

# 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.

NIH U.S. National Library of Medicine  
National Center for Biotechnology Information

Log in

COVID-19 Information

Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS) | Español

BLAST®

Home Recent Results Saved Strategies Help

**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

N E W S

BLAST+ 2.12.0 is here!

We have made some improvements to how BLAST multi-threads and the amount of memory required by makeblastdb.

Tue, 13 Jul 2021 12:00:00 EST

More BLAST news...

**Web BLAST**

**Nucleotide BLAST** (nucleotide ▶ nucleotide)

**blastx** (translated nucleotide ▶ protein)

**tblastn** (protein ▶ translated nucleotide)

**Protein BLAST** (protein ▶ protein)

## Enter Query Sequence

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

2401 TGCTTGCTT TGAGCAAGGT CTGGTGAGTG ATTGAAAAAA  
ATGTTGTTGC AAGCTGTACC  
2461 TTGTATGTTT TTCAACAGGT GAATCTCACG TTTGATGCAT  
TGGATCAGAC

Query subrange [?](#)

From

To

Or, upload file

No se eligió archivo [?](#)

Job Title

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

Align two or more sequences [?](#)

## Choose Search Set

Database

Standard databases (nr etc.):  rRNA/ITS databases  Genomic + tra

[?](#)

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

Limit to

Optional

Sequences from type material

Entrez Query

Optional

You can also enter an Entrez query to limit search [?](#)

Enter an Entrez query to limit search [?](#)

## Program Selection

Optimize for

Highly similar sequences (megablast)  
 More dissimilar sequences (discontiguous megablast)  
 Somewhat similar sequences (blastn)

Choose a BLAST algorithm [?](#)

**BLAST**

Search **database Nucleotide collection (nr/nt)** using **Megablast (Optimize)**

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

## Sequences producing significant alignments

Download ▾ Now Select columns ▾ Show 100 ▾ ?

 select all 100 sequences selected

GenBank Graphics Distance tree of results New MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Zea mays phenylalanine ammonia-lyase (LOC100281532), mRNA	Zea mays	4636	4636	100%	0.0	100.00%	2510	NM_001154450.2
Zea mays full-length cDNA clone ZM_BFb0098J03 mRNA. Go to alignment for Zea mays phenylalanine ammonia-lyase (LOC100281532), mRNA	Zea mays	4595	4595	99%	0.0	99.96%	2491	BT068983.1
Zea mays clone 1580529 phenylalanine ammonia-lyase mRNA, complete cds	Zea mays	4337	4337	97%	0.0	98.61%	2479	EU957015.1
PREDICTED: Sorghum bicolor phenylalanine ammonia-lyase (LOC8066170), mRNA	Sorghum bicolor	3583	3583	98%	0.0	93.06%	2512	XM_002452434.2
Zea mays phenylalanine ammonia-lyase (LOC100381820), mRNA	Zea mays	3373	3373	94%	0.0	92.37%	2424	NM_001174615.1
PREDICTED: Sorghum bicolor phenylalanine ammonia-lyase (LOC8066169), mRNA	Sorghum bicolor	3356	3356	86%	0.0	94.77%	2539	XM_002454157.2
PREDICTED: Sorghum bicolor phenylalanine ammonia-lyase (LOC8066167), mRNA	Sorghum bicolor	3350	3350	86%	0.0	94.72%	2561	XM_002454155.2
Miscanthus x giganteus phenylalanine ammonia-lyase 2 (PAL2) mRNA, complete cds	Miscanthus x gi...	3319	3319	86%	0.0	94.41%	2157	KX084998.1
PREDICTED: Zea mays phenylalanine ammonia-lyase (LOC103627433), mRNA	Zea mays	3192	3192	84%	0.0	93.87%	2513	XM_008647730.4

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

[Download](#) ▾ [GenBank](#) [Graphics](#)
**Zea mays phenylalanine ammonia-lyase (LOC100281532), mRNA**
Sequence ID: **NM\_001154450.2** Length: 2510 Number of Matches: 1

Show report for NM\_001154450.2

Range 1: 1 to 2510 [GenBank](#) [Graphics](#)

▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
4636 bits(2510)	0.0	2510/2510(100%)	0/2510(0%)	Strand Plus/Plus
Query 1	GCACCATCCAGTCATCACGAGCTTCTGCACCAGATTAGCAGGCCATGCCCTACTTTT			60
Sbjct 1	GCACCATCCAGTCATCACGAGCTTCTGCACCAGATTAGCAGGCCATGCCCTACTTTT			60
Query 61	GGCTTCAAATCATTATTTACGGCGTACGTGCCTCTGTTCAAACCCCAGCCCCGCTGC			120
Sbjct 61	GGCTTCAAATCATTATTTACGGCGTACGTGCCTCTGTTCAAACCCCAGCCCCGCTGC			120
Query 121	AATGGAGTGCAGACAACGGCCGCGTGCCTGCTACCAACGGCGACTCCCTGTGCATGGCGCT			180
Sbjct 121	AATGGAGTGCAGACAACGGCCGCGTGCCTGCTACCAACGGCGACTCCCTGTGCATGGCGCT			180
Query 181	GCCCCGCGCCGCCGACCCGTTAAGTGGGGAAAGGC GGAGGAGATGATGGCAGCCA			240
Sbjct 181	GCCCCGCGCCGCCGACCCGTTAAGTGGGGAAAGGC GGAGGAGATGATGGCAGCCA			240
Query 241	CCTCGACGAGGTGAAGCGGATGGTGGCCGAGTACCGCCAGCCCCCTGGTGAAGATCGAGGG			300
Sbjct 241	CCTCGACGAGGTGAAGCGGATGGTGGCCGAGTACCGCCAGCCCCCTGGTGAAGATCGAGGG			300
Query 301	CGCCAGCCTCCGCATCGCGCAGGTGGCCGCTGTCGCCGCCGGCGCAGGGCGAGGCCGGGT			360
Sbjct 301	CGCCAGCCTCCGCATCGCGCAGGTGGCCGCTGTCGCCGCCGGCGCAGGGCGAGGCCGGGT			360
Query 361	CGAGCTCGACGAGTCCGCCGGCCGGTCAAGGC GAGCAGCGACTGGTCAGGGACAG			420
Sbjct 361	CGAGCTCGACGAGTCCGCCGGCCGGTCAAGGC GAGCAGCGACTGGTCAGGGACAG			420
Query 421	CATGATGAACGGCACCGACAGCTACGGCGTACCCACCGGTTGGCGCCACCTCCCACCG			480
Sbjct 421	CATGATGAACGGCACCGACAGCTACGGCGTACCCACCGGTTGGCGCCACCTCCCACCG			480

