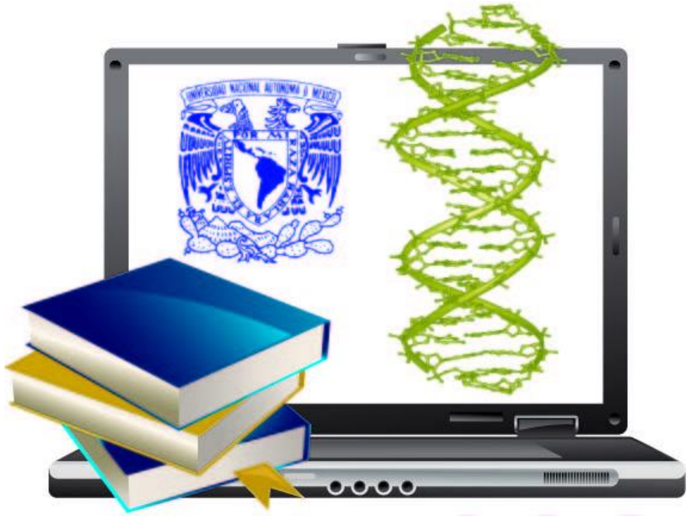
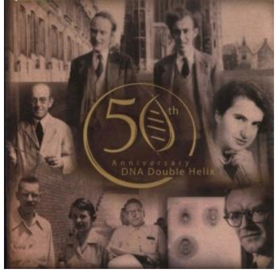
BANCO DE DADOS BIOLÓGICOS

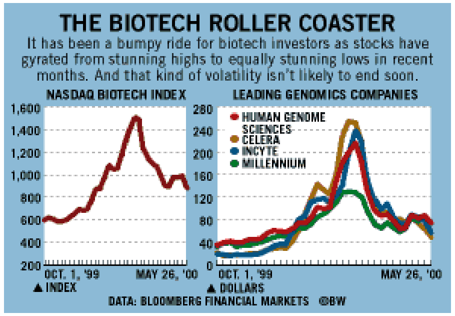
Aula 11

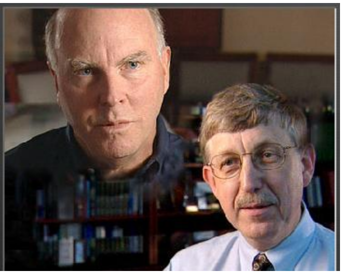
LGN232 – Genética Molecular



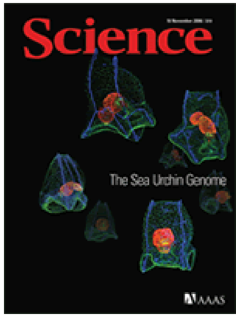
Maria Carolina Quecine
Departamento de Genética
mquecine@usp.br

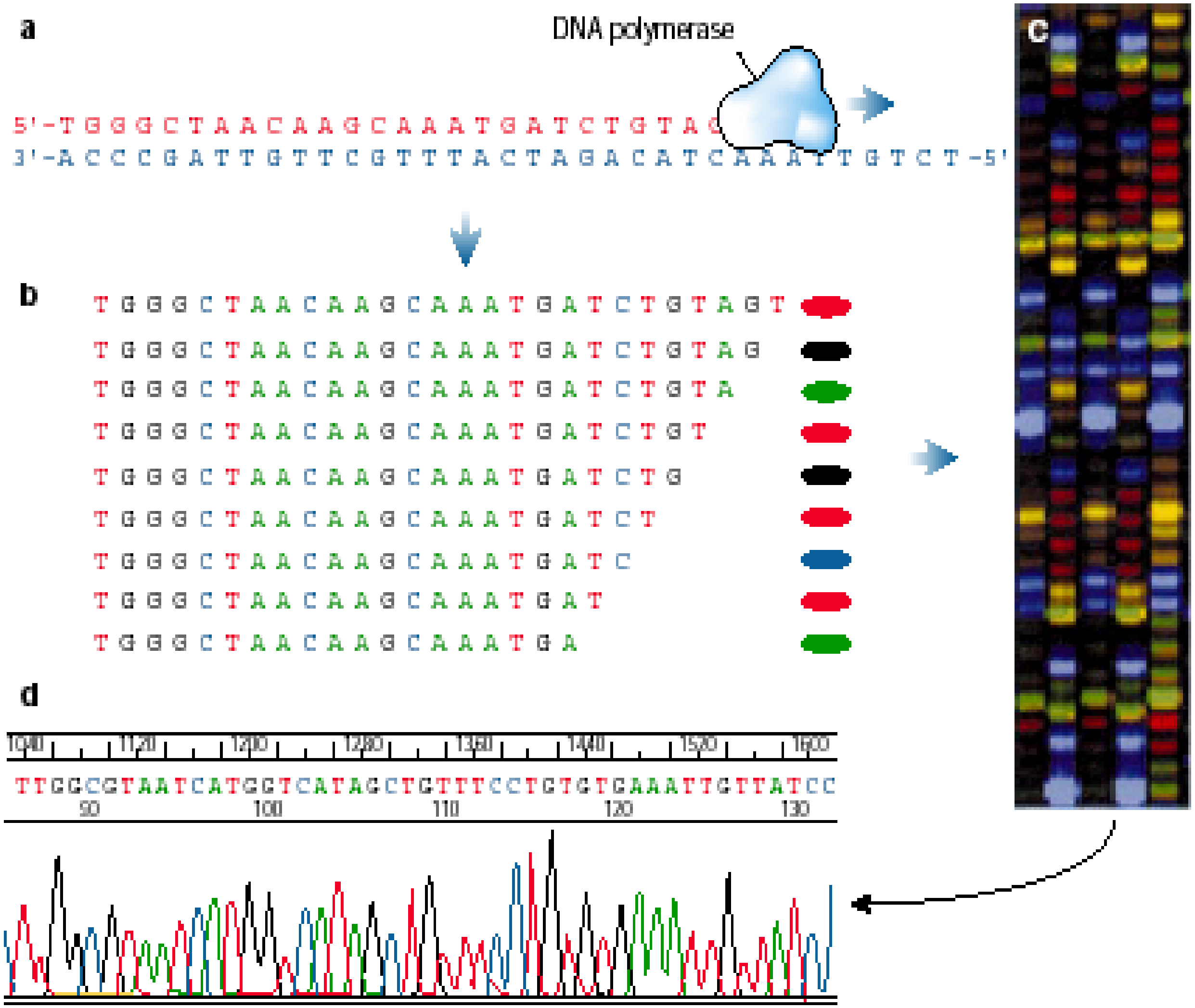






1866	Experimentos de Mendel
1871	Descoberta dos ácidos nucléicos
1930	**Introdução da palavra genoma**
1944	DNA material genético em todos os seres vivos
1951	Primeira proteína (insulina) sequenciada (**Fred Sanger**)
1953	Dupla hélice do DNA (**Watson** & Crick)
1960s	Elucidação do código genético
1977	**Sequenciamento do DNA (Fred Sanger)**
1986	Sequenciamento do DNA automatizado
1995	Primeiro genoma de bactéria sequenciado (*Haemophilus influenza*)
1998	Google
1999	Primeiro cromossomo humano sequenciado (cromossomo 22)
2000	Genomas *Drosophia* / *Arabidopsis* / ***Xylella***
2001	Genomas humano e camundongo
2003	99% genoma humano
2004	Metagenômica (*Environmental genomics*)
2005	Sequenciar um genoma de procarioto por US$1 000.
2007	**DNA de James Watson sequenciado**
2009	Sequenciar seu próprio genoma por US$ 50 000.





MAS PORQUE EU FAÇO SEQUENCIAMENTO?

Arroz

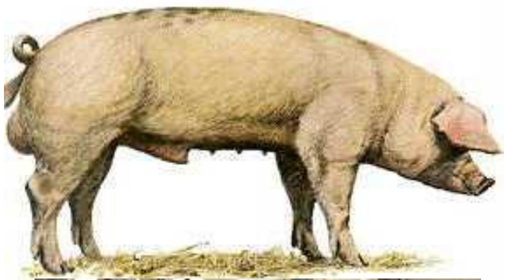
Soja

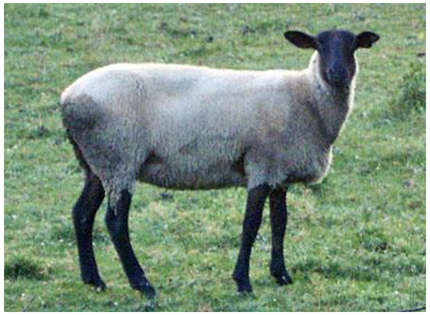
Arabdopsis

Milho

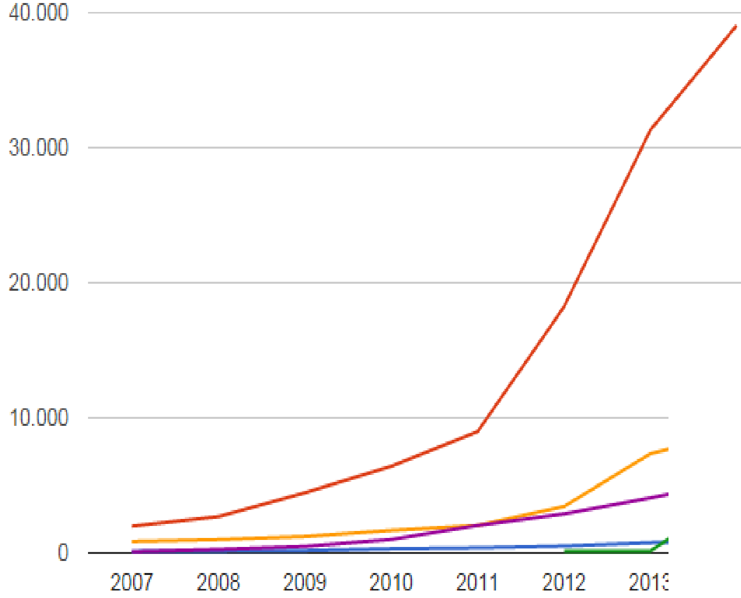
Tomate

Poplar Genome Sequenced and Published;
Model Crop for Biofuels





Total de Projetos no GOLD (domínios da vida)

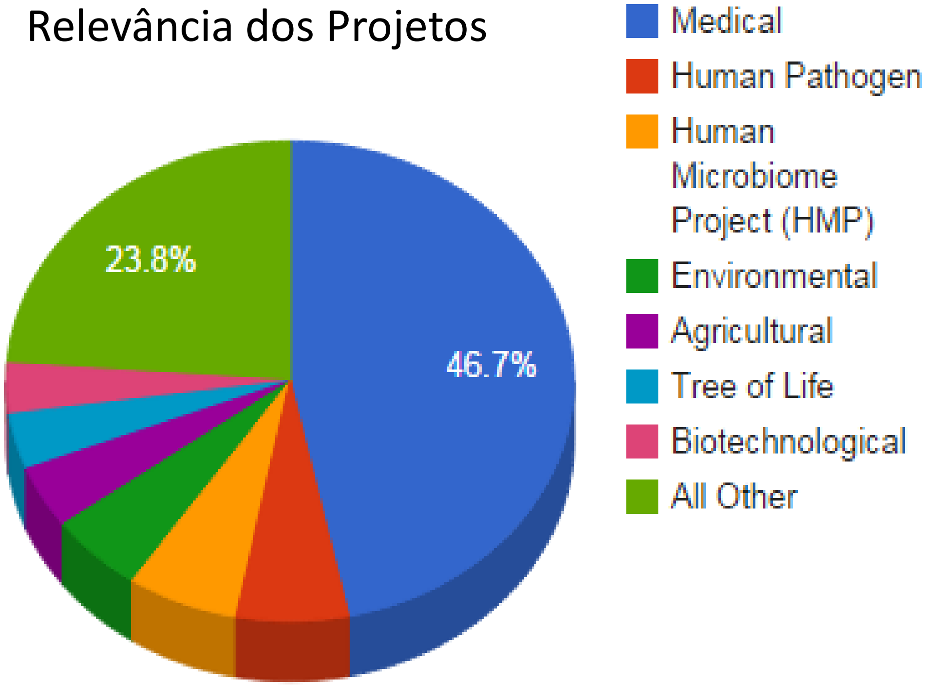


BACTERIA : 38,872

EUKARYA : 9,324

https://gold.jgi-psf.org/statistics

Relevância dos Projetos



MONTAGEM DE GENOMAS

DNA genômico

Fragmentos (enzimas de restrição) que serão **clonados** e **sequenciados**

GCATTTCGAGTTACCTGGACAACCAGTG

CCAGTGGTACTGAGGACGCAAGAGGCTTGA

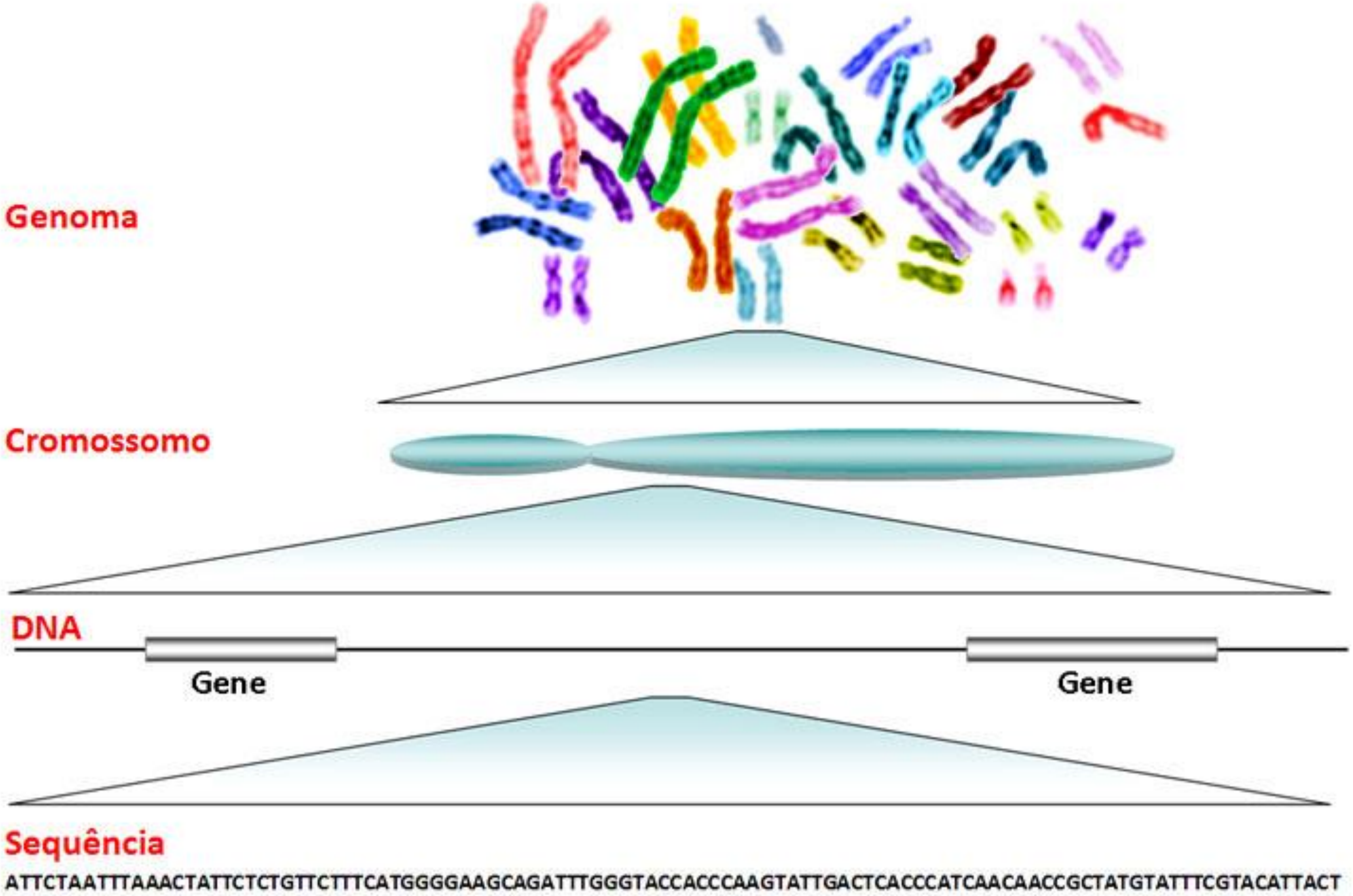
GCTTGATTGGCCAATAATAGTATAT

Seqências Contínuas Alinhadas

GCATTTCGAGTTACCTGGACAACCAGTGGTACTGAGGACGCAAGAGGCTTGATTGGCCAATAATAGTATAT

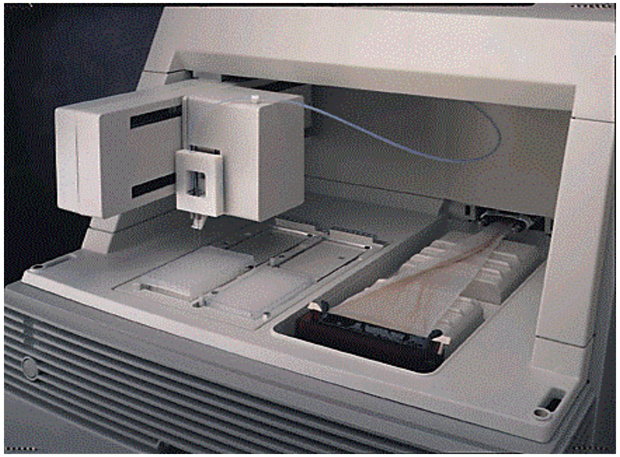
Sequência Final Consenso

GENE X GENOMA



TTCATACTTGGTTAAGACCTTTACAAGCCGACCAACGTGGTGACAGTGTCGTCCTTTA
CGCACCGAATCCCTTTATCATTGAATTAGTAGAAGAGCGATACTTAGGACGTCTTCGG
ATGGAATCTTGGTCCCGTTGCCTGGAACGTCTTGAAACTGAATTCCCGCCAGAAGATG
TTCATACTTGGTTAAGACCTTTACAAGCCGACCAACGTGGTGACAGTGTCGTCCTTTA
CGCACCGAATCCCTTTATCATATTGAATTAGTAGAAGAGCGATACTTAGGACGTCTTC
GGGAATTGTTATCCTATTTCTCAGGAATACGTGAAGTAGTCCTTGCAATTGGCTCACG
ACCTAAAACAACAGAACTACCCGTACCAGTAGACACTACAGGACGTTTGTCTTCAACA
GTCCCATTTAACGGAAATCTCGACACACACTATAACTTTGATAATTTTGTTGAGGGAC
GAAGCAATCAACTCGCTCGTGCTGCAGCTTGGCAAGCGGCACAGAAACCGGGAGACCG
TACTCACAACCCTCTATTGCTCTATGGTGGGACTGGTTTGGGTAAAACCCATTTAATG
TTTGCTGCAGGTAACGTAATGCGGCAAGTAAACCCAACTTATAAAGTAATGTATCTTC
GTTCGGAACAGTTTTTCAGCGCCATGATAAGAGCGTACAAGATAAAAGTATGGATCAT
AAGGGTAAAACCCATTTAATGTTTGCTGCAGGTAACGTAATGCGGCAAGTAAACCCAA
CTTATAAAGTAATGTATCTTCGTTCGGAACAGTTTTTCAGCGCCATGATAAGAGCGTA
CAAGATAAAAGTATGGATCATAAGGGTAAAACCCATTTAATGTTTGCTGCAGGTAACG
TAATGCGGCAAGTAAACCCAACTTATAAAGTAATGTATCTTCGTTCGGAACAGTTTTT
CAGCGCCATGATAAGAGCGTACAAGATAAAAGTATGGATCATAAAACGTAATGCGGCA
AGTAAACCCAACTTATAAAGTAATGTATCTTCGTTCGGAACAGGGTAAAACCCATTTA
ATGTTTGCTGCAGGTAACGTAATGCGGCAAGTAAACCCAACTTATAAAGTAATGTATC
TTCGTTCGGAACAGTTTTTCAGCGCCATGATAAGAGCGTACAAGATAAAAGTATGGAT
CATAAAACGTAATGCGGCAAGTAAACCCAACTTATAAAGTAATGTATCTTCGTTCGGA
ACAAAAACGTAATGCGGCAAGTAAACCCAACTTATAAAGTAATGTATCTTCGTTCGGA
AG

A evolução da técnica...



ABI 3700 Applied Biosystems - ~150.000 bases/dia

3730 – Applied Biosystems – ~1.000.000 bases/dia

454 – Roche – ~100 M bases/7,5 h

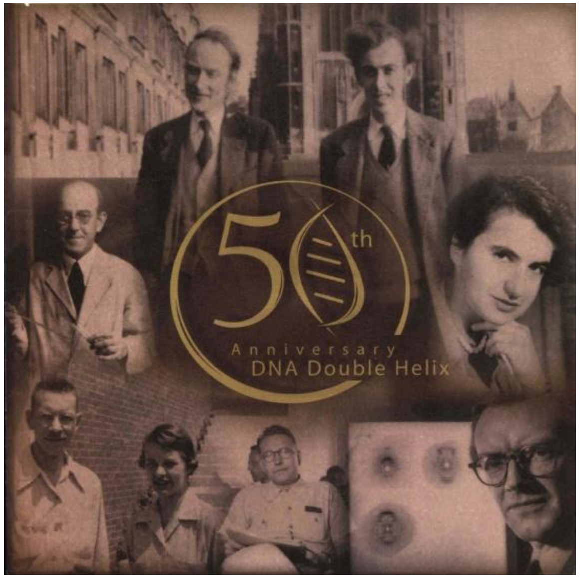
SOLiD – Applied Biosystems –
~3 000 M bases/corrida/ 6 dias (3 bilhões)

Solexa – Illumina – ~2 000 M bases/corrida /6,5 dia (2 bilhões)

TSMS – Helicos – ~1 000 M bases/hora (1 bilhão = 1Gb)
***T**rue **S**ingle **M**olecule **S**equencing*

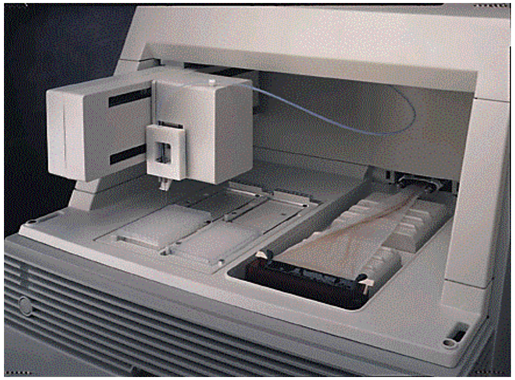
QUER SEQUENCIAR SEU GENOMA?

O que fazer com a informação:

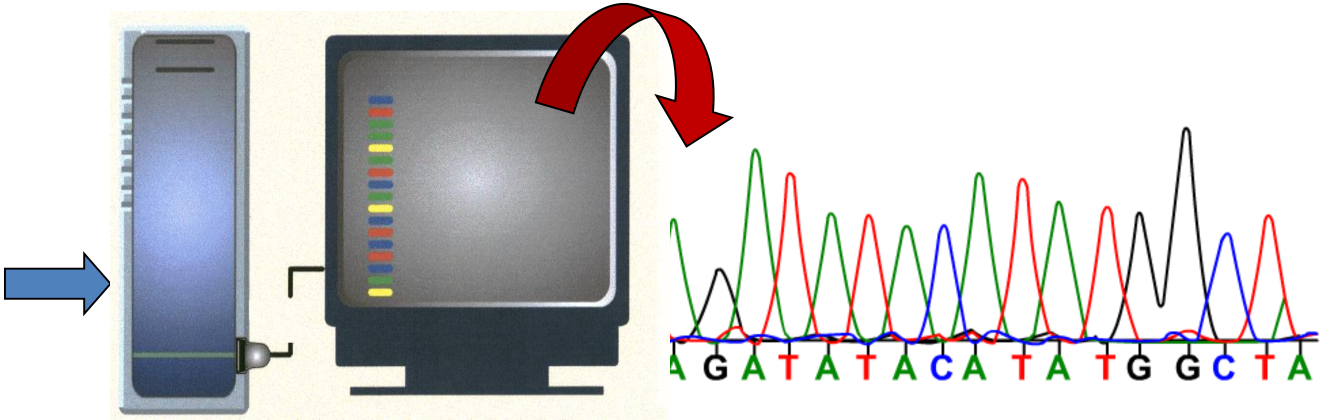


1953-2003

- Produção de Informações
- Sequenciamento de DNA



ABI 3700 Applied Biosystems



BIOINFORMÁTICA

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

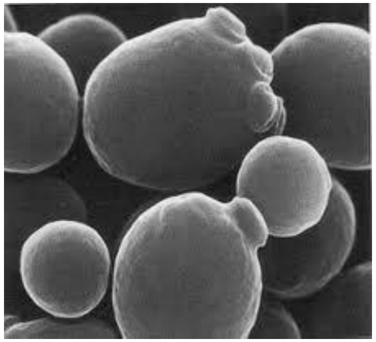
Organização e Armazenamento → **- Bancos de Dados Biológicos**

Visualização e Análise → **- Ferramentas computacionais**
- Compreensão do significado biológico

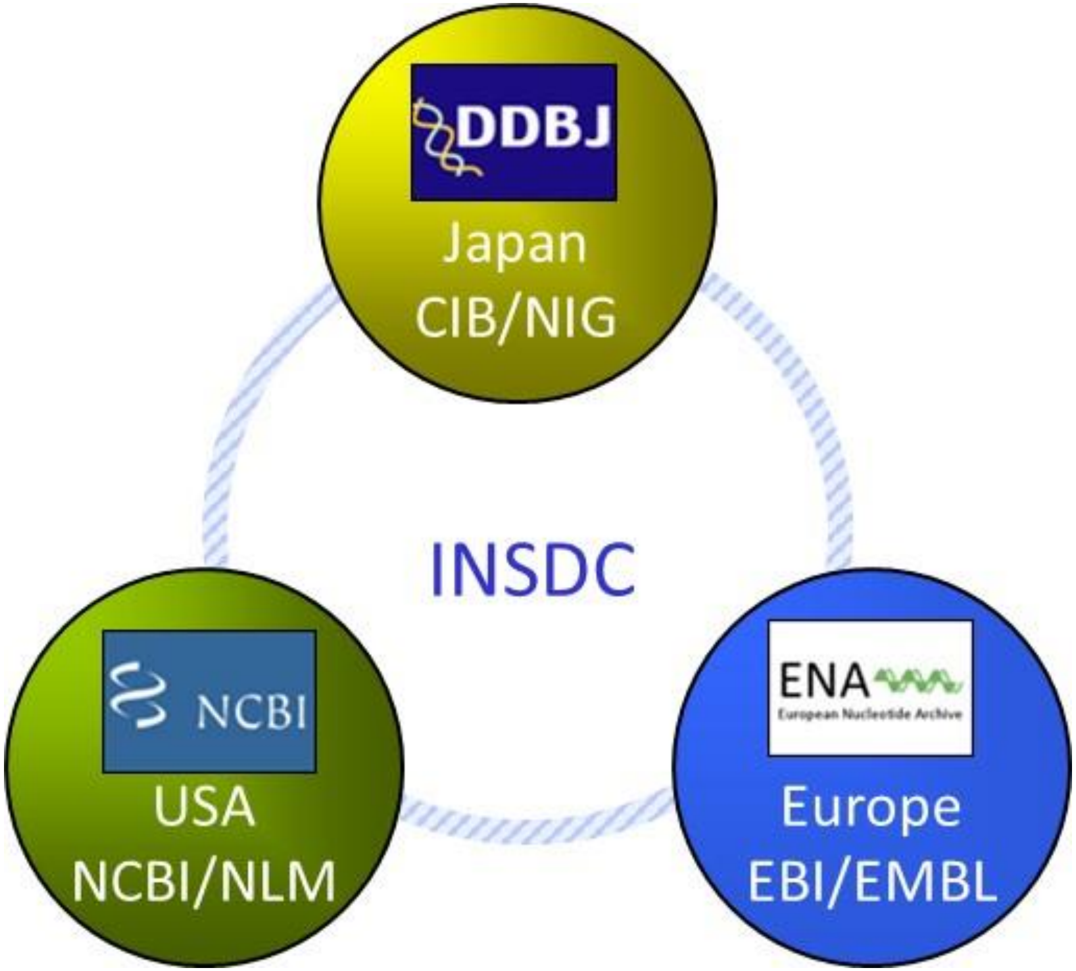
BANCO DE DADOS BIOLÓGICOS

Nucleotídeos :

- International Collaboration
  - NCBI(EUA), EMBL(Europa), DDBJ (Japão)
- Organismo-específico
  - http://flybase.org/
  - http://poultry.mph.msu.edu/
  - http://www.maizegdb.org/
  - http://rice.plantbiology.msu.edu/
  - http://www.yeastgenome.org/
  - http://soybeangenome.siu.edu/
  - http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml



BANCOS DE DADOS MUNDIAIS



DDBJ
Japan
CIB/NIG

INSDC

NCBI
USA
NCBI/NLM

ENA
European Nucleotide Archive
Europe
EBI/EMBL

Sequências de > 300,000 espécies
genoma completo de > 1,500 organismos.

