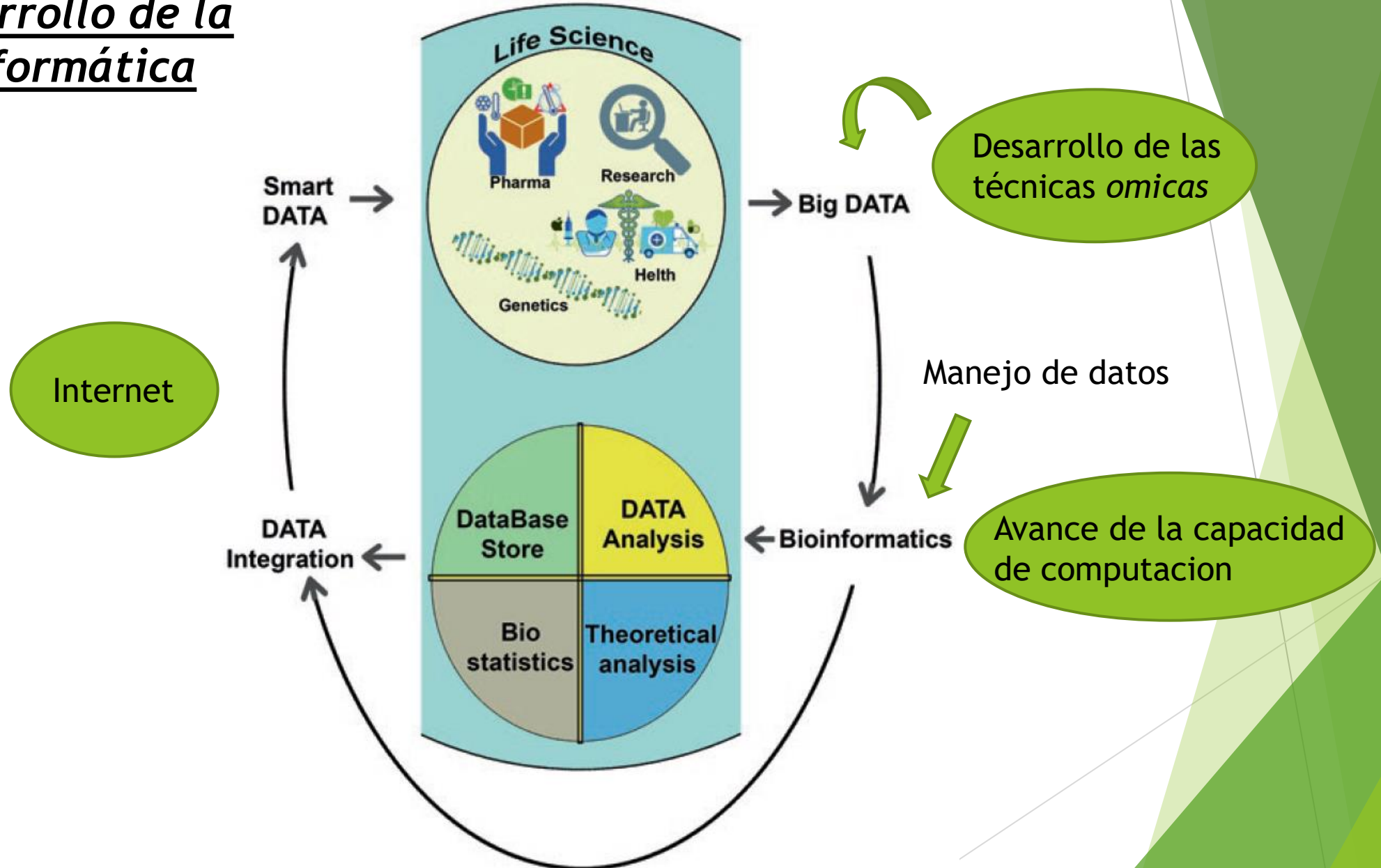


ELEMENTOS DE GENETICA VEGETAL EN LA PROTECCION DE CULTIVOS

Bioinformática - Bases de Datos y Servidores

Dr. Elias Mongiardini
IBBM - CCT La Plata - CONICET

El desarrollo de la Bioinformática



Bases de datos

Def. : Fuentes computalizadas donde la información esta guardada de manera estructurada lo que facilita su acceso

Clasificación de las bases de datos biológicas desde un punto de vista informatico

- Primarias: archivos que sirve como depósitos de los datos crudos (Genbank y Protein Data Bank)
- Secundarias: son bases que usan los datos de la bases de datos primarias para generar nuevos sub-set de datos (InterPro, Swiss-Prot o Ensembl)
- Compuestas o especializadas: combina varias bases de datos primarias de manera que se puedan hacer búsquedas simultaneas (NCBI)

Clasificación de las bases de datos biológicas en base al tipo de datos y funciones

