

LGN0232 - Genética Molecular

Bancos de Dados Biológicos

Antonio Figueira

CENA

figueira@cena.usp.br

Roteiro da Aula

1. Definição de Banco de Dados Biológicos

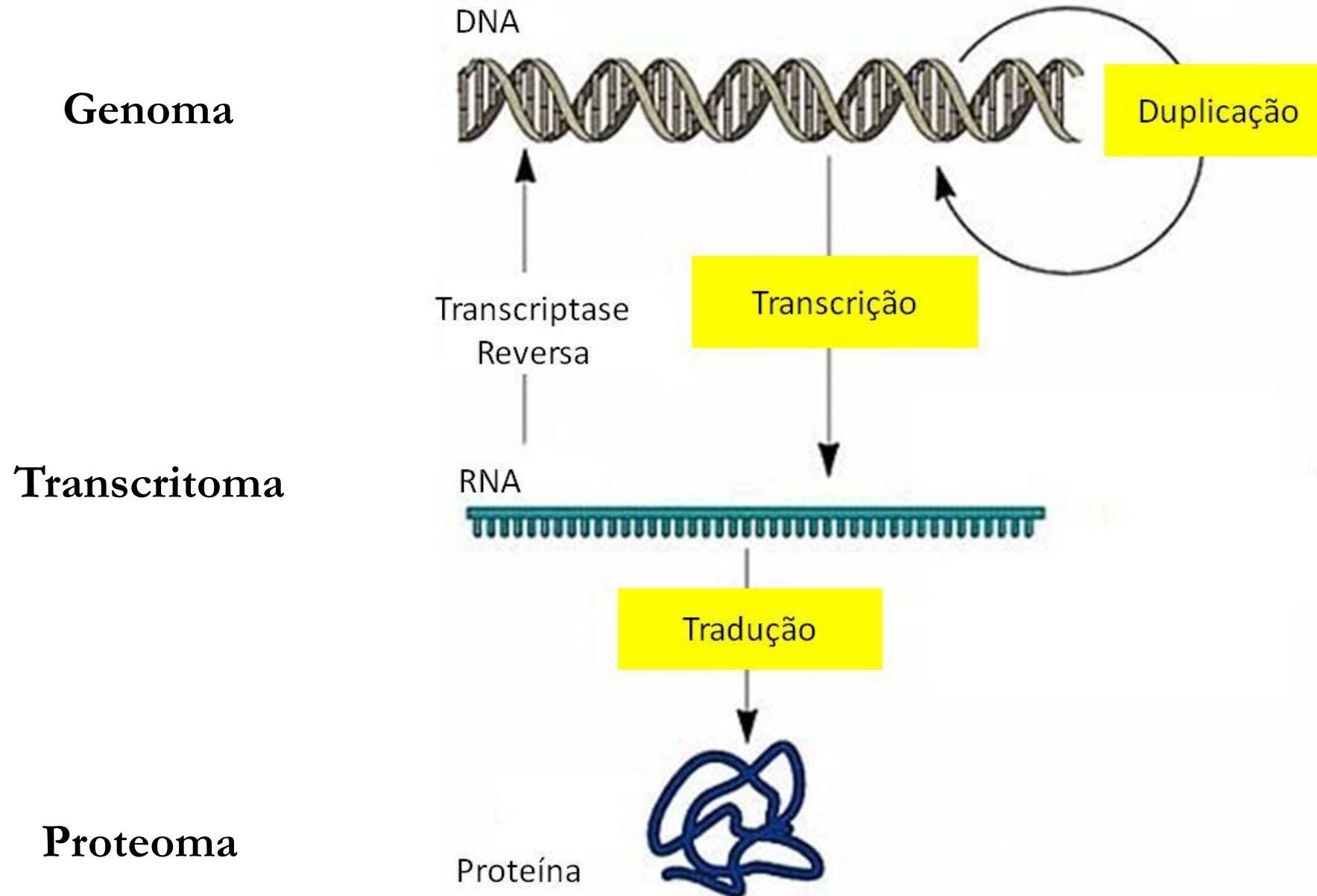
2. Bioinformática

3. Recursos oferecidos pelo NCBI

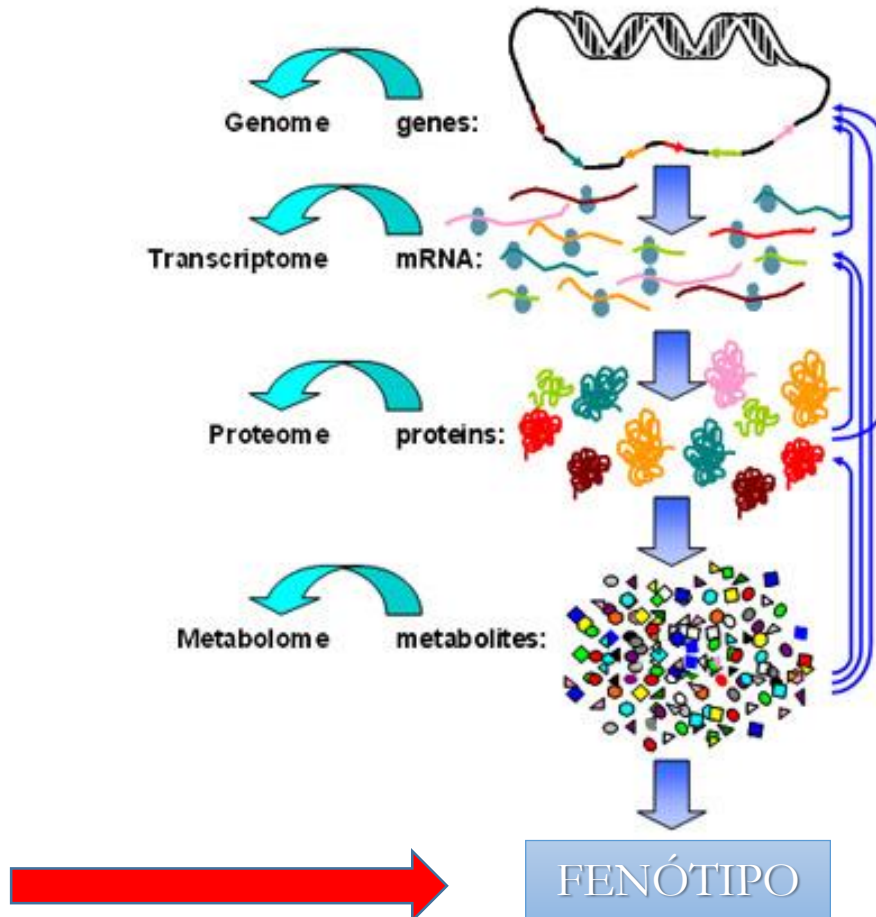
Forma de busca de informações: palavra-chave, sequências de nucleotídeos ou amino ácidos, espécies, artigos, autores,...

4. Utilização da plataforma BLAST

Dogma Central da Biologia Molecular



Avanços tecnológicos recentes permitiram o surgimento da Era das Ômicas



Ambiente



Projetos de Sequenciamento

Aumento do Número de Projetos de Sequenciamento
Novas tecnologias e redução de custos

+

Compartilhamento das Informações Obtidas



Banco de Dados Biológicos



Banco de Dados Biológicos (BDB)

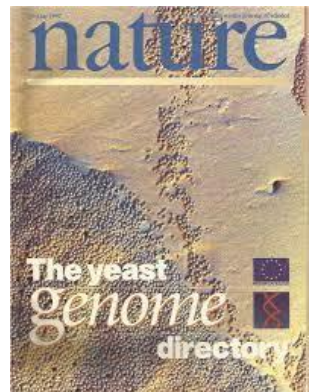
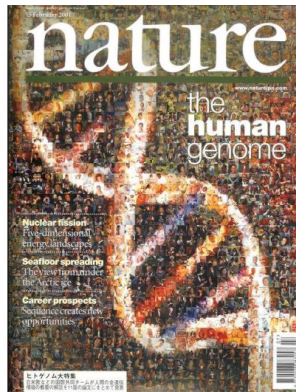
➤ O que são?

São repositórios *online* que **centralizam** as informações genéticas (sequências) de DNA, RNA ou proteína, dentre outros

➤ **Objetivos do BDS:** Centralizar os dados, torná-los públicos e permitir o acesso a informações geradas



Permite, por exemplo, comparar genes/genomas de espécies distintas.



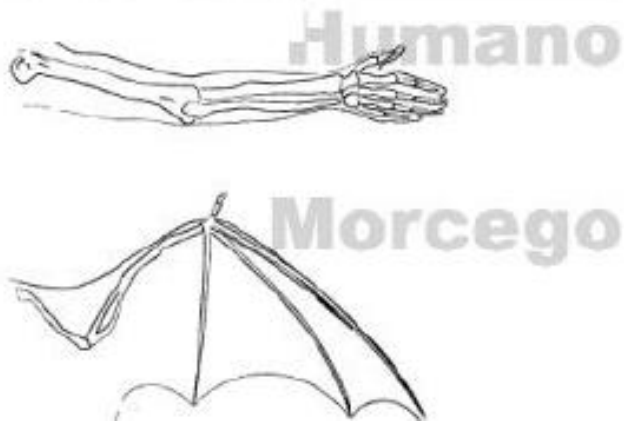
Homologia

Homologia – conceito fundamental na biologia

- **Homologia**: dividem a mesma ancestralidade com significado evolutivo
- **Permite inferências sobre a funcionalidade das sequencias identificadas**

Nature is prodigal in variety, but niggard in inovation - Charles Darwin

A análise de sequências objetiva encontrar similaridades importantes que permitam inferir sobre homologia



Exemplos:

Órgãos homólogos – asas de morcego e mãos de humanos (mesma origem)

Órgãos similares – asas de morcego e asas de borboleta (mesma função)

Bioinformática

Produção massiva de sequências de DNA, mRNA, proteínas

- A bioinformática consiste no desenvolvimento de métodos computacionais, matemáticos e estatísticos para **organizar e analisar** informações biológicas em grande escala e de maneira integrada