# 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: <a href="#">4577</a> " /chromosome="4" /map="4"		
gene		1..2510 /gene="LOC100281532" /gene_synonym="GRMZM2G063917" <del>/note="phenylalanine ammonia-lyase"</del> /db_xref="GeneID: <a href="#">100281532</a> "		
exon		1..2510 /gene="LOC100281532" /gene_synonym="GRMZM2G063917" /inference="alignment:Splign:2.1.0"		
CDS	122 2272	 /gene="LOC100281532" /gene_synonym="GRMZM2G063917" /EC_number=" <a href="#">4.3.1.5</a> " /note="phenylalanine ammonia lyase4" /codon_start=1 /product="phenylalanine ammonia-lyase" /protein_id=" <a href="#">NP_001147922.2</a> " /db_xref="GeneID: <a href="#">100281532</a> " /translation="MECDNGRVAATNGDSLCLMALPRAADPLNWGKAAEEMMGSHLDEV KRMVAEYRQPLVKIEGASLRIAQVAAVAAGAGEARVELDESARGRVKASSDWVRDSMM NGTDSYGVTGFGATSHRRRTKEGGALQRELIRFLNAGAFGIGTDAGHVLPAAETRAAM LVRINTLLQGYSGIRFEILEAIVKLLNANVTPCLPLRGTVTASGDLVPLSYIAGLVTG RENAVAVAPDGTKVNAAEAFRIADIQSGFFELQPKEGLAMVNGTAGVSGLASTVLFEA NVLAVALAEVLSAVFCCEVMNGKPEYTDHLTHKLKHHPQIEAAAIMEHILEGSSYMKLA KKLGELDPLMKPKQDRYALRTSPQWLGPQIEVIRASTKSIEREINSVNDNPLIDVARS KALHGGNFQGTPIGVSMNDNTRLAVAAIGKLMFAQFSELVNDYNNGLPSNLGGRNPS LDYGFKGAEIAMASYCSELQFLGNPVTNHVQSAEQHNQDVNSLGLISSRKTAEAIEIL KLMSSSTFLIALCQAVDRLRHIENENVSAVKSCVMTVAKKTLSTNSTGGLHVARFCEKDL LQEIEREAVFAYADPPCSANYPLMKKLRNVLVERALANGTAEFDAETSVFAKVAQFEE ELRTALPSAVEAARADEVNTAAIPNRIAECRSYPLYRFVREELGAVYLGEKTRSPG EELNKVLVAINQGKHIDPLLECLKEWNGEPLPIC"		

