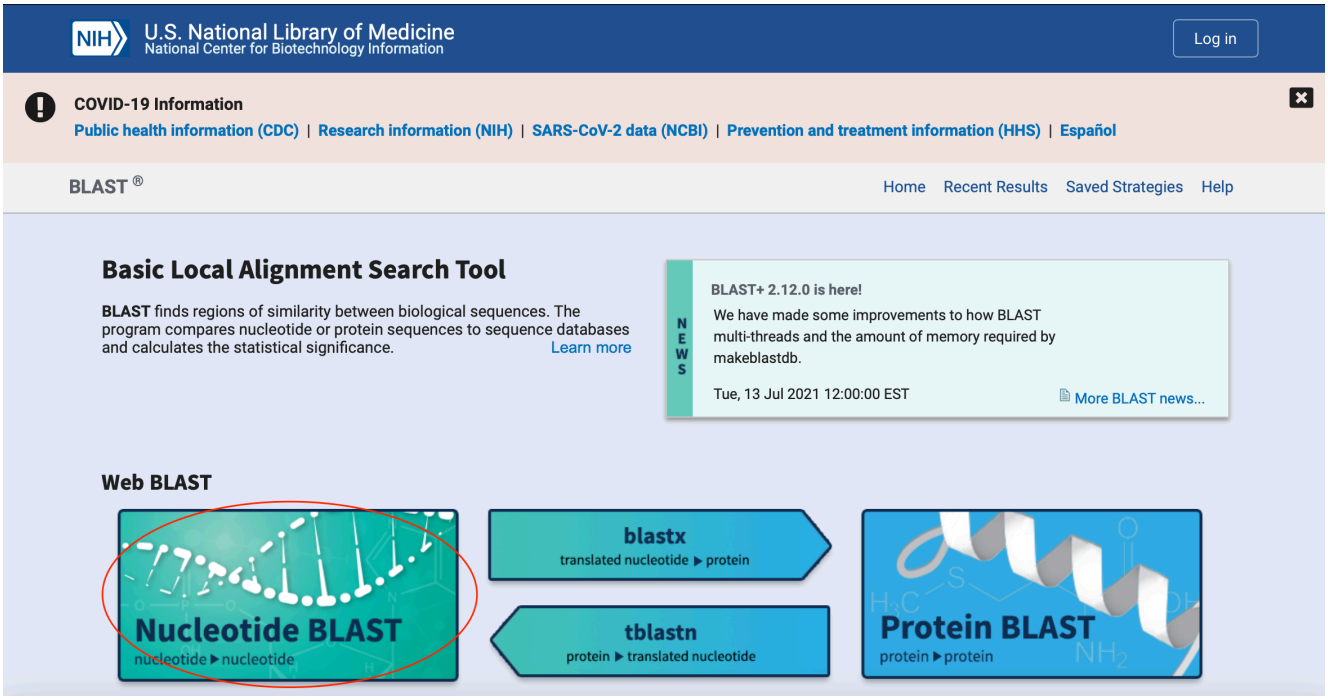
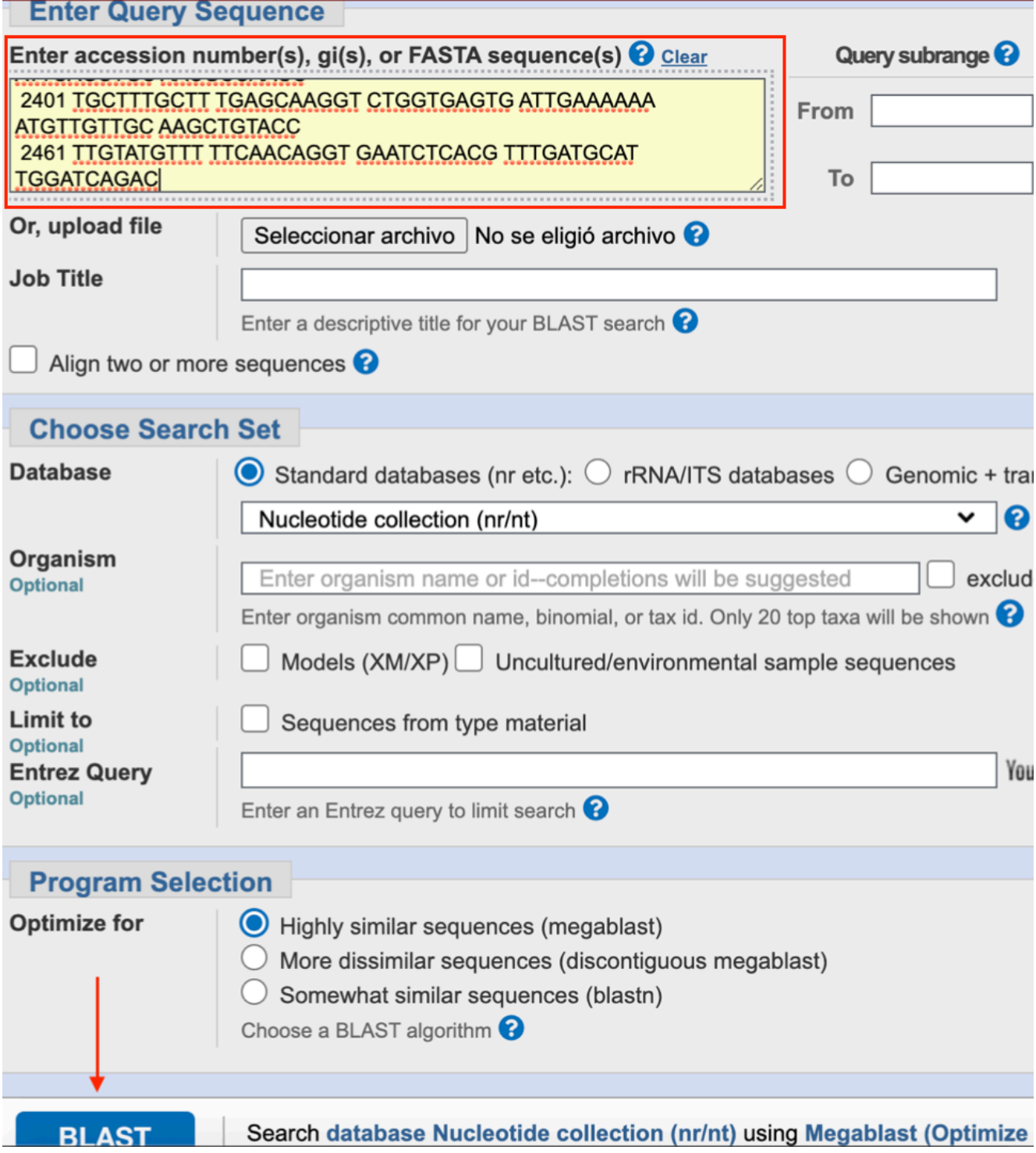
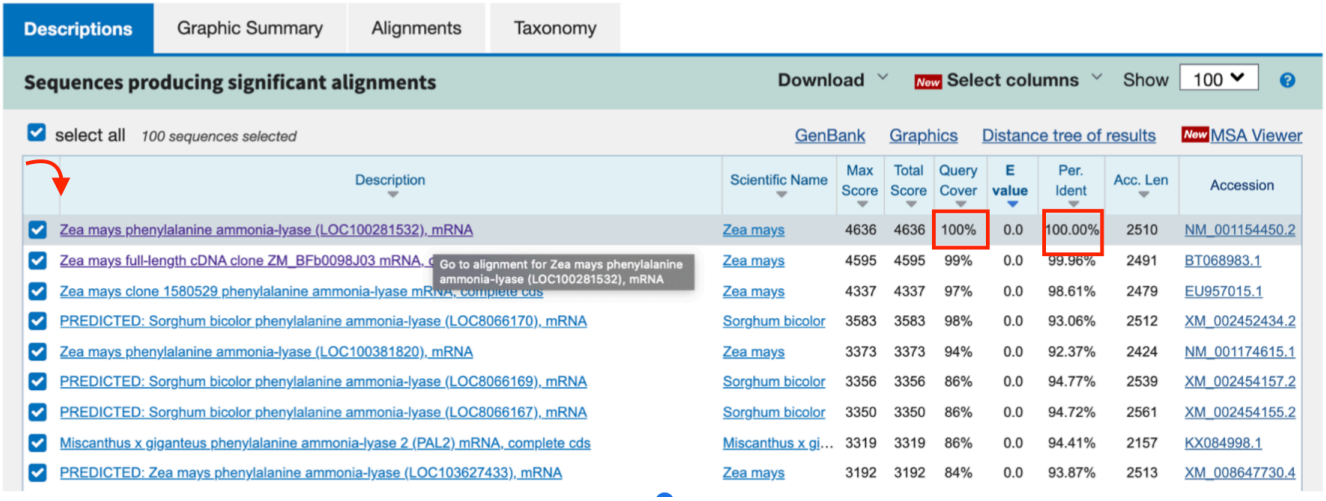
Cómo identificar y caracterizar un gen a partir de su secuencia de ADN

1. Nuestra query es una secuencia de ADN proveniente del organismo (maíz). Las plantas sensibles a ciertas enfermedades poseen un defecto en esta secuencia, lo cual inhabilita la síntesis de , un asociado a la resistencia de enfermedades . Identificar y caracterizar el gen que codifica nos permitara analizar su importancia biológica.





