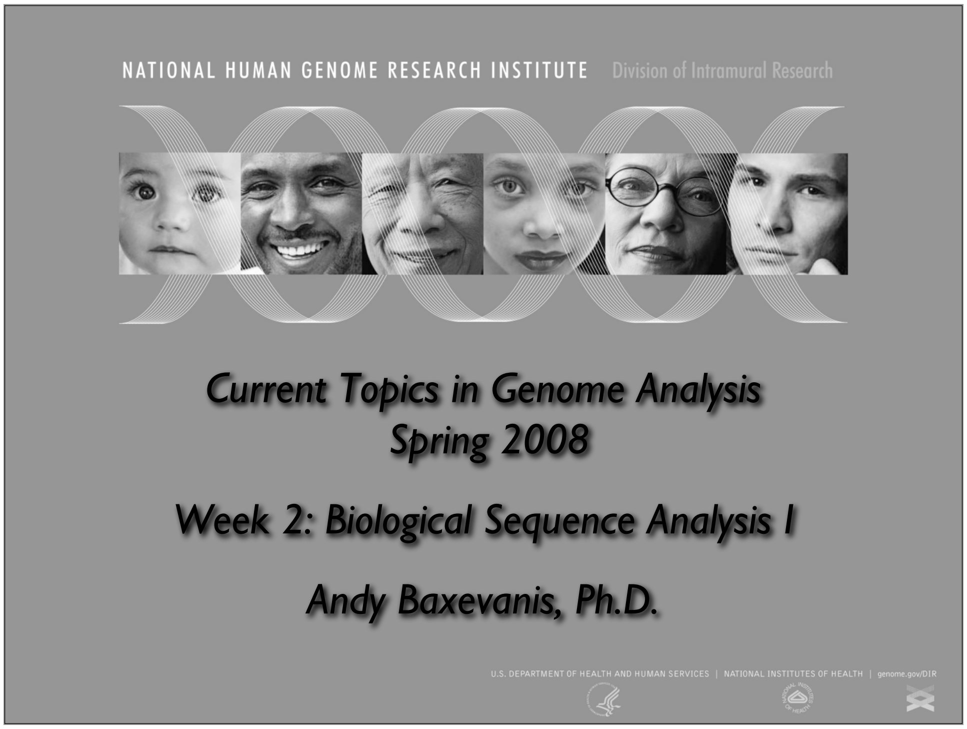


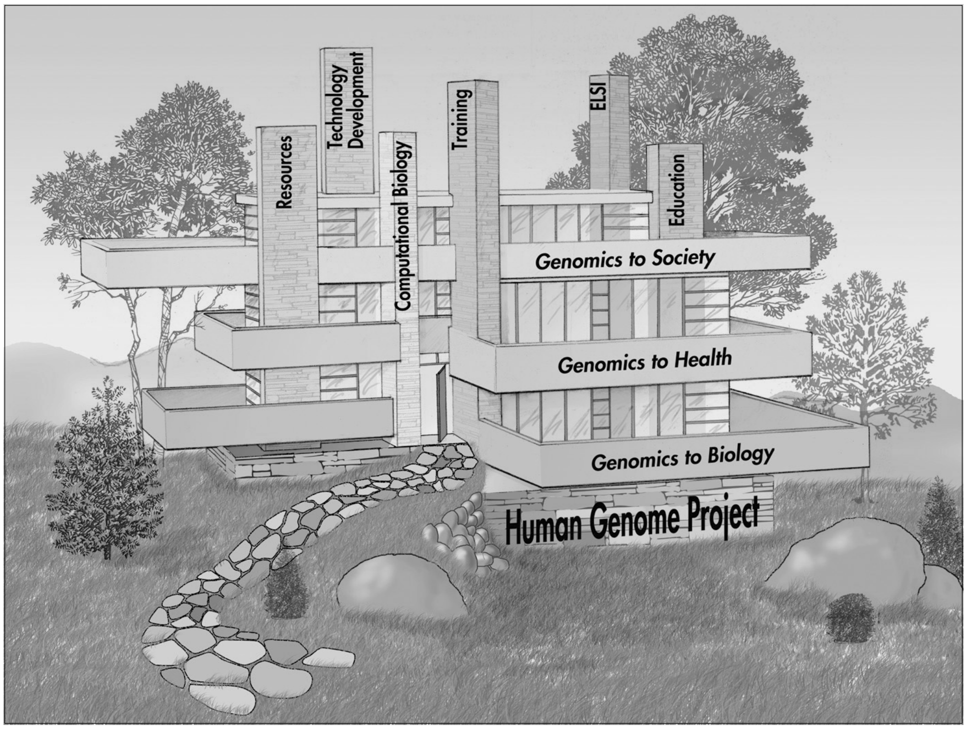
NATIONAL HUMAN GENOME RESEARCH INSTITUTE Division of Intramural Research

Current Topics in Genome Analysis Spring 2008

Week 2: Biological Sequence Analysis I

Andy Baxevanis, Ph.D.

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



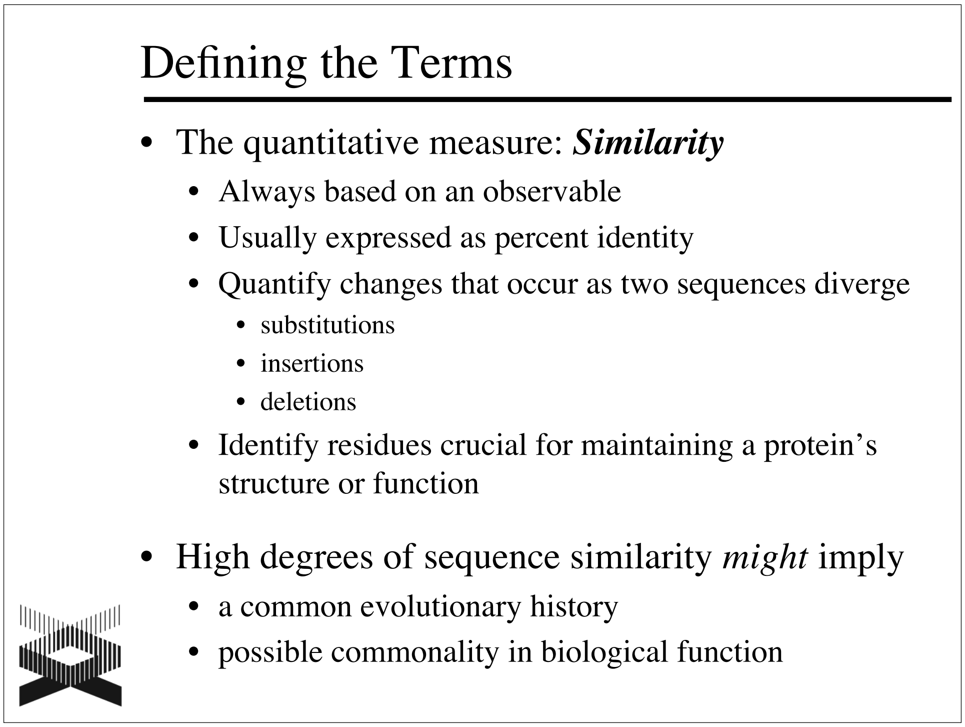
Overview

- Week 2
  - Similarity *vs*. Homology
  - Global *vs*. Local Alignments
  - Scoring Matrices
  - BLAST
  - BLAT
- Week 3
  - Profiles, Patterns, Motifs, and Domains
  - Structures: VAST, Cn3D, and *de novo* Prediction
  - Multiple Sequence Alignment

Why do sequence alignments?

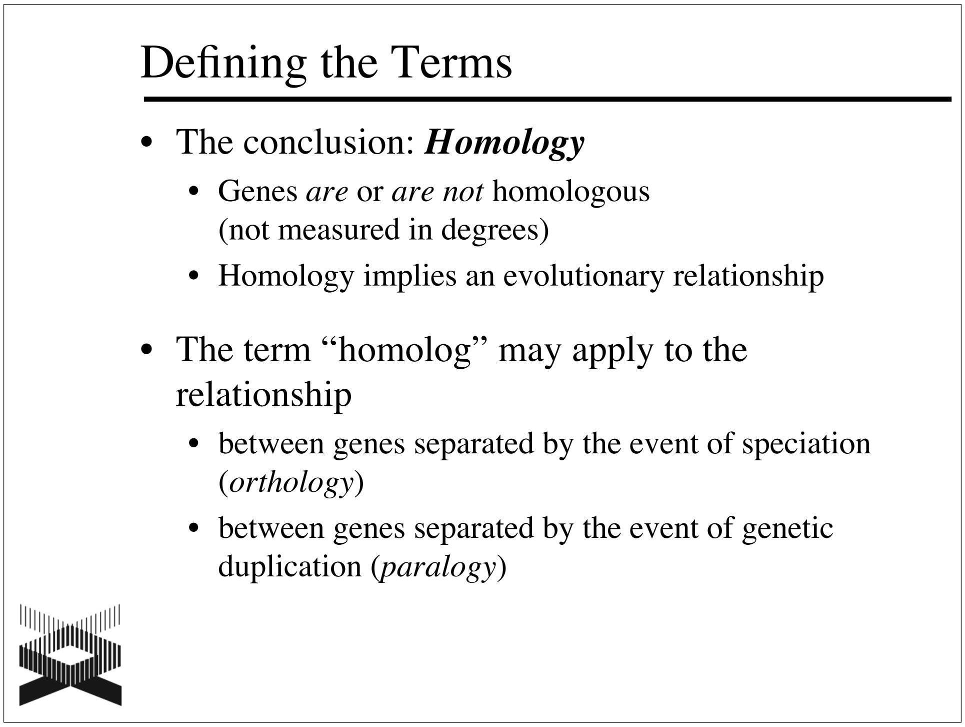
- Provide a measure of relatedness between nucleotide or amino acid sequences
- Determining relatedness allows one to draw biological inferences regarding
  - structural relationships
  - functional relationships
  - evolutionary relationships

→ *importance of using correct terminology*



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
    - 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



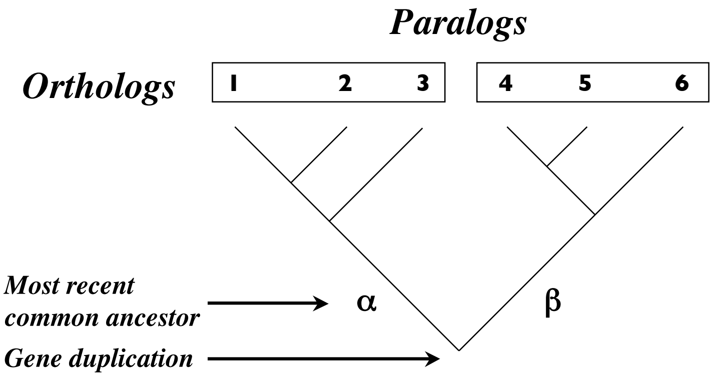
Defining the Terms

- The conclusion: ***Homology***
  - Genes *are* or *are not* homologous
    (not measured in degrees)
  - Homology implies an evolutionary relationship
- The term "homolog" may apply to the relationship
  - between genes separated by the event of speciation (*orthology*)
  - between genes separated by the event of genetic duplication (*paralogy*)

Defining the Terms

- Orthologs
  - Sequences are direct descendants of a sequence in a common ancestor
  - Most likely have similar domain structure, three-dimensional structure, and biological function
- Paralogs
  - Related through a gene duplication event
  - Provides insight into “evolutionary innovation” (adapting a pre-existing gene product for a new function)

