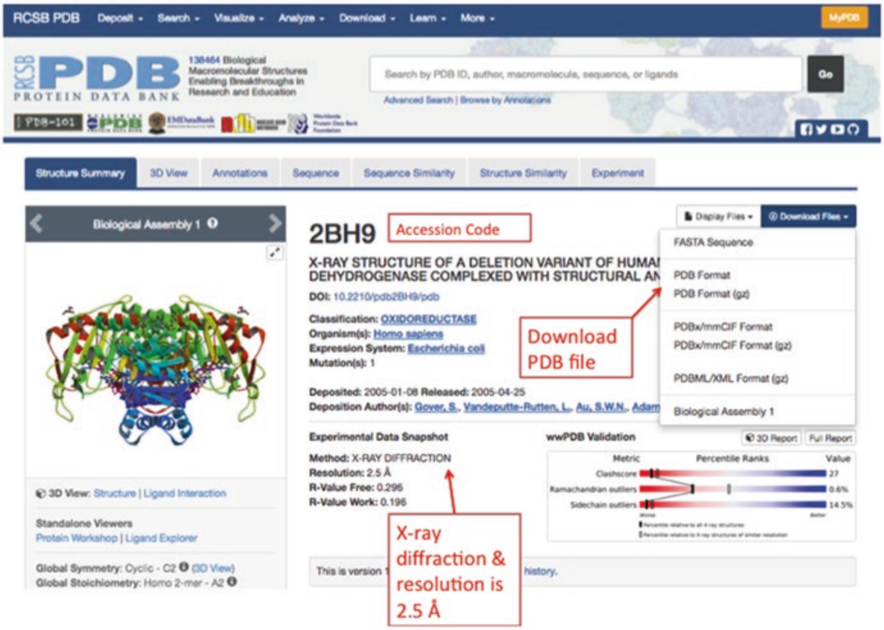
The wwPDB is composed of four partners:

(i) Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB) (Berman et al. 2000)
(ii) Protein Data Bank in Europe (PDBe) (Velankar et al. 2010)
(iii) Protein Data Bank Japan (PDBj) (Kinjo et al. 2016)
(iv) Biological Magnetic Resonance Data Bank (BMRB) (Markley et al. 2008)

4.2.1.1 RCSB PDB

The RCSB Protein Data Bank (RCSB PDB, http://www.rcsb.org) is the US partner of wwPDB, and it presents the PDB archive in an organized and easy way to explore. The PDB archive is accessed by the public and serves diverse disciplines that encompass agricultural, pharmaceutical, and biotechnological applications. It is worth mentioning that a majority of PDB users are of limited expertise in structural biology. The design of the RCSB PDB webpage allows easy navigation and provides different options to find the structure of interest, facilitate in finding similar structures, and jump to related contents in different databases. Figure 4.1 shows a screenshot of RCSB PDB webpage viewing the accession code 2BH9. The page is organized into sections that include different types of information as indicated briefly as follows:

1. The front of the page shows the accession code 2BH9 and information about the authors and the deposition date.
2. The right corner contains a hyperlink to downloadable structural files of the 2BH9 as PDB file extension in addition to other types such as PDBx/mmCIF files. The typical form of storing 3D structure information is PDB file format. These files are typically opened with specific molecular visualization software such as PyMOL or YASARA (DeLano 2002; Krieger and Vriend 2014). However, the file can also be opened and edited – though is not advised for novice users – with text editor software programs such as Notepad or Microsoft Word. Figure 4.2 shows the PDB file for 2BH9 entry as an example; the file lists all the atoms present in the macromolecule (protein) and its coordinates as *X*, *Y*, and *Z*. A typical PDB file includes a header that gives a summary of the protein in terms of its source, author details, and the experimental techniques used. Since the size of



2BH9

X-RAY STRUCTURE OF A DELETION VARIANT OF HUMAN DEHYDROGENASE COMPLEXED WITH STRUCTURAL AN

DOI: 10.2210/pdb2BH9/pdb

Classification: OXIDOREDUCTASE
Organism(s): Homo sapiens
Expression System: Escherichia coli
Mutation(s): 1

Deposited: 2005-01-08 Released: 2005-04-25
Deposition Author(s): Gover, S., Vandeputte-Rutten, L., Au, S.W.N., Adam

Experimental Data Snapshot

Method: X-RAY DIFFRACTION
Resolution: 2.5 Å
R-Value Free: 0.296
R-Value Work: 0.196

Global Symmetry: Cyclic - C2 (3D View)
Global Stoichiometry: Homo 2-mer - A2

This is version 1 ... history.

Accession Code

Download PDB file

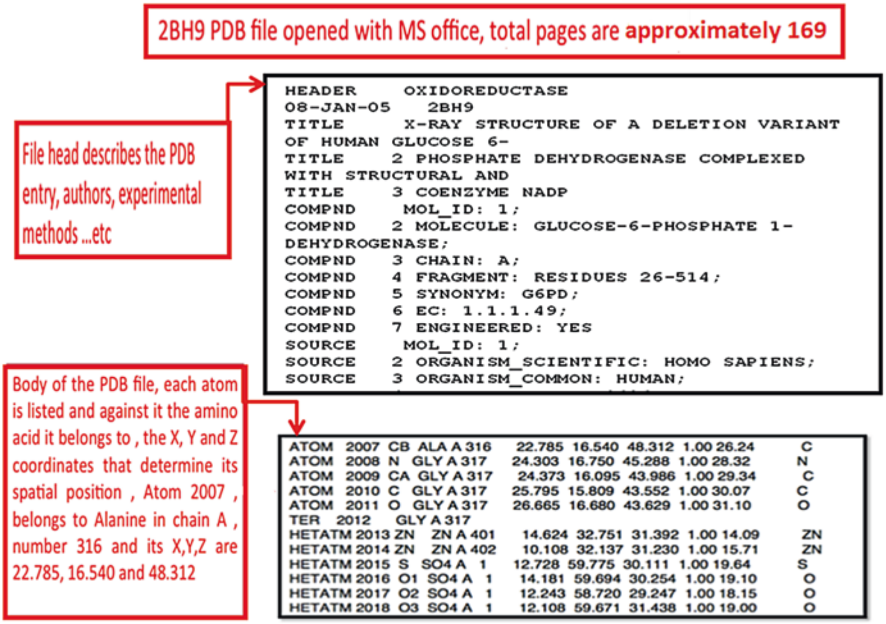
X-ray diffraction & resolution is 2.5 Å

Fig. 4.1 A screenshot of PDB webpage interface for an oxidoreductase protein, deposited under the accession code (2BH9), structure determined by X-ray diffraction technique at a resolution of 2.5 Å. The source organism is *Homo sapiens* and overexpressed in *E. coli*. It also provides different types of downloadable file formats for the user to choose from, e.g., FASTA sequence, PDB, and mmCIF file formats

3D structure information is too big in few cases like virus capsid, a new file format – PDBx/mmCIF – is introduced to accommodate such large files.

3. Information about the peer-reviewed publications linked to 2BH9 and the citation information.
4. Macromolecule section that shows the CATH classification of 2BH9 and the accession code of 2BH9 at UniProt database.
5. Experimental data snapshot: this section is devoted to the X-ray crystallography experiment and the statistical data revealing the resolution of the structure. In case of 2BH9, the structure was determined at a resolution of 2.5 Å, which is not a very good resolution. Resolution refers to the quality of the experimental data generated by X-ray crystallography. High-resolution structures will be determined at values of less than 1.5 Å or so; this level of accuracy of determining the atomic positions is high. Conversely, at a resolution of 3 Å or higher, the structure shape as global will be inferred; however the accurate positioning of the individual atoms is poor.

PDBe Protein Data Bank in Europe, (http://www.ebi.ac.uk/pdbe/) is the European equivalent to RCSB PDB. The PDBe home page provides an organized structure to



2BH9 PDB file opened with MS office, total pages are approximately 169

File head describes the PDB entry, authors, experimental methods ...etc

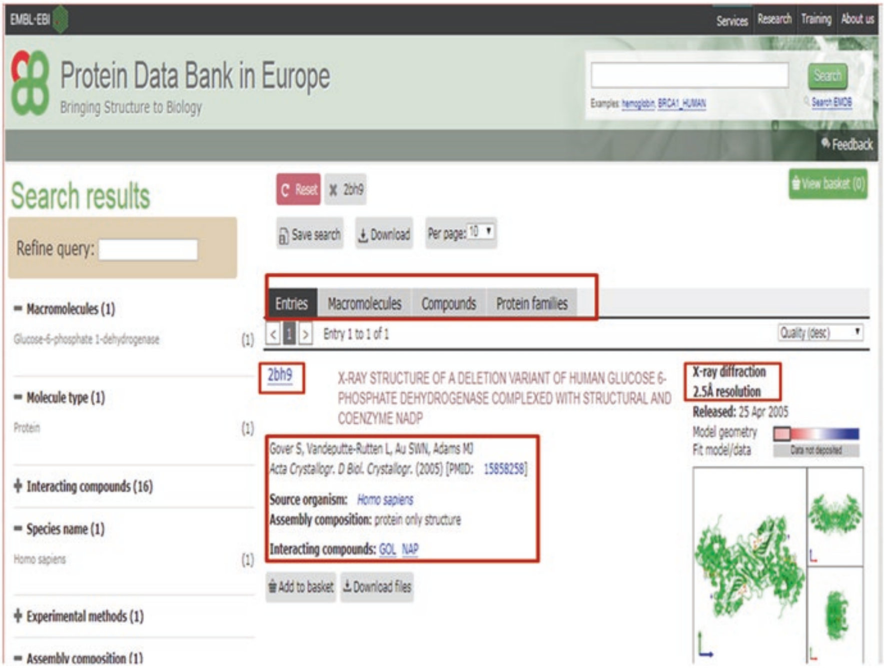
```
HEADER    OXIDOREDUCTASE
08-JAN-05   2BH9
TITLE     X-RAY STRUCTURE OF A DELETION VARIANT
OF HUMAN GLUCOSE 6-
TITLE    2 PHOSPHATE DEHYDROGENASE COMPLEXED
WITH STRUCTURAL AND
TITLE    3 COENZYME NADP
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: GLUCOSE-6-PHOSPHATE 1-
DEHYDROGENASE;
COMPND   3 CHAIN: A;
COMPND   4 FRAGMENT: RESIDUES 26-514;
COMPND   5 SYNONYM: G6PD;
COMPND   6 EC: 1.1.1.49;
COMPND   7 ENGINEERED: YES
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: HOMO SAPIENS;
SOURCE   3 ORGANISM_COMMON: HUMAN;
```

Body of the PDB file, each atom is listed and against it the amino acid it belongs to , the X, Y and Z coordinates that determine its spatial position , Atom 2007 , belongs to Alanine in chain A , number 316 and its X,Y,Z are 22.785, 16.540 and 48.312

```
ATOM   2007 CB  ALA A 316    22.785 16.540 48.312 1.00 26.24        C
ATOM   2008 N   GLY A 317    24.303 16.750 45.288 1.00 28.32        N
ATOM   2009 CA  GLY A 317    24.373 16.095 43.986 1.00 29.34        C
ATOM   2010 C   GLY A 317    25.795 15.809 43.552 1.00 30.07        C
ATOM   2011 O   GLY A 317    26.665 16.680 43.629 1.00 31.10        O
TER    2012     GLY A 317
HETATM 2013 ZN   ZN A 401    14.624 32.751 31.392 1.00 14.09       ZN
HETATM 2014 ZN   ZN A 402    10.108 32.137 31.230 1.00 15.71       ZN
HETATM 2015 S   SO4 A   1    12.728 59.775 30.111 1.00 19.64        S
HETATM 2016 O1  SO4 A   1    14.181 59.694 30.254 1.00 19.10        O
HETATM 2017 O2  SO4 A   1    12.243 58.720 29.247 1.00 18.15        O
HETATM 2018 O3  SO4 A   1    12.108 59.671 31.438 1.00 19.00        O
```

Fig. 4.2 PDB file format of the entry 2BH9 as an example, opened with Microsoft Word. A PDB file provides a full description of the entry such as a list of protein atoms and their 3D arrangement in space. For 2BH9 the header section provides information about the entry citation, authors, source of enzyme, and the experimental technique used in solving the structure. The file body provides information about the protein's atoms; each atom is listed opposite the amino acid it belongs to; moreover, it provides data about *X*, *Y*, and *Z* coordinates that determine its spatial position. For instance, atom number 2007 belongs to alanine in chain a number 316, and its *X*, *Y*, and *Z* coordinates are 22.785, 16.540, and 48.312, respectively

ease the browsing and exploration of the content. The tab (PDBe services) allows the access to categorize resources according to the user's interest and background; these categorized tabs are structural biologists, bioinformaticians, life scientists, students and teachers, medicinal chemists, journal editors and referees, and all services tab. For example, a good training and educational material are available under the PDBe training tab. Among the popular services that are provided by PDBe is FASTA protein sequence search that enables using protein sequence in the searching box. PDBeFold is a tool that finds similar structures starting from a PDB accession code as a query entry or via uploading a PDB file. To address the challenge of slow networking time, PDBe has developed a customized server named CoordinateServer that enables extraction of specific data for a given structure providing an advantage of high-speed exploration of PDB files and reduces the limitation of network file transfers. The server can provide several types of data extraction options such as finding residues interacting with a certain ligand and others. The server can be accessed via the link www.ebi.ac.uk/pdbe/coordinates/. PDBe has developed its own molecular visualization software LiteMole 3D viewer. The tool is



2bh9 X-RAY STRUCTURE OF A DELETION VARIANT OF HUMAN GLUCOSE 6-PHOSPHATE DEHYDROGENASE COMPLEXED WITH STRUCTURAL AND COENZYME NADP

X-ray diffraction
2.5Å resolution
Released: 25 Apr 2005

Gover S, Vandeputte-Rutten L, Au SWN, Adams MJ
Acta Crystallogr. D Biol. Crystallogr. (2005) [PMID: 15858258]

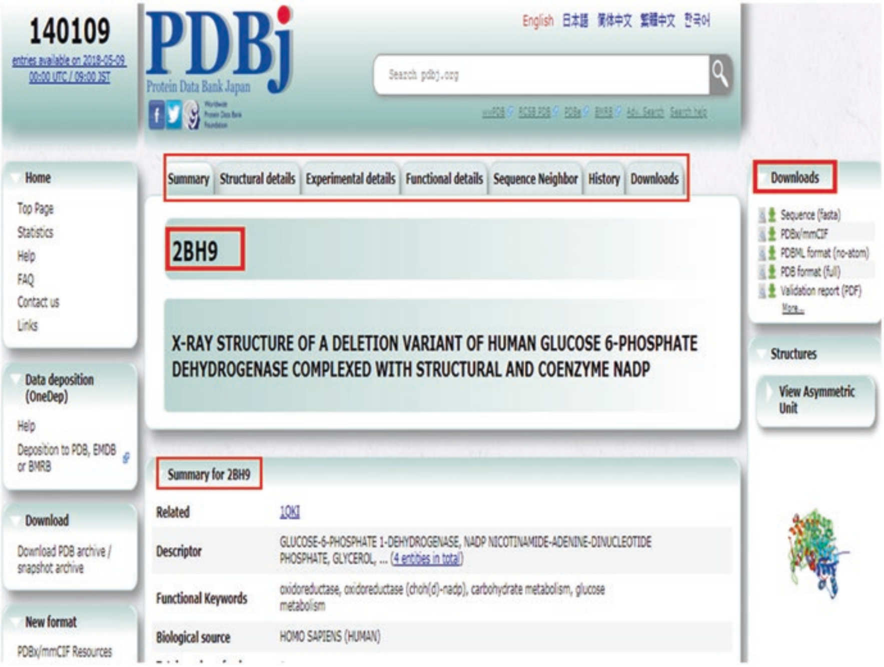
Source organism: Homo sapiens
Assembly composition: protein only structure

Interacting compounds: GOL NAP

Fig. 4.3 A screenshot of Protein Data Bank in Europe (PDBe) webpage interface of 2BH9 entry; the structure determined by X-ray diffraction technique at a resolution of 2.5 Å. In addition to that, binding ligands NAP (nicotinamide adenine dinucleotide phosphate) and GOL (glycerol); literature; the source organism – *Homo sapiens* – and protein assembly composition are also provided besides other buttons to access a plethora of information about the entry, e.g., binding ligands and protein family

compatible with many Internet browsers. The tool is WebGL-based viewer with too little memory footprint. PDBe has also developed a server that enables users to take part in developing their own search queries to meet their needs. The server is known as RESTful application programming interface (API) (Representational State Transfer) and is accessed via pdbe.org/api. Figure 4.3 shows a screenshot of PDBe webpage presenting search results for the entry 2BH9 including the experimental methods, the source organism, the assembly composition, and the interacting compounds (ligands). Additional details are available via other tabs such as macromolecules, compounds, and protein families.

PDBj The Protein Data Bank Japan (https://pdbj.org/), is one of the consortium members of wwPDB; the database is continuously updated to meet the user requirements with a focus on Asian and Middle Eastern users. The database offers a bunch of tools and services that assist the analysis and interpretation of structural data. These services include PDB deposition via an updated tool that supports X-ray, NMR, and EM structures. Group deposition is also available where a group ID is given to a set of structures that are related to each other and have been deposited at the same time. PDBj also provides a tool for easy exploration of the PDB files via PDBj Mine, a tool for searching PDB using either accession codes, keywords, or via the advanced search function. Sequence-based structural alignment is also available via the tool known as SeSAW. The tool allows annotation of the conserved sequences and structural motifs found in the query proteins. eF-seek is a relatively new tool at PDBj that searches similar PDB files with a focus on the ligand binding sites. Omokage is another web-based tool for searching three-dimensional density maps and atomic models, with a focus on global shape similarities. ProMode Elastic database allows inspection of the PDB files regarding the dynamic rather than the static status. The database provides dynamic analysis for the PDB structures, and animations can be generated for PDB structure. PDBj has also developed its own molecular visualization graphic software known as Molmil that enables fast and enhanced graphics and is compatible with JavaScript and WebGL. Figure 4.4 shows a screenshot of PDBj showing summary for the entry 2BH9 including information about the related 3D structure 1QKI, functional keywords, and biological source; also other buttons are found for structural details, experimental details, functional details, sequence neighbor, history, and downloads. In the right side, download format options are available and structure view asymmetric unit.



2BH9

X-RAY STRUCTURE OF A DELETION VARIANT OF HUMAN GLUCOSE 6-PHOSPHATE DEHYDROGENASE COMPLEXED WITH STRUCTURAL AND COENZYME NADP

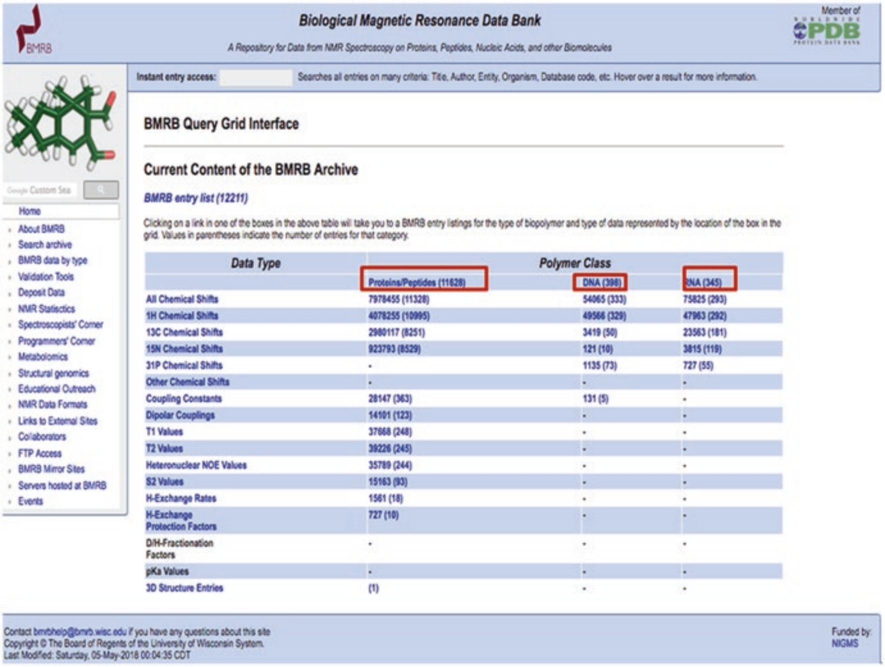
Summary for 2BH9

Related	1QKI
Descriptor	GLUCOSE-6-PHOSPHATE 1-DEHYDROGENASE, NADP NICOTINAMIDE-ADENINE-DINUCLEOTIDE PHOSPHATE, GLYCEROL, ... (4 entities in total)
Functional Keywords	oxidoreductase, oxidoreductase (choh(d)-nadp), carbohydrate metabolism, glucose metabolism
Biological source	HOMO SAPIENS (HUMAN)

Fig. 4.4 A screenshot of Protein Data Bank Japan (PDBj) webpage interface shows detailed informative data about the entry 2BH9, represented in the main navigation menu containing many buttons which provide information about the entry's summary, structural details, experimental details, and functional details. Moreover, it also provides different types of downloadable file formats such as FASTA sequence, PDB format, PDBx/mmCIF file formats, and others

BMRB Biological Magnetic Resonance Data Bank, aims to archive and annotate the nuclear magnetic resonance data obtained from macromolecules and their metabolites. The database is unique and provides an important repository for NMR data for peptides, proteins, and nucleic acids. The current content (May 2018) of BMRB archive includes 11,628 entries of proteins/peptides, 398 entries of DNA, and 345 entries of RNA (Fig. 4.5). BMRB can be accessed via the URL http://www.bmrb.wisc.edu/, which is sponsored by the University of Wisconsin-Madison, the National Library of Medicine, and National Institutes of Health. The website is organized into different tabs such as search archive, validation tools, deposit data, NMR statistics, programmers' corner, spectroscopists' corner, educational outreach, etc. (Ulrich et al. 2007).

NCBI Structure Resources The NCBI devotes one of its databases to the structure information. NCBI provides ENTREZ search function that allows searching keywords all over its databases including the structure database. The structure database is available in the link https://www.ncbi.nlm.nih.gov/structure/, accessed on March 2018. Figure 4.6 is a screenshot of structure summary MMDB webpage using the PDB ID 2BH9 (MMDB ID 33089) as an example. The page displays information about the experimental method, resolution, source organism, similar structures, and biological unit (molecular graphic, interactions) for 2BH9.



Biological Magnetic Resonance Data Bank

A Repository for Data from NMR Spectroscopy on Proteins, Peptides, Nucleic Acids, and other Biomolecules

Instant entry access: Searches all entries on many criteria: Title, Author, Entity, Organism, Database code, etc. Hover over a result for more information.

BMRB Query Grid Interface

Current Content of the BMRB Archive

BMRB entry list (12211)

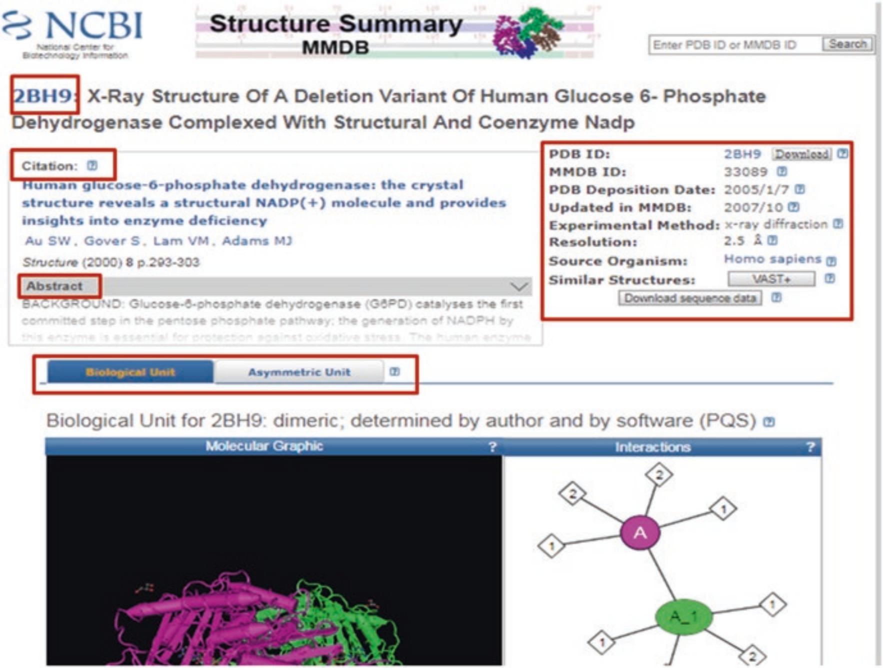
Clicking on a link in one of the boxes in the above table will take you to a BMRB entry listings for the type of biopolymer and type of data represented by the location of the box in the grid. Values in parentheses indicate the number of entries for that category.

Data Type	Polymer Class		
	Proteins/Peptides (11628)	DNA (398)	RNA (345)
All Chemical Shifts	7978455 (11328)	54065 (333)	75825 (293)
1H Chemical Shifts	4078255 (10995)	49566 (329)	47963 (292)
13C Chemical Shifts	2980117 (8251)	3419 (50)	23563 (181)
15N Chemical Shifts	923793 (8529)	121 (10)	3815 (119)
31P Chemical Shifts	-	1135 (73)	727 (55)
Other Chemical Shifts	-	-	-
Coupling Constants	28147 (363)	131 (5)	-
Dipolar Couplings	14101 (123)	-	-
T1 Values	37668 (248)	-	-
T2 Values	39226 (245)	-	-
Heteronuclear NOE Values	35789 (244)	-	-
S2 Values	15163 (93)	-	-
H-Exchange Rates	1561 (18)	-	-
H-Exchange Protection Factors	727 (10)	-	-
D/H-Fractionation Factors	-	-	-
pKa Values	-	-	-
3D Structure Entries	(1)	-	-

Contact bmrbhelp@bmrb.wisc.edu if you have any questions about this site
Copyright © The Board of Regents of the University of Wisconsin System.
Last Modified: Saturday, 05-May-2018 00:04:35 CDT

Funded by:
NIGMS

Fig. 4.5 A screenshot of Biological Magnetic Resonance Data Bank (BMRB) webpage interface shows the recent content of the three major classes of biomacromolecules' structures, determined by nuclear magnetic resonance spectroscopy, 11,628 protein/peptide entries, 398 DNA entries, and 345 RNA entries, and the derived information: coupling constants, chemical shifts, dipolar coupling, etc



2BH9: X-Ray Structure Of A Deletion Variant Of Human Glucose 6- Phosphate Dehydrogenase Complexed With Structural And Coenzyme Nadp

Citation:

Human glucose-6-phosphate dehydrogenase: the crystal structure reveals a structural NADP(+) molecule and provides insights into enzyme deficiency

Au SW, Gover S, Lam VM, Adams MJ

Structure (2000) 8 p.293-303

Abstract

BACKGROUND: Glucose-6-phosphate dehydrogenase (G6PD) catalyses the first committed step in the pentose phosphate pathway; the generation of NADPH by this enzyme is essential for protection against oxidative stress. The human enzyme

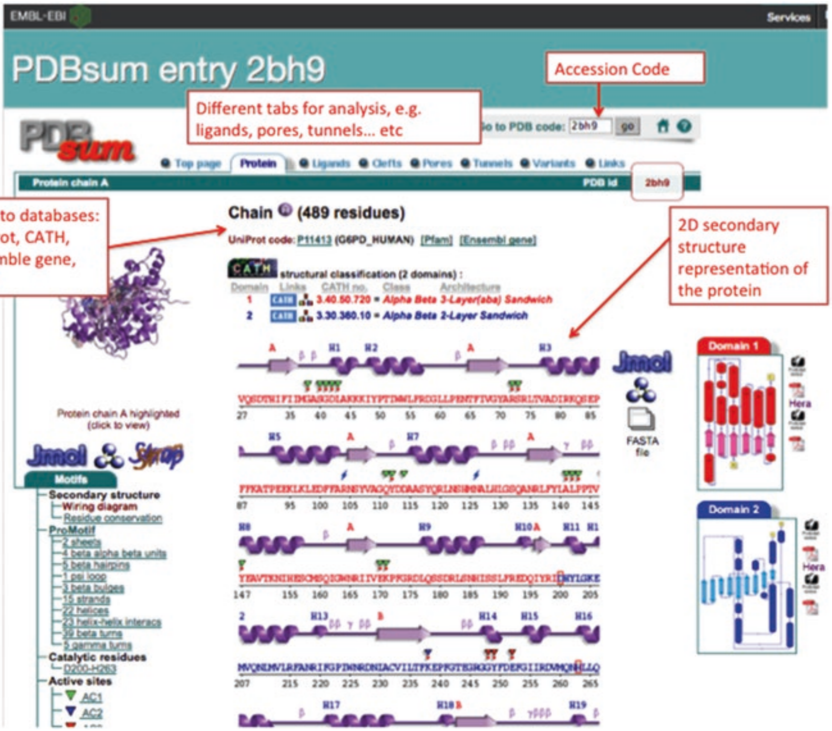
PDB ID: 2BH9
MMDB ID: 33089
PDB Deposition Date: 2005/1/7
Updated in MMDB: 2007/10
Experimental Method: x-ray diffraction
Resolution: 2.5 Å
Source Organism: Homo sapiens
Similar Structures:

Biological Unit for 2BH9: dimeric; determined by author and by software (PQS)

Fig. 4.6 A screenshot of Molecular Modeling Database (MMDB) webpage interface of 2BH9, MMDB ID (33089). The structure is resolved by X-ray diffraction technique at a resolution of 2.5 Å, and the source organism is *Homo sapiens*. Besides, it provides a chemical graph, links to literature, and compact structures (3D structure domains) that help with identifying similar structures

4.3 PDBsum: Structural Summaries of PDB Entries

PDBsum available at https://www.ebi.ac.uk/pdbsum is an atlas of proteins and web server that helps to present the PDB entries in a visualized form. It was developed at the University College London (UCL) in 1995 and is aimed to provide a largely graphic compendium of the proteins and their complexes (Laskowski 2007; Babajan et al. 2011). The server can be accessed freely and is maintained by Laskowski and collaborators at the European Bioinformatics Institute (EBI) (Laskowski et al. 2018). PDBsum provides many different analytical tools for the content of the protein structure including the ligand interaction, protein-protein interaction, and CATH classification. The 3D structures are viewed interactively in PyMOL and RasMol, and users have the ability to upload their own PDB files – could be a homology model – and get them analyzed. Figure 4.7 illustrates some of the pictorial analyses presented by PDBsum. The example given is for PDB entry 2BH9 (G6PD-human) solved by X-ray crystallography at 2.5 Å resolution. The page shows different sections among which the 3D structures are presented interactively using molecular visualization JavaScript viewer called 3Dmol.js. This generated



Accession Code

Different tabs for analysis, e.g. ligands, pores, tunnels… etc

Links to databases: Uniprot, CATH, Ensemble gene, Pfam

2D secondary structure representation of the protein

PDBsum entry 2bh9

Go to PDB code: 2bh9

Chain A (489 residues)

UniProt code: P11413 (G6PD_HUMAN) [Pfam] [Ensembl gene]

CATH structural classification (2 domains) :

Domain	Links	CATH no.	Class	Architecture
1	CATH	3.40.50.720 =	Alpha Beta	3-Layer(aba) Sandwich
2	CATH	3.30.360.10 =	Alpha Beta	2-Layer Sandwich

Protein chain A highlighted (click to view)

- Secondary structure
  - Wiring diagram
  - Residue conservation
- ProMotif
  - 2 sheets
  - 4 beta alpha beta units
  - 5 beta hairpins
  - 1 psi loop
  - 3 beta bulges
  - 15 strands
  - 22 helices
  - 23 helix-helix interacs
  - 39 beta turns
  - 5 gamma turns
- Catalytic residues
  - D200-H263
- Active sites
  - ▼ AC1
  - ▼ AC2

Fig. 4.7 A screenshot of PDBsum webpage interface of 2BH9 entry; a 2D secondary structure representation is shown in the figure. Tabs for protein-protein interactions, ligands, pores, tunnels, and others are seen in the figure. Hyperlinks to r databases like UniProt, Pfam, and Ensembl gene are also provided

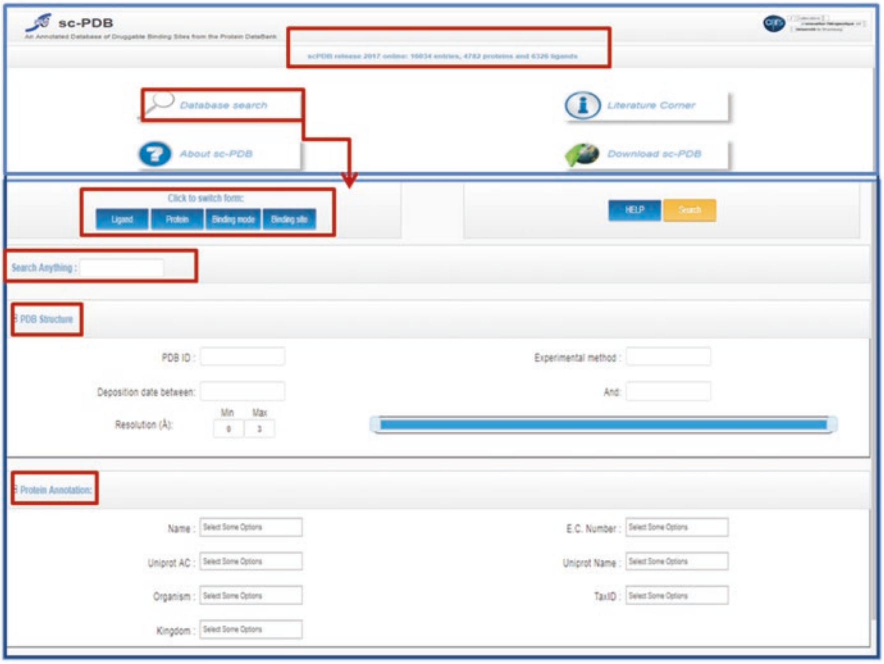
image automatically gives only a rough idea of the sizes and locations of the clefts. Using the RasMol or Jmol options on the clefts tab, an idea about the clefts found in the structure can be obtained. PDBsum webpage also hosts useful links to databases and servers such as:

1. EC-PDB, Enzyme Structure Database, database includes approximately 73,000 PDB enzyme structures. The database classifies the entries according to the Enzyme Commission (EC) as EC1, EC2, EC3, EC4, EC5, and EC6. EC3 – the hydrolase family – is the highest represented family among others in this database including over 27,000 PDB structures.
2. Drug port is the second server which identifies all “drug targets” in the PDB and any drug fragments that exist as ligands in PDB structures. The server lists all the drugs in alphabetical order; therefore, for example, if you are looking for acetaminophen, you will find it under the alphabet A in the list, and visiting its page will show the information of the protein targets of this specific drug and hyperlinks to other related resources such as DrugBank and others.

3. ProFunc server: the server aims to help in the identification of protein of related biochemical function based on the 3D structure. The algorithm of ProFunc uses information such as the active site, fold matching, residue conservation, and surface analysis to do the task. The server allows to look for existing PDB file or to upload custom PDB file (Laskowski et al. 2018).
4. SAS, sequence annotated by the structure, is a tool by PDBsum; the tool allows multiple sequence alignment of a query protein that entered in different forms such as FASTA sequence, PDB accession code, PDB file, or UniProt accession code. The obtained multiple alignments can be color-coded according to different criteria, such as the secondary structure assignment, ligand binding site, and number of hydrogen bonds to ligands or residue similarity. The alignment can be adjusted according to the user needs using selection and sequence similarity filters.

4.4 sc-PDB: A 3D Database of Ligandable Binding Sites

The protein-ligand interaction is very important in determining the critical amino acids in the protein structure that interact with ligands, and based on this information, designing new ligands (drugs) is possible. The sc-PDB database archives and illustrates the ligandable binding sites found in protein structures that are listed in the PDB repository. The database was launched in 2004 and is accessible at http://bioinfopharma.u-strasbg.fr/scPDB/. The Sc-PDB provides specialized structure files that serve the need to do receptor-ligand docking studies. Currently, the sc-PDB stores 16,034 entries (binding sites) extracted from 4782 unique proteins and 6326 exclusive ligands. The sc-PDB database provides annotated druggable binding sites, the coordinates for protein-ligand complexes, and the physicochemical and geometrical properties of the ligands. It also provides a chemical description of ligands and functional explanation of the proteins. Metal ions are not included in sc-PDB, and the ligands included are classified into four main categories: (i) nucleotides of size <4 bases, (ii) peptides <9 amino acids, (iii) cofactors, and (iv) organic compounds. The binding site can be defined as the protein residues (including amino acids, cofactors, and important metal ions) that are in contact with one atom of the ligand within a distance of 6.5 Å. The sc-PDB is very useful in drug design tasks since it can predict receptors for any ligand and it can analyze different structural cavities and establish the interacting points between a ligand and the active site of the receptor (Desaphy et al. 2014; Kellenberger et al. 2006). Ligands can be searched using the chemical structure draw applet provided by ChemAxon. Figure 4.8 is a screenshot of the sc-PDB webpage showing the total number of entries (16034) including 4782 proteins and 6326 ligands. The database home page shows four buttons: ligand, protein, binding mode, and binding site. The database archive can be searched using the search anything box, PDB ID box, or protein UniProt accession code.


An Annotated Database of Druggable Binding Sites from the Protein DataBank

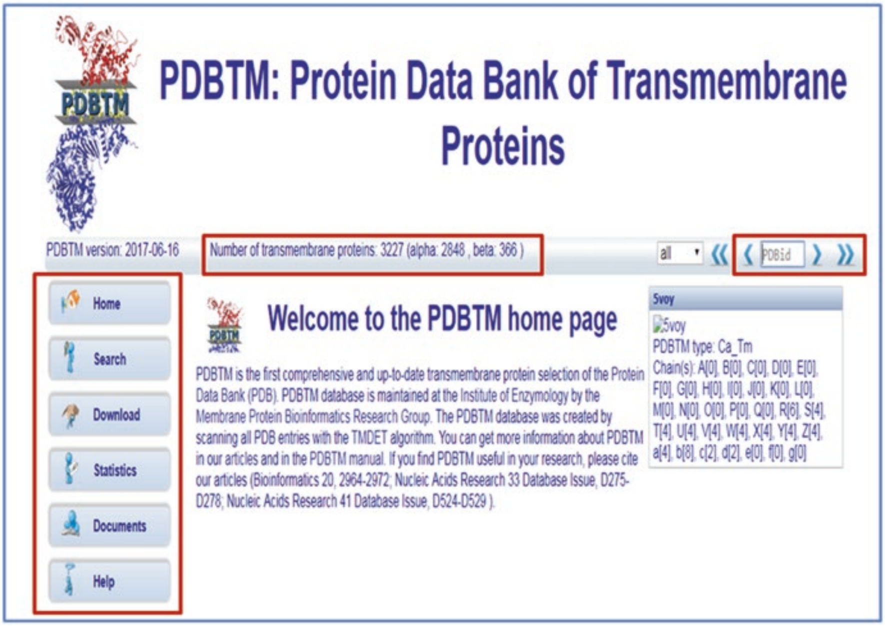
scPDB release 2017 online: 16034 entries, 4782 proteins and 6326 ligands

Resolution (Å): Min 0 Max 3

Fig. 4.8 A screenshot of sc-PDB webpage interface. It shows (16034) three-dimensional structures of binding sites found in the Protein Data Bank (PDB) and includes (4782) unique proteins and (6326) unique ligands. In addition, it provides the main navigation window for the user to navigate and switch views directly (ligand, protein, binding mode, and binding site)

4.5 PDBTM: Protein Data Bank of Transmembrane Proteins

Membrane proteins account for 20–30% of the all human proteins which participate in vital cellular processes and enzymatic reactions. Membrane proteins represent 60% of all druggable proteins in human (Yin and Flynn 2016). The experimental 3D structure determination of these proteins is difficult due to the complexity of obtaining soluble expressed proteins. Since the publication of the first membrane protein 3D structure in 1985, the number of membrane proteins in wwPDB is increasing slowly but steadily. Still, the current representation of the membrane proteins in PDB is low. There was a need to have specialized databases for membrane proteins. The PDBTM database is the first up-to-date and inclusive TM protein consisting of a list of PDB files of transmembrane proteins (Kozma et al. 2012). The database was launched in 2004 and is available at http://pdbtm.enzim.hu; PDBTM archives more than 3000 transmembrane proteins; most of them have the well-known alpha-helical structures. PDBTM is utilizing a special algorithm named TMDET to find transmembrane proteins found in the PDB based on the structural information. The algorithm is also able to determine the spatial arrangement of these proteins inside the lipid bilayer. PDBTM


PDBTM: Protein Data Bank of Transmembrane Proteins

PDBTM version: 2017-06-16

Number of transmembrane proteins: 3227 (alpha: 2848 , beta: 366)

Welcome to the PDBTM home page

PDBTM is the first comprehensive and up-to-date transmembrane protein selection of the Protein Data Bank (PDB). PDBTM database is maintained at the Institute of Enzymology by the Membrane Protein Bioinformatics Research Group. The PDBTM database was created by scanning all PDB entries with the TMDET algorithm. You can get more information about PDBTM in our articles and in the PDBTM manual. If you find PDBTM useful in your research, please cite our articles (Bioinformatics 20, 2964-2972; Nucleic Acids Research 33 Database Issue, D275-D278; Nucleic Acids Research 41 Database Issue, D524-D529).

5voy

5voy

PDBTM type: Ca_Tm

Chain(s): A[0], B[0], C[0], D[0], E[0], F[0], G[0], H[0], I[0], J[0], K[0], L[0], M[0], N[0], O[0], P[0], Q[0], R[6], S[4], T[4], U[4], V[4], W[4], X[4], Y[4], Z[4], a[4], b[8], c[2], d[2], e[0], f[0], g[0]

Fig. 4.9 A screenshot of Protein Data Bank of Transmembrane Protein (PDBTM). It shows a number of transmembrane proteins deposited in PDBTM; total number is 3227 entries: 2848 alpha structured and 366 beta structured

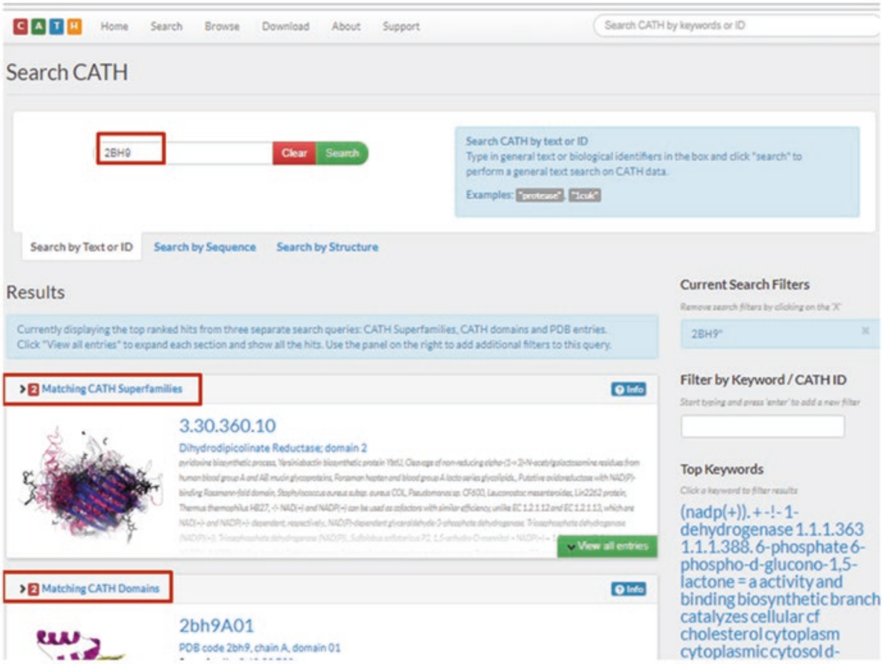
website allows to browse its content by the type of the membranes (alpha or beta structures), and it also permits to download datasets of TM protein structures. Figure 4.9 shows the home page of PDBTM and the number of transmembrane proteins that is archived until May 2018 (a total of 3227, including 2848 alpha structure and 366 beta structure). The search field using PDB ID exists in the right side, while the left side includes six vertical tabs (home, search, download, statistics, documents, and help).

4.6 CATH Database

CATH (Class, Architecture, Topology, Homology) database classifies the protein domains according to the amino acid sequence and the structural and the functional properties. CATH provides a big deal of help for researchers with proteins that have insignificant similarity in sequences yet can be functionally and structurally related. CATH is also a valuable destination for both bioinformatician and biologists. Inexperienced users benefit from the user-friendly web interface; on the other hand, bioinformaticians seeking for analysis of a huge number of domains can find complete downloadable datasets. Therefore CATH has the potentials to be a really valuable and promising recourse. In CATH, domains are classified hierarchically into four levels named as class (C), architecture (A), topology (T), and homologous superfamily (H), hence giving the acronym CATH (Knudsen and Wiuf 2010):

(i) C level: categorize domains into four main groups according to secondary structures as alpha mainly, beta mainly, α-β mixed, and finally category group domains with few alpha and beta structures.
(ii) A level: categorize domains by the general orientation of the secondary structures.
(iii) T level: categorization depends upon the connectivity of secondary structures.
(iv) H level: categorization depends upon a combination of sequence similarity and structural similarity.

Exploration of the contents of the databases can also be done via different links given in the web server, for example, (1) searching by domain ID or keywords, (2) searching by the sequence in FASTA format, and (3) exploring the database from the hierarchy top and download datasets. A list encompasses the names of all domains in CATH – along with their individual groupings – which is likewise accessible, and the amino acid sequences of all domains ordered in CATH are open for download in the FASTA file format (Knudsen and Wiuf 2010). Figure 4.10 is the search results for the PDB ID (2BH9); the figure shows the matching CATH superfamilies and the matching CATH domains.



Search CATH

2BH9

Search CATH by text or ID
Type in general text or biological identifiers in the box and click "search" to perform a general text search on CATH data.
Examples: "protease" "1cuk"

Results

Currently displaying the top ranked hits from three separate search queries: CATH Superfamilies, CATH domains and PDB entries.
Click "View all entries" to expand each section and show all the hits. Use the panel on the right to add additional filters to this query.

2 Matching CATH Superfamilies

3.30.360.10
Dihydrodipicolinate Reductase; domain 2
pyridoxine biosynthetic process, Yersiniabactin biosynthetic protein YbtU, Cleavage of non-reducing alpha-(1->3)-N-acetylgalactosamine residues from human blood group A and AB mucin glycoproteins, Forssman hapten and blood group A lacto series glycolipids., Putative oxidoreductase with NAD(P)-binding Rossmann-fold domain, Staphylococcus aureus subsp. aureus COL, Pseudomonas sp. CF600, Leuconostoc mesenteroides, Lin2262 protein, Thermus thermophilus HB27, -!- NAD(+) and NADP(+) can be used as cofactors with similar efficiency, unlike EC 1.2.1.12 and EC 1.2.1.13, which are NAD(+)- and NADP(+)- dependent, respectively, NAD(P)-dependent glyceraldehyde-3-phosphate dehydrogenase, Triosephosphate dehydrogenase (NAD(P))-1, Triosephosphate dehydrogenase (NAD(P)), Sulfolobus solfataricus P2, 1,5-anhydro-D-mannitol + NADP(+) + 1

2 Matching CATH Domains

2bh9A01
PDB code 2bh9, chain A, domain 01

Current Search Filters
Remove search filters by clicking on the 'X'
2BH9* ×

Filter by Keyword / CATH ID
Start typing and press 'enter' to add a new filter

Top Keywords
Click a keyword to filter results
(nadp(+)). + -!- 1- dehydrogenase 1.1.1.363 1.1.1.388. 6-phosphate 6-phospho-d-glucono-1,5-lactone = a activity and binding biosynthetic branch catalyzes cellular cf cholesterol cytoplasm cytoplasmic cytosol d-

Fig. 4.10 A screenshot of CATH/Gene3D webpage interface of the entry 2BH9. The websites provide different ways of search: text or ID, search by sequence, or search by the structure. In the current example, the screenshot shows the matching CATH superfamilies and domains related to 2BH9

CATH/Gene3D database is complementary to the original CATH database; it is available at http://www.cathdb.info/; it classifies 95 million protein domains into 6119 superfamilies (Dawson et al. 2016). CATH/Gene3D scans the protein sequence information found in UniProt database; it also classifies the structural domains found in the structural files in wwPDB. Annotation of the structure is created using hidden Markov models making use of the domain families deposited in CATH. Moreover, all information is downloadable in an XML file format, enabling users to perform a complex search at their computers (Yeats et al. 2006). Furthermore, Gene3D exploits the data in CATH to predict the position of structural domains on a host of protein sequences available at wwPDB which allows inclusion of informative annotations such as information, function, and residues of the active site. It also provides a broad prediction of globular domains in proteins (Dawson et al. 2016; Dawson et al. 2017).

4.7 SCOP (Structural Classification of Proteins) Database

Structural Classification of Protein (SCOP), available at http://scop.mrc-lmb.cam.ac.uk/scop/, is a database with a focus on structure and evolutionary classifications of proteins. SCOP adopts the following hierarchical scheme to classify protein structures:

A. Family: similar protein structures are assembled into families based on two criteria that suggest a common evolutionary source; the first criterion is a similarity in protein sequence, and the second criterion is a similarity in structure and function.
B. Superfamily: families whose proteins have little sequence similarity yet their function and structure imply typical evolutionary origin are clustered together in superfamilies.
C. Common fold: protein families and subfamilies that have similar secondary structures and same topological associations are assigned to have a common fold.
D. Class: the distinctive folds have been gathered into classes.

The majority of the folds are grouped into one of the five structural classes:

1. All –α: structures that are basically formed of α-helices.
2. All –β: structures that are basically formed of β-sheets.
3. α/β: structures formed of α-helices and β-strands.
4. α + β: structures formed of α-helices and β-strands are to a great extent segregated.
5. Multi-domain: structures with domains of various classes and for which no homologs are yet known.

SCOP is updated into the new version SCOP2, where improvements in the classification criteria were done. SCOP2 classification is based on four criteria, i.e., the protein type, the evolutionary analysis, the structure class, and the protein relationships. The protein types indicate four possible types of proteins, i.e., membrane, soluble, fibrous, and intrinsically disordered proteins. The evolutionary analysis considers the classification of proteins according to the major evolutionary events that had have happened to certain protein class. The third criterion is the secondary structure arrangement of the protein as an efficient way in the classification of protein structures. The protein relationships are unique to SCOP2 compared to SCOP. The database is accessible via the link http://scop2.mrc-lmb.cam.ac.uk/. SCOP2 can be explored in two different ways: SCOP2-graph and SCOP2-browser. SCOP2-graph shows graphical representation for the database entries, while SCOP2-browser allows the exploration of the SCOP2 contents according to the four classification criteria mentioned above in addition to a possibility of keyword search. The SCOP2 additionally provides hyperlinks whenever possible to each entry archived to the external databases such as UniProt and PDB and the original SCOP record (Andreeva et al. 2007; Hubbard et al. 1997). Figure 4.11 is a screenshot of SCOP2-graph database webpage interface. It illustrates a hierarchal classification of protein domains by the structure and evolutionary relevance.

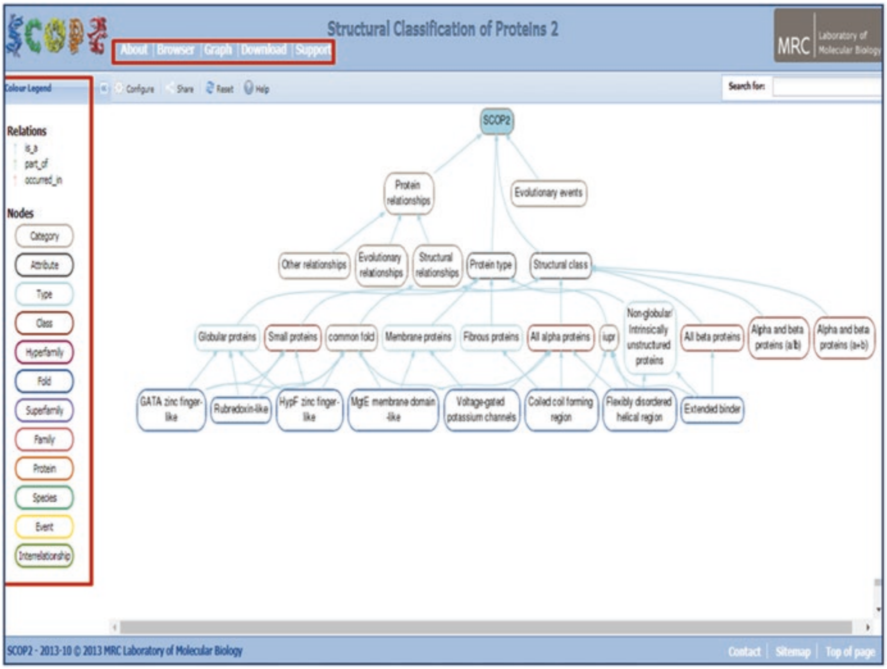


Fig. 4.11 A screenshot of SCOP2-graph database webpage interface. It illustrates a hierarchal classification of protein domains in accordance with the structure and evolutionary relevance; these relationships appear as compound node networks; also, it provides accessible links to the SCOP entries and hence provides a possibility for the users to compare both databases

4.8 Structure Comparison Servers

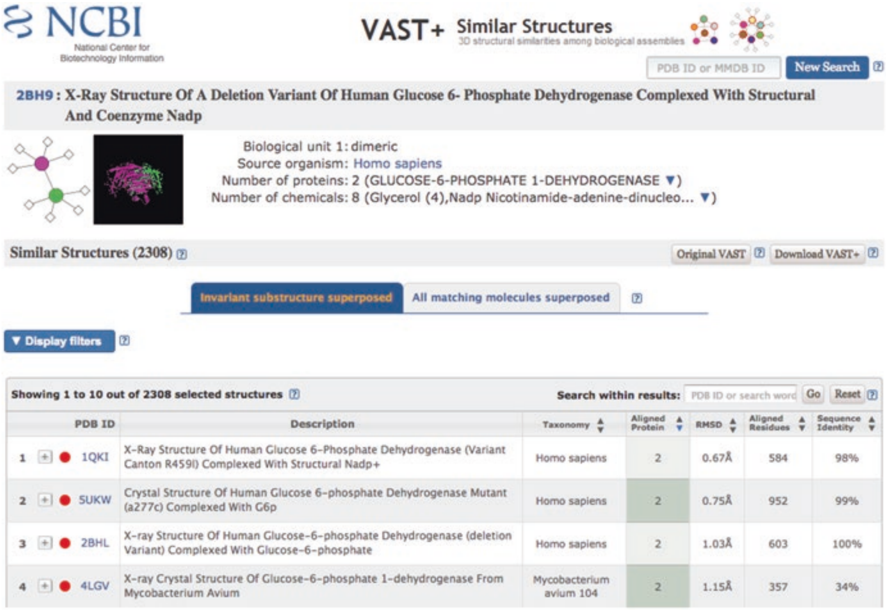
Finding homologous protein structure is very important in the area of structural bioinformatics. Therefore early efforts were carried out to device algorithms for structural alignment; in 1960, Perutz et al. described the structure similarity of hemoglobin and myoglobin (Perutz et al. 1960). It is known that protein structures are more conserved compared to protein sequences; this is the base of evolutionary analysis of related protein structures. It is important to differentiate between two terms, i.e., structure superposition and structure alignment. Structure superposition refers to the spatial fitting of two structures that already have similar starting points – usually in the C-alpha backbone – which work as guiding points in the process of fitting these two structures over each other. The aim is to find the best match between the two structures as judged by the root-mean-square deviation (RMSD) value. RMSD is a measure of the average distance between atoms of two or more superimposed protein structures and is measured in angstrom. Structure alignment does not require prior information of equivalent spatial positions of two structures. However, the alignment algorithm tries to find structures between two 3D structures or more based on the 3D information. There are few clear reasons behind the effort for finding similar protein structures:

1. To help in structure classification and fold assignment
2. To aid the process of function identification, since similar protein structures can provide a wealth of information about the function of an unknown protein
3. To aid, in the process of homology or comparative modeling, the process of predicting protein 3D structure based on similarity to already known 3D structure
4. To aid in the tasks of protein engineering (Gaber 2016; Pavelka et al. 2009)

CATH and SCOP databases were used in the endeavors of finding similar structures based on detection of similar structural domains. Currently, some online servers and tools are used in finding homologous 3D structures of proteins; among these servers are:

1. Combinatorial Extension (CE) is a tool for aligning and comparing protein structures deposited into RCSB PDB (Shindyalov and Bourne 1998). CE is an indispensable part of identifying and annotating protein structures with unknown function. The comparison can be performed on a complete PDB or on structurally representative subsets of proteins. Also, it can be performed in two ways either using a structural representative subset of protein or on the full PDB records. The most direct task is to locate every single similar structure to a starting protein that exceeds 30 residues long and exists in the wwPDB. The superimposed structures can be visualized with programs such as RasMol and Protein Explorer (utilizing Chime) or in an exceptionally outlined Java applet Compare3D. The applet enables the user to investigate the two similarities and differences between the aligned structures both from a sequence and structure viewpoint. It is worth mentioning that the site is always subjected to modification and editing by the Bourne Laboratory staff to keep it up to date (Shindyalov and Bourne 2001).

2. PDBeFold is an online server that is provided by EMBL-EBI (European Molecular Biology Laboratory-European Bioinformatics Institute). PDBeFold can be accessed from the PDBe webpage and is considered a structure alignment server that allows both pairwise or multiple 3D alignments. Searching homologous structures can be initiated by providing the PDB accession code.
3. VAST+ is online server hosted by NCBI and is devoted to finding similar 3D structures; the server does not rely on sequence comparison; hence it can find 3D structures of too low sequence similarity. Figure 4.12 shows the interface of VAST+ using the PDB entry 2BH9.
4. DALI web server was established in 2000 at Helsinki Lab; the server aims to compare 3D structures of proteins to those found in the Protein Data Bank. A new version of DALI known as DALI Lite has been released to do pairwise structural superimposition. Figure 4.13 shows a screenshot of DALI structure comparison server exemplified by a search using the entry 2BH9. DALI website displays nine horizontal tabs as follows: about, PDB search, PDB25, pairwise, all against all, gallery, references, statistics, and tutorial.


NCBI
National Center for Biotechnology Information
VAST+ Similar Structures
3D structural similarities among biological assemblies
PDB ID or MMDB ID
New Search

2BH9 : X-Ray Structure Of A Deletion Variant Of Human Glucose 6- Phosphate Dehydrogenase Complexed With Structural And Coenzyme Nadp

Biological unit 1: dimeric
Source organism: Homo sapiens
Number of proteins: 2 (GLUCOSE-6-PHOSPHATE 1-DEHYDROGENASE ▼)
Number of chemicals: 8 (Glycerol (4),Nadp Nicotinamide-adenine-dinucleo... ▼)

Similar Structures (2308)

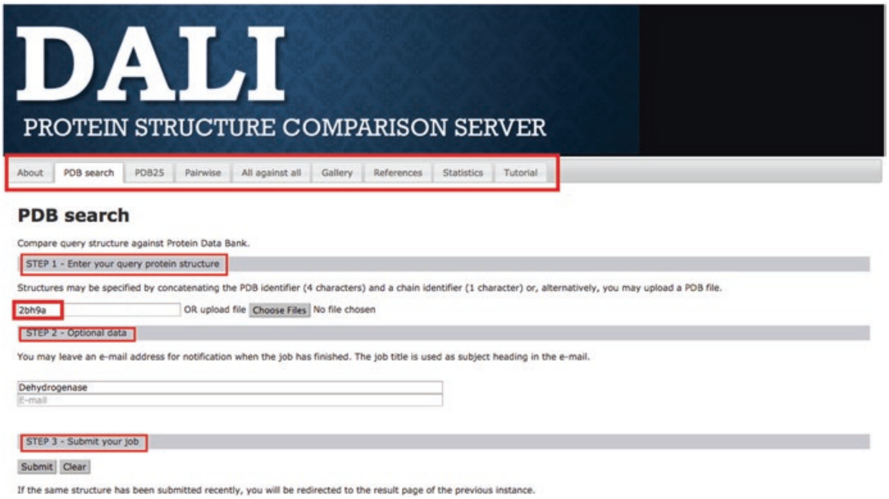
Original VAST
Download VAST+
Invariant substructure superposed
All matching molecules superposed
▼ Display filters

Showing 1 to 10 out of 2308 selected structures

Search within results: PDB ID or search word Go Reset

	PDB ID	Description	Taxonomy	Aligned Protein	RMSD	Aligned Residues	Sequence Identity
1	● 1QKI	X-Ray Structure Of Human Glucose 6-Phosphate Dehydrogenase (Variant Canton R459l) Complexed With Structural Nadp+	Homo sapiens	2	0.67Å	584	98%
2	● 5UKW	Crystal Structure Of Human Glucose 6-phosphate Dehydrogenase Mutant (a277c) Complexed With G6p	Homo sapiens	2	0.75Å	952	99%
3	● 2BHL	X-ray Structure Of Human Glucose-6-phosphate Dehydrogenase (deletion Variant) Complexed With Glucose-6-phosphate	Homo sapiens	2	1.03Å	603	100%
4	● 4LGV	X-ray Crystal Structure Of Glucose-6-phosphate 1-dehydrogenase From Mycobacterium Avium	Mycobacterium avium 104	2	1.15Å	357	34%

Fig. 4.12 A screenshot of VAST+ webpage interface of the entry 2BH9. It provides information about macromolecules that share similar three-dimensional structures. Concerning 2BH9, there is 2308 structure similar to it. It is worth noting that filters can be used to limit the number of matching molecules at will. The RMSD values shown indicate the structural similarity between the query 2BH9 and the retrieved hits; lower RMSD values indicate high structural similarity



PDB search

Compare query structure against Protein Data Bank.

STEP 1 - Enter your query protein structure

Structures may be specified by concatenating the PDB identifier (4 characters) and a chain identifier (1 character) or, alternatively, you may upload a PDB file.

2bh9a OR upload file Choose Files No file chosen

STEP 2 - Optional data

You may leave an e-mail address for notification when the job has finished. The job title is used as subject heading in the e-mail.

Dehydrogenase

E-mail

STEP 3 - Submit your job

If the same structure has been submitted recently, you will be redirected to the result page of the previous instance.

Fig. 4.13 A screenshot of DALI server webpage interface and example input of the entry 2BH9 is shown. The website provides three different types of searches: PDB search, pairwise comparison, and all-against-all comparison which performs a database search comparing a query structure supplied by the user against the database of known structures (PDB) and returns the list of structural neighbors using the e-mail

4.9 Conclusion

Structural databases are providing essential information not only to the scientific community but also to the public. The content of such databases is a really precious information; precious is not just a metaphor; to explain, solving 1000 protein structures costs 150 million USD and the effort of 180 scientists (Ledford 2010). Fortunately, the advancement in the computational sciences allowed structural databases to be explored by both experts and novice users to navigate and easily extract the required information from its content. It is also very feasible to find related contents in the different database based on the interconnectedness between the different databases. The availability of such data allowed new generations of databases to evolve and to provide new layers of information that help in solving serious problems such as designing new drugs or engineering new proteins for different purposes.

References

Acharya C, Kufareva I, Ilatovskiy AV, Abagyan R (2014) PeptiSite: a structural database of peptide binding sites in 4D. Biochem Biophys Res Commun 445(4):717–723

An J, Totrov M, Abagyan R (2005) Pocketome via comprehensive identification and classification of ligand binding envelopes. Mol Cell Proteomics 4(6):752–761

Andreeva A, Howorth D, Chandonia J-M, Brenner SE, Hubbard TJ, Chothia C, Murzin AG (2007) Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res 36(suppl_1):D419–D425

Babajan B, Chaitanya M, Rajsekhar C, Gowsia D, Madhusudhana P, Naveen M et al (2011) Comprehensive structural and functional characterization of Mycobacterium tuberculosis UDP-NAG enolpyruvyl transferase (Mtb-MurA) and prediction of its accurate binding affinities with inhibitors. Interdiscip Sci 3(3):204–216. https://doi.org/10.1007/s12539-011-0100-y

Bagchi A (2012) A brief overview of a few popular and important protein databases. Computat Mol Biosci 2(04):115

Berman HM, Bhat TN, Bourne PE, Feng Z, Gilliland G, Weissig H, Westbrook J (2000) The Protein Data Bank and the challenge of structural genomics. Nat Struct Mol Biol 7:957–959

Berman HM, Kleywegt GJ, Nakamura H, Markley JL (2012) The Protein Data Bank at 40: reflecting on the past to prepare for the future. Structure 20(3):391–396

Carpenter EP, Beis K, Cameron AD, Iwata S (2008) Overcoming the challenges of membrane protein crystallography. Curr Opin Struct Biol 18(5):581–586

Coimbatore Narayanan B, Westbrook J, Ghosh S, Petrov AI, Sweeney B, Zirbel CL, Leontis NB, Berman HM (2013) The nucleic acid database: new features and capabilities. Nucleic Acids Res 42(D1):D114–D122

Craveur P, Rebehmed J, de Brevern AG (2014) PTM-SD: a database of structurally resolved and annotated posttranslational modifications in proteins. Database:2014

Dawson NL, Lewis TE, Das S, Lees JG, Lee D, Ashford P et al (2016) CATH: an expanded resource to predict protein function through structure and sequence. Nucleic Acids Res 45(D1):D289–D295

Dawson NL, Sillitoe I, Lees JG, Lam SD, Orengo CA (2017) CATH-Gene3d: generation of the resource and its use in obtaining structural and functional annotations for protein sequences. Protein Bioinforma 1558:79–110

DeLano WL (2002) The PyMOL molecular graphics system. http://pymol.org

Desaphy J, Bret G, Rognan D, Kellenberger E (2014) sc-PDB: a 3D-database of ligandable binding sites—10 years on. Nucleic Acids Res 43(D1):D399–D404

Emsley P, Lohkamp B, Scott WG, Cowtan K (2010) Features and development of Coot. Acta Crystallogr D Biol Crystallogr 66(4):486–501

Gaber Y (2016) In-silico smart library design to engineer a xylosetolerant hexokinase variant. Afr J Biotechnol 15(21):910–916

Gaber Y, Mekasha S, Vaaje-Kolstad G, Eijsink VG, Fraaije MW (2016) Characterization of a chitinase from the cellulolytic actinomycete Thermobifida fusca. Biochim Biophys Acta 1864(9):1253–1259

Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, Mutowo P, Atkinson F, Bellis LJ, Cibrián-Uhalte E (2016) The ChEMBL database in 2017. Nucleic Acids Res 45(D1):D945–D954

Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, Turner S, Swainston N, Mendes P, Steinbeck C (2015) ChEBI in 2016: improved services and an expanding collection of metabolites. Nucleic Acids Res 44(D1):D1214–D1219

Holm L, Rosenström P (2010) Dali server: conservation mapping in 3D. Nucleic Acids Res 38(suppl_2):W545–W549

Hubbard TJ, Murzin AG, Brenner SE, Chothia C (1997) SCOP: a structural classification of proteins database. Nucleic Acids Res 25(1):236–239

Jo S, Im W (2012) Glycan fragment database: a database of PDB-based glycan 3D structures. Nucleic Acids Res 41(D1):D470–D474

Joosten RP, Te Beek TA, Krieger E, Hekkelman ML, Hooft RW, Schneider R et al (2010) A series of PDB related databases for everyday needs. Nucleic Acids Res 39(suppl_1):D411–D419

Kellenberger E, Muller P, Schalon C, Bret G, Foata N, Rognan D (2006) sc-PDB: an annotated database of druggable binding sites from the Protein Data Bank. J Chem Inf Model 46(2):717–727

Kinjo AR, Bekker G-J, Suzuki H, Tsuchiya Y, Kawabata T, Ikegawa Y, Nakamura H (2016) Protein Data Bank Japan (PDBj): updated user interfaces, resource description framework, analysis tools for large structures. Nucleic Acids Res 45:D282–D288. https://doi.org/10.1093/nar/gkw962

Knudsen M, Wiuf C (2010) The CATH database. Hum Genomics 4(3):207

Kozma D, Simon I, Tusnady GE (2012) PDBTM: Protein Data Bank of transmembrane proteins after 8 years. Nucleic Acids Res 41(D1):D524–D529

Krieger E, Vriend G (2014) YASARA View—molecular graphics for all devices—from smartphones to workstations. Bioinformatics 30(20):2981–2982

Laskowski RA (2007) Enhancing the functional annotation of PDB structures in PDBsum using key figures extracted from the literature. Bioinformatics 23(14):1824–1827

Laskowski RA, Jabłońska J, Pravda L, Vařeková RS, Thornton JM (2018) PDBsum: structural summaries of PDB entries. Protein Sci 27(1):129–134

Ledford H (2010) Big science: the cancer genome challenge. Nat News 464(7291):972–974

Lo Conte L, Ailey B, Hubbard TJ, Brenner SE, Murzin AG, Chothia C (2000) SCOP: a structural classification of proteins database. Nucleic Acids Res 28(1):257–259

Lobanov MY, Shoemaker BA, Garbuzynskiy SO, Fong JH, Panchenko AR, Galzitskaya OV (2009) ComSin: database of protein structures in bound (complex) and unbound (single) states in relation to their intrinsic disorder. Nucleic Acids Res 38(suppl_1):D283–D287

Madej T, Lanczycki CJ, Zhang D, Thiessen PA, Geer RC, Marchler-Bauer A, Bryant SH (2013) MMDB and VAST+: tracking structural similarities between macromolecular complexes. Nucleic Acids Res 42(D1):D297–D303

Markley JL, Ulrich EL, Berman HM, Henrick K, Nakamura H, Akutsu H (2008) BioMagResBank (BMRB) as a partner in the Worldwide Protein Data Bank (wwPDB): new policies affecting biomolecular NMR depositions. J Biomol NMR 40(3):153–155

Mewes H-W, Frishman D, Güldener U, Mannhaupt G, Mayer K, Mokrejs M, Morgenstern B, Münsterkötter M, Rudd S, Weil B (2002) MIPS: a database for genomes and protein sequences. Nucleic Acids Res 30(1):31–34

Pavelka A, Chovancova E, Damborsky J (2009) HotSpot Wizard: a web server for identification of hot spots in protein engineering. Nucleic Acids Res 37(suppl_2):W376–W383

Perutz MF, Rossmann MG, Cullis AF, Muirhead H, Will G, North A (1960) Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-Å resolution obtained by X-ray analysis. Nature 185(4711):416

Popenda M, Szachniuk M, Blazewicz M, Wasik S, Burke EK, Blazewicz J, Adamiak RW (2010) RNA FRABASE 2.0: an advanced web-accessible database with the capacity to search the three-dimensional fragments within RNA structures. BMC Bioinformatics 11(1):231

Putignano V, Rosato A, Banci L, Andreini C (2017) MetalPDB in 2018: a database of metal sites in biological macromolecular structures. Nucleic Acids Res 46(D1):D459–D464

Rosenbaum DM, Rasmussen SG, Kobilka BK (2009) The structure and function of G-protein-coupled receptors. Nature 459(7245):356

Shindyalov IN, Bourne PE (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 11(9):739–747

Shindyalov IN, Bourne PE (2001) A database and tools for 3-D protein structure comparison and alignment using the Combinatorial Extension (CE) algorithm. Nucleic Acids Res 29(1):228–229

Stansfeld PJ, Goose JE, Caffrey M, Carpenter EP, Parker JL, Newstead S, Sansom MS (2015) MemProtMD: automated insertion of membrane protein structures into explicit lipid membranes. Structure 23(7):1350–1361

Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J et al (2007) BioMagResBank. Nucleic Acids Res 36(suppl_1):D402–D408

Varadi M, Tompa P (2015) The protein ensemble database. Intrinsically disordered proteins studied by NMR spectroscopy. Springer, pp 335–349

Velankar S, Alhroub Y, Alili A, Best C, Boutselakis HC, Caboche S et al (2010) PDBe: protein data bank in Europe. Nucleic Acids Res 39(suppl_1):D402–D410

Velankar S, van Ginkel G, Alhroub Y, Battle GM, Berrisford JM, Conroy MJ, Dana JM, Gore SP, Gutmanas A, Haslam P (2015) PDBe: improved accessibility of macromolecular structure data from PDB and EMDB. Nucleic Acids Res 44(D1):D385–D395

Wang XT, Chan TF, Lam V, Engel PC (2008) What is the role of the second "structural" NADP+-binding site in human glucose 6-phosphate dehydrogenase? Protein Sci 17(8):1403–1411

Wlodawer A, Minor W, Dauter Z, Jaskolski M (2008) Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. FEBS J 275(1):1–21

Yeats C, Maibaum M, Marsden R, Dibley M, Lee D, Addou S, Orengo CA (2006) Gene3D: modelling protein structure, function and evolution. Nucleic Acids Res 34(suppl_1):D281–D284

Yin H, Flynn AD (2016) Drugging membrane protein interactions. Annu Rev Biomed Eng 18:51

Zanegina O, Kirsanov D, Baulin E, Karyagina A, Alexeevski A, Spirin S (2015) An updated version of NPIDB includes new classifications of DNA–protein complexes and their families. Nucleic Acids Res 44(D1):D144–D153

Chapter 5
Other Biological Databases

Divya Mishra, Vivek Kumar Chaturvedi, V. P. Snijesh, Noor Ahmad Shaik, and M. P. Singh

Contents

D. Mishra
Centre of Bioinformatics, University of Allahabad, Allahabad, India

V. K. Chaturvedi · M. P. Singh (✉)
Centre of Biotechnology, University of Allahabad, Allahabad, India

V. P. Snijesh
Innov4Sight Health and Biomedical System Private Limited, Bangalore, India

N. A. Shaik
Department of Genetic Medicine, Faculty of Medicine, King Abdulaziz University,
Jeddah, Saudi Arabia
e-mail: nshaik@kau.edu.sa

N. A. Shaik et al. (eds.), *Essentials of Bioinformatics, Volume I*,
https://doi.org/10.1007/978-3-030-02634-9_5

5.1 Introduction

The current era of next-generation genome sequencing demands the storage of huge amount of biological data in specific categorized manner. As biology has progressively revolutionized to a data-rich science, the requirement for storing and communicating large datasets has grown tremendously. Biological databases developed as a response to the massive data generated by DNA sequencing technologies. They are complex, heterogeneous, and dynamic. Biological databases can be further classified into sequence, structure, and functional databases. Sequence database stores nucleic acid and protein sequences, and structure database stores the structures of RNA and proteins. Functional databases deliver data on the functional role of gene products, for instance, enzyme activities or biological pathways.

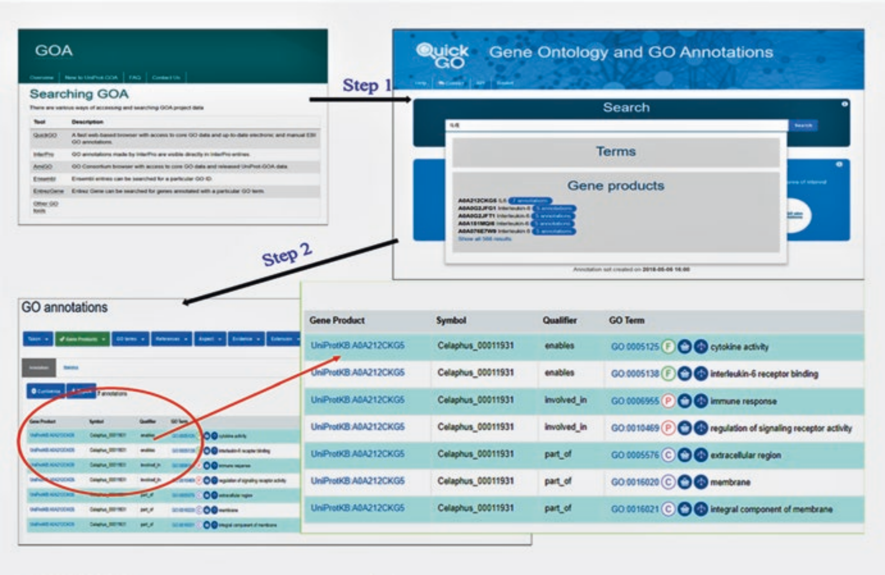
The data can be submitted directly to the database, and the submitted data are indexed, optimized, and organized. The data deposited in biological databases is structured for optimal analysis and comprises of raw and annotated data. Data indexing, organization, and optimization support researchers to identify significant data by making it accessible in a format that is machine or computer readable. Sequences and structures are only among the several different types of data required in the practice of the modern molecular biology. In this chapter, we discuss on protein identification and other biological databases which combine different primary and secondary database sources.

Identifying protein with its specific characteristics is a significant step in the field of proteomics research. It may lead to identification of candidate signatures or novel biomarkers associated with certain diseases based on their characteristics in the sample (McHugh and Arthur 2008). Identification of proteins is taken as a primary step to illuminate the biological information of an organism via studying its protein patterns (Apweiler et al. 2004). Other biological database includes data types like two-dimensional gel electrophoresis images of protein expression, mutations, and polymorphism in molecular sequences and structures, metabolic pathways and molecular interactions, functional enrichment, genetic maps, and physiochemical data.

5.2 Gene or Genome Annotation Databases

5.2.1 GO/GOA Databases

The biological databases may contain diverse kinds of information extracted from various molecular resources; hence there is a need to integrate the biological information in a way that it gives meaningful insight to biologists (Ashburner et al. 2000). The major part of the data integration efforts is the development and use of observation guideline such as ontologies. The availability of an overwhelming amount of protein data generated from different experimental projects necessitates the need to organize and describe the data in a unique and standard vocabulary for conducting future investigations (Evelyn et al. 2004). The Gene Ontology Annotation (GOA) databases provide many aspects of information, i.e., it gives consistent terms for the gene product in multiple databases and also provides the standardized classification of sequence. The tools available in the database are QuickGO, InterPro, AmiGO, Ensembl, and EntrezGene. Annotation of the gene using QuickGO is represented in Fig. 5.1. It has three leading objectives as follows: (Apweiler et al. 2004) to establish an array of complete vocabulary, i.e., ontologies to explain primary realm of molecular biology, (Aranda et al. 2009) to employ GO terms to the annotation of the gene or their products in the databases, and (Ashburner et al. 2000) to



There are various ways of accessing and searching GOA project data

Tool	Description
QuickGO	A fast web-based browser with access to core GO data and up-to-date electronic and manual EBI GO annotations.
InterPro	GO annotations made by InterPro are visible directly in InterPro entries.
AmiGO	GO Consortium browser with access to core GO data and released UniProt-GOA data.
Ensembl	Ensembl entries can be searched for a particular GO ID.
EntrezGene	Entrez Gene can be searched for genes annotated with a particular GO term.
Other GO tools	

Gene Product	Symbol	Qualifier	GO Term
UniProtKB:A0A212CKG5	Celaphus_00011931	enables	GO:0005125 F cytokine activity
UniProtKB:A0A212CKG5	Celaphus_00011931	enables	GO:0005138 F interleukin-6 receptor binding
UniProtKB:A0A212CKG5	Celaphus_00011931	involved_in	GO:0006955 P immune response
UniProtKB:A0A212CKG5	Celaphus_00011931	involved_in	GO:0010469 P regulation of signaling receptor activity
UniProtKB:A0A212CKG5	Celaphus_00011931	part_of	GO:0005576 C extracellular region
UniProtKB:A0A212CKG5	Celaphus_00011931	part_of	GO:0016020 C membrane
UniProtKB:A0A212CKG5	Celaphus_00011931	part_of	GO:0016021 C integral component of membrane

Fig. 5.1 Gene ontologies extracted using QuickGO for the protein interleukin-6 (IL-6) where F, P, and C under GO term represent molecular function, biological process, and cellular component, respectively

consolidate the public domain that permits the worldwide access to the ontologies (Camon et al. 2004). The Gene Ontology Annotation database can be accessed at www.ebi.ac.uk/GOA.

5.2.2 UCSC Genome Browser: Annotation Database

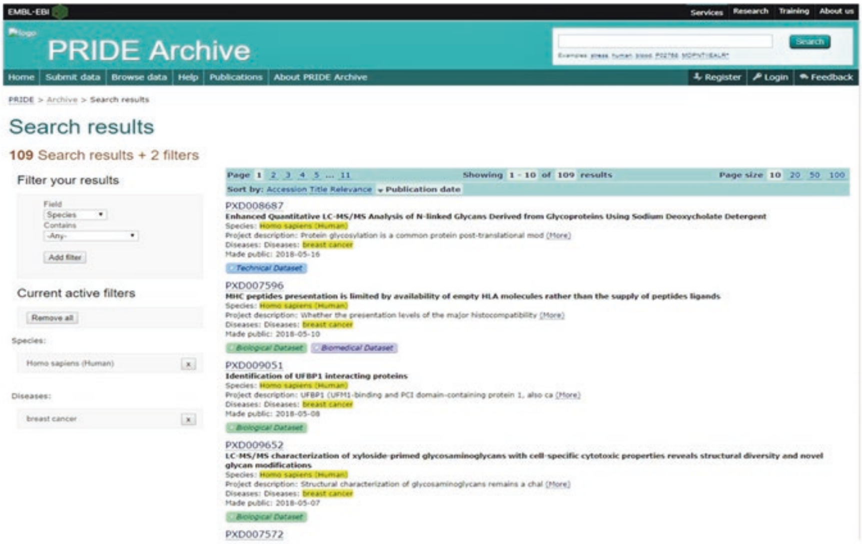
UCSC genome browser is an online as well as stand-alone web-based tool maintained by the University of California. This browser is designed to maintain the user-specified information accessible at any scale to conduct sequence annotations. It has a graphical viewing tool that helps to screen the specified region of the genome. The UCSC genome browser provides the information as in diverse array known as "Tracks," including the gene prediction, mapping, as well as aligned genomic information. This browser site anchors the collection of genomic analysis tools that include the full aspects of GUI web interface to extract the information from this database. It has several online tools such as FAST sequence alignment and BLaTM tool that help to find the sequences from massive genomic sequences.

5.3 Protein Annotation Databases

5.3.1 PRIDE Archive

The PRoteomics IDEntification (PRIDE) database was created in 2004 at the European Bioinformatics Institute (EBI). Since 2014, the original database was renamed to PRIDE Archive and can be accessed via www.ebi.ac.uk/pride/archive. It acts as central resource for deposition of mass spectrometry (MS)-based proteomics data and provides the easy access to experimental data to scientific communities (Jones et al. 2007). A simple search query for the terms breast cancer and *Homo sapiens* is depicted in Fig. 5.2. PRIDE Archive is one of the core members in the ProteomeXchange (PX) consortium (www.proteomexchange.org), which allows the users to submit MS-based proteomics data to the public repository. The data submitted to PRIDE archive via PX are managed and handled by expert biocurators. The public datasets deposited in PRIDE can be explored using ProteomeCentral which is a portal for ProteomeXchange datasets. The implementation of PX with PRIDE has produced an exceptional rise in the number of datasets. Currently, datasets in PRIDE archive contain 9315 projects and 84,479 assays.

PRIDE is the Proteomics Standards Initiative (PSI) submissive public archive for proteomic description in which any proteome laboratory data is accepted for submission. Its main aim is to standardize the data submission as well as dissemination of proteomics information worldwide (Jones et al. 2007). Moreover, PRIDE reposi-



Search results

109 Search results + 2 filters

Filter your results

Field: Species
Contains: -Any-

Current active filters

Species: Homo sapiens (Human)

Diseases: breast cancer

Page 1 2 3 4 5 ... 11 Showing 1 - 10 of 109 results Page size 10 20 50 100

Sort by: Accession Title Relevance ▾ Publication date

PXD008687
Enhanced Quantitative LC-MS/MS Analysis of N-linked Glycans Derived from Glycoproteins Using Sodium Deoxycholate Detergent
Species: Homo sapiens (Human)
Project description: Protein glycosylation is a common protein post-translational mod (More)
Diseases: Diseases: breast cancer
Made public: 2018-05-16
Technical Dataset

PXD007596
MHC peptides presentation is limited by availability of empty HLA molecules rather than the supply of peptides ligands
Species: Homo sapiens (Human)
Project description: Whether the presentation levels of the major histocompatibility (More)
Diseases: Diseases: breast cancer
Made public: 2018-05-10
Biological Dataset *Biomedical Dataset*

PXD009051
Identification of UFBP1 interacting proteins
Species: Homo sapiens (Human)
Project description: UFBP1 (UFM1-binding and PCI domain-containing protein 1, also ca (More)
Diseases: Diseases: breast cancer
Made public: 2018-05-08
Biological Dataset

PXD009652
LC-MS/MS characterization of xyloside-primed glycosaminoglycans with cell-specific cytotoxic properties reveals structural diversity and novel glycan modifications
Species: Homo sapiens (Human)
Project description: Structural characterization of glycosaminoglycans remains a chal (More)
Diseases: Diseases: breast cancer
Made public: 2018-05-07
Biological Dataset

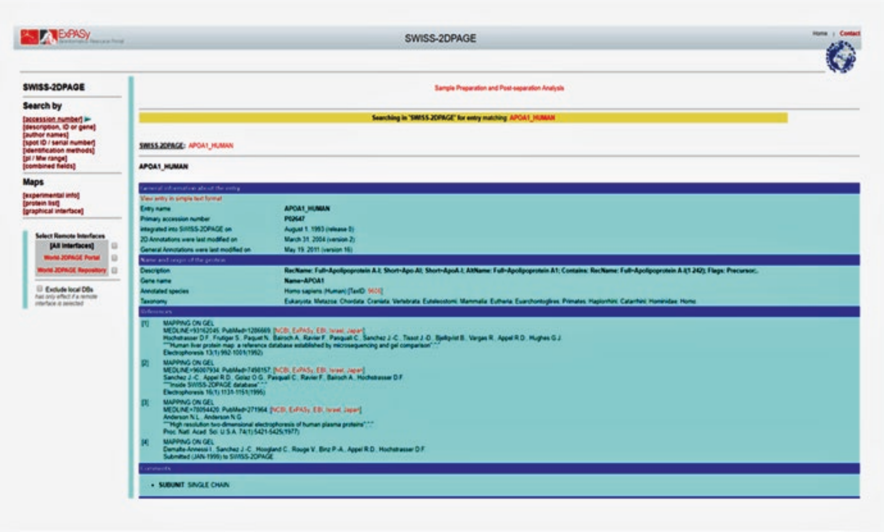
PXD007572

Fig. 5.2 Representation of the query breast cancer and *Homo sapiens* via PRIDE archive

tory contains the information related to peptides, identification of proteins, posttranslational modifications and supporting spectral evidence, relevant measure values, comparable mass spectra, program scripts, and other biological information related to proteome which is contributed by submitters. PRIDE provides inbuilt tools like PRIDE Inspector and PRIDE Converter. PRIDE Inspector is a desktop application that helps researchers to visualize and analyze MS datasets, such as mzML, mzIdentML, and PRIDE XML. PRIDE Converter allows the user to convert common mass spectrometry data formats into PRIDE XML.

5.3.2 SWISS-2DPAGE

The SWISS-2DPAGE database was developed in 1993 and maintained at the clinical laboratory of Geneva University Hospital. The database holds proteins identified on various two-dimensional polyacrylamide gel electrophoresis using microsequencing, immunoblotting, gel comparison, and amino acid composition methods (Hoogland et al. 1999). Each entry in the database has textual information on a protein that include mapping procedures, physiological and pathological information, bibliographical references, and experimental data like isoelectric point, molecular weight, amino acid composition, and peptide masses. Apart from the textual information, the database also provides images of various 2D PAGE and SDS-PAGE which depicts experimental protein localization and theoretical region calculated



Searching in "SWISS-2DPAGE" for entry matching: APOA1_HUMAN

SWISS-2DPAGE: APOA1_HUMAN

APOA1_HUMAN

General information about the entry

View entry in simple text format

Entry name	APOA1_HUMAN
Primary accession number	P02647
integrated into SWISS-2DPAGE on	August 1, 1993 (release 0)
2D Annotations were last modified on	March 31, 2004 (version 2)
General Annotations were last modified on	May 19, 2011 (version 16)

Name and origin of the protein

Description	RecName: Full=Apolipoprotein A-I; Short=Apo-AI; Short=ApoA-I; AltName: Full=Apolipoprotein A1; Contains: RecName: Full=Apolipoprotein A-I(1-242); Flags: Precursor;.
Gene name	Name=APOA1
Annotated species	Homo sapiens (Human) [TaxID: 9606]
Taxonomy	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo

References

[1] MAPPING ON GEL
MEDLINE=93162045; PubMed=1286669; [NCBI, ExPASy, EBI, Israel, Japan]
Hochstrasser D.F., Frutiger S., Paquet N., Bairoch A., Ravier F., Pasquali C., Sanchez J.-C., Tissot J.-D., Bjellqvist B., Vargas R., Appel R.D., Hughes G.J.
""Human liver protein map: a reference database established by microsequencing and gel comparison".";
Electrophoresis 13(1):992-1001(1992).

[2] MAPPING ON GEL
MEDLINE=96007934; PubMed=7498157; [NCBI, ExPASy, EBI, Israel, Japan]
Sanchez J.-C., Appel R.D., Golaz O.G., Pasquali C., Ravier F., Bairoch A., Hochstrasser D.F.
""Inside SWISS-2DPAGE database".";
Electrophoresis 16(1):1131-1151(1995).

[3] MAPPING ON GEL
MEDLINE=78054420; PubMed=271964; [NCBI, ExPASy, EBI, Israel, Japan]
Anderson N.L., Anderson N.G.
""High resolution two-dimensional electrophoresis of human plasma proteins".";
Proc. Natl. Acad. Sci. U.S.A. 74(1):5421-5425(1977).

[4] MAPPING ON GEL
Dematte-Annessi I., Sanchez J.-C., Hoogland C., Rouge V., Binz P.-A., Appel R.D., Hochstrasser D.F.
Submitted (JAN-1999) to SWISS-2DPAGE

Comments

- SUBUNIT: SINGLE CHAIN.

Fig. 5.3 Data explored from SWISS-2DPAGE for the protein query APOA1_Human

from the protein sequence (Fig. 5.3). The data in the textual format follows the parameters used in the Swiss-Prot protein sequence database like ID (identification) and the AC (accession number). However, three parameter types are unique to SWISS-2D PAGE like MasTer (MT), images (IM), and two-dimensional (2D) gel. MasTer (MT) represents the type of map on which proteins are identified and images (IM) represents 2D PAGE image corresponding to the particular entry. Two-dimensional (2D) describes information like mapping procedure, isoelectric point, and molecular weight (Chistine et al. 2000).

5.3.3 *Domain Databases*

Domains are distinct functional or structural units of a protein which can independently fold as a unit of polypeptide chain and carry specific function (Corpet et al. 1999). Protein domains are important to comprehend because it holds the contents related to evolution and protein folding (Majumdar et al. 2009). Many proteins consists of distinct domains. The domains in a protein are conserved from generation to generation, and molecular evolution uses domains as building blocks in different arrangements to perform several distinct functions. Domain database provides inclusive knowledge about protein structures as well as the evolutionary relationship among known structures of proteins. With the use of domain database, it gives insight toward structural prediction as well as protein folding mechanism. Major domain databases are listed in Table 5.1.

Table 5.1 List of domain databases

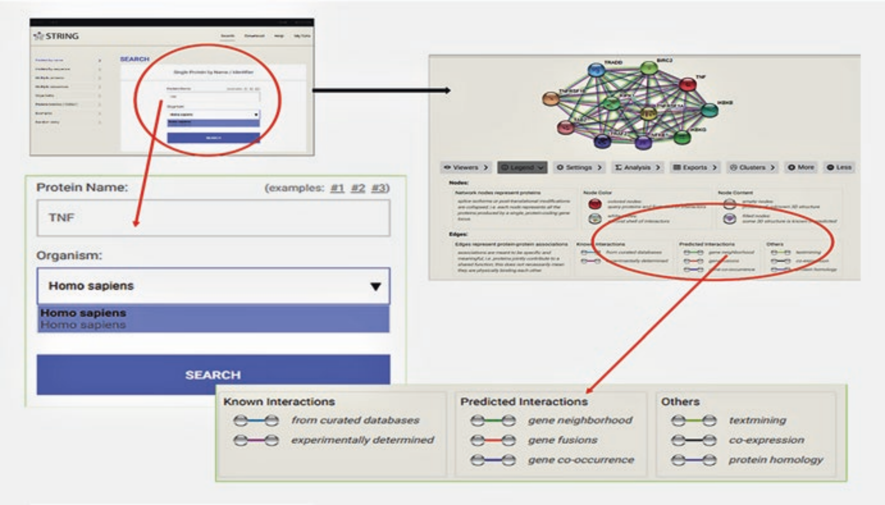
Databases	Links	Description
Pfam	pfam.xfam.org	Family and domain database
ProDom	prodom.prabi.fr	ProDom is a comprehensive set of protein domain families automatically generated from the UniProt Knowledge Database
SCOP	scop.mrc-lmb.cam.ac.uk/scop	SCOP is a (mostly) manually curated ordering of domains from the majority of proteins of known structure in a hierarchy according to structural and evolutionary relationships
CDD	ncbi.nlm.nih.gov/cdd	Family and domain database
CATH	cathdb.info/	The CATH Protein Structure Classification database is a free, publicly available online resource that provides information on the evolutionary relationships of protein domains
DOMINE	manticore.niehs.nih.gov/domine	DOMINE is a database of known and predicted protein domain (domain-domain) interactions. It contains interactions inferred from PDB entries and those that are predicted by 13 different computational approaches using Pfam domain definitions

5.4 Network Databases

Functional characterization of proteins can be done through the interaction models and their relevant position in the corresponding interaction networks. Elucidating the molecular interaction networks of any query protein plays a crucial role in fundamental biological research and also helps to discover new drug targets. Several computational approaches have been developed to predict the interaction between different proteins basing on their sequence and structural features. These computational approaches rely on the homologous sequence analysis, phylogenetic profiling, Bayesian networks, and pattern comparison analysis to build molecular networks for any given protein. Based on these approaches, several types of network databases have been developed like STRING, PIP, BioGRID, MINT (Zanzoni et al. 2002), and IntAct which are depicted in Fig. 5.4.

5.4.1 IntAct

In the present era, protein interaction analysis has become a major focus of proteomics and biomolecular research as protein-protein interaction studies provide valuable information to interpret the cellular activity. IntAct offers an open-source database and tool kit for the storage, presentation, and analysis of protein interactions. An experimental technique such as yeast two-hybrid and affinity purification technique allows generating a large amount of protein-protein interaction data (Hermjakob et al. 2004). The IntAct database offers the extensive knowledge of interactive proteins. Since IntAct is an open-source database, it allows the user to install the database locally according to the requirement of the respective



Protein Name: (examples: #1 #2 #3)

TNF

Organism:

Homo sapiens ▼

Homo sapiens
Homo sapiens

SEARCH

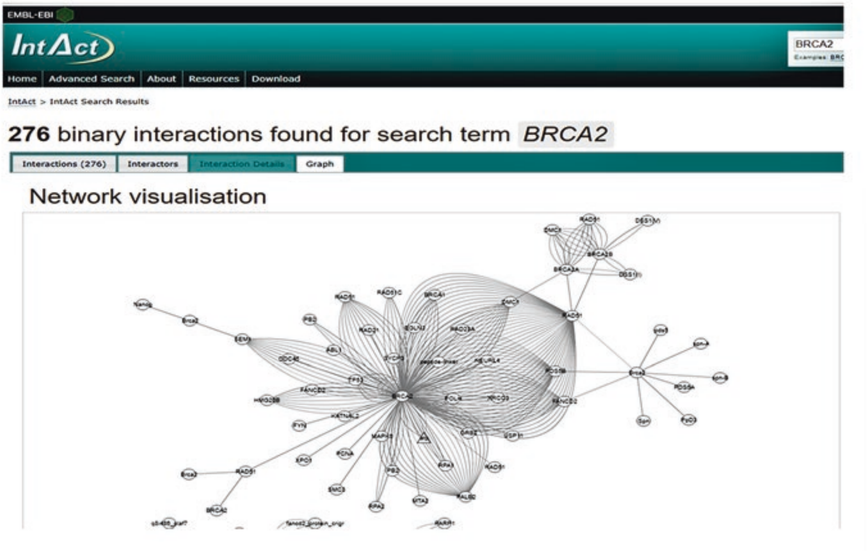
Known Interactions	Predicted Interactions	Others
from curated databases	gene neighborhood	textmining
experimentally determined	gene fusions	co-expression
	gene co-occurrence	protein homology

Fig. 5.4 The protein tumor necrosis factor (TNF) and its functional partners generated from STRING database

organization. Moreover, it lowers the expansion time and provides the consistent information related to interaction datasets through the use of the same framework and curated system.

The IntAct database has three main elements, i.e., experiment, interaction, and interactor. Experiments categorize interactions generally from a publication and group the experimental conditions in which those categorized interactions are reported. An interactor is a biomolecule like protein, DNA, RNA, or small molecule taking part in any biological interaction (Hermjakob et al. 2004). An interaction may consist of one or more than one interactor to participate in interactions which are given in Fig. 5.5.

Generally, an annotated interaction database incorporates the data from several resources, and the key challenge is to ensure data consistency. In the data attributes like experimental methods, source must be curated in a reliable way that data remain accurate and searchable. The IntAct database uses organized vocabularies instead of free-text attributes from existing reference systems like the NCBI and GO database. IntAct gives a simple search interface for investigating in the database via names, accession number, and identifiers like Swiss-Prot and GO terms. It provides two different views, binary and experiment view, for displaying the data of user input. In the case of binary and experimental view, there are collection of particular proteins and their representation in the form of graphics (Aranda et al. 2009). At present, IntAct database consists of 843,123 interactions and 106,978 interactors, principally extracted from large-scale experiments and interactions imported from the literature by the IntAct and Swiss-Prot curation teams.



276 binary interactions found for search term *BRCA2*

Network visualisation

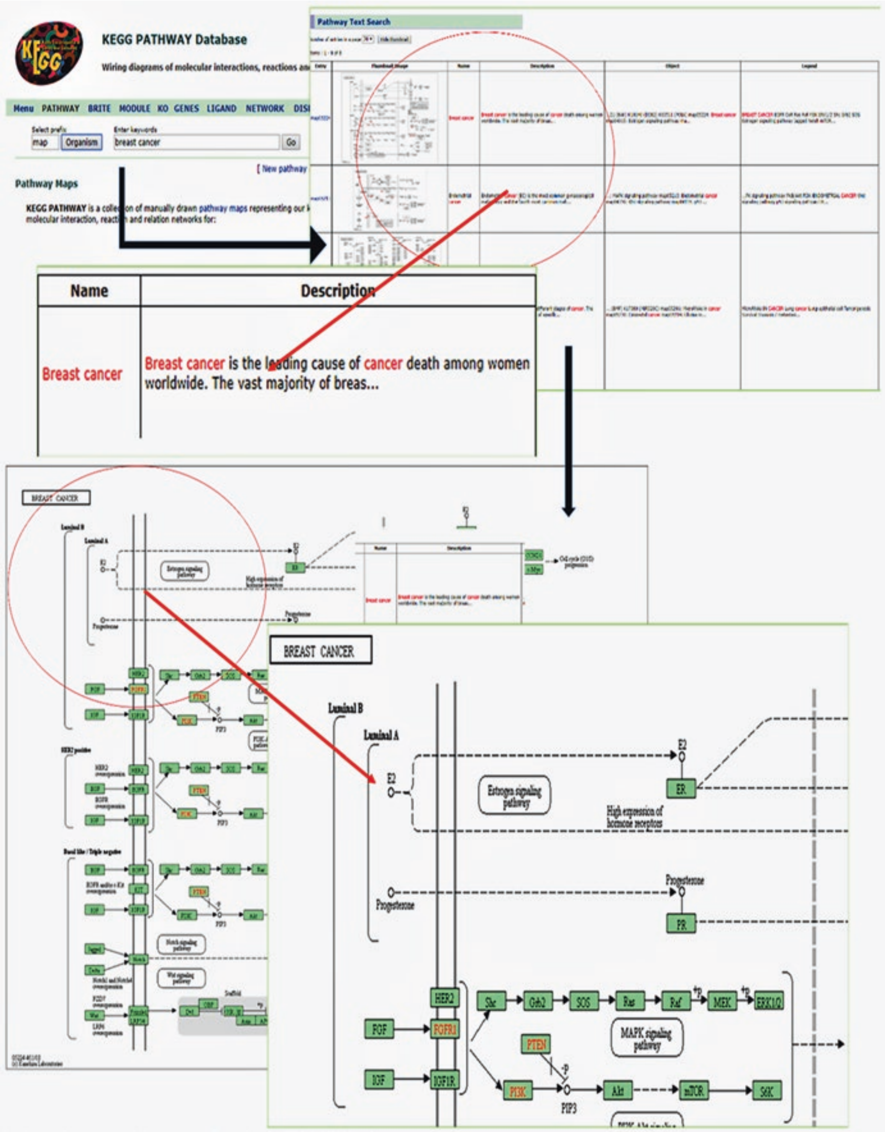
Fig. 5.5 Representation of functional partners of BRCA2 in IntAct database

5.5 Pathway Databases

Data on pathways are available from enormous number of databases developed by expert curators covering a large number of putative pathways, generated using natural language processing and text mining. Due to many changes in quality, size, and property, it is obligatory to use the correct database based on purpose of research. In this section, we introduce some of the major pathway database. Pathway information is often described in the XML (eXtensible Markup Language) data format, which varies from database to database which is read by both humans and computers. The important pathway databases which are widely used are Kyoto Encyclopedia of Genes and Genomes, BioCyc, Ingenuity Pathways Knowledge Base, and Reactome.

5.5.1 Kyoto Encyclopedia of Genes and Genomes

KEGG (Kyoto Encyclopedia of Genes and Genomes) is a series of databases developed by both the Bioinformatics Center of Kyoto University and the Human Genome Center of the University of Tokyo. KEGG contains manually curated pathway maps which signify knowledge on molecular reaction, interaction, and relation networks for metabolism, organismal systems, human diseases, genetic information processing, environmental information processing, cellular process, and drug development (Kanehisa et al. 2016). KEGG collects crucial data relevant to biological systems including phenotype and genotype data and provides necessary information required for system biology understanding of genes, genome sequence, and chemical data. This information is presented as a browser-viewable pathway diagram. For instance, user can search and explore whether metabolic pathways or enzymes exist between molecules X and Y. An overall view of the KEGG pathway database is shown in Fig. 5.6.


KEGG PATHWAY Database
Wiring diagrams of molecular interactions, reactions an
Menu PATHWAY BRITE MODULE KO GENES LIGAND NETWORK DISI
Select prefix
map
Organism
Enter keywords
breast cancer
Go
[New pathway
Pathway Text Search

Pathway Maps

KEGG PATHWAY is a colle on of manually drawn pathway maps representing our k molecular interaction, react n and relation networks for:

Name	Description
Breast cancer	Breast cancer is the leading cause of cancer death among women worldwide. The vast majority of breas...

Fig. 5.6 Breast cancer pathway generated using KEGG database

5.5.2 BioCyc

BioCyc is a high-quality database which focuses on metabolic pathways originally formed by SRI International's Bioinformatics Research Group. BioCyc database provides pathways for eukaryotic and prokaryotic species whose genomes have already been sequenced. The data in the BioCyc are created by software that identify the metabolic pathways of whole-sequenced species and also predict operons and coding genes for missing enzymes in metabolic pathways. BioCyc incorporates data such as GO information and protein features from other bioinformatics databases like UniProt. BioCyc website offers a collection of tools for exploring and visualizing the database for analysis of omics data and comparative genomics. The data in the BioCyc databases are divided into three tiers, based on their quality. Tier 1 databases have the most accurate data and are curated by at least one person a year. Tiers 2 and 3 comprise predicted metabolic pathways using computational methods.

5.5.3 Ingenuity Pathways Knowledge Base

Ingenuity Pathways Knowledge Base (IPKB) is the pathway database created by Ingenuity Systems, Inc. (www.ingenuity.com). It is a central source of functional annotations and biological interactions generated from mass of discretely modeled associations among genes, proteins, cells, metabolites, tissues, complexes, diseases, and drugs. The aforementioned associations or definitions are manually curated for accuracy and comprised of rich contextual information linked to the original article. Overall, the database acts as a starting point for investigation and a bridge between innovative discovery and known biology.

5.5.4 Reactome Pathway Databases

The Reactome pathway databases are freely available online knowledge base of biological pathways. They are largely focusing on human pathway information. These databases consist of curated information including the diverse collection set of data in life sciences. The annotated data include cell signaling, transport, pathogenic interaction with the host, cell cycle as well as biological function, etc. The Reactome pathway database has analysis tool for viewing the interactive pathway graphs, mapping, and overrepresentation onto Reactome pathway database.

5.5.5 Other Pathway Databases

Pathway databases provide information related to the biochemical reactions and the product formed in that reaction. The pathway databases concatenate the information of multiple biological processes that are correlated to each other like biosynthesis, catabolism and tRNA charged with amino acids. A major role of pathway databases is to encode all the metabolic pathways related to an organism, determined either experimentally or computationally. The other alternative function of pathway databases is to provide information for the metabolic processes to the user for a single or set of the related substrate. Some of the pathway databases available at present are given in Table 5.2.

5.6 Drug Databases

Various pharmaceutical systems holds information related to in silico drugs which are required for analytical support as well as research purpose. Pharmaceutical industries build softwares to manage drug data integration like database products, drug allusion results, etc. This system plays an essential role in the drug industry and provides the comprehensive information related to known drugs or about its derivatives. In this section, we discuss few drug databases widely used in both research and clinical settings.

5.6.1 DrugBank

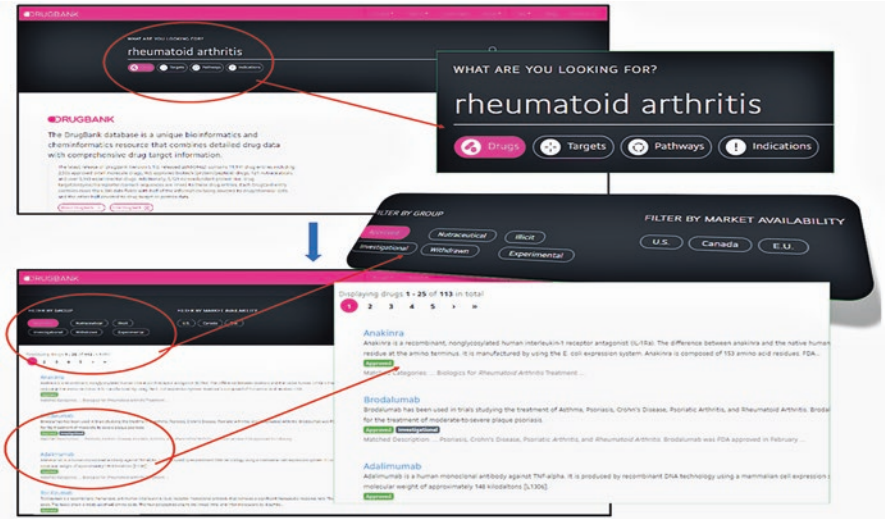
DrugBank is a richly interpreted resource that associates detailed drug data with comprehensive drug target and drug action information. DrugBank has been extensively used to enable in silico drug target discovery, drug design, drug docking or screening, drug metabolism prediction, drug interaction prediction, and general pharmaceutical education (Wishart et al. 2007). DrugBank provides extensive links to major bioinformatics and biomedical databases like GenBank, Swiss-Prot/UniProt, PDB, ChEBI, KEGG, PubChem, and PubMed. It provides different kinds of tools for data maintenance, image processing, and data extraction. DrugBank gives user-friendly web interface for searching, accessing, and exporting the information. The simple search to retrieve drugs for a disease is mentioned in Fig. 5.7. Extensive information of drug and drug target in the DrugBank has enabled the discovery and repurposing of some existing drugs to treat rare and newly identified illnesses. The newest release of DrugBank (version 5.1.0) covers 11,177 drug records including 2560 approved small-molecule drugs, 965 approved biotech (protein/peptide) drugs, 121 nutraceuticals, and over 5160 experimental drugs.

Table 5.2 List of major pathway databases

Databases	Links	Description
Netpath	netpath.org	A curated resource of signal transduction pathways in humans
Reactome	reactome.org	Navigable map of human biological pathways, ranging from metabolic processes to hormonal signaling
WikiPathways	wikipathways.org	Metabolic pathway and protein functional databases
iPath	pathways.embl.de	Interactive Pathways Explorer (iPath) is a web-based tool for the visualization, analysis, and customization of various pathway maps
BioCarta	biocarta.com	BioCarta is a supplier and distributor of characterized reagents and assays for biopharmaceutical and academic research. It catalogs community produced online maps depicting molecular relationships from areas of active research, generating classical pathways as well as suggestions for new pathways
Cancer Cell Map	cancer.cellmap.org	The Cancer Cell Map is a selected set of browsable and searchable human cancer-focused pathways
HumanCyc	humancyc.org	HumanCyc provides an encyclopedic reference on human metabolic pathways. It provides a zoomable human metabolic map diagram, and it has been used to generate a steady-state quantitative model of human metabolism
IntAct	ebi.ac.uk/intact	IntAct provides a freely available, open-source database system and analysis tools for molecular interaction data. All interactions are derived from literature curation or direct user submissions and are freely available
HPRD	hprd.org	The Human Protein Reference Database represents a centralized platform to visually depict and integrate information pertaining to domain architecture, posttranslational modifications, interaction networks, and disease association for each protein in the human proteome
MINT	mint.bio.uniroma2.it	MINT focuses on experimentally verified protein-protein interactions mined from the scientific literature by expert curators
BRENDA	brenda-enzymes.org	BRENDA (the Comprehensive Enzyme Information System) is an information system representing one of the most comprehensive enzyme repositories
TRANSPATH	genexplain.com/transpath	TRANSPATH is a database of mammalian signal transduction and metabolic pathways

5.6.2 *PharmGKB*

The Pharmacogenomics Knowledge Base (PharmGKB) is a publicly available, online knowledge base which collects, annotates, integrates, and distributes the data concerning the impact of human genetic variation on drug response. PharmGKB is managed at Stanford University and funded by the National Institutes of Health (NIH), National Institute of General Medical Sciences (NIGMS), and Pharmacogenomics Research Network (PGRN). An example for annotation of the gene *TP53* in the PharmGKB database is represented in Fig. 5.8.



The DrugBank database is a unique bioinformatics and cheminformatics resource that combines detailed drug data with comprehensive drug target information.

Displaying drugs **1 - 25** of **113** in total

1 2 3 4 5 › »

Anakinra

Anakinra is a recombinant, nonglycosylated human interleukin-1 receptor antagonist (IL-1Ra). The difference between anakinra and the native human residue at the amino terminus. It is manufactured by using the E. coli expression system. Anakinra is composed of 153 amino acid residues. FDA...

Approved

Matched Categories: ... Biologics for *Rheumatoid Arthritis* Treatment ...

Brodalumab

Brodalumab has been used in trials studying the treatment of Asthma, Psoriasis, Crohn's Disease, Psoriatic Arthritis, and Rheumatoid Arthritis. Brodal for the treatment of moderate-to-severe plaque psoriasis.

Approved Investigational

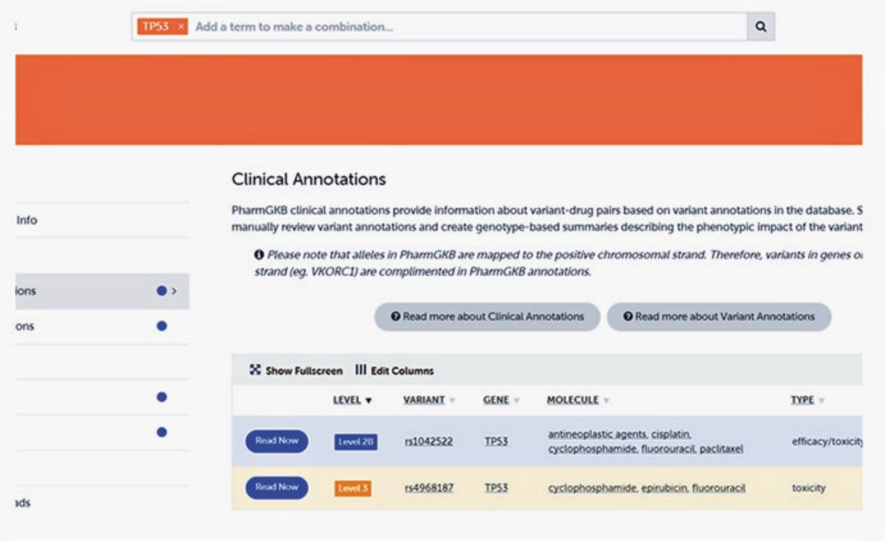
Matched Description: ... Psoriasis, Crohn's Disease, Psoriatic *Arthritis*, and *Rheumatoid Arthritis*. Brodalumab was FDA approved in February ...

Adalimumab

Adalimumab is a human monoclonal antibody against TNF-alpha. It is produced by recombinant DNA technology using a mammalian cell expression molecular weight of approximately 148 kilodaltons [L1306].

Approved

Fig. 5.7 Exploring drug details for the term “rheumatoid arthritis” in DrugBank database


TP53 ×
Add a term to make a combination...
Info
ions
ons
ıds
Read more about Clinical Annotations
Read more about Variant Annotations
Show Fullscreen
Edit Columns

Clinical Annotations

PharmGKB clinical annotations provide information about variant-drug pairs based on variant annotations in the database. S manually review variant annotations and create genotype-based summaries describing the phenotypic impact of the variant

Please note that alleles in PharmGKB are mapped to the positive chromosomal strand. Therefore, variants in genes o strand (eg. VKORC1) are complimented in PharmGKB annotations.

	LEVEL	VARIANT	GENE	MOLECULE	TYPE
Read Now	Level 2B	rs1042522	TP53	antineoplastic agents, cisplatin, cyclophosphamide, fluorouracil, paclitaxel	efficacy/toxicity
Read Now	Level 3	rs4968187	TP53	cyclophosphamide, epirubicin, fluorouracil	toxicity

Fig. 5.8 Clinical annotation of the gene *TP53* in PharmGKB database

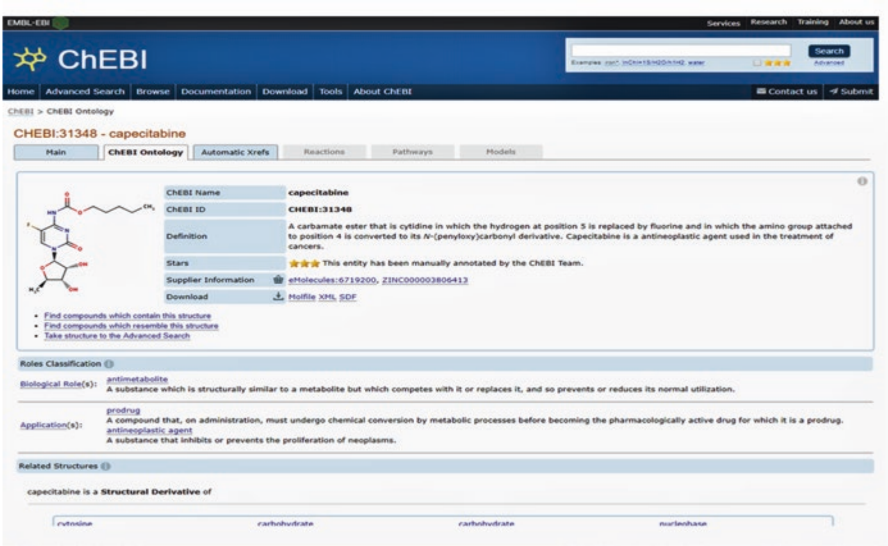
The important objective of PharmGKB is to support the scientists in understanding how genetic variation affects a person and how a person’s body responds to a drug. Technically, this field of study is termed as pharmacogenomics or pharmacogenetics (PGx). To implement this objective, PharmGKB manually annotates and verifies pharmacogenomics data from the primary literature and then stores it in the knowledge base. Identifying consistent genetic variant-drug response interactions with strong supporting evidence can be considered for potential clinical implementation in the future. PharmGKB is collaborated with several international consortia such as Warfarin Pharmacogenetics Consortium (IWPC) and the International Clopidogrel Pharmacogenomics Consortium (ICPC) to support analysis of large pharmacogenomics datasets.

5.6.3 ChEBI

Chemical Entities of Biological Interest (ChEBI) is a database and ontology of molecular entities focused on "small" chemical compounds. The database provides more significant information about molecular entities like distinct isotopic atoms, molecules, ions, and complex atomic information. These molecular entities are either the natural or synthetic products that are involved in the biological processes (Whitfield et al. 2006). ChEBI contains the small molecules' ontology related to each other which helps the users to design the new data depicted in Fig. 5.3. These chemical ontologies also explain the biological role which the small molecules are active in. It also provides the information related to pathways, biochemical reactions, gene expression, as well as protein's structure and function. Java version 6 is required to access the entries in the ChEBI. It consists a number of distinct chemical entities, each of which may manually be curated via expert analyst in the number of knowledge fields. A common ChEBI archive consists of a number of knowledge fields, i.e., characterization of knowledge field given as ChEBI names, descriptions, and ID number.

The submission of compounds in ChEBI database are based on the star system as follows: (Apweiler et al. 2004) three-stars system into which the archive has manually curated or processed via ChEBI organization, (Aranda et al. 2009) two-stars system in which the chemical archive data are manually processed via ChEBI depositors, and (Ashburner et al. 2000) one-star system, where the chemical archive data are automatically curated from a data source. On the other hand, an absence of stars indicates that the chemical archive is wiped out or outdated. ChEBI also represents the information related to chemical structures along with MDL molecular files, interpreted chemical input, formulas, and information related to mass and charge or chemical nomenclature of particular chemical archives. This database is designed as the relational database which makes considerable benefits in ontology interpretations. The ChEBI database is widely used in the field of science and artificial intelligence (Degtyarenko et al. 2007).

