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,...

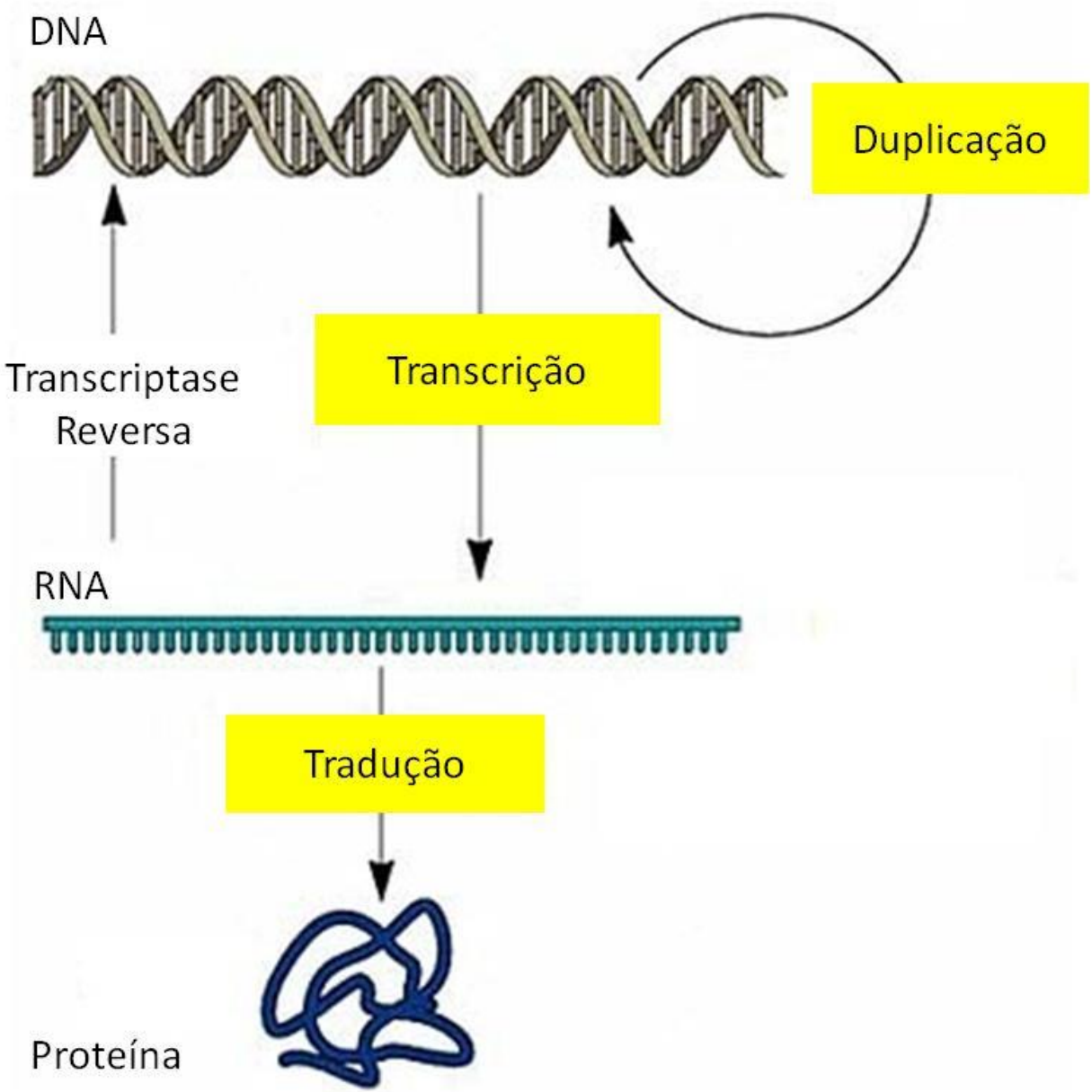
4. Utilização da plataforma BLAST

Dogma Central da Biologia Molecular

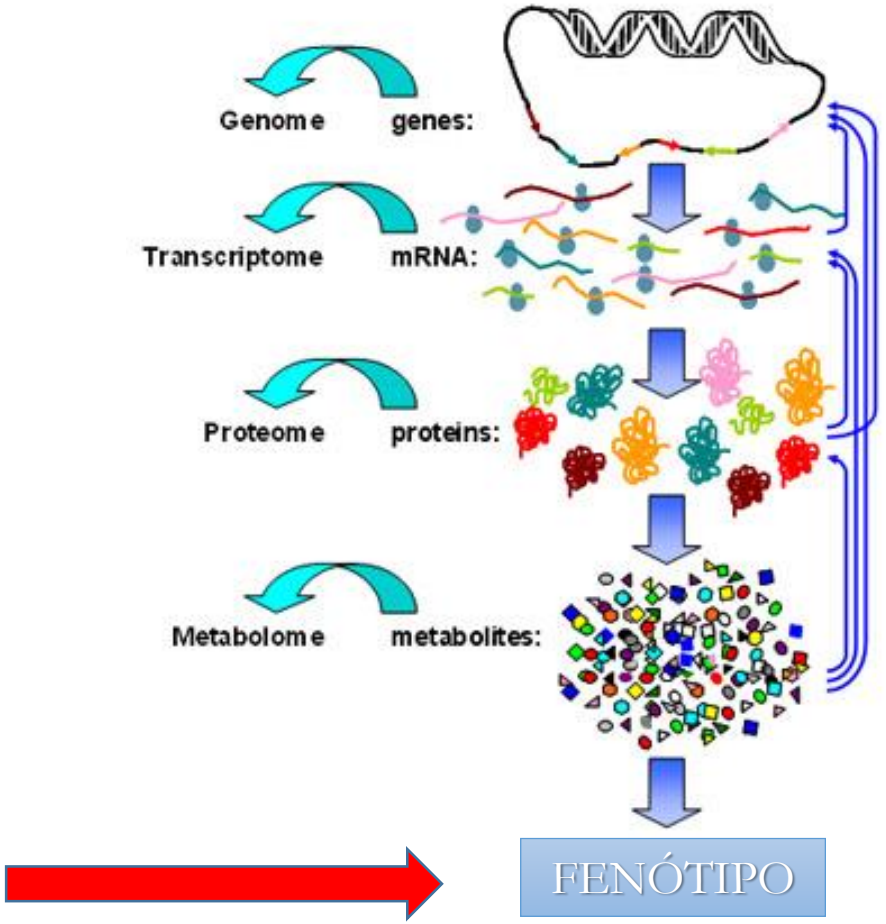
Genoma

Transcritoma

Proteoma



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.

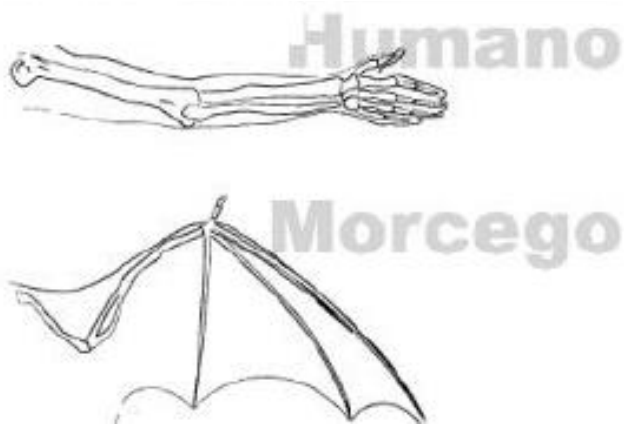
Homologia

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

ATCGGCAGTGACGTTTTATGATGATGAGATCATTGCTCTTGCACAGCCATTTAAACATTCCATGGTAGGA
AAGTTTTCACGTATGCCCCGGTTGAATGACATTAGGGTTGCTTTCAAAGGAATCGGGCTAGTGGGTGCAT
ATGAAATTCGTTGGTTGGATTATAAGCACATCCTGATTCATTTATCTAATGAGCAAGATCTGAATCATTT
ATGGATGCGTCAAGCATGGTTCATTGCAAACCAGAAGATGAGAGTCTTTAAGTGGACTCCGGATTTCCAA
TCGAAAAGGGAATCCTTCTTGGTTCCCGTTTGGGTCTCATTTTCGAACCTGCGGGCTCATCTATATGAAA
AATCGACACTTCCGATGATTGCTAAGTCGGTGGGGAGACCACTTTTTATTGATGAAGCTACGGCAAATGG
CACACGACCAAGTGTGGCCCGAGTGTGTGTTGAGTACGACTGCCAGCAGCCCCCTCTTGAACAGATCTGG
ATCGTGACTAGGGATAGAAGCACAGGAAACATCACTGGAGGATTTCAACAGAAAGTAGAGTTTGCCAGGC
TTCCTGACTATTGCAATCACTGTTGCCATGTGGGACATAGTATTGCAACATGTCTGGTGATGGGTCACAG
TAAGGACAAGCCAAGAAAGGCACGGCCTAAGCCCCTTGTGGATAAAAAGCAGGAAGATGATGATTGGAAA
AGAGAGAAAAGTAAGGAAACAGGTGATCTAATGGTTAATGGCGATAAAAGGAAAAATTCGATCCAAACAG
AATCGAAAAAGCAGAGCGTGAAATGGGTGAAGGTTGAAAAGGGTGGCACAAGCGGGTTCAAGGATGCCCA
CGGCGTAGAAGTCAATCTGGAGAGTAGTGGAGCAGATCCCGTGCAGATCTCGAATGGTTTTAGGGTGCTA
GAAGCAATGGAGGATGGCGGGGATGTTAGATCCGCAAAACAGGGGAGAACAGAGAAGGTGAACAGTACCA
TGCAATTTTTAAAAAATATTTTTAGGGAGAAAGAAAGGCAGTCGACGGAGATGGAAAGATGCTCGGGAAA
GATAAATGGCGACGAAACGACATTAGAAGCTCTACCGATAAAACGGACTGCAGATGGAGTGAATCGGGAC
AAGCTAAAATCTTCTACAGTGGGTGTGATCGAGGGTCCAAAGCAGAAGGAGAGTGAGGTTAAGCAAAGTT
CTGTGCAGACGTTGATGGCTGAAATTTGGCGGACAGGAGCAGATACTCACGAGAGTGTAGAAAATATTGC
AGACTTTGATCGAGTTCAATGGGCGATGGATGCAGGTCGTGTGACGTCCTGGAAGGCAAAAAAAAAGAGC
AACAGAAAACTTGAGGACCGACTGTCGGGGACGGCCGTGCAAGGTGATGGTCAGACAGTACCGGAGGTCG
AACAATGCTTGGGGAGTCCAAAACAGTGGGTGTACCGTCTAAACGTGGACGGTGAAAAGGTGCTGAAGGG
TGGTGAAAATGTGCAGTTGAGTCAACTCGACAGTAATAGTGTAGTGAGTTCTCGTGGCTGTCTTAAACTC
GGTACTGTTCACTCTCATGTAGCCAACTCCCGTGCGGTACATGCAGTGAAAGGAAGTATACACCGGTTGG
AAGAAAATGCTTTACTAGGGGAACCAGCAGCTAGTTCACGTGAAGTGATGGAAGAAAATGCAGAACACGA
TCCAAACTTGGGATCCAACCTGGGTATATGTGGTTACAATAAAGAAATAAGTTCGGTTCCTTCATGTGCA
GGAACTAATTCTGCTGACTTTCACGCACATTTGGAAGCAAACAAACAACAGGAGAACAACAATCGAGGGC
AAGTAAATCAAATCGAAACTGATGATAGCAGTAGATCAGTGCTCCATGTGGACTCGGGAGAGATTTTGGA
CAGCCAGCATATTAAATACCACCCCATGGTTTCCAGGAGAAGAAAATCCGATAGTGAAGTTATATATATC
CCTTCAGAGGATATTCTTTCAGAGAATGATGCTCATATGTTGATGGATGGGTCTGATGAAGAATCCATCT
CCAAGCAATTTACCACTAGAACTTACCCATGATCAGTGCCCTGCTTTGGAATGTAAGGGGAGTGACTGGA
AAAGCAATCCAAAGGAGAATTAAAAAACTGCAGATGATGCACCAAATAAAGATATTGGTTATCCTGGAAC
CAATGGTAACTGTTGATCGAATTGAATTTTTTAGGAGAAAATTAGGCTTTGAGGGGGCGGCCTTTAATTG
TTCTCAAAAAATTTGGATTTTTTGGATGCACGGCATCACTTGCACAACCAGGTTTGATCATCCCCAATGC
TTGCATGTTCAATTATGTTTCCCGTAGCTTCCTGTCCCTATTGAAGCTTCATTTGTTTATGCTAAATGTA
CTAGAATGGAACGACTTGCTTTATGGGATTTTATGAGACGTATTGCAGAGGATGTACAGGGTCCTTGGCT
GGCTGGAGGCGACTTTAATGTTATTTTAAGGTGAGAAGAGAGATTTTTGGGTGCAGACCCACATACTGGA
GCCATGGAAGATTTTGCAAATGCCTTACTTGATTGTGGGTTAGTAGATGCAGGGTTTGAAGGCAACAATT
TTACGTGGACTAACTCCCGGATGTTCCAAAGATTAGATCGGATTCTCTATAACCCACAGTGGGTAGCTCA

