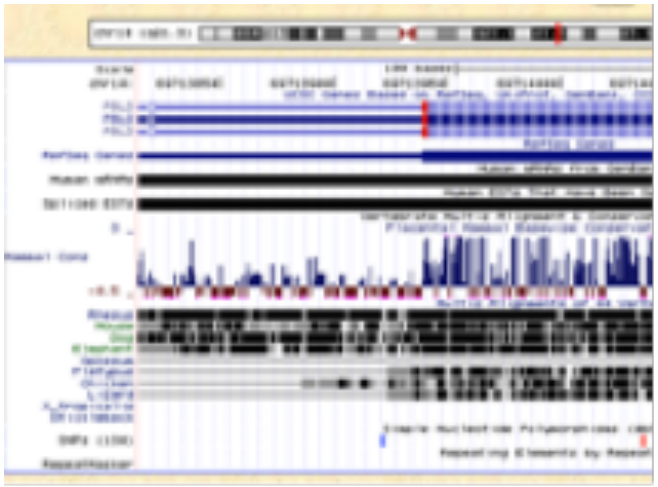
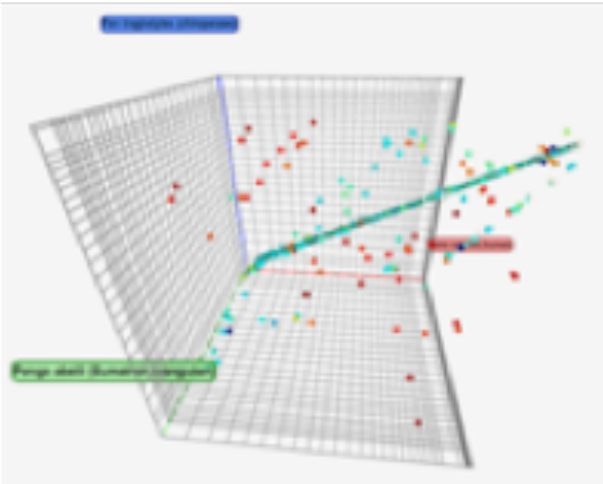
Computational Genomics

Introduction
to
Biological Sequence Analysis





NATIONAL HUMAN GENOME RESEARCH INSTITUTE Division of Intramural Research

Current Topics in Genome Analysis 2016

Week 1: Biological Sequence Analysis I

Andy Baxevanis, Ph.D.

Current Topics in Genome Analysis 2016

Andy Baxevanis, Ph.D.

No Relevant Financial Relationships with Commercial Interests

Sequence Alignments: Determining Similarity and Deducing Homology

Why construct sequence alignments?

- Provide a measure of relatedness between nucleotide or amino acid sequences
- Determining relatedness allows one to draw biological inferences regarding
  - structural relationships
  - functional relationships
  - evolutionary relationships
- Important to use correct terminology when describing phylogenetic relationships

Defining the Terms

- The quantitative measure: ***Similarity***
  - Always based on an observable
  - Usually expressed as percent identity
  - Quantify changes that occur as two sequences diverge (substitutions, insertions, or deletions)
  - Identify residues crucial for maintaining a protein's structure or function

- High degrees of sequence similarity *might* imply
  - a common evolutionary history
  - possible commonality in biological function

Defining the Terms

The conclusion: ***Homology***

- ***Homology***: Implies an evolutionary relationship
- ***Homologs***: Genes that have arisen from a common ancestor
- Genes either *are* or *are not* homologous
  (not measured in degrees)

> It is worth repeating here that homology, like pregnancy, is indivisible[8]. You either are homologous (pregnant) or you are not. Thus, if what one means to assert is that 80% of the character states are identical one should speak of 80% identity, and not 80% homology.

Fitch, Trends Genet. 16: 227-231, 2000

Defining the Terms

Orthologs: Genes that diverged as a result of a speciation event

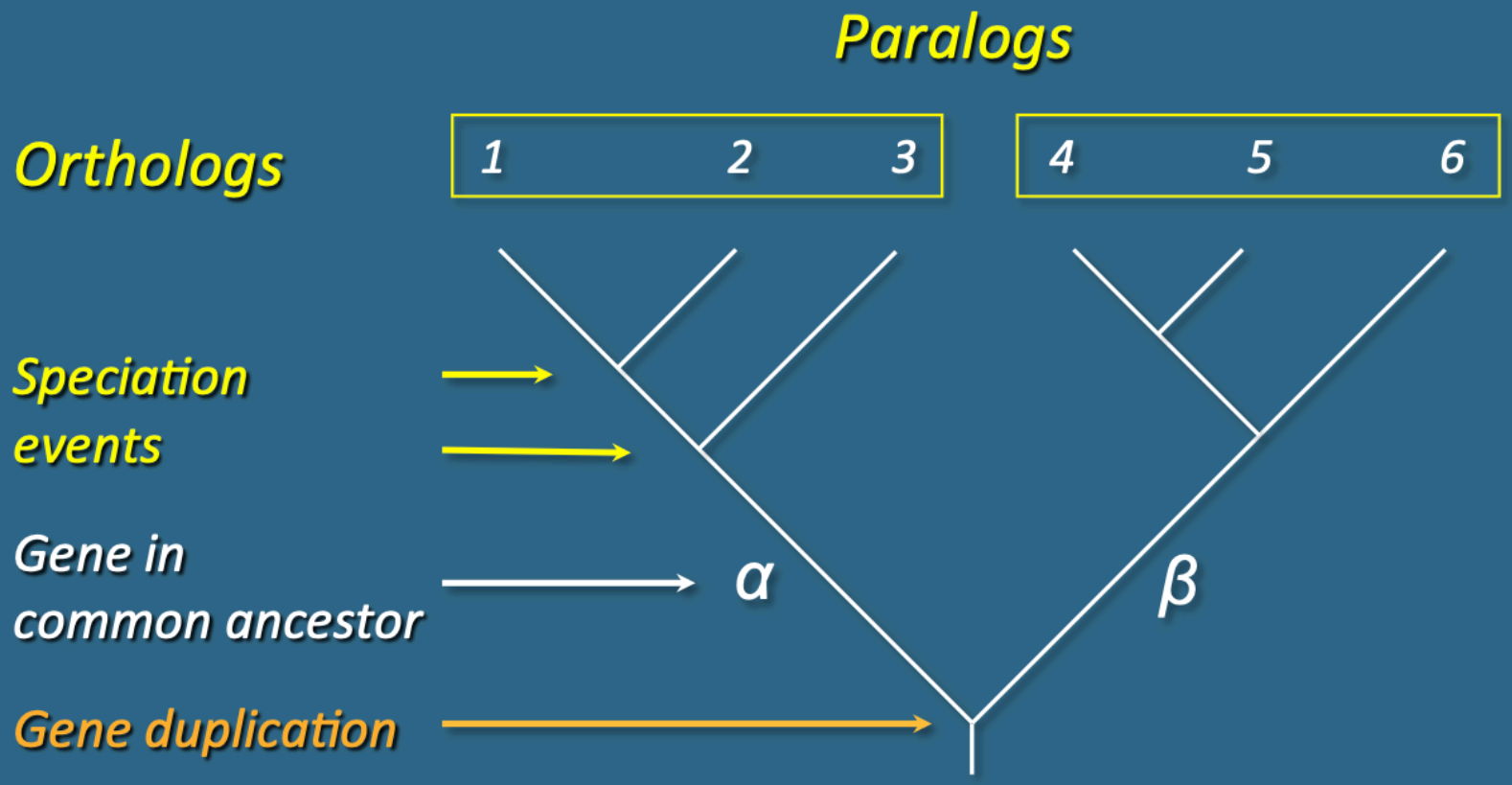
- Sequences are direct descendants of a sequence in a common ancestor (share a common origin)
- Most likely have similar domain and three-dimensional structure
- Usually retain same biological function over evolutionary time
- Can be used to predict gene function in novel genomes

Defining the Terms

Paralogs: Genes that arose by the duplication of a single gene in a particular lineage

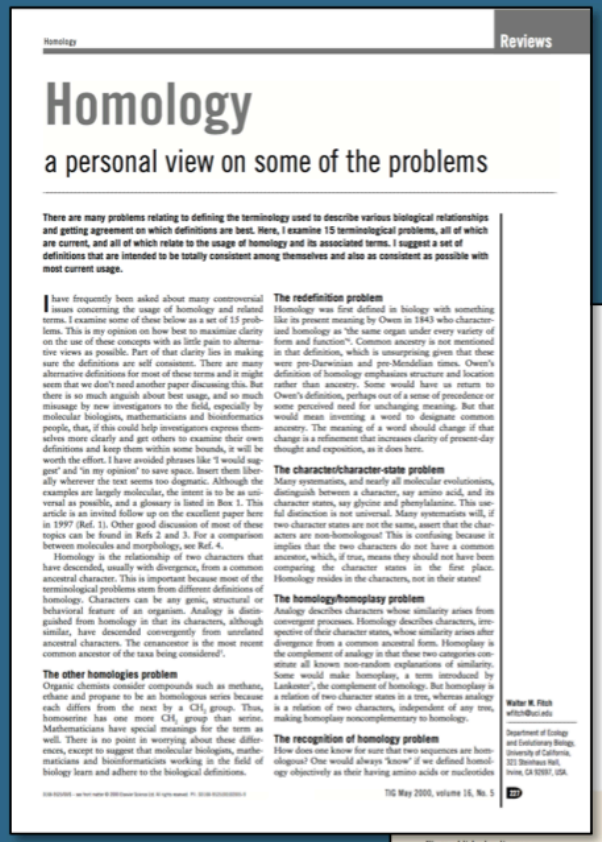
- Perhaps less likely to perform similar functions
- Can take on new functions over evolutionary time
- Provides insight into 'evolutionary innovation'

Defining the Terms



- Genes 1-3 are orthologous
- Genes 4-6 are orthologous
- Any pair of *α* and *β* genes are paralogous (genes related through a gene duplication event)

Orthology and Paralogy: Further Reading



Homology

Reviews

Homology

a personal view on some of the problems

There are many problems relating to defining the terminology used to describe various biological relationships and getting agreement on which definitions are best. Here, I examine 15 terminological problems, all of which are current, and all of which relate to the usage of homology and its associated terms. I suggest a set of definitions that are intended to be totally consistent among themselves and also as consistent as possible with most current usage.

I have frequently been asked about many controversial issues concerning the usage of homology and related terms. I examine some of these below as a set of 15 problems. This is my opinion on how best to maximize clarity on the use of these concepts with as little pain to alternative views as possible. Part of that clarity lies in making sure the definitions are self consistent. There are many alternative definitions for most of these terms and it might seem that we don't need another paper discussing this. But there is so much anguish about best usage, and so much misusage by new investigators to the field, especially by molecular biologists, mathematicians and bioinformatics people, that, if this could help investigators express themselves more clearly and get others to examine their own definitions and keep them within some bounds, it will be worth the effort. I have avoided phrases like 'I would suggest' and 'in my opinion' to save space. Insert them liberally wherever the text seems too dogmatic. Although the examples are largely molecular, the intent is to be as universal as possible, and a glossary is listed in Box 1. This article is an invited follow up on the excellent paper here in 1997 (Ref. 1). Other good discussion of most of these topics can be found in Refs 2 and 3. For a comparison between molecules and morphology, see Ref. 4.

Homology is the relationship of two characters that have descended, usually with divergence, from a common ancestral character. This is important because most of the terminological problems stem from different definitions of homology. Characters can be any genic, structural or behavioral feature of an organism. Analogy is distinguished from homology in that its characters, although similar, have descended convergently from unrelated ancestral characters. The cenancestor is the most recent common ancestor of the taxa being considered[1].

The other homologies problem

Organic chemists consider compounds such as methane, ethane and propane to be an homologous series because each differs from the next by a CH_2 group. Thus, homoserine has one more CH_2 group than serine. Mathematicians have special meanings for the term as well. There is no point in worrying about these differences, except to suggest that molecular biologists, mathematicians and bioinformaticists working in the field of biology learn and adhere to the biological definitions.

The redefinition problem

Homology was first defined in biology with something like its present meaning by Owen in 1843 who characterized homology as "the same organ under every variety of form and function"[5]. Common ancestry is not mentioned in that definition, which is unsurprising given that these were pre-Darwinian and pre-Mendelian times. Owen's definition of homology emphasizes structure and location rather than ancestry. Some would have us return to Owen's definition, perhaps out of a sense of precedence or some perceived need for unchanging meaning. But that would mean inventing a word to designate common ancestry. The meaning of a word should change if that change is a refinement that increases clarity of present-day thought and exposition, as it does here.

The character/character-state problem

Many systematists, and nearly all molecular evolutionists, distinguish between a character, say amino acid, and its character states, say glycine and phenylalanine. This useful distinction is not universal. Many systematists will, if two character states are not the same, assert that the characters are non-homologous! This is confusing because it implies that the two characters do not have a common ancestor, which, if true, means they should not have been comparing the character states in the first place. Homology resides in the characters, not in their states!

The homology/homoplasy problem

Analogy describes characters whose similarity arises from convergent processes. Homology describes characters, irrespective of their character states, whose similarity arises after divergence from a common ancestral form. Homoplasy is the complement of analogy in that these two categories constitute all known non-random explanations of similarity. Some would make homoplasy, a term introduced by Lankester[2], the complement of homology. But homoplasy is a relation of two character states in a tree, whereas analogy is a relation of two characters, independent of any tree, making homoplasy noncomplementary to homology.

The recognition of homology problem

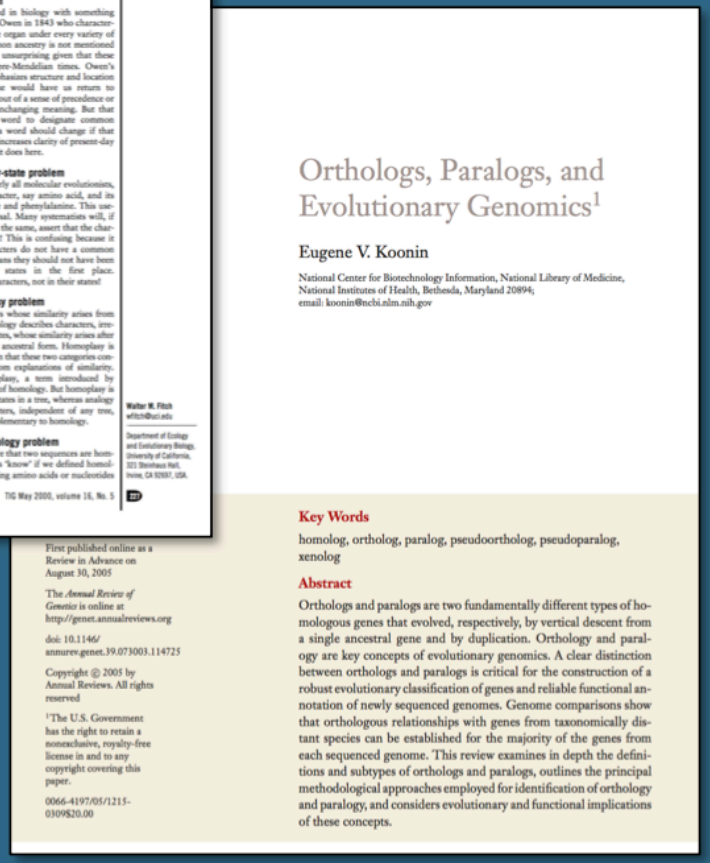
How does one know for sure that two sequences are homologous? One would always 'know' if we defined homology objectively as their having amino acids or nucleotides

Walter M. Fitch
wfitch@uci.edu

Department of Ecology and Evolutionary Biology, University of California, 321 Steinhaus Hall, Irvine, CA 92697, USA.

TIG May 2000, volume 16, No. 5

Walter Fitch
Trends Genet.
16: 227-231, 2000



Orthologs, Paralogs, and Evolutionary Genomics[1]

Eugene V. Koonin

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894; email: koonin@ncbi.nlm.nih.gov

First published online as a Review in Advance on August 30, 2005

The *Annual Review of Genetics* is online at http://genet.annualreviews.org

doi: 10.1146/annurev.genet.39.073003.114725

[1]The U.S. Government has the right to retain a nonexclusive, royalty-free license in and to any copyright covering this paper.

0066-4197/05/1215-0309$20.00

Key Words

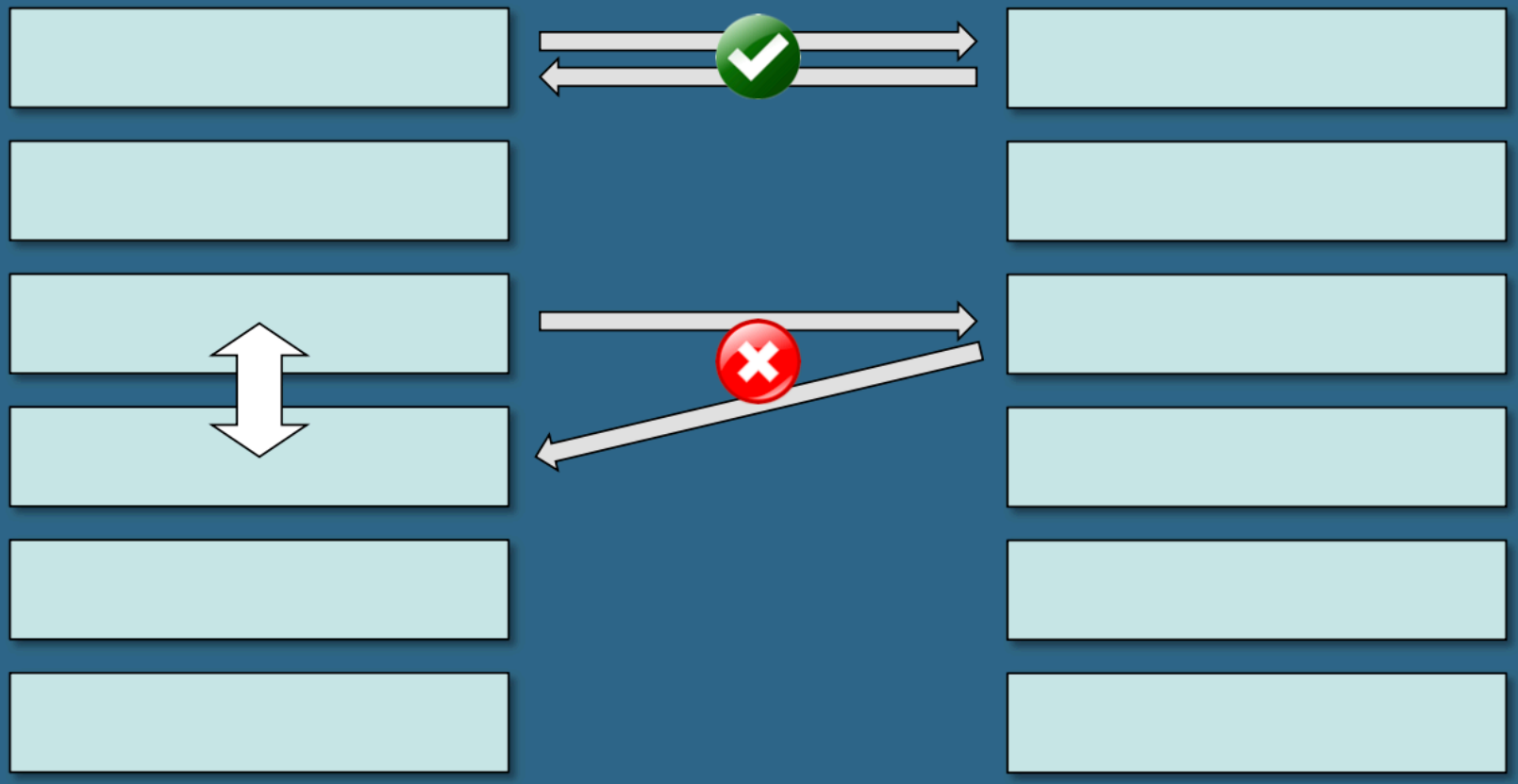
homolog, ortholog, paralog, pseudoortholog, pseudoparalog, xenolog

Abstract

Orthologs and paralogs are two fundamentally different types of homologous genes that evolved, respectively, by vertical descent from a single ancestral gene and by duplication. Orthology and paralogy are key concepts of evolutionary genomics. A clear distinction between orthologs and paralogs is critical for the construction of a robust evolutionary classification of genes and reliable functional annotation of newly sequenced genomes. Genome comparisons show that orthologous relationships with genes from taxonomically distant species can be established for the majority of the genes from each sequenced genome. This review examines in depth the definitions and subtypes of orthologs and paralogs, outlines the principal methodological approaches employed for identification of orthology and paralogy, and considers evolutionary and functional implications of these concepts.

Eugene Koonin
Annu. Rev. Genet.
39: 309-338, 2005

Identifying Candidate Orthologs: Reciprocal Best Hits



Organism 1

Organism 2

Global Sequence Alignments

- Sequence comparison along the entire length of the two sequences being aligned
- Best for highly-similar sequences of similar length
- As the degree of sequence similarity declines, global alignment methods tend to miss important biological relationships

Local Sequence Alignments

- Sequence comparison intended to find the most similar regions in the two sequences being aligned ('paired subsequences')
- Regions outside the area of local alignment are excluded
- More than one local alignment could be generated for any two sequences being compared
- Best for sequences that share some similarity, or for sequences of different lengths

Scoring Matrices: Construction and Proper Selection

Scoring Matrices

- Empirical weighting scheme representing physicochemical and biological characteristics of nucleotides and amino acids
  - Side chain structure and chemistry
  - Side chain function
- Amino acid-based examples of considerations:
  - Cys/Pro are important for structure and function
  - Trp has a bulky side chain
  - Lys/Arg have positively charged side chains

Scoring Matrices

- ***Conservation:*** What residues can substitute for another residue and not adversely affect the function of the protein?
  - `Ile/Val` - both small and hydrophobic
  - `Ser/Thr` - both polar
  - *Conserve charge, size, hydrophobicity, additional physicochemical factors*
- ***Frequency:*** How often does a particular residue occur amongst the entire constellation of proteins?

Scoring Matrices

Why is understanding scoring matrices important?

