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.

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES | NATIONAL INSTITUTES OF HEALTH | genome.gov/DIR

JOHNS HOPKINS MEDICINE
CONTINUING MEDICAL EDUCATION

Current Topics in Genome Analysis 2016

Andy Baxevanis, Ph.D.

No Relevant Financial Relationships with Commercial Interests

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

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

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

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

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Defining the Terms

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

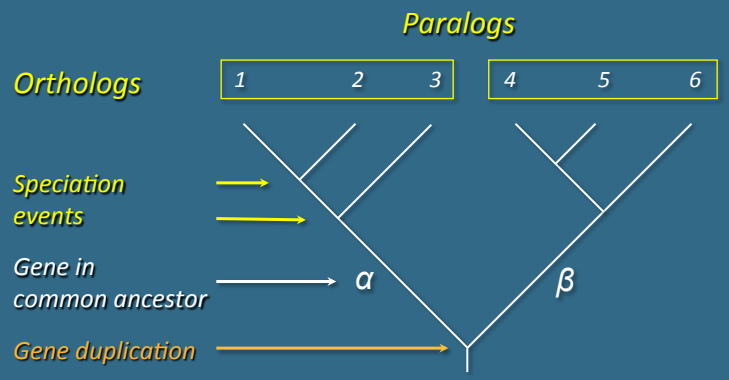
- Sequences are direct descendants of a sequence in a common ancestor (share a common origin)
- Most likely have similar domain and three-dimensional structure
- Usually retain same biological function over evolutionary time
- Can be used to predict gene function in novel genomes

Defining the Terms

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

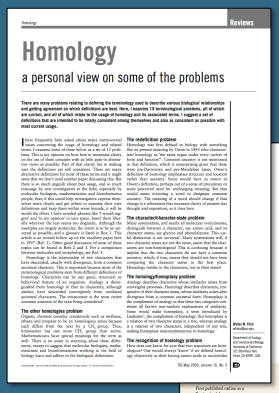
- Perhaps less likely to perform similar functions
- Can take on new functions over evolutionary time
- Provides insight into 'evolutionary innovation'

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)

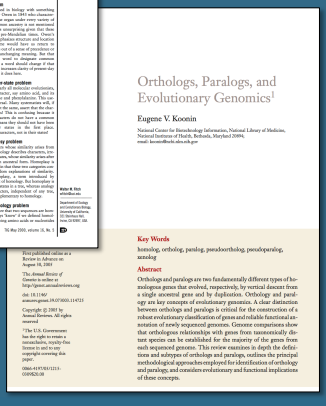
Orthology and Paralogy: Further Reading


Reviews

Homology

a personal view on some of the problems

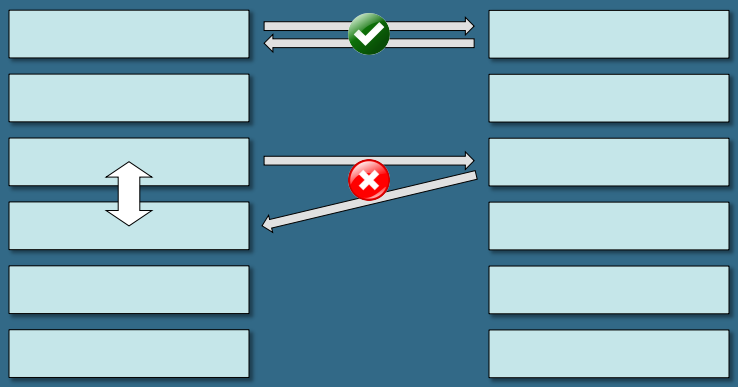
Walter Fitch
Trends Genet.
16: 227-231, 2000


Orthologs, Paralogs, and Evolutionary Genomics[1]

Eugene V. Koonin

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

Identifying Candidate Orthologs: Reciprocal Best Hits



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

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

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?

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

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

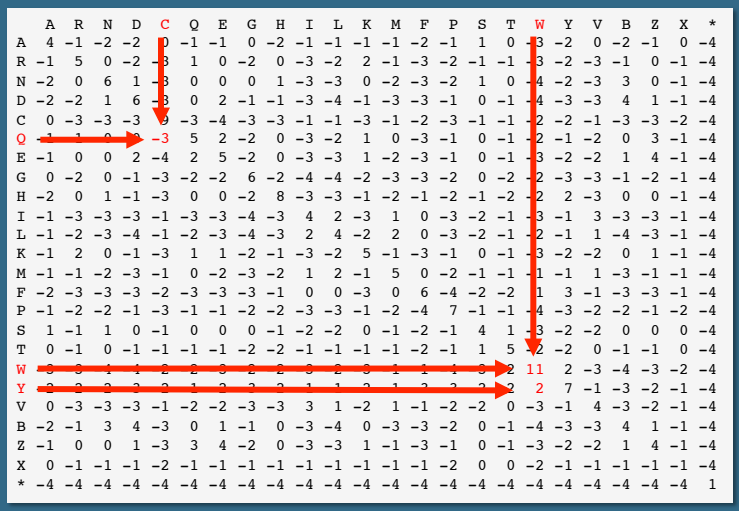
- Simple match/mismatch scoring scheme:

  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

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

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 4.3*). 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 1.1* 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 matrices 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

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

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

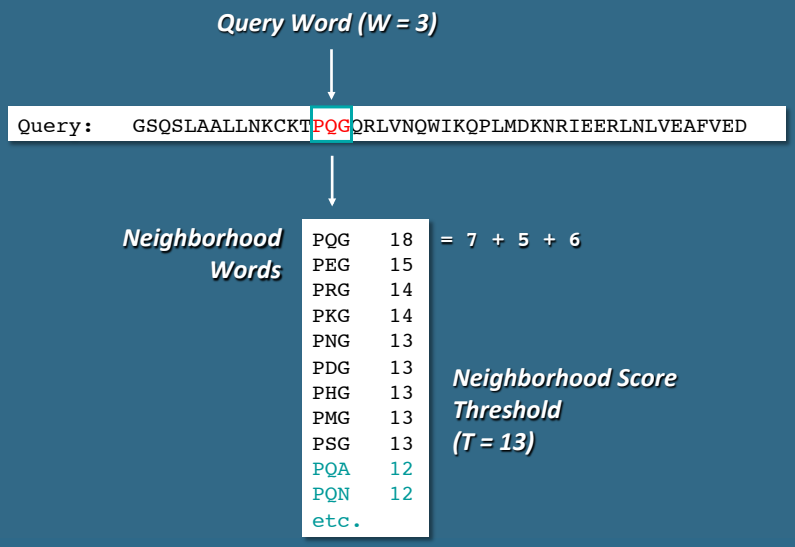
Altschul et al., J. Mol. Biol. 215: 403-410, 1990

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

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Neighborhood Words



NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

High-Scoring Segment Pairs

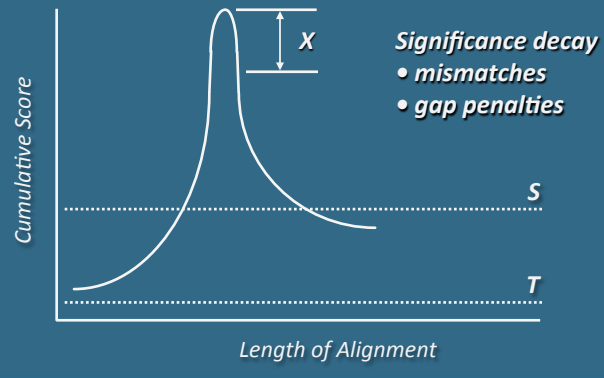
PQG	18
PEG	15
PRG	14
PKG	14
PNG	13
PDG	13
PHG	13
PMG	13
PSG	13
PQA	12
PQN	12
etc.	


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

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Extension

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



Significance decay
- *mismatches*
- *gap penalties*

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Scores and Alignment Length Don't Tell the Whole Story

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

Score = 91

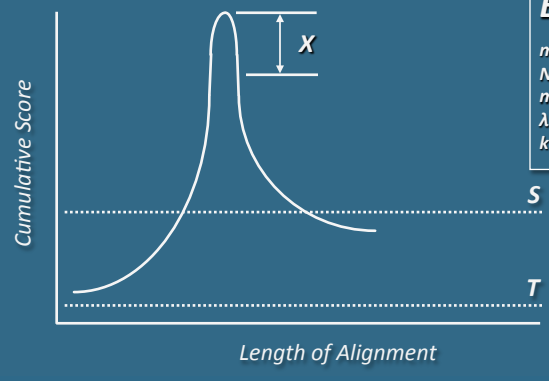
```
Query: 1 CQHMWYQWMIQCIWMYHCMQ 20
         CQHMWYQWMIQCIWMYHCMQ
Sbjct: 1 CQHMWYQWMIQCIWMYHCMQ 20
```

Score = 138

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Scores and Probabilities

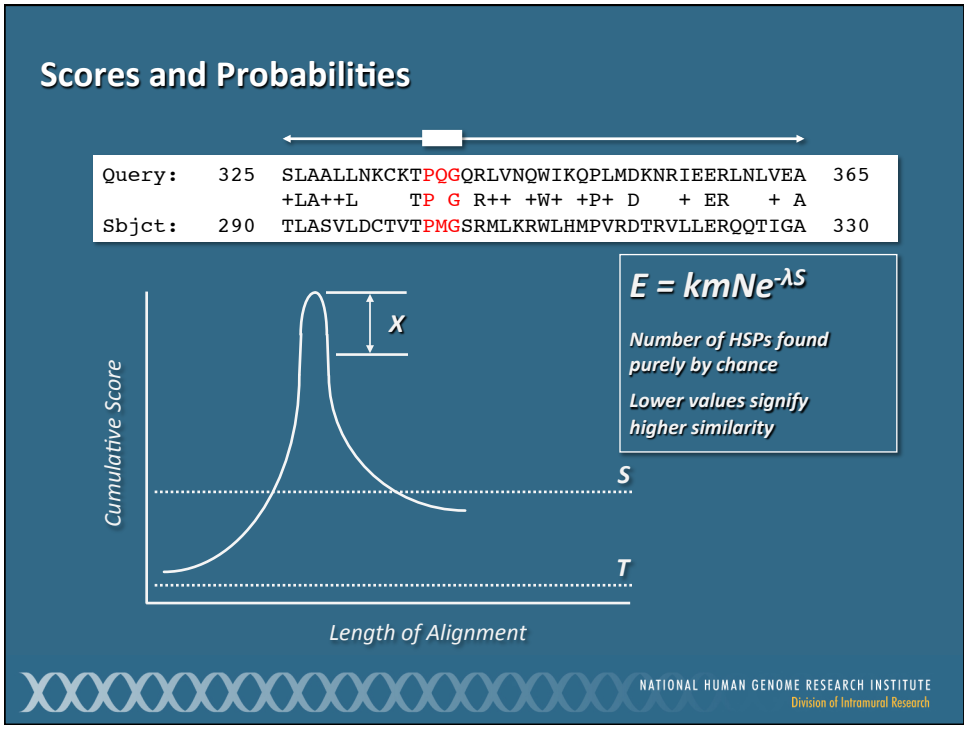
```
Query:   325 SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA 365
             +LA++L    TP G R++ +W+ +P+ D   + ER   + A
Sbjct:   290 TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA 330
```