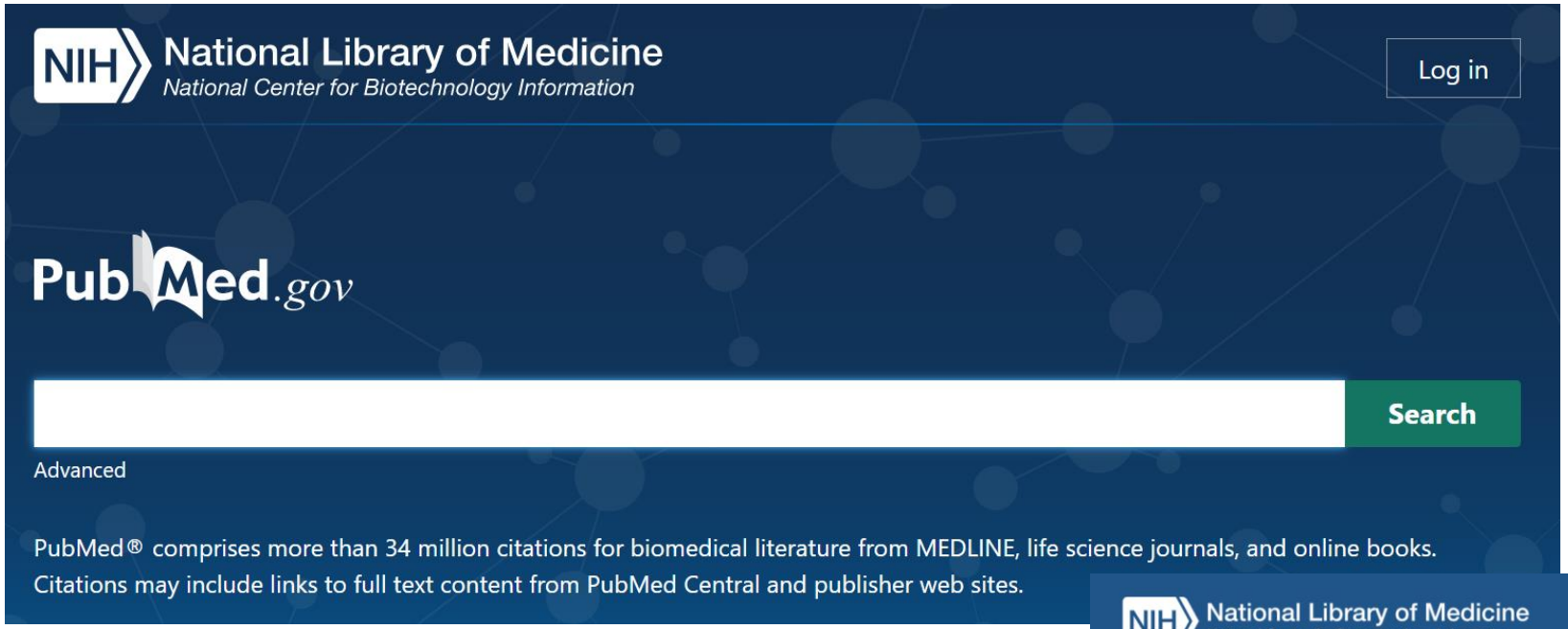
Genoma do cacaueiro

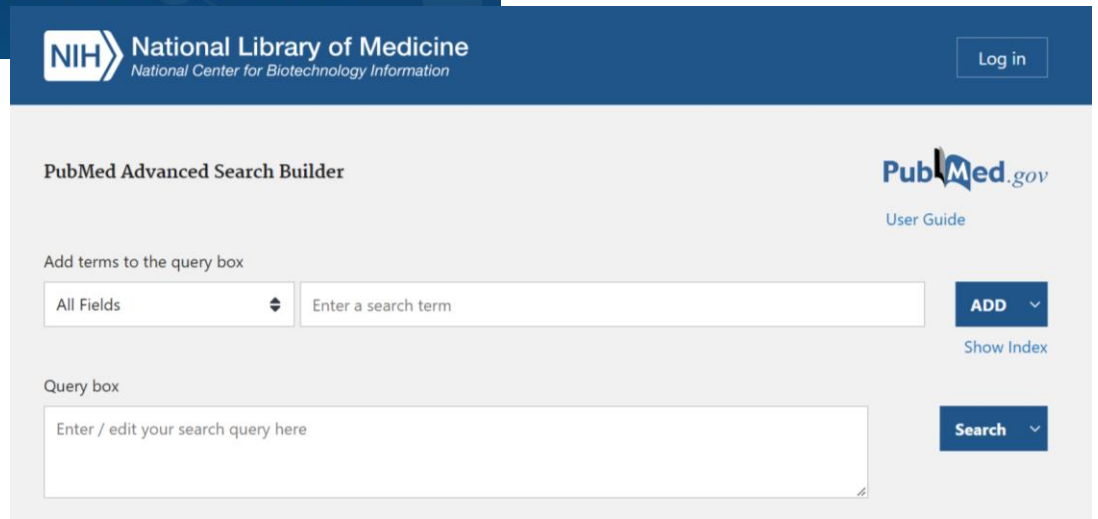
Banco de Dados

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



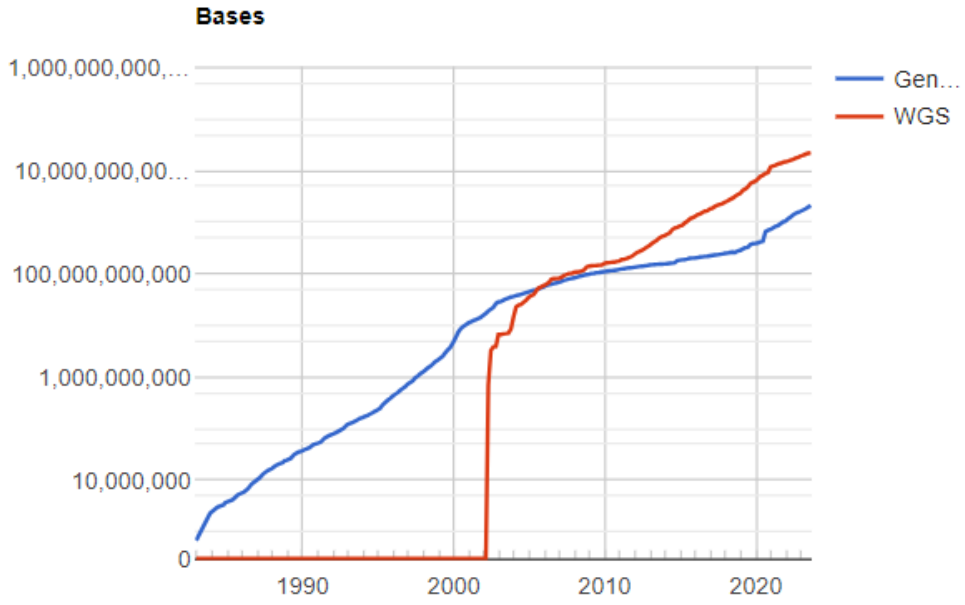


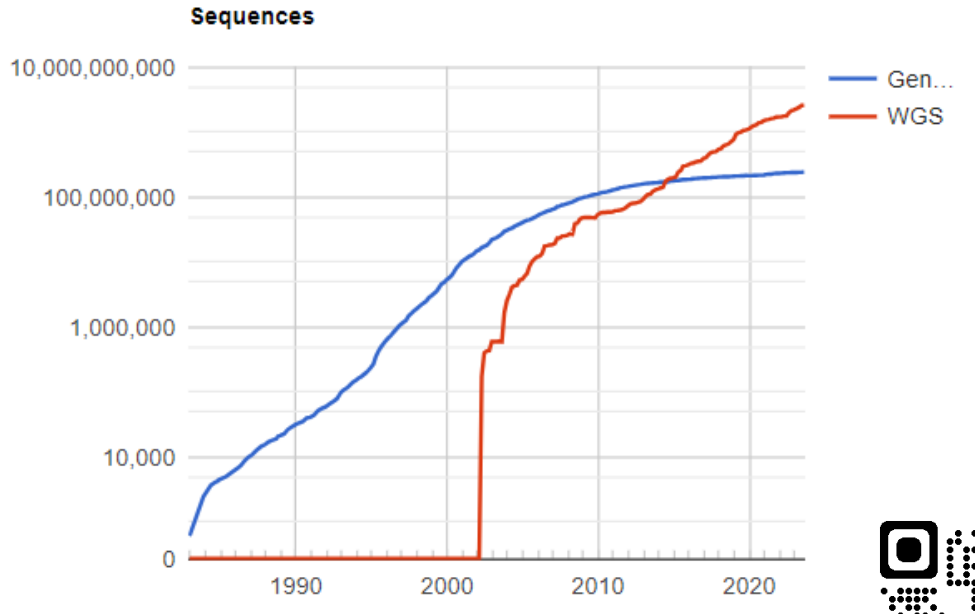
Origem do NCBI -> National Library of Medicine

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

Histórico de Sequências no NCBI

GenBank and WGS Statistics





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

International Nucleotide Sequence Database Collaboration

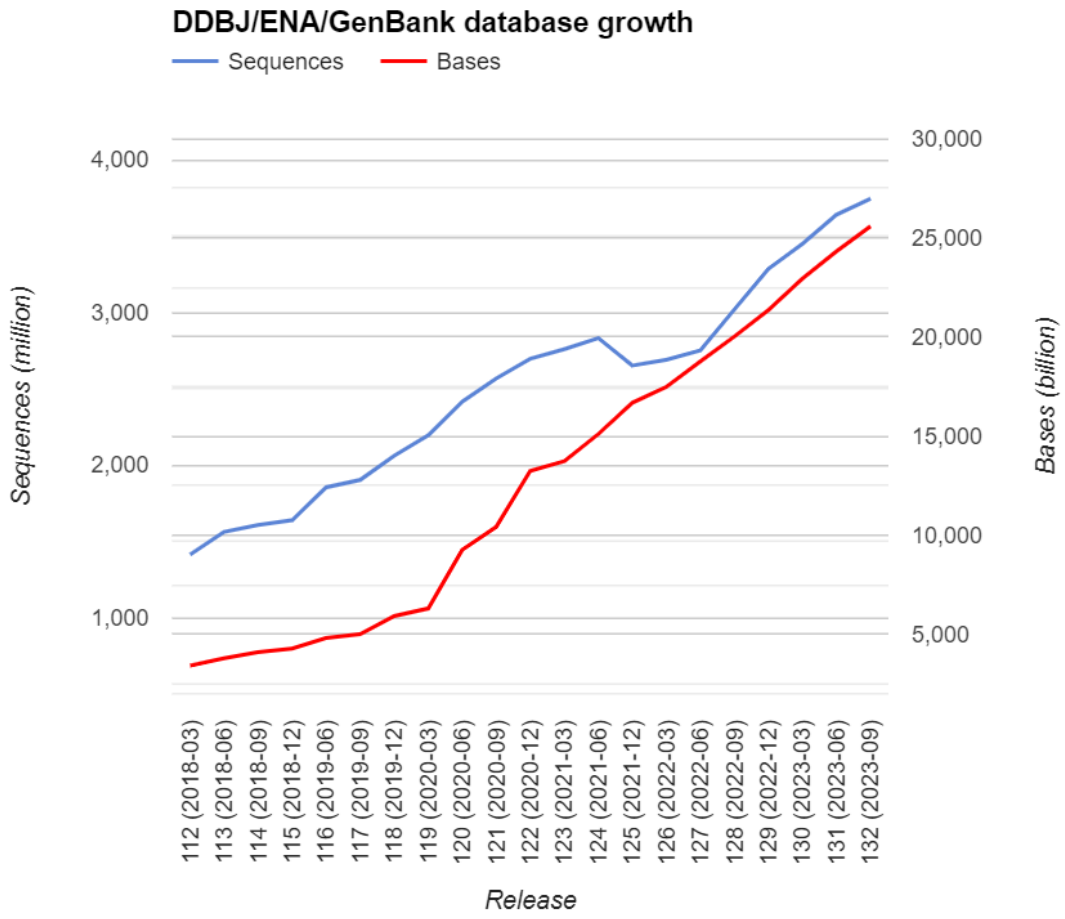
DDBJ DNA Data Bank of Japan NCBI ENA European Nucleotide Archive

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

insdc.org/

International Nucleotide Sequence Database Collaboration



https://www.ddbj.nig.ac.jp/statistics/index-e.html

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

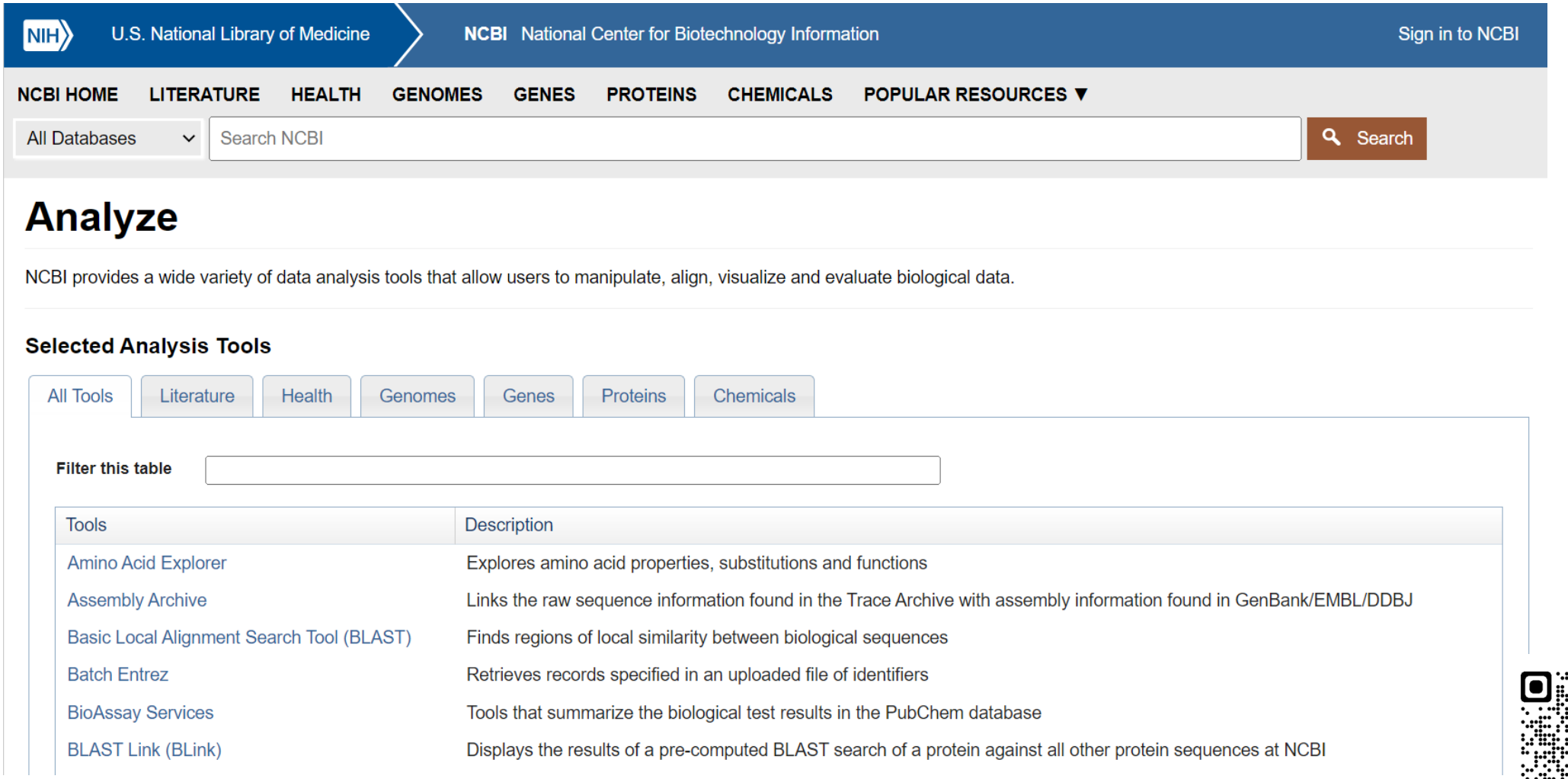
Announcing GenBank release 252.0

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



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

NCBI - TaxBrowser

NCBI Taxonomy Browser

Entrez | PubMed | Nucleotide | Protein | Genome | Structure | PMC | Taxonomy | BioCollections

Search for [ ] as [complete name] [x] lock [Go] [Clear]

Display [3] levels using filter: [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 a 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.