2. **Elegir la secuencia que tenga mejor alineamiento con la query.**



3. **Analizar e interpretar las características del gen identificado.**

Zea mays phenylalanine ammonia-lyase (LOC100281532), mRNA

NCBI Reference Sequence: NM_001154450.2

FASTA Graphics

Go to:

```
LOCUS       NM_001154450            2510 bp    mRNA    linear   PLN 03-JUL-2020
DEFINITION  Zea mays phenylalanine ammonia-lyase (LOC100281532), mRNA.
ACCESSION   NM_001154450
VERSION     NM_001154450.2
KEYWORDS    RefSeq.
SOURCE      Zea mays
  ORGANISM  Zea mays
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliopsida; Liliopsida; Poales; Poaceae; PACMAD
            clade; Panicoideae; Andropogonodae; Andropogoneae; Tripsacinae;
            Zea.
REFERENCE   1  (bases 1 to 2510)
  AUTHORS   Soderlund C, Descour A, Kudrna D, Bomhoff M, Boyd L, Currie J,
            Angelova A, Collura K, Wissotski M, Ashley E, Morrow D, Fernandes
            J, Walbot V and Yu Y.
  TITLE     Sequencing, mapping, and analysis of 27,455 maize full-length cDNAs
  JOURNAL   PLoS Genet. 5 (11), e1000740 (2009)
   PUBMED   19936069
REFERENCE   2  (bases 1 to 2510)
  AUTHORS   Alexandrov NN, Brover VV, Freidin S, Troukhan ME, Tatarinova TV,
            Zhang H, Swaller TJ, Lu YP, Bouck J, Flavell RB and Feldmann KA.
  TITLE     Insights into corn genes derived from large-scale cDNA sequencing
  JOURNAL   Plant Mol. Biol. 69 (1-2), 179-194 (2009)
   PUBMED   18937034
COMMENT     VALIDATED REFSEQ: This record has undergone validation or
            preliminary review. The reference sequence was derived from
            EE041021.2 and BT068983.1.
            On Aug 22, 2017 this sequence version replaced NM_001154450.1.

            ##Evidence-Data-START##
            Transcript is intronless :: BT068983.1, SRR3147047.9503.1
                                        [ECO:0000345]
            ##Evidence-Data-END##
```

```
PRIMARY      REFSEQ_SPAN          PRIMARY_IDENTIFIER PRIMARY_SPAN        COMP
             1-573                EE041021.2         1-573
             574-2510             BT068983.1         555-2491
FEATURES             Location/Qualifiers
     source          1..2510
                     /organism="Zea mays"
                     /mol_type="mRNA"
                     /cultivar="B73"
                     /db_xref="taxon:4577"
                     /chromosome="4"
                     /map="4"
     gene            1..2510
                     /gene="LOC100281532"
                     /gene_synonym="GRMZM2G063917"
                     /note="phenylalanine ammonia-lyase"
                     /db_xref="GeneID:100281532"
     exon            1..2510
                     /gene="LOC100281532"
                     /gene_synonym="GRMZM2G063917"
                     /inference="alignment:Splign:2.1.0"
     CDS             122..2272
                     /gene="LOC100281532"
                     /gene_synonym="GRMZM2G063917"
                     /EC_number="4.3.1.5"
                     /note="phenylalanine ammonia lyase4"
                     /codon_start=1
                     /product="phenylalanine ammonia-lyase"
                     /protein_id="NP_001147922.2"
                     /db_xref="GeneID:100281532"
                     /translation="MECDNGRVAATNGDSLCMALPRAADPLNWGKAAEEMMGSHLDEV
                     KRMVAEYRQPLVKIEGASLRIAQVAAVAAGAGEARVELDESARGRVKASSDWVRDSMM
                     NGTDSYGVTTGFGATSHRRTKEGGALQRELIRFLNAGAFGIGTDAGHVLPAEATRAAM
                     LVRINTLLQGYSGIRFEILEAIVKLLNANVTPCLPLRGTVTASGDLVPLSYIAGLVTG
                     RENAVAVAPDGTKVNAAEAFRIADIQSGFFELQPKEGLAMVNGTAVGSGLASTVLFEA
                     NVLAVLAEVLSAVFCEVMNGKPEYTDHLTHKLKHHPGQIEAAAIMEHILEGSSYMKLA
                     KKLGELDPLMKPKQDRYALRTSPQWLGPQIEVIRASTKSIEREINSVNDNPLIDVARS
                     KALHGGNFQGTPIGVSMDNTRLAVAAIGKLMFAQFSELVNDYYNNGLPSNLSGGRNPS
                     LDYGFKGAEIAMASYCSELQFLGNPVTNHVQSAEQHNQDVNSLGLISSRKTAEAIEIL
                     KLMSSTFLIALCQAVDLRHIEENVKSAVKSCVMTVAKKTLSTNSTGGLHVARFCEKDL
                     LQEIEREAVFAYADDPCSANYPLMKKLRNVLVERALANGTAEFDAETSVFAKVAQFEE
                     ELRTALPSAVEAARAAVENGTAAIPNRIAECRSYPLYRFVREELGAVYLTGEKTRSPG
                     EELNKVLVAINQGKHIDPLLECLKEWNGEPLPIC"
```

¿Que significa el alineamiento? ¿Como se ve si clickeas en proteinas con menor porcentaje de identidad?

El alineamiento de secuencias muestra el parentesco entre la secuencia con la que se trabaja (secuencia '`query`') y la encontrada (secuencia '`subject`'). En el caso de esta secuencia, todos los nucleótidos coinciden base por base con los de la secuencia subject, dándonos un `porcentaje de identidad` del 100%. Si analizamos otros genes que no tienen un porcentaje de identidad del 100%,

podemos ver que la mayoría de los nucleótidos de nuestra secuencia están presentes en la 'subject' pero se encuentran '`gaps`', es decir, se encontró una secuencia similar a la query pero con regiones que no se alinean. A medida que decrece el porcentaje de identidad, comienzan a aparecer más '`mismatches`', además de '`gaps`'.

¿Que proteina es?

La enzima **`Fenilalanina amonio liasa`**

¿Por que la CDS es más chica que la region abarcada por el exon?

Dependiendo del `marco de lectura`, pueden haber diversos exones para una misma secuencia codificante. La ubicación de los `codones de iniciciación` y de `terminación` determina el largo de la secuencia codificante, que va a ser a lo sumo del mismo tamaño que los exones, pero usualmente más corta como en este caso.

¿Donde estan los codones de iniciación y de terminación?

`122 (ATG) - 2272 (TGA).`

Para visualmente identificar estos codones y el marco de lectura, buscamos el codón de iniciación (AUG, en el caso de la secuencia con la que trabajamos seria ATG por ser ADN) y los codones de terminación (TAA, TAG, TGA). Utilizando la herramienta de COMMAND + F y sabiendo

que el marco iba de 122 - 2272 encontramos que la secuencia de terminación era la TGA.

Score	Expect	Identities	Gaps	Strand
4636 bits(2510)	0.0	2510/2510(100%)	0/2510(0%)	Plus/Plus

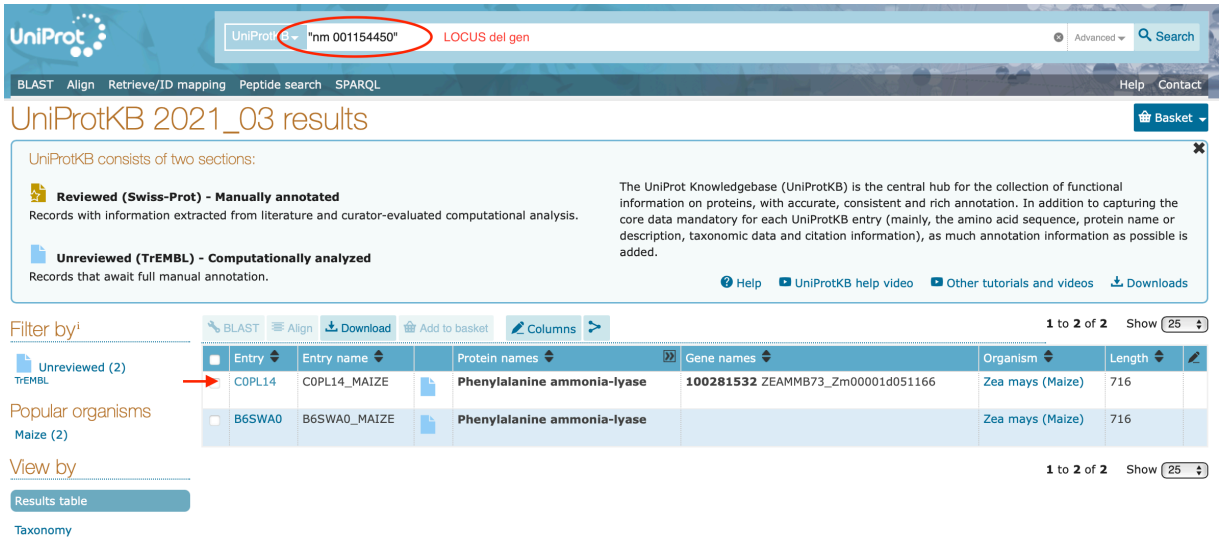
```
Query  1     GCACCATCCAGTGCATCACGAGCTCTTCTGCACCAGATTAGCAGGCCATCGCCTACTTTT  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1     GCACCATCCAGTGCATCACGAGCTCTTCTGCACCAGATTAGCAGGCCATCGCCTACTTTT  60

Query  61    GGCTTCCAAATCATTTATTTACGGCGTACGTGCCTTCTGTTCAAACCCCAGCCCCGCTGC  120
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61    GGCTTCCAAATCATTTATTTACGGCGTACGTGCCTTCTGTTCAAACCCCAGCCCCGCTGC  120

Query  121   AATGGAGTGCGACAACGGCCGCGTCGCTGCTACCAACGGCGACTCCCTGTGCATGGCGCT  180
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121   AATGGAGTGCGACAACGGCCGCGTCGCTGCTACCAACGGCGACTCCCTGTGCATGGCGCT  180

Query  181   GCCCCGCGCCGCCGACCCGCTTAACTGGGGGAAGGCGGCGGAGGAGATGATGGGCAGCCA  240
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  181   GCCCCGCGCCGCCGACCCGCTTAACTGGGGGAAGGCGGCGGAGGAGATGATGGGCAGCCA  240

Query  241   CCTCGACGAGGTGAAGCGGATGGTGGCCGAGTACCGCCAGCCCCTGGTGAAGATCGAGGG  300
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  241   CCTCGACGAGGTGAAGCGGATGGTGGCCGAGTACCGCCAGCCCCTGGTGAAGATCGAGGG  300

Query  301   CGCCAGCCTCCGCATCGCGCAGGTGGCCGCTGTCGCCGCCGGCGCGGGCGAGGCCCGGGT  360
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  301   CGCCAGCCTCCGCATCGCGCAGGTGGCCGCTGTCGCCGCCGGCGCGGGCGAGGCCCGGGT  360

Query  361   CGAGCTCGACGAGTCCGCCCGCGGCCGGGTCAAGGCGAGCAGCGACTGGGTCAGGGACAG  420
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  361   CGAGCTCGACGAGTCCGCCCGCGGCCGGGTCAAGGCGAGCAGCGACTGGGTCAGGGACAG  420

Query  421   CATGATGAACGGCACCGACAGCTACGGCGTCACCACCGGCTTCGGCGCCACCTCCCACCG  480
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  421   CATGATGAACGGCACCGACAGCTACGGCGTCACCACCGGCTTCGGCGCCACCTCCCACCG  480
```

