Trabajo Práctico de Bioinformática

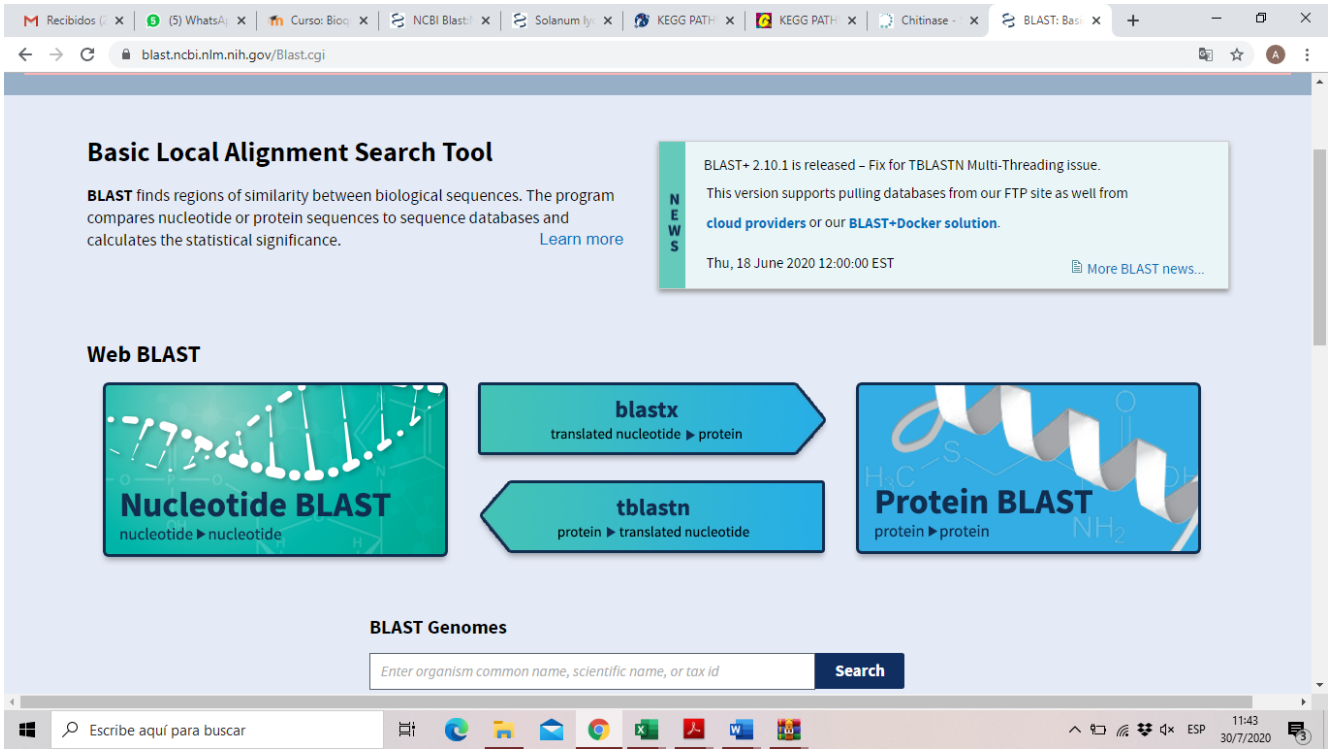
La búsqueda y caracterización de genes y sus productos en bases de datos es una herramienta fundamental de la Biología Molecular y sus aplicaciones. Existen multitud de bases de datos y programas en internet que permiten identificar genes a partir de una secuencia de ADN o bien encontrar secuencias de ADN de genes de interés. Una vez halladas estas secuencias, puede estudiarse la estructura génica, el transcripto producido y el producto final, ya sea éste un polipéptido o un ARN. En este trabajo práctico practicaremos estos aspectos desde los dos puntos de partida mencionados anteriormente.