BANCOS DE DADOS MUNDIAIS

Genbank http://www.ncbi.nlm.nih.gov/
Banco de dados americano de seqüências de DNA e proteínas.
EBI http://www.ebi.ac.uk/
Banco de dados europeu de seqüências de DNA.
DDBJ http://www.ddbj.nig.ac.jp/
Banco de dados japonês de seqüências de DNA.

Atualizações diárias

Troca de informações

Resources

- NCBI Home
- All Resources (A-Z)
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Small Molecules
- 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.

More about the NCBI | Mission | Organization | Research | RSS

Genome

1000 prokaryotic genom
now comp
the Genom

How To...

- Determine
- Find a hom
- Obtain the f
- Design PCR

See all ...

Popular Resources

- BLAST
- Bookshelf
- Gene
- Genome
- Nucleotide
- OMIM
- Protein
- PubChem
- PubMed

EMBL

Databases

Data

- ENA
- UniPro
- ArrayE
- Ensem
- InterPr
- PDBe

European Bioinform

About the EBI

- Research
- PhD Studies *(accepting applications)*
- Training
- Industry Support
- Group & Team Leaders
- EBI Funders
- User Support
- EBI Mission
- People
- Events at the EBI
- Genome Campus Events
- How to Find us
- Jobs

Events

- Conference: Bridging the
27–29 October 2010 more
Registration closes 8 Oc
- EBI Open Day
02 Nov 2010 more
Registration closes 4 Oc
- EMBO Practical Course 'C
determination and analysis:
6-10 Sep 2010 more
Registration closes 23 Ju
- Hands-on training at EBI -

Notice
Termination of a part of DDBJ services (February 3, 2010)

Database Search

getentry
Data retrieval by accession numbers, etc.

ARSA
All-round Retrieval of Sequence and Annotation

TXSearch
Retrieval of unified taxonomy database

Homology Search
BLAST

DDBJ Vector Screening System

Phylogenetics

ClustalW
Multiple alignment and Tree-making

Submission of Gene Expression Data

Genome Analyses

GIB
Genome information broker

GIB-V
GIB for Viruses

GTPS
Reannotation of bacterial genomes using a new common protocol

GTOP
Genome to protein structure and function

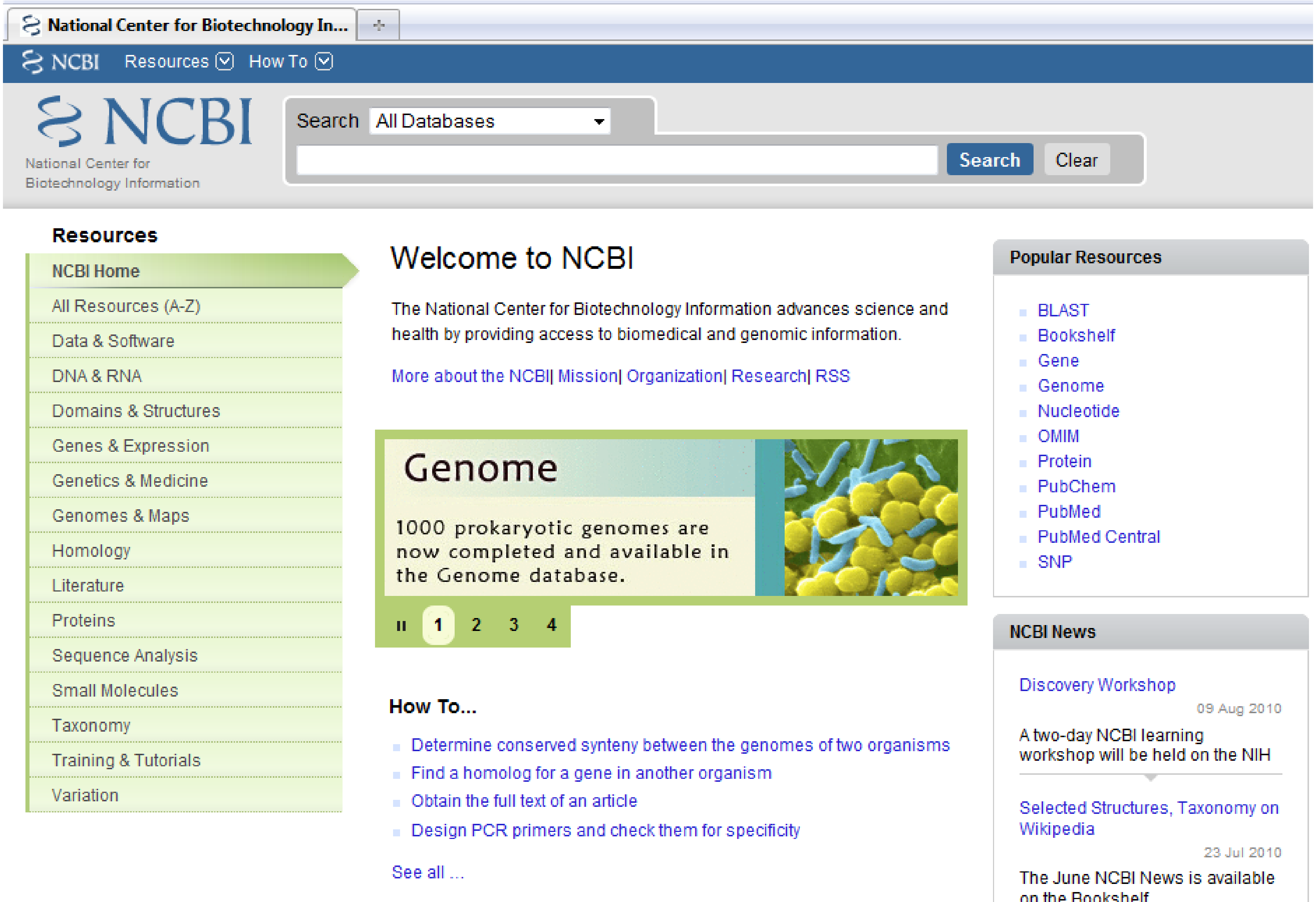
Next Generation Sequence Analysis

DDBJ Read Annotation Pipeline
High-throughput data analysis of next generation sequence data (Login ID is required)

Protein Database and Structure

National Center for Biotechnology Information (NCBI)

http://www.ncbi.nlm.nih.gov/



Welcome to NCBI

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

More about the NCBI| Mission| Organization| Research| RSS

Genome

1000 prokaryotic genomes are now completed and available in the Genome database.

How To...

- Determine conserved synteny between the genomes of two organisms
- Find a homolog for a gene in another organism
- Obtain the full text of an article
- Design PCR primers and check them for specificity

See all ...

NCBI News

Discovery Workshop

09 Aug 2010

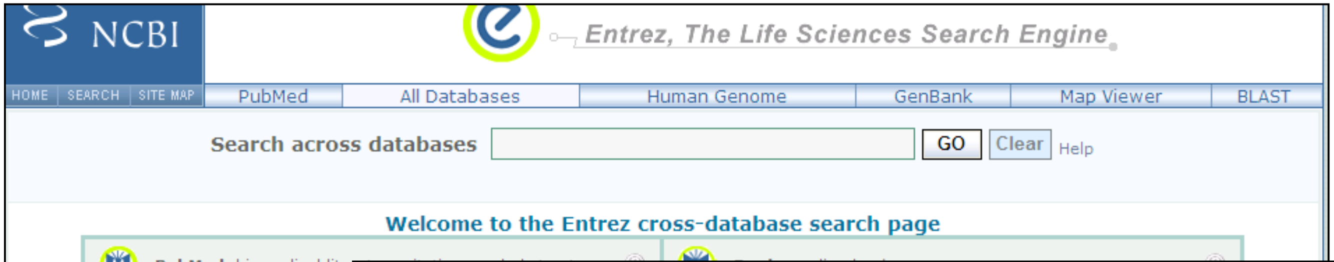
A two-day NCBI learning workshop will be held on the NIH

Selected Structures, Taxonomy on Wikipedia

23 Jul 2010

The June NCBI News is available on the Bookshelf

FERRAMENTAS



Search across databases

Welcome to the Entrez cross-database search page

PubMed: biomedical lite

PubMed Central: free,

Site Search: NCBI web

Nucleotide: Core subse

EST: Expressed Sequen

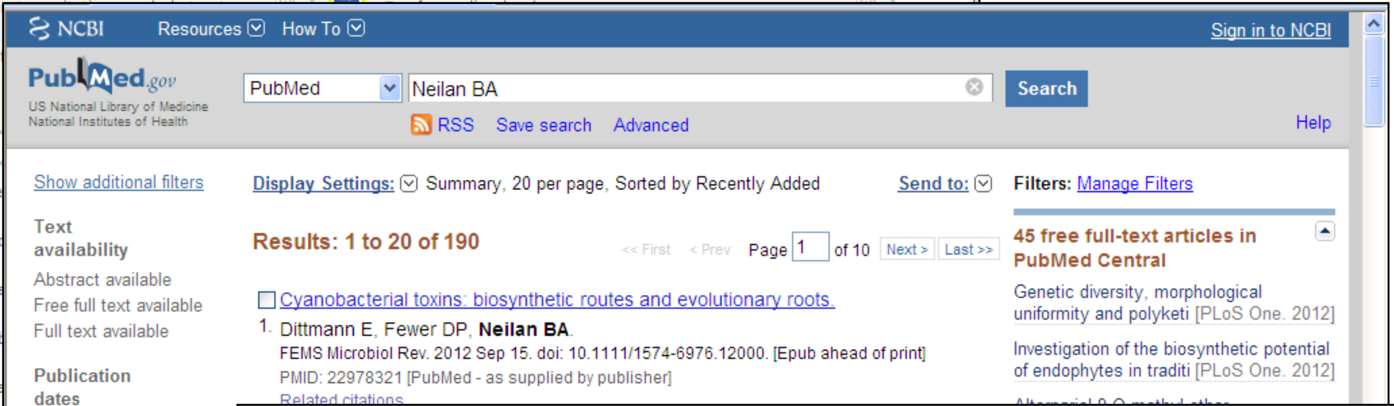
GSS: Genome Survey Se

Protein: sequence datab

Genome: whole genome

Structure: three-dimens

ENTREZ: ferramenta de busca do banco de dados do NCBI



Neilan BA

Display Settings: Summary, 20 per page, Sorted by Recently Added

Results: 1 to 20 of 190

Page 1 of 10

Text availability
Abstract available
Free full text available
Full text available

Publication dates
5 years
10 years
Custom range...

Species
Humans
Other Animals

Article types
Clinical Trial

☐ Cyanobacterial toxins: biosynthetic routes and evolutionary roots.

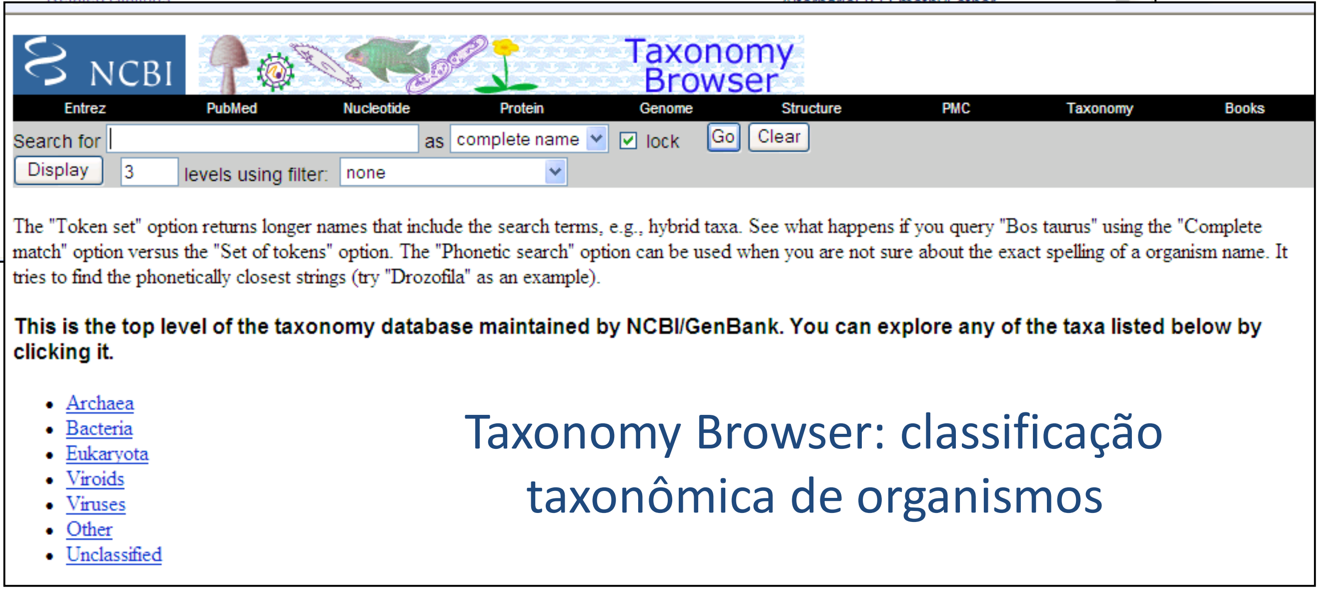
1. Dittmann E, Fewer DP, **Neilan BA**.
FEMS Microbiol Rev. 2012 Sep 15. doi: 10.1111/1574-6976.12000. [Epub ahead of print]
PMID: 22978321 [PubMed - as supplied by publisher]
Related citations

45 free full-text articles in PubMed Central

Genetic diversity, morphological uniformity and polyketi [PLoS One. 2012]

Investigation of the biosynthetic potential of endophytes in traditi [PLoS One. 2012]

PubMed: artigos científicos



Search for ☐ as complete name ☑ lock

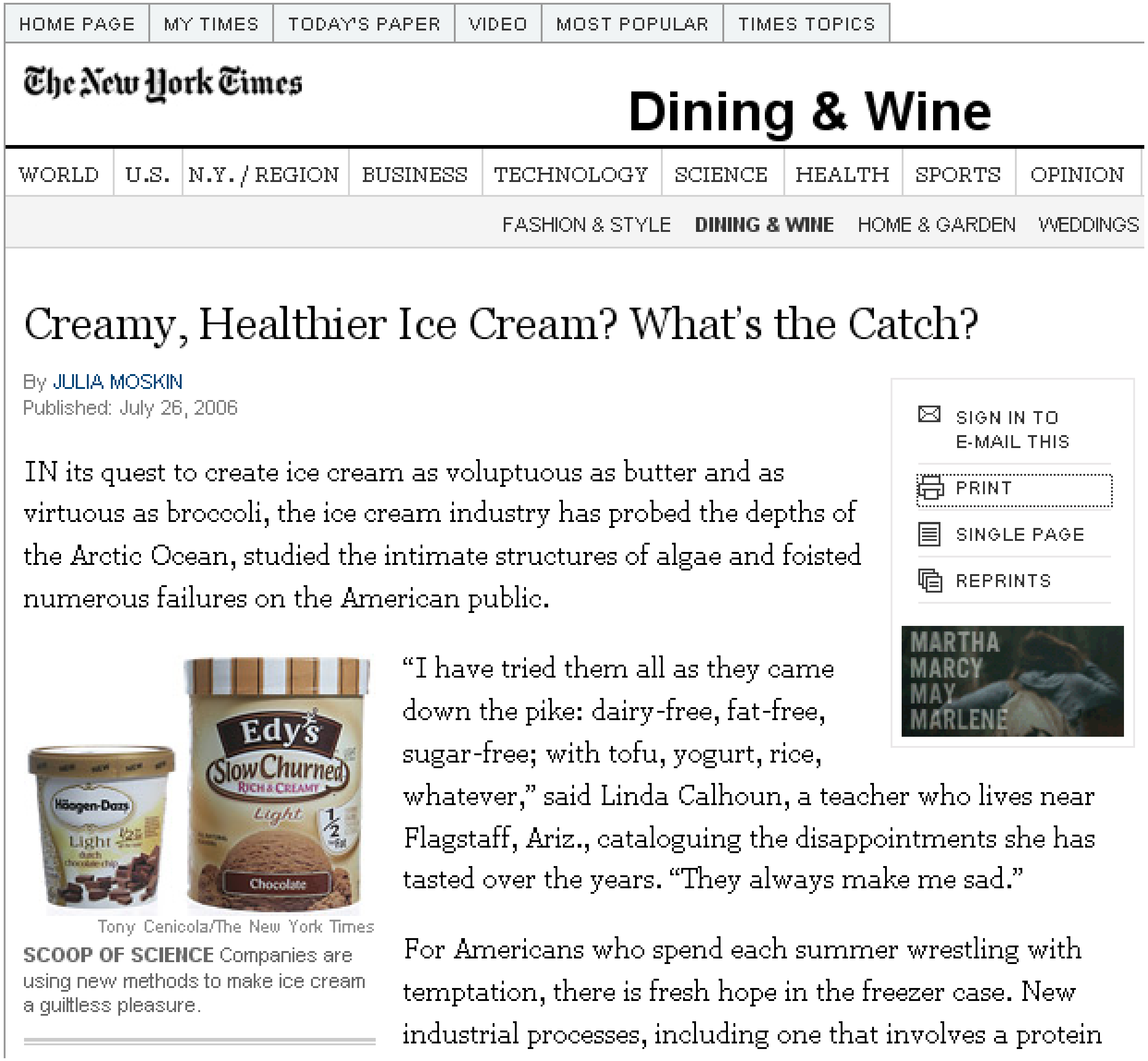
Display 3 levels using filter: none

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
- Viroids
- Viruses
- Other
- Unclassified

Taxonomy Browser: classificação taxonômica de organismos



Creamy, Healthier Ice Cream? What's the Catch?

By JULIA MOSKIN
Published: July 26, 2006

IN its quest to create ice cream as voluptuous as butter and as virtuous as broccoli, the ice cream industry has probed the depths of the Arctic Ocean, studied the intimate structures of algae and foisted numerous failures on the American public.

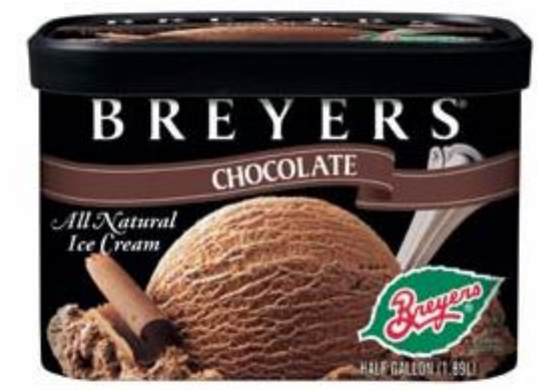
"I have tried them all as they came down the pike: dairy-free, fat-free, sugar-free; with tofu, yogurt, rice, whatever," said Linda Calhoun, a teacher who lives near Flagstaff, Ariz., cataloguing the disappointments she has tasted over the years. "They always make me sad."

Tony Cenicola/The New York Times

SCOOP OF SCIENCE Companies are using new methods to make ice cream a guiltless pleasure.

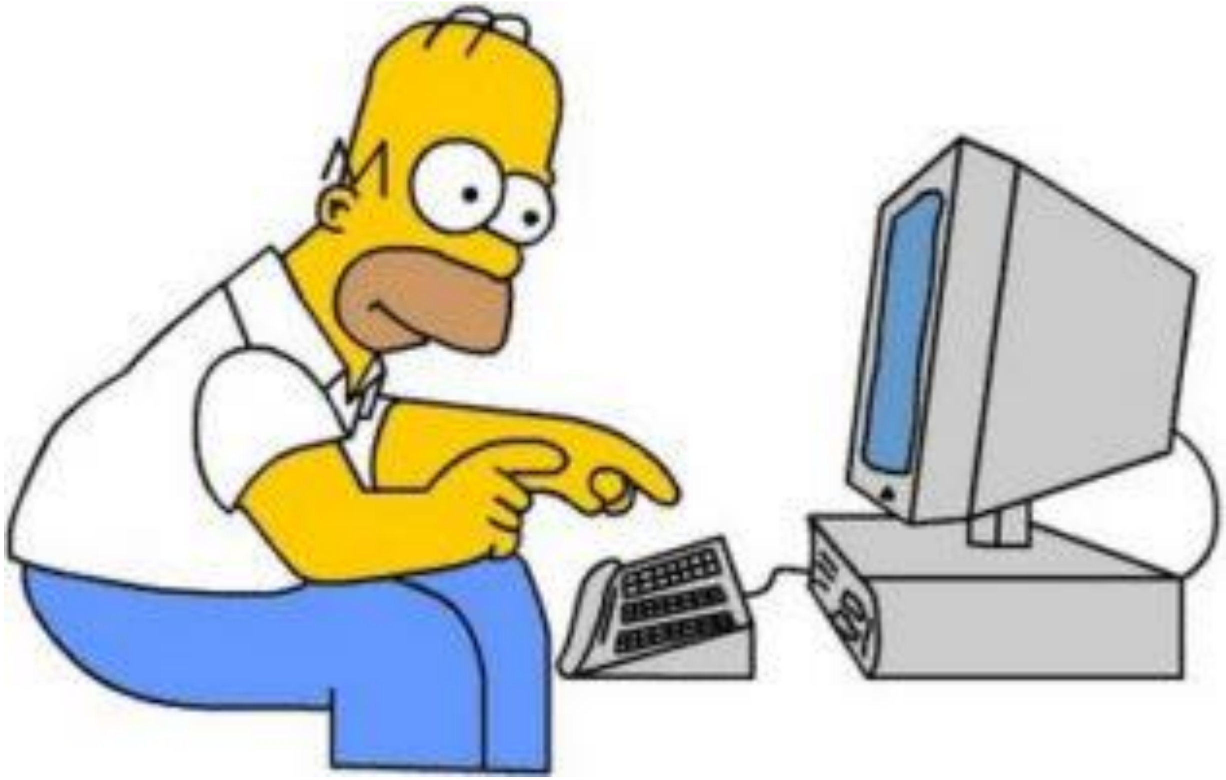
For Americans who spend each summer wrestling with temptation, there is fresh hope in the freezer case. New industrial processes, including one that involves a protein

Ocean pout vive em regiões polares

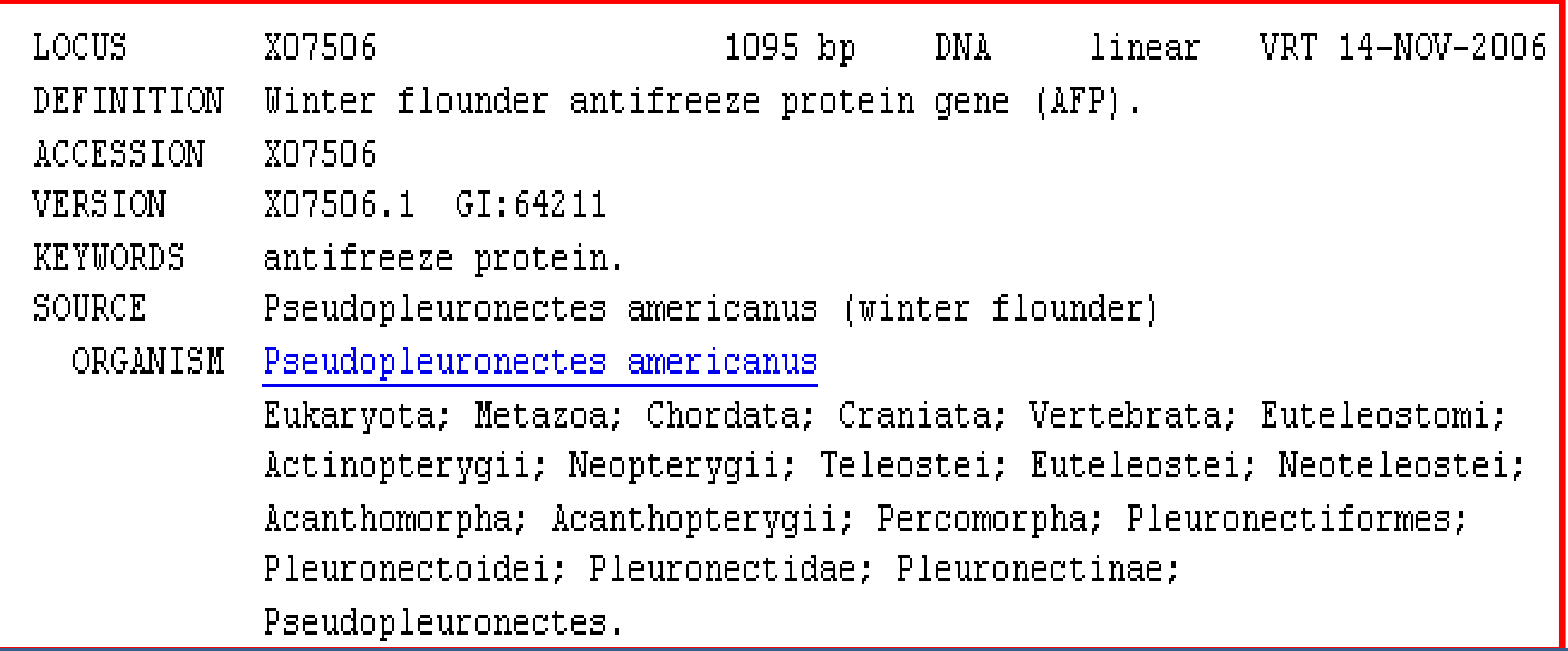


http://www.unilever.com/innovation/productinnovations/coolicecreaminnovations/

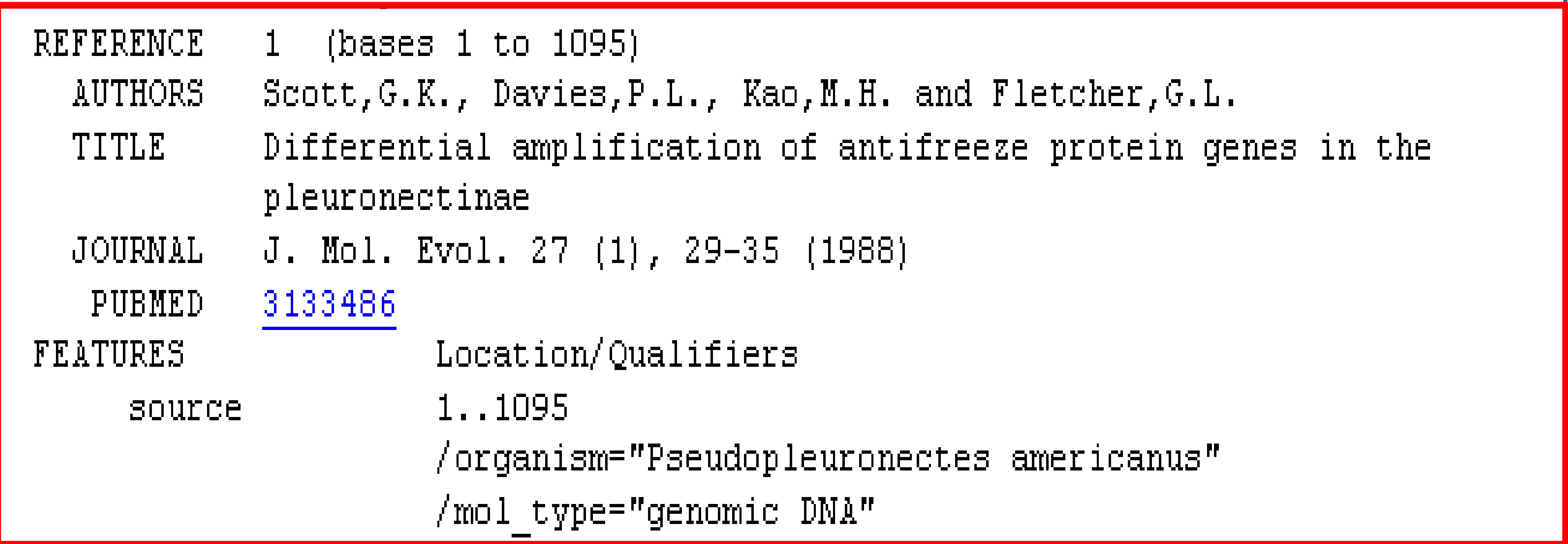
http://academicsreview.org/reviewed-content/genetic-roulette/section-7/part-7-3/