**Organização
e Armazenamento**



Bancos de Dados Biológicos

**Visualização e
Análise**



Ferramentas computacionais
Compreensão do significado biológico

>LT594788.1 Theobroma cacao genome assembly, chromosome: I

ATCGGCAGTGACGTTTTATGATGATGAGATCATTGCTCTTGACAGCCATTTAAACATTCCATGGTAGGA
AAGTTTTACGTATGCCCCGTTGAATGACATTAGGGTTGCTTTCAAAGGAATCGGGCTAGTGGGTGCAT
ATGAAATTCGTTGGTTGGATTATAAGCACATCCTGATTCATTTATCTAATGAGCAAGATCTGAATCATT
ATGGATGCGTCAAGCATGGTTCATTGCAAACCAGAAGATGAGAGTCTTTAAGTGGACTCCGGATTTCCAA
TCGAAAAGGGAATCCTTCTTGGTTCCCGTTTGGGTCTCATTTTCGAACCTGCGGGCTCATCTATATGAAA
AATCGACACTTCCGATGATTGCTAAGTCGGTGGGAGACCCTTTTTATTGATGAAGCTACGGCAAATGG
CACACGACCAAGTGTGGCCGAGTGTGTGTTGAGTACGACTGCCAGCAGCCCCCTTTGAACAGATCTGG
ATCGTGACTAGGGATAGAAGCACAGGAAACATCACTGGAGGATTTCAACAGAAAGTAGAGTTTGCCAGGC
TTCCTGACTATTGCAATCACTGTTGCCATGTGGGACATAGTATTGCAACATGTCTGGTGATGGGTACAG
TAAGGACAAGCCAAGAAAGGCACGGCCTAAGCCCCTGTGGATAAAAAGCAGGAAGATGATGATTGGAAA
AGAGAGAAAAGTAAGGAAACAGGTGATCTAATGGTTAATGGCGATAAAAGGAAAAATTCGATCCAAACAG
AATCGAAAAAGCAGAGCGTGAATGGGTGAAGGTTGAAAAGGGTGGCACAAGCGGGTTCAAGGATGCCCA
CGGCGTAGAAGTCAATCTGGAGAGTAGTGGAGCAGATCCCGTGCAGATCTCGAATGGTTTTAGGGTGCTA
GAAGCAATGGAGGATGGCGGGGATGTTAGATCCGCAAACAGGGGAGAACAGAGAAGGTGAACAGTACCA
TGCAATTTTTAAAAATATTTTTAGGGAGAAAAGAAAGGCAGTCGACGGAGATGGAAAGATGCTCGGGAAA
GATAAATGGCGACGAAACGACATTAGAAGCTTACCGATAAAACGGACTGCAGATGGAGTGAATCGGGAC
AAGCTAAAATCTTCTACAGTGGGTGTGATCGAGGGTCCAAAGCAGAAGGAGAGTGAGTTAAGCAAAGTT
CTGTGCAGACGTTGATGGCTGAAATTTGGCGGACAGGAGCAGATACTCACGAGAGTGTAGAAAAATATTGC
AGACTTTGATCGAGTTCAATGGGCGATGGATGCAGGTCGTGTGACGTCCTGGAAGGCAAAAAAAGAGC
AACAGAAAATTGAGGACCGACTGTCGGGGACGGCCGTGCAAGGTGATGGTCAGACAGTACCGGAGGTGCG
AACAAATGCTTGGGGAGTCCAAAACAGTGGGTGTACCGTCTAAACGTGGACGGTGAAGGTTGCTGAAGGG
TGGTGAATGTGCAGTTGAGTCAACTCGACAGTAATAGTGTAGTGAGTTCTCGTGGCTGTCTTAAACTC
GGTACTGTTCACTCTCATGTAGCCAACTCCCGTGCGGTACATGCAGTGAAGGAAGTATACACCGGTTGG
AAGAAAATGCTTACTAGGGGAACAGCAGCTAGTTCACGTGAAGTGATGGAAGAAAATGCAGAACACGA
TCCAAACTTGGGATCCAACTGGGTATATGTGGTTACAATAAGAAAATAAGTTCGGTTCCTTCATGTGCA
GGAATAATTCTGCTGACTTTCACGCACATTTGGAAGCAAACAACAGGAGAACAACAATCGAGGGC
AAGTAAATCAAATCGAACTGATGATAGCAGTAGATCAGTGCTCCATGTGGACTCGGGAGAGATTTTGG
CAGCCAGCATATTAATACCACCCCATGGTTTCCAGGAGAAGAAAATCCGATAGTGAAGTTATATATATC
CCTTCAGAGGATATTCTTTCAGAGAATGATGCTCATATGTTGATGGATGGGTCTGATGAAGAATCCATCT
CCAAGCAATTTACCACTAGAATTACCCATGATCAGTGCCCTGCTTGGAAATGTAAGGGGAGTACTGGA
AAAGCAATCCAAAGGAGAATTAAAAAACTGCAGATGATGCACCAAATAAAGATATTGGTTATCCTGGAAC
CAATGGTAACTGTTGATCGAATTGAATTTTTAGGAGAAAATTAGGCTTTGAGGGGGCGCCTTTAATTG
TTCTCAAAAAATTTGGATTTTTGGATGCACGGCATCACTTGACAACCAGGTTTGATCATCCCCAATGC
TTGCATGTTCAATTATGTTTCCCGTAGCTTCTGTCCCTATTGAAGCTTCATTTGTTATGCTAAATGTA
CTAGAATGGAACGACTTGCTTTATGGGATTTTATGAGACGTATTGCAGAGGATGTACAGGGTCTTGGCT
GGCTGGAGGGGACTTTAATGTTATTTAAGGTGAGAAGAGAGATTTTGGGTGCAGACCCACATACTGGA
GCCATGGAAGATTTTCAAATGCCTTACTTGATTGTGGGTTAGTAGATGCAGGGTTTGAAGGCAACAATT
TTACGTGGACTAACTCCCGATGTTCCAAAGATTAGATCGGATTCTCTATAACCCACAGTGGGTAGCTCA

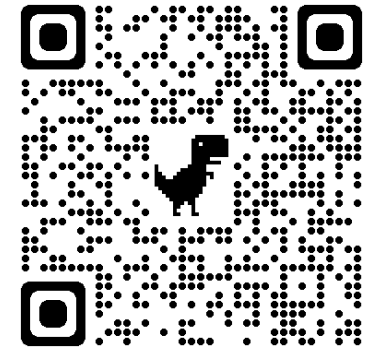
Genoma do cacauero

Banco de Dados

NCBI –National Center for Biotechnology Information

- <https://www.ncbi.nlm.nih.gov/>
- Iniciado em 1988 – ligado a biblioteca de medicina
- **Missão: melhor entendimento dos processos moleculares que afetam a saúde humana**
- *Understanding nature's mute but elegant language of living cells is the quest of modern molecular biology.*
- NCBI cria banco de dados públicos e recursos de biologia computacional e disseminação de informações

NCBI - PubMed



NIH National Library of Medicine
National Center for Biotechnology Information

Log in

PubMed.gov

Search

Advanced

PubMed® comprises more than 34 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full text content from PubMed Central and publisher web sites.

A screenshot of the PubMed.gov homepage. The header features the NIH logo and the text 'National Library of Medicine National Center for Biotechnology Information'. A 'Log in' button is in the top right. The main content area has the 'PubMed.gov' logo, a search bar, and a green 'Search' button. Below the search bar, the word 'Advanced' is visible. At the bottom, a paragraph describes the database's size and content.

Origem do NCBI -> National Library of Medicine

<https://pubmed.ncbi.nlm.nih.gov/>

NIH National Library of Medicine
National Center for Biotechnology Information

Log in

PubMed Advanced Search Builder

PubMed.gov
User Guide

Add terms to the query box

All Fields Enter a search term ADD Show Index

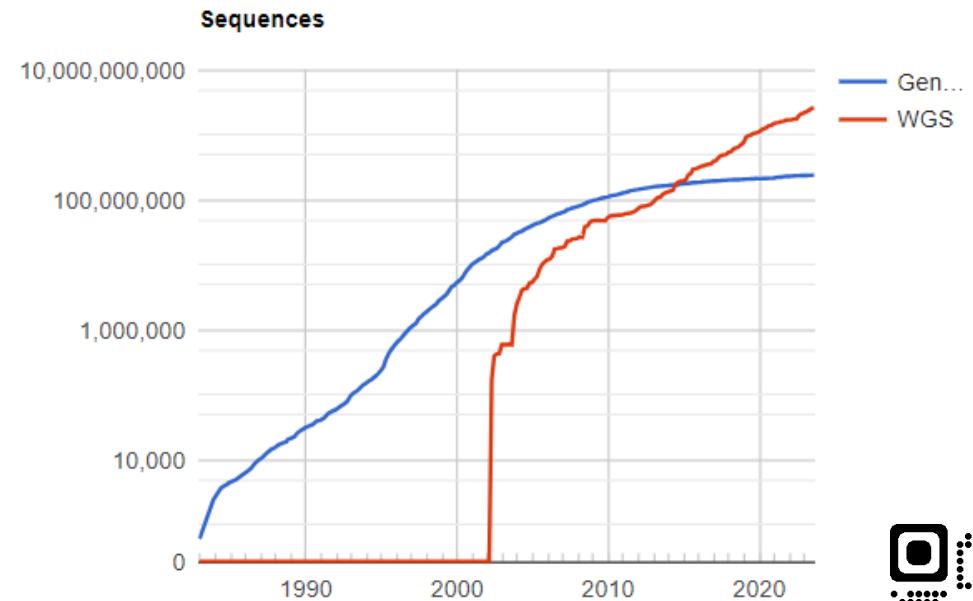
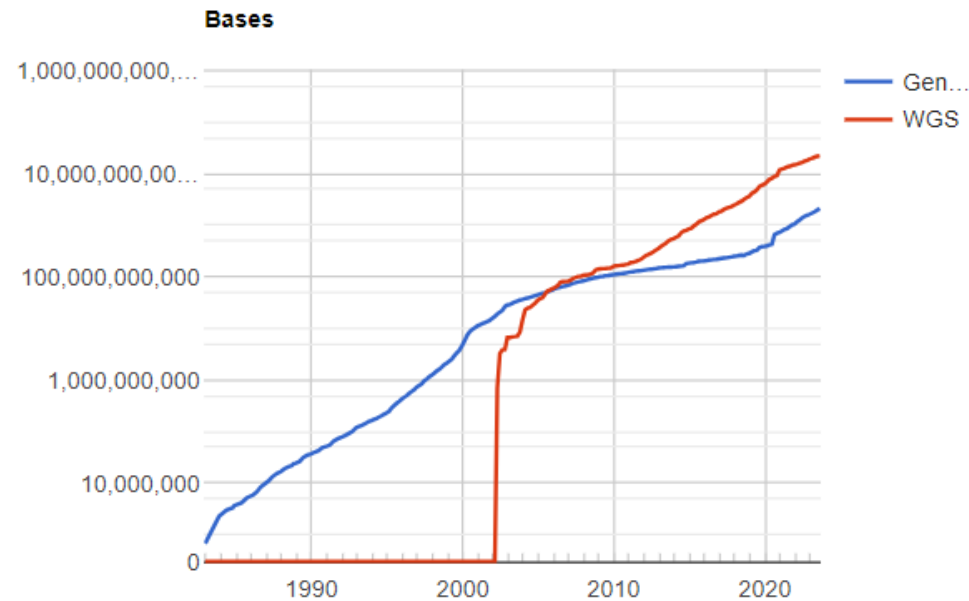
Query box

Enter / edit your search query here Search

A screenshot of the PubMed Advanced Search Builder interface. The header is identical to the homepage screenshot. Below the header, the title 'PubMed Advanced Search Builder' is displayed. To the right is the 'PubMed.gov User Guide' link. The main section is titled 'Add terms to the query box' and contains a dropdown menu set to 'All Fields', a text input field for 'Enter a search term', and an 'ADD' button. Below this is a 'Query box' with a text input field for 'Enter / edit your search query here' and a 'Search' button.

Histórico de Sequências no NCBI

GenBank and WGS Statistics



<https://www.ncbi.nlm.nih.gov/genbank/statistics/>



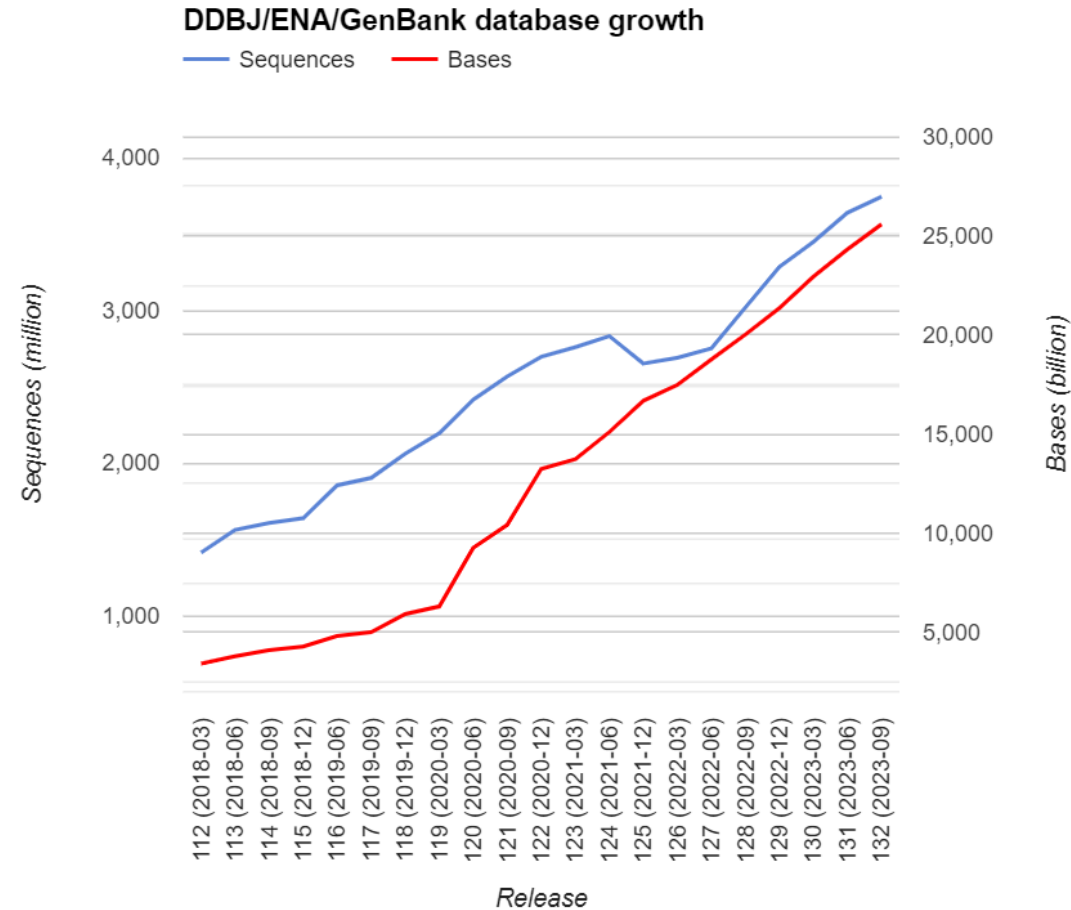
International Nucleotide Sequence Database Collaboration



Databases

Data type	DDBJ	EMBL-EBI	NCBI
Next Generation reads	Sequence Read Archive	European Nucleotide Archive	Sequence Read Archive
Assembled Sequences	DDBJ		GenBank
Samples	BioSample		BioSample
Studies	BioProject		BioProject

International Nucleotide Sequence Database Collaboration



NCBI

NCBI Home

Resource List (A-Z)

All Resources

Chemicals & Bioassays

Data & Software

DNA & RNA

Domains & Structures

Genes & Expression

Genetics & Medicine

Genomes & Maps

Homology

Literature

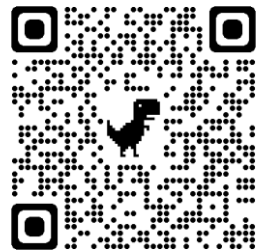
Proteins

Sequence Analysis

Taxonomy

Training & Tutorials

Variation



Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News & Blog](#)

Submit

Deposit data or manuscripts into NCBI databases



Download

Transfer NCBI data to your computer



Learn

Find help documents, attend a class or watch a tutorial



Develop

Use NCBI APIs and code libraries to build applications



Analyze

Identify an NCBI tool for your data analysis task



Research

Explore NCBI research and collaborative projects



Popular Resources

[PubMed](#)

[Bookshelf](#)

[PubMed Central](#)

[BLAST](#)

[Nucleotide](#)

[Genome](#)

[SNP](#)

[Gene](#)

[Protein](#)

[PubChem](#)

NCBI News & Blog

Announcing GenBank release 252.0

19 Oct 2022

Now over 3 billion records! GenBank release 252.0 (10/17/2022) is now available on the NCBI FTP site. This release has 20.35 trillion bases and 3.10 billion records. The current release has 240,539,282 traditional records containing 1,562,963,366,851 base pairs of sequence data. There are also 2,167,900,306 WGS records containing 18,231,960,808,828 base pairs of sequence data, 574,020,080 ... Continue

<https://www.ncbi.nlm.nih.gov/>

NCBI



U.S. National Library of Medicine

NCBI National Center for Biotechnology Information

Sign in to NCBI

NCBI HOME LITERATURE HEALTH GENOMES GENES PROTEINS CHEMICALS POPULAR RESOURCES ▼

All Databases ▼

Search NCBI

Search

Analyze

NCBI provides a wide variety of data analysis tools that allow users to manipulate, align, visualize and evaluate biological data.