When compared with established commercial chemistry resources, ChEBI is a small database (De Matos et al. 2010). However, the asset of ChEBI lies in its quality. ChEBI offers and promotes "gold standard" annotation for molecular entities which contains standard vocabularies, representation of structures as graphs, and well-defined associations between the entities (Fig. 5.9).



CHEBI:31348 - capecitabine

ChEBI Name	capecitabine
ChEBI ID	CHEBI:31348
Definition	A carbamate ester that is cytidine in which the hydrogen at position 5 is replaced by fluorine and in which the amino group attached to position 4 is converted to its *N*-(penyloxy)carbonyl derivative. Capecitabine is a antineoplastic agent used in the treatment of cancers.
Stars	★★★ This entity has been manually annotated by the ChEBI Team.
Supplier Information	eMolecules:6719200, ZINC000003806413
Download	Molfile XML SDF

- Find compounds which contain this structure
- Find compounds which resemble this structure
- Take structure to the Advanced Search

Biological Role(s): antimetabolite
A substance which is structurally similar to a metabolite but which competes with it or replaces it, and so prevents or reduces its normal utilization.

Application(s): prodrug
A compound that, on administration, must undergo chemical conversion by metabolic processes before becoming the pharmacologically active drug for which it is a prodrug.
antineoplastic agent
A substance that inhibits or prevents the proliferation of neoplasms.

capecitabine is a **Structural Derivative** of

cytosine carbohydrate carbohydrate nucleobase

Fig. 5.9 Information extracted from ChEBI ontology for small-molecule capecitabine

5.6.4 PubChem

The PubChem database contains the chemical information at the molecular level, including functionality against the biological aspects. The PubChem database includes the chemical information at the broad range, i.e., name, molecular weight, formula, XlogP, donor as well as acceptor bonding properties, etc. This database has 93.9 million entries of the pure chemical compound as well as 236 million entries of mixed substance or uncharacterized compounds.

5.6.5 ZINC Database

The ZINC database contains the collection of publically available chemical compounds mainly maintained for virtual screening. The goal of ZINC database is to provide the three-dimensional chemical compound to biologically relevant aspects. We can freely download the chemical compound at various file format, i.e., 3D SDF, Dock flexible, as well as SMILES format.

5.7 Specialized Database

5.7.1 Model Organism Databases

Model organism databases (MODs) are biological knowledge base, developed to provide deep in-depth biological data for intensively studied model organisms. MODs support scientists to easily explore contextual information on big sets of genes, plan and conduct experiments more efficiently, integrate their data with prevailing knowledge, and construct novel hypotheses. MODs allow users to analyze results and interpret datasets, and the data they produce are increasingly used to describe less well-studied species. The data derived from MODs are used for the clarifications and understanding of *Homo sapiens*-related data. The well-known model organisms are *Saccharomyces cerevisiae*, *E. coli*, *Drosophila*, and *Mus musculus*. Each organism consists of the genes that encode for proteins which are similar to *Homo sapiens*. In model organisms, the genetic use is the most productive way to understand the human homologs which are affected by the mutation (Engel 2009). The major model organism databases are given in Table 5.3.

5.7.2 IntEnz

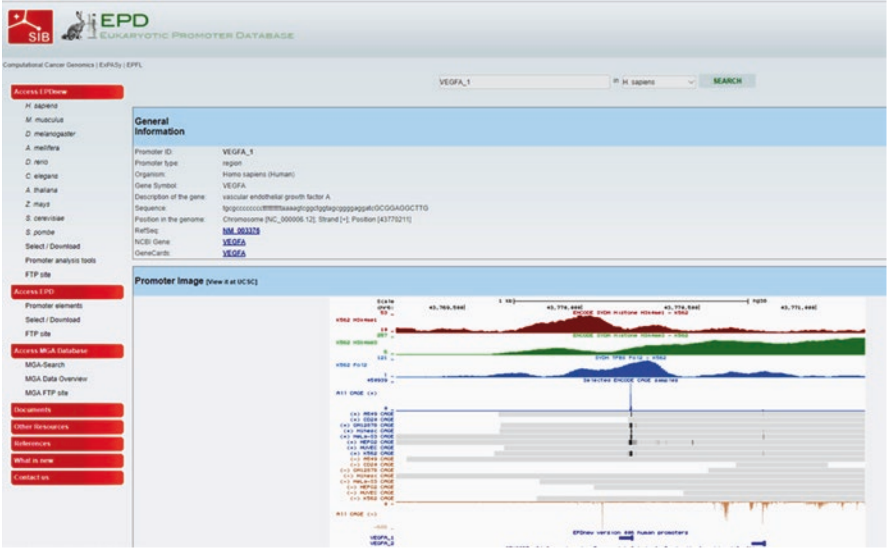
The integrated relational enzyme database (IntEnz) is a publicly available database focused on enzyme nomenclature. IntEnz is supported by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB) and contains enzyme records that are annotated and accepted by the Nomenclature Committee (Whitfield et al. 2006). The IntEnz database contains archive for every enzyme records along with their EC (enzyme classification) number, suggested name, and disease data information (Fleischmann et al. 2004).

5.7.3 EPD

The Eukaryotic Promoter Database (EPD) is created in Weizmann Institute of Science, Israel, based on EMBL information center. The EPD database is a curated collection of eukaryotic POL II promoters, for which the transcription start site has been determined experimentally (Périer et al. 1998). It is developed for comparative sequence analysis and plays a crucial role in determining the eukaryotic transcription regulatory element (Whitfield et al. 2006). EPD is designed in a way that enables dynamic mining of biologically meaningful promoter subsets for comparative sequence analysis. An example for using EPD is represented in Fig. 5.10.

Table 5.3 List of model organism databases

Databases	Links	Description
Saccharomyces Genome Database	yeastgenome.org	The Saccharomyces Genome Database provides comprehensive integrated biological information for the budding yeast *S. cerevisiae* along with search and analysis tools to explore these data, enabling the discovery of functional relationships between sequence and gene products in fungi and higher organisms
PomBase	pombase.org	PomBase is a model organism database that provides online access to the fission yeast *Schizosaccharomyces pombe* genome sequence and its features, together with a wide range of associated biological data and references to supporting literature
Xenbase	xenbase.org	Xenbase is MOD, providing informatics resources, as well as genomic and biological data on *Xenopus* frogs
ZFIN	zfin.org	The Zebrafish Information Network is an online biological database of information about the zebrafish (*Danio rerio*)
TAIR	arabidopsis.org	The Arabidopsis Information Resource (TAIR) maintains a database of genetic and molecular biology data for the model higher plant *Arabidopsis thaliana*
Candida Genome Database	candidagenome.org	The Candida Genome Database is a resource for genomic sequence data, gene and protein information for *Candida albicans* and related species
dictyBase database	dictybase.org	dictyBase is the model organism database for the social amoeba *Dictyostelium discoideum*
Rat Genome Database	rgd.mcw.edu	RGD is responsible for attaching biological information to the rat genome via structured vocabulary, or ontology, annotations assigned to genes and quantitative trait loci (QTL) and for consolidating rat strain data and making it available to the research community
Mouse Genome Database	informatics.jax.org	MGD provides access to data on the genetics, genomics, and biology of the laboratory mouse to facilitate the study of human health and disease
WormBase	wormbase.org	WormBase is an online biological database about the biology and genome of the nematode model *organism Caenorhabditis elegans* and contains information about other related nematodes
EcoCyc	ecocyc.org	The EcoCyc project performs literature-based curation of the *E. coli* genome and of *E. coli* transcriptional regulation, transporters, and metabolic pathways
FlyBase	flybase.org	FlyBase is an online bioinformatics database and the primary repository of genetic and molecular data for the insect family Drosophilidae



VEGFA_1 in H. sapiens

General Information

Promoter ID: VEGFA_1
Promoter type: region
Organism: Homo sapiens (Human)
Gene Symbol: VEGFA
Description of the gene: vascular endothelial growth factor A
Sequence: tgcgcccccccttttttttaaaagtcggctggtagcggggaggatcGCGGAGGCTTG
Position in the genome: Chromosome [NC_000006.12]; Strand [+]; Position [43770211]
RefSeq: NM_003376
NCBI Gene: VEGFA
GeneCards: VEGFA

Fig. 5.10 Exploring promotors of the query VEGFA_1 in EPD

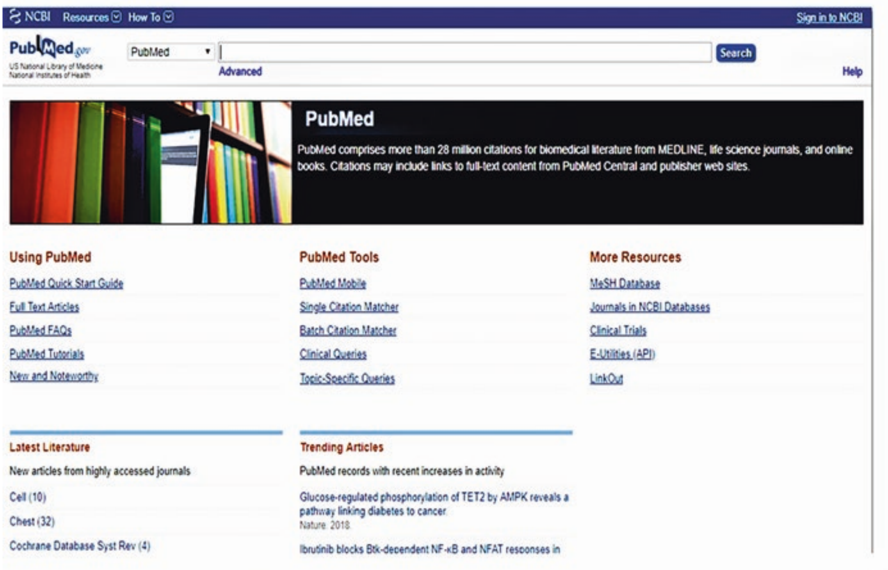
5.7.4 *TRANSFAC*

Transcription factor (TRANSFAC) database is a manually annotated database of eukaryotic transcription factors, their genomic binding sites, and DNA binding profiles. The contents of the database are used as gold standard to predict potential transcription factor binding sites (Wingender et al. 2000). TRANSFAC database can be used to map the individual as well as entire genomic regulatory site and provides extensive knowledge of their transcriptional control.

5.8 Scientific Literature Database

5.8.1 *PubMed*

The PubMed literature database is an open-source database which searches for biomedical and life sciences literature. It provides the access point to search the PMC, NCBI Bookshelf, as well as full text of collected books. The retrieval of information on PubMed is carried via entering the key terms of the object into the PubMed search icon (Fig. 5.11).



PubMed

PubMed comprises more than 28 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher web sites.

Using PubMed

- PubMed Quick Start Guide
- Full Text Articles
- PubMed FAQs
- PubMed Tutorials
- New and Noteworthy

PubMed Tools

- PubMed Mobile
- Single Citation Matcher
- Batch Citation Matcher
- Clinical Queries
- Topic-Specific Queries

More Resources

- MeSH Database
- Journals in NCBI Databases
- Clinical Trials
- E-Utilities (API)
- LinkOut

Latest Literature

New articles from highly accessed journals

Cell (10)

Chest (32)

Cochrane Database Syst Rev (4)

Trending Articles

PubMed records with recent increases in activity

Glucose-regulated phosphorylation of TET2 by AMPK reveals a pathway linking diabetes to cancer
Nature. 2018.

Ibrutinib blocks Btk-dependent NF-κB and NFAT responses in

Fig. 5.11 Web illustration of PubMed

5.8.2 SCI (Science Citation Index)

The Science Citation Index is semantically developed by the Scientific Information Institute. This online web is available through the science web. The database of Science Citation Index provides the platform to the researcher to recognize the citation of the article as well as author's cited information.

5.8.3 Google Scholar

The Google Scholar is an open web-based search engine, which maintained to index the whole literature information across a multitude publishing format. This search engine permits users to extract the creditable scholarly material. The Google Scholar is interdisciplinary and easy to use open source.

5.9 Conclusion

The up-to-date biological relevant data is vital for life sciences research. In this chapter we focused on the protein identification and other biological databases including genome as well as proteome annotation databases and integrated biological information databases including network or pathway databases. Additionally, we have also briefed about drug and scientific literature databases in this chapter.

References

Apweiler R, Bairoch A, Wu CH (2004) Protein sequence databases. J Adv Protein Chem 8(1):76–80

Aranda B, Achuthan P, Alam-Faruque Y, Armean I, Bridge A, Derow C, Feuermann M, Ghanbarian AT et al (2009) The IntAct molecular interaction database in. Nucleic Acids Res 38:525–531

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. J Nat Genet 25(1):25–29

Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, Binns D, Harte N, Lopez R, Apweiler R (2004) The gene ontology annotation (goa) database: sharing knowledge in uniprot with gene ontology. Nucleic Acids Res 32:262–266

Chistine H, Jaen CS, Luisa T, Pierre AB, Amos B, Denis FH, Ron DA (2000) The 1999 SWISS-2DPAGE database updates. Nucleic Acids Res 28(1):286–288

Corpet F, Gouzy J, Kahn D (1999) Recent improvements of the ProDom database of protein domain families. Nucleic Acids Res 27(1):263–267

De Matos P, Dekker A, Ennis M, Hastings J, Haug K, Turner S, Steinbeck C (2010) ChEBI: a chemistry ontology and database. J Cheminform 2:6

Degtyarenko K, De Matos P, Ennis M, Hastings J, Zbinden M, McNaught A, Alcántara R, Darsow M, Guedj M, Ashburner M (2007) ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Res 36:344–350

Engel SR (2009) Using Model Organism Databases (MODs). Curr Protoc Essential Lab Tech 1:1–17

Evelyn C, Michele M, Daniel B, Vivian L, Emily D, John M, David B, Nicola H, Rodrigo L, Rolf A (2004) The gene ontology annotation (GOA) databases: sharing knowledge in Uniprot with gene ontology. Nucleic Acids Res 32:262–266

Fleischmann A, Darsow M, Degtyarenko K, Fleischmann W, Boyce S, Axelsen KB, Bairoch A, Schomburg D, Tipton KF, Apweiler R (2004) IntEnz: the integrated relational databases. Nucleic Acids Res 32:434–437

Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, Vingron M, Roechert B, Roepstorff P, Valencia A, Margalit H (2004) IntAct: an open source molecular interaction database. Nucleic Acids Res 32:452–455

Hoogland C, Sanchez JC, Tonella L, Bairoch A, Hochstrasser DF, Appel RD (1999) The SWISS-2DPAGE database: what has changed during the last year. Nucleic Acids Res 27:289–291

Jones P, Côté RG, Martens L, Quinn AF, Taylor CF, Derache W, Hermjakob H, Apweiler R (2006) PRIDE: a public repository of protein and peptide identifications for the proteomics community. Nucleic Acids Res 34:659–663

Jones P, Cote RG, Cho SY, Klie S, Martens L, Quinn AF, Thorneycroft D, Hermjakob H (2007) PRIDE: new developments and new datasets. Nucleic Acids Res 36:878–883

Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K (2016) KEGG: new perspectives on genomes, pathways, disease and drugs. Nucleic Acids Res 45:353–361

Majumdar I, Kinch LN, Grishin NV (2009) A database of domain definitions for proteins with complex interdomain geometry. PLoS One 4:5084

McHugh L, Arthur JW (2008) Computational methods for protein identification from mass spectrometry data. PLoS Comput Biol 4:12

Périer RC, Junier T, Bucher P (1998) The eukaryotic promoter database EPD. Nucleic Acids Res 26(1):353–357

Whitfield EJ, Pruess M, Apweiler R (2006) Bioinformatics database infrastructure for biotechnology research. J Biotechnol 124(4):629–639

Wingender E, Chen X, Hehl R, Karas H, Liebich I, Matys V, Meinhardt T, Prüß M, Reuter I, Schacherer F (2000) TRANSFAC: an integrated system for gene expression regulation. Nucleic Acids Res 28(1):316–319

Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M (2007) DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res 36:901–906

Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G (2002) MINT: a Molecular INTeraction database. FEBS Lett 513(1):135–140

Chapter 6
Introduction to Nucleic Acid Sequencing

Preetha J. Shetty, Francis Amirtharaj, and Noor Ahmad Shaik

Contents

P. J. Shetty (✉)
Department of Biomedical Sciences, College of Medicine, Gulf Medical University,
Ajman, United Arab Emirates
e-mail: dr.preetha@gmu.ac.ae

F. Amirtharaj
Thumbay Research Institute for Precision Medicine, Gulf Medical University,
Ajman, United Arab Emirates
e-mail: francis@gmu.ac.ae

N. A. Shaik
Department of Genetic Medicine, Faculty of Medicine, King Abdulaziz University,
Jeddah, Saudi Arabia
e-mail: nshaik@kau.edu.sa

N. A. Shaik et al. (eds.), *Essentials of Bioinformatics, Volume I*,
https://doi.org/10.1007/978-3-030-02634-9_6

6.1 Introduction to Biological Sequences

A biological sequence is a single, linear, molecule of nucleic acid or protein referred to as the primary structure of the biological macromolecule. Sequence analysis of this primary structure is a key way of understanding the biology of an organism. Sequence analysis helps to identify the homologous sequences, intrinsic features, sequence variation/differences, the molecular structure of sequence, evolution, and genetic diversity of sequence and organisms. The deoxyribonucleic acid (DNA) is so often called the blueprint of life that contains all the instruction necessary for

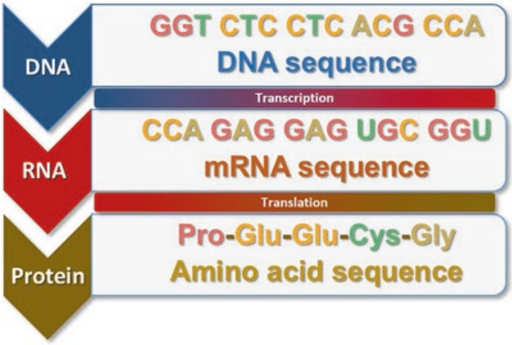


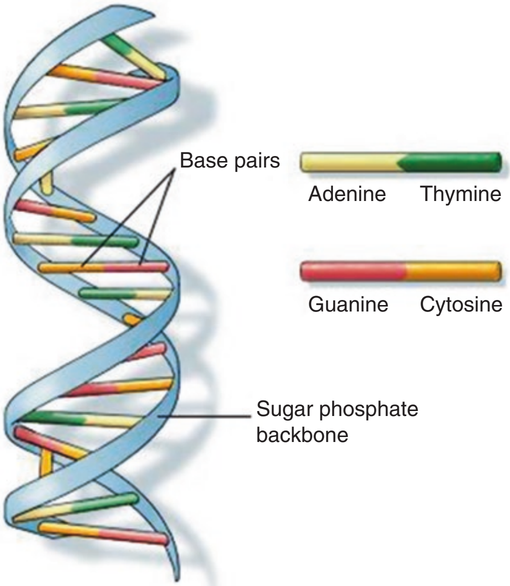
Fig. 6.1 The Central dogma of molecular biology

building an organism. The DNA sequences are transcribed to the mRNA chain which gives the information needed to the ribosome which builds proteins; the polypeptide sequence of amino acids and every part of the body is handled through this system of protein construction (Fig. 6.1).

DNA, the genetic material found in the cells of most living organisms, was first discovered and isolated in 1869 by Friedrich Miescher. It is a polymer of repeating units of nucleotides which are made up of a nitrogenous base, a pentose sugar, and a phosphate group. It is termed as deoxy because of the presence of hydrogen in place of a hydroxyl group (OH) at the second carbon of the deoxyribose (pentose) sugar. DNA has two strands twisted into a double helix. The two strands are made up of simpler molecules called nucleotides. Each nucleotide is made up either of the four nitrogen-containing nucleobases like adenine (A), guanine (G), cytosine (C), and thymine (T) along with deoxyribose and a phosphate group. These nucleotide molecules are interlinked in a chain-like structure by forming covalent bonds between the sugar of one nucleotide and the phosphate of the next, resulting in an alternating sugar phosphate backbone (Fig. 6.2). This pairing contributes in the synthesis of DNA molecules during cell division. There are about 3 billion base pairs which encode different structural and functional proteins in a human being.

DNA sequencing is the experimental method to determine the sequential arrangement of nucleic acid bases (A, T, G, and C) in a polynucleotide which encodes different proteins that are functional in a living cell. The complete set of coding and noncoding sequences in human DNA is referred to as genome. The genome carries the information for all the proteins required for the normal life of an organism. Biological sequences show complex patterns of similarity to each other. This can be identified by searching for similarity among the sequences. For this, we need to know the entire genomic sequence of an organism.

The invention of sequencing technologies along with the bioinformatics tools has contributed a great role in analyzing the genome of an organism. The output file of a sequencer gives us the order of four nucleotides. Maxam, Gilbert, and Sanger's discovery was a landmark discovery which opened the door to develop faster and efficient sequencing technology. Sanger sequencing technology is the most applied technique of sequencing and has been commercialized and automated as the "Sanger sequencing technology," the first-generation sequencing technology.



U.S. National Library of Medicine

Fig. 6.2 Schematic DNA double helix structure and its base complements

Starting from the first-generation sequencing by which the first human genome sequencing was mapped, there have been tremendous growth and innovation in the DNA sequencing technology. The last decade was the new era of sequencing technologies due to the revolutionary changes in new high-throughput sequencing technologies. The DNA sequencing became one of the important research applications in the scientific world in all the fields and revolutionizing many fields of science and is increasingly used in health care to address the diseases as well as the biology of the cell, tissue, and organs.

6.2 DNA Sequencing

The sequencing is an analytical procedure of decoding the DNA sequence to determine the order of the base pairs of the nucleotides present in a stretch of DNA. The central dogma of molecular biology states that DNA can undergo self-replication to form another DNA as well as undergo transcription to form RNA. The order of the nucleotides is essential for the formation of the RNA and its further translation to the coded protein leading to the translation of the genetic information into the structural proteins.

Sanger sequencing developed by Frederick Sanger enabled the sequencing of bacteriophage phi X 174 which contains approximately 5375 nucleotides. This became the first fully sequenced genome in the year 1977. In 2003, the Human

Genome Project (HGP), an international consortium effort, successfully sequenced and mapped the entire human genome, which came to an end after 13 years of research around many laboratories in the world.

A new era of sequencing method commonly described as next-generation sequencing (NGS) technologies with very high throughput and at much lower cost than the first sequencing technologies was launched by Roche's 454 technologies in 2005. The key feature NGS is a parallel sequencing process producing several thousands of sequences simultaneously. These high-throughput sequencers reduced the cost of DNA sequencing. This is achieved by miniaturization of sequencing reactions.

6.2.1 Applications of DNA Sequencing

Since the discovery of first-generation DNA sequencing technology, DNA sequencing has revolutionized numerous fields including biotechnology, forensics, molecular biology, and microbiology. DNA sequencing has helped in the completion of numerous genomes including the human genome. This has enabled researchers to trace the human evolution and at the same time establish evolutionary relationships between species. Apart from this, DNA sequencing has also helped in the identification of genetic variations and is one of the foremost tools used in the study of mutant genes and heritable diseases.

One of the major DNA sequencing applications is in the field of forensic science. DNA sequencing can be used to determine VNTR (variable number tandem repeat) sequences for crime scene profiling or paternity tests. In the recent past, metagenomics has utilized DNA sequencing to explore the ecology of microbes. It has been used in the diagnosis of infectious disease and examination of normal gut flora. DNA sequencing is also being used increasingly in agriculture and animal husbandry, to maximize quantity and quality of yield and determine and establish quality breeding stocks, respectively.

6.3 First-Generation Sequencing

The late 1970–1980s were significant years for genetics and genomics. Invention of polymerase chain reaction (PCR), a process through which amplification of DNA and the development of the first DNA sequencing technologies were made possible, sequencing the entire genome. Sanger sequencing and Maxam-Gilbert sequencing, considered as first-generation sequencing methods, dominated genomics for nearly 40 years. They advanced genomic research by folds and paved a path for subsequent sequencing technologies.

6.3.1 Maxam-Gilbert Sequencing

6.3.1.1 History

Maxam-Gilbert sequencing is one of the earliest DNA sequencing platforms. This sequencing method is popularly known as chemical cleavage method. It was developed in 1977 by Allan Maxam, a Ph.D. student in Harvard University, with Walter Gilbert based on nucleobase-specific partial chemical modifications to the DNA, as well as cleavages of the backbone of the DNA near the modified nucleotides. The method became popular due to the advantage of being able to use purified DNA directly, but it soon faded out of preference because of its technical complexity.

6.3.1.2 Principle

The principle of the Maxam-Gilbert sequencing is based on the displacement of bases as a result of purine (A and G) reactions with dimethyl sulfate and pyrimidine (C and T) reactions with hydrazine. The displacement occurs because of the cleavage of glycosidic linkage between the nitrogenous base and the deoxyribose sugar. In sites of base displacement, piperidine catalyzes the cleavage of the phosphodiester bond. The main aim of the method is to create a single-stranded DNA substrate with a radioactive label on the 5'end, through a series of selective reactions.

6.3.1.3 Procedure

The two polynucleotide strands in DNA are separated into single strand followed by 5′ end radiolabeling with gamma-^{32}P and then cleaved chemically. It is a two-step biocatalytic procedure involving piperidine and two chemicals, namely, dimethyl sulfate and hydrazine, at specific conditions that selectively attack purines and pyrimidines.

Dimethyl sulfate attacks purines, hydrazine attacks pyrimidines, and piperidine catalyzes the cleavage of phosphodiester bond at the site of base displacement. Although both dimethyl sulfate and piperidine specifically cleave the guanine nucleotides, dimethyl sulfate and piperidine in formic acid will cleave both guanine and adenine (G, A + G). Hydrazine and piperidine will cleave both thymine and cytosine nucleotides, whereas hydrazine and piperidine in 1.5 M NaCl will only cleave cytosine nucleotides (C + T, C). This generates series of labeled DNA fragments with specific nucleotides at the 3′ end.

The reaction products are divided by polyacrylamide gel electrophoresis (PAGE) which is based on size. Smallest fragment goes fastest. The labeled fragments in the gel are visualized by autoradiography. The nucleotide sequence is coded from bottom to top of the gel (Fig. 6.3) (Maxam and Gilbert 1977; Gaastra 1985).

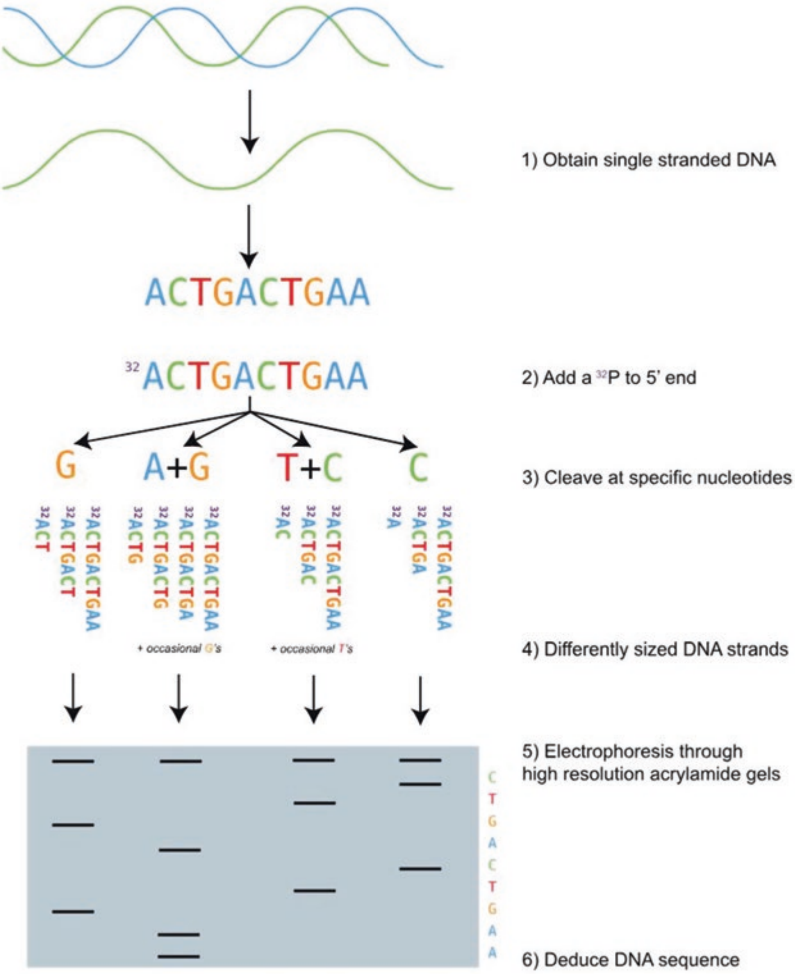


Fig. 6.3 An example of Maxam-Gilbert Sequencing Technique, showing specific cleavage of DNA backbone yielding different sized labelled DNA fragments. Source: Binf snipacdemy

6.3.1.4 Advantages and Disadvantages

The major attraction of the Maxam-Gilbert sequencing procedure is that the DNA template used in the method can be either single-stranded or double-stranded. The Maxam-Gilbert method was preferred over Sanger at a point in time because Sanger method required cloning of the single-stranded DNA for each read start. The Maxam-Gilbert method can also be used for analyzing DNA protein interactions and epigenetic modifications to the DNA.

The main limitation in Maxam-Gilbert sequencing came in the form of usage of harmful chemicals and techniques such as X-rays and radiolabeling. The difficulty in scaling up and handling these techniques, and the requirement of using hydrazine, a known neurotoxin, made the method disadvantageous.

6.3.2 Sanger Sequencing Method

6.3.2.1 History

Sanger sequencing is one of the first methods of DNA sequencing to be discovered. Frederick Sanger along with his colleagues began his research on developing a sequencing technology with sequencing insulin and then RNA, and subsequently DNA. His research paved the way to the Sanger sequencing or chain termination method introduced in the year 1977. It went on to earn Sanger his second Nobel Prize in Chemistry, in 1980, making him one of only two Nobel Laureates to win twice in the same category. The technology was commercialized by Applied Biosystems. It was the method employed to sequence the entire human DNA, in the Human Genome Project, by using hundreds of Sanger sequencing machines across many laboratories in the world.

6.3.2.2 Principle

Chain termination method is also known as dideoxy sequencing method because it involves the use of an analog of normal nucleotide 2′, 3′-dideoxynucleoside triphosphates (ddNTPs). These are chain-terminating nucleotides lacking 3’-OH ends. This method uses single-stranded DNA. This method is based upon the incorporation of ddNTPs into an extending DNA strand to stop chain elongation.

6.3.2.3 Procedure

The chemical reaction is carried out in four separate reaction tubes, with each reaction containing template DNA, primers, DNA polymerase, and four dNTPs with one radiolabeled (Sanger used radio-labeled ddATPs for detection of bands), and additionally, each reaction tube is added with only one of the four ddNTP (ddATP, ddCTP, ddGTP, or ddTTP) in specific concentration. Followed by denaturation and annealing of the primer, the enzyme DNA polymerase starts adding dNTPs to the newly synthesized DNA strand, and if the ddNTP gets incorporated, the reaction terminates. This leads to the separate collection of DNA strands of different sizes in all four different reaction tubes. Individual reaction is then placed into a separate wells of polyacrylamide gel containing urea. Urea helps in the prevention of DNA renaturation during the process if electrophoresis and then the positions of the DNA bands are detected by autoradiography. The radioactive spot indicates the DNA fragments with the ddNTP incorporated at the specific position. The nucleotide sequence in the separating gel is determined from the bottom upward, and the nucleotide sequence of bands in different terminator lanes gives the template nucleotide sequence (Slatko et al. 2001; Men et al. 2008; Sanger et al. 1977; França et al. 2002).

Once the four reactions are complete, gel electrophoresis is performed with SDS-PAGE. All individual reaction mixtures are loaded into a lane to produce four total lanes. The electrophoresis results are transferred onto a polymer sheet and exposed to x-ray autoradiography. This tells us exactly the position of the radioactively labeled fragments. The smaller the fragment is, the farther away it travels. Therefore, the farthest fragment would be the 5’end DNA base. Depending on the size of the fragment, the DNA sequence of the complementary strand can be formed, and this can then be used to sequence the original DNA strand (Fig. 6.4).

6.3.2.4 Advantages and Disadvantages

Sanger sequencing helped researchers to identify mutations and the underlying cause of genetic diseases. It is the best method for identification of short tandem repeats and sequencing single genes. The biggest disadvantage of this method, however, is the amount of time it consumes, which comes with a low throughput. The technique can only process short sequences of DNA (up to 300–1000 base pairs) at a time.

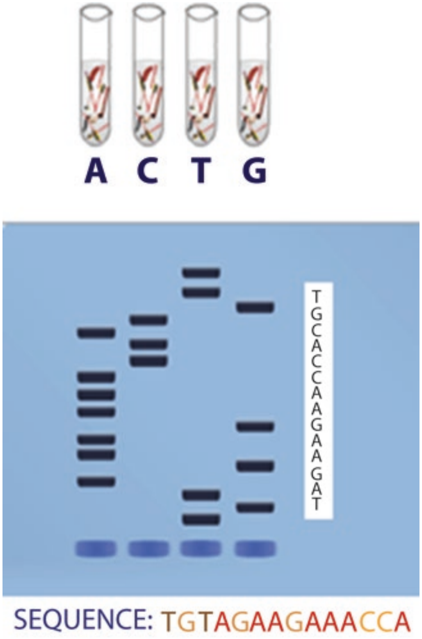


Fig. 6.4 Schematic diagram of PAGE gel reading DNA sequence used in Sanger Sequencing

6.3.3 *Automated DNA Sequencing*

6.3.3.1 History

Both methods – Sanger and Maxam Gilbert – were time-consuming and challenging. In 1986, Leroy Hood and colleagues improved the Sanger sequencing method by using fluorescent labels instead of radiolabels. One of the four fluorescent dyes is used to label the nucleotide primers. Each dye is placed in an individual sequencing reaction with one of the four ddNTPs. Upon the completion of sequencing reactions, all the four reactions are mixed and analyzed together in one lane of a polyacrylamide gel. Later, James M. Prober and colleagues labeled the ddNTPs instead of fluorescence-labeled primers. The use of four different fluorescence-labeled ddNTPs with four different wavelengths permits the sequencing reaction in a single tube instead of four separate reactions. This method was further improved in the early 1990s when Harold Swerdlow and colleagues employed the capillaries in DNA sequencing method. These capillaries are small (with 50 μm inner diameter) and operate with much higher voltages to lower the run times. In 1993, B. L. Karger replaced the polyacrylamide with the low-viscosity separation matrix; later in 1995, Zhang developed a non-cross-linked polymer that is stable even at 60 °C for the high-quality sequence.

6.3.3.2 Principle

This is similar to Sanger sequencing, but the reaction is automated, and the reactions are carried out in a single tube having all four dideoxynucleotide triphosphates each coated with four different fluorescent dyes, each of it emits light at a particular wavelength. The sequence data generated are acquisition and analyzed by the use of a computer (Fig. 6.5).

6.3.3.3 Procedure

In the automated DNA sequencers, sequence reaction is carried out as a single reaction as mentioned in Fig. 6.6. The sequence reaction loaded into the capillary after the dye terminator reaction. The fragments are separated as the constant electrical current applied through the capillary. The fluorescently labeled ddNTPs excite when it passes through the laser. The fluorescent wavelengths are collected by the detector, and the sequence data will be generated using software (Wallis and Morrell 2011).

6.3.3.4 Advantages and Disadvantages

Automated DNA sequencers are expensive, and also it is challenging to sequence repetitive sequence regions. In spite of new discoveries of next-generation sequencers, automated Sanger sequencing is still considered as gold standard due to its accuracy and read length, but it is slow and expensive.

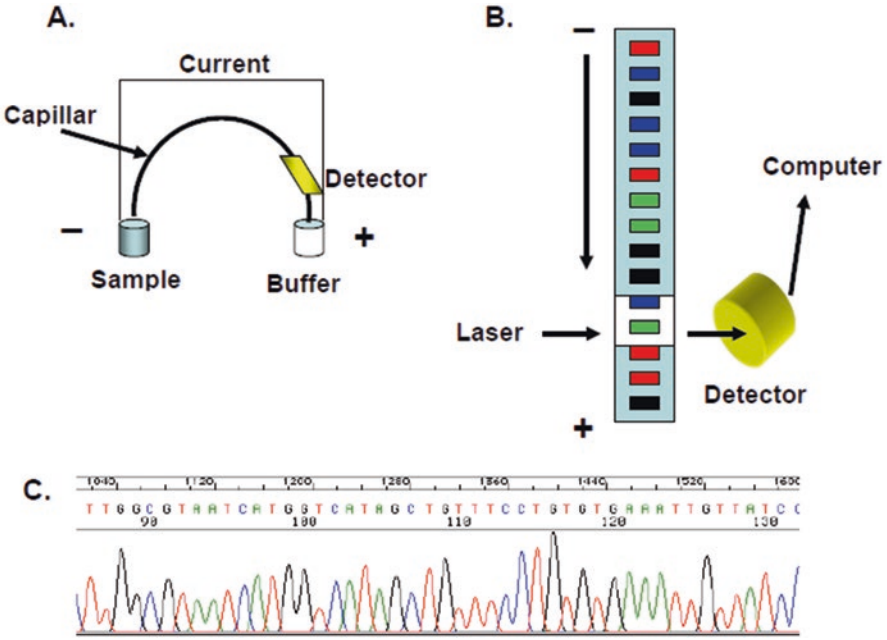
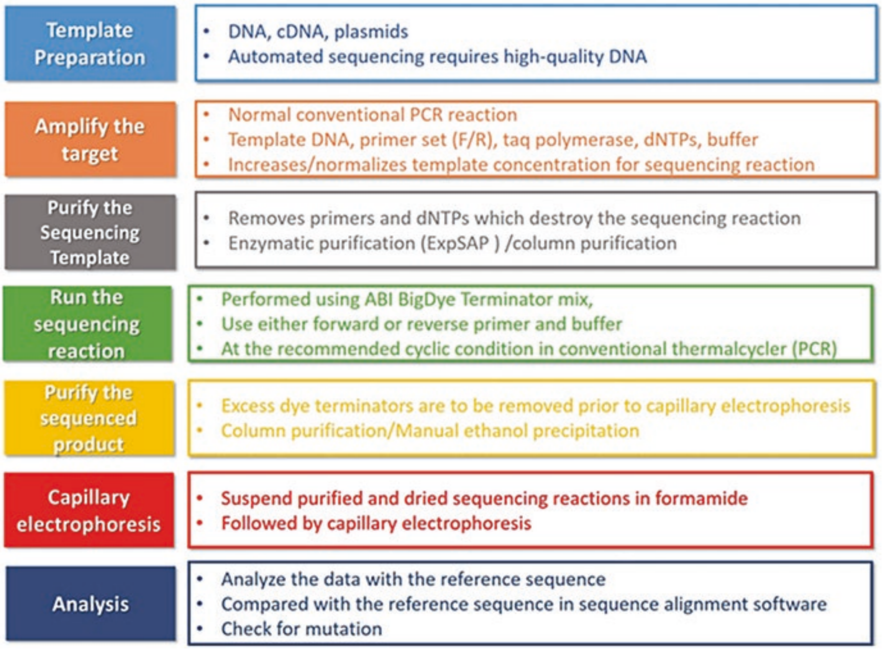


Fig. 6.5 Automated Capillary DNA Sequencing. (**a**). Capillary electrophoresis system; (**b**). Laser detection of the fluorescent labels. (**c**). Electropherogram of the DNA sequence. https://www.prepostseo.com/tmp_imgs/29324374315269O1864.jpg



Template Preparation
- DNA, cDNA, plasmids
- Automated sequencing requires high-quality DNA

Amplify the target
- Normal conventional PCR reaction
- Template DNA, primer set (F/R), taq polymerase, dNTPs, buffer
- Increases/normalizes template concentration for sequencing reaction

Purify the Sequencing Template
- Removes primers and dNTPs which destroy the sequencing reaction
- Enzymatic purification (ExpSAP) /column purification

Run the sequencing reaction
- Performed using ABI BigDye Terminator mix,
- Use either forward or reverse primer and buffer
- At the recommended cyclic condition in conventional thermalcycler (PCR)

Purify the sequenced product
- Excess dye terminators are to be removed prior to capillary electrophoresis
- Column purification/Manual ethanol precipitation

Capillary electrophoresis
- Suspend purified and dried sequencing reactions in formamide
- Followed by capillary electrophoresis

Analysis
- Analyze the data with the reference sequence
- Compared with the reference sequence in sequence alignment software
- Check for mutation

Fig. 6.6 Automated DNA Sequencing workflow

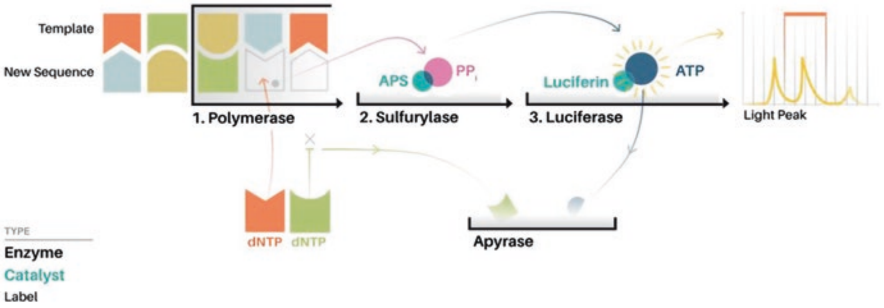


Fig. 6.7 Schematic picture represents the working principle of Pyro Sequencing. (https://commons.wikimedia.org/wiki/File:How_Pyrosequencing_Works.svg)

6.3.4 Pyrosequencing

Pyrosequencing is based on the generation of a light signal by pyrophosphate upon the addition of nucleotide synthesizing complementary strand. This method was developed in the year 1996, and it revolutionized the second-generation sequencers.

6.3.4.1 Procedure

The template DNA is immobilized in the reaction. The nucleotides A, T, G, and C are added and removed sequentially one after the other. The reaction is catalyzed by the use of DNA polymerase, ATP sulfurylase, luciferase, different enzymes, luciferin, and substrates like adenosine 5′ phosphosulfate (APS) and apyrase. Once the primer annealed to the template DNA, DNA polymerase will add one of the complementary nucleotides onto the template strand, and this will release pyrophosphate (PPi). ATP sulfurylase present in the reaction will convert PPi to ATP in the presence of APS. The ATP produced substrates the luciferase and converts luciferin to oxyluciferin which generates light that could be captured by the camera. The reactions start with another nucleotide once the unutilized nucleotides and ATP are degraded by apyrase. Each nucleotide is added in turn so that only one of the four will generate a light signal and eventually captured and recorded (Fig. 6.7) (Harrington et al. 2013; Ravi et al. 2014).

6.4 Second-Generation Sequencing

Sanger sequencing is used exclusively for nearly 30 years after their discovery. The cost and time consumption in the two methods however soon became a concern. The next wave of sequencing technology known as second-generation sequencing emerged in the mid-2000s and aimed at decreasing cost, increasing speed, and eliminating the need for electrophoresis.

6.4.1 Roche 454 Sequencing

6.4.1.1 History

Pyrosequencing began in 1987 as a method utilized for the uninterrupted monitoring of DNA polymerase activity by Nyren and Lundin. In 1988, Edward Hyman continued on the work of Nyren and Lundin to invent a DNA sequencing method. In 1996, the pyrosequencing platform was developed by Ronaghi et al. After almost 10 years, in 2005, Rothberg and his colleagues introduced the first next-generation sequencer to be commercially available based on the pyrosequencing approach done in 1996. Later on, 454 Life Sciences developed a parallelized version of pyrosequencing which since then has been acquired by Roche Diagnostics.

6.4.1.2 Principle

The initial pyrosequencing principle was first described in 1993 by Pål Nyren, Mathias Uhlen, and Bertil Pettersson. During nucleotide incorporation, pyrophosphate is released by the secondary reactions which result in releasing of light. The light is detected, and the light intensity determines the number of the nucleotides added, hence representing the complementary nucleotides on the template strand.

6.4.1.3 Procedure

Hybridization of the sequencing primer to ssDNA template followed by incubation with ATP sulfurylase, DNA polymerase, apyrase, and luciferase, in addition to luciferin and the substrates like adenosine 5′ phosphosulfate (APS), is required for the first step in the solution-based version of pyrosequencing.

Initiation of the second step requires the addition of deoxynucleotide triphosphates (dNTPs), which is incorporated into the template by the action of DNA polymerase. This incorporation releases pyrophosphate (PPi).

In the presence of adenosine 5’phosphosulfate, ATP sulfate converts pyrophosphate to ATP which will act as a substrate for luciferase-mediated conversion of luciferin to oxyluciferin which produces visible light. The amount of light produced is directly proportional to the amount of ATP. A special camera is used to detect the light produced in the luciferase-catalyzed reaction. Apyrase degrades the left-out nucleotides, and ATP restarting the reaction can take place with another nucleotide (Fig. 6.8) (Voelkerding et al. 2009).

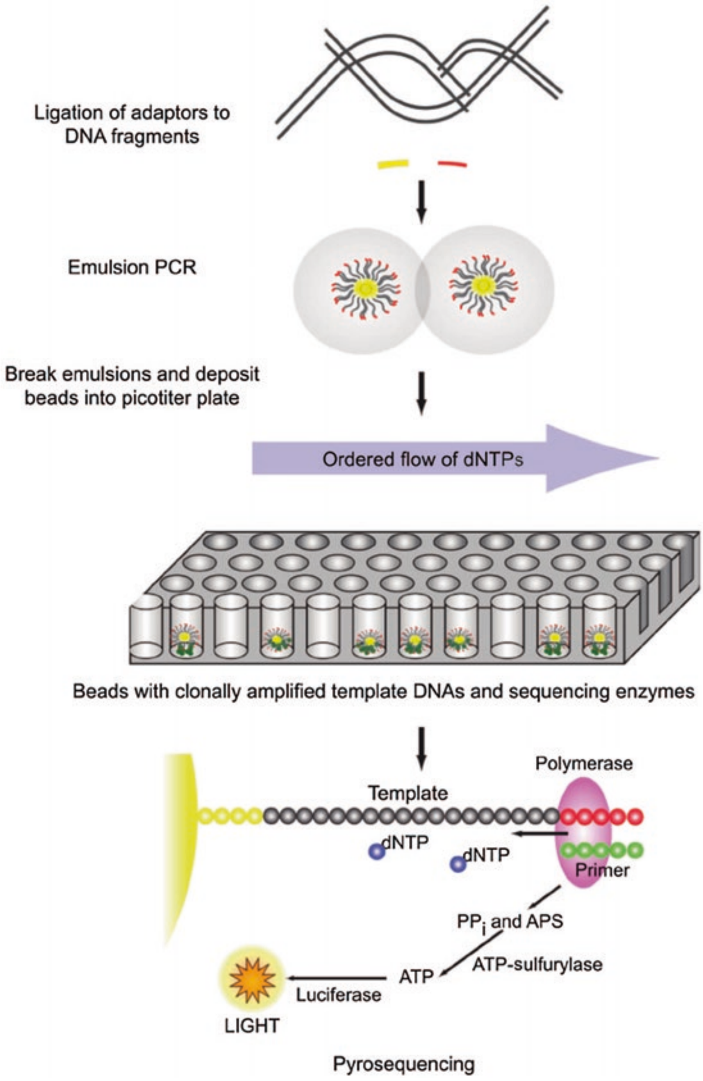

Ligation of adaptors to DNA fragments
Emulsion PCR
Break emulsions and deposit beads into picotiter plate
Ordered flow of dNTPs
Beads with clonally amplified template DNAs and sequencing enzymes
Polymerase
Template
dNTP
dNTP
Primer
PP$_i$ and APS
ATP
ATP-sulfurylase
Luciferase
LIGHT
Pyrosequencing

Fig. 6.8 Roche 454 massive parallel pyrosequencing

6.4.1.4 Advantages and Disadvantages

The main advantage of pyrosequencing is that it is timesaving and can be done in real time. It is cost-effective when compared to dideoxynucleotide chain termination sequencing methods and facilitates haplotype phasing and the identification of structural genetic variation by pairing reads which will span tens of kilobases of the genomic template sequence. The main disadvantage, however, is the occurrence of frameshift errors which are systematic errors in reading (Dewey et al. 2012).

6.4.2 Sequencing by Synthesis: Illumina/Solexa Platform

6.4.2.1 History

The Illumina/Solexa platform is the brainchild of Shankar Balasubramanian and David Klenerman, scientists at Cambridge University, a university that contributed to the first draft of the Human Genome Project. Inspired by the university association with DNA research, both of them used their project on fluorescent-labeled dyes and motion of polymerase to theorize a new sequencing approach known as sequencing by synthesis technology. Later, they formed Solexa Inc. in June 1998 and with investment from Abingworth LLP, established facilities at Chesterford Research Park in 2000. In 2004, Solexa obtained molecular clustering technology from Manteia. In 2006, Solexa launched its first sequencer, the Genome Analyzer, which revolutionized DNA sequencing by enabling sequencing of 1 gigabase of data in a single run. Solexa was acquired by Illumina in 2007, and since then, the Illumina/Solexa platform has remained one of the foremost and widely adopted sequencing technologies in the world.

6.4.2.2 Principle

The Illumina/Solexa platform of sequencing by synthesis (SBS) is based on reversible termination sequencing method. While the principle of SBS technology is very similar to capillary electrophoresis sequencing, the main difference comes in the fact that while Sanger sequencing uses dideoxynucleotides (ddNTPs) to terminate primer extension irreversibly, SBS uses modified nucleotides (i.e., fluorescently labeled deoxyribonucleotide triphosphates) to reversibly terminate primer extension. This has enabled sequencing across millions of DNA fragments in a massively parallel way, instead of single DNA fragment sequencing.

6.4.2.3 Procedure

Illumina sequencing by synthesis has four main steps: sample preparation, cluster generation, sequencing, and data analysis.

Library Preparation The template DNA is randomly fragmented (200–600 base pairs) by an enzyme transposases. This is followed by ligation of adapters (P5/P7) to 5′ and 3′ ends. Alternatively, six-base-pair indices are added which creates the unique barcode for the sample enabling sequencing different samples at the same time. The adapter-ligated fragments are amplified by PCR reaction and subsequently gets purified.

Cluster Generation The sample is loaded on a flow cell with a lawn of two types of oligosides which are complementary to P5/P7 adapter sequence of the DNA fragments. Each hybridized DNA fragments is attached to the complementary oligo, and DNA polymerase enzyme creates a complementary strand. The double-stranded DNA is denatured, and the original template is washed away, while the new fragment which is covalently attached to the flow cell remains. The ssDNA forms a bridge by hybridizing with the adjacent complementary primer and is extended by the polymerase which results in the formation of a dsDNA bridge. The dsDNA bridge is denatured, and the end result is two ssDNA strands covalently attached to the flow cell. The bridge amplification cycle is repeated numerous times. Likewise, each fragment is amplified into distinct, clonal clusters through bridge amplification, leaving a cluster of uniform DNA sequence (Fig. 6.9) (Voelkerding et al. 2009). Now the template is ready for sequencing.

Sequencing After clonal amplification, the reverse ssDNA is cleaved and washed away, leaving only the forward ssDNA attached to the flow cell. The primer anneals to the forward strand and will start adding fluorescently labeled ddNTPs. Only one base pair added at a time with reversible terminator which is to prevent multiple additions in a single time. When a base is incorporated and the fluorophore is excited with a laser and the emission captured. The fluorophore is cleaved off and the terminator removed. The cycle is repeated until the forward strand is completely sequenced, which gives single-end sequencing. For the paired-end sequencing, the sequenced product is washed away. The 3′ ends of the forward strand which were previously blocked are unprotected followed by cluster generation. The primer is introduced to the flow cell and hybridizes to the reverse strand, and a read is generated similar to the forward strand (Ansorge 2009; Guzvic 2013; Buermans and Dunnen 2014; Heather and Chain 2016).

Sequences from the pooled samples are first separated on the basis of the unique indices which were introduced during the sample preparation. Sample reads with similar base calls are clustered, and the forward and reverse reads are paired. The contiguous sequences generated are aligned back to a reference genome. Following alignment variations like single-nucleotide polymorphism, insertion or deletion could be analyzed.

6.4.2.4 Advantages and Disadvantages

The first and foremost advantage of SBS technology is that with standard reagents, it allows up to 96 samples to be sequenced per run. At the same time, SBS technology is better at sequencing homopolymeric sequences in comparison with 454 or ion torrent as it allows incorporation of one nucleotide per reaction.

One of the main limitations of SBS technology remains the limitation of reading length, especially when it comes to de novo sequencing. Substitution errors due to

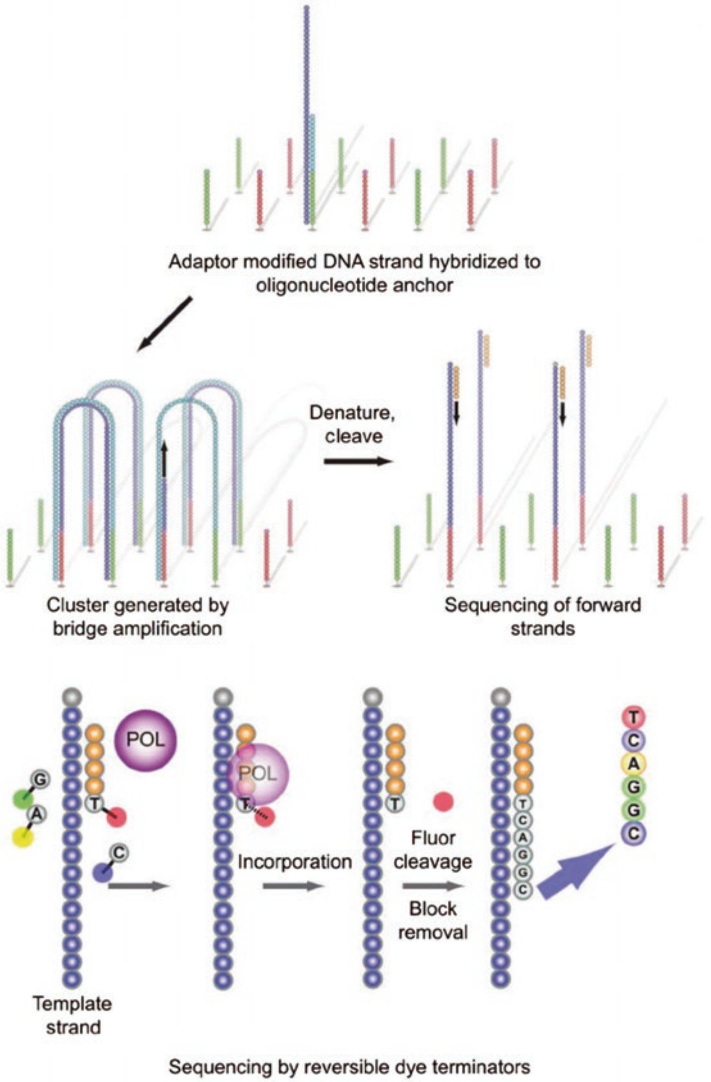

Adaptor modified DNA strand hybridized to oligonucleotide anchor
Denature, cleave
Cluster generated by bridge amplification
Sequencing of forward strands
POL
G
A
C
T
POL
T
Incorporation
T
Fluor cleavage
Block removal
T
C
A
G
G
C
T
C
A
G
G
C
Template strand
Sequencing by reversible dye terminators

Fig. 6.9 Schematic diagram of Illumina sequencing technology

increasing background noise in each cycle, GC bias in bridge amplification, and decreased efficiency of sequencing due to scars caused by unblocking of nucleotides were once major limitations of SBS technology but have hence been reduced as a result of advancement in the field of chemistry (Ari and Arikan 2016).

6.4.3 *Sequencing by Ligation: ABI/Solid*

6.4.3.1 History

ABI Sequencing by Oligonucleotide Ligation and Detection (SOLiD) sequencing is an advanced sequencing method, based on the principle of ligation using DNA ligase instead of DNA polymerase. As of 2008, SOLiD System was said to be the only NGS technology with a 99.94% accuracy. The read length of ABI/SOLiD sequencing is 25 to 35; an approximate number of 40 million beads can be sequenced, and the corresponding output sequence data is 2 to 4 gigabases. It was devised by Applied Biosystems, now known as Life Technologies, Carlsbad, California, United States. It was opened to the market in 2007, but they were discontinued in May 2016 (Huang et al. 2012).

6.4.3.2 Principle

It operates on the principle of constructing the genomic library construction and ligation followed by sequencing reaction. This sequencing technology applies DNA ligase instead of DNA polymerase for sequencing. The genome undergoes random fragmentation and then it attaches to the adapter molecule followed by magnetic bead addition for clonal amplification in such a way that only one DNA fragment will be available on the magnetic bead's surface.

The emulsion PCR is used in amplifying the bead-captured DNA molecules. This amplified bead-captured DNA is anchored to a glass and flooded with fluorescent-labeled oligonucleotides. If the oligonucleotide is complementary to the template, it will be ligated, and then two bases will be detected at one time. The oligonucleotide is then cleaved (Fig. 6.10) (Voelkerding et al. 2009).

6.4.3.3 Procedure

A DNA library is prepared from the sample, and it is utilized in preparing a clonal bead population. Only one DNA fragment is held up on the surface of individual magnetic bead. The P1 adapter is ligated to the starting sequence of every fragment. In a micro reactor containing the entire essential reagent for PCR, an emulsion PCR takes place. The products attached to bead that result from the PCR are then bonded to a glass slide. The primers hybridize the P1 adapter sequence within the library template. A set of four fluorescently labeled di-base probes compete for ligation to the sequencing primer. Interrogating every first and second base in each ligation reaction will specify the di-base probe. Multiple cycles of ligation, detection, and cleavage are performed with the number of cycles determining the eventual read length. After multiple rounds of ligation cycles, the extended product is removed, and this template is again set with primers corresponding to n-1 position

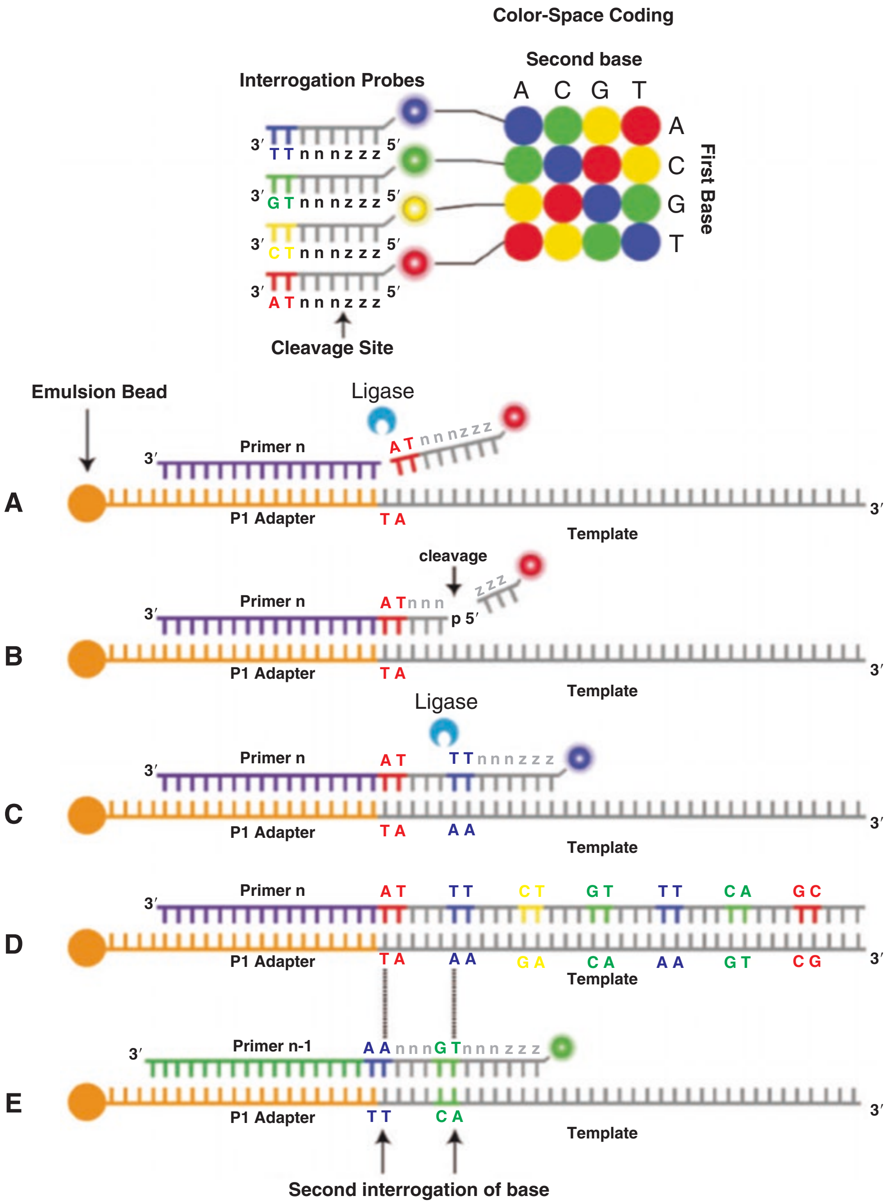


Fig. 6.10 Schematic diagram shows the principle of ABI SOLID sequencing technology

for the second round of ligation cycle. About five rounds of primer reset experiments are executed for sequence tag. During this primer reset process, each nucleotide base is examined in two different independent ligation reactions by two different primer sets.

6.4.3.4 Advantages and Disadvantages

The advantage of ABI/SOLiD sequencing technology is that it is the only next-generation sequencing system that uses ligation-based chemistry with di-base probes which accounts for its high accuracy and downstream data analysis system (Men et al. 2008; Guzvic 2013; Ambardar et al. 2016).

The major limitation of ABI SOLiD Sequencing technology is its shortcoming in sequencing the palindromic sequences, identifying them as miscellaneous random sequences. The obscure error began approx. 2 bps prior to the palindromic sequence.

6.4.4 Ion Torrent Sequencing

6.4.4.1 History

The technology of ion torrent sequencing was licensed from DNA Electronics Ltd., a company focused on innovation in the field of DNA sequencing and in particular next-generation sequencing technology. It was the Ion Torrent Systems Inc. which developed and commercialized the technology in February 2010. The product which came to be known as the Ion Personal Genome Machine (PGM) is commercialized as a rapid but economical sequencer.

6.4.4.2 Principle

Ion torrent platform is the first post light sequencing technology. Rather than using light as an intermediary. It is a semiconductor-based sequencing detection system based on the detection of H_2 ions which are by-products of nucleotide additions to the template strand during polymerization. The beads containing enriched DNA template are added to a microwell in the chip. The microwell will be added with the single type of nucleotide at a single time with the intermittent wash. The complementary nucleotides are incorporated into growing strand of DNA. This will result in the release of hydrogen ion which results in the change in pH that is detected by the sensor on the bottom of the well and converted to an electric signal that is monitored (Fig. 6.11), (Golan and Medvedev 2013).

6.4.4.3 Procedure

The sequencing of DNA is done using a semiconductor chip, which has millions of wells that can capture chemical information and translates them to digital information. The DNA sample is fragmented. Each fragment is attached to a bead and is amplified until it covers the bead by clonal amplification. This automated process covers millions of beads with millions of different fragments. The beads then flow

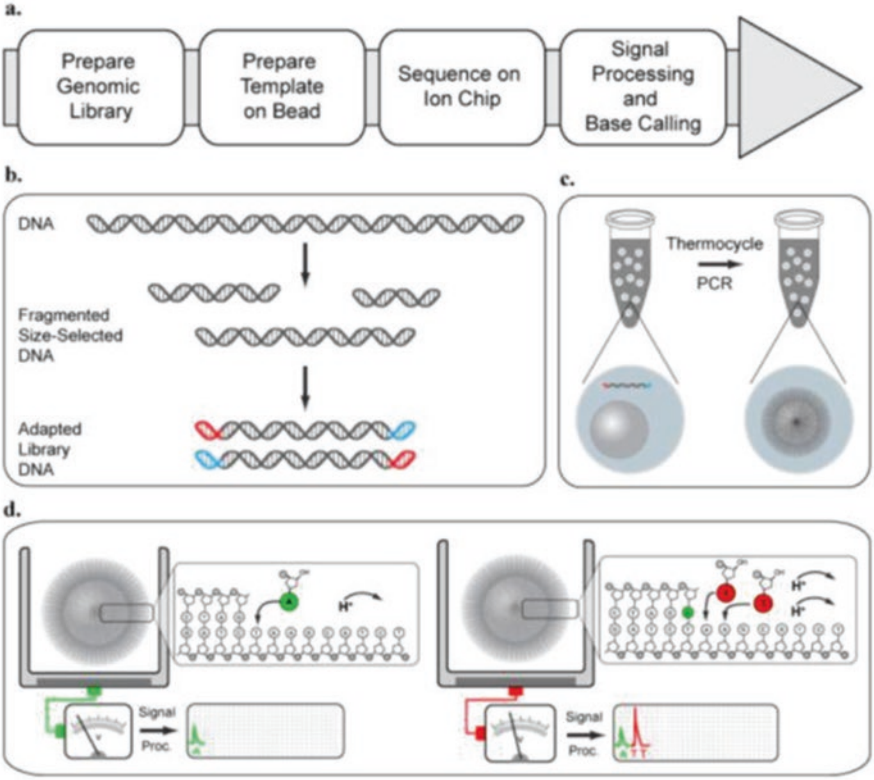


Fig. 6.11 Workflow and principle of Ion Torrent Sequencing technique

across the chip each depositing into a well. Each microwell on the semiconductor chip hosts numerous copies to sequenced ssDNAs molecules. The chip is then flooded with unmodified deoxyribonucleoside triphosphate. The complementary nucleotides are incorporated by DNA polymerase into the growing strand. However, there will be no nucleotide incorporation when noncomplementary strand is found. When a dNTP is incorporated into a single strand of DNA, a hydrogen ion is released. The release of H_2 ion results in changes in the pH of the solution in each well. An ion-sensitive layer beneath the good measures that change in pH and converts it into voltage. The characteristic change in voltage on the incorporation of a nucleotide is recorded (Fig. 6.11) (Ambardar et al. 2016).

6.4.4.4 Advantages and Disadvantages

The main advantage of ion torrent sequencing is that it utilizes relatively simple sequencing chemistry and requires a very small sample size. It takes a maximum of 2–3 hours for sequencing with a rapid turnaround time of approximately 2 days. It also has flexible semiconductor chips on PGM and proton systems, making it easier to operate for different throughput needs.

The disadvantage, however, is that the error rates for some specific regions are higher as seen in the case of homopolymer repeats found in some sequences.

6.5 Third-Generation Sequencing

Even though second-generation sequencing technologies have enabled sequencing several genomes at a reduced cost, analysis of large structural variations and de novo sequencing was challenging. The subsequent era in DNA sequencing focuses on removing the need for DNA amplification and producing longer reads in a single run. The technology is however still under research and development.

6.5.1 Single-Molecule Real-Time Sequencing

6.5.1.1 History

Single-molecule real-time sequencing (SMRT) is a third-generation DNA sequencing method used for long-read sequencing of DNA and enables real-time single-molecule DNA sequence determination. This technology was developed and patented by Pacific Biosciences of California, Inc. in the year of 2011, wherein PacBio RS was their first product sold commercially.

In April 2013, a newer version of the sequencer, PacBio RS II, was released with double the probability of an outcome. At the latest, the company announced in September 2015 the launch of a modified and innovative sequencer called the Sequel System with a sevenfold increase in the outcome capacity when compared to PacBio RS II.

6.5.1.2 Principle

The principle behind the Pacific biosystems SMRT is sequencing by synthesis is quite different from other sequencing techniques. It uses a single molecule for detection, so no amplification step is required to prepare the amplicon library. In this method, the DNA polymerase is immobilized using the biotin-streptavidin system in the bottom of the microwell with zeptoliter (10^{-21}liters); volume waveguides are very small compartments, present on the SMRT cell, which guide attenuated light to pass through wherein the wavelength of light is much larger than the volume of the chamber (Fig. 6.12).

Phospho-linked nucleotides nothing but fluorescently labeled nucleotide are added to the well, and fluorescent label is detached from nucleotide once it is incorporated into the single standard template DNA strand which is coupled with the

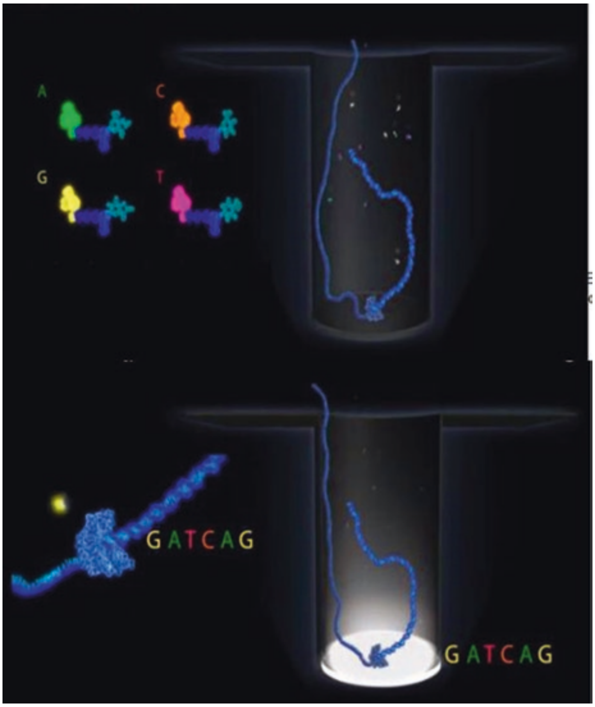


Fig. 6.12 Single Molecule Real Time Sequencing (SMRT) Technology. (SMRT Sequencing - PacBio, https://www.pacb.com/smrtscience/smrt-sequencing/)

immobilized DNA polymerase. The released fluorescent was detected and captured from the bottom of the well. The four nucleotides are added simultaneously; the detection is made in real time with the high-speed sequencing comparatively as individual nucleotides are flushed sequentially.

6.5.1.3 Procedure

The entire workflow of SMRT sequencing is as follows: SMRTbell template preparation. The foremost step is DNA shearing wherein DNA is fragmented using Covaris S2, LE220 system, Covaris g-tube devices, or HydroShear instrument. These devices are variably used depending upon the size of the fragments required. Then the DNA is concentrated using AMPure PB beads. Next, DNA is repaired for any damage like nicks or thymine dimer formation, deaminated cytosine formation, etc. Further end repair is performed to make the ends of the fragments suitable for ligation of adapters. T4 DNA polymerase is used for 5′ filling up of overhanging and 3′ removal. Next, ligation of hairpin looped adapters occurs at the ends of the DNA fragment. These adapters allow both forward and reverse read in the same trace. Next, the purification is done by treating the fragments with exonuclease III and VII. Finally, the primer is annealed to the ligated adapters in the SMRT template, and DNA polymerase is added as well.

The PacBio RS II runs, and then automated primary, secondary, and tertiary analysis occurs. Within the ZMW, the DNA template polymerase complex is fixed using MagBead or diffusion system. Phospho-linked nucleotides are added to the chamber with each of the four nucleotides phosphate backbone tagged with a fluorophore. If a base is complementary and gets attached to the fixated complex, fluorescence occurs, and the tagged phosphate backbone is cleaved and removed. Massively parallel sequencing of SMRT leads to greater output in the single cycle (Buermans and Dunnen 2014; Ambardar et al. 2016).

6.5.1.4 Advantages and Disadvantages

The main advantage of SMRT is that it can efficiently be used for de novo assembly of long-read sequences, and it gives uniform outcomes which are important when it comes to GC-rich region sequencing. The platform can be used for base modification and isoform detection for nearly all organisms.

The disadvantage with SMRT is that it is cost-effective and startup costs and subsequent costs are high. It also has a 5–15% of error rate specifically in cases of insertions or deletions (Ari and Arikan 2016).

6.5.2 Nanopore Sequencing

6.5.2.1 History

Nanopore sequencing technology has been researched on since before next-generation technologies came into play. In the early 1990s, David Dreamer and George Church, independently theorized that a ssDNA could be sequenced by passing through a nanopore. They would later go on to file a patent for nanopore sequencing. In 1996, Dreamer, Branton, and Kasiannowicz published their results on DNA translocation detection through alpha hemolysin nanopore.

The breakthrough in nanopore sequencing technology came in 2001 with the discovery of solid-state nanopore, otherwise, synthetic nanopore which can be fabricated on the Si3N4 membrane. In 2005, Oxford Nanopore Technologies was set by Hagan Bayley along with Spike Willcocks, David Norwood, and Gordon Sanghera. It is the first company to offer commercial sequencers with nanopore-based technology (Fig. 6.13) (Lu et al. 2016).

Fig. 6.13 Nanopore sequencer. (media@nanoporetech.com)

6.5.2.2 Principle and Procedure

Nanopore sequencing relies on porins which are transmembrane proteins that created a porous channel across a membrane. Nanopore sequencing system consists of either biological nanopores (**α**-hemolysin or *Mycobacterium smegmatis* porin A) or solid-state nanopores (Si3N4 and SiO2 nanopores) which are set in a polymer membrane of high electrical resistance. A voltage is maintained across the membrane by passing an ionic current. When an analyte passes through the pore, the current is disrupted, and the disruption is utilized to identify the specific molecule.

The DNA sample to be analyzed is kept intact and mixed with copies of a processive enzyme. DNA enzyme complex approaches the nanopore, the single-stranded DNA is pulled through the aperture, and the enzyme latches the DNA strand through the nanopore one base at a time. The processing of nucleotides through the nanopore creates characteristic disruptions in the flow of electric current. The signal hence generated is used to determine the order of the bases (Fig. 6.14) (Buermans and Dunnen 2014; Feng et al. 2015; Ambardar et al. 2016; Deamer et al. 2016).

6.5.2.3 Advantages and Disadvantages

The main advantage that nanopore technology has to offer is inexpensive sample preparation and elimination of the need for nucleotides or ligases. From this alone, the cost per strand for nanopore technology would be far less than Sanger method or any next-generation sequencing technology. The main limitation of nanopore sequencing is the translocation speed of the nanopore. Nanopore sequencing is still a relatively new technology with a lot of potential and scope for development in the near future (Branton et al. 2008).

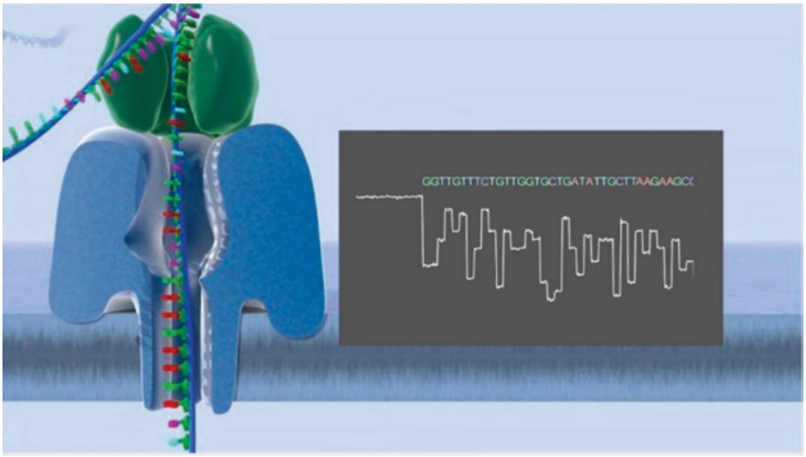


Fig. 6.14 Nanopore sequencing technique. (media@nanoporetech.com)

6.5.3 NGS Data Analysis

Data analysis is important in terms of next-generation sequencing. After sequencing reaction, raw sequence data generated, there are important analysis steps that the data must processed. A generalized data analysis pipeline includes removing the adapter sequences, removing low-quality reads, reference genome mapping or de novo alignment, and analysis of the compiled sequence (Fig. 6.15). Based on the application, different bioinformatics pipelines are used. Variant calling for detection of mutations (SNPs or indels), expression analysis for transcripts, and somatic and germline mutations analysis for clinical diagnosis. Various online and offline bioinformatics tools are available for several different NGS data analysis.

6.5.4 Applications of High-Throughput DNA Sequencing

NGS has enormous application in many fields of genomics. Gene identification in terms of regulatory elements and pathological identification are done through resequencing. Whole genome sequencing of various organisms, as well as bacteria and viruses in the field of public health, is made possible by NGS. Gene expression, noncoding RNA, and epigenetic modification are some of the main application of NGS. Circulating cell-free DNA (cfDNA) and prenatal DNA analysis are the major current application. Moreover, in the future, NGS will be important toward personal genomics and expression studies in the personalized medicine (Grada and Weinbrecht 2013; Ambardar et al. 2016).

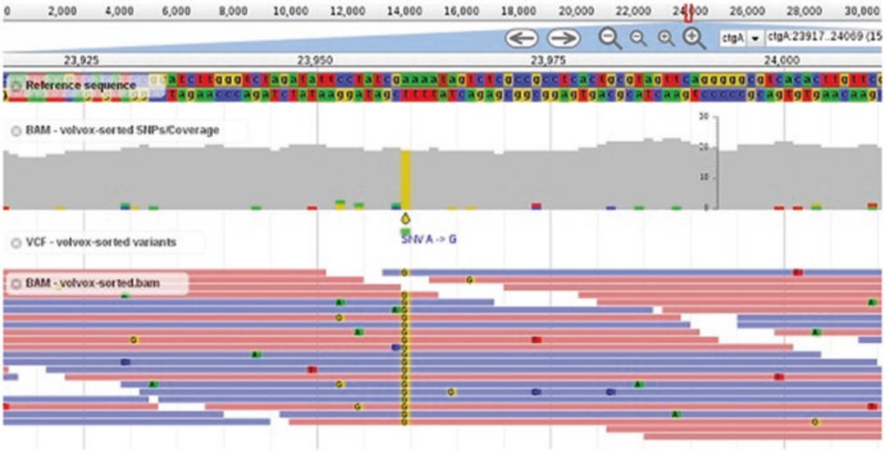


Fig. 6.15 An example of NGS data alignment and analysis tool. http://www.keywordlister.com/bmdzIGRhdGEgYW5hbHlzaXM/

6.5.4.1 Whole-Exome Sequencing (WES)

Sequencing entire human genome is possible with the available NGS technology, but researchers and clinicians are interested in the protein coding regions which are referred to as exome. Only 1% of the genome is coding for protein, and the remaining 99% of the human genome is noncoding regions or sometimes called junk DNA. Mutations occurring in these protein-coding regions give rise to truncated or nonfunctional proteins which may give rise to diseases and various clinical presentations. Exome sequencing provides a fast and affordable way of determining the genetic cause of a disease or a clinical condition. Exome sequencing is very useful in identifying disease-causing mutations in pathogenic conditions when the extract genetic cause is not known (Grada and Weinbrecht 2013).

6.5.4.2 Whole-Genome Sequencing (WGS)

In addition to whole exome sequencing, with the available next-generation sequencing technology, it is possible to sequence the entire 3 billion basepairs of an individual, which is termed as whole genome sequencing (WGS). Sequencing the whole genomes can provide more valuable information on diversity, cancer progression, and other genetic disorders. WGS can capture small and large variant which might otherwise miss by the other methods. Apart from studying human, it is equally useful for sequencing other species such as important livestock, plants, and pathogens. WGS sequencing is possible with Illumina and SMRT sequencing technology. With the long reads in the nanopore sequencing technology, WGS is possible, but it is being refined. But handling the whole genomic data is challenging but with the growing knowledge in the field of bioinformatics is made possible to handle such big data.

6.5.4.3 Targeted Sequencing

Targeted sequencing is another approach in NGS, in which selected specific genes or genomic regions are targeted. Targeted sequencing is affordable and gives higher coverage of genomic regions of interest, and it also narrows down the analysis of specific gene or genomic region. There are plenty of targeted sequencing panels available which target the hotspot regions for specific disease or combination of diseases. Moreover, there are target panels available for the clinical diagnostic purpose, which gives rapid diagnosis of many diseases. Cancer hotspot mutation panel is to target hotspot cancer-causing mutation. The pharmacogenomic panel is another example to detect drug efficacy safely based on the genome (Grada and Weinbrecht 2013).

6.5.4.4 RNA Sequencing/Expression Analysis

In recent scientific world, RNA sequencing is one of the important sequencing applications. In general, RNA sequencing is done by converting it to complementary DNA, also referred to as cDNA by the enzyme reverse transcriptase (RNA-dependent DNA polymerase). cDNA is commonly used in sequencing studies to discover the coding sequence of expressed genes and study the level of gene expression. Microarray technology is one of the common methods used to measure known targeted genome-wide gene expression which is now getting replaced by RNA sequencing. The NGS-based RNA sequencing (RNA-seq) has multiple applications in scientific experiments. It provides an accurate measurement of gene expression for the entire transcriptome than microarray. RNA-seq technique is not only useful in the detection of the mutations, but it also helps in measuring amount of spliced transcripts and in interrogating wide variety of nonprotein-coding RNAs.

The discovery and functional analysis of noncoding RNA (ncRNA) have been the exiting areas of biological research. There are protocols to sequence small nonprotein-coding RNA molecules to understand the function of noncoding RNAs such as miRNA, siRNA, snRNA, snoRNA, piRNA, etc. Most protocols for RNAseq in eukaryotic cells use poly(T) oligonucleotides to isolate mRNA with poly(A) tails or use poly(T) primers in combination with random short oligomers for reverse transcription. After poly(A) enrichment and cDNAsynthesis, most protocols shatter cDNA molecules into small fragments (from 100 to 300 bp) that are then ligated with oligomers specific for the sequencing system (Brown and Goecks 2015).

6.5.4.5 Metagenomics

NGS technology is useful in microbial genomics, especially in the metagenomics measuring the genetic diversity encoded by microbial life in organisms inhabiting a common environment, for example, metagenomic analysis gut microbiome which gives the clear details of the various pathogen infections. Analyzing the collection of microbes in and on the human body will contribute to understanding human health and disease. Changes in the microbial community are linked to the immune system, obesity, and cancer (Bragg and Tyson 2014).

6.5.4.6 Methylation Studies/Bisulfite Sequencing

Nowadays, using NGS technology, it is possible to study genome-wide DNA methylation. The methylation study was normally done using bisulfite whole genome sequencing or methylated CpG island recovery assay. Sequencing both untreated and bisulfite-treated DNA will highlight the C-nucleotides that are methylated and not chemically converted resulting in a T when sequenced (Buermans and Dunnen 2014; Masser et al. 2015).

6.5.4.7 Ancient Genomes

There are challenges in studying the precious ancient DNA samples from a fossil. The advent of NGS made it possible to directly sequence the nuclear genome, which previously permitted to do only mitochondrial DNA (Der Sarkissian et al. 2015).

6.5.4.8 ChIP-seq

Another application by NGS technology is to study the protein binding sites in genomic DNA, especially transcription binding site based on chromatin immune precipitation (ChiP). Initially, it was done using microarray technology. Many proven studies using ChiP-seq by NGS technology reveal genome-wide profiles of protein binding sites with increased coverage (Buermans and Dunnen 2014).

6.5.4.9 Noninvasive Prenatal Testing

It is well-known that DNA of the fetus can be found in maternal blood in very low level, and it is very difficult to differentiate maternal from fetus DNA. But using the power of NGS technology and the analysis method, it is possible to study circulating DNA of the fetus from maternal blood. This is useful in identifying trisomies, trisomy 21 (Down syndrome), trisomy 18 (Edwards syndrome), and trisomy 13 (Patau syndrome). It could be determined using 20 ml of maternal blood from gestation week 10 using various NGS platforms (Buermans and Dunnen 2014).

6.6 Conclusion

The DNA sequencing techniques are key tools in the scientific world revolutionizing many fields of science and are increasingly used in health care especially in the field of oncology, inherited disorders, and infectious diseases. The current chapter traverses in the chronological order, describing different generations of sequencing technology, underlining few key discoveries, scientists, and sequences along the way.

References

Ambardar S, Gupta R et al (2016) High throughput sequencing: an overview of sequencing chemistry. Indian J Microbiol 56(4):394–404

Ansorge WJ (2009) Next-generation DNA sequencing techniques. New Biotechnol 25(4):195–203

Ari Ş, Arikan M (2016) Next-generation sequencing: advantages, disadvantages, and future. In: Plant omics: trends and applications. Springer, Berlin, pp 109–135

Branton D, Deamer D, Marziali A, Bayley H, Benner S, Butler T, Di Ventra M, Garaj S, Hibbs A, Huang X, Jovanovich S, Krstic P, Lindsay S, Ling X, Mastrangelo C, Meller A, Oliver J, Pershin Y, Ramsey J, Riehn R, Soni G, Tabard-Cossa V, Wanunu M, Wiggin M, Schloss J (2008) The potential and challenges of nanopore sequencing. Nat Biotechnol 26(10):1146–1153

Bragg L, Tyson GW (2014) Metagenomics using next-generation sequencing. Methods Mol Biol 1096:183–201

Brown SM, Goecks J (2015) RNA sequencing with next-generation sequencing Chapter 13: RNA sequencing with next-generation sequencing. Cold Spring Harbor Laboratory press, Second Edition

Buermans HPJ, den Dunnen JT (2014) Next generation sequencing technology: advances and applications. Biochim Biophys Acta (BBA) - Mol Basis Dis 1842(10):1932–1941

Deamer D, Akeson M, Branton D (2016) Three decades of nanopore sequencing. Nat Biotechnol 34(5):518–524

Der Sarkissian C, Allentoft ME et al (2015) Ancient genomics. Philos Trans R Soc Lond B Biol Sci 370(1660):20130387

Dewey FE, Pan S et al (2012) DNA sequencing: clinical applications of new DNA sequencing technologies. Circulation 125(7):931–944

Feng Y, Zhang Y et al (2015) Nanopore-based fourth-generation DNA sequencing technology. Genomics Proteomics Bioinformatics 13(1):4–16

França LTC, Carrilho E, Kist TBL (2002) A review of DNA sequencing techniques. Q Rev Biophys 35(02)

Gaastra W (1985) Chemical cleavage (Maxam and Gilbert) method for DNA sequence determination. Methods Mol Biol 2:333–341

Golan D, Medvedev P (2013) Using state machines to model the Ion Torrent sequencing process and to improve read error rates. Bioinformatics 29(13):i344–i351

Grada A, Weinbrecht K (2013) Next-generation sequencing: methodology and application. J Invest Dermatol 133(8):e11

Guzvic M (2013) The History of DNA Sequencing. J Med Biochem 32(4):301–312

Harrington CT, Lin EI et al (2013) Fundamentals of pyrosequencing. Arch Pathol Lab Med 137(9):1296–1303

Heather JM, Chain B (2016) The sequence of sequencers: the history of sequencing DNA. Genomics 107(1):1–8

Huang Y, Chen S, Chiang Y, Chen T, Chiu K (2012) Palindromic sequence impedes sequencing-by-ligation mechanism. BMC Systems Biology 6(Suppl 2):S10

Lu H, Giordano F, Ning Z (2016) Oxford Nanopore MinION sequencing and genome assembly. Genomics Proteomics Bioinformatics 14(5):265–279

Masser DR et al (2015) Targeted DNA methylation analysis by next-generation sequencing. J Vis Exp (96):e52488. https://doi.org/10.3791/52488

Maxam AM, Gilbert W (1977) A new method for sequencing DNA. Proc Natl Acad Sci U S A 74(2):560–564

Men AE, Wilson P, Siemering K, Forrest S (2008) Sanger DNA sequencing. In: Janitz M (ed) Next generation genome sequencing: towards personalized medicine, 1st edn. Wiley-VCH Verlag GmbH & Co, Weinheim

Ravi I, Baunthiyal M, Saxena J (2014) Advances in biotechnology. Springer, New York

Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes JC et al (1977) Nucleotide sequence of bacteriophage |[phi]|X174 DNA. Nature 265(5596):687–695

Slatko BE, Albright LM et al (2001) DNA sequencing by the dideoxy method. Curr Protoc Mol Biol Chapter 7: Unit7 4A

Voelkerding KV, Dames SA, Durtschi JD (2009) Next-generation sequencing: from basic research to diagnostics. Clin Chem 55(4):641–658

Wallis Y, Morrell N (2011) Automated DNA sequencing. Methods Mol Biol 688:173–185

Chapter 7
Tools and Methods in the Analysis of Simple Sequences

Yogesh Kumar, Om Prakash, and Priyanka Kumari

Contents

Y. Kumar (✉) · O. Prakash
Department of Metabolic & Structural Biology, CSIR-Central Institute of Medicinal and Aromatic Plants, Lucknow, India

P. Kumari
Department of Plant Biotechnology, CSIR-Central Institute of Medicinal and Aromatic Plants, Lucknow, India

N. A. Shaik et al. (eds.), *Essentials of Bioinformatics, Volume I*,
https://doi.org/10.1007/978-3-030-02634-9_7

7.1 Introduction

7.1.1 Biological Sequences and Their Analysis

A biological sequence is a single, continuous molecule of nucleic acid or protein. The nucleic acid sequence is composed of nucleotides. The nucleotides adenine, thymine, guanine, and cytosine (ATGC) act as building blocks of deoxyribonucleic acid (DNA), and adenine, uracil, guanine, and cytosine (AUGC) act as building blocks for ribonucleic acid (RNA) molecules. In contrast to this, the primary protein structure is composed of a linear chain of amino acid molecules.

The elucidation of molecular sequence information is the gift of advancements in modern molecular bioanalytical technologies, which has not only made the analysis of biological sequences an quickly achievable task but made it more accurate. The most evident example of these advancements is, the Human Genome Project, which produced an immense amount of data for research in human health care (Chial 2008). The methodologies implemented under sequence analysis include sequence alignment (pairwise sequence and multiple sequence alignment), phylogenetic analysis, motif and domain search/prediction and genome or transcriptome comparative study, and identification of novel genes for the drug. Sequence alignment is also an essential step in molecular phylogenetic, for analysis of homologues, orthologues, and paralogues genes as well as identification of mutations in various leading genetic disorders.

This chapter covers different computational approaches of sequence alignment like (1) pairwise alignments, global and local alignment by dynamic programming with different scoring schemes; (2) sequence profile alignment, where one sequence aligned with a set of query sequences; (3) multiple sequence alignment, which covers several methods for alignment like progressive alignment and iterative and profile alignment; and (4) phylogenetic analysis, which is an integral part of multiple sequence alignment. This chapter covers phylogenetic and sequence evolutionary relationship analysis, using different methods like maximum likelihood and neighbor-joining methods.

7.2 Pairwise Sequence Alignment

7.2.1 An Introduction to Pairwise Alignment

The first question which comes in mind of every biologists after the sequencing process is how similar are the two sequences? This simple question arises in bioinformatics during assembly of overlapping sequence fragment into contigs, and

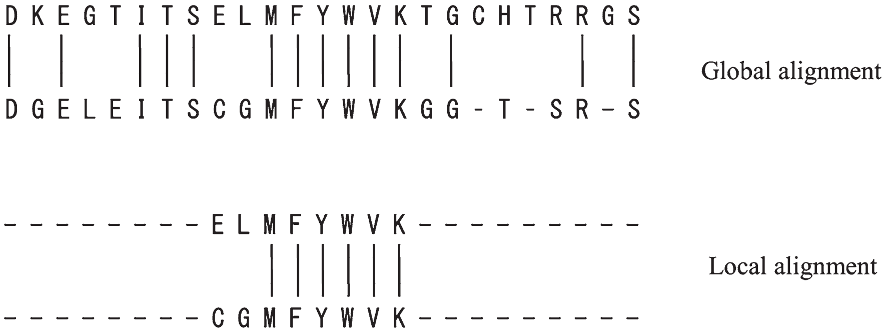
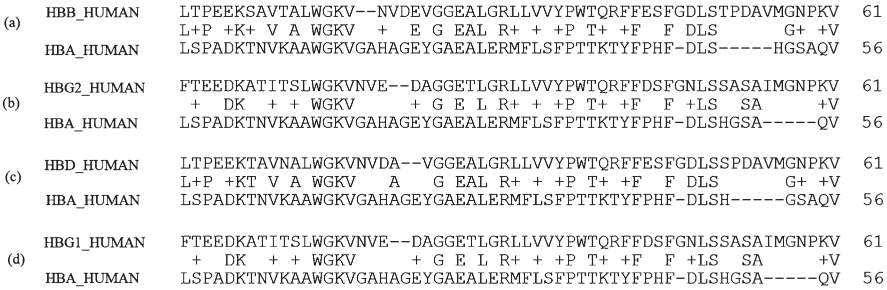

D K E G T I T S E L M F Y W V K T G C H T R R G S
D G E L E I T S C G M F Y W V K G G - T - S R - S
Global alignment
- - - - - - - - E L M F Y W V K - - - - - - - - -
- - - - - - - - C G M F Y W V K - - - - - - - - -
Local alignment

Fig. 7.1 Represents the global and local sequence alignment

alignment of new sequences against reference genomes. After finishing alignment, we have to decide whether the alignment is more likely to have occurred because the sequences are correlated just by chance. These alignments may be global or local (as mentioned in Fig. 7.1). These sequence alignment methods are essential for retrieving essential information from biological sequences, like sequence homology, annotation, pathway identification, phylogenetic relationship, modeling of 3D structures, motif, domain identification, and many more discussed further in this. The goal of pairwise alignment is to find the conserved region (if present) between two or more sequences; these conserved regions are supposed to be an important and functional region (domain or motif) in the sequences. A simple example of a pairwise sequence alignment given in Fig. 7.2. The human hemoglobin subunit alpha (HBA_HUMAN) (P69905) was used as a query (sequence to be aligned with other sequences) sequence, and other four subunit sequences of hemoglobin (HBB_HUMAN) (P68871), (HBG2_HUMAN) (P69892), (HBD_HUMAN) (P02042), and (HBG1_HUMAN) (P69891) were considered as a subject sequence for comparison. The alignment was performed by NCBI BLAST tool using BLOSUM 62 scoring matrix (discussed later). In Fig. 7.2a, HBB_HUMAN was aligned with HBA_HUMAN sequence; in this alignment, we can see that there are many positions at which two corresponding residues are identical; many others are functionally conservative such as T-S representing the alignment of threonine with a serine both polar and tiny residues. Similarly, Fig. 7.2b also represents important alignment, showing the evolutionary relationship among two sequences, having most identical residues with fewer gaps; some of the new residues were showing unmatched which indicates some insertion of new residues during the evolution of time. To distinguish between the alignments of sequences represented in Fig. 7.2a, b, using pairwise alignment method a scoring matrix is often used.



```
(a)  HBB_HUMAN    LTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV  61
                  L+P +K+ V A WGKV  +  E G EAL R+ + +P T+ +F  F DLS      G+ +V
     HBA_HUMAN    LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS-----HGSAQV  56

(b)  HBG2_HUMAN   FTEEDKATITSLWGKVNVE--DAGGETLGRLLVVYPWTQRFFDSFGNLSSASAIMGNPKV  61
                   +  DK  + + WGKV     + G E L R+ + +P T+ +F  F +LS  SA     +V
     HBA_HUMAN    LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLSHGSA-----QV  56

(c)  HBD_HUMAN    LTPEEKTAVNALWGKVNVDA--VGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMGNPKV  61
                  L+P +KT V A WGKV   A   G EAL R+ + +P T+ +F  F DLS      G+ +V
     HBA_HUMAN    LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLSH-----GSAQV  56

(d)  HBG1_HUMAN   FTEEDKATITSLWGKVNVE--DAGGETLGRLLVVYPWTQRFFDSFGNLSSASAIMGNPKV  61
                   +  DK  + + WGKV     + G E L R+ + +P T+ +F  F +LS  SA     +V
     HBA_HUMAN    LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLSHGSA-----QV  56
```

Fig. 7.2 Represents the pairwise sequence alignment of five orthologous sequences

7.2.2 Methods of Sequence Alignment

Sequence alignment has been classified into two major classes namely Global and Local alignment as shown in Fig. 7.1. Global alignment follows an approach of aligning the entire sequences by matching as many characters as possible in both ends of each sequence. The sequences having the same length are best suited for global alignment. In local alignment, sequence matched density will be considered, generating one or more sequence islands in whole sequence stretch (Polyanovsky et al. 2011). Local alignment is, suitable, for sequence alignment that is similar in their lengths, but dissimilar in others, a sequence that differs in length or sequences that share a conserved region or domains.

7.2.2.1 Global Alignment

An alignment is meant to say global alignment when closely related sequences of the same length are aligned together; the alignment of the sequence is carried out from the start to end of the sequence while searching for best possible alignment. The algorithm of aligning two protein sequences, published by Needleman and Wunsch in 1970 (Needleman and Wunsch 1970) was the first dynamic programming application for biological sequence analysis. This algorithm beautifully divides the larger problem (e.g., the full sequence) into a series of smaller problems, which are then solved and reconstructed to the larger problems. It is also known as optimal matching algorithm and the global alignment. The best illustration of the global sequence alignment is represented in Fig. 7.1, where two sequences are aligned with each other. Some software like EMBL-EBI EMBOSS (https://www.ebi.ac.uk/Tools/psa/emboss_needle/), which is based on Needleman-Wunsch algorithm, creates an optimal global alignment of two sequences, where another server EMBOSS Stretcher uses a modification of the Needleman-Wunsch algorithm that allows more massive sequences to be globally aligned. The list of some global alignment software are mentioned in Table 7.1.

7.2.2.2 Local Alignment

Local alignment is mainly used for those sequences which differ in sequence length. This method finds local matches within the sequence stretch instead of looking at the entire sequence; Smith-Waterman algorithm, a dynamic programming algorithm which was developed by Smith and Waterman in 1981, was used for local alignment; this algorithm compares segments of all possible lengths and optimizes the similarity measures. Local alignment uses scoring matrices (PAM, BLOSUM) (*discussed in* Sect. 2.4.1) which give the user a choice to choose the appropriate scoring system based on the goals. It is suggested that the user may try different combinations of scoring matrices while using local alignment. There are many software which use Smith-Waterman algorithm to build alignment of sequences; some of these are mentioned in Table 7.1; here the most popular and widely used software is NCBI-BLAST (basic local alignment search tool); it can be used online or offline on the local machine (Altschul et al. 1990).

BLAST (basic local alignment search tool) is the most commonly used tool for sequence alignment and similarity search. BLAST tool is fast and can be used in analysis of more than 1000s of sequences and even for comparison of two genomes (Altschul et al. 1994). BLAST is freely available for everyone and downloadable (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/). This tool is straightforward to handle and produces very informative data; it can be used on the local machine by merely downloading the setup, or user who is not much handy with command line (available at https://www.ncbi.nlm.nih.gov/BLAST/). The BLAST method is a word search heuristic method which eliminates the irrelevant sequences and saves search time. The program search sequences by setting a word length W (usually 3 for amino acid and11 for nucleotide sequences).This program uses a heuristic approach that approximates the Smith-Waterman algorithm for sequence alignment between the query and subject sequence (existing sequence in the database). BLAST algorithm is a sequential stepwise method as mentioned below.

1. The program removes low-complexity region or sequence repeats in the query sequences.
2. It builds a K letter word list of the query sequence, where $k = 3$ for protein sequence and $k = 11$ for DNA sequence (k means the number of characters of sequence).
3. List of the probable matching word will be created from matched query sequences, and the score was generated using scoring substitution matrix.
4. Organizes the remaining high-scoring words an efficient search tree.
5. Then the step 3 and step 4 are to be repeated for each k-letter word in the query sequence.
6. In next program, scan the database sequences for exact matches with the remaining high-scoring words.
7. Extend the exact matches to high-scoring segment pair (HSP).

Table 7.1 Different computational web tools and software useful in the local and global pairwise sequence alignment

Global alignment software	URL	Local alignment software	URL
GGSEARCH	http://nebc.nerc.ac.uk/bioinformatics/docs/ggsearch.html	BLAST	https://blast.ncbi.nlm.nih.gov/Blast.cgi
HMMER	http://hmmer.org/	HMMER	http://hmmer.org/
G-PAS	http://gpualign.cs.put.poznan.pl/project-gpu-pairAlign.html	PSI-BLAST	https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE=Proteins&PROGRAM=blastp&RUN_PSIBLAST=on
EMBOSS Needle	https://www.ebi.ac.uk/Tools/psa/emboss_needle/	FASTA	https://www.ebi.ac.uk/Tools/sss/fasta/
NW-align	https://zhanglab.ccmb.med.umich.edu/NW-align/	EMBOSS water	https://www.ebi.ac.uk/Tools/psa/emboss_water/
MUMmer	http://mummer.sourceforge.net/	Matcher	https://www.ebi.ac.uk/Tools/psa/emboss_matcher/
MCALIGN2	http://www.homepages.ed.ac.uk/pkeightl//mcalign/mcinstructions.html	SAM	https://web.archive.org/web/20080509161215/http://www.cse.ucsc.edu/research/compbio/sam.html
NW	http://www.bioinf.org.uk/software/nw/	SWIMM	https://github.com/enzorucci/SWIMM
Stretcher	https://galaxy.pasteur.fr/?form=stretcher	ALLALIGN	http://www.allalign.com/
SABERTOOTH	http://www.fkp.tu-darmstadt.de/sabertooth_source/	SWIPE	http://dna.uio.no/swipe/

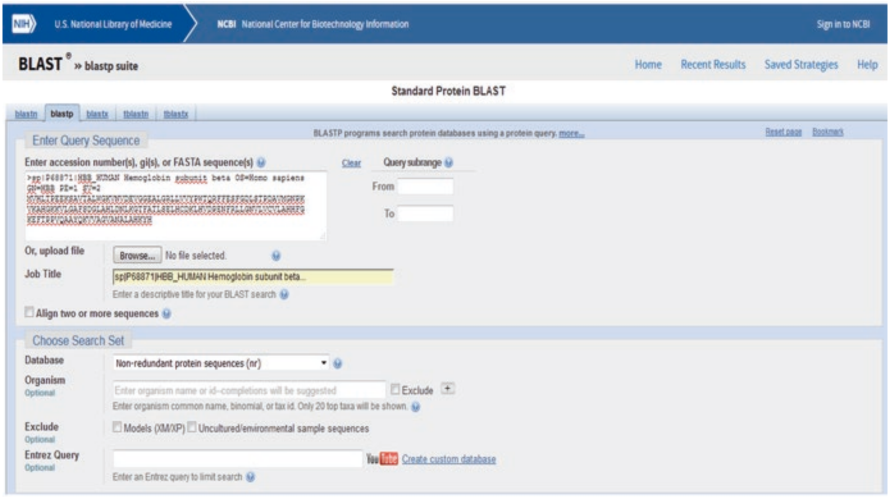
8. Lists out all the HSP matched in the database whose score is high enough to be considered.
9. Evaluate the significance of the HSP score.
10. Make two or more HSP regions into a more extended alignment.
11. The program shows the gapped smith-waterman local alignments of the query sequences and each of the matched database sequences.
12. Finally, the program represents every match, whose expect score is lower than a threshold parameter E (expect value) in the result form.

BLAST has seven subprograms as listed below:

- BLASTn (aligns nucleotide query sequence with nucleotide database).
- BLASTp (aligns protein sequence with protein database).
- PSI-BLAST (protein-specific iterative BLAST) (the program used to find distinctly related protein).
- BLASTx (used to align nucleotide sequence with protein database by comparing six-frame conceptual translation of nucleotide sequence).
- tBLASTx (aligns query nucleotide possible six-frame converted sequence with converted nucleotide six-frame sequences of the database),
- tBLASTn (aligns protein query sequence with translated nucleotide database),
- MegaBLAST (the program used when comparing a large number of input sequences via command line).

There are many software programs and webservers which uses the extended version of BLAST software within their servers and databases. An example has been given below for sequence alignment using the online NCBI BLAST tool in Figs. 7.3, 7.4, 7.5, and 7.6. BLAT is another algorithm which is used in pairwise sequence alignment. This program was developed by Jim Kent to support the annotation of the Human Genome Project (Kent 2002). It can be used to align both DNA and protein sequences and designed to work best of a sequence having more similarity, the sequence of 40 base length that share ≥95% nucleotide identity or ≥ 80% translated protein identity (Bhagwat et al. 2012). BLAT can be used both online and offline available at http://genome.ucsc.edu/cgi-bin/hgBlat. The WU-BLAST (**W**ashington **U**niversity BLAST) version 2.0 is another powerful software package for gene and protein identification; it is based on the public domain NCBI BLAST version 1.4 (Altschul et al. 1990), and it can be downloaded from https://blast.advbiocomp.com/. This program is slightly different from NCBI BLAST, except both versions derived from un-gapped NCBI BLAST 1.4.

Local alignment example: this example shows the sequence alignment using the online NCBI-BLAST tool (https://blast.ncbi.nlm.nih.gov/Blast.cgi); blastp (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome) is selected because we used [P68871] HBB_HUMAN (hemoglobin subunit beta) protein sequence for alignment; in database selection field, we selected NCBI nonredundant protein sequences (nr) database, and the rest of the parameters were kept as default.



Standard Protein BLAST

BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

```
>sp|P68871|HBB_HUMAN Hemoglobin subunit beta OS=Homo sapiens
GN=HBB PE=1 SV=2
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK
VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG
KEFTPPVQAAYQKVVAGVANALAHKYH
```

Query subrange

From

To

Or, upload file: Browse... No file selected.

Job Title: sp|P68871|HBB_HUMAN Hemoglobin subunit beta...

Enter a descriptive title for your BLAST search

☐ Align two or more sequences

Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism Optional: Enter organism name or id--completions will be suggested ☐ Exclude

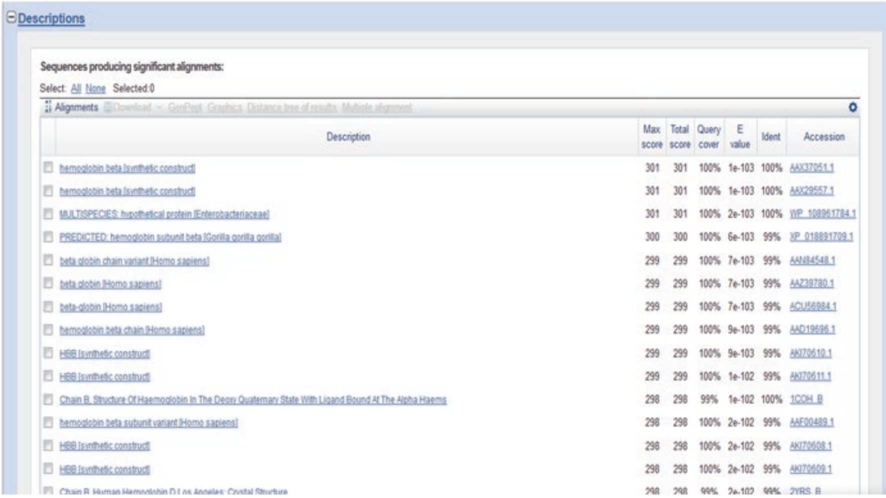
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude Optional: ☐ Models (XM/XP) ☐ Uncultured/environmental sample sequences

Entrez Query Optional

Enter an Entrez query to limit search

Fig. 7.3 NCBI-BLAST window for performing local alignment



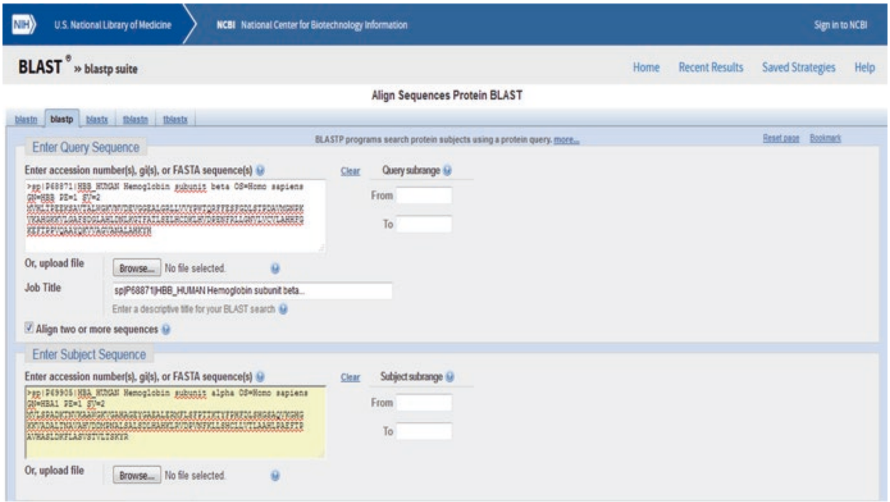
Sequences producing significant alignments:

Select: All None Selected:0

	Description	Max score	Total score	Query cover	E value	Ident	Accession
☐	hemoglobin beta [synthetic construct]	301	301	100%	1e-103	100%	AAX37051.1
☐	hemoglobin beta [synthetic construct]	301	301	100%	1e-103	100%	AAX29557.1
☐	MULTISPECIES: hypothetical protein [Enterobacteriaceae]	301	301	100%	2e-103	100%	WP_108961784.1
☐	PREDICTED: hemoglobin subunit beta [Gorilla gorilla gorilla]	300	300	100%	6e-103	99%	XP_018891709.1
☐	beta globin chain variant [Homo sapiens]	299	299	100%	7e-103	99%	AAN84548.1
☐	beta globin [Homo sapiens]	299	299	100%	7e-103	99%	AAZ39780.1
☐	beta-globin [Homo sapiens]	299	299	100%	7e-103	99%	ACU56984.1
☐	hemoglobin beta chain [Homo sapiens]	299	299	100%	9e-103	99%	AAD19696.1
☐	HBB [synthetic construct]	299	299	100%	9e-103	99%	AKI70610.1
☐	HBB [synthetic construct]	299	299	100%	1e-102	99%	AKI70611.1
☐	Chain B, Structure Of Haemoglobin In The Deoxy Quaternary State With Ligand Bound At The Alpha Haems	298	298	99%	1e-102	100%	1COH_B
☐	hemoglobin beta subunit variant [Homo sapiens]	298	298	100%	2e-102	99%	AAF00489.1
☐	HBB [synthetic construct]	298	298	100%	2e-102	99%	AKI70608.1
☐	HBB [synthetic construct]	298	298	100%	2e-102	99%	AKI70609.1
☐	Chain B, Human Hemoglobin D Los Angeles: Crystal Structure	298	298	99%	2e-102	99%	2YRS_B

Fig. 7.4 Alignment result shows the homologous protein sequences aligned with the query HBB_ HUMAN hemoglobin protein sequence

FASTA (https://fasta.bioch.virginia.edu/) is a first sequence alignment program used for DNA and protein sequence alignment, previously described as FASTP by David J. Lipman and William R. Pearson in 1985 (Lipman and Pearson 1985). It uses FASTA sequence format as an input file which is now standard for every sequence alignment software; it is slow but accurate as compared to BLAST. FASTA is inversely related and depend on the K-tuple variable, which specifies the word



U.S. National Library of Medicine | NCBI National Center for Biotechnology Information | Sign in to NCBI

BLAST® » blastp suite | Home | Recent Results | Saved Strategies | Help

Align Sequences Protein BLAST

blastn | blastp | blastx | tblastn | tblastx

BLASTP programs search protein subjects using a protein query. more... | Reset page | Bookmark

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

```
>sp|P68871|HBB_HUMAN Hemoglobin subunit beta OS=Homo sapiens
GN=HBB PE=1 SV=2
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK
VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG
KEFTPPVQAAYQKVVAGVANALAHKYH
```

Clear | Query subrange | From | To

Or, upload file: Browse... No file selected.

Job Title: sp|P68871|HBB_HUMAN Hemoglobin subunit beta...

Enter a descriptive title for your BLAST search

☑ Align two or more sequences

Enter Subject Sequence

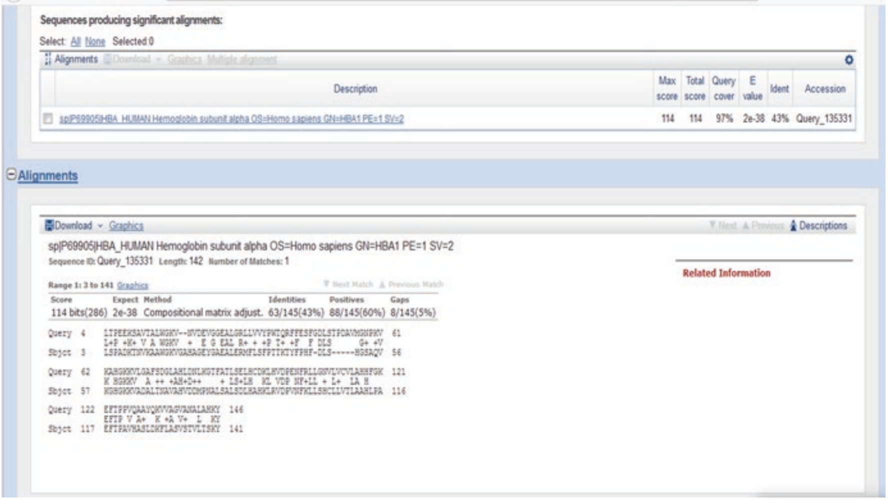
Enter accession number(s), gi(s), or FASTA sequence(s)

```
>sp|P69905|HBA_HUMAN Hemoglobin subunit alpha OS=Homo sapiens
GN=HBA1 PE=1 SV=2
MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
KKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTP
AVHASLDKFLASVSTVLTSKYR
```

Clear | Subject subrange | From | To

Or, upload file: Browse... No file selected.

Fig. 7.5 NCBI-BLAST tool comprises features for aligning two or more sequences with each other. This example shows the alignment of HBB_HUMAN with HBA_HUMAN. For aligning more than two sequences, the user should click on the Align two or more sequences option as shown in the above figure



Sequences producing significant alignments:

Select: All None Selected 0

Alignments | Download | Graphics | Multiple alignment

	Description	Max score	Total score	Query cover	E value	Ident	Accession
□	sp|P69905|HBA_HUMAN Hemoglobin subunit alpha OS=Homo sapiens GN=HBA1 PE=1 SV=2	114	114	97%	2e-38	43%	Query_135331

Alignments

Download | Graphics | Next | Previous | Descriptions

sp|P69905|HBA_HUMAN Hemoglobin subunit alpha OS=Homo sapiens GN=HBA1 PE=1 SV=2

Sequence ID: Query_135331 Length: 142 Number of Matches: 1

Related Information

Range 1: 3 to 141 Graphics | Next Match | Previous Match

Score	Expect	Method	Identities	Positives	Gaps
114 bits(286)	2e-38	Compositional matrix adjust.	63/145(43%)	88/145(60%)	8/145(5%)

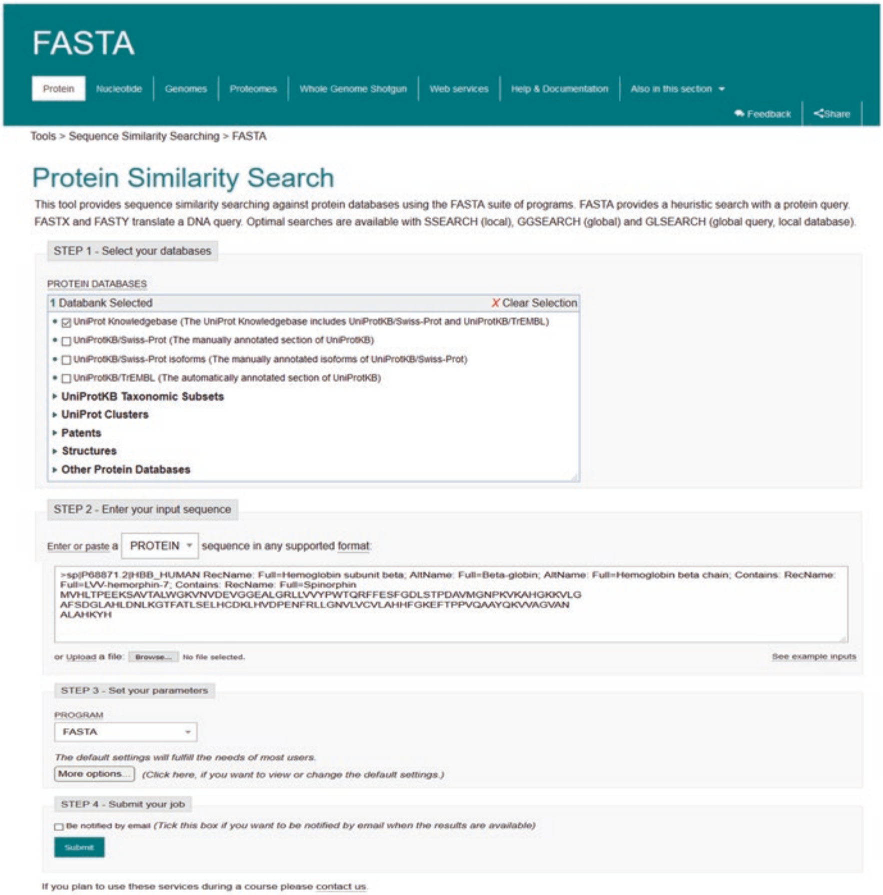
```
Query  4    LTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV  61
            L+P +K+ V A WGKV  +  E G EAL R+ + +P T+ +F  F DLS      G+ +V
Sbjct  3    LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS-----HGSAQV  56

Query  62   KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK  121
            K HGKKV  A ++ +AH+D++    + LS+LH  KL VDP NF+LL + L+  LA H
Sbjct  57   KGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPA  116

Query  122  EFTPPVQAAYQKVVAGVANALAHKY  146
            EFTP V A+  K +A V+  L  KY
Sbjct  117  EFTPAVHASLDKFLASVSTVLTSKY  141
```

Fig. 7.6 The results give the local alignment of two sequences with 97% query covered (means query sequence was covered 97% while aligning with another sequence) and reproduce only 43% identity

size; typically searches are run with word size $k = 3$, but, if high sensitivity at the expense of speed is desired, one may switch to word size $k = 2$. An example is shown below in Fig. 7.7.