1) Cómo identificar y caracterizar un gen a partir de su secuencia de ADN

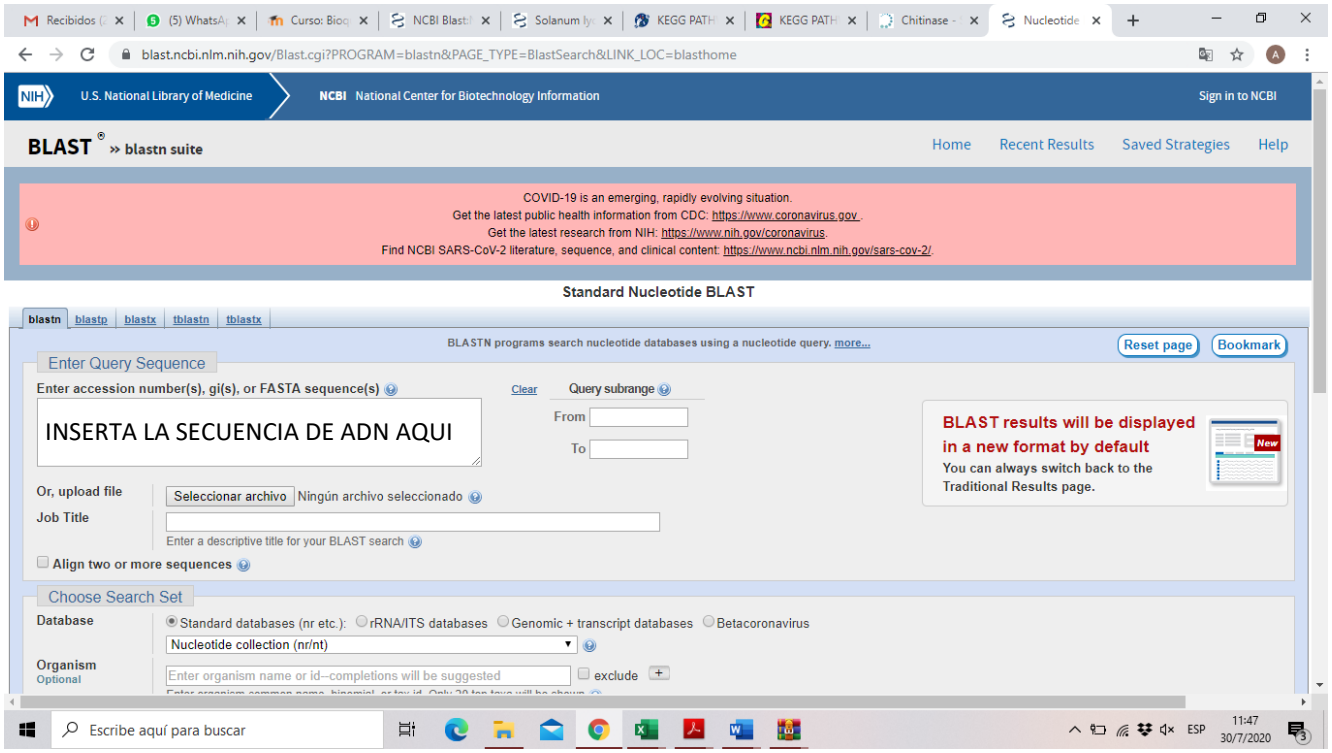
Entre las sustancias que permiten a las plantas resistir enfermedades se encuentran los fenilpropanoides. Estas sustancias, entre las que se encuentran el ácido cinámico, el ácido cumárico, la cumarina y el cinamaldehído, se sintetizan a partir de ciertos aminoácidos, como la fenilalanina. En un estudio de plantas de maíz sensibles a ciertas enfermedades se encontró que **algunas eran incapaces de sintetizar ácido cumárico**. Mediante estudios de biología molecular se encontró que poseían un defecto en una secuencia de ADN, que en las plantas resistentes era la siguiente:

```
   1 GCACCATCCA GTGCATCACG AGCTCTTCTG CACCAGATTA GCAGGCCATC GCCTACTTTT
  61 GGCTTCCAAA TCATTTATTT ACGGCGTACG TGCCTTCTGT TCAAACCCCA GCCCCGCTGC
 121 AATGGAGTGC GACAACGGCC GCGTCGCTGC TACCAACGGC GACTCCCTGT GCATGGCGCT
 181 GCCCCGCGCC GCCGACCCGC TTAACTGGGG GAAGGCGGCG GAGGAGATGA TGGGCAGCCA
 241 CCTCGACGAG GTGAAGCGGA TGGTGGCCGA GTACCGCCAG CCCCTGGTGA AGATCGAGGG
 301 CGCCAGCCTC CGCATCGCGC AGGTGGCCGC TGTCGCCGCC GGCGCGGGCG AGGCCCGGGT
 361 CGAGCTCGAC GAGTCCGCCC GCGGCCGGGT CAAGGCGAGC AGCGACTGGG TCAGGGACAG
 421 CATGATGAAC GGCACCGACA GCTACGGCGT CACCACCGGC TTCGGCGCCA CCTCCCACCG
 481 CCGCACCAAG GAGGGCGGCG CTCTCCAGAG GGAGCTCATC AGGTTCCTCA ACGCCGGCGC
 541 CTTTGGCATC GGCACCGACG CCGGCCACGT CCTGCCGGCC GAGGCCACGC GCGCGGCCAT
 601 GCTCGTCCGC ATCAACACCC TCCTCCAGGG CTACTCCGGT ATCCGCTTCG AGATCCTCGA
 661 GGCCATCGTC AAGCTGCTCA ATGCCAACGT CACGCCGTGC CTGCCGCTGC GCGGCACGGT
 721 CACCGCGTCC GGCGACCTCG TGCCGCTCTC CTACATTGCT GGCCTCGTCA CCGGGCGCGA
 781 GAACGCCGTT GCGGTGGCTC CCGATGGCAC CAAGGTGAAC GCCGCGGAGG CGTTCAGGAT
 841 CGCCGACATC CAAAGCGGCT TCTTCGAGCT GCAGCCCAAG GAAGGTCTCG CCATGGTGAA
 901 CGGCACTGCC GTGGGCTCCG GCCTTGCCTC CACGGTGCTC TTTGAGGCGA ACGTACTTGC
 961 CGTCCTTGCC GAGGTCCTGT CCGCCGTGTT CTGCGAGGTC ATGAACGGCA AGCCGGAGTA
1021 CACCGACCAC CTGACCCACA AGCTGAAGCA CCACCCAGGA CAGATCGAGG CGGCTGCCAT
1081 CATGGAGCAC ATCTTGGAAG GCAGTTCCTA CATGAAGCTT GCTAAGAAGC TCGGTGAGCT
1141 CGACCCGTTG ATGAAGCCCA AGCAGGACAG GTACGCGCTC CGTACGTCGC CGCAGTGGCT
1201 CGGCCCGCAG ATTGAGGTTA TCCGTGCCTC CACCAAGTCC ATTGAGCGCG AGATCAACTC
1261 CGTCAACGAC AACCCGCTCA TCGACGTCGC CCGAAGCAAG GCCCTTCACG GTGGCAACTT
1321 CCAGGGCACG CCCATCGGGG TGTCCATGGA CAACACCCGT CTCGCCGTCG CAGCCATCGG
1381 CAAGCTCATG TTTGCGCAGT TCTCTGAGCT CGTCAACGAC TACTACAACA ACGGCTTGCC
1441 CTCCAACCTG TCCGGCGGGC GCAACCCCAG CTTGGACTAC GGCTTCAAGG GCGCCGAGAT
1501 CGCCATGGCG TCCTACTGCT CTGAGCTGCA GTTCCTGGGG AACCCGGTCA CCAACCACGT
1561 CCAGAGCGCG GAGCAGCACA ACCAGGACGT GAACTCGCTC GGACTCATCT CCTCCAGGAA
1621 GACTGCTGAG GCCATCGAGA TCCTCAAGCT CATGTCCTCC ACGTTCCTGA TCGCCCTGTG
1681 CCAGGCGGTG GACCTGCGCC ACATCGAGGA GAACGTCAAG AGCGCCGTCA AGAGCTGCGT
1741 GATGACGGTG GCCAAGAAGA CTCTGAGCAC CAACTCCACC GGTGGCCTCC ACGTCGCCCG
1801 CTTCTGCGAG AAGGACCTGC TCCAGGAGAT CGAGCGCGAG GCGGTGTTCG CGTATGCTGA
1861 CGACCCCTGC AGCGCTAACT ACCCGCTGAT GAAGAAGCTT CGCAACGTGC TCGTGGAGCG
1921 CGCCCTCGCC AATGGCACCG CCGAGTTCGA CGCCGAGACA TCCGTGTTCG CTAAGGTCGC
1981 CCAGTTCGAG GAGGAGCTGC GCACGGCGCT GCCCAGTGCG GTGGAGGCTG CACGGGCGGC
2041 TGTCGAAAAC GGCACGGCAG CGATACCGAA CAGAATCGCC GAGTGCCGCT CCTACCCGCT
2101 CTACCGCTTC GTGCGCGAGG AGCTCGGAGC AGTGTACCTC ACCGGCGAGA AGACGCGCTC
2161 TCCCGGCGAG GAGCTTAACA AGGTGCTCGT TGCCATCAAC CAGGGCAAGC ACATCGACCC
2221 GCTGCTCGAG TGCCTCAAGG AGTGGAACGG CGAGCCCCTG CCCATCTGCT GAACAGAGAA
2281 AATACAAGGA GCAGAAGACT GTATTTTTTA GCTAATACGC ACTTTTTATT CCTAATTTAT
2341 TTATGTTCGT AAGTTCGTTG ATATGCTACG CAATCTTGTT ATTCAGCTGC AACGCCAACC
2401 TGCTTTGCTT TGAGCAAGGT CTGGTGAGTG ATTGAAAAAA ATGTTGTTGC AAGCTGTACC
2461 TTGTATGTTT TTCAACAGGT GAATCTCACG TTTGATGCAT TGGATCAGAC
```