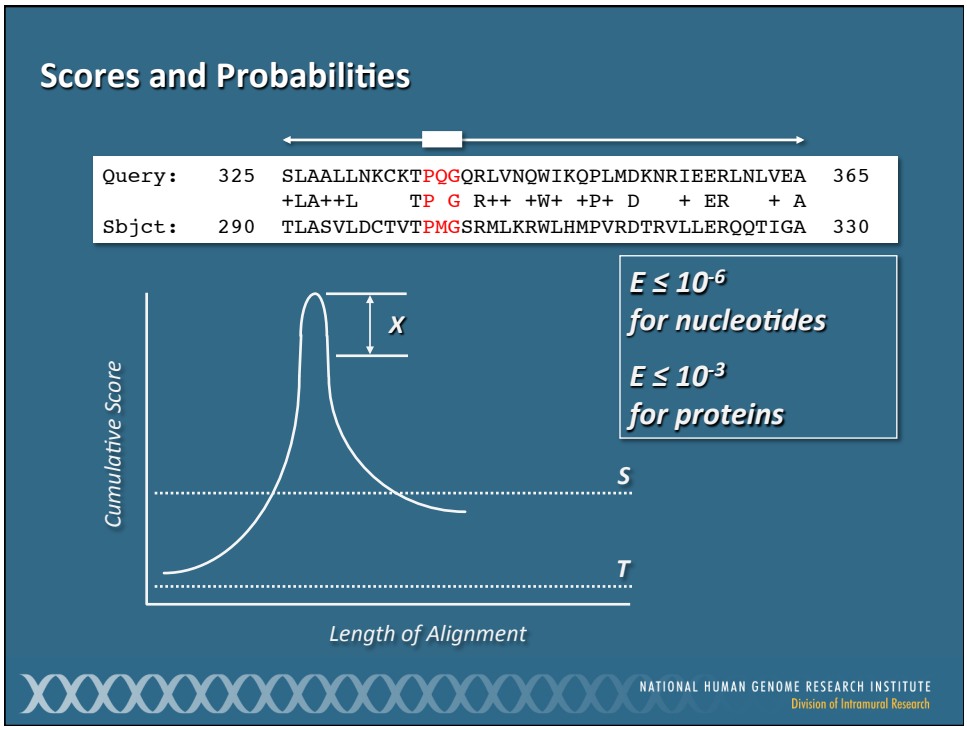


$E = kmNe^{-\lambda S}$

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

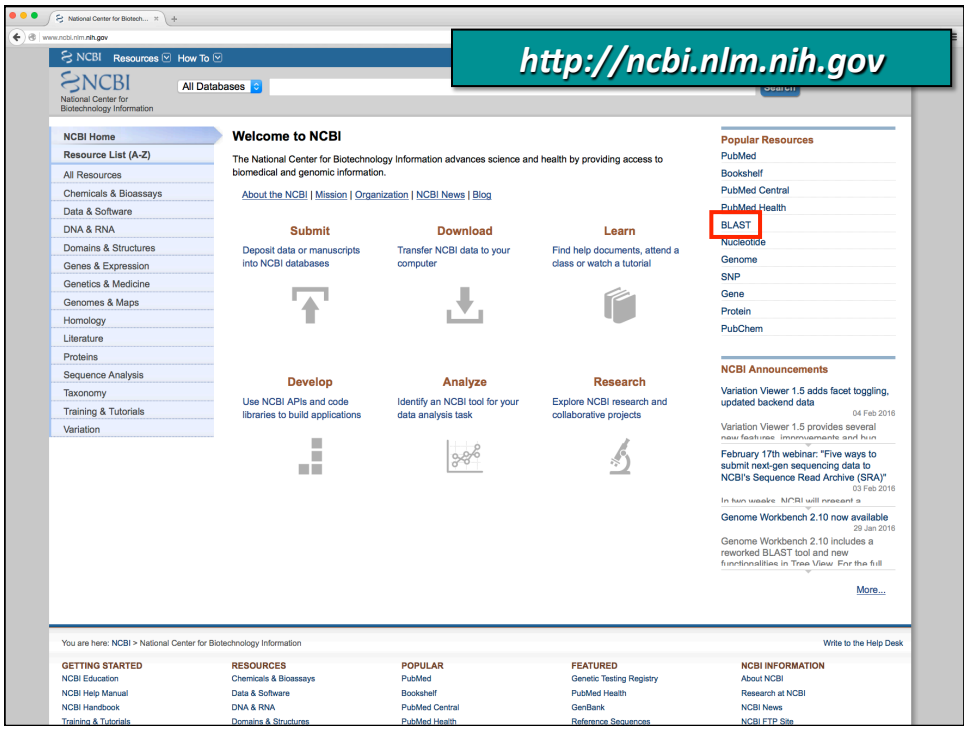
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

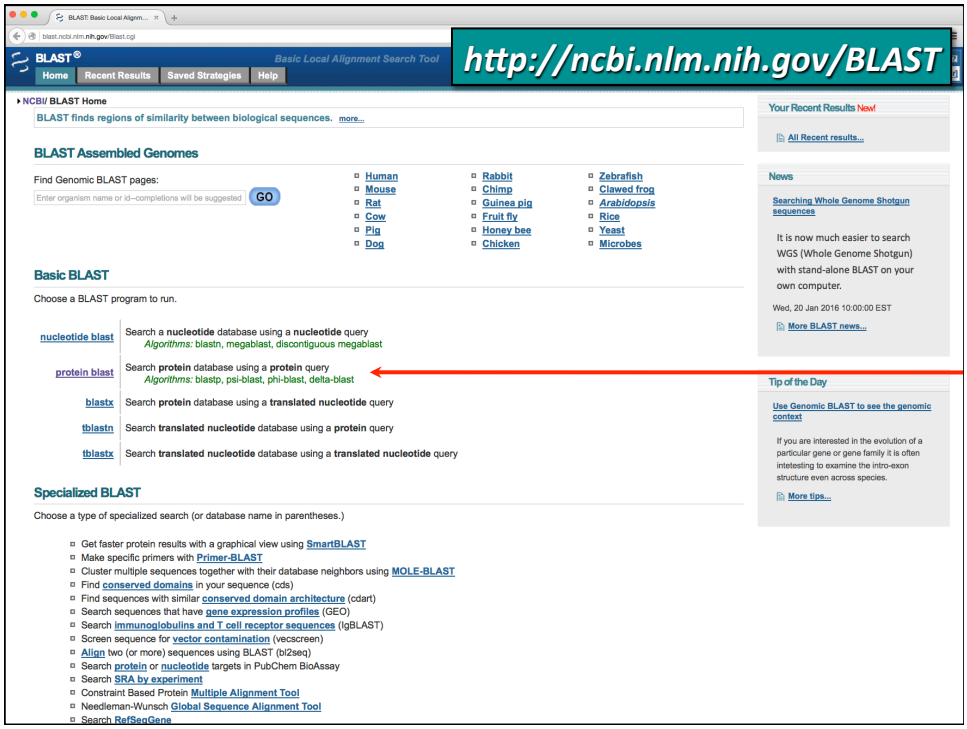

Scores and Probabilities
Query: 325 SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA 365
+LA++L TP G R++ +W+ +P+ D + ER + A
Sbjct: 290 TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA 330
E = kmNe-λS
Number of HSPs found purely by chance
Lower values signify higher similarity
X
S
T
Cumulative Score
Length of Alignment
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

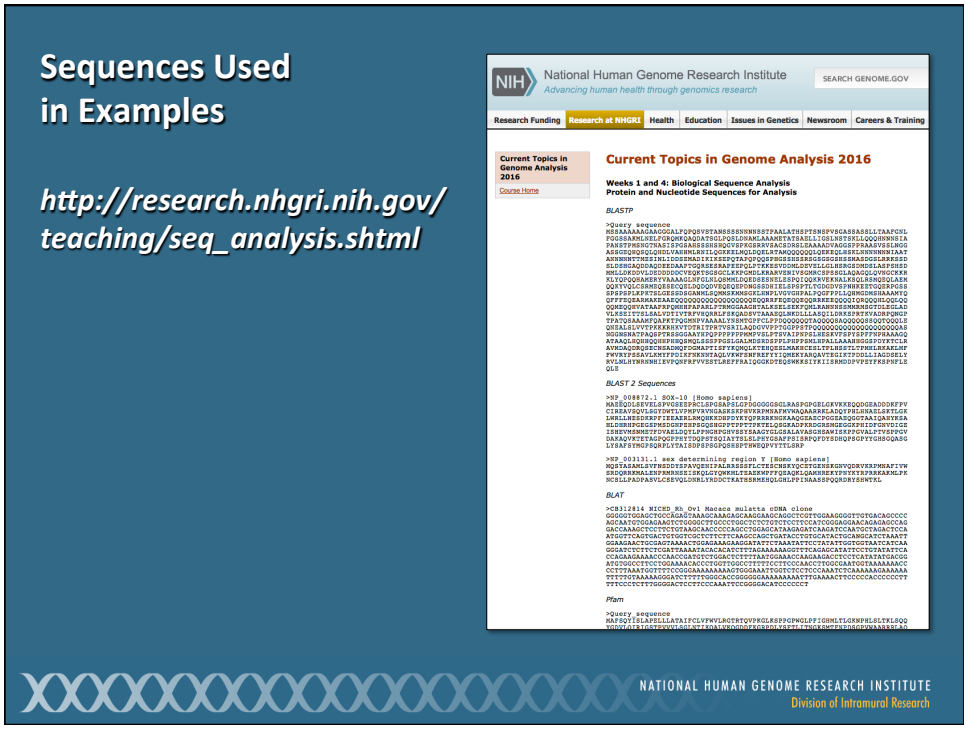

Scores and Probabilities
Query: 325 SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA 365
+LA++L TP G R++ +W+ +P+ D + ER + A
Sbjct: 290 TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA 330
E ≤ 10-6
for nucleotides
E ≤ 10-3
for proteins
X
S
T
Cumulative Score
Length of Alignment
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