Defining the Terms



Paralogs

Orthologs

1 2 3

4 5 6

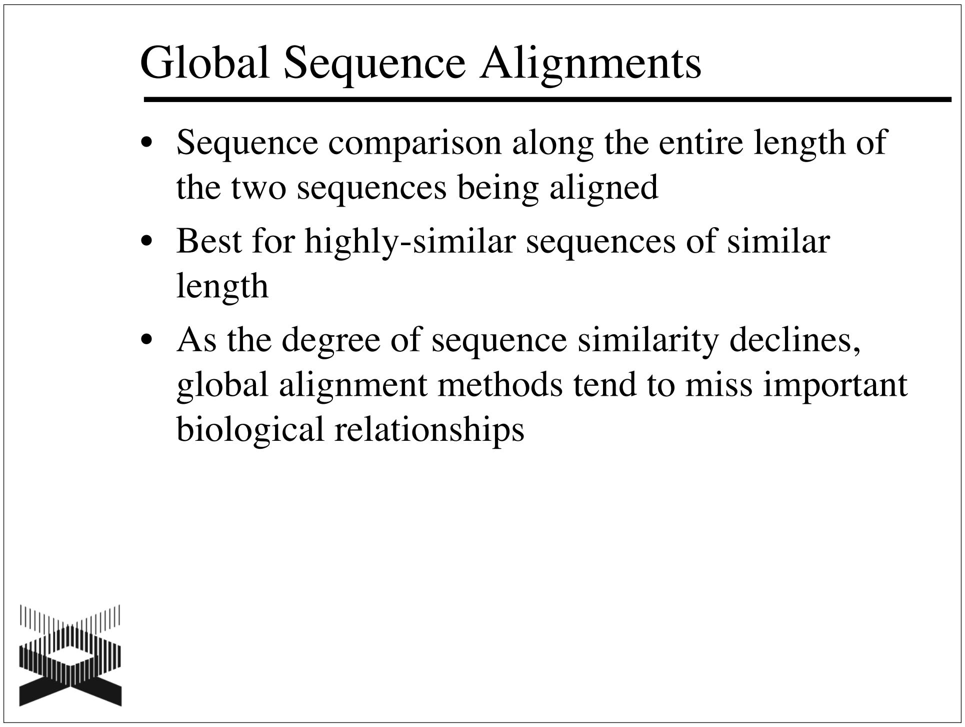
Most recent common ancestor

Gene duplication

α

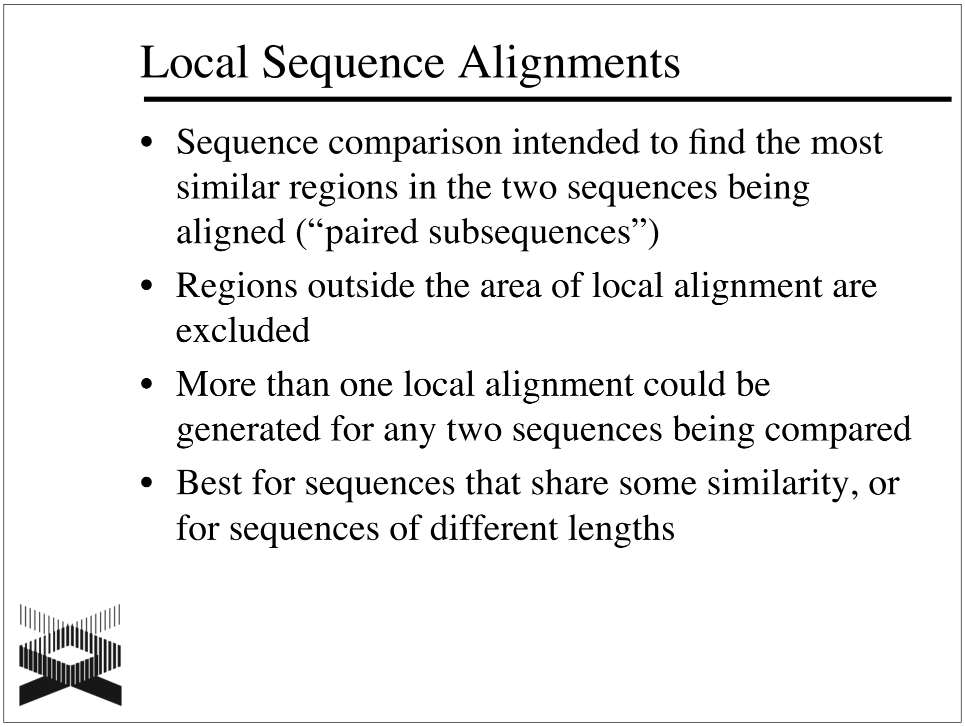
β

- 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)



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

- 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:
  - Cys/Pro important for structure and function
  - Trp has 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, other 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

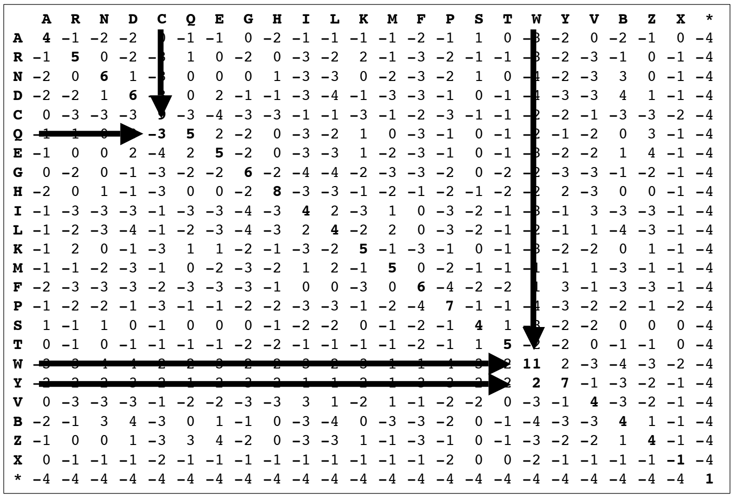
	A	**T**	**G**	**C**	**S**	**W**	**R**	**Y**	**K**	**M**	**B**	**V**	**H**	**D**	**N**
A	**5**	−4	−4	−4	−4	1	1	−4	−4	1	−4	−1	−1	−1	−2
T	−4	**5**	−4	−4	−4	1	−4	1	1	−4	−1	−4	−1	−1	−2
G	−4	−4	**5**	−4	1	−4	1	−4	1	−4	−1	−1	−4	−1	−2
C	−4	−4	−4	**5**	1	−4	−4	1	−4	1	−1	−1	−1	−4	−2
S	−4	−4	1	1	**−1**	−4	−2	−2	−2	−2	−1	−1	−3	−3	−1
W	1	1	−4	−4	−4	**−1**	−2	−2	−2	−2	−3	−3	−1	−1	−1
R	1	−4	1	−4	−2	−2	**−1**	−4	−2	−2	−3	−1	−3	−1	−1
Y	−4	1	−4	1	−2	−2	−4	**−1**	−2	−2	−1	−3	−1	−3	−1
K	−4	1	1	−4	−2	−2	−2	−2	**−1**	−4	−1	−3	−3	−1	−1
M	1	−4	−4	1	−2	−2	−2	−2	−4	**−1**	−3	−1	−1	−3	−1
B	−4	−1	−1	−1	−1	−3	−3	−1	−1	−3	**−1**	−2	−2	−2	−1
V	−1	−4	−1	−1	−1	−3	−1	−3	−3	−1	−2	**−1**	−2	−2	−1
H	−1	−1	−4	−1	−3	−1	−3	−1	−3	−1	−2	−2	**−1**	−2	−1
D	−1	−1	−1	−4	−3	−1	−1	−3	−1	−3	−2	−2	−2	**−1**	−1
N	−2	−2	−2	−2	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	**−1**

- *Simple match/mismatch scoring scheme:*

  Match + 5
  Mismatch – 4

- *Assumes each nucleotide occurs 25% of the time*

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**

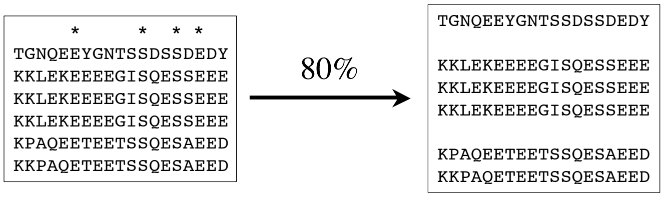
BLOSUM62

BLOSUM Matrices

- Henikoff and Henikoff, 1992
- <u>Blo</u>cks <u>Su</u>bstitution <u>M</u>atrix
    - Look only for differences in conserved, ungapped regions of a protein family (“blocks”)
    - Directly calculated, using no extrapolations
    - More sensitive to detecting structural or functional substitutions
    - Generally perform better than PAM matrices for local similarity searches *(Henikoff and Henikoff, 1993)*

BLOSUM *n*

- Calculated from sequences sharing no more than *n%* identity
- Contribution of sequences > *n%* identical clustered and weighted to 1



```
     *     *  * *
TGNQEEYGNTSSDSSDEDY
KKLEKEEEEGISQESSEEE
KKLEKEEEEGISQESSEEE
KKLEKEEEEGISQESSEEE
KPAQEETEETSSQESAEED
KKPAQETEETSSQESAEED
```

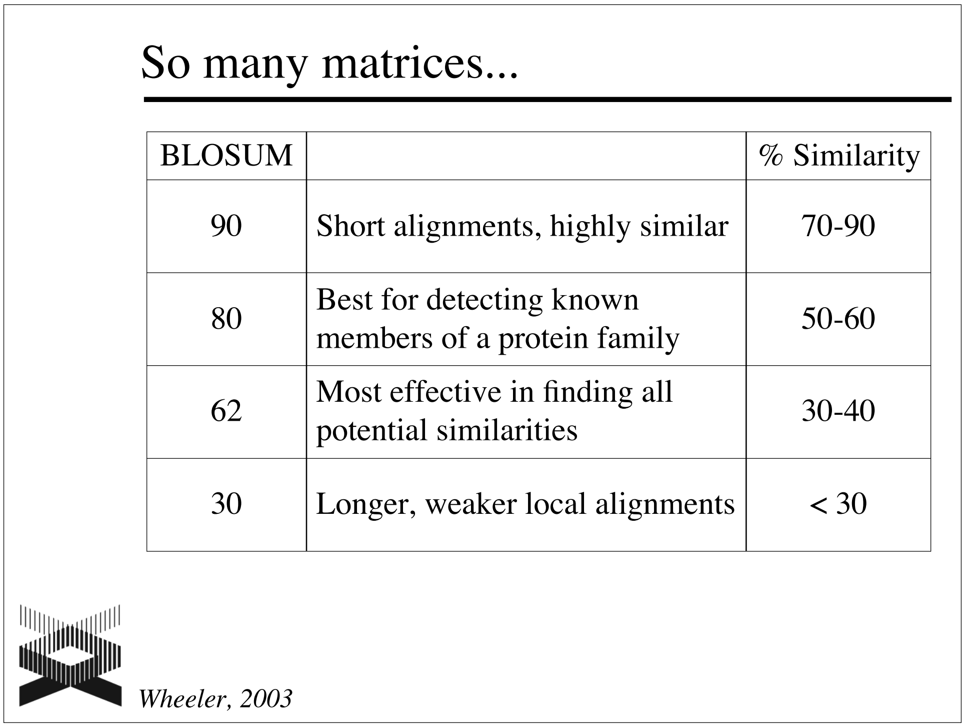
```
TGNQEEYGNTSSDSSDEDY

KKLEKEEEEGISQESSEEE
KKLEKEEEEGISQESSEEE
KKLEKEEEEGISQESSEEE

KPAQEETEETSSQESAEED
KKPAQETEETSSQESAEED
```

A+T Hook Domain (Block IPB000637B)

2,000 blocks representing > 500 groups of related proteins

BLOSUM *n*

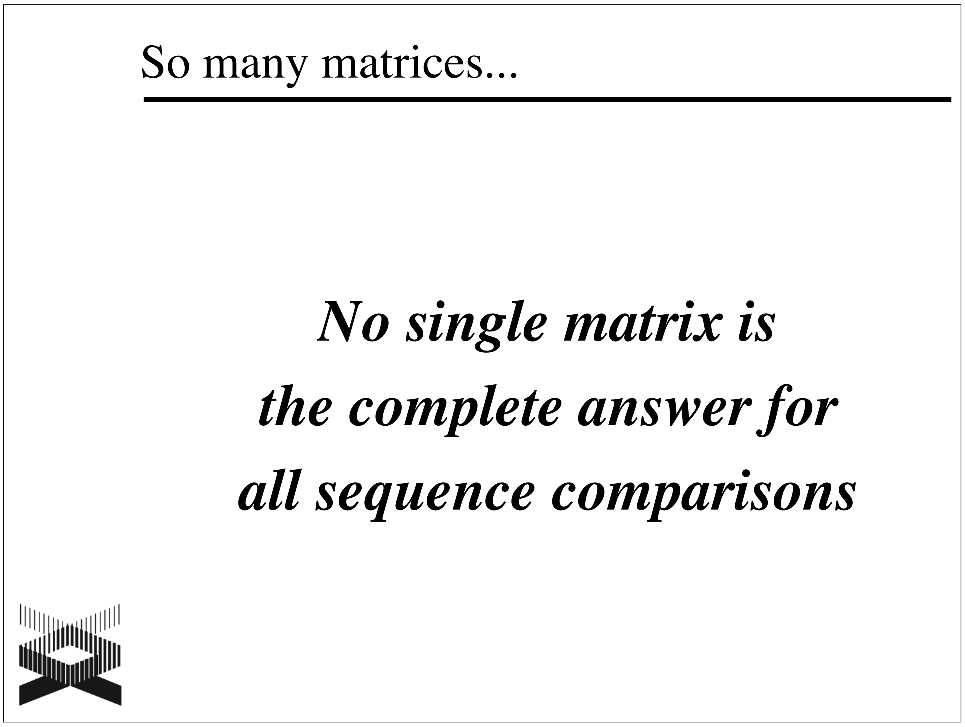
- Clustering reduces contribution of closely-related sequences (less bias towards substitutions that occur in the most closely-related members of a family)
- Substitution frequencies are more heavily-influenced by sequences that are more divergent than this cutoff
- Reducing *n* yields more distantly-related sequences



So many matrices...

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

Wheeler, 2003



So many matrices...