Tools > Sequence Similarity Searching > FASTA

Protein Similarity Search

This tool provides sequence similarity searching against protein databases using the FASTA suite of programs. FASTA provides a heuristic search with a protein query. FASTX and FASTY translate a DNA query. Optimal searches are available with SSEARCH (local), GGSEARCH (global) and GLSEARCH (global query, local database).

STEP 1 - Select your databases

PROTEIN DATABASES

1 Databank Selected X Clear Selection

- ☑ UniProt Knowledgebase (The UniProt Knowledgebase includes UniProtKB/Swiss-Prot and UniProtKB/TrEMBL)
- ☐ UniProtKB/Swiss-Prot (The manually annotated section of UniProtKB)
- ☐ UniProtKB/Swiss-Prot isoforms (The manually annotated isoforms of UniProtKB/Swiss-Prot)
- ☐ UniProtKB/TrEMBL (The automatically annotated section of UniProtKB)
- **UniProtKB Taxonomic Subsets**
- **UniProt Clusters**
- **Patents**
- **Structures**
- **Other Protein Databases**

STEP 2 - Enter your input sequence

Enter or paste a PROTEIN sequence in any supported format:

```
>sp|P68871.2|HBB_HUMAN RecName: Full=Hemoglobin subunit beta; AltName: Full=Beta-globin; AltName: Full=Hemoglobin beta chain; Contains: RecName:
Full=LVV-hemorphin-7; Contains: RecName: Full=Spinorphin
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVAN
ALAHKYH
```

or Upload a file: Browse... No file selected. See example inputs

STEP 3 - Set your parameters

PROGRAM

FASTA

The default settings will fulfill the needs of most users.

More options... *(Click here, if you want to view or change the default settings.)*

STEP 4 - Submit your job

☐ Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

Submit

If you plan to use these services during a course please contact us.

Fig. 7.7 Online FASTA tool, which provides sequence similarity searching against protein databases using the FASTA suite of programs (https://www.ebi.ac.uk/Tools/sss/fasta/). Using this tool, users have to follow four steps. Step 1: Select protein database of choice from the selection checkbox. Step 2: After selecting the database to enter the fast format, query sequence in input text window. Step 3: Set the parameters from the program. Step 4: Click on submit button

7.2.2.3 Significance of Global and Local Alignment

Sequence alignment is an essential post-molecular sequencing step to reveal the structural, evolutionary, and functional information about molecular sequences. While performing sequence alignment, users need to follow some steps; at first, users should be careful about the selection of alignment results, which is produced by sequence alignment tools. Users should not trust blindly on the scoring and identity percent matching obtained from the sequence alignment. The best method for sequence alignment having the same length sequences is global alignment as we discussed previously in this chapter that global alignment program is considered to match sequences from end to end and will search for the best possible score as it does not remove any unmatching regions of the sequence. Local alignment algorithm is used for aligning the sequences which are having different length; this algorithm is essential for searching distinctly related sequences, diverged sequences, and convergent sequences; in exploring the functional region of sequence, domain analysis, and motif analysis; and for molecular structure modeling.

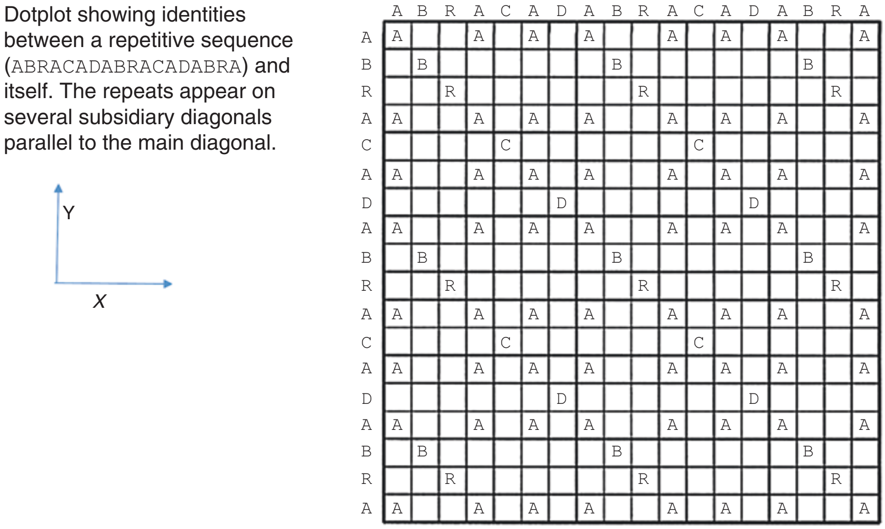
7.2.3 Sequence Similarity and Scoring Methods

7.2.3.1 Dot-Matrix Method

The first method which was applied in sequence comparison was "dot-matrix analysis" or the dot plot method. Gibbs and McIntyre in 1970 first published this method by comparing two sequences (Gibbs and McIntyre 1970). This analysis is done by putting one sequence along the y-axis vertically on the right or left side and another sequence on x-axis horizontally on top as shown in Fig. 7.8. This method generates a simple matrix of sequence, while each item of the matrix is a measure of similarity of those two residues on the horizontal and vertical sequence. In the dot matrix, all identical proteins show a diagonal line in the center of the matrix, and if some insertion or deletion was introduced in the sequence, then it will give rise to disruption in this diagonal. Some more diagonal lines were also seen in the matrix; these diagonal lines are the region which is matched with each other in addition to the central diagonal. The partial diagonal line forming other than the central diagonal line was considered as noise, which can be reduced by only shade runs or "tuples" of residues, e.g., a tuple of three corresponds to three residues in a row. This method is useful because the chance of coming three residues match in continuous form is less than the chance of coming single residue match (Maizel and Lenk 1981).

7.2.3.2 Dynamic Programming

This method was prevalent and used in computer science, mathematics, management science, economics, and bioinformatics. This method was first implemented for global sequence alignment by Needleman and Wunsch (1970) and local alignment by Smith and Waterman (1981), which provides one or more alignments of the sequences. The sequence alignment method is faster in dynamic programming. This method generates a matrix of numbers that represent all possible alignment between the sequences. The score generated by this method defines the quality of alignment; the higher the score, the more optimal the alignment. Two matrices which were designed for amino acid (protein sequence) are PAM 250 (percent accepted mutation 250) (Dayhoff et al. 1978) and BLOSSOM 62 (block substitution matrix 62); they are used for scoring matches and mismatches. Similar matrices are available for aligning DNA sequences (we will



Dotplot showing identities between a repetitive sequence (ABRACADABRACADABRA) and itself. The repeats appear on several subsidiary diagonals parallel to the main diagonal.

	A	B	R	A	C	A	D	A	B	R	A	C	A	D	A	B	R	A
A	A			A		A		A			A		A		A			A
B		B							B							B		
R			R							R							R	
A	A			A		A		A			A		A		A			A
C					C							C						
A	A			A		A		A			A		A		A			A
D							D							D				
A	A			A		A		A			A		A		A			A
B		B							B							B		
R			R							R							R	
A	A			A		A		A			A		A		A			A
C					C							C						
A	A			A		A		A			A		A		A			A
D							D							D				
A	A			A		A		A			A		A		A			A
B		B							B							B		
R			R							R							R	
A	A			A		A		A			A		A		A			A

Fig. 7.8 Represents the simple matrix of amino acids

discuss more substitution matrices further). Dynamic programming can allow a different form of alignments like nucleotide sequence alignment with protein sequences and vice versa. It evaluates frameshift offset by a random number of nucleotides and makes the method suitable for sequences covering a large number of indels, which can be difficult to align with more efficient heuristic methods (Pearson and Miller 1992).

7.2.3.3 Word Methods

Word method is a heuristic method which produces less optimal results for sequence alignment; this method is known as *k*-tuple methods. These are very useful in large-scale database searches. This method was implemented in BLAST and FASTA tools. Word method identifies the short non-overlapping sequences from the query sequence and then matches them with the candidate database sequences. In the FASTA, users can set the value to use it as word length for database search, although this method is slower, but very sensitive at a lower value of *k*, which is preferred for searching of short sequences. In the BLAST word size of *k* have some standard values like for protein sequence it is 2, 3 and 6, and for a nucleotide sequence, it is 16, 20, 24, 28, 32, 48, 64, 128, and 256 the good rule of thumb is that query length must be twice the word size, if your query is a protein sequence of 4 residues, then the Word size should be reduced to 2.

7.2.4 Alignment Scoring Schemes

7.2.4.1 PAM Matrices

PAM matrices, defined as percent or point accepted mutation, were first proposed by Dayhoff and coworkers in 1978. These matrices are used for a sequnce pairs, which are similar to each other. Substitution frequency can be derived for a sequence, which represents the value of frequencies of an evolution. These matrices represent mutations that have been accepted by natural selection. PAM matrices are generally used to score the alignment of amino acid sequences. One PAM of evolution means 1% of amino acid (from 20 amino acids) changed during evolution. This matrix is used in both global and local alignment. PAM1 is the matrix calculated from comparison of sequences with no more than 15% divergence but corresponds to 99% sequence identity. PAM matrices have five sub-matrices, PAM100, PAM120, PAM160, PAM200, and PAM250. If a person wants to compare closely related sequences, lower number PAM is used, while a higher number PAM is used when comparing distantly related protein sequences (Henikoff and Henikoff 1992a, b). The most common PAM matrix is PAM 250; it represents a greater degree of evolutionary divergence and corresponds to multiply the PAM 1 by itself 250 times via a process called dynamic programming.

7.2.4.2 BLOSUM Matrices

BLOSUM (**Blo**ck **Su**bstitution **M**atrices) matrices are used for protein sequence alignment; these matrices were used for alignment of the sequences which are evolutionary diverse. Henikoff and coworkers introduced these substitution matrices in 1992 (Henikoff and Henikoff 1992a, b). It scans the BLOCKS from the database for many conserved regions of protein families that have no gaps in the sequence alignment, and the relative frequency of amino acids and substitution probabilities was counted, and log-odds score for each of the 210 possible substitution pairs of the 20 standard amino acids was calculated. The number added with BLOSUM matrices indicates the percent identity level of sequences in the alignment. For example, BLOSUM62 means that sequences with approximately 62% identity will be counted.

The sequence identity matrix is a simple mathematic formula used for comparing two amino acid sequence residues that are identical with no substitution of a single amino acid residues is acceptable like in substitution matrices. In the process of sequence alignment, some gaps are generated between the characters to align the two sequences, and for matching the conserved portion, these gaps can occur because of single mutations, unequal crossover in meiosis, or gene duplication. These gaps are of two type gap opening and gap extension, and penalty score was generated for both and termed as gap opening penalty and gap extension penalty as proposed by Sellers (1974) and Gotoh (1982). These two penalties have been long used to model insertion and deletions in sequence alignment methods. The penalties were designed to reduce the score when alignment has been disturbed by indels.

7.3 Multiple Sequence Alignment

7.3.1 An Introduction

Advancement of molecular biology contributes to the biology and reveals that all the related or similar sequences contain conserved region across widely divergent species; they often perform a similar or even identical function (Lipman et al. 1989). These conserved regions are almost identical or same in genes despite being in a different organism. During the period, learning of molecular biology and sequence analysis presents potential results using advance tools, techniques, and bioinformatics software. Multiple sequence alignment (MSA) is just an extension of pairwise sequence alignment, where multiple sequences can be aligned at a time. In this, multiple sequences are aligned together by adjusting the characters in a single column and match the maximum character by inserting gaps between the sequences where required, as shown in Fig. 7.9. MSA is useful in identifying the domain and conserved motifs in multiple sequences; these conserved motifs can be used to locate the active catalytic sites of enzymes (Wang and Jiang 1994). In nucleic acids, MSA reveals the structural and functional relationship between the multiple sequences, by which a promoter sequence can identify within the consensus binding sites for regulatory proteins.

7.3.2 Methodologies Used in Multiple Sequence Alignment

7.3.2.1 Progressive Method

This method is known as progressive technique (also known as hierarchical, or tree methods) developed by Paulien Hogeweg and Ben Hesper in 1984. It uses dynamic programming methods to build an MSA starting with the most related sequences and then progressively adding less related sequences or group of sequences to the initial alignment (Waterman and Perlwitz 1984; Higgins et al. 1996; Mount 2009). In pairwise alignment method, insertion in sequences is not distinguished from deletion, but traditional progressive algorithms inherently consider all gaps as deletions and use heuristics to correct for the cost of repeated handling of insertions (Löytynoja and Goldman 2005). The challenge to the MSA method is to use an

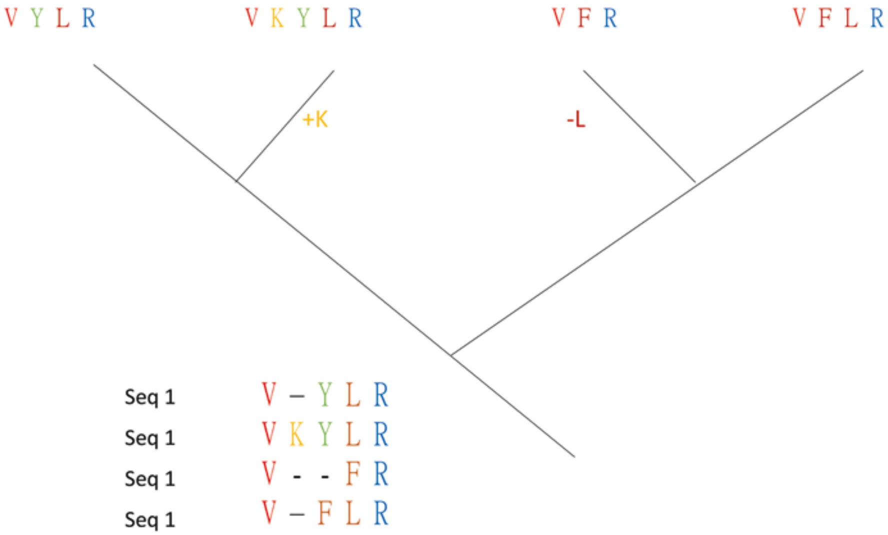


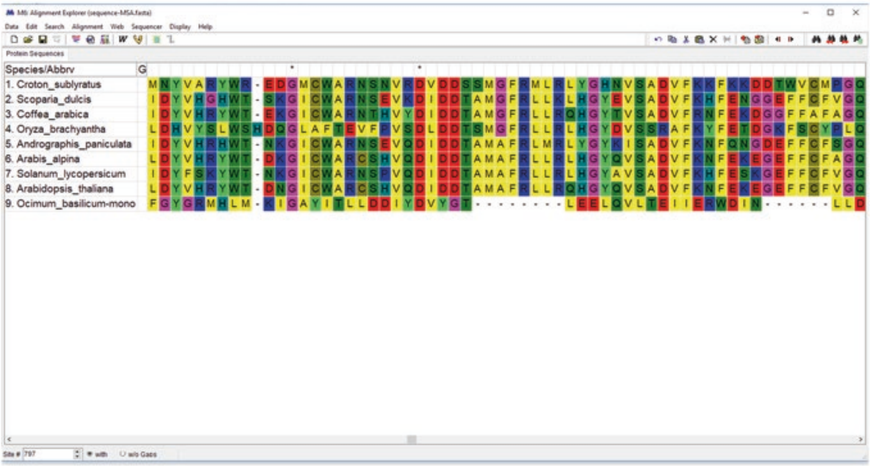
Fig. 7.9 The simple phylogenetic tree shows the relatedness between the sequences

appropriate combination of sequence weighting, scoring matrix, and gap penalties so that the correct series of evolutionary changes may be found (Mount 2009 and Feng and Doolittle 1996).

The method is biologically sound and has the significant advantage of speed and simplicity; using it one can align hundreds of sequences, even on its personal computers. More importantly, the sensitivity of the approach, as judged by the ability to align distantly related sequences, is very high. All of the methods described here are freely available in a computer program called CLUSTALW which can be run under a wide variety of operating systems (Higgins et al. 1996). In progressive alignment, there are two types of problems: (1) the local minimum problem and (2) the parameter choice problem. The local minimum problem is because of the moderate nature of progressive alignment, every time the alignment is carried out, the small proportion of residues is misaligned. This proportion is tiny for the closely related sequence but will increase for diverse sequences. If during alignment the first two closely related sequences were aligned with error, the rest of the alignment will be disturbed, and alignment goes with errors.

This problem can be overcome by computing position-specific gap opening and extension penalties as the alignment proceeds. Different amino acid weight matrices can also be used: "hard" ones for closely related sequences and "softer" ones for more divergent sequences (Higgins et al. 1996).

The software used for MSA are MEGA, T-Coffee, and ClustalX, MEGA (Fig. 7.10). These are the most popular freely available and widely used software (discussed in Sect. 4.3). It can be downloaded and installed on the local machine from http://www.megasoftware.net/. T-Coffee is another server for MSA available at http://tcoffee.crg.cat/, used for aligning DNA, protein, and RNA sequences; it can


M6: Alignment Explorer (sequence-MSA.fasta)
Data Edit Search Alignment Web Sequencer Display Help
Protein Sequences
Site # 797 ● with ○ w/o Gaps

```
Species/Abbrv                 G             *          *
1. Croton_sublyratus            MNYVARYWR-EDGMCWARNSNVRDVDDSSMGFRMLRLYGHNVSADVFKKFKKDDTWVCMPGQ
2. Scoparia_dulcis              IDYVHGHWT-SKGICWARNSEVKDIDDTAMGFRLLKLHGYEVSADVFKHFENGGEFFCFVGQ
3. Coffea_arabica               IDYVHRYWT-EKGICWARNTHVYDIDDTAMGFRLLRQHGYTVSADVFRNFEKDGGFFAFAGQ
4. Oryza_brachyantha            LDHVYSLWSHDQGLAFTEVFPVSDLDDTSMGFRLLRLHGYDVSSRAFKYFETDGKFSCYPLQ
5. Andrographis_paniculata      IDYVHRHWT-NKGICWARNSEVQDIDDTAMAFRLMRLYGYKISADVFKNFQNGDEFFCFSGQ
6. Arabis_alpina                LDYVHRYWT-DKGICWARCSHVQDIDDTAMAFRLLRLHGYQVSADVFKNFEKEGEFFCFAGQ
7. Solanum_lycopersicum         IDYFSKYWT-NKGICWARNSPVQDIDDTAMAFRLLRLHGYAVSADVFKHFESKGEFFCFVGQ
8. Arabidopsis_thaliana         LDYVHRYWT-DNGICWARCSHVQDIDDTAMAFRLLRQHGYQVSADVFKNFEKEGEFFCFVGQ
9. Ocimum_basilicum-mono        FGYGRMHLM-KIGAYITLLDDIYDVYGT........LEELQVLTEIIERWDIN......LLD
```

Fig. 7.10 An example of multiple sequence alignment (MSA): sequences were retrieved from different plants and aligned using the MEGA6 software

be used both online (http://tcoffee.crg.cat/apps/tcoffee/do:regular) and offline on different operating systems (Notredame et al. 2000). ClustalX is used for multiple sequence alignment on every platform; it is downloaded from (http://www.clustal.org/clustal2/). This software also supports phylogenetic tree visualization. Short phylogenetic tree reconstruction was represented below in Fig. 7.12.

In Fig. 7.10, multiple sequence alignment (MSA) is performed taking nine protein sequences from different plant species. These sequences belong to the same protein terpene synthase family; the result shows different color coding for amino acids which means same color amino acid has similar chemical nature; like aspartic acid, glutamic acid which is denoted by symbol **D** and **E** shows red color background, which means these amino acids both have the same acidic property and are polar; every color means the specific nature of amino acid. The single star at the top represents the highly conserved amino acid in all sequences; dash between the sequences represents the gap inserted while aligning the most conserved residues. MSA is useful in motif prediction and conserved pattern identification and for phylogenetic tree construction.

7.3.2.2 Iterative Methods

An iterative method is another method which is used for MSA to reduce the error of progressive alignment. It is an alternative method to overcome the general problem of progressive alignment method, where the error in the initial step produces the wrong alignment for the rest of the sequences; the problem becomes accurate when alignment is between the distinctly related sequences. The iterative alignment method overcomes this problem by repeating and realigning the subgroups of the sequences, and further, this subgroup was aligned into the rest of the sequence; the primary objective is to improve the score of the alignment. Three programs that use iterative methods are MultiAlin, PRRP, and DIALIGN. MultiAlin (Corpet 1988) recalculates pairwise scores during the continuous alignment production; these scores were used for recalculating the tree, which is then used to refine the alignment to improve the score. The program PRRP uses iterative methods to produce an alignment.

7.3.2.3 Profile Sequence Alignment

Profile sequence alignment is a technique for identifying the putative structures and functions of sequences in profile analysis. This sequence comparing method is beneficial in finding the distinctly related sequences by aligning sequences to a family of similar sequences. However, this alignment method is very different from iterative and progressive methods. The protein sequences of the similar family are aligned together by multiple sequence alignment and then represented as a table of position-specific symbol comparison value and gap penalties; the matrix of similar family protein sequences is generated with this method. This matrix further can be used on multiple sequences to search for the protein sequence of the same family whose profile was generated. This method is speedy and widely used and can be run on the whole genome at a time. The use of profile alignment has been much used in PSI-BLAST program by Altschul et al. 1997.

7.4 Molecular Phylogenetic Analysis

7.4.1 Multiple Sequence Alignment and Phylogenetic Analysis

Multiple sequence alignment is the first and essential step for generating the phylogenetic tree (Ortet and Bastien 2010). Every column of alignment predicts mutations that occurred at one site during the time of evolution of the sequence family, revealing which positions in the sequences were conserved and which diverges from a common ancestor sequence as represented in Fig. 7.13. When two sequences found in two organisms are very similar, it is assumed that they have derived from one ancestor (Higgins and Sharp 1988). The sequence alignment reveals the conserved position throughout the ancestor sequence. MSA works in steps as shown in Figs. 7.10, 7.11, and 7.12. The necessary steps followed are: (i) calculate all pairwise alignments and associated distances; (ii) use the distances to build a trial phylogenetic tree; (iii) calculate pair weight based on the tree (divergence among sequences); (iv) produce a heuristic MSA based on the tree, from alignments; and

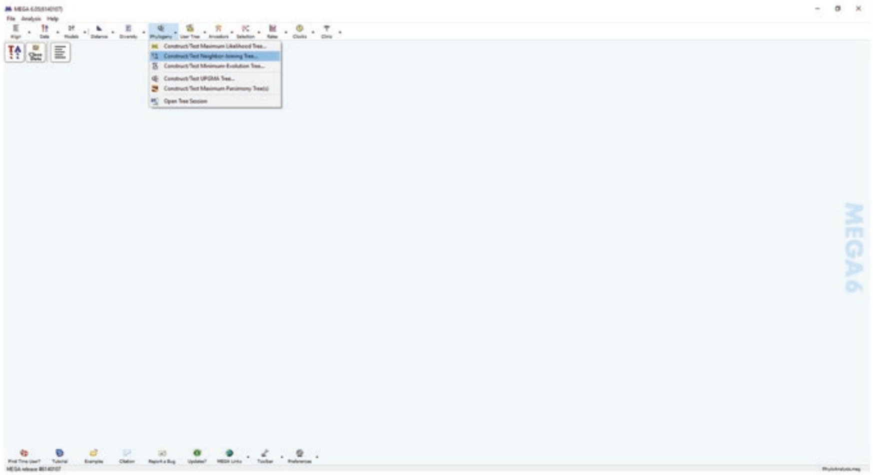

File Analysis Help
Align
Data
Models
Distance
Diversity
Phylogeny
User Tree
Ancestors
Selection
Rates
Clocks
Clinic
Construct/Test Maximum Likelihood Tree...
Construct/Test Neighbor-Joining Tree...
Construct/Test Minimum-Evolution Tree...
Construct/Test UPGMA Tree...
Construct/Test Maximum Parsimony Tree(s)
Open Tree Session
MEGA6

Fig. 7.11 Phylogenetic tree construction using MEGA6: phylogenetic tree as constructed using the neighbor-joining tree method

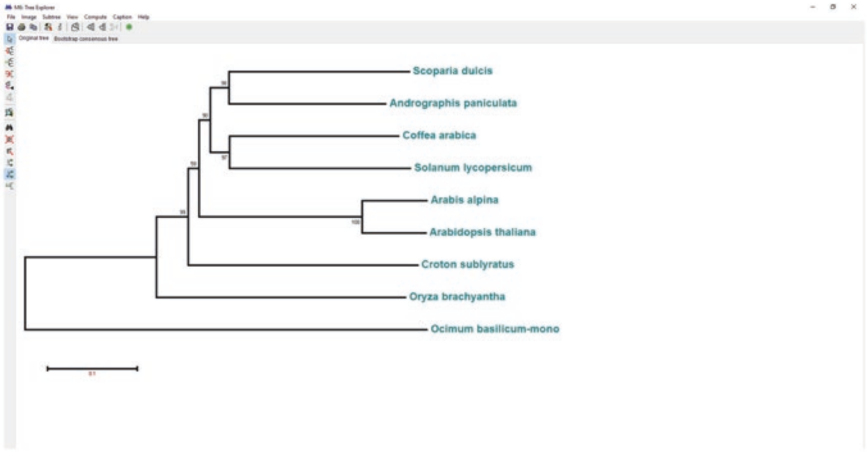


Fig. 7.12 Phylogenetic tree construction using MEGA6: A phylogenetic tree was constructed with sequences which were retrieved from different plants; these sequences were homologue to each other and belong to the same class except last sequence, i.e., *Ocimum basilicum*-mono, this sequence kept as an out-group

(v) determine relevant positions using heuristic MSA followed by scoring of optimal alignment within the relevant space.

The phylogenetic analysis is used for several purposes, including the comparison of more than two sequences, analysis of gene families and subfamilies, including their functional prediction, and estimation of evolutionary relationship among the organisms and homologous sequences (Sokal and Michener 1958). The phylogenetic

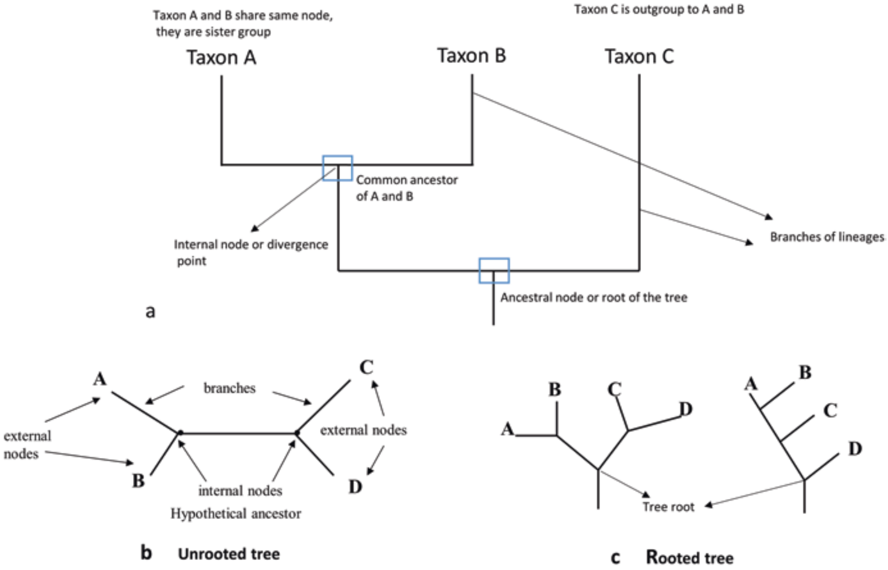

Taxon A and B share same node, they are sister group
Taxon C is outgroup to A and B
Taxon A
Taxon B
Taxon C
Common ancestor of A and B
Internal node or divergence point
Branches of lineages
Ancestral node or root of the tree
a
A
branches
C
external nodes
external nodes
B
internal nodes
Hypothetical ancestor
D
b Unrooted tree
B
C
D
A
A
B
C
D
Tree root
c Rooted tree

Fig. 7.13 Represents the layout of the phylogenetic tree; (**a**) shows the key features of the phylogenetic tree, (**b**) showed the unrooted tree, and (**c**) showed rooted tree

relationship refers to the relative time in the past that species shared common ancestors. The evolutionary relationships between the sequences are represented by placing one unrelated sequence as an outer branch of a tree which is also known as out-group. The rest of the sequences in the inner branch of the tree represents the degree to which different sequences are related. By this analysis, most closely related sequences can be identified by examining the neighboring branches on a tree. The phylogenetic tree generally shows the divergence of lineage through time (i.e., the evolutionary relationship of taxa) among a set of the organism of a group of organisms. The structure of the tree is represented in taxa (singular: taxon). The tips of the tree signify groups of descendent taxa (often species), and the nodes on the tree represent the common ancestors of those descendants. Two descendants that split from the same node are called sister groups. Many phylogenetic trees were represented with out-group of unrelated sequence from a group of sequences. The out-group was used in the larger group of sequence analysis; it gives a sense that the closely related sequences of the same group fall in the same cluster. In Fig. 7.13a, taxon A, B, and C were represented; taxon A and taxon B share the same nod, which shows that they have a common ancestor, where taxon C represented as an out-group of taxon A and taxon B.

Phylogenetic trees generated by computational methods can be either rooted or unrooted depending on the input data, and the algorithm used Fig. 7.13b, c. In a rooted tree, a single node designated as a common ancestor is known as cladogram, and a unique path leads from it through evolutionary time to any other node.

Unrooted trees only specify the relationships between nodes and say nothing about the direction in which evolution occurred known as phenogram. Roots can usually be assigned to unrooted trees through the use of out-group. There are several methods of constructing phylogenetic trees, discussed in next heading; all these can only provide estimates of what a phylogenetic tree might look like for given set of data. Most good methods also provide an indication of how much variation there is in this estimate. The most phylogenetic methods presuppose that every position of a nucleotide/amino acid in a sequence can change independently from the other positions. The gaps in alignment correspond to mutations in sequences at specific positions such as insertion, deletion, or genetic rearrangements, where gaps in phylogenetic methods were treated in various ways. Distance methods use the similarity scores based on scoring matrices (with gap scores). Using multiple sequence alignment and phylogenetic methods, we can identify the orthologous genes (genes related by speciation events, meaning same genes in different species) and paralogs genes (genes related by duplication events).

7.4.2 Methods Used in Phylogenetic Analysis

7.4.2.1 Character-Based Methods (Parsimony and Maximum Likelihood Method)

Parsimony analysis is the second primary way to estimate phylogenetic trees from aligned sequences. The maximum parsimony method generates the evolutionary tree that minimizes the numeral steps required to produce the pragmatic variation in the sequences from common ancestral sequences. For this reason, the method sometimes is referred to as the minimum-evolution method (Mount 2008). The maximum parsimony algorithm is not complicated and produces the best tree, because all possible trees related to a group of sequences were examined concurrently (Felsenstein 1988). Parsimony method is slow in computing but is an accurate method for tree construction; it cannot be used for a more significant number of sequences with a significant variation (Fitch 1971).

Most commonly used program for maximum parsimony methods is integrated into PHYLIP (Plotree and Plotgram 1989; Felsenstein 1996, 2002). For analysis of nucleotide sequence, there are other programs available: DNAPARS (program carries out unrooted parsimony (analogous to Wanger trees) (Eck and Dayhoff 1966; Kluge and Farris 1969), DNAPENNY (Dnapenny is a program that will find all of the most parsimonious trees implied by your data when the nucleic acid sequence parsimony criterion is employed) (Hendy and Penny 1982), DNACOMP (this program implements the compatibility method for DNA sequence data) (Day and Schwartz 1986), and DNAMOVE (Dnamove is an interactive DNA parsimony program). This program uses graphics characters that show the tree to best advantage on some computer systems. This program was developed by Wayne Maddison and David and Wayne Maddison). These programs are an integral part of PHYLIP package.

Maximum likelihood (ML) method is the statistical method which is used to estimate the parameters of the statistical model (Aldrich 1997; Hald 1998). It starts with inscription mathematical expression known as the likelihood function of the sample data (Navidi et al 1991). The method is used for finding the phylogenetic tree, and it involves the finding of topology and branch length of the tree, which gives us the great probability of observing DNA or protein sequence in our data (Tamura and Nei 1993). This method is beneficial for widely divergent group of sequences because it provides choice for the user to choose a model of evolution. The method follows some steps: firstly, the initial phylogenetic tree was built using the fast optimal method. Secondly, its branch lengths are adjusted to maximize the likelihood of the data set for the tree topology under the desired model of evolution (Kuhner et al. 1988). Maximum likelihood method generally works similarly as maximum parsimony. However, the difference is that maximum likelihood has the property to evaluate tree with variation in mutation rate in different lineages and to use evolutionary models like Jukes-Cantor and Kimura models. The main problem with the ML method is that it does not consider many numbers of sequences for analysis as well as it is computationally complex. Beyond this, it performs well with high-performance computing systems and is used for the development of more complex evolutionary models (Schadt et al. 1998). The sequence simulation experiments have shown that this method works better than all others in most cases, but it needs more computational time to construct the tree.

7.4.2.2 Distance-Based Methods (Neighbor-Joining and UPGMA Method)

The distance method works on the number of changes between each pair of sequences in a group of sequences to produce a phylogenetic tree of the group. The pair of sequences which have the smallest number of sequence changes between them is termed “neighbors.” The distance analysis compares two aligned sequences at the same time, and one by one comparison was done to construct a matrix of all possible sequence pairs. Due to the time of comparison, base substitutions and insertion/deletion events were counted and presented as a proportion of overall sequence length (Tamura and Nei 1993). These final estimates of the difference between all possible pairs of sequences are known as pairwise distances. Two primary and widely used distance methods in phylogenetic tree construction are neighbor-joining (Saitou and Nei 1987) method and UPGMA (unweighted pair group method with arithmetic mean). The software CLUSTALW uses the neighbor-joining distance method as a guide to multiple sequence alignments.

Neighbor-joining (NJ) method is used for reconstructing phylogenetic tree for evolutionary distance data for DNA and protein sequences (Saitou and Nei 1987). The objective of the algorithm is to construct the topology of a tree and also the branch length of the final phylogenetic tree. The neighbor-joining method is suitable when the rate of evolution varies for separate lineages. The NJ method algorithm takes input from a distance matrix specifying the distance between the pair of taxa; it finds pairs of OTUs (operational taxonomic units) that minimize total branch length at each stage of clustering starting with a star like a tree (minimum-evolution tree), produced in the assumption that there is no clustering of OTUs. This unresolved tree has topology that corresponds to that of a star network and iterates in multiple steps until the tree is completely resolved, and all branch lengths are known. This method is the fastest method as compared to maximum parsimony and likelihood methods (Kuhner and Felsenstein 1994). It was used to analyze large sequence dataset (hundreds or thousands of taxa).

UPGMA (unweighted pair group method with arithmetic mean) method is merely a bottom-up approach to hierarchical clustering method, which produces a dendrogram from a distance matrix (Sneath and Sokal 1973). It is a statistical method for evaluating systematic relationships developed by Robert R. Sokal and Charles D. Michener (Sokal and Michener 1958). This method corresponds to WPGMA (weight pair group method with arithmetic mean); the algorithm of both methods is similar to its unweighted variant and the UPGMA algorithm. WPGMA algorithm calculates the distance between clusters, as the number of taxa weights a simple average wherein UPGMA algorithm averages in each cluster at each step. In ecology, this method is most popular in classifying the sampling units. The UPGMA method begins with calculating the branch lengths of the most closely related sequences and averages the distance between this pair or sequence cluster. It continues until all sequences are included in the tree. Finally, this method predicts a position for the root of the tree (Sattath and Tversky 1977).

7.4.2.3 Validation Methods: Bootstrap and Jackknife

Multiple sequence alignment and phylogenetic analysis are used for many purposes as discussed earlier in this chapter. The determination of the phylogenetic relationship of the closely related sequence is straightforward as compared to diverse sequences. For obtaining better alignment results, users can decide what type of sequences they are taking for sequence alignment and phylogenetic analysis, and according to the type of sequence, the above-discussed MSA methods were used. Bootstrap is merely a statistical method, which allows assigning measures of accuracy (Efron and Tibshirani. 1994; Efron 2003). This technique was developed in the early 90s, for making certain kinds of statistical inferences. It is mostly used in complex nonparametric estimation problems, where logical methods are not practical. Felsenstein (Efron et al. 1996) introduced the use of the bootstrap in the estimation of phylogenetic trees. Many researchers criticized this method because it is consistently too conservative, but Bradley and team proved it wrong (Efron et al. 1996).

Jackknife is resampling method, which is useful for variance and bias estimation; it is similar to a method like bootstrap. It was developed by Maurice Quenouille (1949, 1956), after that John Tukey expended the technique and named it Jackknife (Tukey 1958). This method was tested on phylogenies which started with the

Mueller and Ayala in 1982, who used Jackknife approach to estimate the variance of the length of a branch in UPGMA phylogeny from gene frequency data (Muller and Ayala 1982; Zuo et al. 2010).

7.4.3 Tools and Software Used for Tree Construction

Phylogenetic analysis programs and tools are widely available, freely or either with little cost. Here, some out of many, well performing online as well as offline tools for generation of the phylogenetic tree has been introduced. The first and most popular software is PHYLIP (phylogenetic inference package) (Felsenstein. 1981, 1996) available from Dr. J. Felsenstein at http://evolution.genetics.washington.edu/phylip.html, and PAUP (phylogenetic analysis using parsimony) available from Sinauer Associates, Sunderland, Massachusetts, and http://www.lms.si.edu/PAUP/. PHYLIP is a free package of the program for inferring phylogenies. It can be installed on Windows, Mac, and Linux systems. It is a most widely distributed phylogenetic package. PAUP (phylogenetic analysis using parsimony and another method) is another computational phylogenetic program; it can be installed on Windows and Linux, UNIX systems, the latest version of this program is PAUP version 4 and can be downloaded from (http://paup.sc.fsu.edu/). Next software is MrBayes, which is a software program for inferring phylogenetic parameters in a Bayesian statistical framework (Huelsenbeck and Ronquist 2001, 2005), and the current version of this software is MrBayes 3.2.3. It can be used online via (http://www.phylogeny.fr/one_task.cgi?task_type=mrbayes&tab_index=2) or offline by downloading and using it via command line. RAxML (randomized axelerated maximum likelihood) is a program used for sequential and parallel maximum likelihood-based inference of large phylogenetic trees. It is used in post-analysis of a set of phylogenetic trees, sequence alignments, and evolutionary placement of short reads. It has initially had been derived from fastDNAml which in turn was derived from Joe Felsenstein's dnaml which is part of the PHYLIP package (Stamatakis 2014), and the user can either download or install on the local machine, or it can be operated online by web server with GUI (graphical user interface) (https://embnet.vital-it.ch/raxml-bb/). PHYML is an online web interface for phylogenetic tree generation which is fast and accurate heuristic for estimating maximum likelihood phylogenetic tree from DNA and protein sequences (Guindon et al. 2005); the latest version of this software is PHYML 3.0 (http://www.atgc-montpellier.fr/phyml/). MEGA (Tamura 2007; Tamura et al. 2011) (Fig. 7.10) (Kumar et al. 2016a), this software is very sophisticated and user-friendly for analyzing DNA and protein sequence data from species and populations. MEGA has perfect GUI and also available in the command line (http://www.megasoftware.net/); anyone can operate this software; it can be downloaded and installed on the local machine on any operating system or platform. Presently, MEGA have two latest versions: MEGA 7 and MEGA 7.1 beta (Kumar et al. 2016b). The beauty of this software is that the user can use different statistical methods for multiple sequence alignment and phylogenetic tree construction. Multiple tools were integrated with this program like for phylogenetic tree visualization (alignment/trace editor, tree explorer, data explorer, gene duplication wizard, and many more). Except for these tools, there are numbers of other tools which used standard methods for phylogenetic tree analysis like TreeViewJ, TreeBest, Philip, FigTree, etc.

7.5 Conclusion

In this chapter, we discussed the most implemented sequence analysis methods, tools, and algorithms. This chapter introduces the biological sequence analysis through different methods including pairwise (global and local); sequence similarity and scoring with dot matrix and dynamic programming; alignment scoring schemes through PAM and BLOSUM matrices; multiple sequence analysis with progressive, iterative, and profile methods; and phylogenetic analysis with character and distance methods along with their validations. In this context, different software tools and web servers were also introduced to the readers for the abovementioned experimental analysis purposes. This chapter enhances the knowledge of the readers who are interested in biological sequence analysis using diverse bioinformatics approaches.

Acknowledgment We are thankful to CSIR-CIMAP, Lucknow, India, Dr. Feroz Khan for his support. Authors express special thanks of gratitude to all the world-wide scientists, professors and their contribution in the research, authors also thankful to all the young research scholar community for giving their life to the research and to all parents who supported them with patience.

References

Aldrich J (1997) RA fisher and the making of maximum likelihood 1912-1922. Stat Sci 12(3):162–176

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410

Altschul SF, Boguski MS, Gish W, Wootton JC (1994) Issues in searching molecular sequence databases. Nat Genet 6(2):119–129

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25(17):3389–3402

Bhagwat M, Young L, Robison RR (2012) Using BLAT to find sequence similarity in closely related genomes. Curr Protoc Bioinformatics:10–18

Chial H (2008) DNA sequencing technologies key to the human genome project. Nature Education 1(1):219

Corpet F (1988) Multiple sequence alignment with hierarchical clustering. Nucleic Acids Res 16(22):10881–10890

Day WH, Sankoff D (1986) The computational complexity of inferring phylogenies by compatibility. Syst Biol 35(2):224–229

Dayhoff MO, Schwartz RM, Orcutt BC (1978) A model of evolutionary change in proteins. In: *Atlas of protein sequence and structure*, vol 5. National Biomedical Research Foundation, Silver Spring, pp 345–352

Eck RV, Dayhoff MO (1966) Evolution of the structure of ferredoxin based on living relics of primitive amino acid sequences. Science 152(3720):363–366

Efron B (2003) Second thoughts on the bootstrap. Stat Sci 18(2):135–140

Efron B, Tibshirani RJ (1994) An introduction to the bootstrap. CRC press

Efron B, Halloran E, Holmes S (1996) Bootstrap confidence levels for phylogenetic trees. Proc Natl Acad Sci 93(23):13429–13429

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17(6):368–376

Felsenstein J (1988) Phylogenies from molecular sequences: inference and reliability. Annu Rev Genet 22(1):521–565

Felsenstein J (1996). [24) Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. Methods Enzymol 266:418–427

Felsenstein J (2002) {PHYLIP}(Phylogeny Inference Package) version 3.6 a3

Feng DF, Doolittle RF (1996). [21) Progressive alignment of amino acid sequences and construction of phylogenetic trees from them. Methods Enzymol 266:368–382

Fitch WM (1971) Toward defining the course of evolution: minimum change for a specific tree topology. Syst Biol 20(4):406–416

Gibbs AJ, McIntyre GA (1970) The diagram, a method for comparing sequences. FEBS J 16(1):1–11

Gotoh O (1982) An improved algorithm for matching biological sequences. J Mol Biol 162(3):705–708

Guindon S, Lethiec F, Duroux P, Gascuel O (2005) PHYML online—a web server for fast maximum likelihood-based phylogenetic inference. Nucleic Acids Res 33(suppl_2):W557–W559

Hald A (1998) A history of mathematical statistics from 1750 to 1930. Wiley

Hendy MD, Penny D (1982) Branch and bound algorithms to determine minimal evolutionary trees. Math Biosci 59(2):277–290

Henikoff S, Henikoff JG (1992a) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci 89(22):10915–10919

Henikoff S, Henikoff JG (1992b) Amino acid substitution matrices from protein blocks. PNAS 89(22):10915–10919

Higgins DG, Sharp PM (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene 73(1):237–244

Higgins DG, Thompson JD, Gibson TJ (1996) Using CLUSTAL for multiple sequence alignments. Methods Enzymol 266:383–402

Hogeweg P, Hesper B (1984) The alignment of sets of sequences and the construction of phyletic trees: an integrated method. J Mol Evol 20(2):175–186

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17(8):754–755

Huelsenbeck JP, Ronquist F (2005) Bayesian analysis of molecular evolution using MrBayes. Statistical methods in molecular evolution:183–226

Kent WJ (2002) BLAT-the BLAST-like alignment tool. Genome Res 12(4):656–664

Kluge AG, Farris JS (1969) Quantitative phyletics and the evolution of anurans. Syst Biol 18(1):1–32

Kuhner MK, Felsenstein J (1994) A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol Biol Evol 11(3):459–468

Kuhner MK, Yamato J, Felsenstein J (1998) Maximum likelihood estimation of population growth rates based on the coalescent. Genetics 149(1):429–434

Kumar S, Stecher G, Tamura K (2016a) MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol 33(7):1870–1874

Kumar S, Stecher G, Tamura K (2016b) MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol 33(7):1870–1874

Lipman DJ, Pearson WR (1985) Rapid and sensitive protein similarity searches. Science 227(4693):1435–1441

Lipman DJ, Altschul SF, Kececioglu JD (1989) A tool for multiple sequence alignment. Proc Natl Acad Sci 86(12):4412–4415

Löytynoja A, Goldman N (2005) An algorithm for multiple progressive alignments of sequences with insertions. Proc Natl Acad Sci U S A 102(30):10557–10562

Maizel JV, Lenk RP (1981) Enhanced graphic matrix analysis of nucleic acid and protein sequences. Proc Natl Acad Sci 78(12):7665–7669

Mount DW (2008) Maximum parsimony method for phylogenetic prediction. Cold Spring Harb Protoc 2008(4):PDB–top32

Mount DW (2009) Using progressive methods for multiple global sequences

Mueller LD, Ayala FJ (1982) Estimation and interpretation of genetic distance in empirical studies. Genet Res 40(2):127–137

Navidi WC, Churchill GA, Von Haeseler A (1991) Methods for inferring phylogenies from nucleic acid sequence data by using maximum likelihood and linear invariants. Mol Biol Evol 8(1):128–143

Needleman SB, Wunsch CD (1970) A general method applies to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48(3):443–453

Notredame C, Higgins DG, Heringa J (2000) T-coffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol 302(1):205–217

Ortet P, Bastien O (2010) Where does the alignment score distribution shape come from. In: Evolutionary bioinformatics online, vol 6, p 159

Pearson WR, Miller W (1992) Dynamic programming algorithms for biological sequence comparison. Methods Enzymol 210:575–601

Plotree DOTREE, Plotgram DOTGRAM (1989) PHYLIP-phylogeny inference package (version 3.2). Cladistics 5(163):6

Polyanovsky VO, Roytberg MA, Tumanyan VG (2011) Comparative analysis of the quality of a global algorithm and a local algorithm for alignment of two sequences. Algorithms Mol Biol 6(1):25

Quenouille MH (1949) Approximate tests of correlation in time series. J R Stat Soc Ser B 11:68–84

Quenouille MH (1956) Notes on bias in estimation. Biometrika 43(3/4):353–360

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4(4):406–425

Sattath S, Tversky A (1977) Additive similarity trees. Psychometrika 42(3):319–345

Schadt EE, Sinsheimer JS, Lange K (1998) Computational advances in maximum likelihood methods for molecular phylogeny. Genome Res 8(3):222–233

Sellers PH (1974) On the theory and computation of evolutionary distances. SIAM J Appl Math 26(4):787–793

Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147(1):195–197

Sneath PH, & Sokal RR (1973) Numerical taxonomy. The principles and practice of numerical classification

Sokal RR (1958) A statistical method for evaluating the systematic relationship. University of Kansas science bulletin 28:1409–1438

Sokal RR, Michener CD (1958) A statistical methods for evaluating relationships. Univ Kansas Sci Bull 38:1409–1448

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30(9):1312–1313

Tamura K (2007) Molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol Biol Evol 24:1596–1599

Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10(3):512–526

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28(10):2731–2739

Tukey JW (1958) Bias and confidence in not quite large samples. Ann Math Statist 29:614

Wang L, Jiang T (1994) On the complexity of multiple sequence alignment. J Comput Biol 1(4):337–348

Waterman MS, Perlwitz MD (1984) Line geometries for sequence comparisons. Bull Math Biol 46(4):567–557

Zuo G, Xu Z, Yu H, Hao B (2010) Jackknife and bootstrap tests of the composition vector trees. Genomics Proteomics Bioinformatics 8(4):262–267

Chapter 8
Tools and Methods in Analysis of Complex Sequences

Noor Ahmad Shaik, Babajan Banaganapalli, Ramu Elango, and Jumana Y. Al-Aama

Contents

8.1 Sequence Technologies

In the early days of gene sequencing, most methods adopted location-specific primer extension strategies, which are not just complex but consumes lot of time, cost, and workforce (Maxam and Gilbert 1977). However, this scenario has dramatically changed with the discovery of fast and accurate DNA sequencing

N. A. Shaik (✉) · J. Y. Al-Aama
Department of Genetic Medicine, King Abdulaziz University, Jeddah, Saudi Arabia
e-mail: nshaik@kau.edu.sa; jalama@kau.edu.sa

B. Banaganapalli · R. Elango
Princess Al-Jawhara Center of Excellence in Research of Hereditary Disorders,
Department of Genetic Medicine, Faculty of Medicine,
King Abdulaziz University, Jeddah, Saudi Arabia
e-mail: bbabajan@kau.edu.sa; relango@kau.edu.sa

N. A. Shaik et al. (eds.), *Essentials of Bioinformatics, Volume I*,
https://doi.org/10.1007/978-3-030-02634-9_8

technique based on the primer extension strategy introduced by Frederick Sanger (Sanger et al. 1977a, b). In 1986, the first fully automated sequence machine was developed by Applied Biosystems Corporation. This automated machine was majorly used in sequencing of human genome by Craig Venter in Human Genome Project (Lander et al. 2001). The recent unprecedented technological advances in both molecular biology and engineering fields have allowed us in further improving the efficiency of Sanger sequencing method. These latest Sanger sequencing machines offer advanced separation techniques, different visualization methods, and parallel sample processing strategies. The current Sanger sequencing machine is able to handle 96 samples in one run. While conventional, gel-based Sanger sequencing machines could only produce 250–500 base pairs of DNA sequence per reaction, the present-day Sanger sequencing machines could produce 750–1000 base pairs of sequence from a single reaction, making sequencing a much less expensive than it used to be in the past (Metzker 2005). Although this gold standard Sanger method is still being used around the world in thousands of laboratories, time and resource limitations brought dramatic next-generation technologies surfaced to sequence whole genomes.

8.2 NGS Technologies and Applications

High efficiency and low budget made NGS technology a highly popular option in current-day genomics research (Metzker 2005). The sequencing technology has immensely improved with the course of the time, having initially started with a 20 megabase pair (Mbp) throughput capacity (GS20 from 454 Life Sciences) (Margulies et al. 2005). Currently, Illumina HiSeq X system can produce 1.8 terabase pair (Tbp) of sequencing data, which is a rise of 100,000 base pair fold in a short span of 10 years (Rhodes et al. 2014). Current NGS platforms generally use following phases to perform sequencing. (1) PCR-based amplification of DNA sequence libraries, (2) synthesis of DNA sequence, i.e., by adding the complimentary nucleotides, and (3) parallel sequencing of the amplified DNA stands. NGS autoparallelization is able to generate enormous amount of nucleotide sequence data (from Mbp to Gbp) in one run. NGS technologies have also significantly reduced the sequencing costs to 500 per exome and 1000 per genome sequencing (Le Gallo et al. 2017). Various sequencing platforms and their main technical features are summarized in Table 8.1. NGS technology has become a major molecular tool in de novo genome sequencing of prokaryotes and viruses, screening for genetic mutation/variants by genome sequencing or exome sequencing, and in exploring the regulatory mechanisms of gene expression in cells and tissues.

Table 8.1 Current NGS technologies

Platform	Types	Base per run	Read length	Principle
Nanopore	MinlON, Pro methlON	42 Gb 12 TB	230–300 Kb 230–300 Kb	(Minion) Real-time portable device for DNA/RNA sequencing (PromethION) Real-time long-read DNA/RNA sequencing
Illumina	Miseq Hiseq	300 Mb–15GB 1.6–1.8 TB	600 bp 300 bp	(Miseq) Single-run NGS analysis instrument, small-scale sequencing (Hiseq), large-scale genome sequencing
PacBio	PacBio RS I I	500 MB–16 GB	60 Kb	Sequence longer reads and give high-throughput sequencing
Life Technologies	Solid 5500	80–320 GB	50–2X 50 bp	Genetic analysis systems are highly accurate, massively parallel next-generation sequencing platforms to perform exome and RNA sequencing

8.3 Experimental Design and Methods of NGS Using Illumina

The latest illumina NGS technology uses sequence-by-synthesis (SBS) principle to sequence millions of copies of DNA fragments in massively parallel fashion. The major advantage of this technology is that it gives error-free reads and >30Q base call quality. Illumina NGS workflow includes three major steps like library preparation, cluster generation, and sequence by synthesis (Kruglyak et al. 2016). (1) Library preparation – The initial step in library preparation is the random fragmentation of DNA samples, and ligation of the DNA at 3′ and 5′ adaptor regions. To increase the productivity of DNA library, "tagmentation" which combines ligation and fragmentations reactions into a single step is followed. In the final step of library preparation, the PCR amplification and gel purifications are carried out on adapter-ligated fragments. (2) Cluster Generation – DNA library is loaded into flow cells, where fragments are captured in a lawn of surface bound with oligos complimentary to the library adapters (Fedurco et al. 2006; Turcatti et al. 2008). Each DNA fragment undergoes amplification step forming a bridge when the strand is bent with the adjacent oligonucleotide strand and hybridize, and the polymerase extends the complementary strand. The process of bridge amplification is repeated multiple times until clonal clusters are produced. Once, the cluster generation is complete, the DNA templates are ready for final step of sequencing. (3) Sequencing by SBS method – uses reversible terminator process, by which each base is recognized as it is inserted into template strands. All the nucleotides that are labeled through fluorescence or terminated through reverse orientation are moved in the lawns of clusters of the cell enabling the incorporation of first nucleotide base followed by laser detection of colored nucleotide.

The fluorophore is broken in the final step and reversal of terminator provides space to new base pair. This process repeats till the predetermined sequence length is generated. The result is highly accurate base-by-base sequencing that virtually eliminates sequence context–specific errors, even within repetitive sequence regions and homopolymers.

8.4 Whole Genome/Whole Exome Sequencing

The WGS and WES methods, owing to their sequencing depth and accuracy, have become the most reliable methods to detect the genetic sequence variations. WGS can determine the total DNA sequence of any organism in a single experiment. The major drawback of WGS is that it provides huge amount of data, whose analysis and interpretation is very complex (Nakagawa et al. 2015). In contrast, WES or target sequencing methods capture only the protein-coding gene sequence. The WES has become an important clinical diagnosis method to identify the disease-causative mutations located in protein-coding genes (Guo et al. 2018). The major focus of this chapter is about WES method and its data analysis (Fig. 8.1).

8.5 Whole Exome Sequencing (WES)

Whole exome sequencing (WES) method starts by capturing and sequencing the coding regions of the genome (Priya et al. 2012). WES has the ability to detect mutations in protein-coding regions of disease-causative mutations (Bamshad et al. 2011). The most popular exome capture/enrichments kits available in market are Agilent SureSelect, NimbleGen SeqCap, Illumina TruSeq, Nextera Rapid, etc. These kits differ in their bait density, bait (capture probe) length, target selection regions, and the nucleotides they target (Puritz and Lotterhos 2018) (Table 8.2).

Agilent SureSelect exome capture protocol requires 100 ng of DNA for library preparation (Al-Aama et al. 2017). The initial phase of DNA fragmentation is done by high shear homogenization technique followed by library preparation with specific sequence adapters. In the second phase, DNA is hybridized with highly specific biotinylated cRNA library baits (120mer RNA) (Chen et al. 2015). In the third phase, the target region is selected using magnetic streptavidin beads. The final phase involves amplification, clustering, and sequencing steps. The major limitation of WES is that its probes cannot cover all the exons listed in Consensus Coding Sequence project (CCDS) and Refseq (Pruitt et al. 2009, 2012) databases. Furthermore, exon capturing is compromised at regions with high or low GC content being more difficult to sequence. Furthermore, WES could only capture up to 92–95% of exons; therefore, mutations in other exons go undetected. These capture kits are also inefficient in determining the structural variants such as inversions and translocations.

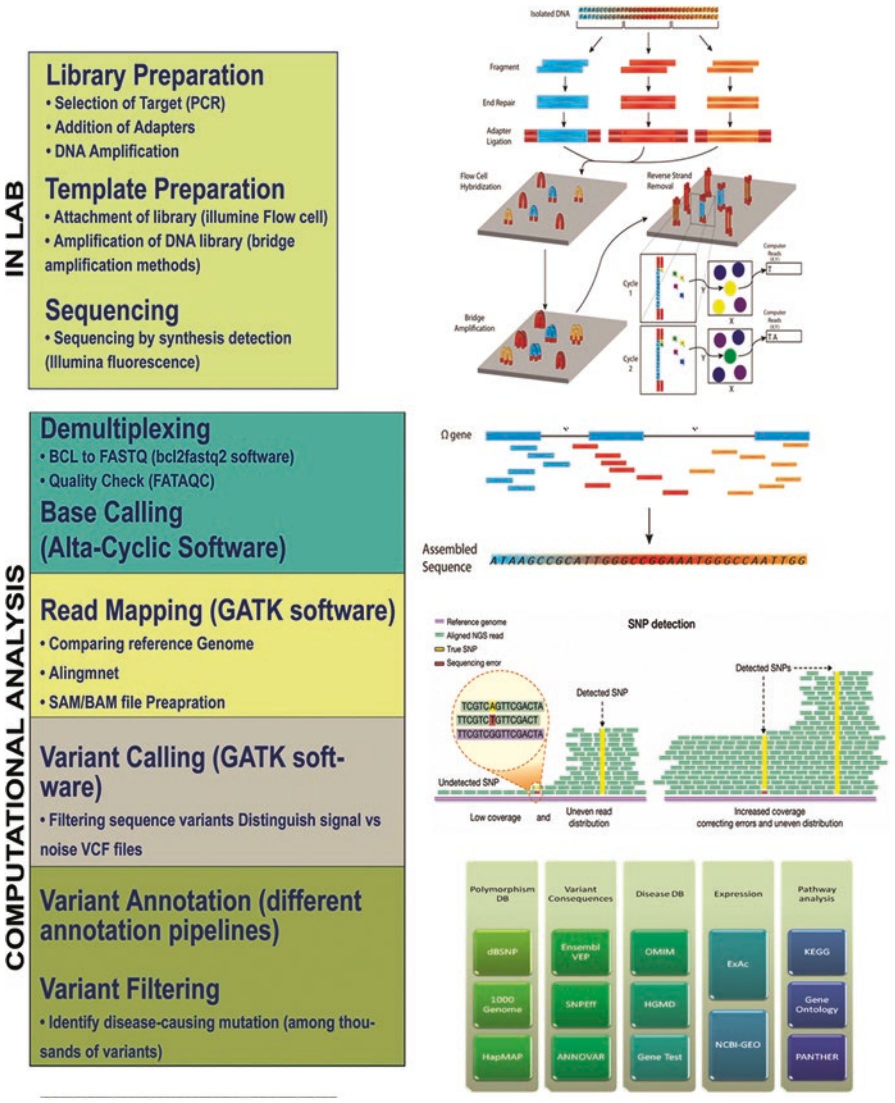


Fig. 8.1 Steps in exome analysis

8.6 WES Data Analysis

8.6.1 Preprocessing of Raw Data Quality

Most of the Illumina platforms generate sequence reads in a binary base call (BCL) file format and most of the software are unable to analyze this format. Therefore, BCL file format is changed into a universally accepted FASTQ format. The reads

Table 8.2 Popular exome-enrichment kits, their target coverage, and advantages

Exome capture kit	Target cover in Mb	Advantage
Agilent SureSelect	51.1 Mb	Able to sequence specific target sites
NimbleGen SeqCap	64.1 Mb	Abolishes thousands of PCR reactions, enabling enrichment of the whole exome or regions of interest in a single test tube
Nextera Rapid	62.08 Mb	Consists of all-in-one library preparation and enrichment
Illumina TruSeq	62.08 Mb	Integrated kit consists of DNA sample preparation, pre-enrichment sample pooling

that successfully pass "Chastity Filtering" are forwarded for further sequence analysis. In the next step, Illumina adapters are eliminated, and sequences such as poly(A) or poly(T) are added to the reads. Trimmomatic Software, Cutadapt, or FASTXToolkit are often recommended for trimming the adapter sequences (Del Fabbro et al. 2013). Different software like NGS QC, PRINSEQ, FastQC are used to produce quality report of generated sequences (Schmieder and Edwards 2011). These software provide information regarding GC content, read length, quality score distribution, and sequence duplication on each read. These tools can also be used for sequence trimming and QC analysis. FastQC generates quality scores for sequences and represents them in the form of box and graph plots.

8.6.2 Alignment with Reference Genome

A mapping tool Bowtie2 is often used to align the short reads against the reference genome (GRCh37 from NCBI, Feb. 2009). The Bowtie tool is considered to be the best tool for sequence alignment as it requires less RAM and is able to perform modest index alignment. Using the Ferragina and Manzini index (FM-index), Bowtie2 aligns reference genome with unpaired reads existing in fastq or fq formats (Langmead and Salzberg 2012). The output of the Bowtie2 is SAM (short alignment summary) format (Li et al. 2009) (Table 8.3). The SAM tool converts the SAM files to BAM (Binary Alignment Mapped) format, and is able to eliminate duplicate and false-positive variants from the WES data. The converted SAM file consists of huge sequence information and requires memory space. BAM refers to the binary format of aligned sequences that will encrypt and reduce the sequence data size. BAM files are usually sorted, filtered for duplicates, locally realigned, and calibrated both for improving the base quality and reduce false-positive variants. However, transcriptomics and epigenomics data do not require these kinds of extended postprocessing realignment steps. In the final step, base quality recalibration utilizes the known variants in the database and readjusts the base quality scores to enhance the accuracy of variants calling.

Table 8.3 Description of each descriptor from exome analysis

File name	Description
FastaQ or fq	Text-based format for storing both DNA and its corresponding quality scores Example of the standard FASTQ file is shown below: Line 1. @EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG Line 2. GAAATCCATTTGTTCAACTTCAACTATCTTGCAAATCCATTTGTTCAACT Line 3. + Line 4. !"*((((***+))%%%++)(%%%%).1***-+*"))**55CCF>>>>>>CCCCCCC65 The character @ in the first line shows the sequence identifier. The biological sequence is shown in the second line and it is composed of four-letter nucleotide (A, T, G, and C). In the third line, again a sequence identifier is shown +. In the third line, there is only + character. Quality scores of corresponding sequence read are shown in the fourth line and it is coded as ASCII characters
SAM	Sequence Alignment/Map format is a genetic format for storing large nucleotide sequence alignments (simple, flexible in storing and aligning the genomic data)
BCL	Base call binary file produced by Illumina sequencing instrument. bcl2fastq tool merges per-cycle BCL files into FASTQ files that are the input of many downstream sequencing analysis tools such as aligners and de novo assemblers
BAM	The compressed format of SAM known as compressed binary alignment matrix (BAM). Both files are interconvertible and help SAM tools as standalone software
Bed	Tab-delimited text file which contains multiple lines each represent a single genomic region or a gene body. In this file, there are three required fields; one is standard BED file, which is named as chrom, chromStart, and chromEnd
gtf/gff	General Transfer Format or Gene Transfer Format (GTF) and General Feature Format (GFF) are text-based annotation files that stores gene structure information of any genome
VCF	VCF is a text format; it contain meta information lines, a header line, and data line, each containing information about position in the genome

8.6.3 Variant Calling (Primary and Secondary Variant Annotations)

Various open source tools like GATK, SAMTools, CRISP, Snver, Varscan etc., are used in variant calling step (Regier et al. 2018). Furthermore, the identified genetic variants will be screened with a statistical formula, i.e. Sensitivity = TP/(TP + FN), Precision = TP/(TP + FP), False discovery rate (FDR) = FP/(TP + FP) and F-Score = 2TP/(2TP + FP + FN). Whereas, TP is true positive variant found in both Varscan or GATK validated dataset and data determined by reference dataset; FP is false positive variant determined by reference dataset but not validated by Varscan or GATK; FN is false negative variant, known as missing variant which is validated by Varscan or GATK but not determined by reference dataset (Stitziel et al. 2011).

The secondary analysis of variant call is performed to further check the performance and genetic priorities of the variants, different metrics such as read depth, genotype quality, genotype concordance, frequency estimation of variant, and customized filtering options, which identify disease-causing variants (variant prioritization), will be analyzed.

8.7 Annotating Variants

In this step, variants are prioritized for the purpose of distinguishing between causative variants and the milieu of polymorphisms and variant calling errors present in a WES dataset. An approximate number of 110,000 to 130,000 variants are generated from the secondary variant calling stage. Therefore, it is important to assign functional information for all these variants. In a clinical genetic analysis, a mutations or variants can have different annotations (Worthey et al. 2011). These variant annotations are majorly classified into six groups (Butkiewicz and Bush 2016):

1. Initial variant data annotation is done by deprioritizing the variants generated through sequence or mapping errors.
2. Annotate and classify (synonymous, nonsynonymous, nonsense codon, or splice site changes in the single nucleotide variant) the variants by their location in the gene.
3. Annotate the variants to classify known clinical variants.
4. Annotate the variants that determine the potential known functional impact of a variant.
5. Annotate the unknown variants for functional predictions, by different pathogenetic predictions tools (variant effect predictor, SnpEff, ANNOVAR, etc.).
6. Annotate the variants to estimate the allele frequency (1000 Genomes Project, ESP6500, dbSNP, ExAC) of the variant in a non-disease or disease population.

In the final step, a customized filtering process is required for determining the disease-causative variants from the exome or genome sequencing data. Briefly, these include the functional relevance of corresponding genetic variant to disease pathology and its mode of inheritance, etc. The specific recommendations and guidelines for discovering disease-causative genetic variants are reviewed elsewhere (Bamshad et al. 2011).

8.8 WES Data Filtration in Mendelian/Monogenic Disease

Traditional disease gene mapping approaches (such as karyotyping, linkage analysis homozygosity mapping, and copy number variation analysis) have no doubt provided deep insights into the molecular causes of different Mendelian diseases (Lee et al. 2014a. However, over the past few decades, these approaches are not fully able to detect all forms of genomic variations, due to the fact that they require minimum number of patients for reaching statistical significance and consume lot of time (Rabbani et al. 2014). Currently, WES and targeted resequencing methods have now empowered researchers to perform quick and efficient analysis of single gene disorders even in small families and even in single affected patient. The disease-causative variants are usually searched based on their minor allele frequency cut-off values and their major effect (missense, truncated, and splice mutations) on the function of query gene (Gilissen et al. 2011). The inheritance pattern of the disease is another feature that can be used in filtering the variants. For example, heterozygote mutations in disease genes are usually seen to cause dominant disorders, whereas homozygous mutations are seen to cause recessive disorders. However, compound heterozygote mutations in recessive genes are an exception to this rule. Ascertaining the lack of rare variants in query genes in large population cohorts (whether sporadic cases or healthy controls) can further improve our search for disease-causative agents (Bamshad et al. 2011).

The present-day technological developments in variant identification methods have enabled us to accurately diagnose disease variants and allowed their utility in clinical practice to initiate precision medicine concept (Lee et al. 2014b). The major challenge in interpreting the effect of variants on disease phenotypes is the unknown and dynamic aspect of genomes. WES analysis of any individual usually reveals the presence of thousands of functionally important variants which may not be actually related to the disease causality (Macarthur 2012). Since dozens to hundreds of disease variants can sometimes pass through different variant filtration steps described above, the prioritization of potential causal variants for functional biology assays should be done very carefully. Mere presence of rare variants in any disease gene is not enough to claim its disease causal role in individuals. Inheritance pattern of disease causative variants can be better understood by examining extended family pedigrees. This WES-variant segregation analysis can be made more reliable by performing linkage or homozygosity mapping methods. In cases, where the family data is missing, one can ascertain the absence of disease-causing rare variants in genome data of sporadic cases with same disease or healthy populations (Bailey-Wilson and Wilson 2011). All the above discussed distinct complementary approaches can help in the identification of candidate variants and genes, which can be further evaluated in functional biology assays (Teare and Santibanez Koref 2014).

The molecular investigation of complex diseases demands unique strategies to filter, analyze, and interpret the NGS data for identifying the disease-causative genetic mutations. From the past one decade, genome-wide association studies (GWAS) have become a most reliable approach that identified hundreds of common risk alleles for complex human diseases (Visscher et al. 2012). These studies were enabled by a combination of the availability of large well-characterized sample collections, advances in genotyping technologies, and advances in methods for the analysis of the resulting genetic data. These studies have provided several biological insights, highlighting the role of the complement genes in age-related macular degeneration, of autophagy in Crohn’s disease or of specific regulatory proteins in blood lipid levels, among others. Recent GWAS approaches identified 200 risk loci in inflammatory bowel disease (IBD) (Ye and McGovern 2016). However, very few have been conclusively resolved to specific functional variants. A recent study genotyped 94 fine mapped loci of IBD in 67,852 individuals, and concluded that 45 variants are highly enriched in immune cells associated with Crohn's disease. Of these 45 variants, 13 variants were enriched for protein changes, 10 for tissue specific epigenetic marks, and 3 for transcription factor binding sites in immune cells. The results of this study suggested that high-resolution fine-mapping in large samples can convert many GWAS discoveries into statistically convincing causal variants, providing a powerful substrate for experimental elucidation of disease mechanisms (Huang et al. 2017). However, these techniques have their own limitations. First, genome-wide linkage search in affected individuals only reveals the genomic region or locus associated with the disease and might not correctly identify the actual disease-causative variant. Second, GWAS relies on the common variants that explain only a modest fraction of the inheritance of genetic diseases. The novel and rare causal variants, which might account for much larger fraction of heritability, remain uncovered in GWAS (Korte and Farlow 2013). In contrast to this, rare variants are most likely to be found in independent genomes and they can be easily detected by whole genome sequencing approach. Therefore, one can argue that a single variant/gene does not produce a large effect and thus their biological pathways should be focused to determine the disease biology.

8.9 Conclusion

NGS approach refers to a wide range of genetic sequencing techniques like WGS, WES, and targeted sequencing. All these NGS techniques enable us to very quickly generate the deeper resolution of genetic sequences with high-throughput capacity. Illumina, is one of the widely used NGS platforms, used in analysis of billions of DNA fragments in one single experiment. In this method, DNA fragments are hybridized with oligonucleotides in the flow cell of glass slide. The experimental design of NGS uses paired end (PE) or single read end (SE) sequencing methods to accurately align, map, and quantify DNA sequences. The quality of raw data is assessed through various software like Trimmomatic Software, FASTXT Toolkit, PRINSEQ, NGS QC, FastQC, etc. WES utilizes different exome capturing kits which differ in target regions, probe type, probe length and number of probes, and required amount of input DNA. The data obtained from WES are processed through rigorous primary analysis in which all the generated sequences are aligned, processed, and then variants analysis is performed to determine the cause of disease of lethality. In the secondary analysis, variants are annotated as per their impact of the genome; then they are prioritized based on the disease they cause. These genetic variants provide detailed information regarding the inheritance pattern and biology of the diseases. However, the ascertainment of disease-causative effect of genetic variants can only be done through functional biology assays.

References

Al-Aama JY, Shaik NA, Banaganapalli B, Salama MA, Rashidi O, Sahly AN, Mohsen MO, Shawoosh HA, Shalabi HA, Edreesi MA, Alharthi SE, Wang J, Elango R, Saadah OI (2017) Whole exome sequencing of a consanguineous family identifies the possible modifying effect of a globally rare AK5 allelic variant in celiac disease development among Saudi patients. PLoS One 12:e0176664

Bailey-Wilson JE, Wilson AF (2011) Linkage analysis in the next-generation sequencing era. Hum Hered 72:228–236

Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J (2011) Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet 12(11):745

Butkiewicz M, Bush WS (2016) In Silico functional annotation of genomic variation. Curr Protoc Hum Genet 88, Unit 6 15

Chen R, Im H, Snyder M (2015) Whole-exome enrichment with the Agilent SureSelect human all exon platform. Cold Spring Harb Protoc 2015:626–633

Del Fabbro C, Scalabrin S, Morgante M, Giorgi FM (2013) An extensive evaluation of read trimming effects on Illumina NGS data analysis. PLoS One 8:e85024

Fedurco M, Romieu A, Williams S, Lawrence I, Turcatti G (2006) BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. Nucleic Acids Res 34(3):e22–e22

Gilissen C, Hoischen A, Brunner HG, Veltman JA (2011) Unlocking Mendelian disease using exome sequencing. Genome Biol 12:228

Guo W, Zhu X, Yan L, Qiao J (2018) The present and future of whole-exome sequencing in studying and treating human reproductive disorders. J Genet Genomics 45:517–525

Huang H, Fang M, Jostins L, Umicevic Mirkov M, Boucher G, Anderson CA, Andersen V, Cleynen I, Cortes A, Crins F, D'amato M, Deffontaine V, Dmitrieva J, Docampo E, Elansary M, Farh KK, Franke A, Gori AS, Goyette P, Halfvarson J, Haritunians T, Knight J, Lawrance IC, Lees CW, Louis E, Mariman R, Meuwissen T, Mni M, Momozawa Y, Parkes M, Spain SL, Theatre E, Trynka G, Satsangi J, Van Sommeren S, Vermeire S, Xavier RJ, International Inflammatory Bowel Disease Genetics, C, Weersma RK, Duerr RH, Mathew CG, Rioux JD, Mcgovern DPB, Cho JH, Georges M, Daly MJ, Barrett JC (2017) Fine-mapping inflammatory bowel disease loci to single-variant resolution. Nature 547:173–178

Korte A, Farlow A (2013) The advantages and limitations of trait analysis with GWAS: a review. Plant Methods 9:29

Kruglyak KM, Lin E, Ong FS (2016) Next-generation sequencing and applications to the diagnosis and treatment of lung cancer. Adv Exp Med Biol 890:123–136

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, Fitzhugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, Levine R, Mcewan P, Mckernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann Y, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, Mcmurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, Mcpherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M et al (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359

Le Gallo M, Lozy F, Bell DW (2017) Next-generation sequencing. Adv Exp Med Biol 943:119–148

Lee S, Abecasis GR, Boehnke M, Lin X (2014a) Rare-variant association analysis: study designs and statistical tests. Am J Hum Genet 95(1):5–23

Lee H, Deignan JL, Dorrani N, Strom SP, Kantarci S, Quintero-Rivera F, Das K, Toy T, Harry B, Yourshaw M, Fox M, Fogel BL, Martinez-Agosto JA, Wong DA, Chang VY, Shieh PB, Palmer CG, Dipple KM, Grody WW, Vilain E, Nelson SF (2014b) Clinical exome sequencing for genetic identification of rare Mendelian disorders. JAMA 312:1880–1887

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing, S (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079

Macarthur DG (2012) Challenges in clinical genomics. Genome Med 4:43

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437(7057):376

Maxam AM, Gilbert W (1977) A new method for sequencing DNA. Proc Natl Acad Sci U S A 74:560–564

Metzker ML (2005) Emerging technologies in DNA sequencing. Genome Res 15:1767–1776

Nakagawa H, Wardell CP, Furuta M, Taniguchi H, Fujimoto A (2015) Cancer whole-genome sequencing: present and future. Oncogene 34:5943–5950

Priya RR, Rajasimha HK, Brooks MJ, Swaroop A (2012) Exome sequencing: capture and sequencing of all human coding regions for disease gene discovery. Methods Mol Biol 884:335–351

Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, Searle S, Farrell CM, Loveland JE, Ruef BJ, Hart E, Suner MM, Landrum MJ, Aken B, Ayling S, Baertsch R, Fernandez-Banet J, Cherry JL, Curwen V, Dicuccio M, Kellis M, Lee J, Lin MF, Schuster M, Shkeda A, Amid C, Brown G, Dukhanina O, Frankish A, Hart J, Maidak BL, Mudge J, Murphy MR, Murphy T, Rajan J, Rajput B, Riddick LD, Snow C, Steward C, Webb D, Weber JA, Wilming L, Wu W, Birney E, Haussler D, Hubbard T, Ostell J, Durbin R, Lipman D (2009) The consensus coding sequence (CCDS) project: identifying a common protein-coding gene set for the human and mouse genomes. Genome Res 19:1316–1323

Pruitt KD, Tatusova T, Brown GR, Maglott DR (2012) NCBI reference sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res 40:D130–D135

Puritz JB, Lotterhos KE (2018) Expressed exome capture sequencing: a method for cost-effective exome sequencing for all organisms. Mol Ecol Resour 18:1209–1222

Rabbani B, Tekin M, Mahdieh N (2014) The promise of whole-exome sequencing in medical genetics. J Hum Genet 59:5–15

Regier AA, Farjoun Y, Larson DE, Krasheninina O, Kang HM, Howrigan DP, Chen BJ, Kher M, Banks E, Ames DC, English AC, Li H, Xing J, Zhang Y, Matise T, Abecasis GR, Salerno W, Zody MC, Neale BM, Hall IM (2018) Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. Nat Commun 9:4038

Rhodes J, Beale MA, Fisher MC (2014) Illuminating choices for library prep: a comparison of library preparation methods for whole genome sequencing of *Cryptococcus neoformans* using Illumina HiSeq. PLoS One 9:e113501

Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M (1977a) Nucleotide sequence of bacteriophage phi X174 DNA. Nature 265:687–695

Sanger F, Nicklen S, Coulson AR (1977b) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74:5463–5467

Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. Bioinformatics 27(6):863–864

Stitziel NO, Kiezun A, Sunyaev S (2011) Computational and statistical approaches to analyzing variants identified by exome sequencing. Genome Biol 12:227

Teare MD, Santibanez Koref MF (2014) Linkage analysis and the study of Mendelian disease in the era of whole exome and genome sequencing. Brief Funct Genomics 13:378–383

Turcatti G, Romieu A, Fedurco M, Tairi AP (2008) A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. Nucleic Acids Res 36(4):e25–e25

Visscher PM, Brown MA, Mccarthy MI, Yang J (2012) Five years of GWAS discovery. Am J Hum Genet 90:7–24

Worthey EA, Mayer AN, Syverson GD, Helbling D, Bonacci BB, Decker B, Serpe JM, Dasu T, Tschannen MR, Veith RL, Basehore MJ, Broeckel U, Tomita-Mitchell A, Arca MJ, Casper JT, Margolis DA, Bick DP, Hessner MJ, Routes JM, Verbsky JW, Jacob HJ, Dimmock DP (2011) Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. Genet Med 13:255–262

Ye BD, Mcgovern DP (2016) Genetic variation in IBD: progress, clues to pathogenesis and possible clinical utility. Expert Rev Clin Immunol 12:1091–1107

Chapter 9
Structural Bioinformatics

Bhumi Patel, Vijai Singh, and Dhaval Patel

Contents

B. Patel · D. Patel (✉)
Department of Bioinformatics & Structural Biology, School of Biological Sciences and Biotechnology, Institute of Advanced Research, Gandhinagar, India

V. Singh
School of Biological Sciences and Biotechnology, Institute of Advanced Research, Gandhinagar, India
e-mail: vijai.singh@iar.ac.in

N. A. Shaik et al. (eds.), *Essentials of Bioinformatics, Volume I*,
https://doi.org/10.1007/978-3-030-02634-9_9

9.1 Introduction

Proteins often mediate the quintessential structure and function of cells, therefore maintaining the integrity of all molecular and biological functions in the entire kingdoms of life. Proteins are complex molecules which exhibit a remarkable versatility regarding sequential and spatial and arrangements of amino acids, which allow them to perform a variety of functions that are fundamental to life. Conceivably it is the only biological macromolecule which has undergone billions of years of evolution and amassed a variety of the functions, a few of which are still unknown to humanity. Through the extensive research in protein science, researchers have agreed to a fundamental principle that the function of proteins is dependent on its structural conformation. The way proteins sequence folds in three-dimensional (3D) conformations leads to unique structures that allow spatial arrangements of chemical group's in particular 3D space. This significant placement of the chemical entity also allows the proteins to play essential distinct structural, regulatory, catalytic, and transport functions in all the kingdom of life.

9.1.1 The Building Blocks of Protein

Amino Acids

The protein sequence consists of 20 different naturally occurring amino acids that serve as building blocks of proteins. Each amino acid contains a central alpha carbon (Cα) which is attached to an amino group (NH2), a hydrogen atom (H), and a carboxyl group (COOH) as shown in Fig. 9.1a. It is diverse from one another due to the presence of side chain which is represented by the R group attached to Cα, and the variations in R group impart specific chemical properties of the residue which governs the function of proteins. Apart from these 20 naturally occurring amino acids, nonnatural amino acids also occur in rare cases because of enzymatic modifications after protein synthesis. The variation in side-chain R group and their propensity for contact with a solvent like water divide these amino acids into broadly three classes – hydrophobic, polar, and charged. There are additional subclasses such as aromatic or aliphatic (Taylor 1986a) (Fig. 9.2). The hydrophobic amino acids have a low propensity for water, which includes lysine, isoleucine, alanine, proline, valine, aromatic amino acids (phenylalanine and tryptophan), and sulfur-containing acids (methionine and cysteine). The charged amino acid includes positively charged (+) lysine, arginine, and histidine and negatively charged (−) aspartate and glutamate. The polar amino acids include asparagine, glutamine, threonine, serine, and proline. Glycine is an exception with only single H atoms in its side chain. Majority of the protein molecules have a hydrophobic core not accessible to the solvent like water and polar amino acids in the surface in contact with the solvent environment with membrane proteins as an exception. The polar surface of the macromolecule is covered by polar and charged amino acids, which are in contact

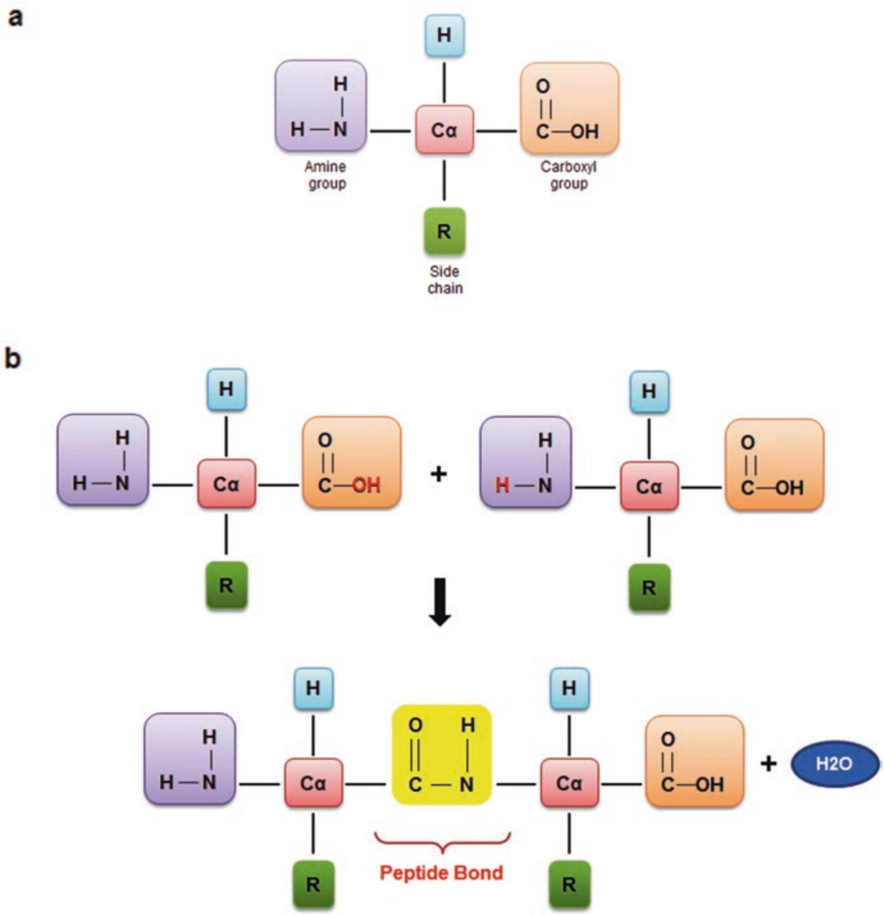


Fig. 9.1 (**a**) Representation of the amino acid structure. The central alpha carbon (Cα) is attached to an amino group (NH2), a hydrogen atom (H), a carboxyl group (COOH), and an R group which varies for 20 different amino acids. (**b**) The representation of peptide bond shown in the yellow box, by the elimination of a water molecule

with the solvent environment due to their ability to form H-bonds. A protein is synthesized by the linear succession of two or more peptide bonds joined end-to-end referring to a polypeptide. The peptide bond is formed through a condensation reaction, eliminating water between the carboxyl group (COOH) of one amino acid and the amino group (NH_2) of other amino acids (Fig. 9.1b). The polypeptide chain is formed by several peptide bond formations between amino acids where the amino group of the first and the carboxyl group of the last amino acid remain intact.

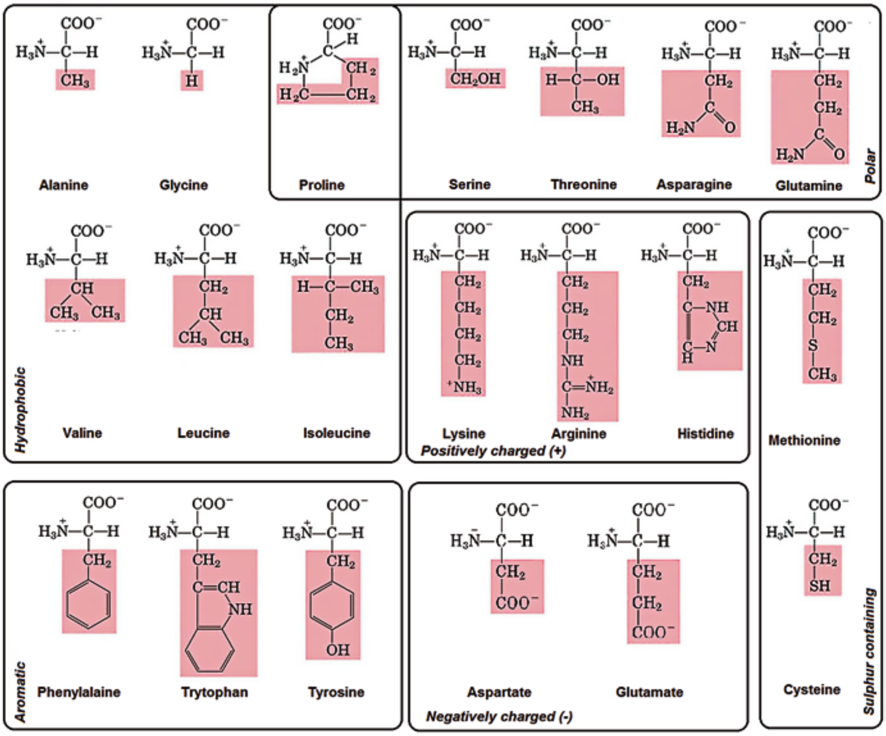


Fig. 9.2 The 20 standard amino acids in proteins labeled with its full name. The square box indicates amino acids which are grouped based on their side-chain properties like hydrophobic, polar, charge, sulfur-containing, and aromatic acids as shown. The variation in R group for each amino acid is marked with a color box. This grouping is used as a guideline but does not convey full complexity of side-chain properties, as it varies according to physiological conditions

9.1.2 The Hierarchal Representation of Proteins

The protein molecules and their complexity in arrangements are described conventionally by four levels of structure, primary, secondary, tertiary, and quaternary (Boyle 2005), as shown in (Fig. 9.3).

9.1.2.1 Primary Structure

The linear sequence of amino acids in the protein is generally referred to as the primary structure of the protein which includes all the covalent bonds between the amino acids. Proteins are connected as a linear polymer of 20 different amino acids (Rödel 1974), by forming a peptide bond between amino acids (Fig. 9.3). The polypeptide sequence of a protein can contain (*n*) number and the combination of 20

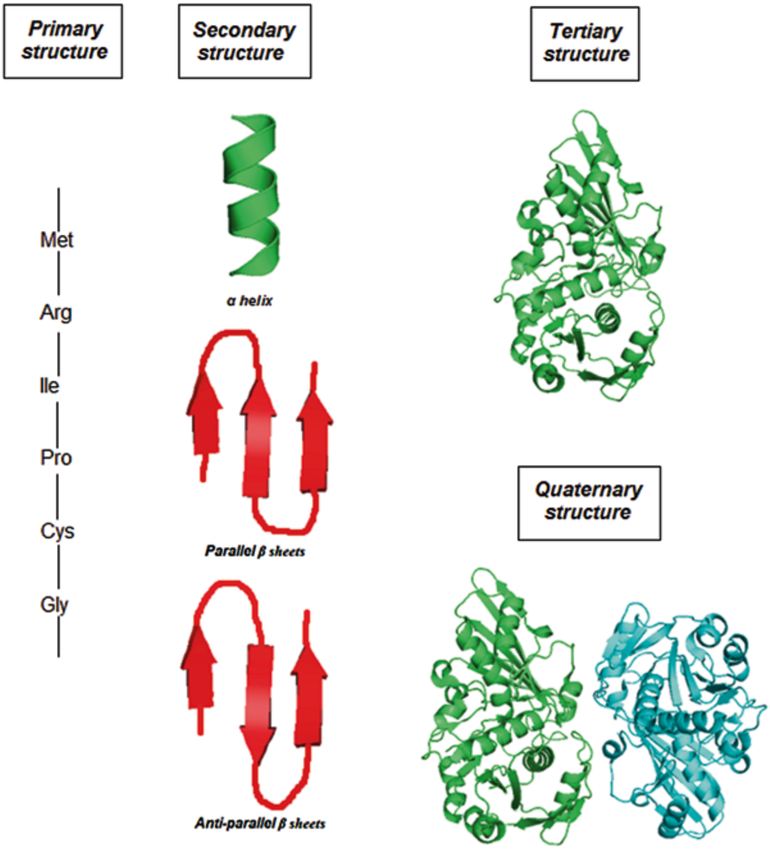


Fig. 9.3 Structural organization of four different protein levels: primary, secondary, tertiary, and quaternary

amino acids, in any order. Similar to the alphabet combinations which form meaningful words and sentences in vocabulary, nature selects the combinations of different amino acids to form polypeptide for their respective functions.

9.1.2.2 Secondary Structure

The secondary structure of a protein refers to recurring and regular spatial arrangements of adjacent linear amino acids as local conformation of the polypeptide chain. The major secondary structural elements which are identified during protein structure research are alpha (α) helix and beta (β) sheets. Secondary structures were predicted by Linus Pauling, Robert Corey, and H. R. Branson before the experimental determination of structures based on the known physical limitations of the polypeptide chain conformation (Pauling et al. 1951; Taylor n.d.). There is a considerable degree of regularity in these secondary structures, particularly the psi (Φ) and phi (Φ) angle combinations which are repeated for the secondary structure in a polypeptide chain. Both helix and sheets are the only regular secondary structural components universally present in proteins as they satisfy the peptide bond geometry constraints as well as due to the H-bond interactions between the backbone atoms of the amino acids in them which help them to make highly favorable and stable conformation, though the irregular structural components such as turns and loops are also observed in protein, mainly in the globular proteins which are vital to both structural integrity and function.

α-Helix

The α-helix was first proposed by Linus Pauling and co-workers in 1951 (1951), and a typical α-helical conformation is generated by curving the polypeptide backbone to produce a regular coil. In this helical structure, the backbone of the polypeptide chain is coiled around the axis of the molecule in such a way that the side-chain R groups of residues project toward outside from the helical backbone. The number of residues required to complete a single turn/coil is 3.6 in α-helix. In a single turn of α alpha helix, there is an H-bond interaction between each of the carbonyl oxygen (*n*) of the backbone and the amide proton of the 4th residue ($n + 4$) toward the C-terminus. These H-bonds stabilize the helical conformation and are almost in parallel orientation to the long axis of the helix. The backbone of the polypeptide can be coiled in two directions (right or left); the helix with a rightward coil is called as a right-handed helix, and the other one is called as a left-handed helix. There is a steric hindrance in the formation of left-handed helices; the majority of the proteins have right-handed helices. Other types of helices have rarely been observed in the proteins, like the 3^{10} helix which have three residues per turn with H-bond between n residue and the 3rd residue ($n + 3$) ahead toward the C-terminus (Taylor n.d.; Huggins 1943). A rare type of pi (π) helix, which is found in some polypeptides, possesses 4.4 residues per turn with H-bond between n residue and the 5th residue ($n + 5$) toward the C-terminus(Low and Baybutt 1952).

β-Sheets

In contrast to α-helix, where H-bond is formed between the neighboring residues within a single chain, β-sheets are formed by H-bond between adjacent polypeptide backbones in chains. These sections of adjacent polypeptide chains are known as β-strands. The β-sheets comprise of H-bonds formed between carbonyl oxygens and amide hydrogen on adjacent β-strands. Unlike the α-helix, the H-bond is almost in perpendicular to the extended β-strands. The β-strands are in two possible configurations; it may be a parallel sheet (same N- to C-terminal direction) or antiparallel sheet (opposite N- to C-terminal direction). A very rare mixed configuration also exists with both parallel and antiparallel sheets. In antiparallel β-sheets, a variant called as β-bulge is characterized by H-bond formation between two residues on one β-strand and one residue on the adjacent β-strand *(*Richardson et al. 1978*;* Chan et al. 1993*)*.

Other Secondary Structures

Although α-helices and β-sheets are considered as major secondary structural elements in proteins, these elements are interspersed in regions of an irregular structure, also called loops or coils. These elements are not only involved in transitions between regular secondary structures but also possess structural significance from function parse and can be the location of the functional active site and usually present at the surfaces of the proteins acting as a mediator of interactors with other biological molecules. Generally, residues with small side chains (R groups) are often found in turns such as proline, cysteine, serine, aspartate, asparagine, and glycine. Systematic analysis of residues in turns has revealed that amino acids have bulky or branched side chains occurring at very low frequencies. Different types of turns have been identified such as Hairpin loops (reverse turns) which often occur between antiparallel β-strands involving minimum number of residues (4–5) required to begin the next strand. An omega (ω) turn involving (6–16) residues is also sometimes observed. Certain structures also have extended loops, involving more than 16 residues and as much as 10 different combinations based on the number of residues in turns and the φ and ψ angles associated with the central residues.

9.1.2.3 Tertiary Structure

The tertiary structure (3°) of a protein is described as the spatial relationship formed as global 3D structure among all the amino acids in a polypeptide chain. The tertiary structure was first described by Alfred Mirsky and Linus Pauling in 1936 as a molecule consisting of one polypeptide chain which is folded into unique configuration throughout the molecule (Mirsky and Pauling 1936). They predicted the role of H-bonds in interactions of side chains of amino acids in protein structure. Subsequently, the determination of hemoglobin (Perutz et al. 1960) and myoglobin (Kendrew et al. 1958) has confirmed that other forces were also important especially noncovalent interaction which also helps in the stabilization of the structure. Thus, the formation of tertiary structure brings the non-neighboring amino acid residues in the primary structure close to each other and helping to generate a protein fold which is a determinant factor for protein functions. The tertiary structure is also commonly referred as protein fold which is the global conformation of all the secondary structures forming a compact globular molecule wherein the secondary structural elements interact via electrostatic interactions, van der Waals interactions, hydrophobic contacts, disulfide and salt bridges, and H-bonds between non-backbone atoms. The efforts in protein structure prediction help to decipher how secondary structural elements unite in three-dimensional space to generate the tertiary structure.

9.1.2.4 Quaternary Structure

In contrast to the tertiary structure which describes the organization of a single polypeptide chain, the quaternary structure is an association of two or more independently folded polypeptides within the protein through noncovalent interactions. Most proteins do not function as a monomer but rather function as multi-subunit or multimeric or oligomeric proteins. In certain types of proteins, the quaternary structure formation is very important from the functional perspective as they allow the formation of binding or catalytic sites between the interfaces of subunits which are not possible in case of single subunit proteins. Enzymes are known to be involved in allosteric regulation which frequently arises due to conformational changes occurring due to ligand/substrate binding in oligomeric proteins. In quaternary structures, the subunits may be identical resulting in a homomeric protein or may be different resulting in a heteromeric association of proteins. They were first observed by The Svedberg in 1926 using analytical ultracentrifuge to determine the molecular weights of proteins which resulted in the separation of multi-subunit proteins (Svedberg and Fåhraeus 1926). The stabilization forces and interactions in the quaternary structure are of the same types as observed in the secondary and tertiary structure stabilization. The surface regions of monomer unit involved in the subunit interactions comprise with nonpolar side-chain residues and residues capable of forming the H-bond and disulfide bonds.

9.2 Protein Structure Predictions

In the past, Anfinsen demonstrated that the unfolded proteins can fold back into their native conformation only by their primary structure or amino acid sequence (Anfinsen et al. 1961). It has laid the foundation of computational methods of predicting tertiary structures from the primary sequence. The ultimate goal of protein structure prediction is to elucidate a structure from its primary sequence, with accuracy comparable to results achieved experimentally using X-ray crystallography and NMR. To achieve the thermodynamically stable fold better than other conformations, the protein must evolve the folding by optimizing the interactions within and between residues and satisfy all the spatial constraints between the atoms of a peptide bond. Though we do have an understanding of the general nature of inter- and intramolecular interactions that determine the protein fold, it is yet challenging to ascertain the structures of protein from basic physiochemical principles.

Why do we need to predict the structures of protein? The answer lies in the fact that the protein structural attributes lead to biological functions, and computational prediction methods are the only way and convenient in all contexts where experimental techniques fail. Many proteins are too large for NMR or lack the propensity to form diffraction quality crystals for X-ray diffraction, so in such cases computational method for structure prediction is the only approach. The 3D structure prediction from its primary structure is the much-debated area of structural bioinformatics. Despite the development of recent enormous algorithms and the computational methods for protein structure prediction, a comprehensive solution for accurate prediction of protein folding still remains elusive. Many structural bioinformatics researchers have introduced several methods and algorithms to solve this problem, but each method has both advantages and disadvantages. Globally, a competition has been set up to evaluate the performance of several structure prediction tools/softwares using blind test on several experimentally predetermined structures of proteins. This competition was started in 1995 as Critical Assessment of Techniques for Protein Structure Prediction (CASP) which provides a global benchmark for this exhaustive computational purpose (Moult et al. 2018).

Decades of research have provided insights into the various ways to accurately predict the 3D structures of proteins like template-based methods (homology modeling), fold recognition (also known as threading), a new fold method, and de novo (ab initio) methods of structure prediction. The homology modeling methods are also known as comparative modeling, and prediction of the query structure is based on close homologs (>25%) of experimentally known structure deposited in the Protein Data Bank (PDB) public domain. The fold recognition method is generally used when a structure with similar folds is available, but lacking a close relative for homology modeling. The new fold method is employed when no structure with the same folding pattern is known, and it requires a priori or knowledge-based methods for prediction. In ab initio methods for structure prediction as the name suggests, prediction is performed from scratch using the amino acid information only. With the introduction of advanced algorithms and availability of experimental solved 3D structures, many software/tools are continuously being developed combining different classical prediction methods described above.

9.2.1 Homology Modeling

This comparative modeling method is based on the fact that when the amino acid sequence of the query structure is homologous to that of the one or more experimentally known structures, the resulting structural fold will also be similar. This relationship was identified first in 1986 by Chothia and Lesk where they have concluded that despite years of evolution, the structure is more stable and less susceptible to changes than the associated sequence and thus similar sequences as well as distantly related sequences to certain extent folds into a similar fold (Chothia and Lesk 1986). If the percentage identity between the query sequence and the known structures falls in the “Safe” region, two sequences may practically have an identical fold. As a rule of thumb, this “Safe” zone should have at least 30–50% identical amino acids in an optimal sequence alignment, and the resulting homology model can be sufficiently used for other application. The homology model output is generated simply by copying aligned regions of the polypeptide from the template/homologous structure, by altering the side chains wherever necessary, and by mutating those residues that differ between sequence alignments. The final model which is created by homology modeling contains enough information about the 3D arrangement of the essential and sufficient residues for the design of subsequent experiments, like structure-based drug discovery and site-directed mutagenesis.

There are numerous tools for generating a homology model, and most of them have consensus steps involved in the generation of structure which are as follows:

1. Template identification
2. Initial sequence alignment
3. Alignment correction
4. Backbone generation
5. Loop modeling and optimization
6. Side-chain modeling and optimization
7. Overall model optimization
8. Model validation

Briefly, the process starts with sequence similarity search for the target sequence and the known structure of a protein using BLAST (Altschul et al. 1990) or PSI-BLAST (Altschul et al. 1997) or fold recognition (Jones et al. 1999) and PDB structure database (www.rcsb.org). In a case where percentage identity is often low, one or more possible template is selected, and further alignment correction is performed. After initial alignment, alignment correction is undertaken wherein the low-percentage regions are more carefully corrected using multiple sequence alignment to generate position-specific scoring matrices also known as profiles (Taylor 1986b). For such alignment refinement and correction, generally preferred tools are MUSCLE (Edgar 2004) and T-COFFEE (Notredame et al. 2000). These corrections and refinements are very crucial to predict the best quality model. Further backbone generation is performed where the coordinates of aligned residues in the template and model are simply copied to the initial backbone model. It is often found that the experimentally determined structures contain errors because of weak electron density map and one can use structure validation tool like PDBREPORT (http://www.cmbi.ru.nl/gv/pdbreport/) for manual inspection. The Swiss Model server (Biasini et al. 2014) (http://swissmodel.expasy.org/) uses multiple templates to create optimum backbone to compensate the missing residues which are not aligned in single template selection. Swiss Model, Phyre, and MODELLER are currently widely used as free homology modeling tools. Modeler is most widely used homology modeling tool which does not create a single backbone; instead, it uses the alignment to derive restraints such as H-bonds, torsion angles to build the model satisfying the restraints (Šali and Blundell 1993).

The next step is the loop modeling of the insertion or deletion in the alignment which is not previously modeled, and the changes are outside the regular secondary structural elements. Two widely used approaches are the knowledge-based (Michalsky et al. 2003) and the energy-based (Xiang et al. 2002). In the first approach, a search in the PDB is performed for known similar loops with matching the endpoints between the loops and simply copying the conformation of loops. In the second approach, the sampling of random loop conformation is performed with energy minimization using the Monte Carlo or molecular dynamics to find maximum energy-minimized loop conformation. Further, side-chain modeling is carried out where the conserved residues from the template are copied because the structurally similar proteins have an identical torsion angle (psi-angle) while comparing the side-chain conformations. This task is generally knowledge-based where the rotamer libraries are built from high-resolution X-ray structures by collecting the stretches of three-seven residues with the query amino acid in the center. After the initial model building, the model optimization takes place where the incorrectly predicted rotamers are optimized by restraining the atom position and applying a few steps of energy minimization using molecular simulations to correct the errors. The final step is model validation where the model is assessed for any kind of errors accumulated during previous steps. In model validation, the method checks for the correct bond angles and lengths; however, one should be cautious as this validation cannot judge whether the model is correctly folded or not. Table 9.1 summarizes the widely used academically free tools and server for homology modeling.

Table 9.1 Summary of widely used academically free tools and server for homology modeling

Program	Web address	Availability	Method
SWISS-MODEL	http://swissmodel.expasy.org/	Free	Rigid-body assembly
MODELLER	https://salilab.org/modeller/	Academically free[b]	Spatial restraints
Phyre2	http://www.sbg.bio.ic.ac.uk/phyre2/	Free	Profile-based alignment
HHpred	https://toolkit.tuebingen.mpg.de/#/tools/hhpred	Free	Profile-based HMM
RaptorX	http://raptorx.uchicago.edu/	Free	Rigid-body assembly
Protein Model Portal	https://www.proteinmodelportal.org/	Free	Metaserver[a]
ROBETTA	http://robetta.bakerlab.org/	Academically free[b]	Metaserver[a]
FoldX	http://foldxsuite.crg.eu/	Academically free[b]	Rigid-body assembly
ESyPred3D	https://www.unamur.be/sciences/biologie/urbm/bioinfo/esypred/	Free	Alignment using Neural Network

[a] Metaservers use existing different servers for final prediction

[b] Academically free requires registration using academic email

9.2.2 *Fold Recognition Method*

Despite more than a decade of research and understanding the protein folding process, still, the current-day knowledge is not sufficient enough to predict protein structures just based on the amino acid sequence. It is well-known fact that the protein which does not have detectable sequence similarity still adopts a similar fold and function. Thus, the fold recognition approach has been developed to examine which protein models have a similar fold of known structures. Unlike homology modeling, where target protein structures are modeled based on homologous template deposited in the Protein Data Bank (PDB), the fold recognition methods work by utilizing statistical knowledge of the relationship between structures deposited in the PDB and the query protein sequence.

These methods have made possible to find a suitable fold of an unknown query protein when the alignment is around the "twilight zone" and with no homology to known 3D structures. The two widely used methods in fold recognition area are 3D profile-based method and threading. The 3D profile-based method works based on the physicochemical attributes of residues of the query protein, which must fit within the environment in the 3D structure. The parameters which relate to the suitable environment are the buried cavities in protein which are inaccessible to solvent, polar atoms (O and N) covering the side-chain area, and the local secondary structure (Bowie et al. 1991). In threading approach, the residues in the sequence are fitted into the backbone coordinates of known structures by comparison in 3D space (Jones et al. 1992). The term "threading" is used since the prediction is made by placing and aligning each amino acid in the query sequence with respect to template structure and later evaluating how well the target fits the template. The threading approach relies on a basic fact: that there is a relatively small amount of different known folds (approximately 1400) and that most of the newly submitted structures in the database have similar structural folds already deposited in the PDB.

The key difference between the homology modeling and fold recognition method is that the homology modeling uses only sequence homology for prediction by treating the matched template in an alignment as a sequence, while in threading structure and sequence in the matching template is used for prediction. In the absence of significant sequence homology, protein threading still makes a prediction based only on the structural information making the threading method more effective than the previous one where the sequence match is below "Safe" zone. One of the significant limitations of the threading method is that the correct prediction is only feasible when an appropriate fold match is present in the library of known structures. The list of widely used freely available prediction methods which uses the threading method is mentioned in Table 9.2.

The basic steps of threading approach are as follows:

- The generation of structure template database: A library is constructed for different protein folds from protein structure databases such as PDB, SCOP, and CATH.

Table 9.2 List of threading software and servers widely used and freely available

Program	Web address	Availability	Method
I-TASSER	https://zhanglab.ccmb.med.umich.edu/I-TASSER/	Academically free[a]	Iterative threading assembly refinement
HHpred	https://toolkit.tuebingen.mpg.de/#/tools/hhpred	Free	Profile-based HMM
pGenTHREADER	http://bioinf.cs.ucl.ac.uk/psipred?pgenthreader=1	Free	Profile-based fold recognition
GenTHREADER	http://bioinf.cs.ucl.ac.uk/psipred?genthreader=1	Free	Rapid fold recognition
IntFOLD	http://www.reading.ac.uk/bioinf/IntFOLD	Free	Multiple-template modeling approach using sequence-structure alignment
Phyre2	http://www.sbg.bio.ic.ac.uk/phyre2/	Free	Profile-based alignment and secondary structure matching
Sparks X	http://sparks-lab.org/yueyang/server/SPARKS-X/	Free	Probabilistic-based sequence-to-structure matching
FUGUE2	http://mizuguchilab.org/fugue/	Free	Sequence-structure compatibility based on databases of structural profiles

[a] Academically free requires registration using academic email

- The designing of the scoring function: The fitness between the query sequence and the templates is measured with scoring functions and knowledge-based relationships between the sequence and structures.
- Optimal fitting into the library: The target sequence is then favorably fitted to each library fold with considerations of insertions and deletions in the loop regions.
- Threading alignment: The alignment of the query sequence with each template structure is generated by optimizing the scoring function. The energy of scoring function of each possible fit is computed by summing the pairwise interactions and energy solvation mainly by two approaches: one by using dynamic programming with a frozen paring environment where interaction partners are chosen from template protein followed by iteration and other by using Monte Carlo method.
- Final threading prediction: The most optimal threading match from the library of folds is chosen based on ascending order of total energy and choosing a template with the lowest energy folds. The final structure model is constructed by placing the backbone atoms of each residue in the query sequence at the aligned backbone spatial coordinates of the selected structural template.

9.2.3 Ab Initio Modeling Method

The ab initio method is also known as de novo prediction that attempts to predict the 3D structure based only on the primary sequence such as amino acid composition. All the information essential for a polypeptide to fold in its native state is already embedded in the protein's amino acid sequence as demonstrated by Anfinsen in 1961 (Anfinsen et al. 1961). Generally, this method is used as a last resort in protein structure prediction when the structural information homolog is missing.

The ab initio method predicts native conformations by computing the most favorable energy conformations. Only few computational programs rely on this method because it requires massive computational power and also due to the limited knowledge available about protein folding patterns. The key areas involved in de novo prediction are accurate energy, scoring functions, and efficient sampling conformation spaces. The Zhang lab's QUARK (Xu and Zhang 2013) and ROSETTA servers (Kim et al. 2004) are the two best prediction methods which make use of ab initio methods combined with other methods, to predict tertiary structure of proteins. The programs and servers which use ab initio method for structure prediction are summarized in Table 9.3.

During the process of ab initio prediction, several sets of candidate structures, also called as decoys, are computed. Furthermore, native-like conformations are then selected from these sets of decoys based on the theory that native protein fold has the lowest entropy and free energy. Several programs which are successful in ab initio prediction generally use knowledge-based methods for prediction of conformation with the lowest free energy followed by fold prediction and threading. A major limitation of this method is a requirement of huge computational power. To solve this issue to a limited extent, Rosetta@home (https://boinc.bakerlab.org/) forum was created which combines individual's home computer idle time for distributed calculations. Another method to overcome the requirement of high computational facilities is involving the use of Monte Carlo models (Jayachandran et al. 2006) by refining computer simulations and also by the use of coarse-grained modeling (Kmiecik et al. 2016).

Table 9.3 List of ab initio-based prediction servers widely used and freely available

Program	Web address	Availability	Method
EVfold	http://evfold.org/evfold-web/evfold.do	Free	Evolutionary couplings calculated from correlated mutations in a protein family
FALCON	http://protein.ict.ac.cn/FALCON/	Free	Position-specific hidden Markov model
QUARK	http://bioinf.cs.ucl.ac.uk/psipred?genthreader=1	Free	Replica-exchange Monte Carlo simulation
ROBETTA	http://robetta.bakerlab.org	Free	Ab initio fragment assembly with Ginzu domain prediction

Table 9.4 List of widely used programs for protein structure model evaluation and quality check

Program	Web address	Availability	Method
PROCHECK	https://www.ebi.ac.uk/thornton-srv/software/PROCHECK/	Free	Checks stereochemical Parameters including the Ramachandran plot
ERRAT	http://services.mbi.ucla.edu/ERRAT/	Free	Analyzes the statistics of nonbonded interactions and plots the value of the error function versus a 9-residue sliding window position
WHAT_CHECK	http://servicesn.mbi.ucla.edu/WHATCHECK/	Free	Checks stereochemical Parameters
Verify3D	http://servicesn.mbi.ucla.edu/Verify3D/	Free	Check 3D model with its own residues based on its location and environment (alpha, beta, loop, polar, nonpolar, etc.)
ProSA web	https://prosa.services.came.sbg.ac.at/prosa.php	Free	Check 3D models of proteins structures for potential errors
WHAT_IF	http://swift.cmbi.ru.nl/whatif/	Free	Protein model assessment especially for a point mutation
MolProbity	http://molprobity.biochem.duke.edu/	Free	Validation using all-atom contact analysis and geometrical criteria for phi/psi, sidechain rotamer, and C-beta deviation
The Protein Model Portal	https://www.proteinmodelportal.org/?pid=quality_estimation	Free	Metaserver[a]
SAVES	http://servicesn.mbi.ucla.edu/SAVES/	Free	Metaserver[a]

[a] Metaserver uses existing different programs for evaluation

9.3 Protein Structure Validations

Proteins are the workhorse of all the biological processes in an organism, and the key to their functions is the 3D structure and dynamics of the protein. To get a better understanding of these functions, we need correct native fold prediction about the protein model. Therefore to increase the reliability, the protein structure prediction should be followed by quality check and assessment. One needs to select the best model from an ensemble of predicted models i.e., either from different modeling methods/servers or based on prediction from different template alignments and structures. The best way is to generate different models using the different methods from the available modeling servers and have a model quality assessment to choose the top stereochemically validated model. The primary reason for not relying on only one prediction method or one template is due to suboptimal methods for target-template alignments, low-resolution template structure, and structural inaccuracies introduced by modeling program. The main objective of model verification programs is to detect unreliable segments in the model by evaluating their stereochemical and geometrical quality so that the models are suitable for subsequent applications. Different packages and online servers are available for modeling quality assessment that is listed in Table 9.4. A number of methods and parameters to be checked in model quality estimation are listed below:

9.3.1 Ramachandran Plot

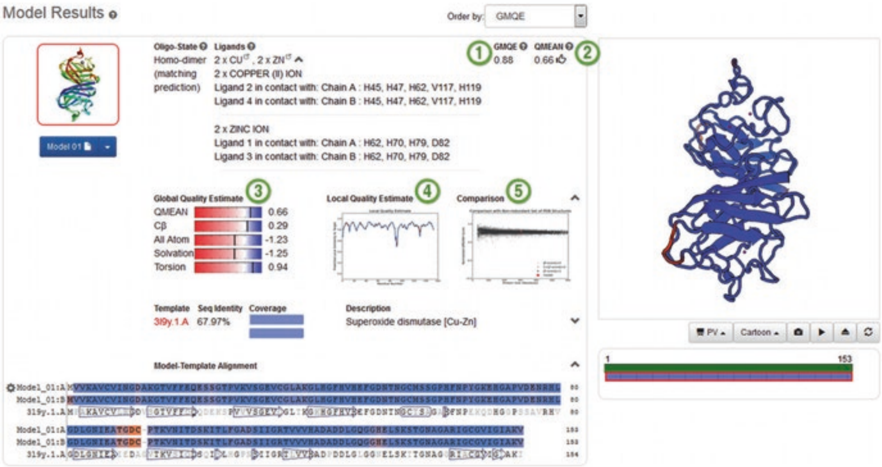
In a peptide bond, the likely conformations for a polypeptide chain are quite restricted due to the limitation of rotational freedom at φ (Cα−N) and ψ (Cα−C) angles by steric hindrance between peptide backbone and the side chains of the residues. Ramachandran plot maps the entire conformational space for a polypeptide (plot of ψ vs φ) and illustrates the allowed and disallowed residues in this conformational space (Ramachandran et al. 1963). One can check the Ramachandran statistics to assess the allowed and disallowed residues in the protein model and select those folds in which more than 90% of residues fall in the allowable region. As a rule of thumb, >90% allowed region criteria should be followed, or at least the residues critical for the function of protein or residues in the active site should be in the allowed region. As it is the most powerful check for protein stereochemical check for protein structure, an attempt should be made to energy minimize the structure and solve the error regions so that it follows the Ramachandran statistics.

9.3.2 CASP: Benchmarking Validation Test

Which server/methods generate the best possible native-like conformation structure? It is a very difficult question to answer as the complexity of protein modeling lies in the method used to predict the correct folding of the sequence from the primary amino acid residues. Any protein prediction server whether it is based on homology modeling, threading, or de novo prediction method has advantages along with certain disadvantages. As a rule of thumb, the combination of multiple methods also called as “metaserver” approach is more reliable prediction than individual methods. An ideal way to select the best prediction server is to model the protein structure using the top 3–5 best performing algorithm from the list of an independent validation benchmark such as CASP, a Critical Assessment of Protein Structure Prediction, followed by model quality assessment and validation. CASP is a biennial competition for benchmarking all the available servers/programs introduced by Moult in 1995 (Moult et al. 2018). The algorithm is trained on databases that already contained the structures so there are chances that a predicted output would be biased; to solve this issue, CASP was introduced where the experimentalist will solve the structure but will not deposit in a public forum and keep it undisclosed to the prediction servers. Once the prediction is made for these sequences, it will be compared with the solved structure to see how correctly the model is predicted.