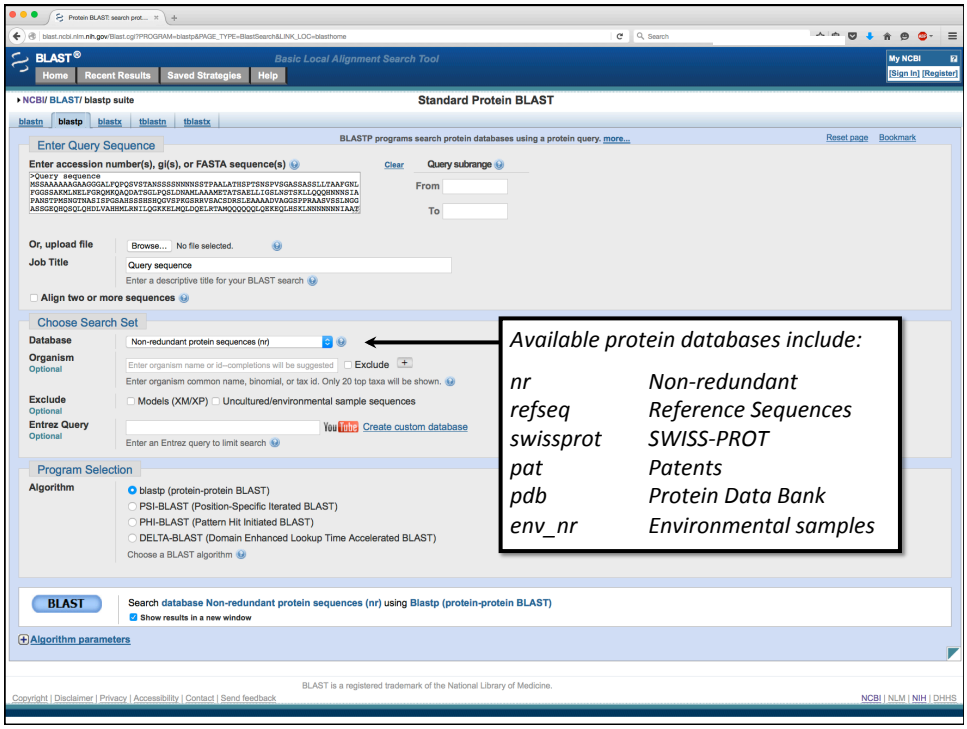
Using BLAST for Protein Similarity Searching

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research




http://ncbi.nlm.nih.gov/BLAST
BLAST®
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
NCBI/ BLAST Home
BLAST finds regions of similarity between biological sequences. more...
BLAST Assembled Genomes
Find Genomic BLAST pages:
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 interesting 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
BLAST 2 Sequences
>NP_008872.1 SOX-10 [Homo sapiens]
>NP_001131.1 sex determining region Y [Homo sapiens]
BLAT
>CB312814 NICED_Rh_Ov1 Macaca mulatta cDNA clone
Pfam
>Query sequence
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



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

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

RefSeq Accession Number Prefixes

From curation of GenBank entries:

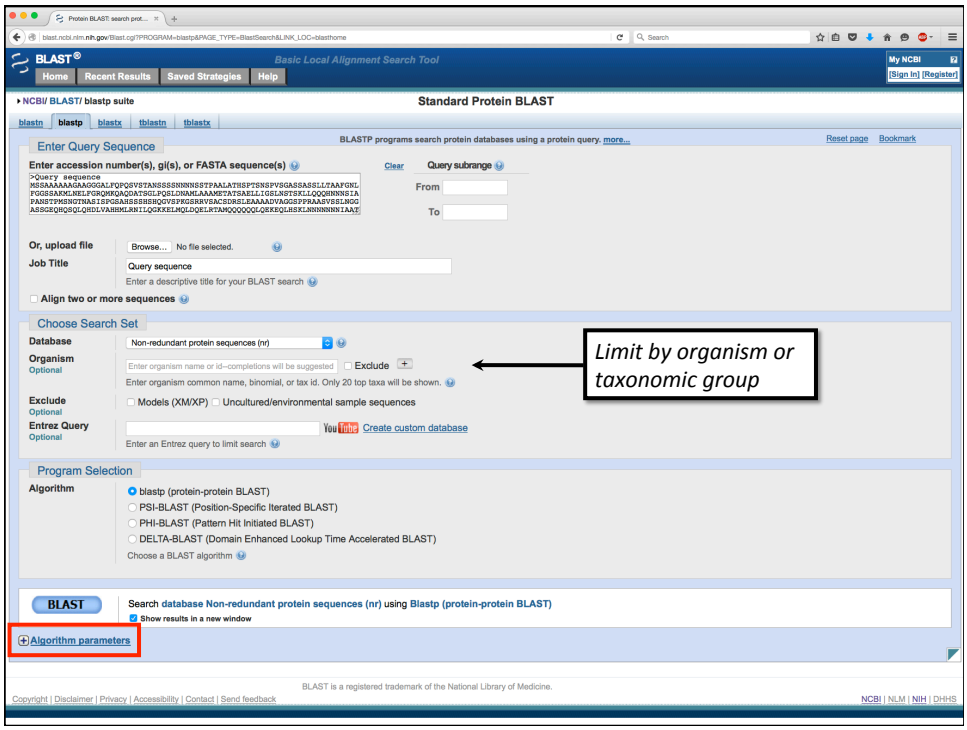
NT_	Genomic contigs
NM_	mRNAs
NP_	Proteins
NR_	Non-coding transcripts

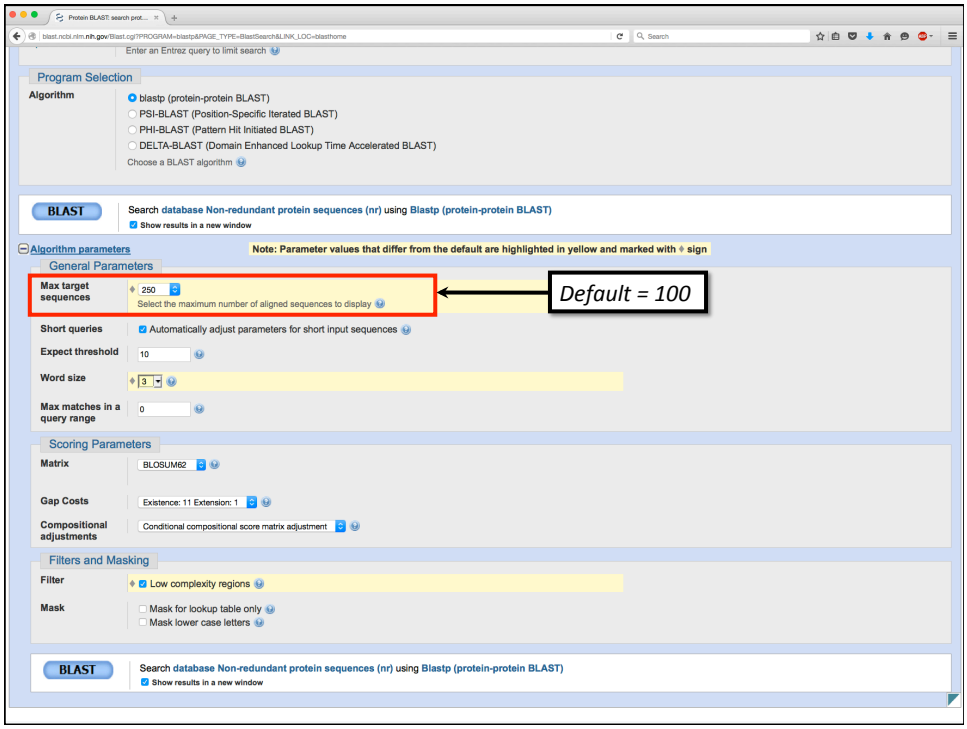
From genome annotation:

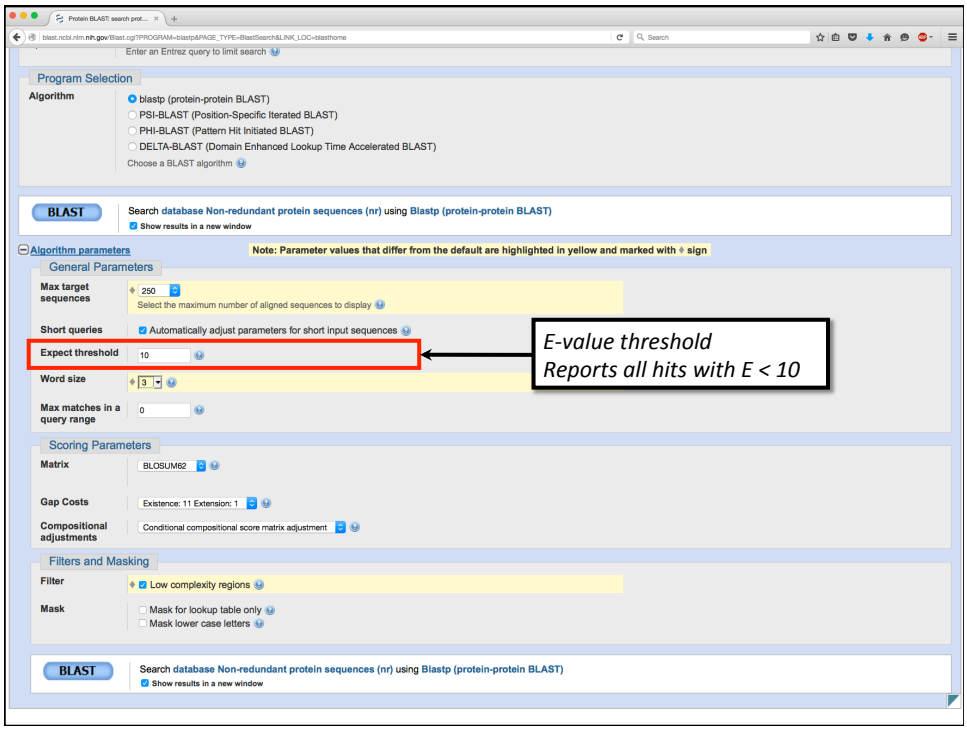
XM_	Model mRNA
XP_	Model proteins

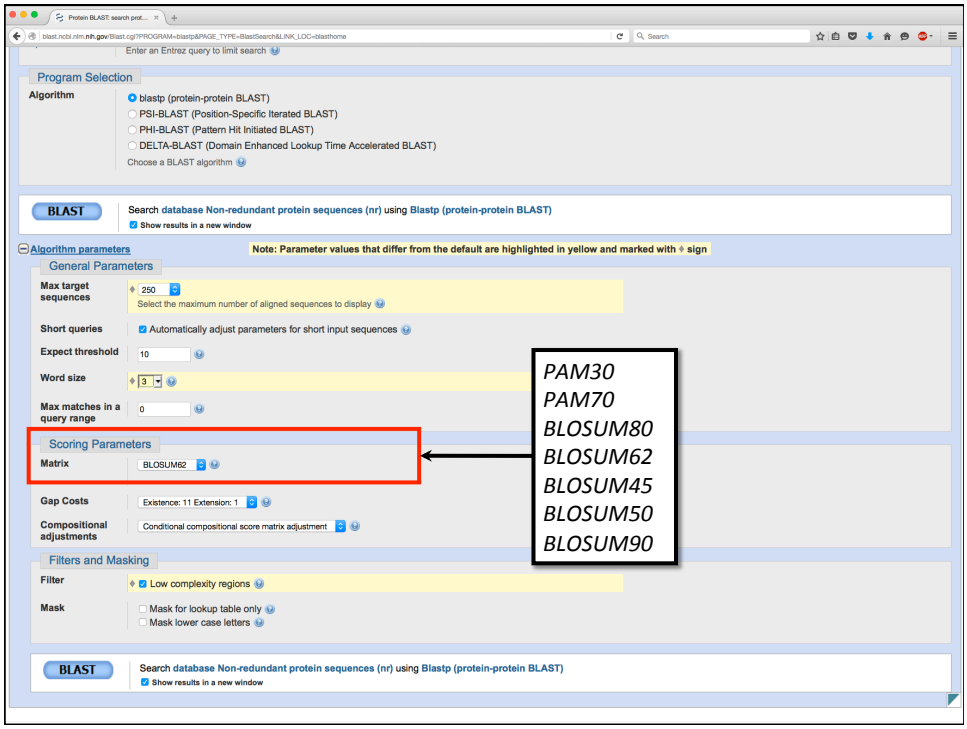
Complete list of molecule types in Chapter 18 of the NCBI Handbook
http://ncbi.nlm.nih.gov/books/NBK21091