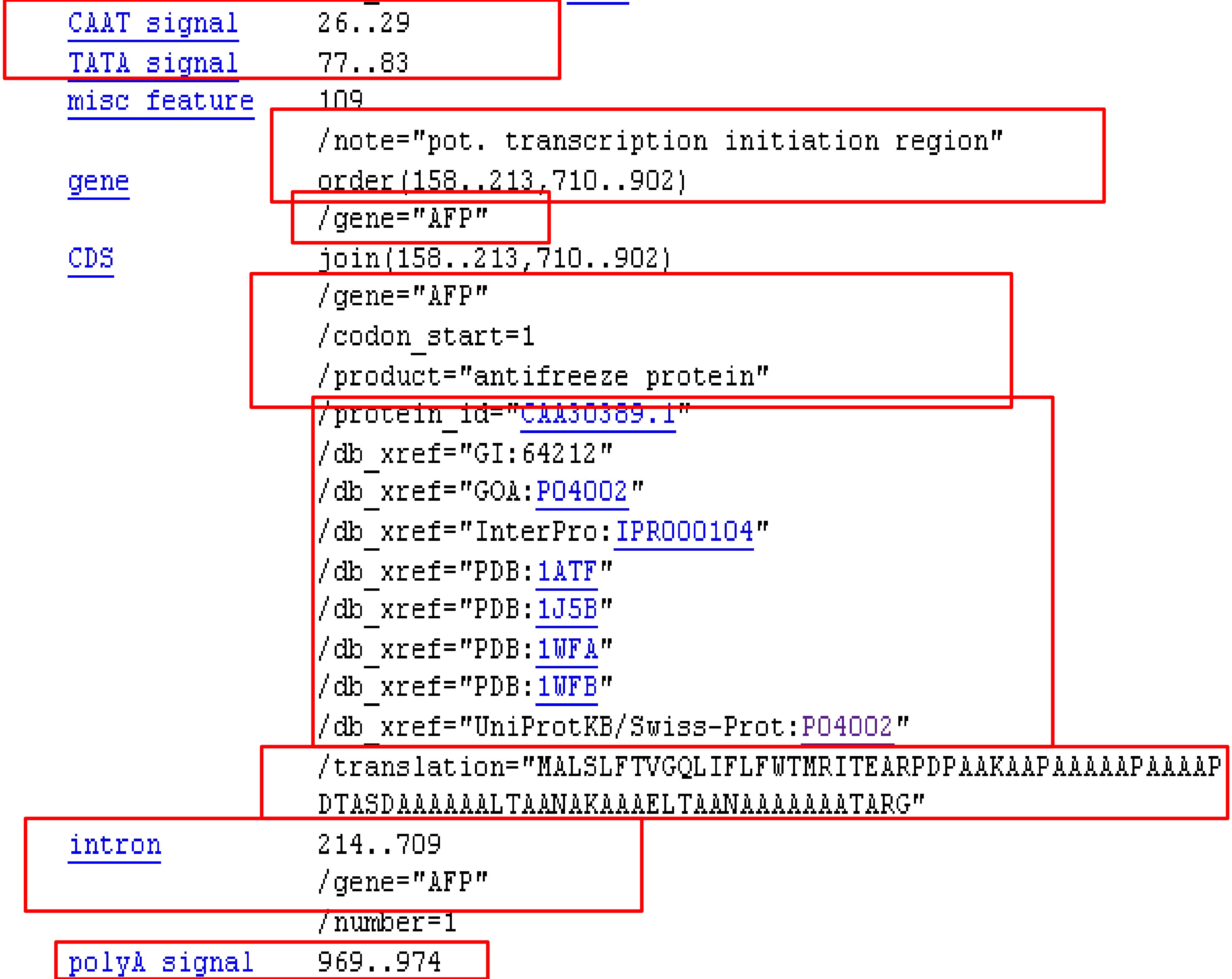
http://www.ncbi.nlm.nih.gov/nuccore/X07506?



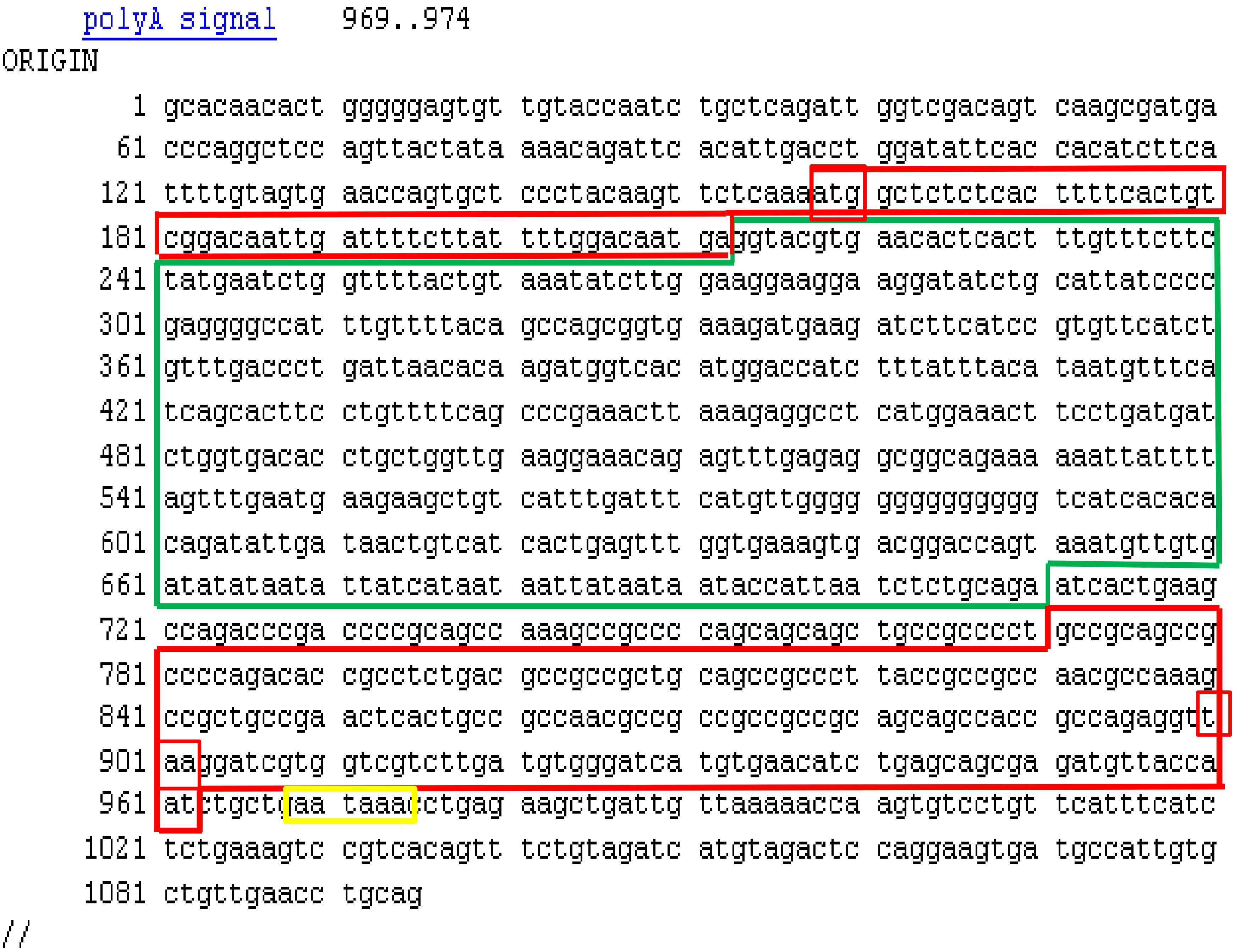
```
LOCUS       X07506                  1095 bp    DNA     linear   VRT 14-NOV-2006
DEFINITION  Winter flounder antifreeze protein gene (AFP).
ACCESSION   X07506
VERSION     X07506.1  GI:64211
KEYWORDS    antifreeze protein.
SOURCE      Pseudopleuronectes americanus (winter flounder)
  ORGANISM  Pseudopleuronectes americanus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Actinopterygii; Neopterygii; Teleostei; Euteleostei; Neoteleostei;
            Acanthomorpha; Acanthopterygii; Percomorpha; Pleuronectiformes;
            Pleuronectoidei; Pleuronectidae; Pleuronectinae;
            Pseudopleuronectes.
```



```
REFERENCE   1  (bases 1 to 1095)
  AUTHORS   Scott,G.K., Davies,P.L., Kao,M.H. and Fletcher,G.L.
  TITLE     Differential amplification of antifreeze protein genes in the
            pleuronectinae
  JOURNAL   J. Mol. Evol. 27 (1), 29-35 (1988)
   PUBMED   3133486
FEATURES             Location/Qualifiers
     source          1..1095
                     /organism="Pseudopleuronectes americanus"
                     /mol_type="genomic DNA"
```



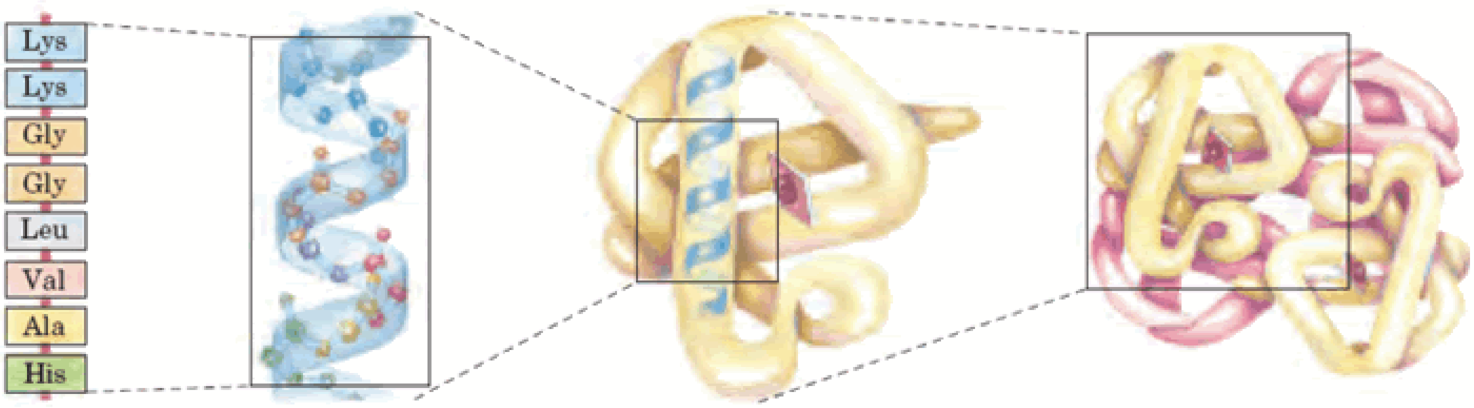
```
CAAT_signal     26..29
TATA_signal     77..83
misc_feature    109
                /note="pot. transcription initiation region"
gene            order(158..213,710..902)
                /gene="AFP"
CDS             join(158..213,710..902)
                /gene="AFP"
                /codon_start=1
                /product="antifreeze protein"
                /protein_id="CAA30389.1"
                /db_xref="GI:64212"
                /db_xref="GOA:P04002"
                /db_xref="InterPro:IPR000104"
                /db_xref="PDB:1ATF"
                /db_xref="PDB:1J5B"
                /db_xref="PDB:1WFA"
                /db_xref="PDB:1WFB"
                /db_xref="UniProtKB/Swiss-Prot:P04002"
                /translation="MALSLFTVGQLIFLFWTMRITEARPDPAAKAAPAAAAAPAAAAP
                DTASDAAAAAALTAANAKAAAELTAANAAAAAAATARG"
intron          214..709
                /gene="AFP"
                /number=1
polyA_signal    969..974
```



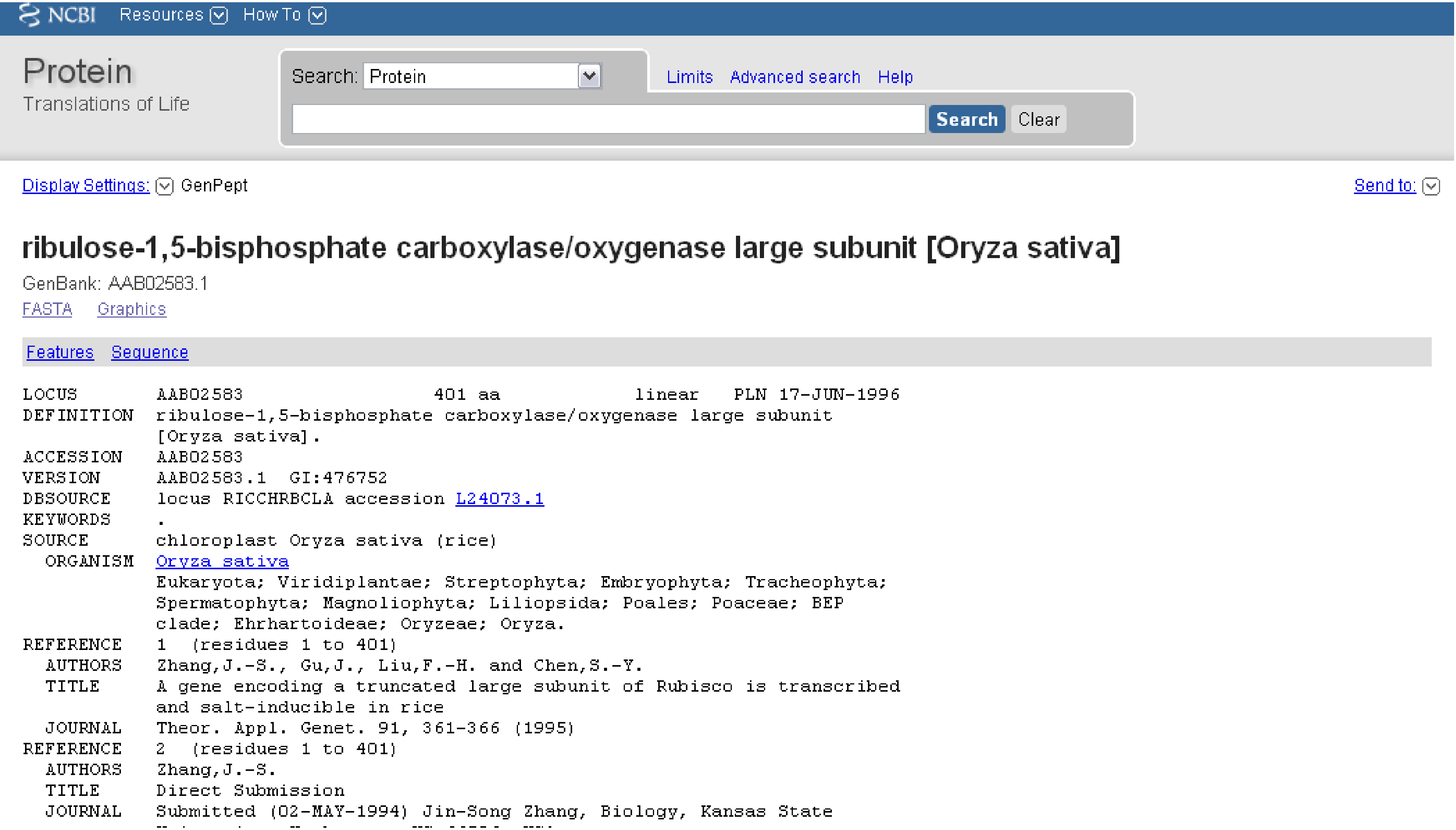
```
     polyA_signal    969..974
ORIGIN
        1 gcacaacact gggggagtgt tgtaccaatc tgctcagatt ggtcgacagt caagcgatga
       61 cccaggctcc agttactata aaacagattc acattgacct ggatattcac cacatcttca
      121 ttttgtagtg aaccagtgct ccctacaagt tctcaaaatg gctctctcac ttttcactgt
      181 cggacaattg attttcttat tttggacaat gaggtacgtg aacactcact ttgtttcttc
      241 tatgaatctg gttttactgt aaatatcttg gaaggaagga aggatatctg cattatcccc
      301 gaggggccat ttgttttaca gccagcggtg aaagatgaag atcttcatcc gtgttcatct
      361 gtttgaccct gattaacaca agatggtcac atggaccatc tttatttaca taatgtttca
      421 tcagcacttc ctgttttcag cccgaaactt aaagaggcct catggaaact tcctgatgat
      481 ctggtgacac ctgctggttg aaggaaacag agtttgagag gcggcagaaa aaattatttt
      541 agtttgaatg aagaagctgt catttgattt catgttgggg gggggggggg tcatcacaca
      601 cagatattga taactgtcat cactgagttt ggtgaaagtg acggaccagt aaatgttgtg
      661 atatataata ttatcataat aattataata ataccattaa tctctgcaga atcactgaag
      721 ccagacccga ccccgcagcc aaagccgccc cagcagcagc tgccgcccct gccgcagccg
      781 ccccagacac cgcctctgac gccgccgctg cagccgccct taccgccgcc aacgccaaag
      841 ccgctgccga actcactgcc gccaacgccg ccgccgccgc agcagccacc gccagaggtt
      901 aaggatcgtg gtcgtcttga tgtgggatca tgtgaacatc tgagcagcga gatgttacca
      961 atctgctgaa taaacctgag aagctgattg ttaaaaacca agtgtcctgt tcatttcatc
     1021 tctgaaagtc cgtcacagtt tctgtagatc atgtagactc caggaagtga tgccattgtg
     1081 ctgttgaacc tgcag
//
```

PROTÉINAS

- NCBI: Protein database (http://www.ncbi.nlm.nih.gov/genbank/)
- UniPRot/Swiss-Prot: (http://www.uniprot.org/)



http://www.ncbi.nlm.nih.gov/protein/AAB02583.1



Display Settings: GenPept

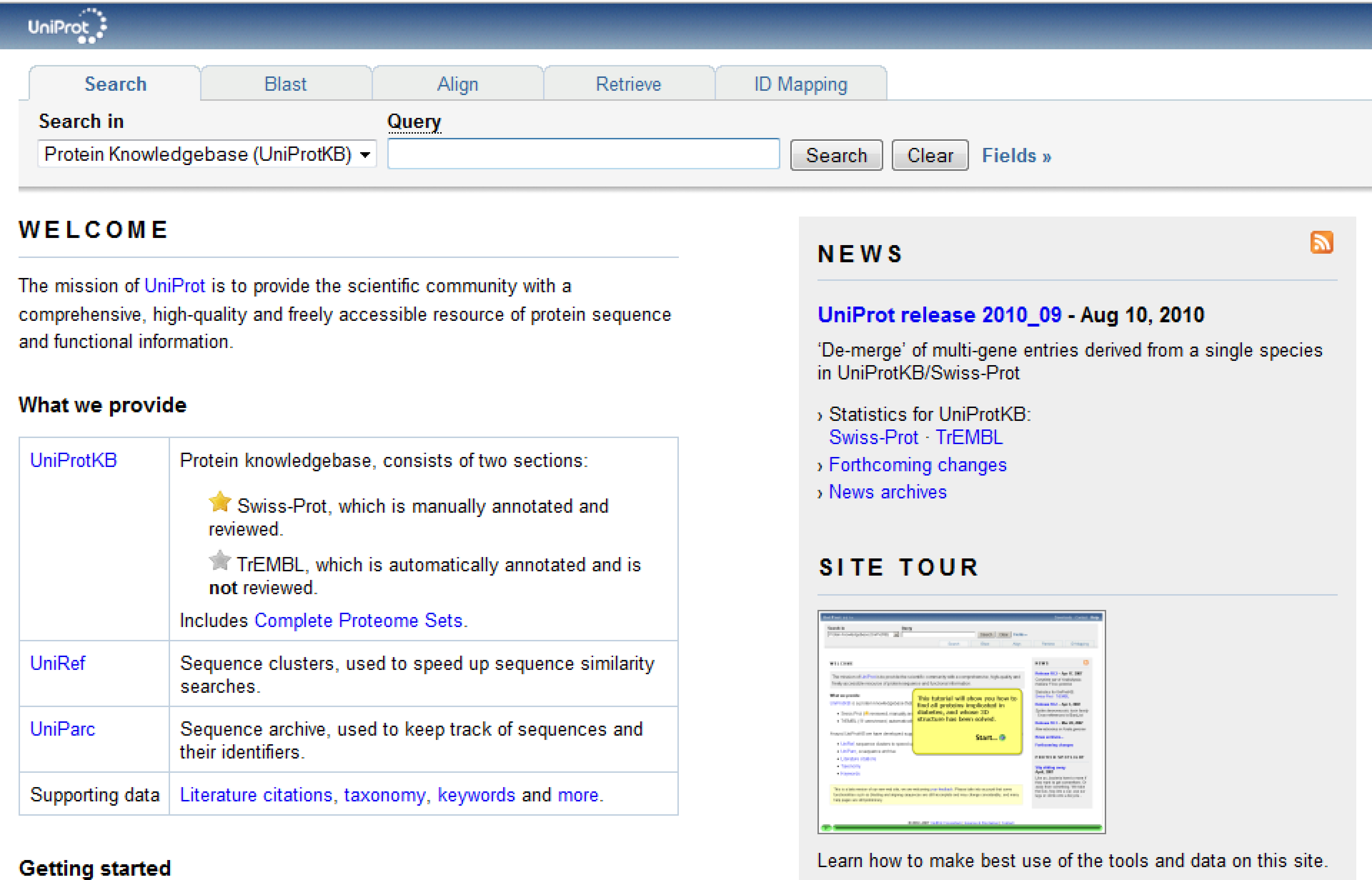
ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit [Oryza sativa]

GenBank: AAB02583.1

```
LOCUS       AAB02583                 401 aa            linear   PLN 17-JUN-1996
DEFINITION  ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit
            [Oryza sativa].
ACCESSION   AAB02583
VERSION     AAB02583.1  GI:476752
DBSOURCE    locus RICCHRBCLA accession L24073.1
KEYWORDS    .
SOURCE      chloroplast Oryza sativa (rice)
  ORGANISM  Oryza sativa
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; Liliopsida; Poales; Poaceae; BEP
            clade; Ehrhartoideae; Oryzeae; Oryza.
REFERENCE   1  (residues 1 to 401)
  AUTHORS   Zhang,J.-S., Gu,J., Liu,F.-H. and Chen,S.-Y.
  TITLE     A gene encoding a truncated large subunit of Rubisco is transcribed
            and salt-inducible in rice
  JOURNAL   Theor. Appl. Genet. 91, 361-366 (1995)
REFERENCE   2  (residues 1 to 401)
  AUTHORS   Zhang,J.-S.
  TITLE     Direct Submission
  JOURNAL   Submitted (02-MAY-1994) Jin-Song Zhang, Biology, Kansas State
```

UniProt Protein knowledgebase

http://www.uniprot.org/



WELCOME

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

What we provide

UniProtKB	Protein knowledgebase, consists of two sections: ★ Swiss-Prot, which is manually annotated and reviewed. ☆ TrEMBL, which is automatically annotated and is **not** reviewed. Includes Complete Proteome Sets.
UniRef	Sequence clusters, used to speed up sequence similarity searches.
UniParc	Sequence archive, used to keep track of sequences and their identifiers.
Supporting data	Literature citations, taxonomy, keywords and more.

Getting started

NEWS

UniProt release 2010_09 - Aug 10, 2010

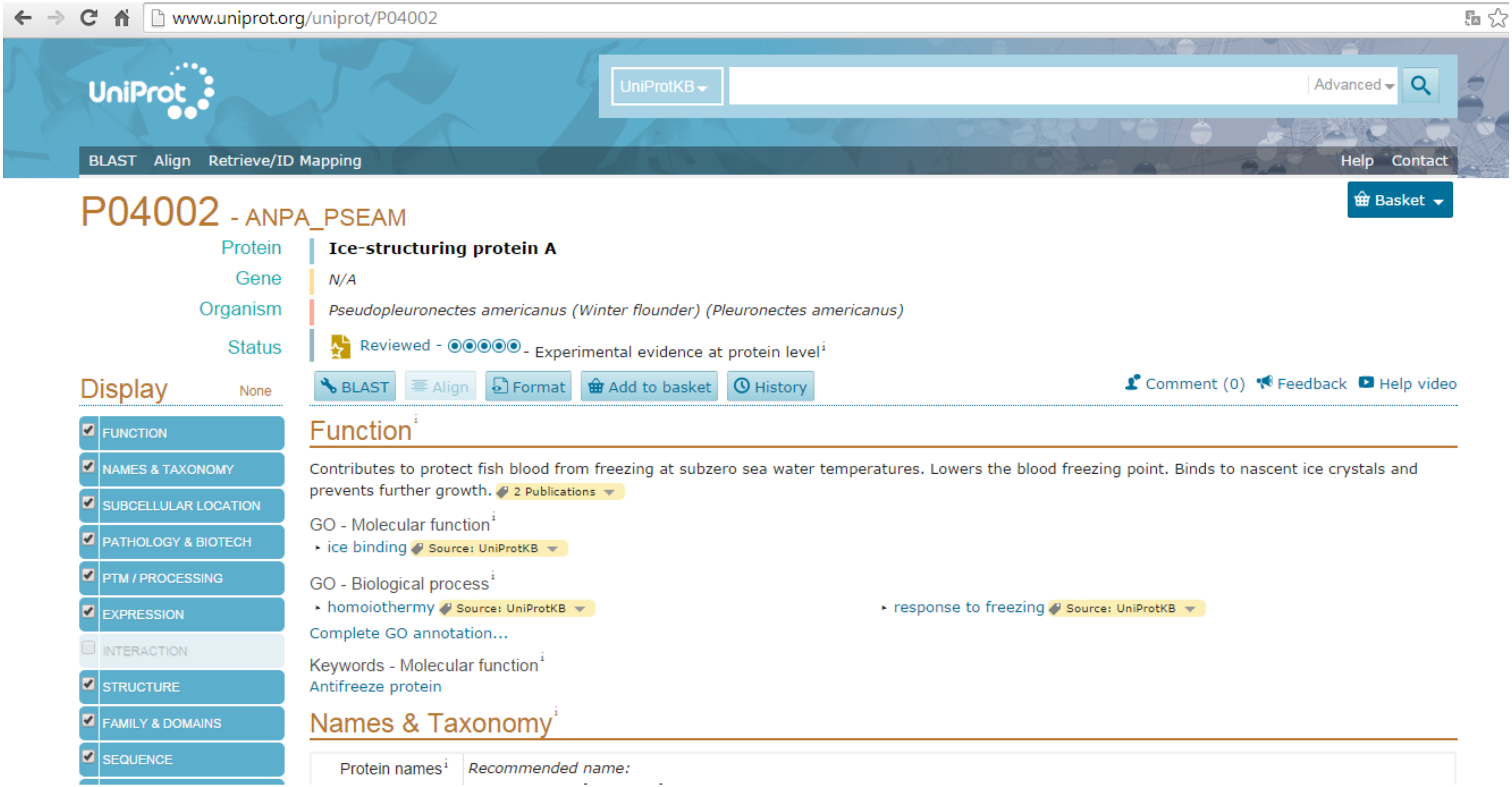
'De-merge' of multi-gene entries derived from a single species in UniProtKB/Swiss-Prot

› Statistics for UniProtKB:
Swiss-Prot · TrEMBL
› Forthcoming changes
› News archives

SITE TOUR

Learn how to make best use of the tools and data on this site.

http://www.uniprot.org/uniprot/P04002



P04002 - ANPA_PSEAM

Protein	**Ice-structuring protein A**
Gene	*N/A*
Organism	*Pseudopleuronectes americanus (Winter flounder) (Pleuronectes americanus)*
Status	Reviewed - ●●●●● - Experimental evidence at protein level

Function

Contributes to protect fish blood from freezing at subzero sea water temperatures. Lowers the blood freezing point. Binds to nascent ice crystals and prevents further growth. 2 Publications

GO - Molecular function

- ice binding Source: UniProtKB

GO - Biological process

- homoiothermy Source: UniProtKB
- response to freezing Source: UniProtKB

Complete GO annotation...