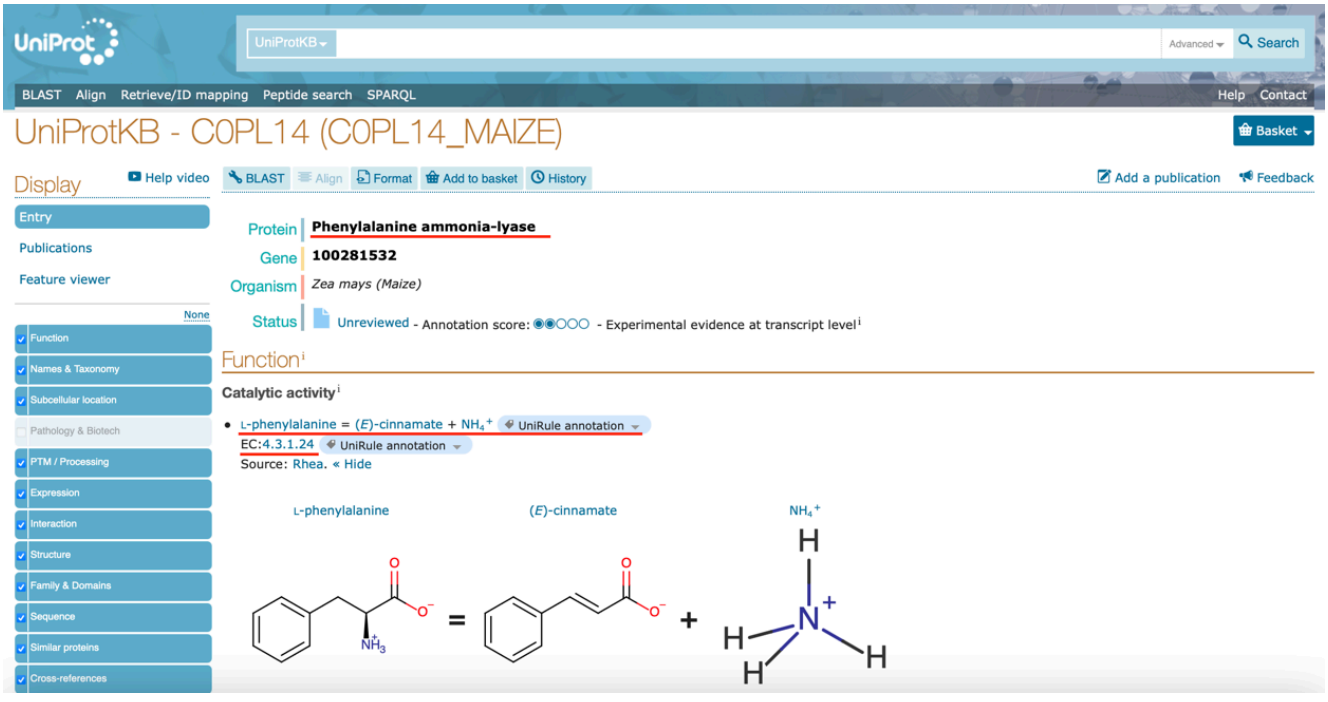
```
Query  2161  TCCCGGCGAGGAGCTTAACAAGGTGCTCGTTGCCATCAACCAGGGCAAGCACATCGACCC  2220
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2161  TCCCGGCGAGGAGCTTAACAAGGTGCTCGTTGCCATCAACCAGGGCAAGCACATCGACCC  2220

Query  2221  GCTGCTCGAGTGCCTCAAGGAGTGGAACGGCGAGCCCCTGCCCATCTGCTGAACAGAGAA  2280
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2221  GCTGCTCGAGTGCCTCAAGGAGTGGAACGGCGAGCCCCTGCCCATCTGCTGAACAGAGAA  2280

Query  2281  AATACAAGGAGCAGAAGACTGTATTTTTTAGCTAATACGCACTTTTTATTCCTAATTTAT  2340
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2281  AATACAAGGAGCAGAAGACTGTATTTTTTAGCTAATACGCACTTTTTATTCCTAATTTAT  2340
```

```
Query  2161  TCCCGGCGAGGAGCTTAACAAGGTGCTCGTTGCCATCAACCAGGGCAAGCACATCGACCC  2220
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2161  TCCCGGCGAGGAGCTTAACAAGGTGCTCGTTGCCATCAACCAGGGCAAGCACATCGACCC  2220

Query  2221  GCTGCTCGAGTGCCTCAAGGAGTGGAACGGCGAGCCCCTGCCCATCTGCTGAACAGAGAA  2280
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2221  GCTGCTCGAGTGCCTCAAGGAGTGGAACGGCGAGCCCCTGCCCATCTGCTGAACAGAGAA  2280

Query  2281  AATACAAGGAGCAGAAGACTGTATTTTTTAGCTAATACGCACTTTTTATTCCTAATTTAT  2340
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2281  AATACAAGGAGCAGAAGACTGTATTTTTTAGCTAATACGCACTTTTTATTCCTAATTTAT  2340
```

```
Query  2161  TCCCGGCGAGGAGCTTAACAAGGTGCTCGTTGCCATCAACCAGGGCAAGCACATCGACCC  2220
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2161  TCCCGGCGAGGAGCTTAACAAGGTGCTCGTTGCCATCAACCAGGGCAAGCACATCGACCC  2220

Query  2221  GCTGCTCGAGTGCCTCAAGGAGTGGAACGGCGAGCCCCTGCCCATCTGCTGAACAGAGAA  2280
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2221  GCTGCTCGAGTGCCTCAAGGAGTGGAACGGCGAGCCCCTGCCCATCTGCTGAACAGAGAA  2280

Query  2281  AATACAAGGAGCAGAAGACTGTATTTTTTAGCTAATACGCACTTTTTATTCCTAATTTAT  2340
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2281  AATACAAGGAGCAGAAGACTGTATTTTTTAGCTAATACGCACTTTTTATTCCTAATTTAT  2340
```

4. **Utilizar UNIPROT para profundizar en la caracterización del gen.** https://www.uniprot.org/





En la sección de `Gene Ontology (GO)` se puede ver la clasificación del gen según su función. Podemos ver que nuestro gen actúa como una enzima **amonio-liasa** y participa en la biosíntesis de **ácido cinámico** a partir de **fenilalalina.** Esto es de particular importancia para nuestro análisis porque el **ácido cinámico** es un **fenilpropanoide**, cómo el **ácido cumárico**.

GO - Molecular function

- ammonia-lyase activity — Source: GO_Central
- phenylalanine ammonia-lyase activity — Source: UniProtKB-EC

Complete GO annotation on QuickGO ...

GO - Biological process

- cinnamic acid biosynthetic process — Source: UniProtKB-UniPathway
- L-phenylalanine catabolic process — Source: InterPro

Complete GO annotation on QuickGO ...