9.3.3 Swiss Model Validation Server

Swiss model output page describes the model quality in two ways: GMQE (Global Model Quality Estimation) and QMEAN. The GMQE is a model quality approximation which is based upon the target-template alignment and the template identification method. The GMQE score as shown ((1) in Fig. 9.4) should be between 0 and 1, where the number close to 1 indicates the higher reliability of the predicted model. The QMEAN score (Benkert et al. 2008) is an estimator based on geometrical properties providing both global (entire structure) and local (per residue-wise) score. The QMEAN scores are transformed into Z-score whose value indicates what one would expect from the experimentally determined X-ray structures. This QMEAN Z-score ((2) in Fig. 9.4) is an estimation of the "degree of nativeness" of the structural features detected in the model on a global scale. QMEAN Z-score around 0 indicates good agreement between the model structure and the known experimental structures of related attributes. The prediction where the Z-score is <= −4.0 indicates low-quality model. There are four individual Z-scores ((3) in Fig. 9.4) which compare the interaction potential between all atoms, Cβ atoms only, solvation, and torsion angle, where positive values indicate model score higher than experimental structure average and vice versa. Besides these, a local quality plot ((4) in Fig. 9.4) is also generated per residue (mentioned on the *x*-axis) and its anticipated similarity to the native structure (mentioned on the *y*-axis). The residues showing score below 0.6 are estimated to be of low quality. The comparison plot ((5) in Fig. 9.4) shows model quality scores of individual models with respect to


Model Results
Model 01
1
2
3
4
5
Local Quality Estimate
Comparison
Model-Template Alignment
PV
Cartoon
1
153
Model_01:A MVVKAVCVINGDAKGTVFFEQESSGTPVKVSGEVCGLAKGLHGFHVHEFGDNTNGCMSSGPHFNPYGKEHGAPVDENRHL 80
Model_01:B MVVKAVCVINGDAKGTVFFEQESSGTPVKVSGEVCGLAKGLHGFHVHEFGDNTNGCMSSGPHFNPYGKEHGAPVDENRHL 80
3l9y.1.A 80
Model_01:A GDLGNIEATGDC-PTKVNITDSKITLFGADSIIGRTVVVHADADDLGQGGHELSKSTGNAGARIGCGVIGIAKV 153
Model_01:B GDLGNIEATGDC-PTKVNITDSKITLFGADSIIGRTVVVHADADDLGQGGHELSKSTGNAGARIGCGVIGIAKV 153
3l9y.1.A 154

Order by: GMQE

Oligo-State: Homo-dimer (matching prediction)

Ligands: 2 x CU, 2 x ZN

2 x COPPER (II) ION
Ligand 2 in contact with: Chain A : H45, H47, H62, V117, H119
Ligand 4 in contact with: Chain B : H45, H47, H62, V117, H119

2 x ZINC ION
Ligand 1 in contact with: Chain A : H62, H70, H79, D82
Ligand 3 in contact with: Chain B : H62, H70, H79, D82

GMQE 0.88 QMEAN 0.66

Global Quality Estimate	
QMEAN	0.66
Cβ	0.29
All Atom	-1.23
Solvation	-1.25
Torsion	0.94

Template	Seq Identity	Coverage	Description
3l9y.1.A	67.97%		Superoxide dismutase [Cu-Zn]

Fig. 9.4 The sample evaluation report of the SWISS-MODEL server for the protein structure. Different scoring parameters are indicated by the number in a green circle. 1 GMQE (global model quality estimation); 2 QMEAN Z-score; 3 four individual Z-scores; 4 per-residue local quality plot; 5 comparison plot with respect to experimentally determined structures. (Source: https://swissmodel.expasy.org/docs/help)

scores of experimental structures of similar size. In the plot, the *x*-axis shows the length of the protein, and the normalized QMEAN score is represented on the *y*-axis where every dot signifies one experimental protein structure. The experimental structures are reported in dots by black color (Z-score between 0 and 1), gray color (Z-score between 1 and 2), and light gray color with further Z-scores from the mean. Red star represents the actual predicted model.

9.3.4 MODELLER Evaluation Criteria

Currently, the most extensively used package for protein structure prediction is MODELLER (Šali and Blundell 1993). It is often called as a comparative model building package as it also builds the model based on the homologous template. If the sequence identity between the template and query sequence is >30%, MODELLER almost predicts the model with higher accuracy. Modeler has an internal evaluation for self-consistency checks to check that whether model satisfies the restraints or not. The model stereochemistry like bonds, dihedral angles, bond angles, and nonbonded atom-atom distances is assessed using PROCHECK and WHATCHECK. The other way to check whether the model is predicted accurately or not is to compare the PROSAII Z-score of the template and model structure, and as the Z-score is a compatibility between the sequence and its structure, the model Z-score should agree with the template Z-score on which the alignment is computed. One can also check the "pseudo energy" profile of a model generated by PROSAII, as the error in the model is reflected by the peak in the energy profile in that region. Another way of evaluating the model prediction is by assessing the DOPE (discrete optimized protein energy) score (Shen and Sali 2006). DOPE is integrated into MODELLER package, and it assesses the energy model computed through iterations by the satisfaction of spatial restraints. The prediction includes generation of many decoys while predicting the native-like model, and DOPE score helps in identifying the native-like models from hundreds of decoys, i.e., the lower the score, the higher the reliability. It is generally used to check the quality of the global model, but alternatively, it can also be used to generate residue-wise energy profile of the predicted model which can be helpful to spot the error region in the model structure.

9.3.5 I-TASSER Model Validation

Apart from the tertiary structure prediction output, I-TASSER also gives secondary structure prediction results where the residue-wise prediction (confidence score) is between 1 and 9 and higher prediction indicates higher confidence. The output also gives predictions for solvent accessibility score for each residue between 0 (buried residue) and 9 (highly exposed residue). The server also gives top 10 threading templates used to construct the template alignment with normalized Z-score where the normalized Z-score > 1 indicates a good alignment. The final prediction gives the top 5 model structure a confidence score called as C-score. The C-score is critical for estimating the quality of the model which is calculated based on the consequence of threading template alignments and convergence parameter. The C-score is generally in the range of [−5,2], where higher values signify a model with high reliability and vice versa. Besides these, each model also has RMSD and the TM-score, where RMSD is an average deviation of all residue pairs in two structures, i.e., local error, and the TM-score is the score for structural similarity between two structures. The lesser the RMSD, the better the model with lesser spatial deviation as compared to template structure and vice versa. While judging the model based on TM-score, one should select a model which TM-score is >0.5 and not consider a model with TM-score <0.17 which indicates a random similarity.

Generally, while performing structure prediction using I-TASSER, it is observed that models have bad Ramachandran statistics as compared to other homology prediction programs. The simple explanation for this is I-TASSER builds a model by reassembling the structural fragments from multiple templates so the model sometimes has more energetically unfavorable regions in the Ramachandran plot. To overcome this problem, one can do post-prediction refinement using simulations, including solvent, or use online servers like FG-MD (Zhang et al. 2011) (https://zhanglab.ccmb.med.umich.edu/FG-MD/) or ModRefiner (Xu and Zhang 2011) (https://zhanglab.ccmb.med.umich.edu/ModRefiner/) to improve overall Ramachandran statistics. The user should keep this in mind that the local structure improvement comes with a cost of deviation in global topology.

9.4 Protein Structure Superimpositions and Deviation

It is true that sequences of similar proteins tend to have similar fold and in turn similar biological functions. It is also often found that two proteins with no detectable sequence similarity also have a similar fold and may function similarly. In such a case, it is necessary to devise a method where we can compare protein structures to elucidate common regions. Thus, programs were developed which perform structure-based multiple sequence alignment to apprehend the influence of similar/dissimilar regions in the structure. But one should keep in mind the differences between structure superposition and structure alignment. Structure superposition refers to the examination of two or more structures to evaluate for similarities in their 3D structure, while structural alignment refers to identifying equivalences between amino acid sequences based on 3D structure. In structure superposition, the C-alpha positions are the anchor points between structure A and B, and a transformation technique is performed which minimized the distance between these aligned residues. The solution to this approach is to produce the lowest value of root-mean-square deviation (RMSD) between A and B.

The root-mean-square deviation (RMSD) is calculated by the squared difference between two sets of atomic coordinates after superposition (Gu and Bourne 2009). In the process of comparing two structures, the coordinates may not be suitable for comparison functions like the translation and rotation which are needed to be applied to one of the two structures to minimize the RMSD, and this method is called as superposition. The RMSD values are also used in model quality evaluation where lower RMSD values indicate a lesser deviation between template and model structure, and eventually it signifies that the model has more nearer native-like fold and also helps in identifying the dissimilarity between them.

Why Is Structure Comparison and Alignment Important?

- To help with the assignment of fold classes and topology of newly determined protein structures/models.
- Structure comparison of fold and function of a known protein with an unknown protein can help to elucidate the function of the unknown protein.
- In the era of next-generation sequencing of genomes, a structural comparison can be helpful to determine the fold and function of protein where no prior knowledge of the biological function exists.
- Structural alignments can help in identifying the distant sequence relationships in a spatial arrangement not available from sequence-based alignment alone and which can be used in protein modeling and engineering.
- Useful in clustering the PDB dataset based on RMSD values so the target is compared only to a subset of all the PDB entries and thus makes the algorithm faster without compromising the accuracy.

Few of the programs and algorithms which are currently widely used for the structure comparison and alignment are mentioned in Table 9.5.

Table 9.5 List of commonly used servers/programs for structural superposition and alignment

Program	Web address	Availability	Method
SuperPose	http://wishart.biology.ualberta.ca/SuperPose/	Free	Protein superposition using modified quaternion approach
VAST	https://www.ncbi.nlm.nih.gov/Structure/VAST	Free	Identify similar protein 3D structures by purely geometric criteria
TM-align	https://zhanglab.ccmb.med.umich.edu/TM-align/	Free	Optimized residue-to-residue alignment based on structural similarity using dynamic programming iterations
MATRAS	http://strcomp.protein.osaka-u.ac.jp/matras/	Free	Markov transition model of structure evolution for homology detection
FATCAT	http://fatcat.burnham.org/	Free	Flexible structure alignment by chaining aligned fragment pairs allowing twists
BioSuper	http://wwww-ablab.ucsd.edu/biosuper	Free	Gaussian-weighted RMSD Superposition of proteins for flexible loops and hinged domain

9.4.1 DALI

DALI algorithm was developed by Holm and Sander in 1995, which uses a distance matrix for representation of each structure to be compared (Holm and Sander 1995). The structure is exemplified as a 2D array of distances between all C-alpha atoms. DALI has an underlying database called as FSSP which comprises the cataloging of 3D protein folds based on an all-against-all comparison of structures deposited in the PDB. The DALI server (http://ekhidna2.biocenter.helsinki.fi/dali/) accepts the coordinates of the query protein and outputs the similarity score, % of identical amino acids in alignment, and RMSD of Cα atoms in superimposition. It also gives Z-scores, which are the standard deviations from the average score from the database.

9.4.2 Combinatorial Extension (CE)

Shindyalov and Bourne have developed the combinatorial extension (CE) algorithm which use a distance approach for structure comparison of C-alpha distance matrices for every combination of eight residues in a polypeptide chain (Shindyalov and Bourne 1998). Then a combinatorial extension is made of an alignment path rather than using Monte Carlo optimization or dynamic programming. This path is generated based on aligned fragment pairs which are constructed on local geometry rather than the orientation of secondary structures and topology. It takes 3D coordinates as input, and the output lists its structural neighbors with Z-score, RMSD, % identity along with the length of the aligned sequence and number of gaps.

9.4.3 SSAP

Sequential structure alignment program (SSAP) was developed by Taylor and Orengo that uses double dynamic programming based on atom-to-atom vectors in structure space using the Cβ atom of each residue. It takes the rotameric state of each residue along with the location of the backbone (Taylor and Orengo 1989). A series of optimal local alignments are generated based on the matrix using dynamic programming which is then summed into a “summary” matrix to which dynamic programming is again applied to determine the final structural alignment. The output gives raw SSAP scores derived from the comparison and is standardized against known comparisons in the CATH (Classification, Architecture, Topology, and Homology) database (Sillitoe et al. 2015). A significant alignment has a raw SSAP score above 70–80% when 60% residues of larger protein are included in the alignment.

9.5 Protein Structure Modeling and Its Applications and Case Studies

Decades of intense research in protein biochemistry and folding process along with the parallel development of high-end computational resources have made the protein prediction possible with high accuracy through the computational modeling. The protein structure determination via the experimental techniques including NMR and X-ray (Fig. 9.5) has became a routine practice. A major application of protein structure determination is in the pharmaceutical segment for drug designing and discovery as well as in molecular biology research and biotechnology (Hillisch et al. 2004; Kopp and Schwede 2004). Homology modeling has also been used in functional annotations of newly sequenced genes (Hermann et al. 2007). Commercially, the homology model has also shown to have a successful application in characterizing the substrate specificity using docking studies for important enzymes with the industrial application and reengineering these enzymes to accept other substrates (Blikstad et al. 2014). The increasing accuracy of protein prediction and availability of high-resolution X-ray structure as templates have driven the homology modeling applications in drug designing as well as in designing of site-directed mutational studies. Many successful examples of homology modeling applications have been recently published and cannot be mentioned here in depth, for example, in antigen-antibody designing (Kuroda et al. 2012), modeling and simulations of ion channels (Maffeo et al. 2012), cystic fibrosis transmembrane conductance regulator (Dalton et al. 2012), inhibitor designing of DNA methyltransferases (Medina-Franco and Caulfield 2011), lead designing in epigenetic targets (Heinke et al. 2011; Andreoli and Del Rio 2015), and many more.

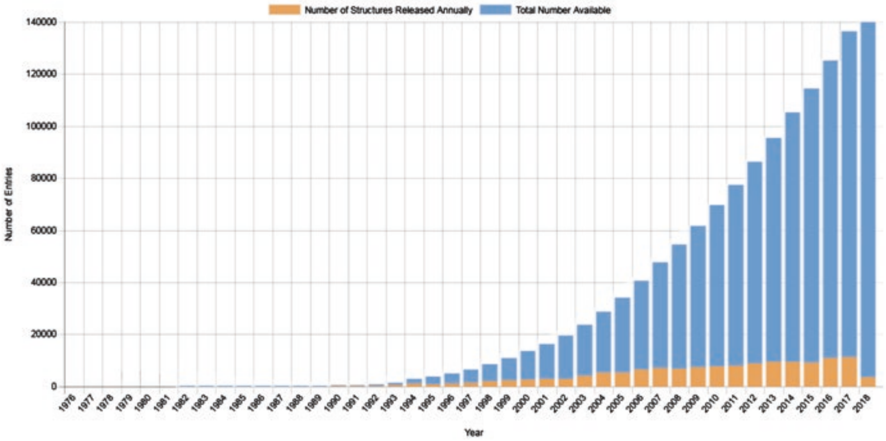


Fig. 9.5 PDB Statistics: Growth of released structures per year. The number of structures released per year is shown in the orange box and total number available is shown with a blue box. (Source: https://www.rcsb.org/stats/)

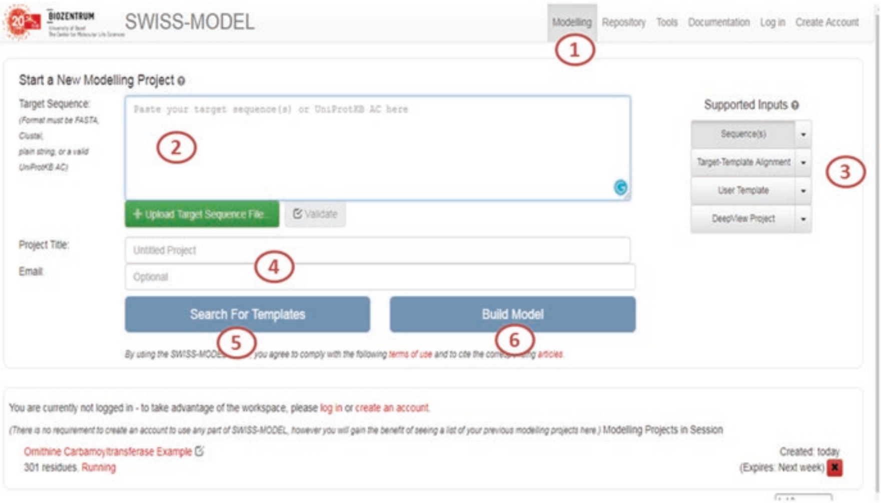
9.6 Tutorials

9.6.1 Homology Modeling of Using SWISS-MODEL Workspace

In this section, we will model the structure of a protein's amino acid sequence using one of its homologs as a template. We will then compare the generated homology model to the actual experimental structure. For this example, we will take insulin protein from Homo sapiens (UniProt ID: P01308).

Steps

1. Go to https://swissmodel.expasy.org/ and click the "Start Modelling" or select "myWorkspace" from the navigation bar (Modelling → myWorkspace) to start a new modeling project ((1) in Fig. 9.6).
2. One can either provide the *UniProtID* (P01308) of the target sequence or paste the protein sequence in FASTA format or in the input form ((2) in Fig. 9.6). Alternatively, different input formats can be selected from the panel using the drop-down menu ((3) in Fig. 9.6). Instead of the protein sequence, you can also input the target-template alignment directly generated from any sequence homology program like BLAST. For more advanced options, the user can directly input the Usertemplate. A project title will be automatically assigned by default, and the user can give the email address where they want the results to be mailed ((4) in Fig. 9.6).
3. The next step is to look for existing template structures; click on the "Search for Templates" button ((5) in Fig. 9.6). It will search for available homologs tem-



By using the SWISS-MODEL server, you agree to comply with the following terms of use and to cite the corresponding articles.

You are currently not logged in - to take advantage of the workspace, please log in or create an account.

(There is no requirement to create an account to use any part of SWISS-MODEL, however you will gain the benefit of seeing a list of your previous modeling projects here.) Modelling Projects in Session

Ornithine Carbamoyltransferase Example
301 residues. Running

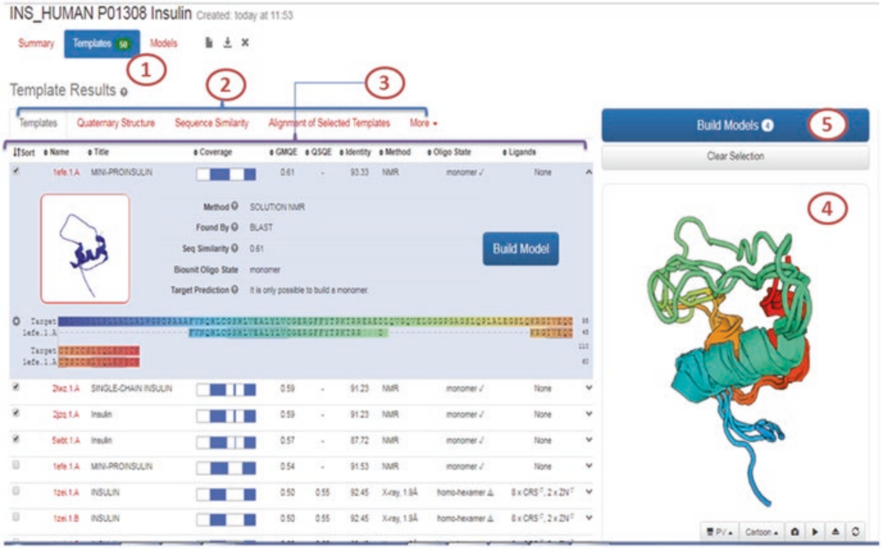
Created: today
(Expires: Next week)

Fig. 9.6 The sequence submission module of SWISS-MODEL for homology modeling

plate structure on which the query model can be built. The user can see the status of the job once the search is started.

Note: If you use “Build Model” option ((6) in Fig. 9.6), an automatic pipeline will be run for both the template search and the template selection steps. The automated mode selects templates that maximize the expected quality of the model but doesn’t guarantee the selection of the best template for modeling. The template selection entirely depends on the intended application of a model, e.g., if the goal is to construct a model of a protein in complex with a ligand/ substrate rather than selecting its apo form as a template, a template with similar ligand should be preferred.

4. After the template search is completed, the output page contains a table showing the list of identified templates (50 templates in the current example ((1) in Fig. 9.7) which are ranked according to the quality of resulting models. The panel ((2) in Fig. 9.7) mentions different tabs: templates, quaternary structures (if one is interested in modeling oligomeric state), sequence similarity (homology between the query and selected template), and alignment of selected templates with the query sequence.
5. For each template, the following information is provided ((3) in Fig. 9.7):
   - A checkbox to select and visualize the template in the 3D panel ((4) in Fig. 9.7).
   - The SMTL ID of the template, the protein name of the template.
   - The coverage of the query sequence (blue shade refers to higher sequence identity).



Sort	Name	Title	Coverage	GMQE	QSQE	Identity	Method	Oligo State	Ligands
☑	1efe.1.A	MINI-PROINSULIN		0.61	-	93.33	NMR	monomer ✓	None
☑	2kqz.1.A	SINGLE-CHAIN INSULIN		0.59	-	91.23	NMR	monomer ✓	None
☑	2jzq.1.A	Insulin		0.59	-	91.23	NMR	monomer ✓	None
☑	5ebt.1.A	Insulin		0.57	-	87.72	NMR	monomer ✓	None
□	1efe.1.A	MINI-PROINSULIN		0.54	-	91.53	NMR	monomer ✓	None
□	1zei.1.A	INSULIN		0.50	0.55	92.45	X-ray, 1.9Å	homo-hexamer ⚠	8 x CRS, 2 x ZN
□	1zei.1.B	INSULIN		0.50	0.55	92.45	X-ray, 1.9Å	homo-hexamer ⚠	8 x CRS, 2 x ZN

Method SOLUTION NMR
Found By BLAST
Seq Similarity 0.61
Biounit Oligo State monomer
Target Prediction It is only possible to build a monomer.

Fig. 9.7 Template identification results for input query sequence in SWISS-MODEL

- The GMQE (global model quality estimation) and QSQE (quaternary structure quality estimation).
- The target-template sequence identity.
- The ligands present in the experimental structure (if any).

6. The next step is to select suitable templates for modeling based on different parameters as mentioned above and the intention and applications of the model. By clicking on the checkbox, various template structures can be visualized and compared to structure superposition in the 3D viewer. One should rank them according to their coverage of the target sequence, resolution of experimentally determined structures, etc. in our example, we can observe that most of the templates share a high sequence identity (>90%). Let's select the four top-ranking templates for further modeling. After selection, one can see the superimposed structures of these templates ((4) in Fig. 9.7).
7. After the template(s) selection, click the "Build Model" button ((5) in Fig. 9.7) to run the final modeling task.
8. In the results page, for, respectively, model generated based on the selected templates (five models in the current example ((1) in Fig. 9.8)), the following information as shown in panel ((2) in Fig. 9.8) is provided:

- A file containing the model coordinates (*x*,*y*,*z*).
- The oligomeric state of the model and the modeled ligands (if any).
- QMEAN model quality estimation results.
- The target-template sequence alignment, the sequence identity to the target, and the target sequence coverage.

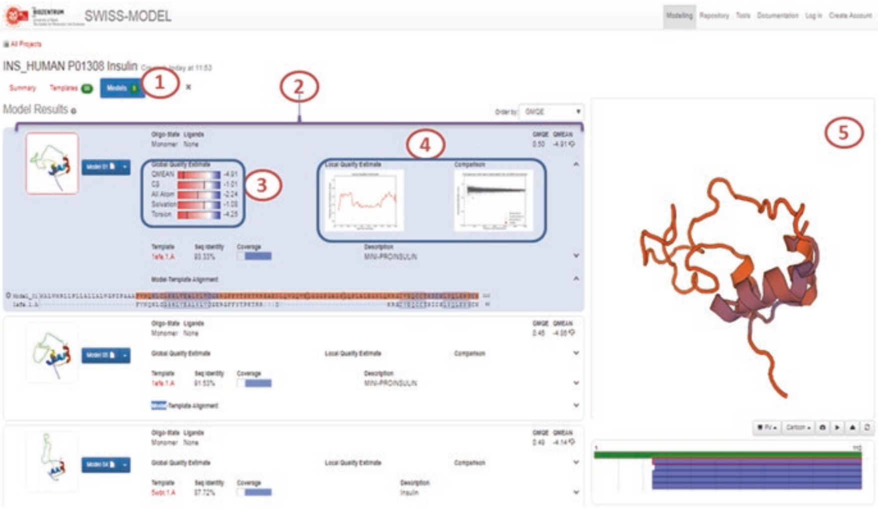


Fig. 9.8 Final output of modeling results and evaluation reports in SWISS-MODEL

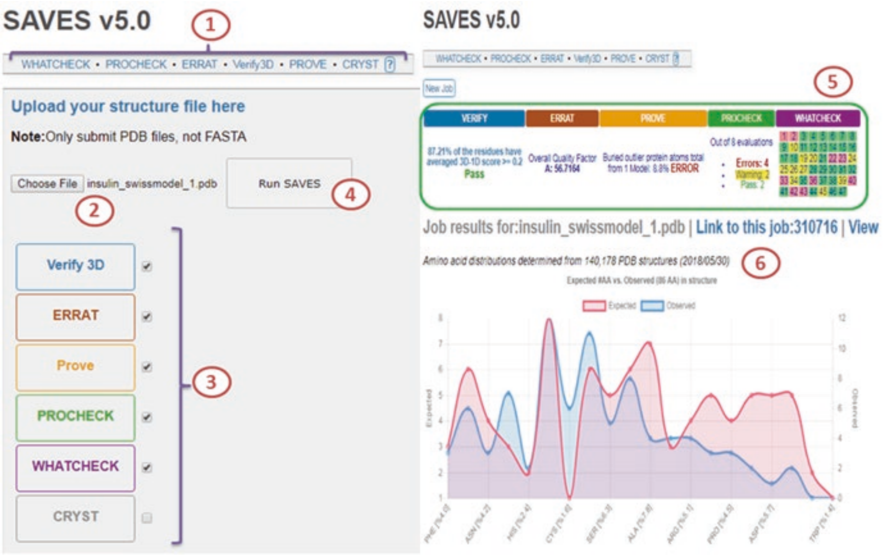
9. The different models and their quality can be evaluated by different local estimates and QMEAN per-residue plot as shown in panel ((3, 4) in Fig. 9.8). The detail evaluation of model is mention above in Sect. 9.3.3 under Swiss model validation server.
10. The models are displayed interactively using the 3D viewer ((5) in Fig. 9.8). Based on the above knowledge and results you obtained, one can assess whether the homology model can contemplate a reliable model or not. If not, choose a different strategy other than mentioned above and compare the different results.

9.6.2 Model Quality Estimation Using SAVES Metaserver and Refinement of the Model

In this section, we will use the SAVES Metaserver to estimate the model quality and refine the structure using MODREFINER until we get the better model which satisfies the parameters. We will use the above-generated model for this task. Download the model 1 from the above SWISS-MODEL output and save as "pdb" format.

Steps:

1. Go to SAVES version 5 homepage (https://servicesn.mbi.ucla.edu/SAVES/). SAVES is a Metaserver which combines different programs/servers available for reliable prediction of quality of generated homology model ((1) in Fig. 9.9).
2. Upload the model pdb file generated through Swiss-Model prediction ((2) in Fig. 9.9). Check the relevant server to be used for quality assessment by clicking the checkbox ((3) in Fig. 9.9). In this example, we will select Verify_3D, ERRAT, PROVE, PROCHECK, and WHATCHECK. Submit the run by clicking on "Run SAVES" button ((4) in Fig. 9.9).
3. After the job is finished, all the results will be displayed in the tab-wise sections, and detailed results from each server can be viewed by clicking on them ((5) in Fig. 9.9). A graph displaying the expected (red color) and observed (blue color) amino acids frequency in the model computed from the total amino acids distribution through entire structures deposited in PDB ((6) in Fig. 9.9).
4. One can analyze the output for each server by clicking them. The detailed output of every program can be found in each of them (Fig. 9.10). In our model evaluation, the Verify_3D shows as PASS ((1) in Fig. 9.10), but the ERRAT quality factor is not reliable ((2) in Fig. 9.10), PROVE output is showing error ((3) in Fig. 9.10), and PROCHECK ((4) in Fig. 9.10) shows unreliable Ramachandran statistics with 72% in allowed and 4.3% in disallowed region ((5) in Fig. 9.10) along with few errors in WHATCHECK output ((6) in Fig. 9.10). One can find the reliable values to be achieved for successful modeling in details for each of the server by individually assessing the complete results.
5. The next step is to either change the template along with different strategies or do the modeling again or refine the existing model to improve the overall score and quality factor of modeling.



Upload your structure file here

Note: Only submit PDB files, not FASTA

insulin_swissmodel_1.pdb

87.21% of the residues have averaged 3D-1D score >= 0.2 Pass

Overall Quality Factor A: 56.7164

Buried outlier protein atoms total from 1 Model: 8.8% ERROR

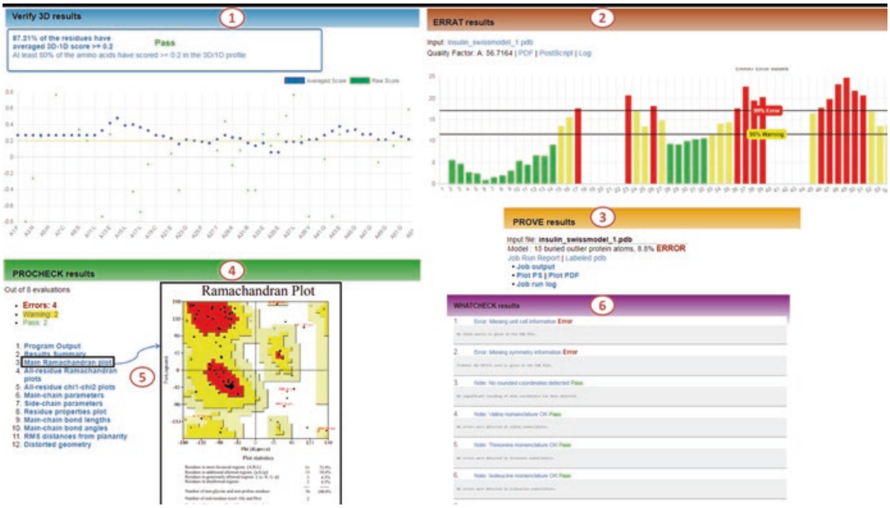
Out of 8 evaluations
- Errors: 4
- Warning: 2
- Pass: 2

Job results for:insulin_swissmodel_1.pdb | Link to this job:310716 | View

Amino acid distributions determined from 140,178 PDB structures (2018/05/30)

Expected #AA vs. Observed (86 AA) in structure

Fig. 9.9 Model evaluation results of SAVES Metaserver



Verify 3D results

87.21% of the residues have averaged 3D-1D score >= 0.2 Pass
At least 80% of the amino acids have scored >= 0.2 in the 3D/1D profile

ERRAT results

Input: insulin_swissmodel_1.pdb
Quality Factor: A: 56.7164 | PDF | PostScript | Log

PROVE results

Input file: insulin_swissmodel_1.pdb
Model : 15 buried outlier protein atoms, 8.8% ERROR
Job Run Report | Labeled pdb
- Job output
- Plot PS | Plot PDF
- Job run log

PROCHECK results

Out of 8 evaluations
- Errors: 4
- Warning: 2
- Pass: 2

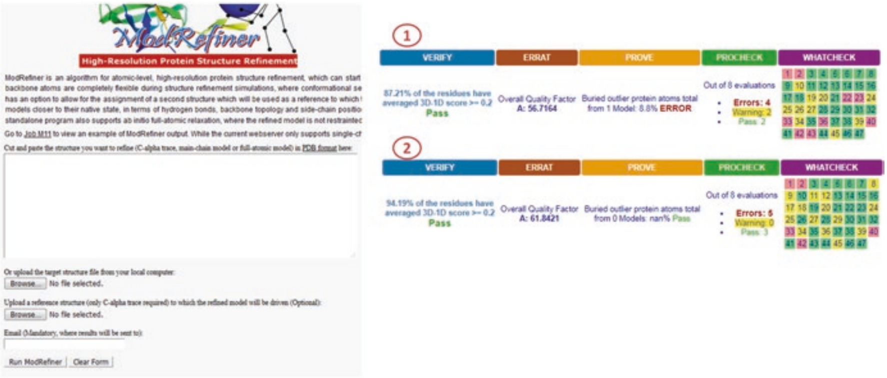
1. Program Output
2. Results Summary
3. Main Ramachandran plot
4. All-residue Ramachandran plots
5. All-residue chi1-chi2 plots
6. Main-chain parameters
7. Side-chain parameters
8. Residue properties plot
9. Main-chain bond lengths
10. Main-chain bond angles
11. RMS distances from planarity
12. Distorted geometry

WHATCHECK results

1. Error: Missing unit cell information Error
2. Error: Missing symmetry information Error
3. Note: No rounded coordinates detected Pass
4. Note: Valine nomenclature OK Pass
5. Note: Threonine nomenclature OK Pass
6. Note: Isoleucine nomenclature OK Pass

Fig. 9.10 Detail evaluation output of different programs in SAVES

6. In the current example, we will use ModRefiner server for refining the homology model to achieve the global minima near-native-like conformation. Open ModRefiner server (https://zhanglab.ccmb.med.umich.edu/ModRefiner/). Upload the homology model ((1) in Fig. 9.11) or paste the content of pdb file in the text box. One can also upload the C-alpha backbone as a reference structure for guided refinement which can be useful if one has a near identical crystal



ModRefiner is an algorithm for atomic-level, high-resolution protein structure refinement, which can start
backbone atoms are completely flexible during structure refinement simulations, where conformational se
has an option to allow for the assignment of a second structure which will be used as a reference to which t
models closer to their native state, in terms of hydrogen bonds, backbone topology and side-chain positio
standalone program also supports ab initio full-atomic relaxation, where the refined model is not restrainted

Go to Job M11 to view an example of ModRefiner output. While the current webserver only supports single-ch

Cut and paste the structure you want to refine (C-alpha trace, main-chain model or full-atomic model) in PDB format here:

Or upload the target structure file from your local computer:
Browse... No file selected.

Upload a reference structure (only C-alpha trace required) to which the refined model will be driven (Optional):
Browse... No file selected.

Email (Mandatory, where results will be sent to):

VERIFY	ERRAT	PROVE	PROCHECK	WHATCHECK
87.21% of the residues have averaged 3D-1D score >= 0.2 Pass	Overall Quality Factor A: 56.7164	Buried outlier protein atoms total from 1 Model: 8.8% ERROR	Out of 8 evaluations • Errors: 4 • Warning: 2 • Pass: 2	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47

VERIFY	ERRAT	PROVE	PROCHECK	WHATCHECK
94.19% of the residues have averaged 3D-1D score >= 0.2 Pass	Overall Quality Factor A: 61.8421	Buried outlier protein atoms total from 0 Models: nan% Pass	Out of 8 evaluations • Errors: 5 • Warning: 0 • Pass: 3	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47

Fig. 9.11 Model refinement server ModRefiner submission page. *1* SAVES output for pre-refined model and *2* SAVES output for refined model

structure available in PDB ((2) in Fig. 9.11). Click the "run ModRefiner" button to submit the job. After the completion of the job, a refined model will be generated with the RMSD and TM-score in reference to the initial pre-refined model.

7. Once the refinement is complete, repeat the steps from 2 to 4 and assess whether the overall quality of the model has improved or not. In our current case, we can clearly see the differences in SAVES output for pre-refined ((3) in Fig. 9.11) and post-refined ((4) in Fig. 9.11) models. The ERRAT quality factor has been improved from 56 to 61, PROVE outputs the model as Pass, and Ramachandran statistics have been improved from 72% in allowed and 4.3% in the disallowed region to 84% in allowed and 1.4% in the disallowed region. This task can be repeated for several cycles until the overall quality of models stops improving without compromising the overall conformation of the model.
8. Alternatively, the model can also be subjected to molecular dynamics simulations for certain timescale until the RMSD values achieve stability. This method also takes solvent and its environment in consideration while refining the model. GROMACS or NAMD package can be used for such refinement, but it requires an advanced stage of computer understanding and is beyond the scope of current section.

9.7 Conclusions

For several years, the protein structure prediction remained to be a challenging goal for structural biologists. Recent technological advances in high-throughput genome sequencing and experimental techniques used in structure determination along with the powerful computational resources have changed this perspective. Nowadays, protein prediction has become a routine bioinformatic practice and applied in various subsequent biological research experiments. Improved algorithms and a better understanding of physiochemical properties for protein stability and folding have increased the accuracy of predictions and enabled the researchers to generate a rational hypothesis for further experiments. Homology modeling has become a significant tool used for structure prediction, while progress in fold recognition and threading for detection of distant homologs for sequence alignments have made structure prediction easier than as thought earlier. The progress in ab initio structure prediction method is relatively slow compared to other methods, but a remarkable achievement made in recent years has enabled us to predict structure for small proteins accurately. Profile-based sequence searches, use of 3D structures in alignment, improved loop and side-chain conformation prediction followed by structure validation, and quality assessment have improved overall accuracy in the process.

Despite several years of understanding and development, researchers still do not have complete knowledge of how a protein folds based on its primary sequence. We still lack the knowledge of sequence/structure/function relationships. For sequence alignment in or below "twilight zone," finding distant homolog still needs improvement because completely unrelated sequences with no detectable homology still fold into similar conformations. Though we are nearer in the prediction of native-like conformation with lower-energy minima, still for certain novel sequences, experimental determination of structure remains the only solution. The ideal way is to predict the structure using different methods, refine until one gets native-like low-energy conformations, and validate those using different approaches. It is highly advisable that while predicting a structure, one should give more emphasis in finding optimal template recognition and hybrid method involving homology modeling and threading followed by simulations to have an accurate prediction which can be further utilized in subsequent experiments. Apart from the remarkable development of algorithms and prediction methods by various groups, more targets and refinement tools along with benchmarking competitions like CASP should be encouraged. The overall community helps to define where efforts should be made to move the field forward progressively. Further advancement in both computational biology and physical sciences will further help in the understanding of biological processes and benefit of human mankind.

References

Altschul SF, Gish W, Miller W et al (1990) Basic local alignment search tool. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Altschul SF, Madden TL, Schäffer AA et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Andreoli F, Del Rio A (2015) Computer-aided molecular Design of Compounds Targeting Histone Modifying Enzymes. Comput Struct Biotechnol J 13:358–365. https://doi.org/10.1016/j.csbj.2015.04.007

Anfinsen CB, Haber E, Sela M, White FH (1961) The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. Proc Natl Acad Sci U S A 47:1309–1314

Benkert P, Tosatto SCE, Schomburg D (2008) QMEAN: a comprehensive scoring function for model quality assessment. Proteins 71:261–277. https://doi.org/10.1002/prot.21715

Biasini M, Bienert S, Waterhouse A et al (2014) SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res 42:W252–W258. https://doi.org/10.1093/nar/gku340

Blikstad C, Dahlström KM, Salminen TA, Widersten M (2014) Substrate scope and selectivity in offspring to an enzyme subjected to directed evolution. FEBS J 281:2387–2398. https://doi.org/10.1111/febs.12791

Bowie JU, Lüthy R, Eisenberg D (1991) A method to identify protein sequences that fold into a known three-dimensional structure. Science 253:164–170

Boyle J (2005) In: Nelson D, Cox M (eds). Biochemistry and Molecular Biology EducationLehninger principles of biochemistry, vol 33, 4th edn, pp 74–75. https://doi.org/10.1002/bmb.2005.494033010419

Chan AWE, Hutchinson EG, Harris D, Thornton JM (1993) Identification, classification, and analysis of beta-bulges in proteins. Protein Sci 2:1574–1590. https://doi.org/10.1002/pro.5560021004

Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. EMBO J 5:823–826

Dalton J, Kalid O, Schushan M et al (2012) New model of cystic fibrosis transmembrane conductance regulator proposes active channel-like conformation. J Chem Inf Model 52:1842–1853. https://doi.org/10.1021/ci2005884

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797. https://doi.org/10.1093/nar/gkh340

Gu J, Bourne PE (2009) Structural bioinformatics. Wiley-Blackwell

Heinke R, Carlino L, Kannan S et al (2011) Computer- and structure-based lead design for epigenetic targets. Bioorg Med Chem 19:3605–3615. https://doi.org/10.1016/j.bmc.2011.01.029

Hermann JC, Marti-Arbona R, Fedorov AA et al (2007) Structure-based activity prediction for an enzyme of unknown function. Nature 448:775–779. https://doi.org/10.1038/nature05981

Hillisch A, Pineda LF, Hilgenfeld R (2004) Utility of homology models in the drug discovery process. Drug Discov Today 9:659–669. https://doi.org/10.1016/S1359-6446(04)03196-4

Holm L, Sander C (1995) Dali: a network tool for protein structure comparison. Trends Biochem Sci 20:478–480

Huggins ML (1943) The structure of fibrous proteins. Chem Rev 32:195–218. https://doi.org/10.1021/cr60102a002

Jayachandran G, Vishal V, Pande VS (2006) Using massively parallel simulation and Markovian models to study protein folding: examining the dynamics of the villin headpiece. J Chem Phys 124:164902. https://doi.org/10.1063/1.2186317

Jones DT, Taylort WR, Thornton JM (1992) A new approach to protein fold recognition. Nature 358:86–89. https://doi.org/10.1038/358086a0

Jones DT, Tress M, Bryson K, Hadley C (1999) Successful recognition of protein folds using threading methods biased by sequence similarity and predicted secondary structure. Proteins Suppl 3:104–111

Kendrew JC, Bodo G, Dintzis HM et al (1958) A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. Nature 181:662–666. https://doi.org/10.1038/181662a0

Kim DE, Chivian D, Baker D (2004) Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res 32:W526–W531. https://doi.org/10.1093/nar/gkh468

Kmiecik S, Gront D, Kolinski M et al (2016) Coarse-grained protein models and their applications. Chem Rev 116:7898–7936. https://doi.org/10.1021/acs.chemrev.6b00163

Kopp J, Schwede T (2004) Automated protein structure homology modeling: a progress report. Pharmacogenomics 5:405–416. https://doi.org/10.1517/14622416.5.4.405

Kuroda D, Shirai H, Jacobson MP, Nakamura H (2012) Computer-aided antibody design. Protein Eng Des Sel 25:507–522. https://doi.org/10.1093/protein/gzs024

Low BW, Baybutt RB (1952) The π helix—a hydrogen bonded configuration of the polypeptide chain. J Am Chem Soc 74:5806–5807. https://doi.org/10.1021/ja01142a539

Maffeo C, Bhattacharya S, Yoo J et al (2012) Modeling and simulation of ion channels. Chem Rev 112:6250–6284. https://doi.org/10.1021/cr3002609

Medina-Franco JL, Caulfield T (2011) Advances in the computational development of DNA methyltransferase inhibitors. Drug Discov Today 16:418–425. https://doi.org/10.1016/j.drudis.2011.02.003

Michalsky E, Goede A, Preissner R (2003) Loops in proteins (LIP)--a comprehensive loop database for homology modelling. Protein Eng 16:979–985. https://doi.org/10.1093/protein/gzg119

Mirsky AE, Pauling L (1936) On the structure of native, denatured, and coagulated proteins. Proc Natl Acad Sci U S A 22:439–447

Moult J, Fidelis K, Kryshtafovych A et al (2018) Critical assessment of methods of protein structure prediction (CASP)-round XII. Proteins 86:7–15. https://doi.org/10.1002/prot.25415

Notredame C, Higgins DG, Heringa J (2000) T-coffee: a novel method for fast and accurate multiple sequence alignment 1 1Edited by J. Thornton. J Mol Biol 302:205–217. https://doi.org/10.1006/jmbi.2000.4042

Pauling L, Corey RB, Branson HR (1951) The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. Proc Natl Acad Sci U S A 37:205–211. https://doi.org/10.1073/PNAS.37.4.205

Perutz MF, Rossmann MG, Cullis AF et al (1960) Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-a. resolution, obtained by X-ray analysis. Nature 185:416–422

Ramachandran GN, Ramakrishnan C, Sasisekharan V (1963) Stereochemistry of polypeptide chain configurations. J Mol Biol 7:95–99

Richardson JS, Getzoff ED, Richardson DC (1978) The beta bulge: a common small unit of nonrepetitive protein structure. Proc Natl Acad Sci U S A 75:2574–2578

Rödel W (1974) J. S. Fruton: molecules and life – historical essays on the interplay of chemistry and biology. 579 Seiten, vol 18. Wiley-Interscience, New York/London/Sydney/Toronto 1972. Preis: 8,95 £. Food / Nahrung, pp 471–472. https://doi.org/10.1002/food.19740180423

Šali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234:779–815. https://doi.org/10.1006/jmbi.1993.1626

Shen M, Sali A (2006) Statistical potential for assessment and prediction of protein structures. Protein Sci 15:2507–2524. https://doi.org/10.1110/ps.062416606

Shindyalov IN, Bourne PE (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 11:739–747

Sillitoe I, Lewis TE, Cuff A et al (2015) CATH: comprehensive structural and functional annotations for genome sequences. Nucleic Acids Res 43:D376–D381. https://doi.org/10.1093/nar/gku947

Svedberg T, Fåhraeus R (1926) A new method for the determination of the molecular weight of the proteins. J Am Chem Soc 48:430–438. https://doi.org/10.1021/ja01413a019

Taylor WR (1986a) The classification of amino acid conservation. J Theor Biol 119:205–218. https://doi.org/10.1016/S0022-5193(86)80075-3

Taylor WR (1986b) Identification of protein sequence homology by consensus template alignment. J Mol Biol 188:233–258

Taylor HS Large molecules through atomic spectacles. Proc Am Philos Soc 85:1–12

Taylor WR, Orengo CA (1989) Protein structure alignment. J Mol Biol 208:1–22

Xiang Z, Soto CS, Honig B (2002) Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction. Proc Natl Acad Sci 99:7432–7437. https://doi.org/10.1073/pnas.102179699

Xu D, Zhang Y (2011) Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. Biophys J 101:2525–2534. https://doi.org/10.1016/j.bpj.2011.10.024

Xu D, Zhang Y (2013) Toward optimal fragment generations for ab initio protein structure assembly. Proteins 81:229–239. https://doi.org/10.1002/prot.24179

Zhang J, Liang Y, Zhang Y (2011) Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling. Structure 19:1784–1795. https://doi.org/10.1016/j.str.2011.09.022

Chapter 10
Protein Structure Annotations

Mirko Torrisi and Gianluca Pollastri

Contents

10.1 Introduction to Protein Structure Annotations

Proteins hold a unique position in structural bioinformatics. In fact, more so than other biological macromolecules such as DNA or RNA, their structure is directly and profoundly linked to their function. Their cavities, protuberances and their overall

M. Torrisi (✉) · G. Pollastri
School of Computer Science, University College Dublin, Dublin, Ireland
e-mail: gianluca.pollastri@ucd.ie

N. A. Shaik et al. (eds.), *Essentials of Bioinformatics, Volume I*,
https://doi.org/10.1007/978-3-030-02634-9_10

shapes determine with what and how they will interact and, therefore, the roles assumed in the hosting organism. Unfortunately, the complexity, wide variability and ultimately the sheer number of diverse structures present in nature make the characterisation of proteins extremely expensive and complex. For this reason, considerable effort has been spent on predicting protein structures by computational means, either directly or in the form of abstractions that simplify the prediction while still retaining structural information. These abstractions, or protein structure annotations, may be one-dimensional when they can be represented by a string or a sequence of numbers, typically of the same length as the protein's primary structure (the sequence of its amino acids). This is the case, for instance, of secondary structure (SS) or solvent accessibility (SA). Another important class of abstractions is composed of two-dimensional properties, that is, features of pairs of amino acids (AA) or SS, such as contact and distance maps, disulphide bonds or pairings of strands into β-sheets.

Machine learning (ML) techniques have been extensively used in bioinformatics and in structural bioinformatics in particular. The abundance of freely available data – such as the Protein Data Bank (PDB) (Berman et al. 2000) and the Universal Protein Resource (The UniProt Consortium 2016) – and their complexity make proteins an ideal domain where to apply the most recent and sophisticated ML techniques, such as deep learning (LeCun et al. 2015). Nonetheless, there are pitfalls to avoid and best practices to follow to correctly train and test any ML method on protein sequences (Walsh et al. 2016).

Deep learning is a collection of methods and techniques to efficiently train nuanced parametric models such as neural networks (NN) with multiple hidden layers (Schmidhuber 2015). These layers contain hierarchical representations of the features of interest extracted from the input. NN are the de facto standard ML method to predict protein structure annotations. They have a central role at the two most important academic assessments of protein structure predictors: CASP and CAMEO (Haas et al. n.d.). Thus, they are widely used to predict protein one-dimensional and two-dimensional structural abstractions.

A typical predictor of protein structure annotations will first look for evolutionary information (PSI-BLAST is commonly used for this task), then will encode the information found, following this will run a ML method (usually a NN) on the encoded information and finally will process the output into a human-readable format. Differently from ab initio methods, template-based predictors directly exploit structural information of resolved proteins alongside evolutionary information (Pollastri et al. 2007).

Position-Specific Iterated Basic Local Alignment Search Tool (PSI-BLAST) (Altschul et al. 1997) is the de facto standard algorithm, released with the BLAST+ suite, to address protein alignment. In particular, it is commonly used in substitution of BLAST, whenever remote homologues have relevance. PSI-BLAST executes a BLAST call to find similar proteins in a given database, and then it either uses the resulting multiple sequence alignment (MSA) to construct a position-specific score matrix (PSSM) or outputs the MSA itself. The entire process is usually iterated few times using the last PSSM as query for the next iteration – in order to improve the PSSM and, thus, maximise the sensitivity of the method. The trade-off for increasing the number of iterations, and the sensitivity of the method, is a higher likelihood of corrupting the PSSM, including false-positive queries into it (Schäffer et al. 2001). For this reason, and the nature itself of the tool, it is fundamental to consider PSI-BLAST as a predicting tool and not as an exact algorithm (Jones and Swindells 2002).

HHblits (Remmert et al. 2012) is a 2011 algorithm to address protein alignment. It focuses on fast iterations and high precision and recall. It obtains these gains by adopting Hidden Markov Models (HMM) to represent both query and database sequences. The overall approach resembles the PSI-BLAST one – except that HMM rather than PSSM are the central entity. In fact, the heuristic algorithm looks for similar proteins in the HMM database at first. Then, it either uses the resulting HMM to improve the HMM query, and iterate, or outputs the MSA found with the last HMM. The same trade-off between number of iterations and likelihood of corrupting the HMM stands for HHblits as it does for PSI-BLAST.

In this chapter we review the main abstractions of protein structures, namely, SS, SA, torsional angles (TA) and distance/contact maps. For each of them, we describe an array of ML algorithms that have been used for their characterisation, point to a set of public tools available to the research community, including some that have been developed in our laboratory, and try to outline the state of the art in their prediction. These structure annotations are complementary with one another as they look at proteins from different views. That said, some annotations received far more interest from the bioinformatics community than others, for reasons such as simplicity or the intrinsic nature of the feature itself. We focus more on these well-assessed annotations, keeping in mind that the main function of protein structure annotations is to facilitate the understanding of the very core of any protein: the three-dimensional structure.

The PSSM built by PSI-BLAST, or the HMM built by HHblits, or the encoded MSA built by either PSI-BLAST or HHblits are generally used as inputs to a protein feature predictor. Different releases of the database used to find evolutionary information may lead to different outcomes. Normally, a computer able to look for evolutionary information (thus, execute PSI-BLAST or HHblits calls successfully) has the right hardware to run the standalones here presented with no problem.

All the predictors described below offer a web server, are free for academic use and provide licenses for commercial users at the time of writing. The web servers described return a result (prediction) in anything between a few minutes and a few hours.

10.2 Secondary Structure

SS prediction is one of the great historical challenges in bioinformatics (Rost 2001; Yang et al. 2016). Its history started in 1951, when Pauling and Corey predicted for the first time the existence of what were later discovered to be the two most common SS conformations: α-helix and β-sheet (Pauling and Corey 1951). Notably, the very first high-resolution protein structure was determined only in 1958 (and led to a Nobel Prize to Kendrew and Perutz) (Kendrew et al. 1960; Perutz et al. 1960). These early successes motivated the first generation of protein predictors, which were able to extrapolate statistical propensities of single AA (or residue) towards certain conformations (Chou and Fasman 1974). The slow but steady growth of available data and more insights on protein structure led to the second generation of predictors, which expanded the input to segments of adjacent residues (3–51 AA) to gather more useful information, and assessed many available theoretical algorithms on SS (Rost 2001). In the 1990s, more available computational power and data allowed the development and implementation of more advanced algorithms, able to look for and take advantage of evolutionary information (Yang et al. 2016). Thus, the third generation of SS predictors was the first able to predict at better than 70% accuracy (Rost and Sander 1993), efficiently exploit PSI-BLAST (Jones 1999) and implement deep NN (Baldi et al. 1999). In 2002, SS was removed from CASP since the few and relatively short targets assessed at the venue were not considered statistically sufficient to evaluate the mature methods available (Aloy et al. 2003).

The intrinsic nature of SS, being an intermediate structural representation between primary and tertiary structure, makes it a strategic and fundamental one-dimensional protein feature. It is often adopted as intermediate step towards more complex and informative features (i.e. contact maps (Jones et al. 2015; Wang et al. 2017; Vullo et al. 2006), the recognition of protein folds (Yang et al. 2011) and protein tertiary structure (Baú et al. 2006)). In other words, a high-quality SS prediction can greatly help to understand the nature of a protein and lead to a better prediction of its structure. For example, SS regularities characterise the proteins in a common fold (Murzin et al. 1995).

The theoretical limit of SS prediction is usually set at 88–90% accuracy per AA (Yang et al. 2016). This limit is mainly derived from the disagreement on how to assign SS and from the intrinsic dynamic nature of protein structure – i.e. the protein structure changes according to the fluid in which the protein is immersed. In particular, define secondary structure of proteins (DSSP) (Kabsch and Sander 1983), the gold standard algorithm to assign SS given the atomic-resolution coordinates of the protein, agrees with the PDB descriptions around 90.8% of the time (Martin et al. 2005). While DSSP aims to provide an unambiguous and physically meaningful assignment, the PDB represents the ground truth in structural proteomics (Berman et al. 2000).

All the SS predictors described in this chapter exploit different architectures of NN to perform their predictions. The list of AA composing the protein of interest is the only input required. The SS is often classified in three states – i.e. helices, sheets and coils – although the DSSP identifies a total of eight different classes. Because of the higher difficulty of the task, compounded also by the rare occurrence of certain classes – i.e. π-helix and β-bridge – only three of the predictors presented here (Porter5, RaptorX-Property and SSpro) can predict in both three states and eight states. The DSSP classifications of SS in eight states are the following:

- G = three-turn helix (310-helix), minimum length three residues
- H = four-turn helix (α-helix), minimum length four residues
- I = five-turn helix (π-helix), minimum length five residues

- T = hydrogen bonded turn (three, four or five turn)
- E = extended strand (in β-sheet conformation), minimum length two residues
- B = residue in isolated β-bridge (single pair formation)
- S = bend (the only non-hydrogen-bond based assignment)
- C = coil (anything not in the above conformations)

When SS is classified in three states, the first three (G, H, I) are generally considered helices, E and B are classified as strands and anything else as coils. SS prediction is evaluated looking at the rate of correctly classified residues (per class) – i.e. Q3 or Q8 for three- or eight-state prediction, respectively – or at the segment overlap score (SOV), i.e. the overlap between the predicted and the real segments of SS (Zemla et al. 1999), for a more biological viewpoint. The best performing ab initio SS predictors are able to predict three-state SS close to 85% Q3 accuracy and SOV score.

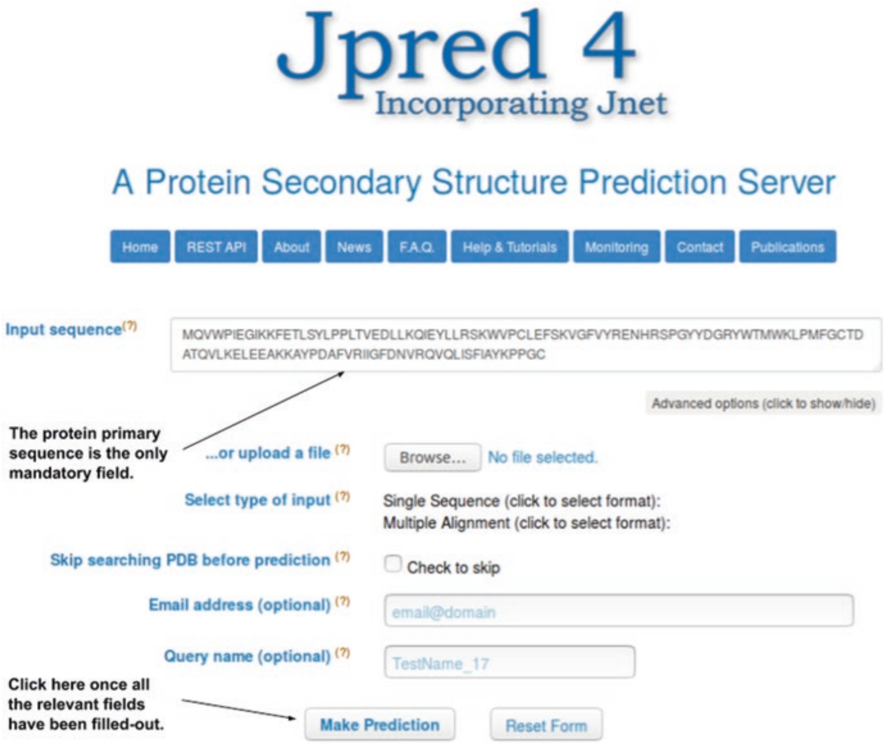
Table 10.1 gathers name, web server and notes on special features of every SS predictor presented in this chapter. A standalone – i.e. downloadable version that can run on a local machine – is currently available for all of them.

10.2.1 Jpred

Jpred is an SS predictor which was initially released in 1998 (Cuff et al. 1998). Jpred4 (Drozdetskiy et al. 2015), the last available version, has been released in 2015 to update HMMer (Finn et al. 2011) and the internal algorithm (a NN). Jpred4 relies on both PSI-BLAST and HMMer to gather evolutionary information, generating a PSSM and a HMM, respectively. It then predicts SS in three states, along with SA and coiled-coil regions. Jpred4 aims to be easily usable also from smartphones and tablets. FAQ and tutorials are available on its website (Fig. 10.1).

Table 10.1 Name, web server and notes on special features of every SS predictor presented in this chapter

Name	Web server	Notes
Jpred (Drozdetskiy et al. 2015)	http://www.compbio.dundee.ac.uk/jpred4/	HHMer, MSA as input, API
PSIPRED (Jones 1999)	http://bioinf.cs.ucl.ac.uk/psipred/	BLAST, cloud version, MSA as input
Porter (Pollastri and McLysaght 2005)	http://distilldeep.ucd.ie/porter/	three- or eight-states, HHblits or PSI-BLAST, light standalone
RaptorX-Property (Wang et al. 2016)	http://raptorx.uchicago.edu/StructurePropertyPred/predict/	three- or eight-states, no PSI-BLAST (only HHblits), option for no evolutionary information
SPIDER3 (Heffernan et al. 2017)	http://sparks-lab.org/server/SPIDER3/	Numpy or Tensorflow, HHblits and PSI-BLAST
SSpro (Magnan and Baldi 2014)	http://scratch.proteomics.ics.uci.edu/	three- or eight-states, BLAST, template-based


Jpred 4
Incorporating Jnet
A Protein Secondary Structure Prediction Server
Home
REST API
About
News
F.A.Q.
Help & Tutorials
Monitoring
Contact
Publications
Input sequence(?)
Advanced options (click to show/hide)
...or upload a file (?)
Browse...
No file selected.
Select type of input (?)
Single Sequence (click to select format):
Multiple Alignment (click to select format):
Skip searching PDB before prediction (?)
□ Check to skip
Email address (optional) (?)
email@domain
Query name (optional) (?)
TestName_17
Make Prediction
Reset Form

MQVWPIEGIKKFETLSYLPPLTVEDLLKQIEYLLRSKWVPCLEFSKVGFVYRENHRSPGYYDGRYWTMWKLPMFGCTD
ATQVLKELEEAKKAYPDAFVRIIGFDNVRQVQLISFIAYKPPGC

The protein primary sequence is the only mandatory field.

Click here once all the relevant fields have been filled-out.

Fig. 10.1 The homepage of Jpred4. The input sequence is the only requirement while more options are made available

The web server of Jpred4 is available at http://www.compbio.dundee.ac.uk/jpred4/. It requires a protein sequence in either FASTA or RAW format. Using the advanced options, it is also possible to submit multiple sequences (up to 200) or MSA as files. An email address and a JobID can be optionally provided. When a single sequence is given, Jpred4 looks for similar protein sequences in the PDB (Berman et al. 2000) and lists them when found. Checking a box, it is possible to skip this step and force an ab initio prediction. Jpred4 relies on a version of UniRef90 (The UniProt Consortium 2016) released in July 2014, while the PDB is regularly updated.

The result page is automatically shown and offers a graphical summary of the prediction along with links to possible views of the result in HTML (simple or full), PDF and Jalview (Waterhouse et al. 2009) (in-browser or not). It is also possible to get an archive of all the files generated or navigate through them in the browser. If an email address is submitted, a link to the result page and a summary containing the query, predicted SS and confidence per AA will be sent. The full result, made available as HTML or PDF, lists the ID of similar sequences used at prediction time, the final and intermediate predictions for SS, the prediction of coiled-coil regions, the prediction of SA with three different thresholds (0, 5 and 25% exposure) and the reliability of such predictions.

Jpred4 is not released as standalone, but it is possible to submit, monitor and retrieve a prediction using the command line software available at http://www.compbio.dundee.ac.uk/jpred4/api.shtml. A second package of scripts is made available at the same address to facilitate the submission, monitoring and retrieving of multiple protein sequences. More instructions and examples on how to use the command line software are presented on the same page.

10.2.2 PSIPRED

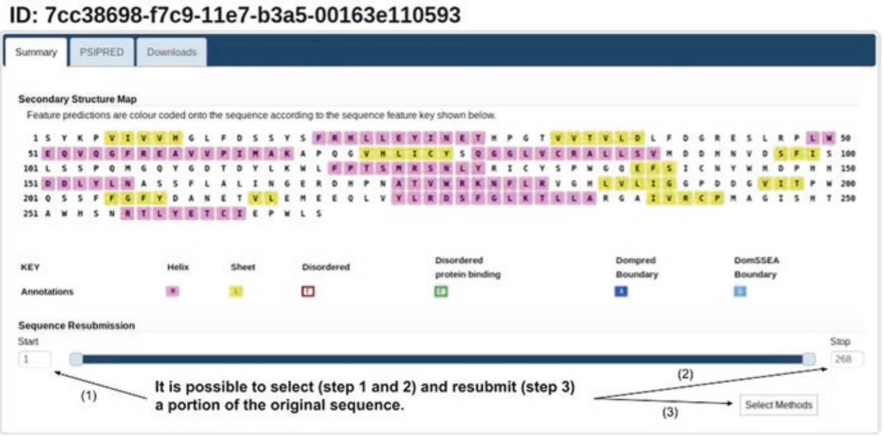
PSIPRED is a high-quality SS predictor freely available since 1999 (Jones 1999). Its last version (v4.01) has been released in 2016. PSIPRED exploits the PSSM of the protein to generate its prediction by neural networks. Like SSpro (described below), it recommends the implementation of the legacy BLAST package (abandoned in 2011) to collect evolutionary information. The BLAST+ package (the active development of BLAST) fixes multiple bugs and provides improvements and new features, but scales by 10 and rounds the PSSM, and thus provides less informative outputs for PSIPRED. BLAST+ is experimentally supported by PSIPRED (Fig. 10.2).

The web server of PSIPRED (Buchan et al. 2010), called the *PSIPRED Protein Sequence Analysis Workbench*, runs a 2012 release of PSIPRED (v3.3) and can be found at http://bioinf.cs.ucl.ac.uk/psipred/. A single sequence (or its MSA) and a short identifier are expected as input. Optionally, an email address can be inserted to receive a confirmation email (with link to the result) when the prediction is ready. Several prediction methods (for other protein features) can be chosen. The default choice (picking only PSIPRED) is sufficient to predict the SS. If the submission proceeds successfully, a courtesy page will be shown until the result is ready.

The result page, organised in tabs, shows the list of AA composing the analysed protein (the query sequence) and the predicted SS class (using different colours). From the same tab, it is possible to select the full query sequence, or a subsequence, to pass it to one of the predictor methods available on the PSIPRED Workbench. The predicted SS is presented in the tab called *PSIPRED* using a diagram. In the same diagram, the confidence of each prediction and the query sequence are included. The *Downloads* tab, the last one, allows the download of the information in the diagram as text or PDF or postscript or of all three versions.

The last release of PSIPRED is typically available as a standalone at http://bioinfadmin.cs.ucl.ac.uk/downloads/psipred/. Once the standalone has been downloaded and extracted, it is sufficient to follow the instructions in the README to perform predictions on any machine. The output will be generated in text format only as horizontal or vertical format. The latter will contain also the individual confidence per helix, strand and coil. Notably, the results obtained from the standalone may very well differ from those obtained from the PSIPRED Workbench. The latter does not implement the last PSIPRED release, at the time of writing.

In 2013, a preliminary package (v0.4) has been released to run PSIPRED on Apache Hadoop. Hadoop is an open-source software to facilitate distributed pro-



ID: 7cc38698-f7c9-11e7-b3a5-00163e110593

Secondary Structure Map

Feature predictions are colour coded onto the sequence according to the sequence feature key shown below.

```
  1 S Y K P V I V V M G L F D S S Y S F R M L L E Y I N E T M P G T V V T V L D L F D G R E S L R P L W  50
 51 E Q V Q G F R E A V V P I M A K A P Q G V M L I C Y S Q G G L V C R A L L S V M D D M N V D S F I S 100
101 L S S P Q M G Q Y G D T D Y L K W L F P T S M R S N L Y R I C Y S P W G Q E F S I C N Y W M D P M M 150
151 D D L Y L N A S S F L A L I N G E R D M P N A T V W R K N F L R V G M L V L I G G P D D G V I T P W 200
201 Q S S F F G F Y D A N E T V L E M E E Q L V Y L R D S F G L K T L L A R G A I V R C P M A G I S M T 250
251 A W M S N R T L Y E T C I E P W L S
```

KEY	Helix	Sheet	Disordered	Disordered protein binding	Dompred Boundary	DomSSEA Boundary
Annotations	H	E	D	B	A	S

Sequence Resubmission

It is possible to select (step 1 and 2) and resubmit (step 3) a portion of the original sequence.

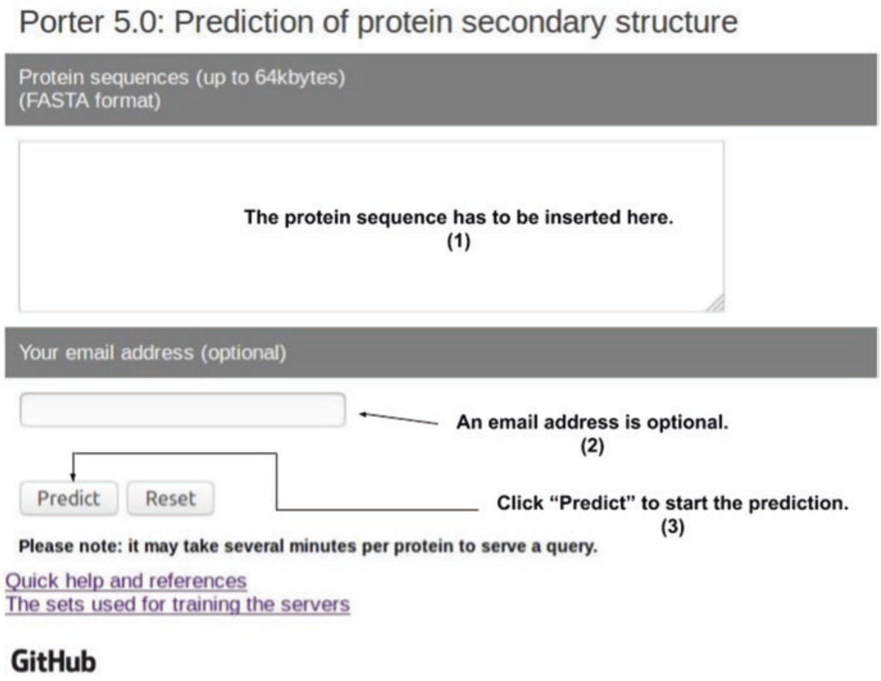
Fig. 10.2 A typical result page of PSIPRED web server. All the AA are listed and coloured according to the predicted SS class

cessing in computer clusters. Although this PSIPRED package is intended as an alpha build, instructions to install it on Hadoop and on AWS (the cloud service of Amazon) are provided. This package does not contain any standalone of PSIPRED. Thus, it is an interface to run the selected PSIPRED release on Hadoop. It can be downloaded at http://bioinfadmin.cs.ucl.ac.uk/downloads/hadoop/.

10.2.3 Porter

Porter is a high-quality SS predictor which has been developed starting in 2005 (Pollastri and McLysaght 2005) and improved since then (Pollastri et al. 2007; Mirabello and Pollastri 2013). Porter is built on carefully tuned and trained ensembles of cascaded bidirectional recurrent neural networks (Baldi et al. 1999). It is typically built on very large datasets, which are released as well. Its last release (v5) is available as web server and standalone (Torrisi et al. 2018). Differently from PSIPRED, it implements BLAST+ to gather evolutionary information. To maximise the gain obtained from evolutionary information, it also adopts HHblits alongside PSI-BLAST. Porter5 is one of the three SS predictors presented here that are able to predict both three-state and eight-state SS (Fig. 10.3).

The web server can be found at http://distilldeep.ucd.ie/porter/. The basic interface asks for protein sequences (in FASTA format) and for an optional email address. Up to 64 KB of protein sequences can be submitted at the same time, which approximately corresponds to 200 average proteins. Differently from other SS web servers, there is no limit of total submissions. The confirmation page will contain a



Porter 5.0: Prediction of protein secondary structure

Protein sequences (up to 64kbytes)
(FASTA format)

The protein sequence has to be inserted here.
(1)

Your email address (optional)

An email address is optional.
(2)

Click “Predict” to start the prediction.
(3)

Please note: it may take several minutes per protein to serve a query.

Quick help and references
The sets used for training the servers

Fig. 10.3 The input form of Porter5. Around 200 proteins can be submitted at once in FASTA format

summary of the job, the server load (how many jobs are to be processed) and the URL to the result page. It is automatically refreshed every minute.

The detailed result page will show the query, the SS prediction and the individual confidence. In other words, the same information shown by the PSIPRED Workbench is given in text format. The time to serve the job is shown as well. Optionally, if an email address has been inserted, all the information in the result page is sent by email. Thus, it can potentially be retrieved at any time. It is possible to predict SS and other protein structure annotations (one-dimensional or not) submitting one job at http://distillf.ucd.ie/distill/.

The very light standalone of Porter5 (7 MB) is available at http://distilldeep.ucd.ie/porter/. It is sufficient to extract the archive on any computer with python3, HHblits and PSI-BLAST to start predicting any SS. Using the parameter *–fast,* it is possible to avoid PSI-BLAST and perform faster but generally slightly less accurate predictions. When the prediction in three states and eight states completes successfully, it is saved in two different files. Each file shows the query, the predicted SS and the individual confidence per class. The datasets adopted for training and testing purposes are available at the same address.

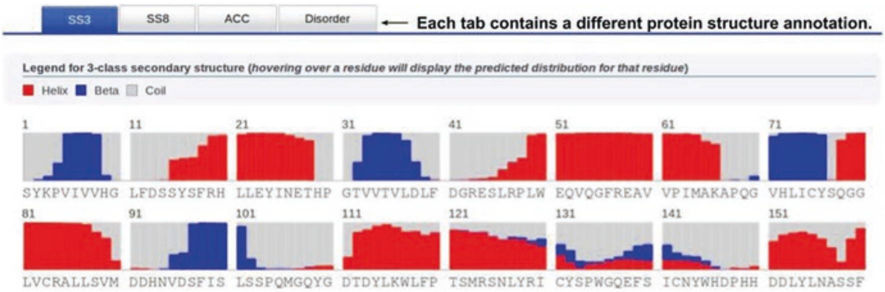
10.2.4 RaptorX-Property

RaptorX-Property, released in 2016, is a collection of methods to predict one-dimensional protein annotations (Wang et al. 2016). Namely, SS, SA and disorder regions are predicted from the same suite. The SS is predicted in both three states and eight states, as with Porter5 and SSpro. At the cost of lower accuracy, evolutionary information can be avoided to perform faster predictions. Its last release substitutes PSI-BLAST with HHblits to get faster protein profiles (Fig. 10.4).

The web server of RaptorX-Property is available at http://raptorx.uchicago.edu/StructurePropertyPred/predict/. Jobname and email address are recommended but not required. Query sequences can be uploaded directly from one's machine. Otherwise, up to 100 protein sequences (in FASTA format) can be passed at the same time through the input form. The system allows up to 500 pending (sequence) predictions at any time. The current server load, shown in the sidebar, tells the pending jobs to complete.

Once the job has been submitted, a courtesy page will provide the URL to the result page, how many pending jobs are ahead and the JobID. Less priority is given to intensive users. The jobs submitted in the previous 60 days are retrievable clicking on "My Jobs". Once the prediction is performed, the result page will show a summary of it using coloured text. At the bottom of the page, the same information is organised in tabs, one tab per feature predicted (SS in three- and eight-state, SA and disorder). The individual confidence is provided in the tabs. All this information is sent by email (in txt and rtf format), if an email address has been provided. Otherwise, it can be downloaded clicking the specific button.

The last standalone of RaptorX-Property (v1.01) can be downloaded at http://raptorx.uchicago.edu/download/. Once it has been extracted, it is sufficient to read and follow the instruction in README to predict SS, SA and disorder regions on one's own machine. As in the web server, it is possible to use or not sequence profiles and the results are saved in txt and rtf format. The disk space required is relatively considerable, 347 MB at the time of writing, almost 50 times the storage required by Porter5.


SS3
SS8
ACC
Disorder
1
11
21
31
41
51
61
71
SYKPVIVVHG LFDSSYSFRH LLEYINETHP GTVVTVLDLF DGRESLRPLW EQVQGFREAV VPIMAKAPQG VHLICYSQGG
81
91
101
111
121
131
141
151
LVCRALLSVM DDHNVDSFIS LSSPQMGQYG DTDYLKWLFP TSMRSNLYRI CYSPWGQEFS ICNYWHDPHH DDLYLNASSF

← Each tab contains a different protein structure annotation.

Legend for 3-class secondary structure (*hovering over a residue will display the predicted distribution for that residue*)

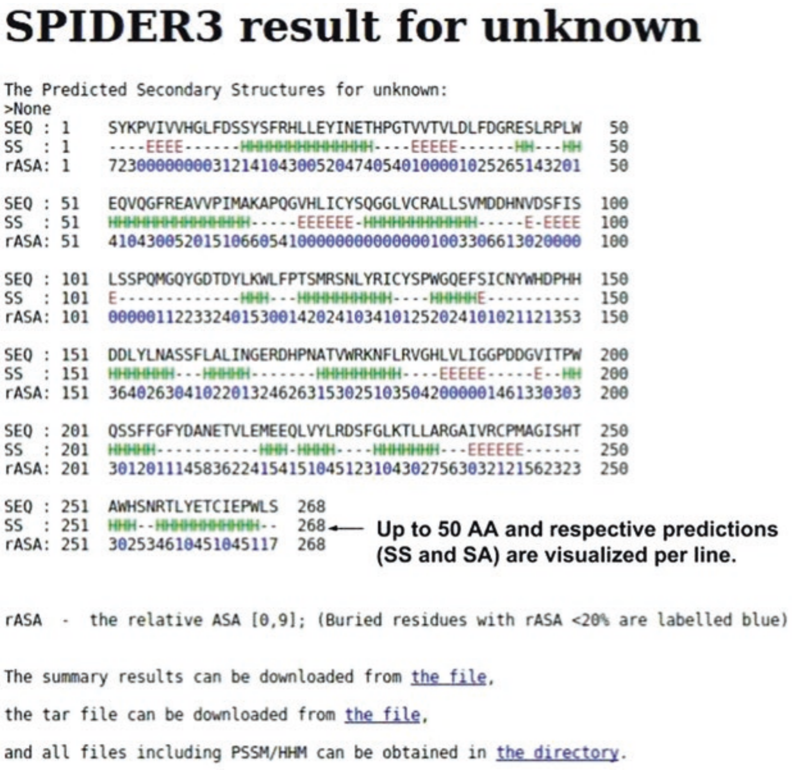
🟥 Helix 🟦 Beta ⬜ Coil

Fig. 10.4 A partial view of the result page of RaptorX-Property. Each bar in the charts represents the individual confidence

10.2.5 SPIDER3

SPIDER3 is the second version of a recent SS predictor first released in 2015 (Heffernan et al. 2015). Its last release is composed by 2 NN, the first of which predicts SS while the second predicts backbone angles, contact numbers and SA (Heffernan et al. 2017). It internally represents each AA using seven representative physiochemical properties (Fauchère et al. 1988). Like Porter5, it implements both HHblits and PSI-BLAST to look for more evolutionary information. SPIDER3 is also described in sections Solvent Accessibility and Torsion Angles, respectively (Fig. 10.5).

The web server of SPIDER3 is available at http://sparks-lab.org/server/SPIDER3/. An email address is required when multiple sequences are submitted or to receive a summary of the prediction. Otherwise, the query sequence is sufficient to submit the job and obtain an URL to the result page. The web server allows up to 100 protein sequences (in FASTA format) at a time and accepts optional JobID. To prevent duplicates, it is possible to visualise the queue of jobs submitted from one's IP address. The result page presents the query sequence and the predicted SS and



SPIDER3 result for unknown

```
The Predicted Secondary Structures for unknown:
>None
SEQ : 1     SYKPVIVVHGLFDSSYSFRHLLEYINETHPGTVVTVLDLFDGRESLRPLW   50
SS  : 1     ----EEEE------HHHHHHHHHHHHHH----EEEEE------HH---HH   50
rASA: 1     72300000000312141043005204740540100001025265143201   50

SEQ : 51    EQVQGFREAVVPIMAKAPQGVHLICYSQGGLVCRALLSVMDDHNVDSFIS  100
SS  : 51    HHHHHHHHHHHHHHH-----EEEEEE-HHHHHHHHHHHH-----E-EEEE  100
rASA: 51    41043005201510660541000000000000001003306613020000  100

SEQ : 101   LSSPQMGQYGDTDYLKWLFPTSMRSNLYRICYSPWGQEFSICNYWHDPHH  150
SS  : 101   E-------------HHH---HHHHHHHHHH----HHHHHE----------  150
rASA: 101   00000112233240153001420241034101252024101021121353  150

SEQ : 151   DDLYLNASSFLALINGERDHPNATVWRKNFLRVGHLVLIGGPDDGVITPW  200
SS  : 151   HHHHHHH---HHHHH-------HHHHHHHHH----EEEEE-----E--HH  200
rASA: 151   36402630410220132462631530251035042000001461330303  200

SEQ : 201   QSSFFGFYDANETVLEMEEQLVYLRDSFGLKTLLARGAIVRCPMAGISHT  250
SS  : 201   HHHHH-----------HHH-HHHH----HHHHHHH---EEEEEE------  250
rASA: 201   30120111458362241541510451231043027563032121562323  250

SEQ : 251   AWHSNRTLYETCIEPWLS  268
SS  : 251   HHH--HHHHHHHHHHH--  268
rASA: 251   302534610451045117  268

rASA  -  the relative ASA [0,9]; (Buried residues with rASA <20% are labelled blue)

The summary results can be downloaded from the file.

the tar file can be downloaded from the file.

and all files including PSSM/HHM can be obtained in the directory.
```

Up to 50 AA and respective predictions (SS and SA) are visualized per line.

Fig. 10.5 An output example of SPIDER3

SA, in a simple and colour-coded text format. In the same page, it is possible to download a summary (containing the same information) or an archive with the four features predicted and the individual confidence for SS. There is also a link to a temporary directory containing all the files created during the prediction, including the HMM and the PSSM.

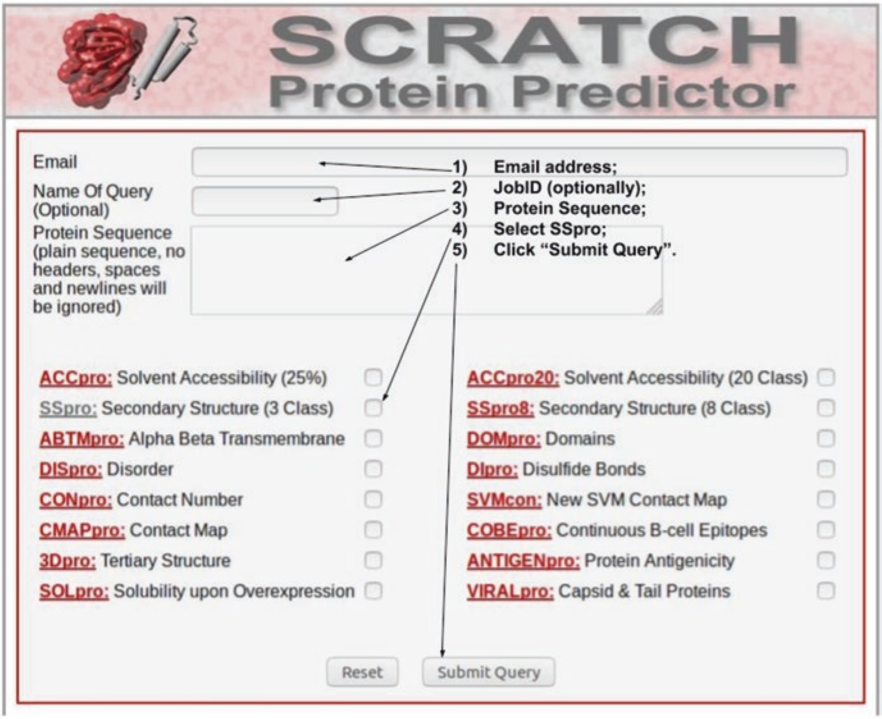
The standalone of SPIDER3, and the dataset used to train and test it, can be downloaded at http://sparks-lab.org/server/SPIDER3/. The main prerequisite is to install a python library of choice between Numpy and Tensorflow r0.11 (an older version). As for Porter5, it is then sufficient to install HHblits and PSI-BLAST to perform SS prediction on one's machine. The outcome of SS, SA, torque angles and contact number prediction will be saved in different columns of just one file. The storage required is 101 MB and 117 MB, respectively, without considering the library of choice.

10.2.6 SSpro

SSpro is a historical SS predictor developed starting in 1999 (Baldi et al. 1999; Magnan and Baldi 2014). Similarly to PSIPRED, it implements the BLAST package rather than the more recent BLAST+. The last version of SSpro (v5) has been released in 2014, together with ACCpro (see Solvent Accessibility, ACCpro), and performs template-based SS predictions (Magnan and Baldi 2014). More specifically, it exploits PSI-BLAST to look for homologues at both sequence and structure level (Pollastri et al. 2007). In other words, SSpro v5 has an additional final step in which it looks for similar proteins in the PDB (Fig. 10.6).

SSpro is available at http://scratch.proteomics.ics.uci.edu/ as part of the SCRATCH protein predictor (Cheng et al. 2005). SS is among the several (one-dimensional or not) protein features predictable on SCRATCH. Like Porter5 and RaptorX-Property, it is possible to predict both three-state (SSpro) and eight-state (SSpro8) predictions. Once SSpro or SSpro8 is selected, an email is required and optionally a JobID. Only one protein (of up to 1500 residues) can be submitted at a time. There are five total slots in the job queue per user. Once ready, the result of the prediction will be sent by email only. It will contain the JobID, the query sequence, the predicted SS (in three or eight classes) and a link to the explanation of the output format.

The standalone of the last SSpro (v5.2) and ACCpro (described in section Solvent Accessibility) compose the SCRATCH suite of 1D predictors available at http://download.igb.uci.edu/. SCRATCH v1.1 is released with all the prerequisites to set up and run SSpro. The BLAST package and the databases with both sequences and structural information are included. Thus, the amount of disk space needed to download and extract SCRATCH v1.1 is considerable (5.7 GB, 97 MB without databases).



Email

Name Of Query (Optional)

Protein Sequence (plain sequence, no headers, spaces and newlines will be ignored)

1) Email address;
2) JobID (optionally);
3) Protein Sequence;
4) Select SSpro;
5) Click "Submit Query".

- ACCpro: Solvent Accessibility (25%) ☐
- SSpro: Secondary Structure (3 Class) ☐
- ABTMpro: Alpha Beta Transmembrane ☐
- DISpro: Disorder ☐
- CONpro: Contact Number ☐
- CMAPpro: Contact Map ☐
- 3Dpro: Tertiary Structure ☐
- SOLpro: Solubility upon Overexpression ☐
- ACCpro20: Solvent Accessibility (20 Class) ☐
- SSpro8: Secondary Structure (8 Class) ☐
- DOMpro: Domains ☐
- DIpro: Disulfide Bonds ☐
- SVMcon: New SVM Contact Map ☐
- COBEpro: Continuous B-cell Epitopes ☐
- ANTIGENpro: Protein Antigenicity ☐
- VIRALpro: Capsid & Tail Proteins ☐

Fig. 10.6 A view of SCRATCH protein predictor

10.3 Solvent Accessibility

SA describes the degree of accessibility of a residue to the solvent surrounding the protein. SA is second only to SS among extensively studied and predicted one-dimensional protein structure annotations. The effort invested into SA predictors has been significant from the early 1990s and highly motivated from the successes obtained developing the third generation of SS predictors (Pascarella et al. 1998). In fact, similarly to SS prediction but sometimes with some time delay, mathematical and statistical methods (Cornette et al. 1987), NN (Rost and Sander 1994), evolutionary information (Holbrook et al. 1990) and deep NN (Pollastri et al. 2002) have been increasingly put to work to predict SA.

Although SA is less conserved than SS in homologous sequences (Rost and Sander 1994), it is typically adopted in parallel with SS in many pipelines towards more complex protein structure annotations such as CM – e.g. SA and SS are predicted for any CM predictor described in section Contact Maps (Jones et al. 2015; Wang et al. 2017; Adhikari et al. 2017; Walsh et al. 2009) – protein fold recognition (Yang et al. 2011) and protein tertiary structure (Mooney and Pollastri 2009). Notably, a strong (negative) correlation of −0.734 between SA and contact numbers

Table 10.2 Solvent Accessibility prediction servers

Name	Web server	Notes
ACCpro (Baldi et al. 1999)	http://scratch.proteomics.ics.uci.edu/	two-state or twenty-states, BLAST, template-based
PaleAle (Pollastri et al. 2007)	http://distilldeep.ucd.ie/paleale/	four-states, HHblits or PSI-BLAST, light standalone
RaptorX-Property (Heffernan et al. 2017)	http://raptorx.uchicago.edu/StructurePropertyPred/predict/	three-states, no PSI-BLAST (only HHblits), option for no evolutionary information
SPIDER3 (Heffernan et al. 2017)	http://sparks-lab.org/server/SPIDER3/	HSE and ASA in R, Numpy or Tensorflow, HHblits and PSI-BLAST

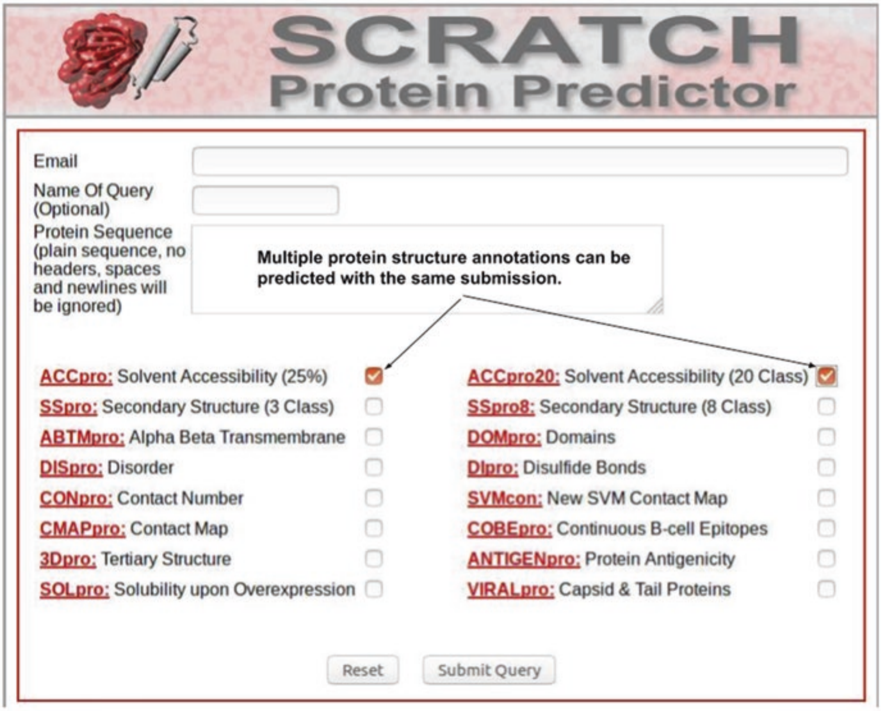
has been observed by Yuan (Yuan 2005) and is motivating the development of predictors for contact number as a possible alternative to SA predictors (Heffernan et al. 2016).

Though there are promising examples of successful NN predictors considering adjacent AA to predict SA since the 1990s (Holbrook et al. 1990), different methods such as linear regression (Xia and Pan 2000) or substitution matrices (Pascarella et al. 1998) have been assessed, but the state of the art has been represented by deep NN since 2002 (Pollastri et al. 2002). Thus, all the SA predictors described below (and summarised in Table 10.2) implement deep NN (Wang et al. 2016; Heffernan et al. 2017; Magnan and Baldi 2014; Mirabello and Pollastri 2013) predicting SA as anything between a two-state problem – i.e. buried and exposed with an average two-state accuracy greater than 80% – and 20-state problem.

SA has been typically measured as accessible surface area (ASA) – i.e. the protein's surface exposed to interactions with the external solvent. ASA is usually obtained normalising the relative SA value observed by the maximum possible value of accessibility for the specific residue according to the DSSP (Kabsch and Sander 1983). The ASA of a protein can be visualised with ASAview, a tool developed in 2004 that requires real values extracted from the PDB or coming from predicted ASA (Ahmad et al. 2004). More recently, a different approach to measuring the SA, called half-sphere exposure (HSE), has been designed by Hamelryck (Thomas 2005). The idea is to split in half the sphere surrounding the Cα atom along the vector of Cα-C_β atoms aiming to provide a more informative and robust measure (Thomas 2005). SPIDER3 can predict both HSE and ASA using real numbers (Heffernan et al. 2017).

10.3.1 ACCpro

ACCpro is a historical SA predictor initially released in 2002 (Pollastri et al. 2002). Since then, it has been developed in parallel with SSpro (see Secondary Structure, SSpro) and last updated to its v5 in 2014, adding support for template-base



SCRATCH
Protein Predictor

Email

Name Of Query (Optional)

Protein Sequence (plain sequence, no headers, spaces and newlines will be ignored)

Multiple protein structure annotations can be predicted with the same submission.

ACCpro: Solvent Accessibility (25%) ☑
SSpro: Secondary Structure (3 Class) □
ABTMpro: Alpha Beta Transmembrane □
DISpro: Disorder □
CONpro: Contact Number □
CMAPpro: Contact Map □
3Dpro: Tertiary Structure □
SOLpro: Solubility upon Overexpression □

ACCpro20: Solvent Accessibility (20 Class) ☑
SSpro8: Secondary Structure (8 Class) □
DOMpro: Domains □
DIpro: Disulfide Bonds □
SVMcon: New SVM Contact Map □
COBEpro: Continuous B-cell Epitopes □
ANTIGENpro: Protein Antigenicity □
VIRALpro: Capsid & Tail Proteins □

Fig. 10.7 A view of SCRATCH protein predictor where both ACCpro predictors have been selected

predictions (Magnan and Baldi 2014). Thus, like SSpro, ACCpro adopts the legacy BLAST to look for evolutionary information at both sequence and structure level. ACCpro predicts whether each residue is more exposed than 25% or not, while ACCpro20, an extension of ACCpro, distinguishes 20 states from 0–95% with incremental steps of 5%, i.e. ACCpro classifies 20 classes, starting from 0–5% to 95–100% of SA (Fig. 10.7).

The web server of ACCpro and ACCpro20 is available at http://scratch.proteomics.ics.uci.edu/ as part of SCRATCH (Cheng et al. 2005). Once an email and the sequence to predict have been inserted, it is possible to select ACCpro or ACCpro20 or any of the available protein predictors (more in Secondary Structure, SSpro).

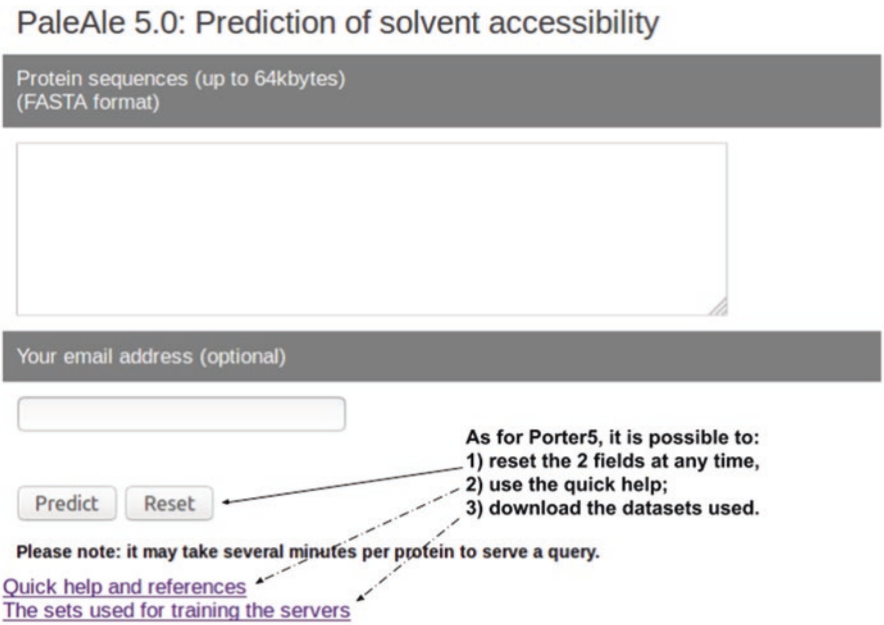
The standalone of ACCpro has been updated in 2015 and is available at http://download.igb.uci.edu/ as part of SCRATCH-1D v1.1. As described above (in Secondary Structure, SSpro), all the requirements are delivered together with the bundled predictors – i.e. ACCpro, ACCpro20, SSpro and SSpro8.

10.3.2 PaleAle

PaleAle is a historical SA predictor developed in parallel with Porter (see Secondary Structure, Porter) since 2007 (Pollastri et al. 2007; Mirabello and Pollastri 2013) and is also based on ensembles of cascaded bidirectional recurrent neural networks (Baldi et al. 1999). PaleAle has been the first template-based SA predictor (Pollastri et al. 2007), while PaleAle (v5) is now able to predict four-state ASA, i.e. exposed at 0–4%, 4–25%, 25–50% or 50 + %. Like Porter5 and Porter+5 (see Torsion Angles), PaleAle5 relies on both HHblits and PSI-BLAST to gather evolutionary information and, thus, improve its predictions (Fig. 10.8).

The web server of PaleAle is available at http://distilldeep.ucd.ie/paleale/. As for Porter and Porter+ (see respective sections), the protein sequence is the only requirement while an email address is optional. More information about these servers is available in the Secondary Structure, Porter subsection.

The light standalone of PaleAle is available at the same address and requires only python3 and HHblits to perform SA predictions. As in Porter, PSI-BLAST can be optionally employed to gather further evolutionary information. The output file presents the confidence per each of the four states predicted. The datasets are released at the same address.


PaleAle 5.0: Prediction of solvent accessibility

Protein sequences (up to 64kbytes)
(FASTA format)

Your email address (optional)

As for Porter5, it is possible to:
1) reset the 2 fields at any time,
2) use the quick help;
3) download the datasets used.

Please note: it may take several minutes per protein to serve a query.

Quick help and references
The sets used for training the servers

Fig. 10.8 A view of PaleAle5 where the reset button and the links are highlighted

10.3.3 RaptorX-Property

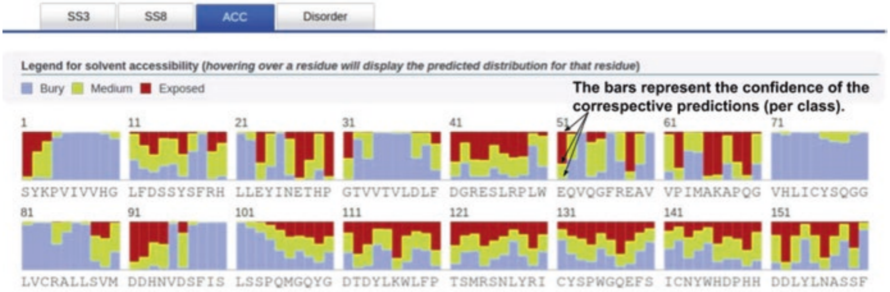
RaptorX-Property, described in section Secondary Structure, is 2016 suite of predictors able to predict SA, SS and disorder regions (Wang et al. 2016). RaptorX-Property predicts SA in three states with thresholds at 10% and 40%, respectively. As for SS predictions, RaptorX-Property can avoid to look for evolutionary information to speed up predictions at the cost of lower accuracy. It relies on HHblits (Remmert et al. 2012) to gather evolutionary information (Fig. 10.9).

The web server of RaptorX-Property is available at http://raptorx.uchicago.edu/StructurePropertyPred/predict/. The result page of RaptorX-Property provides the predicted 1D annotations in different tabs (Fig. 10.9 shows the three-state SA). The web server and the released standalone are described in section Secondary Structure, RaptorX-Property.

10.3.4 SPIDER3

SPIDER has been able to predict SA, SS and TA since 2015 (Heffernan et al. 2015) and was updated in 2017 (Heffernan et al. 2017). SPIDER3, described also in sections Secondary Structure and Torsion Angles, predicts the ASA using real numbers rather than classes, differently from the other predictors here presented (Heffernan et al. 2015). SPIDER2 has been the first HSE predictor (Heffernan et al. 2016), while SPIDER3 predicts HSEα-up and HSEα-down using real numbers, although Heffernan et al. reports result also in HSEβ-up and HSEβ-down (Heffernan et al. 2017).

The web server and the standalone of SPIDER3 are described in Secondary Structure, SPIDER3. As a side note, the result page and the confirmation email of the web server show the predicted SA only as ASA in ten classes – i.e. [0–9] – while the predicted ASA, HSEβ-up and HSEβ-down in real numbers are listed in the output file ("*.spd33") in the temporary directory, along with PSSM/HMM files (see Figs. 10.5 and 10.10).



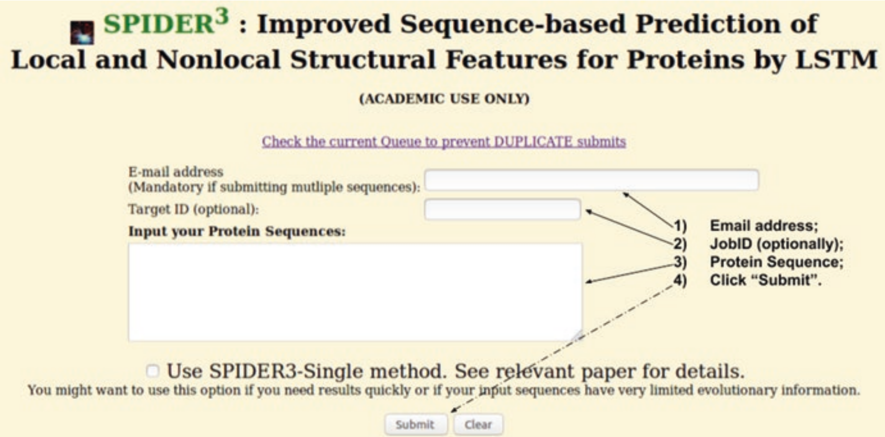
Legend for solvent accessibility *(hovering over a residue will display the predicted distribution for that residue)*

🟦 Bury 🟩 Medium 🟥 Exposed

The bars represent the confidence of the correspective predictions (per class).

```
1          11         21         31         41         51         61         71
SYKPVIVVHG LFDSSYSFRH LLEYINETHP GTVVTVLDLF DGRESLRPLW EQVQGFREAV VPIMAKAPQG VHLICYSQGG
81         91         101        111        121        131        141        151
LVCRALLSVM DDHNVDSFIS LSSPQMGQYG DTDYLKWLFP TSMRSNLYRI CYSPWGQEFS ICNYWHDPHH DDLYLNASSF
```

Fig. 10.9 The view on the predicted three-state SA performed by RaptorX-Property



SPIDER^3 : Improved Sequence-based Prediction of Local and Nonlocal Structural Features for Proteins by LSTM

(ACADEMIC USE ONLY)

Check the current Queue to prevent DUPLICATE submits

E-mail address (Mandatory if submitting mutliple sequences):

Target ID (optional):

Input your Protein Sequences:

□ Use SPIDER3-Single method. See relevant paper for details.

You might want to use this option if you need results quickly or if your input sequences have very limited evolutionary information.

1) Email address;
2) JobID (optionally);
3) Protein Sequence;
4) Click "Submit".

Fig. 10.10 A view of the input window of SPIDER3. The steps to follow to start a prediction are highlighted

10.4 Torsional Angles

Protein torsion (or dihedral or rotational) angles can accurately describe the local conformation of protein backbones. The main protein backbone dihedral angles are phi (ϕ), psi (ψ) and omega (ω). The planarity of protein bonds restricts ω to be either 180° (typical case) or 0° (rarely). Therefore, it is generally sufficient to use ϕ and ψ to accurately describe the local shape of a protein.

TA are highly correlated to protein SS and particularly informative in highly variable loop regions. In fact, while TA of α-helices and β-sheets are mostly clustered and regularly distributed (Kuang et al. 2004), ϕ and ψ can be more effective in describing the local conformation of residues when they are classified as coils (i.e. neither of the other SS classes). When four consecutive residues are considered, a different couple of angles can be observed: theta (θ) and tau (τ) (Lyons et al. 2014). Thus, different annotations (i.e. SS, ϕ/ψ and θ/τ) can be adopted to describe the backbone of a protein (Fig. 10.11).

TA are essentially an alternative representation of local structure with respect to SS. Both TA and SS have been successfully used as restraints towards sequence alignment (Huang and Bystroff 2006), protein folding (Yang et al. 2011) and tertiary structure prediction (Faraggi et al. 2009). HMM (Bystroff et al. 2000), support vector machines (SVM) (Kuang et al. 2004) and several architectures of NN (e.g. iterative (Heffernan et al. 2017; Heffernan et al. 2015) and cascade-correlation (Wood and Hirst 2005)) have been analysed to predict TA since 2000. NN are currently the main tool to predict TA, in parallel with protein SS (Heffernan et al. 2017) or sequentially after it (Wood and Hirst 2005; Mooney et al. 2006).

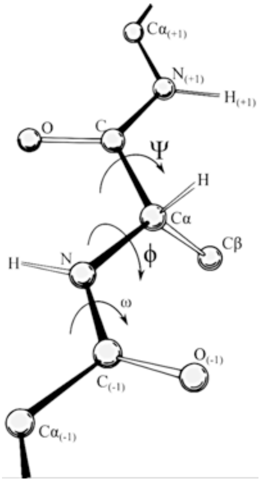


Fig. 10.11 Protein backbone dihedral angles phi, psi and omega; credits: https://commons.wikimedia.org/wiki/File:Protein_backbone_PhiPsiOmega_drawing.svg

Table 10.3 ϕ/ψ angles prediction web server

Name	Web server	Notes
Porter+ (Mooney et al. 2006)	http://distilldeep.ucd.ie/porter+	ϕ/ψ in 16 letters
SPIDER3 (Heffernan et al. 2017)	http://sparks-lab.org/server/SPIDER3/	ϕ/ψ and θ/τ , Numpy or Tensorflow

Table 10.4 Protein contact maps prediction servers

Name	Web server	Notes
DNCON (Adhikari et al. 2017)	http://sysbio.rnet.missouri.edu/dncon2/	Three coevolution algorithms, Computer Vision inspired
MetaPSICOV (Jones et al. 2015)	http://bioinf.cs.ucl.ac.uk/MetaPSICOV/	CCMpred, FreeContact and PSICOV, hydrogen bonds
RaptorX-Contact (Wang et al. 2017)	http://raptorx.uchicago.edu/ContactMap/	Inspired from Computer Vision, CCMpred only
XX-Stout (Walsh et al. 2009)	http://distilldeep.ucd.ie/xxstout/	Contact Density, template-based, multi-class CM

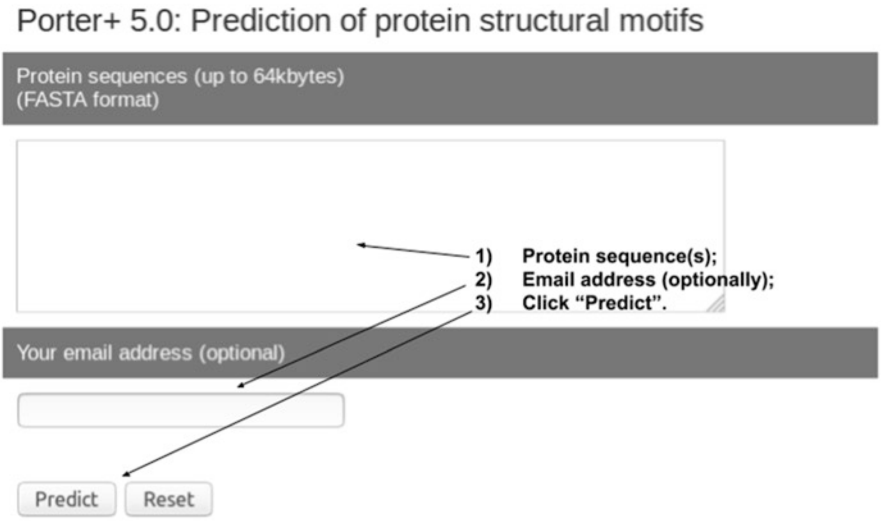
ϕ and ψ can be predicted as real numbers or letters(/clusters). In fact, ϕ and ψ can range from 0° to 360° but are typically observed in certain ranges, given from chemical and physical characteristics of proteins. Bayesian probabilistic (De Brevern et al. 2000; Ting et al. 2010), multidimensional scaling (MDS) (Sims et al. 2005) and density plot (Kuang et al. 2004) approaches have been exploited to define different alphabets of various sizes (Tables 10.3 and 10.4).

10.4.1 Porter+

Porter+ is a TA predictor able to classify the ϕ and ψ angles of a given protein. It was initially developed in 2006 as intermediate step to improve Porter (a SS predictor described in section Secondary Structure) (Mooney et al. 2006). Porter+ adopts an alphabet of 16 letters devised by Sims et al. using MDS on tetrapeptides (four contiguous residues) (Sims et al. 2005). Porter+, similarly to Porter and PaleAle (see Solvent Accessibility, PaleAle), implements BLAST+ to gather evolutionary information and improve the final prediction. As Porter and PaleAle, the most recent version of Porter+ (v5) adopts also HHblits to greatly improve its accuracy.

The web server of Porter+ is available at http://distilldeep.ucd.ie/porter+. The protein sequence is required, while an email address is optional. It will be then sufficient to confirm (clicking "Predict") to view a confirmation page with the overview of the job. Once ready, the prediction will be received by email. It will resemble the format adopted for Porter; see in section Secondary Structure. Porter+ can be executed in parallel with Porter or PaleAle, or several more protein predictors, at http://distillf.ucd.ie/distill/ to predict SS, SA or other protein features, respectively (Fig. 10.12).

The light standalone of Porter+ is available at http://distilldeep.ucd.ie/porter++ and closely resembles the one described in section Secondary Structure, Porter. The output of Porter+ overviews the confidence for all 14 classes predicted. The datasets adopted for training and testing purposes are also released.

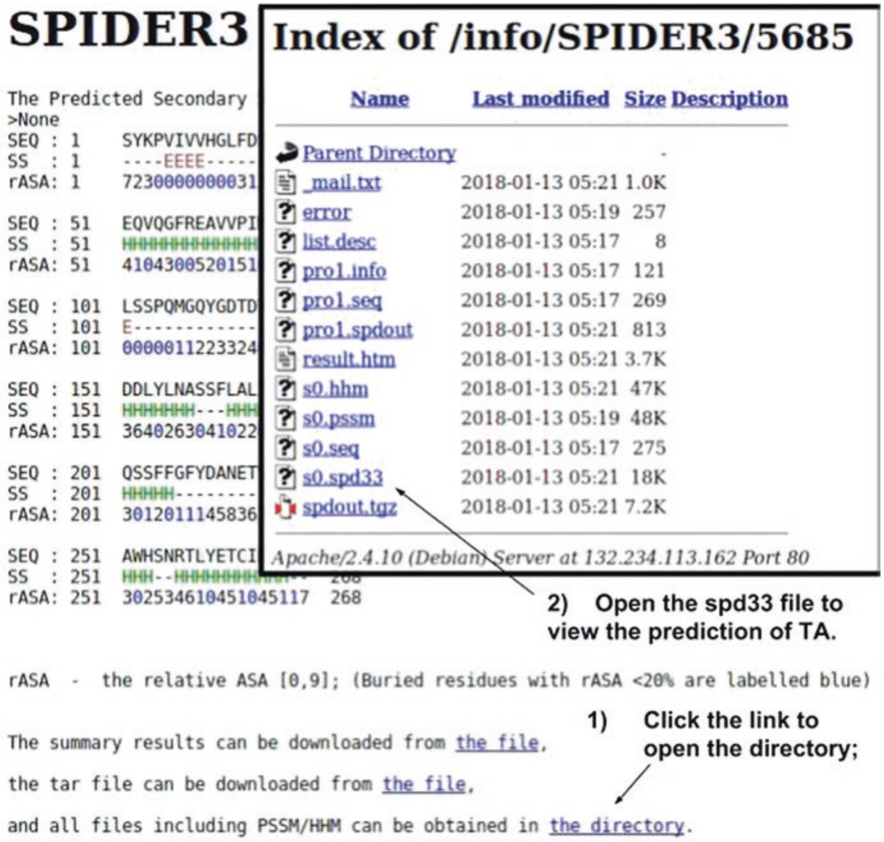

Porter+ 5.0: Prediction of protein structural motifs
Protein sequences (up to 64kbytes)
(FASTA format)
Your email address (optional)
Predict
Reset

1) Protein sequence(s);
2) Email address (optionally);
3) Click "Predict".

Fig. 10.12 A view of Porter+5 where the steps to start a prediction are highlighted

10.4.2 SPIDER3

SPIDER3, also in section Secondary Structure and Solvent Accessibility, predicts TA using real numbers (R). SPIDER was initially released in 2014 to predict only θ/τ (Lyons et al. 2014). It has been further developed to also predict ϕ/ψ, in parallel with SS, SA and contact numbers (see the respective sections) (Heffernan et al. 2017; Heffernan et al. 2015). More details, regarding the pipeline implemented, the web server offered and the standalone available, are outlined in section Secondary Structure (Fig. 10.13).



SPIDER3

```
The Predicted Secondary
>None
SEQ : 1     SYKPVIVVHGLFD
SS  : 1     ----EEEE-----
rASA: 1     7230000000031

SEQ : 51    EQVQGFREAVVPI
SS  : 51    HHHHHHHHHHHHH
rASA: 51    4104300520151

SEQ : 101   LSSPQMGQYGDTD
SS  : 101   E------------
rASA: 101   0000011223324

SEQ : 151   DDLYLNASSFLAL
SS  : 151   HHHHHHH---HHH
rASA: 151   3640263041022

SEQ : 201   QSSFFGFYDANET
SS  : 201   HHHHH--------
rASA: 201   3012011145836

SEQ : 251   AWHSNRTLYETCI
SS  : 251   HHH--HHHHHHHH
rASA: 251   302534610451045117   268

rASA  -  the relative ASA [0,9]; (Buried residues with rASA <20% are labelled blue)

The summary results can be downloaded from the file,

the tar file can be downloaded from the file,

and all files including PSSM/HHM can be obtained in the directory.
```

Index of /info/SPIDER3/5685

Name	Last modified	Size	Description
Parent Directory		-	
_mail.txt	2018-01-13 05:21	1.0K	
error	2018-01-13 05:19	257	
list.desc	2018-01-13 05:17	8	
pro1.info	2018-01-13 05:17	121	
pro1.seq	2018-01-13 05:17	269	
pro1.spdout	2018-01-13 05:21	813	
result.htm	2018-01-13 05:21	3.7K	
s0.hhm	2018-01-13 05:21	47K	
s0.pssm	2018-01-13 05:19	48K	
s0.seq	2018-01-13 05:17	275	
s0.spd33	2018-01-13 05:21	18K	
spdout.tgz	2018-01-13 05:21	7.2K	

Apache/2.4.10 (Debian) Server at 132.234.113.162 Port 80

2) Open the spd33 file to view the prediction of TA.

1) Click the link to open the directory;

Fig. 10.13 A view of the results page of SPIDER3 where the steps to view the predicted TA are highlighted

10.5 Contact Maps

Contact Maps (CM) are the main two-dimensional protein structure annotation tools. A plain 2D representation of protein tertiary structure would describe the distance between all possible pairs of AA using a matrix containing real values. Such dense representation, referred as distance map, is reduced to a more compact abstraction – i.e. CM – by quantising a distance map through a fixed threshold, i.e. describing distances not as real numbers but as contacts (distance smaller than the threshold) or no. This latter abstraction is routinely exploited to reconstruct protein tertiary structures implementing heuristic methods (Vassura et al. 2008; Vendruscolo et al. 1997). Thus, 3D structure prediction being a computationally expensive problem motivates the development of the aforementioned heuristic methods that aim to be both robust against noise in the CM – i.e. to ideally fix CM prediction errors – and computationally applicable on a large scale (Vassura et al. 2011; Kukic et al. 2014). Following closely the development of the third generation of SS predictors, motivated by the same abundance of available data and computational resources, MSA have been thoroughly tested and successfully exploited to extract promising features for CM prediction – e.g. correlated mutations, sequence conservation, alignment stability and family size (Pazos et al. 1997; Olmea and Valencia 1997; Göbel et al. 1994). These initial advancements led to the first generation of ML methods able to predict CM (Vullo et al. 2006; Fariselli et al. 2001; Cheng and Baldi 2007). Given that MSA are replete with useful but noisy information, statistical insights have been necessary to further exploit the growing amount of evolutionary information – e.g. distinguishing between indirect and direct coupling (Jones et al. 2012; Di Lena et al. 2011). The most recent CM predictors gather recent intuitions in both statistics and advanced ML, aiming to collect, clean and employ as much useful data as possible (Jones et al. 2015; Wang et al. 2016; Adhikari et al. 2017). Differently from the other protein annotations in this chapter, CM is currently assessed at CASP (Schaarschmidt et al. 2018) and CAMEO (Haas et al. n.d.).

The intrinsic properties of CM – namely, being compact and discrete two-state annotations, invariant to rotations and translations – make them a more appropriate target for ML techniques than protein tertiary structures or distance maps although still highly informative about the protein 3D structures (Bartoli et al. 2008). CM prediction is a typical intermediate step in many pipelines to predict protein tertiary structure (Mooney and Pollastri 2009; Roy et al. 2010; Kosciolek and Jones 2014). For example, it is a key component for contact-assisted structure prediction (Kinch et al. 2016), contact-assisted protein folding (Wang et al. 2017) and free and template-based modelling (Roy et al. 2010). CM have also been used to predict protein disorder (Schlessinger et al. 2007) and protein function (Pazos et al. 1997) and to detect challenging templates (Mooney and Pollastri 2009). In fact, even partial CM can greatly support robust and accurate protein structure modelling (Kim et al. 2014).

Being a 2D annotation, CM are typically gradually predicted starting from simpler but less informative 1D annotations – e.g. SA, SS and TA (Fariselli et al. 2001; Cheng and Baldi 2007; Pollastri and Baldi 2002). The advantages of this incremental approach lie in the intrinsic nature of protein abstractions – i.e. 1D annotations are easier to predict while providing useful insights. For example, Fig. 10.14 highlights the strong relations between SS conformations and CM. The contact occupancy – i.e. contact number, or number of contacts per AA – is another 1D protein annotation which has been successfully predicted (Heffernan et al. 2017; Pollastri et al. 2001, 2002) to adjust and improve CM prediction (Olmea and Valencia 1997; Fariselli et al. 2001; Pollastri and Baldi 2002). Eigenvector decomposition has been used as a means for template search (Di Lena et al. 2010) and principal eigenvector (PE) prediction as an intermediate step towards CM prediction (Vullo et al. 2006). Finally, correlated mutations appear to be the most informative protein feature for CM prediction – i.e. residues in contact tend to coevolve to maintain the physiochemical equilibrium (Pazos et al. 1997; Olmea and Valencia 1997; Göbel et al. 1994). Thus, statistical methods have been extensively assessed to look for coevolving residues, gathering mutual information from MSA while aiming to discriminate direct from indirect coupling mutations, e.g. implementing sparse inverse covariance estimation to remove indirect coupling (Jones et al. 2012; Kaján et al. 2014; Seemayer et al. 2014).