Keywords - Molecular function

Antifreeze protein

Names & Taxonomy

Protein names | *Recommended name:*

BUSCA EM BLAST

BLAST: Basic Local Alignment Search Tool

- Por sequência de nucleotídeos ou de aminoácidos (proteínas);
- Comparação de sequências a fim de identificar similaridade de DNA ou proteína para inferir origem, função, filogênia;
- Realiza comparações entre pares de sequências, buscando regiões com similaridade local;
- Alinhamento local (segmentos) é a base da busca por BLAST;
- Usa algoritmos para gerar alinhamento de sequências;

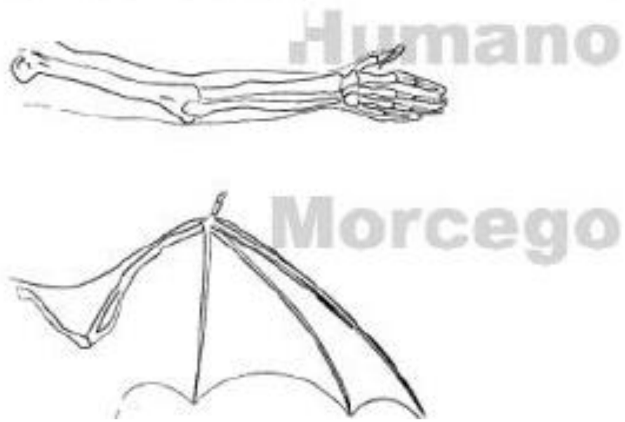
BUSCA EM BLAST

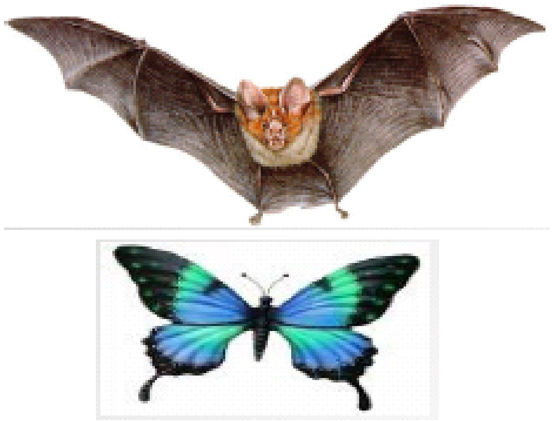
Algoritmos em Blast:

- Não avaliam homologia

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

- Medem similaridade e identidade de seqüências





Exemplos:
Órgãos homólogos – asas de morcego e mãos de humanos (mesma origem)
Órgãos similares – asas de morcego e asas de borboleta (mesma função)

BUSCA EM BLAST

Identidade x Similaridade x Homologia

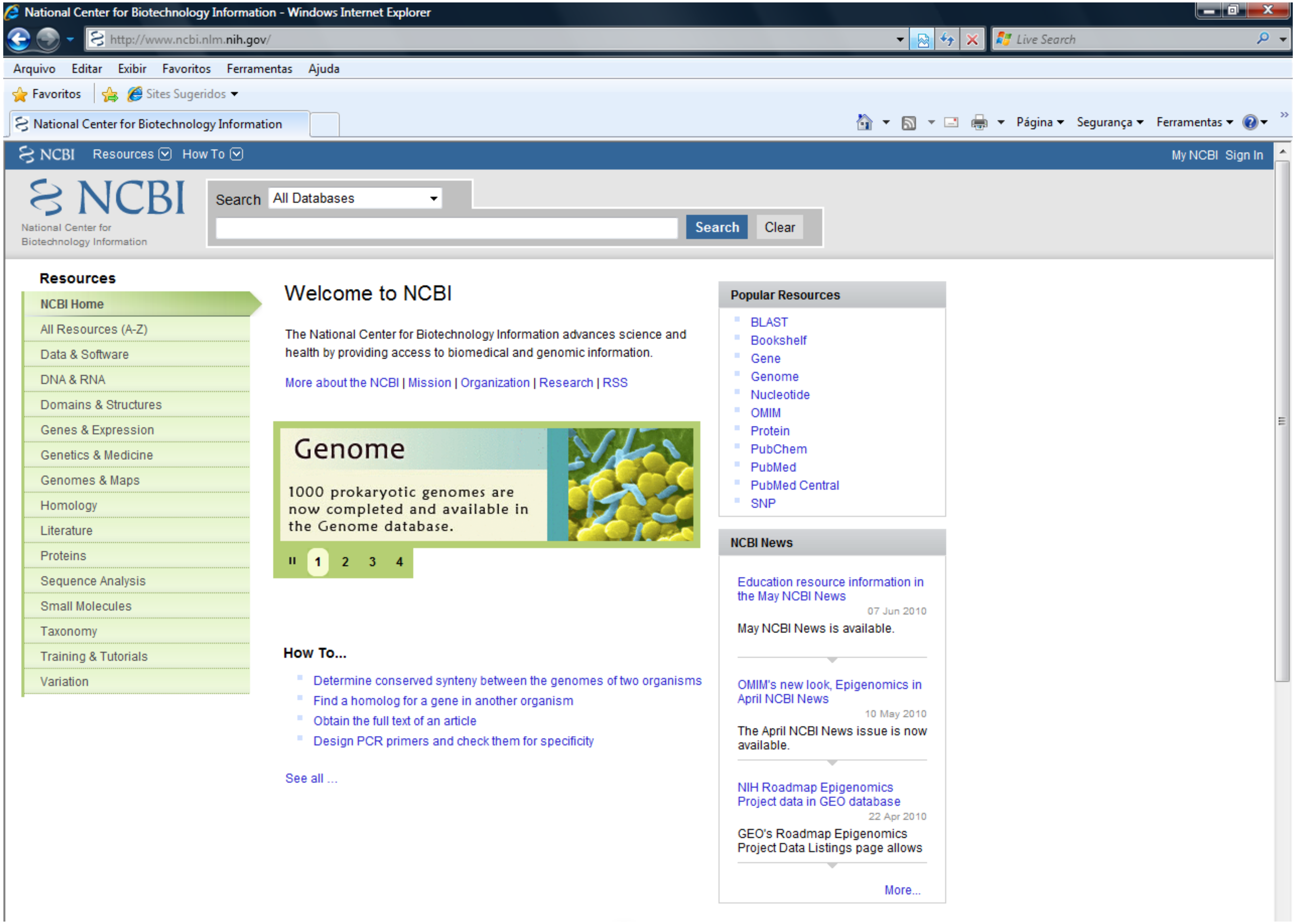
- **Identidade** = ocorrência do mesmo nucleotídeo ou aminoácido na mesma posição nas seqüências alinhadas

- **Similaridade** = considera combinações próximas e avaliada por medidas de diferença/igualdade

- **Homologia** = dividem mesma ancestralidade, com significado evolutivo

Homologia => conceito central de Evolução

BUSCA EM BLAST



NCBI
National Center for Biotechnology Information

Search: All Databases

Resources

- **NCBI Home**
- All Resources (A-Z)
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Small Molecules
- 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.

More about the NCBI | Mission | Organization | Research | RSS

Genome

1000 prokaryotic genomes are now completed and available in the Genome database.

II 1 2 3 4

How To...

- Determine conserved synteny between the genomes of two organisms
- Find a homolog for a gene in another organism
- Obtain the full text of an article
- Design PCR primers and check them for specificity

See all ...

Popular Resources

- BLAST
- Bookshelf
- Gene
- Genome
- Nucleotide
- OMIM
- Protein
- PubChem
- PubMed
- PubMed Central
- SNP

NCBI News

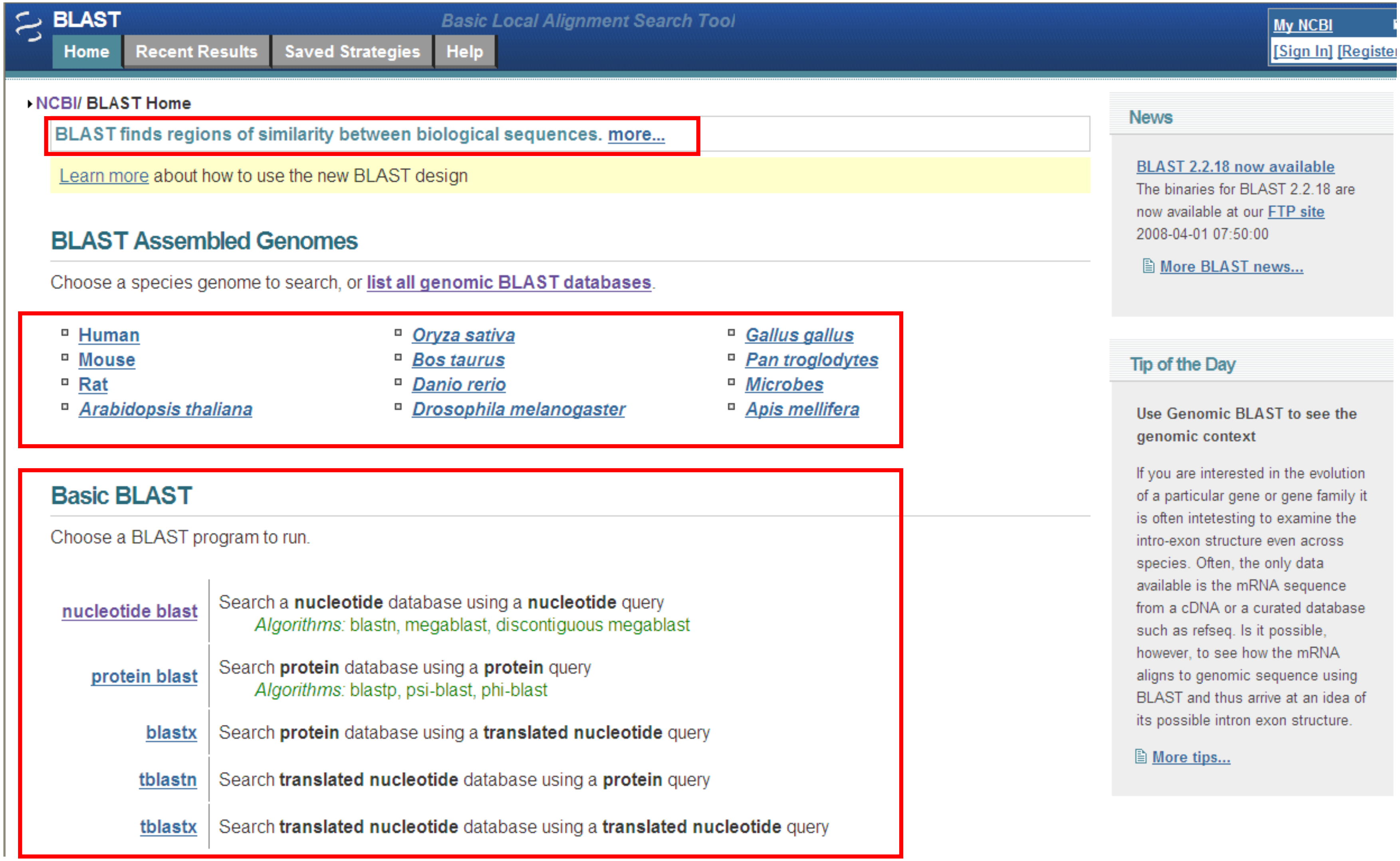
Education resource information in the May NCBI News
07 Jun 2010
May NCBI News is available.

OMIM's new look, Epigenomics in April NCBI News
10 May 2010
The April NCBI News issue is now available.

NIH Roadmap Epigenomics Project data in GEO database
22 Apr 2010
GEO's Roadmap Epigenomics Project Data Listings page allows

More...

BUSCA EM BLAST



NCBI/ BLAST Home

BLAST finds regions of similarity between biological sequences. more...

Learn more about how to use the new BLAST design

BLAST Assembled Genomes

Choose a species genome to search, or list all genomic BLAST databases.

- Human
- Mouse
- Rat
- *Arabidopsis thaliana*
- *Oryza sativa*
- *Bos taurus*
- *Danio rerio*
- *Drosophila melanogaster*
- *Gallus gallus*
- *Pan troglodytes*
- *Microbes*
- *Apis mellifera*

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** query

News

BLAST 2.2.18 now available

The binaries for BLAST 2.2.18 are now available at our FTP site

2008-04-01 07:50:00

More BLAST news...

Tip of the Day

Use Genomic BLAST to see the genomic context

If you are interested in the evolution of a particular gene or gene family it is often intetesting to examine the intro-exon structure even across species. Often, the only data available is the mRNA sequence from a cDNA or a curated database such as refseq. Is it possible, however, to see how the mRNA aligns to genomic sequence using BLAST and thus arrive at an idea of its possible intron exon structure.

More tips...

BUSCA EM BLAST

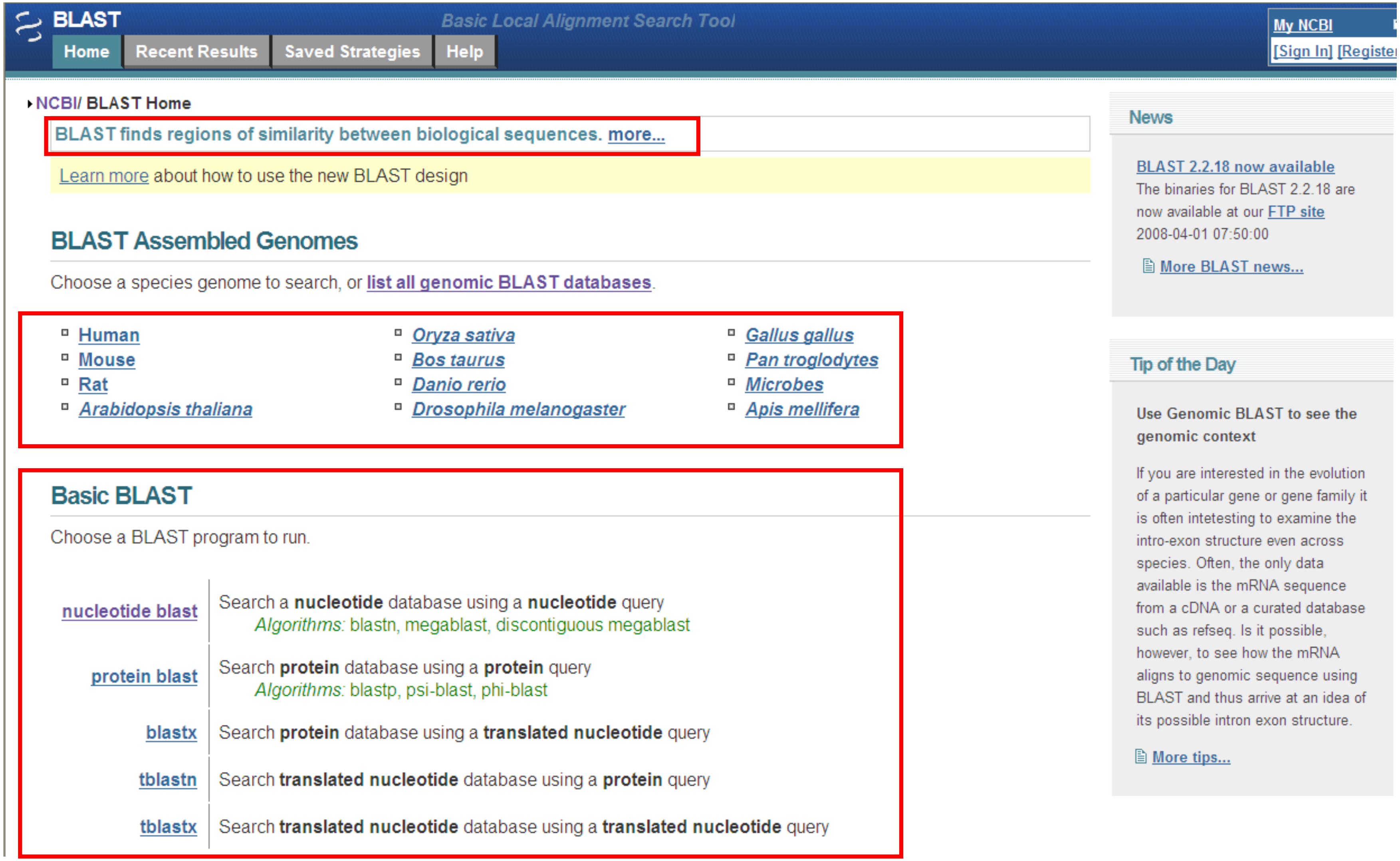
Formato FASTA: formato universalmente aceito para se processado

Identificador

```
>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 EM BLAST



NCBI/ BLAST Home

BLAST finds regions of similarity between biological sequences. more...

Learn more about how to use the new BLAST design

BLAST Assembled Genomes

Choose a species genome to search, or list all genomic BLAST databases.

- Human
- Mouse
- Rat
- *Arabidopsis thaliana*
- *Oryza sativa*
- *Bos taurus*
- *Danio rerio*
- *Drosophila melanogaster*
- *Gallus gallus*
- *Pan troglodytes*
- *Microbes*
- *Apis mellifera*

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** query

News

BLAST 2.2.18 now available

The binaries for BLAST 2.2.18 are now available at our FTP site

2008-04-01 07:50:00

More BLAST news...

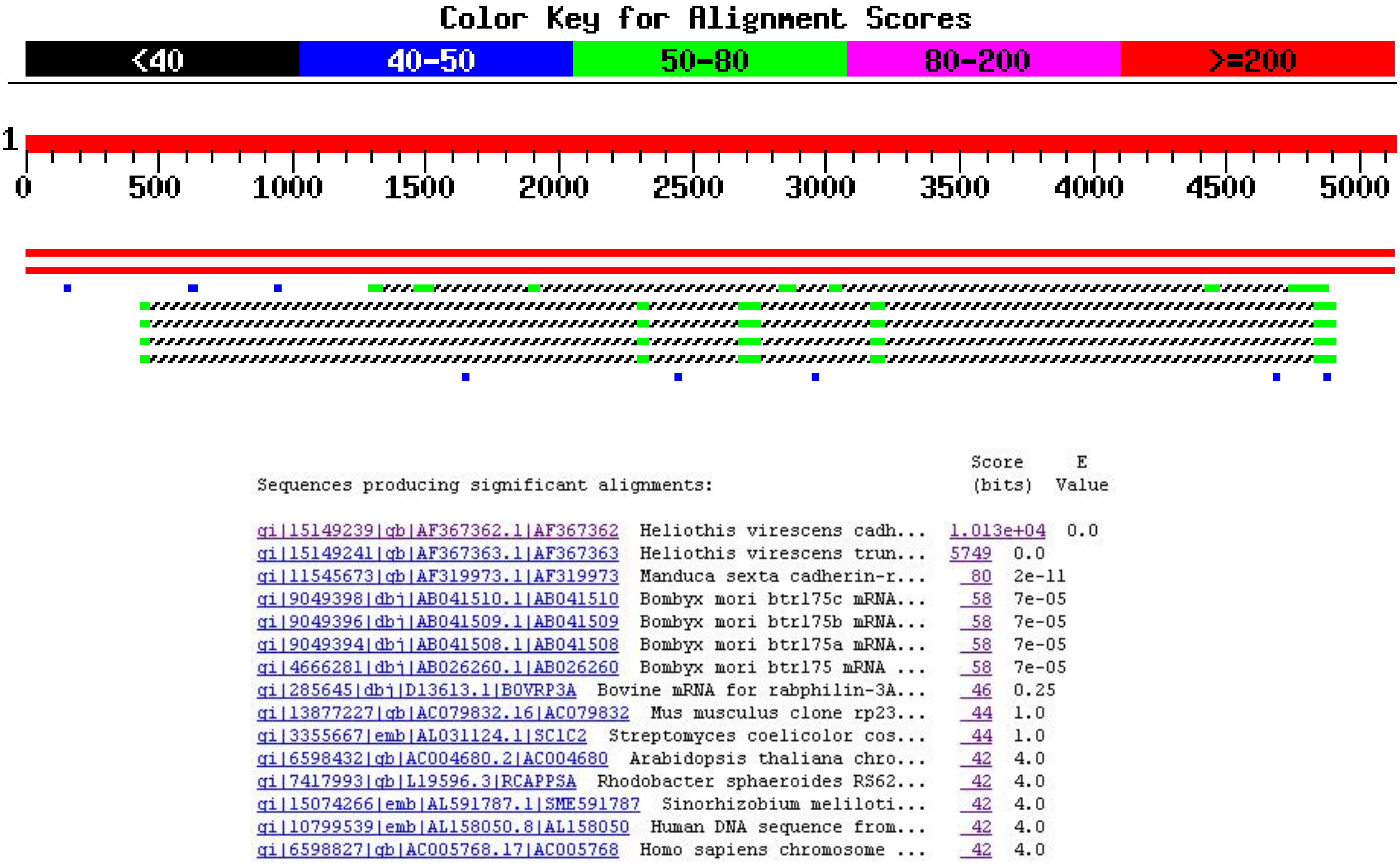
Tip of the Day

Use Genomic BLAST to see the genomic context

If you are interested in the evolution of a particular gene or gene family it is often intetesting to examine the intro-exon structure even across species. Often, the only data available is the mRNA sequence from a cDNA or a curated database such as refseq. Is it possible, however, to see how the mRNA aligns to genomic sequence using BLAST and thus arrive at an idea of its possible intron exon structure.

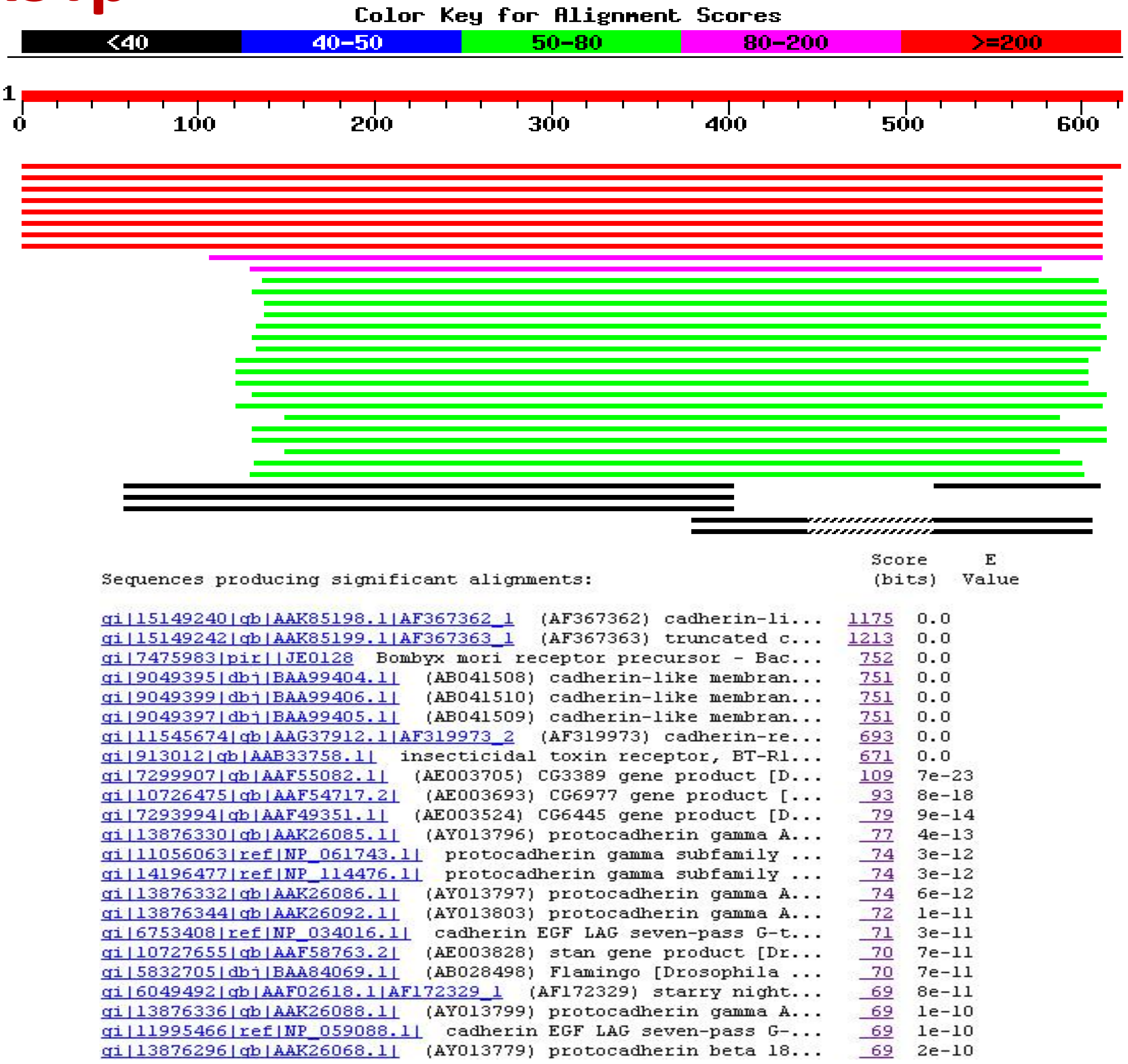
More tips...

BLASTn



```
                                                                   Score     E
Sequences producing significant alignments:                        (bits)  Value

gi|15149239|gb|AF367362.1|AF367362  Heliothis virescens cadh...  1.013e+04  0.0
gi|15149241|gb|AF367363.1|AF367363  Heliothis virescens trun...  5749   0.0
gi|11545673|gb|AF319973.1|AF319973  Manduca sexta cadherin-r...    80   2e-11
gi|9049398|dbj|AB041510.1|AB041510  Bombyx mori btrl75c mRNA...    58   7e-05
gi|9049396|dbj|AB041509.1|AB041509  Bombyx mori btrl75b mRNA...    58   7e-05
gi|9049394|dbj|AB041508.1|AB041508  Bombyx mori btrl75a mRNA...    58   7e-05
gi|4666281|dbj|AB026260.1|AB026260  Bombyx mori btrl75 mRNA ...    58   7e-05
gi|285645|dbj|D13613.1|BOVRP3A  Bovine mRNA for rabphilin-3A...    46   0.25
gi|13877227|gb|AC079832.16|AC079832  Mus musculus clone rp23...    44   1.0
gi|3355667|emb|AL031124.1|SC1C2  Streptomyces coelicolor cos...    44   1.0
gi|6598432|gb|AC004680.2|AC004680  Arabidopsis thaliana chro...    42   4.0
gi|7417993|gb|L19596.3|RCAPPSA  Rhodobacter sphaeroides RS62...    42   4.0
gi|15074266|emb|AL591787.1|SME591787  Sinorhizobium meliloti...    42   4.0
gi|10799539|emb|AL158050.8|AL158050  Human DNA sequence from...    42   4.0
gi|6598827|gb|AC005768.17|AC005768  Homo sapiens chromosome ...    42   4.0
```

BLASTp

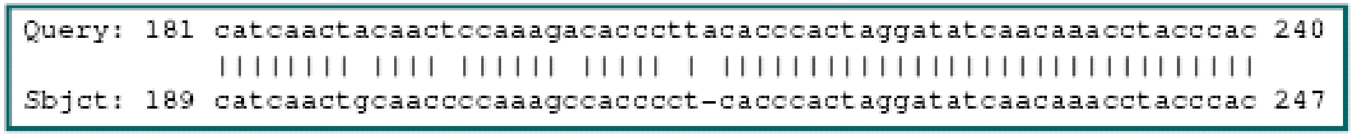