No single matrix is the complete answer for all sequence comparisons

Gaps

- Compensate for insertions and deletions
- Used to improve alignments between two sequences
- Must be kept to a reasonable number, to not reflect a biological implausible scenario (~1 gap per 20 residues good rule-of-thumb)
- Cannot be scored simply as a “match” or a “mismatch”

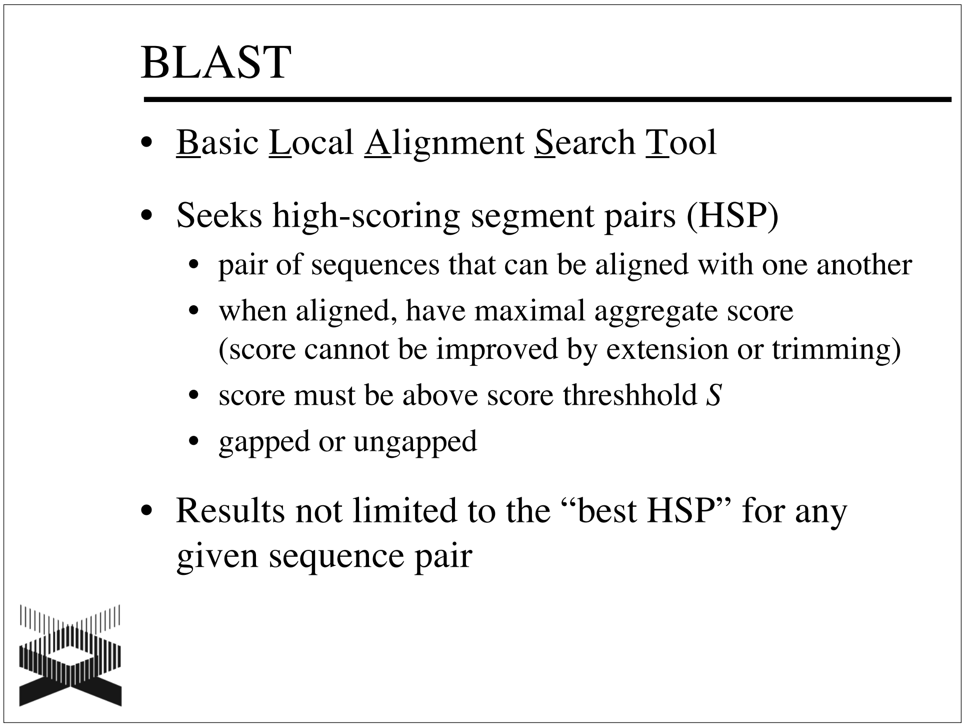
Affine Gap Penalty

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

$$\boxed{\text{Deduction for a gap} = G + Ln}$$

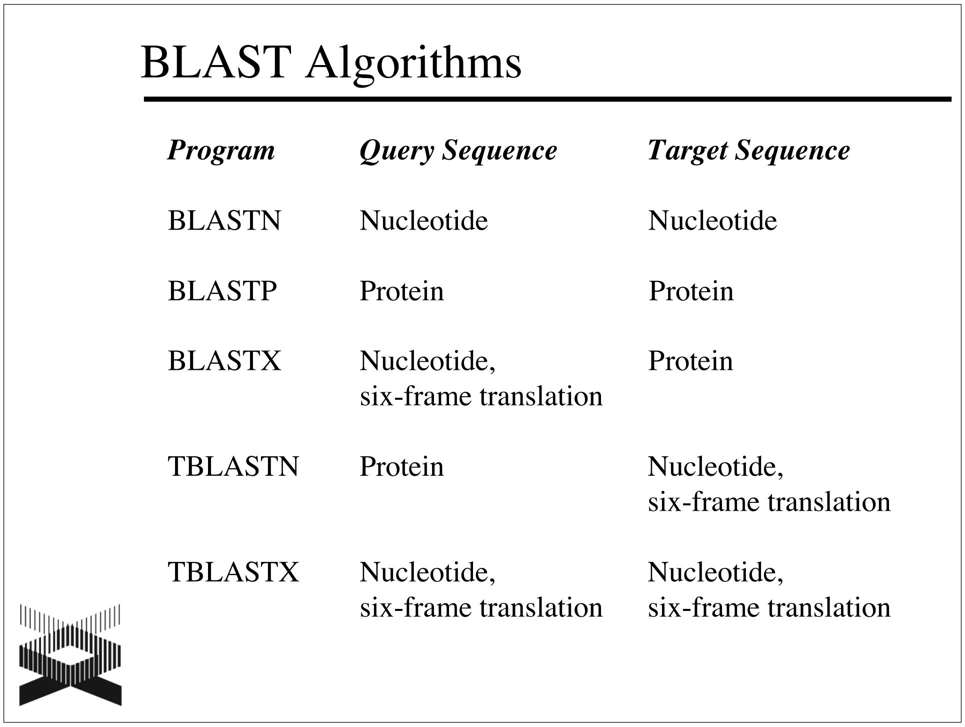
				nuc	pro
where	G	=	gap-opening penalty	5	11
	L	=	gap-extension penalty	2	1
	n	=	length of the gap		
and	G	>	L		

Can adjust scores to make gap insertion more or less permissive, but most programs will use values of G and L most appropriate for the scoring matrix selected



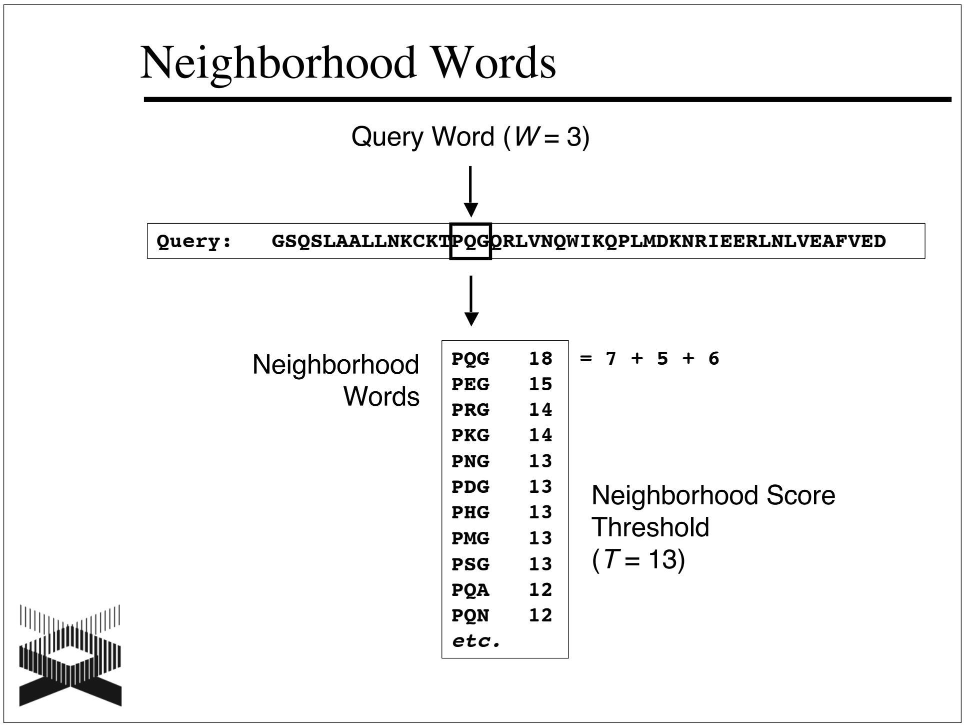
BLAST

- Basic Local Alignment Search Tool
- Seeks high-scoring segment pairs (HSP)
  - 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 threshhold *S*
  - gapped or ungapped
- Results not limited to the "best HSP" for any given sequence pair



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

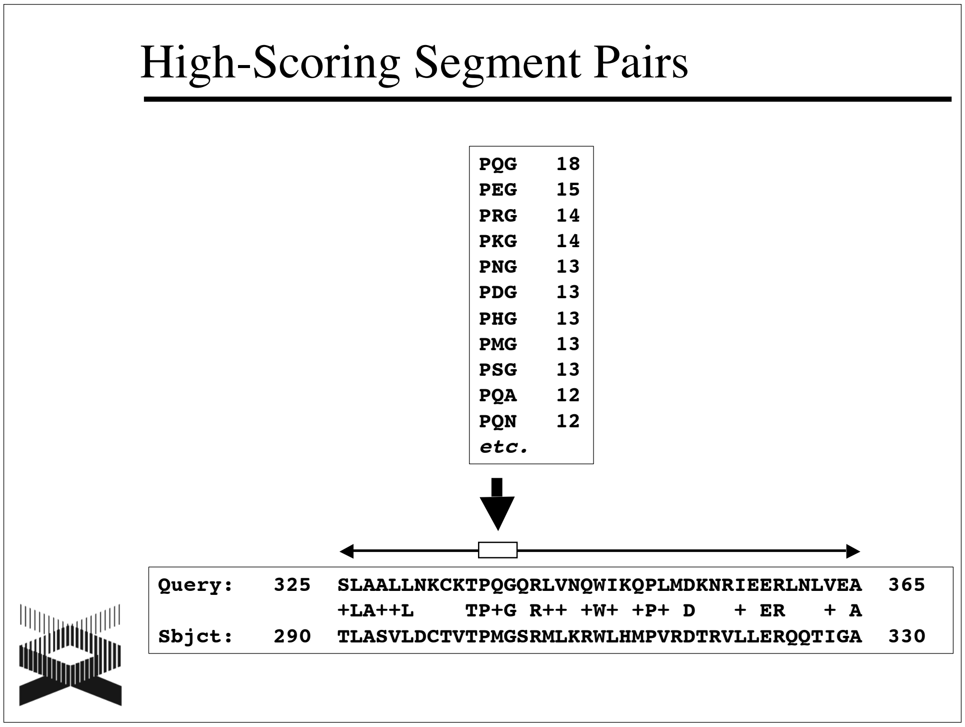
Query Word (*W* = 3)

```
Query:   GSQSLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEAFVED
```

Neighborhood Words

PQG	18	= 7 + 5 + 6
PEG	15	
PRG	14	
PKG	14	
PNG	13	
PDG	13	
PHG	13	
PMG	13	
PSG	13	
PQA	12	
PQN	12	
etc.		

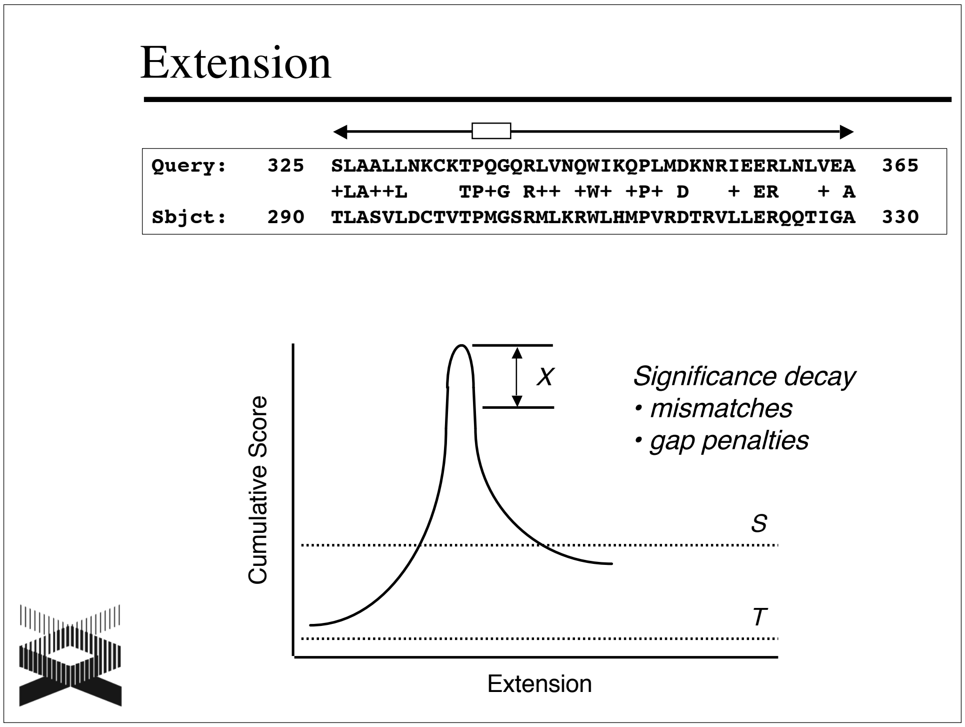
Neighborhood Score Threshold
(*T* = 13)



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.	

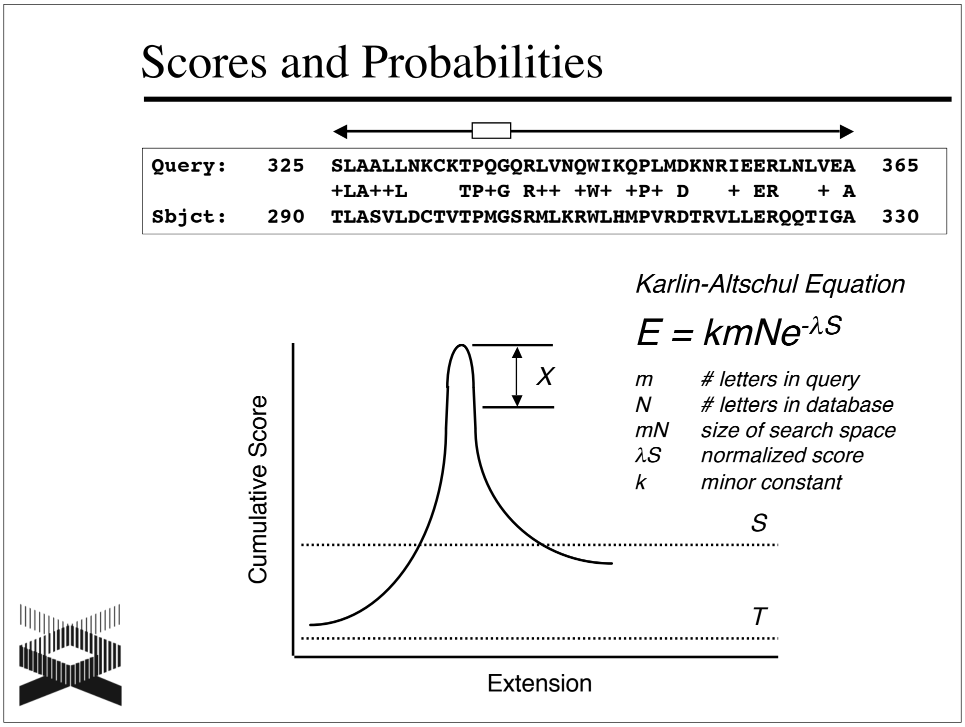
```
Query:   325  SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA  365
              +LA++L    TP+G R++ +W+ +P+ D   + ER   + A
Sbjct:   290  TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA  330
```