Copiá esta secuencia (así como está, con números y todo) e introducila en el programa BLAST, que se encuentra en: https://blast.ncbi.nlm.nih.gov/Blast.cgi. BLAST es una herramienta para realizar alineamientos de secuencia y así identificar una secuencia desconocida. Su nombre significa "Herramienta Básica para la Búsqueda de Alineamientos Locales", en inglés: "Basic Local Alignment Search Tool". Al entrar, vas a ver esta pantalla:



Las distintas opciones te permiten investigar una secuencia de ADN (Nucleotide BLAST, a la izquierda), de proteína (Protein BLAST, a la derecha) o traducir una en otra (opciones del centro). En nuestro caso, vamos entrar en Nucleotide Blast y vamos a pegar la secuencia de ADN en la ventana que dice "Enter Query Sequence"



Una vez hecho eso, apretá el botón que dice BLAST abajo de todo a la izquierda y esperá unos minutos por el resultado. Recordá que es un gen de maíz, y que estas bases de datos están en inglés.

El resultado te va a aparecer como una lista de todas las secuencias de ADN que hay en esa base de datos con un cierto nivel de parecido con la que vos introdujiste. Tu secuencia siempre es la que llaman “query” y las secuencias parecidas encontradas en la base de datos se llaman “subject” (abreviado “Sbjct”). En las columnas a la derecha te dan dos porcentajes importantes: la que dice “Per. Ident.” te da el porcentaje de identidad, es decir cuántos de los nucleótidos de tu secuencia (query) están presentes en la misma posición en cada una de las secuencias subject; la otra columna importante es la que dice “Query Cover”, que te indica si el parecido con la secuencia subject es a lo largo de toda la secuencia (100%) o solamente en una parte de la secuencia (<100%).

Mirá esa lista y clickeá en la que aparece arriba de todo, es decir, la más parecida y con más cobertura. Una vez que hayas entrado allí, vas a ver el alineamiento y una leyenda que dice “Sequence ID” seguida de un número en azul, que es un link. Entrá en ese link y vas a encontrar las características del gen identificado en esa base de datos, entre ellas: “LOCUS”: el número que acabás de picar; “ORGANISM”: la especie de la cual se obtuvo esa secuencia; “AUTHORS, TITLE, JOURNAL”: la publicación donde se describió ese gen; “PUBMED”: el link a la publicación (puede haber más de una, en cuyo caso estos descriptores se van a repetir) y “FEATURES”, donde vas a encontrar aspectos que se resaltarán en la secuencia de ADN o la de la proteína traducida que se reproducen abajo de todo. “Gene” te marcará el gen completo; “Exon” te marcará los exones que pueda haber en ese gen; “CDS” te marcará la secuencia codificante (coding sequence), es decir la que contiene los codones que codifican los aminoácidos de la proteína. Revisá estos aspectos y responde lo siguiente:

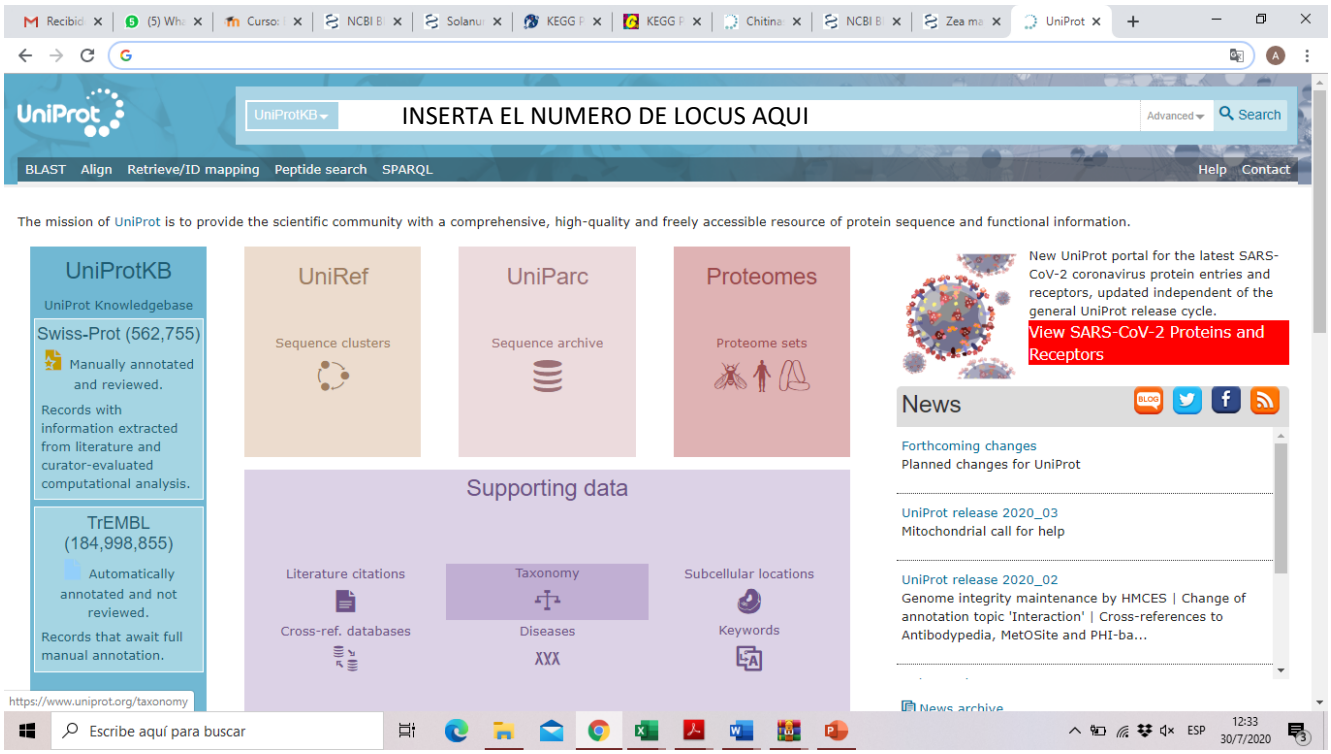
a) ¿Qué significa el alineamiento? ¿Cómo se ve si clickeás en proteínas con menor porcentaje de identidad?