As in Fig. 10.14, CM are represented as (symmetric) matrices or graphs – rather than vectors – where around 2–5% of all possible pairs of AA are "in contact", i.e. an unbalanced problem in ML (Bartoli et al. 2008). Notably, the number of AA in

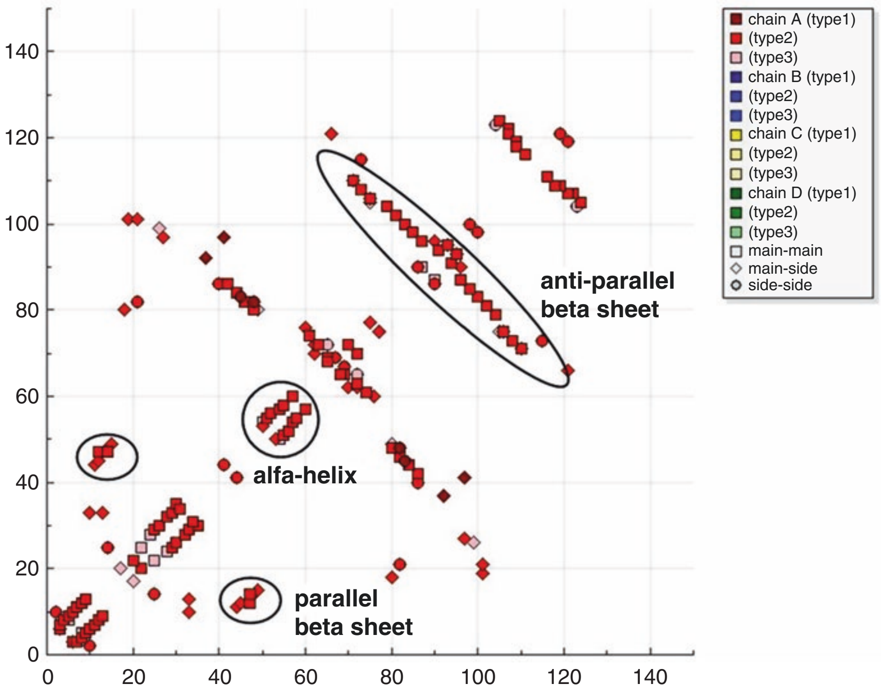


Fig. 10.14 CM with highlighted SS conformations; credits: https://commons.wikimedia.org/wiki/File:Elements_hb2.jpg

contact increases almost linearly with the protein length – i.e. shorter proteins are denser than longer ones (Bartoli et al. 2008). A pair of AA is in contact when the Euclidian distance between their C_β (or $C\alpha$, for glycine) atoms is closer than a given threshold. This threshold is usually set between 6 and 12 Å (8 Å at CASP (Schaarschmidt et al. 2018)), although values in the range of 10–18 Å may lead to better reconstructions (Vassura et al. 2008). In fact, it is arguable whether all predicted "contacts" should be taken in consideration or certain criteria should be applied, such as focusing on those predicted with the highest confidence – i.e. the top 10, $L/5$, $L/2$ or L contacts, with L = protein length – or with a minimum probability threshold (Schaarschmidt et al. 2018). For example, tertiary structure modelling benefits more from well-distributed contacts; thus the entropy score is one of the measures of interest to evaluate CM predictors (Schaarschmidt et al. 2018). Precision – i.e. the ratio between true contact and wrong contact (true contact + wrong contact) – is usually adopted to assess local (short range) contacts, i.e. involving AA within ten positions apart, and non-local (long range) contact, separately. Typically, CM predictors are evaluated at CASP through more complex measures (Schaarschmidt et al. 2018; Kinch et al. 2016; Monastyrskyy et al. 2014), such as z-scores, i.e. weighted sum of energy separation with the true structure for each domain; GDT_TS, i.e. score of optimal superposition between the predicted and the true structure; and root-mean-square deviation (RMSD) or TM-score, i.e. a measure more sensitive at the global (rather than local) structure than RMSD (Zemla 2003). Classic statistical and ML measures, such as the aforementioned precision, recall, F1 score and Matthews correlation coefficient (MCC), are also adopted in parallel with more unusual ones, such as alignment depth or entropy score (Schaarschmidt et al. 2018). The average precision of the top predictors at CASP12 was 47% on L/5 long-range contacts for the difficult category, while the highest GDT_TS for each of the 14 domains assessed went from 12 to 70 (Schaarschmidt et al. 2018).

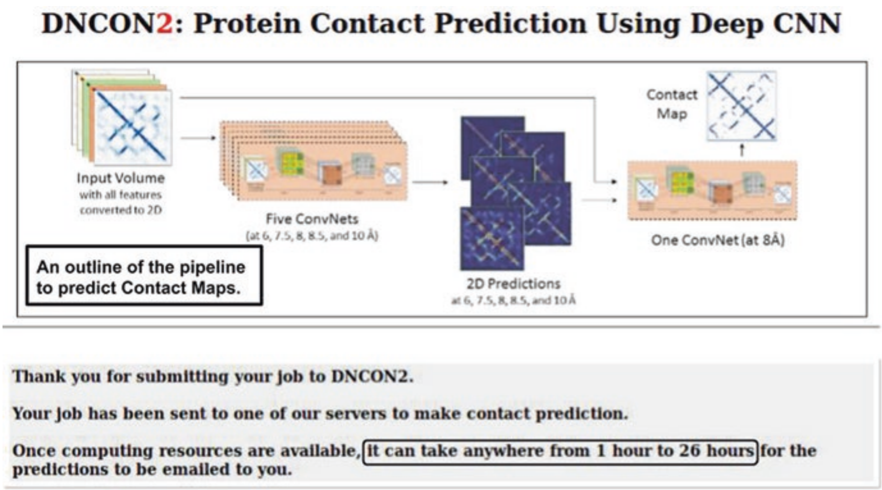
Though correlated mutations and NN have been identified as promising instruments to also predict CM (Fariselli et al. 2001), pairwise contact potential (Schlessinger et al. 2007), self-organising maps (MacCallum 2004) and SVM (Cheng and Baldi 2007) have been used in the past. While 2D-BRNN (Pollastri and Baldi 2002; Tegge et al. 2009), multistage (Vullo et al. 2006; Di Lena et al. 2012) and template-based (Walsh et al. 2009) NN approaches have initially characterised the field (Martin et al. 2010), the most recent CM predictors rely on multiple 1D protein annotation predictors – e.g. predicting SA and SS along with other protein features – two-stage approaches and coevolution information (Adhikari et al. 2017; Buchan and Jones 2018) or multi-class maps (Kukic et al. 2014; Martin et al. 2010). The standard output format of any CM predictor is a text file organised in five columns as follows: the positions of the two AA in contact, a blank column, the set threshold (8 Å) and the confidence of each predicted contact.

10.5.1 DNCON2

DNCON has been initially released in 2012 (Eickholt and Cheng 2012), assessed at CASP10 (Monastyrskyy et al. 2014) and updated in 2017 (Adhikari et al. 2017). DNCON2 gathers coevolution signal along with 1D protein features – e.g. PSIPRED and SSpro (see section Secondary Structure) – with a similar approach to MetaPSICOV2 (see below). It then predicts CM with different thresholds – namely, 6, 7.5, 8, 8.5 and 10 Å – resembling the multi-class maps of XX Stout (see below) and finally refines them generating only one CM at 8 Å. In the described two-stage approach, DNCON2 implements a total of six NN like RaptorX-Contact (Fig. 10.15). Thus, DNCON2 further exploits the most recent intuitions in CM prediction, including recent ML algorithms.

The web server and dataset of DNCON2 are available at http://sysbio.rnet.missouri.edu/dncon2/. JobID and email are required, along with the sequence to predict (up to two sequences at time). Once the prediction is ready, typically in less than 24 h, the predicted CM is sent by email in both text and image format as email content and attachment, respectively. The email content specifies the number of alignments found and the predicted CM (in the standard five columns text format).

The standalone of DNCON2 is available at https://github.com/multicom-toolbox/DNCON2/. The same page lists all the instructions to install every requirement, i.e. CCMpred (Seemayer et al. 2014), FreeContact (Kaján et al. 2014), HHblits (Remmert et al. 2012), JackHMMER (Johnson et al. 2010) and PSICOV (Jones et al. 2012) for coevolution information, python libraries (such as Tensorflow), MetaPSICOV and PSIPRED (see Secondary Structure, PSIPRED) for SS and SA prediction. Once all the requirements are met, it is possible to verify whether


DNCON2: Protein Contact Prediction Using Deep CNN

An outline of the pipeline to predict Contact Maps.

Input Volume
with all features converted to 2D
Five ConvNets
(at 6, 7.5, 8, 8.5, and 10 Å)
2D Predictions
at 6, 7.5, 8, 8.5, and 10 Å
Contact Map
One ConvNet (at 8Å)

Thank you for submitting your job to DNCON2.

Your job has been sent to one of our servers to make contact prediction.

Once computing resources are available, it can take anywhere from 1 hour to 26 hours for the predictions to be emailed to you.

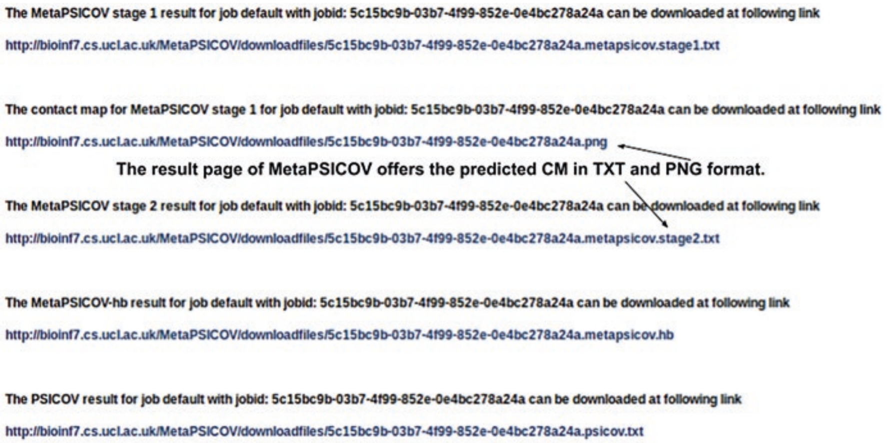
Fig. 10.15 The pipeline of DNCON2 is summarised in the confirmation page

DNCON2 is fully running dealing with the predictions of three proposed sequences. The results of each predictor and package involved are organised in directories.

10.5.2 MetaPSICOV

MetaPSICOV is a CM predictor which has been initially released in 2014 for CASP11 (Kosciolek and Jones 2016) and updated in 2016 for CASP12 (Buchan and Jones 2018). It is recognised as the first CM predictor successfully able to exploit the recent advancements in coevolutionary information extraction (Monastyrskyy et al. 2016). In particular, MetaPSICOV achieved this result implementing three different algorithms to extract coevolution signal from MSA generated with HHblits (Remmert et al. 2012) and HMMER (Finn et al. 2011) – i.e. CCMpred (Seemayer et al. 2014), FreeContact (Kaján et al. 2014) and PSICOV (Jones et al. 2012) – along with other local and global features used for SVMcon (Cheng and Baldi 2007). It relies on PSIPRED (see Secondary Structure, PSIPRED) to predict SS and a similar ML method to predict SA. As a final step, MetaPSICOV adopts a two-stage NN to infer CM from the features described (Jones et al. 2015). The web server and standalone of MetaPSICOV can be used to predict hydrogen-bonding patterns (Jones et al. 2015).

The web server of the 2014 version of MetaPSICOV is available at http://bioinf.cs.ucl.ac.uk/MetaPSICOV. A simple interface, which resembles the web server of PSIPRED (see Secondary Structure, PSIPRED), asks for a single sequence in FASTA format and a short identifier. A confirmation page is automatically shown when the job is completed. If an email address is inserted, an email containing only the permalink to the result page will be sent. As in Fig. 10.16, the result page contains links to the output of MetaPSICOV stage 1 (also as image), of stage 2, of



The MetaPSICOV stage 1 result for job default with jobid: 5c15bc9b-03b7-4f99-852e-0e4bc278a24a can be downloaded at following link

http://bioinf7.cs.ucl.ac.uk/MetaPSICOV/downloadfiles/5c15bc9b-03b7-4f99-852e-0e4bc278a24a.metapsicov.stage1.txt

The contact map for MetaPSICOV stage 1 for job default with jobid: 5c15bc9b-03b7-4f99-852e-0e4bc278a24a can be downloaded at following link

http://bioinf7.cs.ucl.ac.uk/MetaPSICOV/downloadfiles/5c15bc9b-03b7-4f99-852e-0e4bc278a24a.png

The result page of MetaPSICOV offers the predicted CM in TXT and PNG format.

The MetaPSICOV stage 2 result for job default with jobid: 5c15bc9b-03b7-4f99-852e-0e4bc278a24a can be downloaded at following link

http://bioinf7.cs.ucl.ac.uk/MetaPSICOV/downloadfiles/5c15bc9b-03b7-4f99-852e-0e4bc278a24a.metapsicov.stage2.txt

The MetaPSICOV-hb result for job default with jobid: 5c15bc9b-03b7-4f99-852e-0e4bc278a24a can be downloaded at following link

http://bioinf7.cs.ucl.ac.uk/MetaPSICOV/downloadfiles/5c15bc9b-03b7-4f99-852e-0e4bc278a24a.metapsicov.hb

The PSICOV result for job default with jobid: 5c15bc9b-03b7-4f99-852e-0e4bc278a24a can be downloaded at following link

http://bioinf7.cs.ucl.ac.uk/MetaPSICOV/downloadfiles/5c15bc9b-03b7-4f99-852e-0e4bc278a24a.psicov.txt

Fig. 10.16 A typical result page of MetaPSICOV. All the files, except the png, follow PSICOV's format

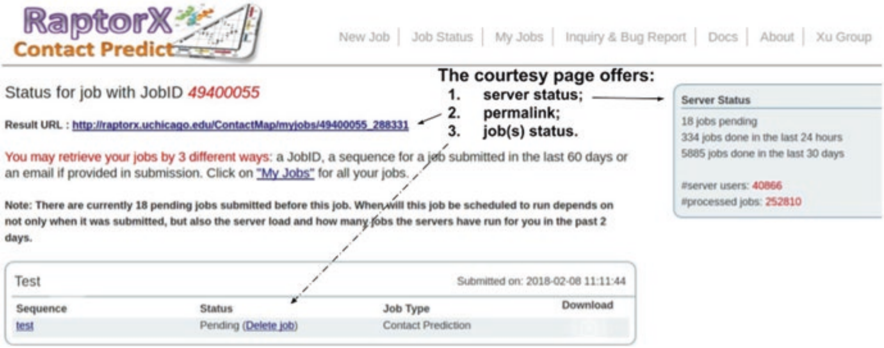
MetaPSICOV-hb (hydrogen bonds) and of PSICOV. A typical CM takes between 20 min and 6 h to be predicted.

The very last version of MetaPSICOV is usually available as standalone at http://bioinfadmin.cs.ucl.ac.uk/downloads/MetaPSICOV/. To run MetaPSICOV2, it is required to install (legacy) BLAST, PSIPRED, PSICOV, FreeContact, CCMpred, HHblits and HMMER, separately. Once the required packages are installed, it is sufficient to follow the README to complete the setup and run MetaPSICOV2. Each run of MetaPSICOV2 will generate the needed features – i.e. the output of the required packages, such as PSIPRED and PSICOV – along with the predicted CM (in standard text format).

10.5.3 RaptorX-Contact

RaptorX-Contact is a 2016 CM predictor which performed well at the last CASP12 (Wang et al. 2017; Schaarschmidt et al. 2018; Wang et al. 2018). RaptorX-Contact aimed to exploit both computer vision (LeCun et al. 2015) and coevolution intuitions to further improve CM prediction. It employs RaptorX-Property (Wang et al. 2016) (see Secondary Structure and Solvent Accessibility) to predict SS and SA, CCMpred (Seemayer et al. 2014) to look for coevolutionary information and in-house algorithms for mutual information and pairwise potential extraction. RaptorX-Contact was trained using MSA generated with PSI-BLAST (Schäffer et al. 2001) while it uses HHblits (Remmert et al. 2012) at prediction time. Thus, the web server and standalone depend on HHblits only.

The web server of RaptorX-Contact is available at http://raptorx.uchicago.edu/ContactMap/. Once a protein sequence (in FASTA format) has been inserted, it is possible to submit it and a result URL will be provided (Fig. 10.17). A JobID is recommended to distinguish among past submissions in the "My Jobs" page, while


RaptorX
Contact Predict
New Job | Job Status | My Jobs | Inquiry & Bug Report | Docs | About | Xu Group

The courtesy page offers:
1. server status;
2. permalink;
3. job(s) status.

Status for job with JobID 49400055

Result URL : http://raptorx.uchicago.edu/ContactMap/myjobs/49400055_288331

You may retrieve your jobs by 3 different ways: a JobID, a sequence for a job submitted in the last 60 days or an email if provided in submission. Click on "My Jobs" for all your jobs.

Note: There are currently 18 pending jobs submitted before this job. When will this job be scheduled to run depends on not only when it was submitted, but also the server load and how many jobs the servers have run for you in the past 2 days.

Server Status
18 jobs pending
334 jobs done in the last 24 hours
5885 jobs done in the last 30 days

#server users: 40866
#processed jobs: 252810

Test — Submitted on: 2018-02-08 11:11:44

Sequence	Status	Job Type	Download
test	Pending (Delete job)	Contact Prediction	

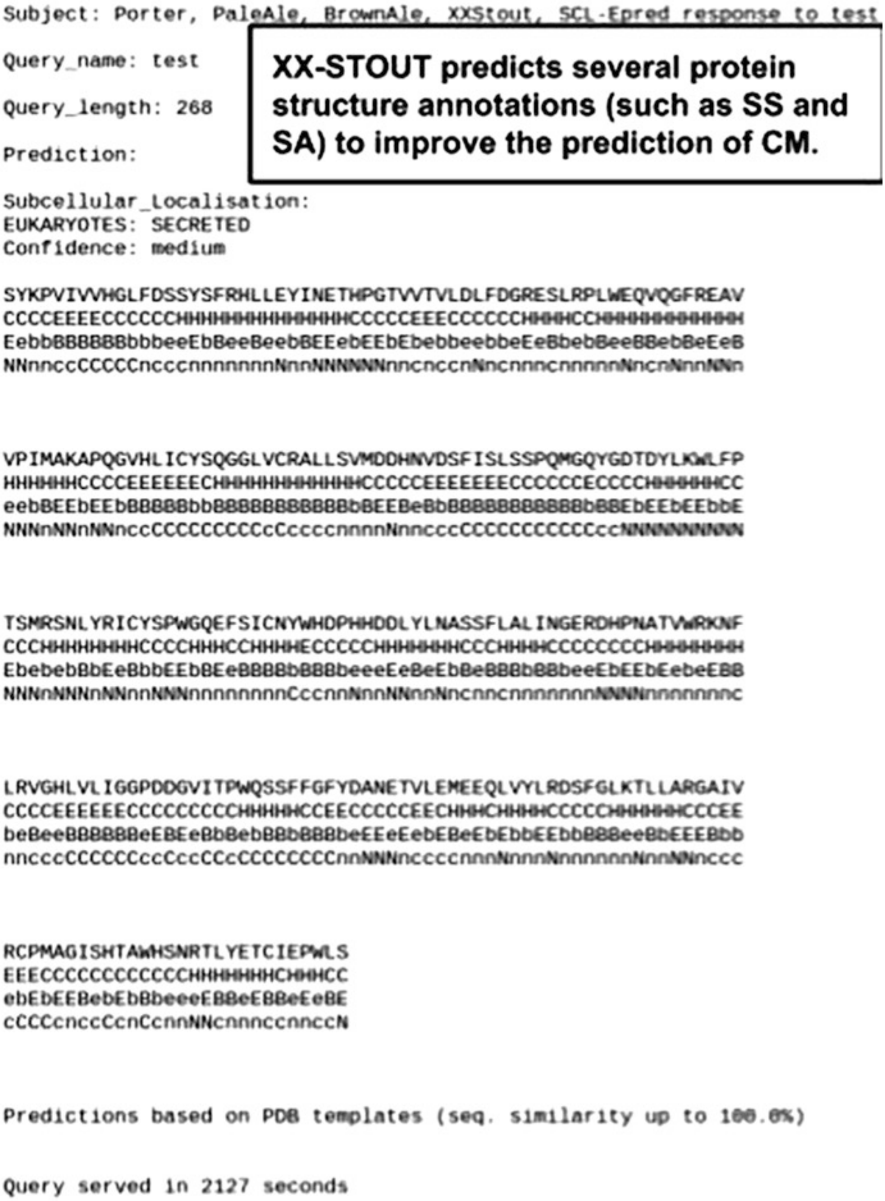
Fig. 10.17 The confirmation page of RaptorX-Contact tells the pending jobs ahead and the result URL

an email address can be specified to receive the outcome of RaptorX-Contact by email – i.e. the result URL and, as attachments, the predicted CM in text and image format. The tertiary structure is also predicted by default, but it is possible to uncheck the respective box to speed up the CM prediction. Up to 50 protein primary structures can be submitted at the same time through the input form or uploaded from one's computer. Optionally, a MSA (of up to 20,000 sequences) can be sent instead of a protein sequence. The result URL links to an interactive page where it is possible to navigate the predicted CM besides downloading it in text or image format. The MSA generated (in A2M format), the CCMpred (Seemayer et al. 2014) output and the 3D models (if requested) are also made available. Finally, it is also possible to query the web server from command line (using curl) as explained at http://raptorx.uchicago.edu/ContactMap/documentation/.

10.5.4 XX-STOUT

XX STOUT is a CM predictor initially released in 2006 (Vullo et al. 2006) and further improved to be template-based (Walsh et al. 2009) and multi-class in 2009 (Martin et al. 2010). XX STOUT employs the predictions by BrownAle, PaleAle and Porter (see Secondary Structure and Solvent Accessibility) – i.e. contact density, SS and SA predictions, respectively – to generate multi-class CM, i.e. CM with four-state annotations. When either PSI-BLAST (Schäffer et al. 2001) or the in-house fold recognition software finds homology information, further inputs are provided to XX STOUT to perform template-based predictions – i.e. greatly improve the prediction quality exploiting proteins in the PDB (Berman et al. 2000; Mooney and Pollastri 2009).

The web server of XX STOUT is available at http://distilldeep.ucd.ie/xxstout/. An email address and the plain protein sequence are required to start the prediction; a JobID is optional. The confirmation page summarises the information provided and the predictors which are going to be used – i.e. the aforementioned 1D predictors and SCL-Epred, a predictor of subcellular localisation (Mooney et al. 2013). The predicted CM (threshold 8 Å), the prediction per residue of SS, SA and contact density and the predicted protein's location are sent by email. The same email describes the confidence of SCL-Epred's prediction and whether the whole prediction has been based on PDB templates and, if found, of which similarity with the query sequence. The standalone of XX STOUT and required 1D predictors are available on request (Fig. 10.18).



```
Subject: Porter, PaleAle, BrownAle, XXStout, SCL-Epred response to test

Query_name: test

Query_length: 268

Prediction:

Subcellular_Localisation:
EUKARYOTES: SECRETED
Confidence: medium

SYKPVIVVHGLFDSSYSFRHLLEYINETHPGTVVTVLDLFDGRESLRPLWEQVQGFREAV
CCCCEEEECCCCCCHHHHHHHHHHHHHHCCCCCEEECCCCCCHHHHCCHHHHHHHHHHHH
EebbBBBBBBbbbeeEbBeeBeebBEEebEEbEbebbeebbeEeBbebBeeBBebBeEeB
NNnnccCCCCCncccnnnnnnnnNnnNNNNNNnncnccnNncnnncnnnnnNncnNnnNNn


VPIMAKAPQGVHLICYSQGGLVCRALLSVMDDHNVDSFISLSSPQMGQYGDTDYLKWLFP
HHHHHHCCCCEEEEEEECHHHHHHHHHHHHCCCCCEEEEEEEECCCCCCECCCCHHHHHHCC
eebBEEbEEbBBBBBbbBBBBBBBBBBBBbBEEBeBbBBBBBBBBBBBBbBBEbEEbEEbbE
NNNnNNnNNnccCCCCCCCCCcCccccnnnnNnncccCCCCCCCCCCCccNNNNNNNNNN


TSMRSNLYRICYSPWGQEFSICNYWHDPHHDDLYLNASSFLALINGERDHPNATVWRKNF
CCCHHHHHHHHCCCCHHHCCHHHHECCCCCHHHHHHHCCCHHHHCCCCCCCCHHHHHHHH
EbebebBbEeBbbEEbBEeBBBBbBBBbeeeEeBeEbBeBBBbBBbeeEbEEbEebeEBB
NNNnNNNnNNnnNNNnnnnnnnnnCccnnNnnNNnnNncnncnnnnnnnnNNNNnnnnnnnnc


LRVGHLVLIGGPDDGVITPWQSSFFGFYDANETVLEMEEQLVYLRDSFGLKTLLARGAIV
CCCCEEEEEEECCCCCCCCCHHHHHCCEECCCCCEECHHHCHHHHCCCCCHHHHHHCCCEE
beBeeBBBBBBeEBEeBbBebBBbBBBbeEEeEebEBeEbEbbEEbbBBBeeBbEEEBbb
nncccCCCCCCcccCccCCcCCCCCCCCCnnNNNnccccnnnNnnnNnnnnnnnNnnNNnccc


RCPMAGISHTAWHSNRTLYETCIEPWLS
EEECCCCCCCCCCCCHHHHHHHCHHHCC
ebEbEEBebEbBbeeeEBBeEBBeEeBE
cCCCcnccCcnCcnnNNcnnnccnnccN


Predictions based on PDB templates (seq. similarity up to 100.0%)


Query served in 2127 seconds
```

XX-STOUT predicts several protein structure annotations (such as SS and SA) to improve the prediction of CM.

Fig. 10.18 XX STOUT sends the predicted protein structure annotations in the body email except the CM (which is attached)

10.6 Conclusions

In this chapter we have discussed the importance of protein structure to understand protein functions and the need for abstractions – i.e. protein structural annotations – to overcome the difficulties of determining such structures in vitro. We have then presented an overview of the role bioinformatics – i.e. in silico biology – has played in advancing such understanding, thanks to one- and two-dimensional abstractions and efficient techniques to predict them that are applicable on a large scale, such as machine learning and deep learning in particular. The typical pipeline to predict protein structure annotations was also presented, highlighting the key tools adopted and their characteristics.

The chapter then described the main one- and two-dimensional protein structure annotations, from their definition to samples of state-of-the-art methods to predict them. We have given a concise introduction to each protein structure annotation trying to highlight what, why and how is predicted. We also tried to give a sense of how different abstractions are linked to one another and how this is reflected in the systems that predict them.

A considerable part of this chapter is dedicated to presenting, describing and comparing state-of-the-art predictors of protein structure annotations. The methods presented are typically available as both web servers and standalone programs and, thus, can be used for small- or large-scale experiments and studies. The general aim of this chapter is to introduce and facilitate the adoption of in silico methods to study proteins by the broader research community.

References

Adhikari B, Hou J, Cheng J (2017) DNCON2: improved protein contact prediction using two-level deep convolutional neural networks. Bioinformatics 34(9):1466–1472

Ahmad S, Gromiha M, Fawareh H, Sarai A (2004) ASAView: database and tool for solvent accessibility representation in proteins. BMC Bioinformatics 5:51

Aloy P, Stark A, Hadley C, Russell RB (2003) Predictions without templates: new folds, secondary structure, and contacts in CASP5. Proteins Struct Funct Bioinforma 53(S6):436–456

Altschul SF et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402

Baldi P, Brunak S, Frasconi P, Soda G, Pollastri G (1999) Exploiting the past and the future in protein secondary structure prediction. Bioinforma Oxf Engl 15(11):937–946

Bartoli L, Capriotti E, Fariselli P, Martelli PL, Casadio R (2008) The pros and cons of predicting protein contact maps. Methods Mol Biol Clifton NJ 413:199–217

Baú D, Martin AJ, Mooney C, Vullo A, Walsh I, Pollastri G (2006) Distill: a suite of web servers for the prediction of one-, two- and three-dimensional structural features of proteins. BMC Bioinformatics 7:402

Berman HM et al (2000) The Protein Data Bank. Nucleic Acids Res 28(1):235–242

Buchan DWA, Jones DT (2018) Improved protein contact predictions with the MetaPSICOV2 server in CASP12. *Proteins* 86(Suppl 1):78–83

Buchan DWA, Ward SM, Lobley AE, Nugent TCO, Bryson K, Jones DT (2010) Protein annotation and modelling servers at University College London. *Nucleic Acids Res* 38(suppl_2):W563–W568

Bystroff C, Thorsson V, Baker D (2000) HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. J Mol Biol 301(1):173–190

Cheng J, Baldi P (2007) Improved residue contact prediction using support vector machines and a large feature set. BMC Bioinformatics 8:113

Cheng J, Randall AZ, Sweredoski MJ, Baldi P (2005) SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Res* 33(suppl_2):W72–W76

Chou PY, Fasman GD (1974) Prediction of protein conformation. Biochemistry (Mosc) 13(2):222–245

Cornette JL, Cease KB, Margalit H, Spouge JL, Berzofsky JA, DeLisi C (1987) Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. J Mol Biol 195(3):659–685

Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ (1998) JPred: a consensus secondary structure prediction server. Bioinformatics 14(10):892–893

De Brevern AG, Etchebest C, Hazout S (2000) Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. Proteins Struct Funct Bioinforma 41(3):271–287

Di Lena P, Fariselli P, Margara L, Vassura M, Casadio R (2010) Fast overlapping of protein contact maps by alignment of eigenvectors. Bioinformatics 26(18):2250–2258

Di Lena P, Fariselli P, Margara L, Vassura M, Casadio R (2011) Is there an optimal substitution matrix for contact prediction with correlated mutations? IEEEACM Trans Comput Biol Bioinforma 8(4):1017–1028

Di Lena P, Nagata K, Baldi P (2012) Deep architectures for protein contact map prediction. Bioinformatics 28(19):2449–2457

Drozdetskiy A, Cole C, Procter J, Barton GJ (2015) JPred4: a protein secondary structure prediction server. Nucleic Acids Res 43(W1):W389–W394

Eickholt J, Cheng J (2012) Predicting protein residue–residue contacts using deep networks and boosting. Bioinformatics 28(23):3066–3072

Faraggi E, Yang Y, Zhang S, Zhou Y (2009) Predicting continuous local structure and the effect of its substitution for secondary structure in fragment-free protein structure prediction. Structure 17(11):1515–1527

Fariselli P, Olmea O, Valencia A, Casadio R (2001) Prediction of contact maps with neural networks and correlated mutations. Protein Eng Des Sel 14(11):835–843

Fauchère JL, Charton M, Kier LB, Verloop A, Pliska V (1988) Amino acid side chain parameters for correlation studies in biology and pharmacology. Int J Pept Protein Res 32(4):269–278

Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* 39(Web Server issue):W29–W37

Göbel U, Sander C, Schneider R, Valencia A (1994) Correlated mutations and residue contacts in proteins. Proteins 18(4):309–317

Haas J et al Continuous Automated Model Evaluation (CAMEO) complementing the critical assessment of structure prediction in CASP12. Proteins: Struct Funct Bioinf p. n/a-n/a

Heffernan R et al (2015) Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. *Sci Rep* 5:11476

Heffernan R et al (2016) Highly accurate sequence-based prediction of half-sphere exposures of amino acid residues in proteins. Bioinformatics 32(6):843–849

Heffernan R, Yang Y, Paliwal K, Zhou Y (2017) Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility. Bioinformatics 33(18):2842–2849

Holbrook SR, Muskal SM, Kim SH (1990) Predicting surface exposure of amino acids from protein sequence. Protein Eng 3(8):659–665

Huang Y, Bystroff C (2006) Improved pairwise alignments of proteins in the Twilight Zone using local structure predictions. Bioinformatics 22(4):413–422

Johnson LS, Eddy SR, Portugaly E (2010) Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics 11:431

Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292(2):195–202

Jones DT, Swindells MB (2002) Getting the most from PSI–BLAST. Trends Biochem Sci 27(3):161–164

Jones DT, Buchan DWA, Cozzetto D, Pontil M (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 28(2):184–190

Jones DT, Singh T, Kosciolek T, Tetchner S (2015) MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. Bioinformatics 31(7):999–1006

Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22(12):2577–2637

Kaján L, Hopf TA, Kalaš M, Marks DS, Rost B (2014) FreeContact: fast and free software for protein contact prediction from residue co-evolution. BMC Bioinformatics 15:85

Kendrew JC et al (1960) Structure of myoglobin: a three-dimensional Fourier synthesis at 2 A. resolution. Nature 185(4711):422–427

Kim DE, DiMaio F, Wang RY-R, Song Y, Baker D (2014) One contact for every twelve residues allows robust and accurate topology-level protein structure modeling. *Proteins* 82(2):208–218

Kinch LN, Li W, Monastyrskyy B, Kryshtafovych A, Grishin NV (2016) Assessment of CASP11 contact-assisted predictions. Proteins 84(Suppl 1):164–180

Kosciolek T, Jones DT (2014) De Novo structure prediction of globular proteins aided by sequence variation-derived contacts. PLoS One 9(3):e92197

Kosciolek T, Jones DT (2016) Accurate contact predictions using covariation techniques and machine learning. *Proteins* 84(Suppl 1):145–151

Kuang R, Leslie CS, Yang A-S (2004) Protein backbone angle prediction with machine learning approaches. Bioinformatics 20(10):1612–1621

Kukic P, Mirabello C, Tradigo G, Walsh I, Veltri P, Pollastri G (2014) Toward an accurate prediction of inter-residue distances in proteins using 2D recursive neural networks. BMC Bioinformatics 15:6

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444

Lyons J et al (2014) Predicting backbone Cα angles and dihedrals from protein sequences by stacked sparse auto-encoder deep neural network. J Comput Chem 35(28):2040–2046

MacCallum RM (2004) Striped sheets and protein contact prediction. *Bioinformatics* 20(suppl_1):i224–i231

Magnan CN, Baldi P (2014) SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. Bioinformatics 30(18):2592–2597

Martin J, Letellier G, Marin A, Taly J-F, de Brevern AG, Gibrat J-F (2005) Protein secondary structure assignment revisited: a detailed analysis of different assignment methods. BMC Struct Biol 5:17

Martin AJ, Mooney C, Walsh I, Pollastri G (2010) Contact map prediction by machine learning. In: Pan Y, Zomaya A, Rangwala H, Karypis G (eds) Introduction to protein structure prediction. Wiley. https://doi.org/10.1002/9780470882207.ch7

Mirabello C, Pollastri G (2013) Porter, PaleAle 4.0: high-accuracy prediction of protein secondary structure and relative solvent accessibility. Bioinformatics 29(16):2056–2058

Monastyrskyy B, D'Andrea D, Fidelis K, Tramontano A, Kryshtafovych A (2014) Evaluation of residue–residue contact prediction in CASP10. Proteins Struct Funct Bioinforma 82:138–153

Monastyrskyy B, D'Andrea D, Fidelis K, Tramontano A, Kryshtafovych A (2016) New encouraging developments in contact prediction: Assessment of the CASP11 results. *Proteins* 84(Suppl 1):131–144

Mooney C, Pollastri G (2009) Beyond the twilight zone: automated prediction of structural properties of proteins by recursive neural networks and remote homology information. Proteins Struct Funct Bioinforma 77(1):181–190

Mooney C, Vullo A, Pollastri G (2006) Protein structural motif prediction in multidimensional ø-ψ space leads to improved secondary structure prediction. J Comput Biol 13(8):1489–1502

Mooney C, Cessieux A, Shields DC, Pollastri G (2013) SCL-Epred: a generalised de novo eukaryotic protein subcellular localisation predictor. Amino Acids 45(2):291–299

Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247(4):536–540

Olmea O, Valencia A (1997) Improving contact predictions by the combination of correlated mutations and other sources of sequence information. Fold Des 2:S25–S32

Pascarella S, Persio RD, Bossa F, Argos P (1998) Easy method to predict solvent accessibility from multiple protein sequence alignments. Proteins Struct Funct Bioinforma 32(2):190–199

Pauling L, Corey RB (1951) Configurations of polypeptide chains with favored orientations around single bonds. Proc Natl Acad Sci U S A 37(11):729–740

Pazos F, Helmer-Citterich M, Ausiello G, Valencia A (1997) Correlated mutations contain information about protein-protein interaction. J Mol Biol 271(4):511–523

Perutz MF, Rossmann MG, Cullis AF, Muirhead H, Will G, North AC (1960) Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-A. resolution, obtained by X-ray analysis. Nature 185(4711):416–422

Pollastri G, Baldi P (2002) Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners. *Bioinformatics* 18(suppl_1):S62–S70

Pollastri G, McLysaght A (2005) Porter: a new, accurate server for protein secondary structure prediction. Bioinformatics 21(8):1719–1720

Pollastri G, Baldi P, Fariselli P, Casadio R (2001) Improved prediction of the number of residue contacts in proteins by recurrent neural networks. *Bioinformatics* 17(suppl_1):S234–S242

Pollastri G, Baldi P, Fariselli P, Casadio R (2002) Prediction of coordination number and relative solvent accessibility in proteins. Proteins 47(2):142–153

Pollastri G, Martin AJ, Mooney C, Vullo A (2007) Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information. BMC Bioinformatics 8:201

Remmert M, Biegert A, Hauser A, Söding J (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods 9(2):173–175

Rost B (2001) Review: protein secondary structure prediction continues to rise. J Struct Biol 134(2):204–218

Rost B, Sander C (1993) Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 232(2):584–599

Rost B, Sander C (1994) Conservation and prediction of solvent accessibility in protein families. Proteins 20(3):216–226

Roy A, Kucukural A, Zhang Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 5(4):725–738

Schaarschmidt J, Monastyrskyy B, Kryshtafovych A, Bonvin AMJJ (2018) Assessment of contact predictions in CASP12: co-evolution and deep learning coming of age. Proteins Struct Funct Bioinforma 86:51–66

Schäffer AA et al (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res 29(14):2994–3005

Schlessinger A, Punta M, Rost B (2007) Natively unstructured regions in proteins identified from contact predictions. Bioinforma Oxf Engl 23(18):2376–2384

Schmidhuber J (2015) Deep learning in neural networks: an overview. Neural Netw 61:85–117

Seemayer S, Gruber M, Söding J (2014) CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. Bioinformatics 30(21):3128–3130

Sims GE, Choi I-G, Kim S-H (2005) Protein conformational space in higher order ϕ-Ψ maps. Proc Natl Acad Sci U S A 102(3):618–621

Tegge AN, Wang Z, Eickholt J, Cheng J (2009) NNcon: improved protein contact map prediction using 2D-recursive neural networks. *Nucleic Acids Res* 37(suppl_2):W515–W518

The UniProt Consortium (2016) UniProt: the universal protein knowledgebase. Nucleic Acids Res 45(D1):D158–D169

Thomas H (2005) An amino acid has two sides: a new 2D measure provides a different view of solvent exposure. Proteins Struct Funct Bioinforma 59(1):38–48

Ting D, Wang G, Shapovalov M, Mitra R, Jordan MI, Jr RLD (2010) Neighbor-dependent Ramachandran probability distributions of amino acids developed from a Hierarchical Dirichlet process model. *PLoS Comput Biol* 6(4):e1000763

Torrisi M, Kaleel M, Pollastri G (2018) Porter 5: state-of-the-art ab initio prediction of protein secondary structure in 3 and 8 classes. *bioRxiv*:289033

Vassura M, Margara L, Di Lena P, Medri F, Fariselli P, Casadio R (2008) Reconstruction of 3D structures from protein contact maps. IEEEACM Trans Comput Biol Bioinforma 5(3):357–367

Vassura M et al (2011) Blurring contact maps of thousands of proteins: what we can learn by reconstructing 3D structure. BioData Min 4:1

Vendruscolo M, Kussell E, Domany E (1997) Recovery of protein structure from contact maps. Fold Des 2(5):295–306

Vullo A, Walsh I, Pollastri G (2006) A two-stage approach for improved prediction of residue contact maps. BMC Bioinformatics 7:180

Walsh I, Baù D, Martin AJ, Mooney C, Vullo A, Pollastri G (2009) Ab initio and template-based prediction of multi-class distance maps by two-dimensional recursive neural networks. BMC Struct Biol 9:5

Walsh I, Pollastri G, Tosatto SCE (2016) Correct machine learning on protein sequences: a peer-reviewing perspective. Brief Bioinform 17(5):831–840

Wang S, Li W, Liu S, Xu J (2016) RaptorX-property: a web server for protein structure property prediction. Nucleic Acids Res 44(W1):W430–W435

Wang S, Sun S, Li Z, Zhang R, Xu J (2017) Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Comput Biol* 13(1):e1005324

Wang S, Sun S, Xu J (2018) Analysis of deep learning methods for blind protein contact prediction in CASP12. *Proteins* 86(Suppl 1):67–77

Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ (2009) Jalview version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics 25(9):1189–1191

Wood MJ, Hirst JD (2005) Protein secondary structure prediction with dihedral angles. Proteins Struct Funct Bioinforma 59(3):476–481

Xia L, Pan X-M (2000) New method for accurate prediction of solvent accessibility from protein sequence. Proteins Struct Funct Bioinforma 42(1):1–5

Yang Y, Faraggi E, Zhao H, Zhou Y (2011) Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. Bioinformatics 27(15):2076–2082

Yang Y et al (2016) Sixty-five years of the long march in protein secondary structure prediction: the final stretch? Brief Bioinform 19(3):482–494

Yuan Z (2005) Better prediction of protein contact number using a support vector regression analysis of amino acid sequence. BMC Bioinformatics 6:248

Zemla A (2003) LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Res 31(13):3370–3374

Zemla A, Venclovas Č, Fidelis K, Rost B (1999) A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. Proteins Struct Funct Bioinforma 34(2):220–223

Chapter 11
Introduction to Functional Bioinformatics

Peter Natesan Pushparaj

Contents

Abbreviations

AE	ArrayExpress
DAVID	Database for Annotation, Visualization and Integrated Discovery
DEGs	Differentially Expressed Genes
GEO	Gene Expression Omnibus
GO	Gene Ontology

P. N. Pushparaj (✉)
Center of Excellence in Genomic Medicine Research, Faculty of Applied Medical Sciences,
King Abdulaziz University, Jeddah, Saudi Arabia

N. A. Shaik et al. (eds.), *Essentials of Bioinformatics, Volume I*,
https://doi.org/10.1007/978-3-030-02634-9_11

GSEA	Gene Set Enrichment Analysis
HPC	High-Performance Computing
IPA	Ingenuity Pathway Analysis
KEGG	Kyoto Encyclopedia of Genes and Genomes
NGS	Next Generation Sequencing
RMA	Robust Multi-array Average
RNAseq	RNA Sequencing
TAC	Transcriptome Analysis Console

11.1 Introduction

The generation of the large scale of biomedical data in the past decade in genomics, proteomics, metabolomics and other “Omics” approaches and the parallel development of innovative computing methodologies have immensely transformed our basic understanding of biology and medicine (Hieter and Boguski 1997). Functional bioinformatics is a subarea of computational biology that utilizes the massive amount of data derived from genomics, transcriptomics, proteomics, glycomics, lipidomics, metabolomics and other large-scale “omics” experiments in interrelated areas, to decipher the complex gene and protein functions and interactions in both health and disease (Fig. 11.1). The number of publications in the area of functional bioinformatics increased significantly from less than 100 in the year 1998 to about 38,000 till the middle of July 2018 (Fig. 11.2).

The whole set of DNA found in each cell is defined as the genome. Each cell contains a complete copy of the genome, distributed along chromosomes (compressed and entwined DNA). About 3.2×10^9 base pairs (*3 billion base pairs*) in the human DNA, which is about 6 feet (2 metres) in each cell if stretched out as a thin thread, encode blueprint for all cellular structures, functions and other activities (Miyaoka et al. 2016). Though the explosion of high-throughput data in molecular biology has befell in the past decade to decode the central dogma (Fig. 11.3), our functional understanding of cellular and molecular processes requires integrated analyses of heterogeneous data sources with robust computational methods.

In functional genomics, the roles of genes are determined using high-throughput technologies such as the microarrays, next-generation sequencing approaches, etc. It decodes how genomes, proteomes and metabolomes result in different cellular phenotypes and analyses differences in how the same genome functions differently in diverse cell types and how changes in genomes alter both cellular and molecular functions through differential expression of transcripts or genes (DEGs) which in turn regulate the expression of proteins and metabolites in the cells (Fig. 11.4). An array of computational tools are used in functional bioinformatics approaches to decipher complex biological information in diverse datasets to generate precise biological understanding and hypotheses about gene functions, protein expression, interactions and regulations in health and disease.

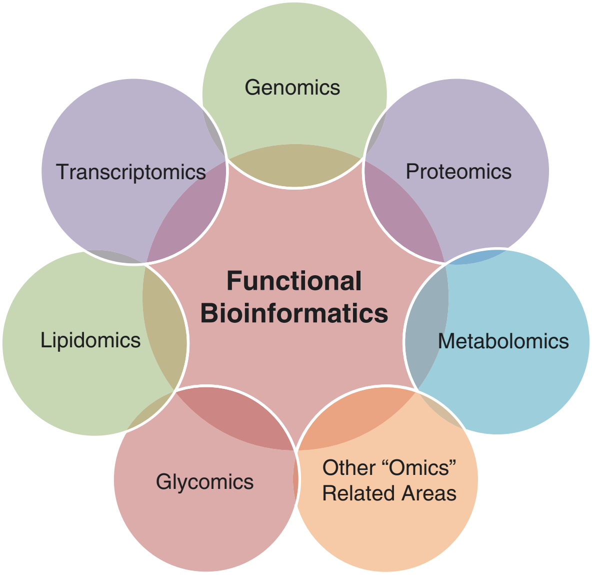


Fig. 11.1 The central paradigm of functional bioinformatics. The functional analyses of genes, proteins, and metabolites using an array of open source and commercial Bioinformatics tools, in contrast to other bioinformatics techniques such as Sequence and Structural Analyses, can be termed as *Functional Bioinformatics*. It is used for molecular profile analysis, pattern detection in multi-variate data, detection of biologically relevant signatures, network analyses etc. Molecular profiles are deciphered from complex data derived from an array of high-throughput "Omics" platforms for their correlation and functional relationships in both health and disease

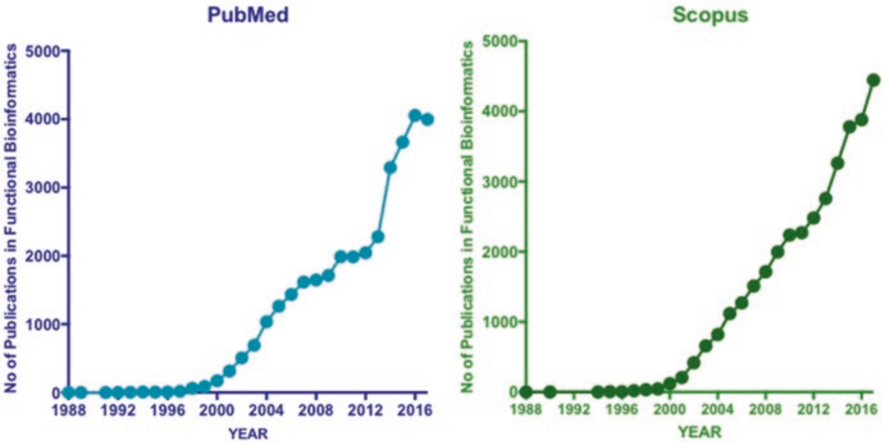


Fig. 11.2 The number of publications in functional bioinformatics. The bibliometric analyses using the largest bibliometric database, Scopus has shown that about 30 documents were published related to the Functional Bioinformatics in the year 1998 and it increased drastically in the past 20 years. In total, 12, 997 publications were published in the area of Functional Bioinformatics till the middle of 2018

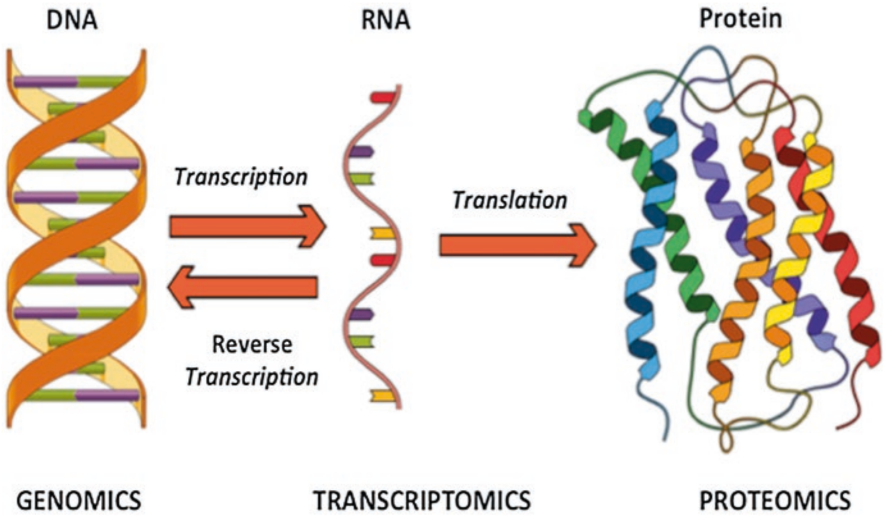


Fig. 11.3 The central dogma of molecular biology

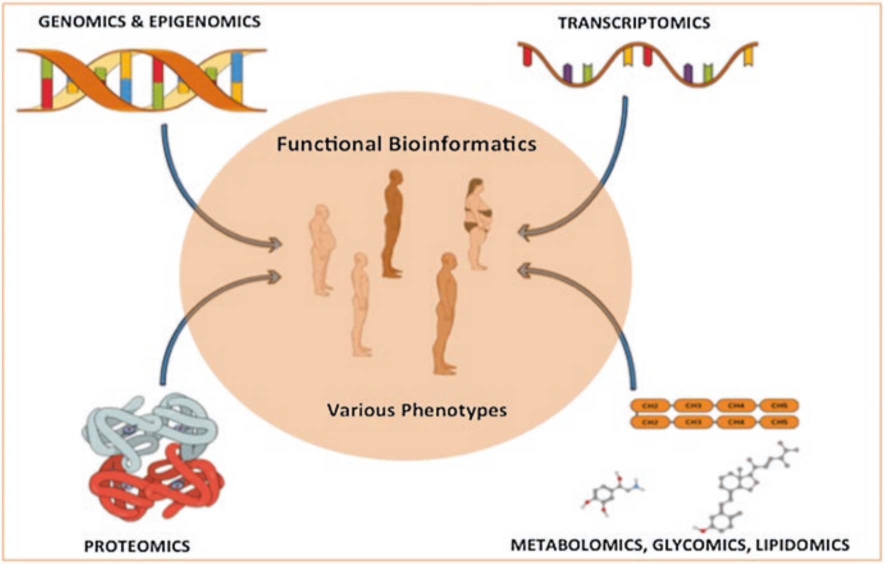


Fig. 11.4 Functional genomics and functional bioinformatics paradigm. Functional Genomics is the study of how the genome, transcripts (genes), proteins and metabolites work together to produce a particular phenotype. Together, transcriptomics, proteomics and metabolomics describe the transcripts, proteins and metabolites of a biological system, and the Functional Bioinformatics analyses and the integration of these processed data is expected to provide a complete model of the biological system under study

11.2 Techniques in Functional Genomics

Techniques used in functional genomics range from low-throughput techniques such as real-time quantitative PCR (SYBR Green and TaqMan methods), digital PCR (dPCR) (Didelot et al. 2013; Zhong et al. 2011) and serial analysis of gene expression (SAGE) to high-throughput technologies such as microarrays and next-generation sequencing technologies (mainly RNASeq) (Sorlie et al. 2006; Wang et al. 2006). Functional genomics experiments measure changes in the genome, transcriptome, proteome, and metabolome (metabolites) or interactions between DNA/RNA/proteins and metabolites that significantly impact or modulate the phenotype of an individual or a biological sample. However, the functional genomics techniques are mainly used for transcription profiling, epigenetic profiling, nucleic acid-protein interactions and genotyping for single-nucleotide polymorphisms (SNP) in biological samples. In this chapter, I will be focusing on the high-throughput microarray technologies and functional bioinformatics strategies to decipher the complex "Omics" data.

11.3 Microarray Technology

Microarrays are made up of short oligonucleotide probes (DNA), which are evenly bound in defined positions onto a solid surface, such as a glass slide, onto which DNA fragments derived from the biological samples will be hybridized (Bunnik and Le Roch 2013). Importantly, the microarrays can further be classified into two types, namely, one-colour arrays (Affymetrix) and two-colour arrays (Agilent). In two-colour arrays, such as Agilent arrays, the oligonucleotides (*probes*) are coated onto the glass slides using inkjet printing technology (*Agilent uses highly efficient SurePrint Technolog*y), and in one-colour arrays (Affymetrix), the probes are synthesized in situ on a solid surface by photolithography. The single-stranded cDNA or antisense RNA molecules derived from biological samples are then hybridized to these DNA microarrays using stringent methods. The quantity of hybridization measured for each specific probe is directly proportional to the number of specific mRNA transcripts present in the biological samples. On the other hand, Illumina uses bead microarray technology that is based on 3-micron silica beads that self-assemble in micro wells either on the fibre-optic bundles or planar silica slides. The self-assembly of these beads has a uniform spacing of ~5.7 microns and is covered with several copies of a specific oligonucleotide acting as the capture sequences in the Illumina microarrays (Eijssen et al. 2015).

It is essential to decide whether to quantify the gene expression levels from each sample on separate microarrays (one-colour array such as Affymetrix) or to calculaterelative gene expression levels between a pair of samples on a single

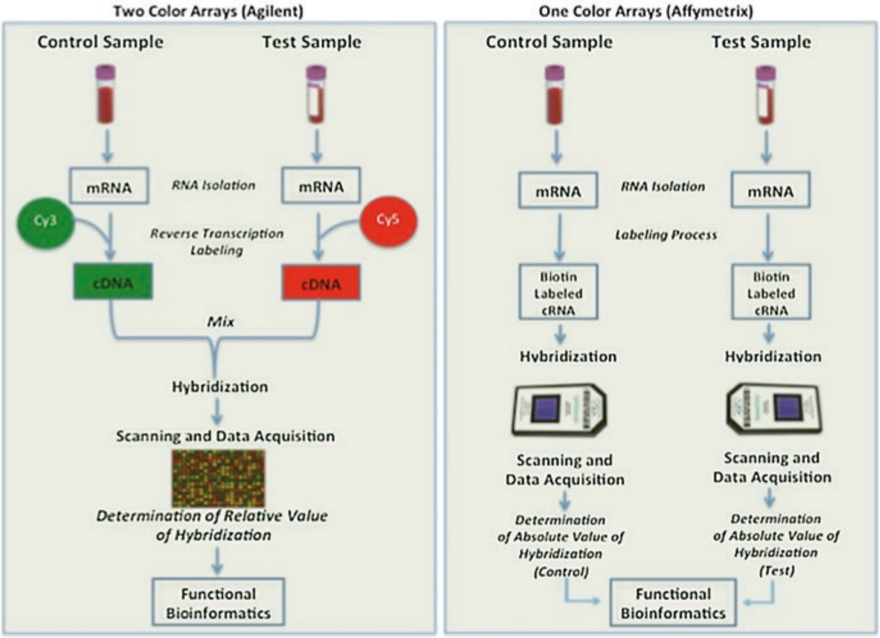

Two Color Arrays (Agilent)
Control Sample
Test Sample
mRNA
RNA Isolation
mRNA
Cy3
Reverse Transcription Labeling
Cy5
cDNA
cDNA
Mix
Hybridization
Scanning and Data Acquisition
Determination of Relative Value of Hybridization
Functional Bioinformatics
One Color Arrays (Affymetrix)
Control Sample
Test Sample
mRNA
RNA Isolation
mRNA
Labeling Process
Biotin Labeled cRNA
Biotin Labeled cRNA
Hybridization
Hybridization
Scanning and Data Acquisition
Scanning and Data Acquisition
Determination of Absolute Value of Hybridization (Control)
Determination of Absolute Value of Hybridization (Test)
Functional Bioinformatics

Fig. 11.5 The comparison of two-colour and one-colour microarray platforms used in functional genomics. In two colour microarrays (Agilent), two biological samples (experimental/test sample and control/reference sample) are labelled with different fluorescent dyes, usually Cyanine 3 (Cy3) and Cyanine 5 (Cy5). Equal amounts of labelled cDNA are then simultaneously hybridized to the same microarray chip. After this competitive hybridization, the fluorescence measurements are made separately for each dye and represent the abundance of each gene in one sample (test sample, Cy5) relative to the other one (control sample, Cy3). The hybridization data are reported as a ratio of the Cy5/Cy3 fluorescent signals at each probe. By contrast, in one colour microarrays (Affymetrix), each sample is labelled and hybridized to a separate microarray and we get an absolute value of fluorescence for each probe

microarray (two-colour array such as Agilent) (Fig. 11.5). Importantly, the efficiency of both one-colour and two-colour arrays is almost identical (Sorlie et al. 2006; Wang et al. 2006).

To obtain sufficient statistical power, a minimum of three replicates is recommended in microarray experiments. Without replicates, the measurement of statistical significance and reliability of the experimental changes is not possible, since an increased number of both false-positive and false-negative errors will result in the detection of differentially expressed genes. There are two types of replicates used in microarray studies such as technical replicates and biological replicates (Grant et al. 2007). At least, three biological replicates are essential in repositories such as Expression Atlas to get sufficient statistical power to derive differentially expressed genes (Petryszak et al. 2014).

11.4 Functional Bioinformatics Analyses of Microarray Data

Microarrays can be used in many types of experiments including genotyping, epigenetics, translation profiling and gene expression profiling. However, microarray technologies are mainly used for gene expression profiling in the past decade. Mostly, both one- and two-colour microarrays are used for the quantification of gene expression in biological samples. The process of analysing gene expression data is similar for both types of microarrays and several predefined steps to derive the differentially expressed genes from the data (Fig. 11.6). The main steps used in the microarray data analysis pipeline are listed below.

1. Feature extraction (image processing)
2. Quality control (QC)
3. Normalization
4. Differential expression analyses
5. Biological interpretation of the results
6. Submission of data to a public repository

11.4.1 Feature Extraction (Image Processing)

Feature extraction or image processing is the process of reading the scanned images of the microarrays into computable values either in binary or text format and interpreting it with corresponding gene IDs, sample names and other related information for further downstream data analysis process. The feature extraction is normally done by the software provided by the microarray manufacturers, which creates a

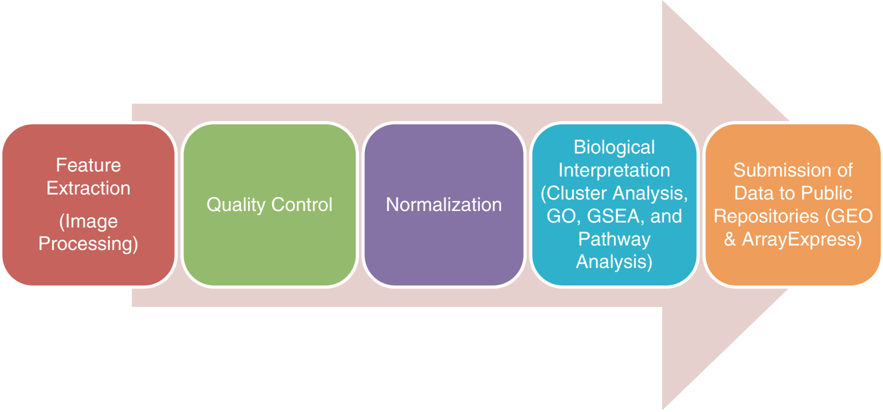

Feature Extraction
(Image Processing)
Quality Control
Normalization
Biological Interpretation (Cluster Analysis, GO, GSEA, and Pathway Analysis)
Submission of Data to Public Repositories (GEO & ArrayExpress)

Fig. 11.6 Microarray analysis pipeline. An overview of microarray data analyses pipeline commonly used in functional bioinformatics

binary file [CEL(Affymetrix), IDAT (Illumina)] or a text file (*Agilent*, *Illumina*, *NimbleGen arrays*, *custom-made cDNA-spotted arrays*, etc.,), at the end of the image-processing stage (Mehta and Rani 2011; Mehta 2011).

The next step is the downstream analysis of the binary or the text files obtained by image processing. There are varieties of microarray analysis software already available from various sources for the microarray data analysis (Koschmieder et al. 2012). However, either commercial software such as GeneSpring GX (Agilent, USA), Partek Genomics Suite 7.0 (Partek Inc., USA), etc. or free open source software such as the Transcriptome Analysis Console (TAC) software for the analysis of CEL files obtained from experiments using Affymetrix microarrays (Thermo Fisher Scientific, USA), GenePattern, R Studio (R packages “limma” and “oligo”), etc. (Mehta and Rani 2011; Mehta 2011) are most commonly used for this purpose.

11.4.2 Quality Control

The quality control (QC) is an important step in the analysis of microarray data. Initial QC step involves the visual inspection of scanned microarray images for any blank areas, local artefacts, marks, abrasions and other physical defects compared to normal scanned microarray images. The background signal, average intensity values and number, as well as the percentage of genes above the background signal, can be plotted for further inspection using the microarray analysis softwares to identify arrays with problems and eliminate these arrays from the analysis.

11.4.3 Normalization

The normalization of microarray data has been performed immediately after the feature extraction or image analysis before commencing the analysis of microarray data to control the technical variations between the arrays without affecting the biological variations (Quackenbush 2002; Irizarry et al. 2003a). The normalization of microarray data is essential to eliminate systematic bias such as sample preparation, spatial effects, variation in hybridisation, bias in the experiments and scanner settings, etc. An array of methods is used to normalize the microarray data, and it primarily depends on the experimental design, type of arrays, type of normalization algorithm and null hypothesis about the differentially expressed genes (Quackenbush 2002). Quantile normalization is used for one-colour arrays (Affymetrix), and loess normalization is used for two-colour arrays (Agilent) (Irizarry et al. 2003b). The Robust Multi-array Average (RMA) algorithm is used to normalize Affymetrix and NimbleGen arrays using “oligo” package in Bioconductor, and the Agilent arrays are mostly normalized using “limma” package in Bioconductor (Irizarry et al. 2003a, b).

11.4.4 Differential Expression Analysis

The analysis of differential expression of genes (DEGs) in a given sample is a comparative measurement related to the corresponding control, different treatments, disease states and so on (Hochberg and Benjamini 1990; Klipper-Aurbach et al. 1995; Tamhane et al. 1996). Multiple comparison procedure (MCP) or multiple testing is a statistical occurrence when the comparison of DEGs across multiple conditions is performed for a small number of samples (most of the microarray experiments involve less than five biological or technical replicates per condition). The multiple testing of DEGs across an array of conditions might lead to false-positive results. Hence, the multiple testing correction is essential when measuring the DEGs (Hochberg and Benjamini 1990; Klipper-Aurbach et al. 1995; Tamhane et al. 1996). The multiple testing correction is done by measuring the log2 fold change ratio between the test and control conditions, and a corrected p-value will also be calculated to identify the statistical significance in many open source and commercial software used to analyse microarray data (Hochberg and Benjamini 1990; Klipper-Aurbach et al. 1995; Tamhane et al. 1996).

11.4.5 Biological Interpretation of Gene Expression Data

The biological interpretation of the massive gene expression data is obtained using high-throughput techniques such as microarrays and RNAseq using heat maps and clustering, gene enrichment analysis, pathway analysis, etc.

11.4.5.1 Heat Maps and Clustering Algorithms

The generation of dendrograms or heat maps is the most common way of representing gene expression data from high-throughput techniques such as microarray (Quackenbush 2002). The heat map may also be combined with clustering methods which group genes and/or samples together based on the similarity of their gene expression pattern. This can be useful for identifying genes that are commonly regulated or biological signatures associated with a particular condition (e.g., a disease, drug treatment, an environmental condition, etc.). There are many open source software freely available for academic purposes such as Genesis provided by the Institute of Computational Biotechnology, Graz University of Technology, Austria, with easy-to-use graphical user interface (GUI), which can be used for the generation of heat maps as well as hierarchical clusters with single linkage, average linkage and complete linkage to deduce gene signatures (Fig. 11.7) (Quackenbush 2002). However, commercial software used for microarray data analysis such as GeneSpring (Agilent, USA), Partek Genomic Suite (Partek Inc., USA), etc. can also generate the heat maps and clusters from the gene expression data. In heat maps,

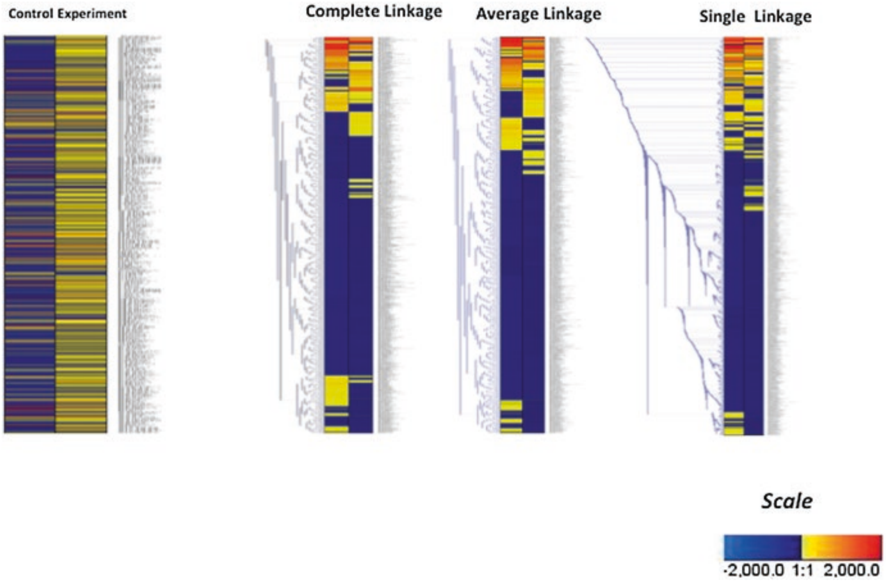


Fig. 11.7 Heatmaps and hierarchical clustering. (**a**) An example of a dendrogram (*Heat Map*) and hierarchical clustering of the control and experimental arrays. (**b–d**) Agglomerative hierarchical clustering of differentially expressed genes in the control and experimental arrays using single linkage, average linkage and complete linkage algorithms by Genesis software. Here, orange represents up-regulated genes and blue represents down-regulated genes and blank or grey represents unchanged expression

each row denotes a gene, and each column denotes a sample. The colour, as well as the intensity of each row (gene), varies based on the changes in the expression of the individual gene. In the example given in Fig. 11.7, orange represents up-regulated genes, blue represents down-regulated genes, and blank or grey represents the unchanged expression.

11.4.5.2 Gene Set Enrichment Analysis and Pathway Analysis

The Gene Ontology (GO) is a complete source of computable knowledge about the functions of genes and gene products, and it is used comprehensively in biomedical research specifically for the analysis of -omics and related data (Ashburner et al. 2000; Ashburner and Lewis 2002; Harris et al. 2004; The Gene Ontology C 2017). The gene set enrichment analysis (GSEA) based on the GO functional annotation of the differentially expressed genes to interpret the differentially expressed gene sets to decipher its association with a particular molecular or biological function or process, chromosomal location, etc. (Subramanian et al. 2005). The GSEA can be performed using the desktop application GSEA-P (Subramanian et al. 2007). The most common tools used for gene set enrichment and pathway enrichment analyses are the Database for Annotation, Visualization and Integrated Discovery (DAVID)

(Sherman and Lempicki 2009a, b; Huang et al. 2009), Ingenuity Pathway Analysis (IPA) (Abu-Elmagd et al. 2017; Kalamegam et al. 2015), Pathway Studio (Rahbar et al. 2017), Reactome (Fabregat et al. 2017), Kyoto Encyclopedia of Genes and Genomes (KEGG) (Yang et al. 2018), STRING (Yang et al. 2018), PathVisio (Kutmon et al. 2015; Fried et al. 2013), etc.

11.4.6 Submission of Data to a Public Repository

The microarray raw and the processed data generated from the experiments should be submitted along with metadata to any of the public repositories such as Gene Expression Omnibus (GEO), ArrayExpress (AE), etc. This is now a mandatory prerequisite to publish papers in journals and submit a progress report to funding bodies. Functional genomics data generated from microarray- and NGS-based methods are difficult to store, retrieve, maintain and transfer; hence, the large publically available repositories such as ArrayExpress and GEO are very important. The stored data in the public repositories are essential and economical since it can be easily retrieved for review, reanalysis and redistribution for the benefit of research and development and the ultimate benefit of mankind. However, the submitted microarray data should be MIAME (*Minimum Information About a Microarray Experiment*) complaint for archiving in these repositories (Brazma et al. 2001).

11.5 Applications and Limitations of Microarrays

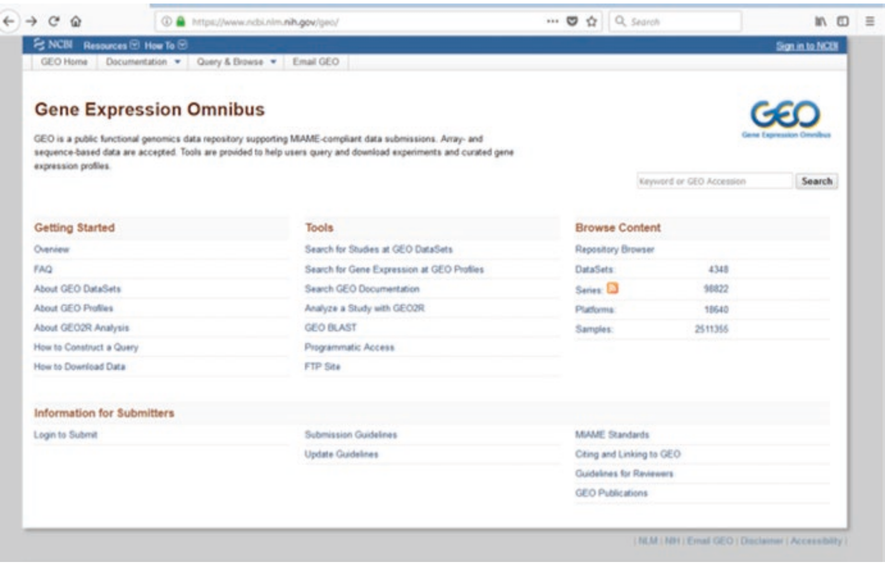
Apart from gene expression studies, microarrays can also be used to evaluate copy number variation, genotypes, epigenetic changes (DNA methylation (bisulfite method) and histone modifications) and DNA/RNA-protein interactions (DNA chip, RIP chip, and cross-linking and immunoprecipitation analysis (CLIP analysis)) in biological samples (Bunnik and Le Roch 2013). Even though the microarrays are comparatively inexpensive, it has limitations like background due to cross-hybridisation, prior knowledge about the genome sequence and lower dynamic range of the signal and requires complex normalization methods to compare different experiments.

11.6 Next-Generation Sequencing Technology and Functional Bioinformatics

Next-generation sequencing (NGS), on the other hand, does not need prior knowledge about the genome, and it is used to analyse DNA and RNA samples with a single nucleotide resolution. It is very easy to study alternatively spliced transcripts, allelic gene variants and single nucleotide polymorphisms (SNPs) with a higher dynamic range of the signal. Furthermore, NGS has higher reproducibility and needs less DNA/RNA concentration (nanograms) compared to the microarrays. RNA sequencing (RNAseq) is one of the applications of NGS to cDNA molecules. This is obtained by reverse transcription from RNA, to get information about the RNA content of a sample. Thus, RNAseq is the set of experimental procedures that generate cDNA molecules derived from RNA molecules, followed by sequencing library construction and massively parallel deep sequencing (Wirka et al. 2018). RNAseq is also used to study differential gene expression, alternative splicing events, allele-specific expression, expression quantitative trait loci analysis and fusion transcript detection (Wirka et al. 2018). In addition, single cells isolated by fluorescence-activated cell sorting (FACS) or magnetic-associated cell sorting (MACS) and other methods can be analysed by RNAseq (*single-cell transcriptomics*) to study differential gene expression, unique cellular processes, cellular diversity and heterogeneity in regenerative medicine, immunology, neurobiology and cardiovascular diseases (Wirka et al. 2018).

11.7 An Overview of Metanalysis Using Functional Bioinformatics Tools and Databases

Meta-analysis is a subarea of functional genomics where data derived from previous experiments may either be analyzed alone or combined with new data to create statistically robust models. The raw as well as processed data deposited in the functional genomics databases like Gene Expression Omnibus (GEO) (Fig. 11.8), ArrayExpress (AE) (Fig. 11.9), etc. can be used for meta-analyses of the high-throughput microarray and NGS data. For example, raw CEL files deposited in the GEO for a specific set of experiments can be analysed using Transcriptome Analysis Console (TAC) software, available for free download at the Thermo Fisher Scientific website. Similarly, several open source software and online resources such as GenePattern (Fig. 11.10) and ArrayAnalysis (Fig. 11.11) are available for the analysis of microarray data. On the other hand, commercial software such as GeneSpring (Agilent, USA), Partek Genomic Suite (Partek Inc., USA), etc. can be used for the analysis of microarray and RNAseq datasets. The differentially expressed genes can be analysed using free online databases such as DAVID for functional annotation and pathway enrichment analysis or using commercially available software such as IPA and Pathway Studio for the identification of differentially regulated pathways, gene networks, biological and toxicological functions, upstream regulators, etc. Furthermore, the role of a specific gene or a group of genes (up to 200) in a disease can be explored using the Open Targets target validation (Koscielny et al. 2017) database (https://www.targetvalidation.org) (Fig. 11.12), and the potential genetic variation and tissue expression can be further explored using the Genotype-Tissue Expression (GTEx) (Consortium 2013; Carithers and Moore 2015; Keen and Moore 2015; Lockhart et al. 2018; Siminoff et al. 2018) online portal (https://www.gtexportal.org) (Fig. 11.13). The regulatory elements, as well as the chromosomal



Gene Expression Omnibus

GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.

Getting Started
- Overview
- FAQ
- About GEO DataSets
- About GEO Profiles
- About GEO2R Analysis
- How to Construct a Query
- How to Download Data

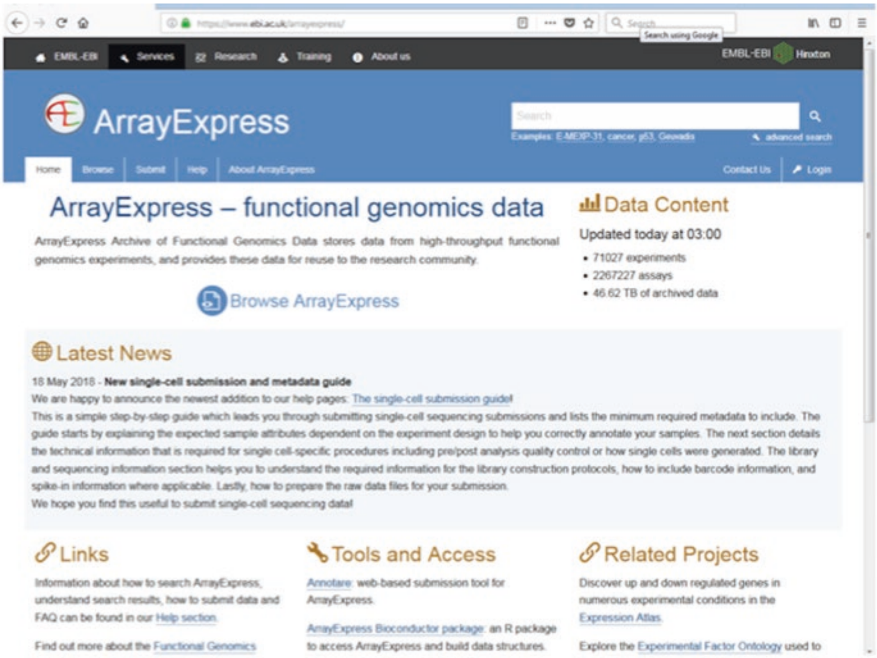
Tools
- Search for Studies at GEO DataSets
- Search for Gene Expression at GEO Profiles
- Search GEO Documentation
- Analyze a Study with GEO2R
- GEO BLAST
- Programmatic Access
- FTP Site

Browse Content

Repository Browser	
DataSets:	4348
Series:	98822
Platforms:	18640
Samples:	2511355

Information for Submitters
- Login to Submit
- Submission Guidelines
- Update Guidelines
- MIAME Standards
- Citing and Linking to GEO
- Guidelines for Reviewers
- GEO Publications

Fig. 11.8 Gene Expression Omnibus (GEO) web interface (https://www.ncbi.nlm.nih.gov/geo)



ArrayExpress – functional genomics data

ArrayExpress Archive of Functional Genomics Data stores data from high-throughput functional genomics experiments, and provides these data for reuse to the research community.

Data Content

Updated today at 03:00

- 71027 experiments
- 2267227 assays
- 46.62 TB of archived data

Latest News

18 May 2018 - **New single-cell submission and metadata guide**

We are happy to announce the newest addition to our help pages: The single-cell submission guide!

This is a simple step-by-step guide which leads you through submitting single-cell sequencing submissions and lists the minimum required metadata to include. The guide starts by explaining the expected sample attributes dependent on the experiment design to help you correctly annotate your samples. The next section details the technical information that is required for single cell-specific procedures including pre/post analysis quality control or how single cells were generated. The library and sequencing information section helps you to understand the required information for the library construction protocols, how to include barcode information, and spike-in information where applicable. Lastly, how to prepare the raw data files for your submission.

We hope you find this useful to submit single-cell sequencing data!

Links

Information about how to search ArrayExpress, understand search results, how to submit data and FAQ can be found in our Help section.

Find out more about the Functional Genomics

Tools and Access

Annotare: web-based submission tool for ArrayExpress.

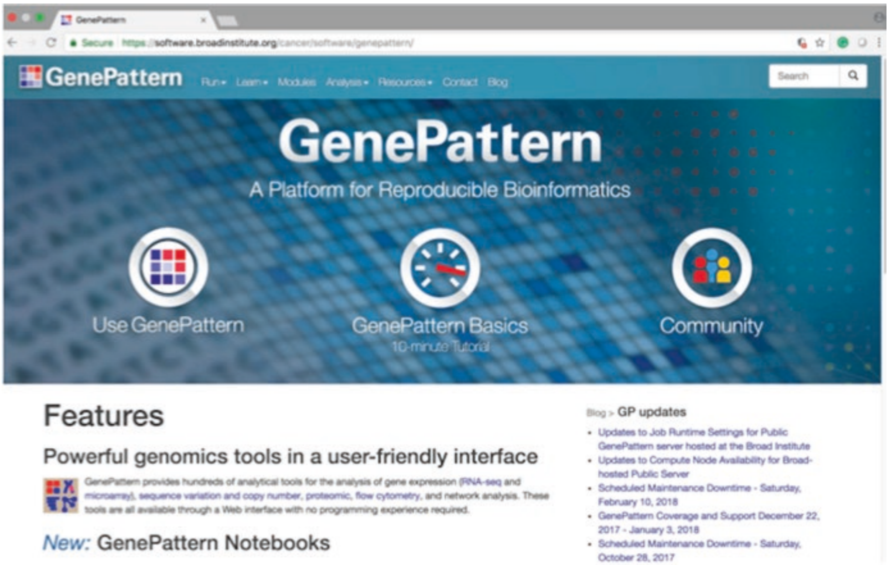
ArrayExpress Bioconductor package: an R package to access ArrayExpress and build data structures.

Related Projects

Discover up and down regulated genes in numerous experimental conditions in the Expression Atlas.

Explore the Experimental Factor Ontology used to

Fig. 11.9 Array express web interface (https://www.ebi.ac.uk/arrayexpress)


GenePattern
https://software.broadinstitute.org/cancer/software/genepattern/
GenePattern
Run▾ Learn▾ Modules Analysis▾ Resources▾ Contact Blog
Search
GenePattern
A Platform for Reproducible Bioinformatics
Use GenePattern
GenePattern Basics
10-minute Tutorial
Community

Features

Powerful genomics tools in a user-friendly interface

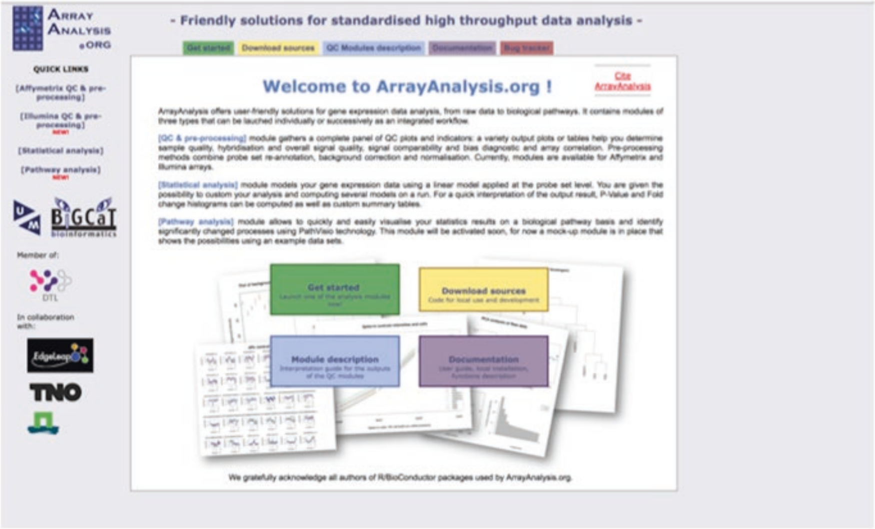
GenePattern provides hundreds of analytical tools for the analysis of gene expression (RNA-seq and microarray), sequence variation and copy number, proteomic, flow cytometry, and network analysis. These tools are all available through a Web interface with no programming experience required.

New: GenePattern Notebooks

Blog > GP updates

- Updates to Job Runtime Settings for Public GenePattern server hosted at the Broad Institute
- Updates to Compute Node Availability for Broad-hosted Public Server
- Scheduled Maintenance Downtime - Saturday, February 10, 2018
- GenePattern Coverage and Support December 22, 2017 - January 3, 2018
- Scheduled Maintenance Downtime - Saturday, October 28, 2017

Fig. 11.10 Gene pattern web interface. GenePattern provides an array of tools (https://software.broadinstitute.org/cancer/software/genepattern) for the analysis of RNA-seq and microarray data, copy number variation, network analysis etc.


ArrayAnalysis.org
QUICK LINKS
[Affymetrix QC & pre-processing]
[Illumina QC & pre-processing]
NEW!
[Statistical analysis]
[Pathway analysis]
NEW!
BiGCaT bioinformatics
Member of:
DTL
In collaboration with:
EdgeLeap
TNO
- Friendly solutions for standardised high throughput data analysis -
Get started
Download sources
QC Modules description
Documentation
Bug tracker
Cite ArrayAnalysis
Get started
Launch one of the analysis modules now!
Download sources
Code for local use and development
Module description
Interpretation guide for the outputs of the QC modules
Documentation
User guide, local installation, functions description

Welcome to ArrayAnalysis.org !

ArrayAnalysis offers user-friendly solutions for gene expression data analysis, from raw data to biological pathways. It contains modules of three types that can be lauched individually or successively as an integrated workflow.

[QC & pre-processing] module gathers a complete panel of QC plots and indicators: a variety output plots or tables help you determine sample quality, hybridisation and overall signal quality, signal comparability and bias diagnostic and array correlation. Pre-processing methods combine probe set re-annotation, background correction and normalisation. Currently, modules are available for Affymetrix and Illumina arrays.

[Statistical analysis] module models your gene expression data using a linear model applied at the probe set level. You are given the possibility to custom your analysis and computing several models on a run. For a quick interpretation of the output result, P-Value and Fold change histograms can be computed as well as custom summary tables.

[Pathway analysis] module allows to quickly and easily visualise your statistics results on a biological pathway basis and identify significantly changed processes using PathVisio technology. This module will be activated soon, for now a mock-up module is in place that shows the possibilities using an example data sets.

We gratefully acknowledge all authors of R/BioConductor packages used by ArrayAnalysis.org.

Fig. 11.11 Array analysis web interface. Array analysis web interface provides tools (http://www.arrayanalysis.org) for the analysis gene expression data, statistical and pathway analyses

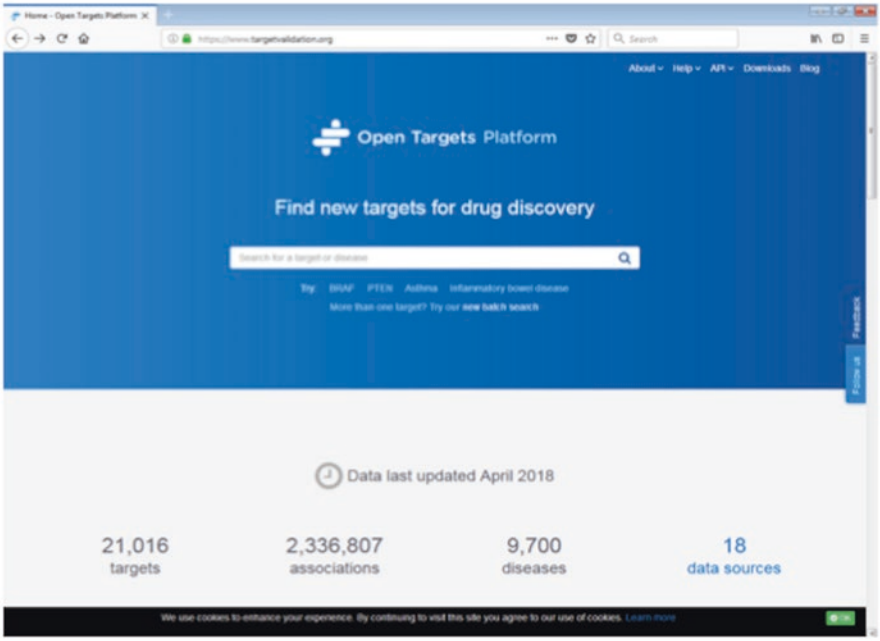

Home - Open Targets Platform
https://www.targetvalidation.org
About
Help
API
Downloads
Blog
Open Targets Platform
Find new targets for drug discovery
Search for a target or disease
Try: BRAF PTEN Asthma Inflammatory bowel disease
More than one target? Try our new batch search
Feedback
Follow us
Data last updated April 2018
21,016
targets
2,336,807
associations
9,700
diseases
18
data sources
We use cookies to enhance your experience. By continuing to visit this site you agree to our use of cookies. Learn more

Fig. 11.12 The open targets database. The open targets is an online tool (https://www.targetvalidation.org) for uncovering potential therapeutic targets and the association between these targets and an array of human diseases. Open Targets Tool can be used to explore the DEGs in a particular microarray or NGS experiment for the potential association with health and disease

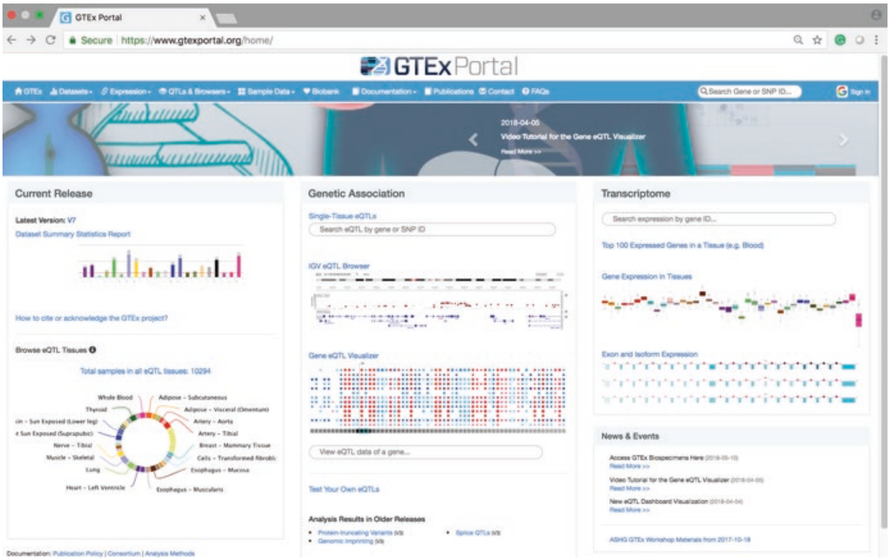
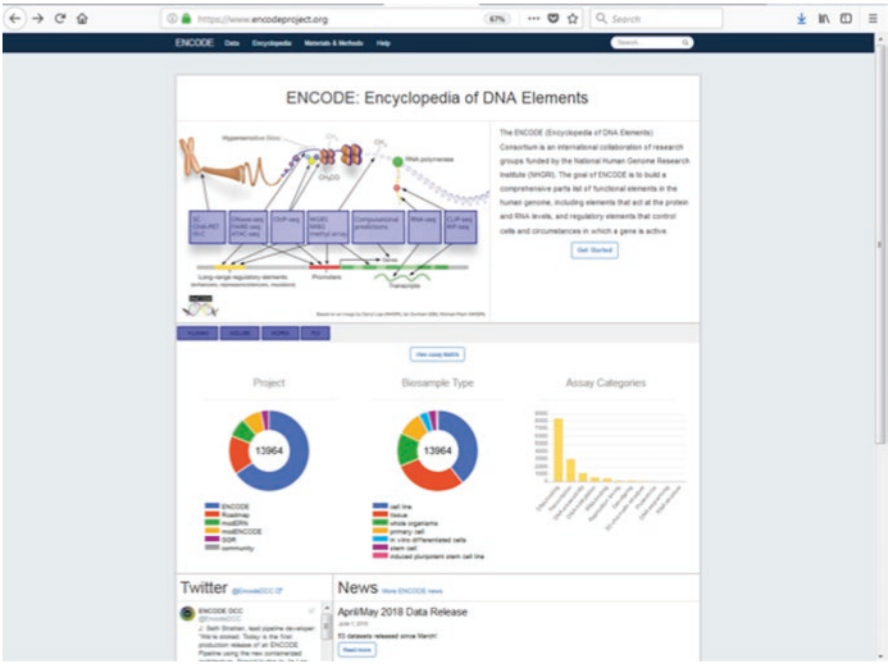

GTEx Portal
Secure
https://www.gtexportal.org/home/
GTExPortal
GTEx
Datasets
Expression
QTLs & Browsers
Sample Data
Biobank
Documentation
Publications
Contact
FAQs
Search Gene or SNP ID...
Sign in
2018-04-05
Video Tutorial for the Gene eQTL Visualizer
Read More >>
Current Release
Latest Version: V7
Dataset Summary Statistics Report
How to cite or acknowledge the GTEx project?
Browse eQTL Tissues
Total samples in all eQTL tissues: 10294
Genetic Association
Single-Tissue eQTLs
Search eQTL by gene or SNP ID
IGV eQTL Browser
Gene eQTL Visualizer
View eQTL data of a gene...
Test Your Own eQTLs
Analysis Results in Older Releases
Transcriptome
Search expression by gene ID...
Top 100 Expressed Genes in a Tissue (e.g. Blood)
Gene Expression in Tissues
Exon and Isoform Expression
News & Events
Access GTEx Biospecimens Here
Video Tutorial for the Gene eQTL Visualizer
New eQTL Dashboard Visualization
Read More >>
ASHG GTEx Workshop Materials from 2017-10-18

Fig. 11.13 Gene type-tissue expression (GTEx) web interface. GTEx is an online database (https://www.gtexportal.org) for finding the potential genetic variation and tissue expression of various genes in health and disease



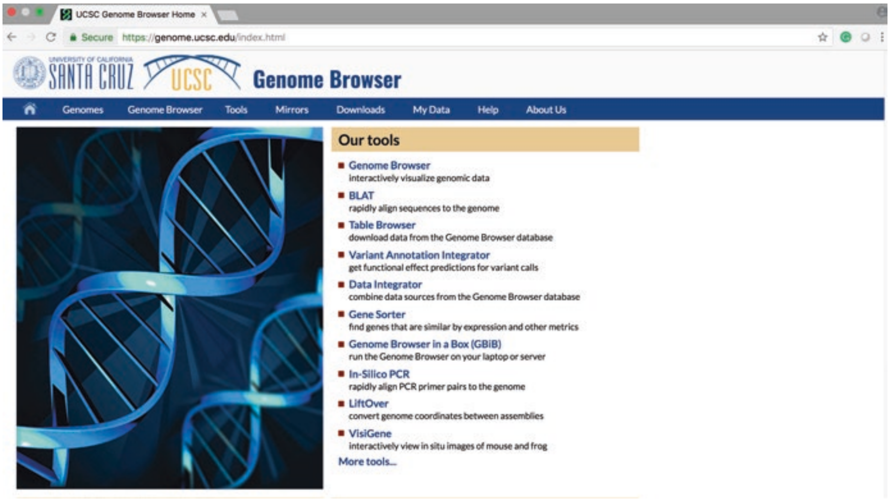
ENCODE: Encyclopedia of DNA Elements

The ENCODE (Encyclopedia of DNA Elements) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.

ENCODE DCC @EncodeDCC
J. Seth Strattan, lead pipeline developer: "We're stoked. Today is the first production release of an ENCODE Pipeline using the new containerized

April/May 2018 Data Release
June 1, 2018
53 datasets released since March!

Fig. 11.14 ENCODE-encyclopedia of DNA elements database. ENCODE is an online database (https://www.encodeproject.org) to unravel the regulatory elements at the protein and RNA level in health and disease


UCSC Genome Browser Home
Secure https://genome.ucsc.edu/index.html
UNIVERSITY OF CALIFORNIA SANTA CRUZ UCSC Genome Browser
Genomes Genome Browser Tools Mirrors Downloads My Data Help About Us

Our tools

- Genome Browser
  interactively visualize genomic data
- BLAT
  rapidly align sequences to the genome
- Table Browser
  download data from the Genome Browser database
- Variant Annotation Integrator
  get functional effect predictions for variant calls
- Data Integrator
  combine data sources from the Genome Browser database
- Gene Sorter
  find genes that are similar by expression and other metrics
- Genome Browser in a Box (GBiB)
  run the Genome Browser on your laptop or server
- In-Silico PCR
  rapidly align PCR primer pairs to the genome
- LiftOver
  convert genome coordinates between assemblies
- VisiGene
  interactively view in situ images of mouse and frog

More tools...

Fig. 11.15 University of California Santa Cruz (UCSC) Genome browser web interface (https://genome.ucsc.edu/)

location of a particular gene, can be elucidated in detail using the Encyclopedia of DNA Elements (ENCODE) database (Elnitski et al. 2007; Gerstein 2012; Lussier et al. 2013; Park et al. 2012; Rosenbloom et al. 2013; Ruau et al. 2013; Sloan et al. 2016; Wang et al. 2013) (https://www.encodeproject.org) (Fig. 11.14) and UCSC Genome Browser (https://genome.ucsc.edu/), respectively (Rosenbloom et al. 2013; Lee 2013; An et al. 2015; Casper et al. 2018; Hung and Weng 2016; Mangan et al. 2014; Speir et al. 2016; Tyner et al. 2017) (Fig. 11.15).

11.8 Conclusions

The type of techniques and the software tools used for the data analysis in functional genomics depend on the scale and objectives of the experiments. The analysis of the complex functional genomics data further depends on various factors such as the research funding and the availability of trained bioinformaticians. The research funding is necessary to purchase commercially available robust software and hardware and hire bioinformaticians with expertise in basic programming skills to utilize the enormous wealth of free and open software tools such as R and Bioconductor for the in-depth analysis of high-throughput functional genomics data. Besides, having a high-performance computing (HPC) facility in the universities and research institutes will be very convenient for the robust analysis and storage of a large amount of data generated from functional bioinformatics analysis.

Acknowledgements This work is funded by the National Plan for Science, Technology and Innovation (MAARIFAH)-King Abdulaziz City for Science and Technology, The Kingdom of Saudi Arabia, award number 12-BIO2719-03. The authors also acknowledge with thanks the Science and Technology Unit (STU), King Abdulaziz University, for their excellent technical support.

References

Abu-Elmagd M, Alghamdi MA, Shamy M, Khoder MI, Costa M, Assidi M et al (2017) Evaluation of the effects of airborne particulate matter on Bone Marrow-Mesenchymal Stem Cells (BM-MSCs): cellular, molecular and systems biological approaches. Int J Environ Res Public Health 14(4):440

An J, Lai J, Wood DL, Sajjanhar A, Wang C, Tevz G et al (2015) RNASeq browser: a genome browser for simultaneous visualization of raw strand specific RNAseq reads and UCSC genome browser custom tracks. BMC Genomics 16:145

Ashburner M, Lewis S (2002) On ontologies for biologists: the Gene Ontology--untangling the web. Novartis Found Symp 247:66–80; discussion −3, 4–90, 244–52

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM et al (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium Nat Genet 25(1):25–29

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C et al (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet 29(4):365–371

Bunnik EM, Le Roch KG (2013) An introduction to functional genomics and systems biology. Adv Wound Care (New Rochelle) 2(9):490–498

Carithers LJ, Moore HM (2015) The genotype-tissue expression (GTEx) project. Biopreserv Biobank 13(5):307–308

Casper J, Zweig AS, Villarreal C, Tyner C, Speir ML, Rosenbloom KR et al (2018) The UCSC Genome Browser database: 2018 update. Nucleic Acids Res 46(D1):D762–D7D9

Consortium GT (2013) The genotype-tissue expression (GTEx) project. Nat Genet 45(6):580–585

Didelot A, Kotsopoulos SK, Lupo A, Pekin D, Li X, Atochin I et al (2013) Multiplex picoliter-droplet digital PCR for quantitative assessment of DNA integrity in clinical samples. Clin Chem 59(5):815–823

Eijssen LM, Goelela VS, Kelder T, Adriaens ME, Evelo CT, Radonjic M (2015) A user-friendly workflow for analysis of Illumina gene expression bead array data available at the arrayanalysis.org portal. BMC Genomics 16:482

Elnitski LL, Shah P, Moreland RT, Umayam L, Wolfsberg TG, Baxevanis AD (2007) The ENCODEdb portal: simplified access to ENCODE consortium data. Genome Res 17(6):954–959

Fabregat A, Sidiropoulos K, Viteri G, Forner O, Marin-Garcia P, Arnau V et al (2017) Reactome pathway analysis: a high-performance in-memory approach. BMC Bioinformatics 18(1):142

Fried JY, van Iersel MP, Aladjem MI, Kohn KW, Luna A (2013) PathVisio-faceted search: an exploration tool for multi-dimensional navigation of large pathways. Bioinformatics 29(11):1465–1466

Gerstein M (2012) Genomics: ENCODE leads the way on big data. Nature 489(7415):208

Grant GR, Manduchi E, Stoeckert CJ Jr (2007) Analysis and management of microarray gene expression data. Curr Protoc Mol Biol Chapter 19:Unit 19 6

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R et al (2004) The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 32(Database issue):D258–D261

Hieter P, Boguski M (1997) Functional genomics: it's all how you read it. Science 278(5338):601–602

Hochberg Y, Benjamini Y (1990) More powerful procedures for multiple significance testing. Stat Med 9(7):811–818

Huang DW, Sherman BT, Zheng X, Yang J, Imamichi T, Stephens R et al (2009) Extracting biological meaning from large gene lists with DAVID. Curr Protoc Bioinformatics Chapter 13:Unit 13 1

Huang DW, Sherman BT, Lempicki RA (2009a) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4(1):44–57

Huang DW, Sherman BT, Lempicki RA (2009b) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37(1):1–13

Hung JH, Weng Z (2016) Visualizing genomic annotations with the UCSC Genome Browser. Cold Spring Harb Protoc 2016(11)

Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP (2003a) Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res 31(4):e15

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U et al (2003b) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4(2):249–264

Kalamegam G, Pushparaj PN, Khan F, Sait KH, Anfinan N, Al-Qahtani M (2015) Primary ovarian cancer cell inhibition by human Wharton's Jelly stem cells (hWJSCs): mapping probable mechanisms and targets using systems oncology. Bioinformation 11(12):529–534

Keen JC, Moore HM (2015) The genotype-tissue expression (GTEx) project: linking clinical data with molecular analysis to advance personalized medicine. J Pers Med 5(1):22–29

Klipper-Aurbach Y, Wasserman M, Braunspiegel-Weintrob N, Borstein D, Peleg S, Assa S et al (1995) Mathematical formulae for the prediction of the residual beta cell function during the first two years of disease in children and adolescents with insulin-dependent diabetes mellitus. Med Hypotheses 45(5):486–490

Koschmieder A, Zimmermann K, Trissl S, Stoltmann T, Leser U (2012) Tools for managing and analyzing microarray data. Brief Bioinform 13(1):46–60

Koscielny G, An P, Carvalho-Silva D, Cham JA, Fumis L, Gasparyan R et al (2017) Open targets: a platform for therapeutic target identification and validation. Nucleic Acids Res 45(D1):D985–DD94

Kutmon M, van Iersel MP, Bohler A, Kelder T, Nunes N, Pico AR et al (2015) PathVisio 3: an extendable pathway analysis toolbox. PLoS Comput Biol 11(2):e1004085

Lee R (2013) An introduction to the UCSC Genome Browser. WormBook:1–2

Lockhart NC, Weil CJ, Carithers LJ, Koester SE, Little AR, Volpi S et al (2018) Development of a consensus approach for return of pathology incidental findings in the genotype-tissue expression (GTEx) project. J Med Ethics 44:643

Lussier YA, Li H, Maienschein-Cline M (2013) Conquering computational challenges of omics data and post-ENCODE paradigms. Genome Biol 14(8):310

Mangan ME, Williams JM, Kuhn RM, Lathe WC 3rd (2014) The UCSC Genome Browser: what every molecular biologist should know. Curr Protoc Mol Biol 107:19 9 1–36

Mehta JP (2011) Microarray analysis of mRNAs: experimental design and data analysis fundamentals. Methods Mol Biol 784:27–40

Mehta JP, Rani S (2011) Software and tools for microarray data analysis. Methods Mol Biol 784:41–53

Miyaoka Y, Chan AH, Conklin BR (2016) Detecting single-nucleotide substitutions induced by genome editing. Cold Spring Harb Protoc 2016(8)

Park E, Williams B, Wold BJ, Mortazavi A (2012) RNA editing in the human ENCODE RNA-seq data. Genome Res 22(9):1626–1633

Petryszak R, Burdett T, Fiorelli B, Fonseca NA, Gonzalez-Porta M, Hastings E et al (2014) Expression Atlas update–a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. Nucleic Acids Res 42(Database issue):D926–D932

Quackenbush J (2002) Microarray data normalization and transformation. Nat Genet 32(Suppl):496–501

Rahbar S, Novin MG, Alizadeh E, Shahnazi V, Pashaei-Asl F, AsrBadr YA et al (2017) New insights into the expression profile of MicroRNA-34c and P53 in infertile men spermatozoa and testicular tissue. Cell Mol Biol (Noisy-le-Grand) 63(8):77–83

Rosenbloom KR, Sloan CA, Malladi VS, Dreszer TR, Learned K, Kirkup VM et al (2013) ENCODE data in the UCSC Genome Browser: year 5 update. Nucleic Acids Res 41(Database issue):D56–D63

Ruau D, Ng FS, Wilson NK, Hannah R, Diamanti E, Lombard P et al (2013) Building an ENCODE-style data compendium on a shoestring. Nat Methods 10(10):926

Siminoff LA, Wilson-Genderson M, Mosavel M, Barker L, Trgina J, Traino HM et al (2018) Impact of cognitive load on family decision Makers' recall and understanding of donation requests for the genotype-tissue expression (GTEx) project. J Clin Ethics 29(1):20–30

Sloan CA, Chan ET, Davidson JM, Malladi VS, Strattan JS, Hitz BC et al (2016) ENCODE data at the ENCODE portal. Nucleic Acids Res 44(D1):D726–D732

Sorlie T, Wang Y, Xiao C, Johnsen H, Naume B, Samaha RR et al (2006) Distinct molecular mechanisms underlying clinically relevant subtypes of breast cancer: gene expression analyses across three different platforms. BMC Genomics 7:127

Speir ML, Zweig AS, Rosenbloom KR, Raney BJ, Paten B, Nejad P et al (2016) The UCSC Genome Browser database: 2016 update. Nucleic Acids Res 44(D1):D717–D725

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA et al (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102(43):15545–15550

Subramanian A, Kuehn H, Gould J, Tamayo P, Mesirov JP (2007) GSEA-P: a desktop application for gene set enrichment analysis. Bioinformatics 23(23):3251–3253

Tamhane AC, Hochberg Y, Dunnett CW (1996) Multiple test procedures for dose finding. Biometrics 52(1):21–37

The Gene Ontology C (2017) Expansion of the gene ontology knowledgebase and resources. Nucleic Acids Res 45(D1):D331–D3D8

Tyner C, Barber GP, Casper J, Clawson H, Diekhans M, Eisenhart C et al (2017) The UCSC Genome Browser database: 2017 update. Nucleic Acids Res 45(D1):D626–DD34

Wang Y, Barbacioru C, Hyland F, Xiao W, Hunkapiller KL, Blake J et al (2006) Large scale real-time PCR validation on gene expression measurements from two commercial long-oligonucleotide microarrays. BMC Genomics 7:59

Wang J, Zhuang J, Iyer S, Lin XY, Greven MC, Kim BH et al (2013) Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. Nucleic Acids Res 41(Database issue):D171–D176

Wirka RC, Pjanic M, Quertermous T (2018) Advances in transcriptomics: investigating cardiovascular disease at unprecedented resolution. Circ Res 122(9):1200–1220

Yang X, Zhu S, Li L, Zhang L, Xian S, Wang Y et al (2018) Identification of differentially expressed genes and signaling pathways in ovarian cancer by integrated bioinformatics analysis. Onco Targets Ther 11:1457–1474

Zhong Q, Bhattacharya S, Kotsopoulos S, Olson J, Taly V, Griffiths AD et al (2011) Multiplex digital PCR: breaking the one target per color barrier of quantitative PCR. Lab Chip 11(13):2167–2174

Chapter 12
Biological Networks: Tools, Methods, and Analysis

Basharat Ahmad Bhat, Garima Singh, Rinku Sharma, Mifftha Yaseen, and Nazir Ahmad Ganai

Contents

B. A. Bhat (✉)
Department of Life Sciences, School of Natural Sciences, Shiv Nadar University,
Greater Noida, UP, India
e-mail: bb284@snu.edu.in

G. Singh · R. Sharma
Department of Life Sciences, Shiv Nadar University, Greater Noida, UP, India

M. Yaseen
School of Interdisciplinary Sciences and Technology, Jamia Hamdard University,
New Delhi, New Delhi, India

N. A. Ganai
Animal Genetics and Breeding, Sher-e-Kashmir University of Agricultural Sciences and
Technology – Kashmir, Srinagar, Jammu and Kashmir, India

N. A. Shaik et al. (eds.), *Essentials of Bioinformatics, Volume I*,
https://doi.org/10.1007/978-3-030-02634-9_12

12.1 Introduction to Biological Networks

The biology of organisms is complex and driven by the interplay of genes, proteins, small molecules, metabolites, and nucleic acids. To understand the biological system, it is important to interpret these interactions. As the genetic code suggests, DNA is transcribed to RNA, and then RNA is translated to proteins (Fig. 12.1), depending on the coding potential of mRNAs. The fundamental objective of systems biology is to comprehend the complete biological system by elucidating the behavior of all components and their interactions.

Over the years, the huge volume of data has been generated by various high-throughput techniques like next-generation sequencing, microarrays, and mass spectrometry to understand the molecular mechanism behind specific diseased state. These techniques provide the expression profiles of proteins and other genomic information for a biological system in one or the other format. However, interpretation of this complex and multidimensional data is a great challenge. In this chapter, we tried to elaborate on the data types from such high-throughput technologies, giving details about the methodologies and software to extract valuable and legible information from such complex data. Network analysis can be one of the promising approaches to address this issue and understand the biology behind the myriad of mechanisms and biological processes.

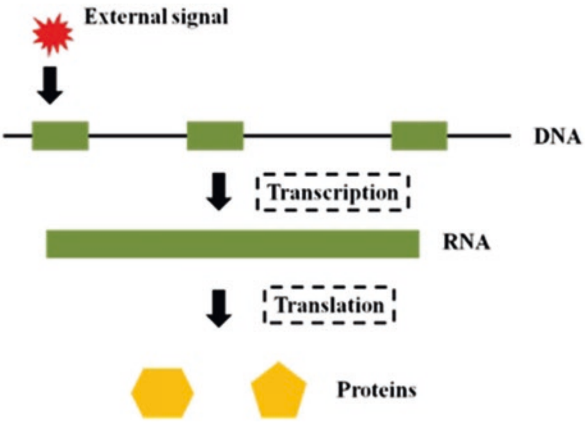


Fig. 12.1 The central dogma of biology. DNA is transcribed to RNA, and RNA is translated to proteins, which are the protagonist in biological systems

12.2 Types of Biological Networks

Biological networks are the mathematical representation of interactions between different types of molecules in a biological system. There are different types of biological networks as described below.

12.2.1 Protein-Protein Interaction Networks (PPIN)

The most important biochemical molecule in the organism is DNA, which stores the genetic information. The central dogma quotes that information from DNA is transferred to RNA and then from RNA to proteins (Fig. 12.1). However, the theory quoted by Beadle and Tatum (Beadle and Tatum 1941) about one gene-one enzyme-one function theory has come a long way. Now the biological processes are more complex, where proteins serve as the major molecule guiding a specific biological pathway. Proteins are long chains of amino acids, which are folded in a particular configuration. It is this specific configuration that enables a protein to physically interact with other proteins to form protein complexes and serve in downstream processes. Since proteins play a principal role in determining the molecular mechanisms and cellular responses, understanding the protein interaction networks is becoming a salient subject in research. Compiling the dense omics data from high-throughput techniques into meaningful biological networks is important to understand the cellular functions in a normal and diseased condition of the organism. This knowledge can further be translated into effective diagnostic strategies.

The reason behind the formation of protein complexes is still enigmatic. Proteins are folded in a specific configuration, which allows them to interact with other proteins via domains. Protein domains are the small conserved sequence of amino acids. These domains can function independently of the chain of protein and interact with other proteins to trigger biochemical processes like posttranslational modification, e.g., phosphorylation, glycosylation, etc. In one way, functional domains bind to other domains via protein interfaces to initiate a cellular response, e.g., interaction between Ras and its GTPase activating protein Ras-GAP, leading to a signaling process (Bader et al. 2008). Such type of interaction has high binding affinity and stability in lower volumes. In another way, domains bind to a stretch of amino acid sequence (3–10 in length) called motifs, present in the disordered region of a protein. For example, PDZ domain binds to C-terminus motifs of interacting proteins. The folds in the protein tertiary structure create active sites or catalytic domains, which interact with other proteins having similar conformations to initiate an enzymatic reaction (an induced-fit model). This model was proposed to be a lock-and-key model (Alberts et al. 2002), where the enzyme and substrate physically interact with each other to stimulate a biochemical reaction. Further, protein interactions in cell signaling pathways help in understanding cellular transports and interconnected modules in a biological process, e.g., p53 pathway.

12.2.1.1 Structure of Protein-Protein Interaction (PPI) Networks

PPI network is an organization of functional modules that comprises of a set of proteins having similar functions. The biological process can be interpreted as a modular network where proteins in a module are densely connected with each other sharing a similar function. Proteins are represented as "nodes" in the PPIN. Some proteins in the network have more interactions than other proteins, and these are called hubs. These nodes have very few interactions outside the module (Yook et al. 2004). PPIN are scale-free networks (Albert 2005). Hubs play a centralized role in scale-free networks and are classified as "party hubs" and "date hubs" (Han et al. 2004). Party hubs function inside the module and bind to interacting partners simultaneously, while date hubs bridge different modules and bind to interacting partners in different time and locations.

Network topology includes modularity and hub-oriented structure. There are four elements which define network topologies: (i) average degree (*K*) which can be calculated as degree distribution *P*(*k*), (ii) clustering coefficient (*C*) calculated as degree distribution of cluster coefficients *C*(*k*), (iii) average path length (*L*) calculated as shortest path distribution SP(*i*), and (iv) diameter (*D*) calculated as topological coefficient distribution TC(*k*). This concept is further explained in the chapter.

To understand the biochemical networks in a particular species, condition, or biological state, scientists are trying to merge the expression data from the myriad of experimental and computational techniques with the existing networks. For example, when expression data of each phase of yeast cell cycle was merged with PPIN in yeast cell cycle, most proteins were expressed continuously and found in the PPIN in each cell cycle, but there were some proteins which are expressed in a specific cell cycle phase and thus present in a PPIN of that phase (Batada et al. 2006). This is how computational algorithms are making the understanding of biological systems in different conditions (species, diseases, drug treatments) much easier than in earlier times. We can translate these results into therapeutic advancements in biomedical science.

12.2.2 Disease-Gene Interaction Networks

A disease is caused by the malfunctioning of any crucial biomolecule of an organism which can be a gene, protein, metabolite, or some unwanted genetic interactions, leading to the structural and functional aberration in the organisms. The genes, proteins, and other cellular components carry out their biological function in a complex network. With the advent of genomic sequencing and large-scale proteomics techniques, abundant genetic information is now available to build interactomes (biological networks). These biological networks help in understanding the pathophysiology of a specific disease and lead to a better understanding of the disease pharmacokinetics. Also, new disease-specific genes are identified which play an important role in disease prognosis.

12.2.2.1 Structure of Disease-Gene Interaction Networks

The important property of molecular networks is that they are dynamic. These networks change with space and time to adapt to different biological conditions. Hence this property of networks can be used to identify disease progression and also prognostic pathways specific for that disease.

Infection or disease progression occurs mainly due to molecular interactions. During host-pathogen interactions, host proteins interact with pathogen's proteins to initiate aberrated pathways. Such networks help researchers in understanding the mechanisms by which pathogens can attack the hosts. These networks are scale-free following the power law.

Recently, a research on human disease network (Goh et al. 2007) has given insight on how diseases are connected to each other through genes associated with them. The diseases are connected to their genes in which the associated mutation was found. This network is called "diseasome." One genetic mutation can be associated with several diseases. This resulted in a bipartite graph.

Diseases are also connected to each other if they have a common linked gene with a mutation, thus leading to human disease network (HDN). Genes are also connected to each other if they are found in the same disorder, thus resulting in disease-gene network (DGN) (Fig. 12.2).

12.2.3 Metabolic Networks

Metabolism is a complex association of metabolic reactions involving substrate, products, molecules, compounds, and cofactors. In general, metabolic reactions are reversible reactions, and they interact with each other, i.e., a product of one reaction can be the reactant of other reaction. The network of these metabolic reactions is called a metabolic network. An example of the metabolic network is the glycolysis process in humans.

12.2.3.1 Structure of Metabolic Networks

Metabolic pathways consist of enzymes, main substances, and co-substances. Main substances are metabolites, and co-substances are molecules like ATP, NADPH, etc., which help in transferring electrons. Metabolic networks have unique properties different from other networks because of (a) conservation constraints at each node and (b) the representation where nodes are metabolites and links are reactions catalyzed by specific gene products. This representation is very different from PPIN, where nodes are gene products and links are interactions. Also, a node in the metabolic network cannot be deleted by genetic techniques but links. A node in PPIN can be deleted using different molecular techniques, but it can result in a lethal phenotype. Metabolic networks have flux distribution with average path length longer, and their functional state does not have scale-free characteristics (Arita 2004).

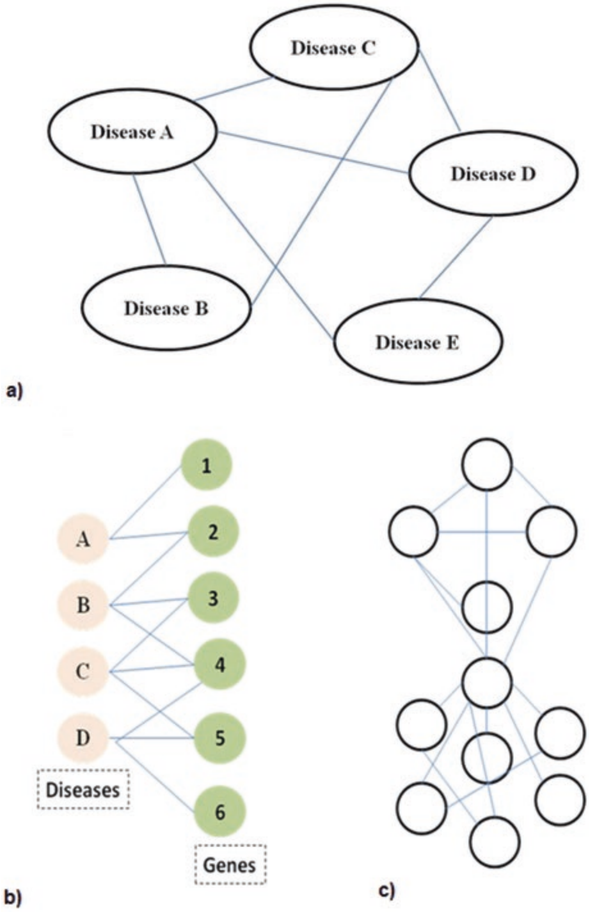


Fig. 12.2 (**a**) Human disease network (HDN): Different types of disease nodes are connected to each other if they share a common mutated gene. (**b**) The diseasome: The set of diseases are connected to the associated mutation in a gene. Genes are green in color while disease nodes are in orange color. (**c**) The disease gene network (DGN): The genes are connected to each other if they are associated with the same disorder

The metabolic networks can be of three types:

(a) *Simplified metabolic network:* A network of enzymes, reactions, and main substances but not co-substances (Fig. 12.3).
(b) *Simplified metabolite network:* A network of metabolites only. This kind of network is not always directed, and the metabolites are not directly connected to each other, but such type of interaction can be obtained from correlation analysis (Fig. 12.4).
(c) *Enzyme network:* A network of enzymes only. This kind of network can be obtained from PPIN (Fig. 12.5).

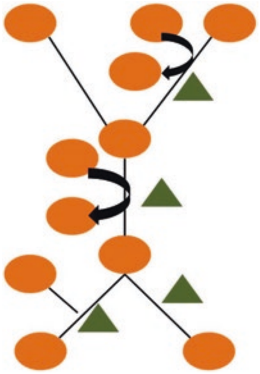


Fig. 12.3 Simplified metabolic network. The circles represent metabolites and the triangles are enzymes

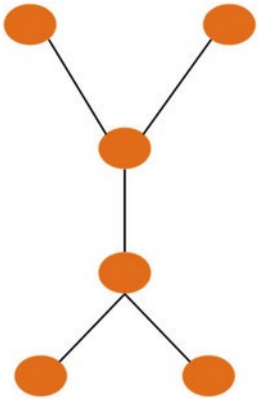


Fig. 12.4 Simplified metabolite network. The circles represent the metabolites

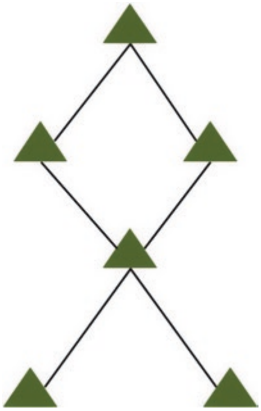


Fig. 12.5 Enzyme network. The triangles represent the enzymes

12.2.4 *Gene Regulatory Networks*

Gene regulation is the control of gene expression and thus the synthesis of proteins at transcription as well as translational level. The biological system is hardwired by the explicitly defined gene regulatory codes that control transcription as well as translation of the gene in a spatial and temporal manner. These control systems consist of transcription factors (TFs), signaling molecules, microRNAs, long noncoding RNAs, and epigenetic modulators. The molecules like TFs are cis-regulatory modules, which control the expression of the neighboring gene. Small RNAs like miRNAs control protein synthesis at the translation levels. Epigenetic modulators control the protein activity. Such kind of association of genes with its regulatory elements forms a gene regulatory network (GRN). GRNs include feedback, feed-forward, and cross-regulatory loops which define the regulation of gene at various levels.