- Appear in all analyses involving sequence comparison
- Implicitly represent particular evolutionary patterns
- Choice of matrix can strongly influence outcomes of analyses

Matrix Structure: Nucleotides

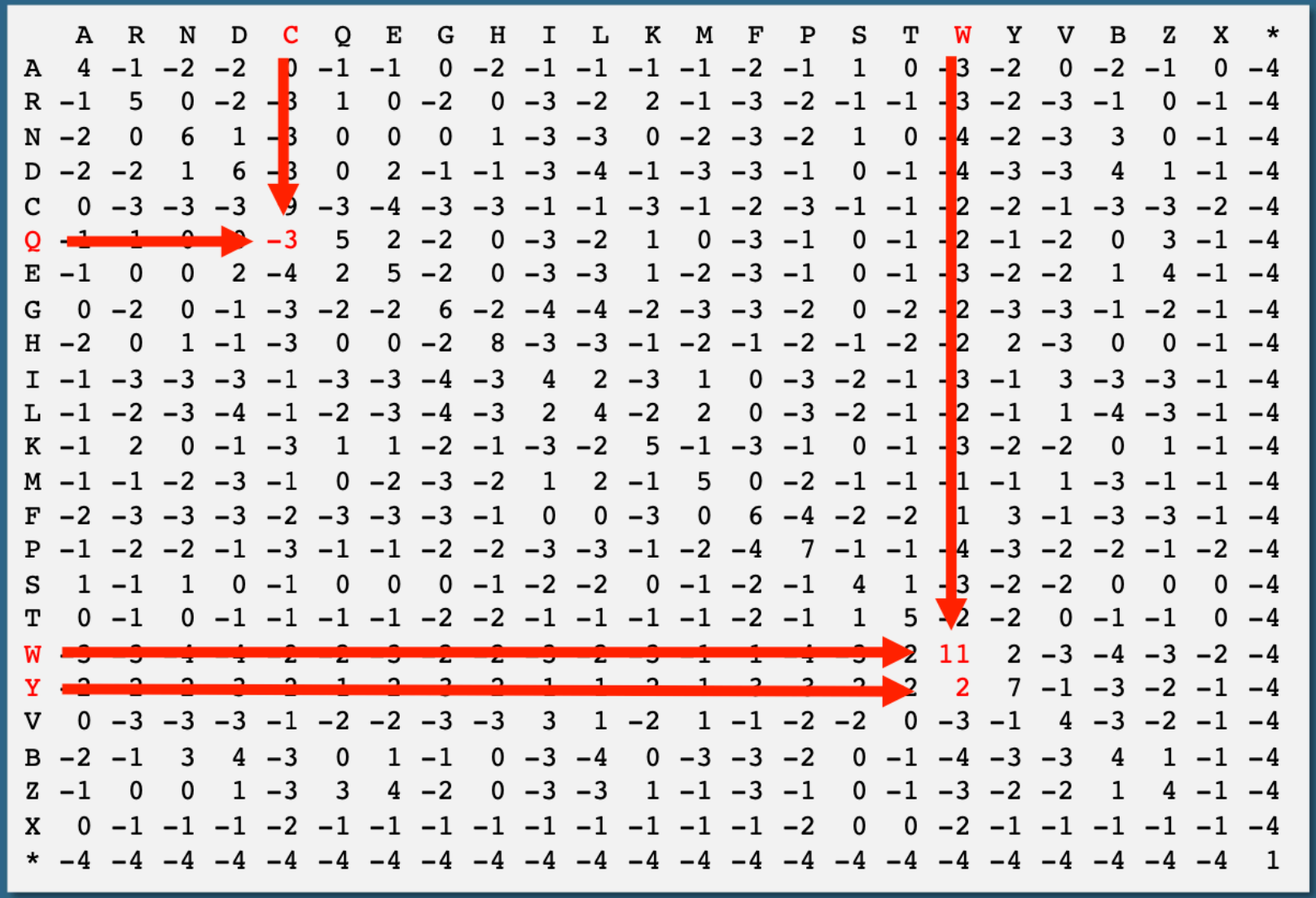
- Simple match/mismatch scoring scheme:

  Match +2
  Mismatch −3

- Assumes each nucleotide occurs 25% of the time

	A	T	G	C
A	2	−3	−3	−3
T	−3	2	−3	−3
G	−3	−3	2	−3
C	−3	−3	−3	2

Matrix Structure: Proteins



	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0	-4
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1	-4
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1	-4
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1	-1	-4
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-3	-2	-4
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	3	-1	-4
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-2	-1	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	0	0	-1	-4
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	-3	-1	-4
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4	-3	-1	-4
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2	0	1	-1	-4
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-3	-1	-1	-4
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-3	-3	-1	-4
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2	-2	-1	-2	-4
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2	0	0	0	-4
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0	-1	-1	0	-4
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3	-4	-3	-2	-4
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1	-3	-2	-1	-4
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	-3	-2	-1	-4
B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4	1	-1	-4
Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	-4
*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

BLOSUM62

BLOSUM Matrices

- Look only for differences in conserved, ungapped regions of a protein family ('blocks')
- Directly calculated based on local alignments
  - Substitution probabilities (*conservation*)
  - Overall *frequency* of amino acids
- Sensitive to detecting structural or functional substitutions
- Generally perform better than PAM matrices for local similarity searches *(Henikoff and Henikoff, 1993)*
- BLOSUM series can be used to identify both closely and distantly related sequences

BLOSUM *n*

- Built using sequences sharing no more than *n*% identity
- Contribution of sequences > *n*% identical clustered and replaced by a sequence that represents the cluster

BLOSUM *n*

- Clustering reduces contribution of closely related sequences (less bias towards substitutions that occur in the most closely related members of a family)
- Reducing *n* yields more distantly related sequences
- Increasing *n* yields more closely related sequences

Which one to choose?

BLOSUM		% Similarity
90	Short alignments, highly similar	70-90
80	Best for detecting known members of a protein family	50-60
62	Most effective in finding all potential similarities	30-40
30	Longer, weaker local alignments	< 30

The takeaway…

No single matrix is the complete answer for all sequence comparisons

David Wheeler
Curr. Protoc. Bioinformatics
3.5.1 – 3.5.6, 2003

Selecting the Right Protein-Scoring Matrix

UNIT 3.5

OVERVIEW

Every program for searching protein sequences against a database includes a choice of a "protein-scoring matrix," also called a "weight matrix." Weight matrices add sensitivity to the search, while statistical significance adds selectivity (see *UNIT 4.1*). Virtually every user chooses the default, typically PAM 250 or BLOSUM62. Despite the fact that the choice of matrix can strongly influence the outcome of the analysis, most users do not know why a particular matrix should be used. In general, scoring matrices implicitly represent a particular theory of protein sequence evolution. This unit provides guidance in the choice of a scoring matrix, as understanding the assumptions underlying the PAM and BLOSUM scoring matrices can aid in making the proper choice. The selection of PAM matrices is covered first, after which the selection of BLOSUM matrices is discussed, and finally a brief overview of the wide variety of specialized scoring matrices is provided.

PAM MATRICES

PAM, a rearranged acronym derived from Accepted Point Mutation (Dayhoff, 1978) is a probabilistic model for amino acid replacement derived by comparing the frequencies of replacement in closely related sequences to the frequency expected from the completely random replacement of amino acids. The basis of this scoring system is the observation that the evolution of protein sequences is a nonrandom process—i.e., some amino acid replacements occur much more frequently than others, especially in related sequences. Amino acid substitutions tend to conserve charge, size, and hydrophobicity among other characteristics. One would expect that the substitution of glycine for alanine (CH_3 versus H) would have less of an effect on a protein's structure and function than the substitution of alanine for threonine (CH_3 versus substituted indole ring). The inference is that if two aligned sequences manifest a higher than expected prevalence of these characteristic replacements, the sequences are related. An excellent discussion of the derivation and use of the PAM matrices is given in George et al. (1990).

PAM matrices are the result of computing the probability of one substitution per 100 amino acids, called the PAM 1 matrix. Higher PAM matrices are derived by multiplying the PAM 1 matrix by itself a defined number of times. Thus, a PAM 160 matrix is the result of performing 160 matrix multiplications of the PAM 1 matrix against itself. Similarly, the PAM 250 matrix is derived by multiplying the PAM 1 matrix against itself 250 times.

Biologically, the PAM 50 matrix means that in 100 amino acids there have been 50 substitutions, while the PAM 250 matrix means there have been 2.5 amino acid replacements at each site (see *UNIT 3.1* regarding insertions and deletions). This sounds unusual, but remember that over evolutionary time, it is possible that an alanine was changed to a glycine, then to a valine, and then back to an alanine. These silent substitutions are derived from observed amino acid frequency data in protein families and superfamilies.

Choosing a PAM Matrix

It is extremely important to note that PAM matrixes are derived from protein sequence data available in the late 1960s and early 1970s. Most proteins known at that time were small, globular, and hydrophilic. If the researcher believes their protein contains substantial hydrophobic regions, such as membrane-spanning helices or sheets, the PAM matrices are less useful than others described in this unit. Dayhoff et al. (1978) were the first to define the terms protein family and superfamily. A protein *family* is defined as sequences 85% identical or greater to each other. A protein *superfamily* is defined as sequences related from 30% identical or greater to each other. A protein superfamily may contain many protein families. The user should be aware that while the terms "family" and "superfamily" are widely used in biology, most of the time the original definition of Dayhoff and collaborators is not being used (see below).

Locating all potential similarities: PAM 250

The most widely used PAM matrix is PAM 250 (Fig. 3.5.1). It has been chosen because it is capable of accurately detecting similarities in the 30% range (i.e., superfamilies), that is, when the two proteins are up to 70% different from each other (George et al., 1990). Another way to think about this is that the PAM 250

Contributed by David Wheeler
Current Protocols in Bioinformatics (2003) 3.5.1-3.5.6

Gaps

- Used to improve alignments between two sequences
  - Compensate for insertions and deletions
  - As such, *gaps represent biological events*
- Gaps must be kept to a reasonable number, to not reflect a biologically implausible scenario. About one gap per 20 residues is a good rule-of-thumb.
- Cannot be scored simply as a ‘match’ or a ‘mismatch’

Affine Gap Penalty

Fixed deduction for introducing a gap *plus*
an additional deduction proportional to the length of the gap

Deduction for a gap = $G + Ln$

				nucleotide	protein
where	G	=	gap-opening penalty	5	11
	L	=	gap-extension penalty	2	1
	n	=	length of the gap		
and	G	>	L		

BLAST:
The Basic Local Alignment Search Tool

BLAST

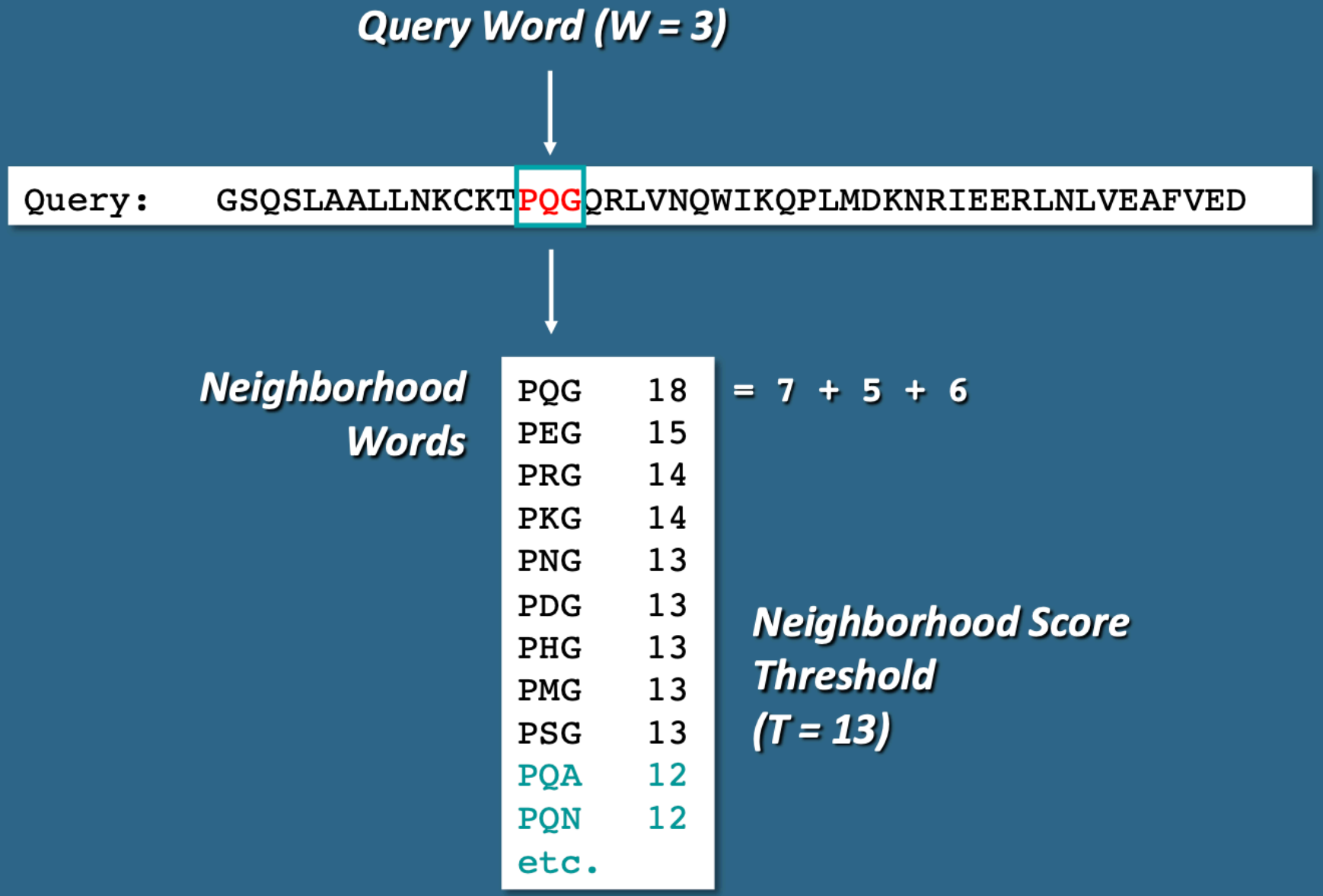
- Seeks high-scoring segment pairs (HSPs)
  - Pair of sequences that can be aligned with one another
  - When aligned, have maximal aggregate score (score cannot be improved by extension or trimming)
  - Score must be above score threshold (*S*)
  - Gapped or ungapped
- Results not limited to the 'best' high-scoring segment pair for the two sequences being aligned

Altschul et al., J. Mol. Biol. 215: 403-410, 1990

BLAST Algorithms

Program	*Query Sequence*	*Target Sequence*
BLASTN	Nucleotide	Nucleotide
BLASTP	Protein	Protein
BLASTX	Nucleotide, six-frame translation	Protein
TBLASTN	Protein	Nucleotide, six-frame translation
TBLASTX	Nucleotide, six-frame translation	Nucleotide, six-frame translation

Neighborhood Words



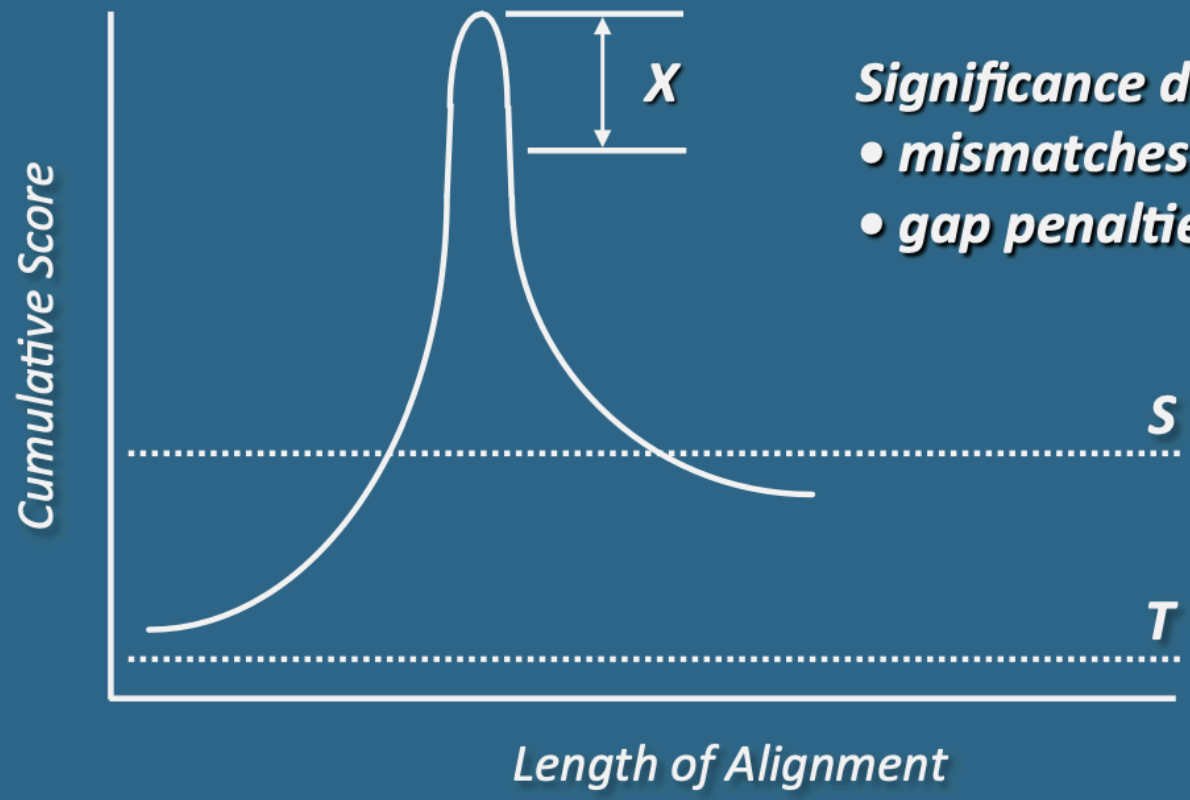
High-Scoring Segment Pairs

PQG	18
PEG	15
PRG	14
PKG	14
PNG	13
PDG	13
PHG	13
PMG	13
PSG	13
PQA	12
PQN	12
etc.	

```
Query:   325  SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA  365
              +LA++L     TP G R++ +W+ +P+ D    + ER    + A
Sbjct:   290  TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA  330
```

Extension

```
Query:   325  SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA  365
              +LA++L    TP G R++ +W+ +P+ D   + ER   + A
Sbjct:   290  TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA  330
```



Significance decay

- ***mismatches***
- ***gap penalties***

Scores and Alignment Length Don't Tell the Whole Story

```
Query:  1  SGLKSLVGKTALLSGTSSKL  20
           SGLKSLVGKTALLSGTSSKL
Sbjct:  1  SGLKSLVGKTALLSGTSSKL  20
```

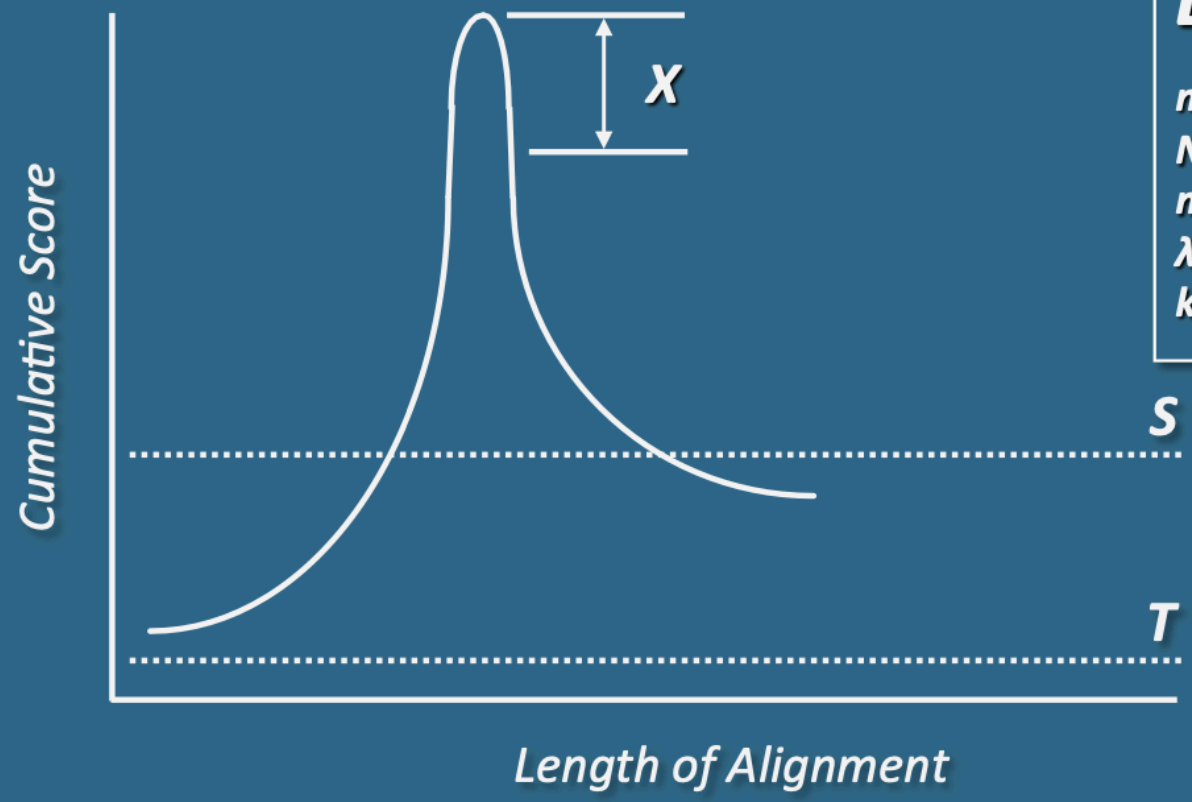
```
Score =  91
```

```
Query:  1  CQHMWYQWMIQCIWMYHCMQ  20
           CQHMWYQWMIQCIWMYHCMQ
Sbjct:  1  CQHMWYQWMIQCIWMYHCMQ  20
```

```
Score = 138
```

Scores and Probabilities

```
Query:    325   SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA   365
                +LA++L     TP G R++ +W+ +P+ D    + ER    + A
Sbjct:    290   TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA   330
```

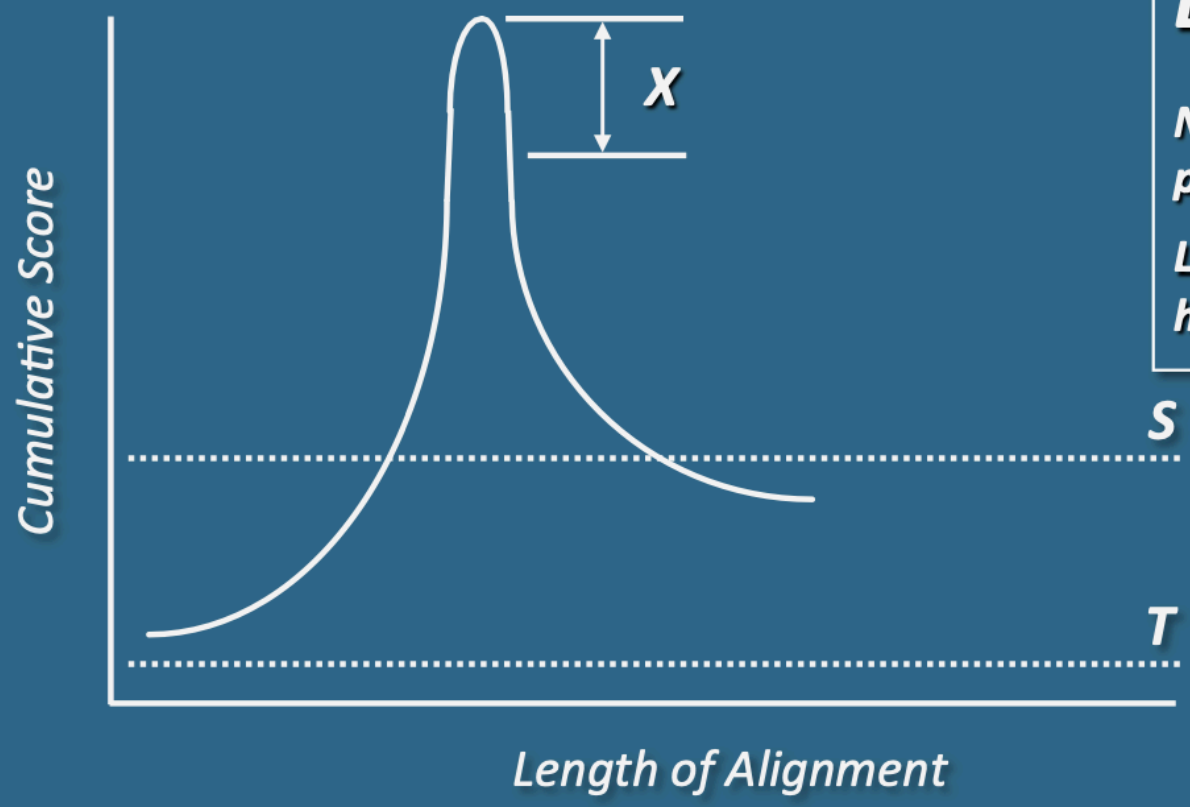


$$E = kmNe^{-\lambda S}$$

m	*# letters in query*
N	*# letters in database*
mN	*size of search space*
λS	*normalized score*
k	*minor constant*