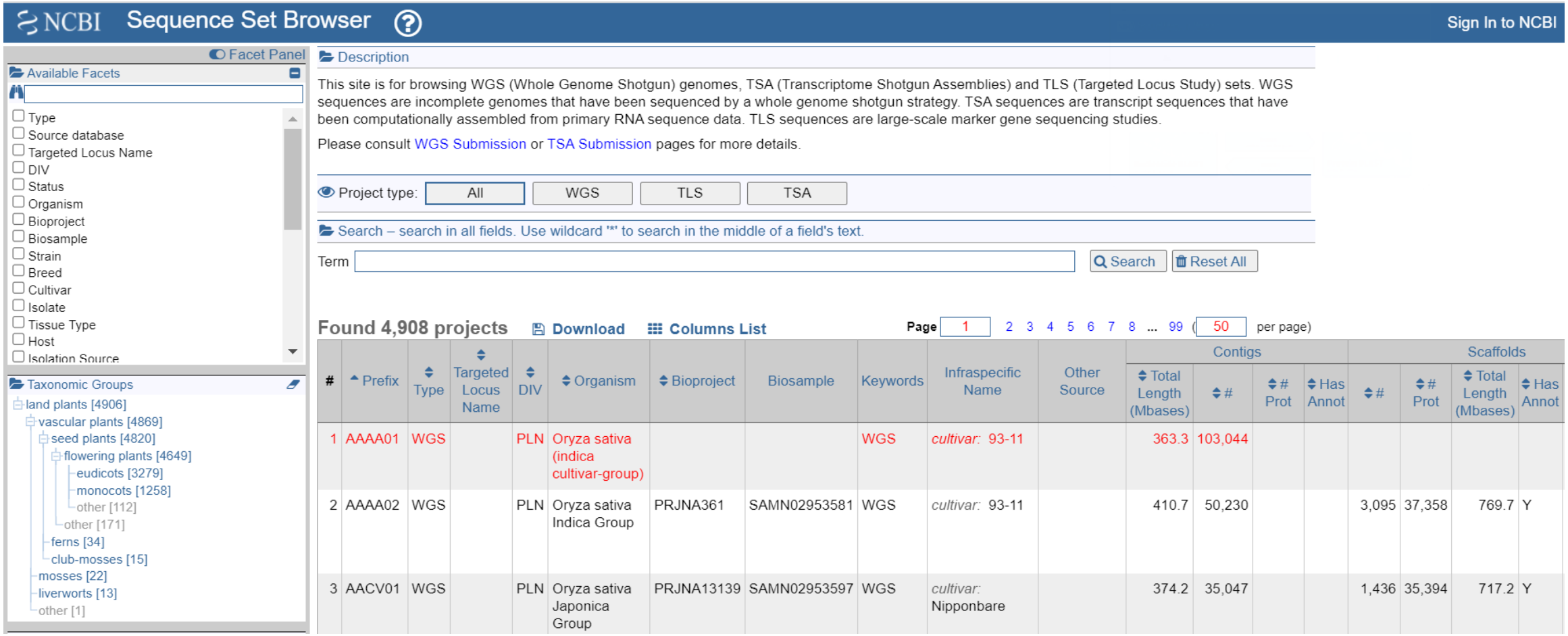
- Archaea
- Bacteria
- Eukaryota
- Viruses
- Other
- Unclassified

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

Arabidopsis thaliana	Escherichia coli	Pneumocystis carinii
Bos taurus	Hepatitis C virus	Rattus norvegicus
Caenorhabditis elegans	Homo sapiens	Saccharomyces cerevisiae
Chlamydomonas reinhardtii	Mus musculus	Schizosaccharomyces pombe
Danio rerio (zebrafish)	Mycoplasma pneumoniae	Takifugu rubripes
Dictyostelium discoideum	Oryza sativa	Xenopus laevis
Drosophila melanogaster	Plasmodium falciparum	Zea mays

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

NCBI - Genomas

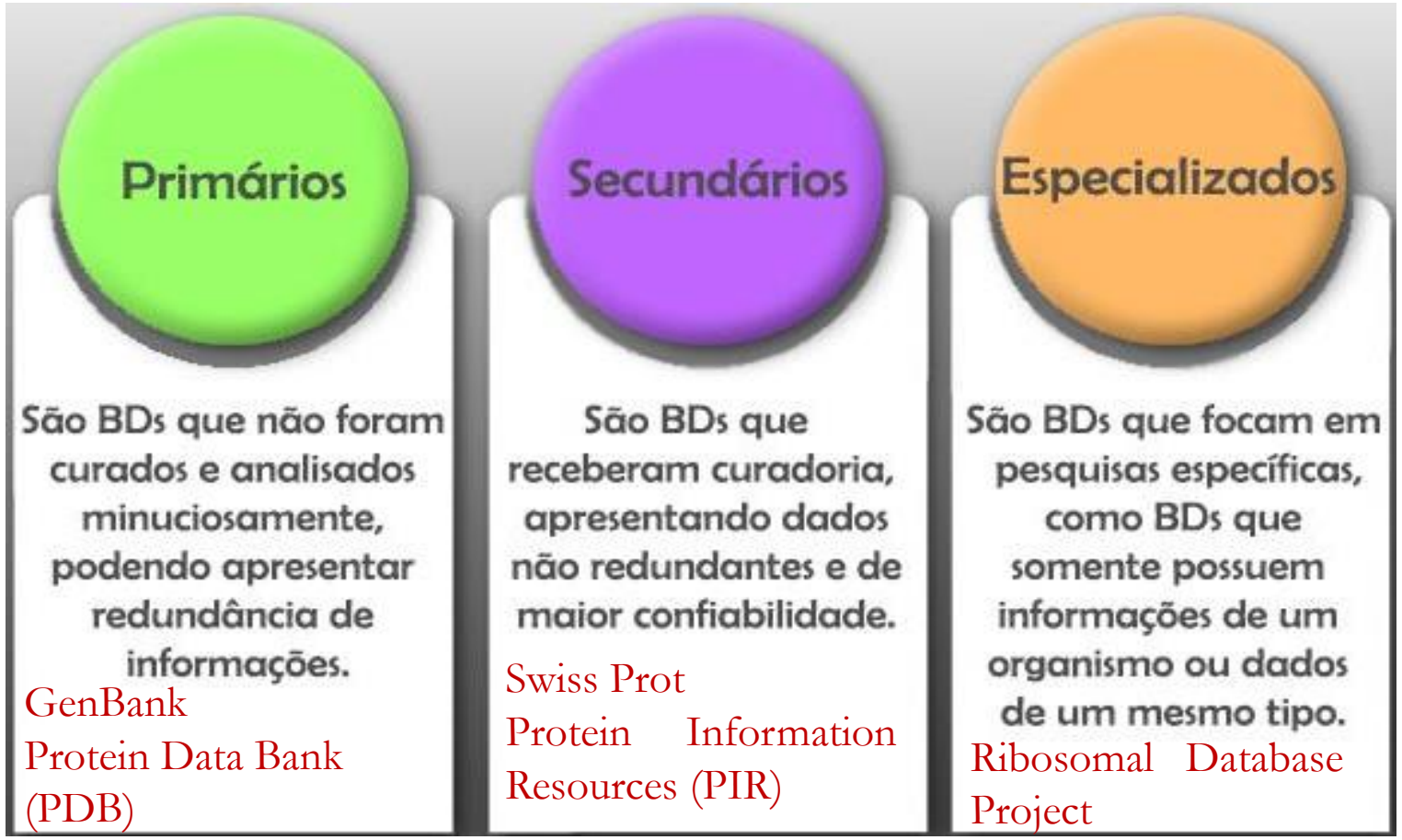


#	Prefix	Type	Targeted Locus Name	DIV	Organism	Bioproject	Biosample	Keywords	Infraspecific Name	Other Source	Contigs				Scaffolds			
											Total Length (Mbases)	#	# Prot	Has Annot	#	# Prot	Total Length (Mbases)	Has Annot
1	AAAA01	WGS		PLN	Oryza sativa (indica cultivar-group)			WGS	*cultivar:* 93-11		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



https://medium.com/omixdata/bancos-de-dados-biol%C3%B3gicos-parte-i-o-ncbi-c16dfc1b0a84

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

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/

Banco de Dados Biológicos (BDB)

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

https://medium.com/omixdata/bancos-de-dados-biol

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

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

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

National Library of Medicine
National Center for Biotechnology Information

Log in

Search NCBI | cp4 epsps agrobacterium | Search

Results found in 6 databases

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

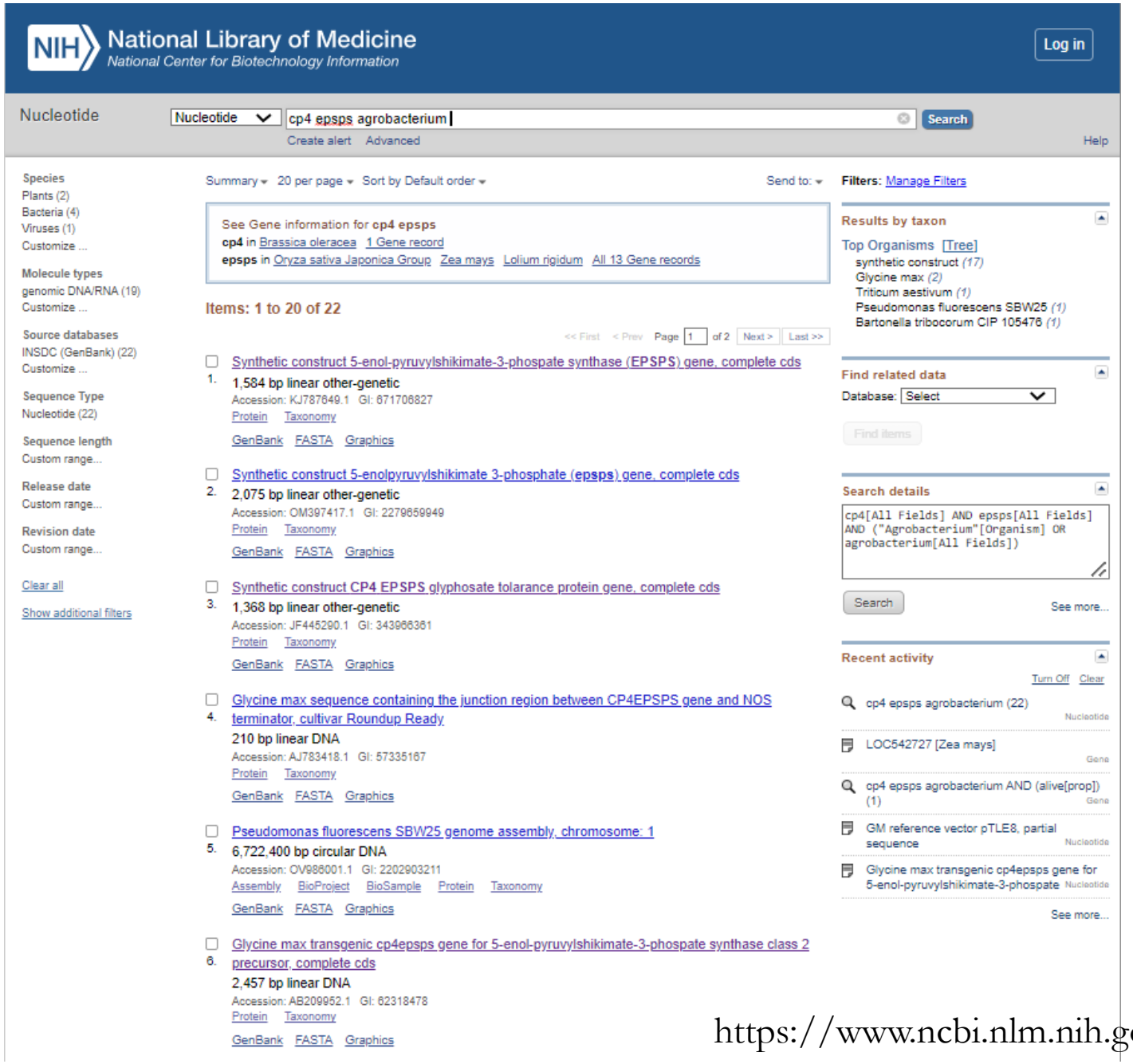
Genes	
Gene	1
GEO DataSets	0
GEO Profiles	0
HomoloGene	0
PopSet	0