¿Que significa el alineamiento? ¿Como se ve si clickeas en proteínas 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 iniciación 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 4636 bits(2510)	Expect 0.0	Identities 2510/2510(100%)	Gaps 0/2510(0%)	Strand Plus/Plus
Query 1	GCACCATCCAGTGCATCACGAGCTCTTCGACCAGATTAGCAGGCCATGCCACTTTT			60
Sbjct 1	GCACCATCCAGTGCATCACGAGCTCTTCGACCAGATTAGCAGGCCATGCCACTTTT			60
Query 61	GGCTTCAAATCATTATTTACGGCGTACGTGCCTCTGTTCAAACCCCAGCCCCGCTGC			120
Sbjct 61	GGCTTCAAATCATTATTTACGGCGTACGTGCCTCTGTTCAAACCCCAGCCCCGCTGC			120
Query 121	AATGGAGTGCACACGGCGCGTCGCTGCTACCAACGGCGACTCCCTGTGCATGGCGCT			180
Sbjct 121	AATGGAGTGCACACGGCGCGTCGCTGCTACCAACGGCGACTCCCTGTGCATGGCGCT			180
Query 181	GCCCCGCGCCGCCGACCCGCTTAACACTGGGGAAAGCGGGGAGGAGATGATGGCAGCCA			240
Sbjct 181	GCCCCGCGCCGCCGACCCGCTTAACACTGGGGAAAGCGGGGAGGAGATGATGGCAGCCA			240
Query 241	CCTCGACGAGGTGAAGCGGATGGTGGCCGAGTACCGCCAGCCCCCTGGTGAAGATCGAGGG			300
Sbjct 241	CCTCGACGAGGTGAAGCGGATGGTGGCCGAGTACCGCCAGCCCCCTGGTGAAGATCGAGGG			300
Query 301	CGCCAGCCTCCGCATCGCGCAGGTGGCCGCTGTCGCCGCCGGCGGGCGAGGCCCGGGT			360
Sbjct 301	CGCCAGCCTCCGCATCGCGCAGGTGGCCGCTGTCGCCGCCGGCGGGCGAGGCCCGGGT			360
Query 361	CGAGCTCGACGAGTCCGCCGCCGGGTCAAGGGAGCAGCGACTGGTCAGGGACAG			420
Sbjct 361	CGAGCTCGACGAGTCCGCCGCCGGGTCAAGGGAGCAGCGACTGGTCAGGGACAG			420
Query 421	CATGATGAAACGGCACCGACAGCTACGGCGTCACCACCGGTTGGGCCACCTCCACCG			480
Sbjct 421	CATGATGAAACGGCACCGACAGCTACGGCGTCACCACCGGTTGGGCCACCTCCACCG			480
Query 2161	TCCCGGGAGGAGCTTAACAAGGTGCTCGTTGCCATCAACCAGGGCAAGCACATCGACCC			2220
Sbjct 2161	TCCCGGGAGGAGCTTAACAAGGTGCTCGTTGCCATCAACCAGGGCAAGCACATCGACCC			2220
Query 2221	GCTGCTCGAGTGCCTCAAGGAGTGGAACGGCGAGCCCCCTGCCCATCTGCTGAACAGAGAA			2280
Sbjct 2221	GCTGCTCGAGTGCCTCAAGGAGTGGAACGGCGAGCCCCCTGCCCATCTGCTGAACAGAGAA			2280
Query 2281	AATACAAGGAGCAGAAGACTGTATTTTAGCTAACGCACTTTTATTCTAATTAT			2340
Sbjct 2281	AATACAAGGAGCAGAAGACTGTATTTTAGCTAACGCACTTTTATTCTAATTAT			2340
Query 2161	TCCCGGGAGGAGCTTAACAAGGTGCTCGTTGCCATCAACCAGGGCAAGCACATCGACCC			2220
Sbjct 2161	TCCCGGGAGGAGCTTAACAAGGTGCTCGTTGCCATCAACCAGGGCAAGCACATCGACCC			2220
Query 2221	GCTGCTCGAGTGCCTCAAGGAGTGGAACGGCGAGCCCCCTGCCCATCTGCTGAACAGAGAA			2280
Sbjct 2221	GCTGCTCGAGTGCCTCAAGGAGTGGAACGGCGAGCCCCCTGCCCATCTGCTGAACAGAGAA			2280
Query 2281	AATACAAGGAGCAGAAGACTGTATTTTAGCTAACGCACTTTTATTCTAATTAT			2340
Sbjct 2281	AATACAAGGAGCAGAAGACTGTATTTTAGCTAACGCACTTTTATTCTAATTAT			2340
Query 2161	TCCCGGGAGGAGCTAAACAAGGTGCTCGTTGCCATCAACCAGGGCAAGCACATCGACCC			2220
Sbjct 2161	TCCCGGGAGGAGCTAAACAAGGTGCTCGTTGCCATCAACCAGGGCAAGCACATCGACCC			2220
Query 2221	GCTGCTCGAGTGCCTCAAGGAGTGGAACGGCGAGCCCCCTGCCCATCTGCTGAACAGAGAA			2280
Sbjct 2221	GCTGCTCGAGTGCCTCAAGGAGTGGAACGGCGAGCCCCCTGCCCATCTGCTGAACAGAGAA			2280
Query 2281	AATACAAGGAGCAGAAGACTGTATTTTAGCTAACGCACTTTTATTCTAATTAT			2340
Sbjct 2281	AATACAAGGAGCAGAAGACTGTATTTTAGCTAACGCACTTTTATTCTAATTAT			2340

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

The screenshot shows the UniProtKB 2021\_03 results page. At the top, a search bar contains the query "nm 001154450" with the label "LOCUS del gen". Below the search bar, there are links for BLAST, Align, Retrieve/ID mapping, Peptide search, and SPARQL. A navigation bar includes Help, Contact, Basket, and a search icon. The main content area displays "UniProtKB 2021\_03 results" and information about the database's two sections: Reviewed (Swiss-Prot) - Manually annotated and Unreviewed (TrEMBL) - Computationally analyzed. A table lists entries for COPL14 and B6SWAO, both identified as Phenylalanine ammonia-lyase from Zea mays (Maize). A red arrow points to the COPL14 entry.

The screenshot shows the detailed view for the COPL14 protein entry. The top navigation bar includes UniProtKB, BLAST, Align, Retrieve/ID mapping, Peptide search, SPARQL, Help, Contact, and Basket. The main content area is titled "UniProtKB - COPL14 (COPL14\_MAIZE)". On the left, a sidebar provides filtering options for Entry, Publications, Feature viewer, Function, PTM / Processing, Expression, Interaction, Structure, Family & Domains, Sequence, Similar proteins, and Cross-references. The central panel displays the protein's name (Phenylalanine ammonia-lyase), gene ID (100281532), organism (Zea mays (Maize)), and status (Unreviewed). It also shows its catalytic activity: L-phenylalanine + (E)-cinnamate + NH<sub>4</sub><sup>+</sup> → (E)-cinnamate + NH<sub>4</sub><sup>+</sup>. Below this, a chemical reaction diagram illustrates the enzyme's function, showing the conversion of L-phenylalanine and (E)-cinnamate to (E)-cinnamate and NH<sub>4</sub><sup>+</sup>.

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 **fenilalanina**. Esto es de particular importancia para nuestro análisis porque el **ácido cinámico** es un **fenilpropanoide**, como el **ácido cumárico**.

## GO - Molecular function<sup>i</sup>

- ammonia-lyase activity
- phenylalanine ammonia-lyase activity

Complete GO annotation on QuickGO ...

## GO - Biological process<sup>i</sup>

- cinnamic acid biosynthetic process
- L-phenylalanine catabolic process

Complete GO annotation on QuickGO ...