Scores and Probabilities

```
Query:   325  SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA  365
              +LA++L    TP G R++ +W+ +P+ D   + ER   + A
Sbjct:   290  TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA  330
```

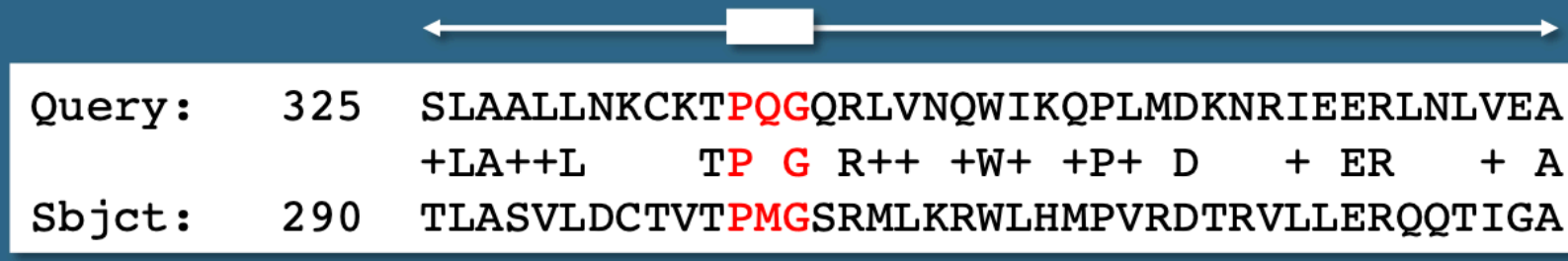


$$E = kmNe^{-\lambda S}$$

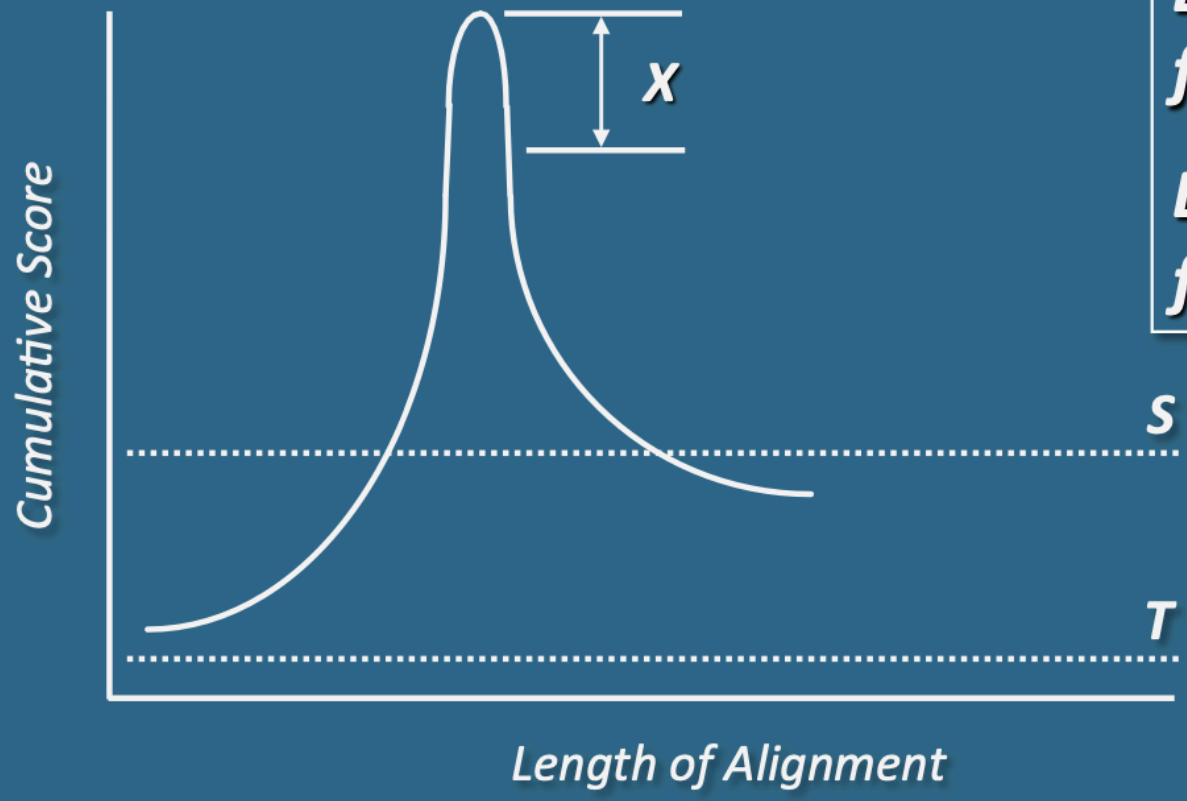
Number of HSPs found purely by chance

Lower values signify higher similarity

Scores and Probabilities



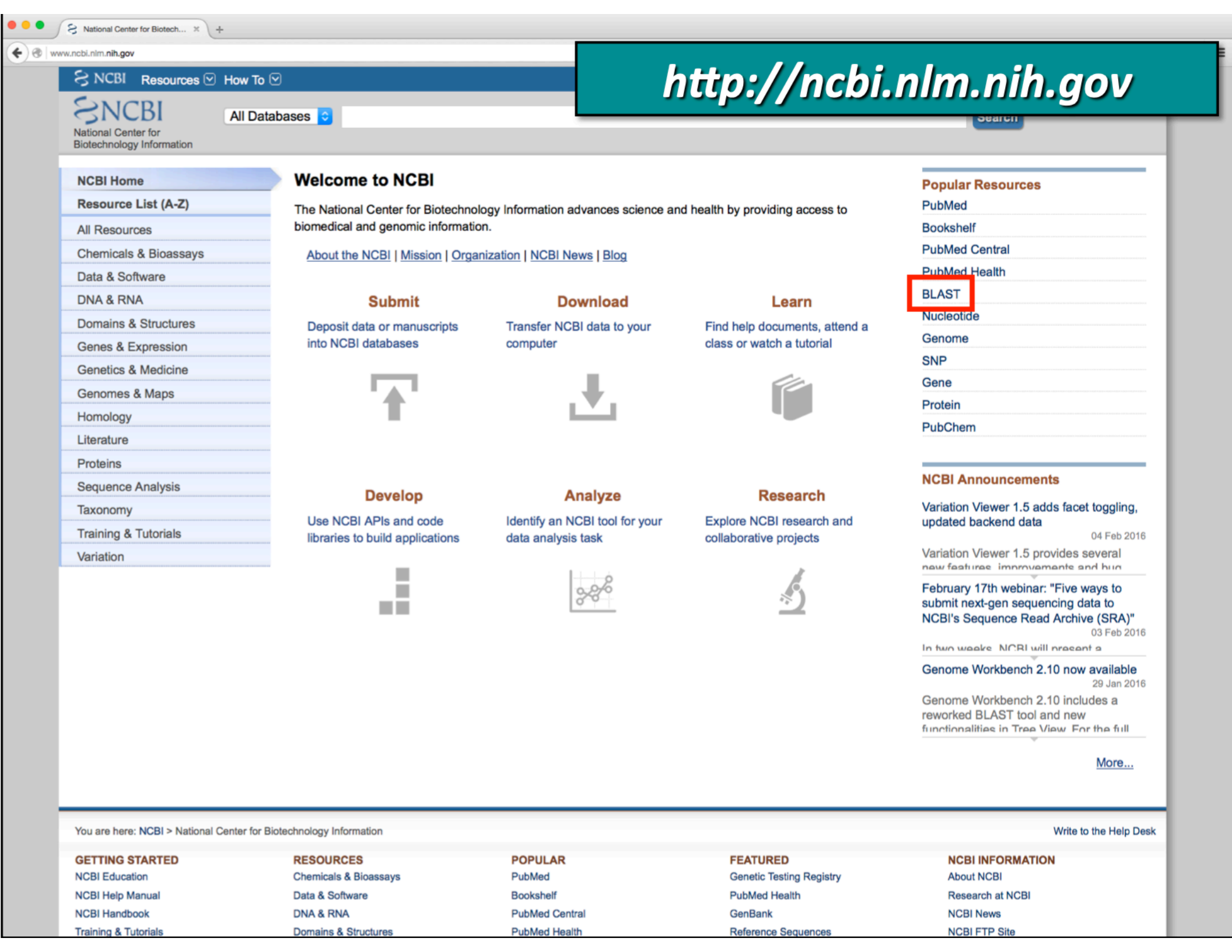
```
Query:   325  SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA  365
              +LA++L     TP G R++ +W+ +P+ D    + ER    + A
Sbjct:   290  TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA  330
```

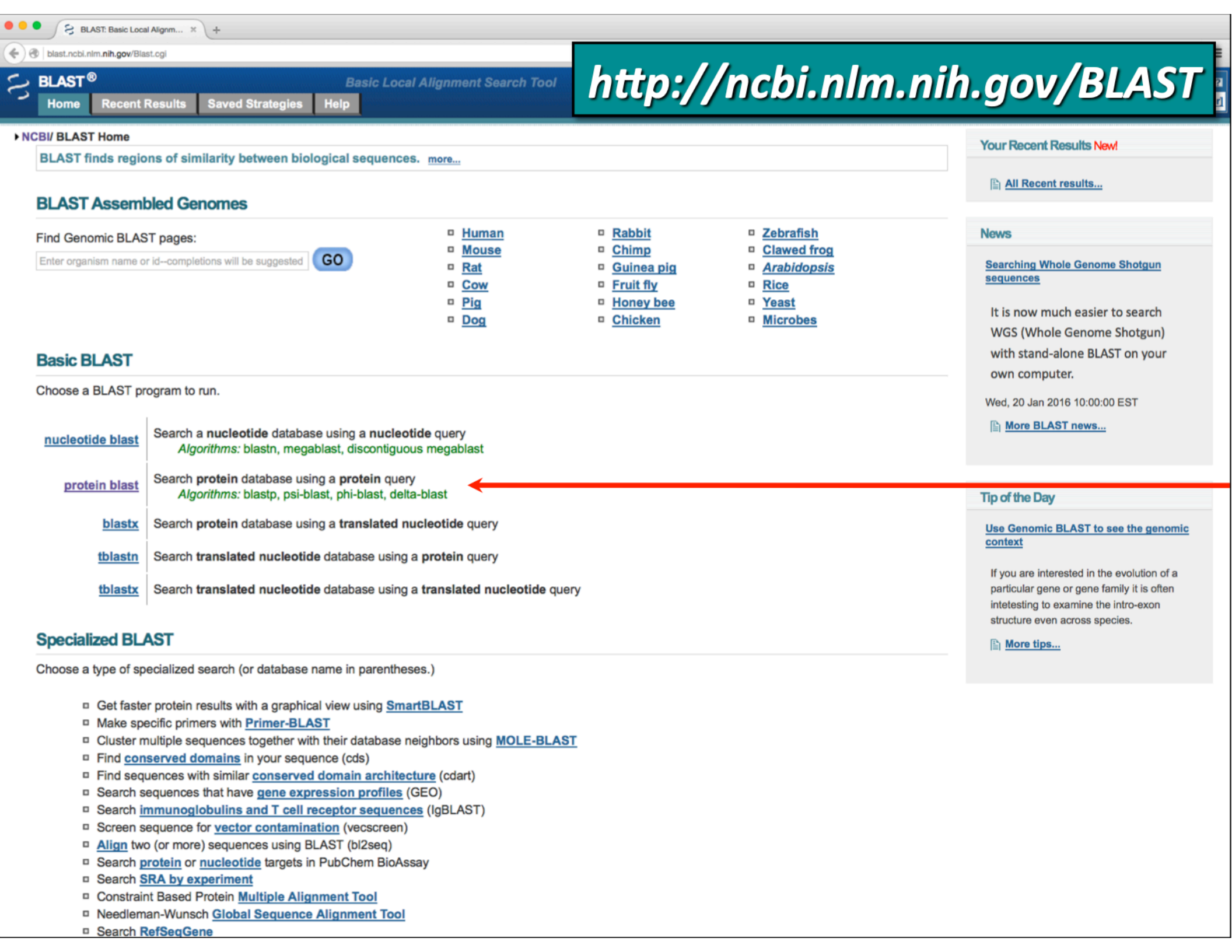


$E \leq 10^{-6}$
for nucleotides

$E \leq 10^{-3}$
for proteins

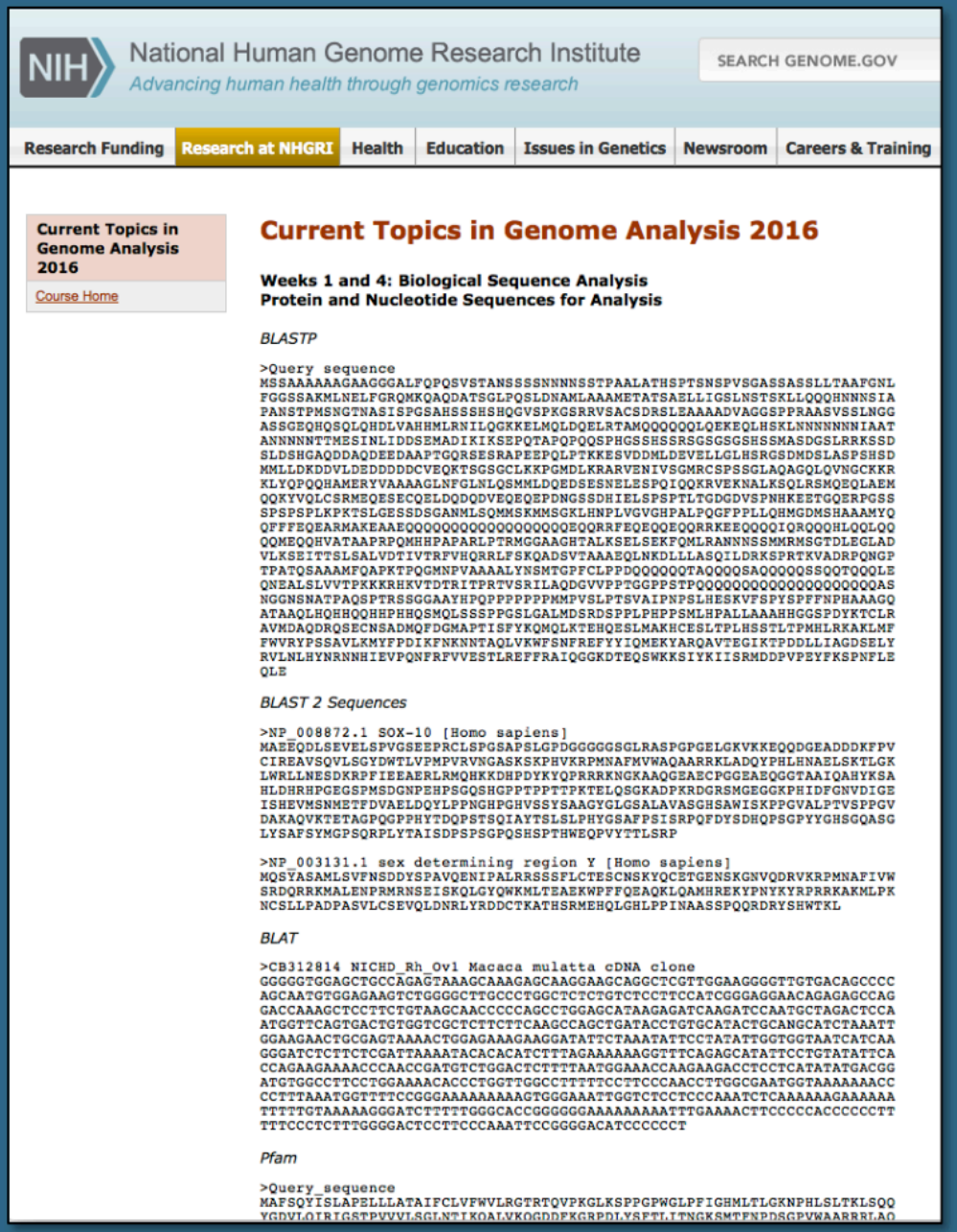
Using BLAST for Protein Similarity Searching


National Center for Biotech...
www.ncbi.nlm.nih.gov
http://ncbi.nlm.nih.gov
NCBI Resources How To
NCBI
National Center for Biotechnology Information
All Databases
Search
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
PubMed Health
BLAST
Nucleotide
Genome
SNP
Gene
Protein
PubChem
NCBI Announcements
Variation Viewer 1.5 adds facet toggling, updated backend data
04 Feb 2016
Variation Viewer 1.5 provides several
February 17th webinar: "Five ways to submit next-gen sequencing data to NCBI's Sequence Read Archive (SRA)"
03 Feb 2016
Genome Workbench 2.10 now available
29 Jan 2016
Genome Workbench 2.10 includes a reworked BLAST tool and new
More...
You are here: NCBI > National Center for Biotechnology Information
Write to the Help Desk
GETTING STARTED
NCBI Education
NCBI Help Manual
NCBI Handbook
Training & Tutorials
RESOURCES
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
POPULAR
PubMed
Bookshelf
PubMed Central
PubMed Health
FEATURED
Genetic Testing Registry
PubMed Health
GenBank
Reference Sequences
NCBI INFORMATION
About NCBI
Research at NCBI
NCBI News
NCBI FTP Site


BLAST: Basic Local Alignm...
blast.ncbi.nlm.nih.gov/Blast.cgi
BLAST®
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
http://ncbi.nlm.nih.gov/BLAST
NCBI/ BLAST Home
BLAST finds regions of similarity between biological sequences. more...
BLAST Assembled Genomes
Find Genomic BLAST pages:
Enter organism name or id--completions will be suggested
GO
Human
Mouse
Rat
Cow
Pig
Dog
Rabbit
Chimp
Guinea pig
Fruit fly
Honey bee
Chicken
Zebrafish
Clawed frog
Arabidopsis
Rice
Yeast
Microbes
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, delta-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
Specialized BLAST
Choose a type of specialized search (or database name in parentheses.)
Get faster protein results with a graphical view using SmartBLAST
Make specific primers with Primer-BLAST
Cluster multiple sequences together with their database neighbors using MOLE-BLAST
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 and T cell receptor sequences (IgBLAST)
Screen sequence for vector contamination (vecscreen)
Align two (or more) sequences using BLAST (bl2seq)
Search protein or nucleotide targets in PubChem BioAssay
Search SRA by experiment
Constraint Based Protein Multiple Alignment Tool
Needleman-Wunsch Global Sequence Alignment Tool
Search RefSeqGene
Your Recent Results New!
All Recent results...
News
Searching Whole Genome Shotgun sequences
It is now much easier to search WGS (Whole Genome Shotgun) with stand-alone BLAST on your own computer.
Wed, 20 Jan 2016 10:00:00 EST
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.
More tips...