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



https://www.ncbi.nlm.nih.gov/nuccore/AB209952.1

GenBank ▾ Send to: ▾

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

cp4 epsps agrobacterium (22) Nucleotide

LOC542727 [Zea mays] Gene

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

https://www.ncbi.nlm.nih.gov/nuccore/AB209952.1

```
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-phospate synthase (EPSPS)
                     class 2 precursor"
                     /codon_start=1
                     /product="5-enol-pyruvylshikimate-3-phospate synthase
                     class 2 precursor"
                     /protein_id="BAD94823.1"
                     /translation="MAQINNMAQGIQTLNPNSNFHKPQVPKSSSFLVFGSKKLKNSAN
                     SMLVLKKDSIFMQKFCSFRISASVATACMLHGASSRPATARKSSGLSGTVRIPGDKSI
                     SHRSFMFGGLASGETRITGLLEGEDVINTGKAMQAMGARIRKEGDTWIIDGVGNGGLL
                     APEAPLDFGNAATGCRLTMGLVGVYDFDSTFIGDASLTKRPMGRVLNPLREMGVQVKS
                     EDGDRLPVTLRGPKTPTPITYRVPMASAQVKSAVLLAGLNTPGITTVIEPIMTCDHTE
                     KMLQGFGANLTVETDADGVRTIRLEGRGKLTGQVIDVPGDPSSTAFPLVAALLVPGSD
                     VTILNVLMNPTRTGLILTLQEMGADIEVINLRLAGGEDVADLRVRSSTLKGVTVPEDR
                     APPMIDEYPILAVAAAFAEGATVMNGLEELRVKESDRLSAVANGLKLNGVDCDEGETS
                     LVVRGRPDGKGLGNASGAAVATHLDHRIAMSFLVMGLVSENPVTVDDATMIATSFPEF
                     MDLMAGLGAKIELSDTKAA"
```

```
                tumefaciens"
     repeat_region   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 caaacccttа atcccaattc caatttccat
      361 aaaccccaag ttcctaaatc ttcaagtttt cttgtttttg gatctaaaaa actgaaaaat
      421 tcagcaaatt ctatgttggt tttgaaaaaa gattcaattt ttatgcaaaa gttttgttcc
      481 tttaggattt cagcatcagt ggctacagcc tgcatgcttc acggtgcaag cagccggccc
      541 gcaaccgccc gcaaatcctc tggcctttcc ggaaccgtcc gcattcccgg cgacaagtcg
      601 atctcccacc ggtccttcat gttcggcggt ctcgcgagcg gtgaaacgcg catcaccggc
      661 cttctggaag gcgaggacgt catcaatacg ggcaaggcca tgcaggccat gggcgccagg
      721 atccgtaagg aaggcgacac ctggatcatc gatggcgtcg gcaatggcgg cctcctggcg
      781 cctgaggcgc cgctcgattt cggcaatgcc gccacgggct gccgcctgac catgggcctc
      841 gtcggggtct acgatttcga cagcaccttc atcggcgacg cctcgctcac aaagcgcccg
      901 atgggccgcg tgttgaaccc gctgcgcgaa atgggcgtgc aggtgaaatc ggaagacggt
      961 gaccgtcttc ccgttacctt gcgcgggccg aagacgccga cgccgatcac ctaccgcgtg
     1021 ccgatggcct ccgcacaggt gaagtccgcc gtgctgctcg ccggcctcaa cacgcccggc
     1081 atcacgacgg tcatcgagcc gatcatgacg tgcgatcata cggaaaagat gctgcagggc
     1141 tttggcgcca accttaccgt cgagacggat gcggacggcg tgcgcaccat ccgcctggaa
     1201 ggccgcggca agctcaccgg ccaagtcatc gacgtgccgg gcgacccgtc ctcgacggcc
     1261 ttcccgctgg ttgcggccct gcttgttccg ggctccgacg tcaccatcct caacgtgctg
     1321 atgaacccca cccgcaccgg cctcatcctg acgctgcagg aaatgggcgc cgacatcgaa
     1381 gtcatcaacc tgcgccttgc cggcggcgaa gacgtggcgg acctgcgcgt tcgctcctcc
     1441 acgctgaagg gcgtcacggt gccggaagac cgcgcgcctc cgatgatcga cgaatatccg
     1501 attctcgctg tcgccgccgc cttcgcggaa ggggcgaccg tgatgaacgg tctggaagaa
     1561 ctccgcgtca aggaaagcga ccgcctctcg gccgtcgcca atggcctcaa gctcaatggc
     1621 gtggattgcg atgagggcga gacgtcgctc gtcgtgcgtg gccgccctga cggcaagggg
     1681 ctcggcaacg cctcgggcgc cgccgtcgcc acccatctcg atcaccgcat cgccatgagc
     1741 ttcctcgtca tgggcctcgt gtcggaaaac cctgtcacgg tggacgatgc cacgatgatc
     1801 gccacgagct tcccggagtt catggacctg atggccgggc tgggcgcgaa gatcgaactc
     1861 tccgatacga aggctgcctg atgagctcga attcgagctc ggtaccggat ccaattcccg
     1921 atcgttcaaa catttggcaa taaagtttct taagattgaa tcctgttgcc ggtcttgcga
     1981 tgattatcat ataatttctg ttgaattacg ttaagcatgt aataattaac atgtaatgca
     2041 tgacgttatt tatgagatgg gtttttatga ttagagtccc gcaattatac atttaatacg
     2101 cgatagaaaa caaaatatag cgcgcaaact aggataaatt atcgcgcgcg gtgtcatcta
```

```
298..1881
/gene="cp4epsps"
/note="5-enol-pyruvylshikimate-3-phospate synthase (EPSPS)
class 2 precursor"
/codon_start=1
/product="5-enol-pyruvylshikimate-3-phospate synthase
class 2 precursor"
/protein_id=" BAD94823.1 "
/translation="MAQINNMAQGIQTLNPNSNFHKPQVPKSSSFLVFGSKKLKNSAN
SMLVLKKDSIFMQKFCSFRISASVATACMLHGASSRPATARKSSGLSGTVRIPGDKSI
SHRSFMFGGLASGETRITGLLEGEDVINTGKAMQAMGARIRKEGDTWIIDGVGNGGLL
APEAPLDFGNAATGCRLTMGLVGVYDFDSTFIGDASLTKRPMGRVLNPLREMGVQVKS
EDGDRLPVTLRGPKTPTPITYRVPMASAQVKSAVLLAGLNTPGITTVIEPIMTCDHTE
KMLQGFGANLTVETDADGVRTIRLEGRGKLTGQVIDVPGDPSSTAFPLVAALLVPGSD
VTILNVLMNPTRTGLILTLQEMGADIEVINLRLAGGEDVADLRVRSSTLKGVTVPEDR
APPMIDEYPILAVAAAFAEGATVMNGLEELRVKESDRLSAVANGLKLNGVDCDEGETS
LVVRGRPDGKGLGNASGAAVATHLDHRIAMSFLVMGLVSENPVTVDDATMIATSFPEF
MDLMAGLGAKIELSDTKAA"
```

FASTA ▾ Send to: ▾

Glycine max transgenic cp4epsps gene for 5-enol-pyruvylshikimate-3-phospate 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-
phospate synthase class 2 precursor, complete cds
ATGGCACAAATTAACAACATGGCACAAGGGATACAAACCCTTAATCCCAATTCCAATTTCCATAAACCCC
AAGTTCCTAAATCTTCAAGTTTTCTTGTTTTTGGATCTAAAAAACTGAAAAATTCAGCAAATTCTATGTT
GGTTTTGAAAAAAGATTCAATTTTTATGCAAAAGTTTTGTTCCTTTAGGATTTCAGCATCAGTGGCTACA
GCCTGCATGCTTCACGGTGCAAGCAGCCGGCCCGCAACCGCCCGCAAATCCTCTGGCCTTTCCGGAACCG
TCCGCATTCCCGGCGACAAGTCGATCTCCCACCGGTCCTTCATGTTCGGCGGTCTCGCGAGCGGTGAAAC
GCGCATCACCGGCCTTCTGGAAGGCGAGGACGTCATCAATACGGGCAAGGCCATGCAGGCCATGGGCGCC
AGGATCCGTAAGGAAGGCGACACCTGGATCATCGATGGCGTCGGCAATGGCGGCCTCCTGGCGCCTGAGG
CGCCGCTCGATTTCGGCAATGCCGCCACGGGCTGCCGCCTGACCATGGGCCTCGTCGGGGTCTACGATTT
CGACAGCACCTTCATCGGCGACGCCTCGCTCACAAAGCGCCCGATGGGCCGCGTGTTGAACCCGCTGCGC
GAAATGGGCGTGCAGGTGAAATCGGAAGACGGTGACCGTCTTCCCGTTACCTTGCGCGGGCCGAAGACGC
CGACGCCGATCACCTACCGCGTGCCGATGGCCTCCGCACAGGTGAAGTCCGCCGTGCTGCTCGCCGGCCT
CAACACGCCCGGCATCACGACGGTCATCGAGCCGATCATGACGTGCGATCATACGGAAAAGATGCTGCAG
GGCTTTGGCGCCAACCTTACCGTCGAGACGGATGCGGACGGCGTGCGCACCATCCGCCTGGAAGGCCGCG
GCAAGCTCACCGGCCAAGTCATCGACGTGCCGGGCGACCCGTCCTCGACGGCCTTCCCGCTGGTTGCGGC
CCTGCTTGTTCCGGGCTCCGACGTCACCATCCTCAACGTGCTGATGAACCCCACCCGCACCGGCCTCATC
CTGACGCTGCAGGAAATGGGCGCCGACATCGAAGTCATCAACCTGCGCCTTGCCGGCGGCGAAGACGTGG
CGGACCTGCGCGTTCGCTCCTCCACGCTGAAGGGCGTCACGGTGCCGGAAGACCGCGCGCCTCCGATGAT
CGACGAATATCCGATTCTCGCTGTCGCCGCCGCCTTCGCGGAAGGGGCGACCGTGATGAACGGTCTGGAA
GAACTCCGCGTCAAGGAAAGCGACCGCCTCTCGGCCGTCGCCAATGGCCTCAAGCTCAATGGCGTGGATT
GCGATGAGGGCGAGACGTCGCTCGTCGTGCGTGGCCGCCCTGACGGCAAGGGGCTCGGCAACGCCTCGGG
CGCCGCCGTCGCCACCCATCTCGATCACCGCATCGCCATGAGCTTCCTCGTCATGGGCCTCGTGTCGGAA
AACCCTGTCACGGTGGACGATGCCACGATGATCGCCACGAGCTTCCCGGAGTTCATGGACCTGATGGCCG
GGCTGGGCGCGAAGATCGAACTCTCCGATACGAAGGCTGCCTGA
```

Sequencia formato FASTA

Nucleotide Nucleotide

Advanced

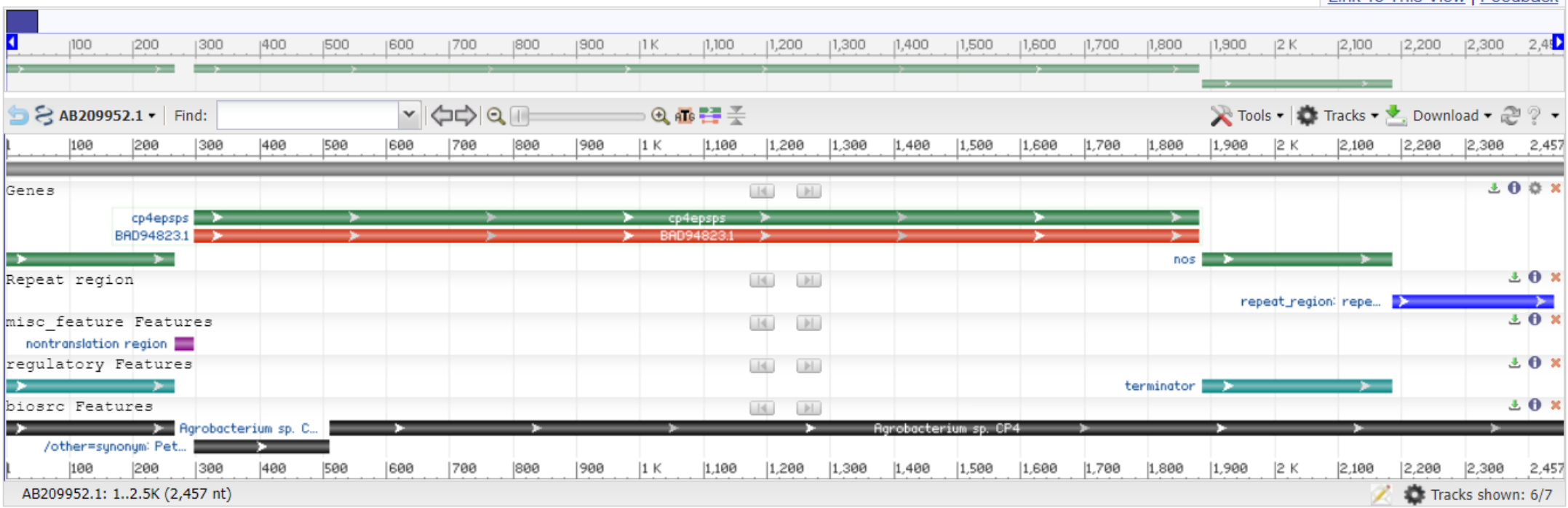
Graphics Send to:

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

GenBank: AB209952.1

GenBank FASTA

Link To This View | Feedback



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

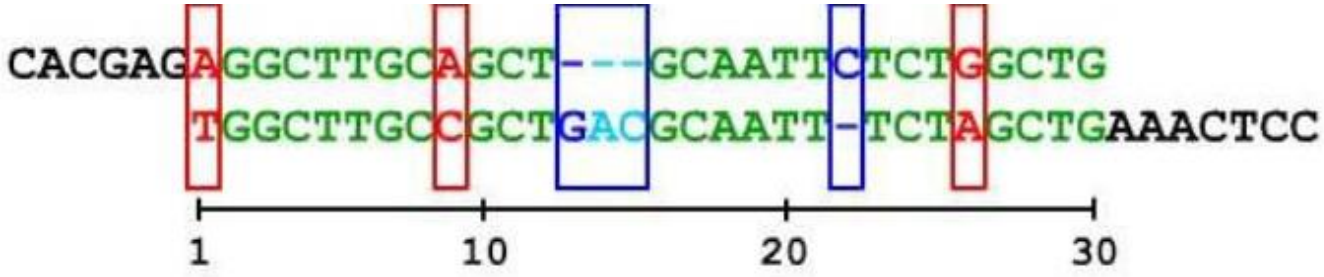
Basic Local Alignment Search Tool

Comparando sequências - Alinhamento

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

```
seq1   GCTCATTCGACCTGACTAG
seq2   GTGACTAAGACCTCTCATT
```

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 sequencias consiste em comparar duas sequencias (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 sequencias de proteínas expressa em percentual de amino ácidos com característica similar alinhados