Selected Analysis Tools

All Tools

Literature

Health

Genomes

Genes

Proteins

Chemicals

Filter this table

Tools	Description
Amino Acid Explorer	Explores amino acid properties, substitutions and functions
Assembly Archive	Links the raw sequence information found in the Trace Archive with assembly information found in GenBank/EMBL/DBJ
Basic Local Alignment Search Tool (BLAST)	Finds regions of local similarity between biological sequences
Batch Entrez	Retrieves records specified in an uploaded file of identifiers
BioAssay Services	Tools that summarize the biological test results in the PubChem database
BLAST Link (BLink)	Displays the results of a pre-computed BLAST search of a protein against all other protein sequences at NCBI



<https://www.ncbi.nlm.nih.gov/home/analyze/>

NCBI - TaxBrowser



The image shows the top navigation bar of the NCBI Taxonomy Browser. It features the NCBI logo on the left, followed by icons for various biological entities: a mushroom, a virus, a fish, a bacterium, and a flower. The text 'Taxonomy Browser' is displayed in blue. Below the icons is a black navigation bar with white text for 'Entrez', 'PubMed', 'Nucleotide', 'Protein', 'Genome', 'Structure', 'PMC', 'Taxonomy', and 'BioCollections'. Underneath is a search bar with a text input field, a dropdown menu set to 'as complete name', a checked 'lock' checkbox, and 'Go' and 'Clear' buttons. Below the search bar is a 'Display' section with a dropdown set to '3' and a 'levels using filter:' dropdown set to 'none'.

The "Token set" option returns longer names that include the search terms, e.g., hybrid taxa. See what happens if you query "Bos taurus" using the "Complete match" option versus the "Set of tokens" option. The "Phonetic search" option can be used when you are not sure about the exact spelling of an organism name. It tries to find the phonetically closest strings (try "Drozofila" as an example).

This is the top level of the taxonomy database maintained by NCBI/GenBank. You can explore any of the taxa listed below by clicking it.

- [Archaea](#)
- [Bacteria](#)
- [Eukaryota](#)
- [Viruses](#)
- [Other](#)
- [Unclassified](#)

These are direct links to some of the organisms commonly used in molecular research projects:

[Arabidopsis thaliana](#)

[Bos taurus](#)

[Caenorhabditis elegans](#)

[Chlamydomonas reinhardtii](#)

[Danio rerio \(zebrafish\)](#)

[Dictyostelium discoideum](#)

[Drosophila melanogaster](#)

[Escherichia coli](#)

[Hepatitis C virus](#)

[Homo sapiens](#)

[Mus musculus](#)

[Mycoplasma pneumoniae](#)

[Oryza sativa](#)

[Plasmodium falciparum](#)

[Pneumocystis carinii](#)

[Rattus norvegicus](#)

[Saccharomyces cerevisiae](#)

[Schizosaccharomyces pombe](#)

[Takifugu rubripes](#)

[Xenopus laevis](#)

[Zea mays](#)



<https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Root>

NCBI - Genomas



NCBI Sequence Set Browser Sign In to NCBI

Facet Panel

Available Facets

- Type
- Source database
- Targeted Locus Name
- DIV
- Status
- Organism
- Bioproject
- Biosample
- Strain
- Breed
- Cultivar
- Isolate
- Tissue Type
- Host
- Isolation Source

Taxonomic Groups

- land plants [4906]
 - vascular plants [4869]
 - seed plants [4820]
 - flowering plants [4649]
 - eu dicots [3279]
 - monocots [1258]
 - other [112]
 - other [171]
 - ferns [34]
 - club-mosses [15]
 - mosses [22]
 - liverworts [13]
 - other [1]

Description

This site is for browsing WGS (Whole Genome Shotgun) genomes, TSA (Transcriptome Shotgun Assemblies) and TLS (Targeted Locus Study) sets. WGS sequences are incomplete genomes that have been sequenced by a whole genome shotgun strategy. TSA sequences are transcript sequences that have been computationally assembled from primary RNA sequence data. TLS sequences are large-scale marker gene sequencing studies.

Please consult [WGS Submission](#) or [TSA Submission](#) pages for more details.

Project type:

Search – search in all fields. Use wildcard "*" to search in the middle of a field's text.

Term

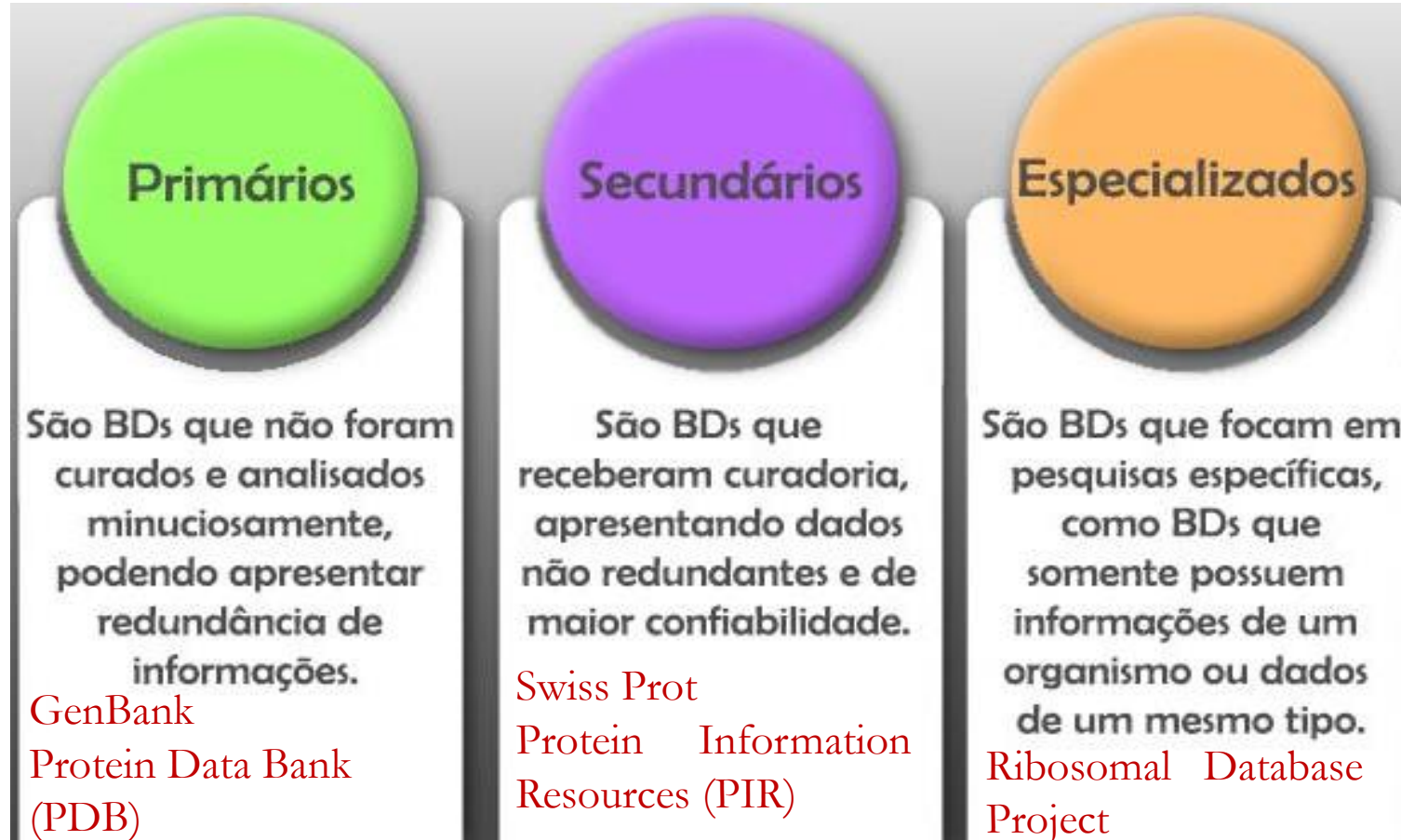
Found 4,908 projects [Download](#) [Columns List](#) Page [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) ... [99](#) (per page)

#	Prefix	Type	Targeted Locus Name	DIV	Organism	Bioproject	Biosample	Keywords	Intraspecific Name	Other Source	Contigs				Scaffolds			
											Total Length (Mbases)	#	# Prot	Has Annot	#	# Prot	Total Length (Mbases)	Has Annot
1	AAAA01	WGS		PLN	Oryza sativa (indica cultivar-group)			WGS	cultivar: 93-11		363.3	103,044						
2	AAAA02	WGS		PLN	Oryza sativa Indica Group	PRJNA361	SAMN02953581	WGS	cultivar: 93-11		410.7	50,230			3,095	37,358	769.7	Y
3	AACV01	WGS		PLN	Oryza sativa Japonica Group	PRJNA13139	SAMN02953597	WGS	cultivar: Nipponbare		374.2	35,047			1,436	35,394	717.2	Y

<https://www.ncbi.nlm.nih.gov/Traces/wgs/>

Banco de Dados Biológicos (BDB)

Tipos de Bancos de Dados Biológicos



Tipos de Bancos de Dados

Nível de Curadoria:

- **Preliminar** – sequências não terminadas

- localizadas nos centros de sequenciamento

- **Arquivo** – repositório da informação

- **redundante** (várias sequências do mesmo gene)

- submissor mantém o controle editorial sobre registros

- **Com Curadoria** – não redundante – ex. RefSeq NCBI

<https://www.ncbi.nlm.nih.gov/refseq/> →

- cada registro pretende conter conhecimento sobre a sequencia

- **Revisado** - Kyoto Encyclopedia of Genes and Genomes - KEGG

Genes, genomas, enzimas, rotas metabólicas

Announcements

September 11, 2023
RefSeq Release 220 is available for FTP

This release includes:

Proteins: 289,333,423

Transcripts: 56,423,426

Organisms: 141,099

Available at: <ftp://ftp.ncbi.nlm.nih.gov/refseq/release/>

Documentation: [Release Notes](#)

See [previous announcements](#), follow [NCBI on Twitter](#), or subscribe to [NCBI's refseq-announce mail list](#) to receive announcements.



Outros bancos específicos..

Proteínas

<https://www.expasy.org/> - Expasy – Instituto Suíço de Bioinformática

<https://proteininformationresource.org/>

<https://www.uniprot.org/>

Estrutura de Proteínas

Protein Data Bank - <https://www.rcsb.org/>

Structure (NCBI) - <https://www.ncbi.nlm.nih.gov/structure>

Outros bancos específicos..

Genomas espécie-específicos

- <http://www.yeastgenome.org/>
- <http://flybase.org/>
- <http://www.maizegdb.org/>
- <http://rice.plantbiology.msu.edu/>
- <https://solgenomics.net/>
- <https://cocoa-genome-hub.southgreen.fr/>

NCBI (*National Center for Biotechnology Information*) – fundado em 1988

O website <https://www.ncbi.nlm.nih.gov/> foi criado em 1994

É de uso gratuito e acolhe diversos BDBs, separados por categorias:

Literature — Repositório de artigos científicos, livros, entre outros. Um dos bancos mais utilizados dessa categoria é o **PubMed Central**

Genes — São encontradas sequências gênicas e **anotações** para estudo de estrutura de ortólogos, expressão e evolução

Genomes — Possui bancos de sequências genômicas, dados de genômica funcional e origem de amostras biológicas. Um dos principais bancos dessa categoria é o **Nucleotide**, que tem o **GenBank** como um dos seus principais componentes

Proteins — Apresenta dados como sequências proteicas, estruturas tridimensionais (3D) e domínios proteicos

BLAST — É uma ferramenta que realiza consultas em diferentes bancos de dados, como **Nucleotide** e **Protein**

PubChem — Repositório de informações químicas, rotas metabólicas e ferramentas para screening de atividade biológica

Busca nos bancos de dados

- **Por texto** – palavra chave, número das sequências, espécie, gênero,...
- **Por sequência de nucleotídeos ou amino ácidos**
 - Uso de programa específico - BLAST
 - *Basic Local Alignment Search Tool*
 - *BLAST* para nucleotídeos e amino ácidos



All Databases ▼

Search NCBI

Search

Analyze

NCBI provides a wide variety of data analysis tools that allow users to manipulate, align, visualize and evaluate biological data.