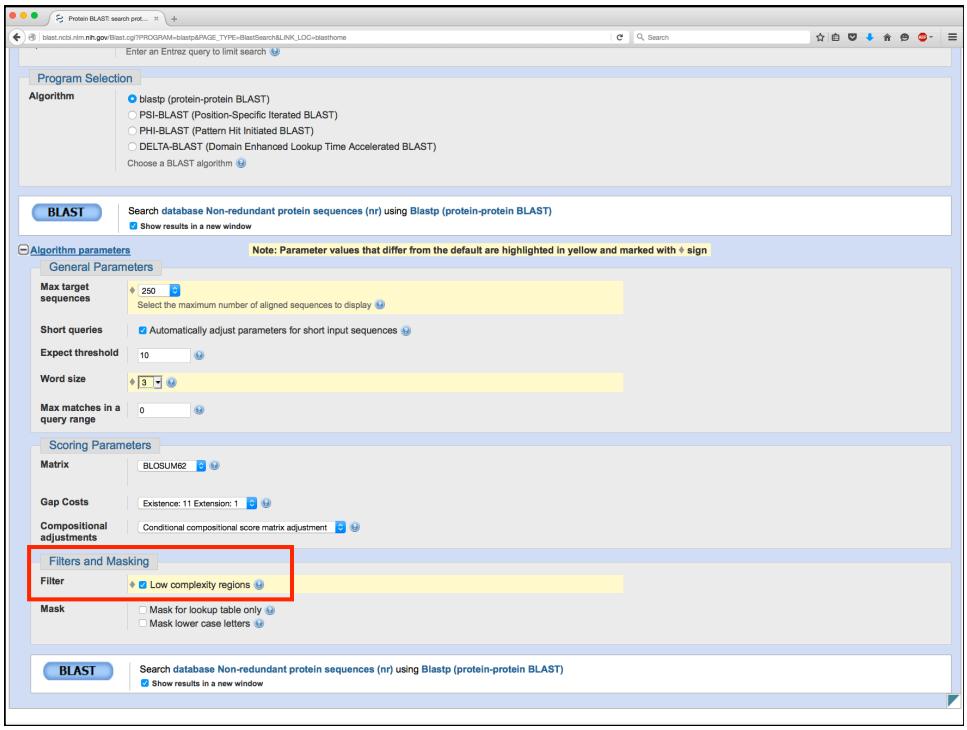
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research




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


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)
BLAST
Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)
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
E-value threshold
Reports all hits with E < 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

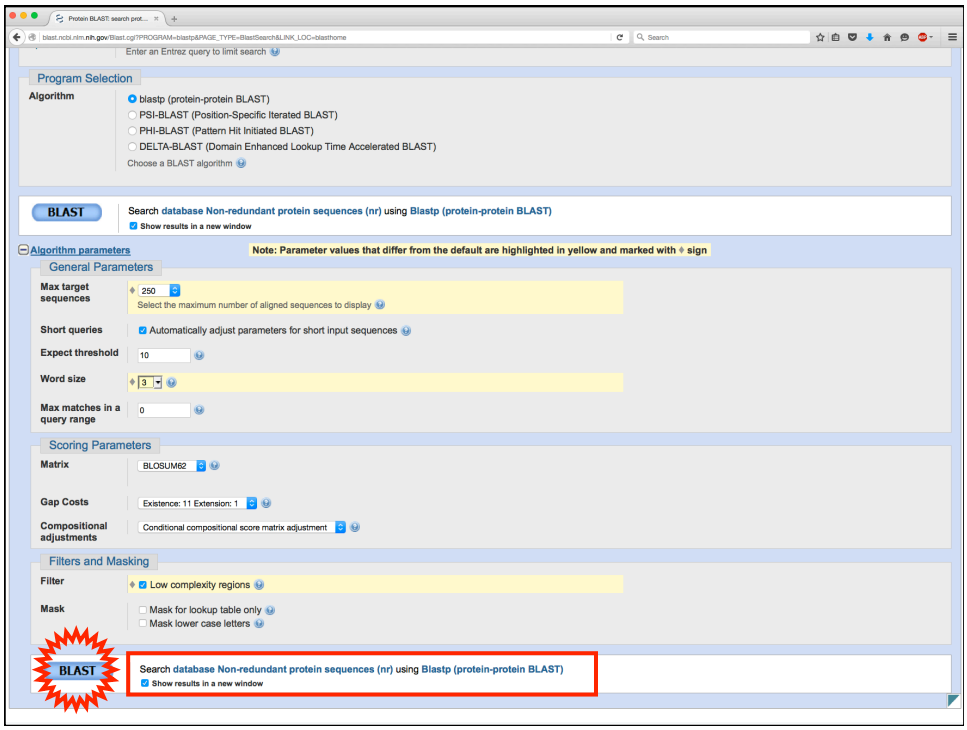

Program Selection
Algorithm
blastp (protein-protein BLAST)
PSI-BLAST (Position-Specific Iterated BLAST)
PHI-BLAST (Pattern Hit Initiated BLAST)
DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
BLAST
Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)
Algorithm parameters
Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign
General Parameters
Max target sequences
Short queries
Expect threshold
Word size
Max matches in a query range
Scoring Parameters
Matrix
BLOSUM62
Gap Costs
Compositional adjustments
Filters and Masking
Filter
Low complexity regions
Mask
Mask for lookup table only
Mask lower case letters
PAM30
PAM70
BLOSUM80
BLOSUM62
BLOSUM45
BLOSUM50
BLOSUM90

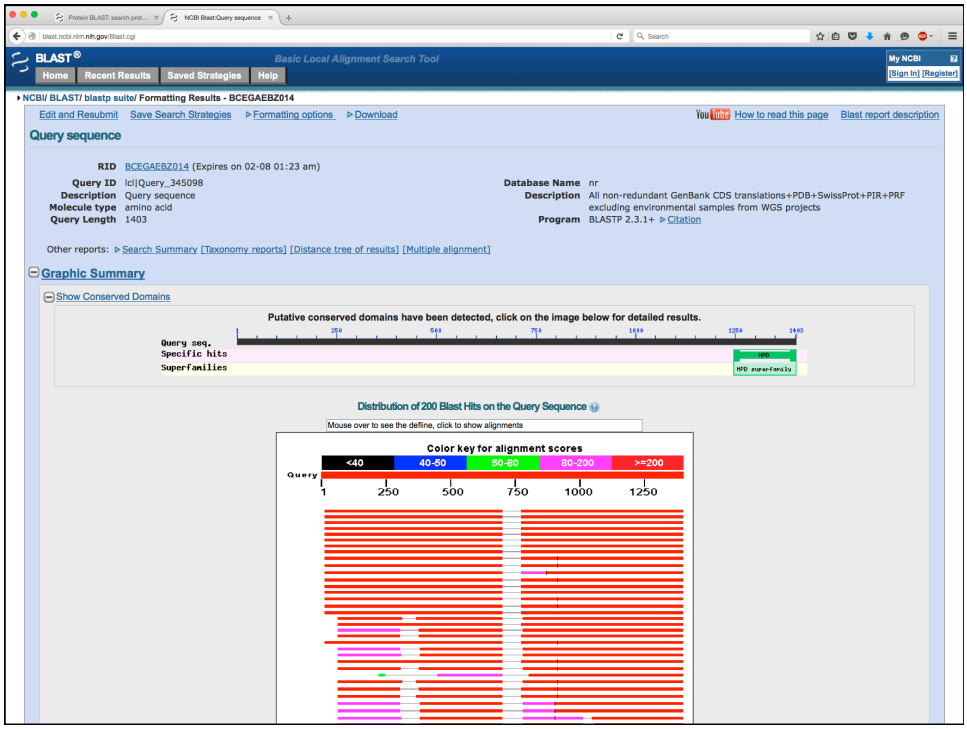

Program Selection
Algorithm
blastp (protein-protein BLAST)
PSI-BLAST (Position-Specific Iterated BLAST)
PHI-BLAST (Pattern Hit Initiated BLAST)
DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
BLAST
Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)
Algorithm parameters
Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign
General Parameters
Max target sequences
Short queries
Expect threshold
Word size
Max matches in a query range
Scoring Parameters
Matrix
BLOSUM62
Gap Costs
Compositional adjustments
Filters and Masking
Filter
Low complexity regions
Mask
Mask for lookup table only
Mask lower case letters

Low-Complexity Regions

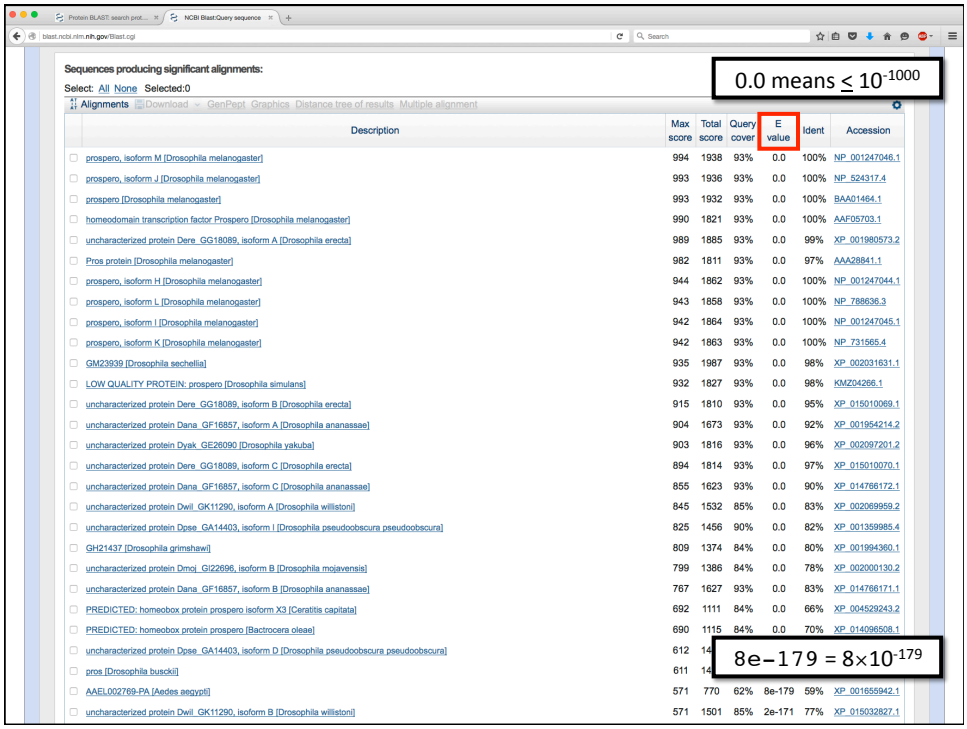
- Defined as regions of 'biased composition'
  - Homopolymeric runs
  - Short-period repeats
  - Subtle over-representation of several residues
- May confound sequence analysis
  - BLAST relies on uniformly-distributed amino acid frequencies
  - Often lead to false positives
- Filtering is advised (but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