```
                                                                   Score     E
Sequences producing significant alignments:                       (bits)  Value

gi|15149240|gb|AAK85198.1|AF367362_1  (AF367362) cadherin-li...   1175    0.0
gi|15149242|gb|AAK85199.1|AF367363_1  (AF367363) truncated c...   1213    0.0
gi|7475983|pir||JE0128  Bombyx mori receptor precursor - Bac...    752    0.0
gi|9049395|dbj|BAA99404.1|  (AB041508) cadherin-like membran...    751    0.0
gi|9049399|dbj|BAA99406.1|  (AB041510) cadherin-like membran...    751    0.0
gi|9049397|dbj|BAA99405.1|  (AB041509) cadherin-like membran...    751    0.0
gi|11545674|gb|AAG37912.1|AF319973_2  (AF319973) cadherin-re...    693    0.0
gi|913012|gb|AAB33758.1|  insecticidal toxin receptor, BT-R1...    671    0.0
gi|7299907|gb|AAF55082.1|  (AE003705) CG3389 gene product [D...    109    7e-23
gi|10726475|gb|AAF54717.2|  (AE003693) CG6977 gene product [...     93    8e-18
gi|7293994|gb|AAF49351.1|  (AE003524) CG6445 gene product [D...     79    9e-14
gi|13876330|gb|AAK26085.1|  (AY013796) protocadherin gamma A...     77    4e-13
gi|11056063|ref|NP_061743.1|  protocadherin gamma subfamily ...     74    3e-12
gi|14196477|ref|NP_114476.1|  protocadherin gamma subfamily ...     74    3e-12
gi|13876332|gb|AAK26086.1|  (AY013797) protocadherin gamma A...     74    6e-12
gi|13876344|gb|AAK26092.1|  (AY013803) protocadherin gamma A...     72    1e-11
gi|6753408|ref|NP_034016.1|  cadherin EGF LAG seven-pass G-t...     71    3e-11
gi|10727655|gb|AAF58763.2|  (AE003828) stan gene product [Dr...     70    7e-11
gi|5832705|dbj|BAA84069.1|  (AB028498) Flamingo [Drosophila ...     70    7e-11
gi|6049492|gb|AAF02618.1|AF172329_1  (AF172329) starry night...     69    8e-11
gi|13876336|gb|AAK26088.1|  (AY013799) protocadherin gamma A...     69    1e-10
gi|11995466|ref|NP_059088.1|  cadherin EGF LAG seven-pass G-...     69    1e-10
gi|13876296|gb|AAK26068.1|  (AY013779) protocadherin beta 18...     69    2e-10
```

BUSCA EM BLAST

- Nossa sequência – *query* (consulta),

- O resultado da busca em BLAST pode ser um ou mais *hits* em sequências-sujeito (*subject*), ou seja, sequências pertecentes o 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!!!

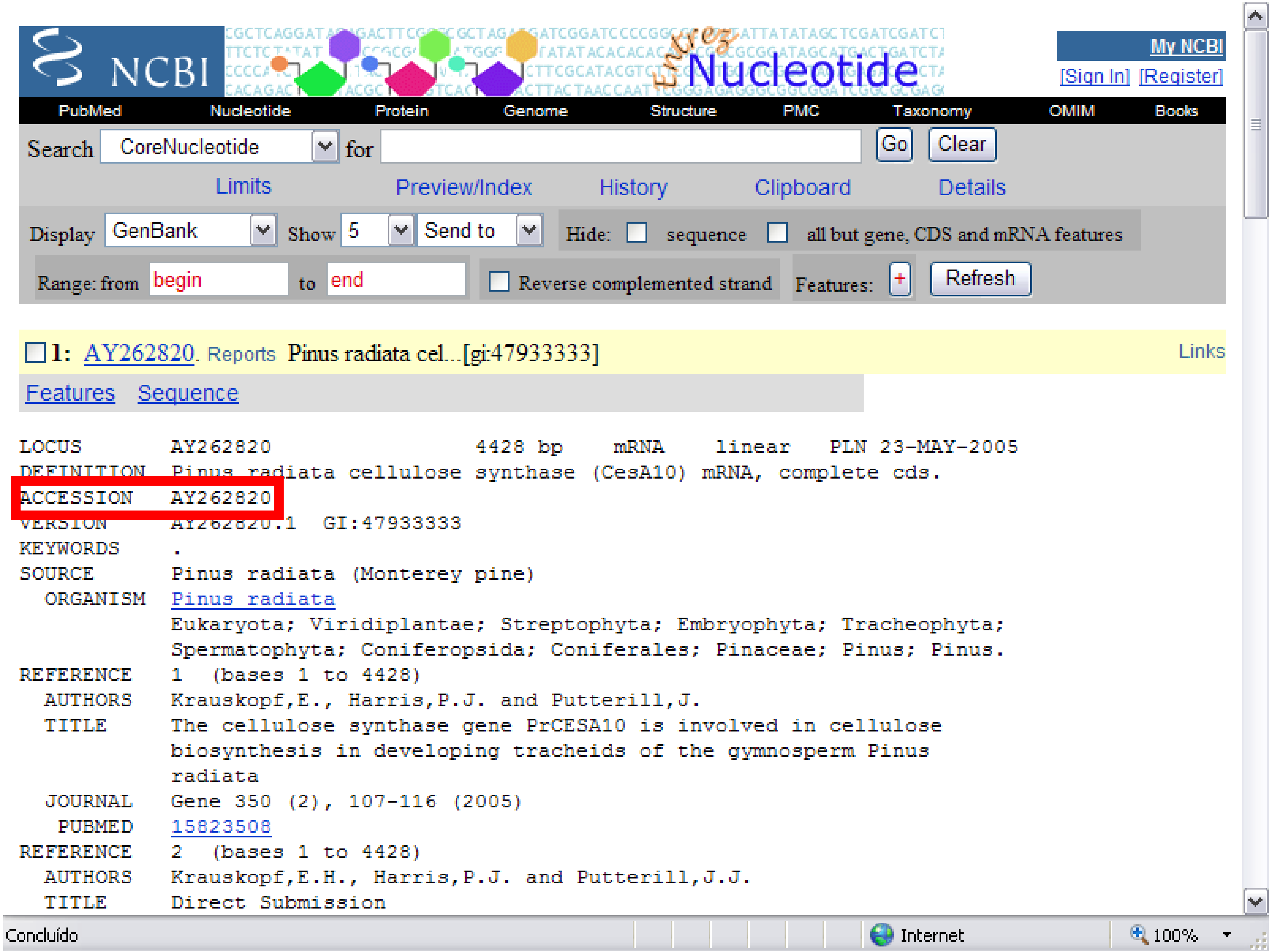
NUCLEOTÍDEOS

FORMATO FASTA

>gi|47933333|gb|AY262820.1| Pinus radiata cellulose synthase (CesA10) mRNA, complete cds
GCACGAGGATTTAATCGAACTCGGTAATTGTTATCATCGTGGTGAGGACTAGTGCTT
GATATTTTAGTTTTATTCTCGAAATTTCATAATAGCTTGGGCTTTCTAAAAAGGGGAAT
GGTGGAATGGGTGTGAGAGTGAAGAGGAATGGTATCGAACCACTAAGAAAAGTAG
TCGTGCAAGTATTAGATGGTTGGCTGTGATAGTTGGAAA...

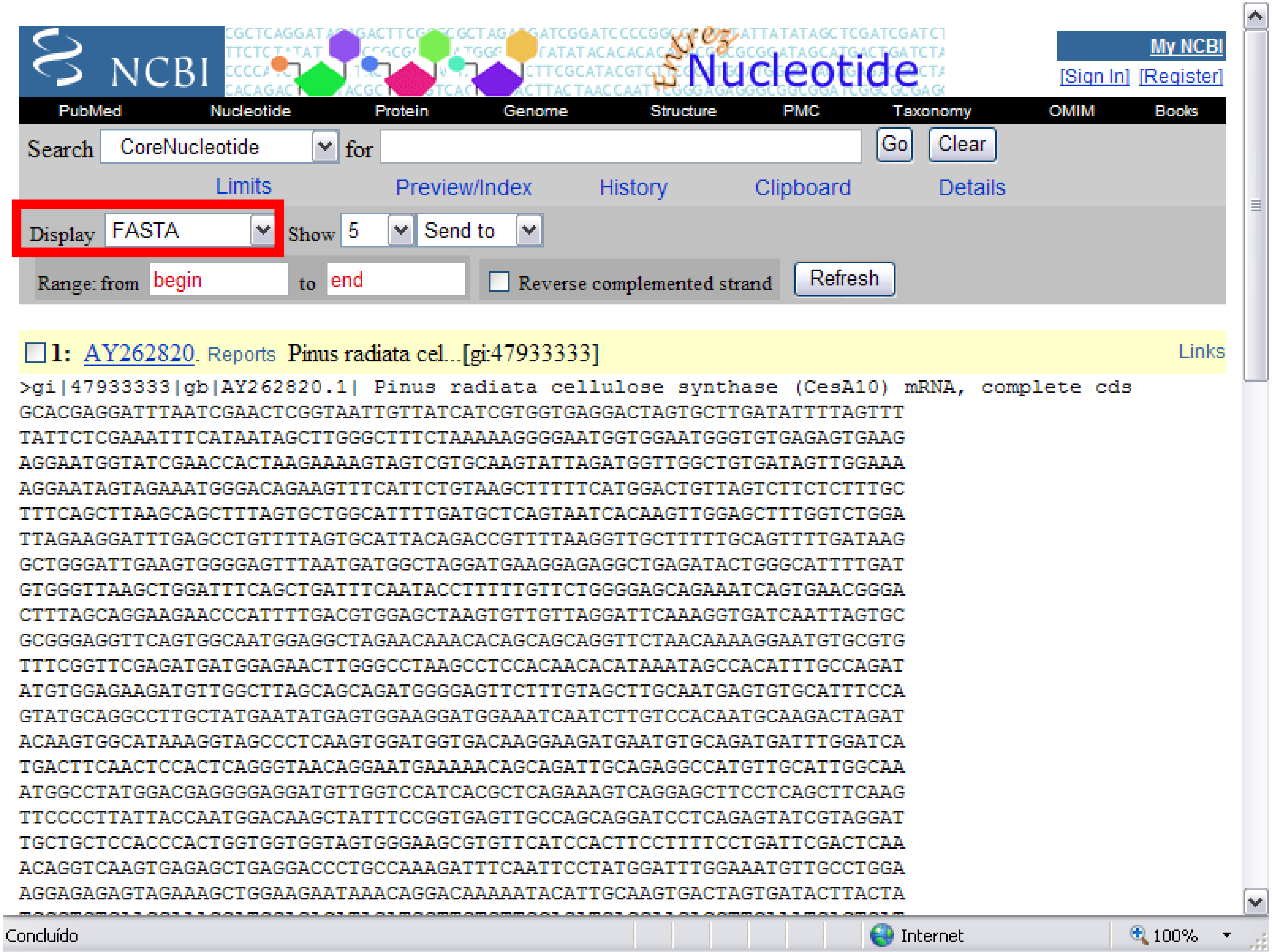
>LINHA DO NOME

MÁXIMA DE 80 CARACTERES POR LINHA



☐ 1: AY262820. Reports Pinus radiata cel...[gi:47933333]

```
LOCUS       AY262820                4428 bp    mRNA    linear   PLN 23-MAY-2005
DEFINITION  Pinus radiata cellulose synthase (CesA10) mRNA, complete cds.
ACCESSION   AY262820
VERSION     AY262820.1  GI:47933333
KEYWORDS    .
SOURCE      Pinus radiata (Monterey pine)
  ORGANISM  Pinus radiata
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Coniferopsida; Coniferales; Pinaceae; Pinus; Pinus.
REFERENCE   1  (bases 1 to 4428)
  AUTHORS   Krauskopf,E., Harris,P.J. and Putterill,J.
  TITLE     The cellulose synthase gene PrCESA10 is involved in cellulose
            biosynthesis in developing tracheids of the gymnosperm Pinus
            radiata
  JOURNAL   Gene 350 (2), 107-116 (2005)
   PUBMED   15823508
REFERENCE   2  (bases 1 to 4428)
  AUTHORS   Krauskopf,E.H., Harris,P.J. and Putterill,J.J.
  TITLE     Direct Submission
```



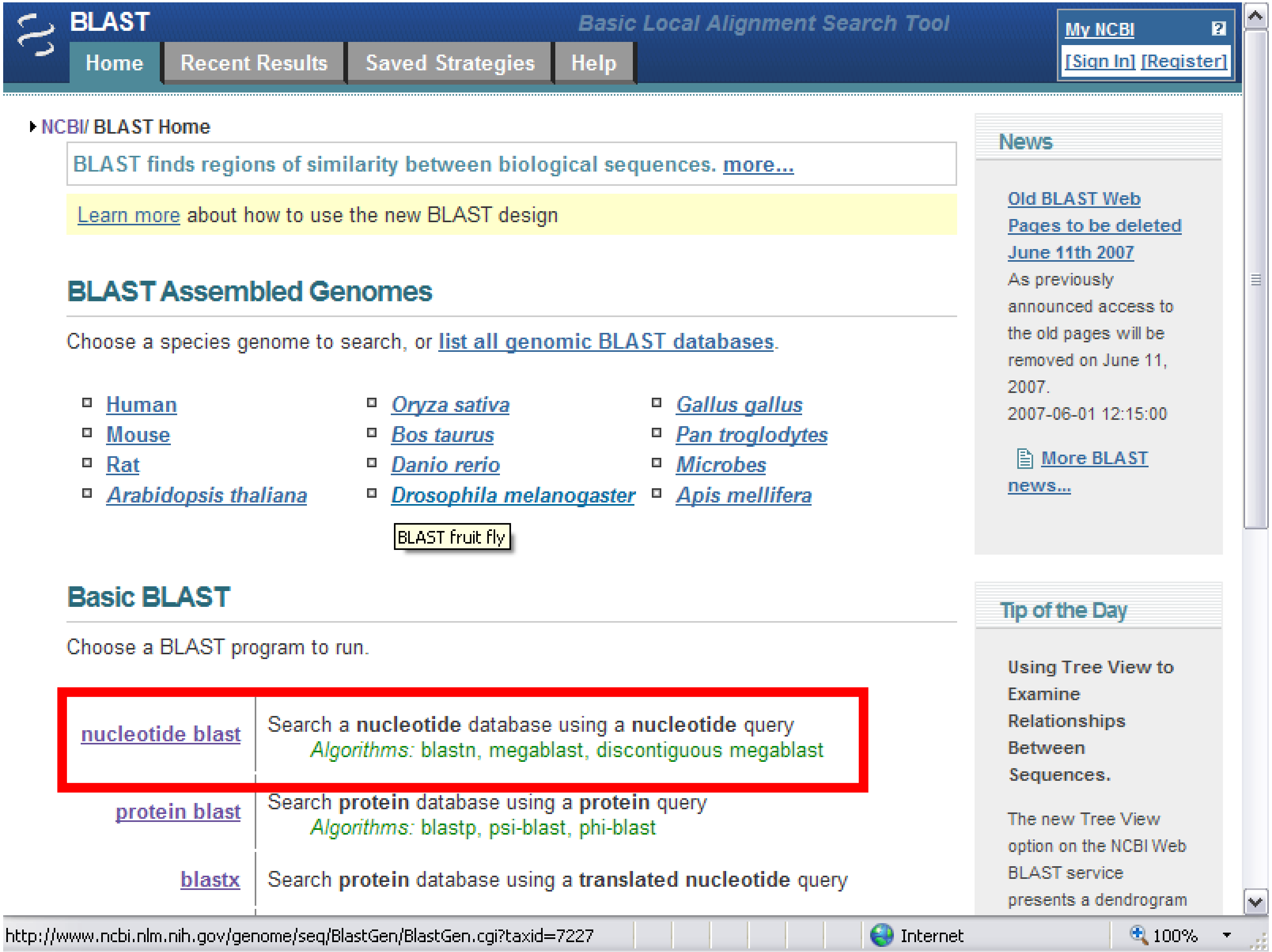
Search CoreNucleotide for

Display FASTA Show 5

Range: from begin to end ☐ Reverse complemented strand

☐ 1: AY262820. Reports Pinus radiata cel...[gi:47933333]

```
>gi|47933333|gb|AY262820.1| Pinus radiata cellulose synthase (CesA10) mRNA, complete cds
GCACGAGGATTTAATCGAACTCGGTAATTGTTATCATCGTGGTGAGGACTAGTGCTTGATATTTTAGTTT
TATTCTCGAAATTTCATAATAGCTTGGGCTTTCTAAAAAGGGGAATGGTGGAATGGGTGTGAGAGTGAAG
AGGAATGGTATCGAACCACTAAGAAAAGTAGTCGTGCAAGTATTAGATGGTTGGCTGTGATAGTTGGAAA
AGGAATAGTAGAAATGGGACAGAAGTTTCATTCTGTAAGCTTTTTCATGGACTGTTAGTCTTCTCTTTGC
TTTCAGCTTAAGCAGCTTTAGTGCTGGCATTTTGATGCTCAGTAATCACAAGTTGGAGCTTTGGTCTGGA
TTAGAAGGATTTGAGCCTGTTTTAGTGCATTACAGACCGTTTTAAGGTTGCTTTTTGCAGTTTTGATAAG
GCTGGGATTGAAGTGGGGAGTTTAATGATGGCTAGGATGAAGGAGAGGCTGAGATACTGGGCATTTTGAT
GTGGGTTAAGCTGGATTTCAGCTGATTTCAATACCTTTTTGTTCTGGGGAGCAGAAATCAGTGAACGGGA
CTTTAGCAGGAAGAACCCATTTTGACGTGGAGCTAAGTGTTGTTAGGATTCAAAGGTGATCAATTAGTGC
GCGGGAGGTTCAGTGGCAATGGAGGCTAGAACAAACACAGCAGCAGGTTCTAACAAAAGGAATGTGCGTG
TTTCGGTTCGAGATGATGGAGAACTTGGGCCTAAGCCTCCACAACACATAAATAGCCACATTTGCCAGAT
ATGTGGAGAAGATGTTGGCTTAGCAGCAGATGGGGAGTTCTTTGTAGCTTGCAATGAGTGTGCATTTCCA
GTATGCAGGCCTTGCTATGAATATGAGTGGAAGGATGGAAATCAATCTTGTCCACAATGCAAGACTAGAT
ACAAGTGGCATAAAGGTAGCCCTCAAGTGGATGGTGACAAGGAAGATGAATGTGCAGATGATTTGGATCA
TGACTTCAACTCCACTCAGGGTAACAGGAATGAAAAACAGCAGATTGCAGAGGCCATGTTGCATTGGCAA
ATGGCCTATGGACGAGGGGAGGATGTTGGTCCATCACGCTCAGAAAGTCAGGAGCTTCCTCAGCTTCAAG
TTCCCCTTATTACCAATGGACAAGCTATTTCCGGTGAGTTGCCAGCAGGATCCTCAGAGTATCGTAGGAT
TGCTGCTCCACCCACTGGTGGTGGTAGTGGGAAGCGTGTTCATCCACTTCCTTTTCCTGATTCGACTCAA
ACAGGTCAAGTGAGAGCTGAGGACCCTGCCAAAGATTTCAATTCCTATGGATTTGGAAATGTTGCCTGGA
AGGAGAGAGTAGAAAGCTGGAAGAATAAACAGGACAAAAATACATTGCAAGTGACTAGTGATACTTACTA
```



NCBI/ BLAST Home

BLAST finds regions of similarity between biological sequences. more...

Learn more about how to use the new BLAST design

BLAST Assembled Genomes

Choose a species genome to search, or list all genomic BLAST databases.

- Human
- Mouse
- Rat
- *Arabidopsis thaliana*
- *Oryza sativa*
- *Bos taurus*
- *Danio rerio*
- *Drosophila melanogaster*
- *Gallus gallus*
- *Pan troglodytes*
- *Microbes*
- *Apis mellifera*

BLAST fruit fly

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

News

Old BLAST Web Pages to be deleted June 11th 2007

As previously announced access to the old pages will be removed on June 11, 2007.

2007-06-01 12:15:00

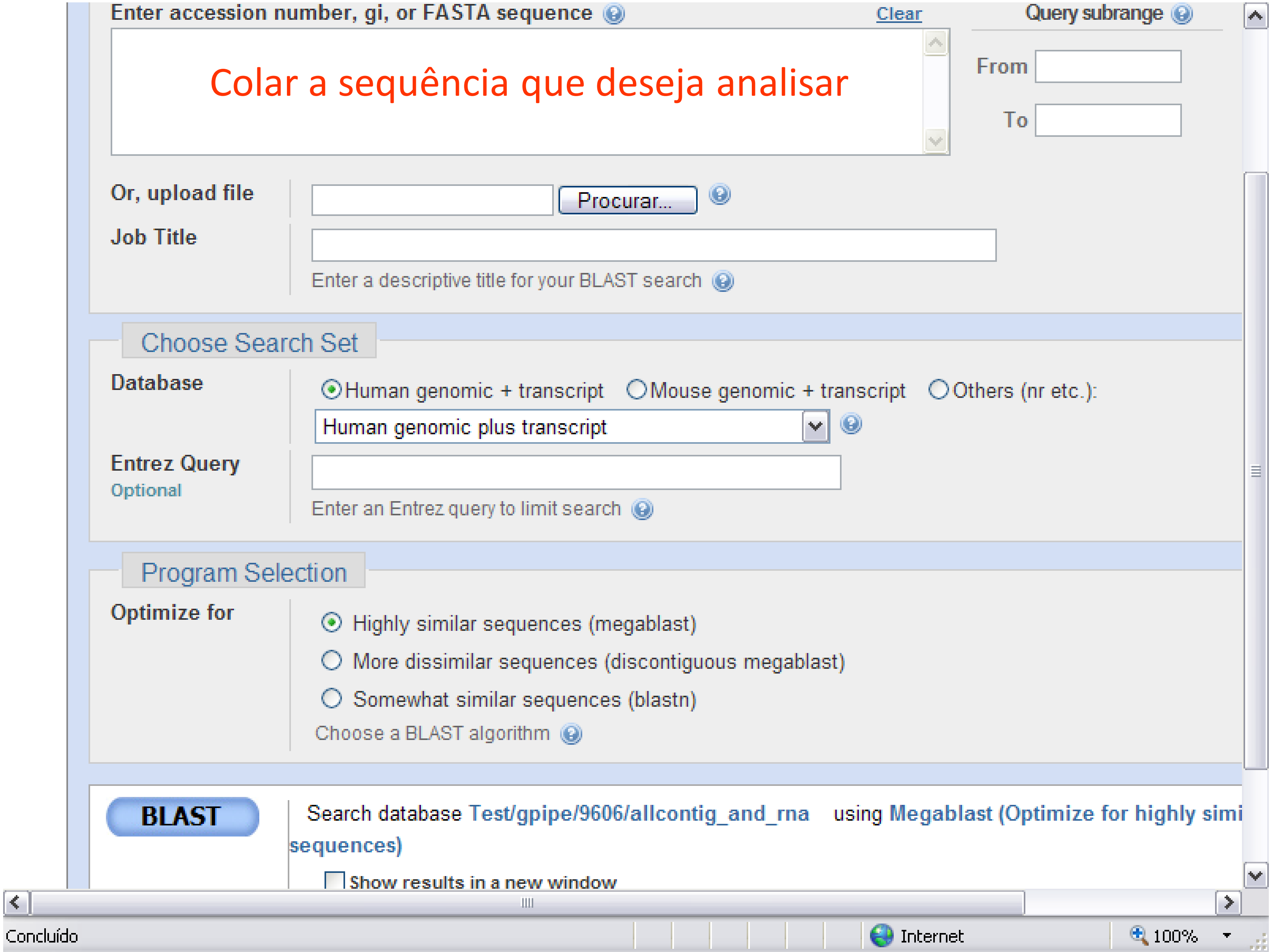
More BLAST news...

Tip of the Day

Using Tree View to Examine Relationships Between Sequences.

The new Tree View option on the NCBI Web BLAST service presents a dendrogram

http://www.ncbi.nlm.nih.gov/genome/seq/BlastGen/BlastGen.cgi?taxid=7227



Enter accession number, gi, or FASTA sequence

Colar a sequência que deseja analisar

Query subrange

From

To

Or, upload file

Job Title

Enter a descriptive title for your BLAST search

Choose Search Set

Database

● Human genomic + transcript ○ Mouse genomic + transcript ○ Others (nr etc.):

Human genomic plus transcript

Entrez Query

Optional

Enter an Entrez query to limit search

Program Selection

Optimize for

● Highly similar sequences (megablast)

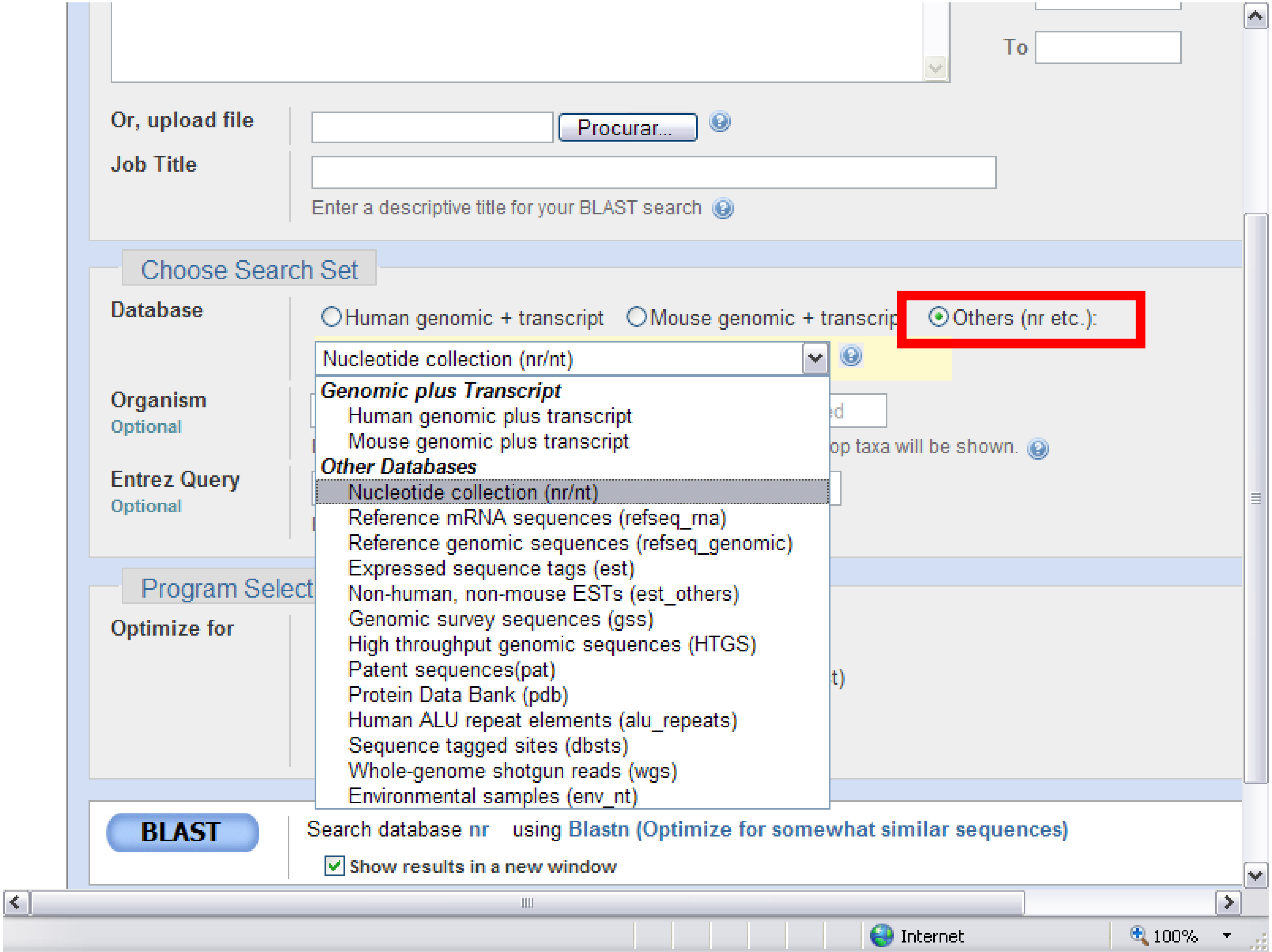
○ More dissimilar sequences (discontiguous megablast)

○ Somewhat similar sequences (blastn)

Choose a BLAST algorithm

Search database Test/gpipe/9606/allcontig_and_rna using Megablast (Optimize for highly simi sequences)

☐ Show results in a new window



Enter a descriptive title for your BLAST search

Database: ○ Human genomic + transcript ○ Mouse genomic + transcrip ● Others (nr etc.):

Nucleotide collection (nr/nt)

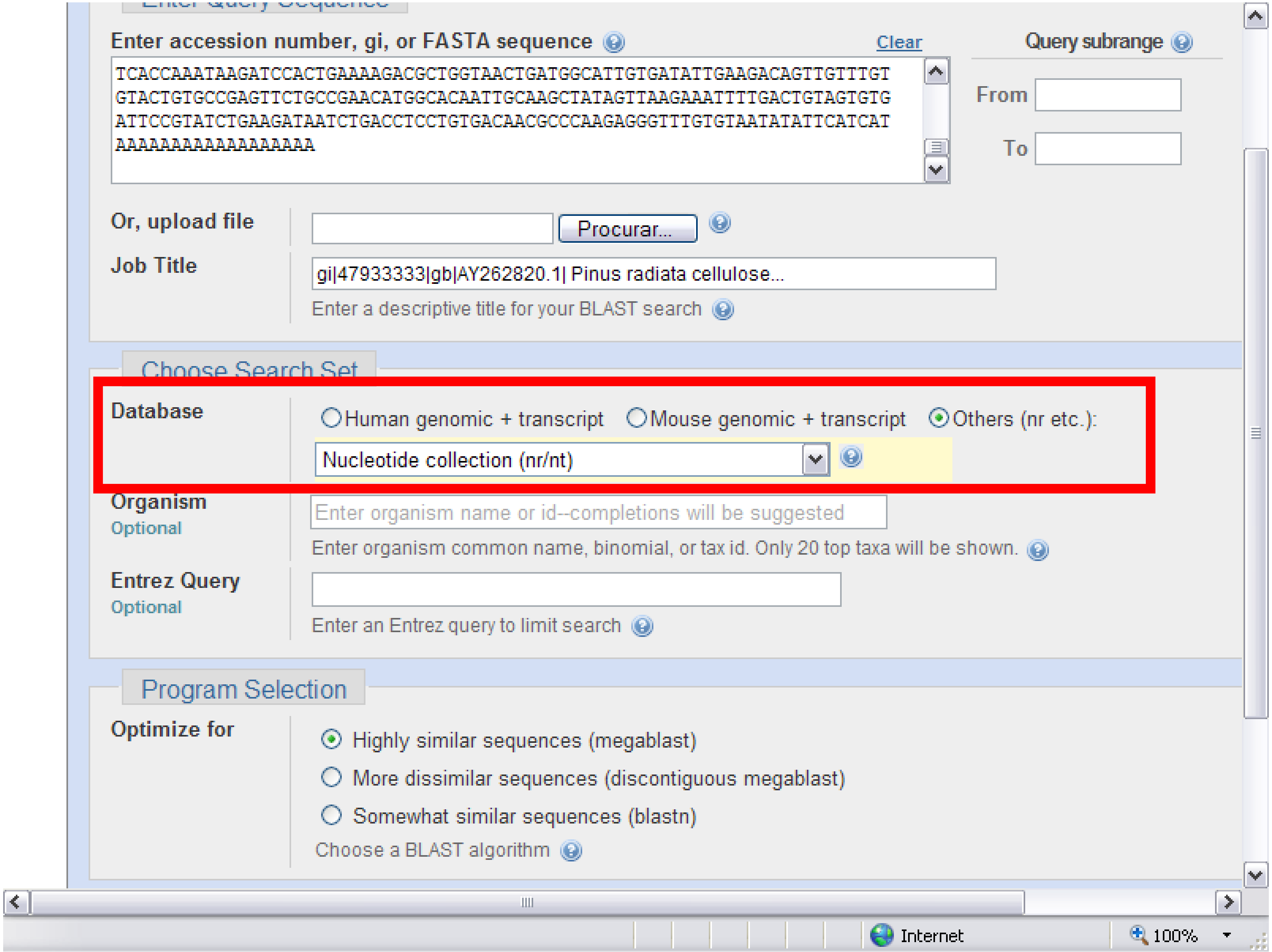
Genomic plus Transcript
- Human genomic plus transcript
- Mouse genomic plus transcript

Other Databases
- Nucleotide collection (nr/nt)
- Reference mRNA sequences (refseq_rna)
- Reference genomic sequences (refseq_genomic)
- Expressed sequence tags (est)
- Non-human, non-mouse ESTs (est_others)
- Genomic survey sequences (gss)
- High throughput genomic sequences (HTGS)
- Patent sequences(pat)
- Protein Data Bank (pdb)
- Human ALU repeat elements (alu_repeats)
- Sequence tagged sites (dbsts)
- Whole-genome shotgun reads (wgs)
- Environmental samples (env_nt)

op taxa will be shown.

Search database nr using Blastn (Optimize for somewhat similar sequences)

☑ Show results in a new window



Enter accession number, gi, or FASTA sequence