Keywords

Molecular function	Lyase — UniRule annotation — Imported
Biological process	Phenylpropanoid metabolism — UniRule annotation — ARBA annotation

Enzyme and pathway databases

UniPathway	UPA00713;UER00725

Names & Taxonomy

Protein names	*Recommended name:* **Phenylalanine ammonia-lyase** — UniRule annotation — EC:4.3.1.24 — UniRule annotation
Gene names	*Name:***100281532** — Imported ORF Names:ZEAMMB73_Zm00001d051166 — Imported
Organism	Zea mays (Maize) — Imported
Taxonomic identifier	4577 [NCBI]
Taxonomic lineage	Eukaryota › Viridiplantae › Streptophyta › Embryophyta › Tracheophyta › Spermatophyta › Magnoliopsida › Liliopsida › Poales › Poaceae › PACMAD clade › Panicoideae › Andropogonodae › Andropogoneae › Tripsacinae › Zea
Proteomes	UP000007305 Component: Chromosome 4

Los números `EC (Enzyme Comission)` clasifican a las enzimas en base a la reacción química asociada. En este caso la enzima pertenece al grupo 4: **liasas**.

EC 4 ***Liasas***	Adición o eliminación no hidrolítica de grupos de los substratos. Pueden romper los enlaces C-C, C-N, C-O o C-S.	RCOCOOH → RCOH + CO_2

La sección de `Interaction`muestra un mapa de las diferentes proteínas que interactuan con nuestra enzima (pal4).

Interaction

Protein-protein interaction databases

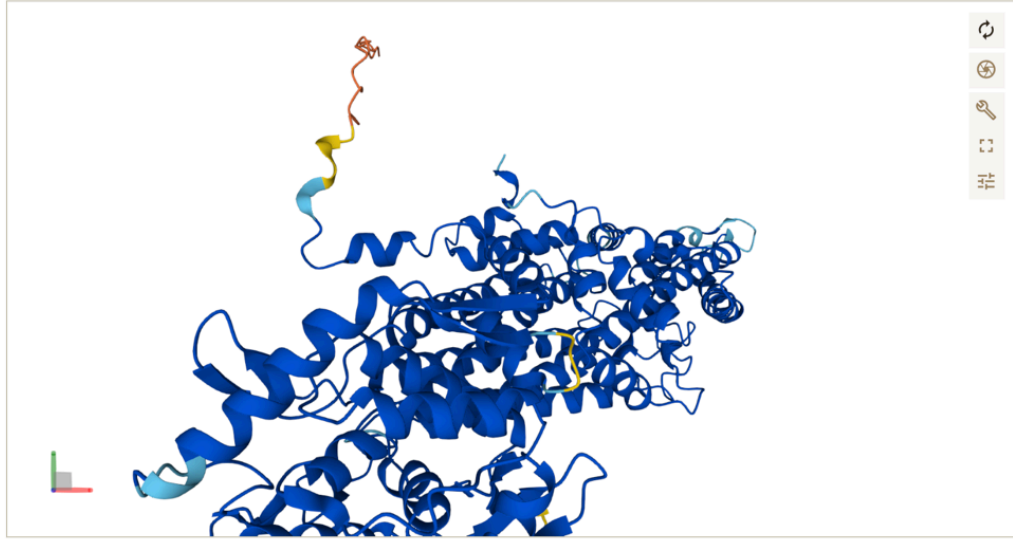
STRING	4577.GRMZM2G063917_P01

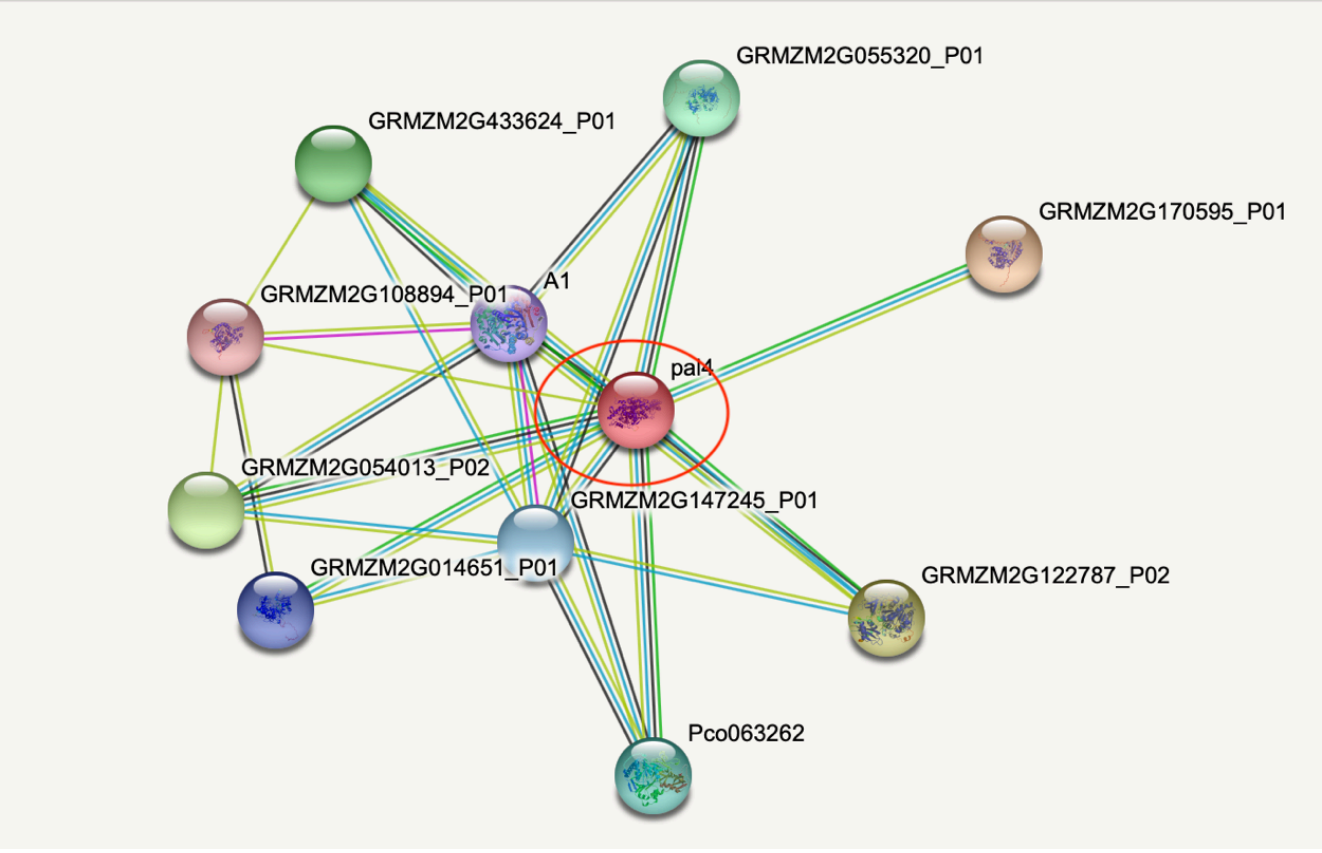
Structure

Model Confidence:

- Very high (pLDDT > 90)
- Confident (90 > pLDDT > 70)
- Low (70 > pLDDT > 50)
- Very low (pLDDT < 50)

AlphaFold produces a per-residue confidence score (pLDDT) between 0 and 100. Some regions with low pLDDT may be unstructured in isolation.