Extension

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

Significance decay
- *mismatches*
- *gap penalties*



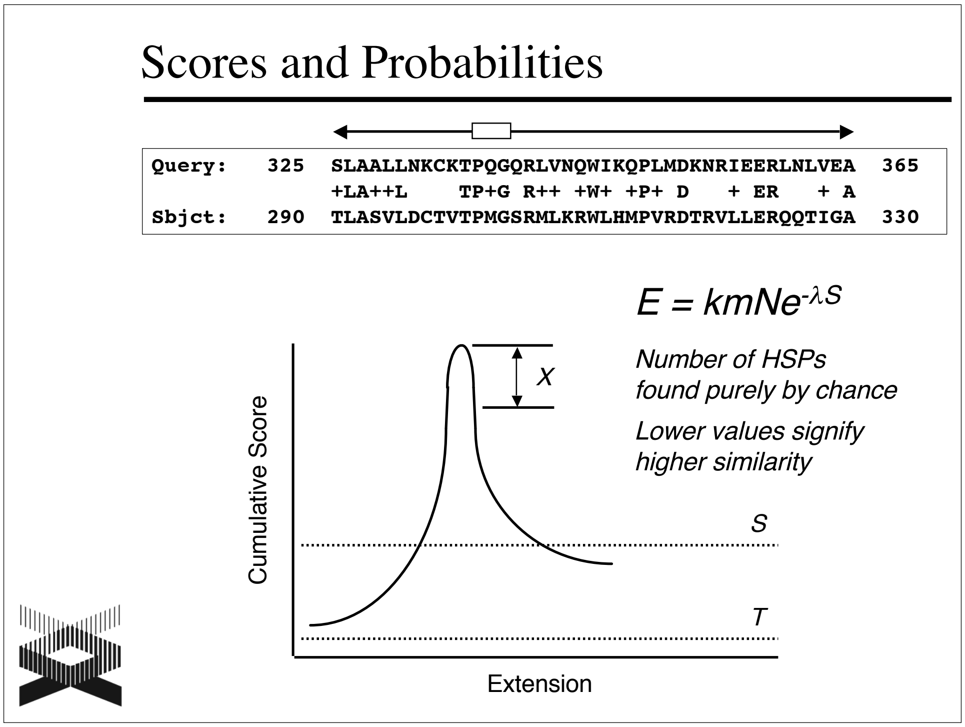
Scores and Probabilities

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

Karlin-Altschul Equation

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

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



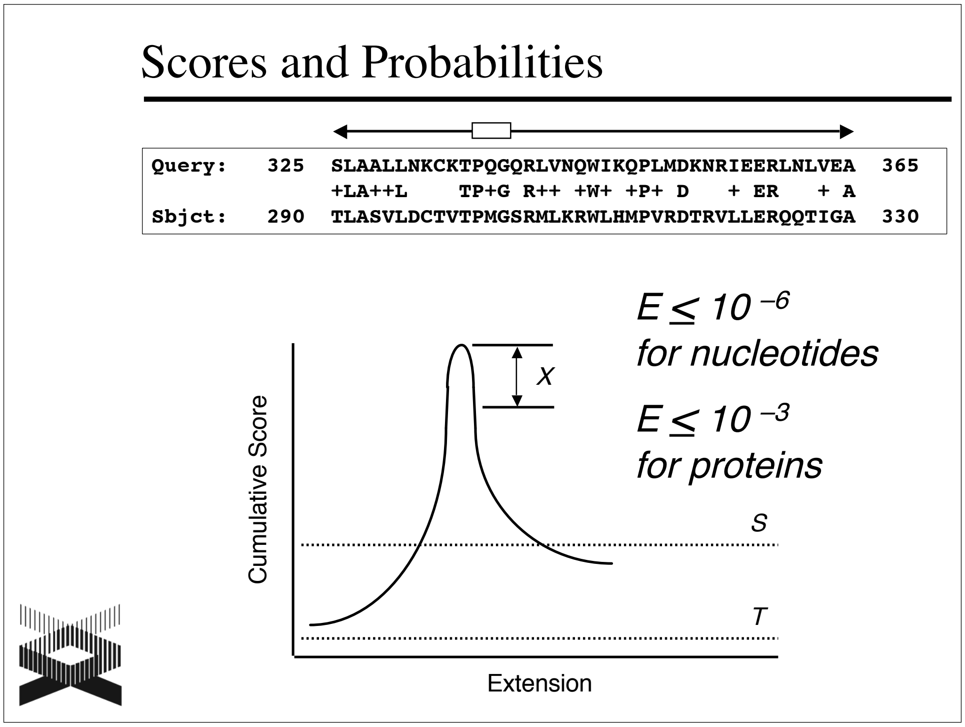
Scores and Probabilities

```
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

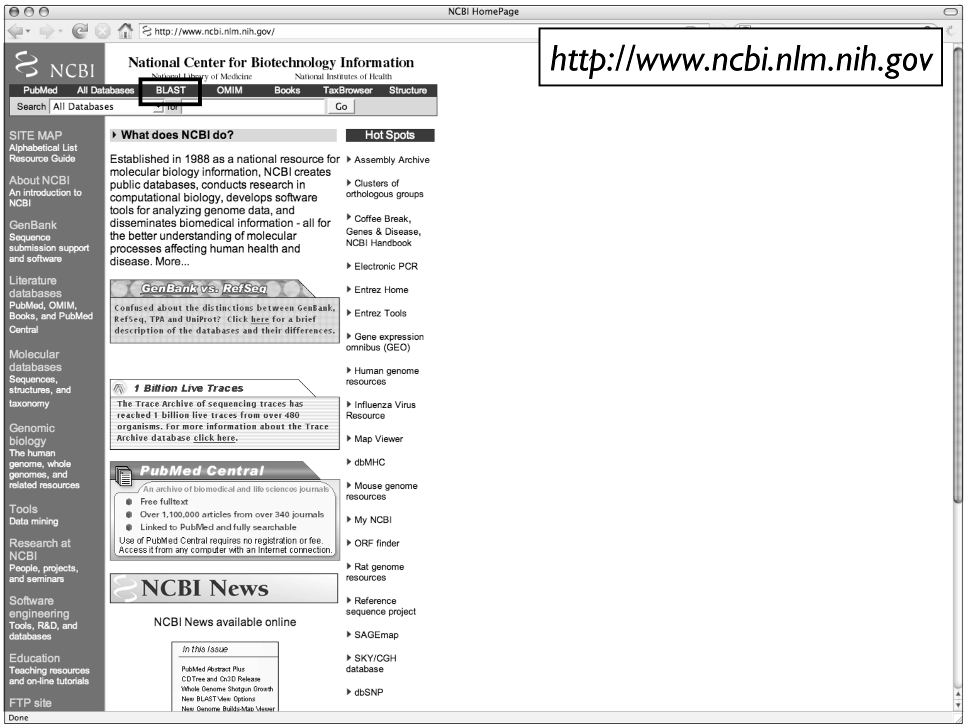


Scores and Probabilities

```
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*



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

National Center for Biotechnology Information
National Library of Medicine National Institutes of Health

SITE MAP
Alphabetical List
Resource Guide

About NCBI
An introduction to NCBI

GenBank
Sequence submission support and software

Literature databases
PubMed, OMIM, Books, and PubMed Central

Molecular databases
Sequences, structures, and taxonomy

Genomic biology
The human genome, whole genomes, and related resources

Tools
Data mining

Research at NCBI
People, projects, and seminars

Software engineering
Tools, R&D, and databases

Education
Teaching resources and on-line tutorials

FTP site

What does NCBI do?

Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. More...

GenBank vs. RefSeq

Confused about the distinctions between GenBank, RefSeq, TPA and UniProt? Click here for a brief description of the databases and their differences.

1 Billion Live Traces

The Trace Archive of sequencing traces has reached 1 billion live traces from over 480 organisms. For more information about the Trace Archive database click here.

PubMed Central

An archive of biomedical and life sciences journals

- Free fulltext
- Over 1,100,000 articles from over 340 journals
- Linked to PubMed and fully searchable

Use of PubMed Central requires no registration or fee. Access it from any computer with an Internet connection.

NCBI News

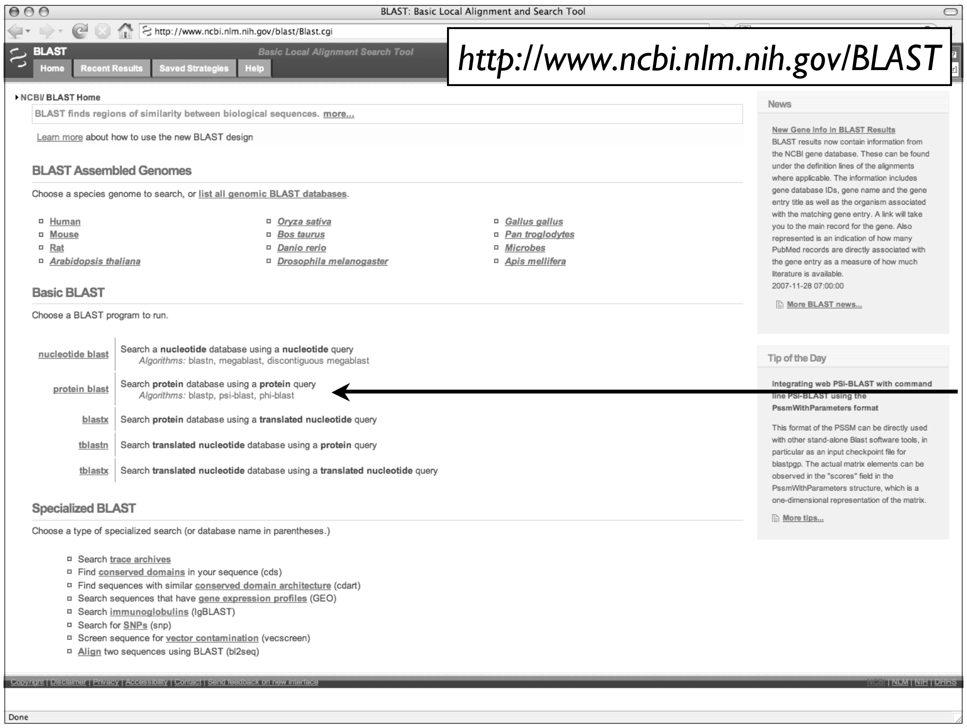
NCBI News available online

In this Issue

PubMed Abstract Plus
CD Tree and Cn3D Release
Whole Genome Shotgun Growth
New BLAST View Options
New Genome Builds-Map Viewer

Hot Spots

- Assembly Archive
- Clusters of orthologous groups
- Coffee Break, Genes & Disease, NCBI Handbook
- Electronic PCR
- Entrez Home
- Entrez Tools
- Gene expression omnibus (GEO)
- Human genome resources
- Influenza Virus Resource
- Map Viewer
- dbMHC
- Mouse genome resources
- My NCBI
- ORF finder
- Rat genome resources
- Reference sequence project
- SAGEmap
- SKY/CGH database
- dbSNP


BLAST: Basic Local Alignment and Search Tool
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi
BLAST
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback on new interface
NCBI | NLM | NIH | DHHS
Done

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

NCBI/ BLAST Home

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

Learn more about how to use the new BLAST design

BLAST Assembled Genomes

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

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

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

Specialized BLAST

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

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

News

New Gene Info in BLAST Results
BLAST results now contain information from the NCBI gene database. These can be found under the definition lines of the alignments where applicable. The information includes gene database IDs, gene name and the gene entry title as well as the organism associated with the matching gene entry. A link will take you to the main record for the gene. Also represented is an indication of how many PubMed records are directly associated with the gene entry as a measure of how much literature is available.
2007-11-28 07:00:00

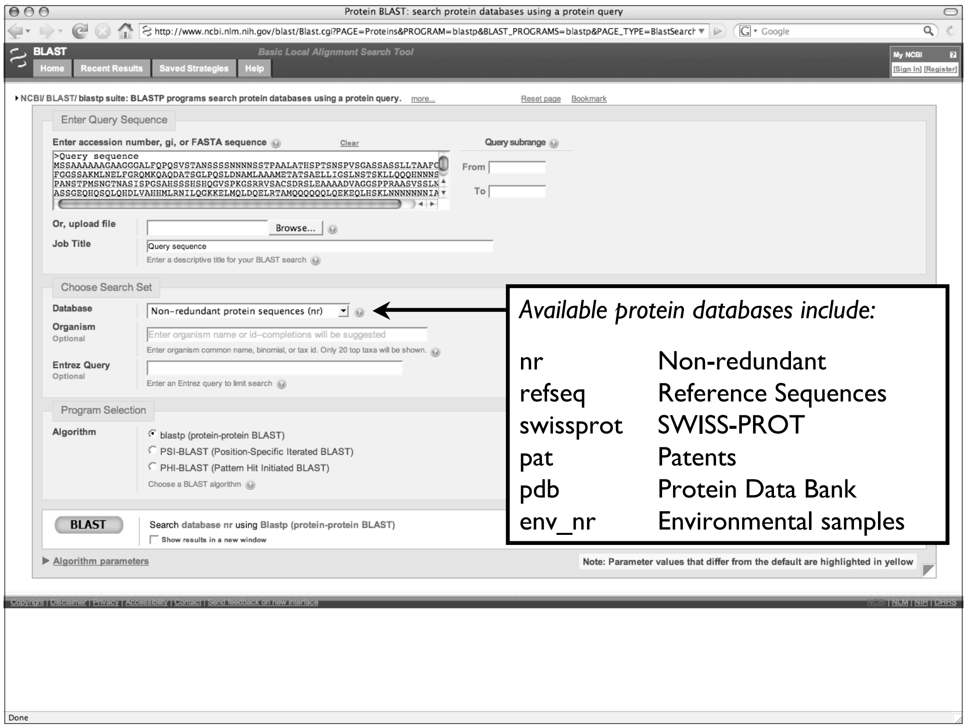
More BLAST news...

Tip of the Day

Integrating web PSI-BLAST with command line PSI-BLAST using the PssmWithParameters format

This format of the PSSM can be directly used with other stand-alone Blast software tools, in particular as an input checkpoint file for blastpgp. The actual matrix elements can be observed in the "scores" field in the PssmWithParameters structure, which is a one-dimensional representation of the matrix.

More tips...