Sequences Used in Examples

http://research.nhgri.nih.gov/teaching/seq_analysis.shtml



National Human Genome Research Institute
Advancing human health through genomics research
SEARCH GENOME.GOV

Research Funding | Research at NHGRI | Health | Education | Issues in Genetics | Newsroom | Careers & Training

Current Topics in Genome Analysis 2016
Course Home

Current Topics in Genome Analysis 2016

Weeks 1 and 4: Biological Sequence Analysis
Protein and Nucleotide Sequences for Analysis

BLASTP

```
>Query sequence
MSSAAAAAAGAAGGGALFQPQSVSTANSSSSNNNNSSTPAALATHSPTSNSPVSGASSASSLLTAAFGNL
FGGSSAKMLNELFGRQMKQAQDATSGLPQSLDNAMLAAAMETATSAELLIGSLNSTSKLLQQQHNNNSIA
PANSTPMSNGTNASISPGSAHSSSHSHQGVSPKGSRRVSACSDRSLEAAAADVAGGSPPRAASVSSLNGG
ASSGEQHQSQLQHDLVAHHMLRNILQGKKELMQLDQELRTAMQQQQQQLQEKEQLHSKLNNNNNNNIAAT
ANNNNNTTMESINLIDDSEMADIKIKSEPQTAPQPQQSPHGSSHSSRSGSGSGSHSSMASDGSLRRKSSD
SLDSHGAQDDAQDEEDAAPTGQRSESRAPEEPQLPTKKESVDDMLDEVELLGLHSRGSDMDSLASPSHSD
MMLLDKDDVLDEDDDDDCVEQKTSGSGCLKKPGMDLKRARVENIVSGMRCSPSSGLAQAGQLQVNGCKKR
KLYQPQQHAMERVVAAAAGLNFGLNLQSMMLDQEDSESNELESPQIQQKRVEKNALKSQLRSMQEQLAEM
QQKYVQLCSRMEQESECQELDQDQDVEQEQEPDNGSSDHIELSPSPTLTGDGDVSPNHKEETGQERPGSS
SPSPSPLKPKTSLGESSDSGANMLSQMMSKMMSGKLHNPLVGVGHPALPQGFPPLLQHMGDMSHAAAMYQ
QFFFEQEARMAKEAAEQQQQQQQQQQQQQQQQQQQEQQRRFEQEQQEQQRRKEEQQQQIQRQQQHLQQLQQ
QQMEQQHVATAAPRPQMHHPAPARLPTRMGGAAGHTALKSELSEKFQMLRANNNSSMMRMSGTDLEGLAD
VLKSEITTSLSALVDTIVTRFVHQRRLFSKQADSVTAAAEQLNKDLLLASQILDRKSPRTKVADRPQNGP
TPATQSAAAMFQAPKTPQGMNPVAAAALYNSMTGPFCLPPDQQQQQQTAQQQQSAQQQQQSSQQTQQQLE
QNEALSLVVTPKKKRHKVTDTRITPRTVSRILAQDGVVPPTGGPPSTPQQQQQQQQQQQQQQQQQQQQAS
NGGNSNATPAQSPTRSSGGAAYHPQPPPPPPPMMPVSLPTSVAIPNPSLHESKVFSPYSPFFNPHAAAGQ
ATAAQLHQHQQHHPHHQSMQLSSSPPGSLGALMDSRDSPPLPHPPSMLHPALLAAAHHGGSPDYKTCLR
AVMDAQDRQSECNSADMQFDGMAPTISFYKQMQLKTEHQESLMAKHCESLTPLHSSTLTPMHLRKAKLMF
FWVRYPSSAVLKMYFPDIKFNKNNTAQLVKWFSNFREFYYIQMEKYARQAVTEGIKTPDDLLIAGDSELY
RVLNLHYNRNNHIEVPQNFRFVVESTLREFFRAIQGGKDTEQSWKKSIYKIISRMDDPVPEYFKSPNFLE
QLE
```

BLAST 2 Sequences

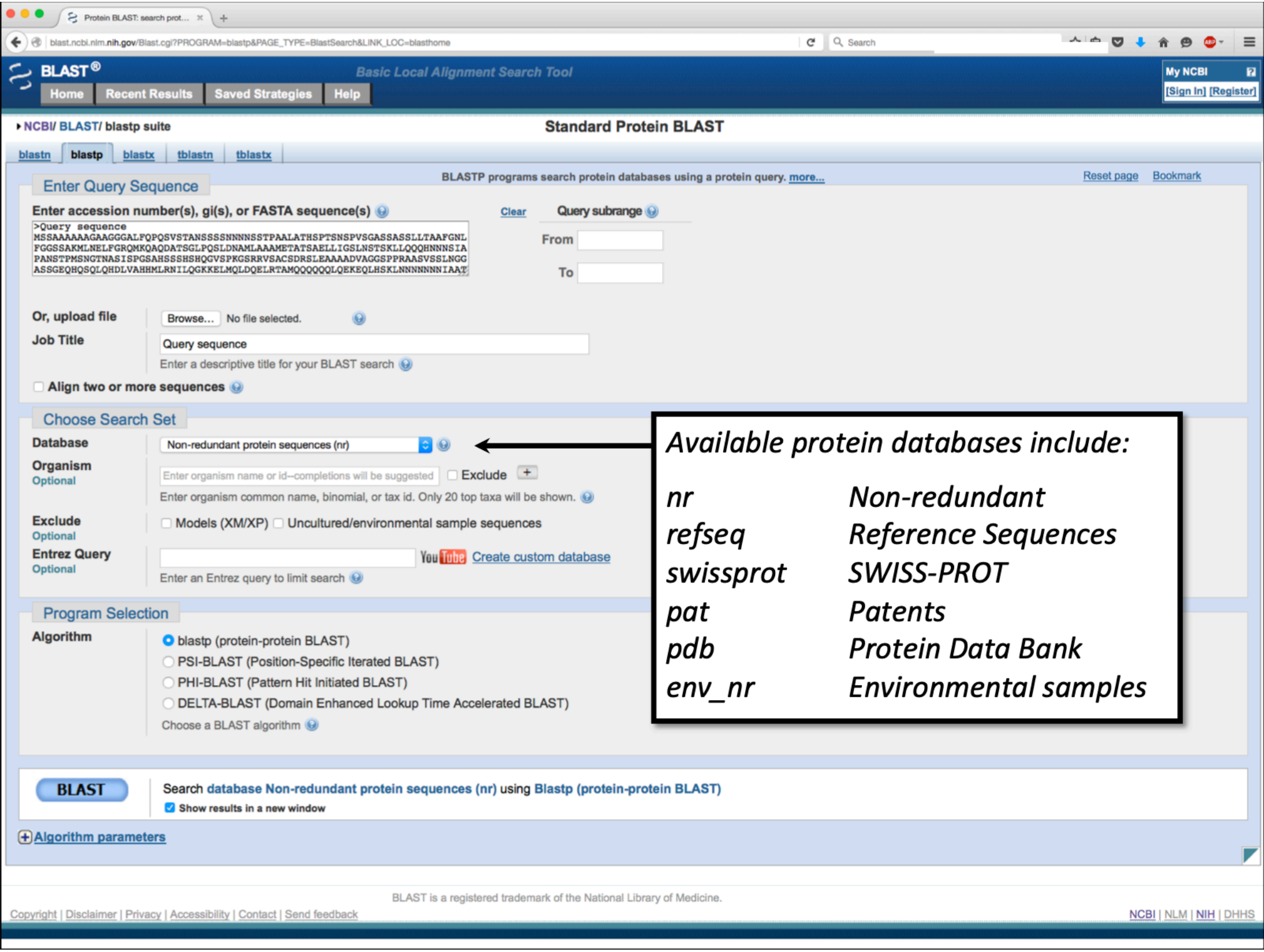
```
>NP_008872.1 SOX-10 [Homo sapiens]
MAEEQDLSEVELSPVGSEEPRCLSPGSAPSLGPDGGGGGSGLRASPGPGELGKVKKEQQDGEADDDKFPV
CIREAVSQVLSGYDWTLVPMPVRVNGASKSKPHVKRPMNAFMVWAQAARRKLADQYPHLHNAELSKTLGK
LWRLLNESDKRPFIEEAERLRMQHKKDHPDYKYQPRRRKNGKAAQGEAECPGGEAEQGGTAAIQAHYKSA
HLDHRHPGEGSPMSDGNPEHPSGQSHGPPTPPTTPKTELQSGKADPKRDGRSMGEGGKPHIDFGNVDIGE
ISHEVMSNMETFDVAELDQYLPPNGHPGHVSSYSAAGYGLGSALAVASGHSAWISKPPGVALPTVSPPGV
DAKAQVKTETAGPQGPPHYTDQPSTSQIAYTSLSLPHYGSAFPSISRPQFDYSDHQPSGPYYGHSGQASG
LYSAFSYMGPSQRPLYTAISDPSPSGPQSHSPTHWEQPVYTTLSRP

>NP_003131.1 sex determining region Y [Homo sapiens]
MQSYASAMLSVFNSDDYSPAVQENIPALRRSSSFLCTESCNSKYQCETGENSKGNVQDRVKRPMNAFIVW
SRDQRRKMALENPRMRNSEISKQLGYQWKMLTEAEKWPFFQEAQKLQAMHREKYPNYKYRPRRKAKMLPK
NCSLLPADPASVLCSEVQLDNRLYRDDCTKATHSRMEHQLGHLPPINAASSPQQRDRYSHWTKL
```

BLAT

```
>CB312814 NICHD_Rh_Ov1 Macaca mulatta cDNA clone
GGGGGTGGAGCTGCCAGAGTAAAGCAAAGAGCAAGGAAGCAGGCTCGTTGGAAGGGGTTGTGACAGCCCC
AGCAATGTGGAGAAGTCTGGGGCTTGCCCTGGCTCTCTGTCTCCTTCCATCGGGAGGAACAGAGAGCCAG
GACCAAAGCTCCTTCTGTAAGCAACCCCAGCCTGGAGCATAAGAGATCAAGATCCAATGCTAGACTCCA
ATGGTTCAGTGACTGTGGTCGCTCTTCTTCAAGCCAGCTGATACCTGTGCATACTGCANGCATCTAAATT
GGAAGAACTGCGAGTAAAACTGGAGAAAGAAGGATATTCTAAATATTCCTATATTGGTGGTAATCATCAA
GGGATCTCTTCTCGATTAAAATACACACATCTTTAGAAAAAAGGTTTCAGAGCATATTCCTGTATATTCA
CCAGAAGAAAACCCAACCGATGTCTGGACTCTTTTAATGGAAACCAAGAAGACCTCCTCATATATGACGG
ATGTGGCCTTCCTGGAAAACACCCTGGTTGGCCTTTTTCCTTCCCAACCTTGGCGAATGGTAAAAAAACC
CCTTTAAATGGTTTTCCGGGAAAAAAAAAGTGGGAAATTGGTCTCCTCCCAAATCTCAAAAAAGAAAAAA
TTTTTGTAAAAAGGGATCTTTTTGGGCACCGGGGGGAAAAAAAAATTTGAAAACTTCCCCCACCCCCCTT
TTTCCCTCTTTGGGGACTCCTTCCCAAATTCCGGGGACATCCCCCCT
```

Pfam

```
>Query_sequence
MAFSQYISLAPELLLATAIFCLVFWVLRGTRTQVPKGLKSPPGPWGLPFIGHMLTLGKNPHLSLTKLSQQ
YGDVLQIRIGSTPVVVLSGLNTIKQALVKQGDDFKGRPDLYSFTLITNGKSMTFNPDSGPVWAARRRLAQ
```


Protein BLAST: search prot...
blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome
Search
BLAST®
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
My NCBI
[Sign In] [Register]
NCBI/ BLAST/ blastp suite
Standard Protein BLAST
blastn
blastp
blastx
tblastn
tblastx
Enter Query Sequence
BLASTP programs search protein databases using a protein query. more...
Reset page
Bookmark
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
Query subrange
From
To
>Query sequence
MSSAAAAAAGAAGGGALFQPQSVSTANSSSSNNNNSSTPAALATHSPTSNSPVSGASSASSLLTAAFGNL
FGGSSAKMLNELFGRQMKQAQDATSGLPQSLDNAMLAAAMETATSAELLIGSLNSTSKLLQQQHNNNSIA
PANSTPMSNGTNASISPGSAHSSSHSHQGVSPKGSRRVSACSDRSLEAAAADVAGGSPPRAASVSSLNGG
ASSGEQHQSQLQHDLVAHHMLRNILQGKKELMQLDQELRTAMQQQQQQLQEKEQLHSKLNNNNNNIAAT
Or, upload file
Browse...
No file selected.
Job Title
Query sequence
Enter a descriptive title for your BLAST search
Align two or more sequences
Choose Search Set
Database
Non-redundant protein sequences (nr)
Organism
Optional
Enter organism name or id--completions will be suggested
Exclude
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.
Exclude
Optional
Models (XM/XP)
Uncultured/environmental sample sequences
Entrez Query
Optional
Create custom database
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)
DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
Choose a BLAST algorithm
Available protein databases include:
nr Non-redundant
refseq Reference Sequences
swissprot SWISS-PROT
pat Patents
pdb Protein Data Bank
env_nr Environmental samples
BLAST
Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)
Show results in a new window
Algorithm parameters
BLAST is a registered trademark of the National Library of Medicine.
Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback
NCBI | NLM | NIH | DHHS

NCBI RefSeq Database

- *Goal:* Provide a single reference sequence for each molecule of the central dogma (DNA, mRNA, and protein)

- Distinguishing features
  - Non-redundancy
  - Updates to reflect the current knowledge of sequence data and biology
  - Includes biological attributes of the gene, gene transcript, or protein
  - Encompasses a wide taxonomic range, with primary focus on mammalian and human species
  - Ongoing updates and curation (both automated and manual review), with review status indicated on each record

Pruitt et al., Nucleic Acids Res. 42: D756-D763, 2014

RefSeq Accession Number Prefixes

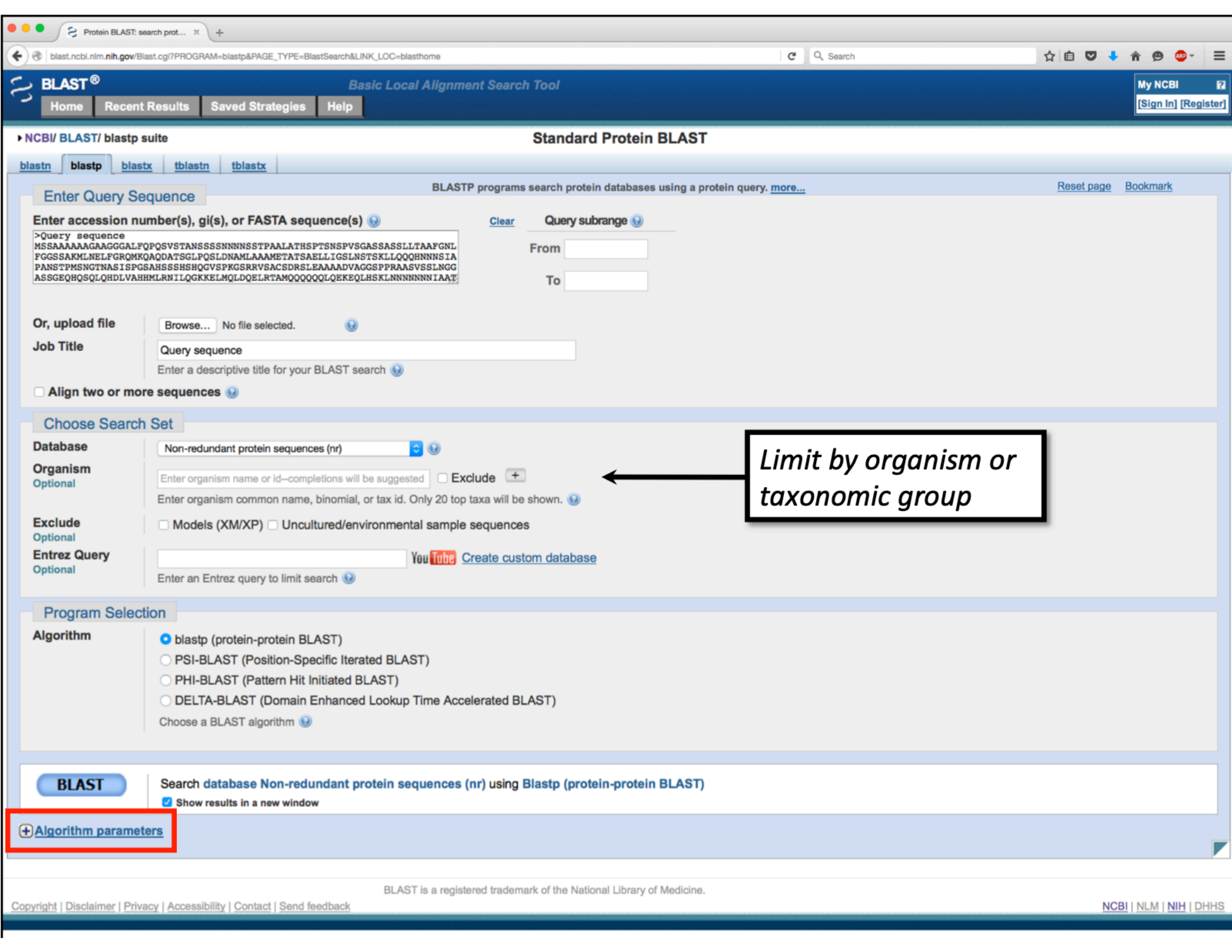
From curation of GenBank entries:

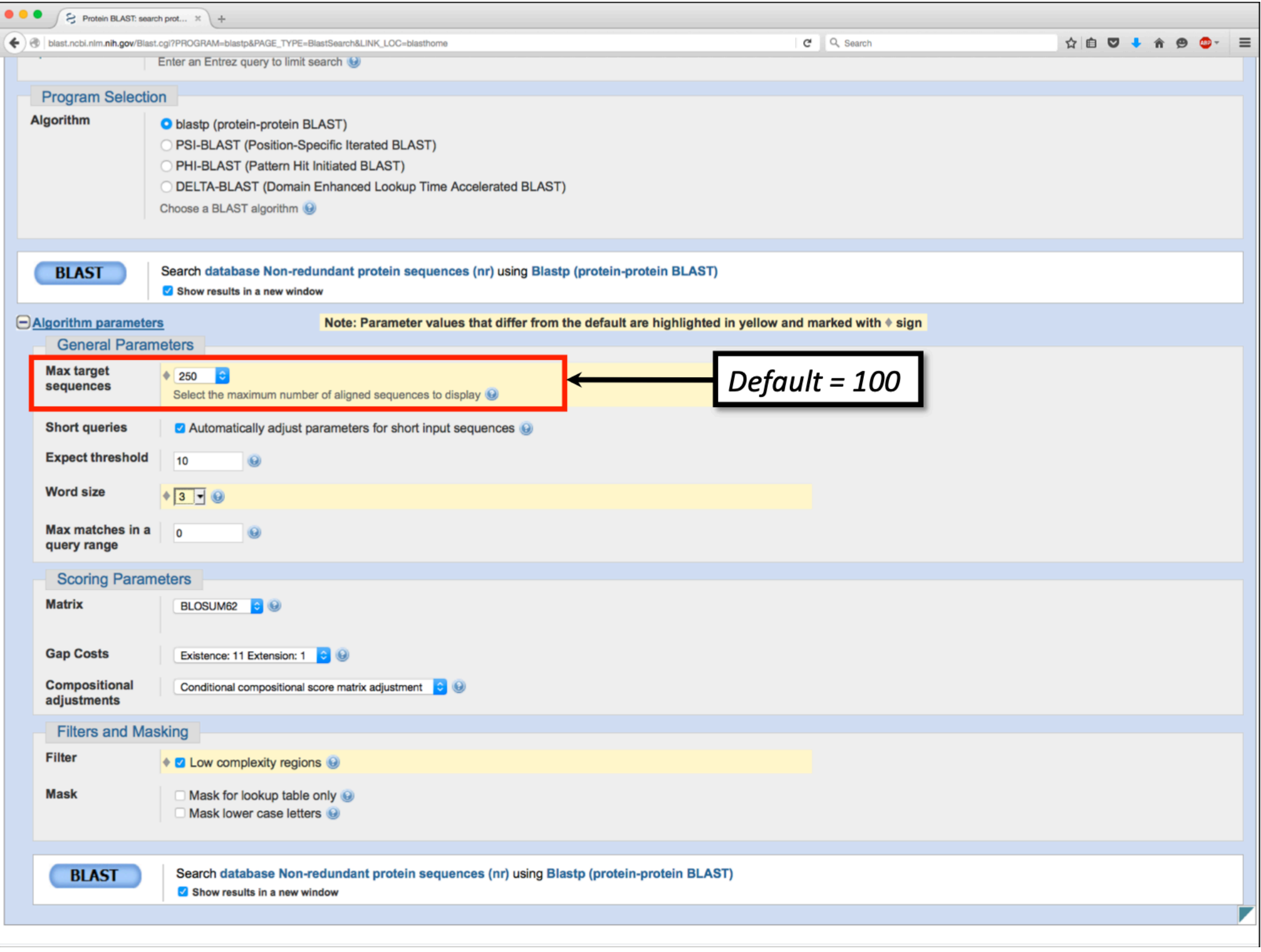
NT_	Genomic contigs
NM_	mRNAs
NP_	Proteins
NR_	Non-coding transcripts

From genome annotation:

XM_	Model mRNA
XP_	Model proteins

Complete list of molecule types in Chapter 18 of the NCBI Handbook
http://ncbi.nlm.nih.gov/books/NBK21091


Protein BLAST: search prot...
blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome
Search
BLAST®
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
My NCBI
[Sign In] [Register]
NCBI/ BLAST/ blastp suite
Standard Protein BLAST
blastn
blastp
blastx
tblastn
tblastx
Enter Query Sequence
BLASTP programs search protein databases using a protein query. more...
Reset page
Bookmark
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
Query subrange
From
To
>Query sequence
MSSAAAAAAGAAGGGALFQPQSVSTANSSSSNNNNSSTPAALATHSPTSNSPVSGASSASSLLTAAFGNL
FGGSSAKMLNELFGRQMKQAQDATSGLPQSLDNAMLAAAMETATSAELLIGSLNSTSKLLQQQHNNNSIA
PANSTPMSNGTNASISPGSAHSSSHSHQGVSPKGSRRVSACSDRSLEAAAADVAGGSPPRAASVSSLNGG
ASSGEQHQSQLQHDLVAHHMLRNILQGKKELMQLDQELRTAMQQQQQQLQEKEQLHSKLNNNNNNNIAAT
Or, upload file
Browse...
No file selected.
Job Title
Query sequence
Enter a descriptive title for your BLAST search
Align two or more sequences
Choose Search Set
Database
Non-redundant protein sequences (nr)
Organism
Optional
Enter organism name or id--completions will be suggested
Exclude
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.
Exclude
Optional
Models (XM/XP)
Uncultured/environmental sample sequences
Entrez Query
Optional
YouTube
Create custom database
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)
DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
Choose a BLAST algorithm
BLAST
Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)
Show results in a new window
Algorithm parameters
Limit by organism or taxonomic group
BLAST is a registered trademark of the National Library of Medicine.
Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback
NCBI | NLM | NIH | DHHS


Protein BLAST: search prot...
blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome
Search
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)
DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
Choose a BLAST algorithm
BLAST
Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)
Show results in a new window
Algorithm parameters
Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign
General Parameters
Max target sequences
250
Select the maximum number of aligned sequences to display
Default = 100
Short queries
Automatically adjust parameters for short input sequences
Expect threshold
10
Word size
3
Max matches in a query range
0
Scoring Parameters
Matrix
BLOSUM62
Gap Costs
Existence: 11 Extension: 1
Compositional adjustments
Conditional compositional score matrix adjustment
Filters and Masking
Filter
Low complexity regions
Mask
Mask for lookup table only
Mask lower case letters
BLAST
Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)
Show results in a new window

Protein BLAST: search prot...

blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome

Search

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)
- ( ) DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

BLAST Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)

[x] Show results in a new window

Algorithm parameters

Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign

General Parameters

Max target sequences: ♦ 250
Select the maximum number of aligned sequences to display