12.2.4.1 Structure of Gene Regulatory Network

GRNs consist of many sub-circuits like signal transduction sub-circuit, metabolic reaction sub-circuit, and protein-protein interaction sub-circuits. Also, there can be a sub-circuit where TFs can regulate the expression of regulatory molecules like miRNAs. These sub-circuits connect to each other along with gene regulatory molecules to design a GRN.

GRNs are used to study the rationale behind the differential expressed genes in various diseased states and also help drug designing. An example of GRN is depicted in Fig. 12.6 where TFs are regulating the genes, which are in turn regulated by miRNAs.

12.2.5 *Gene Co-expression Networks*

A gene co-expression network is a kind of undirected graph where nodes (genes) are linked to each other on the basis of similarity in expression patterns (co-expression) under various experimental conditions. Gene co-expression network

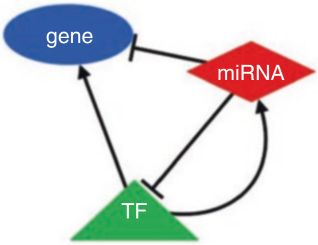


Fig. 12.6 Gene regulatory network

analysis helps in the simultaneous identification and grouping of genes with similar expression profiles. This analysis is of biological importance because co-expressed genes are regulated by the same transcription factors, functionally related or involved in same biological pathway(s). This kind of networks is built using expression data generated from high-throughput techniques such as microarray and RNA-Seq.

The co-expression network construction involves two steps:

1. Co-expression/expression relatedness measure calculation
2. Significant threshold selection

12.2.5.1 Co-expression Measure Calculation

The expression values of a gene for different samples are generally $\log_2$ transformed before co-expression measure calculation in order to scale the values to the same dynamic range. The following are four measures used for co-expression measure (Weirauch 2011) calculation between genes:

- *Pearson’s correlation coefficient*: This measure is widely used for calculating expression similarity among genes for gene co-expression network construction. It gauges the inclination of two vectors to increment or abatement together, rendering a measure of their general relationship. Its value varies from -1 to 1 where absolute values near to 1 represent strong correlation. The positive values represent positive correlation, i.e., activation mechanism where a gene expression value is directly proportional to the expression value of other co-expressed gene and vice versa. When the relation between expression values of co-expressed genes is inverse, it represents the inhibition mechanism, and they will have negative correlation value. Assuming linear correlation, normally distributed values and being sensitive to outliers are some of the drawbacks of the Pearson correlation measure.
- *Mutual Information*: It describes nonlinear relations between genes, which measure the similarity between two genes based on their relations with other genes.
- *Spearman’s rank correlation coefficient*: It is the nonparametric version of Pearson’s correlation which is calculated for the ranks of gene expression values in a gene expression matrix.
- *Euclidean distance*: To calculate the geometric distance between gene pairs, both positive and negative expression values are considered. It is not suitable when the absolute expression values of related genes are highly varying.

12.2.5.2 Threshold Selection

After calculating co-expression measures between all pairs of genes, a cutoff is imposed for selecting the gene pairs that should be connected in the network. Several methods can be used for selecting a threshold for gene co-expression network construction, for example, weighted gene co-expression network analysis (WGCNA) package which follows a power-law distribution approach for threshold selection.

12.2.5.3 WGCNA (Weighted Gene Co-expression Network Analysis)

It is a systems biology approach, which illustrates the correlation gene patterns across a series of microarray samples. It has been widely used in the genomic applications. It can be used to define modules of highly correlated genes, for summarizing such modules based on intra-modular hub genes and for calculating module membership for network nodes, i.e., genes, to study the relationships between co-expressed genes and external sample traits. It can also be used to compare the network topology of different networks. WGCNA (Langfelder and Horvath 2008) can be used as:

1. Data reduction technique
2. Clustering method
3. Feature selection method
4. Framework for integrating genomic data based on expression value.

The WGCNA pipeline is shown in Fig. 12.7.

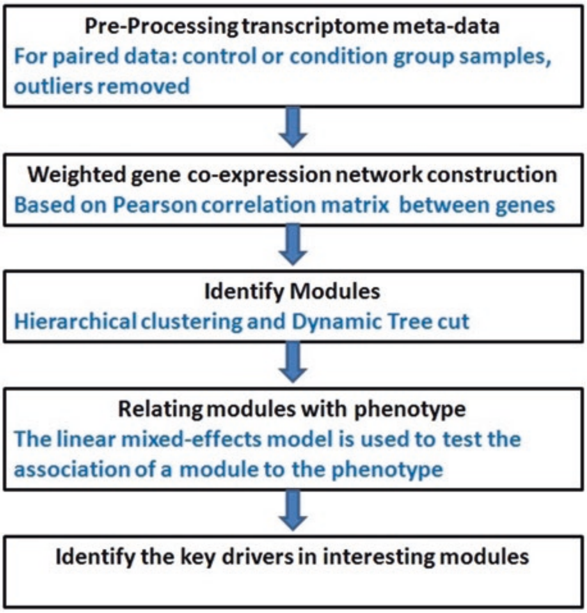

Pre-Processing transcriptome meta-data
For paired data: control or condition group samples, outliers removed
Weighted gene co-expression network construction
Based on Pearson correlation matrix between genes
Identify Modules
Hierarchical clustering and Dynamic Tree cut
Relating modules with phenotype
The linear mixed-effects model is used to test the association of a module to the phenotype
Identify the key drivers in interesting modules

Fig. 12.7 WGCNA pipeline

12.3 Network Measures

A complex biological system can be considered as networks wherein components within a complex system can be represented as nodes and are connected through their interactions, also known as edges. It enables analysis of the network's topology, which gives insight into molecular mechanism operating within a cell under given condition. Network topology considers knowledge about the global and local properties of the network. Graph-theoretic network analysis can be used to measure the topological properties quantitatively (Ma'ayan 2011). Centrality indices are one of the measures which tell about the important nodes or edges, for the connectivity or the information flow within the network. The following are some of the centrality measures which can be calculated to define local properties of a network:

1. *Degree centrality*: It tells about the number of links for each node. The nodes with the highest degree may act as a hub, regulating multiple other nodes in the network.
2. *Node betweenness centrality*: It tells about the number of shortest paths between all possible pairs of nodes. The nodes with high betweenness centrality lie on communication paths and can control information flow.
3. *Closeness centrality*: It is the average shortest path from one node to all other nodes. It estimates how fast the flow of information would be through a given node to other nodes.
4. *Eigenvector centrality*: It accesses the closeness to highly connected nodes.
5. *Edge betweenness centrality*: It is the number of shortest paths that go through an edge among all possible shortest paths between all the pairs of nodes.

The following are some of the global properties of a network:

1. *Degree distribution:* It is the probability distribution of degrees over the whole network. For most of the biological networks, this distribution follows power-law, giving scale-free architecture to the network. It makes network stable and robust to random failures.
2. *Characteristic path length*: It represents the average shortest path between all pairs of nodes.
3. *Clustering coefficient*: It is the local density of interactions by measuring the connectivity of neighbors for each node averaged over the entire network. It demonstrates the tendency of the nodes to cluster together. High clustering coefficient means the presence of communities in a network. The communities are very important in the biological network as they represent functional modules or protein complexes working together to achieve a biological process.

12.4 Gene Ontology

The gene ontology is a cooperative attempt to bring together a consolidated description of gene and gene product for all organisms. It can be a promising approach to decipher key components from complex biological networks and helps in organizing the biological networks in a meaningful way to improve performance and biological interpretability.

Comparative genomics has apparently shown that a vast portion of the genes specifying the major biological functions are common to all organisms. Information of the biological role of such common proteins in one organism can often be exchanged with other organisms. The objective of the Gene Ontology Consortium is to deliver a dynamic, controlled vocabulary that can be connected to all organisms even as information of gene and protein roles in cells is gathering and evolving. The undertaking started in 1998 as a coordinated effort between three model organism databases, the FlyBase (Drosophila), the Saccharomyces Genome Database (SGD), and the Mouse Genome Database (MGD). The GO Consortium (GOC) has since developed to join numerous databases, including a few of the world's significant vaults for the plant, animals, and microbial genomes (Reference Genome Group of the Gene Ontology Consortium 2009).

There are three separate aspects to this effort:

1. The development and maintenance of the ontologies themselves
2. The annotation of gene products, which entails making associations between the ontologies and the genes and gene products in the collaborating databases
3. The development of tools that facilitate the creation, maintenance and use of ontologies

The GO project has created three organized ontologies that associate gene products with their biological processes, cellular components, and molecular functions in a species-independent manner (Botstein et al. 2000).

- Cellular component: The location in the cell where a gene product is functional. In most of the situation, annotations connecting gene product with cellular component types are made on the basis of a direct observation of an instance of the cellular component in a microscope. Cellular component incorporates terms like "ribosome" or "proteasome," specifying where different gene products would be found.
- Molecular function: A molecular function term is an enduring potential of a gene product to act in a certain way or in other words the biochemical activity (including specific binding to ligands or structures) of a gene product. This definition likewise applies to the ability that a gene product conveys as a potential. It portrays just what is done without indicating where or when the occasion really happens. For example, glucose transport, regulation of cell death, etc.

- Biological process: It defines what the gene or gene product contributes. A process is defined by means of at least one requested gathering of molecular functions for example, "cell growth and maintenance", "signal transduction", "cAMP biosynthesis", etc.

Gene ontology (GO) has a graph-like structure where GO terms are nodes and relationships among them are links between nodes. The structure is loosely hierarchical having a parent-child relationship between nodes. Child node terms are more specialized than their parent node terms, but a child may have more than one parent term. For example, "integral component of external side of plasma membrane" is a child of the "integral component of plasma membrane" and "intrinsic component of external side of plasma membrane" (Fig. 12.8).

GO terms are designed with a unique identifier and term name, for example, GO:0015758~ glucose transport. The unique identifier is a zero-padded seven-digit identifier prefixed by "GO:". The link between two nodes represents the relationship between them. For example, in Fig. 12.9, GO term "GO:1900117" has two types of relationship with parent nodes, i.e., "is a" and "regulates" which means GO:1900117 *is a* kind of regulation of apoptotic process (GO:0042981) and it *regulates* execution phase of apoptosis (GO:0097194).

The ontologies are dynamic, as in they exist as a network that is changed as more data gathers yet have adequate uniqueness and accuracy with the goal that databases in light of the ontologies can consequently be refreshed as the ontologies develop. The ontologies are adaptable in another way, so they can reflect the numerous distinctions in the biology of the assorted life forms, such as the breakdown of the nucleus during mitosis. The GO vocabulary is intended to be species-impartial and incorporates terms relevant to prokaryotes and eukaryotes and single and multicellular organisms.

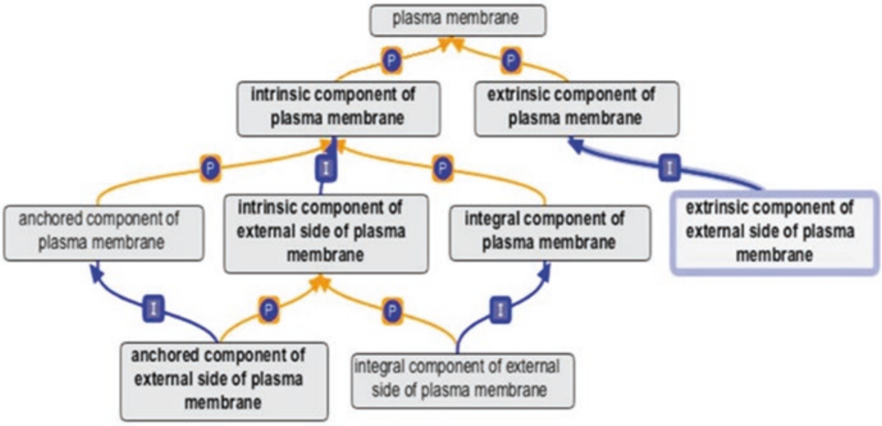
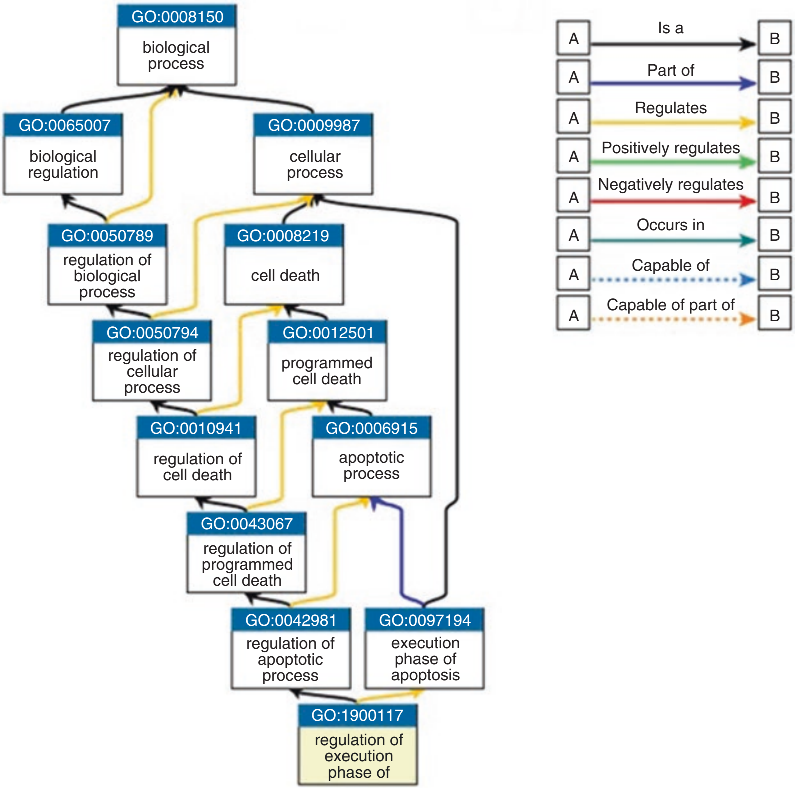

plasma membrane
P
P
intrinsic component of plasma membrane
extrinsic component of plasma membrane
P
I
P
I
anchored component of plasma membrane
intrinsic component of external side of plasma membrane
integral component of plasma membrane
extrinsic component of external side of plasma membrane
I
P
P
I
anchored component of external side of plasma membrane
integral component of external side of plasma membrane

Fig. 12.8 Relationship between GO terms

Ancestor chart for GO:1900117



QuickGO - https://www.tbi.ac.uk/QuichGo

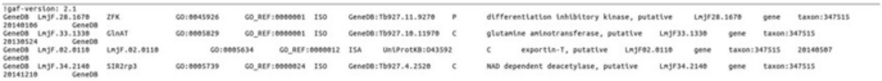
Fig. 12.9 GO ancestor chart

12.4.1 Applications of Gene Ontology

The gene ontology annotation is most widely utilized for deciphering large-scale “omics” data. Gene ontology enrichment analysis is one of the uses of GO annotation which helps in finding the significant clusters of genes associated with biological processes and thus reduce the bulk amount of data to the much smaller number of biological function getting altered under different experimental conditions.

12.5 GO Annotation

GO annotation is a link between the gene product and what that gene product can do, which molecular and biological processes it adds to, and where in the cell it is functioning. The GO annotation focuses on the identification of functional activities



```
!gaf-version: 2.1
GeneDB  LmjF.28.1670    ZFK             GO:0045926      GO_REF:0000001  ISO     GeneDB:Tb927.11.9270    P       differentiation inhibitory kinase, putative     LmjF28.1670     gene    taxon:347515
20140106        GeneDB
GeneDB  LmjF.33.1330    GlnAT           GO:0005829      GO_REF:0000001  ISO     GeneDB:Tb927.10.11970   C       glutamine aminotransferase, putative    LmjF33.1330     gene    taxon:347515
20130524        GeneDB
GeneDB  LmjF.02.0110    LmjF.02.0110            GO:0005634      GO_REF:0000012  ISA     UniProtKB:O43592        C       exportin-T, putative    LmjF02.0110     gene    taxon:347515    20140507
GeneDB
GeneDB  LmjF.34.2140    SIR2rp3         GO:0005739      GO_REF:0000024  ISO     GeneDB:Tb927.4.2520     C       NAD dependent deacetylase, putative     LmjF34.2140     gene    taxon:347515
20141210        GeneDB
```

S.No	Data	Description
1	DB (GeneDB)	Database providing the gene association list
2	DB_Object_ID (LmjF.28.1670)	a primary identifier in the database for the object being annotated
3	DB_Object_Symbol (ZFK)	Contain gene name, used to identify gene annotation.
4	Qualifier	Rarely used; changes interpretation of GO annotation
5	GO ID (GO:0045926)	The GO identifier for the term attributed to the DB_Object_Symbol
6	DB:Reference (GO_REF:0000001)	Literature evidence.
7	Evidence (ISO)	The evidence code is associated with a specific GO annotation term to describe what type of evidence was present in that reference to make the annotation.
8	With (GeneDB:Tb927.11.9270)	Identifier connecting evidence code with annotation
9	Aspect (P)	One of the three ontology classes F: molecular function P: biological process C: cellular component
10	DB_Object_Name	Gene name or gene product name
11	DB_Object_Name_Synonym	Gene symbol aliases
12	DB_Object_Type (gene)	The entity being annotated (gene, protein, exon etc).
13	Taxon_ID (taxon:347515)	Identifier for the species being annotated.
14	Date (YYYY:MM:DD)	Annotation date
15	Assigned_By	Database contributing to the annotation.

Fig. 12.10 Annotation format provided by the GO consortium

of a gene or a protein. GO annotation is principally divided into two parts: first, a link between the gene product and a representative GO term and second is an evidence to establish that link (Weirauch 2011). The annotation data in the GO database is contributed by members of the GO Consortium (GOC); more than 15 major contributing groups are actively working for GOC (Blake 2013). GOC is a dynamic ontology-based resource that contains the most updated and exhaustive set of annotations available in the literature. Keen utilization of GO annotation assures the best result in advancing biological research. GO annotation process follows a basic three-step paradigm in which:

1. Relevant experimental data is being identified from the biomedical literature.
2. Correlation of gene product with GO terms.
3. Finally, annotation quality control and refinement process are employed to ensure that the annotation has a correct formal structure.

GO annotation data file provided to GOC consists of 15 columns (Fig. 12.10). To fully comprehend the GO annotation file, a few important terms are worth to discuss:

An *annotation* is a process of assigning GO terms to the gene product. These assignments are made based on the conclusion drawn from experiments.

A *gene product* is an output generated from RNA or protein molecule that has some defined role in the biology of an organism.

A *molecular function* encompasses activities of a gene product such as catalytic or binding activities, influencing at the molecular level.

A *biological process* is a recognized sequence of molecular events performed by one or more ordered assemblies of molecular functions. For example, the progression of the brain development over time would be an instance of the biological function *brain development*.

A *cellular component* is a part of a cell where a gene product is active.

Curation is the formulation of annotation on the basis of the gene and gene product information from experimental observations.

An *evidence code* is a three-letter code that specifies the type of evidence identified from literature to support the association between gene and gene product. There are 21 (Hill et al. 2008) evidence (Table 12.1) codes classified broadly into five groups.

Table 12.1 Evidence codes classification

Category	Evidence codes
Experimental Evidence codes: literature cited indicates that there is evidence from an experiment directly supporting an association between gene and gene product	Inferred from Experiment (EXP) Inferred from Direct Assay (IDA) Inferred from Physical Interaction (IPI) Inferred from Mutant Phenotype (IMP) Inferred from Genetic Interaction (IGI) Inferred from Expression Pattern (IEP)
Computational Analysis evidence codes: literature cited contains observations from in silico analysis	Inferred from Sequence or structural Similarity (ISS) Inferred from Sequence Orthology (ISO) Inferred from Sequence Alignment (ISA) Inferred from Sequence Model (ISM) Inferred from Genomic Context (IGC) Inferred from Biological aspect of Ancestor (IBA) Inferred from Biological aspect of Descendant (IBD) Inferred from Key Residues (IKR) Inferred from Rapid Divergence(IRD) Inferred from Reviewed Computational Analysis (RCA)
Author statement evidence codes: annotation was made on the basis of declarations made by the author(s) in the literature	Traceable Author Statement (TAS) Non-traceable Author Statement (NAS)
Curator statement evidence codes: when annotation does not support any direct evidence	Inferred by Curator (IC) No biological Data available (ND) evidence code
Electronic Annotation evidence code: specifies that annotation was assigned by automated methods, without curator	Inferred from Electronic Annotation (IEA)

12.5.1 *Utilities for GO Annotation*

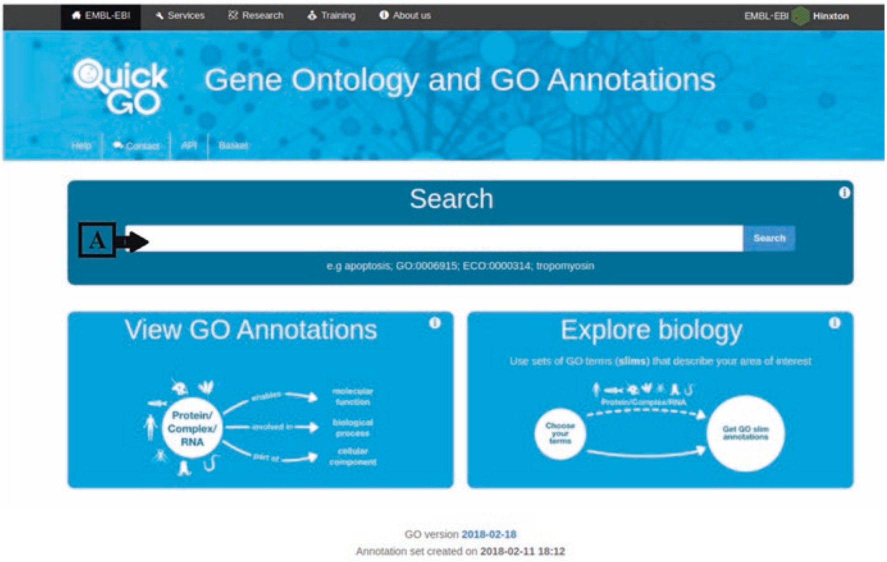
The gene ontology (GO) provides core biological knowledge representation for modern biologists, whether computationally or experimentally based. It has become an extremely useful tool for the analysis of OMICS data and structuring of biological knowledge. With the aim of high-quality annotation and easy access to GO annotation database, a number of online tools are available, such as *QuickGO* (Binns et al. 2009), which have been developed at the EBI, and *AmiGO* (Carbon et al. 2008), which is developed by the GO Consortium.

12.5.1.1 Viewing GO Terms Using QuickGO

A responsive web-based tool that allows easy access to GO annotation. QuickGO can be queried online at https://www.ebi.ac.uk/QuickGO/ or can be downloaded freely from http://www.ebi.ac.uk/QuickGO/installation.html.

The *QuickGO* home page (Fig. 12.11) provides a query box (Fig. 12.11 (A)) to start searching for GO annotation. *QuickGO* takes a wide range of gene identifiers and symbol for annotation retrieval, for example, NCBI Gene IDs, RefSeq accessions, Ensembl Ids, UniProtKB accessions, InterPro IDs, Enzyme Commission (EC) numbers, and GO IDs.

A search for the keyword "apoptosis" retrieves all terms where "apoptosis" is present in the term name and gene product (Fig. 12.12). Here search term "apoptosis" is underlined in red color, and matched terms are shown in green color.

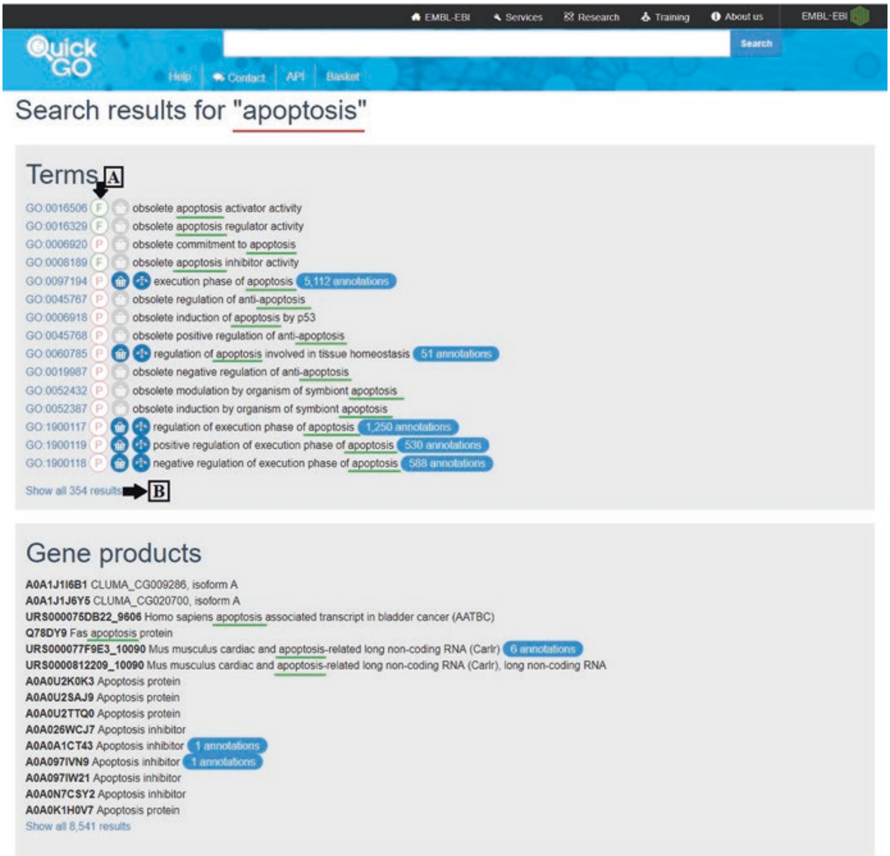


e.g apoptosis; GO:0006915; ECO:0000314; tropomyosin

Use sets of GO terms (**slims**) that describe your area of interest

GO version **2018-02-18**
Annotation set created on **2018-02-11 18:12**

Fig. 12.11 *QuickGO* home page



Search results for "apoptosis"

Terms

GO:0016506 F obsolete apoptosis activator activity
GO:0016329 F obsolete apoptosis regulator activity
GO:0006920 P obsolete commitment to apoptosis
GO:0008189 F obsolete apoptosis inhibitor activity
GO:0097194 P execution phase of apoptosis 5,112 annotations
GO:0045767 P obsolete regulation of anti-apoptosis
GO:0006918 P obsolete induction of apoptosis by p53
GO:0045768 P obsolete positive regulation of anti-apoptosis
GO:0060785 P regulation of apoptosis involved in tissue homeostasis 51 annotations
GO:0019987 P obsolete negative regulation of anti-apoptosis
GO:0052432 P obsolete modulation by organism of symbiont apoptosis
GO:0052387 P obsolete induction by organism of symbiont apoptosis
GO:1900117 P regulation of execution phase of apoptosis 1,250 annotations
GO:1900119 P positive regulation of execution phase of apoptosis 530 annotations
GO:1900118 P negative regulation of execution phase of apoptosis 588 annotations
Show all 354 results

Gene products

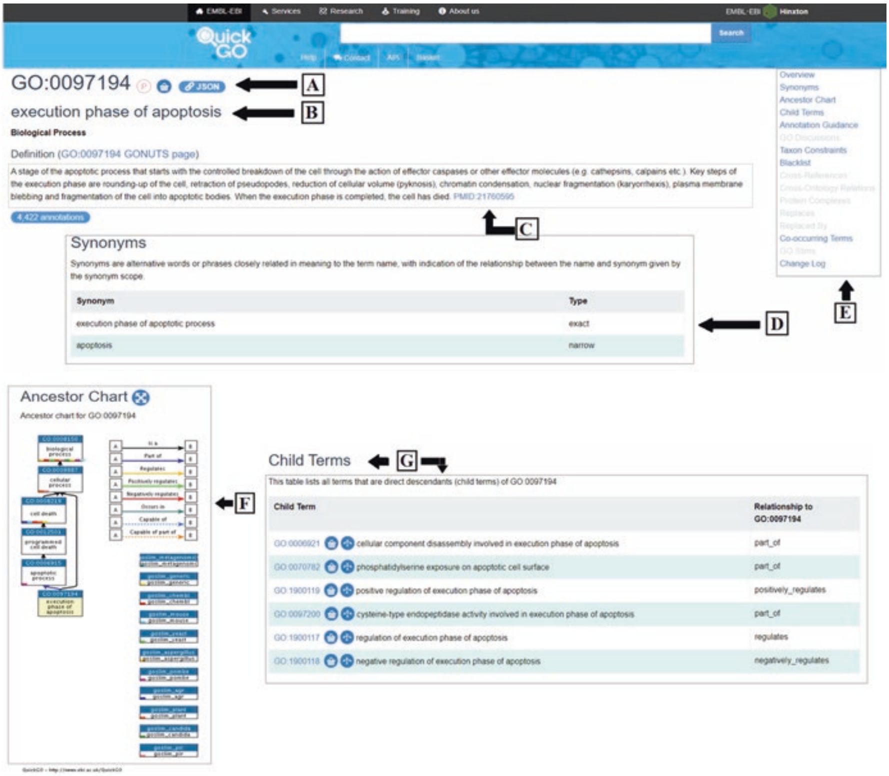
A0A1J1I6B1 CLUMA_CG009286, isoform A
A0A1J1J6Y5 CLUMA_CG020700, isoform A
URS000075DB22_9606 Homo sapiens apoptosis associated transcript in bladder cancer (AATBC)
Q78DY9 Fas apoptosis protein
URS000077F9E3_10090 Mus musculus cardiac and apoptosis-related long non-coding RNA (Carlr) 6 annotations
URS0000812209_10090 Mus musculus cardiac and apoptosis-related long non-coding RNA (Carlr), long non-coding RNA
A0A0U2K0K3 Apoptosis protein
A0A0U2SAJ9 Apoptosis protein
A0A0U2TTQ0 Apoptosis protein
A0A026WCJ7 Apoptosis inhibitor
A0A0A1CT43 Apoptosis inhibitor 1 annotations
A0A097IVN9 Apoptosis inhibitor 1 annotations
A0A097IW21 Apoptosis inhibitor
A0A0N7CSY2 Apoptosis inhibitor
A0A0K1H0V7 Apoptosis protein
Show all 8,541 results

Fig. 12.12 QuickGO: search for keyword “apoptosis”

Clicking on the GO term (e.g., GO:0097194) redirect user to *Term Information Page* (Fig. 12.13), providing complete information for the selected GO term.

12.5.1.2 Viewing GO Terms Using AmiGO

AmiGO is another web-based application provided by the Gene Ontology Consortium for identification and visualization of GO terms related to genes. AmiGO can be accessed from GOC (http://amigo.geneontology.org) or can be downloaded (http://sourceforge.net/projects/geneontology/) to use as the standalone application.



GO:0097194

execution phase of apoptosis

Biological Process

Definition (GO:0097194 GONUTS page)

A stage of the apoptotic process that starts with the controlled breakdown of the cell through the action of effector caspases or other effector molecules (e.g. cathepsins, calpains etc.). Key steps of the execution phase are rounding-up of the cell, retraction of pseudopodes, reduction of cellular volume (pyknosis), chromatin condensation, nuclear fragmentation (karyorrhexis), plasma membrane blebbing and fragmentation of the cell into apoptotic bodies. When the execution phase is completed, the cell has died. PMID:21760595

Synonyms

Synonyms are alternative words or phrases closely related in meaning to the term name, with indication of the relationship between the name and synonym given by the synonym scope.

Synonym	Type
execution phase of apoptotic process	exact
apoptosis	narrow

Ancestor Chart

Ancestor chart for GO:0097194

Child Terms

This table lists all terms that are direct descendants (child terms) of GO:0097194

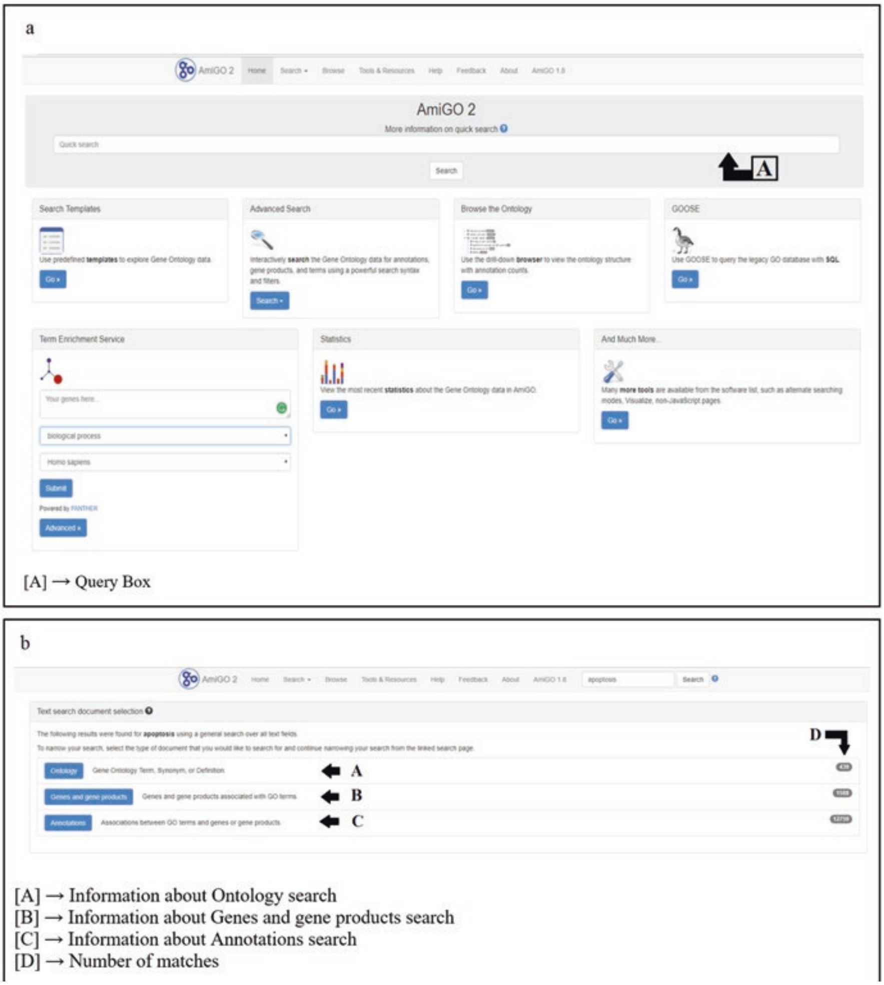
Child Term	Relationship to GO:0097194
GO:0006921 cellular component disassembly involved in execution phase of apoptosis	part_of
GO:0070782 phosphatidylserine exposure on apoptotic cell surface	part_of
GO:1900119 positive regulation of execution phase of apoptosis	positively_regulates
GO:0097200 cysteine-type endopeptidase activity involved in execution phase of apoptosis	part_of
GO:1900117 regulation of execution phase of apoptosis	regulates
GO:1900118 negative regulation of execution phase of apoptosis	negatively_regulates

[A] → A unique, stable identifier for the GO term
[B] → The primary GO term name
[C] → The term definition and evidence information
[D] → Term synonyms
[E] → Output columns
[F] → Ancestor terms to the selected GO term alongwith their relationship
[G] → Terms that are direct descendants of selected GO term.

Fig. 12.13 QuickGO: GO term information page view

The *AmiGO* home page (Fig. 12.14a) provides a search box (Fig. 12.14a (A)) to start searching for GO annotation. *AmiGO* takes a wide range of gene identifiers and symbol for GO annotation retrieval. Search keyword “apoptosis” is used to retrieve all terms where “apoptosis” is present in the GO terms, GO annotation, and gene products (Fig. 12.14b).

Clicking on “*Ontology*” will return all GO IDs containing “apoptosis” keyword in gene ontology term, synonym, or GO definition (Fig. 12.15).



Use predefined **templates** to explore Gene Ontology data.

Interactively **search** the Gene Ontology data for annotations, gene products, and terms using a powerful search syntax and filters.

Use the drill-down **browser** to view the ontology structure with annotation counts.

Use GOOSE to query the legacy GO database with **SQL**.

View the most recent **statistics** about the Gene Ontology data in AmiGO.

Many **more tools** are available from the software list, such as alternate searching modes, Visualize, non-JavaScript pages.

[A] → Query Box

The following results were found for **apoptosis** using a general search over all text fields.

To narrow your search, select the type of document that you would like to search for and continue narrowing your search from the linked search page.

Ontology — Gene Ontology Term, Synonym, or Definition.

Genes and gene products — Genes and gene products associated with GO terms.

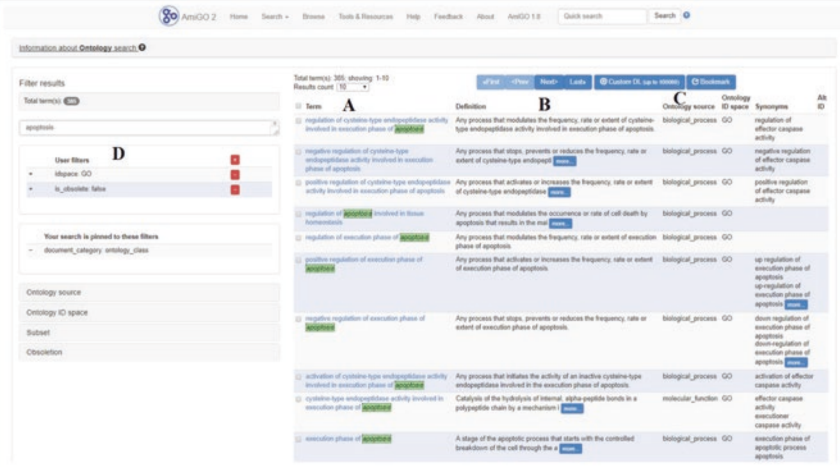
Annotations — Associations between GO terms and genes or gene products.

[A] → Information about Ontology search
[B] → Information about Genes and gene products search
[C] → Information about Annotations search
[D] → Number of matches

Fig. 12.14 AmiGO home page and "apoptosis" keyword search page

12.5.1.3 The Database for Annotation, Visualization, and Integrated Discovery (DAVID)

DAVID (Huang et al. 2008) provides a comprehensive set of functional annotation tools for investigators to comprehend the biological meaning behind large list of gene/protein lists generated from a variety of high-throughput genomic experiments. In this tutorial, given a list of differentially expressed genes, we will use DAVID to identify the enriched GO terms, such that we can have a clue on the role of genes played in the underlying biological processes.



Total term(s): 305; showing: 1-10
Results count 10

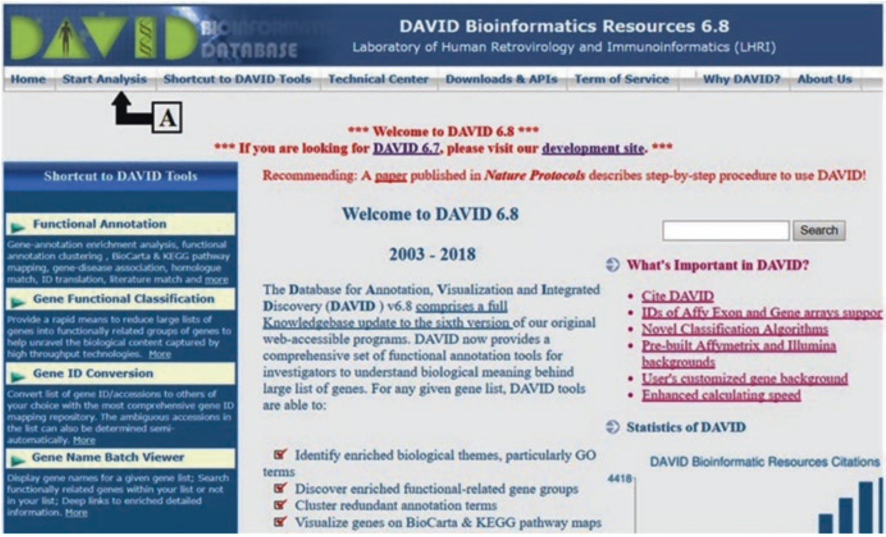
Term	Definition	Ontology source	Ontology ID space	Synonyms	Alt ID
regulation of cysteine-type endopeptidase activity involved in execution phase of apoptosis	Any process that modulates the frequency, rate or extent of cysteine-type endopeptidase activity involved in execution phase of apoptosis.	biological_process	GO	regulation of effector caspase activity	
negative regulation of cysteine-type endopeptidase activity involved in execution phase of apoptosis	Any process that stops, prevents or reduces the frequency, rate or extent of cysteine-type endopept more...	biological_process	GO	negative regulation of effector caspase activity	
positive regulation of cysteine-type endopeptidase activity involved in execution phase of apoptosis	Any process that activates or increases the frequency, rate or extent of cysteine-type endopeptidase more...	biological_process	GO	positive regulation of effector caspase activity	
regulation of apoptosis involved in tissue homeostasis	Any process that modulates the occurrence or rate of cell death by apoptosis that results in the mai more...	biological_process	GO		
regulation of execution phase of apoptosis	Any process that modulates the frequency, rate or extent of execution phase of apoptosis.	biological_process	GO		
positive regulation of execution phase of apoptosis	Any process that activates or increases the frequency, rate or extent of execution phase of apoptosis.	biological_process	GO	up regulation of execution phase of apoptosis up-regulation of execution phase of apoptosis more...	
negative regulation of execution phase of apoptosis	Any process that stops, prevents or reduces the frequency, rate or extent of execution phase of apoptosis.	biological_process	GO	down regulation of execution phase of apoptosis down-regulation of execution phase of apoptosis more...	
activation of cysteine-type endopeptidase activity involved in execution phase of apoptosis	Any process that initiates the activity of an inactive cysteine-type endopeptidase involved in the execution phase of apoptosis.	biological_process	GO	activation of effector caspase activity	
cysteine-type endopeptidase activity involved in execution phase of apoptosis	Catalysis of the hydrolysis of internal, alpha-peptide bonds in a polypeptide chain by a mechanism i more...	molecular_function	GO	effector caspase activity executioner caspase activity	
execution phase of apoptosis	A stage of the apoptotic process that starts with the controlled breakdown of the cell through the a more...	biological_process	GO	execution phase of apoptotic process apoptosis	

Fig. 7 Ontology term information page
[A] → The primary GO term name
[B] → The term definition and evidence information
[C] → Ontology source
[D] → Filters

Fig. 12.15 AmiGO: Ontology term information page

Perform Function Annotation Test

(a) Open the server DAVID 6.8 (https://david.ncifcrf.gov/).
(b) Click “Start Analysis” tab (A) as shown in Fig. 12.16.
(c) Submit a gene list to DAVID using input interface (Fig. 12.17). Paste the Affymetrix_geneID list from (A) to the text box (B), or load a text file containing gene IDs using browse option (C). Select the appropriate gene identifier type for input gene IDs using (D). User can also convert gene IDs to other formats using DAVID “Gene ID conversion” tool (E). Specify input IDs as gene list (i.e., genes to be analyzed) or as background genes (i.e., gene population background) at (F). Finally, click “Submit” (G).
(d) After job submission, the progress bar at the top shows job progress. If >20% of gene_identifiers are ambiguous or unrecognized, user will be redirected automatically to “DAVID Gene ID Conversion Tool” Fig. 12.18 (D). Implicitly, the background is set up to the species that contain majority of genes in the user’s input list (Fig. 12.18 (B)). User can change background using “Background” section as in Fig. 12.18 (A). Run “Functional Annotation chart” (Fig. 12.18 (C)) for functional enrichment analysis and biological knowledge base selection.
(e) Now user needs to input what type of functional annotations are required. For this purpose, the user needs to deselect the “Check Defaults” tab in Fig. 12.19 (A). Then select the GOTERM_BP_FAT (Fig. 12.19 (C)), which is the summarized version of Biological Processes in the GO, by clicking (+) sign as in



DAVID Bioinformatics Resources 6.8
Laboratory of Human Retrovirology and Immunoinformatics (LHRI)

*** Welcome to DAVID 6.8 ***
*** If you are looking for DAVID 6.7, please visit our development site. ***

Shortcut to DAVID Tools

Functional Annotation
Gene-annotation enrichment analysis, functional annotation clustering , BioCarta & KEGG pathway mapping, gene-disease association, homologue match, ID translation, literature match and more

Gene Functional Classification
Provide a rapid means to reduce large lists of genes into functionally related groups of genes to help unravel the biological content captured by high throughput technologies. More

Gene ID Conversion
Convert list of gene ID/accessions to others of your choice with the most comprehensive gene ID mapping repository. The ambiguous accessions in the list can also be determined semi-automatically. More

Gene Name Batch Viewer
Display gene names for a given gene list; Search functionally related genes within your list or not in your list; Deep links to enriched detailed information. More

Recommending: A paper published in *Nature Protocols* describes step-by-step procedure to use DAVID!

Welcome to DAVID 6.8

2003 - 2018

The Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.8 comprises a full Knowledgebase update to the sixth version of our original web-accessible programs. DAVID now provides a comprehensive set of functional annotation tools for investigators to understand biological meaning behind large list of genes. For any given gene list, DAVID tools are able to:

- Identify enriched biological themes, particularly GO terms
- Discover enriched functional-related gene groups
- Cluster redundant annotation terms
- Visualize genes on BioCarta & KEGG pathway maps

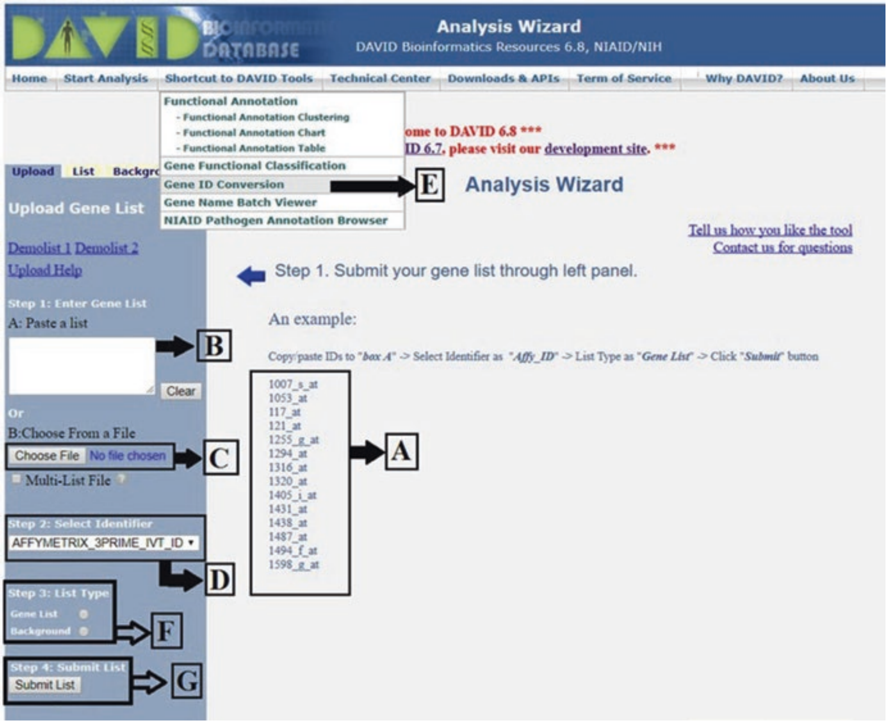
What's Important in DAVID?

- Cite DAVID
- IDs of Affy Exon and Gene arrays suppor
- Novel Classification Algorithms
- Pre-built Affymetrix and Illumina backgrounds
- User's customized gene background
- Enhanced calculating speed

Statistics of DAVID

DAVID Bioinformatic Resources Citations

Fig. 12.16 The DAVID 6.8 home page



Analysis Wizard
DAVID Bioinformatics Resources 6.8, NIAID/NIH

ome to DAVID 6.8 ***
ID 6.7, please visit our development site. ***

Upload Gene List

Demolist 1 Demolist 2
Upload Help

Step 1: Enter Gene List
A: Paste a list

Or
B:Choose From a File
□ Multi-List File

Step 2: Select Identifier
AFFYMETRIX_3PRIME_IVT_ID

Step 3: List Type
Gene List ○
Background ○

Step 4: Submit List

Analysis Wizard

Tell us how you like the tool
Contact us for questions

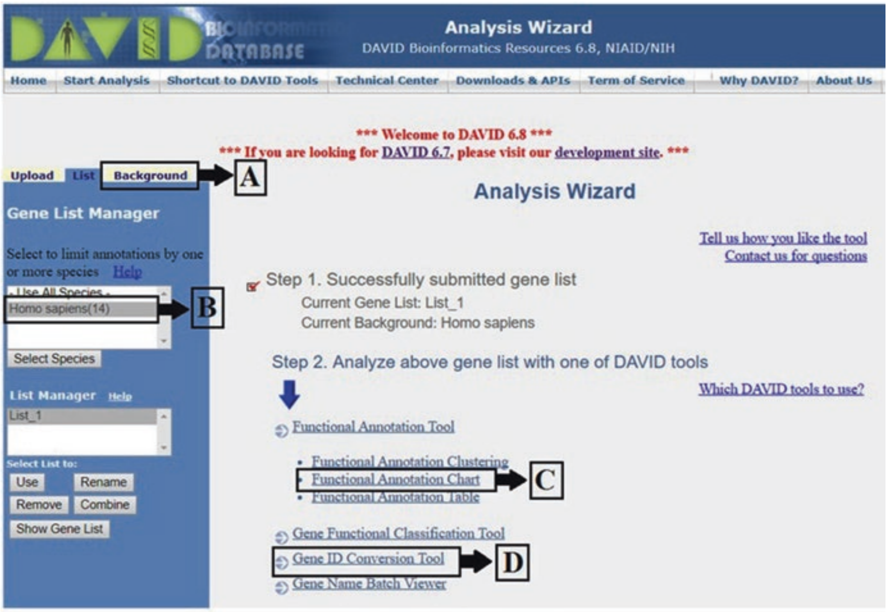
Step 1. Submit your gene list through left panel.

An example:

Copy/paste IDs to "*box A*" -> Select Identifier as "*Affy_ID*" -> List Type as "*Gene List*" -> Click "*Submit*" button

```
1007_s_at
1053_at
117_at
121_at
1255_g_at
1294_at
1316_at
1320_at
1405_i_at
1431_at
1438_at
1487_at
1494_f_at
1598_g_at
```

Fig. 12.17 Gene list submission page to DAVID



Analysis Wizard

DAVID Bioinformatics Resources 6.8, NIAID/NIH

*** Welcome to DAVID 6.8 ***

*** If you are looking for DAVID 6.7, please visit our development site. ***

Analysis Wizard

Select to limit annotations by one or more species Help

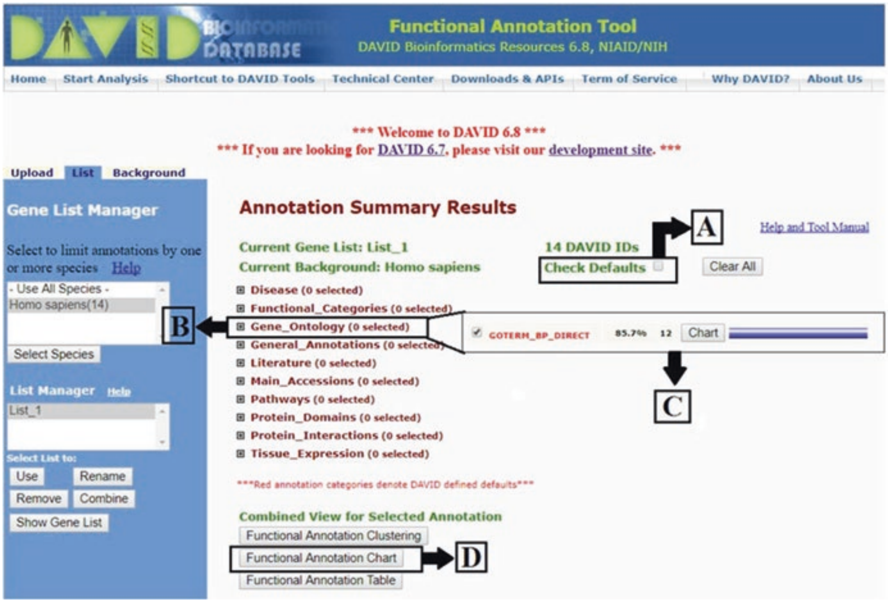
Step 1. Successfully submitted gene list

Current Gene List: List_1

Current Background: Homo sapiens

Step 2. Analyze above gene list with one of DAVID tools

Fig. 12.18 Webpage to access various analytic tools/modules available in DAVID



Functional Annotation Tool

DAVID Bioinformatics Resources 6.8, NIAID/NIH

*** Welcome to DAVID 6.8 ***

*** If you are looking for DAVID 6.7, please visit our development site. ***

Select to limit annotations by one or more species Help

Annotation Summary Results

Current Gene List: List_1

Current Background: Homo sapiens

14 DAVID IDs

- Disease (0 selected)
- Functional_Categories (0 selected)
- Gene_Ontology (0 selected)
- General_Annotations (0 selected)
- Literature (0 selected)
- Main_Accessions (0 selected)
- Pathways (0 selected)
- Protein_Domains (0 selected)
- Protein_Interactions (0 selected)
- Tissue_Expression (0 selected)

Red annotation categories denote DAVID defined defaults

Combined View for Selected Annotation

Fig. 12.19 Layout of DAVID “Functional Annotation Chart”

Fig. 12.19 (B). User can try other annotation categories, for example, classifying genes based on pathways using KEGG database, gene-gene interactions identification using BIOGRID database, domain identification, etc.

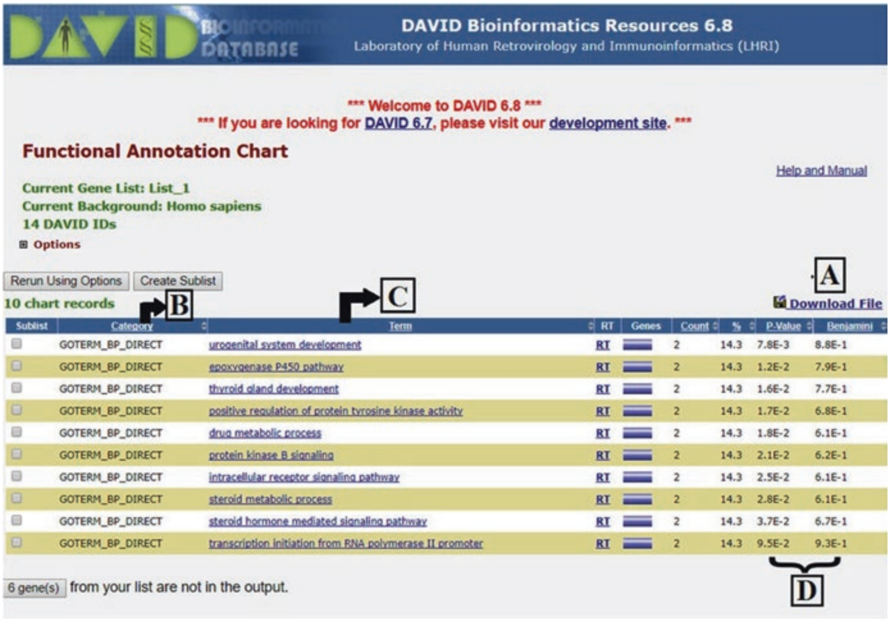
(f) Click on “Functional annotation chart” button (Fig. 12.19 (D)); a window will be prompted to show the results of functional enrichment test. This statistical test identifies the significantly enriched terms in GOTERM_BP_FAT knowledgebase (Fig. 12.20 (B)). Each row represents an enriched functional term (Fig. 12.20 (C)) and is ordered by their significance level; the smaller the score (Fig. 12.20 (D)), the better is the result. User can download the complete annotation file from Fig. 12.20 (A).

When to and Why Use DAVID?

High-throughput techniques like next-generation sequencing and mass spectrometry generate a huge amount of data, which finally yield gene identifiers.

The gene identifier table can be of various types:

- If data is generated from RNA sequencing or MS experiments, these gene identifiers are linked to respective expression values in a particular condition. These expression values can be as FPKM or RPKM units.



DAVID Bioinformatics Resources 6.8
Laboratory of Human Retrovirology and Immunoinformatics (LHRI)

*** Welcome to DAVID 6.8 ***
*** If you are looking for DAVID 6.7, please visit our development site. ***

Functional Annotation Chart

Current Gene List: List_1
Current Background: Homo sapiens
14 DAVID IDs

10 chart records

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
□	GOTERM_BP_DIRECT	urogenital system development	RT	▬	2	14.3	7.8E-3	8.8E-1
□	GOTERM_BP_DIRECT	epoxygenase P450 pathway	RT	▬	2	14.3	1.2E-2	7.9E-1
□	GOTERM_BP_DIRECT	thyroid gland development	RT	▬	2	14.3	1.6E-2	7.7E-1
□	GOTERM_BP_DIRECT	positive regulation of protein tyrosine kinase activity	RT	▬	2	14.3	1.7E-2	6.8E-1
□	GOTERM_BP_DIRECT	drug metabolic process	RT	▬	2	14.3	1.8E-2	6.1E-1
□	GOTERM_BP_DIRECT	protein kinase B signaling	RT	▬	2	14.3	2.1E-2	6.2E-1
□	GOTERM_BP_DIRECT	intracellular receptor signaling pathway	RT	▬	2	14.3	2.5E-2	6.1E-1
□	GOTERM_BP_DIRECT	steroid metabolic process	RT	▬	2	14.3	2.8E-2	6.1E-1
□	GOTERM_BP_DIRECT	steroid hormone mediated signaling pathway	RT	▬	2	14.3	3.7E-2	6.7E-1
□	GOTERM_BP_DIRECT	transcription initiation from RNA polymerase II promoter	RT	▬	2	14.3	9.5E-2	9.3E-1

6 gene(s) from your list are not in the output.

Fig. 12.20 DAVID: Functional annotation chart

- These genes need to be classified according to their molecular functions, biological processes, and cellular locations to identify the major pathways operating in a particular biological condition (e.g., diseased state in which sequencing was performed). Such classification or grouping of genes is called gene enrichment. Genes are also clustered based on their functional annotation. Such functional clustering is essential to identify genes having similar functions. Such kind of functional annotation and clustering can be performed using DAVID.
- Data generated from exome sequencing have gene identifiers linked to respective variant information (e.g., in a diseased state).
- This gene set has to functionally annotate to predict the role of respective variants associated. Also, clustering of genes will recognize the genes with polymorphisms, belonging to similar molecular functions. This will give new leads toward building hypothesis on disease pathogenesis.

12.5.1.4 STRING

STRING (Szklarczyk et al. 2016) is a web-based tool for making protein-protein interaction networks.

Create a PPIN Using STRING

The tutorial is for the set of proteins you have.

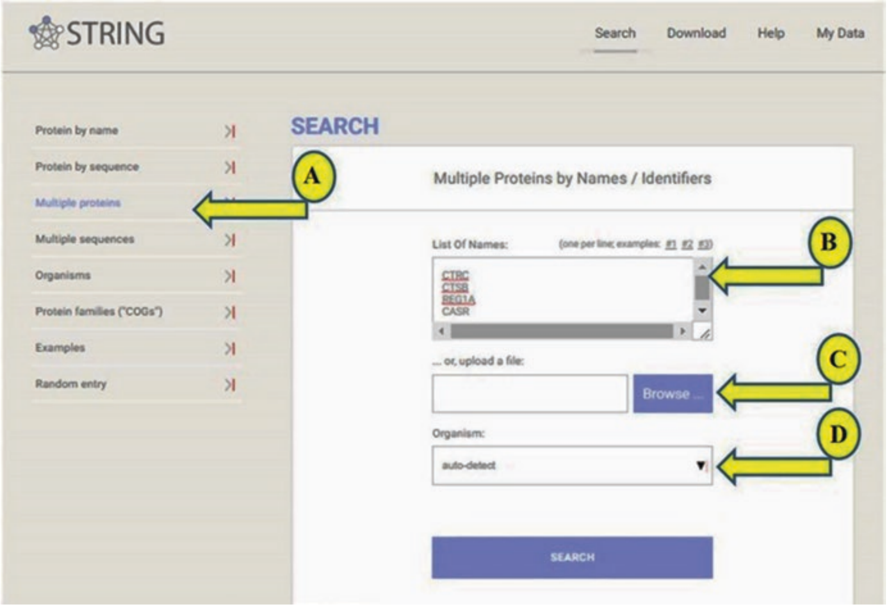
Step 1: You can search interaction network by clicking on “Multiple proteins” (Fig. 12.21 (A)) and paste a list of gene IDs into text box provided (Fig. 12.21 (B)) or load a text file containing gene IDs using “Browse” option (Fig. 12.21 (C)). In the organism field, you can specify organism name explicitly (e.g., Homo sapiens) or leave it to default as “auto-detect” (Fig. 12.21 (D)). Then click the search button (Fig. 12.21).

Step 2: You will be redirected to the page listing the gene symbols you have entered with their alias and function (Fig. 12.22). The user needs to ensure that specific protein of interest being queried. Then click on “Continue” button (Fig. 12.22 (A)).

Step 3: You will be redirected to a network page (Fig. 12.23). In the protein-protein interaction network (Fig. 12.23 (A)), the circles represent the nodes or proteins. The edges represent the associations between nodes. The legend (Fig. 12.23 (B)) section gives information about nodes and interacting partners or edges.

Step 4: User can change the research parameters from “Setting” section (Fig. 12.24 (A–D)).

Step 5: Visualize the Analysis section (Fig. 12.25). The Analysis section provides network statistics (Fig. 12.25 (A)). The functional enrichment analysis of the input gene set is provided in Fig. 12.25 (B). The information about the statistical background used for functional enrichment is also given in Fig. 12.25 (C).



Multiple Proteins by Names / Identifiers

List Of Names: (one per line; examples: #1 #2 #3)

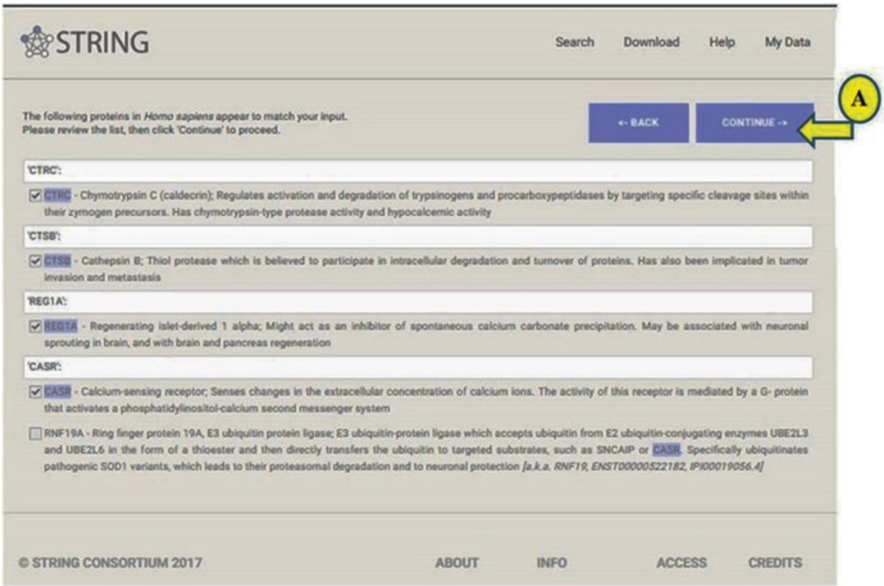
CTRC
CTSB
REG1A
CASR

... or, upload a file:

Organism:

auto-detect

Fig. 12.21 STRING: Use multiple protein identifiers input for PPIN construction



The following proteins in *Homo sapiens* appear to match your input.
Please review the list, then click 'Continue' to proceed.

'CTRC':

☑ CTRC - Chymotrypsin C (caldecrin); Regulates activation and degradation of trypsinogens and procarboxypeptidases by targeting specific cleavage sites within their zymogen precursors. Has chymotrypsin-type protease activity and hypocalcemic activity

'CTSB':

☑ CTSB - Cathepsin B; Thiol protease which is believed to participate in intracellular degradation and turnover of proteins. Has also been implicated in tumor invasion and metastasis

'REG1A':

☑ REG1A - Regenerating islet-derived 1 alpha; Might act as an inhibitor of spontaneous calcium carbonate precipitation. May be associated with neuronal sprouting in brain, and with brain and pancreas regeneration

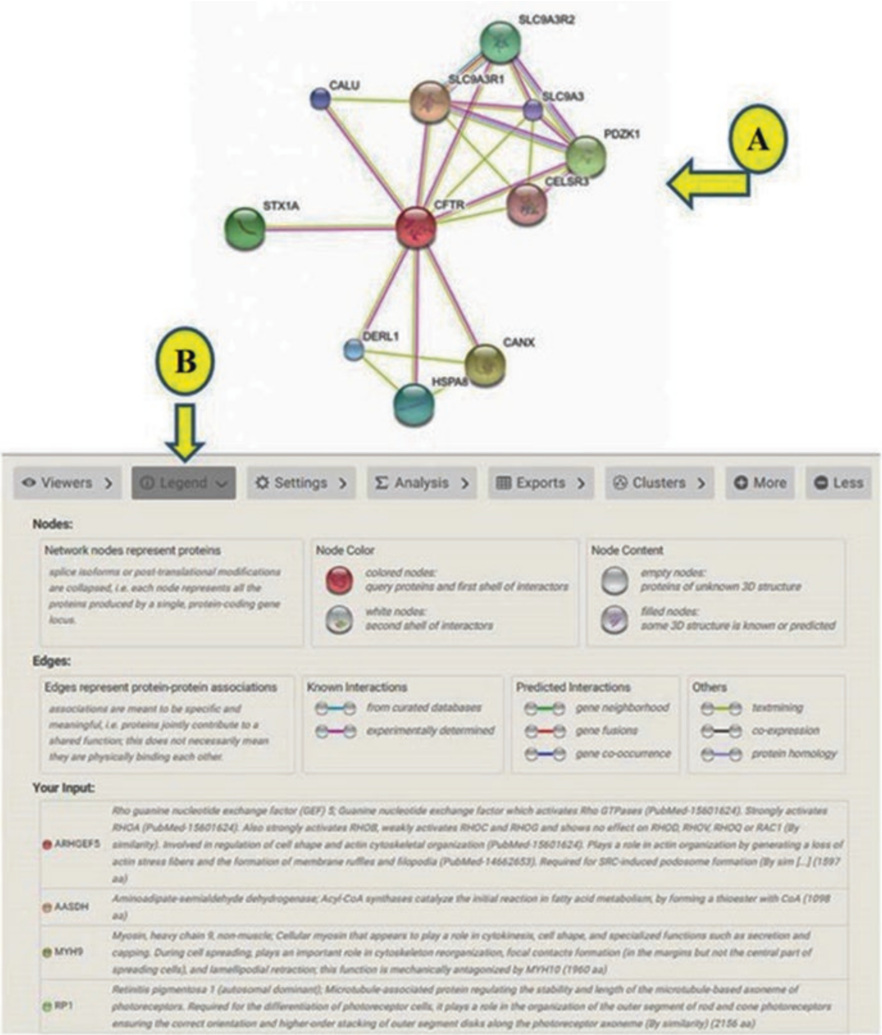
'CASR':

☑ CASR - Calcium-sensing receptor; Senses changes in the extracellular concentration of calcium ions. The activity of this receptor is mediated by a G- protein that activates a phosphatidylinositol-calcium second messenger system

☐ RNF19A - Ring finger protein 19A, E3 ubiquitin protein ligase; E3 ubiquitin-protein ligase which accepts ubiquitin from E2 ubiquitin-conjugating enzymes UBE2L3 and UBE2L6 in the form of a thioester and then directly transfers the ubiquitin to targeted substrates, such as SNCAIP or CASR. Specifically ubiquitinates pathogenic SOD1 variants, which leads to their proteasomal degradation and to neuronal protection *[a.k.a. RNF19, ENST00000522182, IPI00019056.4]*

© STRING CONSORTIUM 2017

Fig. 12.22 STRING: Ensuring the correct protein identifiers are being used for PPIN construction



Nodes:

Network nodes represent proteins

splice isoforms or post-translational modifications are collapsed, i.e. each node represents all the proteins produced by a single, protein-coding gene locus.

Node Color

colored nodes:
query proteins and first shell of interactors

white nodes:
second shell of interactors

Node Content

empty nodes:
proteins of unknown 3D structure

filled nodes:
some 3D structure is known or predicted

Edges:

Edges represent protein-protein associations

associations are meant to be specific and meaningful, i.e. proteins jointly contribute to a shared function; this does not necessarily mean they are physically binding each other.

Known Interactions

from curated databases
experimentally determined

Predicted Interactions

gene neighborhood
gene fusions
gene co-occurrence

Others

textmining
co-expression
protein homology

Your Input:

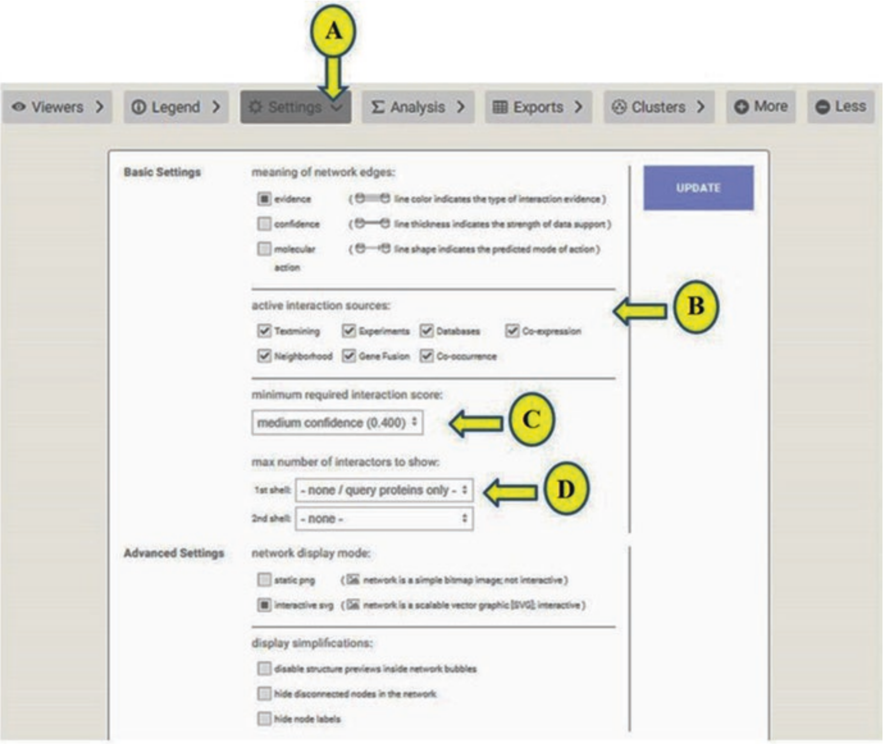
● ARHGEF5	*Rho guanine nucleotide exchange factor (GEF) 5; Guanine nucleotide exchange factor which activates Rho GTPases (PubMed-15601624). Strongly activates RHOA (PubMed-15601624). Also strongly activates RHOB, weakly activates RHOC and RHOG and shows no effect on RHOD, RHOV, RHOQ or RAC1 (By similarity). Involved in regulation of cell shape and actin cytoskeletal organization (PubMed-15601624). Plays a role in actin organization by generating a loss of actin stress fibers and the formation of membrane ruffles and filopodia (PubMed-14662653). Required for SRC-induced podosome formation (By sim [...] (1597 aa)*
● AASDH	*Aminoadipate-semialdehyde dehydrogenase; Acyl-CoA synthases catalyze the initial reaction in fatty acid metabolism, by forming a thioester with CoA (1098 aa)*
● MYH9	*Myosin, heavy chain 9, non-muscle; Cellular myosin that appears to play a role in cytokinesis, cell shape, and specialized functions such as secretion and capping. During cell spreading, plays an important role in cytoskeleton reorganization, focal contacts formation (in the margins but not the central part of spreading cells), and lamellipodial retraction; this function is mechanically antagonized by MYH10 (1960 aa)*
● RP1	*Retinitis pigmentosa 1 (autosomal dominant); Microtubule-associated protein regulating the stability and length of the microtubule-based axoneme of photoreceptors. Required for the differentiation of photoreceptor cells, it plays a role in the organization of the outer segment of rod and cone photoreceptors ensuring the correct orientation and higher-order stacking of outer segment disks along the photoreceptor axoneme (By similarity) (2156 aa)*

Fig. 12.23 STRING: Network visualization

Step 6: Finally you can export the network files (Fig. 12.26 (A)) in different formats (Fig. 12.26 (B)) to analyze it further using Cytoscape or any other network visualization tool(s).

12.5.1.5 Cytoscape

Cytoscape (Shannon et al. 2003) is an open source tool for visualizing biomolecular interaction networks, integrating functional annotations and high-throughput gene expression profiles into a unified conceptual framework, and identifying their



Basic Settings

meaning of network edges:

☒ evidence (line color indicates the type of interaction evidence)
□ confidence (line thickness indicates the strength of data support)
□ molecular action (line shape indicates the predicted mode of action)

active interaction sources:

☑ Textmining ☑ Experiments ☑ Databases ☑ Co-expression
☑ Neighborhood ☑ Gene Fusion ☑ Co-occurrence

minimum required interaction score:

medium confidence (0.400)

max number of interactors to show:

1st shell: - none / query proteins only -
2nd shell: - none -

Advanced Settings

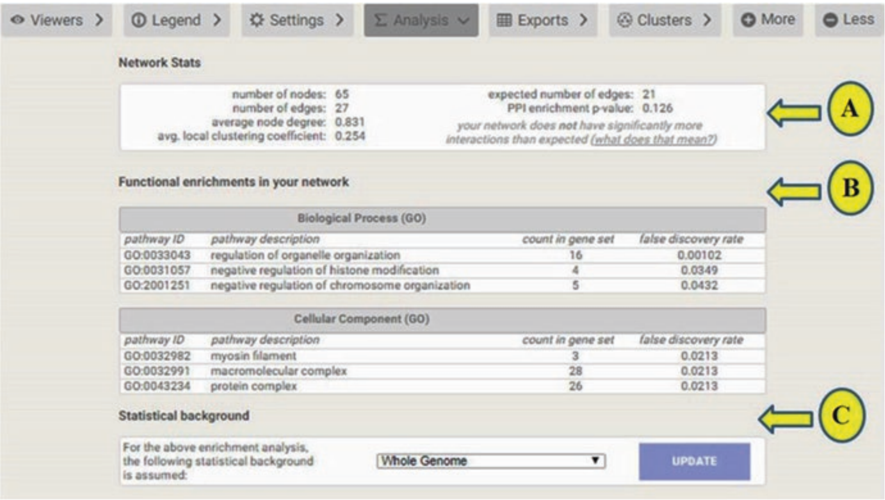
network display mode:

□ static png (network is a simple bitmap image; not interactive)
☒ interactive svg (network is a scalable vector graphic [SVG]; interactive)

display simplifications:

□ disable structure previews inside network bubbles
□ hide disconnected nodes in the network
□ hide node labels

Fig. 12.24 STRING: Change research parameters for PPIN construction



Network Stats

number of nodes: 65
number of edges: 27
average node degree: 0.831
avg. local clustering coefficient: 0.254

expected number of edges: 21
PPI enrichment p-value: 0.126
*your network does **not** have significantly more interactions than expected (what does that mean?)*

Functional enrichments in your network

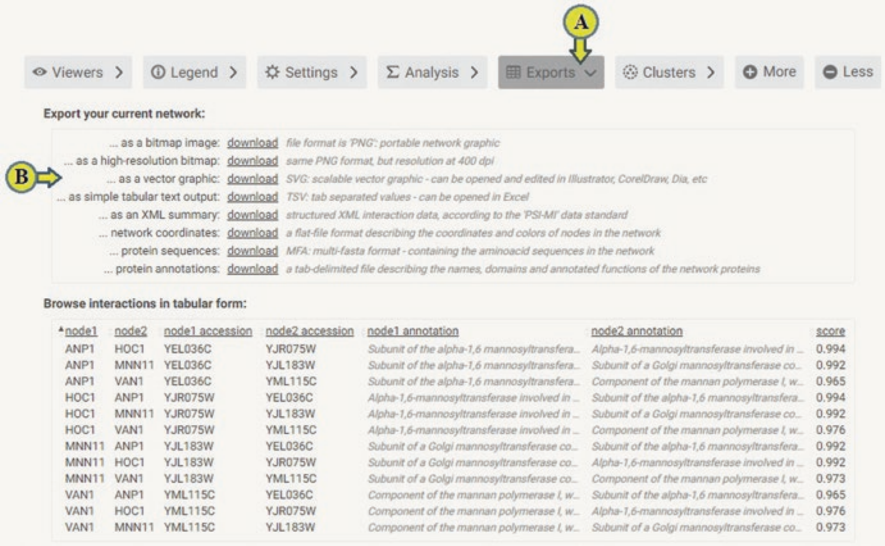
Biological Process (GO)			
pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
GO:0033043	regulation of organelle organization	16	0.00102
GO:0031057	negative regulation of histone modification	4	0.0349
GO:2001251	negative regulation of chromosome organization	5	0.0432

Cellular Component (GO)			
pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
GO:0032982	myosin filament	3	0.0213
GO:0032991	macromolecular complex	28	0.0213
GO:0043234	protein complex	26	0.0213

Statistical background

For the above enrichment analysis, the following statistical background is assumed: Whole Genome

Fig. 12.25 STRING: Analysis section providing network statistics and functional enrichment analysis of input protein identifiers



Export your current network:

... as a bitmap image:	download	*file format is 'PNG': portable network graphic*
... as a high-resolution bitmap:	download	*same PNG format, but resolution at 400 dpi*
... as a vector graphic:	download	*SVG: scalable vector graphic - can be opened and edited in Illustrator, CorelDraw, Dia, etc*
... as simple tabular text output:	download	*TSV: tab separated values - can be opened in Excel*
... as an XML summary:	download	*structured XML interaction data, according to the 'PSI-MI' data standard*
... network coordinates:	download	*a flat-file format describing the coordinates and colors of nodes in the network*
... protein sequences:	download	*MFA: multi-fasta format - containing the aminoacid sequences in the network*
... protein annotations:	download	*a tab-delimited file describing the names, domains and annotated functions of the network proteins*

Browse interactions in tabular form:

▴node1	node2	node1 accession	node2 accession	node1 annotation	node2 annotation	score
ANP1	HOC1	YEL036C	YJR075W	*Subunit of the alpha-1,6 mannosyltransfera...*	*Alpha-1,6-mannosyltransferase involved in ...*	0.994
ANP1	MNN11	YEL036C	YJL183W	*Subunit of the alpha-1,6 mannosyltransfera...*	*Subunit of a Golgi mannosyltransferase co...*	0.992
ANP1	VAN1	YEL036C	YML115C	*Subunit of the alpha-1,6 mannosyltransfera...*	*Component of the mannan polymerase I, w...*	0.965
HOC1	ANP1	YJR075W	YEL036C	*Alpha-1,6-mannosyltransferase involved in ...*	*Subunit of the alpha-1,6 mannosyltransfera...*	0.994
HOC1	MNN11	YJR075W	YJL183W	*Alpha-1,6-mannosyltransferase involved in ...*	*Subunit of a Golgi mannosyltransferase co...*	0.992
HOC1	VAN1	YJR075W	YML115C	*Alpha-1,6-mannosyltransferase involved in ...*	*Component of the mannan polymerase I, w...*	0.976
MNN11	ANP1	YJL183W	YEL036C	*Subunit of a Golgi mannosyltransferase co...*	*Subunit of the alpha-1,6 mannosyltransfera...*	0.992
MNN11	HOC1	YJL183W	YJR075W	*Subunit of a Golgi mannosyltransferase co...*	*Alpha-1,6-mannosyltransferase involved in ...*	0.992
MNN11	VAN1	YJL183W	YML115C	*Subunit of a Golgi mannosyltransferase co...*	*Component of the mannan polymerase I, w...*	0.973
VAN1	ANP1	YML115C	YEL036C	*Component of the mannan polymerase I, w...*	*Subunit of the alpha-1,6 mannosyltransfera...*	0.965
VAN1	HOC1	YML115C	YJR075W	*Component of the mannan polymerase I, w...*	*Alpha-1,6-mannosyltransferase involved in ...*	0.976
VAN1	MNN11	YML115C	YJL183W	*Component of the mannan polymerase I, w...*	*Subunit of a Golgi mannosyltransferase co...*	0.973

Fig. 12.26 STRING: Export required network files

properties. Additional utilities are available in the form of plugins. Plugins are available for network properties and molecular profiling analyses, various layouts for better visualization, additional file format support, and connection with databases and searching in large networks. Cytoscape additionally has a JavaScript-driven sister venture named *Cytoscape.js* that can be utilized to dissect and visualize networks in JavaScript environments through a web browser.

Examples of Uses

Gene function prediction – examining genes (proteins) in a network context shows connections to sets of genes/proteins involved in the same biological process that is likely to function in that process (plugin for analysis: jActiveModules, PiNGO, etc.).

Detection of protein complexes/other modular structures – protein complexes are groups of associated polypeptide chains whose malfunctions play a vital role in disease development. Complexes can perform various functions in the cell, including dynamic signaling, and can serve as cellular machines, rigid structures, and posttranslational modification systems. Many disorders are consequences of changes in a single protein and, thus, in its set of associated partners and functionality (plugin for analysis: Motif Discovery, Mclique, MCODE, PEWCC, etc.).

Identification of disease sub-networks and potential biomarkers – identification of disease network sub-networks that are transcriptionally active in the disease and also provide a rich source of biomarkers for disease classification. These suggest key pathway components in disease progression and provide leads for further study and potential therapeutic targets (plugin for analysis: PhenomeScape, PSFC, etc.).

Dynamics of a biological network – the molecular interactions in a cell vary with time and surrounding environmental conditions. The construction and analysis of dynamic molecular networks can elucidate dynamic cellular mechanisms of different biological functions and provide a chance to understand complex diseases at the system level (plugin for analysis: DyNetViewer, DynNetwork, DynNet, etc.).

INPUT Type

Cytoscape can read network/pathway files written in the following formats:

- Simple interaction file (SIF or .sif format)
- Nested network format (NNF or .nnf format)
- Graph Markup Language (GML or .gml format)
- XGMML (eXtensible Graph Markup and Modeling Language)
- SBML
- BioPAX
- PSI-MI Level 1 and 2.5
- GraphML
- Delimited text
- Excel Workbook (.xls, .xlsx)
- Cytoscape.js JSON
- Cytoscape CX

The SIF format specifies nodes and interactions only, while other formats store additional information about network layout and allow network data exchange with a variety of other network programs and data sources.

Visualization

Substantial progress has been made in the field of "omics" research (e.g., genomics, transcriptomics, proteomics, and metabolomics), leading to a vast amount of biological data generation. In order to represent large biological data sets in an easily interpretable manner, this information is frequently visualized as graphs, i.e., a set of nodes and edges. Cytoscape assists in visual exploration and analysis of biological network in several ways:

- Provides customize network data display using powerful visual styles.
- Helps in integrating gene expression values with the network. This can be done by mapping expression values to network nodes which represent the gene as

color, label, border thickness, etc. according to the user-defined mapping file and provide several layout options in two as well as three dimensions for network visualization, for example, edge-weighted spring-embedded layout, attribute circle layout, etc.

- The network manager can be utilized to manage multiple networks in a single session file. Easily navigate large networks through an efficient rendering engine.

Analysis

- Filter the network to select subsets of nodes and/or interactions based on the current data. For instance, users may select nodes involved in a threshold number of interactions, nodes that share a particular GO annotation, or nodes whose gene expression levels change significantly in one or more conditions according to p-values loaded with the gene expression data.
- Find active sub-networks/pathway modules. The network is screened against gene expression data to identify connected sets of interactions, i.e., interaction sub-networks, whose genes show particularly high levels of differential expression. The interactions contained in each sub-network provide hypotheses for the regulatory and signaling interactions in control of the observed expression changes.
- Find clusters (highly interconnected regions) in any network loaded into Cytoscape. Depending on the type of network, clusters may mean different things. For instance, clusters in a protein-protein interaction network have been shown to be protein complexes and parts of pathways. Clusters in a protein similarity network represent protein families.
- Plugins available for network and molecular profile analysis.

12.6 Conclusion

Complex biological networks are the reservoir for the plethora of biological information about pathways and cellular mechanisms. This chapter summarized different types of biological networks, methodologies to analyze such networks and biological relevance. These networks can provide researchers with critical information about the pathogenesis of diseases (disease-gene networks), identification of drug targets (protein-protein networks, protein-ligand interaction), and biological pathways. Functional and pathway analysis of genes (gene ontology) determine significant gene clusters associated with a specific biological process, molecular function or pathway. This chapter succinctly provides relevant information about the applications of biological networks in the molecular biology field. Our hope is that the tutorials provided in this chapter will guide researchers to annotate genes on gene products and enrich GO annotation both qualitatively and quantitatively on the available tools.

References

Albert R (2005) Scale-free networks in cell biology. J Cell Sci 118:4947–4957

Alberts B et al (2002) Molecular biology of the cell, 4th edn. Garland Science, New York

Arita M (2004) The metabolic world of Escherichia coli is not small. Proc Natl Acad Sci U S A 101:1543–1547

Bader S, Kühner S, Gavin AC (2008) Interaction networks for systems biology. FEBS Lett 582(8):1220–1224

Batada NN, Reguly T, Breitkreutz A, Boucher L, Breitkreutz BJ, Hurst LD et al (2006) Stratus not altocumulus: a new view of the yeast protein interaction network. PLoS Biol 4:e317

Beadle GW, Tatum EL (1941) Genetic control of biochemical reactions in Neurospora. Proc Natl Acad Sci U S A 27:499–506

Binns D et al (2009) QuickGO: a web-based tool for Gene Ontology searching. Bioinformatics 25(22):3045–3046

Blake JA (2013) Ten quick tips for using the gene ontology. PLoS Comput Biol 9(11):e1003343

Botstein D et al (2000) Gene Ontology: tool for the unification of biology. Nat Genet 25(1):25–29

Carbon S et al (2008) AmiGO: online access to ontology and annotation data. Bioinformatics 25(2):288–289

Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL (2007) The human disease network. Proc Natl Acad Sci 104(21):8685–8690

Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJ, Cusick ME, Roth FP, Vidal M (2004) Evidence for dynamically organized modularity in the yeast protein–protein interaction network. Nature 430(6995):88–93

Hill DP et al (2008) Gene Ontology annotations: what they mean and where they come from. BMC Bioinf 9(5):S2; BioMed Central

Huang DW, Sherman BT, Lempicki RA (2008) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4(1):44

Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinf 9(1):559

Ma'ayan A (2011) Introduction to network analysis in systems biology. Sci Signal 4(190):tr5

Reference Genome Group of the Gene Ontology Consortium (2009) The Gene Ontology's Reference Genome Project: a unified framework for functional annotation across species. PLoS Comput Biol 5(7):e1000431

Shannon P et al (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13(11):2498–2504

Szklarczyk D et al (2016) The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic Acids Res. https://doi.org/10.1093/nar/gkw937

Weirauch MT (2011) Gene coexpression networks for the analysis of DNA microarray data. In: Applied statistics for network biology: methods in systems biology. Wiley-Blackwell, Weinheim, pp 215–250

Yook SH, Oltvai ZN, Barabási AL (2004) Functional and topological characterization of protein interaction networks. Proteomics 4(4):928–942

Chapter 13
Metabolomics

Peter Natesan Pushparaj

Contents

Abbreviations

CD	Central Dogma
CE	Capillary Electrophoresis
CP	Chronic Pancreatitis
EI	Electron Impact Ionization

P. N. Pushparaj (✉)
Center of Excellence in Genomic Medicine Research, Faculty of Applied Medical Sciences, King Abdulaziz University, Jeddah, Saudi Arabia

N. A. Shaik et al. (eds.), *Essentials of Bioinformatics, Volume I*,
https://doi.org/10.1007/978-3-030-02634-9_13

ESI	Electrospray Ionization
GC	Gas Chromatography
KEGG	Kyoto Encyclopedia of Genes and Genomes
LC	Liquid Chromatography
MALDI	Matrix-Assisted Laser Desorption and Ionization
MS	Mass Spectrometry
NMR	Nuclear Magnetic Resonance Spectroscopy
OPLS-DA	Orthogonal Partial Least Squares Discriminant Analysis
PCA	Principal Component Analysis
PDAC	Pancreatic Ductal Adenocarcinoma
QTOF	Quadrupole Time of Flight
TW-IMS	Traveling Wave Ion Mobility Spectrometry

13.1 Introduction

Metabolomics is a rapidly evolving "omics" approach to systematically decode the complex small molecules termed as metabolites present in the biological systems (Bartlett and Chen 2016; Jones and Cheung 2007; Schnackenberg 2006; Tan et al. 2016). Metabolomics is the comprehensive study of metabolites generated from the substrates and products of metabolism, within cells, tissues, organisms, and biological fluids. Together, these small molecules and their interactions within a biological system are known as the metabolome. Metabolomics helps to get a panoramic view of an array of metabolites that are implicated in diverse and intricate cellular, molecular, and physiological processes in living systems. Aptly, more focus has been attributed to the initiation and the development of metabolomics-related research in academic institutions, industry, and government organizations around the globe in the past decade. It is clearly evident by just two publications in the year 2000 to nearly 27,000 documents till the year 2018 in metabolomics research (Fig. 13.1). Investigating the metabolome is essential to understand the subtle changes in metabolites and the subsequent impact on molecular networks or pathways in health and several disease conditions. The number of publications using the integration of metabolomics, functional genomics, transcriptomics, and proteomics has been ever increasing from just 1 in the year 2001 to 1221 till the end of 2017 (Fig. 13.2).The rapid emergence of "systems biology" helps to integrate the massive wealth of data derived from these *multi-omics* platforms with metabolomics approach to further interpret or characterize the complex biological processes (Fig. 13.3).

Metabolomics is a robust method since metabolites and their concentrations directly reveal the ongoing biochemical reactions in the cells, tissues, organs, and biological fluids. Importantly, the profile of metabolites in the biological systems depends on both genetic and environmental factors that further influence both anabolism and catabolism. In fact, the metabolomics is an integral part of the central dogma (CD) of molecular biology (Fig. 13.4) and perfectly characterizes the molecular phenotype.

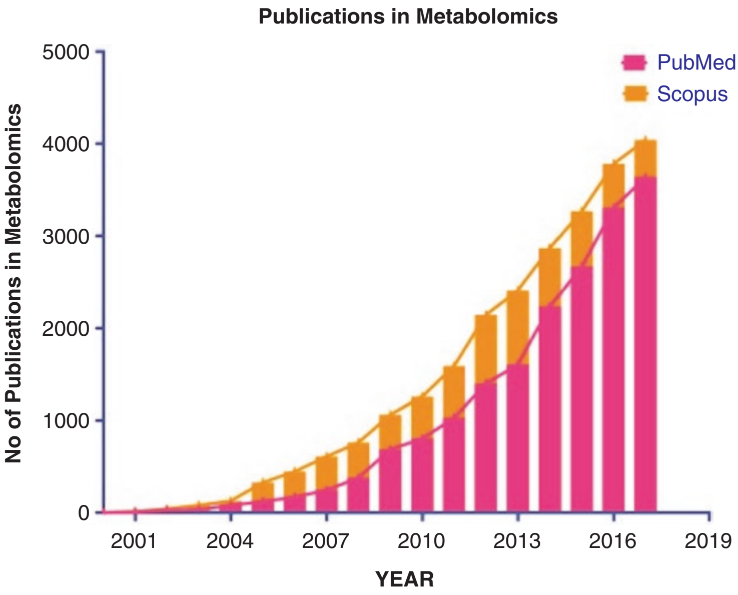
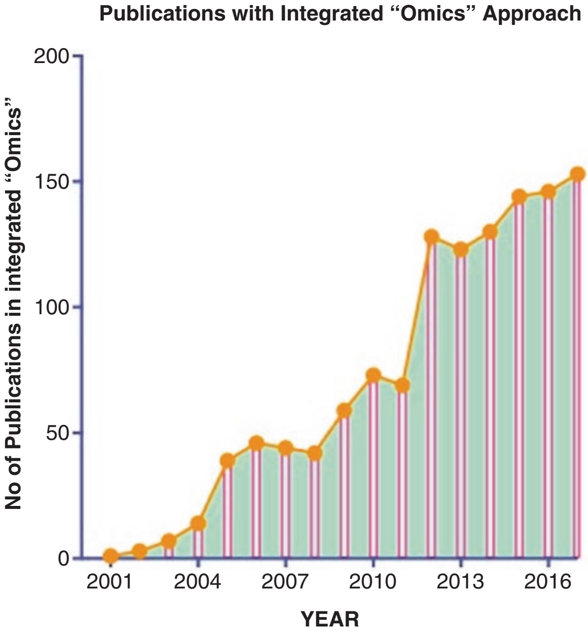


Fig. 13.1 The number of publications in metabolomics. The bibliometric analyses using PubMed and Scopus have shown that only two documents were published related to the metabolomics in the year 2000, and it has increased remarkably to 27000 till the middle of 2018

Fig. 13.2 The number of publications in integrated" omics" approach. The bibliometric analysis using Scopus have shown that only one document was published related to the integrated "omics" approach in the year 2001. However, it has increased to 1221 till the year 2017



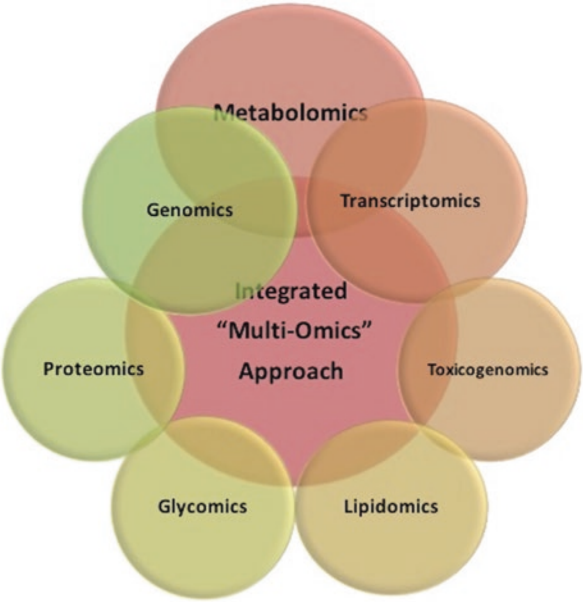


Fig. 13.3 Integrated "multi-omics" approach in precision medicine. The rapid emergence of "systems biology" helps to integrate the massive wealth of data derived from these "multi-omics" platforms with metabolomics approach to further interpret or characterize the complex biological processes

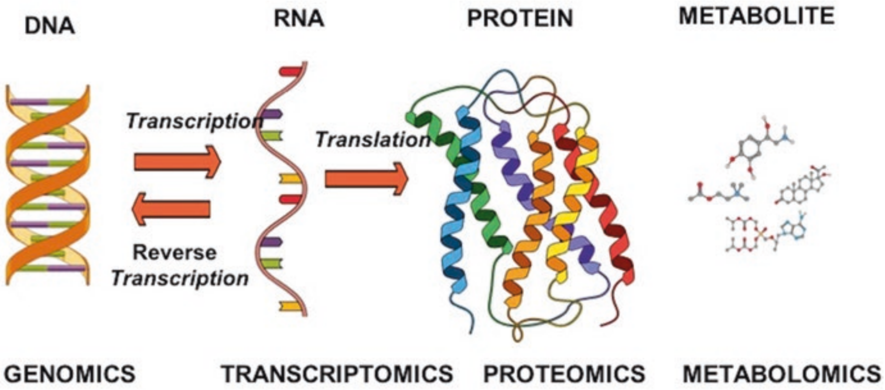


Fig. 13.4 Metabolomics and central dogma. Metabolomics is an integral part of the central dogma (CD) of molecular biology

13.2 Analytical Methods in Metabolomics

A metabolite is a small low molecular compound involved in complex metabolic reactions (Fig. 13.5). The metabolomics typically investigates small molecules that are less than 1500 daltons (Da) in molecular weight such as amino acids, lipids, sugars, fatty acids, phenolic compounds, etc. There are currently about 2900–3000 common or endogenous metabolites identified in the human body participating in complex metabolic reactions. In order to profile these metabolites, three different methods are used, namely, untargeted or global method, targeted method, and imaging method.

13.2.1 Untargeted or Global Metabolomics Method

Global or untargeted method determines most of the metabolites without any bias from the biological samples. It is the method of choice for studying the small molecular or metabolic phenotypes in basic, applied, and translational research programs around the globe.

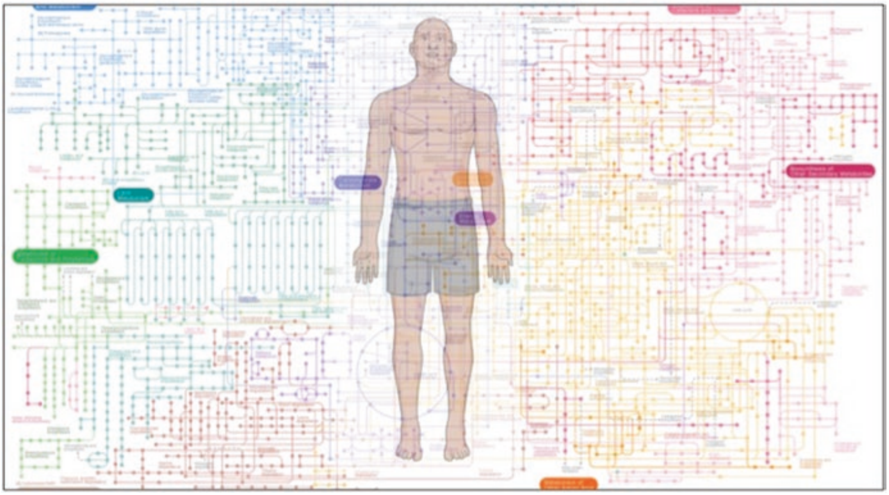


Fig. 13.5 Intricacy of metabolic reactions. Intricate network of metabolites as depicted by KEGG metabolic pathways. Metabolomics typically investigates small molecules that are less than 1500 daltons (Da) in molecular weight such as amino acids, lipids, sugars, fatty acids, phenolic compounds, etc. There are currently about 2900 to 3000 common or endogenous metabolites identified in the human body participating in complex metabolic reactions. (Adapted and modified from KEGG metabolic pathways)

13.2.2 Targeted Metabolomics Method

The targeted method provides the analytical validation of metabolites measured by the global or untargeted method (Johnson et al. 2016). It quantifies the concentrations of a specific or defined set of metabolites implicated in a particular metabolic reaction. Hence, preparation of standard curves for each metabolite of interest is essential for the precise quantitation in biological samples. It is mostly used when investigating pharmacokinetics of drugs, therapeutic outcomes, or genetic modifications of a protein or an enzyme implicated in health and disease (Johnson et al. 2016).

13.2.3 Imaging Metabolomics Method

Nanostructure imaging mass spectrometry (NIMS), desorption electrospray ionization mass spectrometry (DESI), secondary ion mass spectrometry (SIMS), and matrix-assisted laser desorption ionization (MALDI) are used in imaging the location of the metabolites in tissues (Johnson et al. 2016).

However, there is no difference in the metabolomics pipeline between the global and untargeted methods. The mass spectrometry-based global or untargeted metabolomics methods typically consist of steps such as experimental design, sample preparation strategy, injection of sample either direct infusion or mainly via gas or liquid chromatography (GC/LC), mass spectrometric (MS) analysis, acquisition and processing of MS data, and analysis and biological interpretation of MS data (Fig. 13.6).

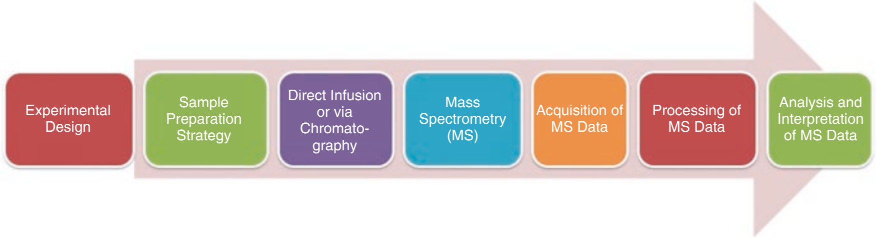

Experimental Design
Sample Preparation Strategy
Direct Infusion or via Chromato-graphy
Mass Spectrometry (MS)
Acquisition of MS Data
Processing of MS Data
Analysis and Interpretation of MS Data

Fig. 13.6 Targeted or untargeted metabolomics methods. An overview of the mass spectrometry-based metabolomics pipeline. There is no difference in the metabolomics pipeline between the global and untargeted methods. The mass spectrometry-based global or untargeted metabolomics methods typically consist of steps such as experimental design, sample preparation strategy, injection of sample either direct infusion or mainly via gas or liquid chromatography (GC/LC), mass spectrometric (MS) analysis, acquisition and processing of MS data, and analysis and biological interpretation of MS data

13.3 Experimental Design

Experimental design is very much important before the initiation of the metabolomics studies. The collection of appropriate type of samples, sample size, random sampling (randomization), sample preparation strategies, and suitable storage of samples are essential before the metabolomics analysis using either MS or NMR. Importantly, identification of external as well as internal factors that affect the results is essential and has to be taken into consideration during the data processing, analysis, and interpretation. This will certainly reduce error and increase the data reproducibility in metabolomics experiments.

13.4 Preparation of Samples for Metabolomics Experiment

The sample preparation in metabolomics involves many steps such as sample collection, storage, extraction, and preparation (Fig. 13.7).

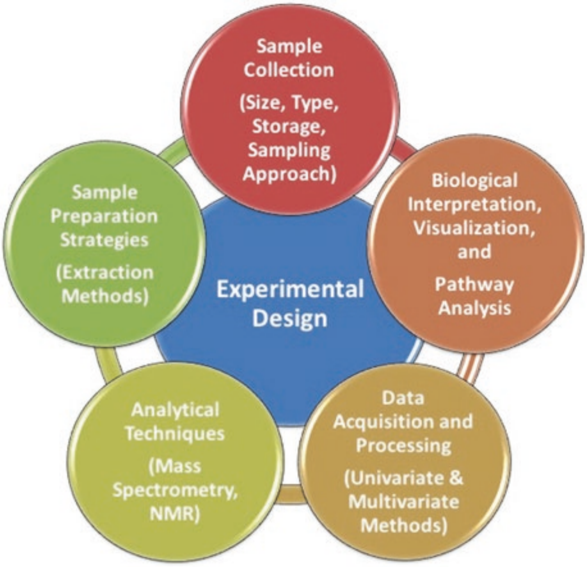

Sample Collection
(Size, Type, Storage, Sampling Approach)
Biological Interpretation, Visualization, and
Pathway Analysis
Data Acquisition and Processing
(Univariate & Multivariate Methods)
Analytical Techniques
(Mass Spectrometry, NMR)
Sample Preparation Strategies
(Extraction Methods)
Experimental Design

Fig. 13.7 Experimental design in metabolomics. Experimental design is very much important before the initiation of the metabolomics studies. The collection of appropriate type of samples based on the objectives of the experiments; sample size, random sampling (randomization), sample preparation strategies, and suitable storage of samples before the metabolomics analysis either in MS or NMR. Importantly, identification of external as well as internal factors that affect the results is essential and has to be taken into consideration during the data processing, analysis, and interpretation. This will certainly reduce error and increase the data reproducibility in metabolomics experiments.

13.4.1 *Extraction Methods*

In metabolomics, the sample extraction can be performed using solid-phase extraction, gas chromatography, and liquid chromatography.

13.4.1.1 Solid-Phase Extraction

In solid-phase extraction (SPE), the metabolites suspended in a liquid mobile phase are separated from other compounds based on their affinity and other physiochemical interactions with a sorbent or a solid separation media. SPE is mostly used to concentrate, clean up, and partially purify a sample before further clarification using either gas or liquid chromatographic methods and analysis using either MS or NMR techniques (Fig. 13.8). Based on the solid separation media used, the SPE can be classified as normal-phase SPE, reversed-phase SPE, and ion-exchange SPE.

A recent technique termed as microextraction by packed sorbent (MEPS) is used for the isolation of drugs and metabolites from the biological fluids such as the whole blood or plasma or serum, etc. MEPS can be used for small sample volumes such as 10 μL. MEPS has been used in a number of latest research investigations in preclinical, clinical, and environmental analysis, and it also has advantages compared to other extraction techniques such as solid-phase microextraction method (SPME) (Abdel-Rehim 2011; Ma and Ouyang 2016).

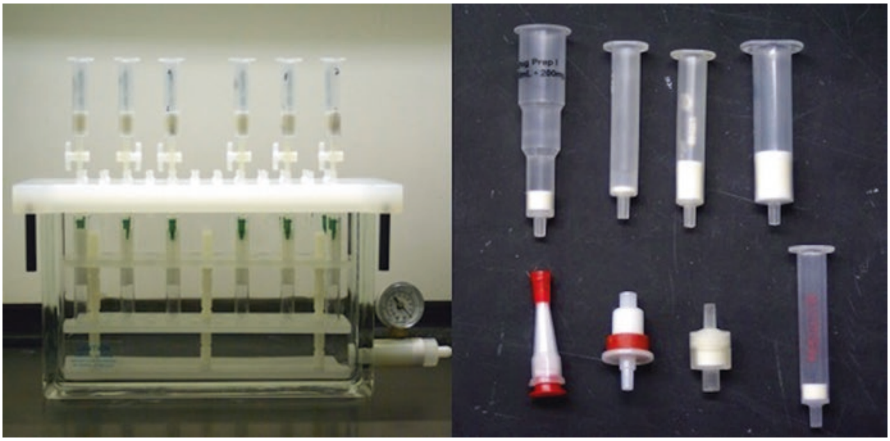


Fig. 13.8 Solid-phase extraction apparatus and a variety of column types used in the preprocessing of samples in metabolomics experiments. (Image courtesy: Jeff Dahl shared through Wikipedia Commons by CC BY-SA 3.0 license)

13.4.1.2 Chromatographic Methods

Several types of chromatographic methods (*low, medium, or high-throughput*) are used to separate metabolites from a mixture, and this is a key step before infusing the samples into MS and NMR. Gas and liquid chromatography are mostly coupled with MS analysis in metabolomics experiments. The separation is based on the interactions of individual metabolites in a sample with the mobile phase and the stationary solid phase and eluted from the solid phase using their physiochemical properties. In high-throughput liquid chromatography (HPLC), the column with a varied width and length can be packed with reversed phase (RP), normal phase, and ion-exchange stationary phases based on the objectives of the metabolomics analysis.

13.5 Mass Spectrometry

Mass spectrometry (MS) is used to measure small molecules or metabolites that are injected either directly (*direct infusion method*) or indirectly through a coupled method such as chromatography. The direct infusion methods such as electrospray ionization (ESI), electron impact ionization (EI), quadrupole time of flight (QTOF), and traveling wave ion mobility spectrometry (TW-IMS) as in Synapt G2 MS system (Waters Corporation, USA) are being used in metabolomics methods. On the other hand, a coupled gas (GC) or liquid chromatography (LC) or capillary electrophoresis (CE) can be used to isolate or purify metabolites in samples before infusing (*indirect infusion method*) into the MS system. The coupled techniques with MS, such as LC-MS/MS in RPLC or HILIC mode, GC-MS/MS, and CE-MS/MS, are especially used in targeted analysis (Drouin et al. 2017).

Basically, after infusion (direct or indirect), the metabolites in the sample are ionized using an ion source before the detection of metabolites in the mass detector (Fig. 13.9). The resulting MS data consists of mass-to-charge (m/z), time, and intensity triplets giving the mass, the strength of the ion beam, and the time of detection in the MS for each ion.

(a) *Sample Inlet*: The samples are injected through the sample inlet into the MS either direct infusion without prior separation of metabolites or coupled with a chromatographic system for the purification of samples, before entering the ion source. Importantly, the choice of reconstitution of samples using solvents, such as water, methanol, etc., after SPE or MEPS or any other extraction process significantly influences the number of metabolites identified using MS system (Lindahl et al. 2017a).

(b) *Ionization Source*: The sample is ionized using a high-energy electron beam into cations by the loss of an electron in vacuum. Ionization methods can be classified into hard ionization and soft ionization. In hard ionization method,

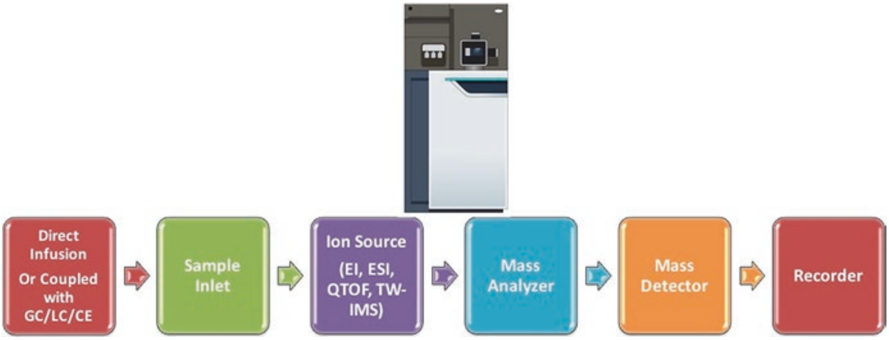

Direct Infusion Or Coupled with GC/LC/CE
Sample Inlet
Ion Source (EI, ESI, QTOF, TW-IMS)
Mass Analyzer
Mass Detector
Recorder

Fig. 13.9 Synapt G2 mass spectrometer (Waters Corporation, USA) and the components of a typical mass spectrometer. The sample can be injected into the inlet of the MS either directly or indirectly for ionization, mass analyses, mass detection, and recording of the resultant MS profile based on the m/z ratio of the molecular ion (M+) and other smaller molecular fragments. Please note that the latest Synapt G2-Si uses TW-IMS for the ionization of metabolites

such as electron impact ionization (EI), the high-energy electrons interact with the metabolites to generate heavy fragments. In contrast, soft ionization, such as electrospray ionization (ESI) as well as matrix-assisted laser desorption and ionization (MALDI) (a solid-phase technique that uses laser for ionization), ionizes metabolites and produces only few fragments. John Fenn and Koichi Tanaka won a share of the 2002 Nobel Prize in Chemistry for their discovery of soft ionization methods, ESI and MALDI, respectively (Cook 2002).

When a high-energy electron beam hits or collides with a molecule (M), it ionizes it to generate a molecular ion (M+). In addition, neutral pieces and smaller fragment ions are generated from the fragmentation of M+ ion due to residual energy. The molecular ion is a radical cation, but the fragment ions may either be radical cations or carbocations, depending on the nature of the neutral fragment. The MS spectrum of hexanoylcarnitine (C13H25NO4 molecular weight: 259.3419) showing the molecular ion (M+ 260.1) is given as an example in Fig. 13.10.

(c) *Mass Analyzer*: The ions produced in the ionizer are separated based on their m/z ratio in the mass analyzer where the m/z ration is equal to the molecular mass of the ion (charge is mostly equal to 1). All types of MS utilize the mass and electrical charge properties of ions but the separation techniques might vary.

(d) *Detector*: The ions separated based on the m/z ratio in the mass analyzer are then detected based on the m/z ratio in a mass detector.

(e) *Recorder*: The mass spectrophotometric profile of an ion, produced as a continuous ion beam, is recorded at different time intervals.

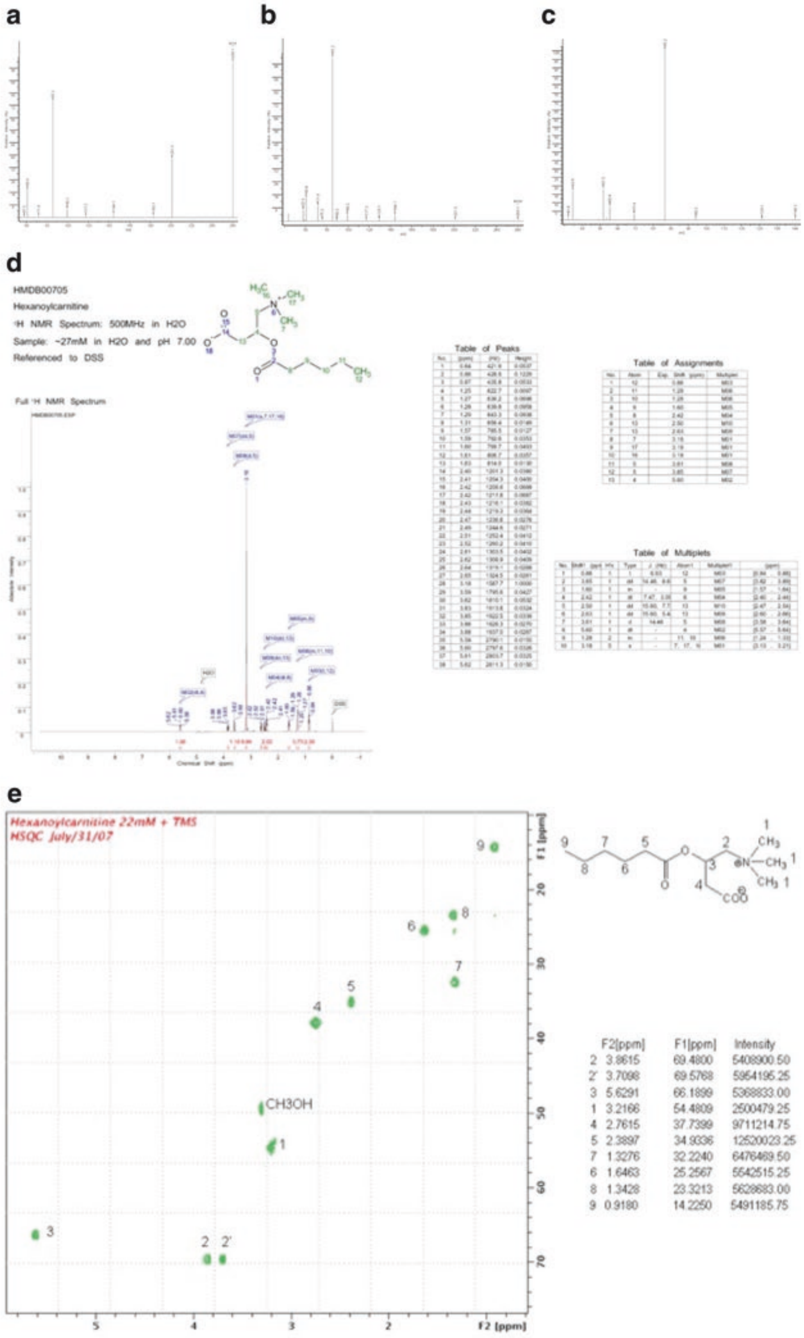


Fig. 13.10 (**a**) LC-MS/MS spectrum of hexanoylcarnitine using Quattro_QQQ 10V, (**b**) LC-MS/MS spectrum of hexanoylcarnitine using Quattro_QQQ 25V, (**c**) LC-MS/MS spectrum of hexanoylcarnitine using Quattro_QQQ 40V by positive ionization method, (**d**) 1H-NMR (1D) spectrum of hexanoylcarnitine, (**e**) 1H-13C NMR (2D) spectrum of hexanoylcarnitine (all the spectra (MS and NMR) were obtained from the HMDB version 4.0, Wishart et al. 2018)

13.6 Nuclear Magnetic Resonance (NMR)

NMR is the only experimental technique that can determine the structures and dynamics of biological molecules and their molecular complexes with atomic resolution. Kurt Wuthrich shared the 2002 Nobel Prize for Chemistry for his pioneering efforts in developing and applying NMR (Palmer and Patel 2002). It is a noninvasive or nondestructive analytical technique for the detection of both organic and inorganic compounds in the biological samples (solid and liquid). It is a robust method for the identification of new metabolites or biomarkers in health and disease. In the presence of an external magnetic field, an atom in a sample absorbs radio-frequency photon which promotes a nuclear spin from its ground state to its excited state. Hence, in NMR, the atoms reemit electromagnetic radiation with a specific resonating frequency termed as chemical shifts (δ). Importantly, the chemical shifts depend on an array of factors such as the magnetic properties of the atoms' isotopes, strength of the magnetic field, sample integrity, etc. In the case of metabolomics, proton atoms from small molecules are investigated using ^{1}H-NMR (1D) and ^{1}H-^{13}C-NMR (2D). In NMR, the resulting signal from small molecules' protons resonating within a magnetic field will be measured. The NMR can also be hyphenated or coupled with HPLC, SPE, CE, etc. for increasing the sensitivity and the detection of metabolites from the biological samples.