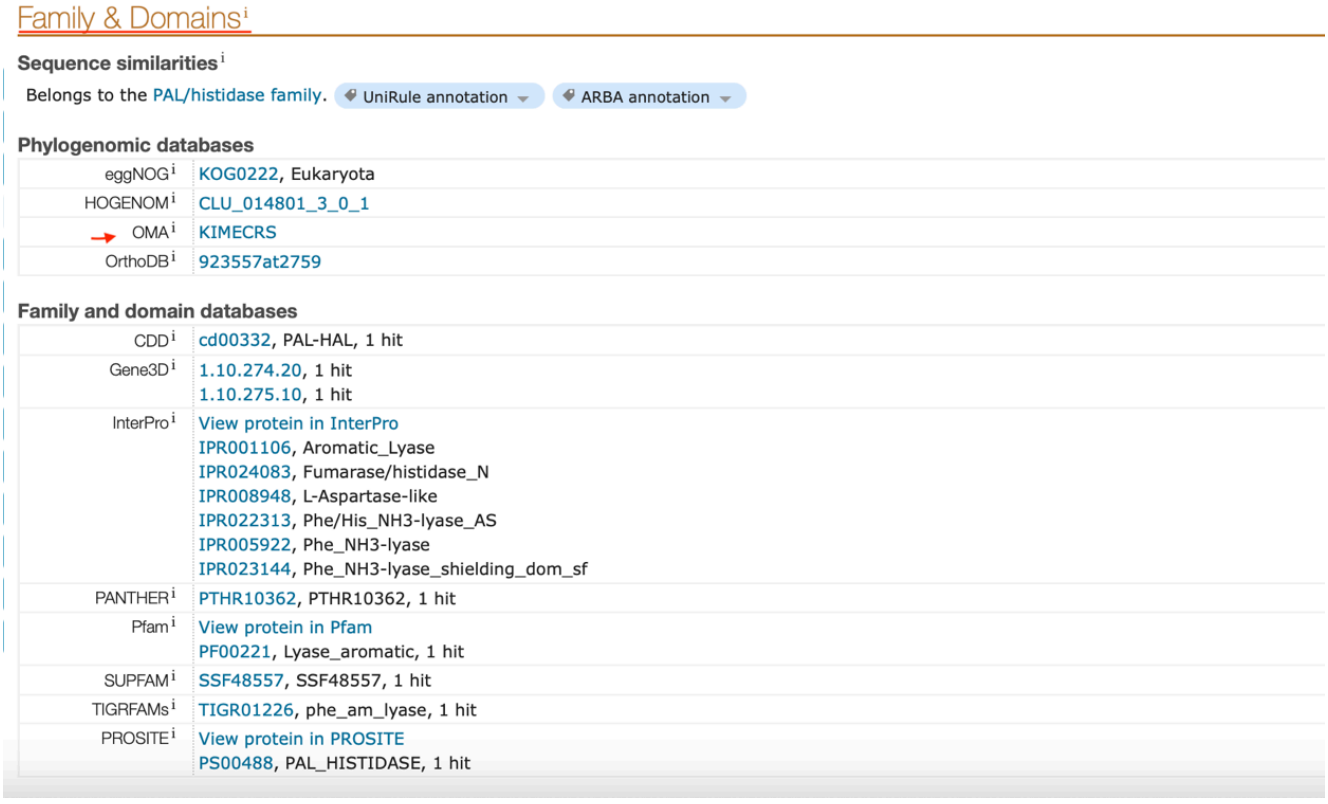




GRMZM2G170595_P01	*Histidinol-phosphate aminotransferase 2 chloroplastic*	0.941
GRMZM2G122787_P02	*4-coumarate--CoA ligase-like 7*	0.879
GRMZM2G054013_P02	*Putative AMP-dependent synthetase and ligase superfamily protein*	0.874
GRMZM2G433624_P01	*annotation not available*	0.865
GRMZM2G055320_P01	*Putative AMP-dependent synthetase and ligase superfamily protein*	0.862
Pco063262	*4-coumarate--CoA ligase 1*	0.858
GRMZM2G147245_P01	*Trans-cinnamate 4-monooxygenase; Putative cytochrome P450 superfamily protein; Uncharacterized pro...*	0.826
GRMZM2G014651_P01	*Uncharacterized protein*	0.792
A1	*Dihydroflavonol 4-reductase; Bifunctional enzyme involved in flavonoid metabolism; Belongs to the NAD(...*	0.777
GRMZM2G108894_P01	*Type III polyketide synthase B; Chalcone synthase; Uncharacterized protein ; Belongs to the chalcone/stil...*	0.772

Luego en la sección de `Family & Domains` se puede ver la similitud estructural y secuencial con otras proteínas y los dominios funcionales presentes.

En este caso entramos a `OMA`, una base de datos filogenómica.



Family & Domains

Sequence similarities

Belongs to the PAL/histidase family. UniRule annotation ARBA annotation

Phylogenomic databases

eggNOG	KOG0222, Eukaryota
HOGENOM	CLU_014801_3_0_1
OMA	KIMECRS
OrthoDB	923557at2759

Family and domain databases

CDD	cd00332, PAL-HAL, 1 hit
Gene3D	1.10.274.20, 1 hit 1.10.275.10, 1 hit
InterPro	View protein in InterPro IPR001106, Aromatic_Lyase IPR024083, Fumarase/histidase_N IPR008948, L-Aspartase-like IPR022313, Phe/His_NH3-lyase_AS IPR005922, Phe_NH3-lyase IPR023144, Phe_NH3-lyase_shielding_dom_sf
PANTHER	PTHR10362, PTHR10362, 1 hit
Pfam	View protein in Pfam PF00221, Lyase_aromatic, 1 hit
SUPFAM	SSF48557, SSF48557, 1 hit
TIGRFAMs	TIGR01226, phe_am_lyase, 1 hit
PROSITE	View protein in PROSITE PS00488, PAL_HISTIDASE, 1 hit

Podemos ver 6 organismos diferentes que poseen esta proteína en su genoma, y sus respectivos dominios.

Domains	Taxon	Protein ID	Cross reference	Domain Architectures
E	Triticum aestivum	WHEAT111616	A0A3B6QI72	
E	Zea mays	MAIZE48728	C0PL14	
E	Setaria italica	SETIT02528	K3YQD2	
E	Cucumis sativus	CUCSA18623	A0A0A0KEL3	
E	Manihot esculenta	MANES03985	A0A2C9VYR2	
E	Selaginella moellendorffii	SELML11505	D8SER5	

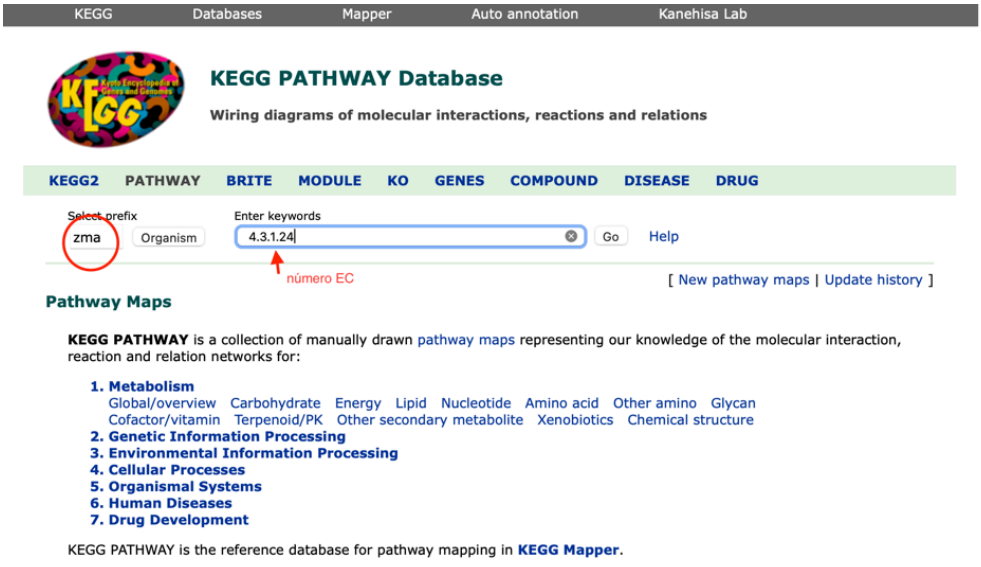
Luego entrando a otra plataforma, como `eggNOG`, encontramos 900 proteínas similares a nuestra secuencia: **ortólogos.**

Ortólogos son genes que comparten el último ancestro común y cuya divergencia se debe a la especiación. Es decir, el mismo gen en diferentes especies .[Tress, M. (2005). Análisis de Secuencias, Familias de Proteínas. Masters En Bioinformática Madrid 2005. http://www.pdg.cnb.uam.es/cursos/Master2005/Fam_theory/familias.pdf]

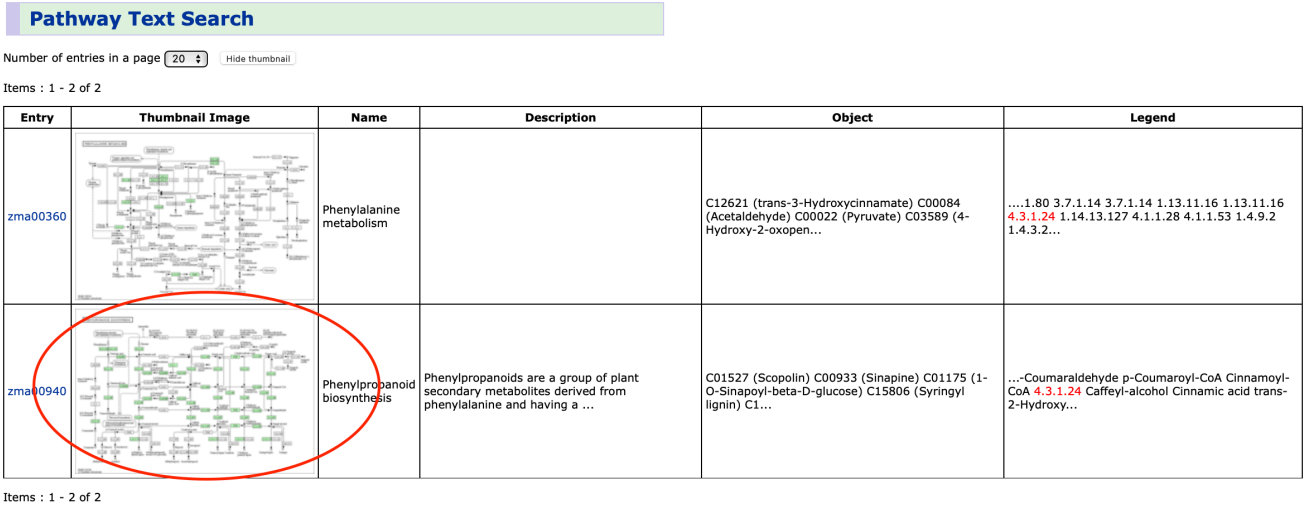
Podemos ver que en la lista de ortólogos de `eggNOG` aparecen los mismos organismos que en `OMA`. Setaria italica es mijo y triticum aestivum es trigo común.