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

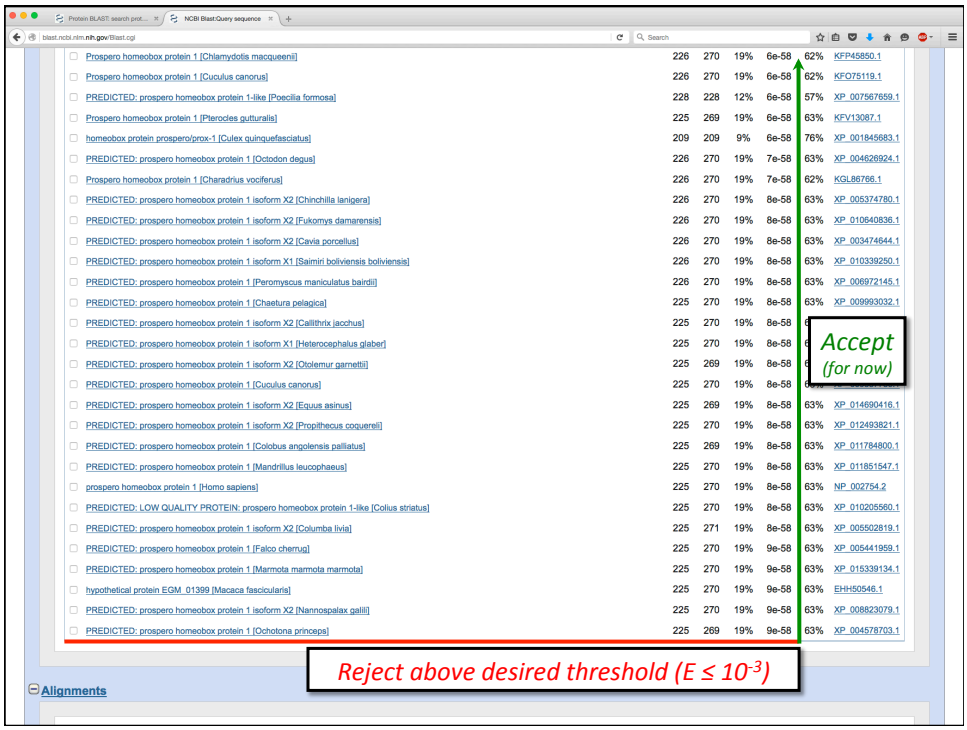



BLAST®
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
Edit and Resubmit
Save Search Strategies
Formatting options
Download
How to read this page
Blast report description
Query sequence
Graphic Summary
Show Conserved Domains
Putative conserved domains have been detected, click 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
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1
250
500
750
1000
1250


Distribution of 200 Blast Hits on the Query Sequence
Mouse-over to show define and scores, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1
250
500
750
1000
1250
Color key
Gap in alignment with subject
• >1 HSP
• Masked region
Descending score order
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



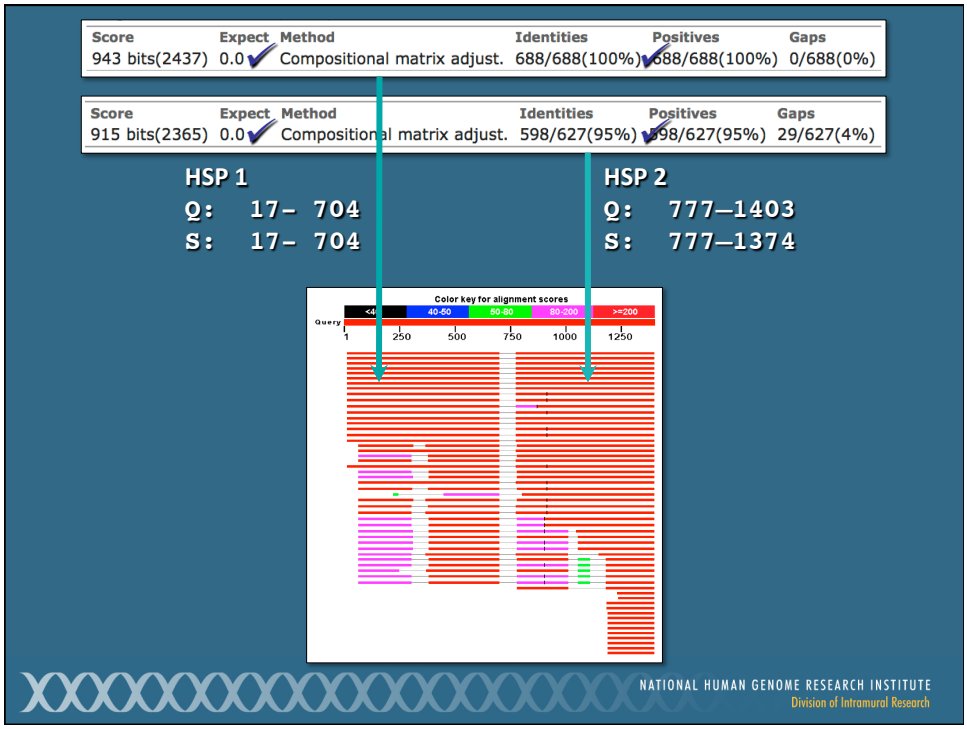
Description	Max score	Total score	Query cover	E value	Ident	Accession
prospero, isoform M [Drosophila melanogaster]	994	1938	93%	0.0	100%	NP_001247046.1
prospero, isoform J [Drosophila melanogaster]	993	1936	93%	0.0	100%	NP_524317.4
prospero [Drosophila melanogaster]	993	1932	93%	0.0	100%	BAA01464.1
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



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	63%	
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_011764800.1
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


Identities:
≥ 25% for proteins
≥ 70% for nucleotides
a Low complexity


Second HSP identified
– Gap



Suggested BLAST Cutoffs

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

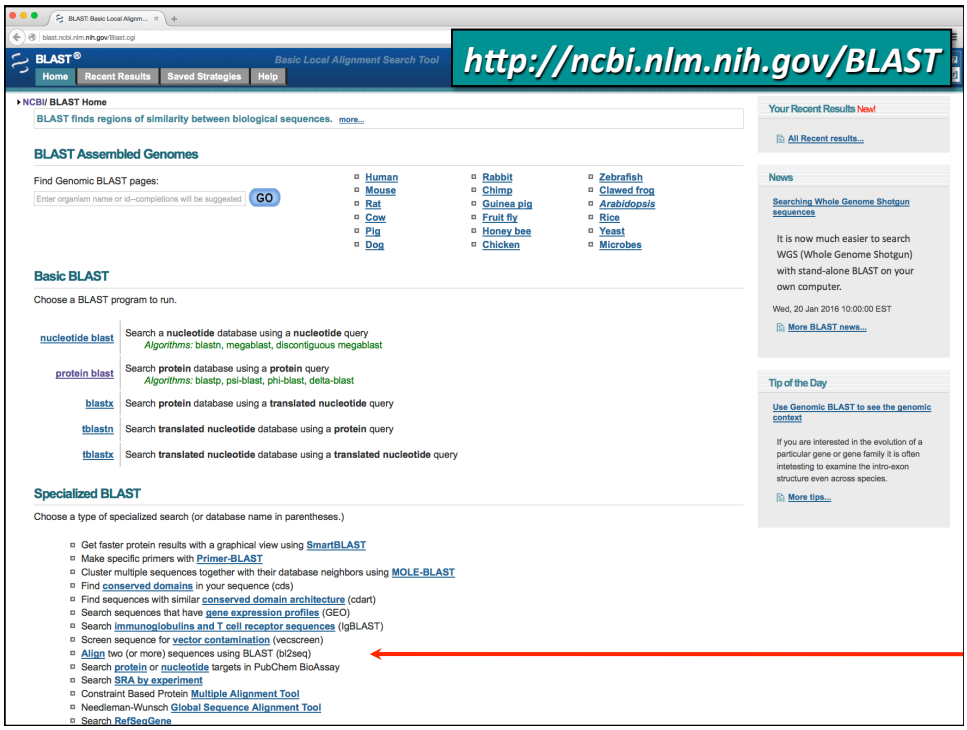
- ***Do not use these cutoffs blindly***
- ***Pay attention to alignments on either side of the dividing line***
- ***Do not ignore biology!***

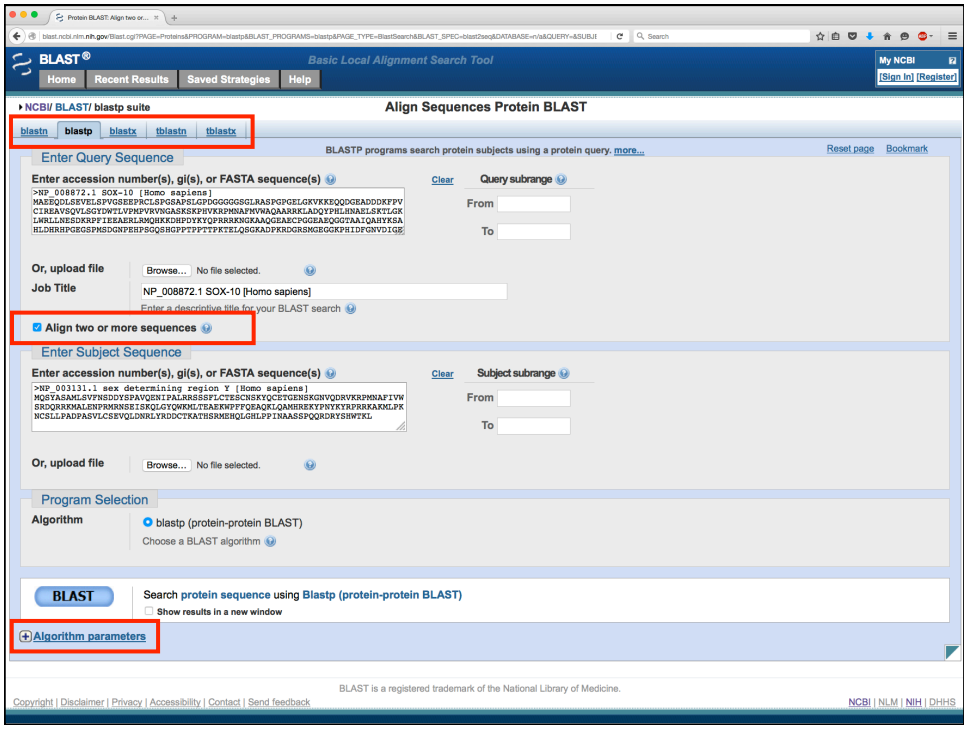
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