```
TCACCAAATAAGATCCACTGAAAAGACGCTGGTAACTGATGGCATTGTGATATTGAAGACAGTTGTTTGT
GTACTGTGCCGAGTTCTGCCGAACATGGCACAATTGCAAGCTATAGTTAAGAAATTTTGACTGTAGTGTG
ATTCCGTATCTGAAGATAATCTGACCTCCTGTGACAACGCCCAAGAGGGTTTGTGTAATATATTCATCAT
AAAAAAAAAAAAAAAAA
```

Query subrange

From

To

Or, upload file

Job Title: gi|47933333|gb|AY262820.1| Pinus radiata cellulose...

Enter a descriptive title for your BLAST search

Choose Search Set

Database: ○ Human genomic + transcript ○ Mouse genomic + transcript ● Others (nr etc.):

Nucleotide collection (nr/nt)

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

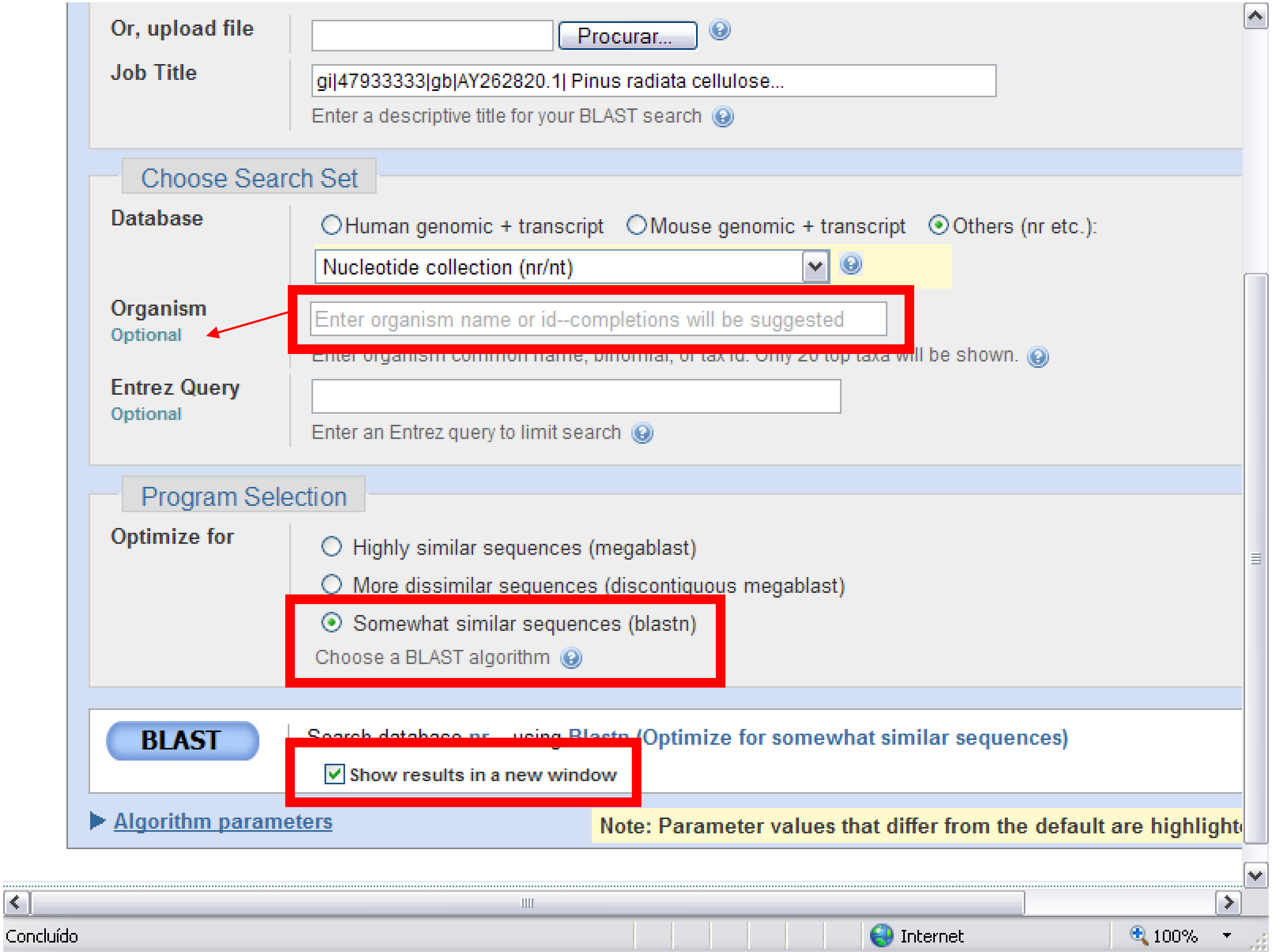
Optimize for

● Highly similar sequences (megablast)

○ More dissimilar sequences (discontiguous megablast)

○ Somewhat similar sequences (blastn)

Choose a BLAST algorithm



Or, upload file Procurar...

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

Enter a descriptive title for your BLAST search

Choose Search Set

Database ○ Human genomic + transcript ○ Mouse genomic + transcript ● Others (nr etc.):

Nucleotide collection (nr/nt)

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

Optimize for

○ Highly similar sequences (megablast)

○ More dissimilar sequences (discontiguous megablast)

● Somewhat similar sequences (blastn)

Choose a BLAST algorithm

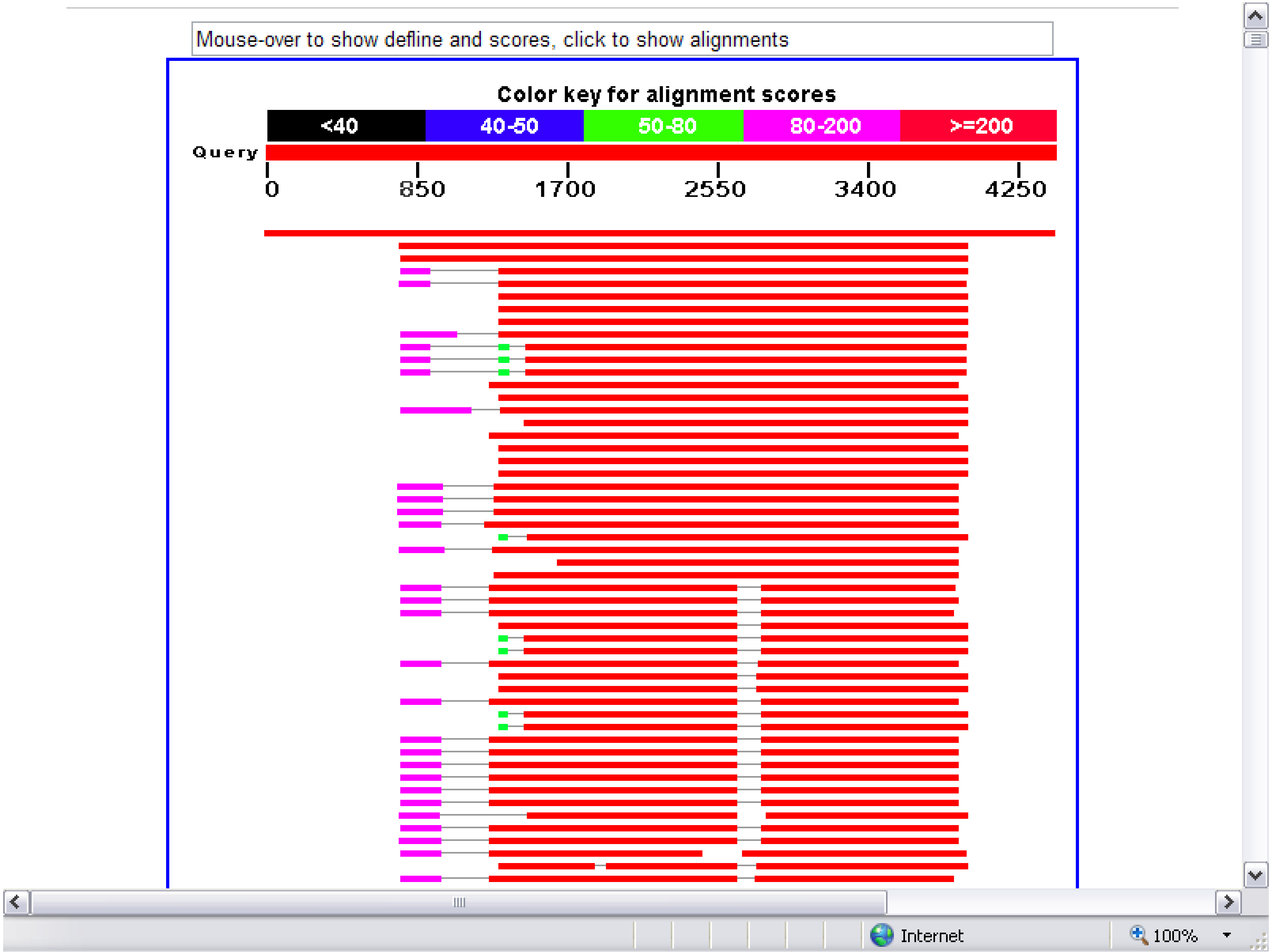
BLAST Search database nr using Blastn (Optimize for somewhat similar sequences)

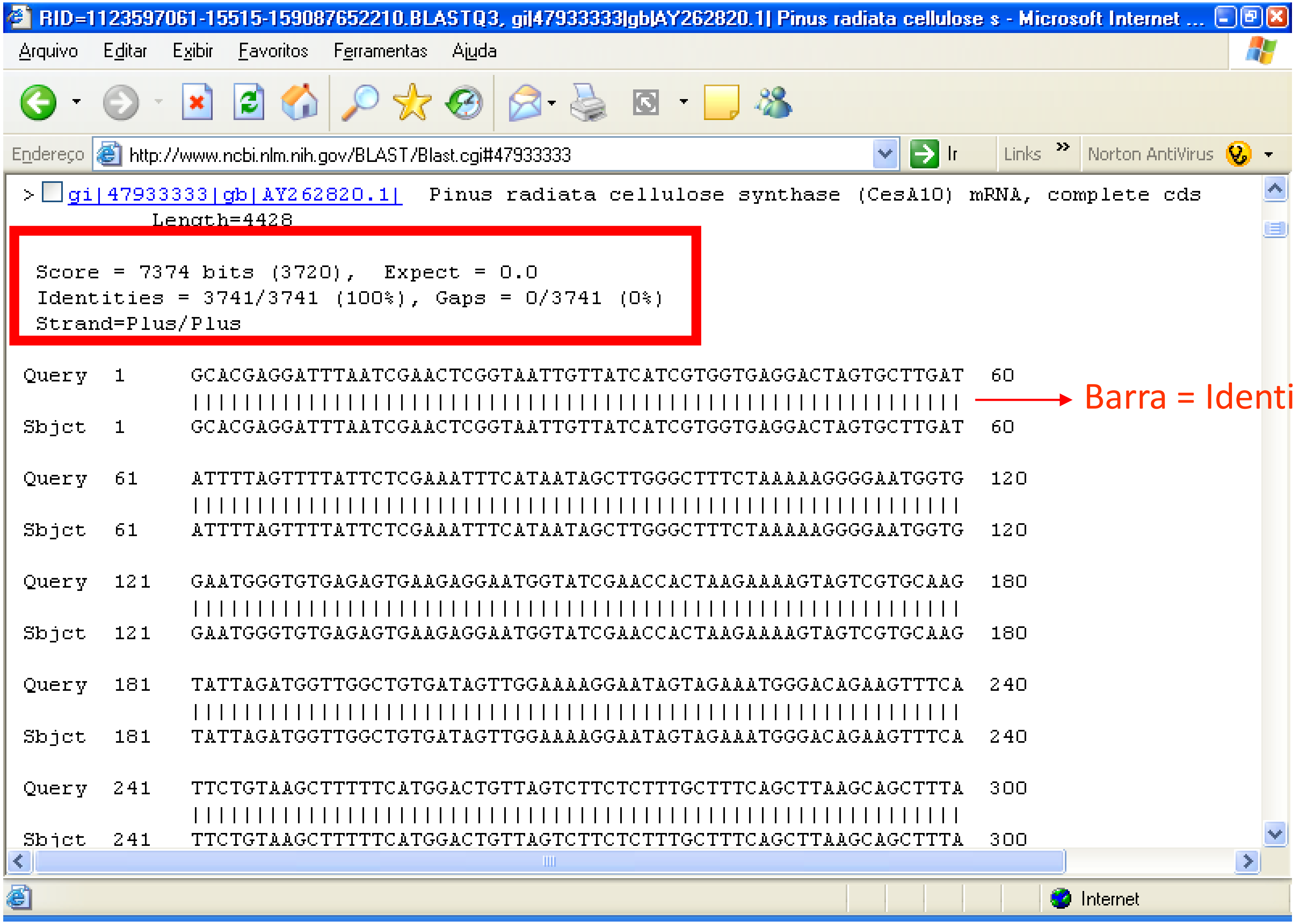
☑ Show results in a new window

Algorithm parameters

Note: Parameter values that differ from the default are highlight

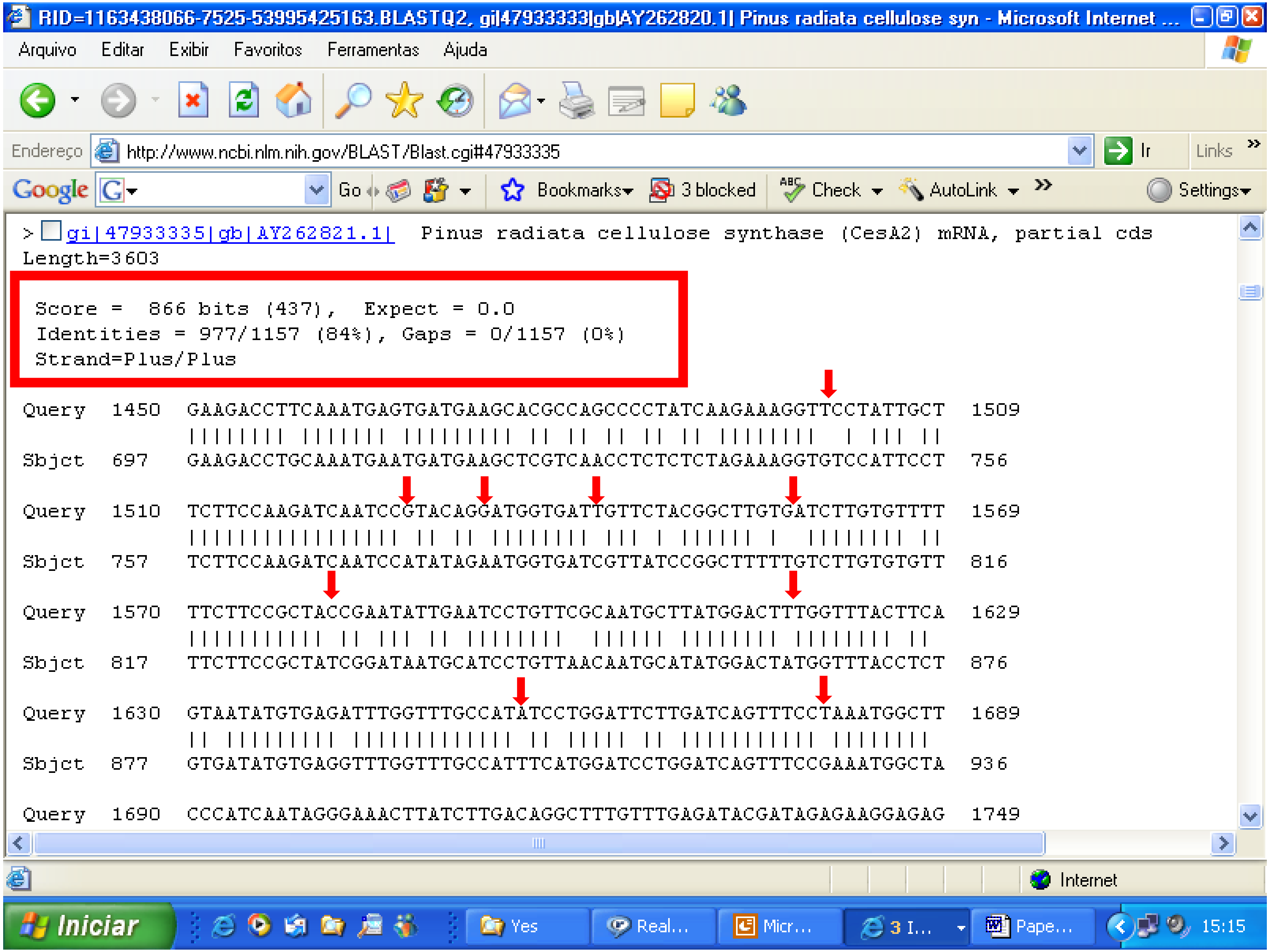
Concluído Internet 100%


Mouse-over to show defline and scores, click to show alignments

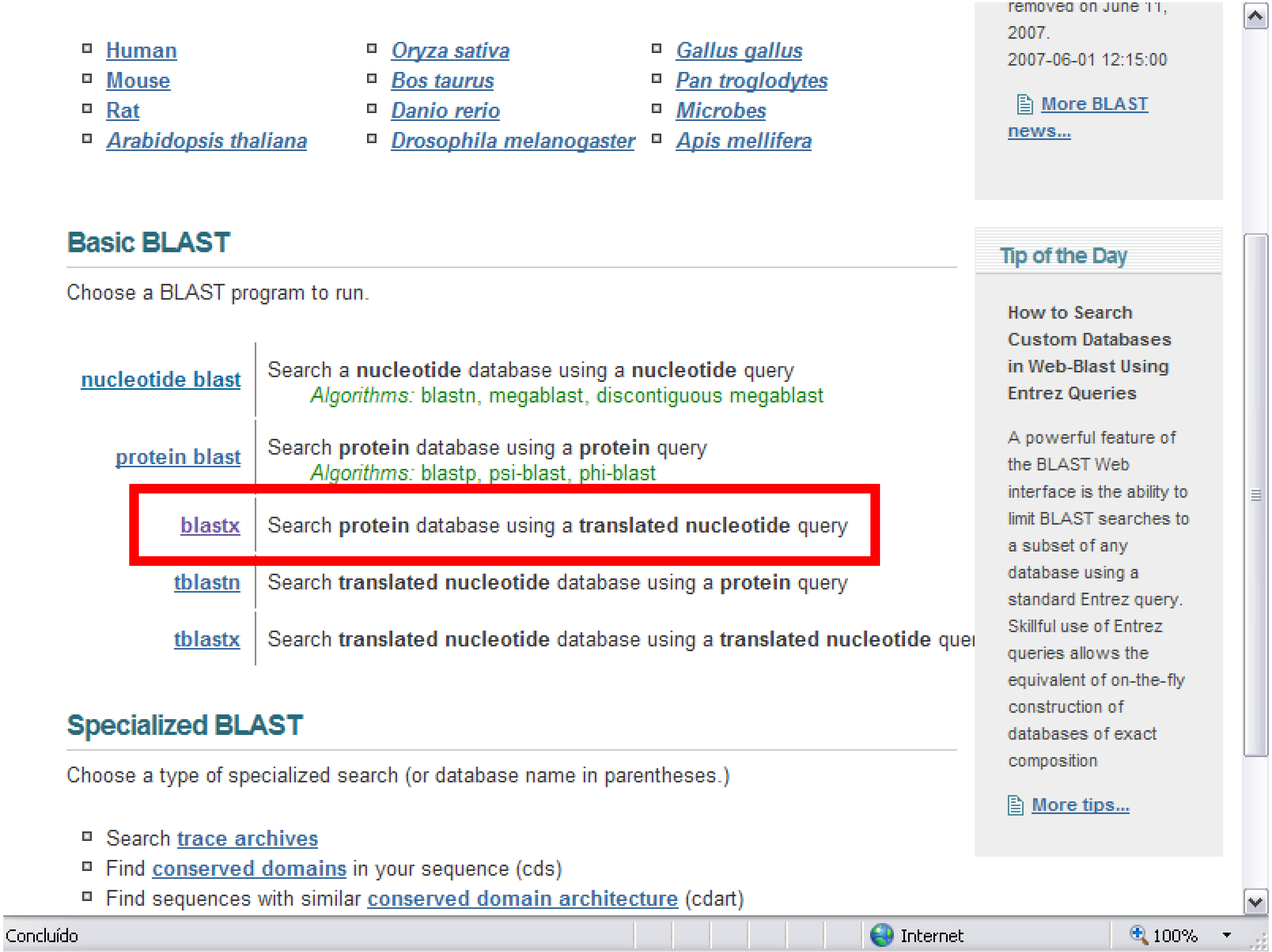


```
> ☐ gi|47933333|gb|AY262820.1|  Pinus radiata cellulose synthase (CesA10) mRNA, complete cds
          Length=4428