Protein BLAST: search protein databases using a protein query
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?PAGE=Proteins&PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch
Google
BLAST
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
My NCBI
[Sign In] [Register]
Reset page
Bookmark
Clear
Browse...
BLAST
Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback on new interface
NCBI | NLM | NIH | DHHS
Done

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

Enter Query Sequence

Enter accession number, gi, or FASTA sequence

```
>Query sequence
MSSAAAAAAGAAGGGALFQPQSVSTANSSSSNNNNSSTPAALATHSPTSNSPVSGASSASSLLTAAFG
FGGSSAKMLNELFGRQMKQAQDATSGLPQSLDNAMLAAAMETATSAELLIGSLNSTSKLLQQQHNNNS
PANSTPMSNGTNASISPGSAHSSSHSHQGVSPKGSRRVSACSDRSLEAAAADVAGGSPPRAASVSSLN
ASSGEQHQSQLQHDLVAHHMLRNILQGKKELMQLDQELRTAMQQQQQLQEKEQLHSKLNNNNNNNIA
```

Query subrange

From

To

Or, upload file

Job Title: Query sequence

Enter a descriptive title for your BLAST search

Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism Optional: Enter organism name or id--completions will be suggested

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

Entrez Query Optional

Enter an Entrez query to limit search

Program Selection

Algorithm

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

Choose a BLAST algorithm

Search database nr using Blastp (protein-protein BLAST)

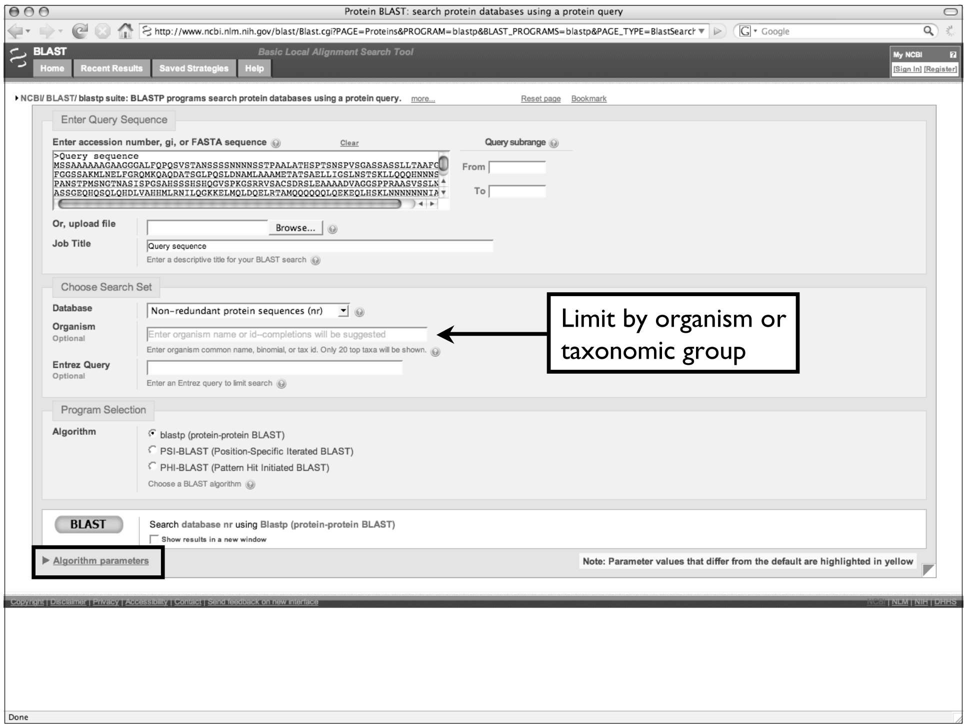
□ Show results in a new window

▶ Algorithm parameters

Note: Parameter values that differ from the default are highlighted in yellow

Available protein databases include:

nr	Non-redundant
refseq	Reference Sequences
swissprot	SWISS-PROT
pat	Patents
pdb	Protein Data Bank
env_nr	Environmental samples


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

Enter Query Sequence

Enter accession number, gi, or FASTA sequence

```
>Query sequence
MSSAAAAAAGAAGGGALFQPQSVSTANSSSSNNNNSSTPAALATHSPTSNSPVSGASSASSLLTAAFG
FGGSSAKMLNELFGRQMKQAQDATSGLPQSLDNAMLAAAMETATSAELLIGSLNSTSKLLQQQHNNNS
PANSTPMSNGTNASISPGSAHSSSHSHQGVSPKGSRRVSACSDRSLEAAAADVAGGSPPRAASVSSLN
ASSGEQHQSQLQHDLVAHHMLRNILQGKKELMQLDQELRTAMQQQQQLQEKEQLHSKLNNNNNNNIA
```

Query subrange

From

To

Or, upload file

Job Title: Query sequence

Enter a descriptive title for your BLAST search

Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism Optional: Enter organism name or id--completions will be suggested

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

Entrez Query Optional

Enter an Entrez query to limit search

Program Selection

Algorithm

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

Choose a BLAST algorithm

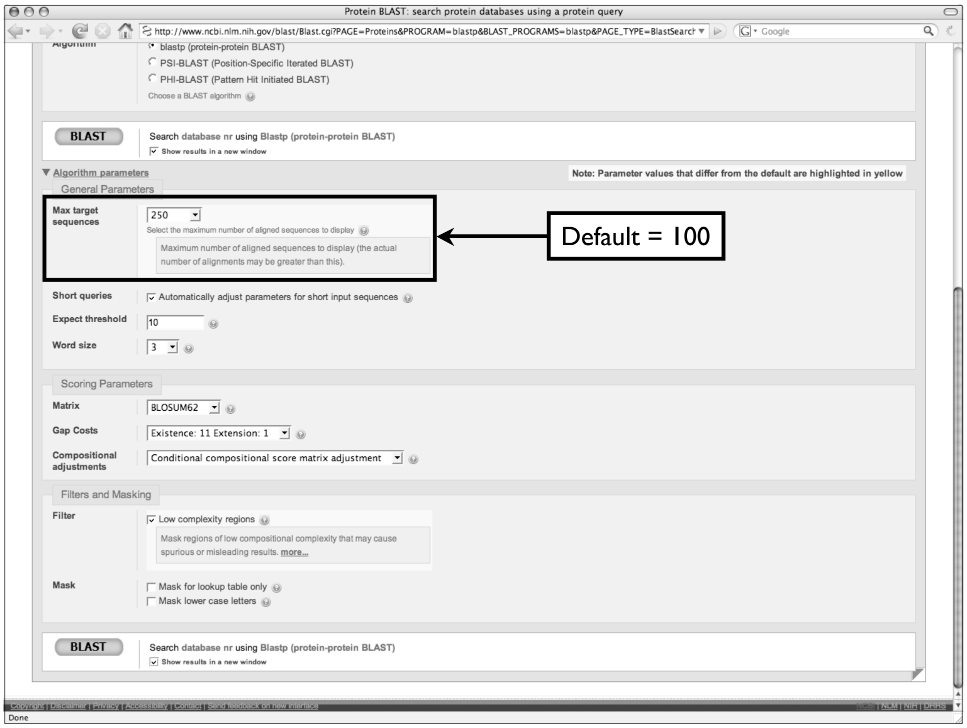
Search database nr using Blastp (protein-protein BLAST)

□ Show results in a new window

▶ Algorithm parameters

Note: Parameter values that differ from the default are highlighted in yellow

Limit by organism or taxonomic group



Protein BLAST: search protein databases using a protein query

http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?PAGE=Proteins&PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch

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

BLAST — Search database nr using Blastp (protein-protein BLAST)
☑ Show results in a new window

▼ Algorithm parameters — Note: Parameter values that differ from the default are highlighted in yellow

General Parameters

Max target sequences: 250
Select the maximum number of aligned sequences to display
Maximum number of aligned sequences to display (the actual number of alignments may be greater than this).

Short queries: ☑ Automatically adjust parameters for short input sequences
Expect threshold: 10
Word size: 3

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 regions of low compositional complexity that may cause spurious or misleading results. more...

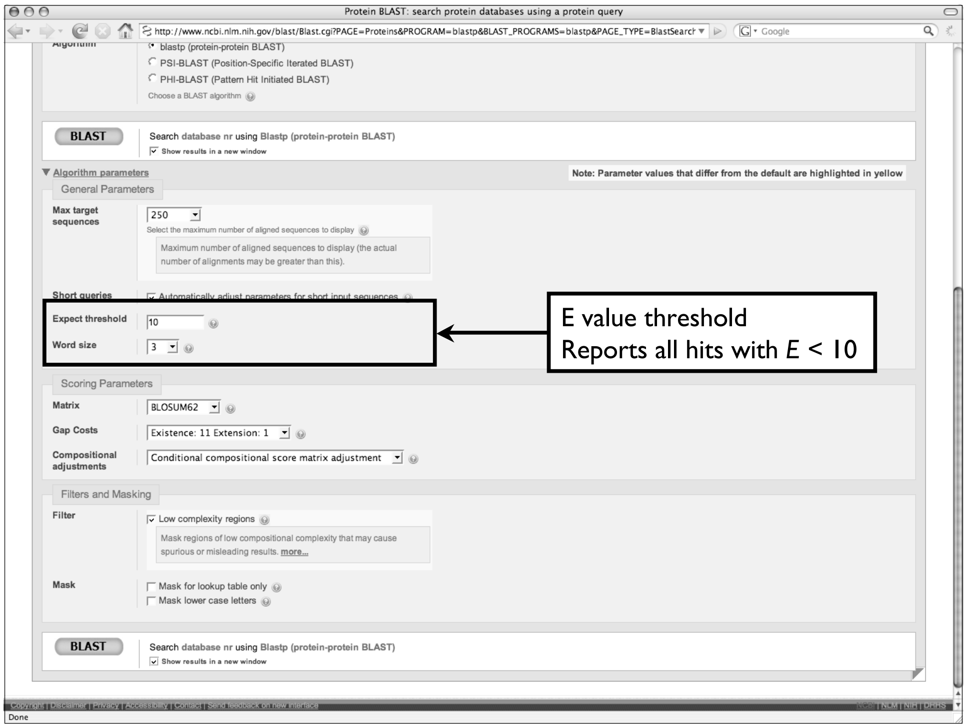
Mask: ☐ Mask for lookup table only
☐ Mask lower case letters

BLAST — Search database nr using Blastp (protein-protein BLAST)
☑ Show results in a new window

Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback on new interface — NCBI | NLM | NIH | DHHS

Done

Default = 100



Protein BLAST: search protein databases using a protein query

http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?PAGE=Proteins&PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch

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

BLAST — Search database nr using Blastp (protein-protein BLAST)
☑ Show results in a new window

▼ Algorithm parameters — Note: Parameter values that differ from the default are highlighted in yellow

General Parameters

Max target sequences: 250
Select the maximum number of aligned sequences to display
Maximum number of aligned sequences to display (the actual number of alignments may be greater than this).

Short queries: ☑ Automatically adjust parameters for short input sequences
Expect threshold: 10
Word size: 3

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 regions of low compositional complexity that may cause spurious or misleading results. more...

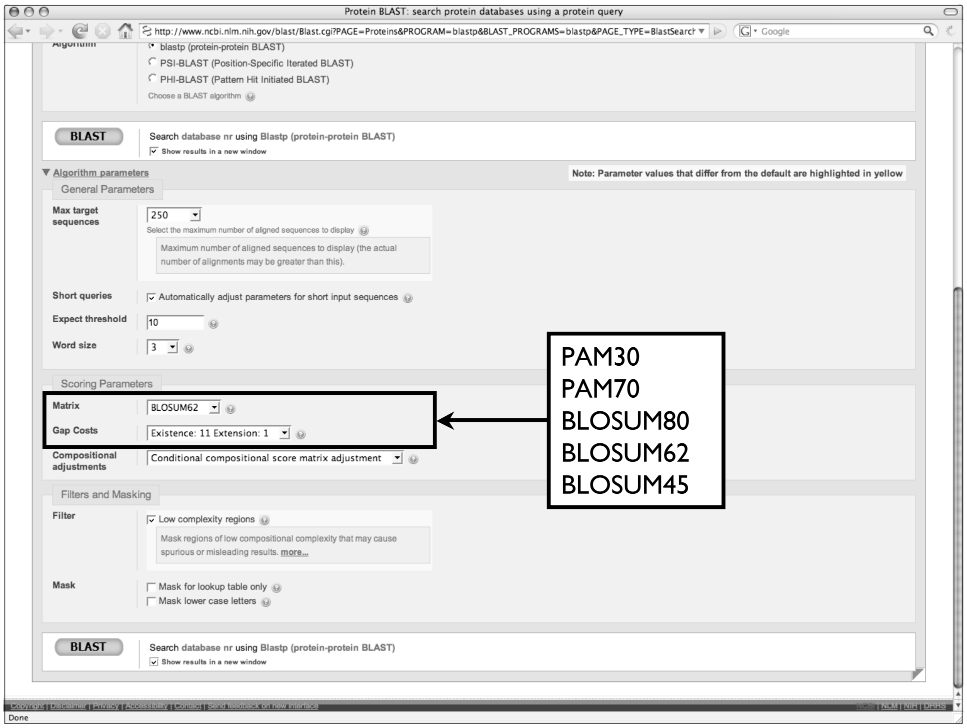
Mask: ☐ Mask for lookup table only
☐ Mask lower case letters

BLAST — Search database nr using Blastp (protein-protein BLAST)
☑ Show results in a new window

Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback on new interface — NCBI | NLM | NIH | DHHS

Done

E value threshold
Reports all hits with *E* < 10



Protein BLAST: search protein databases using a protein query

http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?PAGE=Proteins&PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch

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

BLAST — Search database nr using Blastp (protein-protein BLAST)
☑ Show results in a new window

Algorithm parameters

Note: Parameter values that differ from the default are highlighted in yellow

General Parameters

Max target sequences: 250
Select the maximum number of aligned sequences to display
Maximum number of aligned sequences to display (the actual number of alignments may be greater than this).