```
RBP:          26  RV**K**ENFDKARFS**GTW**YA**MA**KKDPEGLF**L**QDNIV**A** 59
                  + **K**++ + ++  **GTW**++**MA**   +    **L** +  **A**
glicodelina:  23  QT**K**QDLELPKLA**GTW**HS**MA**MA-TNNIS**L**MATLK**A** 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

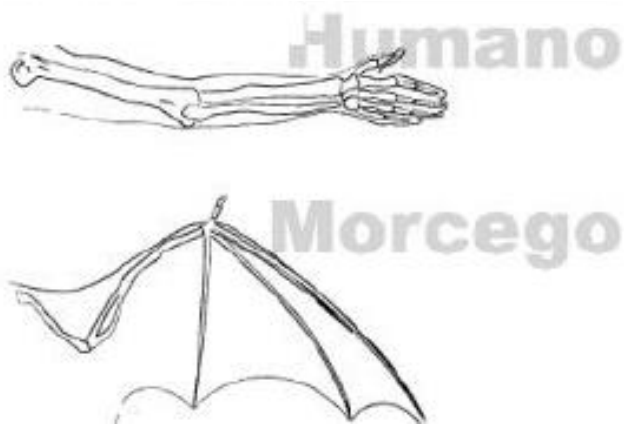
Homologia – conceito fundamental na biologia

Algoritmos em Blast:

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

- **Não avaliam homologia**

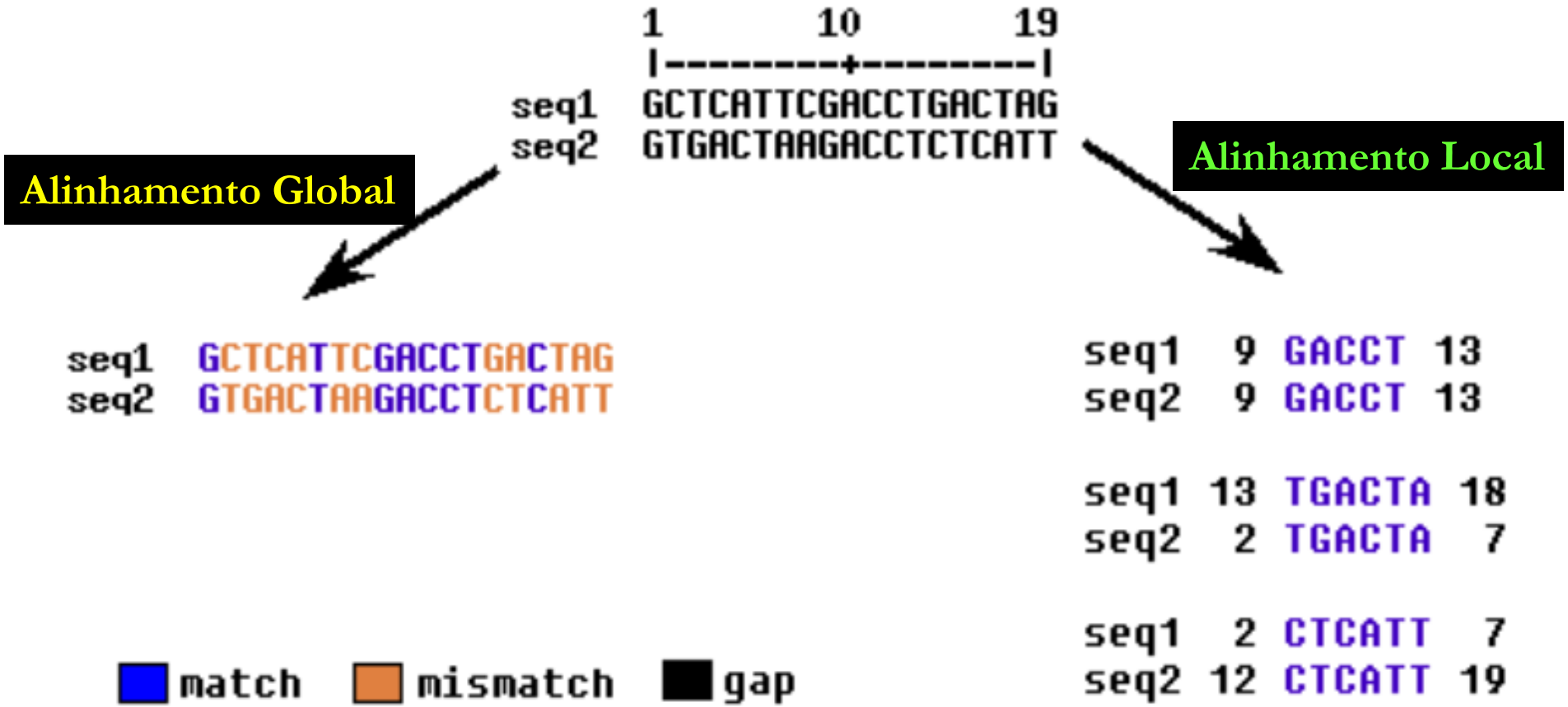
- Medem similaridade e identidade entre sequências



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

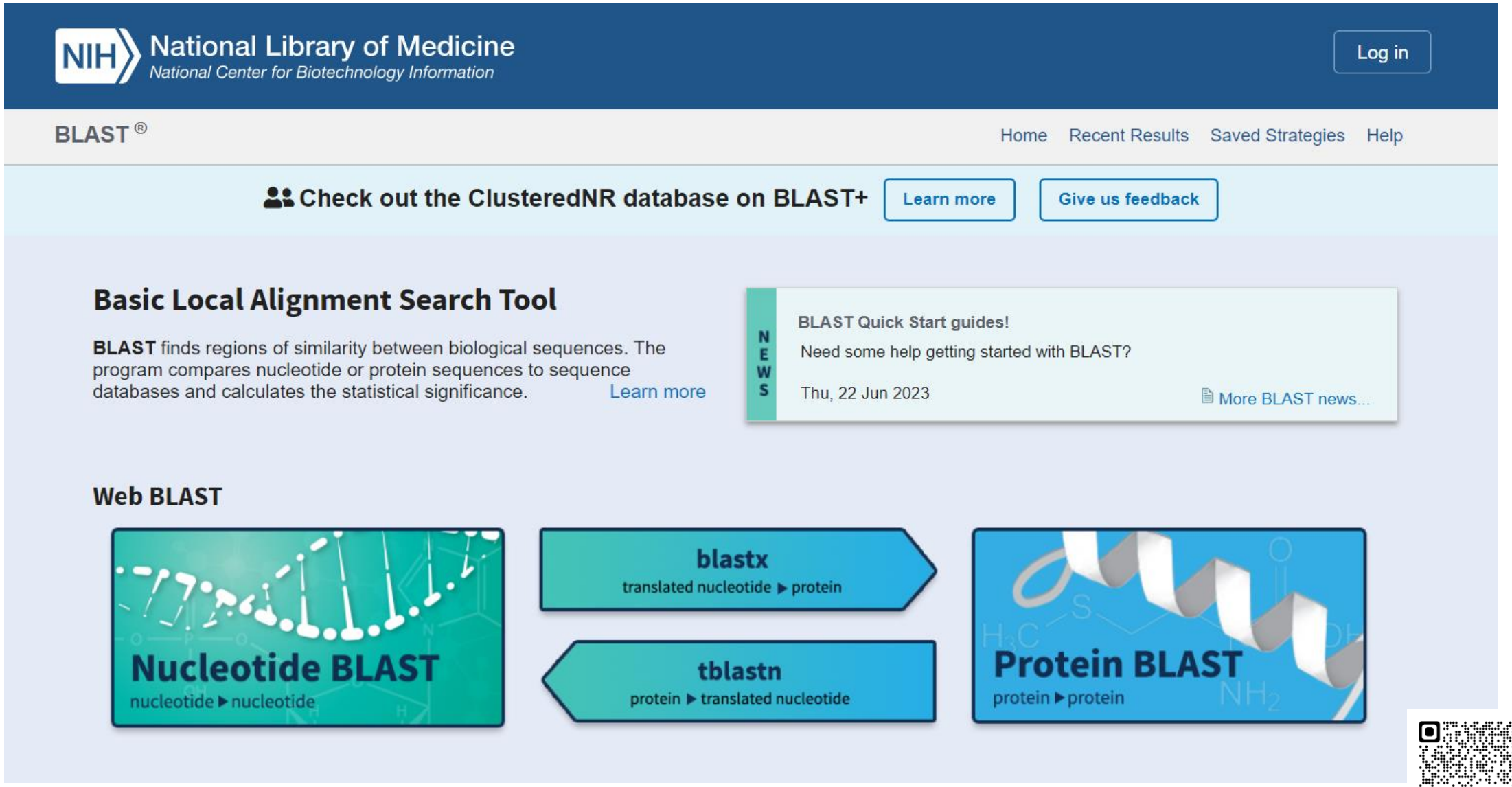
PREDICTED: Syzygium oleosum rubisco accumulation factor 1.1, chloroplastic (LOC115690120),

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    aataaaaaaaTC-AAAGCCATCAGTGAACTTCAGTTCAGATAAGGTTTCACTGAAAACCC  82
             || ||||| ||| |||||||||| | || ||||||||||||||||||||||||||| || |
Sbjct  56    AAGAAAAATATCAAAAGCCATCATTCAATTTCAGTTCAGATAAGGTTTCACTGAAGACTC  115

Query  83    ACAGTGACCTCCCTCCACTCCTCCCTGAACCACCAACTCCCGCCACCGCCACCGCC---T  139
             |||||||||  ||||||| |||| ||| ||||||  |||||||| ||||||||||||
Sbjct  116   ACAGTGACCCTCCTCCAC-CCTCACTGCACCACCGCCTCCCGCCGCCGCCACCGCCGCCG  174

Query  140   CCG---CCACCATGCTCTCTCTAGCCCTCATCAAcaccacccac--cc-ccaccaccacc  193
             |||   ||||||||||||||||||||||||||||||||||||||  || ||||||||||||
Sbjct  175   CCGCCTCCACCATGCTCTCTCTAGCCCTCATCAACACCACCCACACCCACCACCACCACC  234

Query  194   aattcttcaatcccaccaccaccatcctcaaaccctcacttcctctcccctccaccacc  253
             |||||||||| |||||||||||| ||  |  | |||| |   || ||| | ||| |   |
Sbjct  235   AATTCTTCAACCCCACCACCACCA-CCAC-CA-CCCTTA-AGCCCCTCACTTCCCCTTTC  290

Query  254   cttgccgccaccaccgccgccaccactccgccggcccaaggcccatctccgccaGCCTCA  313
             | | || ||||| || || ||||||||| |||||||| ||||||||||||||||||||||
Sbjct  291   C-T-CCACCACCCCCACCACCACCACTCGGCCGGCCCGAGGCCCATCTCCGCCAGCCTCA  348

Query  314   ACCCGAACCCTCTAAtcccggcctcccctccgccgcagcagcagcggccctaccagccct  373
             |||||| ||||||||||||||||||||||||||||||||||||||   ||||||||||||
Sbjct  349   ACCCGAGCCCTCTAATCCCGGCCTCCCCTCCGCCGCAGCAGCAGCGTGTCTACCAGCCCT  408