## Keywords<sup>i</sup>

Molecular function Lyase Imported

Biological process Phenylpropanoid metabolism ARBA annotation

## Enzyme and pathway databases

UniPathway<sup>i</sup> UPA00713; UER00725

## Names & Taxonomy<sup>i</sup>

Protein names <sup>i</sup>	Recommended name: <b>Phenylalanine ammonia-lyase</b> EC:4.3.1.24
Gene names <sup>i</sup>	Name: <b>100281532</b>
	ORF Names: ZEAMMB73_Zm00001d051166
Organism <sup>i</sup>	Zea mays (Maize)
Taxonomic identifier <sup>i</sup>	<b>4577</b> [NCBI]
Taxonomic lineage <sup>i</sup>	Eukaryota > Viridiplantae > Streptophyta > Embryophyta > Tracheophyta > Spermatophyta > Magnoliopsida > Liliopsida > Poales > Poaceae > PACMAD clade > Panicoideae > Andropogonodea > Andropogoneae > Tripsacinae > Zea
Proteomes <sup>i</sup>	UP000007305 Component <sup>i</sup> : 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: **líasas**.

**EC 4  
Líasas**

Adición o eliminación no hidrolítica de grupos de los substratos. Pueden romper los **enlaces C-C, C-N, C-O o C-S**.

$\text{RCOCOOH} \rightarrow \text{RCOH} + \text{CO}_2$

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

## Interaction<sup>i</sup>

### Protein-protein interaction databases

→ STRING<sup>i</sup> 4577.GRM2G063917\_P01

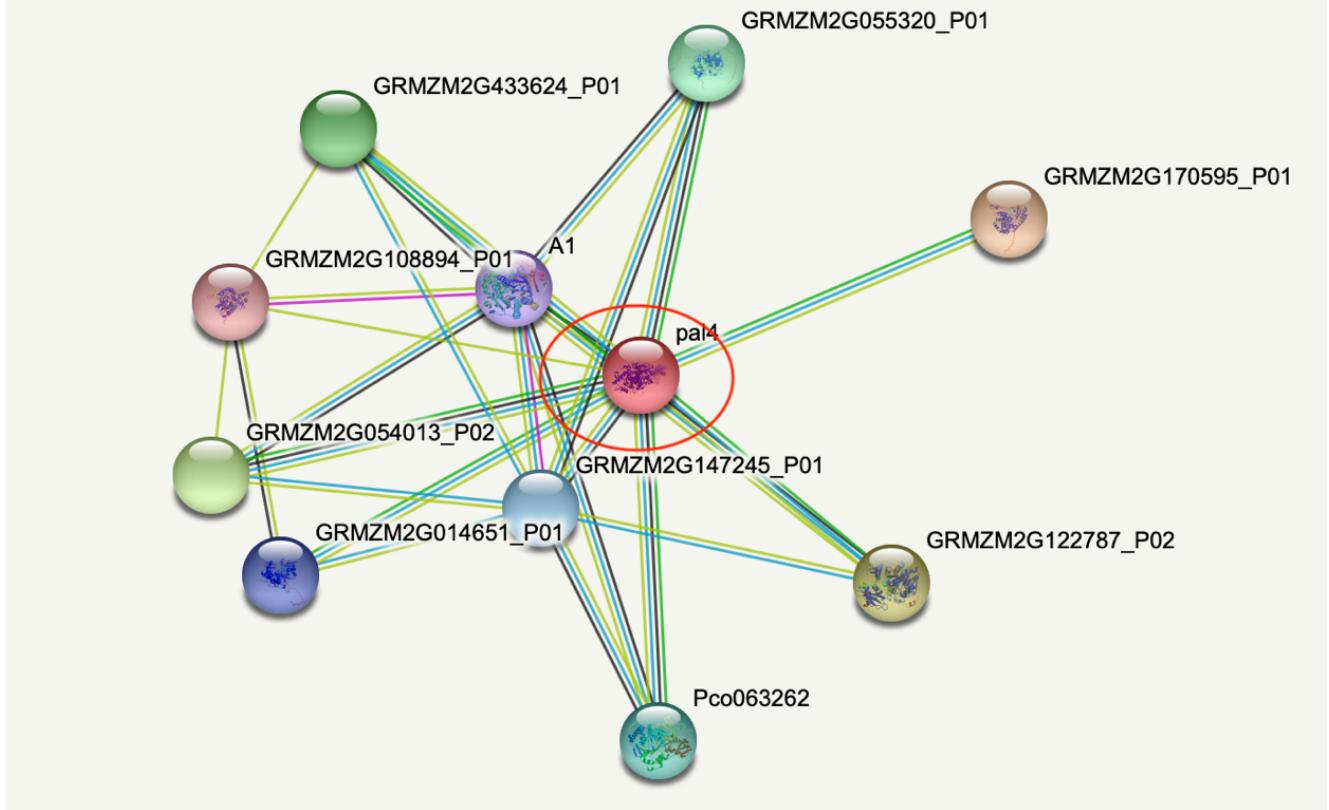
## Structure<sup>i</sup>

### 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-monoxygenase; Putative cytochrome P450 superfamily protein; Uncharacterized protein	0.826
GRMZM2G014651_P01	Uncharacterized protein	0.792
A1	Dihydroflavonol 4-reductase; Bifunctional enzyme involved in flavonoid metabolism; Belongs to the NAD(P)H... GRMZM2G108894_P01 Type III polyketide synthase B; Chalcone synthase; Uncharacterized protein ; Belongs to the chalcone/stil...	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.