Short queries: ☑ Automatically adjust parameters for short input sequences

Expect threshold: 10

Word size: 3

Scoring Parameters

Matrix: BLOSUM62

Gap Costs: Existence: 11 Extension: 1

Compositional adjustments: Conditional compositional score matrix adjustment

Filters and Masking

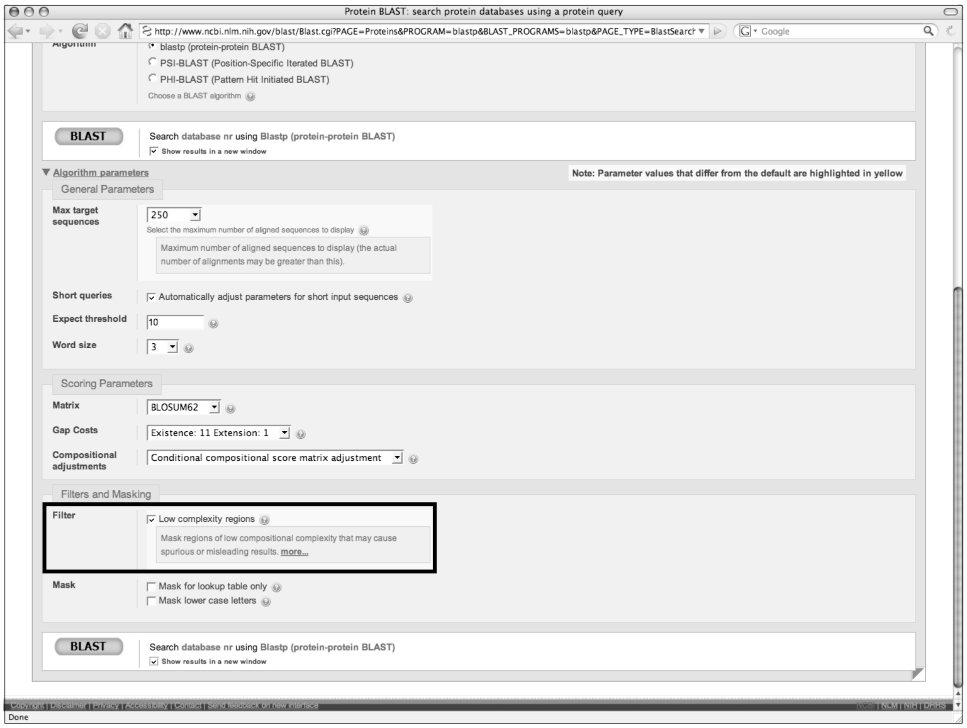
Filter: ☑ Low complexity regions
Mask regions of low compositional complexity that may cause spurious or misleading results. more...

Mask: □ Mask for lookup table only
□ Mask lower case letters

BLAST — Search database nr using Blastp (protein-protein BLAST)
☑ Show results in a new window

Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback on new interface
NCBI | NLM | NIH | DHHS
Done

PAM30
PAM70
BLOSUM80
BLOSUM62
BLOSUM45



Protein BLAST: search protein databases using a protein query

http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?PAGE=Proteins&PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch

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

BLAST — Search database nr using Blastp (protein-protein BLAST)
☑ Show results in a new window

Algorithm parameters

Note: Parameter values that differ from the default are highlighted in yellow

General Parameters

Max target sequences: 250
Select the maximum number of aligned sequences to display
Maximum number of aligned sequences to display (the actual number of alignments may be greater than this).

Short queries: ☑ Automatically adjust parameters for short input sequences

Expect threshold: 10

Word size: 3

Scoring Parameters

Matrix: BLOSUM62

Gap Costs: Existence: 11 Extension: 1

Compositional adjustments: Conditional compositional score matrix adjustment

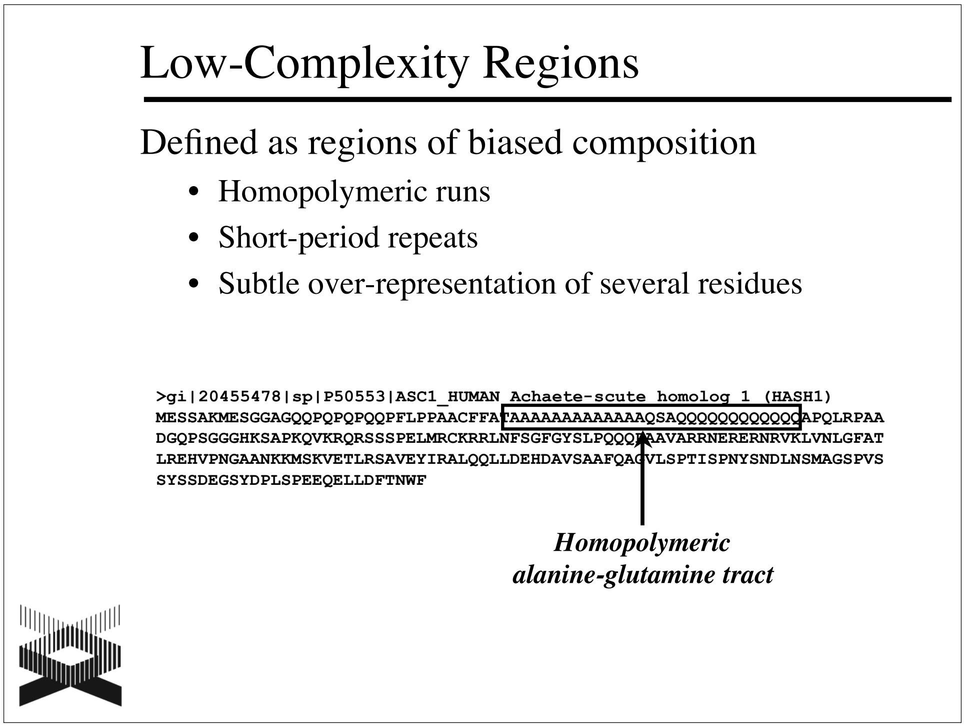
Filters and Masking

Filter: ☑ Low complexity regions
Mask regions of low compositional complexity that may cause spurious or misleading results. more...

Mask: □ Mask for lookup table only
□ Mask lower case letters

BLAST — Search database nr using Blastp (protein-protein BLAST)
☑ Show results in a new window

Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback on new interface
NCBI | NLM | NIH | DHHS
Done



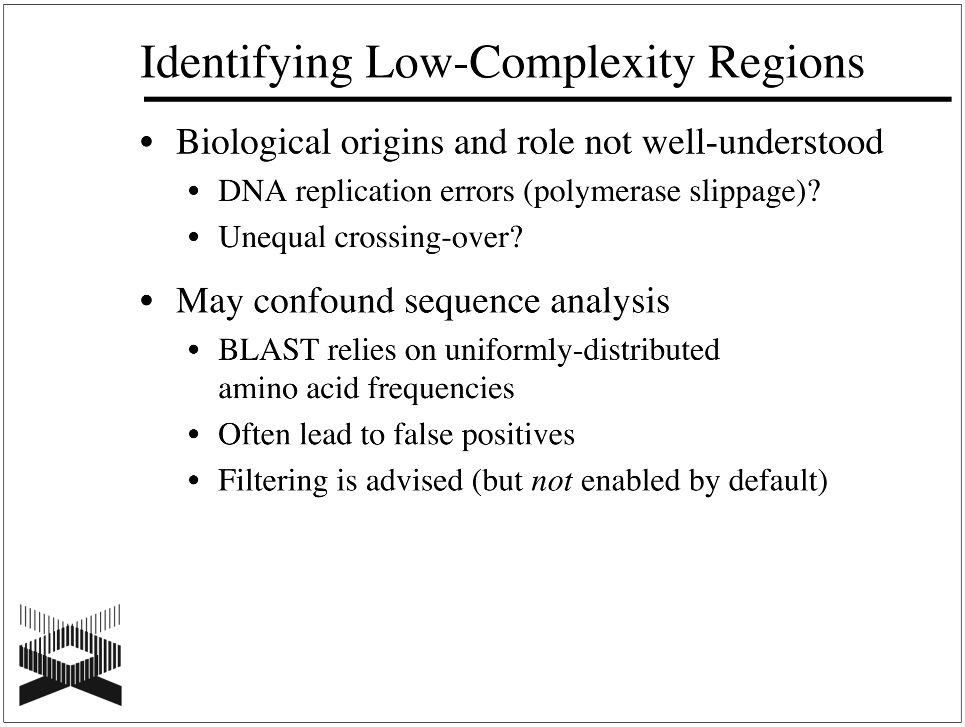
Low-Complexity Regions

Defined as regions of biased composition

- Homopolymeric runs
- Short-period repeats
- Subtle over-representation of several residues

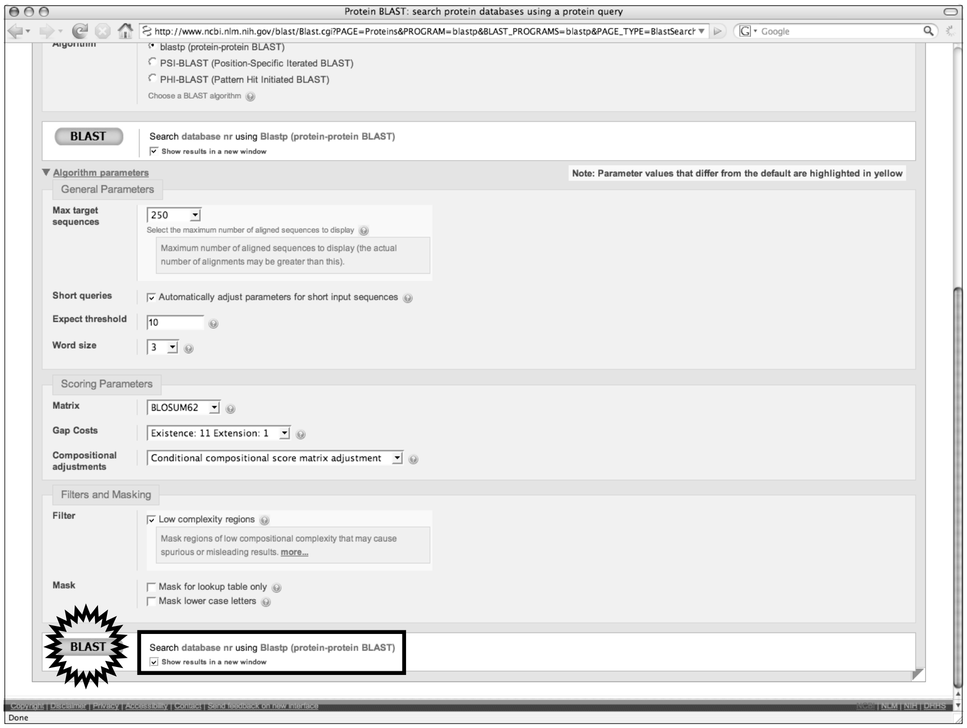
```
>gi|20455478|sp|P50553|ASC1_HUMAN Achaete-scute homolog 1 (HASH1)
MESSAKMESGGAGQQPQPQPQQPFLPPAACFFATAAAAAAAAAAAAAQSAQQQQQQQQQQQQAPQLRPAA
DGQPSGGGHKSAPKQVKRQRSSSPELMRCKRRLNFSGFGYSLPQQQPAAVARRNERERNRVKLVNLGFAT
LREHVPNGAANKKMSKVETLRSAVEYIRALQQLLDEHDAVSAAFQAGVLSPTISPNYSNDLNSMAGSPVS
SYSSDEGSYDPLSPEEQELLDFTNWF
```

Homopolymeric alanine-glutamine tract



Identifying Low-Complexity Regions

- Biological origins and role not well-understood
  - DNA replication errors (polymerase slippage)?
  - Unequal crossing-over?
- 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 protein databases using a protein query
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?PAGE=Proteins&PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch
Google
BLAST
Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback on new interface
NCBI | NLM | NIH | DHHS
Done

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

Search database nr using Blastp (protein-protein BLAST)
☑ Show results in a new window

▼ Algorithm parameters

Note: Parameter values that differ from the default are highlighted in yellow

General Parameters

Max target sequences: 250
Select the maximum number of aligned sequences to display
Maximum number of aligned sequences to display (the actual number of alignments may be greater than this).

Short queries: ☑ Automatically adjust parameters for short input sequences

Expect threshold: 10

Word size: 3

Scoring Parameters

Matrix: BLOSUM62

Gap Costs: Existence: 11 Extension: 1

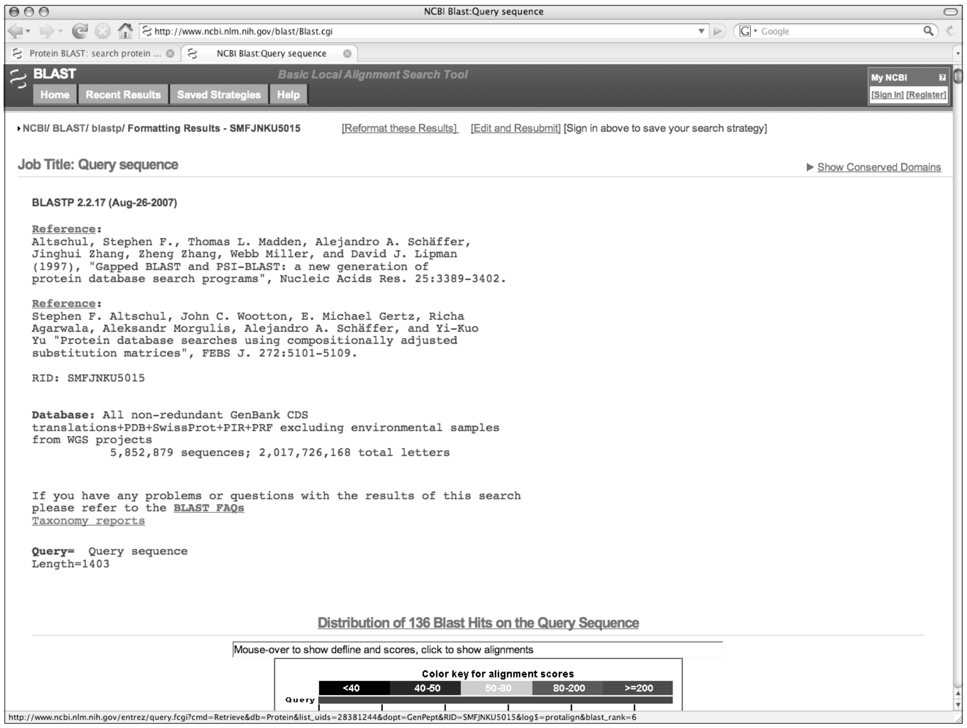
Compositional adjustments: Conditional compositional score matrix adjustment

Filters and Masking

Filter: ☑ Low complexity regions
Mask regions of low compositional complexity that may cause spurious or misleading results. more...