- 1 - De secuencias
- 2 - De estructuras
- 3 - Funcionales

Bases de datos de secuencias de nucleótidos

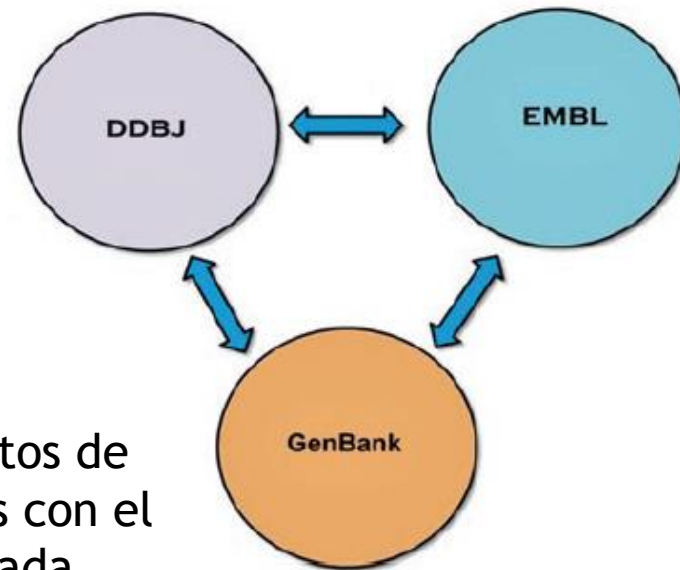
Tres bases mas importantes

EMBL-Bank -> mantenida por el EBI (European Bioinformatics Institute en Europa)

DDBJ -> mantenida por el NIG (National Institute of Genetics en Japón)

GenBank -> mantenida por el NCBI (National Center in Biotechnology en USA)

Las tres bases permiten el envío de nuevas secuencias



Otras bases imp.

RefSeq: es una base que incorpora a NCBI que toma los datos de GenBank y de bases de proteínas y proyectos genomas con el fin de hacer una anotación ordenada. Es una base curada

Ensembl: reúne varios genomas de vertebrados. Tiene un sistema propio de anotación de genomas aunque no hace el ensamblado del mismo. Provee datos de calidad curados

Bases de datos de secuencias de nucleótidos especializadas

<u>Name</u>	<u>Link</u>	<u>Description</u>
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	eugen.es.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

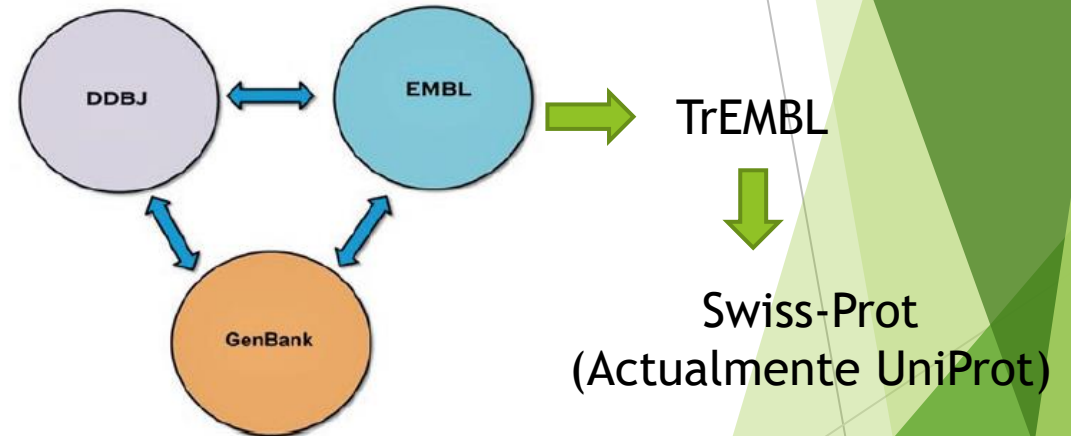
Bases de datos de secuencias de proteínas

TrEMBL: incluye todas las secuencias de DDBJ/EMBL/GenBank traducidas. Es automática
(Swiss-Prot: toma de TrEMBL las anotaciones y las cura manualmente)

GenPept -> es derivado de la anotación de GenBank

Entrez Protein -> es el servicio de anotación de NCBI

UniProt -> mantenida por NIH
(Combinación de Swiss-Prot, TrEMBL y PIR-PSD)





UniProtKB ▾


BLAST Align Retrieve/ID mapping Peptide search SPARQL

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and f

UniProtKB


UniProt Knowledgebase

Swiss-Prot (562,755)

 Manually annotated and reviewed.

Records with information extracted from literature and curator-evaluated computational analysis.

TrEMBL (184,998,855)

 Automatically annotated and not reviewed.

Records that await full manual annotation.

UniRef



The UniProt Reference Clusters (UniRef) provide clustered sets of sequences from the UniProt Knowledgebase (including isoforms) and selected UniParc records.

UniParc



UniParc is a comprehensive and non-redundant database that contains most of the publicly available protein sequences in the world.

Proteomes



A proteome is the set of proteins thought to be expressed by an organism. UniProt provides proteomes for species with completely sequenced genomes.

Supporting data

Literature citations



Cross-ref. databases

Taxonomy



Diseases

Subcellular locations



Keywords

Bases de datos de secuencias de proteínas especializadas

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

Bases de datos de estructuras 3D

En 1971, Brookhaven National Laboratory -> PDB

Diversas bases de datos de estructuras

Bases de datos primarias de estructuras

- RCSB PDB (<https://www.rcsb.org/>): Research Collaboratory for Structural Bioinformatics Protein Data Bank
- PDBe (<http://www.ebi.ac.uk/pdbe/>) del EBI
- PDBj (<https://pdbj.org/>) en Japón

Bases de datos de clasificación de proteínas

- CATH (<http://www.cathdb.info/>)
- SCOP (<http://scop2.mrc-lmb.cam.ac.uk/>)

Bases de Ácidos Nucléicos

- NDB (<http://ndbserver.rutgers.edu/>) -> ácidos nucleicos
- RNA FRABASE (<http://rnafrabase.cs.put.poznan.pl/>) -> fragmentos de RNA
- NPIDB (<http://npidb.belozersky.msu.ru/>) -> complejos ácidos nucleicos y proteínas