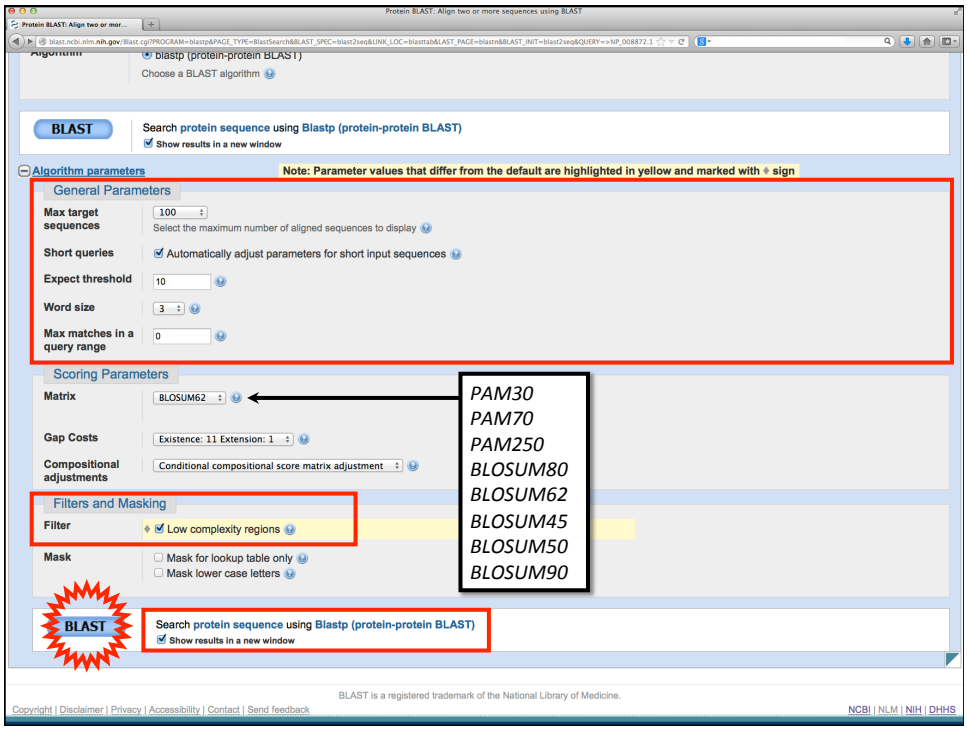
BLAST 2 Sequences

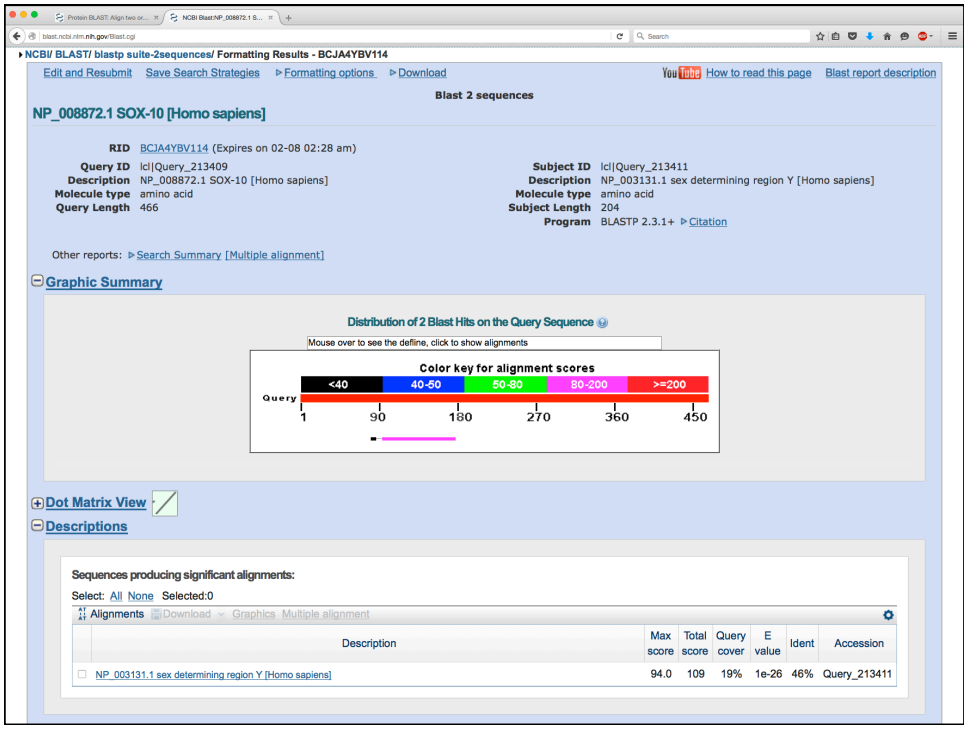
- Finds local alignments between two protein or nucleotide sequences of interest
- All BLAST programs available
- Select BLOSUM and PAM matrices available for protein comparisons
- Same affine gap costs (adjustable)
- Input sequences can be masked

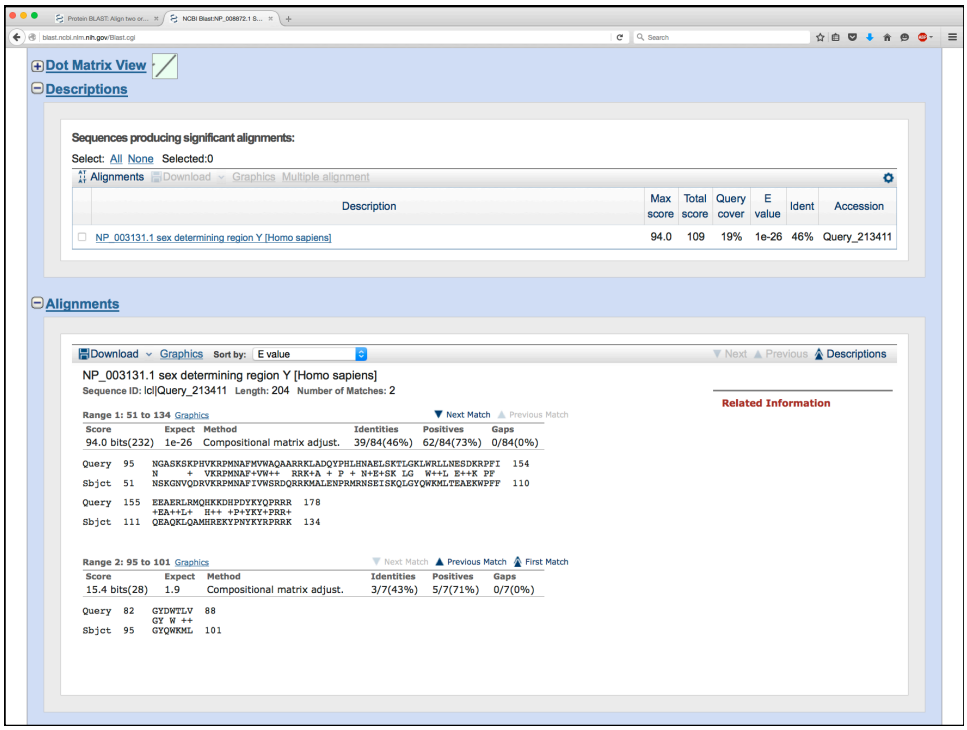
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research




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
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
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

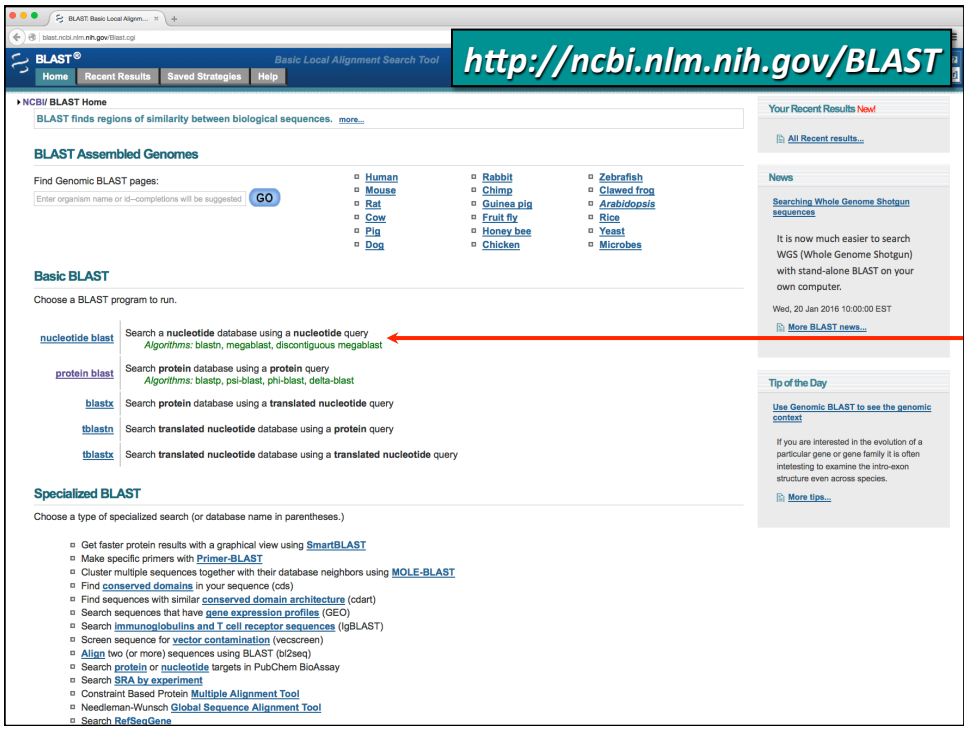

Protein BLAST: Align two or more sequences using BLAST
blastp (protein-protein BLAST)
Choose a BLAST algorithm
BLAST
Search protein sequence using Blastp (protein-protein BLAST)
Show results in a new window
Algorithm parameters
Note: Parameter values that differ from the default are highlighted in yellow 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
Filters and Masking
Filter
Low complexity regions
Mask
Mask for lookup table only
Mask lower case letters
PAM30
PAM70
PAM250
BLOSUM80
BLOSUM62
BLOSUM45
BLOSUM50
BLOSUM90
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