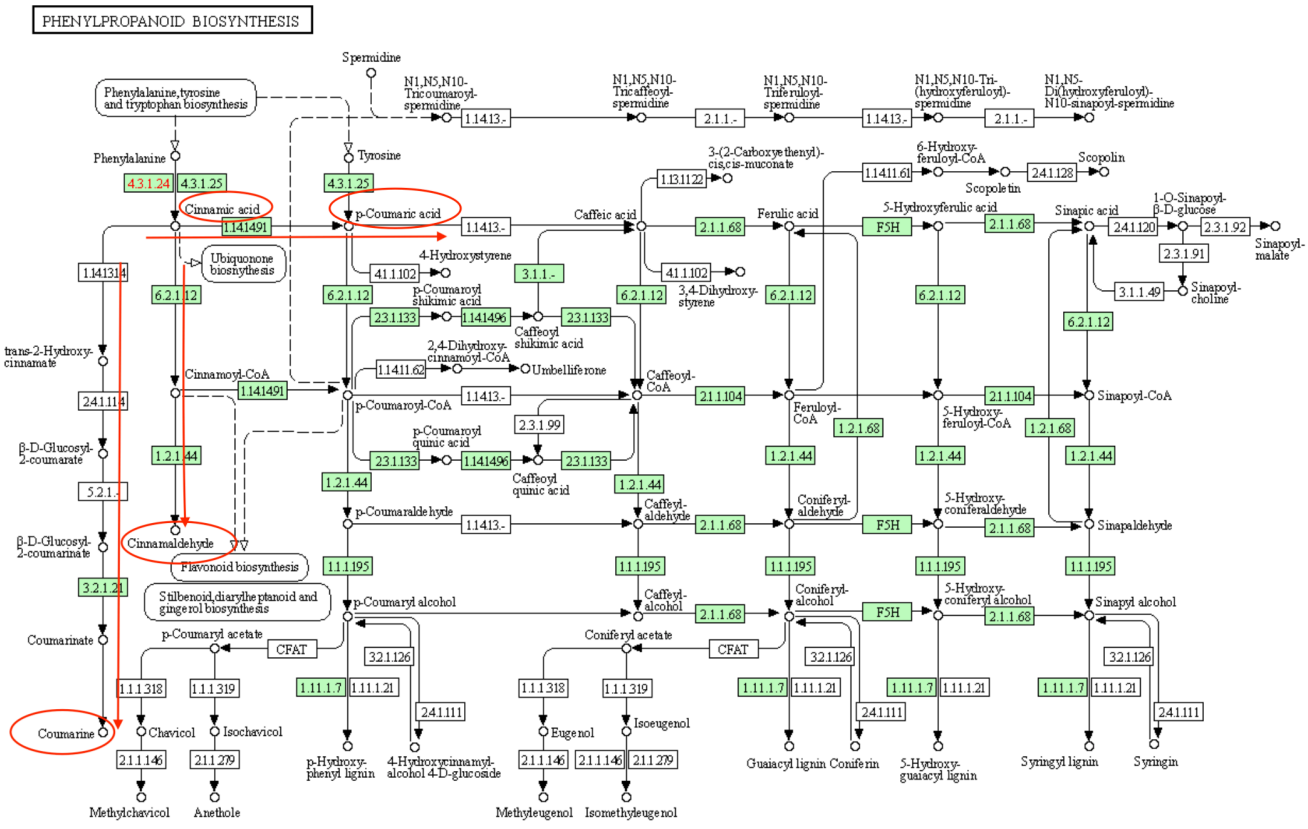
Setaria italica	10 seqs	4555.Si016504m, 4555.Si009509m, 4555.Si016467m, 4555.Si013348m, 4555.Si009345m, 4555.Si029093m, 4555.Si016475m, 4555.Si019385m, 4555.Si016478m, 4555.Si012256m	4555.Si016504m, 4555.Si009509m, 4555.Si016467m, 4555.Si013348m, 4555.Si009345m, 4555.Si029093m, 4555.Si016475m, 4555.Si019385m, 4555.Si016478m, 4555.Si012256m
Sorghum bicolor	9 seqs	4558.Sb01g014020.1, 4558.Sb06g022740.1, 4558.Sb04g026510.1, 4558.Sb04g026530.1, 4558.Sb04g026540.1, 4558.Sb04g026520.1, 4558.Sb04g026550.1, 4558.Sb06g022750.1, 4558.Sb04g026560.1	4558.Sb01g014020.1, 4558.Sb06g022740.1, 4558.Sb04g026510.1, 4558.Sb04g026530.1, 4558.Sb04g026540.1, 4558.Sb04g026520.1, 4558.Sb04g026550.1, 4558.Sb06g022750.1, 4558.Sb04g026560.1
Triticum aestivum	72 seqs	4565.Traes_1DS_F14FF0AC4.2, 4565.Traes_5BL_AB528184A.1, 4565.Traes_2DL_1AF45DD78.1, 4565.Traes_2AS_9711A3BEE.1, 4565.Traes_6BL_4A69B0CDF.2, 4565.Traes_2BL_5A50FDA1A.1, 4565.Traes_2DS_3791A8A36.1, 4565.Traes_2AL_C07505E42.1, 4565.Traes_1BS_C26CCE0E9.1, 4565.Traes_2AS_0F971F527.1, 4565.Traes_2DL_44DA40D1B.1, 4565.Traes_2BS_DF9FFA81F.1, 4565.Traes_1DS_A8BD91E4A.1, 4565.Traes_2DL_342D2421E.1, 4565.Traes_2DL_8BA1CE63D.1, 4565.Traes_6AL_77FE7A3DC.1, 4565.Traes_6BL_C96F626D9.1, 4565.Traes_2DL_BF4C00019.1, 4565.Traes_2BL_13D5272D7.1, 4565.Traes_6AL_67ED9DB4D.1, 4565.Traes_5BL_F50A47503.1, 4565.Traes_1DS_A171C7D59.1, 4565.Traes_6AL_BF1A559FF.1, 4565.Traes_2AL_A2BB17CE1.1, 4565.Traes_1BS_AEA817616.1, 4565.Traes_2AL_9EC3226F7.1, 4565.Traes_2AS_5EB59FFC0.1, 4565.Traes_1BS_BD86C90A7.1, 4565.Traes_1DS_A21FCBA3D.1, 4565.Traes_6BL_F321C9050.1, 4565.Traes_2DL_8C21AAAEF.1, 4565.Traes_1AS_0B9295A68.1, 4565.Traes_2DL_9A0F096AD.1, 4565.Traes_1AS_F9013A945.1, 4565.Traes_2DL_4D00BC16D.1, 4565.Traes_2DL_03E388B44.2, 4565.Traes_2DS_28CA50371.1, 4565.Traes_1BS_C3AA04888.1, 4565.Traes_2AS_9ACCB9C97.2, 4565.Traes_2AS_958327519.3, 4565.Traes_6AL_691F5142C.1, 4565.Traes_1BS_A3B2A0DAD.1, 4565.Traes_6AL_EB6D01B8E.1, 4565.Traes_2DL_130D6031E.1, 4565.Traes_1BS_723922D171.1, 4565.Traes_3B_DBB113CF1.1,	4565.Traes_1DS_F14FF0AC4.2, 4565.Traes_5BL_AB528184A.1, 4565.Traes_2DL_1AF45DD78.1, 4565.Traes_2AS_9711A3BEE.1, 4565.Traes_6BL_4A69B0CDF.2, 4565.Traes_2BL_5A50FDA1A.1, 4565.Traes_2DS_3791A8A36.1, 4565.Traes_2AL_C07505E42.1, 4565.Traes_1BS_C26CCE0E9.1, 4565.Traes_2AS_0F971F527.1, 4565.Traes_2DL_44DA40D1B.1, 4565.Traes_2BS_DF9FFA81F.1, 4565.Traes_1DS_A8BD91E4A.1, 4565.Traes_2DL_342D2421E.1, 4565.Traes_2DL_8BA1CE63D.1, 4565.Traes_6AL_77FE7A3DC.1, 4565.Traes_6BL_C96F626D9.1, 4565.Traes_2DL_BF4C00019.1, 4565.Traes_2BL_13D5272D7.1, 4565.Traes_6AL_67ED9DB4D.1, 4565.Traes_5BL_F50A47503.1, 4565.Traes_1DS_A171C7D59.1, 4565.Traes_6AL_BF1A559FF.1, 4565.Traes_2AL_A2BB17CE1.1, 4565.Traes_1BS_AEA817616.1, PAL, 4565.Traes_2AS_5EB59FFC0.1, 4565.Traes_1BS_BD86C90A7.1, 4565.Traes_1DS_A21FCBA3D.1, 4565.Traes_6BL_F321C9050.1, 4565.Traes_2DL_8C21AAAEF.1, 4565.Traes_1AS_0B9295A68.1, 4565.Traes_2DL_9A0F096AD.1, 4565.Traes_1AS_F9013A945.1, 4565.Traes_2DL_4D00BC16D.1, 4565.Traes_2DL_03E388B44.2, 4565.Traes_2DS_28CA50371.1, 4565.Traes_1BS_C3AA04888.1, 4565.Traes_2AS_9ACCB9C97.2, 4565.Traes_2AS_958327519.3, 4565.Traes_6AL_691F5142C.1, 4565.Traes_1BS_A3B2A0DAD.1, 4565.Traes_6AL_EB6D01B8E.1, 4565.Traes_2DL_130D6031E.1, 4565.Traes_1BS_723922D171.1, 4565.Traes_3B_DBB113CF1.1,

5. **Analizar las vías metabólicas en las que participa el gen identificado.** En este caso se usa el programa KEGG PATHWAY: https://www.genome.jp/kegg/pathway.html .



Seleccionamos el mapa de vías metabólicas relevante, en este caso el asociado a la síntesis de fenilpropanoides.