Short queries: [x] Automatically adjust parameters for short input sequences

Expect threshold: 10

Word size: ♦ 3

Max matches in a query range: 0

E-value threshold
Reports all hits with E < 10

Scoring Parameters

Matrix: BLOSUM62

Gap Costs: Existence: 11 Extension: 1

Compositional adjustments: Conditional compositional score matrix adjustment

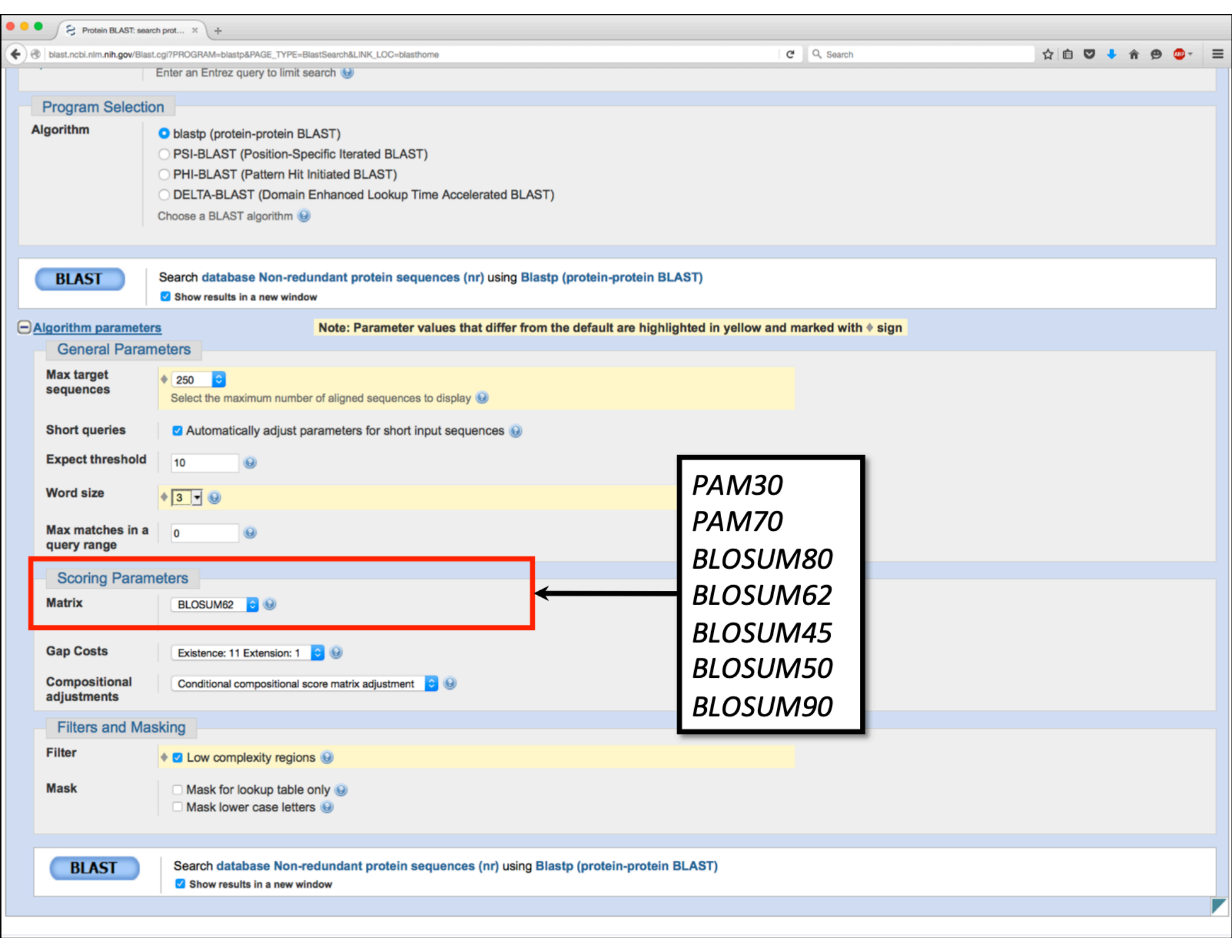
Filters and Masking

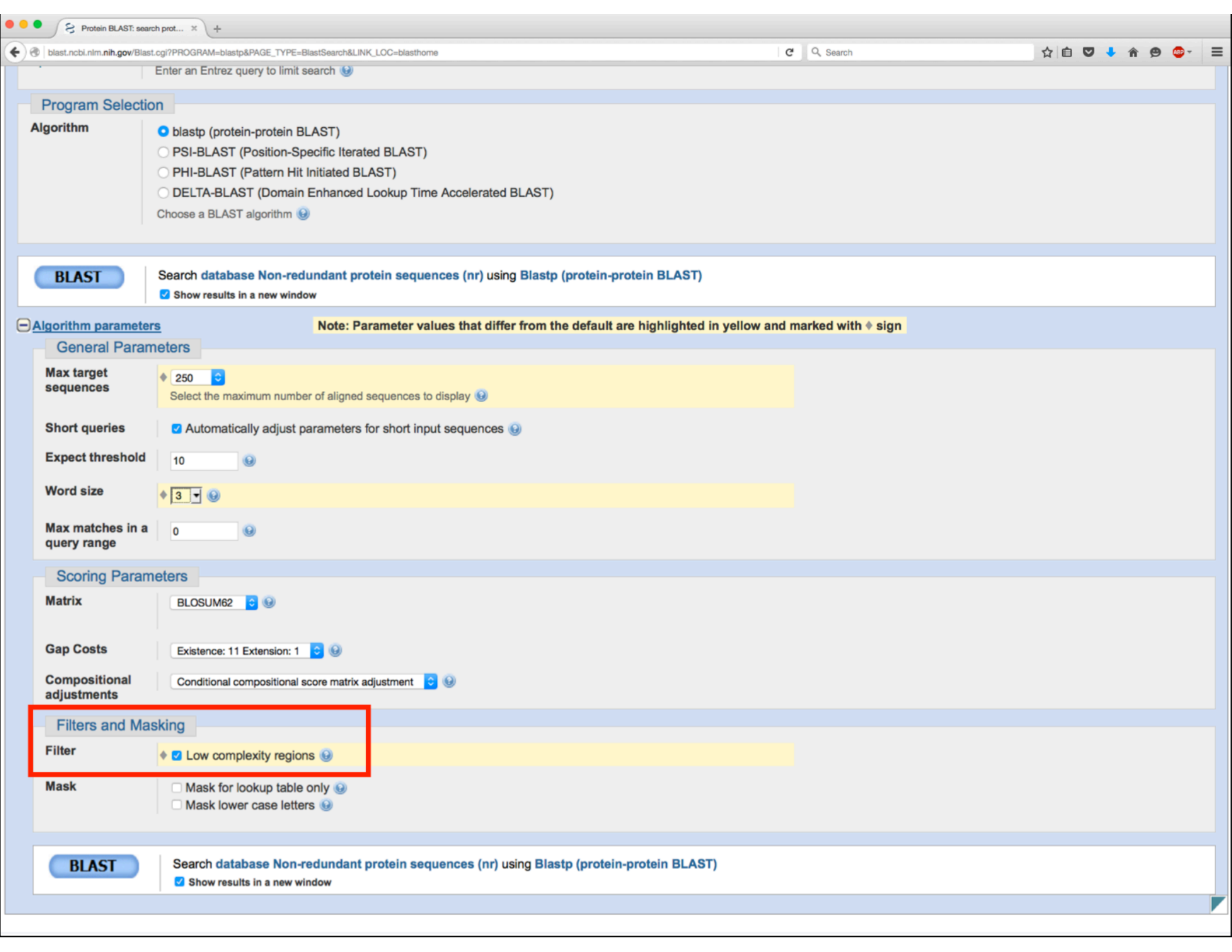
Filter: ♦ [x] Low complexity regions

Mask:
- [ ] Mask for lookup table only
- [ ] Mask lower case letters

BLAST Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)

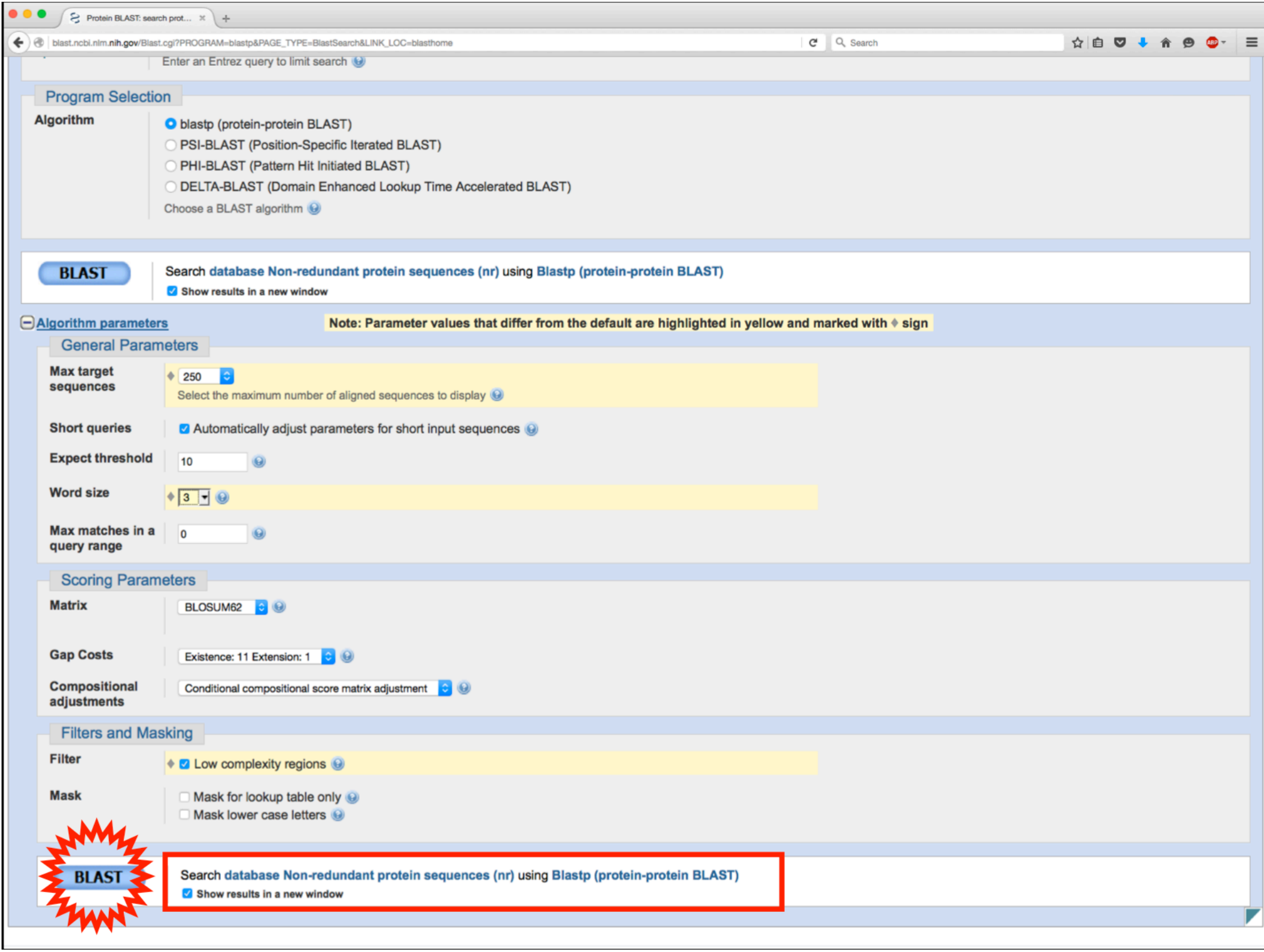
[x] Show results in a new window


Protein BLAST: search prot...
blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome
Search
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)
DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
Choose a BLAST algorithm
BLAST
Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)
Show results in a new window
Algorithm parameters
Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign
General Parameters
Max target sequences
250
Select the maximum number of aligned sequences to display
Short queries
Automatically adjust parameters for short input sequences
Expect threshold
10
Word size
3
Max matches in a query range
0
Scoring Parameters
Matrix
BLOSUM62
Gap Costs
Existence: 11 Extension: 1
Compositional adjustments
Conditional compositional score matrix adjustment
Filters and Masking
Filter
Low complexity regions
Mask
Mask for lookup table only
Mask lower case letters
BLAST
Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)
Show results in a new window
PAM30
PAM70
BLOSUM80
BLOSUM62
BLOSUM45
BLOSUM50
BLOSUM90


Protein BLAST: search prot...
blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome
Search
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)
DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
Choose a BLAST algorithm
BLAST
Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)
Show results in a new window
Algorithm parameters
Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign
General Parameters
Max target sequences
250
Select the maximum number of aligned sequences to display
Short queries
Automatically adjust parameters for short input sequences
Expect threshold
10
Word size
3
Max matches in a query range
0
Scoring Parameters
Matrix
BLOSUM62
Gap Costs
Existence: 11 Extension: 1
Compositional adjustments
Conditional compositional score matrix adjustment
Filters and Masking
Filter
Low complexity regions
Mask
Mask for lookup table only
Mask lower case letters
BLAST
Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)
Show results in a new window

Low-Complexity Regions

- Defined as regions of 'biased composition'
  - Homopolymeric runs
  - Short-period repeats
  - Subtle over-representation of several residues
- May confound sequence analysis
  - BLAST relies on uniformly-distributed amino acid frequencies
  - Often lead to false positives
- Filtering is advised (but *not* enabled by default)


Protein BLAST: search prot...
blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome
Search
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)
DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
Choose a BLAST algorithm
BLAST
Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)
Show results in a new window
Algorithm parameters
Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign
General Parameters
Max target sequences
250
Select the maximum number of aligned sequences to display
Short queries
Automatically adjust parameters for short input sequences
Expect threshold
10
Word size
3
Max matches in a query range
0
Scoring Parameters
Matrix
BLOSUM62
Gap Costs
Existence: 11 Extension: 1
Compositional adjustments
Conditional compositional score matrix adjustment
Filters and Masking
Filter
Low complexity regions
Mask
Mask for lookup table only
Mask lower case letters
BLAST
Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)
Show results in a new window

Edit and Resubmit | Save Search Strategies | Formatting options | Download | How to read this page | Blast report description

Query sequence

RID BCEGAEBZ014 (Expires on 02-08 01:23 am)

Query ID lcl|Query_345098
Description Query sequence
Molecule type amino acid
Query Length 1403

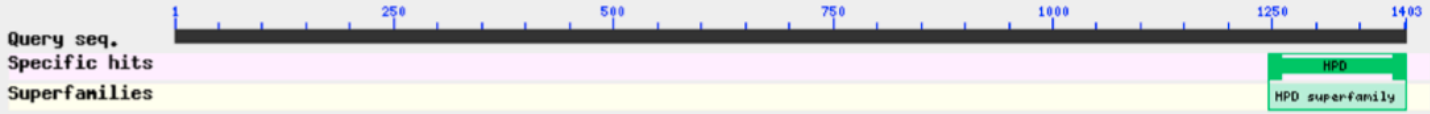
Database Name nr
Description All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Program BLASTP 2.3.1+ Citation

Other reports: Search Summary [Taxonomy reports] [Distance tree of results] [Multiple alignment]

Graphic Summary

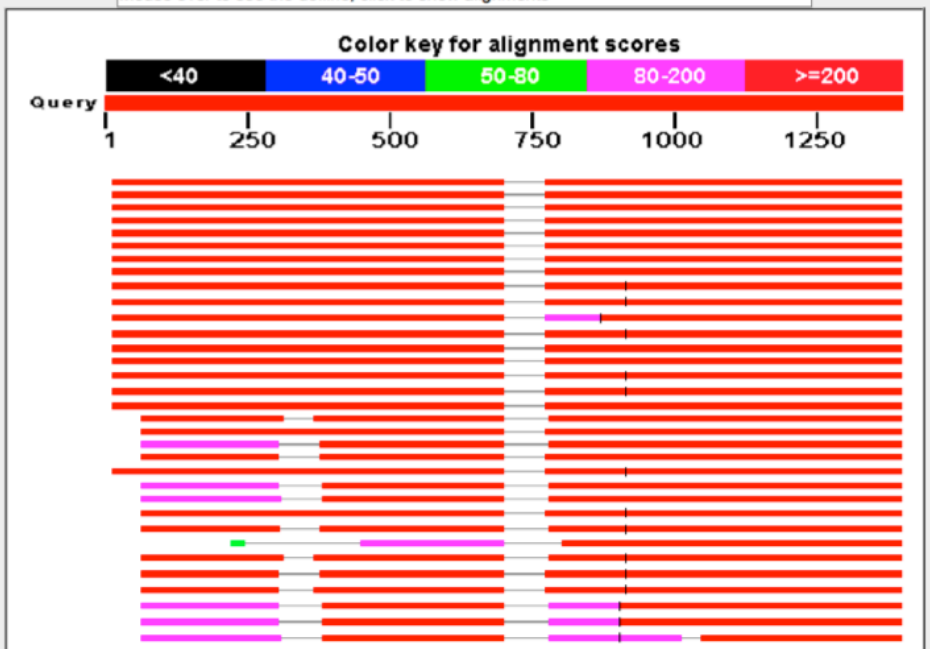
Show Conserved Domains

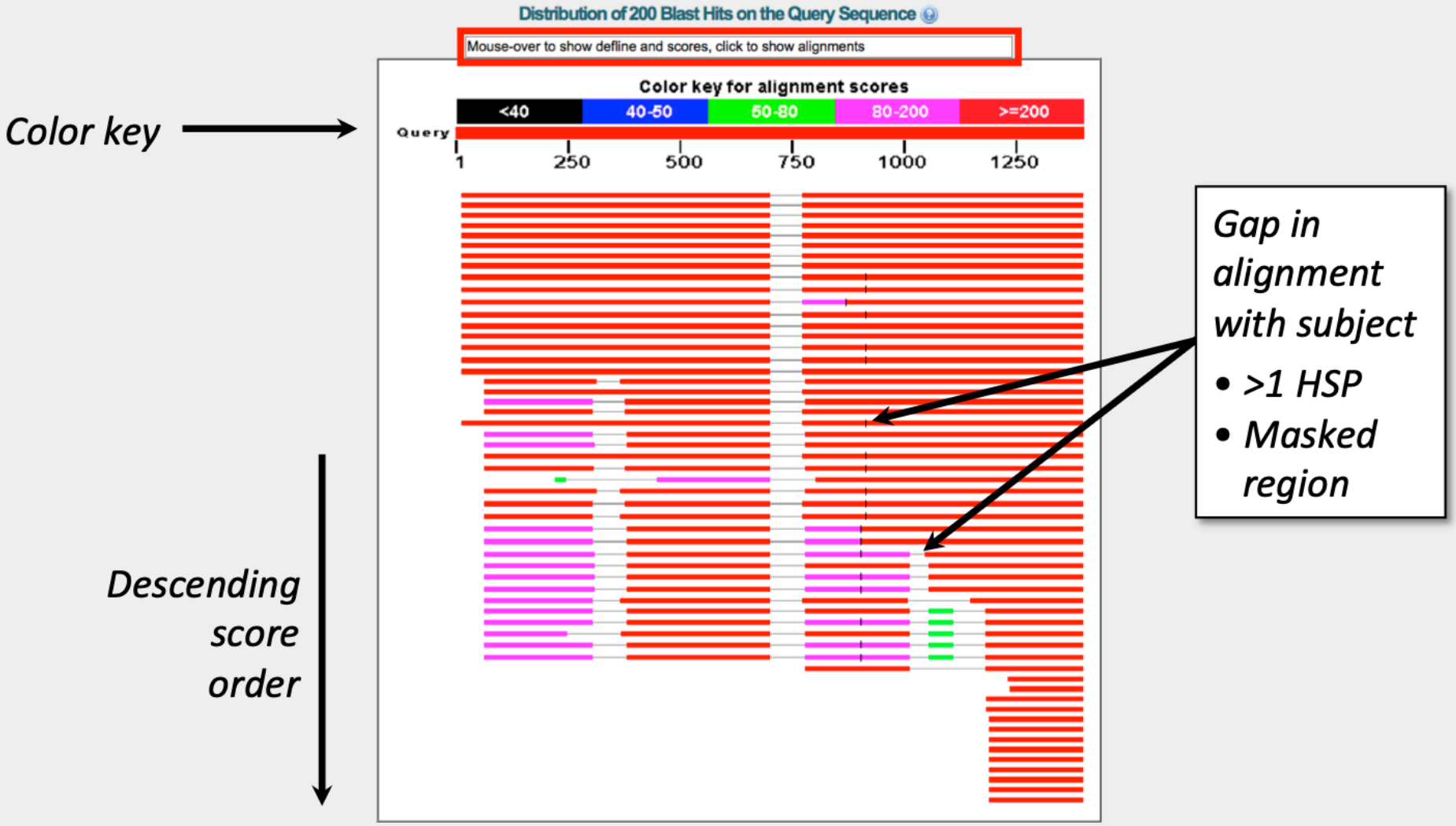
Putative conserved domains have been detected, click on the image below for detailed results.



Distribution of 200 Blast Hits on the Query Sequence

Mouse over to see the defline, click to show alignments





Descriptions

Sequences producing significant alignments:

Select: All None Selected:0

Alignments | Download | GenPept | Graphics | Distance tree of results | Multiple alignment

	Description	Max score	Total score	Query cover	E value	Ident	Accession
☐	prospero, isoform M [Drosophila melanogaster]	994	1938	93%	0.0	100%	NP_001247046.1
☐	prospero, isoform J [Drosophila melanogaster]	993	1936	93%	0.0	100%	NP_524317.4

Sequences producing significant alignments:

Select: All None Selected:0

0.0 means ≤ 10^{-1000}

Alignments | Download | GenPept | Graphics | Distance tree of results | Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession
prospero, isoform M [Drosophila melanogaster]	994	1938	93%	0.0	100%	NP_001247046.1
prospero, isoform J [Drosophila melanogaster]	993	1936	93%	0.0	100%	NP_524317.4
prospero [Drosophila melanogaster]	993	1932	93%	0.0	100%	BAA01464.1
homeodomain transcription factor Prospero [Drosophila melanogaster]	990	1821	93%	0.0	100%	AAF05703.1
uncharacterized protein Dere_GG18089, isoform A [Drosophila erecta]	989	1885	93%	0.0	99%	XP_001980573.2
Pros protein [Drosophila melanogaster]	982	1811	93%	0.0	97%	AAA28841.1
prospero, isoform H [Drosophila melanogaster]	944	1862	93%	0.0	100%	NP_001247044.1
prospero, isoform L [Drosophila melanogaster]	943	1858	93%	0.0	100%	NP_788636.3
prospero, isoform I [Drosophila melanogaster]	942	1864	93%	0.0	100%	NP_001247045.1
prospero, isoform K [Drosophila melanogaster]	942	1863	93%	0.0	100%	NP_731565.4
GM23939 [Drosophila sechellia]	935	1987	93%	0.0	98%	XP_002031631.1
LOW QUALITY PROTEIN: prospero [Drosophila simulans]	932	1827	93%	0.0	98%	KMZ04266.1
uncharacterized protein Dere_GG18089, isoform B [Drosophila erecta]	915	1810	93%	0.0	95%	XP_015010069.1
uncharacterized protein Dana_GF16857, isoform A [Drosophila ananassae]	904	1673	93%	0.0	92%	XP_001954214.2
uncharacterized protein Dyak_GE26090 [Drosophila yakuba]	903	1816	93%	0.0	96%	XP_002097201.2
uncharacterized protein Dere_GG18089, isoform C [Drosophila erecta]	894	1814	93%	0.0	97%	XP_015010070.1
uncharacterized protein Dana_GF16857, isoform C [Drosophila ananassae]	855	1623	93%	0.0	90%	XP_014766172.1
uncharacterized protein Dwil_GK11290, isoform A [Drosophila willistoni]	845	1532	85%	0.0	83%	XP_002069959.2
uncharacterized protein Dpse_GA14403, isoform I [Drosophila pseudoobscura pseudoobscura]	825	1456	90%	0.0	82%	XP_001359985.4
GH21437 [Drosophila grimshawi]	809	1374	84%	0.0	80%	XP_001994360.1
uncharacterized protein Dmoj_GI22696, isoform B [Drosophila mojavensis]	799	1386	84%	0.0	78%	XP_002000130.2
uncharacterized protein Dana_GF16857, isoform B [Drosophila ananassae]	767	1627	93%	0.0	83%	XP_014766171.1
PREDICTED: homeobox protein prospero isoform X3 [Ceratitis capitata]	692	1111	84%	0.0	66%	XP_004529243.2
PREDICTED: homeobox protein prospero [Bactrocera oleae]	690	1115	84%	0.0	70%	XP_014096508.1
uncharacterized protein Dpse_GA14403, isoform D [Drosophila pseudoobscura pseudoobscura]	612	14				
pros [Drosophila busckii]	611	14				
AAEL002769-PA [Aedes aegypti]	571	770	62%	8e-179	59%	XP_001655942.1
uncharacterized protein Dwil_GK11290, isoform B [Drosophila willistoni]	571	1501	85%	2e-171	77%	XP_015032827.1