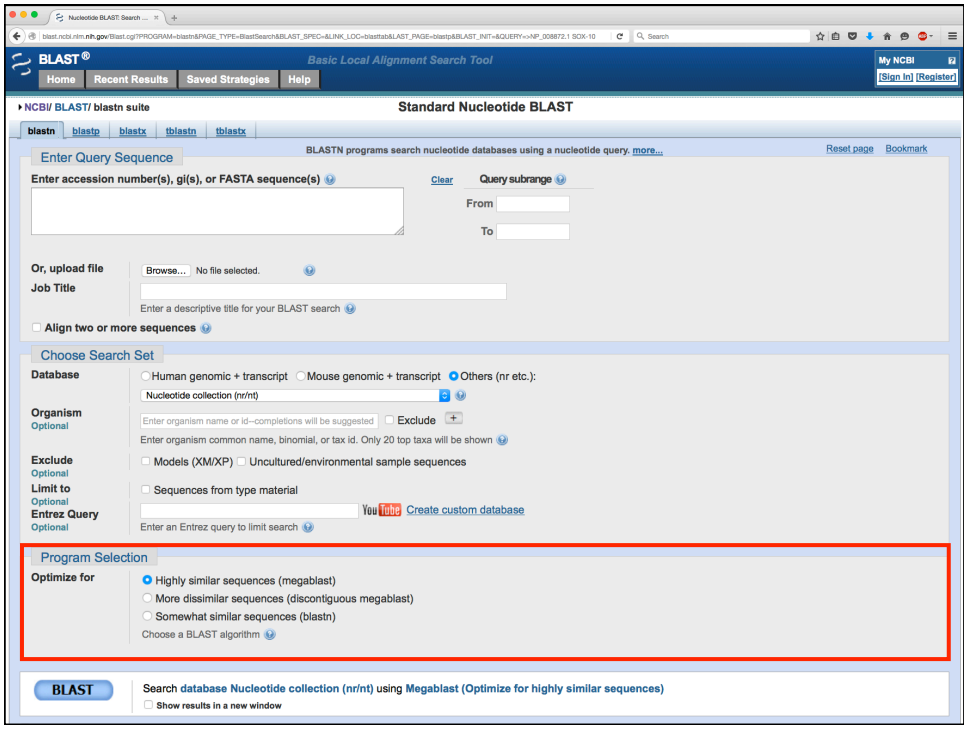

NCBI/ BLAST/ blastp suite-2sequences/ Formatting Results - BCJA4YBV114
Edit and Resubmit
Save Search Strategies
Formatting options
Download
How to read this page
Blast report description
Blast 2 sequences
NP_008872.1 SOX-10 [Homo sapiens]
RID BCJA4YBV114 (Expires on 02-08 02:28 am)
Query ID lcl|Query_213409
Description NP_008872.1 SOX-10 [Homo sapiens]
Molecule type amino acid
Query Length 466
Subject ID lcl|Query_213411
Description NP_003131.1 sex determining region Y [Homo sapiens]
Molecule type amino acid
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 define, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1
90
180
270
360
450
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

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research





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

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

BLAT

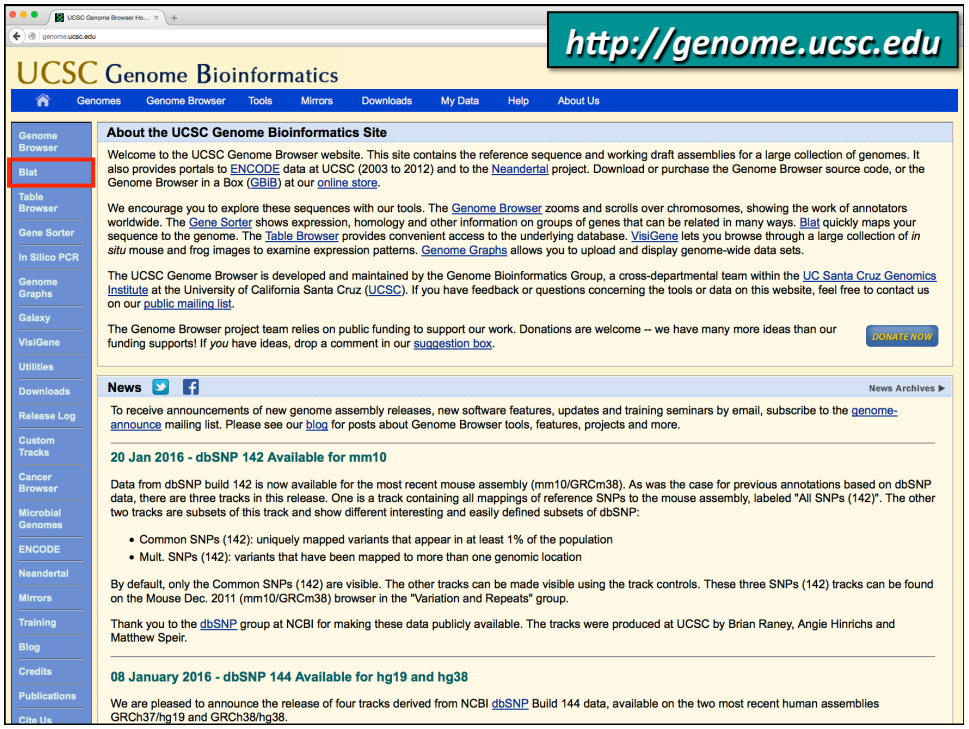
- "BLAST-Like Alignment Tool"
- Designed to rapidly align longer nucleotide sequences ($L \geq 40$) having ≥ 95% sequence similarity
- Can find exact matches reliably down to $L = 33$
- Method of choice when looking for exact matches in nucleotide databases
- 500 times faster than BLAST for mRNA/DNA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

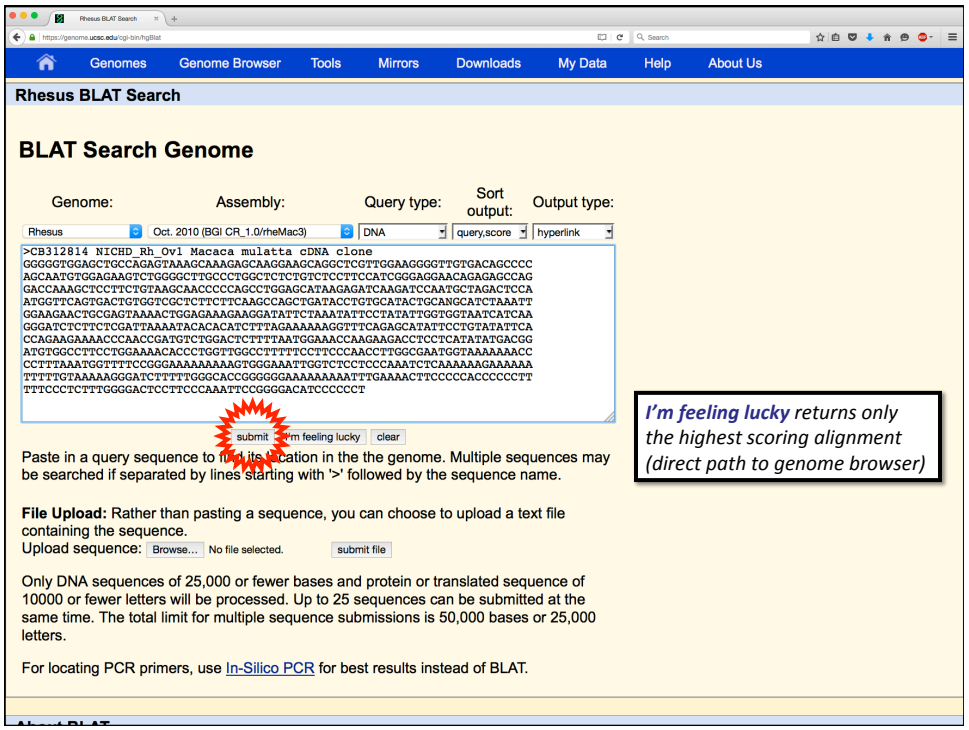
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

When to Use BLAT

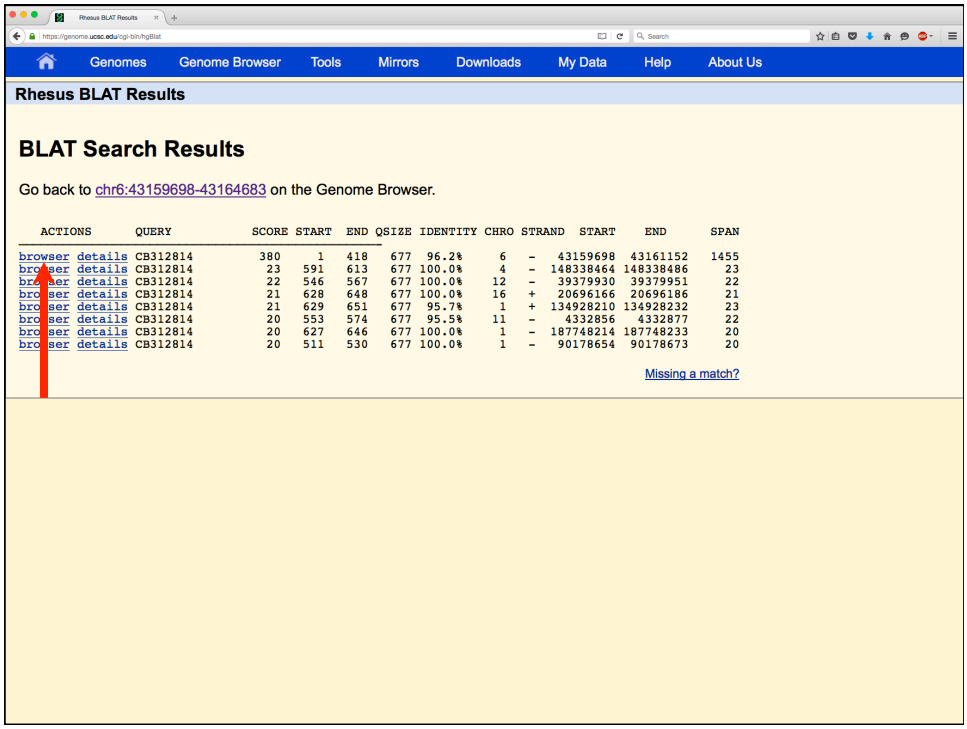
- To characterize an unknown gene or sequence fragment
  - Find its genomic coordinates
  - Determine gene structure (the presence and position of exons)
  - Identify markers of interest in the vicinity of a sequence
- To find highly similar (or identical) sequences
  - Alignment of mRNA sequences onto 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