13.7 Comparison of MS and NMR

MS and NMR are the analytical tools that are routinely used in metabolomics experiments. NMR is quantitative and highly sensitive and requires less amount of sample, and tissues can also be analyzed. NMR is limited to detect abundant metabolites ($\geq$1 μM), a lower resolution, and dynamic range (up to 10^{2}). Moreover, NMR instrument occupies more space in the laboratory and the cost of the instrumentation is very high. However, the MS has the ability to measure metabolites at very low concentrations (femtomolar to attomolar) and has a higher resolution ($\sim 10^{3}$–10^{4}) and dynamic range ($\sim 10^{3}$–10^{4}), but quantitation is a challenge and sample complexity may limit metabolite detection because of ion suppression. MS instrumentation is cheaper, but the cost of analyses per sample is much higher and the sample preparation is more complex compared to NMR.

The MS and NMR (1D and 2D) spectra of hexanoylcarnitine (HMDB000705) have been given in the Fig. 13.10a–e. Please note the change in the MS spectra of hexanoylcarnitine based on the collision voltage strength (Fig. 13.10a (*10 V*), Fig. 13.10b (*25 V*), and Fig. 13.10c (*40 V*)) in the ionization process.

13.8 Coupling of MS and NMR Techniques

The coupling or hyphenation of NMR and MS greatly augments the results of metabolomics studies (Marshall and Powers 2017). The development of shielded magnets has greatly helped to couple LC, NMR, and MS instrumentation and increases the sensitivity and coverage of metabolite detection in biological samples (Fig. 13.11a, b). The number of publications cited in Scopus for “mass spectrometry” and metabolomics, “nuclear magnetic resonance” and metabolomics, and the combined search term “mass spectrometry” and “nuclear magnetic resonance” and metabolomics till the year 2017 is 36,926, 12,553, and 7412, respectively (Fig. 13.12).The LC-NMR-MS coupled method has greatly enhanced the concentration sensitivity by tenfold and improved the mass sensitivity by 1000-fold (Marshall and Powers 2017; Lin et al. 2008). A direct infusion FT-ICR-MS coupled with 1D and 2D NMR techniques was successfully used to distinguish isotopomers of glycerophospholipids (GPL) derived from [U-13C]-glucose in the extracts of MCF7-LCC2 cells (Marshall and Powers 2017; Lane et al. 2009a). Recently, SUMMIT MS/NMR, a direct infusion combined high-throughput approach, has been developed for the rapid and accurate identification of metabolites in complex biological samples (Bingol and Bruschweiler 2015a, b; Bingol et al. 2015a).

13.9 Processing and Analysis of Metabolomics Data

The peak heights for the internal standards should be continuously monitored during MS experiments. However, the presence of noise has to be taken into consideration since it distorts the signal from the MS. There are two types of noise present in the metabolomics data, namely, random noise and systematic noise. The random noise occurs due to the presence of contaminants and general technical problems in the system. It causes non-specific spikes and discontinuous or aberrant data. On the other hand, the systematic noise results from external factors like the baseline shift or drift (uneven baseline) observed in LC-MS caused by the gradient of the mobile solvent phase. Hence, the proper identification and reduction of noise in the MS data is essential to get reproducible data in the metabolomics experiments.

Data processing includes several similar steps in MS and NMR to extract biological relevant information from the metabolomics datasets. A feature matrix contains relative intensities of m/z ratios of ions and chemical shifts (ppm) from MS and NMR experiments, respectively. The statistical concepts used to analyze the high-throughput metabolomics data are broadly divided into univariate and multivariate approaches. Both approaches are used to analyze the metabolomics data in tandem and each provides exclusive information about the metabolomics datasets.

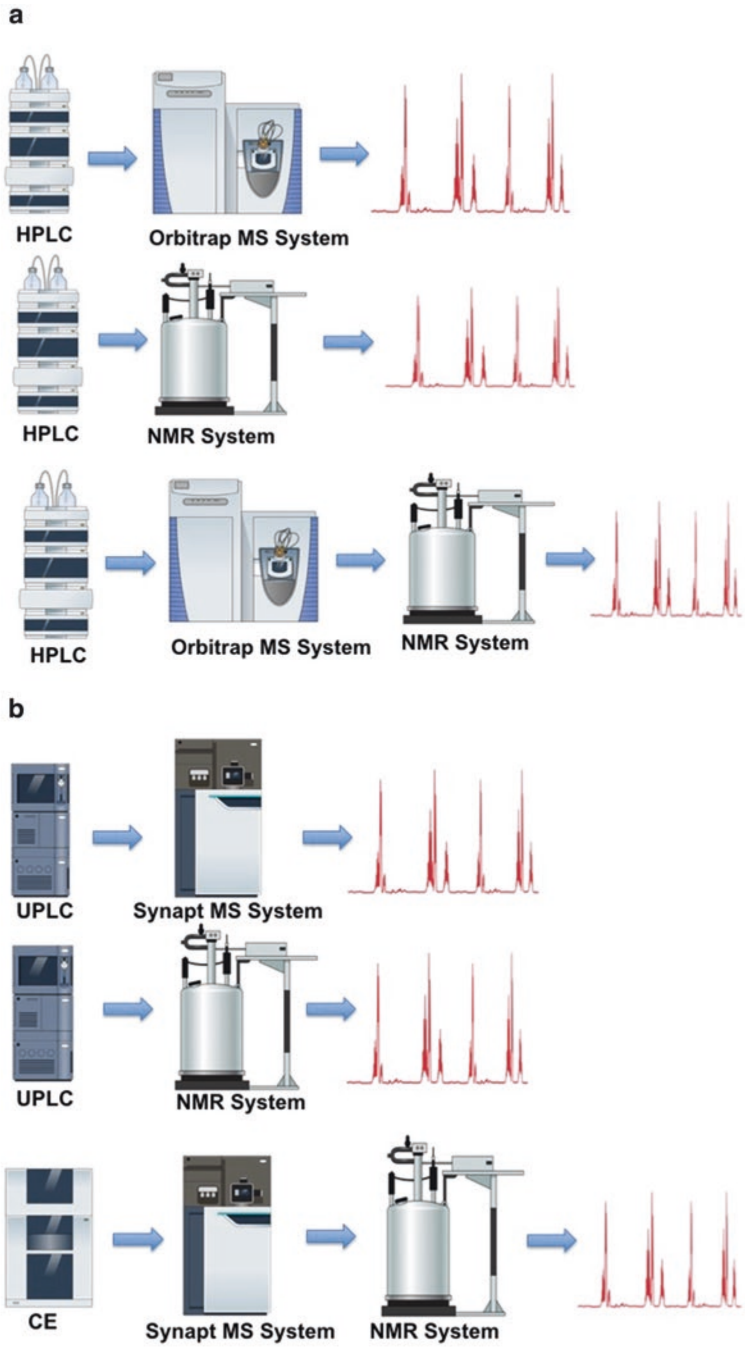


Fig. 13.11 Coupled or hyphenated techniques in metabolomics. (**a**) The coupling or hyphenation of LC-MS, LC-NMR, LC-MS-NMR, (**b**) UPLC-MS, UPLC-NMR, CE-MS-NMR, etc. greatly augments the results of metabolomics studies. The development of shielded magnets has greatly helped to couple LC, NMR, and MS instrumentation and increases the sensitivity and coverage of metabolite detection in biological samples

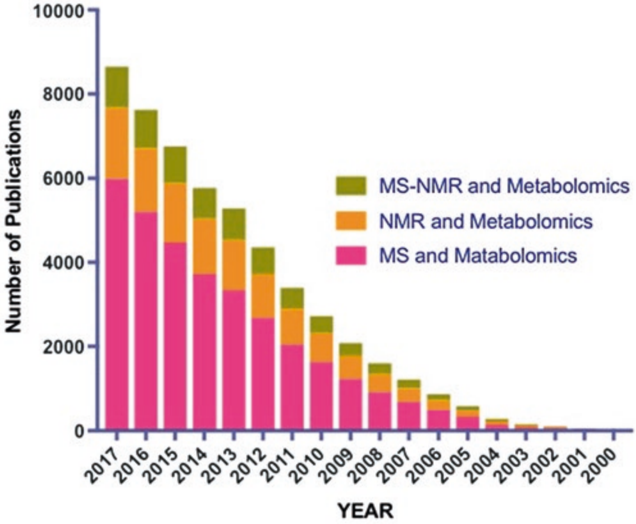


Fig. 13.12 The number of publications cited in Scopus for "mass spectrometry" and metabolomics, "nuclear magnetic resonance" and metabolomics, and the combined search term "mass spectrometry" and "nuclear magnetic resonance" and metabolomics till the year 2017 is 36926, 12553, and 7412, respectively

Univariate analysis such as student's t-test, fold change, etc. takes only one variable into account, whereas the multivariate analysis works on an array of variables and their association with other variables. Lindahl et al. (2017a) have recently shown that univariate and multivariate analyses were successfully used in tandem for comparing metabolomics data obtained using LC-ESI-MS platform from the blood of patients with pancreatic ductal adenocarcinoma (PDAC) and chronic pancreatitis (CP). The metabolites were investigated using a discovery cohort and a validation cohort for each set of disease. The data analysis revealed the presence of large number of metabolite features (n = 4578) (Fig. 13.13). These features were further filtered using both student's t-test and orthogonal partial least squares discriminant analysis (OPLS-DA) leading to the identification of 17 and 19 metabolites, respectively, in the discovery cohort. Further applications of univariate and multivariate methods provided 11 and 19 metabolites in the validation cohort. Finally, using the fold change approach, the metabolites were filtered yielding three (univariate) and five (multivariate) metabolites, respectively (Lindahl et al. 2017b).

Hence, in the multivariate analyses, both principal component analysis (PCA) and partial least squares (PLS) are established methods to understand the pattern hidden in the metabolomics data. PCA helps us to deduce major trends and features in the metabolomics data, thus reducing the dimensionality of the data (Fig. 13.14). The data preprocessing, data scaling, and data normalization strategies are essential to properly infer the metabolomics data. Consequently, successful data analysis needs careful investigation of several models for arriving at a consensus on the potential metabolite biomarkers in health and disease (Fig. 13.15). Lindahl et al. (2017b) have used the PCA

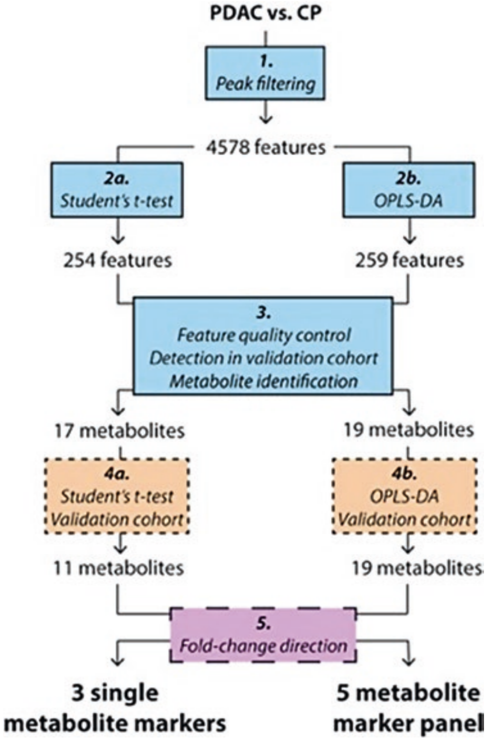


Fig. 13.13 Metabolomics data analysis pipeline. Univariate and multivariate analyses comparing PDAC and CP were done in tandem. The data analysis revealed the presence of large number of metabolite features (*n* = 4578). These features were further filtered using both student's t-test and orthogonal partial least squares discriminant analysis (OPLS-DA) leads to the identification of 17 and 19 metabolites, respectively, in the discovery cohort. Further applications of univariate and multivariate methods provided 11 and 19 metabolites in the validation cohort. Finally, using the fold change, the metabolites were filtered yielding three (univariate) and five (multivariate) metabolites, respectively. (Adapted from Lindahl et al. 2017a, Metabolomics, 13:61 and shared based on Creative Commons CC BY 4.0) (http://creativecommons.org/licenses/by/4.0/)

and OPLS-DA (Fig. 13.16) methods to successfully identify five different metabolites that were significantly present in PDAC compared to CP (Lindahl et al. 2017b). Besides, they obtained precise mass measurements using the public databases such as METLIN (Smith et al. 2005) and Human Metabolome Database (Wishart et al. 2007, 2009, 2013, 2018) as well as their own in-house library comprising 384 synthetic standards (Fig. 13.17). Furthermore, the univariate analyses showed that hexanoylcarnitine, glycocholic acid, and N-palmitoyl glutamic acid were significantly higher in PDAC compared to CP in both discovery and validation cohorts (Fig. 13.18). In the multivariate analyses, the metabolite-metabolite correlation analysis (MMCA) heat maps are successfully used (Fig. 13.19) for extracting biologically relevant information from the high-throughput metabolomics datasets (Jauhiainen et al. 2014; Madhu et al. 2017).

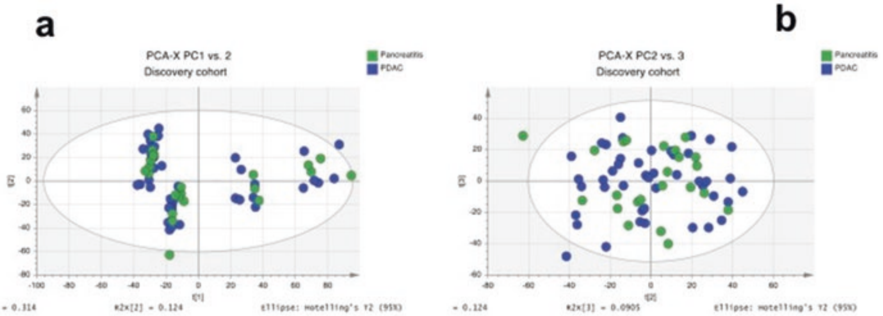


Fig. 13.14 (**a**) Principal component analysis (PCA) score (PC1 vs PC2) plot based on the 4578 variables identified using XCMS processing. (**b**) Score plot of PC2 vs PC3 showing two potential outliers in the metabolomics. (Adapted from Lindahl et al. 2017a, Metabolomics, 13:61 and shared based on Creative Commons CC BY 4.0) (http://creativecommons.org/licenses/by/4.0/)

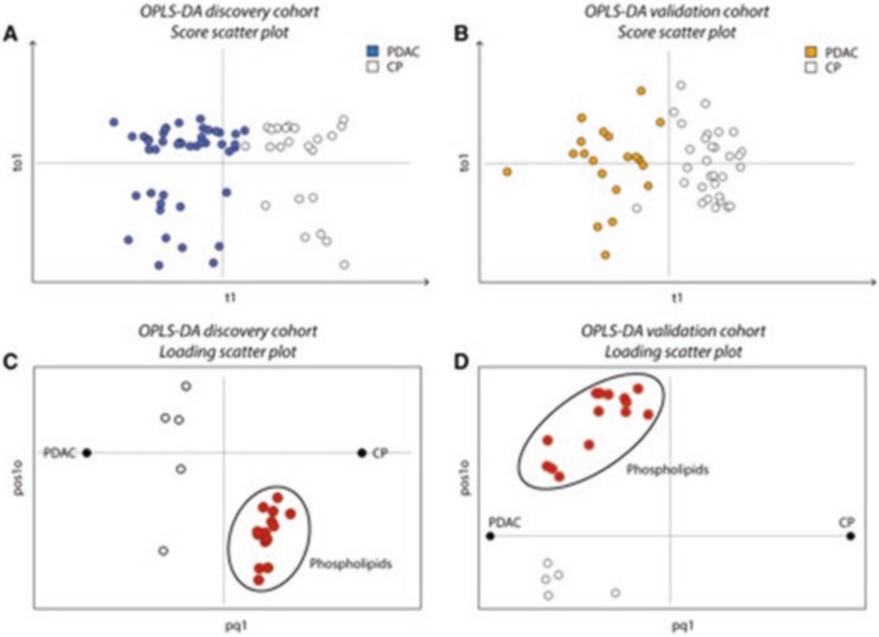


Fig. 13.15 Reversed levels of metabolites in the discovery and validation cohorts. (**a**) Score scatter plot for the initial OPLS-DA model of 4578 metabolite features in the discovery cohort and (**b**) validation cohort; (**c**) fold change of phospholipids in the discovery cohort (downregulated) and (**d**) validation cohorts (upregulated). (Adapted from Lindahl et al. 2017a, Metabolomics, 13:61 and shared based on Creative Commons CC BY 4.0) (http://creativecommons.org/licenses/by/4.0/)

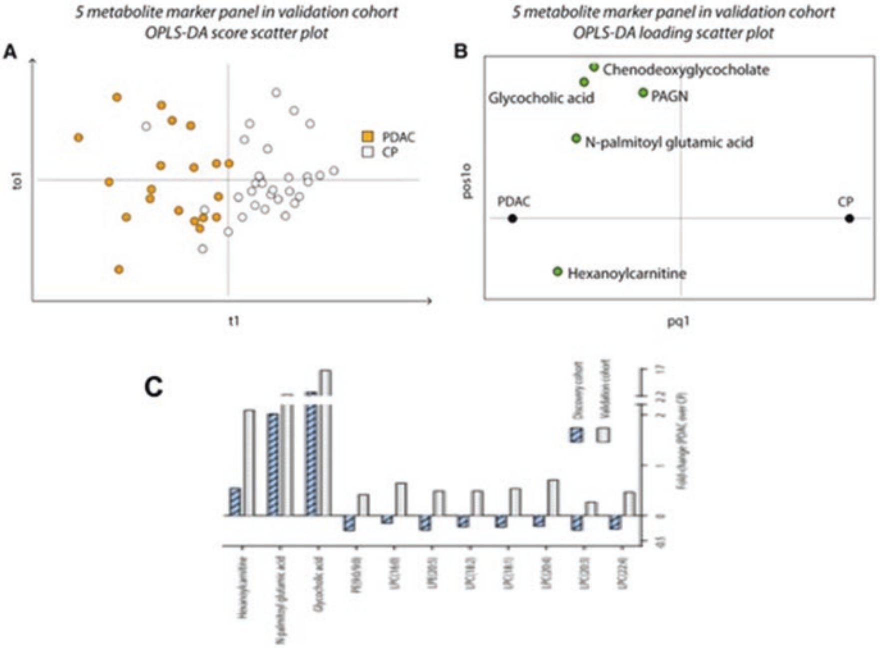


Fig. 13.16 A marker panel of five metabolites differentiates PDAC and CP. (**a**) Score scatter plot and (**b**) loading scatter plot for the refined OPLS-DA model of the five discriminative metabolites with consistent fold-change directions in the validation cohort. (**c**) Phospholipids were removed from further analysis due to their differential regulation in discovery and validation cohorts. However, the levels of hexanoylcarnitine, N-palmitoyl glutamic acid, and glycocholic acid, hexanoylcarnitine were found to be similar in discovery and validation cohorts. (Adapted and modified from Lindahl et al. 2017b, Metabolomics, 13:61 and shared based on Creative Commons CC BY 4.0) (http://creativecommons.org/licenses/by/4.0/)

13.10 Reporting Standards in Metabolomics Data

The metabolomics standards initiative (MSI) has provided several guidelines for journals, academia, industry, government organizations, etc. (Members et al. 2007). It requires minimum reporting standards that describe the experiments in order to disseminate and reuse metabolomics data (Members et al. 2007). Similarly, COSMOS (Coordination of Standards in Metabolomics) is a FP7 EU Initiative (http://cosmos-fp7.eu) that has robust data infrastructure and supports workflows for an array of metabolomics applications (Salek et al. 2015; Salek et al. 2013a, b). The open source ISA software suite offered by ISA framework (http://isa-tools.org) helps to standardize metadata for scientific experiments and manage an increasingly diverse set of basic, translational, and clinical research data to provide a rich description of the experimental metadata such as sample type, characteristics, metabolomics approach used, etc. to make the resulting data and discoveries reproducible and reusable. The tool is based on standard ontology for a range of biological approaches, including the

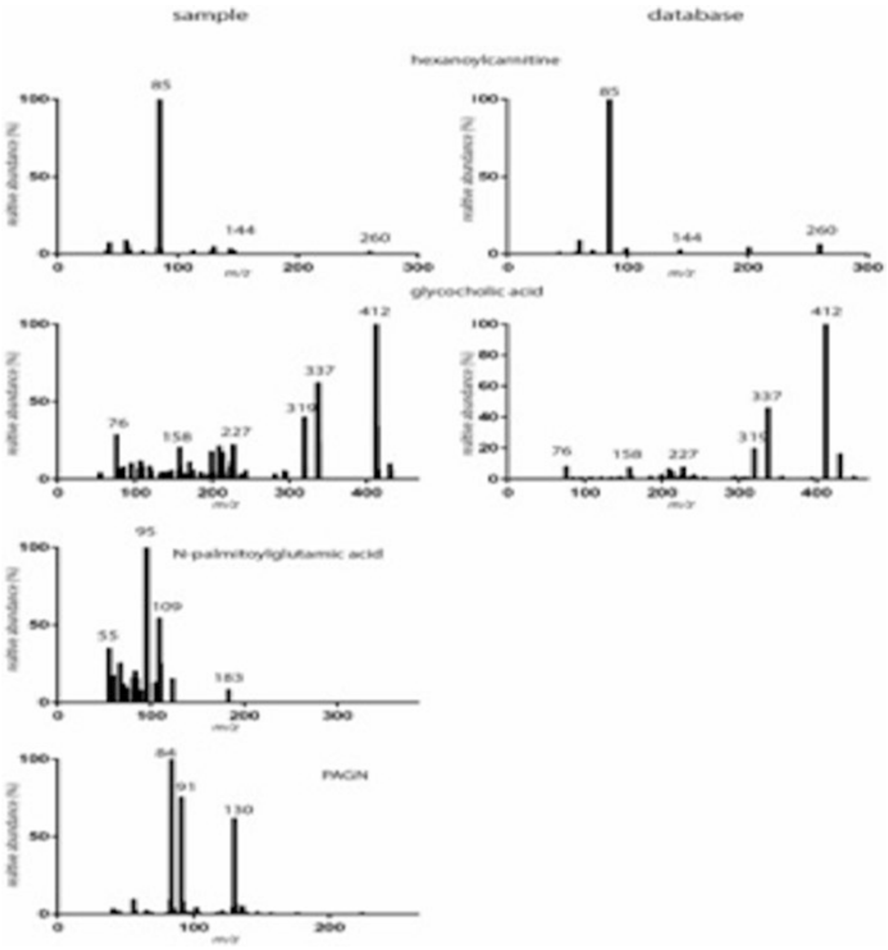


Fig. 13.17 Metabolites MS data matching with MS spectral library. LC/MS spectra hexanoylcarnitine, glycocholic acid, N-palmitoyl glutamic acid, and PAGN. No database spectra was available for N-palmitoyl glutamic acid and PAGN test. (Adapted from Lindahl et al. 2017a, Metabolomics, 13:61 and shared based on Creative Commons CC BY 4.0) (http://creativecommons.org/licenses/by/4.0/)

MSI initiative and COSMOS, and built around the concept of investigation (I), study (S), and assay (A) strategies adopted in experiments (Sansone et al. 2012).

13.11 Metabolomics Databases and Repositories

(a) *Human Metabolome Database*

The Human Metabolome Database (HMDB) is a free database (current version: HMDB version 4.0, www.hmdb.ca) containing huge wealth of information about

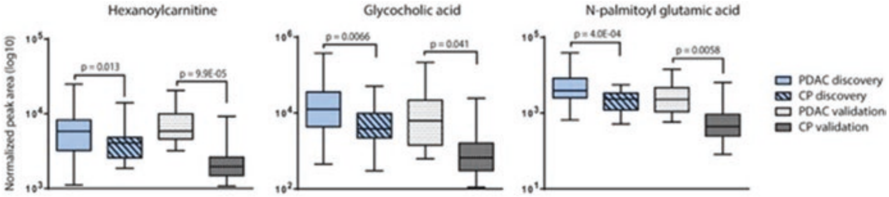


Fig. 13.18 Univariate analysis of three metabolite biomarkers for differentiating PDAC and CP. Single metabolite markers discriminating PDAC and CP. All three were upregulated in PDAC. Statistical test: Welch's unequal variances t test. (Adapted from Lindahl et al. 2017a, Metabolomics, 13:61 and shared based on Creative Commons CC BY 4.0) (http://creativecommons.org/licenses/by/4.0/)

small molecule metabolites experimentally uncovered in the human body (Wishart et al. 2018). The database contains chemical, biochemical, molecular, and clinical data. The HMDB (version 4.0) database contains 114,100 metabolite entries including both water-soluble and lipid-soluble metabolites as well as abundant (>1 μM) or relatively rare (<1 nM) metabolites (Wishart et al. 2018). HMDB has hyperlinks to various free databases such as PubChem, KEGG, PDB, GenBank, MetaCyc, ChEBI, and Swiss-Prot (Wishart et al. 2007, 2009, 2013, 2018). Users can search for metabolites using text, sequence, chemical structure, etc. The HMDB has the provision for the search LC-MS, GC-MS, 1D, and 2D NMR data derived from biological samples (Fig. 13.20). In the MS search option, users can submit mass spectral files (MoverZ format) for searching against the HMDB's library of MS spectra for the identification of unknown metabolites from the LC-MS and LC-MS/MS spectra. Similarly, the peak lists from ^{1}H NMR, ^{13}C NMR, 2D TOCSY, or ^{13}C HSQC spectra can be searched using the NMR libraries contained in the HMDB for the identification of unknown metabolites.

(b) *METLIN*

The METLIN was created in 2004 at the Scripps Research Institute (https://metlin.scripps.edu). It is a free repository for various types of MS data (MS/MS, LC-MS, etc.) obtained using various types of instruments such as Agilent, Bruker, SCIEX, and Waters QTOF mass spectrometers (Smith et al. 2005). It has the largest collection of MS data generated using multiple collision energies and in both positive and negative ionization modes (Guijas et al. 2018; Sana et al. 2008; Tautenhahn et al. 2012; Zhu et al. 2013). MS data can be searched using METLIN by mass range, peak lists, disease, and biological source (Fig. 13.21). Over 14,000 metabolites have been individually analyzed and another 220,000 have in silico MS/MS data (Smith et al. 2005; Guijas et al. 2018; Sana et al. 2008; Tautenhahn et al. 2012; Zhu et al. 2013).

(c) *XCMS*

XCMS is an open source software package (latest version: 3.7.0) (https://xcmsonline.scripps.edu) that has been developed to analyze MS data. XCMS allows

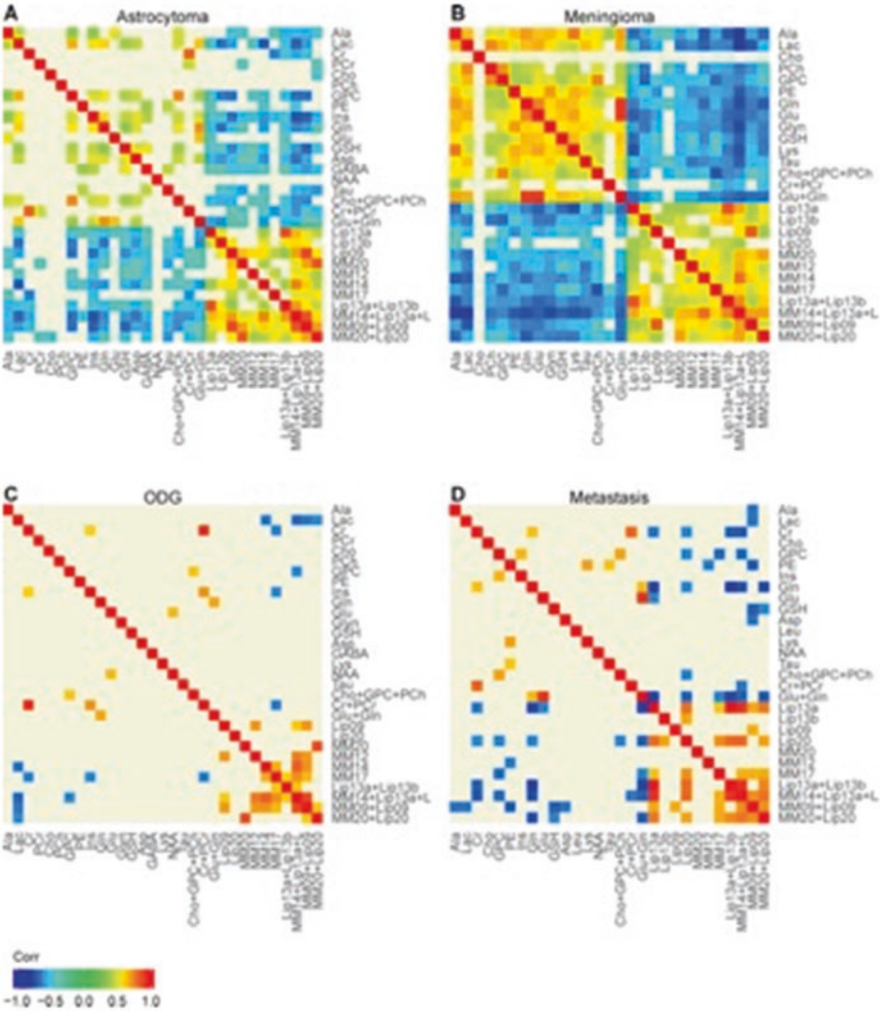


Fig. 13.19 Metabolite-metabolite correlation snalysis (MMCA) heat maps based on datasets from different types of brain tumors such as astrocytomas, meningiomas, and oligodendrogliomas and metastases. (Adapted from Madhu et al. (2017) and shared based on Creative Commons License) (http://creativecommons.org/licenses/by/4.0/)

users to perform pathway analyses directly from their raw metabolomic data, and it enables proteomic and genomic data integration (Fig. 13.22). The output can be visualized in table form or through Pathway Cloud Plot (Forsberg et al. 2018; Gowda et al. 2014; Huan et al. 2017; Mahieu et al. 2016).

(d) *MetaboLights Database*

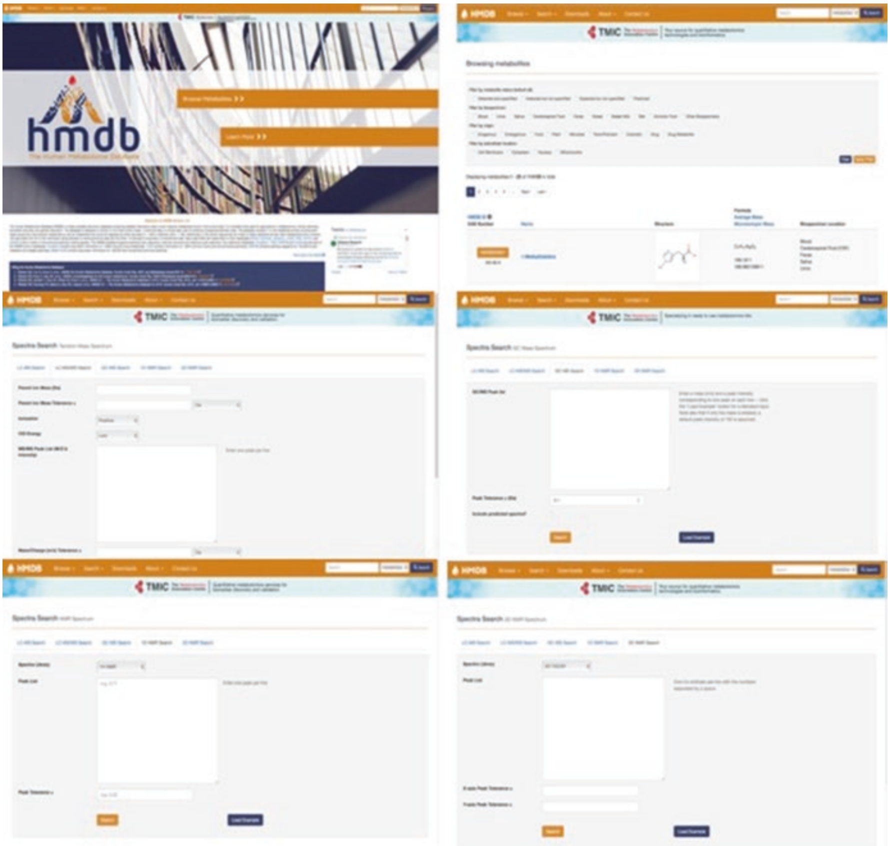


Fig. 13.20 The Human Metabolome Database (HMDB, current version 4.0) is a free database (www.hmdb.ca) containing huge wealth of information about small molecule metabolites experimentally uncovered in the human body. Users can search for metabolites using text, sequence, chemical structure, etc. The HMDB has the provision for the search MS, GC, 1D, and 2D NMR data for the unknown metabolites. In the MS search option, users can submit mass spectral files (MoverZ format) for searching against the HMDB's library of MS spectra for the identification of unknown metabolites from the LC-MS and LC-MS/MS spectra. Similarly, the peak lists from 1H NMR, 13C NMR, 2D TOCSY, or 13C HSQC spectra can be searched using the NMR libraries contained in the HMDB for the identification of unknown metabolites

MetaboLights is an open source database (https://www.ebi.ac.uk/metabolights) with experimental data derived from metabolomics approaches (Salek et al. 2013a). It is an ELIXIR-recommended repository and the most preferred depository of various eminent journals for the metabolomics data (Salek et al. 2013a, b; Haug et al. 2013; Kale et al. 2016; Steinbeck et al. 2012). It includes the metabolite structures and their reference spectra with their location, biological roles, etc. from metabolomics experiments (Salek et al. 2013a, b; Haug et al. 2013; Kale et al. 2016; Steinbeck et al. 2012).

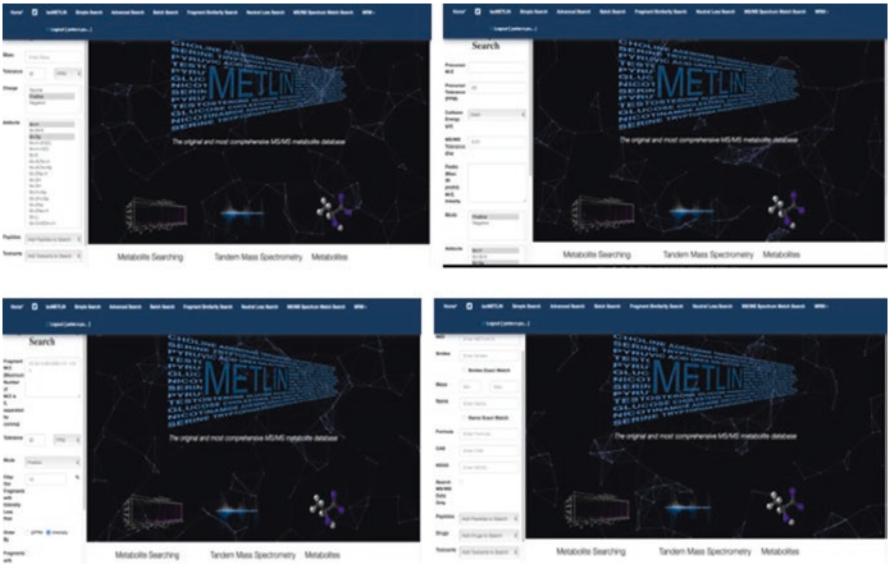


Fig. 13.21 The METLIN (https://metlin.scripps.edu) is a free repository for various types of MS data (MS/MS, LC-MS, etc.). It has the largest collection of MS data generated using multiple collision energies and in both positive and negative ionization modes. MS data can be searched using METLIN by mass range, peak lists, disease, and biological source. Over 14,000 metabolites have been individually analyzed and another 220,000 have in silico MS/MS data

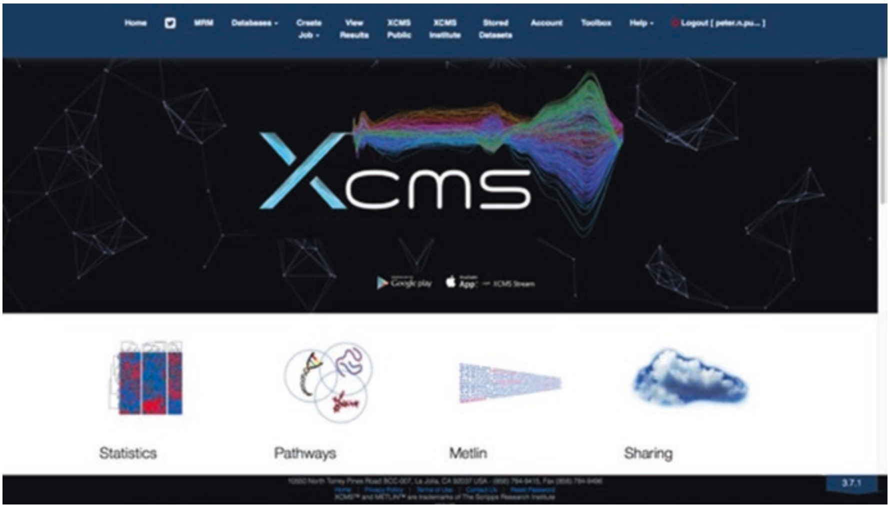


Fig. 13.22 XCMS is an open-source software package (latest version: 3.7.0) (https://xcmsonline.scripps.edu) that has been developed to analyze MS data. XCMS allows users to perform pathway analyses directly from their raw metabolomic data, and it enables proteomic and genomic data integration. The output can be visualized in table form or through Pathway Cloud Plot

(e) *Biological Magnetic Resonance Data Bank*

The Biological Magnetic Resonance Data Bank (BMRB) is the central repository (http://www.bmrb.wisc.edu) for experimental NMR spectral data for macromolecules, and it has an additional analysis option for metabolite data (Smelter et al. 2017; Ulrich et al. 2008; Wishart et al. 1997).

(f) *Madison Metabolomics Consortium Database (MMCD)*

The Madison Metabolomics Consortium Database (MMCD) is a database (http://mmcd.nmrfam.wisc.edu) on small molecules of biological interest from MS and NMR experiments compiled from various metabolomics databases and the scientific literature (Cui et al. 2008). It has 19,700 metabolites and experimental spectral data (Cui et al. 2008).

13.12 Metabolomics Data Analysis Software and Servers

(a) *COLMAR*

COLMAR is a webserver (http://spin.ccic.ohio-state.edu/index.php/colmarm/index) for deducing the structures of metabolites from NMR data (Bingol et al. 2015b; Robinette et al. 2008; Zhang et al. 2008, 2009). The COLMAR metabolomics data analysis web portal is used to analyze both 1D NMR and 2D NMR data (Bingol et al. 2015b; Robinette et al. 2008; Zhang et al. 2008, 2009).

(b) *Fragment iDentificator*

Fragment iDentificator (FiD) (https://www.cs.helsinki.fi/group/sysfys/software/fragid) is a software tool, free for academic use, for the identification of metabolites from MS data (Heinonen et al. 2008). Graphical user interface of FiD is easy to use with visualization capabilities and provides information about the metabolite structures (Heinonen et al. 2008).

(c) *MeltDB*

MeltDB (version 2.0) is a web-based software for the analysis of metabolomics data (https://meltdb.cebitec.uni-bielefeld.de) (Kessler et al. 2013, 2015; Neuweger et al. 2008). MeltDB can analyze netCDF, mzXML, and mzDATA files (Kessler et al. 2013, 2015; Neuweger et al. 2008). The system provides comprehensive data analysis and visualization tools to the researchers and stores their experimental datasets (Kessler et al. 2013, 2015; Neuweger et al. 2008).

(d) *MetaboAnalyst*

MetaboAnalyst (version 4.0) is a web-based MS and NMR data processing tool (http://www.metaboanalyst.ca) (Chong et al. 2018; Xia et al. 2009, 2012, 2015; Xia and Wishart 2011a, b, 2016). The metabolomic data can be processed and normalized, and various statistical tests (univariate and multivariate analysis) can be performed with both MS and NMR datasets (Chong et al. 2018; Xia et al. 2009, 2012, 2015; Xia and Wishart 2011a, b; Xia and Wishart 2016).

(e) *MetaboMiner*

MetaboMiner is used to identify metabolites from 2D NMR spectra (http://wishart.biology.ualberta.ca/metabominer) (Xia et al. 2008) and consists of reference spectra of about 500 pure metabolites for comparison and identification (Xia et al. 2008).

(f) *MolFind*

MolFind is a free Java-based software package for identifying unknown chemical structures in complex mixtures using HPLC/MS data (http://metabolomics.pharm.uconn.edu/Software.html), and the web interface is very easy to use (Menikarachchi et al. 2012).

(g) *MVAPACK*

MVAPACK is an open source toolkit (http://bionmr.unl.edu/mvapack.php) for data handling in NMR and MS metabolic profiling experiments with easy to use analytical tools (Worley and Powers 2014; Marshall et al. 2015).

13.13 Metabolic Pathway Databases

(a) *Kyoto Encyclopedia of Genes and Genomes Pathway*

Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway is a part of KEGG database, which consists of manually drawn pathways from a wide variety of organisms (https://www.genome.jp/kegg/pathway.html) (Aoki and Kanehisa 2005; Kanehisa 2013, 2016; Kanehisa et al. 2010, 2014, 2017; Kotera et al. 2012; Okuda et al. 2008; Tanabe and Kanehisa 2012). KEGG pathway database shows the molecular interaction, reaction, and associated networks for metabolism, genetic and environmental information processing, cellular and molecular processes, human diseases, drug development, organismal systems, etc. KEGG pathways are linked to metabolite, protein, enzyme, and other related molecular information (Aoki and Kanehisa 2005; Kanehisa 2013, 2016; Kanehisa et al. 2010, 2014, 2017; Kotera et al. 2012; Okuda et al. 2008; Tanabe and Kanehisa 2012).

(b) *MetaCyc*

MetaCyc (https://metacyc.org) has experimentally clarified pathways (Caspi et al. 2006, 2008, 2012; Caspi and Karp 2007; Karp and Caspi 2011; Karp et al. 2002, 2013; Krieger et al. 2004). MetaCyc has 2642 pathways from 2941 different organisms (Caspi et al. 2006, 2008, 2012; Caspi and Karp 2007; Karp and Caspi 2011; Karp et al. 2002, 2013; Krieger et al. 2004; Caspi et al. 2016). MetaCyc has the pathway prediction tools such as PathoLogic for the computational reconstruction of metabolic networks from sequenced genomes (Caspi et al. 2006, 2008, 2012;

Caspi and Karp 2007; Karp and Caspi 2011; Karp et al. 2002, 2013; Krieger et al. 2004; Caspi et al. 2016).

(c) *HumanCyc: Encyclopedia of Human Genes and Metabolism*

HumanCyc (https://humancyc.org) is a database for human pathways (Romero et al. 2005). It has information about 28,783 genes and their protein products, metabolic reactions, and other interrelated pathways (Romero et al. 2005).

(d) *BioCyc*

BioCyc (https://biocyc.org) consists of 13,075 Pathway and Genome Databases (PGDBs) (Caspi et al. 2008, 2012; Caspi and Karp 2007; Caspi et al. 2014, 2016; Krummenacker et al. 2005; Latendresse et al. 2012; Walsh et al. 2014) and various software tools to visualize and navigate various databases and analyze multi-omics data (Caspi et al. 2008, 2012; Caspi and Karp 2007; Caspi et al. 2014; 2016; Krummenacker et al. 2005; Latendresse et al. 2012; Walsh et al. 2014). Based on the quality, the databases in BioCyc are classified into Tier 1, Tier 2, and Tier 3 (Caspi et al. 2008, 2012; Caspi and Karp 2007; Caspi et al. 2014, 2016; Krummenacker et al. 2005; Latendresse et al. 2012; Walsh et al. 2014).

(e) *The Reactome Pathway Knowledgebase*

The Reactome (https://reactome.org) is a free, curated, peer-reviewed online database (latest version: 65) of biological pathways, including metabolic pathways as well as protein trafficking and signaling pathways (Croft et al. 2014; Fabregat et al. 2016, 2018). The primary goal of the Reactome project is to provide intuitive bioinformatics tools for the biological interpretation, data analysis, and visualization for basic research, genome analysis, systems biology, modeling, and education (Fabregat et al. 2018). The Reactome has about 2222 human pathways, 1880 small molecules, 11,896 reactions, 10,763 proteins, and 28,436 literature references (Croft et al. 2014; Fabregat et al. 2016, 2018).

(f) *WikiPathways*

WikiPathways (https://www.wikipathways.org) is an open source, collaborative platform for capturing and disseminating models of biological pathways for high-throughput “Omics” data visualization and analysis (Bohler et al. 2016; Kelder et al. 2009, 2012; Kutmon et al. 2016; Pico et al. 2008; Slenter et al. 2018; Waagmeester et al. 2016). The database currently has 1570 pathways, covers 11,532 human genes, and has links to many metabolomics databases (Bohler et al. 2016; Kelder et al. 2009, 2012; Kutmon et al. 2016; Pico et al. 2008; Slenter et al. 2018; Waagmeester et al. 2016).

In addition to the above, there are several open source databases as well as commercial databases and also the free software tools available at the disposal of researchers to extract, analyze, and interpret the high-throughput metabolomics data derived from MS or NMR experiments (Bingol et al. 2015a; Tsugawa 2018; Ellinger et al. 2013; Jeffryes et al. 2015; Gil de la Fuente et al. 2017).

13.14 Potential Applications of Metabolomics

Metabolomics helps to discover potential metabolites that may be used as biomarkers to differentiate health and disease in medicine. Biological samples such as plasma, serum, saliva, tears, seminal plasma, bile, sweat, etc. can be rapidly obtained from patients to study the profiles of metabolites. Besides, metabolomics has an array of applications in agriculture like studying the metabolite profiling in wild type and genetically modified plants and also in natural products research (Johnson and Lange 2015). Individualized or precision medicine, a term used for personalized therapy, requires metabolomics for rapid diagnosis of a specific disease (Tan et al. 2016). For decades, in a typical healthcare setting, classical biochemical tests are used to precisely measure individual metabolites such as blood glucose, creatinine, bilirubin, urea, uric acid, antioxidants, adenosine triphosphate (ATP), redox compounds (NAD+, NADP), etc. Metabolomics offers the potential for the rapid identification of hundreds of metabolites, enabling us to identify numerous disease states such as cancer (Lane et al. 2009b; Fan et al. 2009). For example, the PDAC has shown a distinct urinary metabolome (Davis et al. 2013) and PDAC can be differentiated from CP using LC-MS metabolomics approach (Lindahl et al. 2017b). A recent study used LC-MS approach to uncover the changes in blood metabolites due to ageing in humans (Chaleckis et al. 2016).

13.15 Conclusions

Metabolomics is an essential “omics” approach to decipher the differential expression of metabolites in health and disease. Till now, most of the metabolomics data were generated using either MS or NMR. Recent advancements in shielded magnets help in coupling both MS and NMR together (MS-NMR) to increase the sensitivity, reliability, reproducibility, coverage of the metabolome, and quality of the metabolomics data. The advancements in the development of analytical tools and freely available databases for metabolomics have greatly increased the accuracy of data analysis and biological interpretation. Hence, it is used in various areas such as agriculture, precision medicine, biomarker discovery, drug discovery, food science, veterinary science, environmental studies, and other interlinked areas.

Acknowledgments This work is funded by the National Plan for Science, Technology, and Innovation (MAARIFAH), King Abdulaziz City for Science and Technology, the Kingdom of Saudi Arabia, award number 12-BIO2267-03. The author also acknowledges with thanks the Science and Technology Unit (STU), King Abdulaziz University, for their excellent technical support.

References

Abdel-Rehim M (2011) Microextraction by packed sorbent (MEPS): a tutorial. Anal Chim Acta 701(2):119–128

Aoki KF, Kanehisa M (2005) Using the KEGG database resource. Curr Protoc Bioinformatics Chapter 1:Unit 1 12

Bartlett MG, Chen B (2016) Editor-in-chief editorial and introduction to ‘Metabolomics and biomarkers’ special issue. Biomed Chromatogr 30(1):5–6

Bingol K, Bruschweiler R (2015a) Two elephants in the room: new hybrid nuclear magnetic resonance and mass spectrometry approaches for metabolomics. Curr Opin Clin Nutr Metab Care 18(5):471–477

Bingol K, Bruschweiler R (2015b) NMR/MS translator for the enhanced simultaneous analysis of metabolomics mixtures by NMR spectroscopy and mass spectrometry: application to human urine. J Proteome Res 14(6):2642–2648

Bingol K, Bruschweiler-Li L, Yu C, Somogyi A, Zhang F, Bruschweiler R (2015a) Metabolomics beyond spectroscopic databases: a combined MS/NMR strategy for the rapid identification of new metabolites in complex mixtures. Anal Chem 87(7):3864–3870

Bingol K, Li DW, Bruschweiler-Li L, Cabrera OA, Megraw T, Zhang F et al (2015b) Unified and isomer-specific NMR metabolomics database for the accurate analysis of (13)C-(1)H HSQC spectra. ACS Chem Biol 10(2):452–459

Bohler A, Wu G, Kutmon M, Pradhana LA, Coort SL, Hanspers K et al (2016) Reactome from a WikiPathways perspective. PLoS Comput Biol 12(5):e1004941

Caspi R, Karp PD (2007) Using the MetaCyc pathway database and the BioCyc database collection. Curr Protoc Bioinformatics Chapter 1:Unit1 17

Caspi R, Foerster H, Fulcher CA, Hopkinson R, Ingraham J, Kaipa P et al (2006) MetaCyc: a multiorganism database of metabolic pathways and enzymes. Nucleic Acids Res 34(Database issue):D511–D516

Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, Latendresse M et al (2008) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res 36(Database issue):D623–D631

Caspi R, Altman T, Dreher K, Fulcher CA, Subhraveti P, Keseler IM et al (2012) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res 40(Database issue):D742–D753

Caspi R, Altman T, Billington R, Dreher K, Foerster H, Fulcher CA et al (2014) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res 42(Database issue):D459–D471

Caspi R, Billington R, Ferrer L, Foerster H, Fulcher CA, Keseler IM et al (2016) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res 44(D1):D471–D480

Chaleckis R, Murakami I, Takada J, Kondoh H, Yanagida M (2016) Individual variability in human blood metabolites identifies age-related differences. Proc Natl Acad Sci U S A 113(16):4252–4259

Chong J, Soufan O, Li C, Caraus I, Li S, Bourque G et al (2018) MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. Nucleic Acids Res 46(W1):W486–WW94

Cook KD (2002) ASMS members John Fenn and Koichi Tanaka share Nobel: the world learns our “secret”. American Society for Mass Spectrometry. J Am Soc Mass Spectrom 13(12):1359

Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G et al (2014) The Reactome pathway knowledgebase. Nucleic Acids Res 42(Database issue):D472–D477

Cui Q, Lewis IA, Hegeman AD, Anderson ME, Li J, Schulte CF et al (2008) Metabolite identification via the Madison Metabolomics Consortium Database. Nat Biotechnol 26(2):162–164

Davis VW, Schiller DE, Eurich D, Bathe OF, Sawyer MB (2013) Pancreatic ductal adenocarcinoma is associated with a distinct urinary metabolomic signature. Ann Surg Oncol 20(Suppl 3):S415–S423

Drouin N, Rudaz S, Schappler J (2017) Sample preparation for polar metabolites in bioanalysis. Analyst 143(1):16–20

Ellinger JJ, Chylla RA, Ulrich EL, Markley JL (2013) Databases and software for NMR-based metabolomics. Curr Metabolomics (1):1, 28–40

Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R et al (2016) The Reactome pathway knowledgebase. Nucleic Acids Res 44(D1):D481–D487

Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P et al (2018) The Reactome pathway knowledgebase. Nucleic Acids Res 46(D1):D649–DD55

Fan TW, Lane AN, Higashi RM, Farag MA, Gao H, Bousamra M et al (2009) Altered regulation of metabolic pathways in human lung cancer discerned by (13)C stable isotope-resolved metabolomics (SIRM). Mol Cancer 8:41

Forsberg EM, Huan T, Rinehart D, Benton HP, Warth B, Hilmers B et al (2018) Data processing, multi-omic pathway mapping, and metabolite activity analysis using XCMS online. Nat Protoc 13(4):633–651

Gil de la Fuente A, Grace Armitage E, Otero A, Barbas C, Godzien J (2017) Differentiating signals to make biological sense – a guide through databases for MS-based non-targeted metabolomics. Electrophoresis 38(18):2242–2256

Gowda H, Ivanisevic J, Johnson CH, Kurczy ME, Benton HP, Rinehart D et al (2014) Interactive XCMS online: simplifying advanced metabolomic data processing and subsequent statistical analyses. Anal Chem 86(14):6931–6939

Guijas C, Montenegro-Burke JR, Domingo-Almenara X, Palermo A, Warth B, Hermann G et al (2018) METLIN: a technology platform for identifying knowns and unknowns. Anal Chem 90(5):3156–3164

Haug K, Salek RM, Conesa P, Hastings J, de Matos P, Rijnbeek M et al (2013) MetaboLights–an open-access general-purpose repository for metabolomics studies and associated meta-data. Nucleic Acids Res 41(Database issue):D781–D786

Heinonen M, Rantanen A, Mielikainen T, Kokkonen J, Kiuru J, Ketola RA et al (2008) FiD: a software for ab initio structural identification of product ions from tandem mass spectrometric data. Rapid Commun Mass Spectrom 22(19):3043–3052

Huan T, Forsberg EM, Rinehart D, Johnson CH, Ivanisevic J, Benton HP et al (2017) Systems biology guided by XCMS online metabolomics. Nat Methods 14(5):461–462

Jauhiainen A, Madhu B, Narita M, Narita M, Griffiths J, Tavare S (2014) Normalization of metabolomics data with applications to correlation maps. Bioinformatics 30(15):2155–2161

Jeffryes JG, Colastani RL, Elbadawi-Sidhu M, Kind T, Niehaus TD, Broadbelt LJ et al (2015) MINEs: open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics. J Cheminform 7:44

Johnson SR, Lange BM (2015) Open-access metabolomics databases for natural product research: present capabilities and future potential. Front Bioeng Biotechnol 3:22

Johnson CH, Ivanisevic J, Siuzdak G (2016) Metabolomics: beyond biomarkers and towards mechanisms. Nat Rev Mol Cell Biol 17(7):451–459

Jones OA, Cheung VL (2007) An introduction to metabolomics and its potential application in veterinary science. Comp Med 57(5):436–442

Kale NS, Haug K, Conesa P, Jayseelan K, Moreno P, Rocca-Serra P et al (2016) MetaboLights: an open-access database repository for metabolomics data. Curr Protoc Bioinformatics 53:14 3 1–8

Kanehisa M (2013) Molecular network analysis of diseases and drugs in KEGG. Methods Mol Biol 939:263–275

Kanehisa M (2016) KEGG bioinformatics resource for plant genomics and metabolomics. Methods Mol Biol 1374:55–70

Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res 38(Database issue):D355–D360

Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M (2014) Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res 42(Database issue):D199–D205

Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 45(D1):D353–DD61

Karp PD, Caspi R (2011) A survey of metabolic databases emphasizing the MetaCyc family. Arch Toxicol 85(9):1015–1033

Karp PD, Riley M, Paley SM, Pellegrini-Toole A (2002) The MetaCyc database. Nucleic Acids Res 30(1):59–61

Karp PD, Paley S, Altman T (2013) Data mining in the MetaCyc family of pathway databases. Methods Mol Biol 939:183–200

Kelder T, Pico AR, Hanspers K, van Iersel MP, Evelo C, Conklin BR (2009) Mining biological pathways using WikiPathways web services. PLoS One 4(7):e6447

Kelder T, van Iersel MP, Hanspers K, Kutmon M, Conklin BR, Evelo CT et al (2012) WikiPathways: building research communities on biological pathways. Nucleic Acids Res 40(Database issue):D1301–D1307

Kessler N, Neuweger H, Bonte A, Langenkamper G, Niehaus K, Nattkemper TW et al (2013) MeltDB 2.0-advances of the metabolomics software system. Bioinformatics 29(19):2452–2459

Kessler N, Bonte A, Albaum SP, Mader P, Messmer M, Goesmann A et al (2015) Learning to classify organic and conventional wheat – a machine learning driven approach using the MeltDB 2.0 metabolomics analysis platform. Front Bioeng Biotechnol 3:35

Kotera M, Hirakawa M, Tokimatsu T, Goto S, Kanehisa M (2012) The KEGG databases and tools facilitating omics analysis: latest developments involving human diseases and pharmaceuticals. Methods Mol Biol 802:19–39

Krieger CJ, Zhang P, Mueller LA, Wang A, Paley S, Arnaud M et al (2004) MetaCyc: a multiorganism database of metabolic pathways and enzymes. Nucleic Acids Res 32(Database issue):D438–D442

Krummenacker M, Paley S, Mueller L, Yan T, Karp PD (2005) Querying and computing with BioCyc databases. Bioinformatics 21(16):3454–3455

Kutmon M, Riutta A, Nunes N, Hanspers K, Willighagen EL, Bohler A et al (2016) WikiPathways: capturing the full diversity of pathway knowledge. Nucleic Acids Res 44(D1):D488–D494

Lane AN, Fan TW, Xie Z, Moseley HN, Higashi RM (2009a) Isotopomer analysis of lipid biosynthesis by high resolution mass spectrometry and NMR. Anal Chim Acta 651(2):201–208

Lane AN, Fan TW, Higashi RM, Tan J, Bousamra M, Miller DM (2009b) Prospects for clinical cancer metabolomics using stable isotope tracers. Exp Mol Pathol 86(3):165–173

Latendresse M, Paley S, Karp PD (2012) Browsing metabolic and regulatory networks with BioCyc. Methods Mol Biol 804:197–216

Lin Y, Schiavo S, Orjala J, Vouros P, Kautz R (2008) Microscale LC-MS-NMR platform applied to the identification of active cyanobacterial metabolites. Anal Chem 80(21):8045–8054

Lindahl A, Saaf S, Lehtio J, Nordstrom A (2017a) Tuning metabolome coverage in reversed phase LC-MS metabolomics of MeOH extracted samples using the reconstitution solvent composition. Anal Chem 89(14):7356–7364

Lindahl A, Heuchel R, Forshed J, Lehtio J, Lohr M, Nordstrom A (2017b) Discrimination of pancreatic cancer and pancreatitis by LC-MS metabolomics. Metabolomics 13(5):61

Ma X, Ouyang Z (2016) Ambient ionization and miniature mass spectrometry system for chemical and biological analysis. Trends Anal Chem 85(A):10–19

Madhu B, Jauhiainen A, McGuire S, Griffiths JR (2017) Exploration of human brain tumour metabolism using pairwise metabolite-metabolite correlation analysis (MMCA) of HR-MAS 1H NMR spectra. PLoS One 12(10):e0185980

Mahieu NG, Genenbacher JL, Patti GJ (2016) A roadmap for the XCMS family of software solutions in metabolomics. Curr Opin Chem Biol 30:87–93

Marshall DD, Powers R (2017) Beyond the paradigm: combining mass spectrometry and nuclear magnetic resonance for metabolomics. Prog Nucl Magn Reson Spectrosc 100:1–16

Marshall DD, Lei S, Worley B, Huang Y, Garcia-Garcia A, Franco R et al (2015) Combining DI-ESI-MS and NMR datasets for metabolic profiling. Metabolomics 11(2):391–402

Members MSIB, Sansone SA, Fan T, Goodacre R, Griffin JL, Hardy NW et al (2007) The metabolomics standards initiative. Nat Biotechnol 25(8):846–848

Menikarachchi LC, Cawley S, Hill DW, Hall LM, Hall L, Lai S et al (2012) MolFind: a software package enabling HPLC/MS-based identification of unknown chemical structures. Anal Chem 84(21):9388–9394

Neuweger H, Albaum SP, Dondrup M, Persicke M, Watt T, Niehaus K et al (2008) MeltDB: a software platform for the analysis and integration of metabolomics experiment data. Bioinformatics 24(23):2726–2732

Okuda S, Yamada T, Hamajima M, Itoh M, Katayama T, Bork P et al (2008) KEGG Atlas mapping for global analysis of metabolic pathways. Nucleic Acids Res 36(Web Server issue):W423–W426

Palmer AG, Patel DJ (2002) Kurt Wuthrich and NMR of biological macromolecules. Structure 10(12):1603–1604

Pico AR, Kelder T, van Iersel MP, Hanspers K, Conklin BR, Evelo C (2008) WikiPathways: pathway editing for the people. PLoS Biol 6(7):e184

Robinette SL, Zhang F, Bruschweiler-Li L, Bruschweiler R (2008) Web server based complex mixture analysis by NMR. Anal Chem 80(10):3606–3611

Romero P, Wagg J, Green ML, Kaiser D, Krummenacker M, Karp PD (2005) Computational prediction of human metabolic pathways from the complete human genome. Genome Biol 6(1):R2

Salek RM, Haug K, Conesa P, Hastings J, Williams M, Mahendraker T et al (2013a) The MetaboLights repository: curation challenges in metabolomics. Database (Oxford) 2013:bat029

Salek RM, Haug K, Steinbeck C (2013b) Dissemination of metabolomics results: role of MetaboLights and COSMOS. Gigascience 2(1):8

Salek RM, Neumann S, Schober D, Hummel J, Billiau K, Kopka J et al (2015) COordination of Standards in MetabOlomicS (COSMOS): facilitating integrated metabolomics data access. Metabolomics 11(6):1587–1597

Sana TR, Roark JC, Li X, Waddell K, Fischer SM (2008) Molecular formula and METLIN personal metabolite database matching applied to the identification of compounds generated by LC/TOF-MS. J Biomol Tech 19(4):258–266

Sansone SA, Rocca-Serra P, Field D, Maguire E, Taylor C, Hofmann O et al (2012) Toward interoperable bioscience data. Nat Genet 44(2):121–126

Schnackenberg LK (2006) Metabolomics special focus: an introduction. Pharmacogenomics 7(7):1053–1054

Slenter DN, Kutmon M, Hanspers K, Riutta A, Windsor J, Nunes N et al (2018) WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. Nucleic Acids Res 46(D1):D661–D6D7

Smelter A, Astra M, Moseley HN (2017) A fast and efficient python library for interfacing with the Biological Magnetic Resonance Data Bank. BMC Bioinformatics 18(1):175

Smith CA, O'Maille G, Want EJ, Qin C, Trauger SA, Brandon TR et al (2005) METLIN: a metabolite mass spectral database. Ther Drug Monit 27(6):747–751

Steinbeck C, Conesa P, Haug K, Mahendraker T, Williams M, Maguire E et al (2012) MetaboLights: towards a new COSMOS of metabolomics data management. Metabolomics 8(5):757–760

Tan SZ, Begley P, Mullard G, Hollywood KA, Bishop PN (2016) Introduction to metabolomics and its applications in ophthalmology. Eye (Lond) 30(6):773–783

Tanabe M, Kanehisa M (2012) Using the KEGG database resource. Curr Protoc Bioinformatics Chapter 1:Unit1 12

Tautenhahn R, Cho K, Uritboonthai W, Zhu Z, Patti GJ, Siuzdak G (2012) An accelerated workflow for untargeted metabolomics using the METLIN database. Nat Biotechnol 30(9):826–828

Tsugawa H (2018) Advances in computational metabolomics and databases deepen the understanding of metabolisms. Curr Opin Biotechnol 54:10–17

Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J et al (2008) BioMagResBank. Nucleic Acids Res 36(Database issue):D402–D408

Waagmeester A, Kutmon M, Riutta A, Miller R, Willighagen EL, Evelo CT et al (2016) Using the semantic web for rapid integration of WikiPathways with other biological online data resources. PLoS Comput Biol 12(6):e1004989

Walsh JR, Sen TZ, Dickerson JA (2014) A computational platform to maintain and migrate manual functional annotations for BioCyc databases. BMC Syst Biol 8:115

Wishart DS, Watson MS, Boyko RF, Sykes BD (1997) Automated 1H and 13C chemical shift prediction using the BioMagResBank. J Biomol NMR 10(4):329–336

Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N et al (2007) HMDB: the human metabolome database. Nucleic Acids Res 35(Database issue):D521–D526

Wishart DS, Knox C, Guo AC, Eisner R, Young N, Gautam B et al (2009) HMDB: a knowledgebase for the human metabolome. Nucleic Acids Res 37(Database issue):D603–D610

Wishart DS, Jewison T, Guo AC, Wilson M, Knox C, Liu Y et al (2013) HMDB 3.0–the human metabolome database in 2013. Nucleic Acids Res 41(Database issue):D801–D807

Wishart DS, Feunang YD, Marcu A, Guo AC, Liang K, Vazquez-Fresno R et al (2018) HMDB 4.0: the human metabolome database for 2018. Nucleic Acids Res 46(D1):D608–DD17

Worley B, Powers R (2014) MVAPACK: a complete data handling package for NMR metabolomics. ACS Chem Biol 9(5):1138–1144

Xia J, Wishart DS (2011a) Web-based inference of biological patterns, functions and pathways from metabolomic data using MetaboAnalyst. Nat Protoc 6(6):743–760

Xia J, Wishart DS (2011b) Metabolomic data processing, analysis, and interpretation using MetaboAnalyst. Curr Protoc Bioinformatics Chapter 14:Unit 14 0

Xia J, Wishart DS (2016) Using MetaboAnalyst 3.0 for comprehensive metabolomics data analysis. Curr Protoc Bioinformatics 55:14 0 1–0 91

Xia J, Bjorndahl TC, Tang P, Wishart DS (2008) MetaboMiner–semi-automated identification of metabolites from 2D NMR spectra of complex biofluids. BMC Bioinformatics 9:507

Xia J, Psychogios N, Young N, Wishart DS (2009) MetaboAnalyst: a web server for metabolomic data analysis and interpretation. Nucleic Acids Res 37(Web Server issue):W652–W660

Xia J, Mandal R, Sinelnikov IV, Broadhurst D, Wishart DS (2012) MetaboAnalyst 2.0–a comprehensive server for metabolomic data analysis. Nucleic Acids Res 40(Web Server issue):W127–W133

Xia J, Sinelnikov IV, Han B, Wishart DS (2015) MetaboAnalyst 3.0–making metabolomics more meaningful. Nucleic Acids Res 43(W1):W251–W257

Zhang F, Bruschweiler-Li L, Robinette SL, Bruschweiler R (2008) Self-consistent metabolic mixture analysis by heteronuclear NMR. Application to a human cancer cell line. Anal Chem 80(19):7549–7553

Zhang F, Robinette SL, Bruschweiler-Li L, Bruschweiler R (2009) Web server suite for complex mixture analysis by covariance NMR. Magn Reson Chem 47(Suppl 1):S118–S122

Zhu ZJ, Schultz AW, Wang J, Johnson CH, Yannone SM, Patti GJ et al (2013) Liquid chromatography quadrupole time-of-flight mass spectrometry characterization of metabolites guided by the METLIN database. Nat Protoc 8(3):451–460

Chapter 14
Drug Discovery: Concepts and Approaches

Varalakshmi Devi Kothamuni Reddy, Babajan Banaganapalli, and Galla Rajitha

Contents

V. D. K. Reddy (✉)
SKU College of Pharmaceutical Sciences, Sri Krishnadevaraya University, Ananthapuramu, AP, India

B. Banaganapalli
Princess Al-Jawhara Center of Excellence in Research of Hereditary Disorders, Department of Genetic Medicine, Faculty of Medicine, King Abdulaziz University, Jeddah, Saudi Arabia
e-mail: bbabajan@kau.edu.sa

Galla Rajitha
Institute of Pharmaceutical Technology, Sri Padmavathi Mahila Viswavidyalayam,
Tirupati, AP, India

N. A. Shaik et al. (eds.), *Essentials of Bioinformatics, Volume I*,
https://doi.org/10.1007/978-3-030-02634-9_14

14.1 Introduction

The process of drug discovery and development is arduous and has a long history dating back to the early days of human civilization. Throughout history, humans have searched either incidentally or intentionally for remedies to fight against diseases, exploring nature to meet two major needs – food and herbs for mitigating diseases. By this process new drugs were discovered from nature and later through scientific experimentation. In general, a disease or clinical condition that lacks suitable drug treatment triggers initiation of the drug discovery process.

The research and development of a new drug is an expensive and complex process, which, on average, takes around 10–15 years (Fig. 14.1) and costs around $1.8 billion (Zhong and Zhou 2014; Xiao et al. 2015). In this process, drugs were discovered by synthesizing novel compounds; this discovery involved multi-step procedures, in which lack of potency was determined (30%) and toxicity studies (11%) were carried out, with drug failures attributed to poor pharmacokinetic parameters (39%), side effects in humans (10%), and various commercial factors. Now the drug discovery process has been transformed, with advances in genomics, proteomics, informatics, high-throughput and virtual screening methods, quantitative structure-activity relationships, and structure-based drug design.

The process of drug discovery and the development of a new drug may be summarized as follows:

1. Design and synthesis of novel compounds and the study of their physicochemical parameters.
2. Preliminary biological evaluation, followed by specific biological evaluation of the compounds.
3. Study and analysis of toxicological parameters.
4. Target organ toxicological studies.
5. Acute and sub-acute toxicological studies.
6. Metabolic studies.
7. Synthesis and quality control of bulk material.
8. Phase I clinical evaluation, which includes the study of human toxicity and metabolic studies.

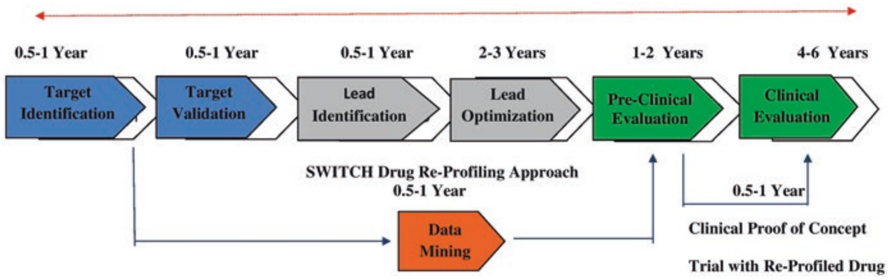


Fig. 14.1 Traditional drug discovery methods

9. Final formulation and physicochemical evaluation.
10. Phase II clinical evaluation, which includes dose titration and limited efficacy studies.
11. Phase III clinical evaluation, performed to study broad efficacy and tolerance in a large population of patients, as well as chronic toxicological studies.
12. Phase IV clinical evaluation, which includes post-marketing surveillance studies during general clinical use.

In the above summary, steps 1, 7, and 9 involve techniques/advances in medicinal chemistry and pharmaceutics, whereas steps 2–6 are covered by preclinical pharmacological evaluation of the drug in animals. Steps 8, 10, and 11 are in the domain of clinical pharmacological evaluation of the drug in humans. Step 12 is an ongoing process of surveillance to ensure safe use of the drug. If all these steps are satisfactorily passed by the candidate drug, it is granted registration as a new drug designed for the treatment of a specific disease for use in humans.

14.2 Drug Discovery Process

14.2.1 Target Identification

Drugs that are discovered sometimes fail clinically, for two reasons; the first being pharmacological failure and the second being their adverse effects. Hence, the foremost key during the process of drug discovery and development is the identification and verification of the site of action (target), this is briefly outlined in the Fig. 14.2. The word 'target' is a broad term that covers a huge range of biological moieties such as proteins, genes and nucleic acids. Amalgamating expertise in medicinal chemistry and biological sciences has resulted in redefined criteria for target selection by considering a drug's biological and pharmacological properties (Zhou and Zhong 2017; Hughes et al. 2011). The selected target should be essential for the growth or survival of the organism under a chosen condition that is considered as most essential, for example the main promiscuous targets choosen in the treatment of tuberculosis are ATP synthase, RNA polymerase, gyrase etc. of the Mycobacterium tuberculosis. The target should be vulnerable to chemical inhibition – a property commonly referred to as druggability, which means that it should be able to bind with high affinity to the presumed drug moiety; after forming a bond, or on interaction. The presumed drug moiety should develop pharmacological activity that can be established by both in-vitro and in-vivo methods. The refinement of this step helps to determine the correlation between the specific target and the disease, and hence facilitates the assessment of whether or not the mechanism of the target modulation causes unwanted effects.

The data mining process represents a milestone in target identification. It is an essential approach that is frequently applied to determine patterns from large data sets and hence to create a structure for future use (Yang et al. 2009). The data mining approach uses bioinformatics tools that aid in identifying, selecting, and priori-

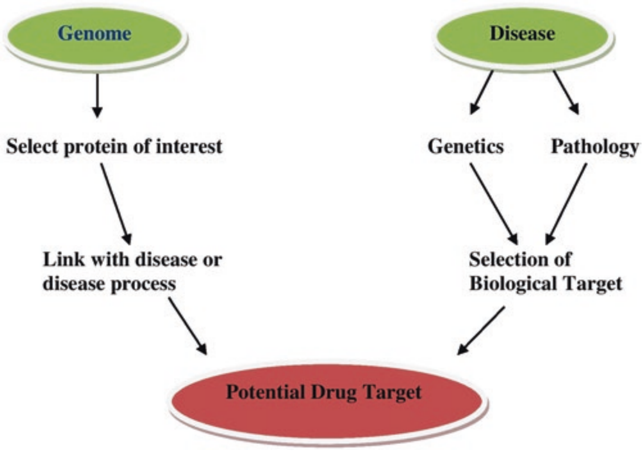

Genome
Select protein of interest
Link with disease or disease process
Disease
Genetics
Pathology
Selection of Biological Target
Potential Drug Target

Fig. 14.2 Approaches in drug target identification

tizing potential drug targets. The data available is obtained from various sources, which include research papers and filed patents, proteomics data, transcription data, and transgenic phenotypic statistics. Another approach includes the examination of the mRNA or protein to analyze its manifestation during the disease; the results of this examination are correlated to find the relation between the protein and disease progression. The target process with the most fruitful results is to consider genetic associations, the best paradigm being in the characterization of familial Alzheimer's disease.

14.2.2 Target Validation

The next step after target identification is the unambiguous validation of the target, which involves demonstration of the pertinence of the site of action and thus can confirm the causes of the specific disease; modification of the validated target may probably be the reason for the drug's pharmacological efficacy or activity. This process of validation involves exhaustive functional group characterization, authentication of the pathway, and modification of the activity of the protein to establish its association with the disease phenotype. The process of target identification and validation is depicted in Fig. 14.3.

The basic concept of antisense therapeutics is to involve in the binding of antisense moieties, forming double-stranded series, thereby inhibiting the translation or promote degradation of the targeted mRNA. In this technique, antisense DNA/RNA and RNA interference (RNAi), which are complementary to a distinct series of RNA and DNA, are used (Henning and Beste 2002). The best example of the result of this technique is fomiviresin, an antisense drug for viral retinitis which has been

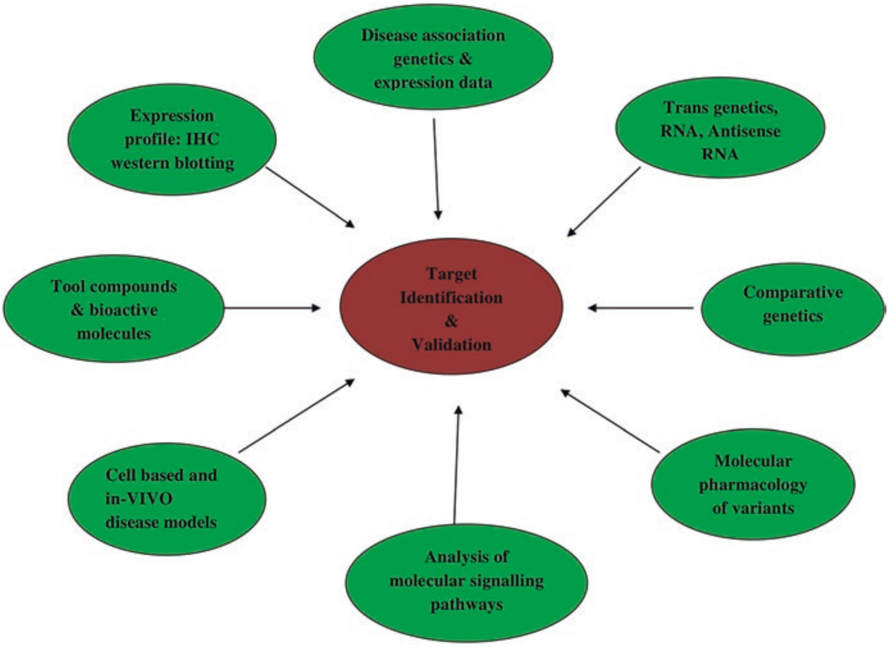

Disease association genetics & expression data
Expression profile: IHC western blotting
Trans genetics, RNA, Antisense RNA
Tool compounds & bioactive molecules
Target Identification & Validation
Comparative genetics
Cell based and in-VIVO disease models
Molecular pharmacology of variants
Analysis of molecular signalling pathways

Fig. 14.3 Multifunctional process of target identification and validation

approved by U.S Food and Drug Administration (FDA). But the process integrated while generating the oligonucleotides emanated the moieties with good amount reaching the systemic circulation and even noticeable harmful effects, creating their in vivo usage doubtful and even insufficiency in distinctiveness in choosing the suitable nucleotide probes limited their utilization.

Another attractive validation tool is the use of transgenic animals to explain the functional consequence of gene manipulation and to observe the phenotypic endpoints. In this approach, during the initial period of gene targeting, animals are produced that lack a gene for a specific process/activity from birth to the end of their life. One good example is the use of the P2X7 knockout mouse to establish the role of ion channels in the progress of neuropathic and inflammatory pain (Chassell et al. 2005). Monoclonal antibodies are exceptional aids for site evaluation. During the process of target validation, there will be interface to a considerable extent on the site of action. This allows effective distinction among the equivalent sites. In divergence to this, small molecules often cannot interact with the protected site of action.

More recently, chemical genomics, which is the study of genomic responses to chemical compounds, has emerged as a tool for target identification and validation. This approach facilitates the rapid identification of novel drugs and targets during the early phase of drug discovery, with the goal, in anticipation, of providing moieties that will target every protein encrypted by the genome.

14.2.3 Lead Discovery

The identification of hits and lead compounds is crucial in the drug discovery process. There are several approaches for lead identification. To identify lead compounds from arrays of synthesized compounds, they need to be screened for biological activity. Once the lead is identified, it can be structurally modulated to improve its potency.

A ‘hit’ can be defined as a chemical compound that exhibits the desired biological activity during the screening process and reproduces the activity upon retesting. Various screening paradigms are available for screening and identifying hit molecules. High-throughput screening (HTS) is the most popular, authentic, and rational approach for identifying new chemical entities with therapeutic efficacy to prove the distinctiveness of the action of chemical molecules against a specific site. This is a robust assay method for the recognition of actual hits. It is carried out in 96- or 384-well plates; this helps in screening huge numbers of molecules, about 10,000, concurrently (Fox et al. 2006). Some researchers use an ultra-HTS approach to evaluate up to 100,000 molecules per day. Once a library is entrenched, then it can be utilized for various assays. When the exact hit molecule is identified, it is refined further to improve its selectivity, binding properties, potency, and physicochemical properties by a process called hit-to-lead development. Other screening techniques are also used for identifying lead molecules; for example, focused screen, fragment screen, structural-aided drug design, physiological evaluation, and nuclear magnetic resonance screening. In recent years, pharmaceutical companies have become associated with large institutes, with the aim of discovering targets and gathering molecular data and infrastructure to screen novel compounds and to optimize the screening ‘hits’ into clinical candidates.

The probability of effectiveness of small-molecular-weight molecules assembled in compound libraries should obey biological variables such as Lipinski’s rule of five – the molecule should have a molecular weight of 500 or less; c log P value less than 5 (to estimate the lipophilic nature of the drug, which alters the absorption parameters), should have fewer than 5 hydrogen bond donors, also fewer than 10 hydrogen bond acceptors and not more than 3 rotatable bonds. The term ‘drug-like’ implies molecules with these features; most marketed drugs have a molecular weight of less than 350 and a c log P value of less than 3.

The following are a few important approaches for lead identification:

14.2.3.1 Random Screening

Random screening is considered to be a valuable approach when a known chemical entity or any other compound with the desired activity is unavailable. This method involves no intellectualization. The new chemical moieties are subjected to a battery of screening tests to establish their different biological activities. This was the only approach in 1935 and even now it is an important approach, considered as the method of choice to discover drugs or leads and when nothing is known about the receptor target (Giridhar 2012). The screening tests include studies of animal behavior, isolated tissues, intact animals, and some animal models of the disease of interest. Random screening is a sort of blind hitting to hit a nail head. Streptomycin and the tetracyclines are drugs that were discovered during a random screening process.

14.2.3.2 Serendipity

Serendipity refers to an accidental discovery that is made while searching for something else (Baumeister et al. 2013). Serendipity has led to the introduction of many useful drugs in the past; for example, penicillin as an antibacterial agent, and lignocaine and phenytoin as antiarrhythmics. Serendipity also leads to new uses being found for old drugs and to drug side effects being employed as therapeutic applications.

14.2.3.3 Molecular Modelling

Molecular modeling is an essential and established computational tool box for medicinal chemists that assists in early drug discovery and development. Nowadays molecular modeling has become an integrated part of investigating, predicting, and explaining the molecular and biological properties of organic molecules, thus establishing them as potential drug candidates. The incorporation of this method can bring about helpful insights into the behavior of the molecules and make the drug discovery process more efficient and fruitful.

14.2.3.4 Molecular Manipulation

In this method analogues of existing established drugs are synthesized and evaluated for their biological activity. This is a more scientific and logical approach than other approaches discussed and may yield newer moieties with added advantages such as increased rates of absorption, greater potency, and enhanced selectivity, and thus fewer side effects. Most of the synthetic penicillins and cephalosporins were developed in this manner.

14.2.3.5 Molecular Designing

The molecular designing approach aims to design compounds to fulfill a specific biological activity; hence, it is considered to be the most rational form of drug research and development. This method may involve the synthesis of naturally existing substances, such as a precursor of a neurotransmitter like dopamine for cardiac shock, a hormone, or a vitamin.

14.2.3.6 Metabolites of Drugs

The active metabolite formed in the body after the metabolism of a drug continues to produce therapeutic effects in the body and sometimes shows advantages over the parent compound. The pharmacologically significant metabolites may be subjected for structural modification to attain the required chemical stability, efficay and selectivity. The simplest and best example is nortriptyline, a an active metabolite of amitriptyline, which is more effective, potent and with enhanced selectivity.

14.2.3.7 Combinatorial Chemistry

In the present era of medicinal chemistry, combinatorial chemistry has emerged as a technique that generates billions of new compounds to produce libraries, which are originally screened by employing robotic high-throughput screening tools. A compound with a positive response is examined by applying conventional research approaches; then the moiety is subjected to additional modification to intensify its effectiveness (Liu et al. 2017).

14.2.3.8 Genetic Medicines

Synthetic oligonucleotides are being developed to target sites on nucleic acids especially on DNA sequences or genes or messenger RNA so that the generation of disease related protiens is blocked. This approach is worthwhile in the treatment of cancers and viral diseases without harming healthy tissues (Cohen and Hogan 1994).

14.2.3.9 Gene Therapy

Gene therapy is a promising therapeutic option for many diseases, including some cancers, as well as genetic disorders. In this strategy a nucleic acid, generally in the form of DNA, is given to alter the genetic repertoire for the treatment of diseases.

14.2.4 Lead Optimization

The molecules (i.e., hit molecules) that are considered to meet the basic objective of the lead optimization step are then considered for characterization before being subjected to preclinical studies. From this stage researchers proceed with their discovery work to generate potential backup molecules. Lead optimization is accomplished through the synthetic manipulation of the hit molecule the structural activity relationship (SAR) approach and a structure-based method, if structural data about the site of action is ready for use. All the data gathered will be used in the development of the target data, along with information about toxic effects and chemical manufacture. This information will be further utilised in the preparation of regulatory compliance which in turn permits the use of this molecule in humans.

14.3 Drug Development Process

A new chemical entity is recognized and optimized at the drug development stage when it has been proven to have drug-like properties and potency in the in-vitro studies. Despite the laboratory results, the differences between these investigations and results in humans should be diminished. Now the chemical entity must be subjected to the developmental process, a precarious step in the drug discovery procedure. The SARs need to be determined. After synthesis, the structure of the new compound and its purity is determined and confirmed by analytical techniques. Further, investigation of the promising moiety may be done in two phases: animal studies – preclinical pharmacology, and human studies – clinical pharmacology. Then the novel drug candidate is subjected to validation of its pharmacological actions and toxicity studies. Once the drug's potency is established, detailed toxicity studies – acute, subacute, and chronic – and metabolic studies are carried out.

The preclinical data are scrupulously screened, scrutinized, and analyzed by the drug control authority of the country, and if the drug is considered to be safe that authority then issues permission for human trials. The candidate drug is then subjected to clinical trials; once the candidate drug clears the strict evaluation channel it then passes from the laboratory to the market for use in humans for the treatment of disease.

14.3.1 Preclinical Studies

Preclinical studies involve extensive pharmacological testing of the drug candidate by in-vivo studies (i.e., in an animal population) and in-vitro studies (i.e., in test tubes or a research laboratory). These studies are designed to estimate the initial potency, harmful effects, and ADME (absorption, distribution, metabolism, and excretion) parameters to enable the relevant pharmaceutical company to understand whether or not it is valuable to proceed with the evaluation process. Based on the results of these studies, further specific tests are performed to screen for anticancer, antiarrhythmic, anti-inflammatory, anticonvulsant, anti-depressive, and tranquillizing effects, and other pharmacological properties of the drug. The experimental animals used include rats, mice, guinea pigs, dogs, and sometimes monkeys.

The major areas covered under the category of preclinical evaluation of the drug are its toxicity profile, safety and efficacy evaluation, and pharmacokinetics profile (ADME studies).

14.3.1.1 Acute Toxicity Testing

Acute toxicity studies are most commonly conducted to determine the effect of a single dose on a specific animal species; this testing permits the median lethal dose (LD_{50}; i.e., the dose that is lethal to 50% of the test population of animals) of the investigational product to be established (Parasuraman 2011). The studies are designed to take place for a period of 14 days in two distinct species (one rodent and one non-rodent). Toxic symptoms such as convulsions, tremors, and hyperactivity are examined, and all mortalities that occur during the drug study are documented, and morphological, biochemical, pathological, and histological changes in the dead animals are examined . The drawbacks of this method are the large numbers of animals involved and their high mortality rate. To overcome these limitations, a few modified methods, such as the fixed dose procedure (FDP), the acute toxicity category method (ATC), and the up and down method (UDP) have been developed.

14.3.1.2 Sub-acute Toxicity Testing

Sub-acute toxicity tests are designed to estimate the toxicity of the drug under investigation after repeated administration; these tests assist in establishing doses for long-term sub-chronic studies. Various laboratory studies, including hematologic examinations and hepatic and renal function tests are conducted, and the results are carefully observed. The animals are maintained at the maximum tolerated dose for about 2–3 weeks for observation of the development of pathological changes. If any changes are noted, the animals are killed and complete histopathological examination is performed.

14.3.1.3 Chronic Toxicity Testing

The goal of performing chronic toxicity studies is to determine the adverse effects of long-term exposure to the investigational drug/chemical. These tests usually use two species of animals, one rodent and one non-rodent. The study period comprises many months and during this period detailed biological and histopathological parameters are evaluated.

Today, toxicological studies of the effects of all new drugs on reproduction and development have become mandatory. These studies assumed prime importance after the thalidomide disaster of the late 1950s, which left more than 10,000 infants congenitally deformed and crippled. The tests carried out are:

(a) *Tests of fertility and reproductive performance*, which are usually carried out in rats, where they are treated with the new drug before and after the mating period. The effects on the early and late stages of embryonic and fetal development and lactation are analyzed and documented.
(b) *Teratological studies*, which are usually carried out in two animal species to ascertain the effects of the drug on the process of organogenesis. The drug is given after mating, during the period of organogenesis. The fetuses are carefully examined for visceral and skeletal abnormalities; the number of live and dead fetuses is recorded; and resorption sites in the uteri and corpora lutea are examined.
(c) *Studies of the adverse effects on the mother and offspring,* which are carried out in the perinatal and postnatal periods by administering the new drug during the last third of pregnancy, up to the time of weaning. Observations are made for adverse effects on labor and lactation and for direct toxic effects on the newborn.

In all these tests control groups of untreated animals in sufficient numbers must be studied to accurately assess the drug effects, if any.

14.3.1.4 Therapeutic Index (TI)

The therapeutic index (TI) , also referred to as the therapeutic ratio, is the relative margin of safety of a drug. In the early days of pharmaceutical toxicology, first the LD_{50} for the drug was determined and then the dose that was effective in 50% of the test population, termed the median effective dose (ED_{50}), was estimated. The objective was to elucidate the benefit/risk ratio. However, in clinical pharmacology/medicine the TI based on the LD_{50} and ED_{50} is not valid. Instead, in the clinic a dose that has toxic effects in fewer than 50% of humans (e.g., a specified increase in heart rate in the case of an adrenoceptor agonist) can be related to that dose which is effective in 50% of patients with bronchial asthma (e.g., a specified decrease in airway resistance of an adrenoceptor agonist).

In animals, the ratio of the LD_{50} to the ED_{50} is the TI:

$$\text{TI in animals} = LD_{50} / ED_{50}$$
$$\text{TI in Humans} = TD_{50} / ED_{50}$$

In humans, the ratio of the toxic dose in 50% of patients (TD_{50}) to the ED_{50} is the TI.

For a drug to have a greater safety profile a high TI value is preferable. In animals a modification of this concept can be applied. The TI can be calculated as:

$$\text{TI} = \frac{\text{Plasma concentration causing adverse effect}}{\text{Plasma concentration causing therapeutic effect}}$$

In humans the TI data are not available for many drugs; however, this concept provides a sensible way of correlating the versatility of one drug with other parameters, i.e., safety in relation to efficacy.

Thus, the TI has not been regarded as having much significance, and it has little value as a measure of the clinical usefulness of a drug. It is often pointed out that digoxin is a very useful drug despite its low TI. The benzodiazepines have replaced barbiturates as hypnotic drugs because their TI is very high compared with that of the barbiturates. To sum up, the TI provides a valid general concept, but it provides no measure of the actual usefulness of a drug.

14.3.1.5 Pharmacokinetic Parameters

All promising new compounds that has proven to be worthwhile are further subjected to pharmacokinetic studies. These are performed in several species of animals to establish the relative bioavailability of the compound upon oral or parenteral administration. Information regarding the elimination half-life is useful for estimating optimal dosage intervals. The ADME data obtained in animal studies are cautiously applied to humans. This information is useful if further testing is warranted.

14.3.2 Clinical Trials

The preclinical data obtained from animal studies provides complete information regarding the pharmacological, toxicological, and pharmacokinetic parameters of the new drug to the pharmaceutical companies. The data obtained is scrutinized by expert government bodies in each country to confirm whether or not to proceed with clinical drug trials. Once the data is approved, then, with great care and meticulous planning, the methodology adopted will be implemented. The new drug application, in the prescribed format, with all the relevant literature and preclinical data, must be submitted to the relevant drug control authority for scrutiny, and after approval clinical evaluation studies are initiated.

In the United Kingdom the introduction of new drugs is regulated by the Committee on Safety of Medicines (CSM), and in the United States this regulation is done by the Food and Drug Administration (FDA). In India the Drug Controller, Government of India, based in New Delhi, is responsible for the organization of this system. Only when approval is given by these organizations can the drug can be administered to humans for clinical evaluation.

To design a perfect clinical trial much thought and expertise is needed, and perfect team work is involved. Some salient guidelines are:

(i) Ethics and patient selection:

The clinical research carried out by the scientists/doctors working with patients or healthy volunteers should follow the recommendations of the Declaration of Helsinki of the World Medical Association. Consent must be obtained in writing from the subjects (patients or volunteers), or their guardians if the patient is incapable of giving consent. The subjects must be informed that a new drug is being tried that may be beneficial; however, a calculated risk is involved. The new treatment is compared with the known conventional treatment. Use of placebo controls is unethical and is not permissible if an effective remedy is available for the disease. Criteria for the selection of patients should be well thought out and defined. Special care must be taken if more than one doctor is involved in the selection of patients in the trial, especially in multicentric trials.

(ii) Response measurements. The end-points should be clearly defined. It is useful to define non-responders. Side effects should be carefully observed and recorded.

(iii) Experimental design. The design of the trial must be statistically sound, for which, preferably, a biostatistician should be consulted. In general, controlled clinical trials must include four safeguards against bias: (a) double- blind technique; (b) randomization of treatment; (c) matching of patients; and (d) cross-over technique.

The final stage of a clinical trial is the statistical analysis of the data obtained. Relatively simple tests like the Student's *t*-test or the Chi-square test may be sufficient to determine the significance of results. Complex statistical methods include non-parametric tests, analysis of variance and covariance, and dispersion and sequential techniques.

14.3.2.1 Phases of a Clinical Trial

Phase I: Clinical Pharmacologic Evaluation

Phase I trials are dose escalation studies, considered as the initial stage for testing of the drug in groups of about 20–80 healthy volunteers or patients, depending on the class of drug and its safety. These studies are designed to evaluate the safety, toxicity, tolerability, pharmacological actions, pharmacodynamics, and pharmacokinetics of the drug. The subject is observed until the drug is completely eliminated from the body, which permits in designing the therapeutic dose of the drug half-lives. Under some circumstances patients who are in the final stage of a disease and for whom there is no alternative medication are used in Phase I trials, especially in various cancer and HIV drug trials. Phase I clinical trials may be single ascending dose (SAD) studies or multiple ascending dose (MAD) studies based on the nature of the study design.

Phase II: Controlled Clinical Evaluation

After confirming the fundamental efficacy and safety of the investigational chemical moiety during the Phase I clinical studies, Phase II studies are performed in large groups of 30–300 patients. In fact, these studies are extensions of Phase I studies that are intended to ascertain the effectiveness of the drug as well as establishing the safety of the drug under strictly controlled conditions. This step is crucial in the drug development process. The drug under investigation fails if it does not elicit the expected activity or if it elicits unwanted results. Late Phase II trials are done in a controlled double-blind manner, and at the end of these studies one should be convinced about the therapeutic usefulness of the drug.

Phase III: Extended Clinical Evaluation

Phase III trials are formal therapeutic trials that are carried out, preferably in a double-blind controlled manner, in 300–1000 patients and are considered as randomized controlled multicenter trials. Phase III trials are of relatively long duration, and are costlier and more laborious to plan and implement than Phase II trials. If, after Phase III studies, the drug control authority is satisfied regarding the safety and efficacy of the drug then it will be approved by that authority. Most of the drugs that meet these trial requirements will be marketed under FDA norms with proper recommendations through a New Drug Application comprising all the data regarding manufacturing details, preclinical and clinical studies. After this step, the drug is marketed for general use.

Phase IV: Post-marketing Surveillance

Phase IV trials include pharmacovigilance and ongoing technical support for the drug after its sale has been authorized, for example after getting approval under the FDA Accelaerated Approval Program. On clinical use over many years unexpected harmful reactions may occur. Harmful effects discovered may result in the withdrawal of the drug from the market or the drug's restricted use. Current examples of such drug withdrawals are cerivastatin and troglitazone.

Investigational New Drug (IND)

An investigational new drug application needs to comply with suitable regulatory authorities for procuring consent to conduct investigational research, which includes the evaluation of a new dosage form or new pharmacological activity of a drug that has already gained permission to be marketed. The protocol of testing needs to be approved by suitable regulatory authorities or by an independent review board (IRB) or ethical advisory board. An IRB committee is an independent one consisting of physicians, community advocates, and others, to ensure that the clinical trial is ethical.

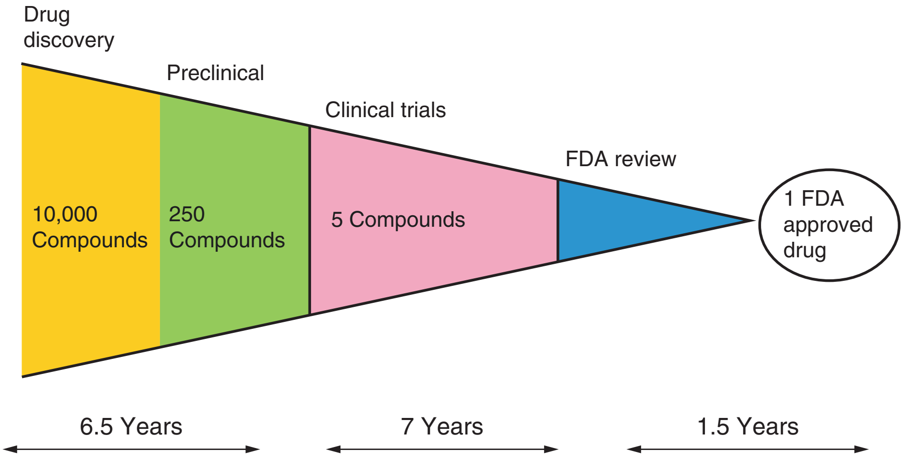


Fig. 14.4 Drug discovery time line

New Drug Application (NDA)

An NDA is an application to market a novel drug. This application is a record with the safety and potency details of the investigational drug and it contains all the data gathered at the time of the drug development process. After the successful completion of the preclinical and clinical testing, all these sequences of reports are submitted to the FDA in the United States, or to the respective regulatory authorities in other countries. The application provides substantial evidence regarding the usage of the drug and the conditions for which it is specified, along with other considerations as specified on the label. The entire process of drug discovery is briefly depicted in Fig. 14.4.

14.4 Conclusion

The drug discovery and development process is remarkably an interesting and challenging area because of the emergence of different new diseases. Drug discovery in modern medicine is a costly, laborious process with low rate of success; it requires huge investments from pharmaceutical companies, as well as grants from national governments. According to the statistical data in the year 2010, the drug discovery process cost around $1.8 billion. A 'hit' molecule that clears preclinical and clinical trials may sometimes be withdrawn from the market owing to its adverse effects. Hence, developing a safe and effective drug involves the understanding of clinical strategies and legal and regulatory matters. Despite all the challenges, drug discovery has been revolutionized, converting many fatal ailments into diseases that can be treated with routine therapeutic practices (Landau et al. 1999).

References

Arun B (2009) Challenges in drug discovery: can we improve drug development. J Bioanal Biomed 1:050–053

Baumeister AA, Pow JL, Henderson K, Lopez-Munoz F (2013) On the exploitation of serendipity in drug discovery. Clin Exp Pharmacol 3:3

Chassell IP, Hatcher JP, Bountra C, Michel AD, Hughes JP, Green P (2005) Disruption of the P2X7 purinoceptor gene abolishes chronic inflammatory and neuropathic pain. Pain 114:386–396

Cohen JS, Hogan ME (1994) The new genetic medicines. Sci Am 271:50–55

Fox S, Farr-Jones S, Sopchak L, Boggs A, Nicely AW, Khoury R et al (2006) High-throughput screening; update on practices and success. J Biomol Screen 11:864–869

Giridhar R (2012) Drug discovery: past and present. J Adv Pharm Technol Res 3(1):2

Henning SW, Beste G (2002) Loss of function strategies in drug target validation. Curr Drug Discov 2:17–21

Hughes JP, Rees S, Kalindjian SB, Philpott KL (2011) Principles of early drug discovery. Br J Pharmacol 162:1239–1249

Landau R, Achilladelis B, Scriabine (1999) Pharmaceutical Innovation, Philadelphia: Chemical Hertage Press

Liu R, Li X, Lam KS (2017) Combinatorial chemistry in drug discovery. Curr Opin Chem Biol 38:117–126

Medina-Franco JL (2012) Drug discovery with novel chemical libraries. Drug Des 1:e105

Nicolson TJ (2010) The post transcriptional regulator EIF 2S3 and gender differences in the dog. Implications for drug development, drug efficacy and safety profiles. J Drug Metab Toxicol 1:101

Parasuraman S (2011) Toxicological screening. J Pharmacother 2(2):74–79

Sanchez HEP (2012) Exploitation of massively parallel architectures of drug discovery. Drug Des 2:e108

Sang N (2011) Biochemistry, drug development and open access. Biochem Pharmacol 1:e101

Tamimi NAM, Ellis P (2009) Drug development: from concept to marketing. Nephron Clin Pract 113:c125–c131

Torzewski J (2011) Road map to drug discovery and development – inhibiting C-reactive protein for the treatment of cardiovascular disease. J Bioequiv Availab S1:001

Xiao X, Min JL, Lin WZ, Liu Z (2015) iDrug-Target: predicting the interactions between drug compounds and target proteins in cellular networking via the benchmark dataset optimization approach. J Biomol Struct Dyn 33:2221–2233

Yang Z, Marotta F (2012) Pharmacometabolomics in drug discovery and development; applications and challenges. Metabolomics 2:e122

Yang Y, Adelstein SJ, Kassis AL (2009) Target discovery from data mining approaches. Drug Discov Today 14:147–154

Zhong WZ, Zhou SF (2014) Molecular science for drug development and biomedicine. Int J Mol Sci 15:20072–20078

Zhou S-F, Zhong W-Z (2017) Drug design and discovery – principles and applications. Molecules 22:279

Chapter 15
Molecular Docking

Babajan Banaganapalli, Fatima A. Morad, Muhammadh Khan, Chitta Suresh Kumar, Ramu Elango, Zuhier Awan, and Noor Ahmad Shaik

Contents

B. Banaganapalli · R. Elango
Princess Al-Jawhara Center of Excellence in Research of Hereditary Disorders, Department of Genetic Medicine, Faculty of Medicine, King Abdulaziz University, Jeddah, Saudi Arabia
e-mail: bbabajan@kau.edu.sa; relango@kau.edu.sa

F. A. Morad · M. Khan
Princess Al-Jawhara Al-Brahim Center of Excellence in Research of Hereditary Disorders, King Abdulaziz University, Jeddah, Saudi Arabia

C. S. Kumar
Department of Biochemistry, SK University, Sri Krishnadevaraya University, Anantapur, India

Z. Awan
Department of Clinical Biochemistry, Faculty of Medicine, King Abdulaziz University, Jeddah, Saudi Arabia

N. A. Shaik (✉)
Department of Genetic Medicine, Faculty of Medicine, King Abdulaziz University, Jeddah, Saudi Arabia
e-mail: nshaik@kau.edu.sa

N. A. Shaik et al. (eds.), *Essentials of Bioinformatics, Volume I*,
https://doi.org/10.1007/978-3-030-02634-9_15

15.1 Introduction to Molecular Docking

Molecular docking (MD) is a quick way to predict the orientation of ligand-receptor complex, taking into consideration the known structure of the receptor (Taylor et al. 2002). MD can potentially explore different aspects of ligand-receptor binding characteristics like complementarity and affinity, etc. The techniques that are commonly used in MD are genetic algorithms, molecular dynamics, simulated hardening, Monte Carlo algorithms, purpose complementary ways, distance pure mathematics ways, fragment-based ways, tabu searches, and systematic searches. There are two major steps involved in docking procedure, first is a search algorithm and second is a scoring function. The search algorithm can differentiate between conformational changes of the ligand through one of the techniques mentioned above. Different possibilities of binding between the ligand and the receptor can be listed by systematic searches. This can be a great time-consuming process especially in case of large substrates with elastic shape; therefore, time rate between exploring the conformational domains and adequate computational time for searching should be measured. Scoring function usually classify different shapes retrieved by the search algorithm. Effective scoring function can easily highlight the experimentally obtained structure from other predicted structures retrieved through the search algorithm. Commonly used techniques are empirical free energy scoring functions, molecular mechanics force fields, and knowledge-based functions (Taylor et al. 2002). Several computational servers like DOCK (Kuntz 1992; Ewing et al. 1997), GOLD (Vicente et al. 1997), AutoDock (Morris et al. 2008), Surflex (Spitzer & Jain 2012), and FlexX (Schellhammer & Rarey 2004; Kramer et al. 1999; Kramer et al. 1997) are available to execute the MD procedures. These programs differ from each other in searching and implementation of algorithms and scoring methods. These programs mostly hold the receptor in rigid form, allowing a certain degree of flexibility to the ligands.

15.2 Genetic Algorithms

The genetic algorithms are known to be stochastic universal optimization strategies (Judson 1996). These algorithms can be customized for a variety of optimization issues as they don't employ gradient information as an input. And it also explores the parameters that form the three-dimensional structure at the same time. The genetic algorithms are simplified in Fig. 15.1.

Three local minima named as I, II (the global minimum), and III are illustrated in the f(x) function, where the terms of genetic algorithm were inserted as fitness, populations, and chromosomes. The fitness is represented in function f(x) where the populations are a group of individuals representing the conformational space. The x refers to chromosomes which are parameters forming each individual. Additional expression terms for genetic algorithm are mutations, selection, crossovers, and

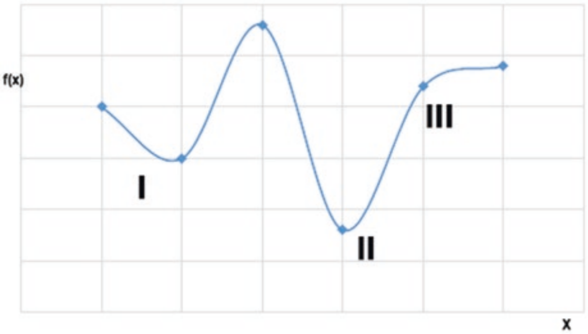


Fig. 15.1 A sample one-dimensional fitness function illustrating local minima labeled as I, II, and III

migrations (Westhead et al. 1997). The operator of the mutation chooses individuals through random changes of the chromosomes. The best individuals based on the fitness function are then chosen for crossover, which basically permits the swapping of chromosome sets between parents. In migration process, chromosomes of individuals are transmitted between subpopulations. All the above techniques persist to the point where certain stopping criteria run across. In the following sections, an introduction to most commonly used three docking programs is described.

15.3 Molecular Docking Tools

15.3.1 FlexX

FlexX tool characterizes the interaction characteristics between protein and ligand molecules (Bohm 1992). The interacting group in any molecule which is to be docked, will be assigned with interaction type and corresponding compatibility. Some examples for interaction types are geometrically restricted hydrogen bonds, metal and metal acceptor interactions, and hydrophobic interactions, for instance, phenyl ring and methyl group interactions. The special contact geometry for each group which forms an interaction, is defined by allowing the interaction surface to prevail over the centre of the molecule, as part of the sphere. An interaction takes place once the center of one group coincides with the interaction surface of an opposite group. FlexX docking algorithm is mainly built on three stages of gradual construction design as follows: (1) base selection, where the base part of the ligand is chosen, and (2) position in the active site of the protein (base fragment placement). At the end, and starting from another position of the base fragment, (3) the ligand is gradually reconstructed (complex construction). Once new fragments to the ligand are added, additional interactions appear, and the highest scoring function rank is marked until the ligand is fully structured.

In FlexX docking algorithm is partially sensitive to the first two stages: selection and placement of base portion. If the structure of a fragment in a molecule used in docking is known previously, then the most useful and time-saving option is to pose that portion manually in the binding site through mapref command. Thus, besides minimizing the docking period, this step will also increase the chance of foretelling the highest binding mode for the ligand. The updated version of FlexX tool is FlexX-Pharm (Hindle et al. 2002), which can facilitate recruiting more information about protein-ligand interactions prior to the docking procedure. Note that constraints are specified by chosen FlexX interactions and volumes that are included, leading the docking procedure to result in a group of docking solutions with specific attributes. By checking a spectrum of predicted phenomena throughout the elastic built of ligand fragments among the active site, FlexX-Pharm recognizes the specific construct docking solutions that are likely to follow the constraints. Those that are not following the constraints are mostly excluded, lessening time consumption and giving a chance for new solutions of docking to appear.

15.3.2 AutoDock

AutoDock (Adeniyi & Ajibade 2013; El-Hachem et al. 2017; Jiang et al. 2015) is an application made to set an automated operation in ligand-biomacromolecular interaction prediction. The simulation of docking procedure here uses one of the many available search methods. The genetic algorithm used in AutoDock program is Lamarckian, and Monte Carlo simulations. Ligand elasticity reaches to 52 dihedral angles and receptor is maintained in rigid form. Four phases AutoDock based docking procedure are ligand and receptor preparation, the AutoTors and AutoGrid procedure, the application of the genetic algorithm, and the fitness function plus free energy evaluation

15.3.2.1 Preparation of the Ligand and Receptor

In the following example, a docking box built of a grid with default measurements of 60 × 60 × 60 points and 0.575 Å grid spacing was used and positioned in the active side of the receptor. The long side of the box is placed in the direction of the binding site center covering the entry of that binding site. This way, the docking box covered the binding site completely along with some region beyond the binding entrance. Other possible sizes of the box that can be used in calculations are 82 × 60 × 60 and 110 × 80 × 80 with four grid point additions. In fact, the optimum parameter of the box that has been detected is 92 × 70 × 70. It provides the maximum flexibility for the ligand to have various conformations in the binding cleft and a very minimum time consumption of calculation due to its small size.

15.3.2.2 AutoGrid Procedure

AutoDock intakes pre-computed grid maps for each atom type that exists in the docked ligand to make the simulations fast. Grid maps are generated by AutoGrid and composed of a 3D structured lattice of evenly distributed points, surrounded entirely or partly, and placed in the center of some important sites of the studied macromolecule. All points in the lattice of the map stores the potential energy of a "probe atom" or functional group in each atom of the macromolecule. Figure 15.2 shows the main features of a grid map.

The illustration in Fig. 15.2 presents the docking box with the entire protein inside. The box size is determined by the lattice points with grid spacing of the user's setting. The substrate configuration energetics are calculated through a trilinear interpolation of affinity values for those grid points around each atom. The required time for energy calculation in the grid depends on the atom amount in the substrate, apart from the atoms in the protein.

15.3.2.3 AutoDock Genetic Algorithm Implementation

The configuration of the ligand-protein is marked by state variables that are composed of a group of variables representing the translation, orientation, and conformation of ligands with respect to proteins. Every state variable indicates a gene, and the ligand's state corresponds to a genotype, whereas its atomic coordinates correspond to the phenotype. In AutoDock application, the chromosome consists of real valued genes as following, three Cartesian coordinates for translation of the ligand, four elements that determine a quaternion that can specify the orientation of the ligand, and one real value for every ligand dihedral angle. The genetic algorithm starts with picking a number of individuals and initiating a group randomly. Every individual in the group is given hypothetically a certain value of genes. A series of generations occur and cycle to the farthest possible number or to the most amount

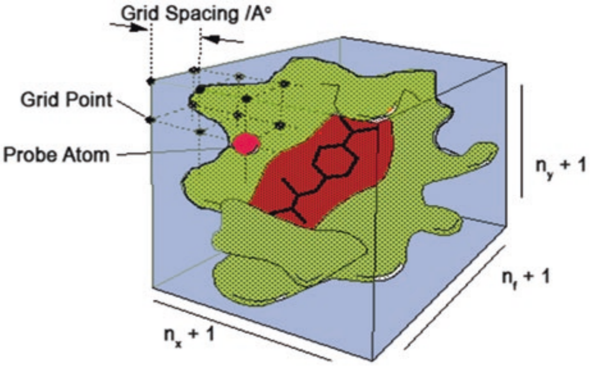


Fig. 15.2 Diagrammatic presentation of docking box and grid points

of energy available. One generation is composed of five sequential phases including mapping and fitness evaluation, selection, crossover, mutation, and elitist selection. The genotype of individuals translates into its parallel phenotype through the mapping stage, evaluating therefore the fitness of every individual. The function of fitness and the evaluation of energy are demonstrated in the following subsection, calculating the energy of each individual in every step. Then a proportional selection occurs to choose individuals who can reproduce. Some members are randomly chosen by the user parameters of crossover and mutation from the population to go under these stages. Once the crossover is performed, the next production of members will immediately replace their ancestors to remain the size of population firm. The next phase is mutation. Electively, some elite standards defined by the user automatically select superior members to move on the next generation. The algorithm repeats along the generations to the stage of meeting one of the termination criteria.

15.3.2.4 The AutoDock Fitness Function and Free Energy Calculation

The fitness is a result from the combined energies of the intermolecular interaction of ligand-protein and the intramolecular energy of the ligand. AutoDock presents at the final docking procedure the following: the fitness (the docked energy), the state variables, the coordinates of the docked conformation, and the estimated free energy of binding ΔG):

$$\begin{aligned}\Delta G = \Delta G_{\mathrm{vdw}} \sum_{i,j}\left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right) + \Delta G_{\mathrm{hbond}} E(t)\left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}}\right) \\ + \Delta G_{\mathrm{elec}} \sum_{i,j} \frac{q_i q_j}{\varepsilon(r_{ij})(r_{ij})} + \Delta G_{\mathrm{tor}} + \Delta G_{\mathrm{sol}} \sum_{i,j}\left(S_i V_j + S_j V_i\right) e^{\left(r_{ij}^2/2\sigma^2\right)}\end{aligned} \tag{15.2}$$

ΔG calculation including the solvation effect in AutoDock is shown in Eq. 15.2.

The first term is the Lennard-Jones 12–6 dispersion-repulsion. The second term is *a*, *t*, based on the angle *t*, between the probe and the target atom. The third term is a screened Coulombic electrostatic potential. The fourth term is a measure of the unfavorable entropy of ligand binding due to the restriction of conformational degrees of freedom that depends on the number of sp5 bonds in the ligand. Finally, the last term is Ntor, which indicates the desolvation effect. In desolvation, the approach utilized is the pair-wise, volume-based method of Stouten et al. The feature that distinguishes this approach is that it suits with the pre-calculated affinity grid formulation applied in AutoDock. Partial volumes of nearby protein fragments around each atom of the ligand are calculated using an exponential function and then summed. This can evaluate and measure the percentage of volume surrounding the atom of ligand that is occupied by protein atoms (Morris et al. 2008, 1996). The energy of desolvation is calculated after weighting the percentage using the atomic solvation parameter of the ligand atom. The complete procedure can be divided into four separate parts: burial of polar atoms in the ligand, burial of a polar protein atom, burial of polar and charged atoms in the ligand, and burial of polar and charged protein atoms. Major significant results have been presented in many experiments through measuring the "hydrophobic effect." They examined several formulations that contained only the volume lost around ligand carbon atoms (Morris et al. 2008, 1996). Some issues arise from buried polar atoms. Besides the volume-based approach, a simple formulation for solvent transferring of polar atoms was applied. A fixed term corresponding to the proper free energy of interaction of a polar atom with solvent is estimated, and this is subtracted from the binding free energy.

15.3.3 Gold

GOLD application has been used widely in molecular docking since 1997 (Genetic algorithms and their use in chemistry, reviews in computational chemistry 1999). GOLD stands for Genetic Optimization for Ligand Docking, and it employs a genetic algorithm to explore the structural scope. GOLD permits a complete elasticity for noncyclic ligands and the flexibility of partial protein in the area around the binding cleft. The docking strategy in GOLD will be evaluated in three stages: protein and ligand initialization, implementation of genetic algorithm, and the fitness function.

15.3.3.1 Protein and Ligand Initialization

GOLD software accepts the inputs like a point or an atom, and radius, from the users to determine the center of protein molecules. It also intakes the docking sphere that lay on the binding cleft of that particular protein. The binding mode of some receptors is well studied by X-ray crystallography, such as the structure of HLA-A2.1 receptors. The other software which can easily point the center of the binding cleft is CHARMM (Brooks et al. 2009). Within 10 Å distances around the ligand atoms, receptor atoms are determined, and the center of the whole chosen volume is found through "stats" option. The coordinates of the center are located as 4.00, 16.1, and − 6.70 in x, y, and z directions, respectively. The closest radii suggested to the center were 20, 25, and 50 Å. The 20 Å radius was favored due to the ability of balancing computational time and accuracy. Figure 3.12 illustrates the calculated dock sphere. All proteins are counted as rigid except OH groups of SER, THR, and TYR and NH4+ group of LYS around the active site. The ligands can be prepared fully flexible. Modest constraints should be considered while prepping the ligands, retaining the ring corners, amide bonds, planar nitrogens, and/or internal hydrogen bonds that are constant. Other possible constraints are the covalent constraints, distance constraints, H-bond constraints, structure-based constraints, and similarity constraints. The preferred number of runs is 10, although it ranges usually from 1 to 50. The favored format of the protein and ligand files is tripos mol2, although pdb can be used sometimes. In case of using pdf format, the program will specify partial charges using a modified Tripos force field. But mol2 files have the partial charge information; thus, other force fields are applied to prepare the mol2 files.

15.3.3.2 The Implementation of Genetic Algorithm

The genetic algorithm employed by GOLD is steady-state operator-based to represent the structural shape and binding modes of the ligand. The following 7 steps summarize the genetic algorithm used by GOLD: 1. A group of reproduction operators like crossover or mutation is selected. Every operator is assigned with a certain weight. 2. A random population is initially produced and the fitnesses of its members determined. 3. Depending on the weight, a specific operator is selected through roulette wheel selection. 4. The parents chosen by the operator are selected using roulette wheel selection based on the fitness grads. 5. Running the operator will result in new offspring chromosomes. The fitness is evaluated for each. 6. If not exist formerly in the population, the children substitute the least members fitting in the group. 7. Terminate after running 100000 operations, else return to step 3 once again. Mutation, crossover, and migration are the operators that have been used. The mutation operator brings individuals to the group/population through random changes of the rotatable bonds in the protein and ligand. Torsion angle values vary between −180° and 180° in step-sizes of 1.4°. Five groups of population are automatically set. Each consists of 100 individuals. The crossover operator obtains the exchange of chromosomes between the individuals. The migration operator creates copies from individuals in many groups. Operators were selected using roulette wheel selection depending on the operator weights. Weights were selected based on the occurrence of crossover and mutation in similar probability, while the migration only applied 5% of the time. The genetic algorithm stops once it reached the optimum number of operators (which is mostly 100,000). The ligand then positions in the active site using a least-squares fitting procedure, after the binding cleft is ready. The final step will evaluate and determine the fitness score.