 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
```



- 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.)

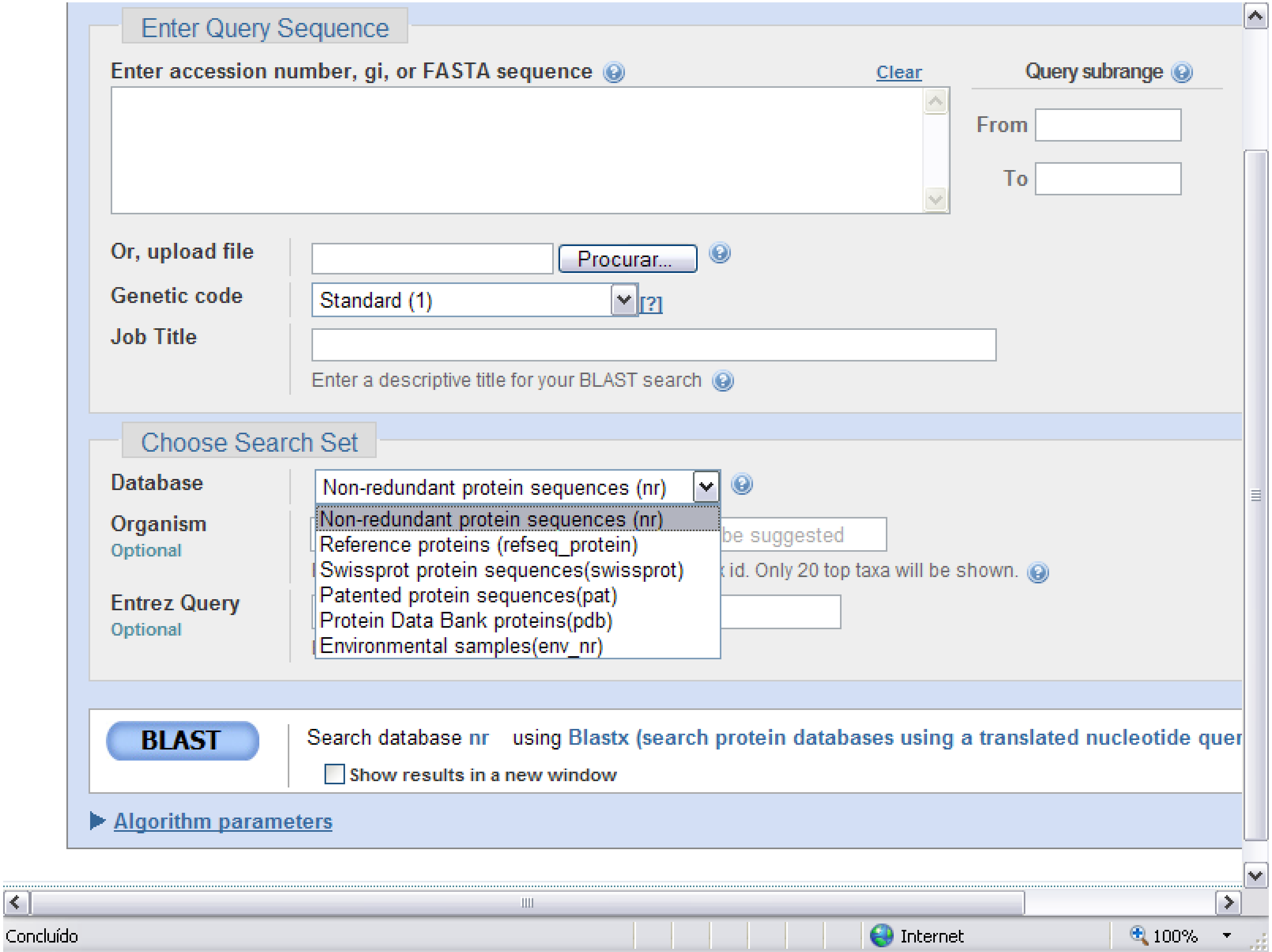
- Search trace archives
- Find conserved domains in your sequence (cds)
- Find sequences with similar conserved domain architecture (cdart)

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...



Genetic code: Standard (1)

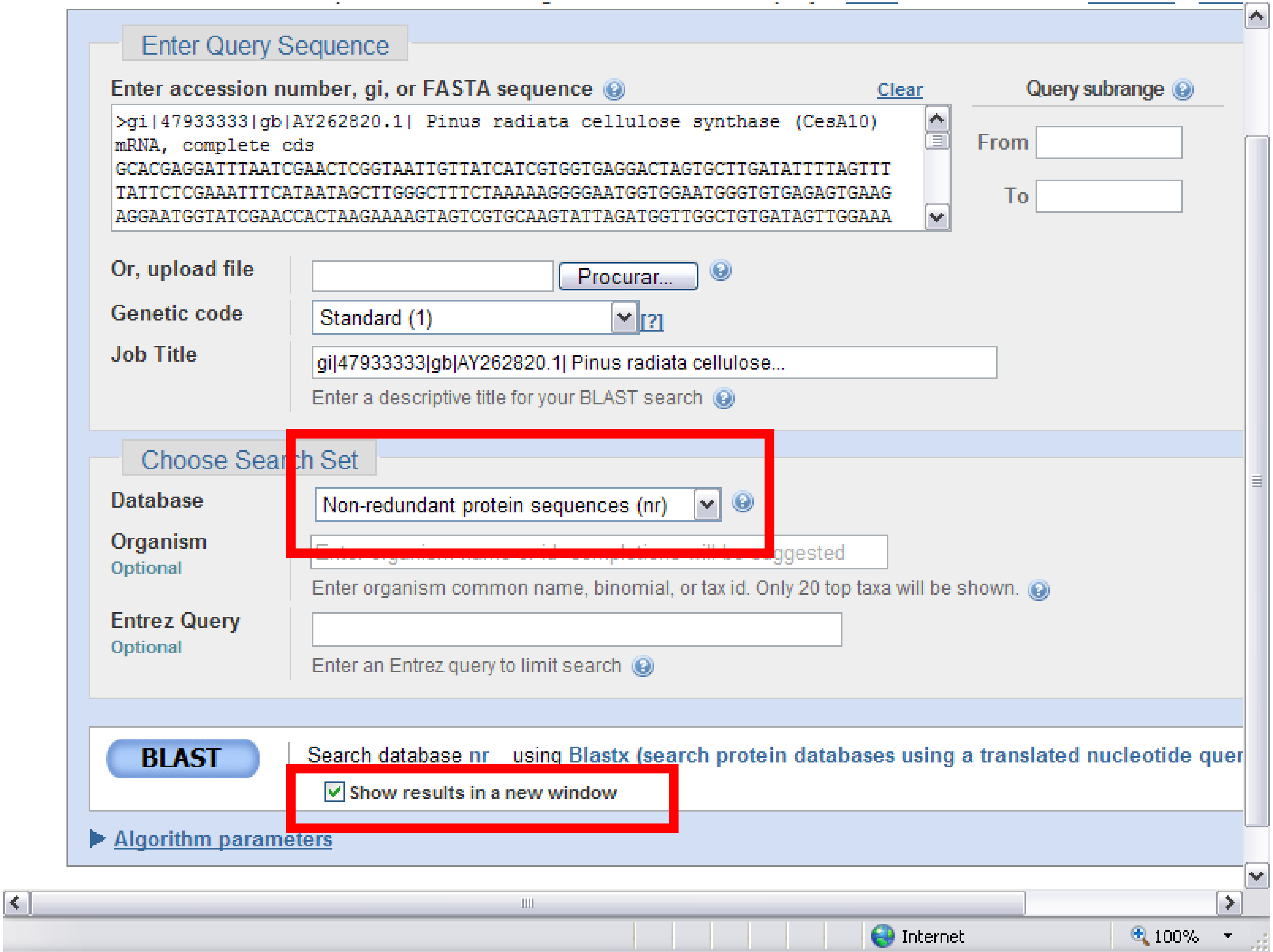
Enter a descriptive title for your BLAST search

Database: Non-redundant protein sequences (nr)

id. Only 20 top taxa will be shown.

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

☐ Show results in a new window



Enter Query Sequence

Enter accession number, gi, or FASTA sequence Clear

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

Query subrange

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 or id - completions will be suggested

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

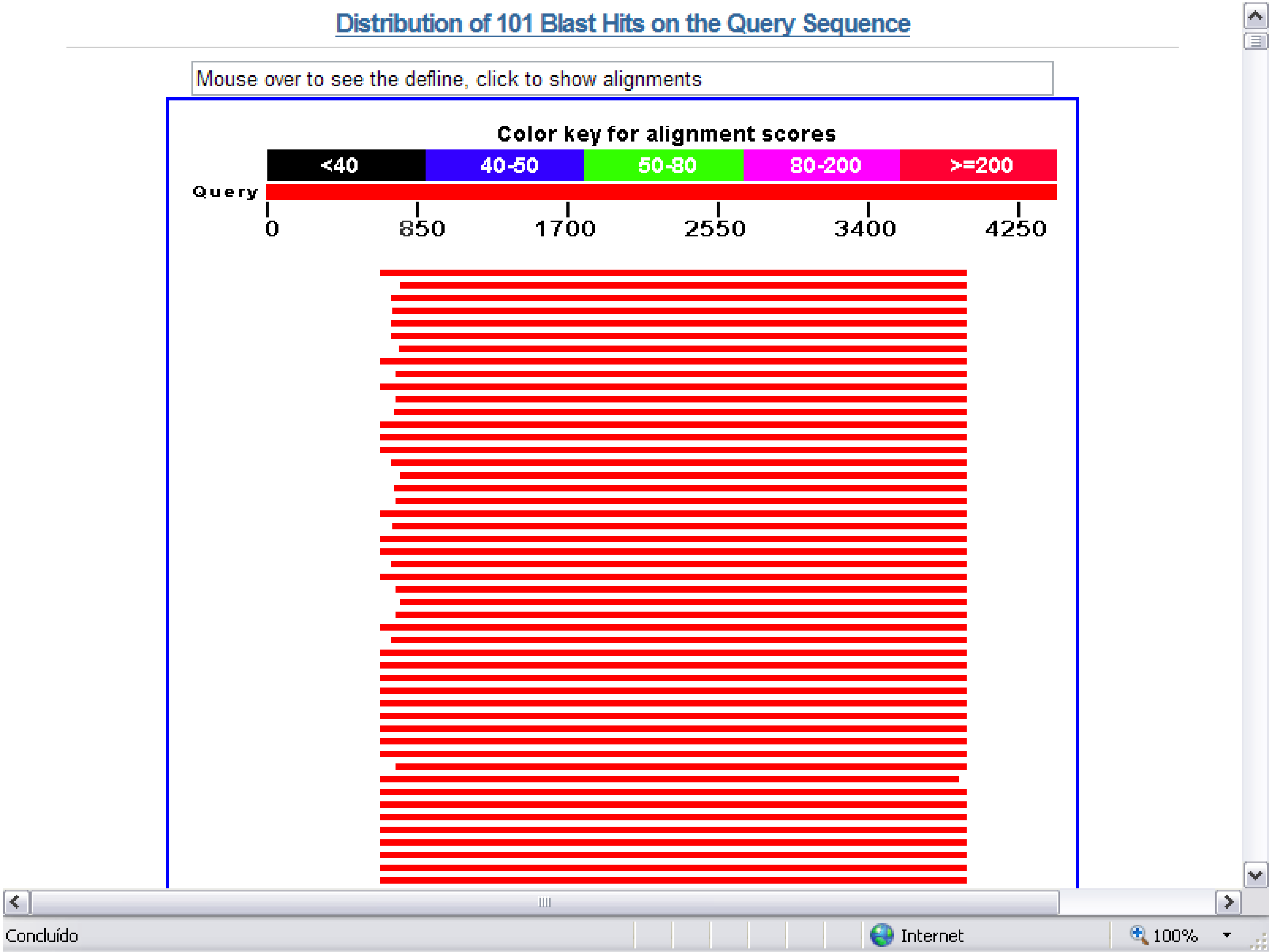
Entrez Query
Optional

Enter an Entrez query to limit search

BLAST Search database nr using Blastx (search protein databases using a translated nucleotide quer

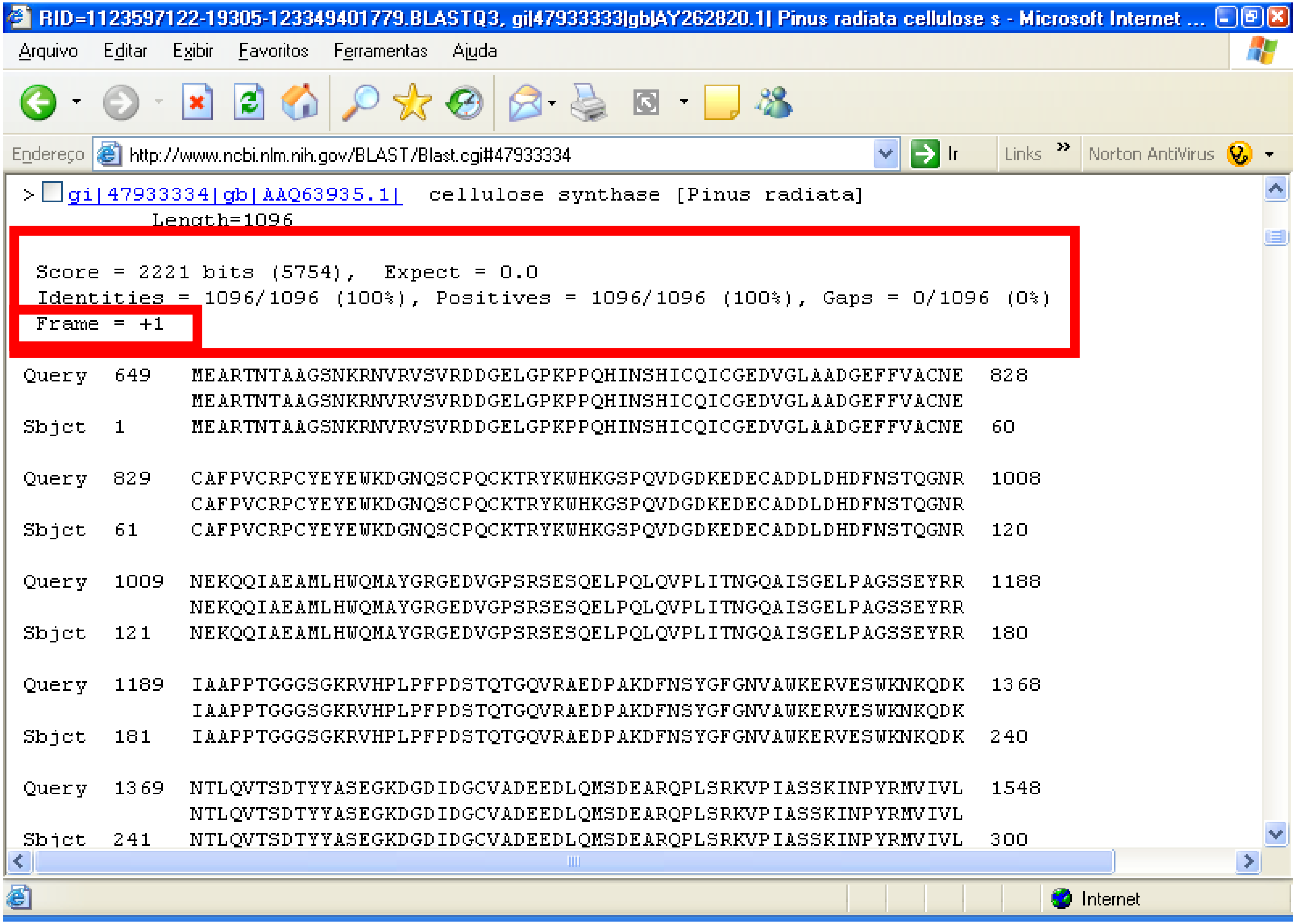
☑ Show results in a new window

Algorithm parameters

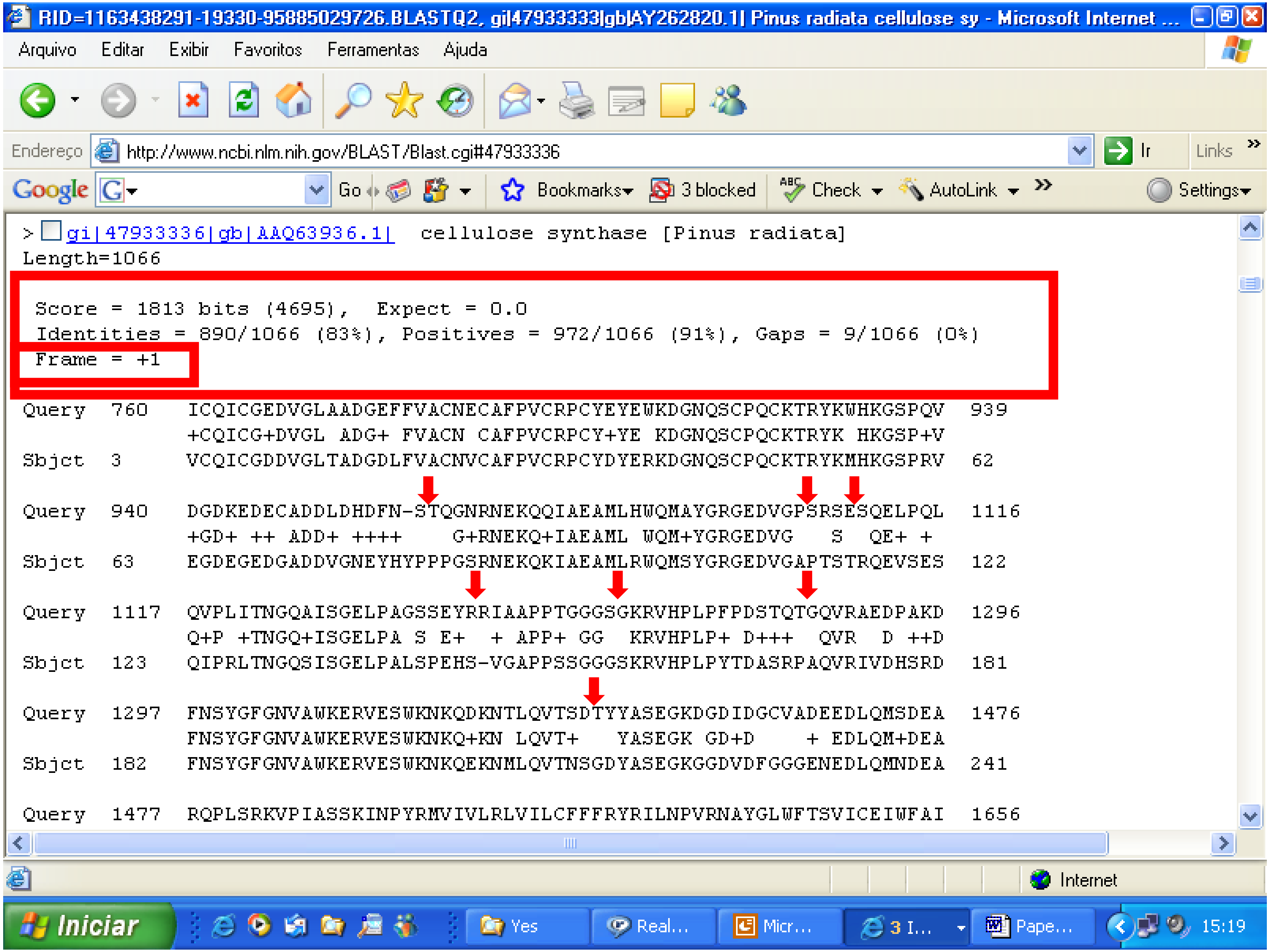


Distribution of 101 Blast Hits on the Query Sequence

Mouse over to see the defline, click to show alignments



```
> ☐ 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
```



RID=1163438291-19330-95885029726.BLASTQ2, gi|47933333|gb|AY262820.1| Pinus radiata cellulose sy - Microsoft Internet ...

Endereço http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi#47933336

```
> ☐ 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
```

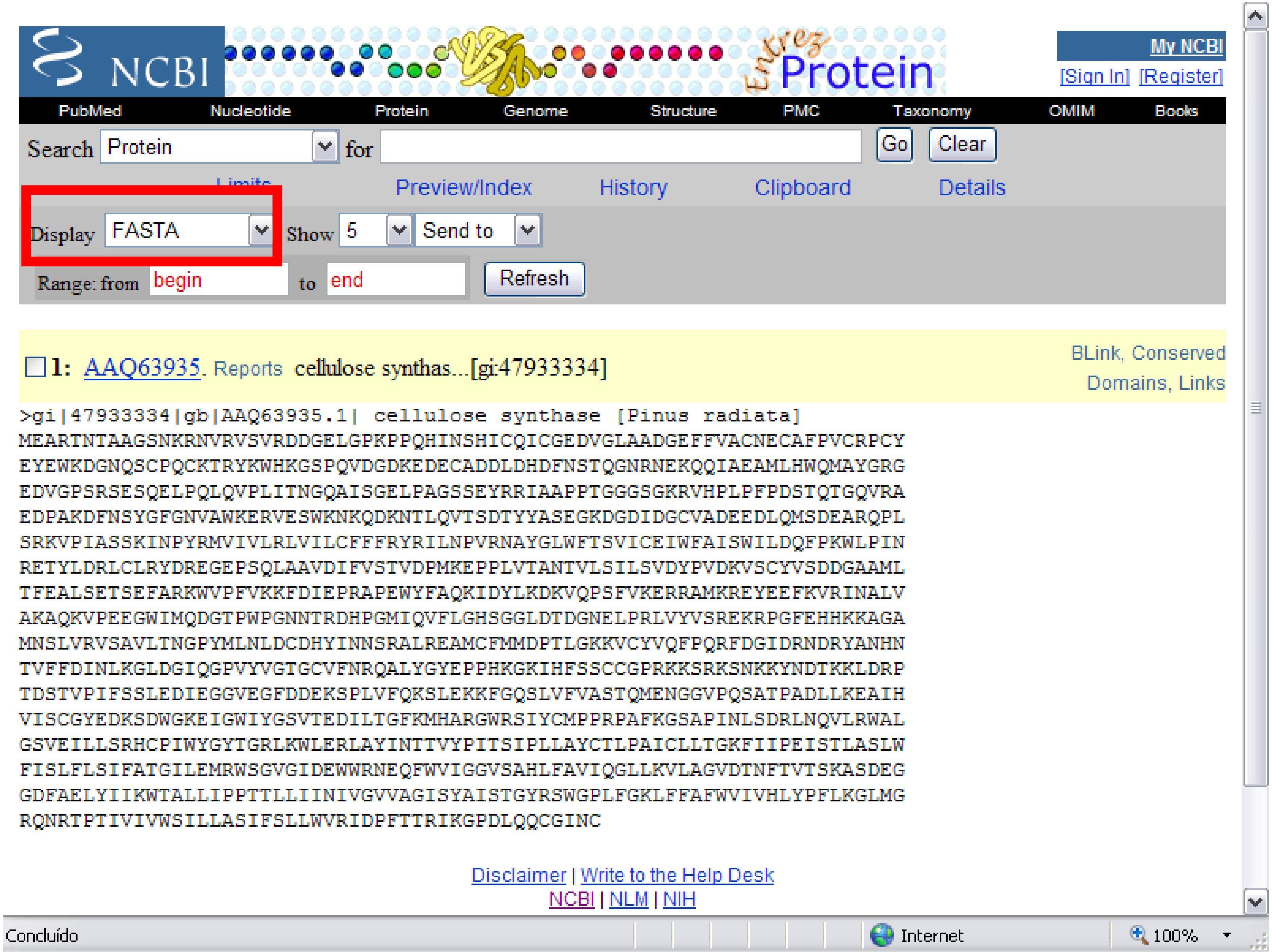
PROTEÍNAS

FORMATO FASTA

```
>gi|47933334|gb|AAQ63935.1| cellulose synthase [Pinus radiata]
MEARTNTAAGSNKRNVRVSVRDDGELGPKPPQHINSHICQICGEDV
GLAADGEFFVACNECAFPVCRPCYEYEWKDGNQSCPQCKTRYKWH
KGSPQVDGDKEDECADDLDHDFNSTQGNRNEKQQIAEAMLHWQM
AYGRGEDVGPSRSESQELPQLQVPLITNGQAISGELPAGSSEYRRIA
APPTGGGSGKRVHPLPFPDSTQTGQVRA
```

>LINHA DO NOME

MÁXIMA DE 80 CARACTERES POR LINHA



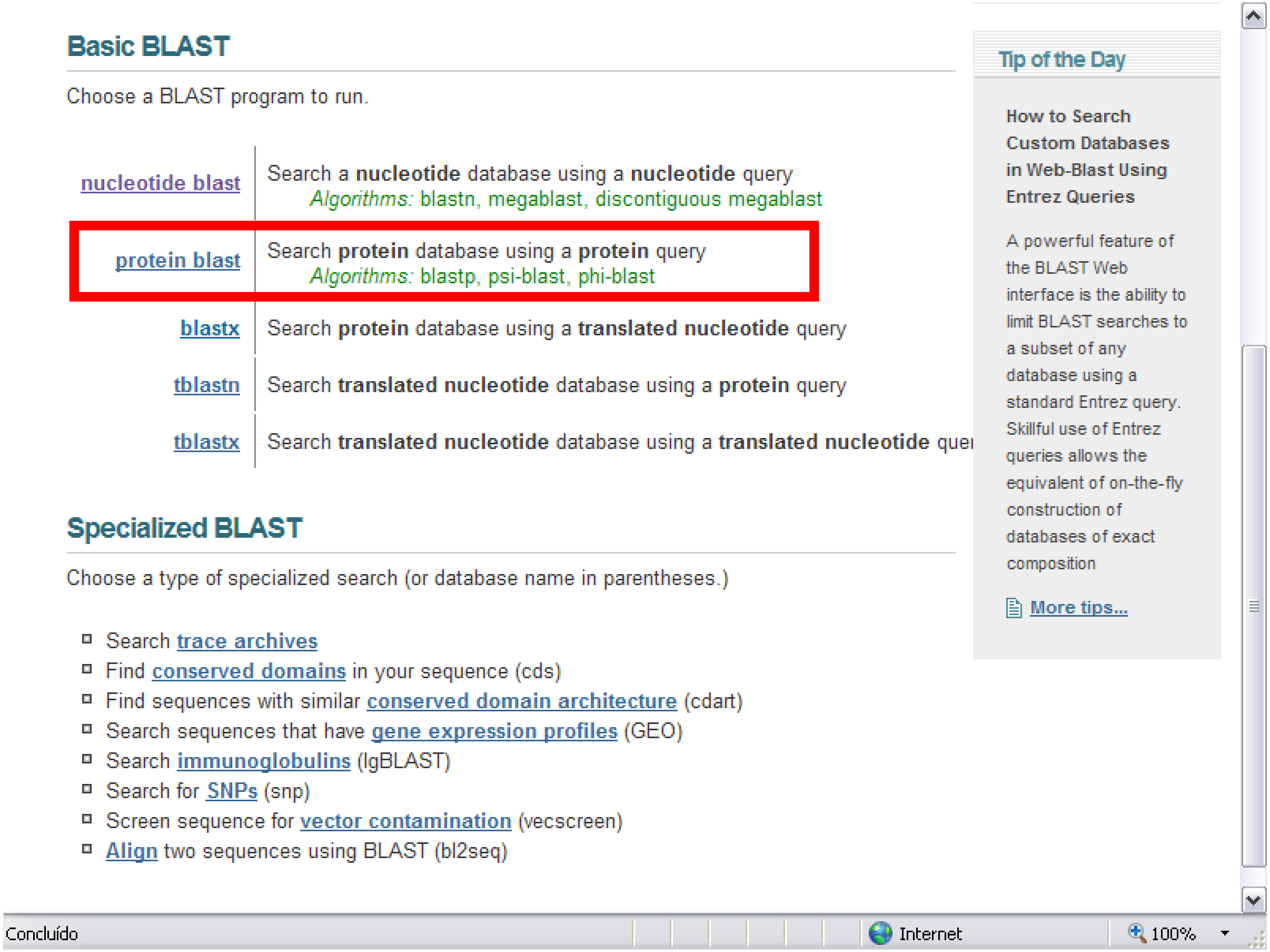
Search Protein for

Display FASTA Show 5

Range: from begin to end

☐ 1: AAQ63935. Reports cellulose synthas...[gi:47933334]