Bases de datos de proteínas de membrana

- MemProtMD (<http://sbcb.bioch.ox.ac.uk/memprotmd/>)

Bases de sitios activos, de unión de ligandos y metaloproteínas

- PeptiSite (<http://peptisite.ucsd.edu/>)
- ComSin (<http://antares.protres.ru/comsin/>)

Servidores para comparación de estructuras

- DALI (<http://ekhidna2.biocenter.helsinki.fi/dali/>)
- VAST+ (<https://structure.ncbi.nlm.nih.gov/Structure/VAST/vastsearch.html>)

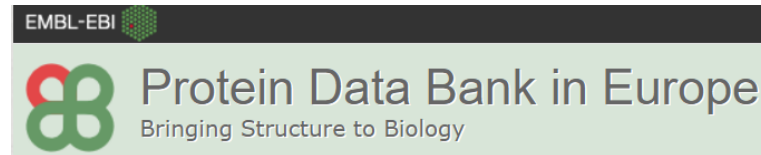
Otras Bases de datos

- PTM-SD (http://www.dsimb.inserm.fr/dsimb_tools/PTM-SD/) -> modificaciones post-traduccionales
- GFDB (<http://www.glycanstructure.org/>) -> restos glicosídicas y estructuras de carbohidratos
- ChEMBL (<https://www.ebi.ac.uk/chembl/>) -> moléculas pequeñas bioactivas

Bases de datos primarias de estructuras 3D

Datos de experimentos de: Difracción de rayos X
Resonancia Magnética Nuclear (NMR)
Cryo-EM

wwPDB (www.wwpdb.org) -> 167,132 anotaciones



Search History Browse Annotations MyPDB Help

QUERY: Full Text = ""covid-19"" Open In Query Builder MyPDB Login

Advanced Search Query Builder

- Attribute
- Sequence
- Sequence Motif
- Structure Similarity
- Chemical

Display Results as Structures Count Clear

Refinements Clear All

Summary Gallery Compact -- Tabular Report -- Score Download Selected Files Select All

- SCIENTIFIC NAME OF SOURCE ORGANISM Clear
- Severe acute respiratory syndrome coronavirus 2 (315)
 - Homo sapiens (45)
 - synthetic construct (13)
 - Lama glama (7)
 - Severe acute respiratory syndrome-related coronavirus (4)
 - Escherichia virus T4 (3)
 - Foot-and-mouth disease virus (2)
 - Gallus gallus (1)
 - Middle East respiratory syndrome-related coronavirus (1)
 - Mus musculus (1)
- More...

- TAXONOMY Clear
- Riboviria (319)
 - Eukaryota (52)

Displaying 1 to 25 of 324 Structures Page 1 of 13 Previous Next Display 25 per page

7BZ5 Download File View File

Structure of COVID-19 virus spike receptor-binding domain complexed with a neutralizing antibody

Wu, Y., Qi, J., Gao, F.

(2020) Science 368: 1274-1278

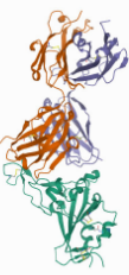
Released 2020-05-13

Method X-RAY DIFFRACTION 1.84 Å

Organisms Homo sapiens Severe acute respiratory syndrome coronavirus 2

Macromolecule Heavy chain of B38 (protein) Light chain of B38 (protein) Spike protein S1 (protein)

Unique Ligands NAG



3D View

Biological Assembly 1 ?

3D View: [Structure](#) | [Ligand Interaction](#)

Global Symmetry: Cyclic - C3 (3D View)

6VXX

Structure of the SARS-CoV-2 spike glycoprotein (closed state)

DOI: [10.2210/pdb6VXX/pdb](https://doi.org/10.2210/pdb6VXX/pdb) EMDDataResource: [EMD-21452](https://www.ebi.ac.uk/emdb/EMD-21452)

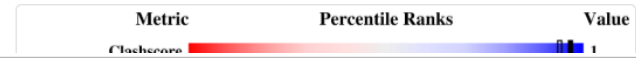
Classification: VIRAL PROTEIN
Organism(s): Severe acute respiratory syndrome coronavirus 2
Expression System: Homo sapiens
Mutation(s): No

Deposited: 2020-02-25 **Released:** 2020-03-11
Deposition Author(s): Walls, A.C., Park, Y.J., Tortorici, M.A., Wall, A., Seattle Structural Genomics Center for Infectious Disease (SSGICD), McGuire, A.T., Veesler, D.
Funding Organization(s): National Institutes of Health/National Institute of General Medical Sciences (NIH/NIGMS)

Experimental Data Snapshot

Method: ELECTRON MICROSCOPY
Resolution: 2.80 Å
Aggregation State: PARTIAL
Reconstruction Method: SINGLE PARTICLE

wwPDB Validation



Display Files **Download Files**

- FASTA Sequence
- PDB Format
- PDB Format (gz)
- PDBx/mmCIF Format
- PDBx/mmCIF Format (gz)
- PDBML/XML Format (gz)
- Biological Assembly 1
- Download EM Map

Sequence of 6VXX | Structure of 6VXX

Literature Download Primary Citation

Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein.
[Walls, A.C., Park, Y.J., Tortorici, M.A., Wall, A., McGuire, A.T., Veesler, D.](#)
 (2020) Cell **181**: 281

PubMed: [32155444](https://pubmed.ncbi.nlm.nih.gov/32155444/) [Search on PubMed](#)
 DOI: [10.1016/j.cell.2020.02.058](https://doi.org/10.1016/j.cell.2020.02.058)
[Structures With Same Primary Citation](#)

PubMed Abstract:
 The emergence of SARS-CoV-2 has resulted in >90,000 infections and >3,000 deaths. Coronavirus spike (S) glycoproteins promote entry into cells and are the main target of antibodies. We show that SARS-CoV-2 S uses ACE2 to enter cells and that the rece ...