La proteína de interés está remarcada en rojo con su número EC correspondiente (4.3.1.24). Con este mapa podemos visualizar las diferentes vías metabólicas en la que participa. Como habíamos visto previamente, nuestra enzima participa en la biosíntesis de **ácido cinámico.**

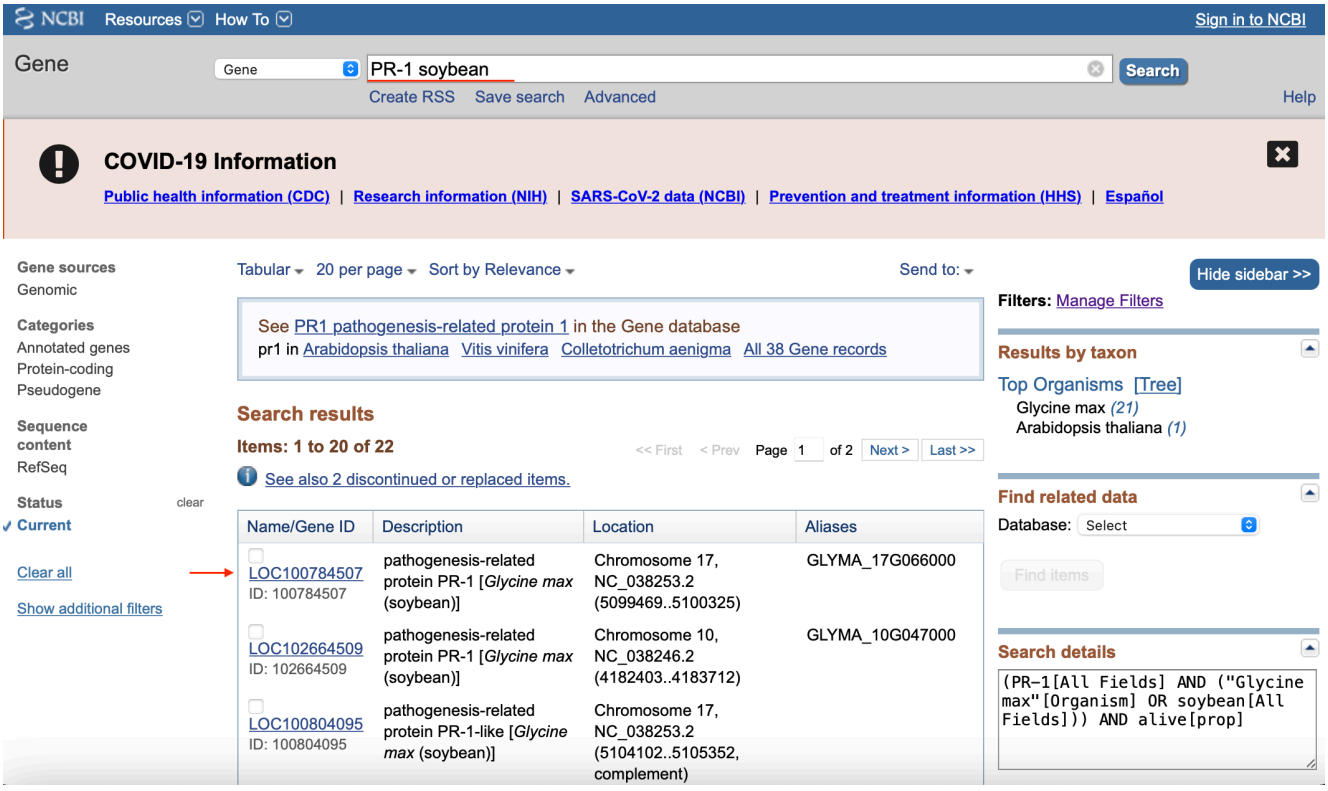
Un defecto en la secuencia que codifica para la enzima **Fenilalanina amonio liasa** lleva a que esta no pueda sintetizar **ácido cinámico** a partir de la **fenilalanina**, que luego no puede sintetizar **ácido cumárico**, cuyo rol es importante en la resistencia a enfermedades en las plantas.

A partir del **ácido cinámico**, además del **ácido cumárico**, se sintetizan la **cumarina** y el **cinamaldehído**, que también son **fenilpropanoides** que permiten a las plantas resistir enfermedades.

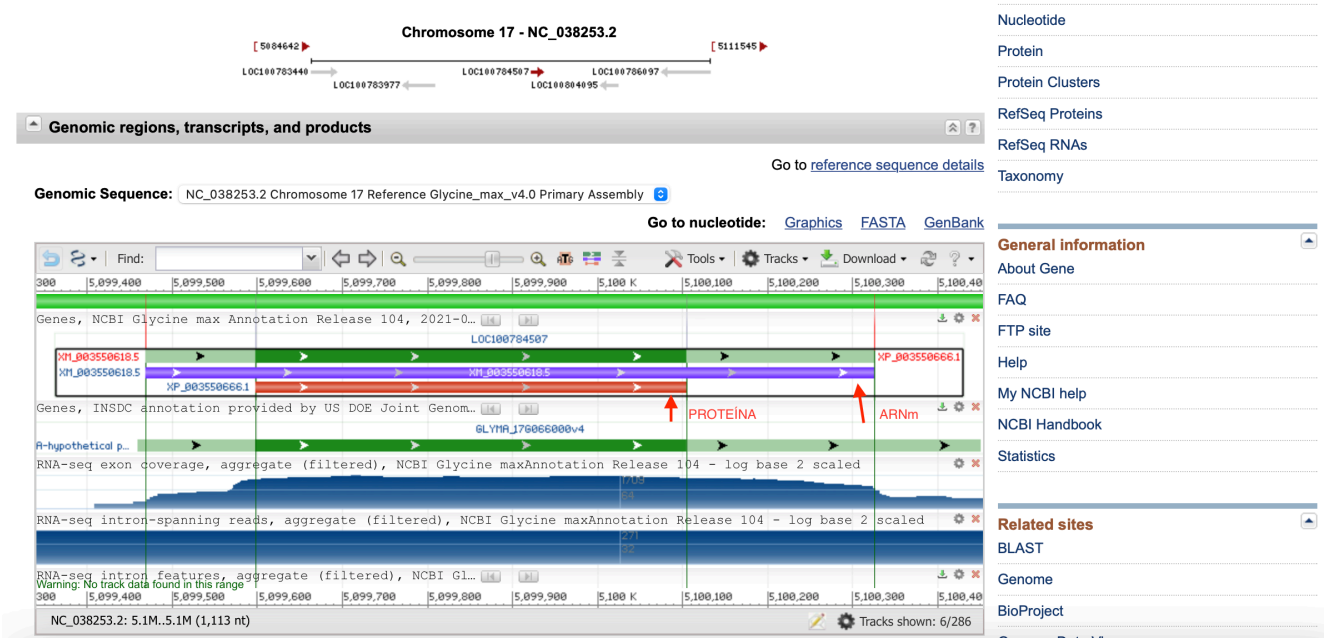
Cómo encontrar secuencias de ADN de genes de interés

1. **Usar una plataforma para buscar la proteína o gen de interés y encontrar secuencias de ADN asociadas**. Usamos la plataforma `Gene` de `NCBI`: https://www.ncbi.nlm.nih.gov/gene/.

   Nos interesa una serie de proteínas llamadas "**proteínas relacionadas con patogénesis**" , las cuales se abrevian como **`PR`** seguidas de un número, por ejemplo PR-1. Para acotar la búsqueda, conviene indicar una especie, como soybean en este caso.

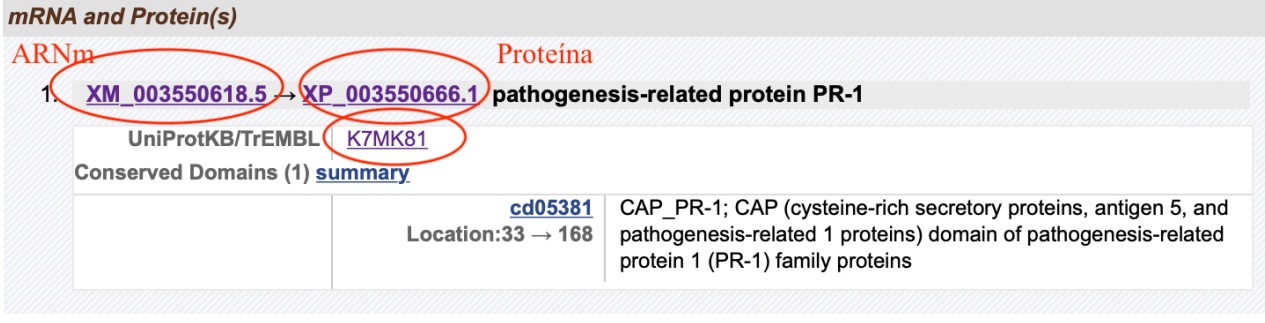


2. **Elegir una proteína y explorar**



¿Por qué la barra roja es más corta que la azul?

La barra roja representa la proteína, que se tradujo a partir del **codón de inicio** del **ARNm** hasta el **codón de terminación**, por eso la secuencia violeta del ARNm es más larga, ya que esta presenta todos los nucleótidos, más allá del marco de lectura.