Sequence similarities<sup>i</sup>

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

## Phylogenomic databases

eggNOG <sup>i</sup>	<a href="#">KOG0222</a> , Eukaryota
HOGENOM <sup>i</sup>	<a href="#">CLU_014801_3_0_1</a>
→ OMA <sup>i</sup>	<a href="#">KIMECRS</a>
OrthoDB <sup>i</sup>	<a href="#">923557at2759</a>

## Family and domain databases

CDD <sup>i</sup>	<a href="#">cd00332</a> , PAL-HAL, 1 hit
Gene3D <sup>i</sup>	<a href="#">1.10.274.20</a> , 1 hit <a href="#">1.10.275.10</a> , 1 hit
InterPro <sup>i</sup>	<a href="#">View protein in InterPro</a> <a href="#">IPR001106</a> , Aromatic_Lyase <a href="#">IPR024083</a> , Fumarase/histidase_N <a href="#">IPR008948</a> , L-Aspartase-like <a href="#">IPR022313</a> , Phe/His_NH3-lyase_AS <a href="#">IPR005922</a> , Phe_NH3-lyase <a href="#">IPR023144</a> , Phe_NH3-lyase_shielding_dom_sf
PANTHER <sup>i</sup>	<a href="#">PTHR10362</a> , PTHR10362, 1 hit
Pfam <sup>i</sup>	<a href="#">View protein in Pfam</a> <a href="#">PF00221</a> , Lyase_aromatic, 1 hit
SUPERFAM <sup>i</sup>	<a href="#">SSF48557</a> , SSF48557, 1 hit
TIGRFAMs <sup>i</sup>	<a href="#">TIGR01226</a> , phe_am_lyase, 1 hit
PROSITE <sup>i</sup>	<a href="#">View protein in PROSITE</a> <a href="#">PS00488</a> , 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	<a href="#">Triticum aestivum</a>	<a href="#">WHEAT111616</a>	<a href="#">A0A3B6QI72</a>	
E	<a href="#">Zea mays</a>	<a href="#">MAIZE48728</a>	<a href="#">C0PL14</a>	
E	<a href="#">Setaria italica</a>	<a href="#">SETIT02528</a>	<a href="#">K3YQD2</a>	
E	<a href="#">Cucumis sativus</a>	<a href="#">CUCSA18623</a>	<a href="#">A0A0A0KEL3</a>	
E	<a href="#">Manihot esculenta</a>	<a href="#">MANES03985</a>	<a href="#">A0A2C9VYR2</a>	
E	<a href="#">Selaginella moellendorffii</a>	<a href="#">SELML11505</a>	<a href="#">D8SER5</a>	

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

Ortholog	Organism
<b>pal4</b>	<i>Zea mays</i>
HAL	<i>Cricetulus griseus</i>
HAL	<i>Mesocricetus auratus</i>
XP_006988520.1	<i>Peromyscus maniculatus</i>
HAL	<i>Mus musculus</i>
FPSE_09706	<i>Fusarium pseudograminearum</i>
HAL	<i>Rattus norvegicus</i>
HAL	<i>Cavia porcellus</i>

909 more...

**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](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.

<i>Setaria italica</i>	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
<i>Sorghum bicolor</i>	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

## 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>.

KEGG Databases Mapper Auto annotation Kanehisa Lab

KEGG PATHWAY Database  
Wiring diagrams of molecular interactions, reactions and relations

KEGG2 PATHWAY BRITE MODULE KO GENES COMPOUND DISEASE DRUG

Select prefix zma Enter keywords 4.3.1.24 Go Help

número EC [ New pathway maps | Update history ]

**Pathway Maps**

KEGG PATHWAY is a collection of manually drawn pathway maps representing our knowledge of the molecular interaction, reaction and relation networks for:

- 1. Metabolism
- 2. Genetic Information Processing
- 3. Environmental Information Processing
- 4. Cellular Processes
- 5. Organismal Systems
- 6. Human Diseases
- 7. Drug Development

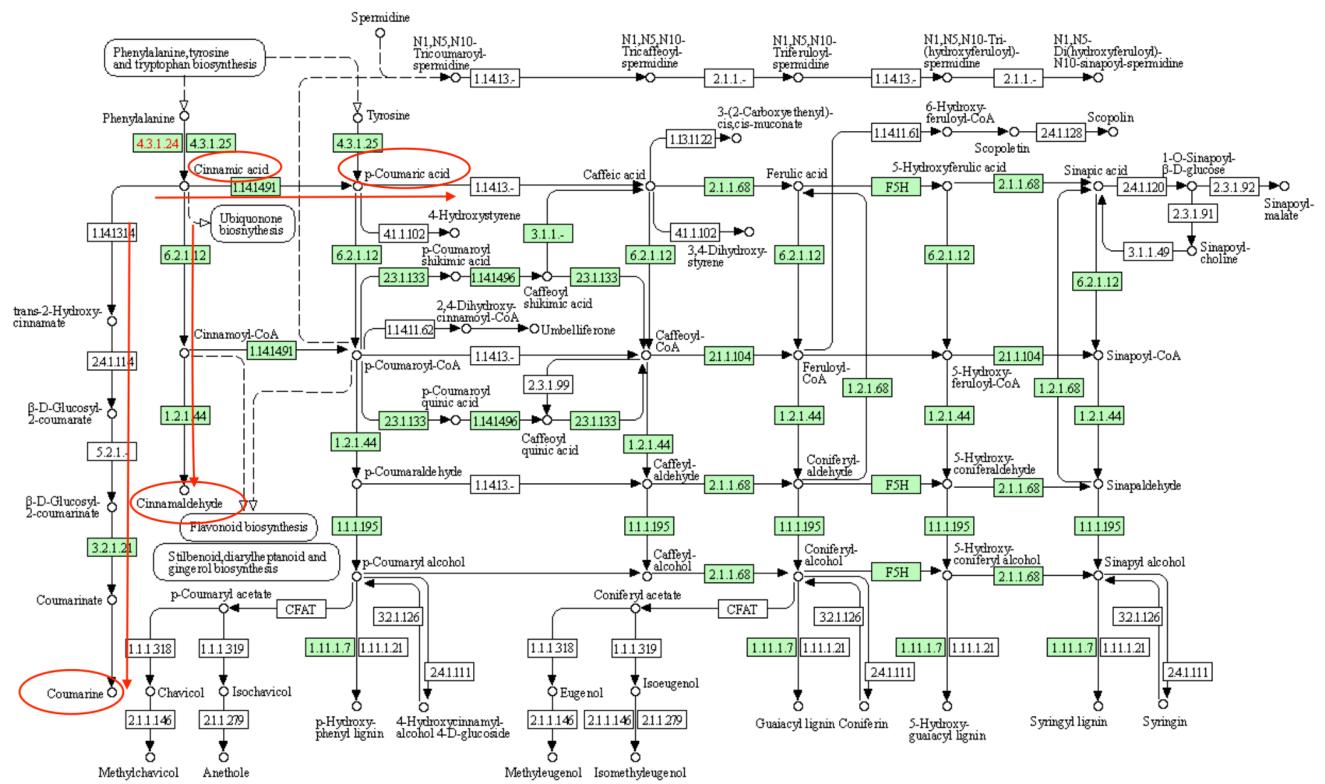
KEGG PATHWAY is the reference database for pathway mapping in [KEGG Mapper](#).