8e−179 = 8×10^{-179}

Prospero homeobox protein 1 [Chlamydotis macqueenii]	226	270	19%	6e-58	62%	KFP45850.1
Prospero homeobox protein 1 [Cuculus canorus]	226	270	19%	6e-58	62%	KFO75119.1
PREDICTED: prospero homeobox protein 1-like [Poecilia formosa]	228	228	12%	6e-58	57%	XP_007567659.1
Prospero homeobox protein 1 [Pterocles gutturalis]	225	269	19%	6e-58	63%	KFV13087.1
homeobox protein prospero/prox-1 [Culex quinquefasciatus]	209	209	9%	6e-58	76%	XP_001845683.1
PREDICTED: prospero homeobox protein 1 [Octodon degus]	226	270	19%	7e-58	63%	XP_004626924.1
Prospero homeobox protein 1 [Charadrius vociferus]	226	270	19%	7e-58	62%	KGL86766.1
PREDICTED: prospero homeobox protein 1 isoform X2 [Chinchilla lanigera]	226	270	19%	8e-58	63%	XP_005374780.1
PREDICTED: prospero homeobox protein 1 isoform X2 [Fukomys damarensis]	226	270	19%	8e-58	63%	XP_010640836.1
PREDICTED: prospero homeobox protein 1 isoform X2 [Cavia porcellus]	226	270	19%	8e-58	63%	XP_003474644.1
PREDICTED: prospero homeobox protein 1 isoform X1 [Saimiri boliviensis boliviensis]	226	270	19%	8e-58	63%	XP_010339250.1
PREDICTED: prospero homeobox protein 1 [Peromyscus maniculatus bairdii]	226	270	19%	8e-58	63%	XP_006972145.1
PREDICTED: prospero homeobox protein 1 [Chaetura pelagica]	225	270	19%	8e-58	63%	XP_009993032.1
PREDICTED: prospero homeobox protein 1 isoform X2 [Callithrix jacchus]	225	270	19%	8e-58	6	
PREDICTED: prospero homeobox protein 1 isoform X1 [Heterocephalus glaber]	225	270	19%	8e-58	6	
PREDICTED: prospero homeobox protein 1 isoform X2 [Otolemur garnettii]	225	269	19%	8e-58	6	
PREDICTED: prospero homeobox protein 1 [Cuculus canorus]	225	270	19%	8e-58	6	
PREDICTED: prospero homeobox protein 1 isoform X2 [Equus asinus]	225	269	19%	8e-58	63%	XP_014690416.1
PREDICTED: prospero homeobox protein 1 isoform X2 [Propithecus coquereli]	225	270	19%	8e-58	63%	XP_012493821.1
PREDICTED: prospero homeobox protein 1 [Colobus angolensis palliatus]	225	269	19%	8e-58	63%	XP_011784800.1
PREDICTED: prospero homeobox protein 1 [Mandrillus leucophaeus]	225	270	19%	8e-58	63%	XP_011851547.1
prospero homeobox protein 1 [Homo sapiens]	225	270	19%	8e-58	63%	NP_002754.2
PREDICTED: LOW QUALITY PROTEIN: prospero homeobox protein 1-like [Colius striatus]	225	270	19%	8e-58	63%	XP_010205560.1
PREDICTED: prospero homeobox protein 1 isoform X2 [Columba livia]	225	271	19%	8e-58	63%	XP_005502819.1
PREDICTED: prospero homeobox protein 1 [Falco cherrug]	225	270	19%	9e-58	63%	XP_005441959.1
PREDICTED: prospero homeobox protein 1 [Marmota marmota marmota]	225	270	19%	9e-58	63%	XP_015339134.1
hypothetical protein EGM_01399 [Macaca fascicularis]	225	270	19%	9e-58	63%	EHH50546.1
PREDICTED: prospero homeobox protein 1 isoform X2 [Nannospalax galili]	225	270	19%	9e-58	63%	XP_008823079.1
PREDICTED: prospero homeobox protein 1 [Ochotona princeps]	225	269	19%	9e-58	63%	XP_004578703.1

Accept (for now)

Reject above desired threshold ($E \leq 10^{-3}$)

Alignments


Protein BLAST: search prot...
NCBI Blast:Query sequence
blast.ncbi.nlm.nih.gov/Blast.cgi#386765570
prospero, isoform L [Drosophila melanogaster]
Sequence ID: ref|NP_788636.3| Length: 1374 Number of Matches: 2
See 2 more title(s)
Range 1: 17 to 704 GenPept Graphics
Next Match
Previous Match
Score Expect Method Identities Positives Gaps
943 bits(2437) 0.0 Compositional matrix adjust. 688/688(100%) 688/688(100%) 0/688(0%)
Query 17 LFQPQSVSTAnssssnnnnssTPAALATHsptsnspvsgassasslltaaFGNLFGGSSA 76
LFQPQSVSTANSSSSNNNNSSTPAALATHSPTSNSPVSGASSASSLLTAAFGNLFGGSSA
Sbjct 17 LFQPQSVSTANSSSSNNNNSSTPAALATHSPTSNSPVSGASSASSLLTAAFGNLFGGSSA 76
Query 77 KMLNELFGRQMKQAQDATSGLPQSLDNAMLAAAMETATSAELLIGSLNSTSKLLQQQHNN 136
KMLNELFGRQMKQAQDATSGLPQSLDNAMLAAAMETATSAELLIGSLNSTSKLLQQQHNN
Sbjct 77 KMLNELFGRQMKQAQDATSGLPQSLDNAMLAAAMETATSAELLIGSLNSTSKLLQQQHNN 136
Query 137 NSIAPANSTPMSNGTNasispgsahssshshqgvspKGSRRVSACSDRSLEAAAADVAGG 196
NSIAPANSTPMSNGTNASISPGSAHSSSHSHQGVSPKGSRRVSACSDRSLEAAAADVAGG
Sbjct 137 NSIAPANSTPMSNGTNASISPGSAHSSSHSHQGVSPKGSRRVSACSDRSLEAAAADVAGG 196
Query 197 SPPRAASVSSLNGGASSGEQHQSQLQHDLVAHHMLRNILQGKKELMQLDQELRTAMqqqq 256
SPPRAASVSSLNGGASSGEQHQSQLQHDLVAHHMLRNILQGKKELMQLDQELRTAMQQQQ
Sbjct 197 SPPRAASVSSLNGGASSGEQHQSQLQHDLVAHHMLRNILQGKKELMQLDQELRTAMQQQQ 256
Query 257 qqlqekeqlHSKLnnnnnnniaatannnnnttMESINLIDDSEMADIKIKSEPQTAPQPQ 316
QQLQEKEQLHSKLNNNNNNNIAATANNNNNTTMESINLIDDSEMADIKIKSEPQTAPQPQ
Sbjct 257 QQLQEKEQLHSKLNNNNNNNIAATANNNNNTTMESINLIDDSEMADIKIKSEPQTAPQPQ 316
Query 317 QsphgsshssrsgsgsgshssmasdgslrrkssdsldsHGaqddaqdeedaaPTGQRSES 376
QSPHGSSHSSRSGSGSGSHSSMASDGSLRRKSSDSLDSHGAQDDAQDEEDAAPTGQRSES
Sbjct 317 QSPHGSSHSSRSGSGSGSHSSMASDGSLRRKSSDSLDSHGAQDDAQDEEDAAPTGQRSES 376
Query 377 RAPEEPQLPTKKESVDDMLDEVELLGLHSRGSDMDSLASPSHSdmmlldkddvldedddd 436
RAPEEPQLPTKKESVDDMLDEVELLGLHSRGSDMDSLASPSHSDMMLLDKDDVLDEDDDD
Sbjct 377 RAPEEPQLPTKKESVDDMLDEVELLGLHSRGSDMDSLASPSHSDMMLLDKDDVLDEDDDD 436
Query 437 dCVEQKTSGSGCLKKPGMDLKRARVENIVSGMRCSPSSGLAQAGQLQVNGCKKRKLYQPQ 496
DCVEQKTSGSGCLKKPGMDLKRARVENIVSGMRCSPSSGLAQAGQLQVNGCKKRKLYQPQ
Sbjct 437 DCVEQKTSGSGCLKKPGMDLKRARVENIVSGMRCSPSSGLAQAGQLQVNGCKKRKLYQPQ 496
Query 497 QHAMERYVaaaaGLNFGLNLQSMMLDQEDSESNELESPQIQQKRVEKNALKSQLRSMQEQ 556
QHAMERYVAAAAGLNFGLNLQSMMLDQEDSESNELESPQIQQKRVEKNALKSQLRSMQEQ
Sbjct 497 QHAMERYVAAAAGLNFGLNLQSMMLDQEDSESNELESPQIQQKRVEKNALKSQLRSMQEQ 556
Query 557 LAEMQQKYVQLCSRMEQESecqeldqdqdveqeqepdNGSSDHIELSPSPTLTGDGDVSP 616
LAEMQQKYVQLCSRMEQESECQELDQDQDVEQEQEPDNGSSDHIELSPSPTLTGDGDVSP
Sbjct 557 LAEMQQKYVQLCSRMEQESECQELDQDQDVEQEQEPDNGSSDHIELSPSPTLTGDGDVSP 616
Query 617 NHKEETGQERpgssspspsplkpktslgESSDSGANMLSQMMSKMMSGKLHNPLVGVGHP 676
NHKEETGQERPGSSSPSPSPLKPKTSLGESSDSGANMLSQMMSKMMSGKLHNPLVGVGHP
Sbjct 617 NHKEETGQERPGSSSPSPSPLKPKTSLGESSDSGANMLSQMMSKMMSGKLHNPLVGVGHP 676
Query 677 ALPQGFPPLLQHMGDMSHAAAMYQQFFF 704
ALPQGFPPLLQHMGDMSHAAAMYQQFFF
Sbjct 677 ALPQGFPPLLQHMGDMSHAAAMYQQFFF 704
Identities:
≥ 25% for proteins
≥ 70% for nucleotides
a Low complexity


NCBI Blast:Query sequence
Protein BLAST: search protein d...
NCBI Blast:Query sequence
blast.ncbi.nlm.nih.gov/Blast.cgi#386765570
Query 617 NHKEETGQERpgssspspsplkpktslgESSDSGANMLSQMMSKMMSGKLHNPLVGVGHP 676
NHKEETGQERPGSSSPSPSPLKPKTSLGESSDSGANMLSQMMSKMMSGKLHNPLVGVGHP
Sbjct 617 NHKEETGQERPGSSSPSPSPLKPKTSLGESSDSGANMLSQMMSKMMSGKLHNPLVGVGHP 676
Query 677 ALPQGFPPLLQHMGDMSHAAAMYQQFFF 704
ALPQGFPPLLQHMGDMSHAAAMYQQFFF
Sbjct 677 ALPQGFPPLLQHMGDMSHAAAMYQQFFF 704
Second HSP identified
Range 2: 777 to 1374 GenPept Graphics
Score Expect Method Identities Positives Gaps
915 bits(2365) 0.0 Compositional matrix adjust. 598/627(95%) 598/627(95%) 29/627(4%)
Query 777 HVATAAPRPQMHHPAPARLPTRMGGAAGHTALKSELSEKFQMLRANNNSSMMRMSGTDLE 836
HVATAAPRPQMHHPAPARLPTRMGGAAGHTALKSELSEKFQMLRANNNSSMMRMSGTDLE
Sbjct 777 HVATAAPRPQMHHPAPARLPTRMGGAAGHTALKSELSEKFQMLRANNNSSMMRMSGTDLE 836
Query 837 GLADVLKSEITTSLSALVDTIVTRFVHQRRLFSKQADSVTAAAEQLNKDLLLASQILDRK 896
GLADVLKSEITTSLSALVDTIVTRFVHQRRLFSKQADSVTAAAEQLNKDLLLASQILDRK
Sbjct 837 GLADVLKSEITTSLSALVDTIVTRFVHQRRLFSKQADSVTAAAEQLNKDLLLASQILDRK 896
Query 897 SPRTKVADRPQNGPTPATQSAAAMFQAPKTPQGMNPVAAAALYNSMTGPFCLPPDqqqqq 956
SPRTKVADRPQNGPTPATQSAAAMFQAPKTPQGMNPVAAAALYNSMTGPFCLPPDQQQQQ
Sbjct 897 SPRTKVADRPQNGPTPATQSAAAMFQAPKTPQGMNPVAAAALYNSMTGPFCLPPDQQQQQ 956
Query 957 qtaqqqqsaqqqqqssqqtqqqLEQNEALSLVVTPKKKRHKVTDTRITPRTVSRILAQDg 1016
QTAQQQQSAQQQQQSSQQTQQQLEQNEALSLVVTPKKKRHKVTDTRITPRTVSRILAQDG
Sbjct 957 QTAQQQQSAQQQQQSSQQTQQQLEQNEALSLVVTPKKKRHKVTDTRITPRTVSRILAQDG 1016
Query 1017 vvpptgqppstpqqqqqqqqqqqqqqqqqqqqqASNGGNSNATPAQSPTRSSGGAAYHpqp 1076
VVPPTGQPPSTPQQQQQQQQQQQQQQQQQQQQASNGGNSNATPAQSPTRSSGGAAYHPQP
Sbjct 1017 VVPPTGQPPSTPQQQQQQQQQQQQQQQQQQQQASNGGNSNATPAQSPTRSSGGAAYHPQP 1076
Query 1077 pppppmmpVSLPTSVAIPNPSLHESKVFSPYSPFFNPhaaaqqataaqlhqhhqqhhph 1136
PPPPPPMMPVSLPTSVAIPNPSLHESKVFSPYSPFFNPHAAAGQATAAQLHQHHQQHHPH
Sbjct 1077 PPPPPPMMPVSLPTSVAIPNPSLHESKVFSPYSPFFNPHAAAGQATAAQLHQHHQQHHPH 1136
Query 1137 hqsmqlsssppgslgALMDSRDspplphppsmlhpallaaahhggspDYKTCLRAVMDAQ 1196
HQSMQLSSSPPGSLGALMDSRDSPPLPHPPSMLHPALLAAAHHGGSPDYKTCLRAVMDAQ
Sbjct 1137 HQSMQLSSSPPGSLGALMDSRDSPPLPHPPSMLHPALLAAAHHGGSPDYKTCLRAVMDAQ 1196
Query 1197 DRQSECNSADMQFDGMAPTISFYKQMQLKTEHQESLMAKHCESTPLHSSTLTPMHLRKA 1256
DRQSECNSADMQFDGMAPT SSTLTPMHLRKA
Sbjct 1197 DRQSECNSADMQFDGMAPT----------SSTLTPMHLRKA 1227
Gap
Query 1257 KLMFFWVRYPSSAVLKMYFPDIKFNKNNTAQLVKWFSNFREFYYIQMEKYARQAVTEGIK 1316
KLMFFWVRYPSSAVLKMYFPDIKFNKNNTAQLVKWFSNFREFYYIQMEKYARQAVTEGIK
Sbjct 1228 KLMFFWVRYPSSAVLKMYFPDIKFNKNNTAQLVKWFSNFREFYYIQMEKYARQAVTEGIK 1287
Query 1317 TPDDLLIAGDSELYRVLNLHYNRNNHIEVPQNFRFVVESTLREFFRAIQGGKDTEQSWKK 1376
TPDDLLIAGDSELYRVLNLHYNRNNHIEVPQNFRFVVESTLREFFRAIQGGKDTEQSWKK
Sbjct 1288 TPDDLLIAGDSELYRVLNLHYNRNNHIEVPQNFRFVVESTLREFFRAIQGGKDTEQSWKK 1347
Query 1377 SIYKIISRMDDPVPEYFKSPNFLEQLE 1403
SIYKIISRMDDPVPEYFKSPNFLEQLE
Sbjct 1348 SIYKIISRMDDPVPEYFKSPNFLEQLE 1374

Score	Expect	Method	Identities	Positives	Gaps
943 bits(2437)	0.0	Compositional matrix adjust.	688/688(100%)	688/688(100%)	0/688(0%)

Score	Expect	Method	Identities	Positives	Gaps
915 bits(2365)	0.0	Compositional matrix adjust.	598/627(95%)	598/627(95%)	29/627(4%)

HSP 1

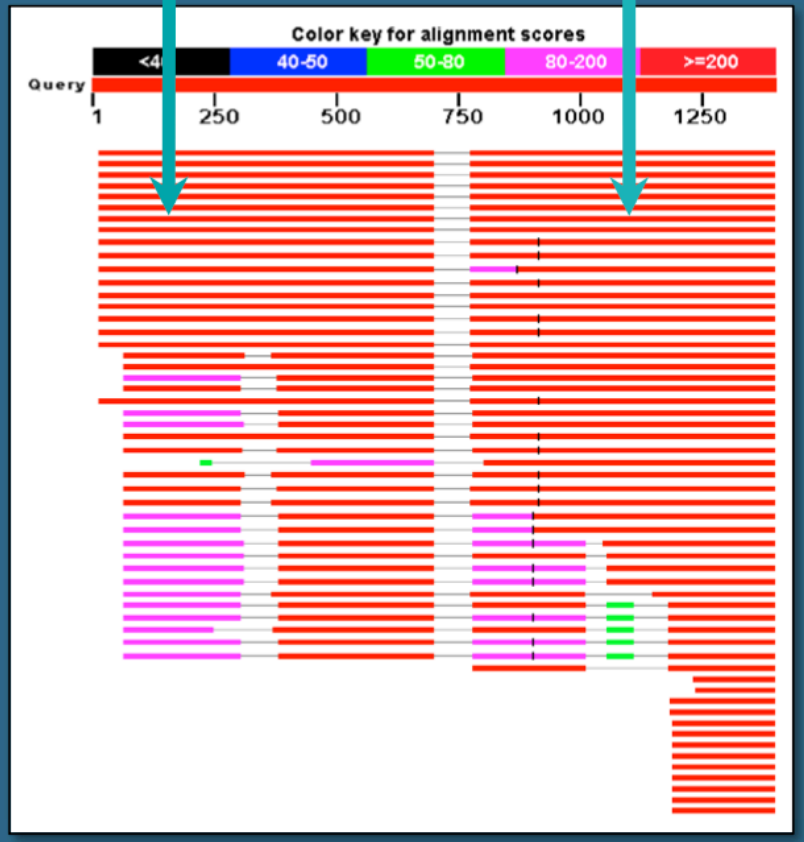
Q: 17– 704

S: 17– 704

HSP 2

Q: 777–1403

S: 777–1374



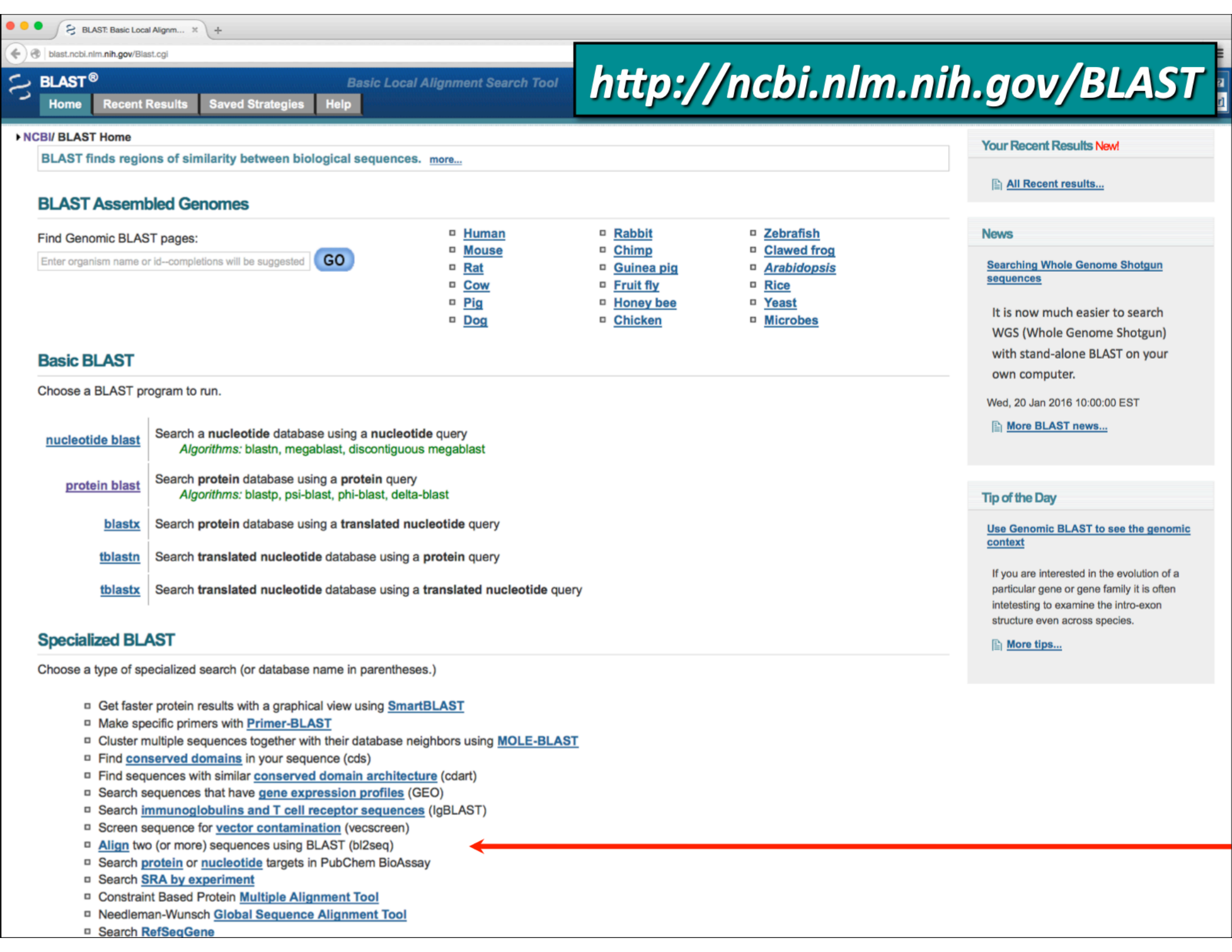
Suggested BLAST Cutoffs

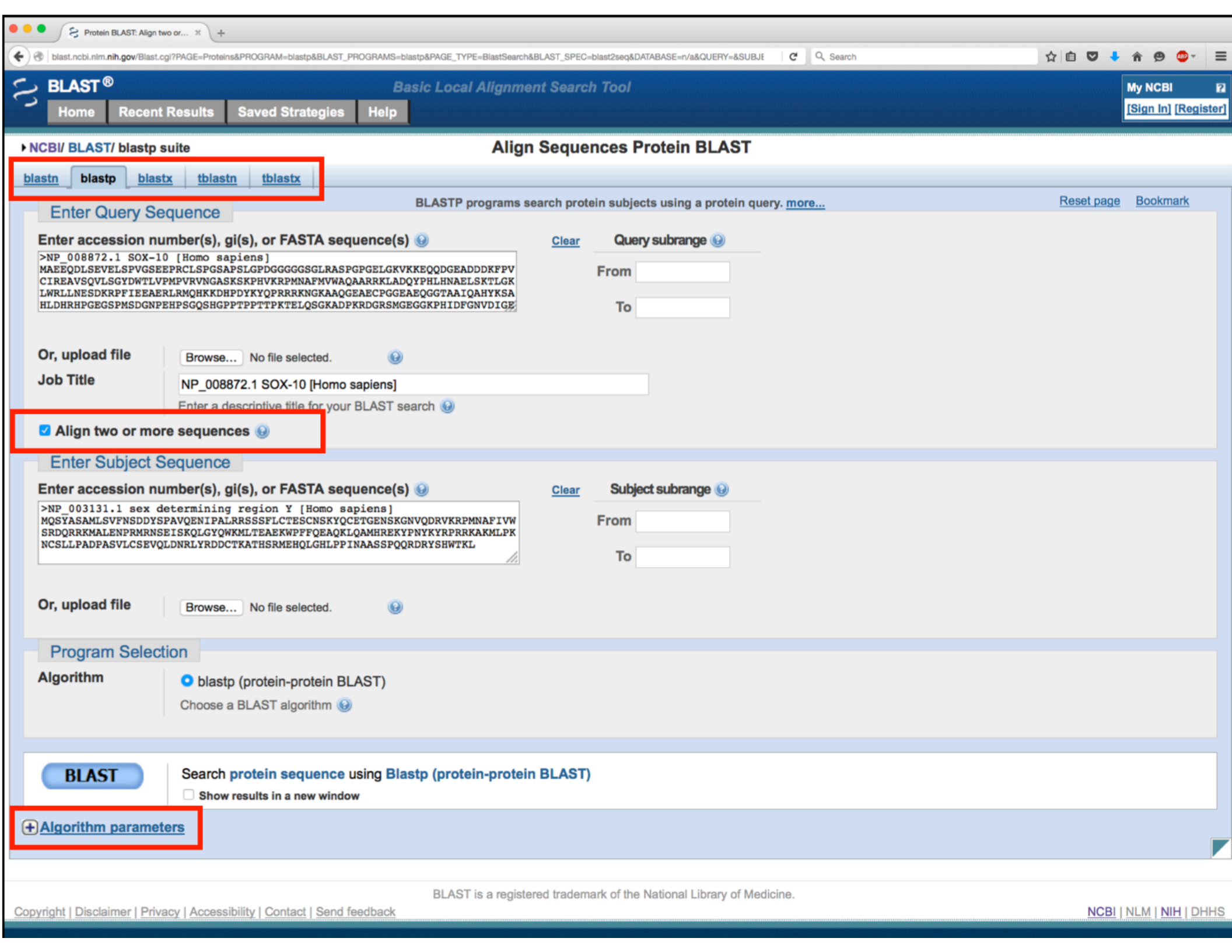
	E-value	Sequence Identity
Nucleotide	$\leq 10^{-6}$	$\geq 70\%$
Protein	$\leq 10^{-3}$	$\geq 25\%$