Selected Analysis Tools

All Tools

Literature

Health

Genomes

Genes

Proteins

Chemicals

Filter this table

Tools	Description
Amino Acid Explorer	Explores amino acid properties, substitutions and functions
Assembly Archive	Links the raw sequence information found in the Trace Archive with assembly information found in GenBank/EMBL/DDBJ
Basic Local Alignment Search Tool (BLAST)	Finds regions of local similarity between biological sequences
Batch Entrez	Retrieves records specified in an uploaded file of identifiers
BioAssay Services	Tools that summarize the biological test results in the PubChem database
BLAST Link (BLink)	Displays the results of a pre-computed BLAST search of a protein against all other protein sequences at NCBI
BLAST Microbial Genomes	Finds regions of local similarity between query sequences and sequences from complete microbial genomes



All Databases ▼

cp4 epsps agrobacterium |

Search

Analyze

NCBI provides a wide variety of data analysis tools that allow users to manipulate, align, visualize and evaluate biological data.

Selected Analysis Tools

All Tools

Literature

Health

Genomes

Genes

Proteins

Chemicals

Filter this table

Tools	Description
Amino Acid Explorer	Explores amino acid properties, substitutions and functions
Assembly Archive	Links the raw sequence information found in the Trace Archive with assembly information found in GenBank/EMBL/DDBJ
Basic Local Alignment Search Tool (BLAST)	Finds regions of local similarity between biological sequences
Batch Entrez	Retrieves records specified in an uploaded file of identifiers
BioAssay Services	Tools that summarize the biological test results in the PubChem database
BLAST Link (BLink)	Displays the results of a pre-computed BLAST search of a protein against all other protein sequences at NCBI

Search NCBI

cp4 epsps agrobacterium

✕ Search

Results found in 6 databases

Literature	
Bookshelf	0
MeSH	0
NLM Catalog	0
PubMed	44
PubMed Central	212

Genes	
Gene	1
GEO DataSets	0
GEO Profiles	0
HomoloGene	0
PopSet	0

Proteins	
Conserved Domains	0
Identical Protein Groups	0
Protein	12
Protein Family Models	0
Structure	9

Genomes	
Assembly	0
BioCollections	0
BioProject	0
BioSample	0
Genome	0
Nucleotide	22
SRA	0
Taxonomy	0

Clinical	
ClinicalTrials.gov	0
ClinVar	0
dbGaP	0
dbSNP	0
dbVar	0
GTR	0
MedGen	0
OMIM	0

PubChem	
BioAssays	0
Compounds	0
Pathways	0
Substances	0

Nucleotide

Nucleotide

cp4 epsps agrobacterium

Search

Create alert Advanced

Help

Species

Plants (2)
Bacteria (4)
Viruses (1)
Customize ...

Molecule types

genomic DNA/RNA (19)
Customize ...

Source databases

INSDC (GenBank) (22)
Customize ...

Sequence Type

Nucleotide (22)

Sequence length

Custom range...

Release date

Custom range...

Revision date

Custom range...

[Clear all](#)

[Show additional filters](#)

Summary 20 per page Sort by Default order

Send to:

Filters: [Manage Filters](#)

See Gene information for cp4 epsps
cp4 in [Brassica oleracea](#) 1 Gene record
epsps in [Oryza sativa Japonica Group](#) [Zea mays](#) [Lolium rigidum](#) All 13 Gene records

Items: 1 to 20 of 22

<< First < Prev Page 1 of 2 Next > Last >>

[Synthetic construct 5-enol-pyruvylshikimate-3-phosphate synthase \(EPSPS\) gene, complete cds](#)

1. **1,584 bp linear other-genetic**
Accession: KJ787849.1 GI: 871708827
[Protein](#) [Taxonomy](#)
[GenBank](#) [FASTA](#) [Graphics](#)

[Synthetic construct 5-enolpyruvylshikimate 3-phosphate \(epsps\) gene, complete cds](#)

2. **2,075 bp linear other-genetic**
Accession: OM397417.1 GI: 2279859949
[Protein](#) [Taxonomy](#)
[GenBank](#) [FASTA](#) [Graphics](#)

[Synthetic construct CP4 EPSPS glyphosate tolerance protein gene, complete cds](#)

3. **1,368 bp linear other-genetic**
Accession: JF445290.1 GI: 343988361
[Protein](#) [Taxonomy](#)
[GenBank](#) [FASTA](#) [Graphics](#)

[Glycine max sequence containing the junction region between CP4EPSPS gene and NOS](#)

4. **210 bp linear DNA**
Accession: AJ783418.1 GI: 57335187
[Protein](#) [Taxonomy](#)
[GenBank](#) [FASTA](#) [Graphics](#)

[Pseudomonas fluorescens SBW25 genome assembly, chromosome: 1](#)

5. **6,722,400 bp circular DNA**
Accession: OV988001.1 GI: 2202903211
[Assembly](#) [BioProject](#) [BioSample](#) [Protein](#) [Taxonomy](#)
[GenBank](#) [FASTA](#) [Graphics](#)

[Glycine max transgenic cp4epsps gene for 5-enol-pyruvylshikimate-3-phosphate synthase class 2](#)

6. **2,457 bp linear DNA**
Accession: AB209952.1 GI: 62318478
[Protein](#) [Taxonomy](#)
[GenBank](#) [FASTA](#) [Graphics](#)

Results by taxon

Top Organisms [Tree](#)
synthetic construct (17)
Glycine max (2)
Triticum aestivum (1)
Pseudomonas fluorescens SBW25 (1)
Bartonella tribocorum CIP 105478 (1)

Find related data

Database: [Select](#)

[Find items](#)

Search details

cp4[All Fields] AND epsps[All Fields]
AND ("Agrobacterium"[Organism] OR
agrobacterium[All Fields])

[Search](#)

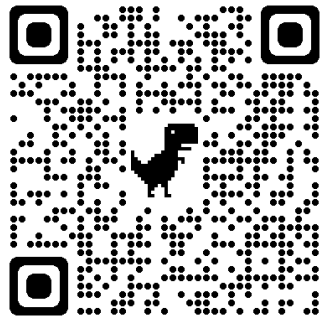
[See more...](#)

Recent activity

[Turn Off](#) [Clear](#)

- [cp4 epsps agrobacterium \(22\)](#) Nucleotide
- [LOC542727 \[Zea mays\]](#) Gene
- [cp4 epsps agrobacterium AND \(alive\[prop\]\) \(1\)](#) Gene
- [GM reference vector pTLE8, partial sequence](#) Nucleotide
- [Glycine max transgenic cp4epsps gene for 5-enol-pyruvylshikimate-3-phosphate](#) Nucleotide

[See more...](#)



Glycine max transgenic cp4epsps gene for 5-enol-pyruvylshikimate-3-phosphate synthase class 2 precursor, complete cds

GenBank: AB209952.1

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

LOCUS **AB209952** 2457 bp DNA linear SYN 07-APR-2005

DEFINITION Glycine max transgenic cp4epsps gene for 5-enol-pyruvylshikimate-3-phosphate synthase class 2 precursor, complete cds.

ACCESSION **AB209952**

VERSION AB209952.1

KEYWORDS .

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.

REFERENCE 1

AUTHORS Sotoshiro,H., Myouga,H., Kawata,M. and Tominaga,H.
 TITLE DNA sequence of introduced gene in glyphosate tolerant-GM soybeans
 JOURNAL Unpublished

REFERENCE 2 (bases 1 to 2457)

AUTHORS Sotoshiro,H., Myouga,H. and Tominaga,H.
 TITLE Direct Submission
 JOURNAL Submitted (30-MAR-2005) Hisashi Myouga, Musashigaoka College, Laboratory of Food Science; 111-1 Minamiyoshimi, Yoshimi-machi, Hiki-gun, Saitama 355-0154, Japan (E-mail:ofhmyoga@musashigaoka.ac.jp, URL:http://www.musashigaoka.ac.jp, Tel:81-493-54-6406(ex.2305), Fax:81-493-54-6756)

Change region shown

Customize view

Analyze this sequence

Run BLAST

Pick Primers

Highlight Sequence Features

Find in this Sequence

Related information


Protein


Taxonomy


Full text in PMC


Recent activity

[Turn Off](#) [Clear](#)

 Glycine max transgenic cp4epsps gene for 5-enol-pyruvylshikimate-3-phosphate Nucleotide

 cp4 epsps agrobacterium (22) Nucleotide

 LOC542727 [Zea mays] Gene

 cp4 epsps agrobacterium AND (alive[prop]) (1) Gene

FEATURES	Location/Qualifiers
source	1..2457 /organism="Glycine max" /mol_type="genomic DNA" /cultivar="Roundup Ready 30-4-2" /db_xref="taxon:3847" /clone="pCR4-TOPO" /transgenic /country="Japan"
source	1..265 /organism="Cauliflower mosaic virus" /mol_type="genomic DNA" /db_xref="taxon:10641"
source	298..510 /organism="Petunia x hybrida" /mol_type="genomic DNA" /db_xref="taxon:4102" /note="synonym: Petunia hybrida"
source	511..2457 /organism="Agrobacterium sp. CP4" /mol_type="genomic DNA" /strain="CP4" /db_xref="taxon:268951"
gene	1..265 /gene="CaMV35S"
regulatory	1..265 /regulatory_class="promoter" /gene="CaMV35S" /note="Cauliflower mosaic virus 35S promoter"
misc_feature	266..297 /note="nontranslation region"
gene	298..1881 /gene="cp4epsps"
CDS	298..1881 /gene="cp4epsps" /note="5-enol-pyruvylshikimate-3-phosphate synthase (EPSPS) class 2 precursor" /codon_start=1 /product="5-enol-pyruvylshikimate-3-phosphate synthase class 2 precursor" /protein_id="BAD94823.1" /translation="MAQINNMAQGIQTLNPNNSNFHKPQVPKSSSFLVFGSKLKNNSAN SMLVLKKDSIFMQKFCFRISASVATACMLHGASSRPATARKSSGLSGTVRIPGDKSI SHRSFMFGGLASGETRITGLLEGEDVINTGKAMQAMGARIRKEGDTWIIDGVNGGGLL APEAPLDFGNAATGCRITMGLVGVYDFDSTFIGDASLTKRPMGRVNLNPLREMGVQVKS EDGDRLPVTLRGPKTPTPIYRVPMASQVKSAVLLAGLNTPGITTVIEPIIMTCDHTE KMLQGFGANLTVETDADGVRTIRLEGRKLTGQVIDVPGDPSSTAFPLVAALLVPGSD VTILNVLMPNTRTGLILTQEMGADIEVINLRAGGEDVADLRVRSSTLKGVTVPEDR APPMIDEYPI LAVAAAF AEGATVMNGLEEELRVKESDRLSAVANGKLNQVDCDEGETS LVVRGRPDGKGLGNASGAAVATHLDHRIAMSLVLMGLVSENPVTVDATMIATSFPEF MDLMAGLGAKEI ELSDTKAA"



repeat_region tumefaciens"
2186..2439
/note="repeated fragment of cp4epsps;
truncate cp4epsps"