Mask: ☐ Mask for lookup table only
☐ Mask lower case letters

Search database nr using Blastp (protein-protein BLAST)
☑ Show results in a new window


NCBI Blast:Query sequence
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi
Google
Protein BLAST: search protein ...
NCBI Blast:Query sequence
BLAST
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
My NCBI
[Sign In] [Register]
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=Protein&list_uids=28381244&dopt=GenPept&RID=SMFJNKU5015&log$=protalign&blast_rank=6

NCBI/ BLAST/ blastp/ Formatting Results - SMFJNKU5015 [Reformat these Results] [Edit and Resubmit] [Sign in above to save your search strategy]

Job Title: Query sequence

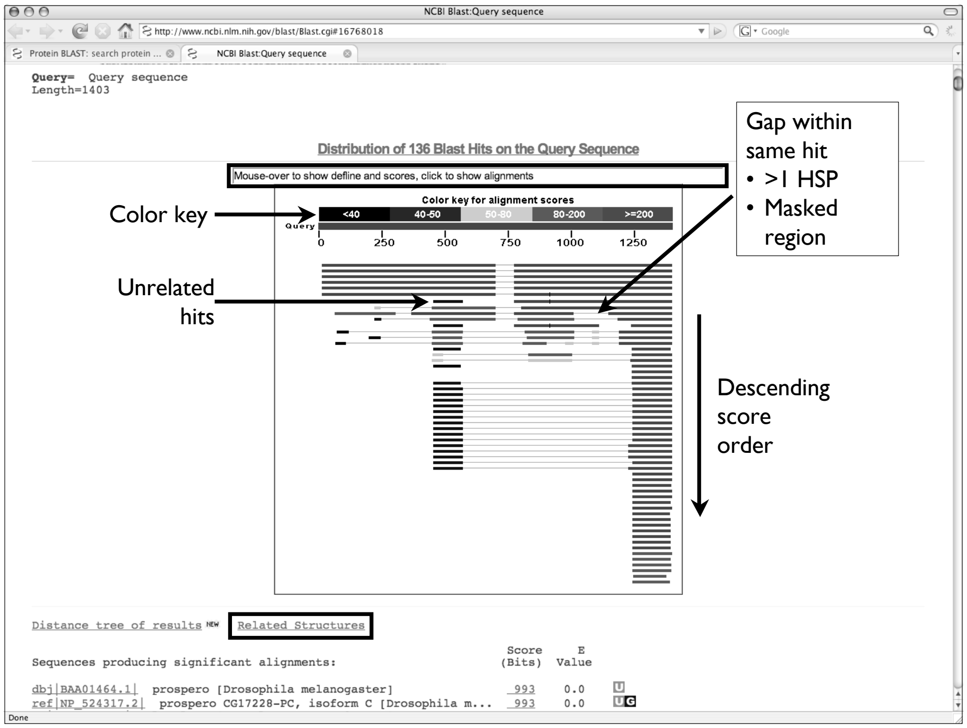
▶ Show Conserved Domains

BLASTP 2.2.17 (Aug-26-2007)

```
Reference:
Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman
(1997), "Gapped BLAST and PSI-BLAST: a new generation of
protein database search programs", Nucleic Acids Res. 25:3389-3402.

Reference:
Stephen F. Altschul, John C. Wootton, E. Michael Gertz, Richa
Agarwala, Aleksandr Morgulis, Alejandro A. Schäffer, and Yi-Kuo
Yu "Protein database searches using compositionally adjusted
substitution matrices", FEBS J. 272:5101-5109.

RID: SMFJNKU5015


Database: All non-redundant GenBank CDS
translations+PDB+SwissProt+PIR+PRF excluding environmental samples
from WGS projects
           5,852,879 sequences; 2,017,726,168 total letters


If you have any problems or questions with the results of this search
please refer to the BLAST FAQs
Taxonomy reports

Query=  Query sequence
Length=1403
```

Distribution of 136 Blast Hits on the Query Sequence

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


NCBI Blast:Query sequence
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi#16768018
Google
Protein BLAST: search protein ...
NCBI Blast:Query sequence
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
0
250
500
750
1000
1250
Done

```
Query=  Query sequence
Length=1403
```

Distribution of 136 Blast Hits on the Query Sequence

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

Gap within same hit
- >1 HSP
- Masked region

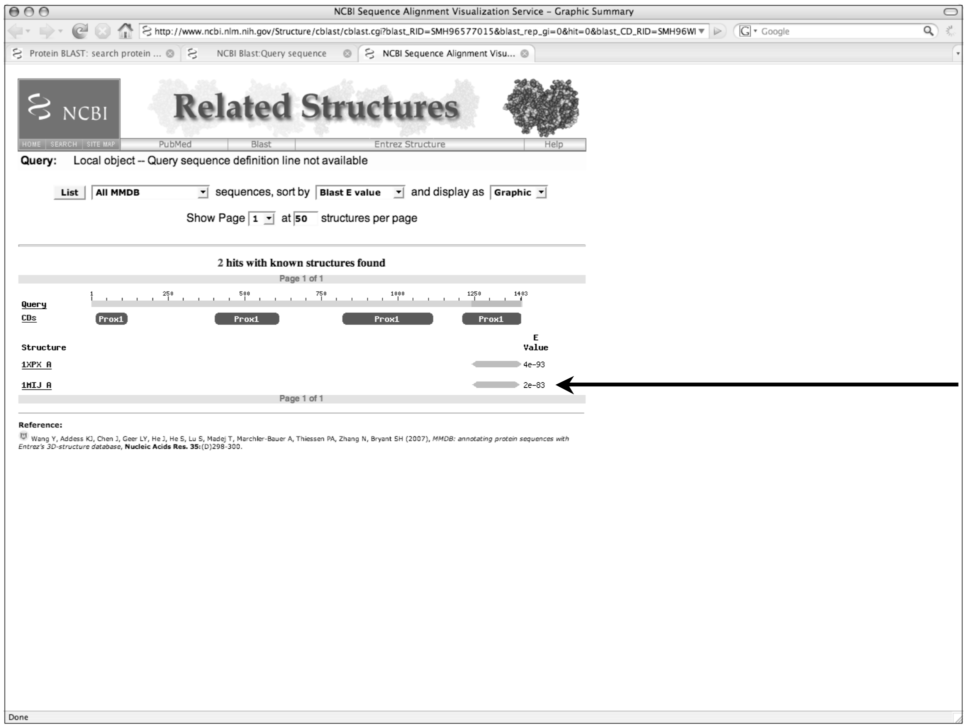
Color key

Unrelated hits

Descending score order

Distance tree of results NEW Related Structures

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

dbj|BAA01464.1|  prospero [Drosophila melanogaster]             993    0.0
ref|NP_524317.2|  prospero CG17228-PC, isoform C [Drosophila m...  993    0.0
```


NCBI Sequence Alignment Visualization Service – Graphic Summary
http://www.ncbi.nlm.nih.gov/Structure/cblast/cblast.cgi?blast_RID=SMH96577015&blast_rep_gi=0&hit=0&blast_CD_RID=SMH96Wl
Google
Protein BLAST: search protein ...
NCBI Blast:Query sequence
NCBI Sequence Alignment Visu...
NCBI
Related Structures
HOME
SEARCH
SITE MAP
PubMed
Blast
Entrez Structure
Help
1
250
500
750
1000
1250
1403
Query
CDs
Prox1
Prox1
Prox1
Prox1
Done

Query: Local object -- Query sequence definition line not available

List | All MMDB | sequences, sort by | Blast E value | and display as | Graphic

Show Page 1 at 50 structures per page

2 hits with known structures found

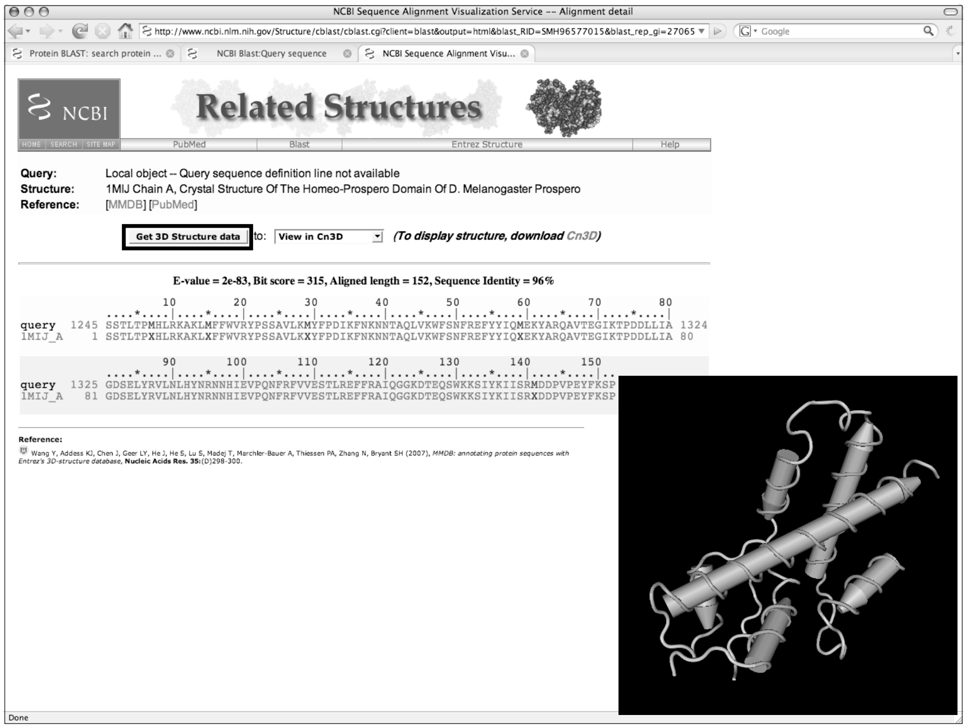
Page 1 of 1

Structure	E Value
1XPX A	4e-93
1MIJ A	2e-83

Page 1 of 1

Reference:

Wang Y, Addess KJ, Chen J, Geer LY, He J, He S, Lu S, Madej T, Marchler-Bauer A, Thiessen PA, Zhang N, Bryant SH (2007), *MMDB: annotating protein sequences with Entrez's 3D-structure database*, **Nucleic Acids Res. 35:**(D)298-300.


NCBI Sequence Alignment Visualization Service -- Alignment detail
http://www.ncbi.nlm.nih.gov/Structure/cblast/cblast.cgi?client=blast&output=html&blast_RID=SMH96577015&blast_rep_gi=27065
Google
Protein BLAST: search protein ...
NCBI Blast:Query sequence
NCBI Sequence Alignment Visu...
NCBI
Related Structures
HOME
SEARCH
SITE MAP
PubMed
Blast
Entrez Structure
Help
Get 3D Structure data
View in Cn3D
Done

Query: Local object -- Query sequence definition line not available
Structure: 1MIJ Chain A, Crystal Structure Of The Homeo-Prospero Domain Of D. Melanogaster Prospero
Reference: [MMDB] [PubMed]

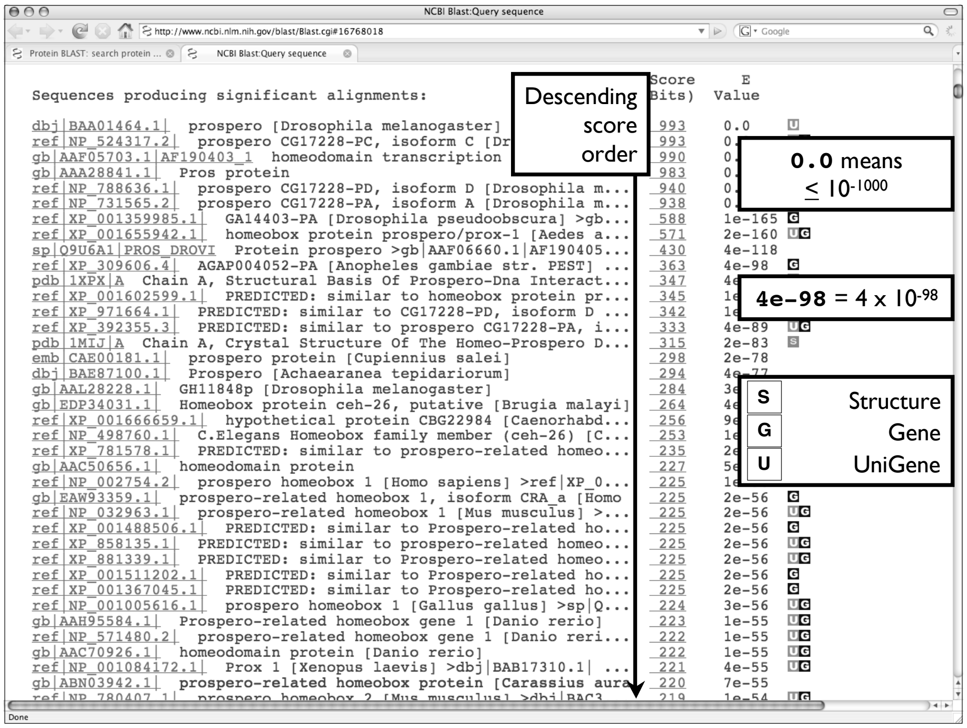
Get 3D Structure data to: View in Cn3D *(To display structure, download Cn3D)*

E-value = 2e-83, Bit score = 315, Aligned length = 152, Sequence Identity = 96%

```
                  10        20        30        40        50        60        70        80
          ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
query  1245 SSTLTPMHLRKAKLMFFWVRYPSSAVLKMYFPDIKFNKNNTAQLVKWFSNFREFYYIQMEKYARQAVTEGIKTPDDLLIA 1324
1MIJ_A    1 SSTLTPXHLRKAKLXFFWVRYPSSAVLKXYFPDIKFNKNNTAQLVKWFSNFREFYYIQXEKYARQAVTEGIKTPDDLLIA 80