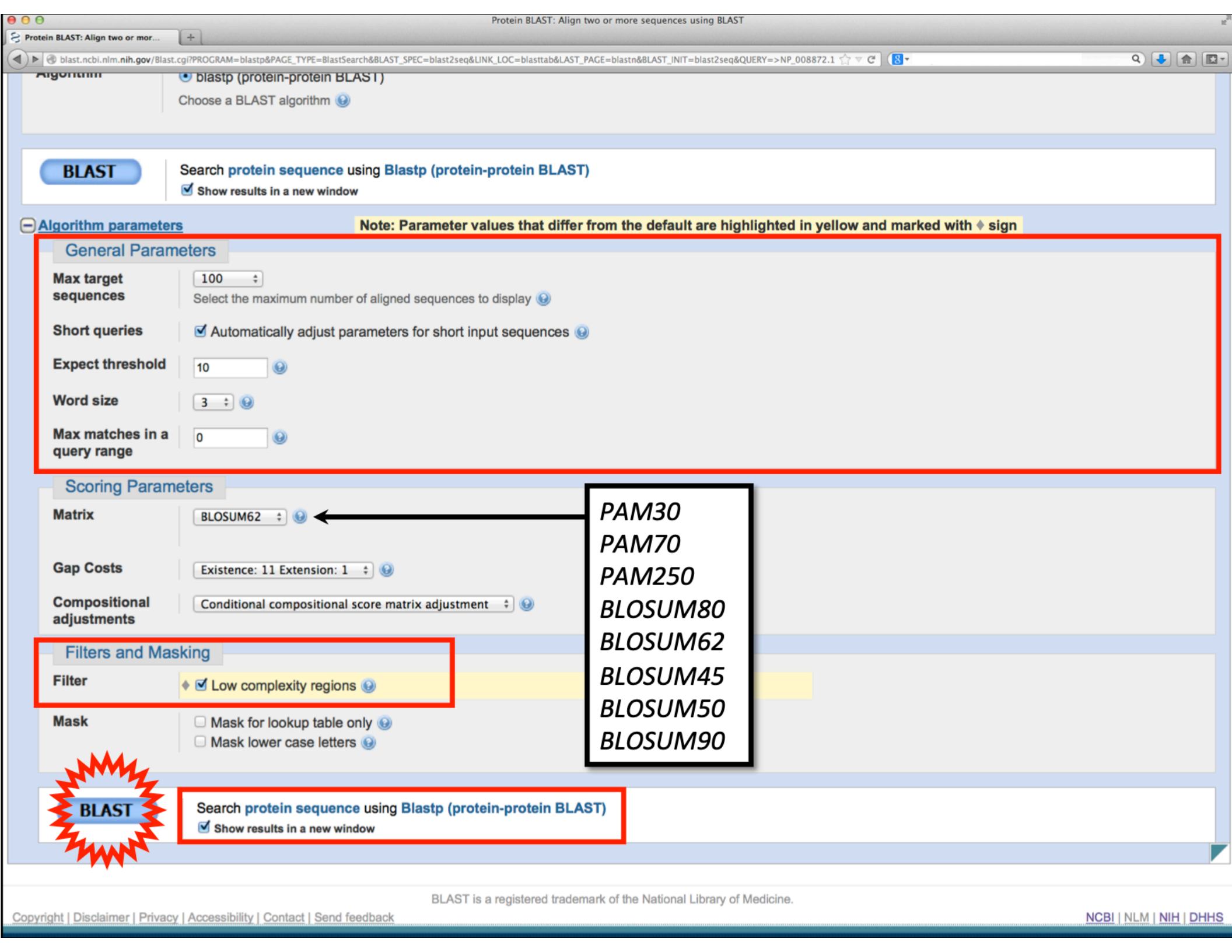
- ***Do not use these cutoffs blindly***
- ***Pay attention to alignments on either side of the dividing line***
- ***Do not ignore biology!***

BLAST 2 Sequences

- Finds local alignments between two protein or nucleotide sequences of interest
- All BLAST programs available
- Select BLOSUM and PAM matrices available for protein comparisons
- Same affine gap costs (adjustable)
- Input sequences can be masked


BLAST: Basic Local Alignm...
blast.ncbi.nlm.nih.gov/Blast.cgi
BLAST®
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
http://ncbi.nlm.nih.gov/BLAST
NCBI/ BLAST Home
BLAST finds regions of similarity between biological sequences. more...
BLAST Assembled Genomes
Find Genomic BLAST pages:
Enter organism name or id--completions will be suggested
GO
Human
Mouse
Rat
Cow
Pig
Dog
Rabbit
Chimp
Guinea pig
Fruit fly
Honey bee
Chicken
Zebrafish
Clawed frog
Arabidopsis
Rice
Yeast
Microbes
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, delta-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
Specialized BLAST
Choose a type of specialized search (or database name in parentheses.)
Get faster protein results with a graphical view using SmartBLAST
Make specific primers with Primer-BLAST
Cluster multiple sequences together with their database neighbors using MOLE-BLAST
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 and T cell receptor sequences (IgBLAST)
Screen sequence for vector contamination (vecscreen)
Align two (or more) sequences using BLAST (bl2seq)
Search protein or nucleotide targets in PubChem BioAssay
Search SRA by experiment
Constraint Based Protein Multiple Alignment Tool
Needleman-Wunsch Global Sequence Alignment Tool
Search RefSeqGene
Your Recent Results New!
All Recent results...
News
Searching Whole Genome Shotgun sequences
It is now much easier to search WGS (Whole Genome Shotgun) with stand-alone BLAST on your own computer.
Wed, 20 Jan 2016 10:00:00 EST
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.
More tips...


Protein BLAST: Align two or...
blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins&PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch&BLAST_SPEC=blast2seq&DATABASE=n/a&QUERY=&SUBJE
Search
BLAST®
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
My NCBI
[Sign In] [Register]
NCBI/ BLAST/ blastp suite
Align Sequences Protein BLAST
blastn
blastp
blastx
tblastn
tblastx
BLASTP programs search protein subjects using a protein query. more...
Reset page
Bookmark
Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
Query subrange
From
To
>NP_008872.1 SOX-10 [Homo sapiens]
MAEEQDLSEVELSPVGSEEPRCLSPGSAPSLGPDGGGGGSGLRASPGPGELGKVKKEQQDGEADDDKFPV
CIREAVSQVLSGYDWTLVPMPVRVNGASKSKPHVKRPMNAFMVWAQAARRKLADQYPHLHNAELSKTLGK
LWRLLNESDKRPFIEEAERLRMQHKKDHPDYKYQPRRRKNGKAAQGEAECPGGEAEQGGTAAIQAHYKSA
HLDHRHPGEGSPMSDGNPEHPSGQSHGPPTPPTTPKTELQSGKADPKRDGRSMGEGGKPHIDFGNVDIGE
Or, upload file
Browse...
No file selected.
Job Title
NP_008872.1 SOX-10 [Homo sapiens]
Enter a descriptive title for your BLAST search
Align two or more sequences
Enter Subject Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
Subject subrange
From
To
>NP_003131.1 sex determining region Y [Homo sapiens]
MQSYASAMLSVFNSDDYSPAVQENIPALRRSSSFLCTESCNSKYQCETGENSKGNVQDRVKRPMNAFIVW
SRDQRRKMALENPRMRNSEISKQLGYQWKMLTEAEKWPFFQEAQKLQAMHREKYPNYKYRPRRKAKMLPK
NCSLLPADPASVLCSEVQLDNRLYRDDCTKATHSRMEHQLGHLPPINAASSPQQRDRYSHWTKL
Or, upload file
Browse...
No file selected.
Program Selection
Algorithm
blastp (protein-protein BLAST)
Choose a BLAST algorithm
BLAST
Search protein sequence using Blastp (protein-protein BLAST)
Show results in a new window
Algorithm parameters
BLAST is a registered trademark of the National Library of Medicine.
Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback
NCBI | NLM | NIH | DHHS


Protein BLAST: Align two or more sequences using BLAST
Protein BLAST: Align two or mor...
blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&BLAST_SPEC=blast2seq&LINK_LOC=blasttab&LAST_PAGE=blastn&BLAST_INIT=blast2seq&QUERY=>NP_008872.1
Algorithm
blastp (protein-protein BLAST)
Choose a BLAST algorithm
BLAST
Search protein sequence using Blastp (protein-protein BLAST)
Show results in a new window
Algorithm parameters
Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign
General Parameters
Max target sequences
100
Select the maximum number of aligned sequences to display
Short queries
Automatically adjust parameters for short input sequences
Expect threshold
10
Word size
3
Max matches in a query range
0
Scoring Parameters
Matrix
BLOSUM62
Gap Costs
Existence: 11 Extension: 1
Compositional adjustments
Conditional compositional score matrix adjustment
PAM30
PAM70
PAM250
BLOSUM80
BLOSUM62
BLOSUM45
BLOSUM50
BLOSUM90
Filters and Masking
Filter
♦ Low complexity regions
Mask
Mask for lookup table only
Mask lower case letters
BLAST
Search protein sequence using Blastp (protein-protein BLAST)
Show results in a new window
BLAST is a registered trademark of the National Library of Medicine.
Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback
NCBI | NLM | NIH | DHHS

Blast 2 sequences

NP_008872.1 SOX-10 [Homo sapiens]

RID BCJA4YBV114 (Expires on 02-08 02:28 am)

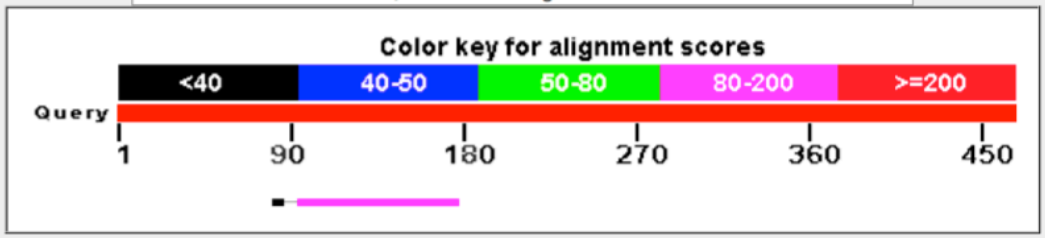
Query ID	lcl|Query_213409	**Subject ID**	lcl|Query_213411
Description	NP_008872.1 SOX-10 [Homo sapiens]	**Description**	NP_003131.1 sex determining region Y [Homo sapiens]
Molecule type	amino acid	**Molecule type**	amino acid
Query Length	466	**Subject Length**	204
		Program	BLASTP 2.3.1+ Citation

Other reports: Search Summary [Multiple alignment]

Graphic Summary

Distribution of 2 Blast Hits on the Query Sequence

Mouse over to see the defline, click to show alignments



Dot Matrix View

Descriptions

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download Graphics Multiple alignment

	Description	Max score	Total score	Query cover	E value	Ident	Accession
☐	NP_003131.1 sex determining region Y [Homo sapiens]	94.0	109	19%	1e-26	46%	Query_213411

Dot Matrix View

Descriptions

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download Graphics Multiple alignment

	Description	Max score	Total score	Query cover	E value	Ident	Accession
☐	NP_003131.1 sex determining region Y [Homo sapiens]	94.0	109	19%	1e-26	46%	Query_213411

Alignments

Download Graphics Sort by: E value Next Previous Descriptions

NP_003131.1 sex determining region Y [Homo sapiens]

Sequence ID: lcl|Query_213411 Length: 204 Number of Matches: 2

Related Information

Range 1: 51 to 134 Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
94.0 bits(232)	1e-26	Compositional matrix adjust.	39/84(46%)	62/84(73%)	0/84(0%)

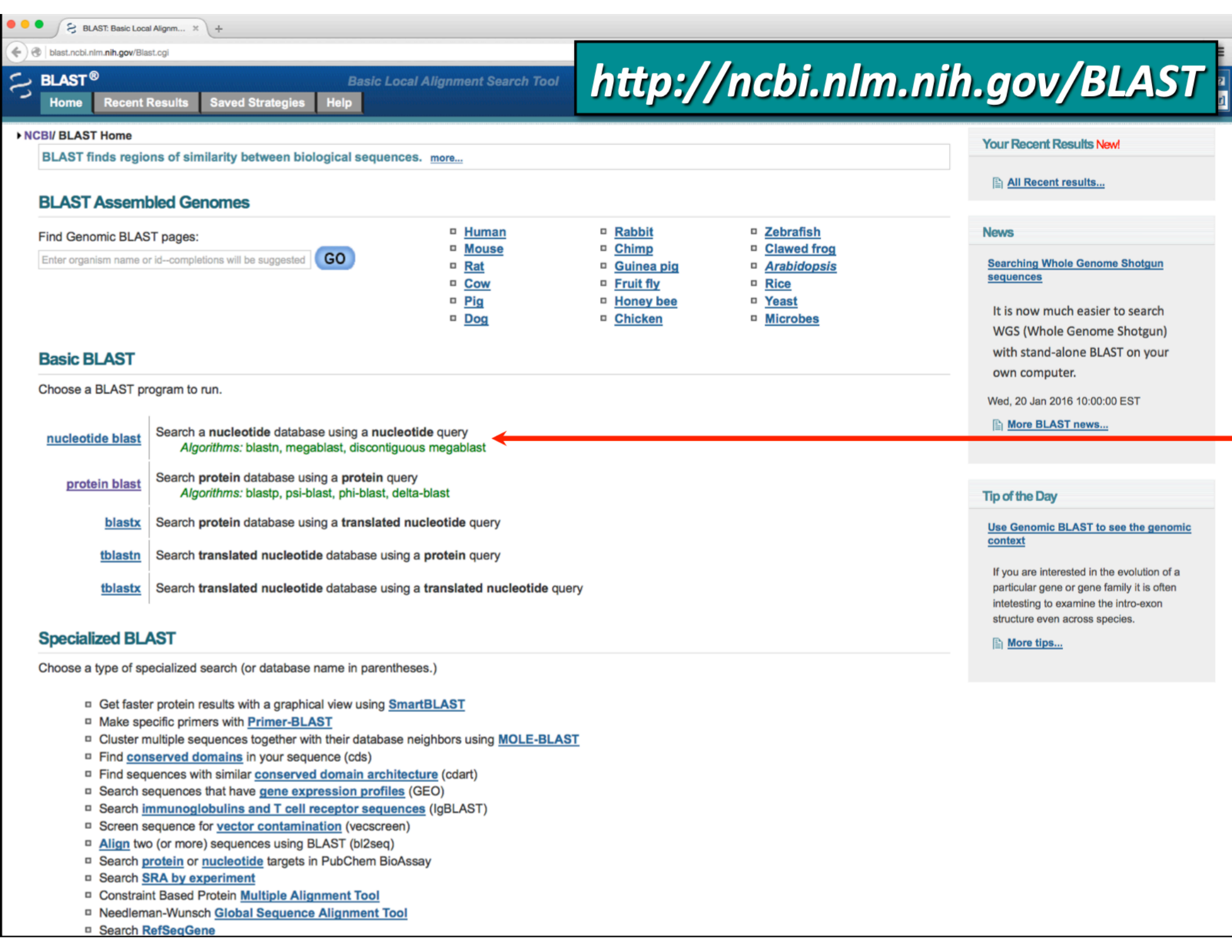
```
Query  95   NGASKSKPHVKRPMNAFMVWAQAARRKLADQYPHLHNAELSKTLGKLWRLLNESDKRPFI  154
            N     +  VKRPMNAF+VW++  RRK+A + P + N+E+SK LG  W++L E++K PF
Sbjct  51   NSKGNVQDRVKRPMNAFIVWSRDQRRKMALENPRMRNSEISKQLGYQWKMLTEAEKWPFF  110

Query  155  EEAERLRMQHKKDHPDYKYQPRRR  178
            +EA++L+  H++ +P+YKY+PRR+
Sbjct  111  QEAQKLQAMHREKYPNYKYRPRRK  134
```

Range 2: 95 to 101 Graphics Next Match Previous Match First Match

Score	Expect	Method	Identities	Positives	Gaps
15.4 bits(28)	1.9	Compositional matrix adjust.	3/7(43%)	5/7(71%)	0/7(0%)

```
Query  82   GYDWTLV  88
            GY W ++
Sbjct  95   GYQWKML  101
```

Nucleotide Similarity Searching: MegaBLAST, BLASTN, and BLAT


BLAST: Basic Local Alignm...
blast.ncbi.nlm.nih.gov/Blast.cgi
BLAST®
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
http://ncbi.nlm.nih.gov/BLAST
NCBI/ BLAST Home
BLAST finds regions of similarity between biological sequences. more...
BLAST Assembled Genomes
Find Genomic BLAST pages:
Enter organism name or id--completions will be suggested
GO
Human
Mouse
Rat
Cow
Pig
Dog
Rabbit
Chimp
Guinea pig
Fruit fly
Honey bee
Chicken
Zebrafish
Clawed frog
Arabidopsis
Rice
Yeast
Microbes
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, delta-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
Specialized BLAST
Choose a type of specialized search (or database name in parentheses.)
Get faster protein results with a graphical view using SmartBLAST
Make specific primers with Primer-BLAST
Cluster multiple sequences together with their database neighbors using MOLE-BLAST
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 and T cell receptor sequences (IgBLAST)
Screen sequence for vector contamination (vecscreen)
Align two (or more) sequences using BLAST (bl2seq)
Search protein or nucleotide targets in PubChem BioAssay
Search SRA by experiment
Constraint Based Protein Multiple Alignment Tool
Needleman-Wunsch Global Sequence Alignment Tool
Search RefSeqGene
Your Recent Results New!
All Recent results...
News
Searching Whole Genome Shotgun sequences
It is now much easier to search WGS (Whole Genome Shotgun) with stand-alone BLAST on your own computer.
Wed, 20 Jan 2016 10:00:00 EST
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.
More tips...

Nucleotide BLAST: Search ...

blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&BLAST_SPEC=&LINK_LOC=blasttab&LAST_PAGE=blastp&BLAST_INIT=&QUERY=>NP_008872.1 SOX-10

Search

BLAST® *Basic Local Alignment Search Tool*

Home | Recent Results | Saved Strategies | Help

My NCBI [Sign In] [Register]

NCBI/ BLAST/ blastn suite

Standard Nucleotide BLAST

blastn | blastp | blastx | tblastn | tblastx

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

Reset page | Bookmark

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

Clear

Query subrange

From

To

Or, upload file Browse... No file selected.