ORIGIN

```
1 tggaaaagga aggtggctcc tacaaatgcc atcattgcga taaaggaaag gccatcgttg
61 aagatgcctc tgccgacagt ggtcccaaag atggaccccc acccacgagg agcatcgtgg
121 aaaaagaaga cgttccaacc acgtcttcaa agcaagtgga ttgatgtgat atctccactg
181 acgtaaggga tgacgcacaa tcccactatc cttcgcaaga cccttcctct atataaggaa
241 gttcatttca tttggagagg acacgctgac aagctgactc tagcagatct ttcaagaatg
301 gcacaaatta acaacatggc acaagggata caaacctta atccaatic caatttccat
361 aaacccaag ttctaaatc ttcaagtttt cttgtttttg gatctaaaaa actgaaaaat
421 tcagcaaat ctatgttggg ttgaaaaaa gattcaattt ttatgcaaaa gttttgttcc
481 tttaggattt cagcatcagt ggctacagcc tgcattgctc acpgtgcaag cagccggccc
541 gcaaccgccc gcaaatcctc tggcctttcc ggaaccgtcc gatttcccgg cgacaagtcg
601 atctcccacc ggtccttcat gttcggcggg ctcgagcg gtgaaacg cgaccggc
661 cttctggaag gcgaggcgt catcaatc ggcaaggcca tgcaggccat gggcgcagg
721 atccgtaagg aaggcgacac ctggatcctc gatggcgtcg gcaatggcgg cctcctggcg
781 cctgaggcgc cgctcgattt cggcaatgcc gccacgggct gccgcctgac catgggcctc
841 gtcggggtct acgatttcca cagcaccttc atcggcgacg cctcgtcac aaagcggccg
901 atggcgcg tgttgaacc gctgcgcgaa atggcgtgc aggtgaaatc ggaagcggg
961 gaccgtcttc ccgttacctt gcgcgggccc aagacgcca cgccgatcac ctaccgctg
1021 ccgatggcct ccgcacaggt gaagtccgc gtgctgctcg ccggcctcaa cagcccggc
1081 atcacgacgg tcatcgagcc gatcatgacg tgcgatcata cggaaaaagat gctgcagggc
1141 tttggccca acctaccgt cgagacggat gcggacggcg tgcgaccat ccgcctggaa
1201 ggccgcccga agctcaccgg ccaagtcac gacgtgccgg gcgacccgct ctcgacggcc
1261 ttcccgtgg ttgcggcct cttgttccg ggctccgacg tcaccatcct caactgtctg
1321 atgaacccca cccgaccgg cctcatcctg acgctgcagg aaatggcgc cgacatcgaa
1381 gtcatacaacc tgcgcttgc cggcggcgaa gacgtggcgg acctgcgctg tcgctcctc
1441 acgctgaagg gcgtcacggt gccggaagac gcgcgcctc cgatgatcga gaaatattcg
1501 attctcgtc tgcgcggcg cttcgcggaa gggcgaccg tgatgaacgg tctggaagaa
1561 ctccgctca aggaagcga ccgcctctc gccgtgccca atggcctcaa gctcaatggc
1621 gtggattgcg atgagggcga gacgtcgtc gtcgtgctg gcccccctga cggcaagggg
1681 ctggcaacg cctcggcgc cgccgtgcc acctatctc atcaccgat cgccatgagc
1741 ttctctgca tggcctcgt gtcggaaaac cctgtcacgg tggacgatgc cacgatgatc
1801 gccacgact tcccggatt catggacctg atggccggc tggcgcgaa gatcgaactc
1861 tccgatacga aggtgcctg atgagctcga attcagctc ggtaccggat ccaattcccg
1921 atcgttcaaa catttggca taaagtttct taagattgaa tcctgttggc ggtccttgcga
1981 tgattatcat ataatttctg ttgaattacg ttaagcatgt aataattaac atgtaatgca
2041 tgacgttatt tatgagatgg gtttttatga ttagagtccc gcaattatac atttaatagc
2101 gctagaaa caaatatag cagcaact agaatatt atcggcggc atatcctc
```

```
298..1881
/gene="cp4epsps"
/note="5-enol-pyruvylshikimate-3-phosphate synthase (EPSPS)
class 2 precursor"
/codon_start=1
/product="5-enol-pyruvylshikimate-3-phosphate synthase
class 2 precursor"
/protein_id=" BAD94823.1 "
/translation="MAQINNMAQGIQLNPNFNHFKPQVPKSSFLVFGSKLKNNSAN
SMLVLKSDSIFMQKFCFRISASVATACMLHGASSRPATARKSSGLSGTVRIPGDKSI
SHRSFMFGGLASGETRITGLLEGEDVINTGKAMQAMGARIRKEGDTWIDGVNGGLL
APEAPLDFGNAATGCRLTMGLVGVYDFDSTFIGDASLTKRPMGRVNLPLREMGVQVKS
EDGDRPLVPTLRGPKTPTPTIYRVPMSAQVKSAVLLAGLNPGLITVIEPIMTCDHTE
KMLQGFGANLTVETDADGVRTIRLEGRGLTGQVIDVPGDPSSTAFPLVAALLVPGSD
VTILNVLMPNTRTGLILTLQEMGADIEVINLR LAGGEDVADLRVRSSTLKGVTVPEDR
APPMIDEYPI LAVAAAFAEGATVMNGLLELRVKESDRLSAVANGLKLNVDCEGETS
LVVRGRPDGKGLGNASGAAVATHLDHRIAMSFLVMGLVSENPTVDDATMIATSFPEF
MDLMAGLGAKIELSDTKAA"
```

Glycine max transgenic cp4epsps gene for 5-enol-pyruvylshikimate-3-phosphate synthase class 2 precursor, complete cds

GenBank: AB209952.1

[GenBank](#) [Graphics](#)

>AB209952.1:298-1881 Glycine max transgenic cp4epsps gene for 5-enol-pyruvylshikimate-3-phosphate synthase class 2 precursor, complete cds

```
ATGGCACAAATTAACAACATGGCACAAAGGGATACAAACCCCTTAATCCCAATTCCAATTTCCATAAACCCC
AAGTTCCTAAATCTTCAAGTTTTCTTGTTTTGGATCTAAAAAACTGAAAAATTCAGCAAATTCATATGTT
GGTTTTGAAAAAAGATTCAATTTTTATGCAAAAGTTTTGTTTCCTTTAGGATTTTACGCATCAGTGGCTACA
GCCTGCATGCTTACGGTGCAAGCAGCCGGCCGCAACCGCCCGCAAATCCTCTGGCCTTTCCGGAACCG
TCCGCATTCCCGGCGACAAGTCGATCTCCACCGGTCTTCATGTTTCGGCGGTCTCGCGAGCGGTGAAAC
GCGCATCACCGGCCCTTCTGGAAGGCGAGGACGTCATCAATACGGGCAAGGCCATGCAGGCCATGGGCGCC
AGGATCCGTAAGGAAGGCGACACCTGGATCATCGATGGCGTCGGCAATGGCGGCCCTCTGGCGCCTGAGG
CGCCGCTCGATTTGGCAATGCCGCCACGGGCTGCCGCTGACCATGGGCCTCGTCGGGGTCTACGATTT
CGACAGCACCTTCATCGGCGACGCTCGCTCACAAAGCGCCCGATGGGCCGCGTGTGAACCCGCTGCGC
GAAATGGGCGTGCAGGTGAAATCGGAAGACGGTGACCGTCTTCCCGTTACCTTGC CGGGCCGAAGACGC
CGACGCCGATCACCTACCGCGTGCCGATGGCCTCCGCACAGGTGAAGTCCGCCGTGCTGCTCGCCGGCCT
CAACACGCCCGGCATCACGACGGTCATCGAGCCGATCATGACGTGCGATCATACGGAAAAGATGCTGCAG
GGCTTTGGCGCAACCTTACCGTCGAGACGGATGCGGACGGCGTGCGCACCATCCGCCTGGAAGGCCGCG
GCAAGCTCACCGGCCAAGTCATCGACGTGCCGGGCGACCCGTCCTCGACGGCCTTCCCGCTGGTTGCGGC
CCTGCTTGTTCGGGCTCCGACGTCACCATCCTCAACGTGCTGATGAACCCACCCGCACCGGCCTCATC
CTGACGCTGCAGGAAATGGGCGCCGACATCGAAGTCATCAACCTGCGCCTTGGCGGGCGGAAGACGTGG
CGGACCTGCGCGTTCGCTCCTCCACGCTGAAGGGCGTCACGGTGCCGGAAGACCGCGCGCCTCCGATGAT
CGACGAATATCCGATTCTCGCTGTGCCGCCGCTTCGCGGAAGGGGCGACCGTGATGAACGGTCTGGAA
GAACTCCGCGTCAAGGAAAAGCGACCGCCTCTCGGCCGTGCGCAATGGCCTCAAGCTCAATGGCGTGGATT
GCGATGAGGGCGAGACGTGCTCGTGTGCGTGGCCGCCCTGACGGCAAGGGGCTCGGCAACGCCTCGGG
CGCCGCCGTGCCACCCATCTCGATCACCGCATCGCCATGAGCTTCTCGTCATGGGCCTCGTGTTCGGAA
AACCTGTACGGTGGACGATGCCACGATGATCGCCACGAGCTTCCCGGAGTTCATGGACCTGATGGCCG
GGCTGGGCGCGAAGATCGAACTCTCCGATACGAAGGCTGCCTGA
```

Sequencia formato FASTA

Busca por BLAST

- Busca por nucleotídeos ou amino ácidos (proteínas)
- Comparação de sequencias para identificar similaridade significativa de DNA ou PTN para inferir **função, origem, filogenia**
- Realiza comparações entre pares de sequencias, buscando regiões com similaridade local
- Alinhamento local (segmentos) é a base da busca por BLAST
- Usa algoritmos para gerar alinhamento de sequências

Basic Local Alignment Search Tool

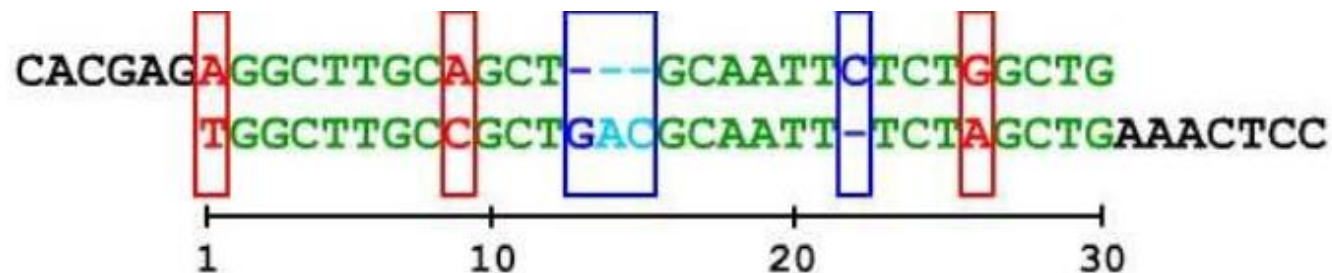
Comparando sequências - Alinhamento

Alinhamento Global: é feito quando comparamos uma sequência de aminoácidos ou nucleotídeos com outra **ao longo de toda sua extensão**

```
seq1  GCTCATTGACCTGACTAG
seq2  GTGACTAAGACCTTCATT
```

Alinhamento Local: a comparação entre duas sequências não é feita ao longo de toda sua extensão, mas sim através de pequenas regiões destas.

O *BLAST* é o principal programa para realizar o alinhamento local



BLAST

Comparando sequências - Alinhamento

O alinhamento de sequências consiste em comparar duas sequências (de nucleotídeos ou aminoácidos) de forma a identificarmos o **grau de identidade/similaridade** entre elas

Identidade

Número de posições invariáveis em duas sequências (nucleotídeos ou aminoácidos) alinhadas.

Similaridade

Grau de semelhança entre duas sequências de proteínas expressa em percentual de amino ácidos com característica similar alinhados

```
RBP:          26  RVKENFDKARFSGTWYMAKKDPEGLFLQDNIVA 59
                + K++ + ++ GTW++MA + L + A
glicodelina: 23  QTKQDLELPKLAGGTWHSMAMA-TNNISLMATLKA 55
```

Busca por BLAST

- **Identidade:** ocorrência do exato mesmo nucleotídeo ou amino ácido na mesma posição em sequencias alinhadas
- **Similaridade:** ocorrência de amino ácidos equivalentes (quimicamente) na mesma posição
- **Homologia:** dividem a mesma ancestralidade com significado evolutivo

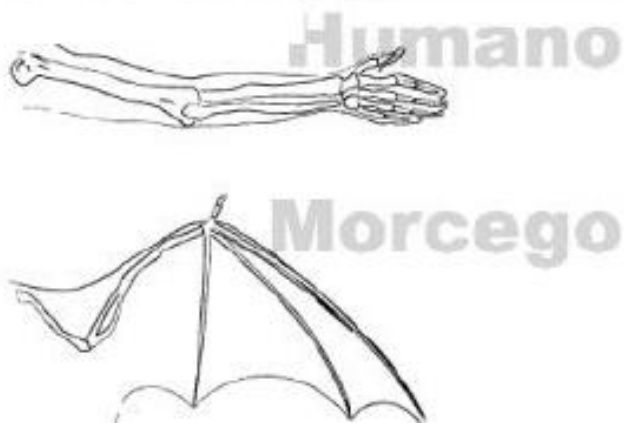
Homologia – conceito fundamental na biologia

Algoritmos em Blast:

- **Não avaliam homologia**

A análise de seqüências objetiva encontrar similaridades importantes que permitam inferir sobre homologia