The screenshot displays the PDBsum website interface. At the top, the EMBL-EBI logo is on the left, and navigation links for 'Services', 'Research', 'Training', and 'About us' are on the right. The main header features the 'PDBsum' logo and the tagline 'Pictorial database of 3D structures in the Protein Data Bank'. Below this, a breadcrumb trail reads 'Databases > Structure Databases > PDBsum'. The main content area contains a descriptive paragraph about PDBsum and three search methods: 'PDB code' (with an input field and 'Find' button), 'Text search' (with an input field and 'Search' button), and 'Search by sequence' (with a large text area and 'Search' button). A fourth search method is provided below, allowing searches by 'UniProt id', 'Pfam id', and 'Ensembl id', each with its own input field and 'Search' button. On the left side, a vertical menu lists 'Browse options' (List of PDB codes, Het Groups, Ligands, Drugs, Enzymes) and 'Generate' (Figures from Papers, Gallery, Figure stats), along with 'Documentation', 'Downloads', and 'Contact us'. At the bottom left of the menu area is a 3D protein structure visualization. On the right side, a 'Contents' box states 'PDBsum contains 171,343 entries, including 2,067 superseded. Last update: 22 July, 2020'. Below it is an 'In-house version' section with a 'PDBsum Proprietary' logo and a description. The 'Related databases' section includes 'EC-PDB' and 'DrugPort', each with a dropdown arrow.

Funciona a modo de atlas que compila las anotaciones de todas las bases y facilita la búsqueda entre todas la bases

Bases de datos de clasificación de proteínas

CATH database (Class, Architecture, Topology, Homology)

clasifica los dominios en 4 niveles de jerarquía

- **C** level: de acuerdo a estructura secundaria
- **A** level: orientación de estructura secundaria
- **T** level: relación entre estructuras secundarias
- **H** level: combinación de similitud de secuencia y estructura

Dentro de CATH se encuentra la **CATH/Gene3D database** que es complementaria

Utiliza las secuencias depositadas en UniProt y la PDB para clasificar las proteínas en familias

Hay 95 millones de dominios de proteínas clasificados en 6119 superfamilias

SCOP database

Base de datos enfocada en estructura y evolución de proteínas

Servidores para comparación de estructuras

Estos servidores tratan de encontrar proteínas con estructuras 3D similares sin basarse en la secuencia lineal de aa

Estos servidores

- ayudan en la clasificación de proteínas basadas en el folding
- colaboran en el proceso de identificación de función basada en estructura
- aportan en los métodos de modelado por homología

Los dos más importantes son:

VAST+ -> de NCBI (no busca por comparación de secuencia sino por similitud 3D por lo tanto tiene utilidad en los casos de baja homología).

DALI web server -> Helsinki Lab. Esta basado en clasificar las estructuras de la PDB basado en la comparación de sus estructuras.

Ambos se pueden acceder a partir de códigos PDB y proveen información de estructuras similares a la que se está buscando

Bases de datos funcionales

GO/GOA databases: gene ontology annotation -> creada para unificar y organizar los datos referidos a anotación de proteínas

PRIDE Archive -> depositorio de espectros de MS de identificación de proteínas

Swiss 2d Prot -> repositorio de experimentos de geles 2D

Network Databases -> bases de datos basadas en modelos de interacción de proteínas

- IntAct -> reúne datos experimentales de interacción de proteínas y facilita su búsqueda

Bases de datos de vías metabólica

- KEGG (Kyoto Encyclopedia of Genes and Genomes -> contiene vías metabólicas curadas manualmente

Bases de datos de drogas

- DrugBank -> la ultima versión contiene 11,177 compuestos que incluyen moléculas pequeñas, péptidos y otros.

Bases de datos de organismos modelos

- Saccharomyces Genome Database (yeastgenome.org)
- ZFIN (zfin.org) -> Zebrafish
- TAIR (arabidopsis.org) -> Arabidopsis
- Rat Genome Database (rgd.mcw.edu) -> rata
- Mouse Genome Database (informatics.jax.org) -> ratón
- FlyBase (flybase.org) -> drosophila

Bases de datos de literatura

- PubMed -> NCBI
- Google Scholar



Bases de datos complejas o combinadas -> NCBI

The image shows a screenshot of the NCBI (National Center for Biotechnology Information) website. A dropdown menu is open from the 'All Databases' link in the top navigation bar. The menu lists the following databases: All Databases, Assembly, Biocollections, BioProject, BioSample, BioSystems, Books, ClinVar, Conserved Domains, dbGaP, dbVar, Gene, Genome, GEO DataSets, GEO Profiles, GTR, HomoloGene, Identical Protein Groups, MedGen, and MeSH. The right side of the dropdown menu is partially obscured by another list of links: NCBI Web Site, NLM Catalog, Nucleotide, OMIM, PMC, PopSet, Protein, Protein Clusters, PubChem BioAssay, PubChem Compound, PubChem Substance, PubMed, SNP, Sparcle, SRA, Structure, Taxonomy, ToolKit, ToolKitAll, and ToolKitBookgh. The main content area of the website includes a 'Welcome to NCBI' message, a 'Submit' button for depositing data, and a 'Download' button for transferring data. A sidebar on the left contains a 'Resource List (A-Z)' with links to various categories like All Resources, Chemicals & Bioassays, Data & Software, etc.