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

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 | Exclude | +

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

Exclude Optional

Models (XM/XP) | Uncultured/environmental sample sequences

Limit to Optional

Sequences from type material

Entrez Query Optional

YouTube Create custom database

Enter an Entrez query to limit search

Program Selection

Optimize for

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

Choose a BLAST algorithm

BLAST

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

Show results in a new window

Nucleotide-Based BLAST Algorithms

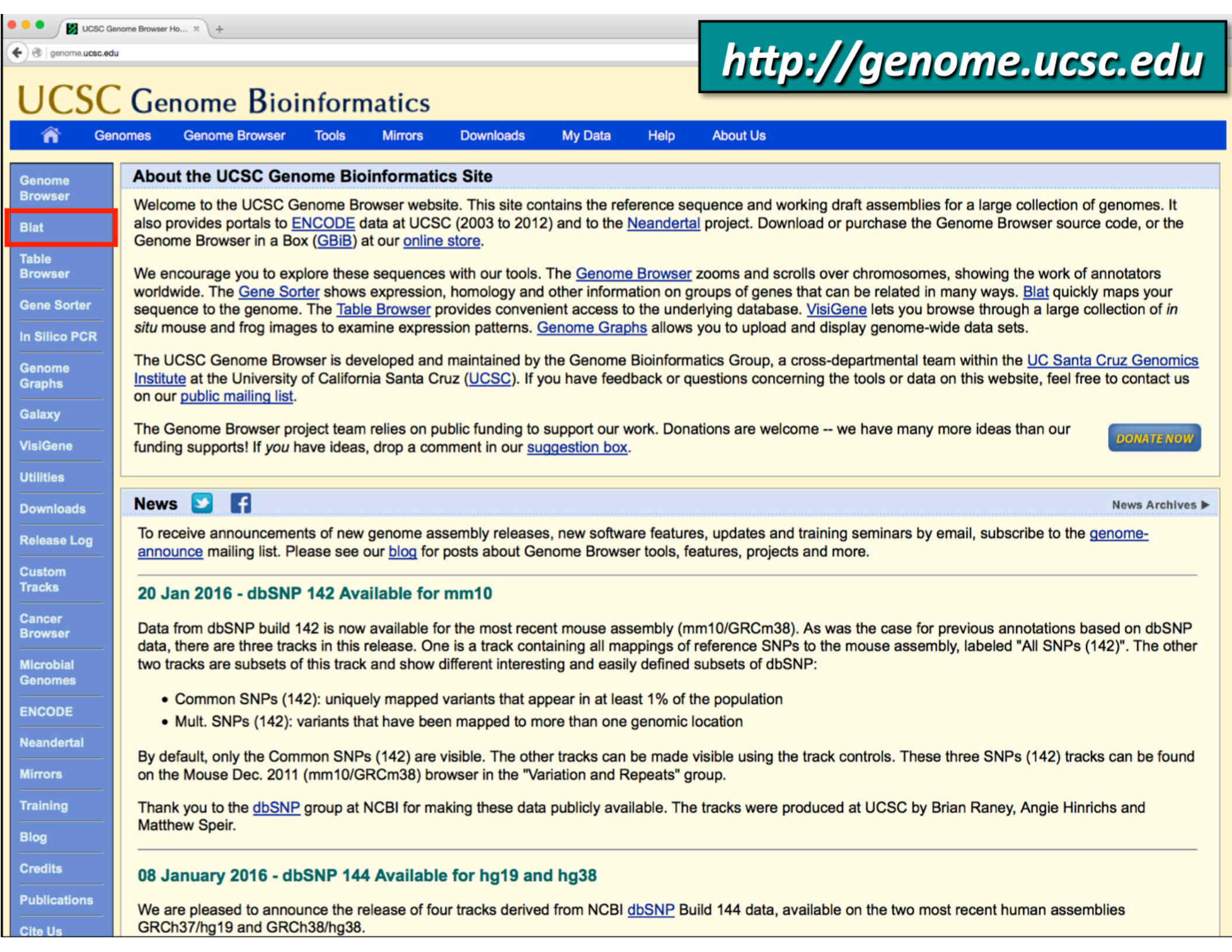
	W	*+/–*	*Gaps*
Optimized for aligning very long and/or highly similar sequences (> 95%)			
MegaBLAST *(default)*	28	1, –2	Linear
Better for diverged sequences and/or cross-species comparisons (< 80%)			
Discontiguous MegaBLAST	11	2, –3	Affine
BLASTN	11	2, –3	Affine
Finding short, nearly exact matches (< 20 bases)			
BLASTN	7	2, –3	Affine

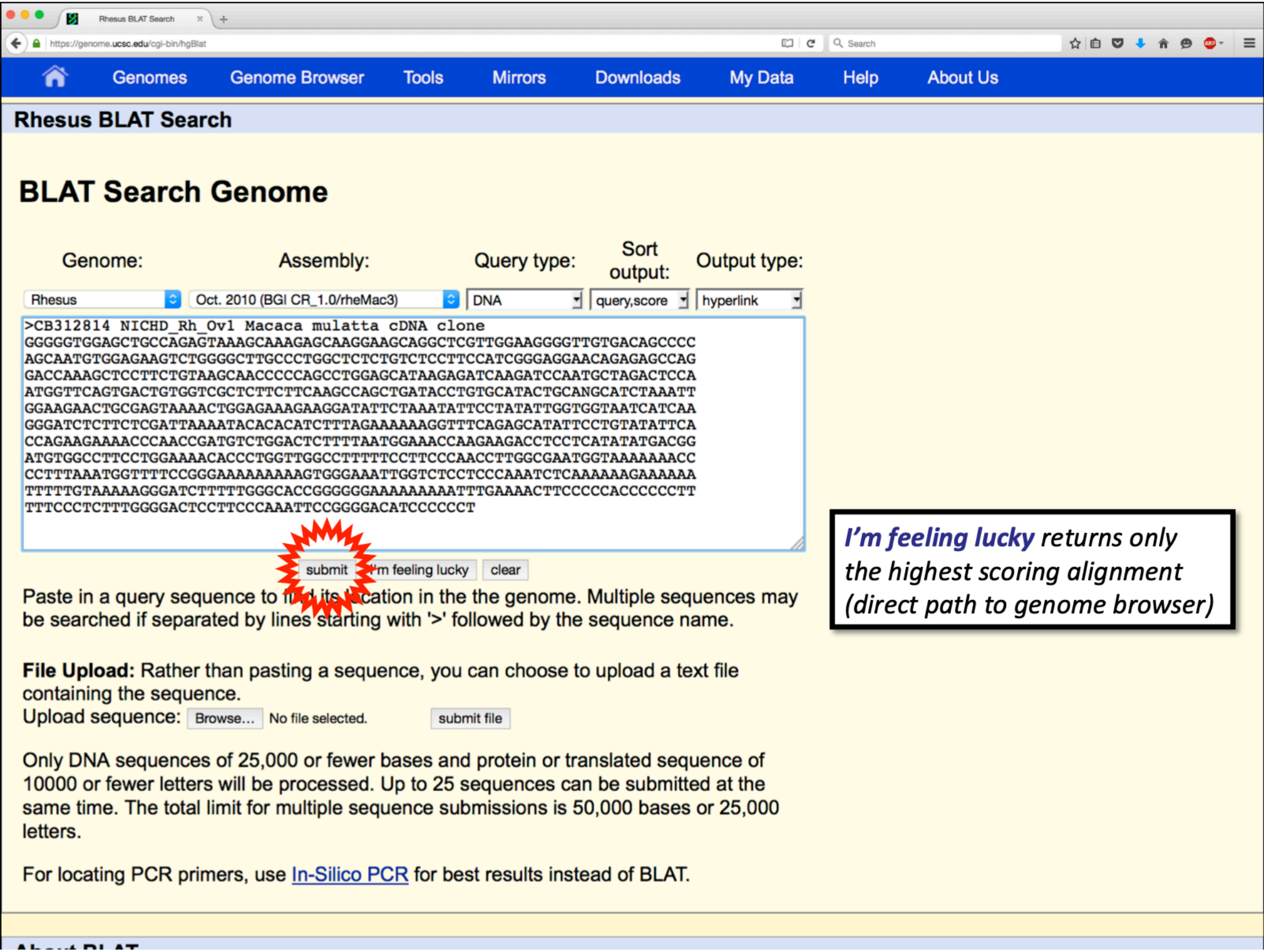
BLAT

- "BLAST-Like Alignment Tool"
- Designed to rapidly align longer nucleotide sequences ($L \geq 40$) having ≥ 95% sequence similarity
- Can find exact matches reliably down to $L = 33$
- Method of choice when looking for exact matches in nucleotide databases
- 500 times faster than BLAST for mRNA/DNA searches
- May miss divergent or shorter sequence alignments
- Can be used on protein sequences, but BLASTP is more efficient

When to Use BLAT

- To characterize an unknown gene or sequence fragment
  - Find its genomic coordinates
  - Determine gene structure (the presence and position of exons)
  - Identify markers of interest in the vicinity of a sequence
- To find highly similar (or identical) sequences
  - Alignment of mRNA sequences onto a genome assembly
  - Identification of gene family members
  - Cross-species alignment to identify putative homologs
- To display a specific sequence as a separate track within the UCSC Genome Browser


UCSC Genome Browser Ho...
genome.ucsc.edu
http://genome.ucsc.edu
UCSC Genome Bioinformatics
Genomes
Genome Browser
Tools
Mirrors
Downloads
My Data
Help
About Us
Genome Browser
Blat
Table Browser
Gene Sorter
In Silico PCR
Genome Graphs
Galaxy
VisiGene
Utilities
Downloads
Release Log
Custom Tracks
Cancer Browser
Microbial Genomes
ENCODE
Neandertal
Mirrors
Training
Blog
Credits
Publications
Cite Us
About the UCSC Genome Bioinformatics Site
Welcome to the UCSC Genome Browser website. This site contains the reference sequence and working draft assemblies for a large collection of genomes. It also provides portals to ENCODE data at UCSC (2003 to 2012) and to the Neandertal project. Download or purchase the Genome Browser source code, or the Genome Browser in a Box (GBiB) at our online store.
We encourage you to explore these sequences with our tools. The Genome Browser zooms and scrolls over chromosomes, showing the work of annotators worldwide. The Gene Sorter shows expression, homology and other information on groups of genes that can be related in many ways. Blat quickly maps your sequence to the genome. The Table Browser provides convenient access to the underlying database. VisiGene lets you browse through a large collection of in situ mouse and frog images to examine expression patterns. Genome Graphs allows you to upload and display genome-wide data sets.
The UCSC Genome Browser is developed and maintained by the Genome Bioinformatics Group, a cross-departmental team within the UC Santa Cruz Genomics Institute at the University of California Santa Cruz (UCSC). If you have feedback or questions concerning the tools or data on this website, feel free to contact us on our public mailing list.
The Genome Browser project team relies on public funding to support our work. Donations are welcome -- we have many more ideas than our funding supports! If you have ideas, drop a comment in our suggestion box.
DONATE NOW
News
News Archives
To receive announcements of new genome assembly releases, new software features, updates and training seminars by email, subscribe to the genome-announce mailing list. Please see our blog for posts about Genome Browser tools, features, projects and more.
20 Jan 2016 - dbSNP 142 Available for mm10
Data from dbSNP build 142 is now available for the most recent mouse assembly (mm10/GRCm38). As was the case for previous annotations based on dbSNP data, there are three tracks in this release. One is a track containing all mappings of reference SNPs to the mouse assembly, labeled "All SNPs (142)". The other two tracks are subsets of this track and show different interesting and easily defined subsets of dbSNP:
Common SNPs (142): uniquely mapped variants that appear in at least 1% of the population
Mult. SNPs (142): variants that have been mapped to more than one genomic location
By default, only the Common SNPs (142) are visible. The other tracks can be made visible using the track controls. These three SNPs (142) tracks can be found on the Mouse Dec. 2011 (mm10/GRCm38) browser in the "Variation and Repeats" group.
Thank you to the dbSNP group at NCBI for making these data publicly available. The tracks were produced at UCSC by Brian Raney, Angie Hinrichs and Matthew Speir.
08 January 2016 - dbSNP 144 Available for hg19 and hg38
We are pleased to announce the release of four tracks derived from NCBI dbSNP Build 144 data, available on the two most recent human assemblies GRCh37/hg19 and GRCh38/hg38.



I'm feeling lucky *returns only the highest scoring alignment (direct path to genome browser)*

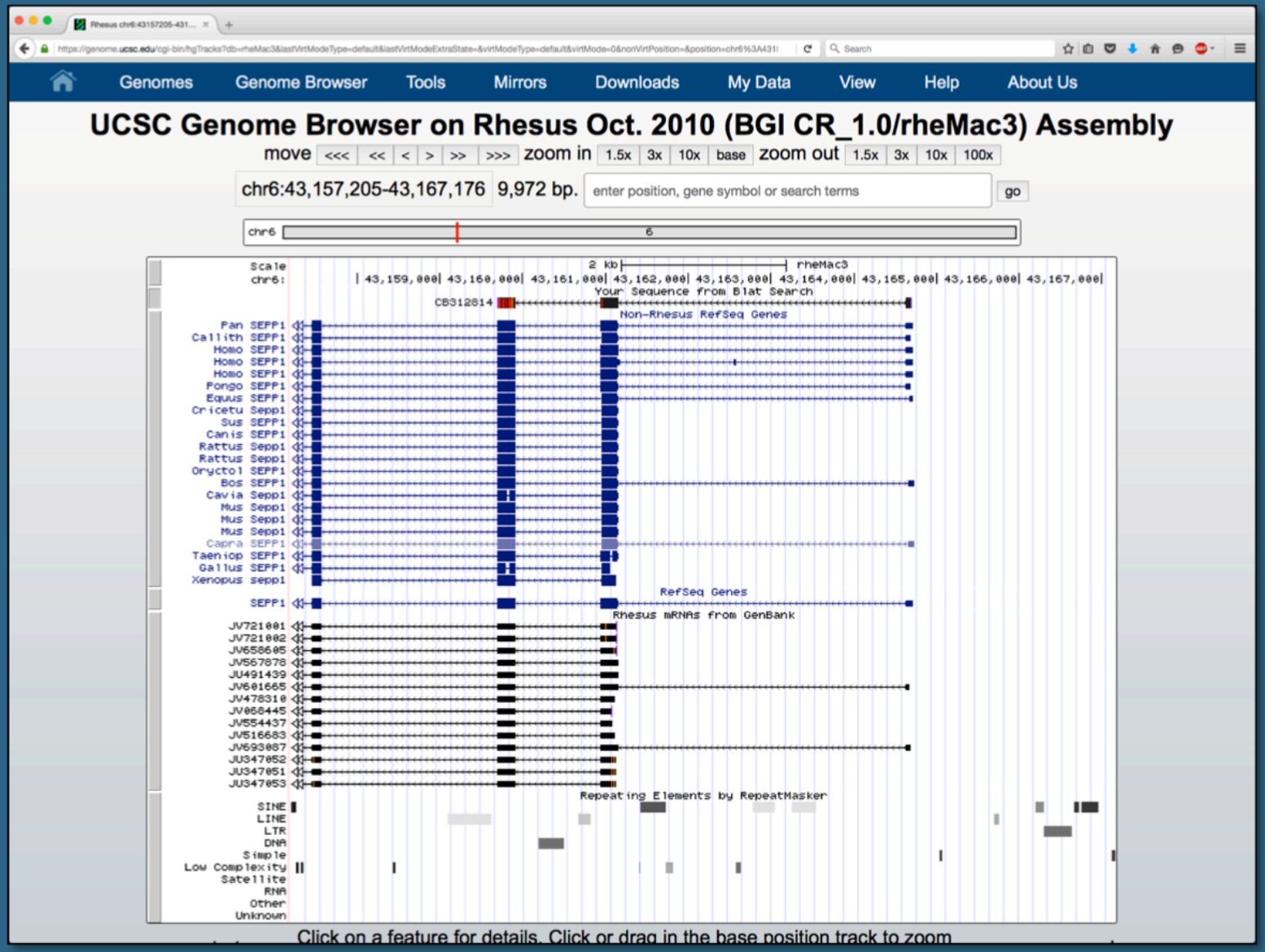
Rhesus BLAT Results

BLAT Search Results

Go back to chr6:43159698-43164683 on the Genome Browser.

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
browser details	CB312814	380	1	418	677	96.2%	6	-	43159698	43161152	1455
browser details	CB312814	23	591	613	677	100.0%	4	-	148338464	148338486	23
browser details	CB312814	22	546	567	677	100.0%	12	-	39379930	39379951	22
browser details	CB312814	21	628	648	677	100.0%	16	+	20696166	20696186	21
browser details	CB312814	21	629	651	677	95.7%	1	+	134928210	134928232	23
browser details	CB312814	20	553	574	677	95.5%	11	-	4332856	4332877	22
browser details	CB312814	20	627	646	677	100.0%	1	-	187748214	187748233	20
browser details	CB312814	20	511	530	677	100.0%	1	-	90178654	90178673	20

Missing a match?



- **red:** Genome and query sequence have different bases at this position.
- **orange:** The query sequence has an insertion (or genome has a deletion / alignment gap) at this point.
- **purple:** The query sequence extends beyond the end of the alignment.
- **green:** The query sequence appears to have a polyA tail which is not aligned to the genome.

Rhesus BLAT Results

BLAT Search Results

Go back to chr6:43159698-43164683 on the Genome Browser.

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
browser details	CB312814	380	1	418	677	96.2%	6	-	43159698	43161152	1455
browser details	CB312814	23	591	613	677	100.0%	4	-	148338464	148338486	23
browser details	CB312814	22	546	567	677	100.0%	12	-	39379930	39379951	22
browser details	CB312814	21	628	648	677	100.0%	16	+	20696166	20696186	21
browser details	CB312814	21	629	651	677	95.7%	1	+	134928210	134928232	23
browser details	CB312814	20	553	574	677	95.5%	11	-	4332856	4332877	22
browser details	CB312814	20	627	646	677	100.0%	1	-	187748214	187748233	20
browser details	CB312814	20	511	530	677	100.0%	1	-	90178654	90178673	20

Missing a match?

Alignment of CB312814 and chr6:43159698-43161152

Click on links in the frame to the left to navigate through the alignment. Matching bases in cDNA and genomic sequences are colored blue and capitalized. Light blue bases mark the boundaries of gaps in either sequence (often splice sites).

---

cDNA CB312814

```
AGCAATGTGG AGAAGTCTGG GGCTTGCCCT GGCTCTCTGT CTCCTTCCAT  50
CGGGAGGAAC AGAGAGCCAG GACCAAAGCT CCTTCTGTAA GCAACCCCCA  100
GCCTGGAGCA TAAGAGATCA AGATCCAATG CTAGACTCCA ATGGTTCAGT  150
GACTGTGGTC GCTCTTCTTC AAGCCAGCTG ATACCTGTGC ATACTGCAnG  200
CATCTAAATT GGAAGAACTG CGAGTAAAAC TGGAGAAAGA AGGATATTCT  250
AAaTATTcCT ATATTGgTGg TAATCATCAA GGgATCTCTT CTCGATTAAA  300
ATACACACAT CTTtAGAAaa AAGGTTTCAG AGCATATTCC TGTATATtcA  350
CcAGAAGAAA ACCcAACcGA TGTCTGGACT CTTTTAATGG AAaCcAAGAa  400
GACcTCCTCA TATATGACgg atgtggcctt cctggaaaac accctggttg  450
gccttttcc ttcccaacct tggcgaatgg taaaaaaacc cctttaaatg  500
gttttccggg aaaaaaaaag tgggaaattg gtctcctccc aaatctcaaa  550
aaagaaaaaa tttttgtaaa aagggatctt tttgggcacc gggggggaaaa  600
aaaaatttga aaacttcccc caccccccctt tttccctctt tggggactcc  650
ttcccaaatt ccggggacat cccccct
```

Genomic chr6 (reverse strand):

```
agtggaatta tgtctgcagg atttatagaa attcatagtt aggactgtga  43161203
agttaactat gaagaagagt gacaggtttt ctcttttaca ggacagcccc  43161153
AGCAATGTGG AGAAGTCTGG GGCTTGCCCT GGCTCTCTGT CTCCTTCCAT  43161103
CGGGAGGAAC AGAGAGCCAG GACCAAAGCT CCTTCTGTAA GCAACCCCCA  43161053
GCCTGGAGCA TAAGAGATCA AGATCCAATG CTAGACTCCA ATGGTTCAGT  43161003
GACTGTGGTC GCTCTTCTTC AAGCCAGCTG ATACCTGTGC ATACTGCAgG  43160953
CATCTAAgta agacagtctt tctgtggctt aaaatacctt aaggggaagg  43160903
ttattagata cacacatgca tatagacata aaagtgtaaa caataattta  43160853
agtcacactt ttggaaaaac tatgtgtttg cacagaacat attagaagag  43160803
agaaactagg atgatacaca ccaaaatgtt tacagtgggt ttcctagggt  43160753
tatggaaatt ttttcttctt tttgtggacc tatattttat aacttcctac  43160703
taaatatgta ttacttgtgt aatgaaaatg ttctaaaatg tacttctgca  43160653
aatagaacag ttacttcaat acaagaagca gagaagactt ttgtcagagt  43160603
agaaagaaca ctaggcttgc tatcaaaggt tttgggttat ttaacaaata  43160553
aatatcaaat atatttgttg agttattcac cagatattga ttgggtggaa  43160503
ttctccttac agtcatccaa tggtatctgt gcaggattgg ttccaggatc  43160453
cccctcacac accaaaatcc atggagctca aatcccttat ataaaatggc  43160403
atatataacc tatgcacata ttttcacatg ttttaatcat ctctagatta  43160353
cttataatac ctaacagaat gtaaatgcta tgtaagtaat ttttatactg  43160303
tattgtttag tgaataatga atgacatttt aaaaagtcta catgttcact  43160253
```

Alignment of CB312814


```
tgtta
```

Side by Side Alignment

```
00000001 agcaatgtggagaagtctggggcttgccctggctctctgtctccttccat 00000050
<<<<<<<< |||||||||||||||||||||||||||||||||||||||||||||||||| <<<<<<<<
43161152 agcaatgtggagaagtctggggcttgccctggctctctgtctccttccat 43161103

00000051 cgggaggaacagagagccaggaccaaagctccttctgtaagcaaccccca 00000100
<<<<<<<< |||||||||||||||||||||||||||||||||||||||||||||||||| <<<<<<<<
43161102 cgggaggaacagagagccaggaccaaagctccttctgtaagcaaccccca 43161053

00000101 gcctggagcataagagatcaagatccaatgctagactccaatggttcagt 00000150
<<<<<<<< |||||||||||||||||||||||||||||||||||||||||||||||||| <<<<<<<<
43161052 gcctggagcataagagatcaagatccaatgctagactccaatggttcagt 43161003

00000151 gactgtggtcgctcttcttcaagccagctgatacctgtgcatactgcang 00000200
<<<<<<<< |||||||||||||||||||||||||||||||||||||||||||||||||  <<<<<<<<
43161002 gactgtggtcgctcttcttcaagccagctgatacctgtgcatactgcagg 43160953

00000201 catctaa 00000207
<<<<<<<< ||||||| <<<<<<<<
43160952 catctaa 43160946
```

```
00000208 attggaagaactgcgagtaaaactggagaaagaaggatattctaaatatt 00000257
<<<<<<<< |||||||||||||||||||||||||||||||||||||||||||||||||| |||| <<<<<<<<
43159908 attggaagaactgcgagtaaaactggagaaagaaggatattctaa.tatt 43159860

00000258 cc.tatattggtggtaatcatcaagggatctcttctcgattaaaatacac 00000306
<<<<<<<<  | ||||||| || |||||||||||| |||||||||||||||||||||| <<<<<<<<
43159859 tcttatattgttgttaatcatcaaggaatctcttctcgattaaaatacac 43159810

00000307 acatctttagaa 00000318
<<<<<<<< ||||||| |||| <<<<<<<<
43159809 acatcttaagaa 43159798
```

```
00000321 aaggtttcagagcatattcctgtatattcaccagaagaaaacccaaccga 00000370
<<<<<<<< |||||||||||||||||||||||||  || |||||||||||| ||| || <<<<<<<<
43159796 aaggtttcagagcatattcctgtatatcaacaagaagaaaaccaaacaga 43159747

00000371 tgtctggactcttttaa.tggaaaccaagaagacctcctcatatatgac 00000418
<<<<<<<< |||||||||||||||| ||||| | |||| ||| |||||||||||||| <<<<<<<<
43159746 tgtctggactcttttaaatggaagcaaagatgacttcctcatatatgac 43159698
```

**Aligned Blocks with gaps <= 8 bases are merged for this display when only one sequence has a gap, or when gaps in both sequences are of the same size.*