Query  374   tccgccctccccc-tccccgct--cctcccagttccgctccctcGACACCGCCAGCCGCC  430
             |||||||||||| |||||||  ||||||||||||||||||||||||||||| ||||||
Sbjct  409   TCCGCCCTCCCCCCTCCCCGCTCCCCTCCCAGTTCCGCTCCCTCGACACCGCCGGCCGCC  468
```


National Library of Medicine
National Center for Biotechnology Information
Log in
BLAST®
Home
Recent Results
Saved Strategies
Help
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

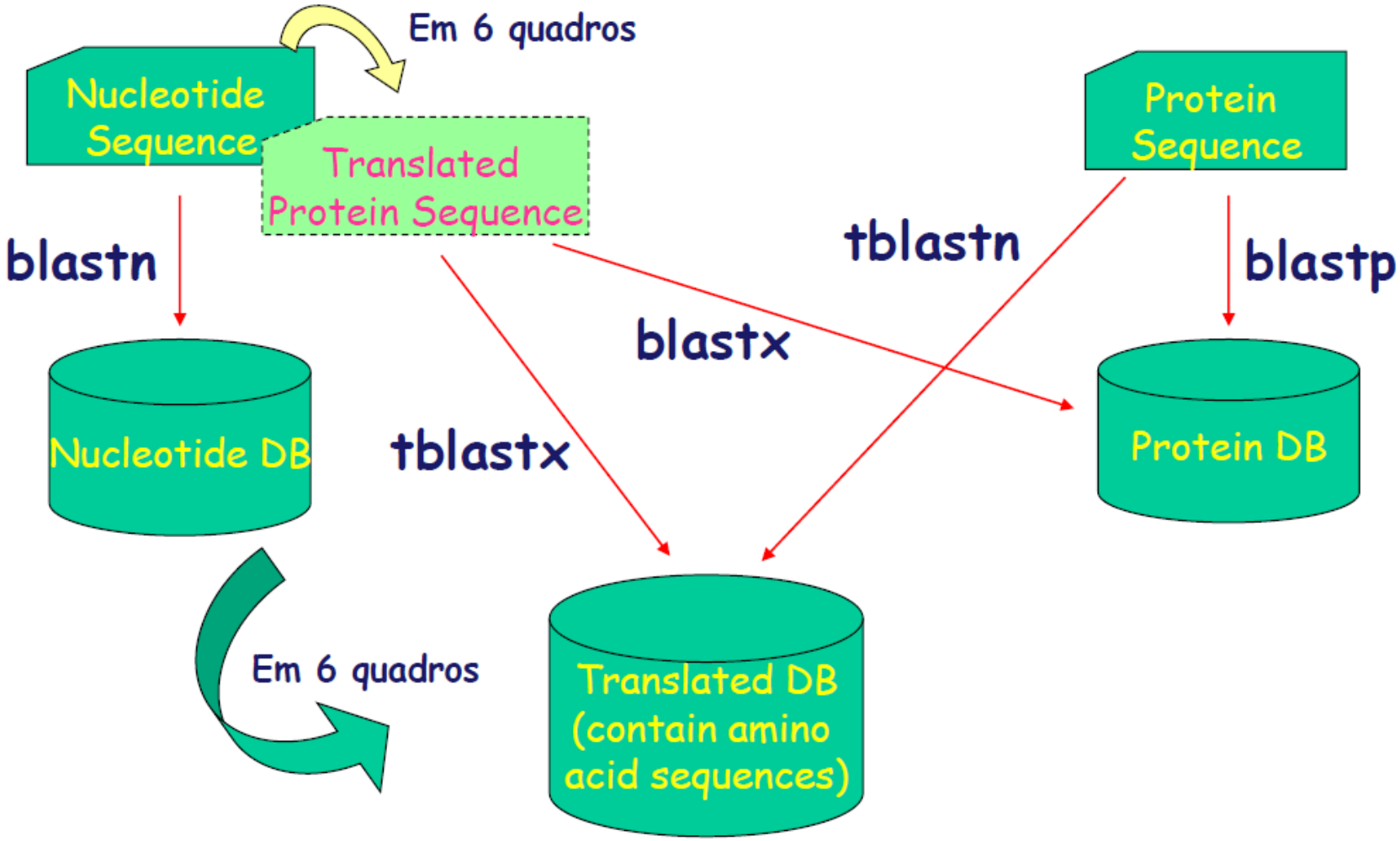
Tipos de BLAST de acordo com o tipo de sequencia fornecida e qual o tipo buscado

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

Programa	Sequência	Base	Comparação
blastn	DNA	DNA	DNA
blastp	PTN	PTN	PTN
blastx	DNA	PTN	PTN
tblastn	PTN	DNA	PTN
tblastx	DNA	DNA	PTN

Artigo BLAST

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
CCGGCGAAATTAAAGGTCACTACCTCAACGTTACCGCTCCTACCTGCGAAGAAATGTTGAAACGGGCTGA
GTACGCTAAAGAACTCAAAATGCCCATCATCATGCACGACTACCTAACCGCAGGTTTCACCGCTAACACC
ACATTGGCTCGTTGGTGTCGTGATAACGGTATTTTATTGCACATTCACCGTGCTATGCACGCTGTAATTG
ACCGTCAAAAAAATCACGGTATCCACTTCCGCGTATTAGCTAAAGCCCTCCGCTTGTCCGGTGGTGATCA
CATCCACACTGGTACAGTTGTTGGTAAGTTAGAAGGTGAACGCGGTATTACCATGGGCTTCGTTGACTTA
TTACGTGAAAACTACGTTGAGCAAGACAAGTCTCGCGGTATTTACTTTACCCAAGATTGGGCGTCTCTAC
CTGGTGTAATGGCCGTTGCTTCTGGTGGTATCCACGTATGGCATATGCCCGCGTTGGTTGAGATCTTCGG
TGATGACTCCGTATTACAATTCGGTGGTGGTACACTCGGACATCCTTGGGGTAACGCTCCTGGTGCTACA
GCTAACCGCGTAGCTCTAAAAGCAGTTGTTCAAGCTCGTAACGAAGGCCGTAACTTAGCTCGTGAAGGTA
ACGATATTATCCGCGAAGCTGCTAAGTGGTCTCCTGAGTTGGCTGTTGCTTGCGAACTG


>gi|226347323|gb|ACO50079.1| ribulose-1,5-bisphosphate
carboxylase/oxygenase large subunit [Anabaena planctonica CENA210]
GEIKGHYLNVTAPTCEEMLKRAEYAKELKMPIIMHDYLTAGFTANTTLARWCRDNGILLHIHRAMHAVID
RQKNHGIHFRVLAKALRLSGGDHIHTGTVVGKLEGERGITMGFVDLLRENYVEQDKSRGIYFTQDWASLP
GVMAVASGGIHVWHMPALVEIFGDDSVLQFGGGTLGHPWGNAPGATANRVALKAVVQARNEGRNLAREGN
DIIREAAKWSPELAVACEL
```

Busca por BLAST

- Nossa sequência –> ***query* (consulta)**

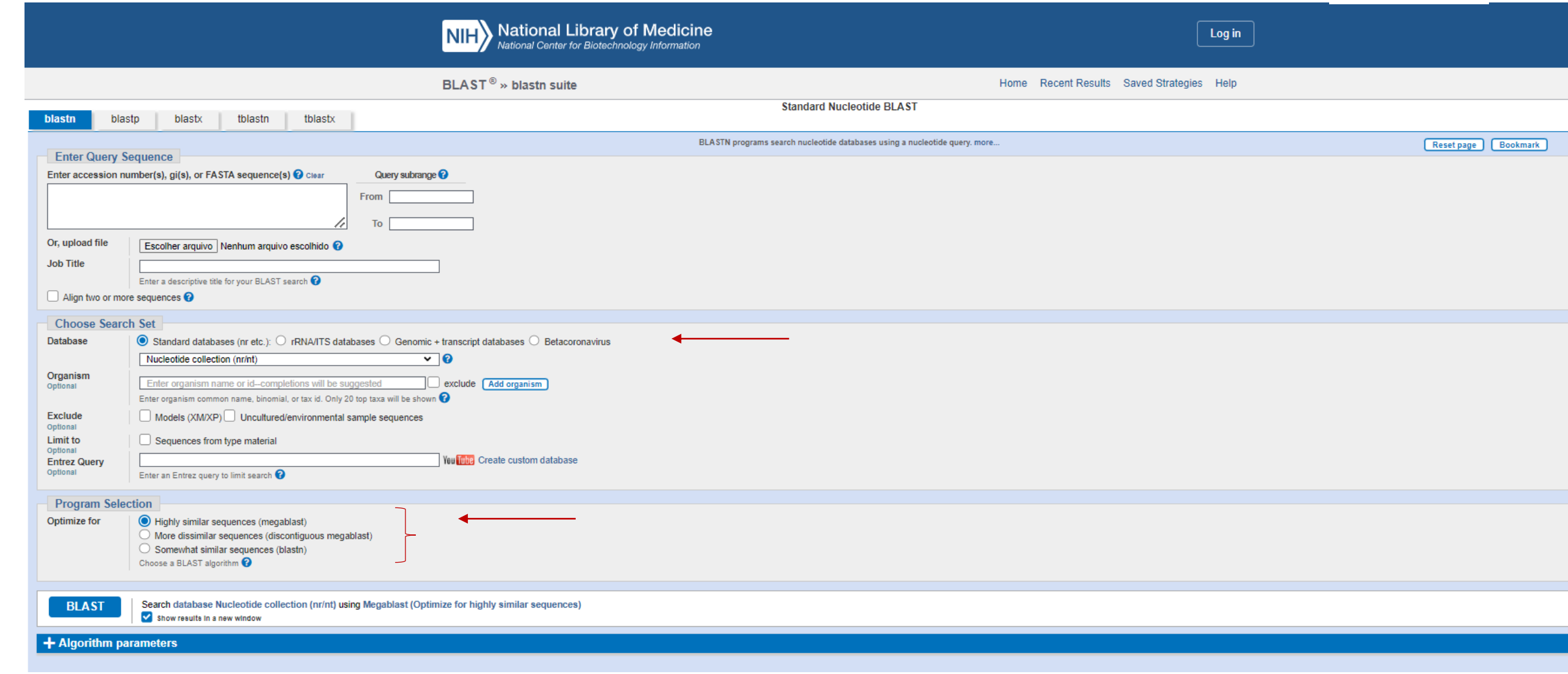
- O resultado da busca em BLAST pode ser um ou mais *hits* em sequências-sujeito (*subject*), ou seja, sequências pertencentes ao banco

```
Query: 181 catcaactacaactccaaagacacccttacacccactaggatatcaacaaacctacccac 240
           |||||||| |||| |||||| ||||| | ||||||||||||||||||||||||||||||
Sbjct: 189 catcaactgcaaccccaaagccacccct-cacccactaggatatcaacaaacctacccac 247
```

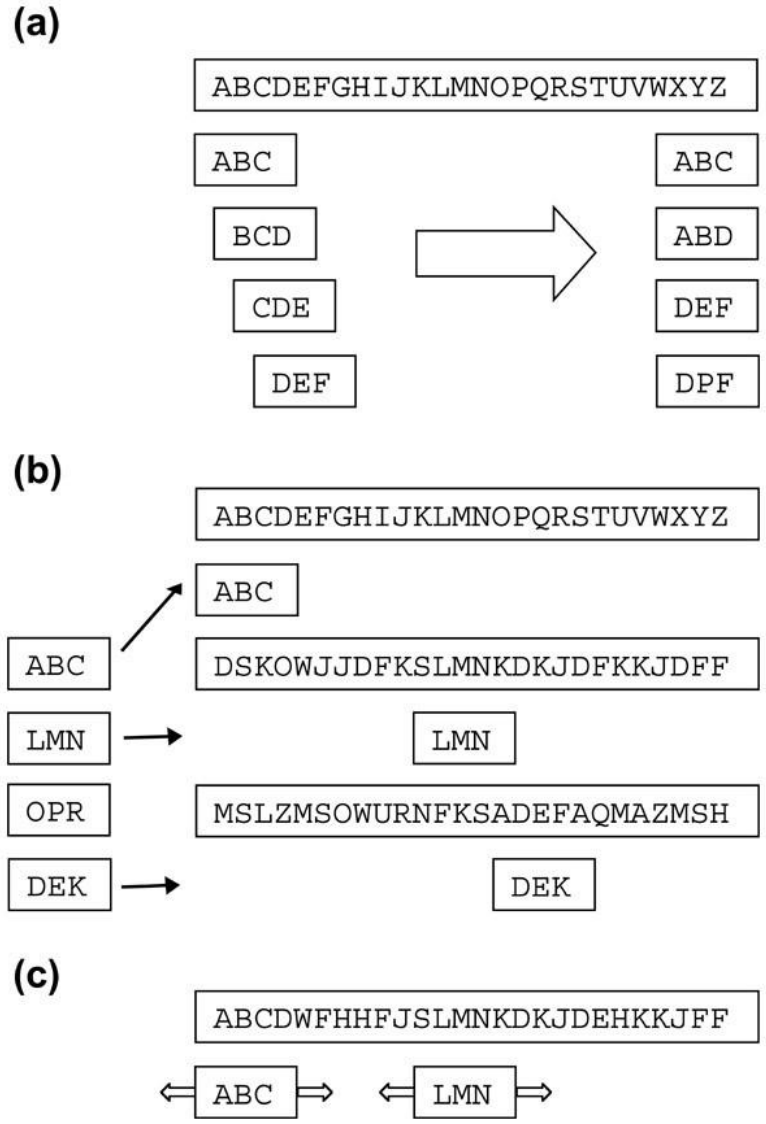
- Os melhores resultados de escores são relatados,
- usar valor *E*
  - valor *E* <0.01

Quanto menor o *e-value*, mais significativo o alinhamento!!!