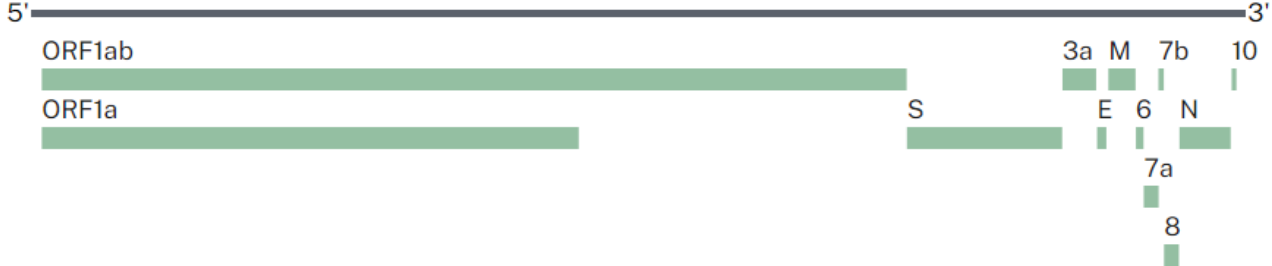
Una búsqueda en NCBI en todas las bases

Search NCBI x Search

Results found in 26 databases

REFERENCE GENOME SEQUENCE Was this helpful?  


Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) reference genome
Severe acute respiratory syndrome coronavirus 2 (Host: human,vertebrates)
ssRNA(+)
RefSeq: NC_045512.2
[RefSeq genome \(1\)](#) [RefSeq Proteins \(38\)](#) [NCBI SARS-CoV-2 resources](#)




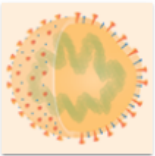
5' ————— 3'

ORF1ab
ORF1a
S
3a M 7b 10
E 6 N
7a
8

[Download](#)

BLAST
Use our new Betacoronavirus database for SARS-CoV-2 genome sequence analysis 

NCBI Virus
The most up-to-date set of SARS-CoV-2 nucleotide and protein sequences 

LitCovid
A curated literature hub for the latest scientific information on COVID-19 

Literature	
Bookshelf	123
MeSH	36
NLM Catalog	35
PubMed	36,505
PubMed Central	37,526

Genes	
Gene	55
GEO DataSets	130
GEO Profiles	0
HomoloGene	0
PopSet	66

Proteins	
Conserved Domains	17
Identical Protein Groups	14,935
Protein	137,271
Protein Clusters	0
Sparcle	0
Structure	300

Genomes	
Assembly	92
BioCollections	0
BioProject	158
BioSample	38,561
Genome	1
Nucleotide	13,459
SRA	36,350
Taxonomy	1

Genetics	
ClinVar	0
dbGaP	0
dbSNP	0
dbVar	3,387
GTR	8
MedGen	1
OMIM	0

PubChem	
BioAssays	275
Compounds	1,032
Pathways	1,876
Substances	50



Items: 1 to 20 of 137271

<< First < Prev Page 1 of 686

 [Chain A, Spike protein S1](#)

1. 229 aa protein

Accession: 7BZ5_A GI: 1841213709

[PubMed](#) [Taxonomy](#)[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#) [Chain A, main protease](#)

2. 306 aa protein

Accession: 6LU7_A GI: 1806061810

[PubMed](#) [Taxonomy](#)[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#) [Chain B, SARS-Cov-2 NSP 8](#)

3. 198 aa protein

Accession: 6M71_B GI: 1827515550

[PubMed](#) [Taxonomy](#)[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#) [Chain D, SARS-Cov-2 NSP 8](#)

4. 198 aa protein

Accession: 6M71_D GI: 1827515549

[PubMed](#) [Taxonomy](#)[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#) [Chain C, SARS-Cov-2 NSP 7](#)

Chain A, Spike protein S1

PDB: 7BZ5_A

[Identical Proteins](#) [FASTA](#) [Graphics](#)Go to:

LOCUS 7BZ5_A 229 aa linear VRL 24-JUN-2020

DEFINITION Chain A, Spike protein S1.

ACCESSION 7BZ5_A

VERSION 7BZ5_A

DBSOURCE pdb: molecule 7BZ5, chain 65, release Jun 24, 2020;
deposition: Apr 26, 2020;
class: VIRAL PROTEIN/IMMUNE SYSTEM;
source: Mmdb_id: [187670](#), Pdb_id 1: 7BZ5;
Exp. method: X-ray Diffraction.

KEYWORDS .

SOURCE Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)

ORGANISM [Severe acute respiratory syndrome coronavirus 2](#)
Viruses; Riboviria; Orthornavirae; Pisuviricota; Pisoniviricetes;
Nidovirales; Coronidovirineae; Coronaviridae; Orthocoronavirinae;
Betacoronavirus; Sarbecovirus.

REFERENCE 1 (residues 1 to 229)

AUTHORS Wu,Y., Wang,F., Shen,C., Peng,W., Li,D., Zhao,C., Li,Z., Li,S.,
Bi,Y., Yang,Y., Gong,Y., Xiao,H., Fan,Z., Tan,S., Wu,G., Tan,W.,
Lu,X., Fan,C., Wang,Q., Liu,Y., Zhang,C., Qi,J., Gao,G.F., Gao,F.
and Liu,L.