b) ¿Qué proteína es?

c) ¿Por qué la CDS es más chica que la región abarcada por el exón?

d) ¿Dónde están los codones de iniciación y de terminación?

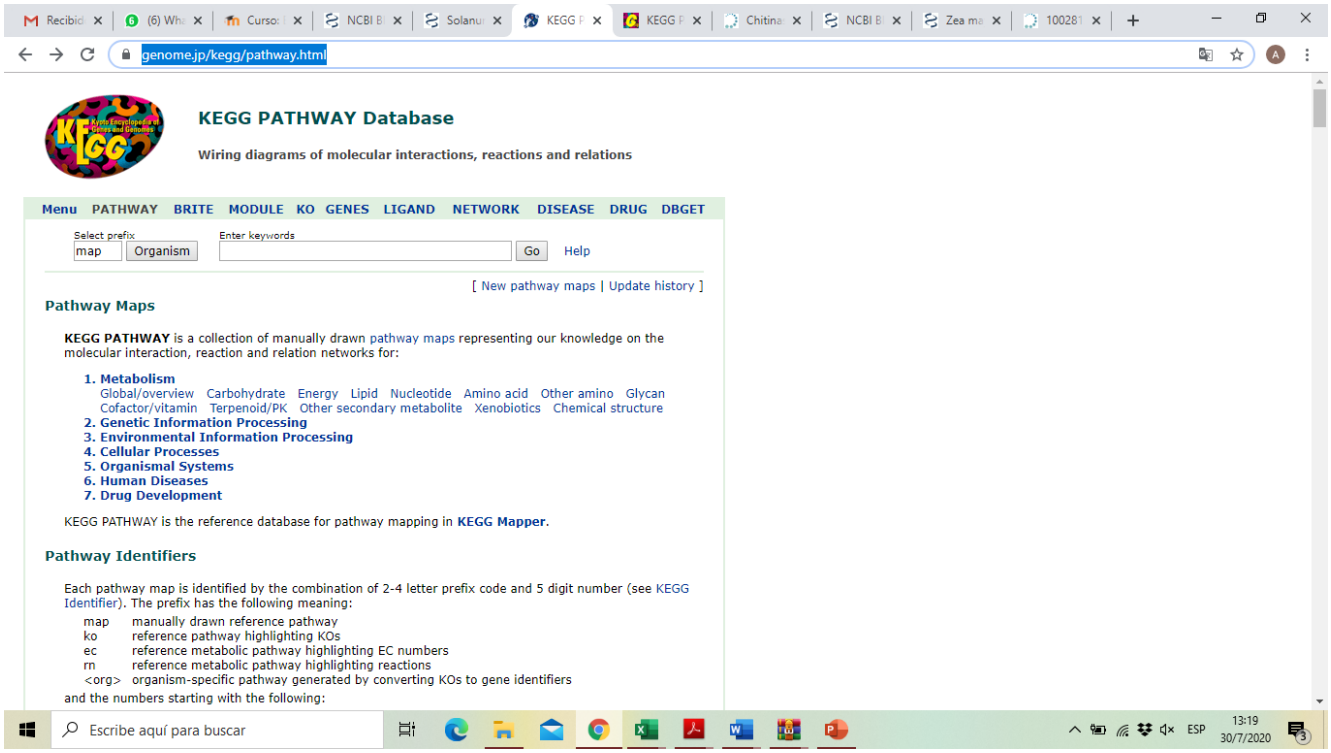
A continuación, copiá el número de locus e insertalo en el programa Uniprot, que se encuentra en: https://www.uniprot.org/, y te va a abrir la siguiente pantalla:



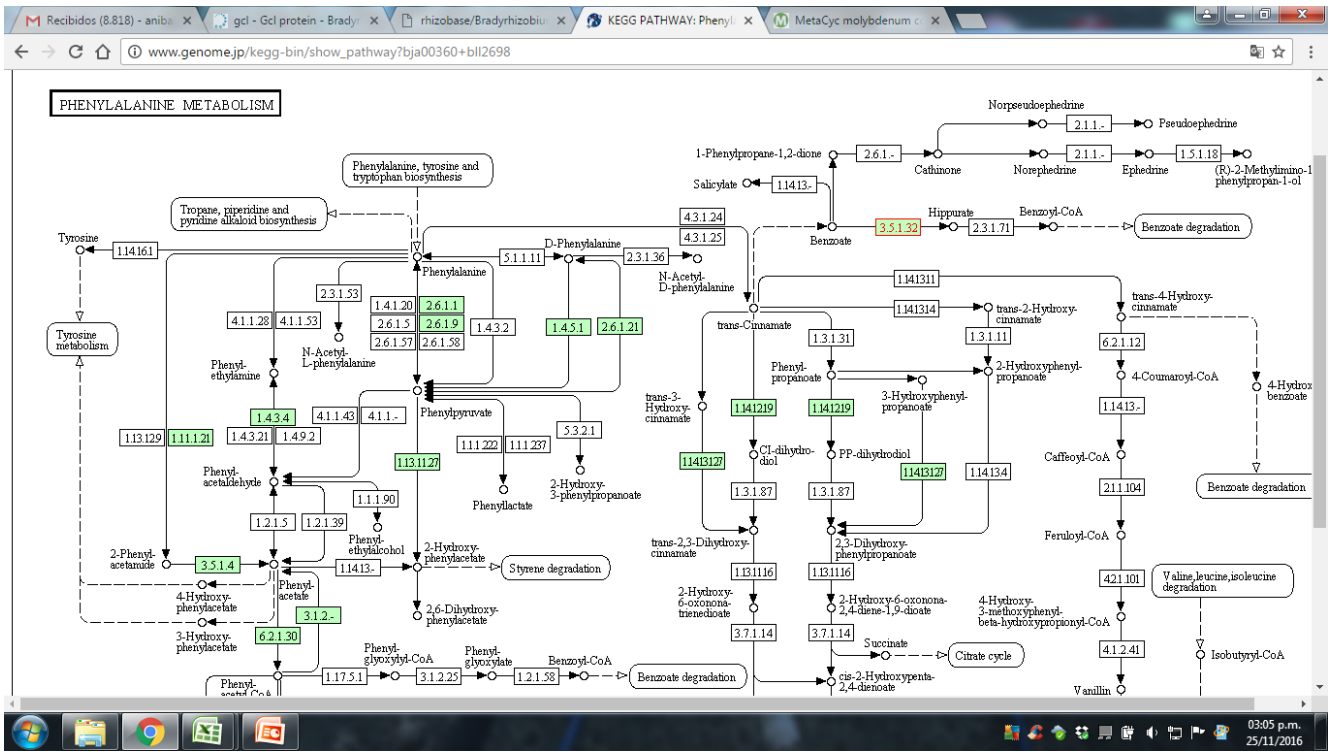
Una vez que hayas insertado el número de locus, apretá “search” (a la derecha) y fijate que te va a dar las proteínas con ese nombre que haya en su base de datos. En este punto es importante aclarar que estas bases de datos son redundantes, es decir que contienen todo lo que los investigadores hayan incorporado allí, de modo que si dos o más investigadores estaban trabajando con el mismo gen, puede haber más de una entrada para el mismo gen o proteína. En nuestro caso, andá a la primera que aparece, clickeá en ella y fijate lo que encontrás.

Uniprot, como lo sugiere su nombre, reúne información de varias plataformas diferentes, las cuales están como links en la página que se te va a abrir. Estaría bueno hacer una exploración a través de estos recursos. Por ejemplo, donde dice “GO-…” se refiere a “Gene Ontology”, que es algo así como una clasificación de los genes de acuerdo a su función. Fijate en las funciones que pone para este gen tanto en la función molecular como en el proceso biológico. Más abajo vas a encontrar los nombres y taxonomía tanto para la proteína como para la especie vegetal de donde proviene. Con respecto a la taxonomía de la proteína, es importante recalcar una clasificación que aparece como “EC: …” donde los … son unos números separados por puntos, por ejemplo: 1.1.1.1. Ese nombre corresponde a una clasificación de las enzimas de acuerdo a la reacción que catalizan, realizada por la Enzyme Commision (de donde viene la abreviatura EC; para una descripción, podés ver https://es.wikipedia.org/wiki/N%C3%BAmero_EC y links allí adentro). Continuando con las descripciones, vas a encontrar más abajo otra que te da la ubicación subcelular, y a continuación otra que se llama “Interaction” y que si entrás allí te va a dar un mapa de todas las otras proteínas que interactúan con la que encontraste. Esto es muy útil para intentar armarse una idea de en qué procesos puede intervenir esta proteína. Por su parte, en “Family and Domains” vas a encontrar una cantidad de plataformas que explican la estructura de la proteína y la distribución de sus dominios funcionales. Entrá en esos links a ver qué podés descubrir.

Finalmente, vamos a explorar otro programa, llamado KEGG Pathway, que está en https://www.genome.jp/kegg/pathway.html y que te va a permitir estudiar la vía metabólica donde actúa la proteína que acabás de descubrir. Allí tenés que seleccionar el organismo (en nuestro caso, zma, por *Zea mays*, pero si no sabés la sigla la encontrás apretando donde dice “Organism”) y luego en la ventana donde dice “Enter keywords”, es decir, “Ingresar palabras-clave”, ingresá el nombre de la enzima con el número EC que tenés que copiar desde Uniprot.



Una vez que ingreses, te va a mostrar unos mapas metabólicos en miniatura. Seleccioná el más relevante de acuerdo a lo que queremos saber de este gen (para eso releé el párrafo introductorio, donde te dice por qué estamos estudiando esta secuencia de ADN). Una vez que entres a la vía metabólica vas a ver algo como esto:

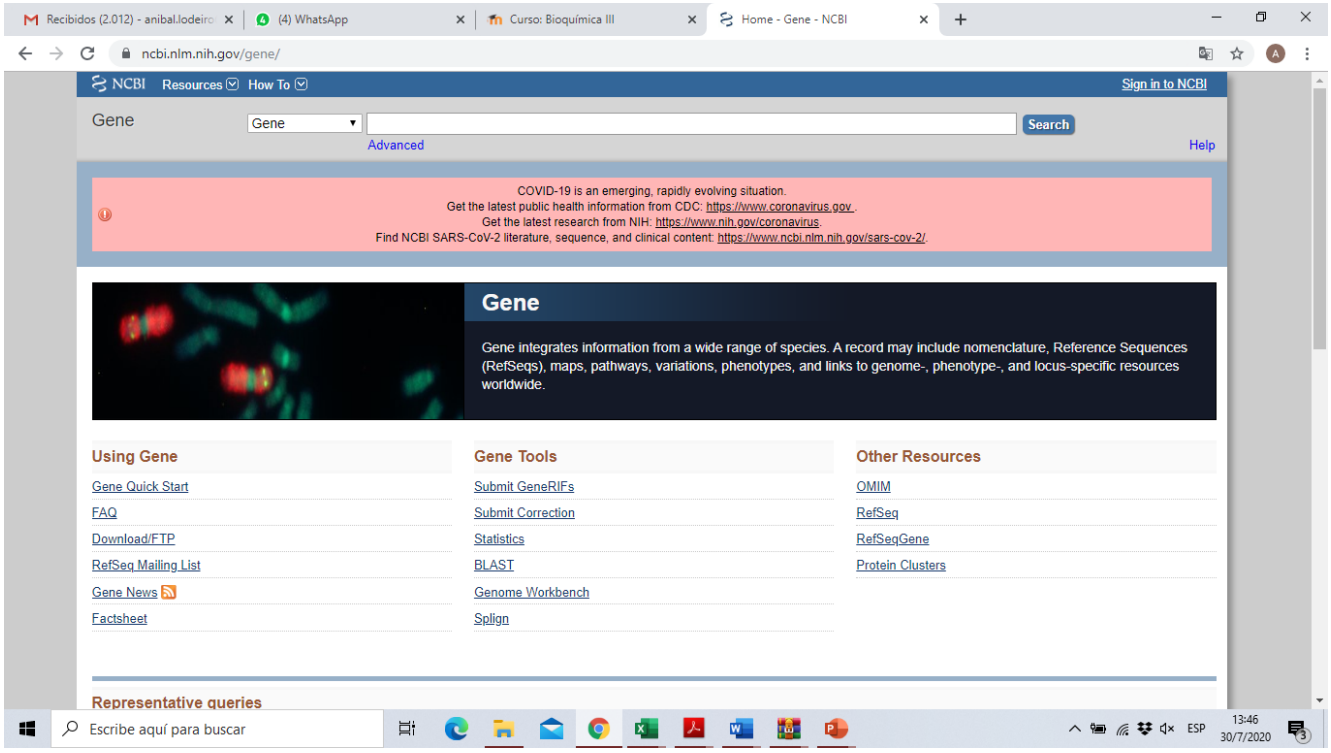


donde tu proteína de interés está remarcada en rojo (tené en cuenta que este **NO ES** el mapa de la proteína que estamos estudiando, es solo a título de ejemplo). En el mapa, todas las enzimas están indicadas con sus números EC y los recuadros verdes indican que esa proteína se encuentra en el genoma del organismo que se seleccionó para estudiar. Los rectángulos blancos, en cambio, indican que esas enzimas se conocen pero no se pudieron localizar en el genoma de la especie que se seleccionó para estudiar. Las flechas indican el sentido preferencial de las reacciones (recordá que todas son reversibles) y los circulitos con un nombre al lado indican los metabolitos. Si clickeás en un círculo te muestra la fórmula química de cada metabolito, y si clickeás en una enzima va a otras páginas donde te da características de las enzimas y de las reacciones catalizadas.

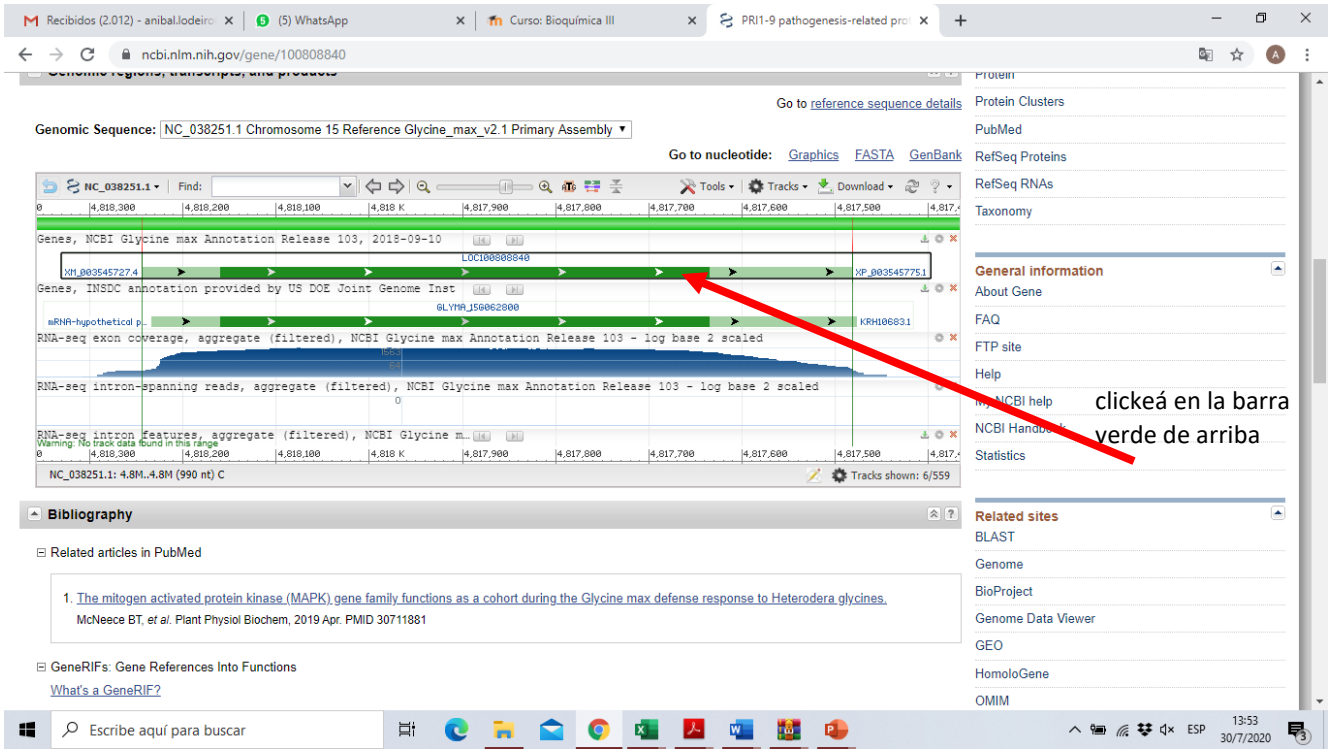
Una vez que hayas encontrado la vía metabólica donde interviene nuestra proteína (recordá que puede haber más de una vía metabólica donde interviene la misma proteína y tenés que elegir la más relevante para nuestros objetivos) explorala y responde por qué la planta sensible a la enfermedad que originó este estudio era incapaz de sintetizar ácido cumárico, y qué otras sustancias relacionadas con la resistencia a enfermedades podrían estar ausentes en este tipo de plantas.