ARNm:

PREDICTED: Glycine max pathogenesis-related protein PR-1 (LOC100784507), mRNA

NCBI Reference Sequence: XM_003550618.5

FASTA Graphics

Go to:

```
LOCUS       XM_003550618             857 bp    mRNA    linear   PLN 19-APR-2021
DEFINITION  PREDICTED: Glycine max pathogenesis-related protein PR-1
            (LOC100784507), mRNA.
ACCESSION   XM_003550618
VERSION     XM_003550618.5
DBLINK      BioProject: PRJNA48389
KEYWORDS    RefSeq.
SOURCE      Glycine max (soybean)
  ORGANISM  Glycine max
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliopsida; eudicotyledons; Gunneridae;
            Pentapetalae; rosids; fabids; Fabales; Fabaceae; Papilionoideae; 50
            kb inversion clade; NPAAA clade; indigoferoid/millettioid clade;
            Phaseoleae; Glycine; Glycine subgen. Soja.
COMMENT     MODEL REFSEQ:  This record is predicted by automated computational
            analysis. This record is derived from a genomic sequence
            (NC_038253.2) annotated using gene prediction method: Gnomon,
            supported by EST evidence.
            Also see:
                Documentation of NCBI's Annotation Process

            On Apr 19, 2021 this sequence version replaced XM_003550618.4.
```

Proteína:

```
LOCUS       XP_003550666             168 aa            linear   PLN 19-APR-2021
DEFINITION  pathogenesis-related protein PR-1 [Glycine max].
ACCESSION   XP_003550666
VERSION     XP_003550666.1
DBLINK      BioProject: PRJNA48389
DBSOURCE    REFSEQ: accession XM_003550618.5
KEYWORDS    RefSeq.
SOURCE      Glycine max (soybean)
  ORGANISM  Glycine max
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliopsida; eudicotyledons; Gunneridae;
            Pentapetalae; rosids; fabids; Fabales; Fabaceae; Papilionoideae; 50
            kb inversion clade; NPAAA clade; indigoferoid/millettioid clade;
            Phaseoleae; Glycine; Glycine subgen. Soja.
COMMENT     MODEL REFSEQ:  This record is predicted by automated computational
            analysis. This record is derived from a genomic sequence
            (NC_038253.2) annotated using gene prediction method: Gnomon,
            supported by EST evidence.
            Also see:
                Documentation of NCBI's Annotation Process

            ##Genome-Annotation-Data-START##
            Annotation Provider         :: NCBI
            Annotation Status           :: Full annotation
            Annotation Name             :: Glycine max Annotation Release 104
            Annotation Version          :: 104
            Annotation Pipeline         :: NCBI eukaryotic genome annotation
                                           pipeline
            Annotation Software Version :: 8.6
            Annotation Method           :: Best-placed RefSeq; Gnomon
            Features Annotated          :: Gene; mRNA; CDS; ncRNA
            ##Genome-Annotation-Data-END##
            COMPLETENESS: full length.
```

¿Qué es el péptido señal y cuál es su función?

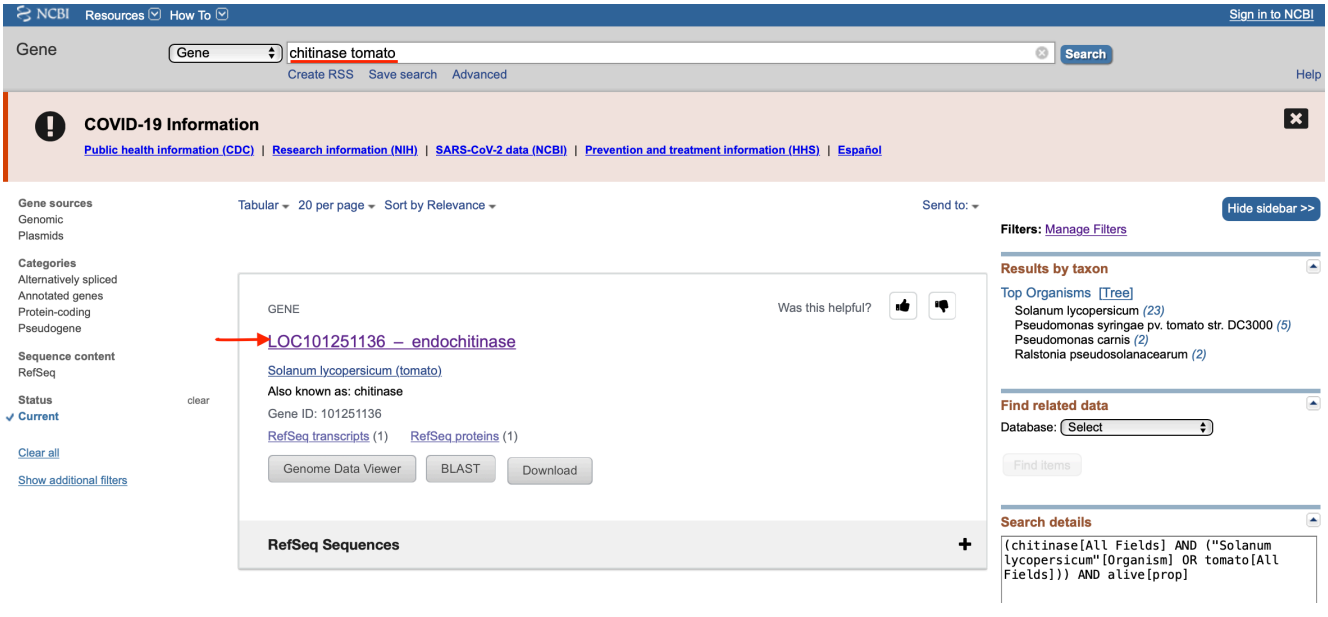
Cuando el ARNm llega al citoplasma unido a un ribosoma, comienza a traducirse a la proteína que corresponda al marco de lectura correspondiente. Una vez comenzada la traducción, a esta proteína se le asigna una **péptido señal** (secuencia de aminoácidos) la cual actúa como una especie de etiqueta que indica donde debe seguir traduciéndose esta proteína, ya sea acoplada al retículo endoplasmático rugoso, en el citoplasma, núcleo, etc. para que así la proteína llegue al lugar correcto.

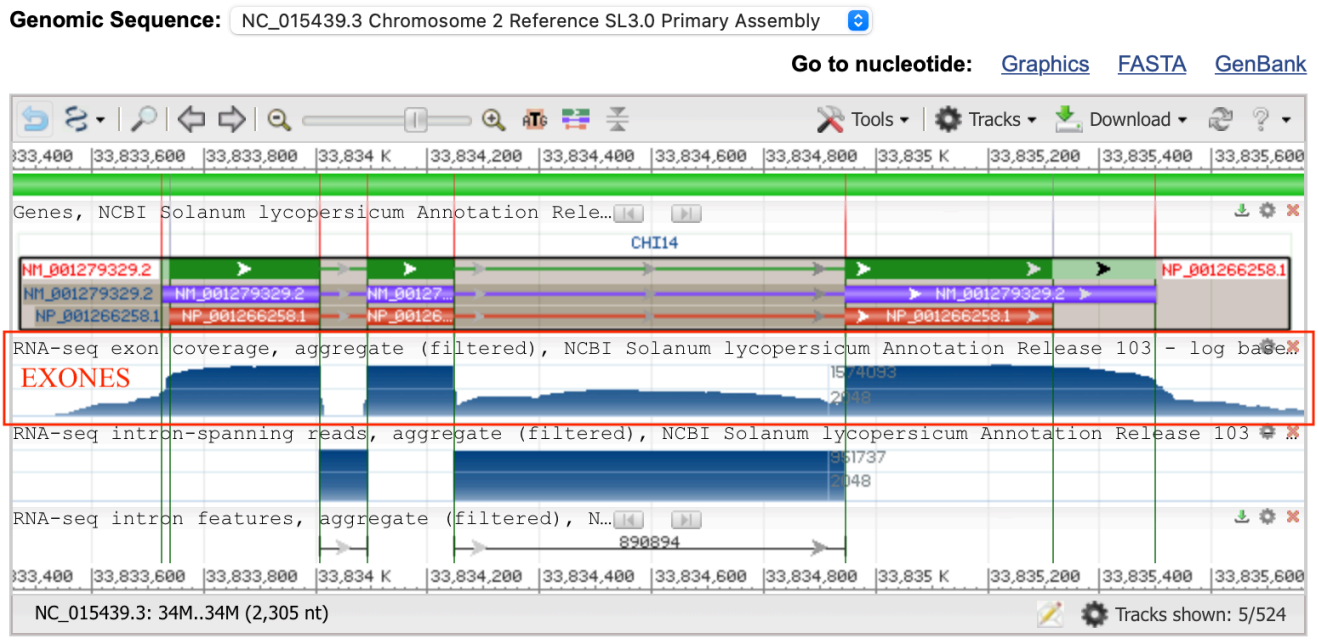
3. **Explorar UNIPROT**

Quitinasa del tomate:

Otras proteínas de interés son las **quitinasas** que son enzimas capaces de digerir las paredes celulares de los hongos.

Buscando nuevamente en el NCBI, podemos ver el esquema de la estructura de la quitinasa y en particular sus **exones e intrones**.





Las partes de la secuencia de genes que contienen la información para producir las proteínas se llaman **exones**, ya que se expresan, mientras que las partes de la secuencia del gen que no codifican se llaman **intrones**, porque están en medio o interfieren con los exones. El proceso por el cual los intrones son escindidos del transcripto de ARNm y los exones se unen para generar el **ARNm maduro** se denomina **Splicing**.