TITLE A noncompeting pair of human neutralizing antibodies block COVID-19
virus binding to its receptor ACE2

JOURNAL Science 368 (6496), 1274-1278 (2020)

PUBMED [32404477](#)

REFERENCE 2 (residues 1 to 229)

AUTHORS Wu,Y., Qi,J. and Gao,F.

Herramientas y métodos en el manejo de secuencias

Alineamiento de secuencias

- 2 tipos: - Globales
- Locales

Blast: es uno de los algoritmos mas utilizados para hacer alineamientos y búsquedas de secuencias por similitud

Servidor de NCBI con el Blast tools

Basic Local Alignment Search Tool

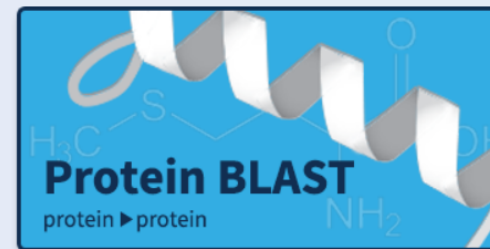
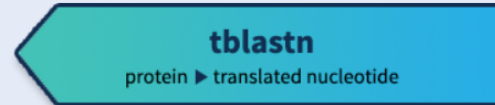
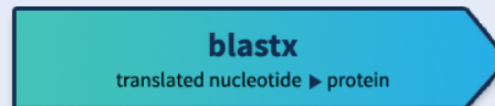
BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

NEWS
BLAST+ 2.10.1 is released - Fix for TBLASTN Multi-Threading issue.
This version supports pulling databases from our FTP site as well from [cloud providers](#) or our [BLAST+Docker solution](#).

Thu, 18 June 2020 12:00:00 EST

[More BLAST news...](#)

Web BLAST



BlastP

blastn **blastp** blastx tblastn tblastx

BLASTP programs search protein databases using a protein query. [more...](#)

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#) [Query subrange](#)

Secuencia de proteína

From

To

Or, upload file Ningún archivo seleccionado [+](#)

Job Title

Enter a descriptive title for your BLAST search [?](#)

Align two or more sequences [?](#)

Choose Search Set

Database [+](#)

Organism exclude [+](#)

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown. [?](#)

Exclude Models (XM/XP) Non-redundant RefSeq proteins (WP) Uncultured/environmental sample sequences [?](#)

Program Selection

Algorithm

- Quick BLASTP (Accelerated protein-protein BLAST)
- blastp (protein-protein BLAST)
- PSI-BLAST (Position-Specific Iterated BLAST)
- PHI-BLAST (Pattern Hit Initiated BLAST)
- DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm [?](#)

- Non-redundant protein sequences (nr)
- Reference proteins (refseq_protein)
- Model Organisms (landmark)
- UniProtKB/Swiss-Prot (swissprot)
- Patented protein sequences (pataa)
- Protein Data Bank proteins (pdb)
- Metagenomic proteins (env_nr)
- Transcriptome Shotgun Assembly proteins (tsa_nr)

Limitar o excluir un organismo de la búsqueda

Resultado del BlastP

Descriptions		Graphic Summary	Alignments	Taxonomy			
Sequences producing significant alignments				Download	Manage Columns	Show 100	?
<input checked="" type="checkbox"/> select all 100 sequences selected				GenPept	Graphics	Distance tree of results	Multiple alignment
	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
<input checked="" type="checkbox"/>	SARS_CoV_2RBD_his [synthetic construct]	478	478	100%	2e-170	100.00%	QJE37811.1
<input checked="" type="checkbox"/>	Chain E, Spike receptor binding domain [Severe acute respiratory syndrome coronavirus 2]	477	477	100%	4e-170	100.00%	6M0J_E
<input checked="" type="checkbox"/>	Chain C, Spike protein S1 [Severe acute respiratory syndrome coronavirus 2]	472	472	100%	5e-168	99.13%	6W41_C
<input checked="" type="checkbox"/>	Chain E, Spike glycoprotein [Severe acute respiratory syndrome coronavirus 2]	466	466	100%	1e-165	91.24%	6XDG_E
<input checked="" type="checkbox"/>	Chain E, Spike protein S1 [Severe acute respiratory syndrome coronavirus 2]	462	462	98%	1e-163	98.67%	6XE1_E
<input checked="" type="checkbox"/>	Chain E, SARS-coV-2 Receptor Binding Domain [Severe acute respiratory syndrome coronavirus 2]	460	460	97%	1e-163	100.00%	6M17_E
<input checked="" type="checkbox"/>	surface glycoprotein, partial [Severe acute respiratory syndrome coronavirus 2]	465	465	97%	2e-158	99.55%	QKK14051.1
<input checked="" type="checkbox"/>	Chain E, SARS-CoV-2 receptor binding domain [Severe acute respiratory syndrome coronavirus 2]	444	444	100%	2e-157	94.76%	7BJW_E
<input checked="" type="checkbox"/>	surface glycoprotein, partial [Severe acute respiratory syndrome coronavirus 2]	469	469	97%	2e-155	99.55%	QJD23474.1
<input checked="" type="checkbox"/>	surface glycoprotein, partial [Severe acute respiratory syndrome coronavirus 2]	469	469	97%	2e-155	99.55%	QJD25214.1
<input checked="" type="checkbox"/>	surface glycoprotein, partial [Severe acute respiratory syndrome coronavirus 2]	463	463	96%	6e-155	99.55%	QKV38614.1
<input checked="" type="checkbox"/>	Chain A, SARS-CoV-2 Spike glycoprotein [Severe acute respiratory syndrome coronavirus 2]	469	469	97%	2e-154	99.55%	7BYR_A
<input checked="" type="checkbox"/>	surface glycoprotein, partial [Severe acute respiratory syndrome coronavirus 2]	469	469	97%	2e-154	99.55%	QLL35949.1
<input checked="" type="checkbox"/>	surface glycoprotein, partial [Severe acute respiratory syndrome coronavirus 2]	469	469	97%	3e-154	99.55%	QMI96595.1
<input checked="" type="checkbox"/>	surface glycoprotein, partial [Severe acute respiratory syndrome coronavirus 2]	469	469	97%	3e-154	99.55%	QMI90362.1