- Medem similaridade e identidade entre seqüências



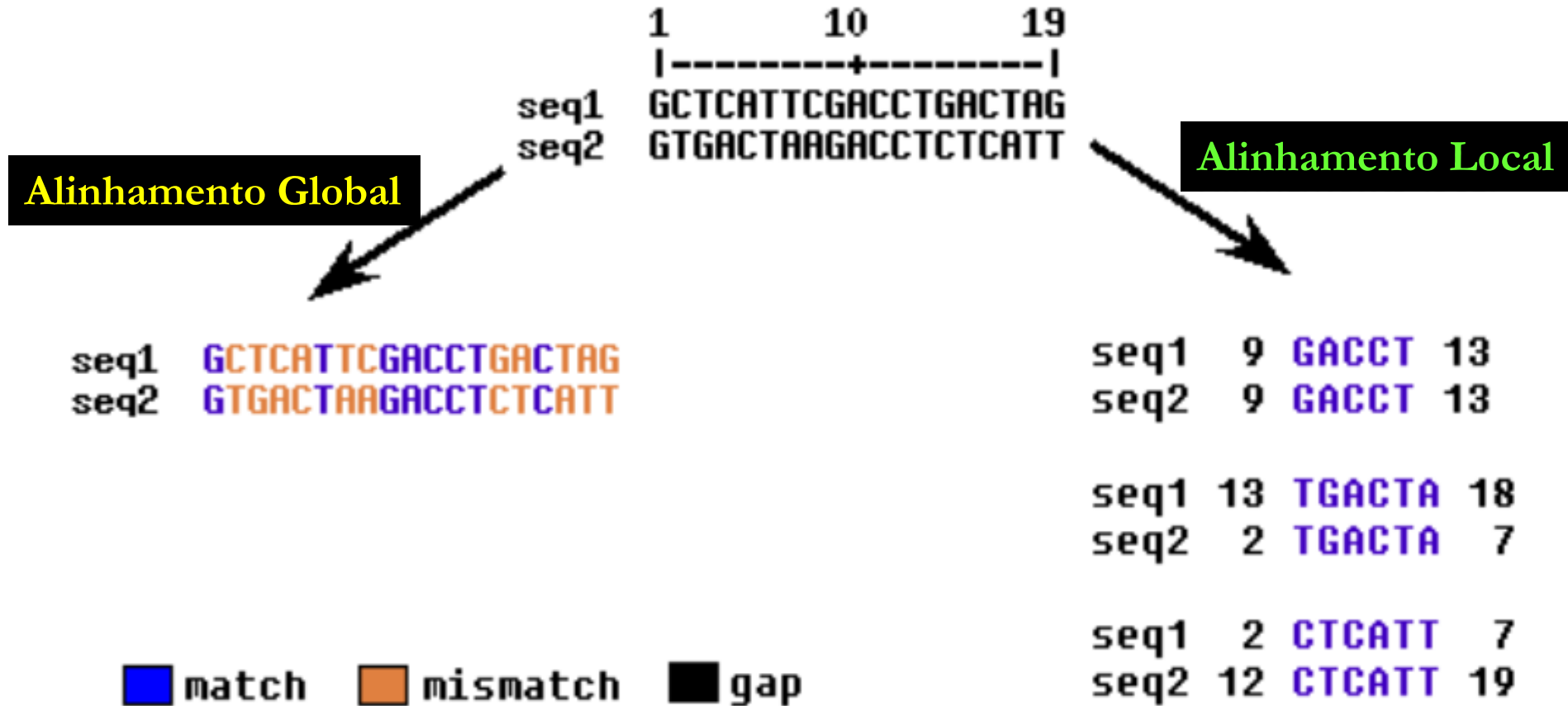
Exemplos:

Órgãos homólogos – asas de morcego e mãos de humanos (mesma origem)

Órgãos similares – asas de morcego e asas de borboleta (mesma função)

BLAST

Comparando sequências - Alinhamento



XM_010026669.3

PREDICTED: Syzygium oleosum rubisco accumulation factor 1.1, chloroplastic (LOC115690120), i

Sequence ID: [XM_030616373.2](#) Length: 1930 Number of Matches: 1

Range 1: 56 to 1928 [GenBank](#) [Graphics](#)

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

Score	Expect	Identities	Gaps	Strand
2082 bits(1127)	0.0	1643/1886(87%)	60/1886(3%)	Plus/Plus
Query 24	aataaaaaaaaaTC-AAAGCCATCAGTGAACCTCAGTTCAGATAAGGTTTCACTGAAAACCC	82		
Sbjct 56	AAGAAAAATATCAAAAGCCATCATTCAATTTAGTTCAGATAAGGTTTCACTGAAGACTC	115		
Query 83	ACAGTGACCTCCCTCCACTCCTCCCTGAACCACCAACTCCCGCCACCGCCACCGCC---T	139		
Sbjct 116	ACAGTGACCCTCCTCCAC-CCTCACTGCACCACCGCCTCCCGCCGCGCCACCGCCGCGC	174		
Query 140	CCG---CCACCATGCTCTCTAGCCCTCATCAAcaccaccac---cc-ccaccaccacc	193		
Sbjct 175	CCGCCTCCACCATGCTCTCTAGCCCTCATCAACACCACCCACACCACCACCACCACC	234		
Query 194	aattcttcaatcccaccaccaccatcctcaaaccctcacttctctcccctccaccacc	253		
Sbjct 235	AATTCTTCAACCCACCACCACCA-CCAC-CA-CCCTTA-AGCCCCTCACTTCCCCTTTC	290		
Query 254	cttgccgccaccaccgccgccaccactccgccggcccaaggcccatctccgcccaGCCTCA	313		
Sbjct 291	C-T-CCACCACCCACCACCACCCTCGGCCGGCCCGAGGCCATCTCCGCCAGCCTCA	348		
Query 314	ACCCGAACCTCTAAtcccggcctcccctccgccgcagcagcagcggccctaccagccct	373		
Sbjct 349	ACCCGAGCCCTCTAATCCCGGCCTCCCCTCCGCCGCAGCAGCAGCGTGTCTACCAGCCCT	408		
Query 374	tccgccctcccc-tccccgct--cctcccagttccgctccctcGACACCGCCAGCCGCC	430		
Sbjct 409	TCCGCCCTCCCCCTCCCCGCTCCCCTCCCAGTTCGGCTCCCTCGACACCGCCGGCCGCC	468		



Check out the ClusteredNR database on BLAST+

Learn more

Give us feedback

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 Quick Start guides!

Need some help getting started with BLAST?

Thu, 22 Jun 2023

[More BLAST news...](#)

Web BLAST



Nucleotide BLAST
nucleotide ▶ nucleotide



blastx
translated nucleotide ▶ protein

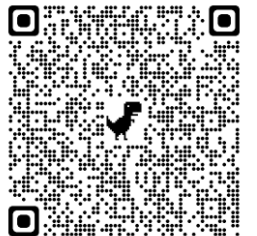


tblastn
protein ▶ translated nucleotide



Protein BLAST
protein ▶ protein

BLAST



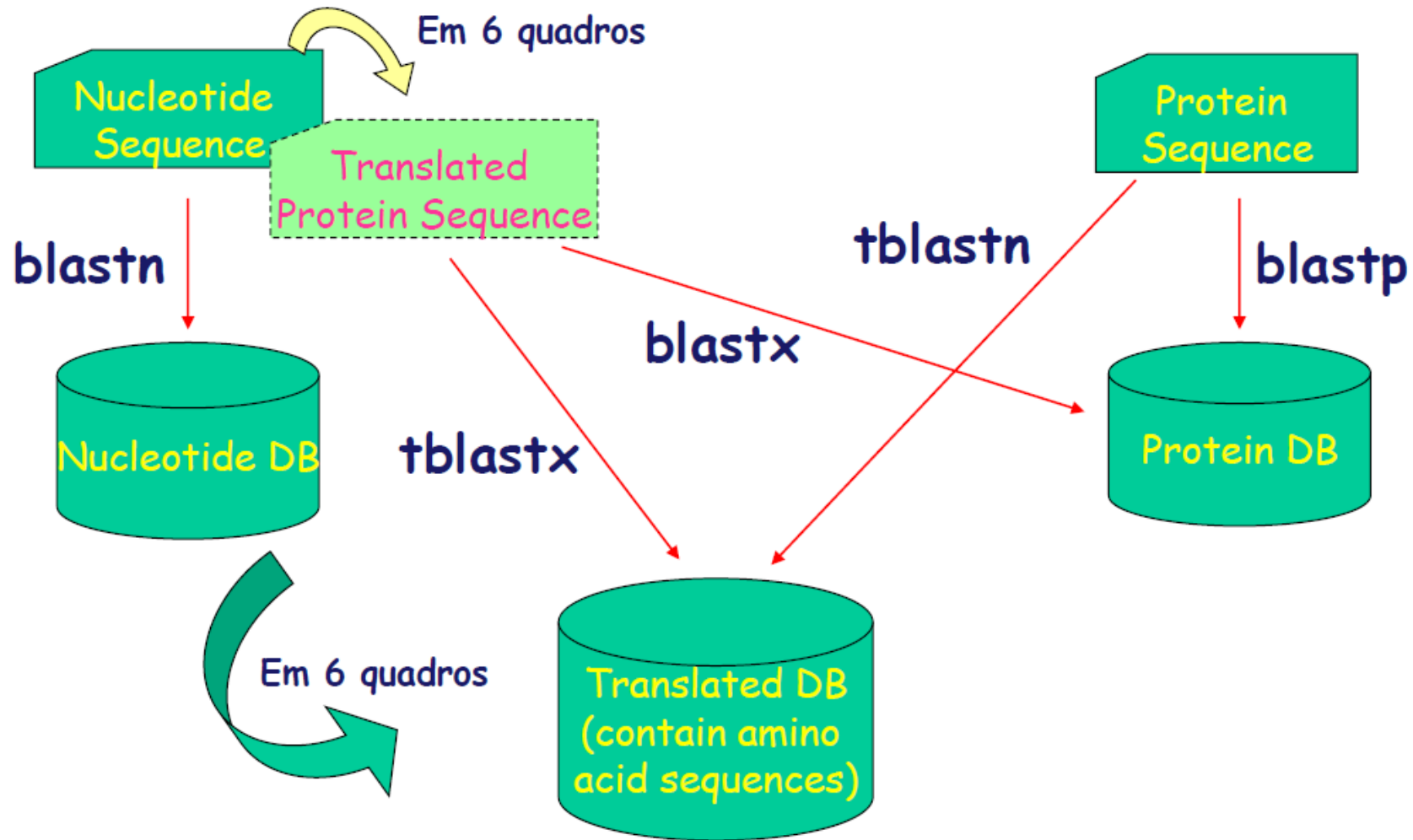
Busca por BLAST

Tipos de BLAST de acordo com o tipo de sequencia fornecida e qual o tipo buscado

Program	Query sequence type	Target sequence type	
BLASTP	Protein	Protein	Compares an amino acid query sequence against a protein sequence database
BLASTN	Nucleotide	Nucleotide	Compares a nucleotide query sequence against a nucleotide sequence database
BLASTX	Nucleotide (translated)	Protein	Compares a nucleotide query sequence translated in all reading frames against a protein sequence database
TBLASTN	Protein	Nucleotide (translated)	Compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames
TBLASTX	Nucleotide (translated)	Nucleotide (translated)	Compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database

<u>Programa</u>	<u>Sequência</u>	<u>Base</u>	<u>Comparação</u>
blastn	DNA	DNA	DNA
blastp	PTN	PTN	PTN
blastx	DNA	PTN	PTN
tblastn	PTN	DNA	PTN
tblastx	DNA	DNA	PTN

Busca por BLAST



Formato FASTA: formato universalmente aceito para ser processado

Identificador - linha do nome (máximo 80 caracteres por linha)

```
>gi|226347322|gb|FJ830553.1| Anabaena planctonica CENA210 ribulose-1,5-  
bisphosphate carboxylase/oxygenase large subunit (rbcL) gene, partial cds  
CCGGCGAAATTAAAGGTCACCTCAACGTTACCGCTCCTACCTGCGAAGAAATGTTGAAACGGGCTGA  
GTACGCTAAAGAACTCAAAATGCCCATCATCATGCACGACTACCTAACCGCAGGTTTCACCGCTAACACC  
ACATTGGCTCGTTGGTGTCTGATAACGGTATTTTATTGCACATTCACCGTGCTATGCACGCTGTAATTG  
ACCGTCAAAAAAATCACGGTATCCACTTCCGCGTATTAGCTAAAGCCCTCCGCTTGTCCGGTGGTGATCA  
CATCCACACTGGTACAGTTGTTGGTAAGTTAGAAGGTGAACGCGGTATTACCATGGGCTTCGTTGACTTA  
TTACGTGAAAACACTACGTTGAGCAAGACAAGTCTCGCGGTATTTACTTTACCCAAGATTGGGCGTCTCTAC  
CTGGTGTAATGGCCGTTGCTTCTGGTGGTATCCACGTATGGCATATGCCCGCGTTGGTTGAGATCTTCGG  
TGATGACTCCGTATTACAATTCGGTGGTGGTACACTCGGACATCCTTGGGGTAACGCTCCTGGTGCTACA  
GCTAACCGCGTAGCTCTAAAAGCAGTTGTTCAAGCTCGTAACGAAGGCCGTAACCTAGCTCGTGAAGGTA  
ACGATATTATCCGCGAAGCTGCTAAGTGGTCTCCTGAGTTGGCTGTTGCTTGCGAACTG
```

```
>gi|226347323|gb|AC050079.1| ribulose-1,5-bisphosphate  
carboxylase/oxygenase large subunit [Anabaena planctonica CENA210]  
GEIKGHYLNVTAPTCEEMLKRAEYAKELKMPIIMHDYLTAGFTANTTLARWCRDNGILLHIHRAMHAVID  
RQKNHGIHFRVLAKALRLSGGDHIHTGTVVGKLEGERGITMGFVDLLRENYVEQDKSRGIYFTQDWASLP  
GMAVASGGIHVWHMPALVEIFGDDSVLQFGGGTLGHPWGNAPGATANRVALKAVVQARNEGRNLAREGN  
DIIREAAKWSPELAVACEL
```


Busca por BLAST



blastn blastp blastx tblastn tblastx

BLASTN programs search nucleotide databases using a nucleotide query. more...