Busca por BLAST

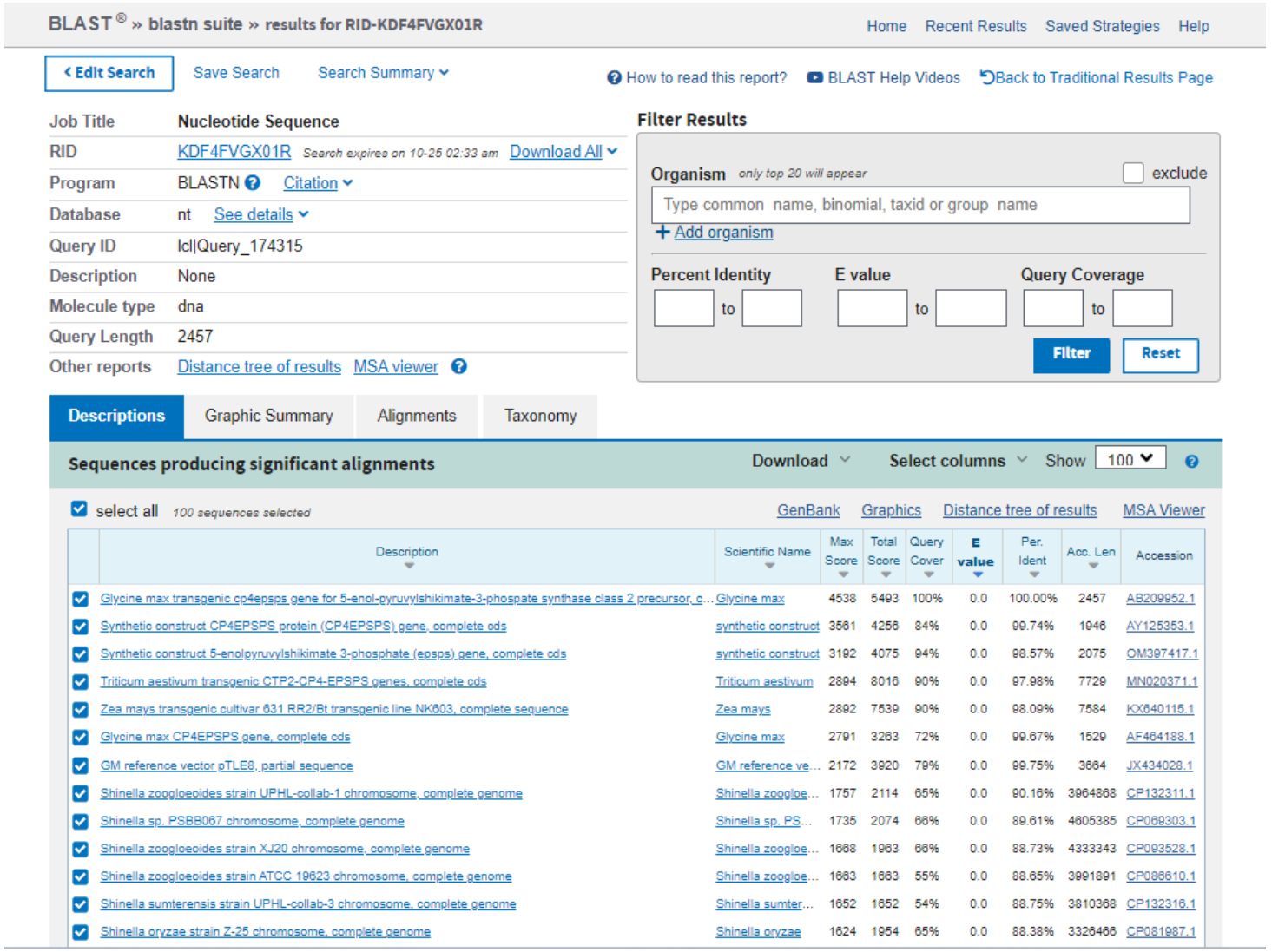


Busca por BLAST



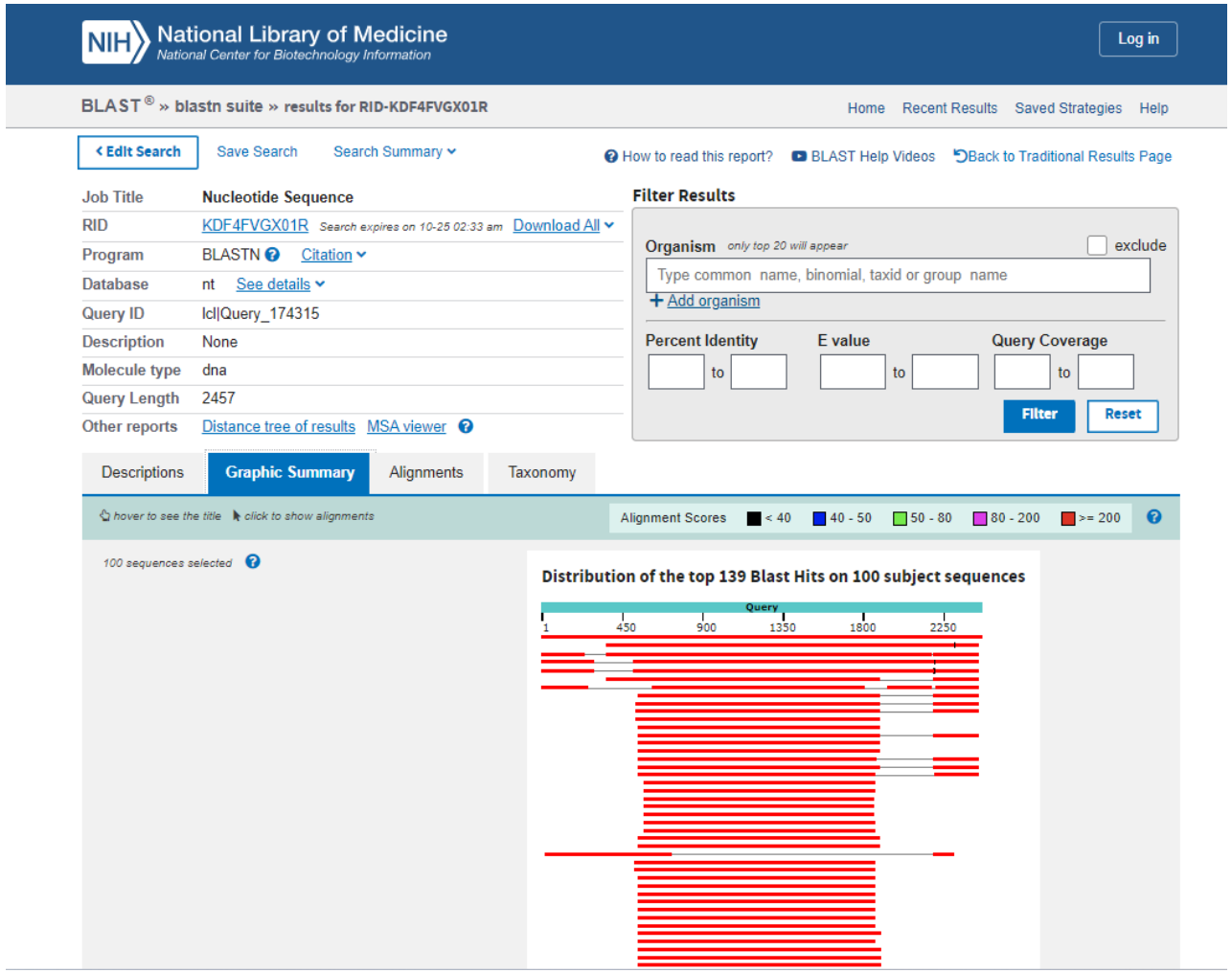
Glycine max transgenic cp4epsps gene for 5-enol-pyruvylshikimate-3-phospate synthase class 2 precursor, complete cds

Busca por BLAST



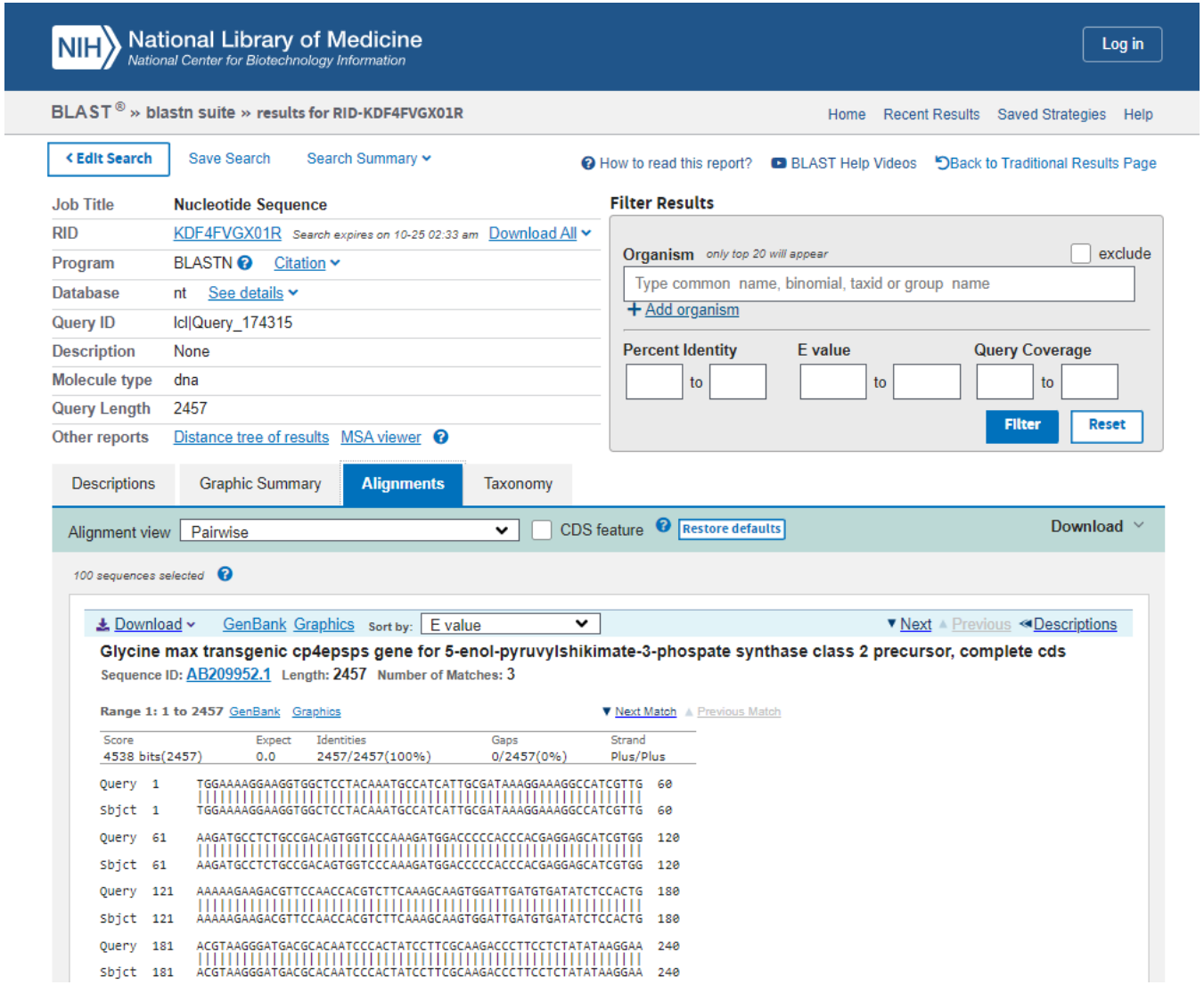
Glycine max transgenic cp4epsps gene for 5-enol-pyruvylshikimate-3-phospate synthase class 2 precursor, complete cds

Busca por BLAST



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

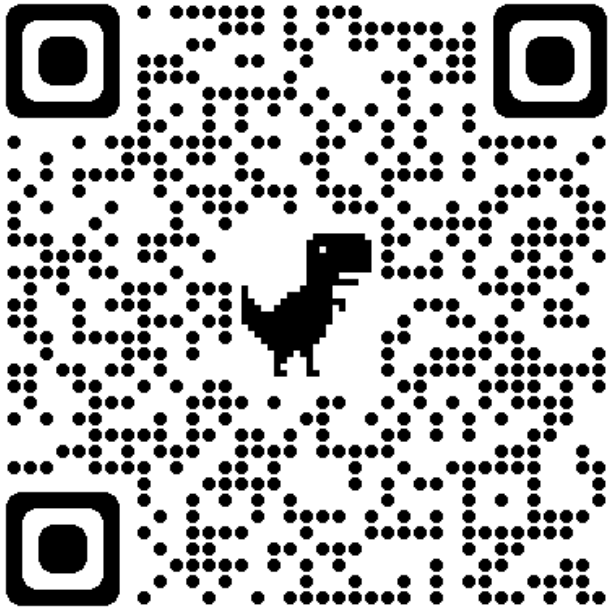
Busca por BLAST



Glycine max transgenic cp4epsps gene for 5-enol-pyruvylshikimate-3-phospate 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 ATTTTAGTTTTATTCTCGAAATTTCATAATAGCTTGGGCTTTCTAAAAAGGGGAATGGTG 120
Sbjct 61 ATTTTAGTTTTATTCTCGAAATTTCATAATAGCTTGGGCTTTCTAAAAAGGGGAATGGTG 120
Query 121 GAATGGGTGTGAGAGTGAAGAGGAATGGTATCGAACCACTAAGAAAAGTAGTCGTGCAAG 180
Sbjct 121 GAATGGGTGTGAGAGTGAAGAGGAATGGTATCGAACCACTAAGAAAAGTAGTCGTGCAAG 180
Query 181 TATTAGATGGTTGGCTGTGATAGTTGGAAAAGGAATAGTAGAAATGGGACAGAAGTTTCA 240
Sbjct 181 TATTAGATGGTTGGCTGTGATAGTTGGAAAAGGAATAGTAGAAATGGGACAGAAGTTTCA 240
Query 241 TTCTGTAAGCTTTTTCATGGACTGTTAGTCTTCTCTTTGCTTTCAGCTTAAGCAGCTTTA 300
Sbjct 241 TTCTGTAAGCTTTTTCATGGACTGTTAGTCTTCTCTTTGCTTTCAGCTTAAGCAGCTTTA 300
Barra = Identidade

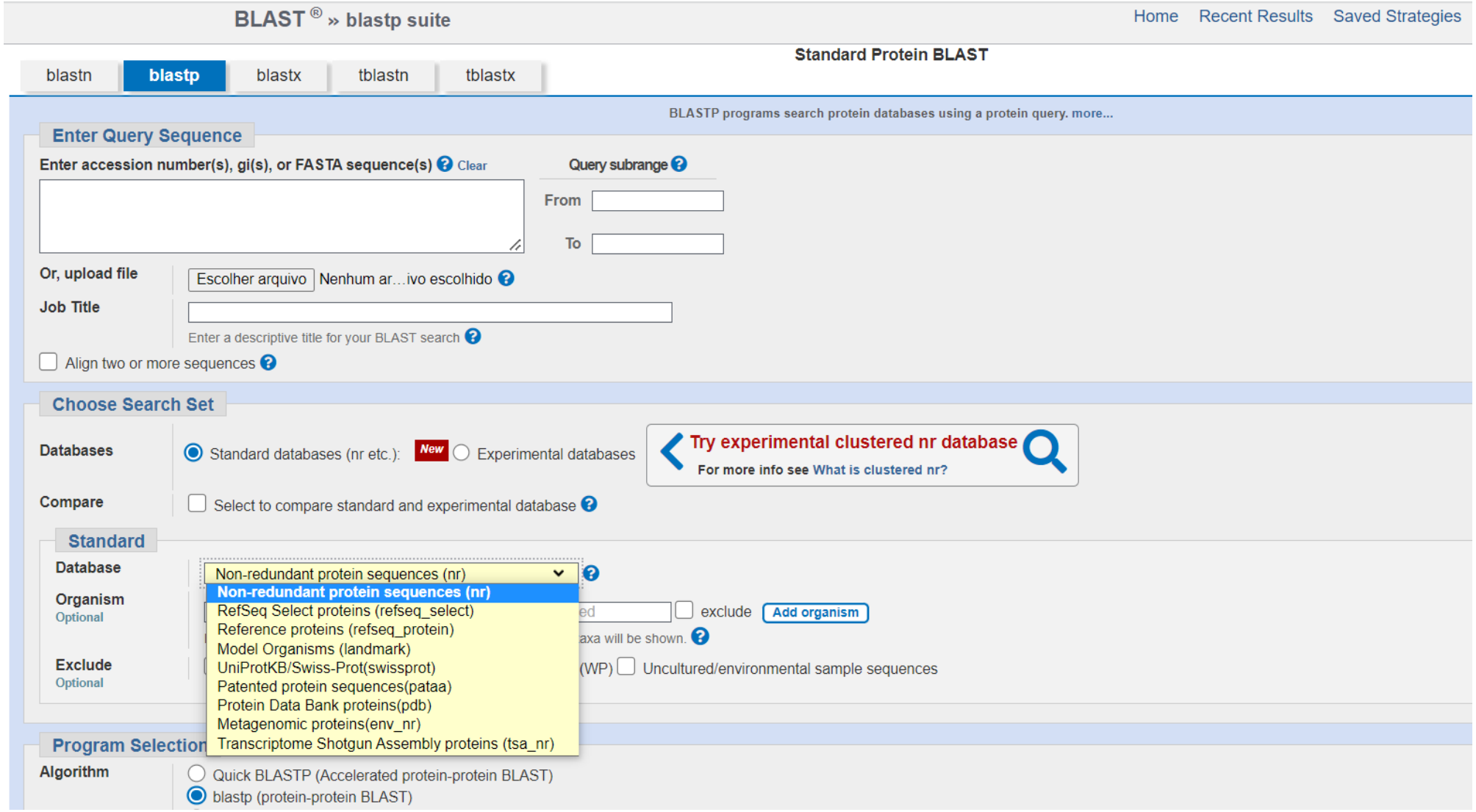
> gi|47933335|gb|AY262821.1| Pinus radiata cellulose synthase (CesA2) mRNA, partial cds