Descriptions

Graphic Summary

Alignments

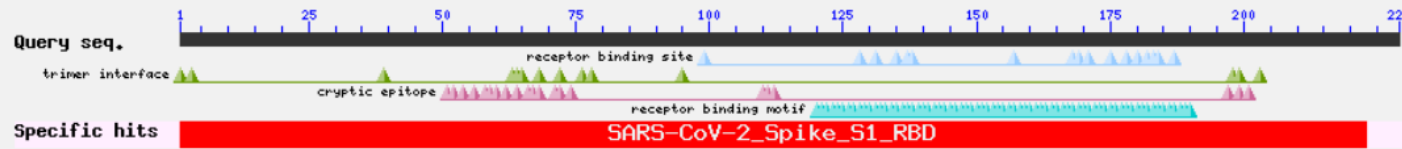
Taxonomy

hover to see the title click to show alignments Show Conserved Domains

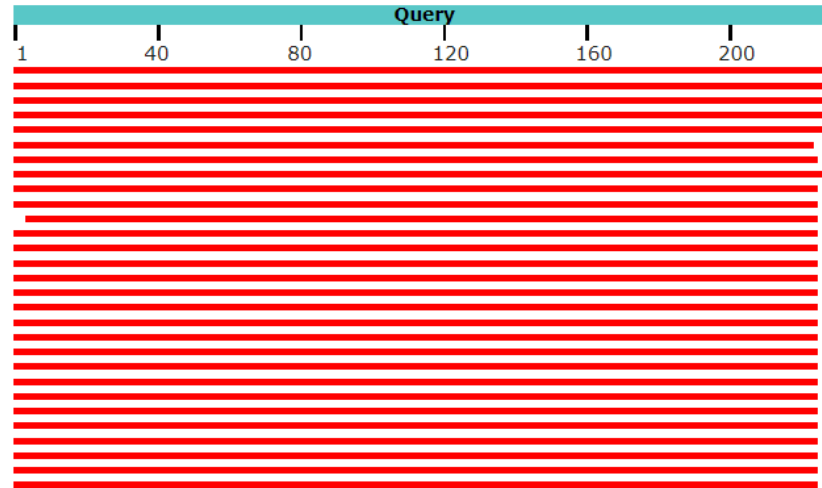
Alignment Scores < 40 40 - 50 50 - 80 80 - 200 >= 200

100 sequences selected

Putative conserved domains have been detected, click on the image below for detailed results.



Distribution of the top 100 Blast Hits on 100 subject sequences



Descriptions

Graphic Summary

Alignments

Taxonomy

Alignment view

[Restore defaults](#)

[Download](#)

100 sequences selected

[Download](#)

[GenPept](#) [Graphics](#)

[Next](#) [Previous](#) [Descriptions](#)

SARS_CoV_2RBD_his [synthetic construct]

Sequence ID: **QJE37811.1** Length: 243 Number of Matches: 1

Range 1: 15 to 243 [GenPept](#) [Graphics](#)

[Next Match](#) [Previous Match](#)

Score	Expect	Method	Identities	Positives	Gaps
478 bits(1230)	2e-170	Compositional matrix adjust.	229/229(100%)	229/229(100%)	0/229(0%)
Query 1		RVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFK			60
Sbjct 15		RVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFK			74
Query 61		CYGVSP TKLNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNS			120
Sbjct 75		CYGVSP TKLNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNS			134
Query 121		NNLDSKVGGNLYRLFRKSNLKPFRDISTEIQAGSTPCNGVEGFNCYFPLQSYGFQ			180
Sbjct 135		NNLDSKVGGNLYRLFRKSNLKPFRDISTEIQAGSTPCNGVEGFNCYFPLQSYGFQ			194
Query 181		PTNGVGYQPYRWWLSFELLHAPATVCGPKKSTNLVKNKCVNFHHHHHH		229	
Sbjct 195		PTNGVGYQPYRWWLSFELLHAPATVCGPKKSTNLVKNKCVNFHHHHHH		243	

Alineamientos multiples

Métodos progresivo -> clustalW

Método iterativo -> MultiAlin

Alineamiento por Profile -> PSI-BLAST



Análisis filogenético

ClustalW se puede ejecutar online en distintos servidores -> EBI

tb. se puede ejecutar en el paquete MEGA (software free)

PSI-BLAST -> NCBI

Servidores de bioinformática estructural

Predicción de estructuras 3D

3 métodos:

- Modelado por homología requiere de un homólogo resuelto en la PDB
- Por reconocimiento de plegado: requiere de la presencia de estructuras secundaria similares resueltas
- De novo (ab inicio): solo se utiliza la información de la estructura primaria

Modelado por homología

Supone que secuencias de aa similares se van a plegar de la misma manera

Como regla se utiliza que al menos debería haber entre un 30%-50% de identidad para utilizar este método

Flujo de trabajo en modelado por homología

Sequence

Database Search

Identificación de un molde o templado -> BLAST, PSI-BLAST, método de reconocimiento de plegado o una búsqueda en la PDB.