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

Pathway Text Search					
Number of entries in a page 20 Hide thumbnail					
Items : 1 - 2 of 2					
Entry	Thumbnail Image	Name	Description	Object	Legend
zma00360		Phenylalanine metabolism		C12621 (trans-3-Hydroxycinnamate) C00084 (Acetaldehyde) C00022 (Pyruvate) C03589 (4-Hydroxy-2-oxopen...)	...1.80 3.7.1.14 3.7.1.14 1.13.11.16 1.13.11.16 4.3.1.24 1.14.13.127 4.1.1.28 4.1.1.53 1.4.9.2 1.4.3.2...
zma00940		Phenylpropanoid biosynthesis	Phenylpropanoids are a group of plant secondary metabolites derived from phenylalanine and having a ...	C01527 (Scopolin) C00933 (Sinapine) C01175 (1-O-Sinapoyl-beta-D-glucose) C15806 (Syringyl lignin) C1...	...-Cumaraldehyde p-Coumaroyl-CoA Cinnamoyl-CoA 4.3.1.24 Caffeyl-alcohol Cinnamic acid trans-2-Hydroxy...

Items : 1 - 2 of 2

PHENYLPROPANOID BIOSYNTHESIS



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**.

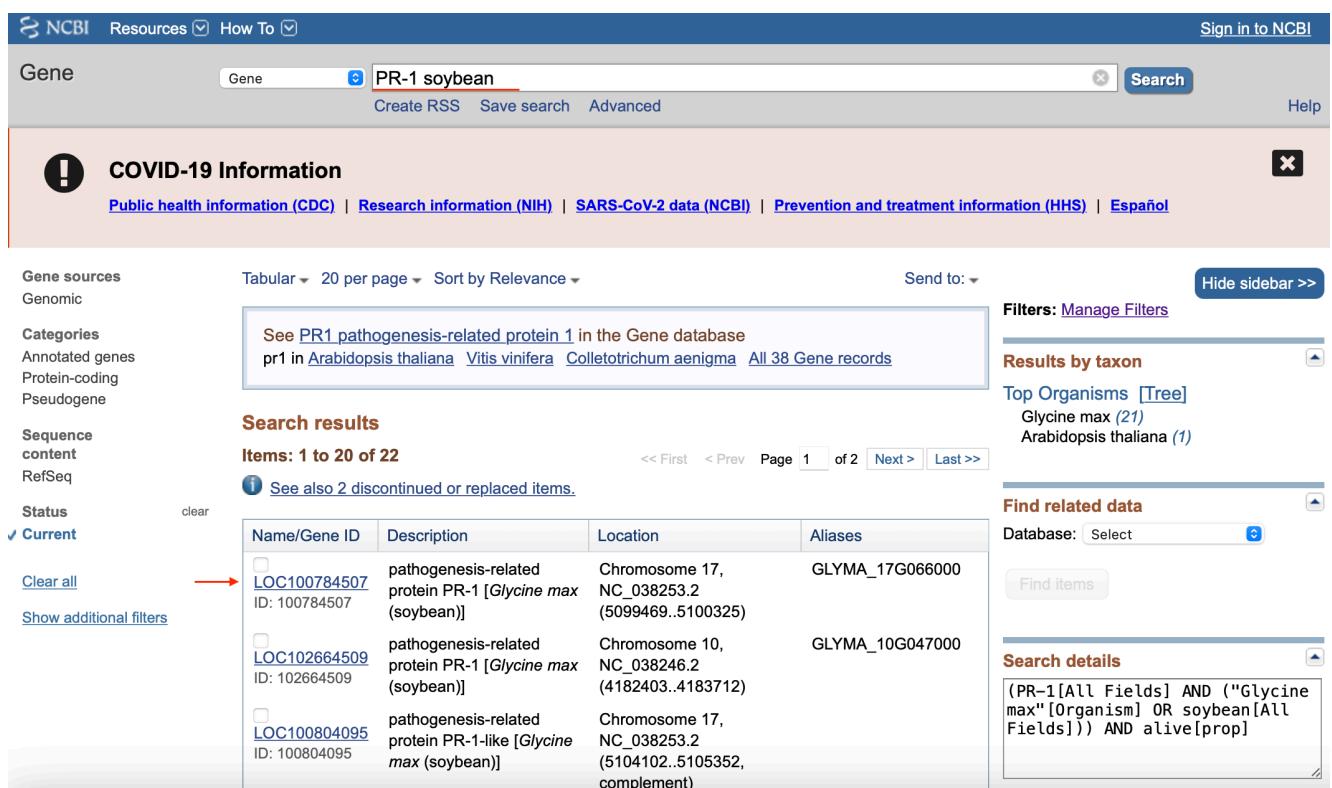
Un defecto en la secuencia que codifica para la enzima **Fenilalanina amonio liasa** lleva a que esta no pueda sintetizar **ácido cinámico** a partir de la **fenilalanina**, que luego no puede sintetizar **ácido cumárico**, cuyo rol es importante en la resistencia a enfermedades en las plantas.