Reset page Bookmark

Enter Query Sequence

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

Query subrange [?](#)

From

To

Or, upload file

Escolher arquivo Nenhum arquivo escolhido [?](#)

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 + transcript databases Betacoronavirus

Nucleotide collection (nr/nt) [?](#)

Organism
Optional

exclude [Add organism](#)

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

[YouTube](#) [Create custom database](#)

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

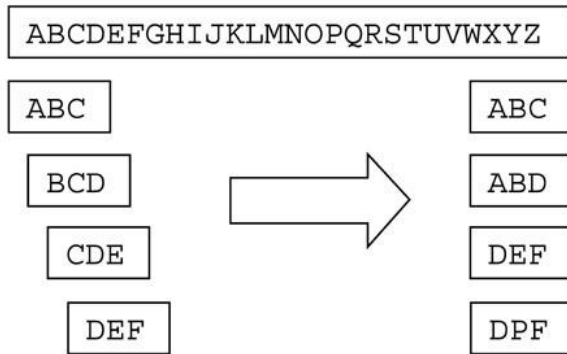
Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)

Show results in a new window

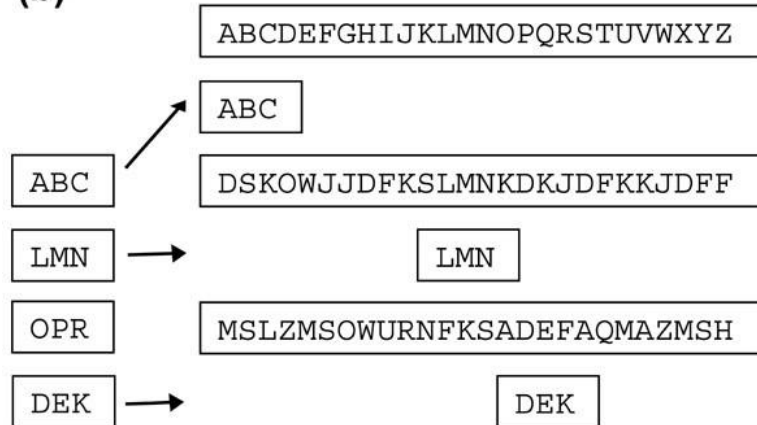
+ Algorithm parameters

Busca por BLAST

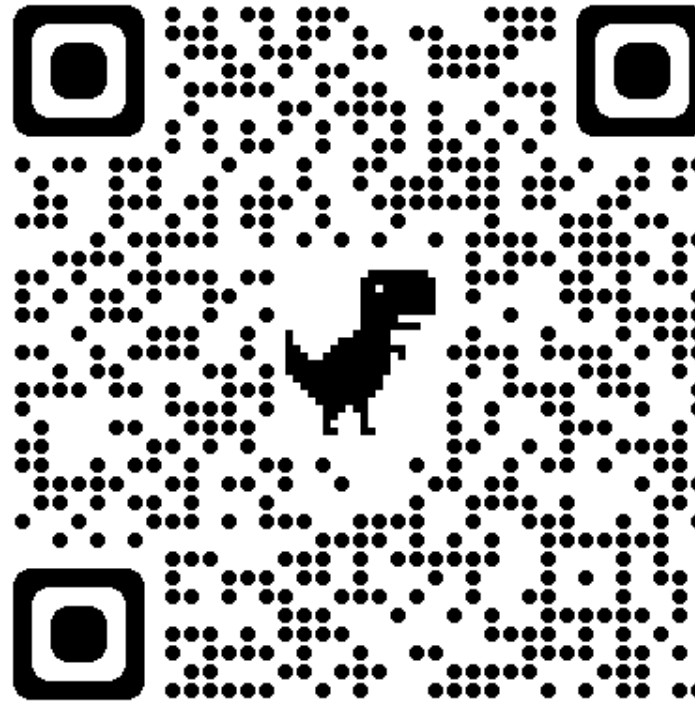
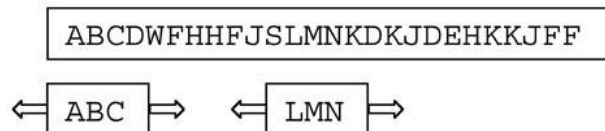
(a)



(b)



(c)



Glycine max transgenic
cp4epsps gene for 5-enol-
pyruvylshikimate-3-phosphate
synthase class 2 precursor,
complete cds

Busca por BLAST

BLAST® » blastn suite » results for RID-KDF4FVGX01R [Home](#) [Recent Results](#) [Saved Strategies](#) [Help](#)

[← Edit Search](#) [Save Search](#) [Search Summary](#) [How to read this report?](#) [BLAST Help Videos](#) [Back to Traditional Results Page](#)

Job Title Nucleotide Sequence

RID [KDF4FVGX01R](#) Search expires on 10-25 02:33 am [Download All](#) ▼

Program BLASTN [Citation](#) ▼

Database nt [See details](#) ▼

Query ID lcl|Query_174315

Description None

Molecule type dna

Query Length 2457

Other reports [Distance tree of results](#) [MSA viewer](#) ?

Filter Results

Organism only top 20 will appear exclude

Type common name, binomial, taxid or group name

[+ Add organism](#)

Percent Identity to **E value** to **Query Coverage** to

[Filter](#) [Reset](#)

Descriptions [Graphic Summary](#) [Alignments](#) [Taxonomy](#)

Sequences producing significant alignments [Download](#) ▼ [Select columns](#) ▼ [Show](#) ?

select all 100 sequences selected [GenBank](#) [Graphics](#) [Distance tree of results](#) [MSA Viewer](#)

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/>	Glycine max transgenic cp4epsps gene for 5-enol-pyruvylshikimate-3-phosphate synthase class 2 precursor, c...	Glycine max	4538	5493	100%	0.0	100.00%	2457	AB209952.1
<input checked="" type="checkbox"/>	Synthetic construct CP4EPSPS protein (CP4EPSPS) gene, complete cds	synthetic construct	3561	4256	84%	0.0	99.74%	1946	AY125353.1
<input checked="" type="checkbox"/>	Synthetic construct 5-enolpyruvylshikimate 3-phosphate (epsps) gene, complete cds	synthetic construct	3192	4075	94%	0.0	98.57%	2075	OM397417.1
<input checked="" type="checkbox"/>	Triticum aestivum transgenic CTP2-CP4-EPSPS genes, complete cds	Triticum aestivum	2804	8016	90%	0.0	97.98%	7729	MN020371.1
<input checked="" type="checkbox"/>	Zea mays transgenic cultivar 631 RR2/Bt transgenic line NK603, complete sequence	Zea mays	2892	7539	90%	0.0	98.00%	7584	KX640115.1
<input checked="" type="checkbox"/>	Glycine max CP4EPSPS gene, complete cds	Glycine max	2791	3283	72%	0.0	99.67%	1529	AF464188.1
<input checked="" type="checkbox"/>	GM reference vector pTLE8, partial sequence	GM reference ve...	2172	3920	79%	0.0	99.75%	3664	JX434028.1
<input checked="" type="checkbox"/>	Shinella zoogloeoides strain UPHL-collab-1 chromosome, complete genome	Shinella zoogloe...	1757	2114	65%	0.0	90.16%	3964868	CP132311.1
<input checked="" type="checkbox"/>	Shinella sp. PSBB067 chromosome, complete genome	Shinella sp. PS...	1735	2074	66%	0.0	89.61%	4605385	CP069303.1
<input checked="" type="checkbox"/>	Shinella zoogloeoides strain XJ20 chromosome, complete genome	Shinella zoogloe...	1688	1983	66%	0.0	88.73%	4333343	CP093528.1
<input checked="" type="checkbox"/>	Shinella zoogloeoides strain ATCC 19623 chromosome, complete genome	Shinella zoogloe...	1683	1683	55%	0.0	88.65%	3991891	CP086610.1
<input checked="" type="checkbox"/>	Shinella sumterensis strain UPHL-collab-3 chromosome, complete genome	Shinella sumter...	1652	1652	54%	0.0	88.75%	3810368	CP132316.1
<input checked="" type="checkbox"/>	Shinella oryzae strain Z-25 chromosome, complete genome	Shinella oryzae	1624	1954	65%	0.0	88.38%	3326466	CP081967.1

Glycine max transgenic cp4epsps gene for 5-enol-pyruvylshikimate-3-phosphate synthase class 2 precursor, complete cds

Busca por BLAST

NIH National Library of Medicine
National Center for Biotechnology Information

BLAST® » blastn suite » results for RID-KDF4FVGX01R

Home Recent Results Saved Strategies Help

< Edit Search Save Search Search Summary

How to read this report? BLAST Help Videos Back to Traditional Results Page

Job Title Nucleotide Sequence

RID [KDF4FVGX01R](#) Search expires on 10-25 02:33 am [Download All](#)

Program BLASTN Citation

Database nt See details

Query ID Icl|Query_174315

Description None

Molecule type dna

Query Length 2457

Other reports [Distance tree of results](#) [MSA viewer](#)

Filter Results

Organism only top 20 will appear exclude

Type common name, binomial, taxid or group name

+ Add organism

Percent Identity to to E value to to Query Coverage to to

Filter Reset

Descriptions **Graphic Summary** Alignments Taxonomy

hover to see the title click to show alignments

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

100 sequences selected

Distribution of the top 139 Blast Hits on 100 subject sequences

Glycine max transgenic
cp4epsps gene for 5-enol-
pyruvylshikimate-3-phosphate
synthase class 2 precursor,
complete cds

Busca por BLAST

NIH National Library of Medicine
National Center for Biotechnology Information

BLAST® » blastn suite » results for RID-KDF4FVGX01R

Home Recent Results Saved Strategies Help

[< Edit Search](#) Save Search Search Summary

How to read this report? BLAST Help Videos Back to Traditional Results Page

Job Title: Nucleotide Sequence
RID: [KDF4FVGX01R](#) Search expires on 10-25 02:33 am [Download All](#)
Program: BLASTN [Citation](#)
Database: nt [See details](#)
Query ID: Icl|Query_174315
Description: None
Molecule type: dna
Query Length: 2457
Other reports: [Distance tree of results](#) [MSA viewer](#)

Filter Results

Organism only top 20 will appear exclude
Type common name, binomial, taxid or group name
[+ Add organism](#)

Percent Identity: to E value: to Query Coverage: to
[Filter](#) [Reset](#)

Descriptions Graphic Summary **Alignments** Taxonomy

Alignment view: Pairwise CDS feature [Restore defaults](#) Download

100 sequences selected

[Download](#) [GenBank](#) [Graphics](#) Sort by: E value [Next](#) [Previous](#) [Descriptions](#)

Glycine max transgenic cp4epsps gene for 5-enol-pyruvylshikimate-3-phosphate synthase class 2 precursor, complete cds
Sequence ID: [AB209952.1](#) Length: 2457 Number of Matches: 3

Range 1: 1 to 2457 [GenBank](#) [Graphics](#) [Next Match](#) [Previous Match](#)

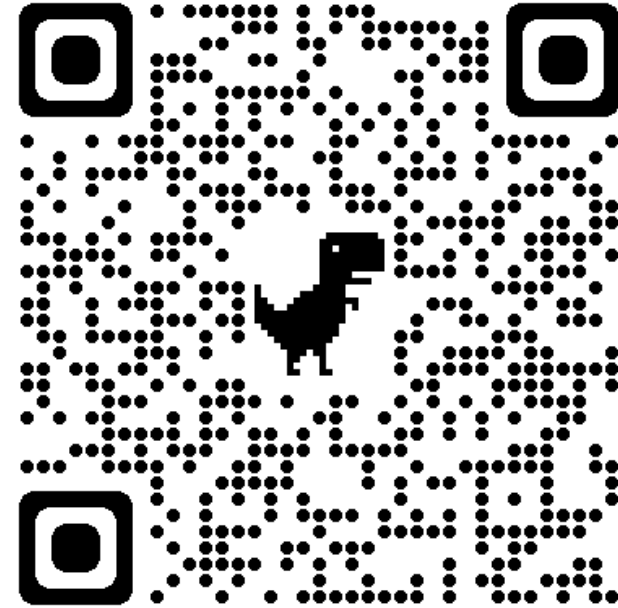
Score	Expect	Identities	Gaps	Strand
4538 bits(2457)	0.0	2457/2457(100%)	0/2457(0%)	Plus/Plus
Query 1	TGGAAAAGGAAGGTGGCTCTCAAAATGCCATCATTGCGATAAAGGAAAGGCCATCGTTG	60		
Sbjct 1	TGGAAAAGGAAGGTGGCTCTCAAAATGCCATCATTGCGATAAAGGAAAGGCCATCGTTG	60		
Query 61	AAGATGCCTCTGCCGACAGTGGTCCAAAGATGGACCCCAACCCACGAGGAGCATCGTGG	120		
Sbjct 61	AAGATGCCTCTGCCGACAGTGGTCCAAAGATGGACCCCAACCCACGAGGAGCATCGTGG	120		
Query 121	AAAAAGAAGACGTTCCAACCACGCTTCAAAGCAAGTGGATTGATGTGATATCTCCACTG	180		
Sbjct 121	AAAAAGAAGACGTTCCAACCACGCTTCAAAGCAAGTGGATTGATGTGATATCTCCACTG	180		
Query 181	ACGTAAGGGATGACGCACAATCCCACTATCCTTCGCAAGACCCCTTCTCTATATAAGGAA	240		
Sbjct 181	ACGTAAGGGATGACGCACAATCCCACTATCCTTCGCAAGACCCCTTCTCTATATAAGGAA	240		