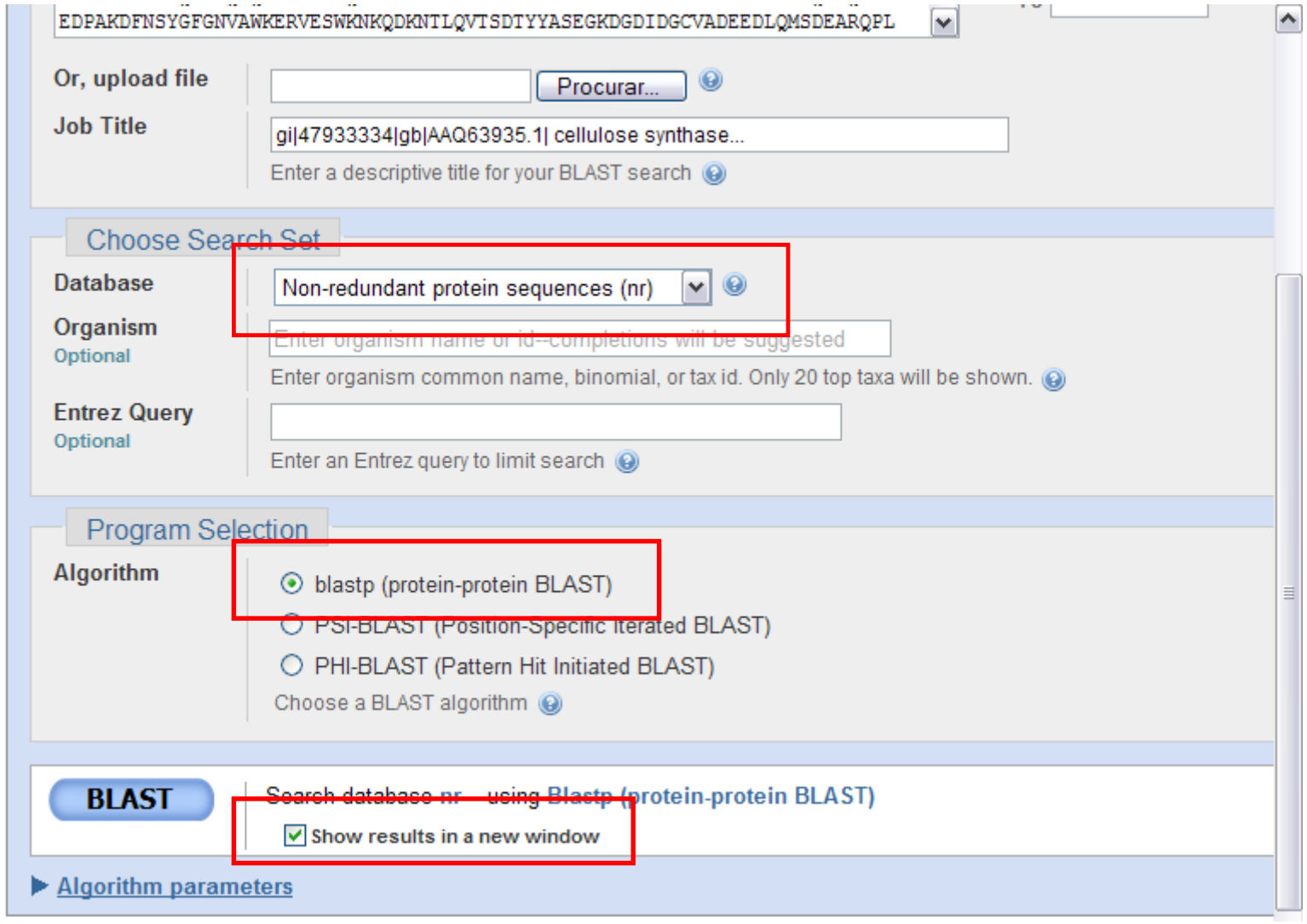
Length=3603

Score = 866 bits (437), Expect = 0.0
Identities = 977/1157 (84%), Gaps = 0/1157 (0%)
Strand=Plus/Plus

```
Query  1450  GAAGACCTTCAAATGAGTGATGAAGCACGCCAGCCCCTATCAAGAAAGGTTCCTATTGCT  1509
             |||||||| ||||||| ||||||||| || || || || || ||||||||  | ||| ||
Sbjct  697   GAAGACCTGCAAATGAATGATGAAGCTCGTCAACCTCTCTCTAGAAAGGTGTCCATTCCT  756

Query  1510  TCTTCCAAGATCAATCCGTACAGGATGGTGATTGTTCTACGGCTTGTGATCTTGTGTTTT  1569
             |||||||||||||||| || || |||||||| ||| | |||||| |  |||||||| ||
Sbjct  757   TCTTCCAAGATCAATCCATATAGAATGGTGATCGTTATCCGGCTTTTTGTCTTGTGTGTT  816

Query  1570  TTCTTCCGCTACCGAATATTGAATCCTGTTCGCAATGCTTATGGACTTTGGTTTACTTCA  1629
             |||||||||||  || ||| || ||||||||  |||||| |||||||| |||||||| ||
Sbjct  817   TTCTTCCGCTATCGGATAATGCATCCTGTTAACAATGCATATGGACTATGGTTTACCTCT  876

Query  1630  GTAATATGTGAGATTTGGTTTGCCATATCCTGGATTCTTGATCAGTTTCCTAAATGGCTT  1689
             || ||||||||| |||||||||||||| || ||||| || ||||||||||| ||||||||
Sbjct  877   GTGATATGTGAGGTTTGGTTTGCCATTTCATGGATCCTGGATCAGTTTCCGAAATGGCTA  936

Query  1690  CCCATCAATAGGGAAACTTATCTTGACAGGCTTTGTTTGAGATACGATAGAGAAGGAGAG  1749
```

Busca por BLASTp




EDPAKDFNSYGFGNVAWKERVESWKNKQDKNTLQVTSDTYYASEGKDGDIDGCVADEEDLQMSDEARQPL
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 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 *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast
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** quer

Specialized BLAST

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

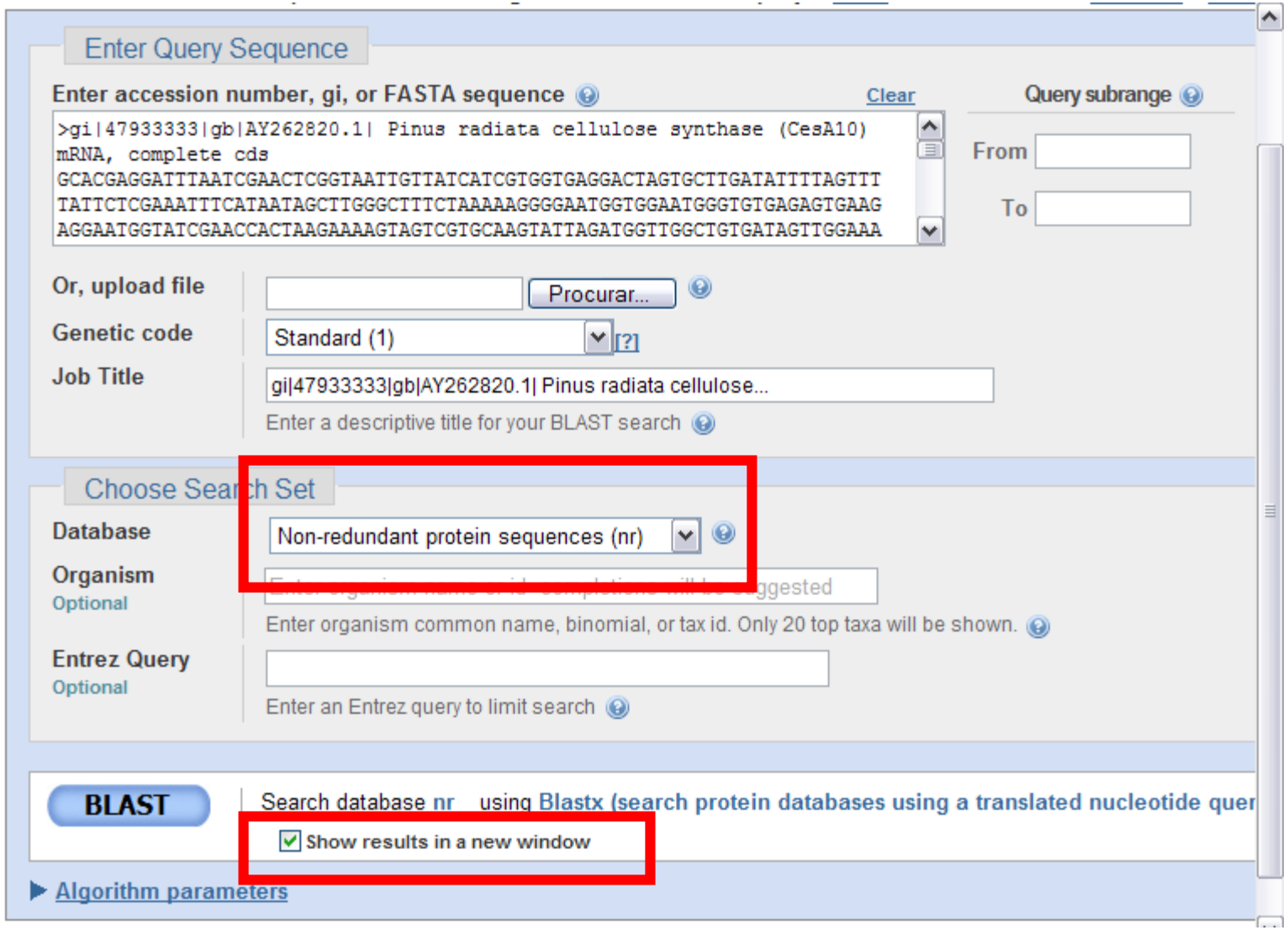
- Search trace archives
- Find conserved domains in your sequence (cds)
- Find sequences with similar conserved domain architecture (cdart)

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...


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
TATTCTCGAAATTTCATAATAGCTTGGGCTTTCTAAAAAGGGGAATGGTGGAATGGGTGTGAGAGTGAAG
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 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 quer
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   CAFPVCRPCYEYEWKDGNQSCPQCKTRYKWHKGSPQVDGDKEDECADDLDHDFNSTQGNR  1008
             CAFPVCRPCYEYEWKDGNQSCPQCKTRYKWHKGSPQVDGDKEDECADDLDHDFNSTQGNR
Sbjct  61    CAFPVCRPCYEYEWKDGNQSCPQCKTRYKWHKGSPQVDGDKEDECADDLDHDFNSTQGNR  120

Query  1009  NEKQQIAEAMLHWQMAYGRGEDVGPSRSESQELPQLQVPLITNGQAISGELPAGSSEYRR  1188
             NEKQQIAEAMLHWQMAYGRGEDVGPSRSESQELPQLQVPLITNGQAISGELPAGSSEYRR
Sbjct  121   NEKQQIAEAMLHWQMAYGRGEDVGPSRSESQELPQLQVPLITNGQAISGELPAGSSEYRR  180

Query  1189  IAAPPTGGGSGKRVHPLPFPDSTQTGQVRAEDPAKDFNSYGFGNVAWKERVESWKNKQDK  1368
             IAAPPTGGGSGKRVHPLPFPDSTQTGQVRAEDPAKDFNSYGFGNVAWKERVESWKNKQDK
Sbjct  181   IAAPPTGGGSGKRVHPLPFPDSTQTGQVRAEDPAKDFNSYGFGNVAWKERVESWKNKQDK  240

Query  1369  NTLQVTSDTYYASEGKDGDIDGCVADEEDLQMSDEARQPLSRKVPIASSKINPYRMVIVL  1548
             NTLQVTSDTYYASEGKDGDIDGCVADEEDLQMSDEARQPLSRKVPIASSKINPYRMVIVL
Sbjct  241   NTLQVTSDTYYASEGKDGDIDGCVADEEDLQMSDEARQPLSRKVPIASSKINPYRMVIVL  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   ICQICGEDVGLAADGEFFVACNECAFPVCRPCYEYEWKDGNQSCPQCKTRYKWHKGSPQV  939
             +CQICG+DVGL ADG+ FVACN CAFPVCRPCY+YE KDGNQSCPQCKTRYK HKGSP+V
Sbjct  3     VCQICGDDVGLTADGDLFVACNVCAFPVCRPCYDYERKDGNQSCPQCKTRYKMHKGSPRV  62

Query  940   DGDKEDECADDLDHDFN-STQGNRNEKQQIAEAMLHWQMAYGRGEDVGPSRSESQELPQL  1116
             +GD+ ++ ADD+ ++++    G+RNEKQ+IAEAML WQM+YGRGEDVG   S  QE+ +
Sbjct  63    EGDEGEDGADDVGNEYHYPPPGSRNEKQKIAEAMLRWQMSYGRGEDVGAPTSTRQEVSES  122

Query  1117  QVPLITNGQAISGELPAGSSEYRRIAAPPTGGGSGKRVHPLPFPDSTQTGQVRAEDPAKD  1296
             Q+P +TNGQ+ISGELPA S E+  + APP+ GG  KRVHPLP+ D+++  QVR  D ++D
Sbjct  123   QIPRLTNGQSISGELPALSPEHS-VGAPPSSGGGSKRVHPLPYTDASRPAQVRIVDHSRD  181

Query  1297  FNSYGFGNVAWKERVESWKNKQDKNTLQVTSDTYYASEGKDGDIDGCVADEEDLQMSDEA  1476
             FNSYGFGNVAWKERVESWKNKQ+KN LQVT+   YASEGK GD+D    + EDLQM+DEA
Sbjct  182   FNSYGFGNVAWKERVESWKNKQEKNMLQVTNSGDYASEGKGGDVDFGGGENEDLQMNDEA  241

Query  1477  RQPLSRKVPIASSKINPYRMVIVLRLVILCFFFRYRILNPVRNAYGLWFTSVICEIWFAI  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.