```
>gi|47933334|gb|AAQ63935.1| cellulose synthase [Pinus radiata]
MEARTNTAAGSNKRNVRVSVRDDGELGPKPPQHINSHICQICGEDVGLAADGEFFVACNECAFPVCRPCY
EYEWKDGNQSCPQCKTRYKWHKGSPQVDGDKEDECADDLDHDFNSTQGNRNEKQQIAEAMLHWQMAYGRG
EDVGPSRSESQELPQLQVPLITNGQAISGELPAGSSEYRRIAAPPTGGGSGKRVHPLPFPDSTQTGQVRA
EDPAKDFNSYGFGNVAWKERVESWKNKQDKNTLQVTSDTYYASEGKDGDIDGCVADEEDLQMSDEARQPL
SRKVPIASSKINPYRMVIVLRLVILCFFFRYRILNPVRNAYGLWFTSVICEIWFAISWILDQFPKWLPIN
RETYLDRLCLRYDREGEPSQLAAVDIFVSTVDPMKEPPLVTANTVLSILSVDYPVDKVSCYVSDDGAAML
TFEALSETSEFARKWVPFVKKFDIEPRAPEWYFAQKIDYLKDKVQPSFVKERRAMKREYEEFKVRINALV
AKAQKVPEEGWIMQDGTPWPGNNTRDHPGMIQVFLGHSGGLDTDGNELPRLVYVSREKRPGFEHHKKAGA
MNSLVRVSAVLTNGPYMLNLDCDHYINNSRALREAMCFMMDPTLGKKVCYVQFPQRFDGIDRNDRYANHN
TVFFDINLKGLDGIQGPVYVGTGCVFNRQALYGYEPPHKGKIHFSSCCGPRKKSRKSNKKYNDTKKLDRP
TDSTVPIFSSLEDIEGGVEGFDDEKSPLVFQKSLEKKFGQSLVFVASTQMENGGVPQSATPADLLKEAIH
VISCGYEDKSDWGKEIGWIYGSVTEDILTGFKMHARGWRSIYCMPPRPAFKGSAPINLSDRLNQVLRWAL
GSVEILLSRHCPIWYGYTGRLKWLERLAYINTTVYPITSIPLLAYCTLPAICLLTGKFIIPEISTLASLW
FISLFLSIFATGILEMRWSGVGIDEWWRNEQFWVIGGVSAHLFAVIQGLLKVLAGVDTNFTVTSKASDEG
GDFAELYIIKWTALLIPPTTLLIINIVGVVAGISYAISTGYRSWGPLFGKLFFAFWVIVHLYPFLKGLMG
RQNRTPTIVIVWSILLASIFSLLWVRIDPFTTRIKGPDLQQCGINC
```



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.)

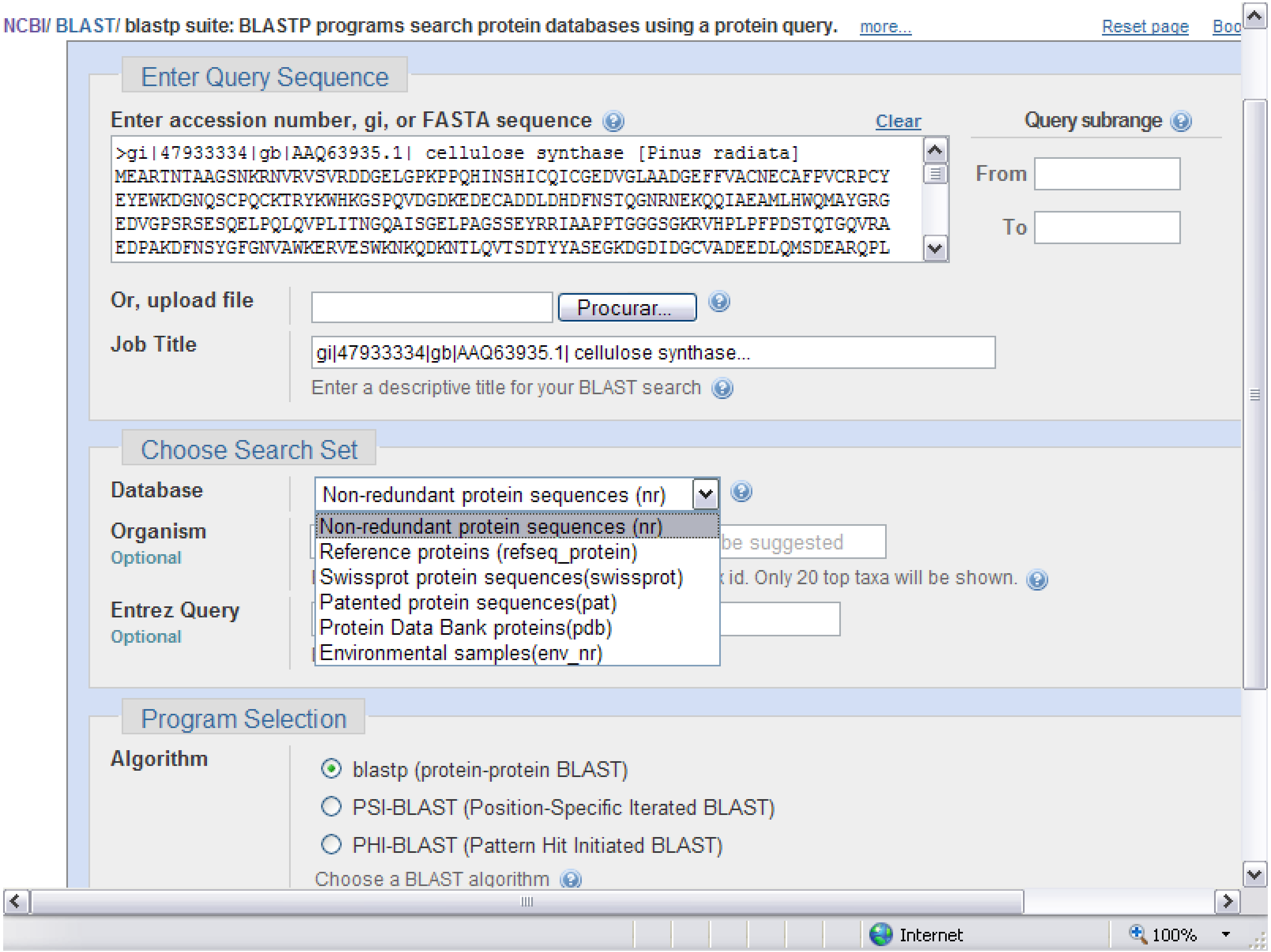
- Search trace archives
- Find conserved domains in your sequence (cds)
- Find sequences with similar conserved domain architecture (cdart)
- Search sequences that have gene expression profiles (GEO)
- Search immunoglobulins (IgBLAST)
- Search for SNPs (snp)
- Screen sequence for vector contamination (vecscreen)
- Align two sequences using BLAST (bl2seq)

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...



NCBI/ BLAST/ blastp suite: BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number, gi, or FASTA sequence

```
>gi|47933334|gb|AAQ63935.1| cellulose synthase [Pinus radiata]
MEARTNTAAGSNKRNVRVSVRDDGELGPKPPQHINSHICQICGEDVGLAADGEFFVACNECAFPVCRPCY
EYEWKDGNQSCPQCKTRYKWHKGSPQVDGDKEDECADDLDHDFNSTQGNRNEKQQIAEAMLHWQMAYGRG
EDVGPSRSESQELPQLQVPLITNGQAISGELPAGSSEYRRIAAPPTGGGSGKRVHPLPFPDSTQTGQVRA
EDPAKDFNSYGFGNVAWKERVESWKNKQDKNTLQVTSDTYYASEGKDGDIDGCVADEEDLQMSDEARQPL
```

Query subrange

From

To

Or, upload file

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)

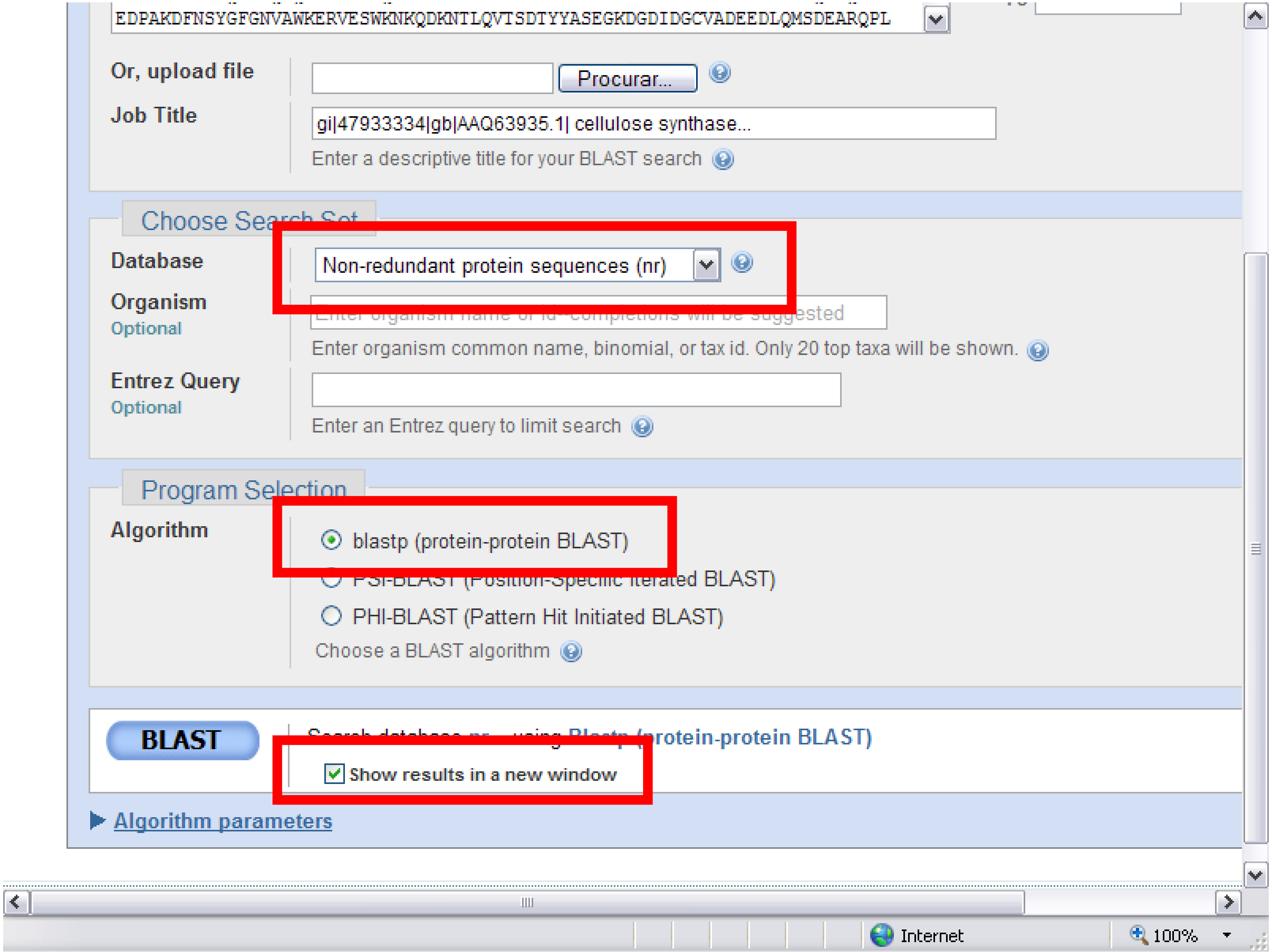
Organism
Optional
be suggested
id. Only 20 top taxa will be shown.

Entrez Query
Optional

Program Selection

Algorithm
● blastp (protein-protein BLAST)
○ PSI-BLAST (Position-Specific Iterated BLAST)
○ PHI-BLAST (Pattern Hit Initiated BLAST)

Choose a BLAST algorithm



```
EDPAKDFNSYGFGNVAWKERVESWKNKQDKNTLQVTSDTYYASEGKDGDIDGCVADEEDLQMSDEARQPL
```

Or, upload file

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

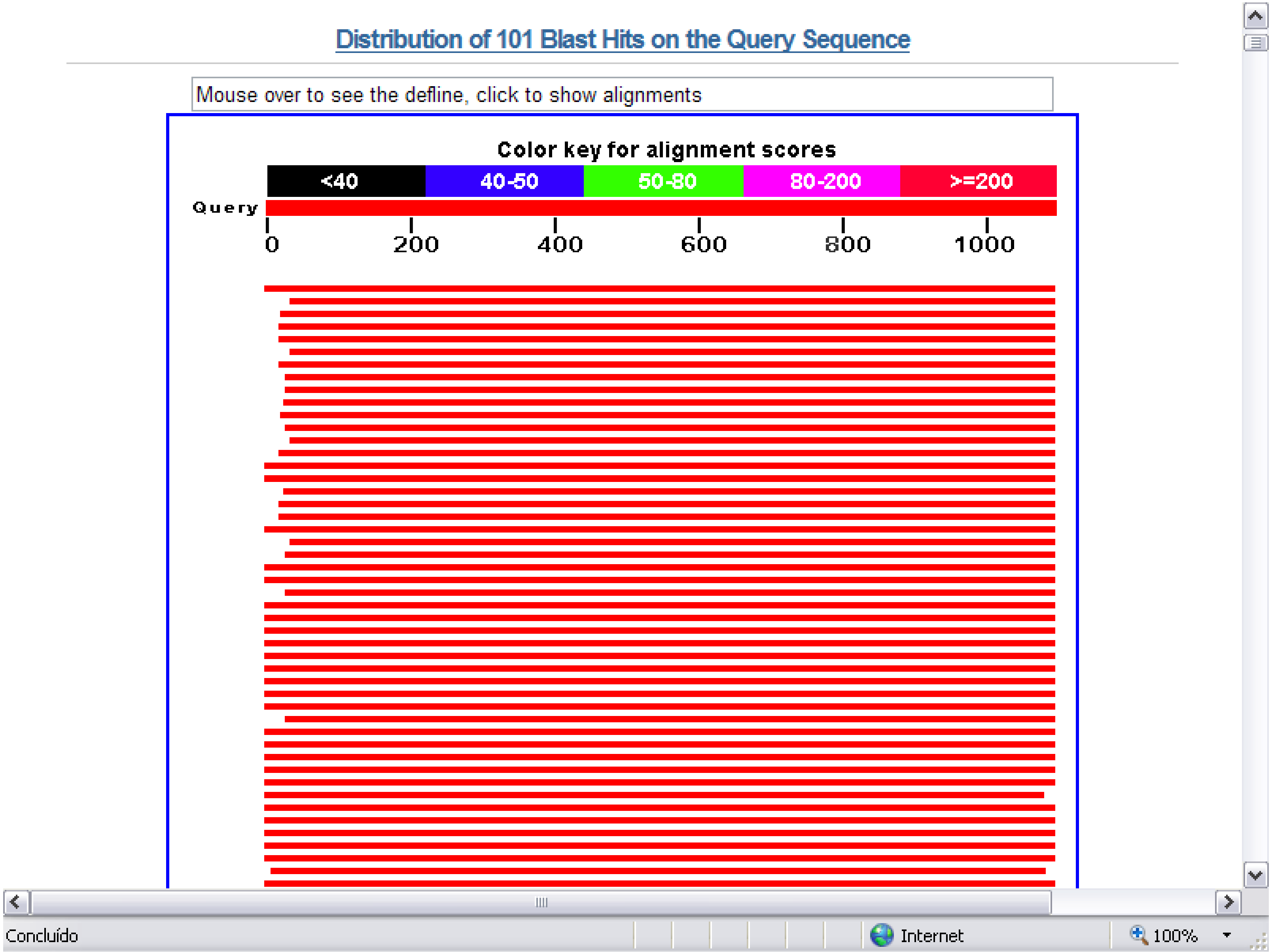
- ● blastp (protein-protein BLAST)
- ○ PSI-BLAST (Position-Specific Iterated BLAST)
- ○ PHI-BLAST (Pattern Hit Initiated BLAST)

Choose a BLAST algorithm

Search database nr using Blastp (protein-protein BLAST)

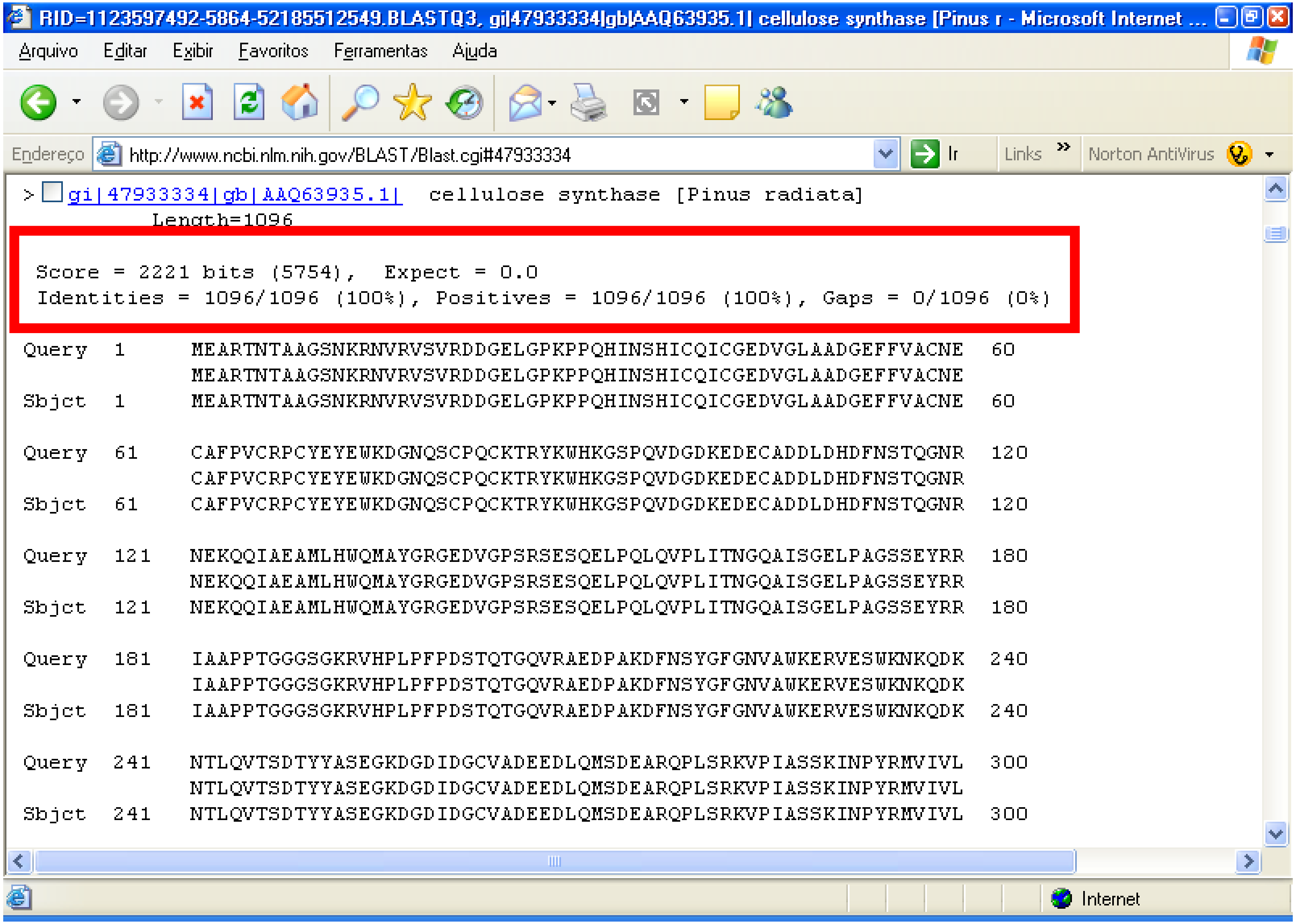
☑ Show results in a new window

Algorithm parameters



Distribution of 101 Blast Hits on the Query Sequence

Mouse over to see the defline, click to show alignments



```
> ☐ 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%)

Query  1     MEARTNTAAGSNKRNVRVSVRDDGELGPKPPQHINSHICQICGEDVGLAADGEFFVACNE  60
             MEARTNTAAGSNKRNVRVSVRDDGELGPKPPQHINSHICQICGEDVGLAADGEFFVACNE
Sbjct  1     MEARTNTAAGSNKRNVRVSVRDDGELGPKPPQHINSHICQICGEDVGLAADGEFFVACNE  60

Query  61    CAFPVCRPCYEYEWKDGNQSCPQCKTRYKWHKGSPQVDGDKEDECADDLDHDFNSTQGNR  120
             CAFPVCRPCYEYEWKDGNQSCPQCKTRYKWHKGSPQVDGDKEDECADDLDHDFNSTQGNR
Sbjct  61    CAFPVCRPCYEYEWKDGNQSCPQCKTRYKWHKGSPQVDGDKEDECADDLDHDFNSTQGNR  120

Query  121   NEKQQIAEAMLHWQMAYGRGEDVGPSRSESQELPQLQVPLITNGQAISGELPAGSSEYRR  180
             NEKQQIAEAMLHWQMAYGRGEDVGPSRSESQELPQLQVPLITNGQAISGELPAGSSEYRR
Sbjct  121   NEKQQIAEAMLHWQMAYGRGEDVGPSRSESQELPQLQVPLITNGQAISGELPAGSSEYRR  180

Query  181   IAAPPTGGGSGKRVHPLPFPDSTQTGQVRAEDPAKDFNSYGFGNVAWKERVESWKNKQDK  240
             IAAPPTGGGSGKRVHPLPFPDSTQTGQVRAEDPAKDFNSYGFGNVAWKERVESWKNKQDK
Sbjct  181   IAAPPTGGGSGKRVHPLPFPDSTQTGQVRAEDPAKDFNSYGFGNVAWKERVESWKNKQDK  240

Query  241   NTLQVTSDTYYASEGKDGDIDGCVADEEDLQMSDEARQPLSRKVPIASSKINPYRMVIVL  300
             NTLQVTSDTYYASEGKDGDIDGCVADEEDLQMSDEARQPLSRKVPIASSKINPYRMVIVL
Sbjct  241   NTLQVTSDTYYASEGKDGDIDGCVADEEDLQMSDEARQPLSRKVPIASSKINPYRMVIVL  300
```

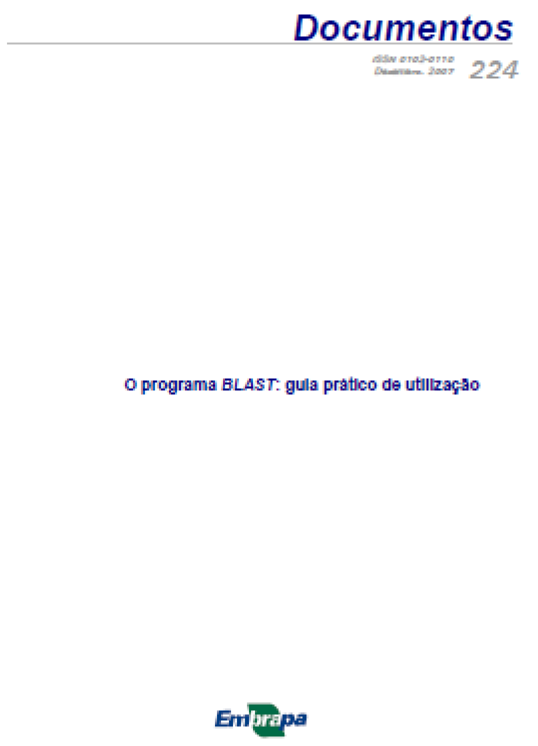
VAMOS PRATICAR…

```
> oquesera?
CAGTACACTCTCAATGGTTCGCGGCTGGTTCAACATACGCCCAAAGCGGCGTCGTCGACCTGCTAC
CAAGACCTGTATTGGAGAATTCCTCAAACTGGATGTGTTCAGTGTTTGGCCGATCAGTTTTCCAAGC
AGACCATCCTTCGGGATGAATTATGTTGGAAAAGTCACAAAATTGATAGGTGACAGCTGCGTATGA
CGACCAAGGTCGTCCAAGGTATGTTGTGTGCGGTGGAGTTCCATTAAGCCCCAGAATTCTGCATTT
ATTGAATACAAACCTTGTCTGAACTGTAGTAGCCACTTTTTTGAGCCGTGATACTTGAAGATTTTGAT
AATTCATAGTATTAGTTAGGATTCGGAAAAAAAAATCCAGAATGATCAAGTTAAACCTTAGTTCAGG
CTTCAAAAGATGACATGAGAGAAAAAGATATCCAATCATGCAAATTTGACCCACGTGCTAGAGTGG
AGCGGCTTTTTGACTGCAATTTGTGCCTTTTCAAACCAAGCCGTCCCAGGACCACAAATGAAATCT
TCAAGGATGATGGATTTCAGGTTGGAATAGAAAACGAAATTGAATATCTTACTTGGGTAAGCCTAA
ACCACTTACCAACAGTTCCTTGTATCAGGCATTTGTTAAAGTAGGCTCGAGTTGCTTTGATATACAAC
GTATCTTGGTAACTGCTAAAAACGCAGGCTTTGAAGACTGGGTGTTTAGAAGAGACAGCCAAGAT
TTGGGTTATTGAGAAAGAATAGGTTGAATCATCTTTAGGAGAAAGTACAGTATAATTACCCTGTTGT
TCTCCTTCAGCAGTTAAGGCTACTGCTTGAGTATCACTTCCCGAACCGAAAGAATTGATCACGTTAC
TACGTTACGATTTTAAATGGTGTTTTTCGAAAAAAATCCGTCAACGCTGTTATAATATGTAACGAATT
TCTTGTAAACCCAAGGCTCAAACGCCACTCACACATTGCGCATTTCAAAGCTTTCAACGTTTCGCAA
AAATTAGTATCCATTGAAAAGAATTTGTCCAAAACCATAAGTTTTAAGCACTCGTTACGCTTTTTCAG
CATCAAGAAGACGAAGCGCGTTGTAGCCTACCAGGAGCTTTTGGCCCACAACACTGAGCTTTTTG
AATCAGAACCTGTAGAACCCAGGCTCCGAGCCTCTGTGATAAACACCGTATTCCGAACAAAGGATC
GGGTGCTTGGTGAGAATCCTTGAATTAACAACGGCTGGAAAATTTGAGAAAA
```

VAMOS PRATICAR…

```
>oqueserá?
MVYGSVGFFPVMATPASRMVQKSKSAMVFWNISSQRSNMNRPTHFDHISMTTPSGERIDRLPAWI
RKYLAAVFLSCPNIFFLKMFYLMIGQACHSGSICLASLCDLLGPSTKISMEKYPSTFGFSGLSTWARNAV
FTITRISSFGFLFNSTFPSGYFNPGSSRKLQRIYRQRSSSYQSFGLRNSDEWVQSLQMIFANQLLFGHDS
GHHGTRLLCFWVSSYKLLLHQSLWLVSDEFSKKYMRRALYISWLVPFFFAGTPEDLMELIDTAHGMGL
TVLLDVVHSHACKNVLDGINMFDGTDHCYFHEGGKGRHDLWDRYVLNNLEWTHQRYYSVSALYSV
VCLIMVIMKFDSYFPIDFSWSSTS
```

BIBLIOGRAFIA RECOMENDADA

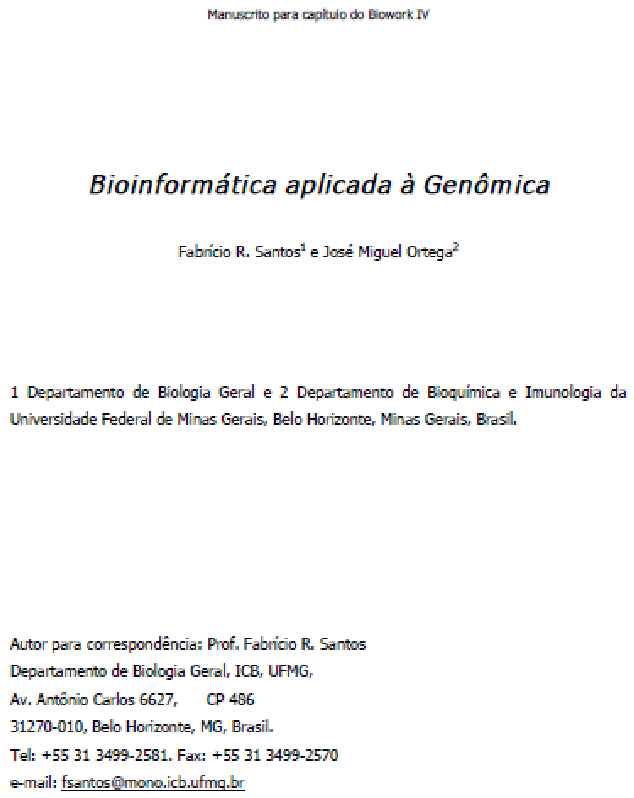


Documentos

224

O programa *BLAST*: guia prático de utilização

Embrapa



Manuscrito para capítulo do Biowork IV

Bioinformática aplicada à Genômica

Fabrício R. Santos[1] e José Miguel Ortega[2]

1 Departamento de Biologia Geral e 2 Departamento de Bioquímica e Imunologia da Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brasil.

Autor para correspondência: Prof. Fabrício R. Santos
Departamento de Biologia Geral, ICB, UFMG,
Av. Antônio Carlos 6627, CP 486
31270-010, Belo Horizonte, MG, Brasil.
Tel: +55 31 3499-2581. Fax: +55 31 3499-2570
e-mail: fsantos@mono.icb.ufmg.br

ESTUDO DIRIGIDO

1. O que fazer com uma sequência de DNA?
2. Bancos de dados públicos e internacionais: NCBI, EMBL, DDBJ;
3. Definição de Bioinformática;
4. Análise da sequência no NCBI;
5. Busca de sequências por similaridade;
6. BLAST e Banco de dados de sequências.