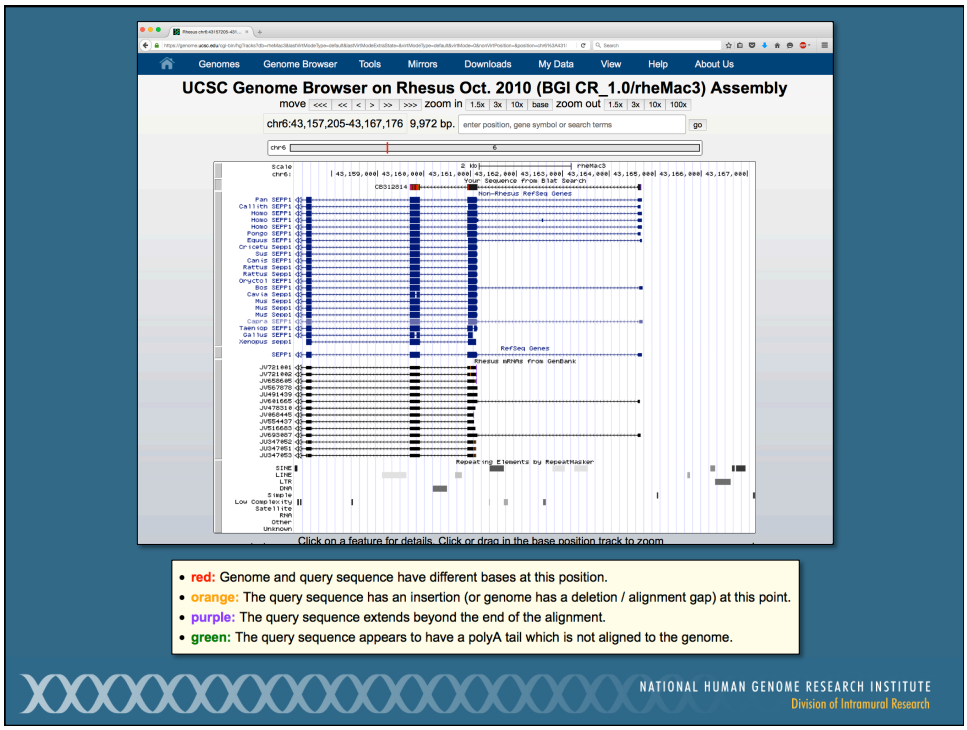
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

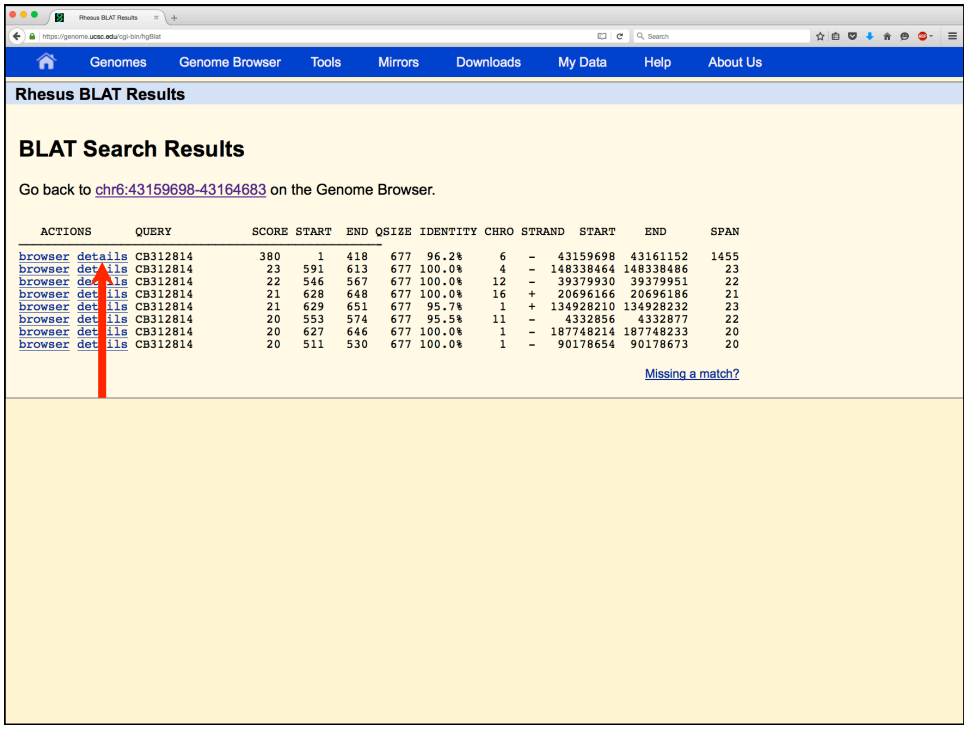




I'm feeling lucky returns only the highest scoring alignment (direct path to genome browser)


Genomes
Genome Browser
Tools
Mirrors
Downloads
My Data
Help
About Us
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?


Genomes
Genome Browser
Tools
Mirrors
Downloads
My Data
View
Help
About Us
UCSC Genome Browser on Rhesus Oct. 2010 (BGI CR_1.0/rheMac3) Assembly
move
zoom in
zoom out
chr6:43,157,205-43,167,176 9,972 bp.
enter position, gene symbol or search terms
go
Click on a feature for details. Click or drag in the base position track to zoom.
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.
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



Rhesus BLAT Results

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Help | About Us

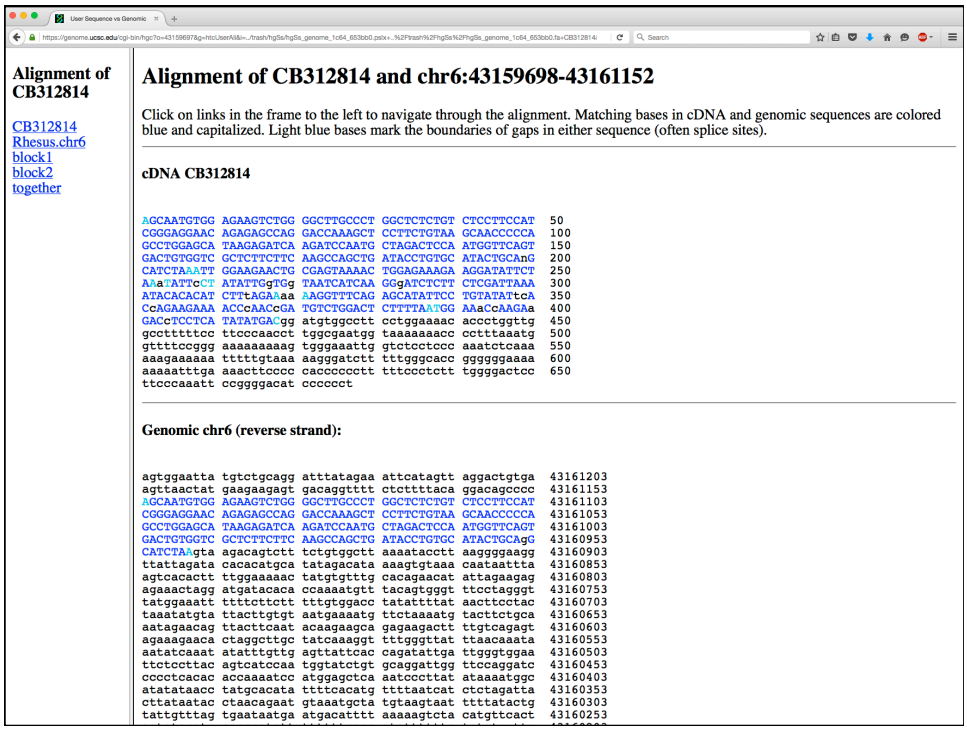
Rhesus BLAT Results

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

CB312814
Rhesus.chr6
block1
block2
together

Alignment of CB312814 and chr6:43159698-43161152

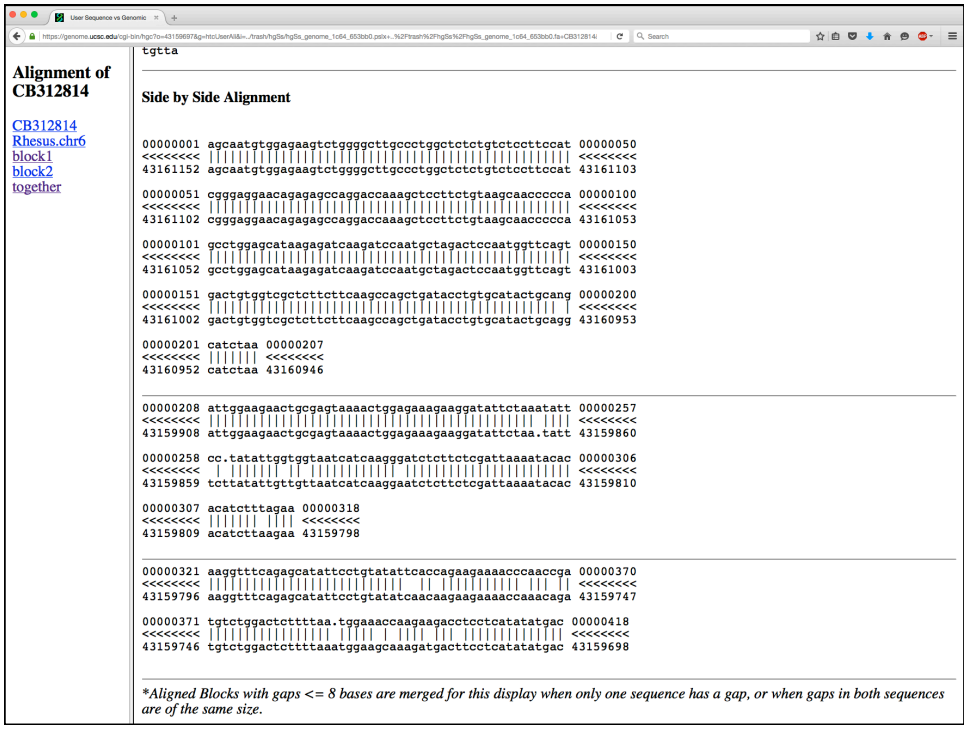
Click on links in the frame to the left to navigate through the alignment. Matching bases in cDNA and genomic sequences are colored blue and capitalized. Light blue bases mark the boundaries of gaps in either sequence (often splice sites).

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
gcctttttcc ttcccaacct tggcgaatgg taaaaaaacc cctttaaatg  500
gttttccggg aaaaaaaaag tgggaaattg gtctcctccc aaatctcaaa  550
aaagaaaaaa tttttgtaaa aagggatctt tttgggcacc gggggggaaaa  600
aaaaatttga aaacttcccc caccccctt tttccctctt tggggactcc  650
ttcccaaatt ccggggacat cccccct
```

Genomic chr6 (reverse strand):

```
agtggaatta tgtctgcagg atttatagaa attcatagtt aggactgtga  43161203
agttaactat gaagaagagt gacaggtttt ctctttttaca 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
cttatattac ctaacagaat gtaaatgcta tgtaagtaat ttttatactg  43160303
tattgtttag tgaataatga atgacatttt aaaaagtcta catgttcact  43160253
```



Alignment of CB312814

CB312814
Rhesus.chr6
block1
block2
together

tgtta

Side by Side Alignment

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

00000051 cgggaggaacagagagccaggaccaaagctccttctgtaagcaacccccca 00000100
<<<<<<<< |||||||||||||||||||||||||||||||||||||||||||||||||| <<<<<<<<
43161102 cgggaggaacagagagccaggaccaaagctccttctgtaagcaacccccca 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 aaggtttcagagcatattcctgtatattcaccagaagaaaaacccaaccga 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.*