Alineamiento de las dos proteínas (puede ser multiple) -> ClustalW

Sequence Alignment

Structural Alignment

Curado manual del alineamiento (se pueden utilizar alineamientos estructurales (MUSCLE o T-COFFEE))

Structural Based Multiple Seq Alignment

- Generación del modelo -> *modeller* (<https://salilab.org/modeller/>)

- Modelado de loops

- Optimización de cadenas secundarias (Ramachandran plots)

Homology Modelling

Validación: hay diversos servidores -> modeller calcula un índice DOPE que hace referencia a la calidad del modelo (energético)

- PROCHECK

- Swiss Model Validation Service

- Calculo del RMSD (root-mean-square deviation)

3D Protein Structure

Modelado por reconocimiento de plegamiento

Se utiliza cuando el % de homología esta dentro de los límites de aplicabilidad del homolgy modelling

Se basa en la búsqueda de plegamientos similares en la PDB a partir de métodos estadísticos

Hay varios servidores:

- I-TASSER
- Hhpred
- Phyre2

Ab initio

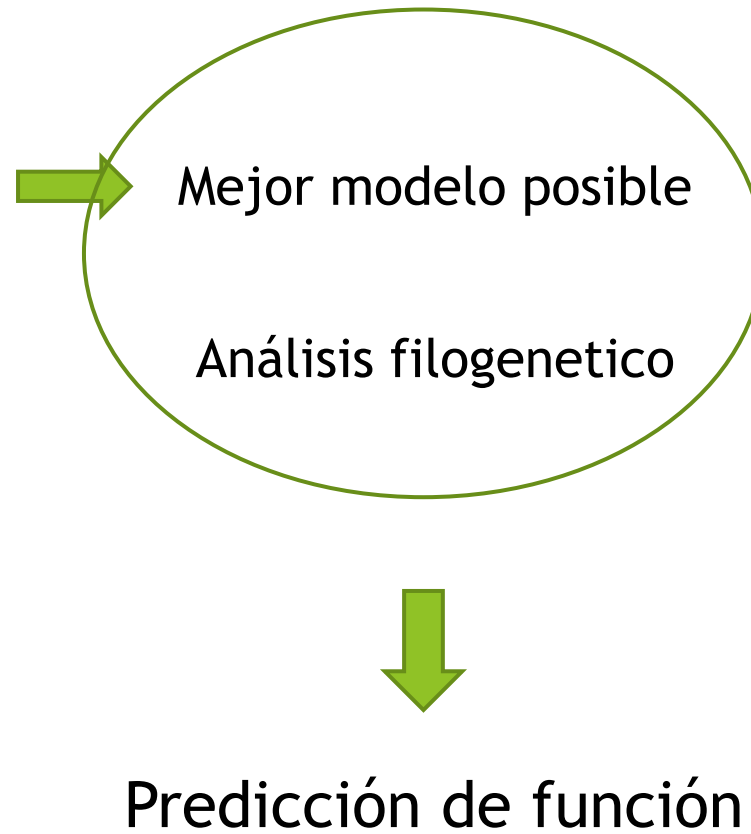
Solo se utiliza la secuencia de aa de la proteína como información

Se basa en predecir la estructura nativa a partir de encontrar la conformación de energía mas favorable

Hay varios servidores:

- QUARK
- ROSETTA

En general es conveniente utilizar todos los 3 métodos y evaluar los resultados que se encuentran con cada uno de ellos



Nucleic Acids Research, 2020, Vol. 48, Database issue **D1–D8**
doi: 10.1093/nar/gkz1161

The 27th annual Nucleic Acids Research database issue and molecular biology database collection

Daniel J. Rigden^{1,*} and Xosé M. Fernández²

¹Institute of Integrative Biology, University of Liverpool, Crown Street, Liverpool L69 7ZB, UK and ²Institut Curie, 25 rue d'Ulm, 75005 Paris, France

The NAR online

Molecular Biology Database Collection has been revised, updating 305 entries, adding 65 new resources and eliminating 125 discontinued URLs; so bringing the current total to **1637 databases**

<http://www.oxfordjournals.org/nar/database/c/>.

THANK YOU

GRACIAS
ARIGATO
SHUKURIA

THANKS
DANKSCHEEN
TASHAKKUR ATU
GRAZIE
MEHRBANI
PALSIES

YOU
BOLZIN
MERCERCI

BIYAN
SHUKRIA

JUSPAXAR
TAVTAPUCH
MEDAWAGSE

GOZAIMASHITA
EFCHARISTO

KOMAPSUMNIDA
MERASTAWHY
GAEJTHO
FAKAABE

MAAKE
LAH

YAQHANYELAY
CHALTU
SNACHALHUVA
MUHUN
SNACHALHUVA
YUSPAGADATAM

SUKSAMA
EKHMET
DENKAUJA
MENACHALHYA

TINGKI
HUI
GUR
HATUR
UNALCHEEN
EKOJU
SIKOMO
MIMMONCHAR

SPASSIBO
DANKSCHEEN

BAINA
JUSPAXAR

YUSPAGADATAM
YUSPAGADATAM

UNALCHEEN
UNALCHEEN

SIKOMO
SIKOMO

MIMMONCHAR
MIMMONCHAR