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.

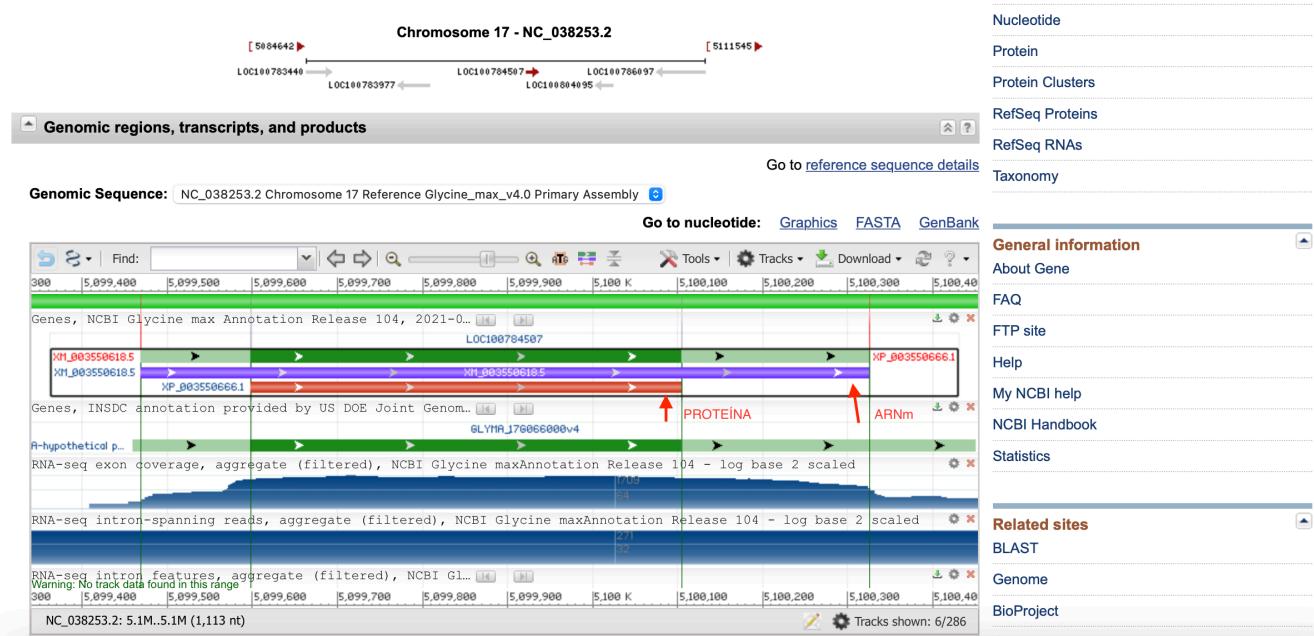


The screenshot shows the NCBI Gene search interface. The search term "PR-1 soybean" is entered in the search bar. The results page displays a table of gene records, with three entries shown:

Name/Gene ID	Description	Location	Aliases
<a href="#">LOC100784507</a> ID: 100784507	pathogenesis-related protein PR-1 [Glycine max (soybean)]	Chromosome 17, NC_038253.2 (5099469..5100325)	GLYMA_17G066000
<a href="#">LOC102664509</a> ID: 102664509	pathogenesis-related protein PR-1 [Glycine max (soybean)]	Chromosome 10, NC_038246.2 (4182403..4183712)	GLYMA_10G047000
<a href="#">LOC100804095</a> ID: 100804095	pathogenesis-related protein PR-1-like [Glycine max (soybean)]	Chromosome 17, NC_038253.2 (5104102..5105352, complement)	

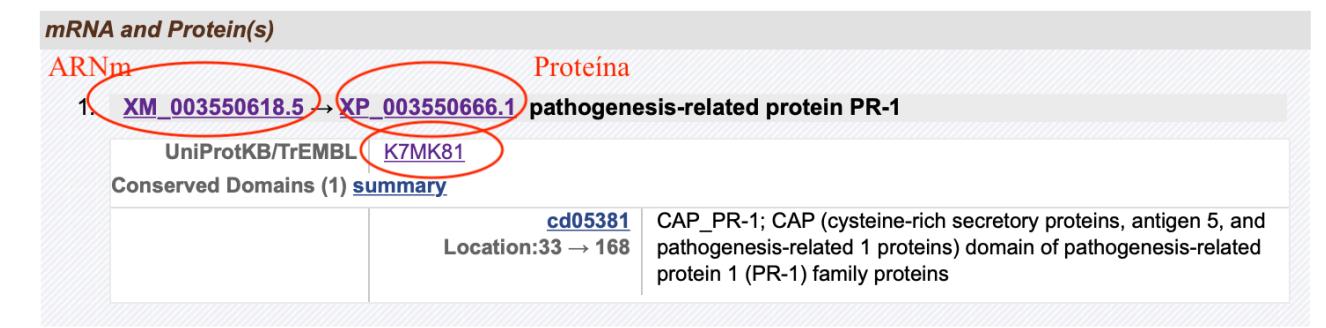
On the left sidebar, there are filters for Gene sources (Genomic), Categories (Annotated genes, Protein-coding, Pseudogene), Sequence content (RefSeq), and Status (Current). A red arrow points from the "Current" status filter to the "See also 2 discontinued or replaced items." link. The right sidebar includes sections for COVID-19 Information, Results by taxon (Top Organisms: Glycine max (21), Arabidopsis thaliana (1)), Find related data, and Search details.