Glycine max transgenic cp4epsps gene for 5-enol-pyruvylshikimate-3-phosphate synthase class 2 precursor, complete cds

Busca por BLAST



pesticidal protein (plasmid)
[Bacillus thuringiensis]
GenBank: AKJ62760.1



Bacillus thuringiensis strain
T29 Cry1Ac gene, partial cds
GenBank: MK882923.1

> [gi|47933333|gb|AY262820.1](#) Pinus radiata cellulose synthase (CesA10) mRNA, complete cds
Length=4428

Score = 7374 bits (3720), Expect = 0.0
Identities = 3741/3741 (100%), Gaps = 0/3741 (0%)
Strand=Plus/Plus

```
Query 1 GCACGAGGATTTAATCGAACTCGGTAATTGTTATCATCGTGGTGAGGACTAGTGCTTGAT 60
      |||
Sbjct 1 GCACGAGGATTTAATCGAACTCGGTAATTGTTATCATCGTGGTGAGGACTAGTGCTTGAT 60

Query 61 ATTTTAGTTTTATTCTCGAAATTTTCATAATAGCTTGGGCTTTCTAAAAAGGGGAATGGTG 120
      |||
Sbjct 61 ATTTTAGTTTTATTCTCGAAATTTTCATAATAGCTTGGGCTTTCTAAAAAGGGGAATGGTG 120

Query 121 GAATGGGTGTGAGAGTGAAGAGGAATGGTATCGAACCCTAAGAAAAGTAGTCGTGCAAG 180
      |||
Sbjct 121 GAATGGGTGTGAGAGTGAAGAGGAATGGTATCGAACCCTAAGAAAAGTAGTCGTGCAAG 180

Query 181 TATTAGATGGTTGGCTGTGATAGTTGGAAAAGGAATAGTAGAAATGGGACAGAAGTTTCA 240
      |||
Sbjct 181 TATTAGATGGTTGGCTGTGATAGTTGGAAAAGGAATAGTAGAAATGGGACAGAAGTTTCA 240

Query 241 TTCTGTAAGCTTTTTTCATGGACTGTTAGTCTTCTCTTTGCTTTCAGCTTAAGCAGCTTTA 300
      |||
Sbjct 241 TTCTGTAAGCTTTTTTCATGGACTGTTAGTCTTCTCTTTGCTTTCAGCTTAAGCAGCTTTA 300
```

→ Barra = Identidade

Busca por BLASTp

BLAST® » blastp suite

[Home](#) [Recent Results](#) [Saved Strategies](#)

Standard Protein BLAST

[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 [?](#)

From

To

Or, upload file

Nenhum ar...ivo escolhido [?](#)

Job Title

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

Align two or more sequences [?](#)

Choose Search Set

Databases

Standard databases (nr etc.): **New** Experimental databases

[< Try experimental clustered nr database](#) [Q](#)
For more info see [What is clustered nr?](#)

Compare

Select to compare standard and experimental database [?](#)

Standard

Database

Organism

Optional

Exclude

Optional

- Non-redundant protein sequences (nr) [?](#)
- Non-redundant protein sequences (nr)**
- RefSeq Select proteins (refseq_select)
- 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)

exclude

taxa will be shown. [?](#)

Uncultured/environmental sample sequences

Program Selection

Algorithm

Quick BLASTP (Accelerated protein-protein BLAST)
 blastp (protein-protein BLAST)

EDPAKDFNSYGFGNVAWKERVE SWKKNKQDKNTLQVTSDTYYASEGKGDIDGCVADEEDLQMSDEARQPL

Or, upload file

Procurar...

Job Title

gi|47933334|gb|AAQ63935.1| cellulose synthase...

Enter a descriptive title for your BLAST search

Choose Search Set

Database

Non-redundant protein sequences (nr)

Organism

Optional

Enter organism name or id—completions will be suggested

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

Entrez Query

Optional

Enter an Entrez query to limit search

Program Selection

Algorithm

blastp (protein-protein BLAST)

PSI-BLAST (Position-Specific Iterated BLAST)

PHI-BLAST (Pattern Hit Initiated BLAST)

Choose a BLAST algorithm

BLAST

Search database nr using Blastp (protein-protein BLAST)

Show results in a new window

▶ [Algorithm parameters](#)

- [Human](#)
- [Mouse](#)
- [Rat](#)
- [Arabidopsis thaliana](#)
- [Oryza sativa](#)
- [Bos taurus](#)
- [Danio rerio](#)
- [Drosophila melanogaster](#)
- [Gallus gallus](#)
- [Pan troglodytes](#)
- [Microbes](#)
- [Apis mellifera](#)

removed on June 11,
2007.
2007-06-01 12:15:00

[More BLAST news...](#)

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a nucleotide database using a nucleotide query <i>Algorithms: blastn, megablast, discontinuous megablast</i>
protein blast	Search protein database using a protein query <i>Algorithms: blastp, psi-blast, phi-blast</i>
blastx	Search protein database using a translated nucleotide query
tblastn	Search translated nucleotide database using a protein query
tblastx	Search translated nucleotide database using a translated nucleotide query

Tip of the Day

How to Search Custom Databases in Web-Blast Using Entrez Queries

A powerful feature of the BLAST Web interface is the ability to limit BLAST searches to a subset of any database using a standard Entrez query. Skillful use of Entrez queries allows the equivalent of on-the-fly construction of databases of exact composition

[More tips...](#)

Specialized BLAST

Choose a type of specialized search (or database name in parentheses.)

- Search [trace archives](#)
- Find [conserved domains](#) in your sequence (cds)
- Find sequences with similar [conserved domain architecture](#) (cdart)

Enter Query Sequence

Enter accession number, gi, or FASTA sequence

Clear

Query subrange

```
>gi|47933333|gb|AY262820.1| Pinus radiata cellulose synthase (CesA10)  
mRNA, complete cds  
GCACGAGGATTTAATCGAACTCGGTAATTGTTATCATCGTGGTGAGGACTAGTGCTTGATATTTTAGTTT  
TATTCTCGAAATTTTATAATAGCTTGGGCTTTCTAAAAAGGGGAATGGTGGAAATGGGTGTGAGAGTGAAG  
AGGAATGGTATCGAACCACTAAGAAAAGTAGTCGTGCAAGTATTAGATGGTTGGCTGTGATAGTTGGAAA
```

From

To

Or, upload file

Procurar...

Genetic code

Standard (1)

Job Title

gi|47933333|gb|AY262820.1| Pinus radiata cellulose...

Enter a descriptive title for your BLAST search

Choose Search Set

Database

Non-redundant protein sequences (nr)

Organism

Optional

Enter organism name and completions will be suggested

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

Entrez Query

Optional

Enter an Entrez query to limit search

BLAST

Search database nr using Blastx (search protein databases using a translated nucleotide query)

Show results in a new window

Algorithm parameters

> [gi|47933334|gb|AAQ63935.1|](#) cellulose synthase [Pinus radiata]

Length=1096

Score = 2221 bits (5754), Expect = 0.0
Identities = 1096/1096 (100%), Positives = 1096/1096 (100%), Gaps = 0/1096 (0%)
Frame = +1

Query	649	MEARTNTAAGSNKRNVRVSVRDDGELGPKPPQHINSHICQICGEDVGLAADGEFFVACNE	828
		MEARTNTAAGSNKRNVRVSVRDDGELGPKPPQHINSHICQICGEDVGLAADGEFFVACNE	
Sbjct	1	MEARTNTAAGSNKRNVRVSVRDDGELGPKPPQHINSHICQICGEDVGLAADGEFFVACNE	60
Query	829	CAFPVCRPCYEYEWKDGNOQSCPCQCKTRYKWHKGSPOVDGDKEDCACDDLDFNSTQGMR	1008
		CAFPVCRPCYEYEWKDGNOQSCPCQCKTRYKWHKGSPOVDGDKEDCACDDLDFNSTQGMR	
Sbjct	61	CAFPVCRPCYEYEWKDGNOQSCPCQCKTRYKWHKGSPOVDGDKEDCACDDLDFNSTQGMR	120
Query	1009	NEKQQIAEAMLHWQMAVGRGEDVGPSRSESQELPQLQVPLITNGQAISGELPAGSSEYRR	1188
		NEKQQIAEAMLHWQMAVGRGEDVGPSRSESQELPQLQVPLITNGQAISGELPAGSSEYRR	
Sbjct	121	NEKQQIAEAMLHWQMAVGRGEDVGPSRSESQELPQLQVPLITNGQAISGELPAGSSEYRR	180
Query	1189	IAAPPTGGGSGKRVHPLPFPDSTQTGQVRAEDPAKDFNSYGFGNVANKERVESWKNKQDK	1368
		IAAPPTGGGSGKRVHPLPFPDSTQTGQVRAEDPAKDFNSYGFGNVANKERVESWKNKQDK	
Sbjct	181	IAAPPTGGGSGKRVHPLPFPDSTQTGQVRAEDPAKDFNSYGFGNVANKERVESWKNKQDK	240
Query	1369	NTLQVTSPTYASEGKDGDIDGCVADEEDLQMSDEARQPLSRKVP IASSKINPYRMVIVL	1548
		NTLQVTSPTYASEGKDGDIDGCVADEEDLQMSDEARQPLSRKVP IASSKINPYRMVIVL	
Sbjct	241	NTLQVTSPTYASEGKDGDIDGCVADEEDLQMSDEARQPLSRKVP IASSKINPYRMVIVL	300

> [gi|47933336|gb|AAQ63936.1|](#) cellulose synthase [Pinus radiata]
Length=1066

Score = 1813 bits (4695), Expect = 0.0
Identities = 890/1066 (83%), Positives = 972/1066 (91%), Gaps = 9/1066 (0%)
Frame = +1

```
Query 760 ICQICGEDVGLAADGEGFFVACNECAFVPCRPCYEYEWKDGNSQCPQCKTRYKWHKGSPOV 939
          +CQICG+DVGL ADG+ FVACN CAFVPCRPCY+YE KDGNSQCPQCKTRYK HKGSP+V
Sbjct 3 VCQICGDDVGLTADGDLFVACNVCAFVPCRPCYDYERKDGNSQCPQCKTRYKMHKGSPOV 62

Query 940 DGDKEDECADDLHDHFN-STQGNRNEKQKIAEAMLHWQMA YGRGEDVGP SRSESQELPQL 1116
          +GD+ ++ ADD+ ++++ G+RNEKQ+IAEAML WQM+YGRGEDVG S QE+ +
Sbjct 63 EGDEGEDGADDVGNEYHYPPPGSRNEKQKIAEAMLRWQMSYGRGEDVGA PTSTRQEVSES 122

Query 1117 QVPLITNGQAI SGELPAGSSEYRRIAAPPTGGGSGKRVHPLPFPDSTQTGQVRAEDPAKD 1296
          Q+P +TNGQ+ISGELPA S E+ + APP+ GG KRVHPLP+ D+++ QVR D ++D
Sbjct 123 QIPRLTNGQSI SGELPALSPEHS-VGAPPSSGGGSKRVHPLPYTDASRPAQVRIVDHSRD 181

Query 1297 FNSYGFGNVAWKERVESWKNKQDKNTLQVTS DTY YASEGKDGDIDGCVADEEDLQMSDEA 1476
          FNSYGFGNVAWKERVESWKNKQ+KN LQVT+ YASEGK GD+D + EDLQM+DEA
Sbjct 182 FNSYGFGNVAWKERVESWKNKQEKNTLQVTSNGDYASEGKGDVDFGGGENEDLQMNDEA 241

Query 1477 RQPLSRKVP IASSKINPYRMVIVLRLVILCFFFRYRILNPVRNAYGLWFTSVICEIWF AI 1656
```


ESTUDO DIRIGIDO

1. Bancos de dados públicos e internacionais: NCBI, EMBL, DDBJ;
2. Definição de Bioinformática;
3. Análise da sequência no NCBI;
4. Busca de sequências por similaridade;
5. BLAST e Banco de dados de sequências.