15.3.3.3 The Fitness Function

Scoring function in GOLD is presented in two styles: ChemScore and GoldScore. In ChemScore, the scoring system is built empirically on the existing measures of binding affinities for a group of 82 protein-ligand complexes. It is trained by regression against measured affinity data. Equation 1 illustrates the free energy calculation of the binding (ΔGbinding):

$$\Delta G_{\text{binding}} = \Delta G_0 + \Delta G_{\text{hbond}} + \Delta G_{\text{metal}} + \Delta G_{\text{lipo}} + \Delta G_{\text{rot}}$$

Every element in this equation is a final outcome of an expression depending on the size of a specific physical contribution to free energy and a scale factor determined by regression. The final result of ChemScore is found after combining a clash penalty and internal torsion terms that can block close contacts in docking and poor internal conformations. Covalent and constraint scores can be added as well:

$$\mathrm{ChemScore} = \Delta G_{\mathrm{binding}} + P_{\mathrm{clash}} + C_{\mathrm{internal}} P_{\mathrm{internal}} + \left(C_{\mathrm{covalent}} P_{\mathrm{covalent}} + P_{\mathrm{constraint}} \right)$$

Four main objects build up the functional system of the GoldScore fitness: protein-ligand hydrogen bond energy (external H-bond), protein-ligand van der Waals (vdw) energy (external vdw), ligand internal vdw energy (internal vdw), and ligand torsional strain energy (internal torsion). A fifth additional element may be included, which is the ligand intramolecular hydrogen bond energy (internal H-bond). GOLD parameter files (editable by users) contain the fitness function empirical parameters: hydrogen bond energies, atom radii and polarizabilities, torsion potentials, hydrogen bond directionalities, etc. The score of external vdw is multiplied by a factor of 1.575 when the score of the total fitness is calculated. This is used as an empirical correction to encourage hydrophobic contact of protein-ligand complex. The final value of GoldScore is represented in the next equation:

$$\text{Gold Score} = -\left(\mathrm{H} - \text{Bond} - \text{Energy} + \text{Internal} - \text{Energy} + \text{Complex Energy} \right)$$

The first element refers to the hydrogen binding energy that can be found by summing all donor and acceptor atoms that are able to form hydrogen bond. The second term expresses the internal energy of the ligand, which is the combination of the ligand steric and torsional energies. Some molecular mechanics applications are used to compute this value. The form (4) with 6–12 potential is used to calculate the steric energy:

$$E_{ij} = \frac{C}{d_{ij}^{12}} - \frac{D}{d_{ij}^{6}}$$

The Tripos force field of the following form is used to calculate the torsional energy Eijkl:

$$E_{ijkl} = \frac{1}{2} V_{ijkl} \left[1 + \frac{n_{ijkl}}{\left| n_{ijkl} \right|} \cos\left(\left| n_{ijkl} \right| \cdot \omega_{ijkl} \right) \right]$$

The final element is an energy gathered from the steric energy of interaction between the protein and the ligand. For this calculation, the following formula of 4–8 potential with linear cutoff is recruited:

$$E_{ij} = \frac{A}{d_{ij}^{8}} - \frac{B}{d_{ij}^{4}}$$

The cutoff distance used was 1.5 times the sum of the van der Waals radii of the atoms. The 4–8 potential was parametrized to reproduce the minimum of the more usual 6–12 potential.

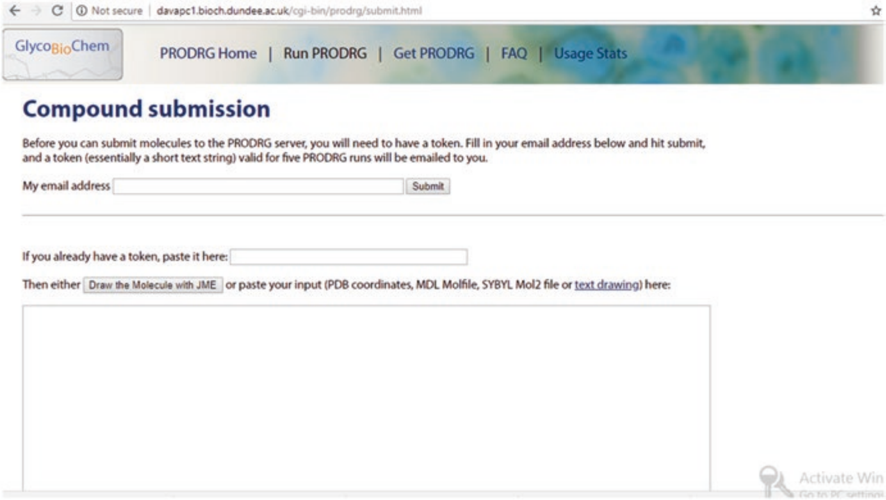
15.4 Applications of Docking

In the past decade, a huge amount of effort has been invested in the field of computational prediction and scoring function tools to provide efficient results for docking and other molecular interactions. Major progress has been achieved in predicting ligand-target binding modes via computer programs, and several review articles discussed this emerging field of science. DOCK is successfully applied when two main subjects are fulfilled: the in silico virtual screen with high throughput for substrates with high-affinity receptor and the computational design of selective inhibitors for specific receptor. Former docking experiments determined aminoglycosides for its ability to bond with the standard A-RNA duplex not the B-DNA form. Conformational evidence of NMR solvent isotope shift parameters shows that lividomycin (a proposed calculated compound) significantly enhances the results of docking through binding to the major groove of RNA, thus increasing the stability of RNA duplex. A study published by Filikov (Filikov et al. 2000; James et al. 2000) stated that lead compounds damage the binding of HIV-1 TART, which is an important interaction in viral replication. The newly published work from James et al. used DOCK and ICM with more advanced scoring scheme to produce a sub-micro-molar lead with a novel chemotype that showed anti-HIV activity in a cellular assay (Banaganapalli et al. 2013a; Banaganapalli et al. 2013b).

15.5 Protocol of AutoDock

Steps in Ligand-Receptor Docking Using AutoDock Tool
Lead compounds can be designed by drawing two-dimensional structure and then converting it into three-dimensional structure through the online server *PRODRG*. Lead compounds should be drawn on a plain white background and transferred to PRODRG server. Now the protocol given below provides complete information and demonstrates how to convert a 2D lead molecule into 3D structure.

1. Click the Internet explorer icon on the desktop and type the website at the address toolbar as www.google.co.in.
2. Type the word PRODRG server or directly go to the following link: http://davapc1.bioch.dundee.ac.uk/cgi-bin/prodrg/submit.html (Fig. 15.3).



Not secure | davapc1.bioch.dundee.ac.uk/cgi-bin/prodrg/submit.html

GlycoBioChem

PRODRG Home | Run PRODRG | Get PRODRG | FAQ | Usage Stats

Compound submission

Before you can submit molecules to the PRODRG server, you will need to have a token. Fill in your email address below and hit submit, and a token (essentially a short text string) valid for five PRODRG runs will be emailed to you.

My email address [ ] Submit

If you already have a token, paste it here: [ ]

Then either Draw the Molecule with JME or paste your input (PDB coordinates, MDL Molfile, SYBYL Mol2 file or text drawing) here:

Activate Win

Fig. 15.3 Home page of PRODRG web server

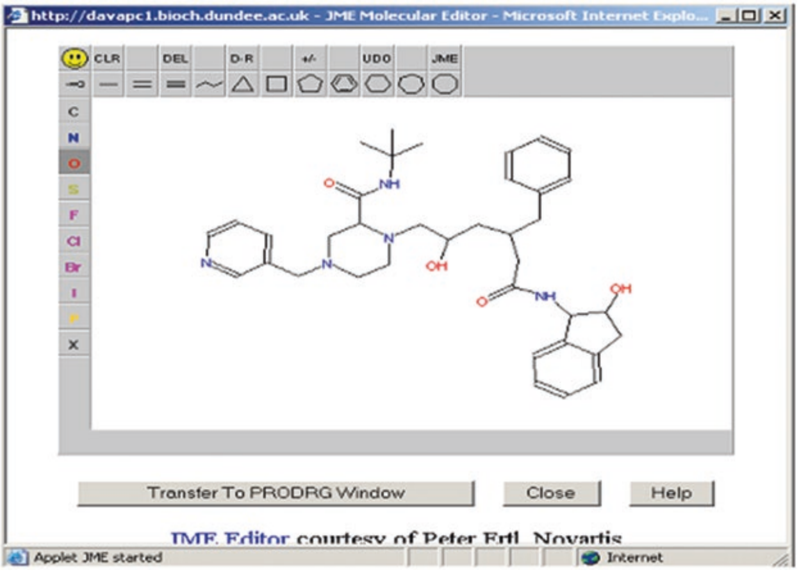
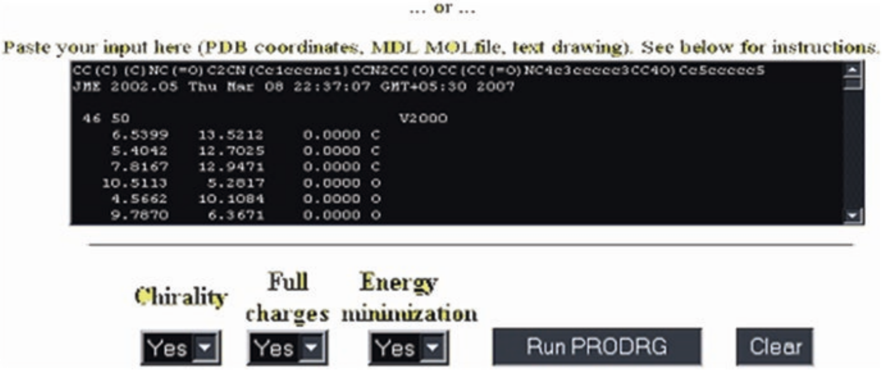

http://davapc1.bioch.dundee.ac.uk - JME Molecular Editor - Microsoft Internet Explo...
CLR
DEL
D-R
+/-
UDO
JME
C
N
O
S
F
Cl
Br
I
X
O
NH
N
N
N
OH
O
NH
OH
Transfer To PRODRG Window
Close
Help
JME Editor courtesy of Peter Ertl, Novartis
Applet JME started
Internet

Fig. 15.4 Java run-time view from PRODRG

3. Click on Draw Molecule with JME (before that you need to register by giving an email address). A new small window will open with white background. This window called as drawing window needs a Java run-time environment. Before drawing any molecule, you should install this Java software. You can install the Java run-time software provided in the workshop CD. Just double-click on the setup icon and follow the instructions (Fig. 15.4).



... or ...

Paste your input here (PDB coordinates, MDL MOLfile, text drawing). See below for instructions.

```
CC(C)(C)NC(=O)C2CN(Cc1cccnc1)CCN2CC(O)CC(CC(=O)NC4c3ccccc3CC4O)Cc5ccccc5
JME 2002.05 Thu Mar 08 22:37:07 GMT+05:30 2007

 46 50                                 V2000
    6.5399    13.5212     0.0000 C
    5.4042    12.7025     0.0000 C
    7.8167    12.9471     0.0000 C
   10.5113     5.2817     0.0000 O
    4.5662    10.1084     0.0000 O
    9.7870     6.3671     0.0000 O
```

Chirality: Yes
Full charges: Yes
Energy minimization: Yes

Fig. 15.5 PRODRG home page window and coordinate pasting window

4. A small window appears with basic structures of chemicals. Now you have to draw a ligand molecule as shown below in JME. Select suitable molecules and add one by one to the basic lead molecule. If you need N, O, and P, get it from the left bar containing atoms. Put the suitable atoms with minimum knowledge in chemistry and chemoinformatics. If you do any mistake, click DEL or CLR. Take care while using these buttons. With right-click you can rotate the molecule and left-click to move the molecule in all four sides.

Now click the button transfer to PRODRG. You can observe the coordinates in the main window. Select chirality, full charges, and energy minimization as yes as shown below. Now click on Run PRODRG. Wait for 2 min to get 3D coordinate files for AutoDock 3.0 (Fig. 15.5).

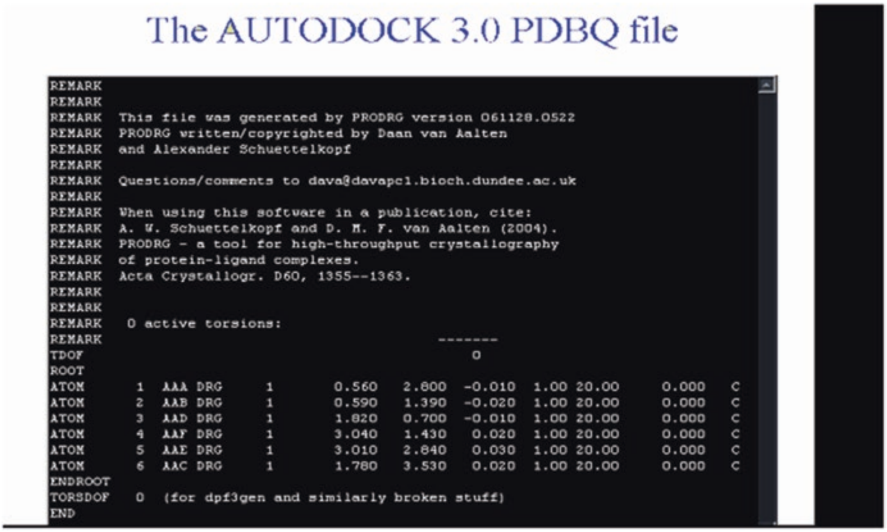
Copy the total matter into the new text document and save as “drg.pdbq” onto the desktop. This file will be used for docking purpose.

You can prepare such type molecules by just altering functional groups; halogens and alkyl groups with minimum knowledge in chemistry will give very good lead molecules (Fig. 15.6).

The drg.pdbq file can be used as lead compound to dock into the protein molecules. We will use this molecule in the next session for protein and drug interaction studies.

Overview of the method in flow chart

Open **www.google.co.in**>>type.**PRODRG SERVER**> > Select **prodrg**> > **Click Draw molecule JME**> > *Draw the molecule*> > **Transfer to Prodrg**> > **chirality** > YES> > **Full charges** > YES> > **energy minimization** > YES> > **Run PRODRG**> > *select***AUTODOCK3.0 PDBQ FILE**> > *copy the content*> > *open text document*> > *paste the content into the text document*> > *save as* “**drg.pdbq**”on desktop> > close the new text document



The AUTODOCK 3.0 PDBQ file

```
REMARK
REMARK
REMARK  This file was generated by PRODRG version 061128.0522
REMARK  PRODRG written/copyrighted by Daan van Aalten
REMARK  and Alexander Schuettelkopf
REMARK
REMARK  Questions/comments to dava@davapc1.bioch.dundee.ac.uk
REMARK
REMARK  When using this software in a publication, cite:
REMARK  A. W. Schuettelkopf and D. M. F. van Aalten (2004).
REMARK  PRODRG - a tool for high-throughput crystallography
REMARK  of protein-ligand complexes.
REMARK  Acta Crystallogr. D60, 1355--1363.
REMARK
REMARK
REMARK   0 active torsions:
REMARK                                         -------
TDOF                                                 0
ROOT
ATOM      1  AAA DRG     1       0.560   2.800  -0.010  1.00 20.00     0.000   C
ATOM      2  AAB DRG     1       0.590   1.390  -0.020  1.00 20.00     0.000   C
ATOM      3  AAD DRG     1       1.820   0.700  -0.010  1.00 20.00     0.000   C
ATOM      4  AAF DRG     1       3.040   1.430   0.020  1.00 20.00     0.000   C
ATOM      5  AAE DRG     1       3.010   2.840   0.030  1.00 20.00     0.000   C
ATOM      6  AAC DRG     1       1.780   3.530   0.020  1.00 20.00     0.000   C
ENDROOT
TORSDOF   0  (for dpf3gen and similarly broken stuff)
END
```

Fig. 15.6 Output file of PRODRG in AutoDock format

Preparation of Receptor Files for AutoDock Tools

AutoDock tools which you are using in the present workshop are having academic license only. It cannot be distributed by any institute without permission from software owners. This software is used for validation and investigation of interaction between the ligand and receptor. The ligand may be any lead compound or drug molecule in the other receptor with protein or enzyme. You have already prepared the ligand molecule with *PRODRG* server suitable for AutoDock tool 3.0 pdbq.

Now this is the time to prepare the receptor file for docking purpose, but there is a need to convert the receptor file from *protein data bank format to pdbqs format*. First you have to do all the following steps as shown below. You should be very careful in doing each step and do not miss any step. If you miss any step, you may not get docking and again you have to start from the beginning.

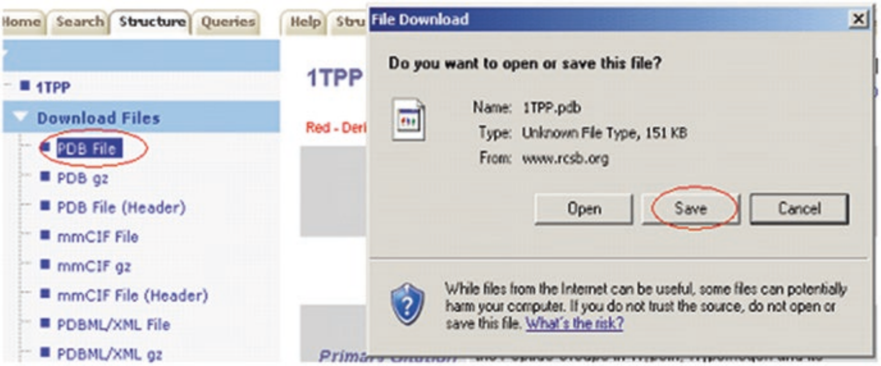
Download the receptor as follows:

Open www.google.co.in>>type rcsb> > select protein data bank> > type 1TPP in search box> > click search> > wait> > click download files at left side> > click on Pdb file> > save > > on desktop (Fig. 15.7).

The file will be saved onto the desktop with the name 1TPP in the pdb format. This file can be viewed with WordPad and delete so4, HOH, and Ca molecules.

Copy the APA and connect series and paste into new text document and save as apa. Now save the modified pdb file with trp.pdb. This file contains only amino acids but not any heteroatoms.

As we cut the heteroatoms, the missing C-terminal oxygen atom issue can be rectified by spdbv software which has very good option to do this.

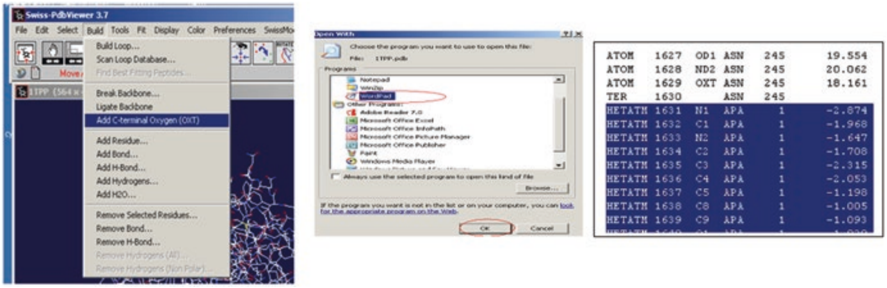


Do you want to open or save this file?

Name: 1TPP.pdb
Type: Unknown File Type, 151 KB
From: www.rcsb.org

While files from the Internet can be useful, some files can potentially harm your computer. If you do not trust the source, do not open or save this file. What's the risk?

Fig. 15.7 Retriving receptor file from RCSB database



Choose the program you want to use to open this file:
File: 1TPP.pdb

☐ Always use the selected program to open this kind of file

If the program you want is not in the list or on your computer, you can look for the appropriate program on the Web.

```
ATOM    1627  OD1 ASN   245         19.554
ATOM    1628  ND2 ASN   245         20.062
ATOM    1629  OXT ASN   245         18.161
TER     1630      ASN   245
HETATM  1631  N1  APA     1         -2.874
HETATM  1632  C1  APA     1         -1.968
HETATM  1633  N2  APA     1         -1.647
HETATM  1634  C2  APA     1         -1.708
HETATM  1635  C3  APA     1         -2.315
HETATM  1636  C4  APA     1         -2.053
HETATM  1637  C5  APA     1         -1.198
HETATM  1638  C8  APA     1         -1.005
HETATM  1639  C9  APA     1         -1.093
```

Fig. 15.8 Editing PDB file using SWISS-MODEL software

Double-click on spdbv folder> > double-click spdbv icon> > new window will appear> > file> > open pdb file> > select all> > build>add C-terminal oxygen (OXT).> > file > > save> > layer> > save as trp.Pdb. Trp file is completely modified without ligands and water molecules. This molecule is read to input into AutoDock software (Fig. 15.8)

Copy the complete apa file and paste in the PRODRG server for the generation of pdbq file for AutoDock.

Select yes option for each one like chirality, full charges, energy minimization> > run prodrg> > click enter key.

Now the lead molecule or drug molecule is ready for dock for protein is required for this purpose.
Open AutoDock tools from the desktop, and you can visualize as follows:

Convert trp.pdb file to pdbqs file

Select the trp.pdb file from the desktop and load the repair commands for histidine use.

File> > load module> > repair commands> > load> > dismiss
Edit> Hydrogens>Add> Select "Ploar Only"
Fix the histidine residues
Edit>Hydrogens>Edit Histidine Hydrogens, and change the selection from +0 to 0, HD1, and then click apply following by dismiss

Add the partial atomic charges
Edit> charges> add kollman charges
Edit> charges> check total on residues – and fix using spread charge deficit

Here choose…(AG3) means you are selecting for AutoDock tools 3.0 version. So every time choose the ligand and receptor for AG3 or AD3 only.
Automatically solvent parameters are added to receptor and charges; the molecule is ready to be saved into trp.pdbqs format.
Grid> > set map types> > choose ligand (AG3)> > select ligand apa and accept the atom types.
Grid> > Grid box> > *enter the X,Y and Z axis values***-1.573,14.473,19.036,** respectively, and leave all the values as default > > **file> > close saving current**.
Now save the grid file in GPF format, **Grid> > Output> > save GPF (AG3)**

Making the docking parameter file (dpf)

Docking> > Macromolecule> > set File name..(AD3)> > select trp.pdbqs file

Docking> > ligand> > choose….(AD3)

Accept all the default values and continue as procedure below.

Docking> > output> > Genetic Algorithm..(AD3)> > save the file trp.dpf

Now this is the time to run the docking files from the command line

Close all the windows, and click left corner windows start button> > run> > type cmd> > ok
Type as follows:
Cd\
Cd autodock
Now you can run AutoGrid program (AD3)

<autodock>autogrid3 –p trp.gpf –l trp.glg

Wait for 5 min to complete the program, and you can fine message indication that program is successfully completed (Fig. 15.9).

Now run again on command line for AutoDock program (AD3)
<autodock>autodock3 –p trp.dpf –l trp.dlg
Wait for 30 min to complete the program.

Then type exit and close all the windows.

To analyze the results, click on AutoDock tools, and follow the steps

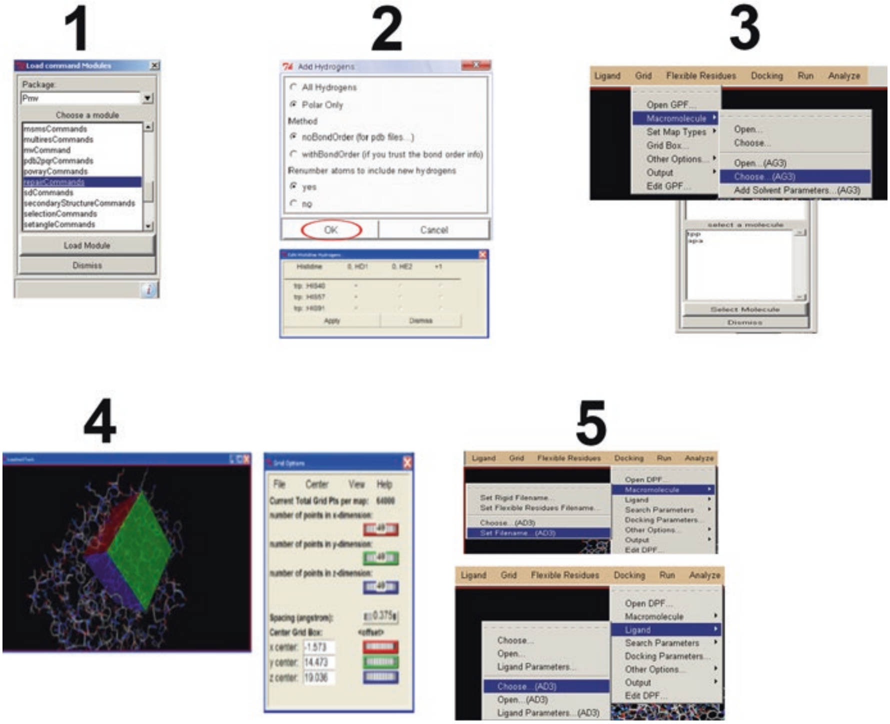


Fig. 15.9 Steps in preparation of receptor and ligand files for docking

File> > read molecule> > trp.pdbqs

Analyze> > docking> > open> > select trp.dlg file.

Color > > by atom type

Analyze> > conformations> > play

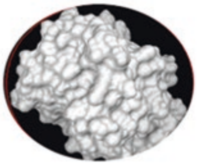
You can see the docking hits with amino acids within the protein molecule.

Analyze> > docking> > open> > ligand molecule will appear on the screen

Analyze> > macromolecule> > open> > select tpp.pdbqs or open from AutoDock folder.

To view molecular surface of the macromolecule, just click on MS option as shown in the above figure. You can also remove lines by clicking on Lines once. Now rotate the molecule with the right mouse. To rotate the molecule, hold the right-click, and move the mouse; if you want to zoom, just hold on the shift key, and right-click move. Select apa as S&B >> color by atom type.

Receptor can be colored by selecting tpp, and color > > color by molecule> > give tick on MSMS-mol> > color molecule will appear.



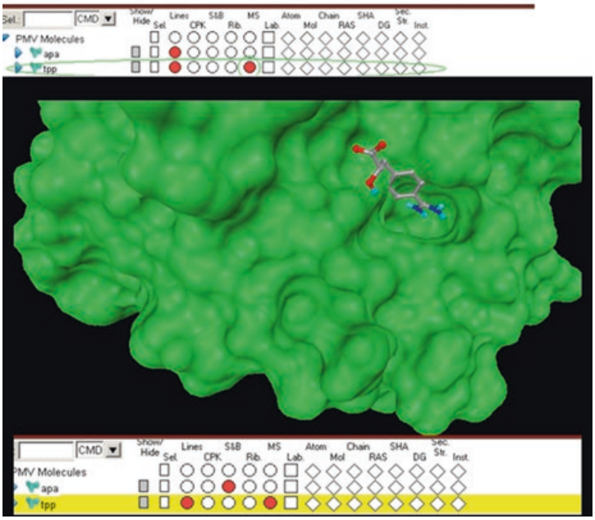


Fig. 15.10 Molecular visulation of receptor and ligand in MGL tools

Conformations > > play> > just click on & icon and give all ticks at a time, and observe the docking energies and conformation. At this stage you can see animation of ligand and protein interaction of all calculated values (Fig. 15.10).

15.6 Conclusion

Docking is a molecular modeling approach which examines how a protein interacts with small molecules. The molecular interaction potential of between protein or nucleic acids with small molecules forms a supramolecular complex whose stability could turn influence their biological functions. This docking method has potential uses and applications in different phases of drug discovery like in conducting structure-activity studies, in discovering potential leads and their optimization and also in exploring binding affinities between query molecules. In the current chapter, we have described the basic principle and fundamental steps involved in molecular docking. We have also provided an overview of some commonly used computational programs in molecular docking and structural analysis. Although molecular docking can predict the potential interactions between small molecules and receptors, the elucidation of accurate interactions can only be confirmed by doing laboratory based experiments.

References

Adeniyi AA, Ajibade PA (2013) Comparing the suitability of autodock, gold and glide for the docking and predicting the possible targets of Ru(II)-based complexes as anticancer agents. Molecules 18(4):3760–3778

Banaganapalli B, Mulakayala C, D G, Mulakayala N, Pulaganti M, Shaik NA, Cm A, Rao RM, Al-Aama JY, Chitta SK (2013a) Synthesis and biological activity of new resveratrol derivative and molecular docking: dynamics studies on NFkB. Appl Biochem Biotechnol 171(7):1639–1657

Banaganapalli B, Mulakayala C, Pulaganti M, Mulakayala N, Anuradha CM, Suresh Kumar C, Shaik NA, Yousuf Al-Aama J, Gudla D (2013b) Experimental and computational studies on newly synthesized resveratrol derivative: a new method for cancer chemoprevention and therapeutics? OMICS 17(11):568–583

Bohm HJ (1992) The computer program LUDI: a new method for the de novo design of enzyme inhibitors. J Comput Aided Mol Des 6(1):61–78

Brooks BR, Brooks CL 3rd, Mackerell AD Jr, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner AR, Feig M, Fischer S, Gao J, Hodoscek M, Im W, Kuczera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor RW, Post CB, Pu JZ, Schaefer M, Tidor B, Venable RM, Woodcock HL, Wu X, Yang W, York DM, Karplus M (2009) CHARMM: the biomolecular simulation program. J Comput Chem 30(10):1545–1614

El-Hachem N, Haibe-Kains B, Khalil A, Kobeissy FH, Nemer G (2017) AutoDock and AutoDockTools for protein-ligand docking: Beta-site amyloid precursor protein cleaving enzyme 1(BACE1) as a case study. Methods Mol Biol 1598:391–403

Ewing SA, Dawson JE, Panciera RJ, Mathew JS, Pratt KW, Katavolos P, Telford SR 3rd (1997) Dogs infected with a human granulocytotropic Ehrlichia spp. (Rickettsiales: Ehrlichieae). J Med Entomol 34(6):710–718

Filikov AV, Mohan V, Vickers TA, Griffey RH, Cook PD, Abagyan RA, James TL (2000) Identification of ligands for RNA targets via structure-based virtual screening: HIV-1 TAR. J Comput Aided Mol Des 14(6):593–610

Hindle SA, Rarey M, Buning C, Lengaue T (2002) Flexible docking under pharmacophore type constraints. J Comput Aided Mol Des 16(2):129–149

James TL, Lind KE, Filikov AV, Mujeeb A (2000) Three-dimensional RNA structure-based drug discovery. J Biomol Struct Dyn 17(Suppl 1):201–205

Jiang S, Huang K, Liu W, Fu F, Xu J (2015) Combined autodock and comparative molecular field analysis study on predicting 5-lipoxygenase inhibitory activity of flavonoids isolated from Spatholobus suberectus Dunn. Z Naturforsch C 70(3–4):103–113

Judson RS (1996) Genetic algorithms and their uses in chemistry. In: Boyd DB, Lipkowitz K (eds) Reviews in computational chemistry, vol 10. Wiley, New York, pp 1–73

Kramer B, Rarey M, Lengauer T (1997) CASP2 experiences with docking flexible ligands using FlexX. Proteins (1):221–225

Kramer B, Rarey M, Lengauer T (1999) Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking. Proteins 37(2):228–241

Kuntz ID (1992) Structure-based strategies for drug design and discovery. Science 257(5073):1078–1082

Morris GM, Goodsell DS, Huey R, Olson AJ (1996) Distributed automated docking of flexible ligands to proteins: parallel applications of AutoDock 2.4. J Comput Aided Mol Des 10(4):293–304

Morris GM, Huey R, Olson AJ (2008) Using AutoDock for ligand-receptor docking, Curr Protoc bioinformatics Chapter 8. Unitas 8:14

Schellhammer I, Rarey M (2004) FlexX-Scan: fast, structure-based virtual screening. Proteins 57(3):504–517

Spitzer R, Jain AN (2012) Surflex-Dock: Docking benchmarks and real-world application. J Comput Aided Mol Des 26(6):687–699

Taylor RD, Jewsbury PJ, Essex JW (2002) A review of protein-small molecule docking methods. J Comput Aided Mol Des 16(3):151–166

Vicente J, Chicote MT, Guerrero R, Jones PG, Ramirez De Arellano MC (1997) Gold(I) Complexes with N-Donor Ligands. 2.[1] Reactions of Ammonium Salts with $[Au(acac\text{-}\kappa C^2)(PR_3)]$ To Give $[Au(NH_3)L]^+$, $[(AuL)_2(\mu_2\text{-}NH_2)]^+$, $[(AuL)_4(\mu_4\text{-}N)]^+$, or $[(AuL)_3(\mu_3\text{-}O)]^+$. A New and Facile Synthesis of $[Au(NH_3)_2]^+$ Salts. Crystal Structure of $[\{AuP(C_6H_4OMe\text{-}4)_3\}_3(\mu_3\text{-}O)]CF_3SO_3$. Inorg Chem 36(20):4438–4443

Westhead DR, Clark DE, Murray CW (1997) A comparison of heuristic search algorithms for molecular docking. J Comput Aided Mol Des 11(3):209–228

Chapter 16
In Silico PCR

Babajan Banaganapalli, Noor Ahmad Shaik, Omran M. Rashidi, Bassam Jamalalail, Rawabi Bahattab, Hifaa A. Bokhari, Faten Alqahtani, Mohammed Kaleemuddin, Jumana Y. Al-Aama, and Ramu Elango

Contents

16.1 Introduction

PCR is the abbreviation of polymerase chain reaction. It is the most scientific evolutionary technique in the molecular biology. PCR can copy a segment of the DNA into millions of copies (amplicons) in a short period. Amplification process can be carried out via thermocycler, a device that facilitates multiple rounds of temperature-sensitive cycles. In brief, PCR is a thermal technique that requires the complementary sequence of oligonucleotides, desired DNA targeted segment to initiate the process of polymerization by the steady actions of polymerase enzyme, which

B. Banaganapalli · J. Y. Al-Aama · R. Elango (✉)
Princess Al-Jawhara Center of Excellence in Research of Hereditary Disorders,
Department of Genetic Medicine, Faculty of Medicine, King Abdulaziz University,
Jeddah, Saudi Arabia
e-mail: bbabajan@kau.edu.sa; relango@kau.edu.sa

N. A. Shaik · B. Jamalalail · R. Bahattab · H. A. Bokhari · F. Alqahtani · M. Kaleemuddin
Department of Genetic Medicine, Faculty of Medicine, King Abdulaziz University,
Jeddah, Saudi Arabia
e-mail: nshaik@kau.edu.sa

O. M. Rashidi
Princess Al-Jawhara Center of Excellence in Research of Hereditary Disorders,
King Abdulaziz University, Jeddah, Saudi Arabia

N. A. Shaik et al. (eds.), *Essentials of Bioinformatics, Volume I*,
https://doi.org/10.1007/978-3-030-02634-9_16

incorporates the new dNTPs to the template DNA strand, and synthesizes multiple copies of new double-stranded DNAs.

Back in 1983, PCR was originally inspired by Kary Mullis, a Nobel Prize-winning American biochemist who initially started to use fresh polymerase enzyme extracted from an *E. coli* bacteria. Heat cycles would degrade the integrity of the enzyme. At the time, the only plausible solution was to continuously add the polymerase enzyme in each cycle of the reaction to assure the success of the reaction.

Mullis and his team initially presented the idea at the American Society for Human Genetics annual meeting in 1985. Previously available PCR techniques would have taken a week to complete the amplification of the targeted sequence (Saiki et al. 1985). Mullis successfully experimented his technique to amplify the codon 6 of beta globin sequence in less than 1 day. Following that year, in 1986, Henry Erlich had replaced the *E. coli* polymerase with a more heat-stable Taq polymerase from *Thermus aquaticus*, a bacterium that lives in hot springs and can tolerate the heat without any damages to its enzymes (Saiki et al. 1988). Finally, 1987 marks the year when the thermocycler machine was developed and adjusted to be an instrument of a programed heating mechanism. In 1993 Kary Mullis received the Nobel Prize for his great invention of the PCR which would forever reflect a huge impact on the genomics and biotechnology.

Several primary components are necessary for the methodology of the PCR to have successful results, and they are as follows:

(a) Extracted and purified DNA template harboring a target sequence to amplify.
(b) DNA polymerase enzyme, to assemble the DNA strands that are complementary to the targeted sequence.
(c) DNA primers: two separate primer oligonucleotides used (1) in the forward 5′ prime of the DNA strand, more specifically to be annealed to the sense strand and (2) reverse primer that acts primarily on the anti-sense strand of the DNA. Their presence is essential to anchor the polymerase enzyme and guide its action to start and to be the specific targeted region of the DNA.
(d) Deoxynucleotide triphosphates (dNTPs), free single base pairs (A, G, C, T). They serve as building blocks for the newly synthesized strands.
(e) Reaction buffers and thermocycler (Fig. 16.1).

PCR has become a cornerstone to multiple foundations of analytical, diagnostic, and many interdisciplinary sciences of molecular biology and molecular genetics to be precise. It is used in cytogenetic analysis, diagnosis of infections, paternity tests, step in sequencing, and forensics, etc. (Table 16.1).

16.2 Methods of PCR

In typical PCR there are three main steps to the process, repeated 25–30 times to have millions of DNA template copies (Fig. 16.2).

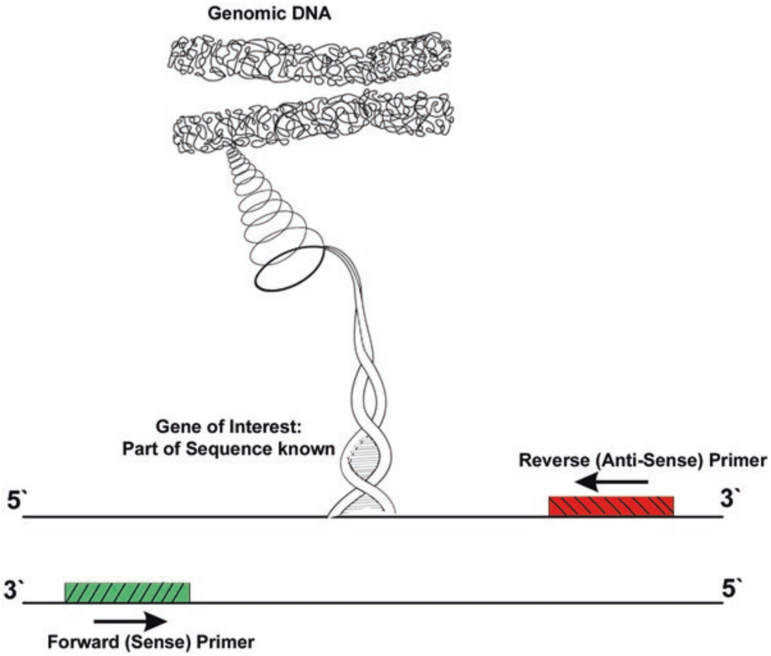


Fig. 16.1 Primer annealing to specific known region of genomic DNA

Step (1): Denaturation

This is the first step where a double-stranded DNA is unzipped into single-stranded DNA by breaking the hydrogen bonds that link the two strands of the DNA. A temperature range of 90 –95 °C is the ideal one to separate the two DNA strands. Usually it starts with initial denaturation step for 3 min and then another 30 s to complete the denaturation, but the timing could be different from one reaction to another, depending on the sequence content. For example, GC-rich sequence takes more time than AT-rich sequence, due to the triple and double bonds between the base pairs, respectively.

Step (2): Annealing

Well-designed primers will bind to the complementary targeted single-stranded DNA after the denaturation step with free 3′ end to serve as starting point for the synthesis to take place. 55–70 °C is the temperature range for the primers to anneal, usually for 30 s. Forward and reverse primer annealing are the most important steps to have an accurate result which is a challenge for the scientists to design the best primers for their experiments.

Step (3): Extension (Elongation)

Extension is the final step to have the amplicons ready. The temperature will be raised up to 70–75 °C, and the polymerase enzyme will start to function by adding the dNTPs to the 3′ end of the primer and making new copies from the

Table 16.1 Types and application of different PCR

Types of PCR	Application
Multiplex PCR	Multiple PCR reactions in one single tube. Key applications include identification of pathogens such as lower respiratory tract infections, detection of RNA, and mutation analysis in the area of genetics. Also used in the process such as mutation detection and gene detection
Nested PCR	Refers to a modified PCR that helps in the reduction of non-specified binding products based on the amplification of different and unexpected binding sites. The two sets of primers in this process are crucial toward reducing the rate of contamination from unwanted products such as primer target sequences and dimers
Inverse PCR	Inverse PCR assay is a useful tool in the analysis of rare structures associated with mutational integrons found in areas of class 1 integrons
Reverse interpretation PCR	Applies a process of grouping two DNA strands to promote the polymerization and duplication of the strand. Mainly used in the area of genetics for cloning research and development.
Reverse transcription PCR (RR-PCR)	Majorly applied in the area of molecular biology in the quantification and detection of RNA expression from a single cell
Real-time PCR (qPCR)	Used to investigate the expression of samples in order to quantitate changes in gene expression. Therefore it is used in basic research and diagnostics such as quantifying different forms of gene expression, diagnosis of gene abnormalities, cancer diagnosis, and detection of infectious diseases including new forms of flu
Arbitrary primer (AP-PCR)	Used in the process of fingerprinting genomes to detect variations in the human DNA and identify various types of bacterial strains present in similar species such as the process of *Streptococcus mutans* genotyping. *Streptococcus mutans* refers to a bacterium present in the oral cavity that contributes to tooth decay (Tabchoury et al. 2008)
Allele-specific PCR	This procedure is important for the determination of various types of single-nucleotide polymorphisms for HLA typing, model and non-model organizations, and paternity testing (Gaudet et al. 2009)
Assembly PCR	Used in the combination of large forms of DNA oligonucleotides and applied in genetics as a mechanism for the amplification of DNA sequences and development of novel synthetic genes
Degenerate PCR	Mainly applied in the area of genetics for the process of amplification and matching of multiple genes from related families. Used as major tool in gene cloning
Dial-out PCR	Used in biological processes to accurately identify and retrieve DNA molecules in all processes involving gene synthesis.
Traditional PCR	Carries out the amplification of various forms of nucleic acid for gene sequencing and cloning by estimating the quantity of PCR product on completion of several PCR cycles
Digital PCR	More advanced than the traditional PCR as it produces absolute minute amounts of nucleic acid where it is applied in the process of rare sequence detection and analysis of rare gene expression

(continued)

Table 16.1 (continued)

Types of PCR	Application
Inter-simple sequence PCR	The process of obtaining multilocus fingerprinting profiles is applied in studies on genetic identity and the process of quantifying issues of genome instability such as deletions and amplifications in cases of human sporadic tumors
Hot start PCR	Improves the process of DNA analysis through the use of polymerase inhibitors in lower temperatures to inactivate the DNA polymerase to improve specificity for target genes where it is mainly applied in long-distance PCR experiments
In silico PCR	Helps in the identification of newly designed primers and efficiently supports the development of primer specificity for multi-exon genes for practical investigation procedures in molecular diagnosis and forensic DNA typing
Suicide PCR	Helps in the process of molecular identification of causative organisms (agents) of infections and diseases through assessment of specific biotypes such as *Yersinia pestis* based on the analysis of intergenic spacer DNA
Late PCR	Utilizes primer pairs intentionally fashioned for use at varied concentrations to generate single-stranded DNA. It resolves the problems associated with conventional PCR primers: optimization difficulties, inefficiencies, and promotion of non-specific amplification
Long-range PCR	Helps to amplify DNA lengths that routine reagents or PCR methods cannot typically amplify. For the simple templates of DNA, polymerase that is optimized for long PCR can amplify equal to 30 kb. For genomic templates that are complex, the typical target is normally 20 kb
In situ PCR	Helps to detect minute amounts of single-copy or rare nucleic acid sequences in paraffin-embedded or frozen tissue sections or cells for the localization of the sequences within the cells. This PCR approach's principle comprises tissue fixing in a bid to preserve the cell morphology and consequent treatment with proteolytic digestion to give the PCR reagents access to the target DNA
Colony PCR	Helps to determine the absence or presence of insert DNA in plasmid constructs. It also assists in determining insert orientation

template (dNTPs are dATPs, dTTPs, dGTPs, and dCTPs). Extra 5-min final extension time can be added to make sure that the polymerase has done its job properly.

Tips in Using PCR

- Primer length: the accuracy and specificity in having good results depend on the primer oligonucleotide length. An 18–24 base pair primers appeared to be the ideal length (Dieffenbach et al. 1993). Many stability complications in the reaction could result from primers shorter or greater than this length range.
- G–C content: it must follow the range of 40–60%.
- 1 ng–1μ of the genomic DNA template is used in the reaction or 1 pg–1 ng in the plasmid or viral template (Rychlik et al. 1990). Higher concentration of DNA template decreases the specificity of the amplicon.

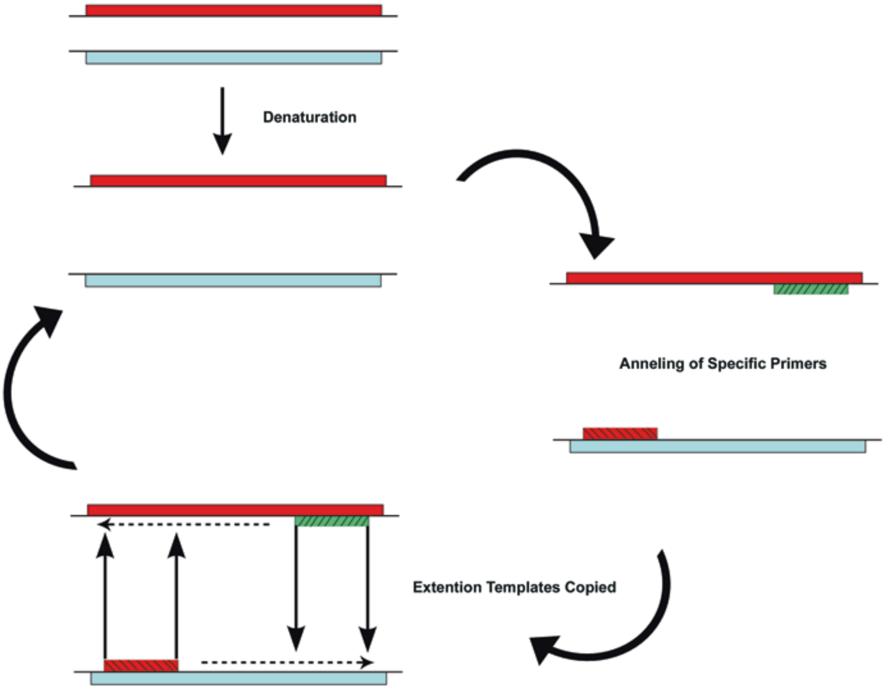


Fig. 16.2 Steps of PCR DNA amplification: denaturation, primer annealing, and synthesis of complementary strands

16.3 Guidelines for Primer Design for PCR

DNA amplification, the process of generating multiple copies of a DNA sequence, can be done by polymerase chain reaction (PCR). The effectiveness of PCR is largely dependent on how efficient the designed primers are (Abd-Elsalam 2003; He et al. 1994). There are multiple factors like the primer-template association kinetics, the stability of the duplex formed between the primer and template in the formation of mismatched nucleotides, and the ability of the polymerase to identify and repair mismatched duplex (Abd-Elsalam 2003; Dieffenbach et al. 1993) that are known to have influenced the oligonucleotides to act as ideal PCR primers. For a primer to be able to amplify a specific target sequence, it must possess specific characteristics, and these include the length of the primer, the percent of guanine and cytosine nucleotides, the five-prime end stability, and the three-prime end specificity (Abd-Elsalam 2003; Dieffenbach et al. 1993). If the PCR primer was poorly designed, this would result in an unsuccessful PCR reaction because the primer didn't work properly; it may result in little or no product formation due to multiple errors that resulted in dimer formation or non-specific sequence amplification (Abd-Elsalam 2003; Dieffenbach et al. 1993). The designed primer plays a major role in the PCR product because it determines its length, melting temperature, and yield

(Abd-Elsalam 2003; Dieffenbach et al. 1993). For a successful PCR reaction, a well-designed primer is critical to ensure high-yield product; following the below simple guidelines, one can ensure to design a good primer.

Primer length The primer length refers to how many nucleotides the primer sequence has. Primer length is a critical parameter for a successful PCR reaction; the specificity, temperature, and time of annealing depend on the length of the primer (Abd-Elsalam 2003; Wu et al. 1991). A typical preferred length of a primer is between 18 and 30 nucleotides. The minimum number of nucleotides for the primer should be 18 so that problems such as secondary hybridization sites can be avoided. Also, it is important to ensure that the primer doesn't contain multiple stretches of a single base; one must avoid four or more guanine and cytosine in a row (Abd-Elsalam 2003).

Melting temperature (T_m) The melting temperature refers to the temperature required to dissociate the DNA duplex to a single strand. For optimal results, it is best that the melting temperature for the primers is to be between 52 and 58 °C. Designing primers that have a melting temperature of 65 °C and above should be avoided because it might initiate a secondary annealing reaction (Abd-Elsalam 2003). Wallace et al. (1979) proposed a formula to calculate the melting temperature of primers based on oligonucleotides between 18 and 30 bases: $T_m = 2(A + T) + 4(G + C)$.

GC content The percentage of guanine and cytosine nucleotides in a primer is an essential component of the primer as it is associated with the annealing strength. Based on various observations, it is best to have a GC primer content between 45% and 60% (Abd-Elsalam 2003; Dieffenbach et al. 1993). If primers had GC content below 50, it is advisable that the primer sequence should be extended above 18 bases, so that melting temperature and annealing temperature requirements are fulfilled (Abd-Elsalam 2003; Rychlik and Rhoads 1989).

3′-end sequence The role of the 3′-terminal position of PCR primers is in controlling mis-priming (Abd-Elsalam 2003; Kwok et al. 1990). When it comes to sticky ends, the 5′-end of the primer should be stickier than the 3′-end. If 3′-ends were sticky with multiple GC contents, this can lead to multiple sites annealing on the template sequence. However, it is advisable that the 3′-end has guanine or cytosine, but one must take into account the sticky-end rule (Abd-Elsalam 2003; Sheffield et al. 1989).

Dimers and false priming cause misleading results When it comes to designing primers, one must be careful of complementary sequences within the primer. The designed primers must not have complementary sequences that result in the primers forming hairpins by folding back on itself. The folding will cause the primer not to work properly, and it will affect the overall PCR reaction (Abd-Elsalam 2003; Breslauer et al. 1986). However, if hairpins were formed below 50 °C, these hairpins can be ignored because they will not cause that much of a problem. As a general guideline, there shouldn't be any sequences that will cause the primer to anneal to itself or other primers in the PCR reaction; this will cause what is known as primer dimer (Abd-Elsalam 2003).

Specificity It is very important to consider specificity when designing a primer because choosing random bases will only result in unsuccessful results when amplified. The primer must contain a specific sequence that will be targeted on the DNA sequence being amplified; designing a primer with multiple repetitive sequences will show a smear of amplified DNA in the results (Abd-Elsalam 2003).

Degenerate primers Degenerate primers are defined as a combination of primers that have multiple substitutions of different bases; they are similar, but not the same. The role of degenerate primers lies in the process of amplifying a sequence that presents different protein sequences. Degenerate primers must be taken into account when designing primers to ensure that the final protein sequences are not altered (Abd-Elsalam 2003). One can use different computer programming to design specific degenerate primers (Abd-Elsalam 2003; Chen and Zhu 1997).

Complementary primer sequences When designing primers, the sequence should not contain sequence homology within itself, which is known as intra-primer homology; designing primers with sequence homology will lead to the occurrence of snapback. Another problem that may occur with the complementary sequence is homology sequence content within the primer sequence itself, occurring in the middle regions of the two primers, which can interfere with hybridization. Primer dimer formation will occur if sequence homology occurred at 3′-end of the primers (Abd-Elsalam 2003).

Other recommendations When it comes to amplification, the primer concentration should be in the range of 0.1–0.5 μm. Before starting the PCR, and after designing a primer, one can use computer analysis software to analyze the designed primer to ensure that the primer will work properly by evaluating the mentioned guidelines (Abd-Elsalam 2003).

16.4 In Silico Designing PCR Primers Using Bioinformatics Tools

Primer-BLAST tool is being used widely for designing optimized primers for a target DNA sequence region. The algorithms of the tool were originally developed at NCBI to assist users in their quest to easily and efficiently design primers that are specific and intended for target region. In principle, Primer-BLAST engages Primer3 software to design PCR primers and then turns to BLAST to execute a global alignment algorithm to screen primers against user-selected database in order to avoid primer pairs (all combinations including forward-reverse primer pair, forward-forward, as well as reverse-reverse pairs) that can cause non-specific amplifications. Currently there are many online tools available to the design and validation of primer (Table 16.2). In this chapter as an example, we will design a primer for the gene *IL10* exon number 5 using NCBI primer-BLAST tool.

Table 16.2 Currently available online tools in design validation of primer sequence

Primer name	Description	Site
CODEHOP	COnsensus-DEgenerate Hybrid Oligonucleotide Primer; design degenerate PCR primer. Will accept unaligned sequences	http://blocks.fhcrc.org/codehop.html
Gene Fisher	A primer design tool for normal or degenerate primers. Will accept unaligned sequences	http://bibiserv.techfak.uni-bielefeld.de/genefisher/
Primer3	Inclusive hybridization probe and PCR primer design tool	http://www.justbio.com/primer/index.php
Web Primer	Design primers for both PCR and sequencing purposes	https://www.yeastgenome.org/cgi-bin/web-primer?name=YML058W
PCR Designer	For restriction analysis of sequence mutations	http://cedar.genetics.soton.ac.uk/public_html/primer.html
Primo Pro 3.4	Decreasing the possibility of random primering which results in reduced PCR noise	http://www.changbioscience.com/primo/primo.html
FAS-DPD	A package to design degenerate primers for PCR	https://omictools.com/fas-dpd-tool
EPRIMER3	Picks PCR primers and hybridization oligos (EMBOSS)	http://bioinfo.nhri.org.tw/cgi-bin/emboss/eprimer3
PrimerQuest	A primer design tool	https://eu.idtdna.com/Primerquest/Home/Index
Tool name	Description	Site
MethPrimer	MethPrimer design primers for methylation PCRs	http://www.urogene.org/methprimer/
MEDUSA	A tool for the assessment of PCR primer and automatic selection	http://www.mybiosoftware.com/medusa-selection-visual-assessment-pcr-primer-pair.html
Eurofins genomics	A tool for designing PCR primer as well as primer for sequencing purposes	https://www.eurofinsgenomics.eu/en/dna-rna-oligonucleotides/oligo-tools/primer-design-tools/
Primer plus3	A new improved web interface to the common used one Primer3 primer design program	http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi
Genscript	GenScript online PCR primers designs tool	https://www.genscript.com/tools/pcr-primers-designer
Primer3:WWW primer tool	This site comprise a very controlling PCR primer design program allowing the control of the size of product desired, primer size and *Tm* range, and presence/absence of a 3’-GC clamp	http://biotools.umassmed.edu/bioapps/primer3_www.cgi
BiSearch	Primer design and search tool. This tool is useful for primer design for any DNA template and particularly for bisulfite-treated genomes	http://bisearch.enzim.hu/?m=genompsearch
Primer-BLAST	Was established at NCBI. Primer-BLAST uses Primer3 to design PCR primers	https://www.ncbi.nlm.nih.gov/tools/primer-blast/

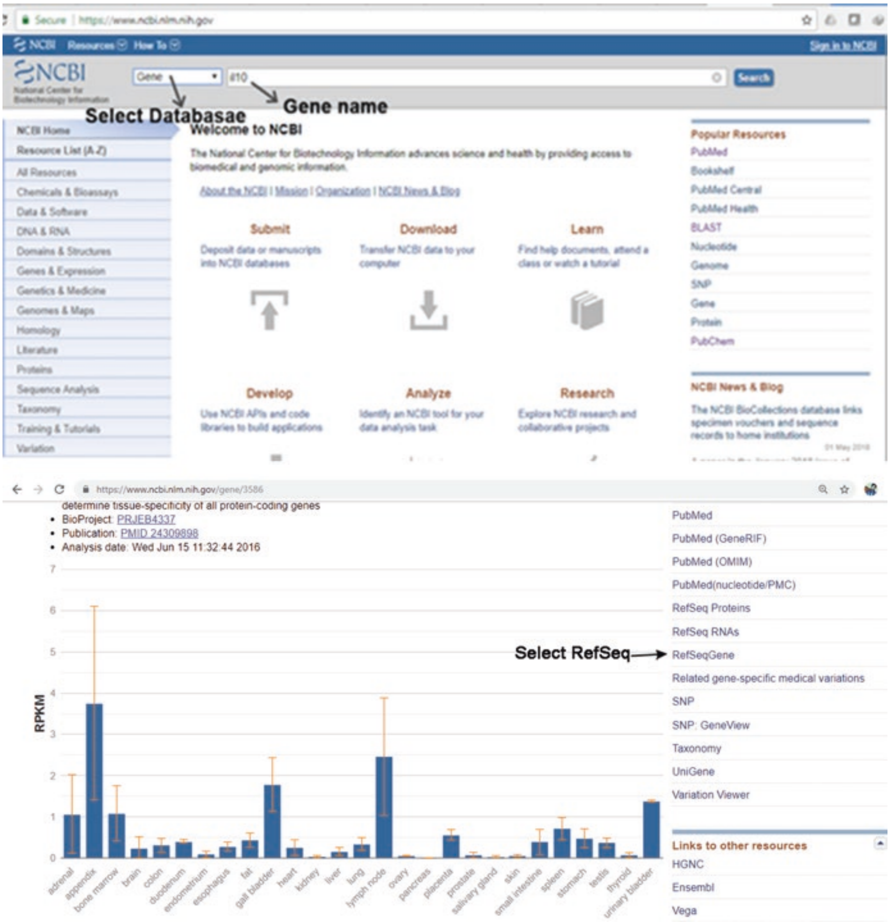
(continued)

Table 16.2 (continued)

Tool name	Description	Site
Primer design-M	Includes some options for multiple-primer design, it can design walking primers that cover long DNA targets. It can also minimize primer dimerization	https://www.hiv.lanl.gov/content/sequence/PRIMER_DESIGN/primer_design.html
Primer Designer Tool	Primer designer tool for PCR and Sanger sequencing	https://www.thermofisher.com/sa/en/home/life-science/sequencing/sanger-sequencing/pre-designed-primers-pcr-sanger-sequencing.html
MFEprimer	MFEprimer utilizes a k-mer index algorithm to accelerate the exploration process for primer binding sites	http://mfeprimer.igenetech.com
Primer4 clades	Design PCR primers for amplification of novel sequences from metagenomics DNA or from uncharacterized organisms	http://maya.ccg.unam.mx/primers4clades/index.html#0
Flexi® Vector Primer Design Tool	This design tool will design PCR primers to use it with Flexi Cloning System Vectors which will amplify a compatible coding region from the input sequence	https://worldwide.promega.com/resources/tools/flexi-vector-primer-design-tool/
primerx	Systematized design of mutagenic primers for site-directed mutagenesis	http://www.bioinformatics.org/primerx/cgi-bin/DNA_1.cgi
Quant prime	Automated tool for primer pair design for qPCR	http://quantprime.mpimp-golm.mpg.de

The first step In order to design a primer for *IL10* gene, it is initially important to import the target DNA sequence from NCBI database. To do this, go to NCBI website, and select the option “Gene” from “All Database,” and then type the gene symbol in the nearby search box which will then direct you to the main entry page of the gene of interest. “Gene” main entry would usually contain additional useful entries that link associations such as genomic context, genomic region, and transcript identification which can add a supportive base to the overall understanding of the sequence structure of the gene (Fig. 16.3).

Second step To the left of the main entry of “Gene,” under the section “Related Information,” the option “RefSeq Gene” will direct the user to an open-access, annotated, and curated form of the genomic sequence as well as the genomic context of IL10. In essence this entry will provide the sequence of IL10 exons to automatically distinguish and easily jump through any of its featured exons, cDNA, or even mRNA equivalent sequences.


Select Databasae

Gene name

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

Select RefSeq

determine tissue-specificity of all protein-coding genes
- BioProject: PRJEB4337
- Publication: PMID 24309898
- Analysis date: Wed Jun 15 11:32:44 2016

Fig. 16.3 PCR primer d s 1 to 4

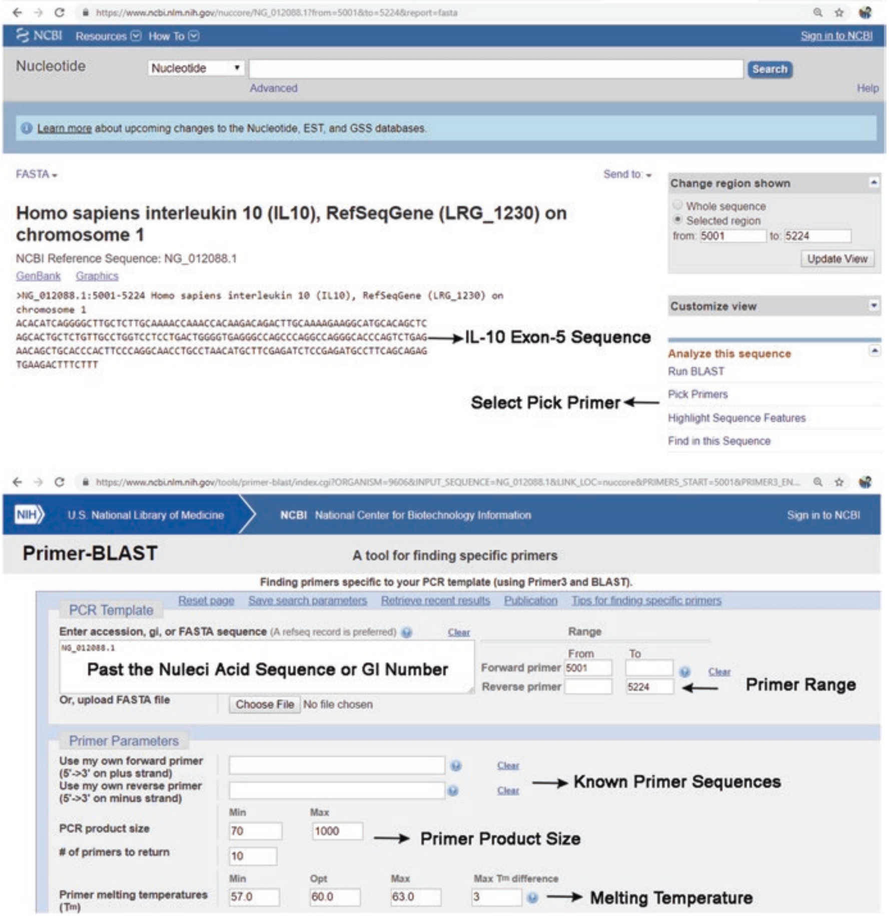
Third step Additional features can be highlighted by clicking on "Highlight Sequence Features" tab, as shown below, to the top right side of the current page. It activates the feature search bar that appears at the bottom of the display which can deliver the option "Exon" to automatically annotate the corresponding base pairs of exon 5 in the display as shown below. Using the left- or right-pointed arrows will facilitate a smooth traveling across the gene sequence of base pairs for faster reaching of the intended exon.

Fourth step By using the **"FASTA"** command at the bottom of the page, user will be given the opportunity to distinguish and display the annotated sequence of exon 5, in FASTA format, with a reference GI entry ID: NG_012088.1, in a separate window. GenInfo Identifier is a simple series of digits that are assigned consecutively to each sequence record processed by NCBI.

Fifth step In the separate window, NCBI established a default algorithm to design and test primers for this sequence using Primer-BLAST, which can be accessed through "Pick Primers" option under the tab "Analyzed Sequence" to the right of the screen (Fig. 16.4).

Sixth step Primer-BLAST is a tool to find specific primers to the desired PCR template using Primer3 and BLAST. When activated, user should paste the FASTA sequence of exon 5 DNA, or there can be the option to paste its unique gi number (NG_012088.1), and in both cases Primer-BLAST settings will learn to generate primers that are specific to exon 5.

There are of course additional features present to restrict the search of Primer-BLAST in order to limit the number of generated primer sets. There is an option to design primers that are site-specific within the target sequence, knowing the nucleo-



https://www.ncbi.nlm.nih.gov/nuccore/NG_012088.1?from=5001&to=5224&report=fasta

Learn more about upcoming changes to the Nucleotide, EST, and GSS databases.

Homo sapiens interleukin 10 (IL10), RefSeqGene (LRG_1230) on chromosome 1

NCBI Reference Sequence: NG_012088.1

```
>NG_012088.1:5001-5224 Homo sapiens interleukin 10 (IL10), RefSeqGene (LRG_1230) on
chromosome 1
ACACATCAGGGGCTTGCTCTTGCAAAACCAAACCACAAGACAGACTTGCAAAAGAAGGCATGCACAGCTC
AGCACTGCTCTGTTGCCTGGTCCTCCTGACTGGGGTGAGGGCCAGCCCAGGCCAGGGCACCCAGTCTGAG
AACAGCTGCACCCACTTCCCAGGCAACCTGCCTAACATGCTTCGAGATCTCCGAGATGCCTTCAGCAGAG
TGAAGACTTTCTTT
```

Change region shown: ○ Whole sequence ● Selected region from: 5001 to: 5224

https://www.ncbi.nlm.nih.gov/tools/primer-blast/index.cgi?ORGANISM=9606&INPUT_SEQUENCE=NG_012088.1&LINK_LOC=nuccore&PRIMER5_START=5001&PRIMER3_EN...

Primer-BLAST A tool for finding specific primers

Finding primers specific to your PCR template (using Primer3 and BLAST).

PCR Template

Enter accession, gi, or FASTA sequence (A refseq record is preferred): NG_012088.1

Or, upload FASTA file: No file chosen

Range	From	To
Forward primer	5001	
Reverse primer		5224

Primer Parameters

Use my own forward primer (5'->3' on plus strand):

Use my own reverse primer (5'->3' on minus strand):

	Min	Max
PCR product size	70	1000
# of primers to return	10	

	Min	Opt	Max	Max Tm difference
Primer melting temperatures (Tm)	57.0	60.0	63.0	3

Fig. 16.4 PCR primer designing steps 5 to 8

tide positions of the desired site by using “Range” dedicated entry. Nucleotide positions refer to the base numbers as per their arrangement in the annotated sequence segment on the plus strand of target template (“From” position should be smaller than “To” position for any given primer). For example, if the desired PCR product is located between nucleotide in position 100 and position 1000 on the template, then, and as for the forward primer, the command “From” can be set to 100 and reverse primer “To” to 1000.

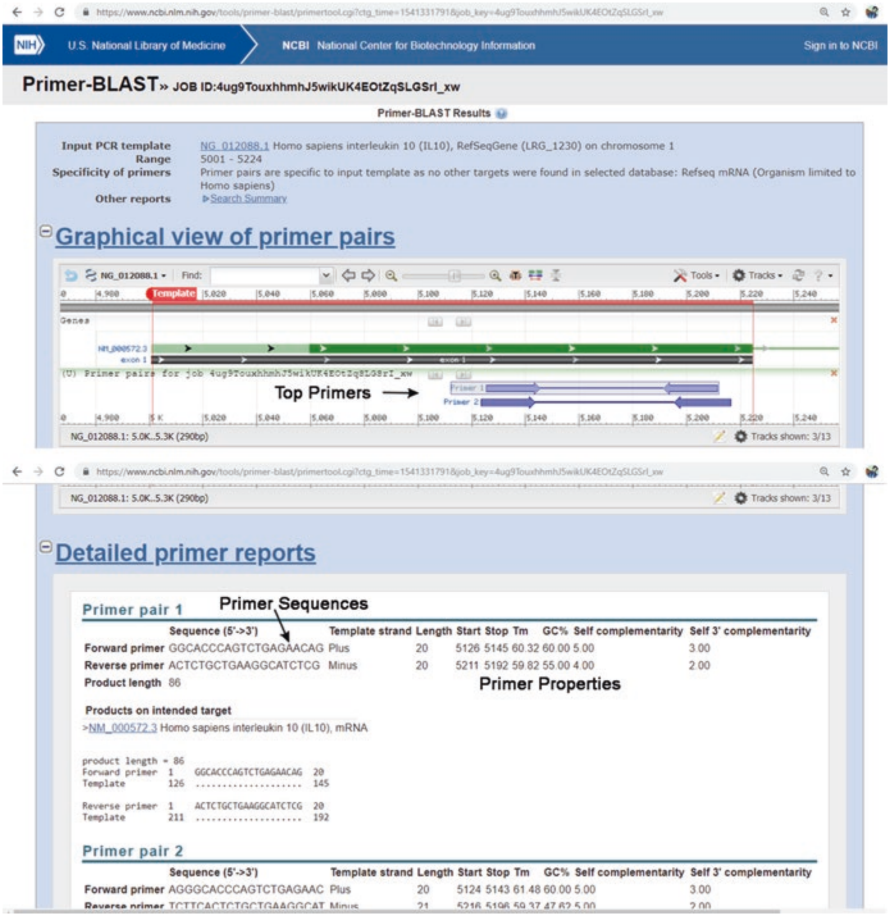
It is also possible to investigate known designed primers to check for their unique actions to anneal exclusively to the area surrounding exon 5 only by entering the actual sequence of the forward primer into the search box designated as “Plus stand” under the section of the “Primer Parameters.” Such analysis can be done separately for each primer, or both primers (forward and reverse) can be entered at once. Reverse sequence of reverse primer can go into the search box designated as “minus Strand” under the same section. Preferred PCR product size can also be predetermined to set Primer-BLAST to identify primers to amplify specific size product. In the entry “PCR product Size” under the section “Primer Parameters,” for example, the minimum and maximum product size can be set to amplify, no more than 400 base pairs and no less than 200 base pairs, assuming that that is where the area of research interest lies.

In the Primer Pair Specificity Checking parameters section, selecting the appropriate source organism, the smallest database in addition to choosing the nonredundant (nr) database will generate precise results and are likely to limit the searching area for Primer-BLAST.

Seventh step After the selection of “Pick Primers,” it is optional to view the generated results in a separate window. The search engine of Primer-BLAST will perform a quick database scan to detect PCR templates that are highly similar to the query sequence (pasted nucleotides sequence), which will be used for the selection of primers.

Eighth step Primer-BLAST has already generated set(s) of primers dedicated to the input PCR template (NG_01288.1) of the gene IL10, exon 5 sequence. By the selection of nr database, generated primers are most likely site-specific and less likely they would have a chance to bind somewhere else in the genome other than the targeted sequence of the exon. Graphical view of designed primer sets provides the proper coverage of the specific area of the research interest within the target sequence of the gene for amplification.

Ninth step Primer-BLAST usually generates sets of primers and arranges them in a way that the first suggested set primer is usually the one with the highest potential to successfully amplify the target sequence. According to the displayed primer reports, each and every generated set should include access details of actual nucleotide sequence, for both forward and reverse primers, potential product size of the PCR target, direction of the template strand, melting temperature and the number of base pairs that would potentially self-complement to each other. Such report should help in selecting the ideal primer for PCR amplification (Fig. 16.5).


U.S. National Library of Medicine
NCBI National Center for Biotechnology Information
Sign in to NCBI
NG_012088.1
Find:
Tools
Tracks
Template
Genes
NM_000572.3
exon 1
(U) Primer pairs for job 4ug9TouxhhmhJ5wikUK4EOtZqSLGSrI_xw
Top Primers
Primer 1
Primer 2
NG_012088.1: 5.0K..5.3K (290bp)
Tracks shown: 3/13
Primer Sequences
Primer Properties

Primer-BLAST» JOB ID:4ug9TouxhhmhJ5wikUK4EOtZqSLGSrl_xw

Primer-BLAST Results

Input PCR template	NG_012088.1 Homo sapiens interleukin 10 (IL10), RefSeqGene (LRG_1230) on chromosome 1
Range	5001 - 5224
Specificity of primers	Primer pairs are specific to input template as no other targets were found in selected database: Refseq mRNA (Organism limited to Homo sapiens)
Other reports	▶Search Summary

Graphical view of primer pairs

Detailed primer reports

Primer pair 1

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	GGCACCCAGTCTGAGAACAG	Plus	20	5126	5145	60.32	60.00	5.00	3.00
Reverse primer	ACTCTGCTGAAGGCATCTCG	Minus	20	5211	5192	59.82	55.00	4.00	2.00
Product length	86								

Products on intended target

>NM_000572.3 Homo sapiens interleukin 10 (IL10), mRNA

```
product length = 86
Forward primer  1    GGCACCCAGTCTGAGAACAG  20
Template        126  ....................  145

Reverse primer  1    ACTCTGCTGAAGGCATCTCG  20
Template        211  ....................  192
```

Primer pair 2

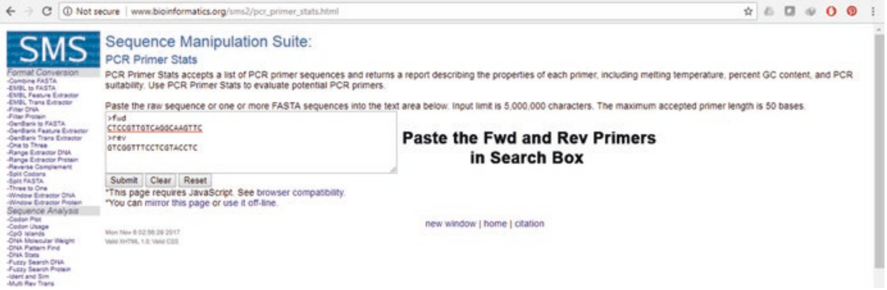
	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	AGGGCACCCAGTCTGAGAAC	Plus	20	5124	5143	61.48	60.00	5.00	3.00
Reverse primer	TCTTCACTCTGCTGAAGGCAT	Minus	21	5216	5196	59.37	47.62	5.00	2.00

Fig. 16.5 PCR primer designing steps 9 to 10

Final step For clarification of the quality of the selected primers, it is recommended to perform a simple primer evaluation test via PCR Primer Stats which accepts a list of PCR primer sequences and returns a report describing the properties of each primer, including melting temperature, percent GC content, and PCR suitability. Use PCR Primer Stats to evaluate potential PCR primers. The raw sequence or one or more FASTA sequences should be pasted into the text area below. The input limit is 5,000,000 characters. The maximum accepted primer length is 50 bases (Fig. 16.6).

PCR Primer Stats results
Global settings:

- The primers do not have a 5'-phosphate group.
- Combined concentration of K+ and Na+ in the reaction = 50 millimolar.



Sequence Manipulation Suite:

PCR Primer Stats

PCR Primer Stats accepts a list of PCR primer sequences and returns a report describing the properties of each primer, including melting temperature, percent GC content, and PCR suitability. Use PCR Primer Stats to evaluate potential PCR primers.

Paste the raw sequence or one or more FASTA sequences into the text area below. Input limit is 5,000,000 characters. The maximum accepted primer length is 50 bases.

```
>fwd
CTCCGTTGTCAGGCAAGTTC
>rev
GTCGGTTTCCTCGTACCTC
```

Paste the Fwd and Rev Primers in Search Box

*This page requires JavaScript. See browser compatibility.
*You can mirror this page or use it off-line.

Fig. 16.6 A home page of Primer Stats – a primer validation tool

– Mg+2 concentration in the reaction = 1.5 millimolar.
– Primer concentration in the reaction = 200 nanomolar.

```
----------------------------------------------------------
General properties:
------------------
Primer name: fwd
Primer sequence: GGCACCCAGTCTGAGAACAG
Sequence length: 20
Base counts: G=6; A=6; T=2; C=6; Other=0;
GC content (%): 60.00
Molecular weight (Daltons): 6136.04
nmol/A260: 5.07
Micrograms/A260: 31.08
Basic Tm (degrees C): 56
Salt adjusted Tm (degrees C): 51
Nearest neighbor Tm (degrees C): 64.0365.31
PCR suitability tests (pass/warning):
----------------------------------
Single base runs: Pass
Dinucleotide base runs: Pass
Length: Pass
Percent GC: Pass
Tm (nearest neighbor): Warning; Tm is greater than 58
GC clamp: Pass
Self-annealing: Pass
Hairpin formation: Pass
----------------------------------------------------------
General properties:
------------------
Primer name: rev
Primer sequence: ACTCTGCTGAAGGCATCTCG
```

```
Sequence length: 20
Base counts: G=5; A=4; T=5; C=6; Other=0;
GC content (%): 55.00
Molecular weight (Daltons): 6093.01
nmol/A260: 5.41
Micrograms/A260: 32.95
Basic Tm (degrees C): 54
Salt adjusted Tm (degrees C): 49
Nearest neighbor Tm (degrees C): 64.49
PCR suitability tests (pass/warning):
----------------------------------
Single base runs: Pass
Dinucleotide base runs: Pass
Length: Pass
Percent GC: Pass
Tm (nearest neighbor): Warning; Tm is greater than 58
GC clamp: Pass
Self-annealing: Pass
Hairpin formation: Pass
--------------------------------------------------------
```

16.5 Conclusion

In the current chapter, we demonstrated a pre-laboratory, computational PCR primer designing and evaluation method which could effectively decrease the chances of synthesizing and optimizing the false PCR primer sequences. This chapter would enable researchers to self-design highly sensitive and specific primers of their choice, using accessible and easy-to-use web resources. The online web server described in this chapter can aid in the rapid search of primer sequences and determine their orientation, location, melting point, secondary structure, and binding potential. The "PCR Primer Stats" program described in this chapter helps in the selection of best primer sets from the predicted ones, by validating them against the core properties of an ideal primer sequence. Moreover, this chapter also underlines the fact that in silico PCR is not just suitable for traditional PCR alone but also to a variety of other PCR methods like Fast PCR, inverse PCR and multiplex PCR, etc.

References

Abd-Elsalam KA (2003) Bioinformatic tools and guideline for PCR primer design. Afr J Biotechnol 2(5):91–95

Breslauer KJ, Frank R, Blöcker H, Marky LA (1986) Predicting DNA duplex stability from the base sequence. Proc Natl Acad Sci U S A 83(11):3746–3750

Chen H, Zhu G (1997) Computer program for calculating the melting temperature of degenerate oligonucleotides used in PCR or hybridization. Biotechniques 22(6):1158–1160

Dieffenbach CW, Lowe TM, Dveksler GS (1993) General concepts for PCR primer design. Genome Res 3(3):S30–S37

Gaudet M, Fara AG, Beritognolo I, Sabatti M (2009) Allele-Specific PCR in SNP Genotyping. In: Komar A (eds) Single Nucleotide Polymorphisms. Methods in Molecular Biology™ (Methods and Protocols), vol 578. Humana Press, Totowa, NJ

He Q, Marjamaki M, Soini H, Mertsola J, Viljanen MK (1994) Primers are decisive for sensitivity of PCR. BioTechniques 17(1):82, 84, 86–82, 84, 87

Kwok S, Kellogg DE, McKinney N, Spasic D, Goda L, Levenson C, Sninsky JJ (1990) Effects of primer-template mismatches on the polymerase chain reaction: human immunodeficiency virus type 1 model studies. Nucleic Acids Res 18(4):999–1005

Rychlik W, Rhoads RE (1989) A computer program for choosing optimal oligonucleotides for filter hybridization, sequencing and in vitro amplification of DNA. Nucleic Acids Res 17(21):8543–8551

Rychlik W, Spencer WJ, Rhoads RE (1990) Optimization of the annealing temperature for DNA amplification in vitro. Nucleic Acids Res 18(21):6409–6412

Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT et al (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science 239(4839):487–491

Saiki RK, Scharf S, Faloona F, Mullis KB, Horn GT, Erlich HA, Arnheim N (1985) Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science 230(4732):1350–1354

Sheffield VC, Cox DR, Lerman LS, Myers RM (1989) Attachment of a 40-base-pair G + C-rich sequence (GC-clamp) to genomic DNA fragments by the polymerase chain reaction results in improved detection of single-base changes. Proc Natl Acad Sci U S A 86(1):232–236

Tabchoury CP, Sousa MC, Arthur RA, Mattos-Graner RO, Del Bel Cury AA, CURY JA (2008) Evaluation of genotypic diversity of Streptococcus mutans using distinct arbitrary primers. J Appl Oral Sci 16:403–407

Wallace RB, Shaffer J, Murphy RF, Bonner J, Hirose T, Itakura K (1979) Hybridization of synthetic oligodeoxyribonucleotides to phi chi 174 DNA: the effect of single base pair mismatch. Nucleic Acids Res 6(11):3543–3557

Wu DY, Ugozzoli L, Pal BK, Qian J, Wallace RB (1991) The effect of temperature and oligonucleotide primer length on the specificity and efficiency of amplification by the polymerase chain reaction. DNA Cell Biol 10(3):233–238. https://doi.org/10.1089/dna.1991.10.233

Chapter 17
Modeling and Optimization of Molecular Biosystems to Generate Predictive Models

Ankush Bansal, Siddhant Kalra, Babajan Banaganapalli, and Tiratha Raj Singh

Contents

17.1 Introduction

A system can be defined as a complex structure in which different components have a specific role, and when they work together, they accomplish tasks in much efficient manner compared to each component separately (Kitano 2002). The system is a collection of elements or components that are organized for a common purpose. The biological system analysis provides us tools and techniques that help in

A. Bansal · S. Kalra · T. R. Singh (✉)
Department of Biotechnology and Bioinformatics, Jaypee University of Information Technology, Solan, Himachal Pradesh, India

B. Banaganapalli
Princess Al-Jawhara Center of Excellence in Research of Hereditary Disorders, Department of Genetic Medicine, Faculty of Medicine, King Abdulaziz University, Jeddah, Saudi Arabia
e-mail: bbabajan@kau.edu.sa

N. A. Shaik et al. (eds.), *Essentials of Bioinformatics, Volume I*,
https://doi.org/10.1007/978-3-030-02634-9_17

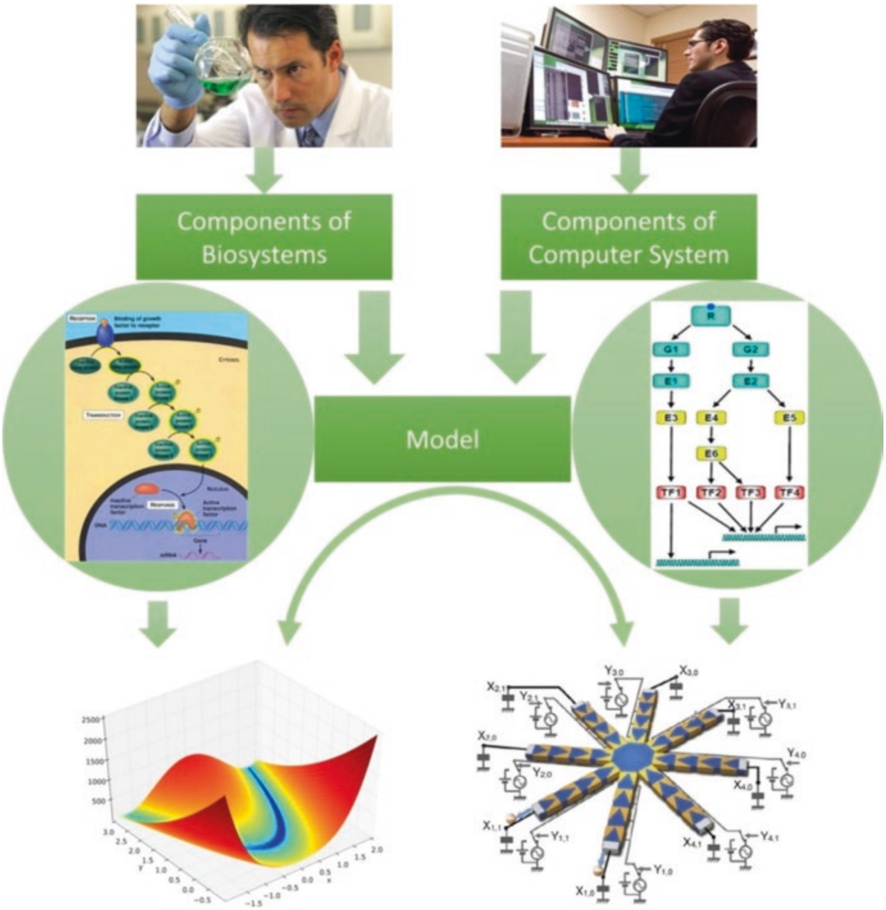


Fig. 17.1 System level understanding of bio-models

organizing the diverse piece of information and data gathered from traditional biological experiments. Development, integration, and experimental testing of hypothesis help us to analyze these systems, as depicted in Fig. 17.1.

Modeling means converting our hypothesis or assumptions into computational programs which further can be used for prediction. Some suitable assumptions are essential for model construction which includes modeling of the system into mathematical form. The mathematical model includes all kind of variables, real numbers, integers, Boolean flags, matrices, and other data structure. Each interaction represents a state in the model, and the final step involves converting the mathematical model into a computer program which is done by suitable genetic algorithms and other differential equation analysis-based algorithm. Once the computational model is built, it requires testing and verification in terms of validation. Models are helpful as they help us to test the different hypothesis, refine and interpret experiment, and integrate knowledge, leading to new approaches by investigating coupling and feedback. The model helps us to unlock biological systems as they offer different perspectives compared to the perspective provided by experiments and theory. Though models cannot replace lab experiments and cannot prove mechanism, still they serve as a standard feature for scientific investigations.

The illustrative models are precise representation of real situations. Here, we specifically focus on the controlling element from the real world which can be used as a deterministic factor to control our modeled system. Mathematics plays a dominant role in defining system using variables which precisely define real-world scenario. Using mathematical equations we can simply find the solution of various common problems. There are various network level studies exist in literature to perform modeling for individual nodes or high throughput data (Bansal and Ramana 2015; Bansal and Srivastava 2018; Davis et al. 2017; Giraud et al. 2017; Griffen et al. 2017; Jindal and Bansal 2016; Jo et al. 2017; Kim et al. 2017; Nordholt et al. 2017; Romero & López 2017; Vreven et al. 2017; Xie et al. 2017).

Once the model has been generated, it should be optimized. Optimization means to find the best solution that helps in better decision making. The key elements in optimization problems are decision variables and objective function (Zheng et al. 2017). Decision variables are the variables that can be varied during the search of the best solution. An objective function helps us to quantify the quality of a solution, and constraints are the conditions that should be fulfilled in order to achieve the desired results. Different optimization techniques such as linear programming, nonlinear programming, parameter estimation, dynamic optimization, etc. are used for different problems.

Modeling is of no use until it is optimized as per the real situation. Thus, it is a key process for any kind of real model establishment. Optimization mainly controlled the principles of machine learning. We are considering one dataset for analysis and splitting it into two datasets, commonly known as training and test dataset. Various modelers use machine learning-based approaches in their algorithms for better efficiency. Henceforth, we can say that model is typically dependent on optimization using set of variables or parameters which can be used to regulate or control the models as per the need of optimizer. Mathematical modeling is compiled with various machine learning approaches in applicable manner which came into market in the form of various development tools and software.

With the increase of computer power and advanced mathematical techniques, mathematics is now playing the prominent role of integrating information and generating predictions, through the generation of the computationally inspired hypothesis. Therefore, mathematical models can be used to understand the complex biological problems to unbind various diseases and drug effects to benefit the society in utmost sophisticated manner. Mathematical model allows a systematic approach for investigating system perturbations and is not limited to experimental constraints (Fan et al. 2017). These models are able to determine the systematic behavior of any real-world disease scenario.

17.2 Development of Concept Map Models

Biological experiments deal with understanding of hidden processes in the layers of various unannotated datasets. The major goal of such analysis is to provide new insight about regulation mechanisms so that the system can be controlled in an efficient manner. A variety of homogeneous and heterogeneous data are generated through various big data approaches using high-throughput methods. Generation of data is not sufficient to perform analysis to reveal the function of the system. There is a great need of concept map modeling to understand the systematic way to deal with such big data (Kumar et al. 2017).

Concept map modeling focuses on understanding and developing concept for development of methods for mathematical analysis. This approach is time-consuming as it involves the development of a model for the process and response for each level. Therefore it is important to derive such models that allow the incorporation of simple as well as complex methods for complete as well as incomplete datasets at defined instance (Sun et al. 2017). The main objective of this approach is to get acquainted with the quantitative formalization of the biological phenomenon by developing mathematical model for the hypothesis.

The initial step of this approach consists of converting or transforming a stable or static map to dynamic biological map. The next step consists of interpretation of local dynamic response under a set of conditions. By following these two steps, one can determine a parameterized model which is further analyzed and refined. A flow diagram of this approach is shown in Fig. 17.2.

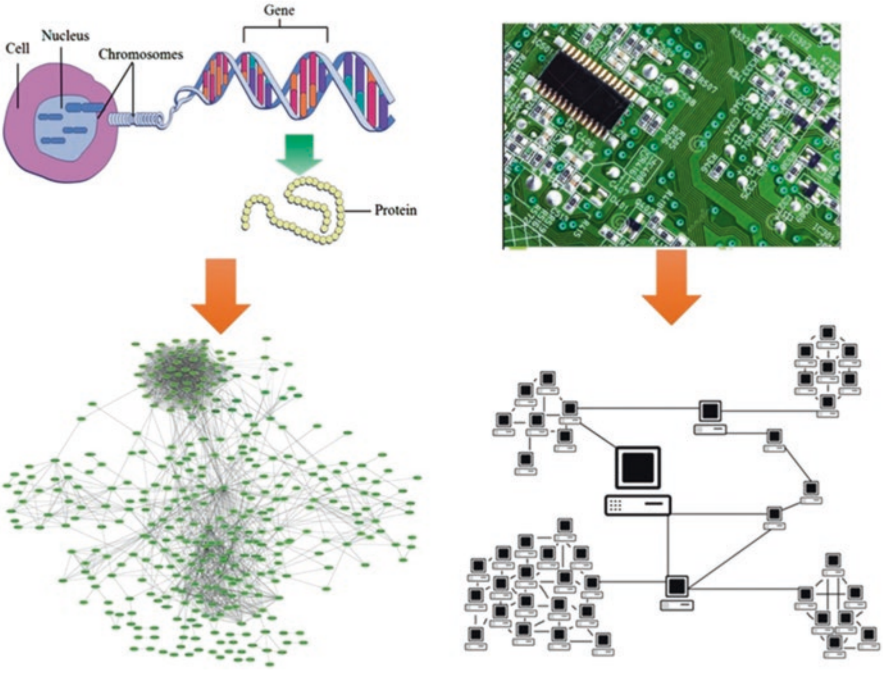


Fig. 17.2 Molecular classification of system to modeling and optimization

One needs to examine how the components and process in a concept map relate to each other and contribute to the overall functioning. Conversion of the map into mathematically testable structures is an essential part of maps as such system cannot provide quantitative analysis themselves. Considering a modeling method, regulatory interactions can be inferred using mathematical variables or symbolic representations (Fig. 17.3).

The static maps can be converted into Boolean or semiquantitative dynamics (SOD) map if a biologist has some prior knowledge about the information contained in the static map such as the type of reaction or time required to convert gene expression (Kumar and Singh 2017; Teku and Vihinen 2017). The Boolean case determines the close relationship of having direct control on the components within the global system. For instance, gene X is essential for process Y to occur. It helps us to determine the accurate function which is applied in inverse methods.

In the real case, the concept of the model represents control about dynamics of each node available rather than the detailed time series. An initial model can be constructed with the help of this minimum information. Once the model is substituted by actual time series, a simple function can be determined that captures the dynamics at each node (Sehgal et al. 2015). The overall mathematical formulation and understanding to develop models are not always a critical task as generalized

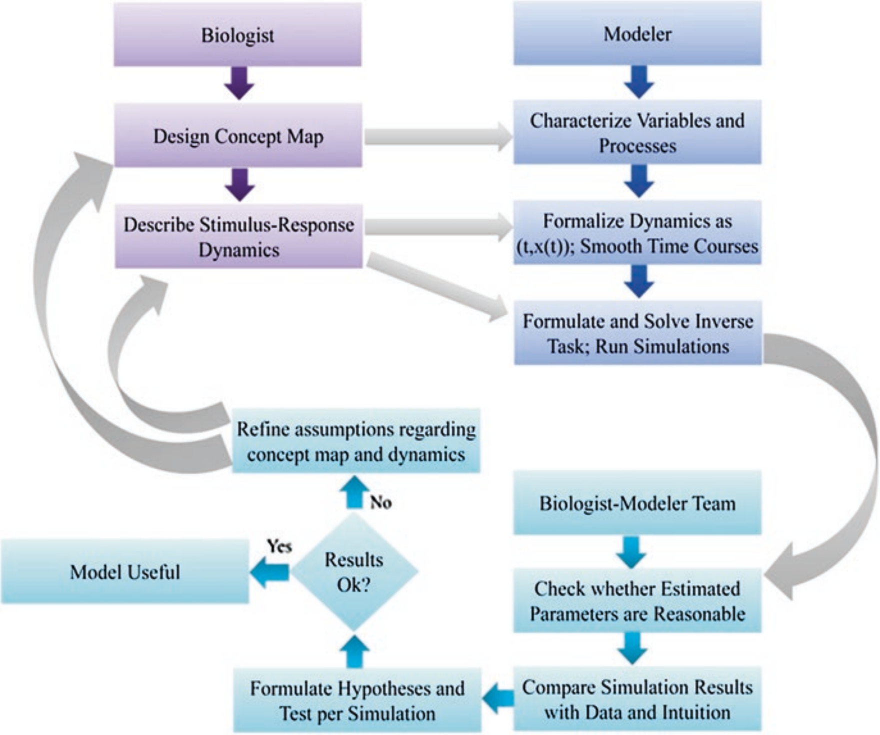

Biologist
Modeler
Design Concept Map
Characterize Variables and Processes
Describe Stimulus-Response Dynamics
Formalize Dynamics as (t,x(t)); Smooth Time Courses
Formulate and Solve Inverse Task; Run Simulations
Refine assumptions regarding concept map and dynamics
No
Biologist-Modeler Team
Yes
Model Useful
Results Ok?
Check whether Estimated Parameters are Reasonable
Formulate Hypotheses and Test per Simulation
Compare Simulation Results with Data and Intuition

Fig. 17.3 Flow diagram of the proposed approach to formalizing biological concept maps

models can be used for depictions of user-specific data. This data can be further customized in the forms of graphs or curves. For instance, dynamicity of the system can be represented in the form of various distributional curves or sigmoid curves. Once we have constructed these model-derived curves, we can switch on or off the functions and change the curves as per the need in the presence or absence of defined variables or parameters. Condition-based approximation and differential analysis on the basis of conditions can be applied on these generated models.

17.3 Network of Networks

A network helps in understanding and combining scattered data at various dimensions. One of the key features of systems biology is focusing on "network of networks." In the human body, *n* number of networks is integrated in such a fashion so that efficient communication can happen at molecular and cellular levels. Generating understandable biosystems may help us to get insights about biological functions and variations and trace out changes at cellular to phenotypic levels. Figure 17.4 represents the structure of "network of networks" which gives an idea about various system biology approaches which differs from traditional biological approaches.

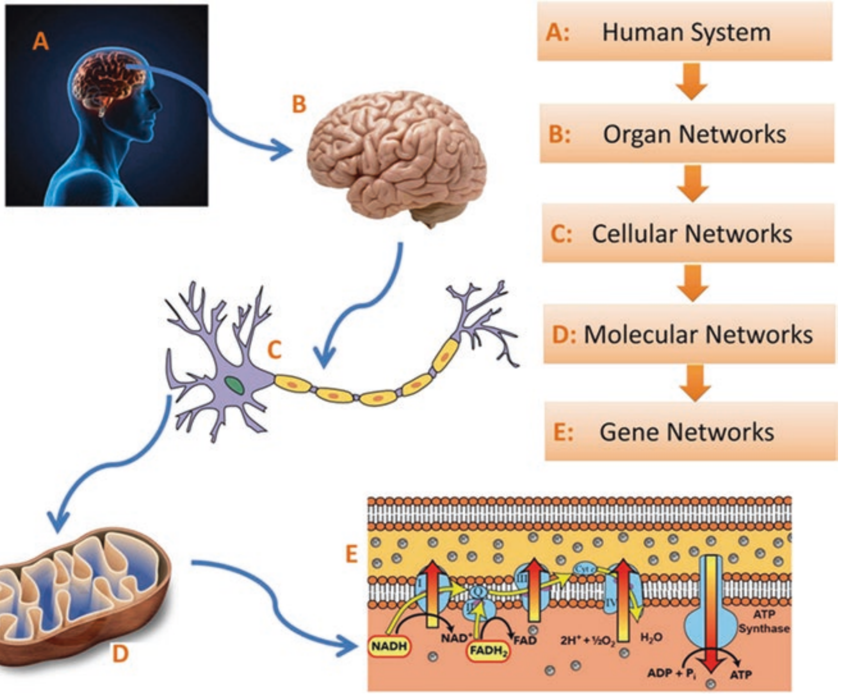


Fig. 17.4 Systems Dissection in terms of networks of networks

17.4 System Dissection into Components

Biological systems can be implemented in various ways; precisely it can be dissected using four components.

1. High-throughput methods for data generation which includes identifying unknown information from the depth of biological aura.
2. Developing concept, logic, and computational methods to combine various biological datasets to infer meaningful information.
3. Hypothesis generation and testing on newly generated data and comparison of the same existing data in various online portals and literature.
4. Understanding global scenario as big data and solving the phenotypic effects related to problems in differential data analysis for new information discovery.

17.5 Types of Modeling

Mathematical modeling is composed of various standard parameters, conceptual framing of tools, and interpretation of any kind of real system in mathematical form to decipher the control mechanics of the system. Mathematical representation of biological systems not only constructs the models but also optimizes and predicts in much efficient way compared to various traditional approaches. Thus, mathematical models can be implemented in terms of stochastic process, continuous process, or any other black box representation which doesn't have well-known information of composition.

For all the cases, the modeling process consists of the following same steps. First, using physical laws from first principles, a symbolic model is constructed which serves as an extension to the already known existing model (Athanasiou et al. 2017). This model consists of variables and parameters. The analysis requires comprehension of all parameter values obtained from biological knowledge. Variables in mathematical modeling can represent anything, whether it is a plant, animal, metabolite, pathway, or gene expression. Approximation and estimation of any parameter in biological terms is quite difficult as biological phenomenon doesn't reveal complete information in one go as other modeled systems do. The analysis of the model is done with the techniques and tricks of mathematics and computer science once the parameters are estimated. Due to the complexity of biological systems, optimization and analysis of differential conditions and large datasets are performed using computational approaches. Interpretation in terms of graphs and matrix provides an edge to scientific community to accurately depict the behavior of the aligned system.

The identification of unknown parameters in terms of biological entities is the genuine deterrent in the progress of biomathematical modeling. A non-specific approach called *biochemical systems theory* is used for biological systems modeling and analysis which is used for the improvements, developments, and applications of thousands of research papers. BST was initially used to study the control systems and biochemical pathways.

The fundamental precepts of BST are very basic and transparent. Every variable that progresses after some time is given a name X and is represented in the form of the different orders of differential equation and depicts the variation in such a way so that it can affect other variables or parameters in positive or negative regulatory ways. BST also addresses the problem where the modeler has some broad data about the procedures but does not know their mathematical representation to develop a structure to solve the complexity of biological systems. Sometimes it is very difficult for a developer to develop a system which doesn't contain absolute values, or sometimes a developer is not having an idea of unknown things in the systems, but logically if we speak about linear regression, we are not sure about what kind of data points are there which need to be included or excluded at initial point. Both approaches are somehow similar while dealing with unknown information and positively providing an edge to mathematical modeling to structure the unstructured data. As biological networks don't follow the Poisson distribution and converge toward scale-free networks which comprises the properties of power law. So, it will not be wrong to say that such approaches can result in successful analysis toward validation of real dataset.

17.5.1 Forward Modeling

Identification of a parameter in a system is based on local information which subsequently deals with small component integration and formation of complete network. For instance, for metabolic pathway construction, there is a need to understand the enzymes involved in pathway, transporters involved, co-factors playing the role in regulation, and ultimately metabolite formation through secondary metabolisms. All these terms need to be integrated to form mathematical equations and depict the understanding of biological phenomenon. Biological modeling is generally dealt with Michaelis-Menten or power law function. Dynamicity of the system is controlled by various rate law and parameter approximations like Km and Vmax, and forward rate of reactions can be controlled on the basis of concentration assigned to each entity defined in the model (Apostolopoulos et al. 2017). In such modeling methods, there is a need to study the direct rate law to control the local parameters and test various hypotheses on the basis of developed models.

The main utilization of this method is the use of kinetic equations, using enzyme concentration for tracing the rate of reaction. Variation in the rate of reaction subsequently leads to variation at phenotypic levels. Construction of such models and their refinements always has been a crucial task for scientists in biological community.

17.5.2 Inverse Modeling

Variables are observed from high end to low end which means reduction approach. The most important advantage of this technique is that data is originated from the same organism, acquired in a similar trial condition, and represented in all the procedures within the organism that could affect the factors of the framework (Kallhovd et al. 2017). Computational time complexity is a major issue with such kind of analysis. Moreover, various biological entities are ignored in case of modeling. The inverse modeling also use time-dependent analysis where pathways information is not absolute.

17.5.3 Partial Modeling

A specific issue with any model building approaches emerges due to the presence of the “omnipresent” metabolites like energy molecules (ATP) which cannot be modeled as they are additionally required in different reactions. As a result, a mathematical buffer is constructed that absorbs the excess material, thus adjusting the dynamic changes in concentration at an already determined rate (Yalçın et al. 2017). Better-characterized statements are defined as differential conditions in BST, and their progression includes energy molecules as factors.

17.6 Inference from Qualitative Data to Computational Simulation

Biological system usually deals with enormous methods and tools whether they are qualitative or quantitative. Sometimes, there are exact implications of a system that are missing, and semiquantitative methods are prioritized over other measures of data segmentation or integration for network model construction. For instance, graphical methods represent directional flow of the information by connecting components of a system in a systematic fashion. Moreover, network construction and hypothesis testing on the basis of available information and predicting the information of missing links in the networks provide more insights about qualitative measurement from raw unstructured data. Various probabilistic measures like Markov chains which are used to represent Hidden Markov models and Bayesian model-based networks deal with graphical presentation of unknown entities in a network through random measure.

Sometimes, these graphical methods do not represent the dynamicity of the network and do not express much detailed information as per real-time scenario;

therefore mechanistic models come to existence where data can be analyzed in an automated manner.

Computer-based models and simulations provide an easy tool to understand biological systems in terms of complex nonlinear dynamics. The first is that "instinctive thinking about MAP kinase pathways led to the long-held view that the obligatory cascade of three sequential kinases serves to provide signal intensification. In contrast, computational studies have suggested that the purpose of such a network is to achieve extreme positive cooperativity so that the pathway behaves in a switch-like, rather than a graded, fashion."

Simulations present an understanding of biological phenomenon over differential time. Using differential equations on the same biological dataset can reveal hidden properties of the systems. But it will be unfair to expect accurate prediction through computational methods as these methods are developed to get insight about candidate entity selection. More data leads to more simulation time and subsequently increases the rate of precise selection of prediction attribute. Optimization can be performed on the basis of simulation measures of selected parameter. Simulation results in certain biological behavior analysis especially can be used in case of complex disease like cancer, diabetes, and neurodegenerative diseases. Simulations are modern and nontraditional techniques. In earlier days, people used conferences, abstract, and poster presentations to grab the idea of one's understanding. With the advancement in the internet world, these techniques can be integrated to form network to get holistic view of understanding of different people across the world (Huang et al. 2017). With the advancement in computational resources, the time and space complexity has not been an issue in the present world. So, mathematical simulations remain as the best alternative to reduce the time, effort, and resources of any wet lab experiments.

17.7 Protein Class Identification

The helix-turn-helix structural motif has an important and crucial role in various cellular pathways that are involved in transcription, DNA recombination and repair, and DNA replication. At present, methods that are used for motif identification are dependent on the amino acid sequence. The major drawback of these methods is that motif members belong to different sequence families that do not share common ancestry or homology, and hence these methods are incapable to identify all motif members (Qing and Gerson 2017).

So to overcome this drawback, a new method based on three-dimensional structure was created that involved the following steps:

1. Selecting a conserved component of the motif.
2. Computing structural features relative to that component.

3. Generating categorization models by comparing the relevant measurements of structures that contain motifs and those structures that do not contain motifs.

With the establishment of classification model, the entire Protein Data Bank of experimentally measured structures was searched, and new examples of motifs were identified that do not show any sequence homology with previously known examples. Two such examples are Esa1 histone acetyltransferase and flavone 4-O-methyltransferase. This result shows the importance of classification-based method that is proven helpful for the two abovementioned examples. The sequence-based methods are used to recognize a functional class of protein which can be improved by using the classification model that is based on three-dimensional structure information.

17.8 Computational Structure and Function Prediction

With the help of X-ray, NMR, and computational method techniques, structural genomics is now showing great enhancement in producing the three-dimensional structures of proteins. The important and crucial step after this is to understand how protein structure and functions are related. Studying protein structure individually impairs the overall understanding of the protein as various missing links will exist while studying a part of the protein. The availability of the expected surfeit protein structures has resulted in the development of computational methods that examines multiple protein structures at once and returns the important biophysical and biochemical features. Apart from this, these methods can also recognize important features in new protein structures (Winter et al. 2015) (Fig. 17.5).

FEATURE is an automated system developed by Wei and Altman. This system applies statistical parameters to study vital functional and structural sites in protein structures such as active sites, binding sites, disulfide bonding sites, and so forth. By collecting all known examples of a type of site and non-site, FEATURE computes the spatial distributions of defined biophysical and biochemical properties. It applies various statistical measures to calculate accurate, active, and binding sites. The use of parametric and nonparametric test provides this tool a high-level sensitivity and specificity.

SBML, Gepasi, and CellML are specialized systems for biological and biochemical modeling (Webb and White 2005). Madonna is a general-purpose system for solving a variety of equations (differential equations, integral equations, and so on). This has been represented in Fig. 17.6.

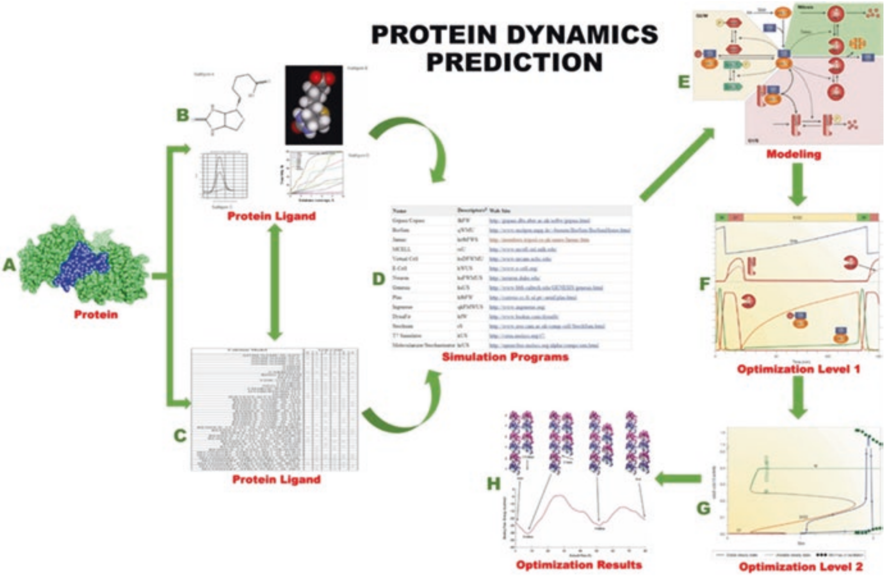


Fig. 17.5 Protein structure, docking and dynamics study

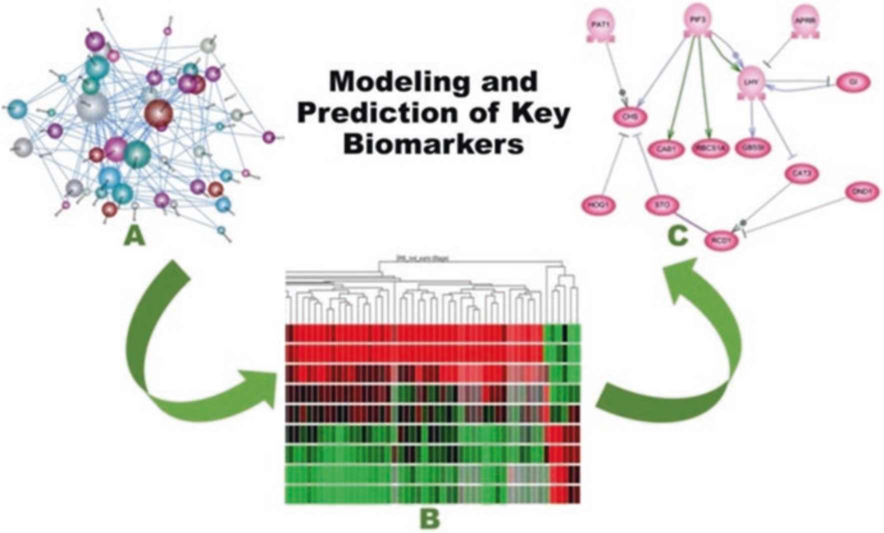


Fig. 17.6 Modeling system and screening key biomarkers

17.9 Forest Dynamics

SORTIE is a stochastic and mechanistic model that has been developed to simulate the growth of northeastern forests. This model mimics the fate of individual tree and its offspring. The model is based on the species-specific information regarding the growth rates, fecundity, mortality, and seed dispersal distances as well as some information regarding local regimes. SORTIE generates dynamic map by following tens of thousands of trees. This dynamic map depicts the distribution of nine dominant or subdominant species of trees that look like real forests. The model also predicts the realistic forest responses to certain minor and major disturbances like destruction of tress within small circle of forest boundary and improved tree mortality.

17.10 Cell Designer: A Computational Tool for Modeling

CellDesigner is a software developed by Systems Biology Institute using Systems Biology Markup Language and graphical notation. Different kinds of boxes were used to represent different kinds of biological entities. And different kinds of flux box reactions are present in the model to define kinetic equations. Interaction between one entity (i.e., node) to another is represented by edges. The graphical design of the software is supported by Jarnac, Plot, and Gibson, while associated databases are BioModels, PubMed, IHOP, KEGG, and SABIO. With the help of all these integrated modules, a user can model biochemical and gene regulatory networks. Using cell designer the user can create graphical notation for gene, RNA, and protein and also make a complex of protein. There are options to import and control the models developed by other people in systems biology field. The major parameter in this software is to perform simulation at molecular level using genes, proteins, or metabolite concentration at different time periods. Ordinary differential equations are used to create the simulation profiles. Simulation profiles can be analyzed and compared within a model, same organism model or other model. Another important feature of this modeling tool is to study the small pathway by considering a system as a whole which implies that the user need not to study complete information at one instance. The user can split their pathway of interests into different modules and later integrate them to reduce the time complexity for the simulation. Apart from this, there are various plugins which can be integrated with this software.

Cytoscape is a similar tool for model development on the basis of topological analysis. This tool lacks the use of simulation to study differential conditions, but statistical analysis and beautiful graphical layouts for representing networks provide an edge for this tool over other modeling softwares.

17.11 Conclusion

Major purpose of modeling and optimization in research is to systematically assemble and simulate all the molecules and their interactions that are occurring inside the living cell. There is a need to understand how these molecular interactions take place and how to determine the function of this complex machinery that cannot be solved only by biotechnology lab experiments. The advancement in the modeling techniques indicates that cellular networks are governed by diverse universal properties and offer a new conceptual structure that could potentially renovate our view of biology and drug therapies.

References

Apostolopoulos Y, Lemke MK, Barry AE, Lich KH (2017) Moving alcohol prevention research forward-part II: new directions grounded in community-based system dynamics modeling. Addiction. https://doi.org/10.1111/add.13953

Athanasiou G, Anastasopoulos GC, Tiritidou E, Lymberopoulos D (2017) A trust model for ubiquitous healthcare environment on the basis of adaptable fuzzy-probabilistic inference system. IEEE J Biomed Health Inform. https://doi.org/10.1109/JBHI.2017.2733038

Bansal A, Ramana J (2015) TCGDB: a compendium of molecular signatures of thyroid cancer and disorders. J Cancer Sci Ther. https://doi.org/10.4172/1948-5956.1000350

Bansal A, Srivastava PA (2018) Transcriptomics to metabolomics: a network perspective for big data. http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-5225-2607-0.ch008:188–206. https://doi.org/10.4018/978-1-5225-2607-0.ch008

Davis LE, Jeng S, Svalina MN, Huang E, Pittsenbarger J, Cantor EL et al (2017) Integration of genomic, transcriptomic and functional profiles of aggressive osteosarcomas across multiple species. Oncotarget. https://doi.org/10.18632/oncotarget.19532

Fan D, Lopez Ruiz L, Gong J, Lach J (2017) EHDC: an energy harvesting modeling and profiling platform for body sensor networks. IEEE J Biomed Health Inform. https://doi.org/10.1109/JBHI.2017.2733549

Giraud S, Brock AM, Macé MJ-M, Jouffrais C (2017) Map learning with a 3D printed interactive small-scale model: improvement of space and text memorization in visually impaired students. Front Psych 8:930. https://doi.org/10.3389/fpsyg.2017.00930

Griffen BD, Riley ME, Cannizzo ZJ, Feller IC (2017) Indirect effects of ecosystem engineering combine with consumer behavior to determine the spatial distribution of herbivory. J Anim Ecol. https://doi.org/10.1111/1365-2656.12730

Huang M, Dissanayake T, Kuechler E, Radak BK, Lee T-S, Giese TJ, York DM (2017) A multidimensional B-spline method for accurate modeling sugar puckering in QM/MM simulations. J Chem Theory Comput. https://doi.org/10.1021/acs.jctc.7b00161

Jindal K, Bansal A (2016) APOEε2 is associated with milder clinical and pathological Alzheimer's disease. Ann Neurosci 23(2):112. https://doi.org/10.1159/000443572

Jo S, Cheng X, Lee J, Kim S, Park S-J, Patel DS et al (2017) CHARMM-GUI 10 years for biomolecular modeling and simulation. J Comput Chem 38(15):1114–1124. https://doi.org/10.1002/jcc.24660

Kallhovd S, Maleckar MM, Rognes ME (2017) Inverse estimation of cardiac activation times via gradient-based optimisation. Int J Numer Methods Biomed Eng. https://doi.org/10.1002/cnm.2919

Kim SY, Kawaguchi T, Yan L, Young J, Qi Q, Takabe K (2017) Clinical relevance of microRNA expressions in breast cancer validated using the cancer genome atlas (TCGA). Ann Surg Oncol. https://doi.org/10.1245/s10434-017-5984-2

Kitano H (2002) Systems biology: a brief overview. Science 295(5560):1662–1664. https://doi.org/10.1126/science.1069492

Kumar A, Singh TR (2017) A new decision tree to solve the puzzle of Alzheimer's disease pathogenesis through standard diagnosis scoring system. Interdiscip Sci Comput Life Sci 9(1):107–115. https://doi.org/10.1007/s12539-016-0144-0

Kumar V, Bansal A, Chauhan RS (2017) Modular design of picroside-II biosynthesis deciphered through NGS transcriptomes and metabolic intermediates analysis in naturally variant chemotypes of a medicinal herb, Picrorhiza kurroa. Front Plant Sci 8:564. https://doi.org/10.3389/fpls.2017.00564

Nordholt N, van Heerden J, Kort R, Bruggeman FJ (2017) Effects of growth rate and promoter activity on single-cell protein expression. Sci Rep 7(1):6299. https://doi.org/10.1038/s41598-017-05871-3

Qing Y, Gerson SL (2017) Mismatch repair deficient hematopoietic stem cells are preleukemic stem cells. PLoS One 12(8):e0182175. https://doi.org/10.1371/journal.pone.0182175

Romero AH, López SE (2017) In silico molecular docking studies of new potential 4-phthalazinyl-hydrazones on selected Trypanosoma cruzi and Leishmania enzyme targets. J Mol Graph Model 76:313–329. https://doi.org/10.1016/j.jmgm.2017.07.013

Sehgal M, Gupta R, Moussa A, Singh TR (2015) An integrative approach for mapping differentially expressed genes and network components using novel parameters to elucidate key regulatory genes in colorectal cancer. PLoS One 10(7):e0133901. https://doi.org/10.1371/journal.pone.0133901

Sun L, Sang M, Zheng C, Wang D, Shi H, Liu K et al (2017) The genetic architecture of heterochronsy as a quantitative trait: lessons from a computational model. Brief Bioinform. https://doi.org/10.1093/bib/bbx056

Teku GN, Vihinen M (2017) Simulation of the dynamics of primary immunodeficiencies in CD4+ T-cells. PLoS One 12(4):e0176500. https://doi.org/10.1371/journal.pone.0176500

Vreven T, Pierce BG, Borrman TM, Weng Z (2017) Performance of ZDOCK and IRAD in CAPRI rounds 28-34. Proteins 85(3):408–416. https://doi.org/10.1002/prot.25186

Webb K, White T (2005) UML as a cell and biochemistry modeling language. Biosystems 80(3):283–302. https://doi.org/10.1016/j.biosystems.2004.12.003

Winter A, Schmid R, Bayliss R (2015) Structural insights into Separase architecture and substrate recognition through computational modelling of Caspase-like and death domains. PLoS Comput Biol 11(10):e1004548. https://doi.org/10.1371/journal.pcbi.1004548

Xie X, Liu Z, Xu C, Zhang Y (2017) A multiple sensors platform method for power line inspection based on a large unmanned helicopter. Sensors (Basel) 17(6). https://doi.org/10.3390/s17061222

Yalçın B, Zhao L, Stofanko M, O'Sullivan NC, Kang ZH, Roost A et al (2017) Modeling of axonal endoplasmic reticulum network by spastic paraplegia proteins. Elife 6. https://doi.org/10.7554/eLife.23882

Zheng J, Fessler JA, Chan H-P (2017) Detector blur and correlated noise modeling for digital breast tomosynthesis reconstruction. IEEE Trans Med Imaging. https://doi.org/10.1109/TMI.2017.2732824

Index

N. A. Shaik et al. (eds.), *Essentials of Bioinformatics, Volume I*,
https://doi.org/10.1007/978-3-030-02634-9