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/millettoid 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/millettoid 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

Display Help video

BLAST Align Format Add to basket History

Add a publication Feedback

Entry

Publications

Feature viewer

Feature table

Protein SCP domain-containing protein

Gene 100784507

Organism Glycine max (Soybean) (Glycine hispida)

Status Unreviewed - Annotation score: 00000 - Protein predicted<sup>i</sup>

None

Function<sup>i</sup>

- Function
- Names & Taxonomy
- Subcell. location
- Pathol./Biotech
- PTM / Processing
- Expression

Probably involved in the defense reaction of plants against pathogens.

ARBA annotation

GO - Biological process<sup>i</sup>

- defense response Source: UniProtKB-KW
- response to biotic stimulus Source: UniProtKB-KW

[Complete GO annotation on QuickGO ...](#)

## Molecule processing

Feature key	Position(s)	Description	Actions	Graphical view	Length
Signal peptide <sup>i</sup>	1 – 17	Sequence analysis	Add  BLAST		17
Chain <sup>i</sup> (PRO_5014581846)	18 – 168	SCP domain-containing protein Sequence analysis	Add  BLAST		151

## 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**.

NCBI Resources How To Sign in to NCBI

Gene Gene chitinase tomato Search

Create RSS Save search Advanced

**COVID-19 Information**

[Public health information \(CDC\)](#) | [Research information \(NIH\)](#) | [SARS-CoV-2 data \(NCBI\)](#) | [Prevention and treatment information \(HHS\)](#) | [Español](#)

Gene sources Genomic Plasmids

Categories Alternatively spliced Annotated genes Protein-coding Pseudogene

Sequence content RefSeq

Status **Current** Clear Show additional filters

Tabular 20 per page Sort by Relevance Send to: Hide sidebar >>

**Filters: Manage Filters**

**Results by taxon**

**Top Organisms** [Tree]

Solanum lycopersicum (23)  
Pseudomonas syringae pv. tomato str. DC3000 (5)  
Pseudomonas carnis (2)  
Ralstonia pseudosolanacearum (2)

**Find related data**

Database: Select Find items

**Search details**

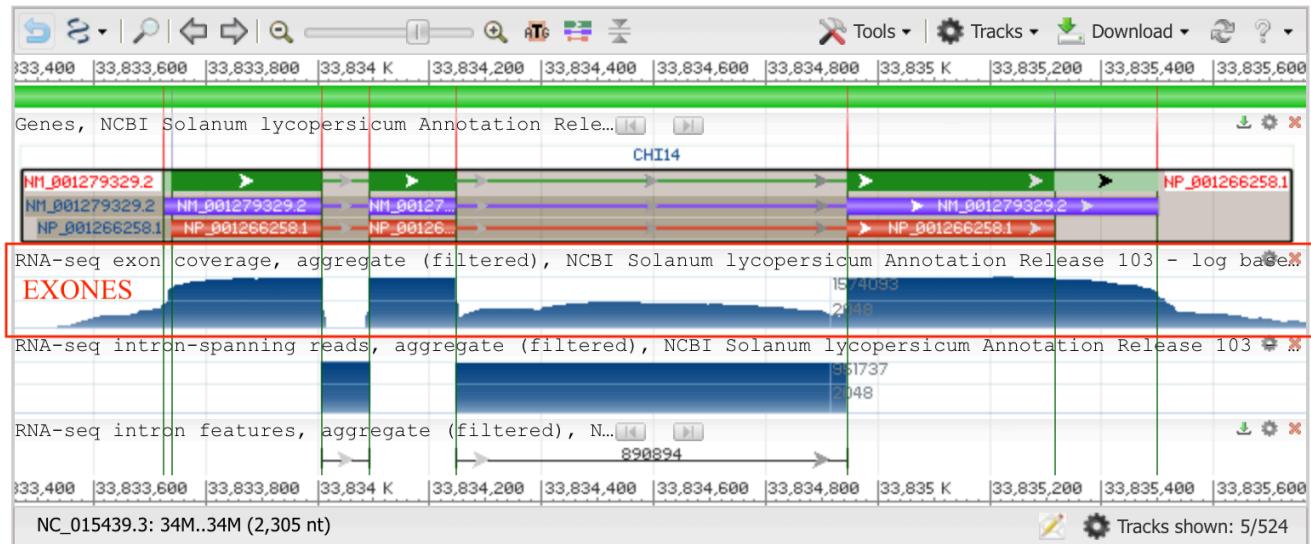
(chitinase[All Fields] AND ("Solanum lycopersicum"[Organism] OR tomato[All Fields])) AND alive[prop]

**RefSeq Sequences**

LOC101251136 – endochitinase

Solanum lycopersicum (tomato)  
Also known as: chitinase  
Gene ID: 101251136  
RefSeq transcripts (1) RefSeq proteins (1)

Genome Data Viewer BLAST Download



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**.