1) Cómo encontrar secuencias de ADN de genes de interés

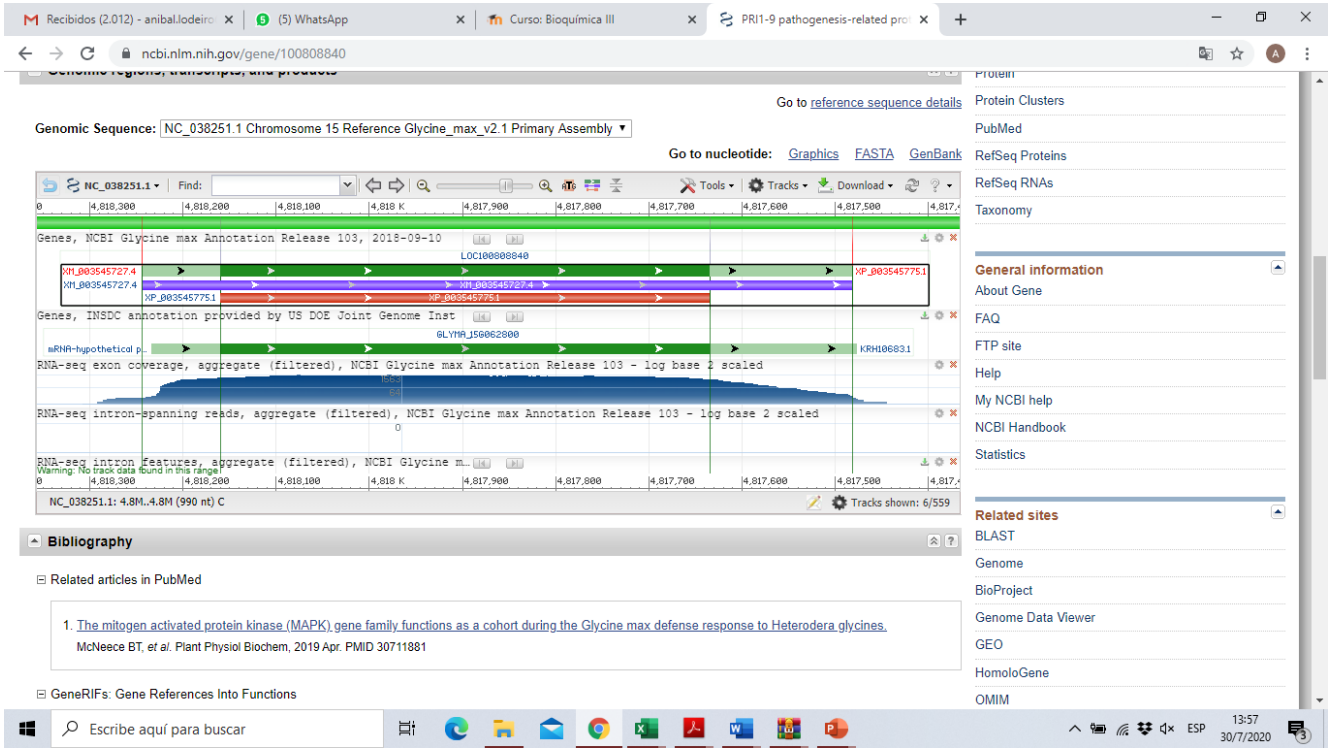
Supongamos que en vez de contar con una secuencia de ADN tenemos la sospecha de que cierta función biológica puede ser interesante con el objeto de entender y mejorar la resistencia a enfermedades. Entre estas, podemos mencionar una serie de proteínas llamadas “proteínas relacionadas con patogénesis” o “pathogenesis-related proteins”, las cuales se abrevian como PR seguidas de un número, por ejemplo PR-1. Otras proteínas de interés son las quitinasas (chitinases en inglés) que son enzimas capaces de digerir las paredes celulares de los hongos. Para encontrar sus secuencias de ADN y proteínas, vamos a ir nuevamente a la plataforma de los Institutos Nacionales de Salud de EEUU (National Institutes of Health, NIH) donde estaba la plataforma BLAST, y vamos a entrar a un sitio diferente, llamado Gene: https://www.ncbi.nlm.nih.gov/gene/



Ahí vamos a tipear el nombre del gen en la parte superior, y después “search”. Hacelo primero para PR-1 y para acotar la búsqueda, elegí una especie, por ejemplo, soja: “PR-1 soybean”. Cuando veas la lista de las proteínas que te devuelve, acordate que la base de datos es redundante, es decir hay varias entradas para la misma proteína, y no todas están completas. Elegí una, por ejemplo PR1-9, e iniciá la exploración. Vas a ver un esquema del gen donde te da la posibilidad de entrar y ver el ARNm y la proteína:

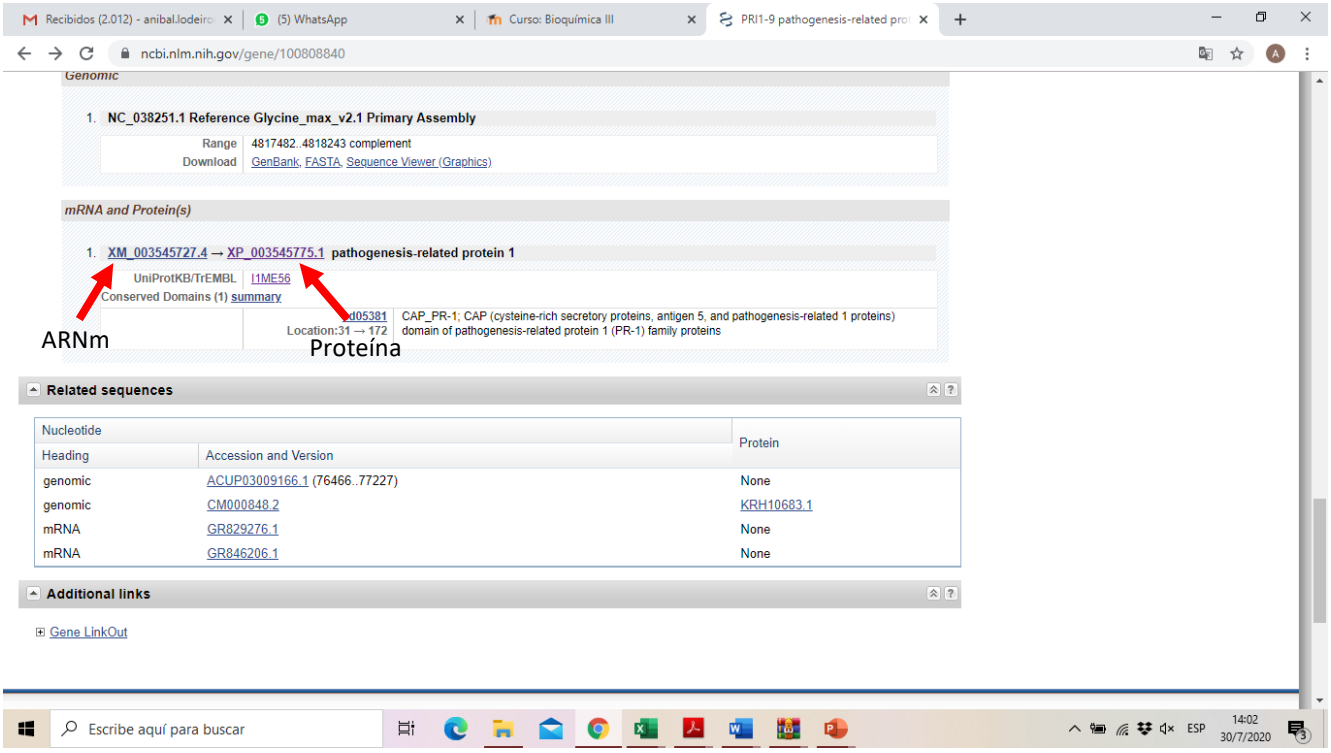


Te va a aparecer el esquema de la siguiente manera:



La barra azul representa el ARNm y la roja, la proteína. ¿Por qué la barra roja es más corta que la azul? Si te parás con el cursor encima de las barras, te va a mostrar la secuencia de proteína o nucleótidos correspondiente a la zona donde te paraste. Además, si movés las barras verticales y apretás la lupa que está encima, podés ver la secuencia de ADN correspondiente a la zona entre esas barras.

Si seguís hacia abajo, vas a encontrar enlaces a la secuencia del ARNm y a la proteína:

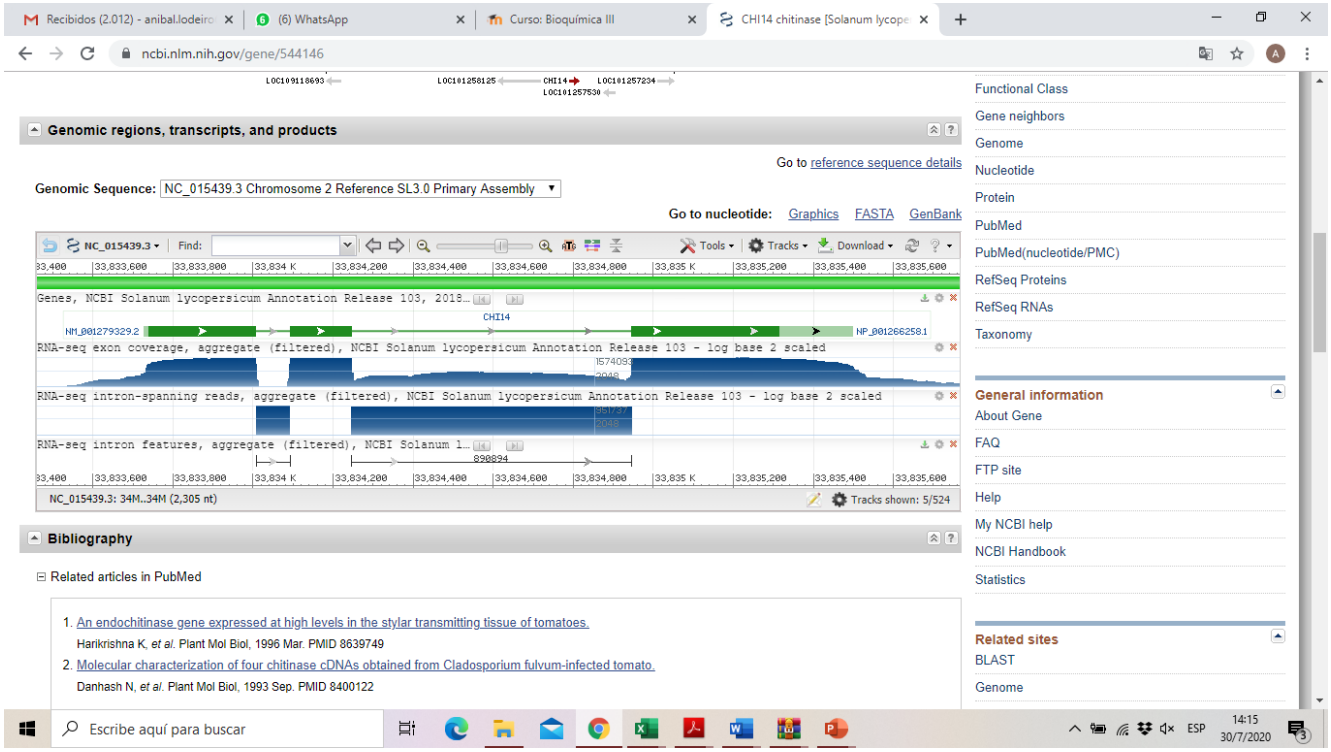


Entrá en ambos, y fijate que en la proteína, además de los atributos que se encuentran en “FEATURES” vas a encontrar uno que se llama “sig_peptide”, que significa “péptido señal”. Si apretás ahí, te va a señalar los aminoácidos que forman ese péptido señal.

Buscá en la literatura y en internet qué es y para qué sirve el péptido señal, y en relación con eso, deducí por qué a esta proteína la nombran como “precursor”.

Luego de esto, entrá nuevamente a Uniprot con el locus correspondiente a esta proteína y fíjate si en la ubicación subcelular encontrás la explicación de para qué sirve el péptido señal.

Finalmente, repetí el ejercicio con la quitinasa de tomate (entrá al mismo sitio Gene con “chitinase tomato”) y fíjate que en la barra verde que vimos antes podés encontrar la estructura de exones e intrones de ese gen.



Buscá qué representan estas estructuras y cómo se procesan durante la maduración del ARNm.

Además de estos ejercicios, aprovechá para navegar por todos estos sitios y los relacionados por los links que encuentres en ellos, y tomate el tiempo para averiguar qué son todos los recursos que hay y para qué sirven. La mejor manera de aprender estas cosas es metiendo la mano.