                  90       100       110       120       130       140       150
          ....*....|....*....|....*....|....*....|....*....|....*....|....*....|..
query  1325 GDSELYRVLNLHYNRNNHIEVPQNFRFVVESTLREFFRAIQGGKDTEQSWKKSIYKIISRMDDPVPEYFKSP
1MIJ_A   81 GDSELYRVLNLHYNRNNHIEVPQNFRFVVESTLREFFRAIQGGKDTEQSWKKSIYKIISRXDDPVPEYFKSP
```

Reference:

Wang Y, Addess KJ, Chen J, Geer LY, He J, He S, Lu S, Madej T, Marchler-Bauer A, Thiessen PA, Zhang N, Bryant SH (2007), *MMDB: annotating protein sequences with Entrez's 3D-structure database*, **Nucleic Acids Res. 35:**(D)298-300.


NCBI Blast:Query sequence
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi#16768018
Google
Protein BLAST: search protein ...
NCBI Blast:Query sequence
Done

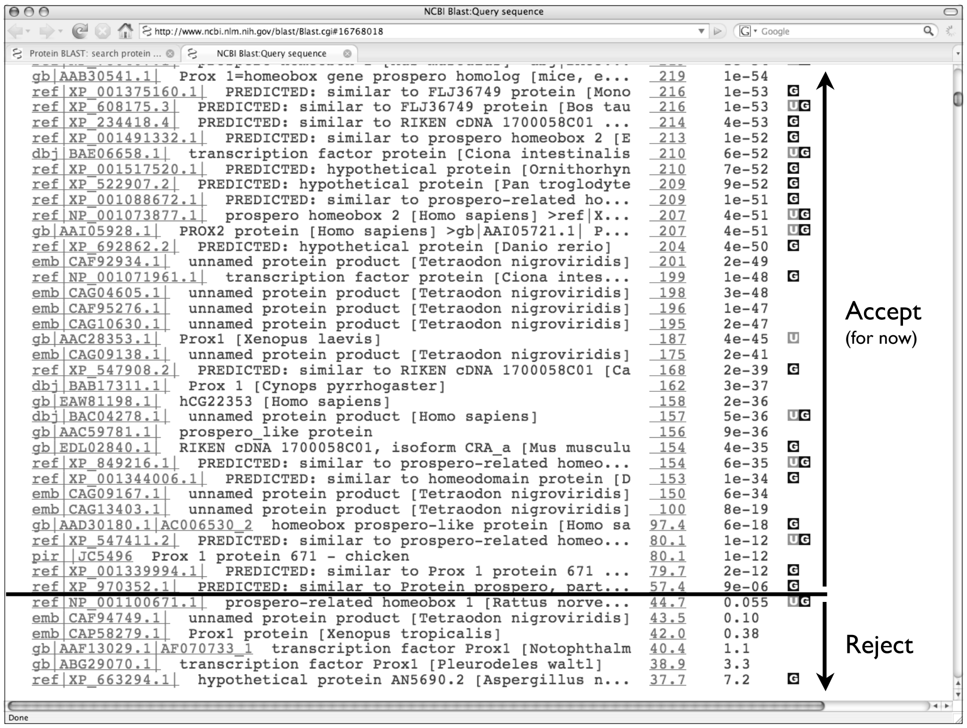
```
                                                               Score     E
Sequences producing significant alignments:                   (Bits)  Value

dbj|BAA01464.1|  prospero [Drosophila melanogaster]             993   0.0     U
ref|NP_524317.2|  prospero CG17228-PC, isoform C [Dr            993   0.
gb|AAF05703.1|AF190403_1  homeodomain transcription             990   0.
gb|AAA28841.1|  Pros protein                                    983   0.
ref|NP_788636.1|  prospero CG17228-PD, isoform D [Drosophila m...  940   0.
ref|NP_731565.2|  prospero CG17228-PA, isoform A [Drosophila m...  938   0.
ref|XP_001359985.1|  GA14403-PA [Drosophila pseudoobscura] >gb...  588   1e-165  G
ref|XP_001655942.1|  homeobox protein prospero/prox-1 [Aedes a...  571   2e-160  UG
sp|Q9U6A1|PROS_DROVI  Protein prospero >gb|AAF06660.1|AF190405...  430   4e-118
ref|XP_309606.4|  AGAP004052-PA [Anopheles gambiae str. PEST] ...  363   4e-98   G
pdb|1XPX|A  Chain A, Structural Basis Of Prospero-Dna Interact...  347   4e
ref|XP_001602599.1|  PREDICTED: similar to homeobox protein pr...  345   1e
ref|XP_971664.1|  PREDICTED: similar to CG17228-PD, isoform D ...  342   1e
ref|XP_392355.3|  PREDICTED: similar to prospero CG17228-PA, i...  333   4e-89   UG
pdb|1MIJ|A  Chain A, Crystal Structure Of The Homeo-Prospero D...  315   2e-83   S
emb|CAE00181.1|  prospero protein [Cupiennius salei]            298   2e-78
dbj|BAE87100.1|  Prospero [Achaearanea tepidariorum]            294   4e-77
gb|AAL28228.1|  GH11848p [Drosophila melanogaster]              284   3e
gb|EDP34031.1|  Homeobox protein ceh-26, putative [Brugia malayi]  264   4e
ref|XP_001666659.1|  hypothetical protein CBG22984 [Caenorhabd...  256   9e
ref|NP_498760.1|  C.Elegans Homeobox family member (ceh-26) [C...  253   1e
ref|XP_781578.1|  PREDICTED: similar to prospero-related homeo...  235   2e
gb|AAC50656.1|  homeodomain protein                             227   5e
ref|NP_002754.2|  prospero homeobox 1 [Homo sapiens] >ref|XP_0...  225   1e
gb|EAW93359.1|  prospero-related homeobox 1, isoform CRA_a [Homo  225   2e-56   G
ref|NP_032963.1|  prospero-related homeobox 1 [Mus musculus] >...  225   2e-56   UG
ref|XP_001488506.1|  PREDICTED: similar to Prospero-related ho...  225   2e-56   G
ref|XP_858135.1|  PREDICTED: similar to prospero-related homeo...  225   2e-56   UG
ref|XP_881339.1|  PREDICTED: similar to Prospero-related homeo...  225   2e-56   UG
ref|XP_001511202.1|  PREDICTED: similar to Prospero-related ho...  225   2e-56   G
ref|XP_001367045.1|  PREDICTED: similar to Prospero-related ho...  225   2e-56   G
ref|NP_001005616.1|  prospero homeobox 1 [Gallus gallus] >sp|Q...  224   3e-56   UG
gb|AAH95584.1|  Prospero-related homeobox gene 1 [Danio rerio]  223   1e-55   UG
ref|NP_571480.2|  prospero-related homeobox gene 1 [Danio reri...  222   1e-55   UG
gb|AAC70926.1|  homeodomain protein [Danio rerio]               222   1e-55   UG
ref|NP_001084172.1|  Prox 1 [Xenopus laevis] >dbj|BAB17310.1| ...  221   4e-55   UG
gb|ABN03942.1|  prospero-related homeobox protein [Carassius aura  220   7e-55
ref|NP_780407.1|  prospero homeobox 2 [Mus musculus] >dbj|BAC3   219   1e-54   UG
```

Descending score order

0.0 means ≤ 10^{-1000}

4e-98 = 4 x 10^{-98}

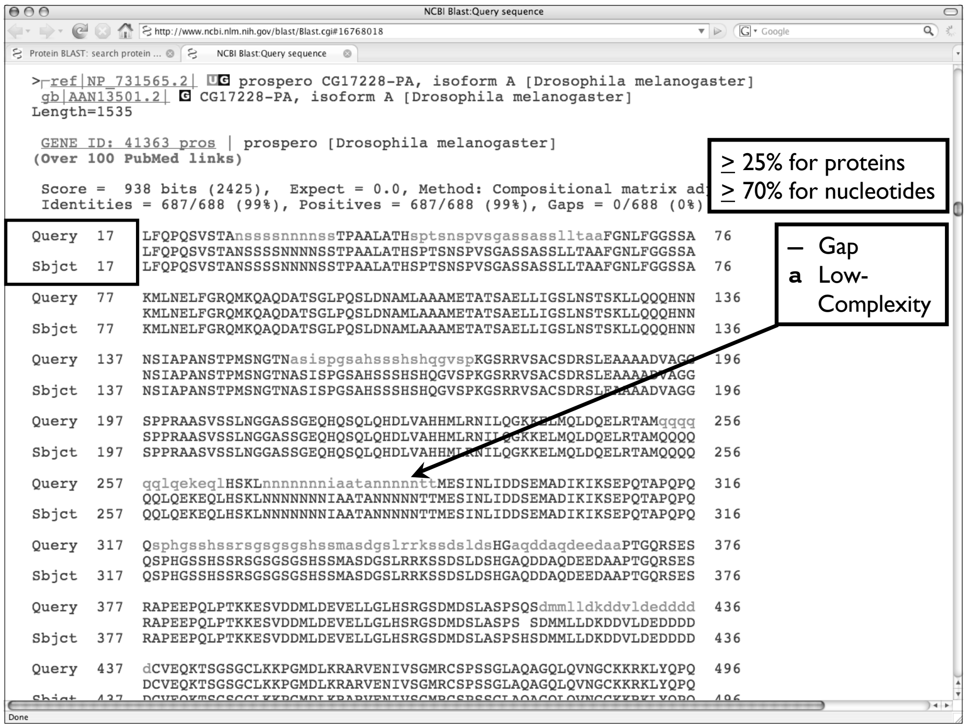
S	Structure
G	Gene
U	UniGene


NCBI Blast:Query sequence
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi#16768018
Google
Protein BLAST: search protein ...
NCBI Blast:Query sequence
Done

```
gb|AAB30541.1|  Prox 1=homeobox gene prospero homolog [mice, e...   219    1e-54
ref|XP_001375160.1|  PREDICTED: similar to FLJ36749 protein [Mono   216    1e-53  G
ref|XP_608175.3|  PREDICTED: similar to FLJ36749 protein [Bos tau   216    1e-53  UG
ref|XP_234418.4|  PREDICTED: similar to RIKEN cDNA 1700058C01 ...   214    4e-53  G
ref|XP_001491332.1|  PREDICTED: similar to prospero homeobox 2 [E   213    1e-52  G
dbj|BAE06658.1|  transcription factor protein [Ciona intestinalis   210    6e-52  UG
ref|XP_001517520.1|  PREDICTED: hypothetical protein [Ornithorhyn   210    7e-52  G
ref|XP_522907.2|  PREDICTED: hypothetical protein [Pan troglodyte   209    9e-52  G
ref|XP_001088672.1|  PREDICTED: similar to prospero-related ho...   209    1e-51  G
ref|NP_001073877.1|  prospero homeobox 2 [Homo sapiens] >ref|X...   207    4e-51  UG
gb|AAI05928.1|  PROX2 protein [Homo sapiens] >gb|AAI05721.1| P...   207    4e-51  UG
ref|XP_692862.2|  PREDICTED: hypothetical protein [Danio rerio]     204    4e-50  G
emb|CAF92934.1|  unnamed protein product [Tetraodon nigroviridis]   201    2e-49
ref|NP_001071961.1|  transcription factor protein [Ciona intes...   199    1e-48  G
emb|CAG04605.1|  unnamed protein product [Tetraodon nigroviridis]   198    3e-48
emb|CAF95276.1|  unnamed protein product [Tetraodon nigroviridis]   196    1e-47
emb|CAG10630.1|  unnamed protein product [Tetraodon nigroviridis]   195    2e-47
gb|AAC28353.1|  Prox1 [Xenopus laevis]                              187    4e-45  U
emb|CAG09138.1|  unnamed protein product [Tetraodon nigroviridis]   175    2e-41
ref|XP_547908.2|  PREDICTED: similar to RIKEN cDNA 1700058C01 [Ca   168    2e-39  G
dbj|BAB17311.1|  Prox 1 [Cynops pyrrhogaster]                       162    3e-37
gb|EAW81198.1|  hCG22353 [Homo sapiens]                             158    2e-36
dbj|BAC04278.1|  unnamed protein product [Homo sapiens]             157    5e-36  UG
gb|AAC59781.1|  prospero_like protein                               156    9e-36
gb|EDL02840.1|  RIKEN cDNA 1700058C01, isoform CRA_a [Mus musculu   154    4e-35  G
ref|XP_849216.1|  PREDICTED: similar to prospero-related homeo...   154    6e-35  UG
ref|XP_001344006.1|  PREDICTED: similar to homeodomain protein [D   153    1e-34  G
emb|CAG09167.1|  unnamed protein product [Tetraodon nigroviridis]   150    6e-34
emb|CAG13403.1|  unnamed protein product [Tetraodon nigroviridis]   100    8e-19
gb|AAD30180.1|AC006530_2  homeobox prospero-like protein [Homo sa  97.4    6e-18  G
ref|XP_547411.2|  PREDICTED: similar to prospero-related homeo...  80.1    1e-12  UG
pir||JC5496  Prox 1 protein 671 - chicken                          80.1    1e-12
ref|XP_001339994.1|  PREDICTED: similar to Prox 1 protein 671 ...  79.7    2e-12  G
ref|XP_970352.1|  PREDICTED: similar to Protein prospero, part...  57.4    9e-06  G
ref|NP_001100671.1|  prospero-related homeobox 1 [Rattus norve...  44.7    0.055  UG
emb|CAF94749.1|  unnamed protein product [Tetraodon nigroviridis]  43.5    0.10
emb|CAP58279.1|  Prox1 protein [Xenopus tropicalis]                42.0    0.38
gb|AAF13029.1|AF070733_1  transcription factor Prox1 [Notophthalm  40.4    1.1
gb|ABG29070.1|  transcription factor Prox1 [Pleurodeles waltl]     38.9    3.3
ref|XP_663294.1|  hypothetical protein AN5690.2 [Aspergillus n...  37.7    7.2    G
```

Accept
(for now)

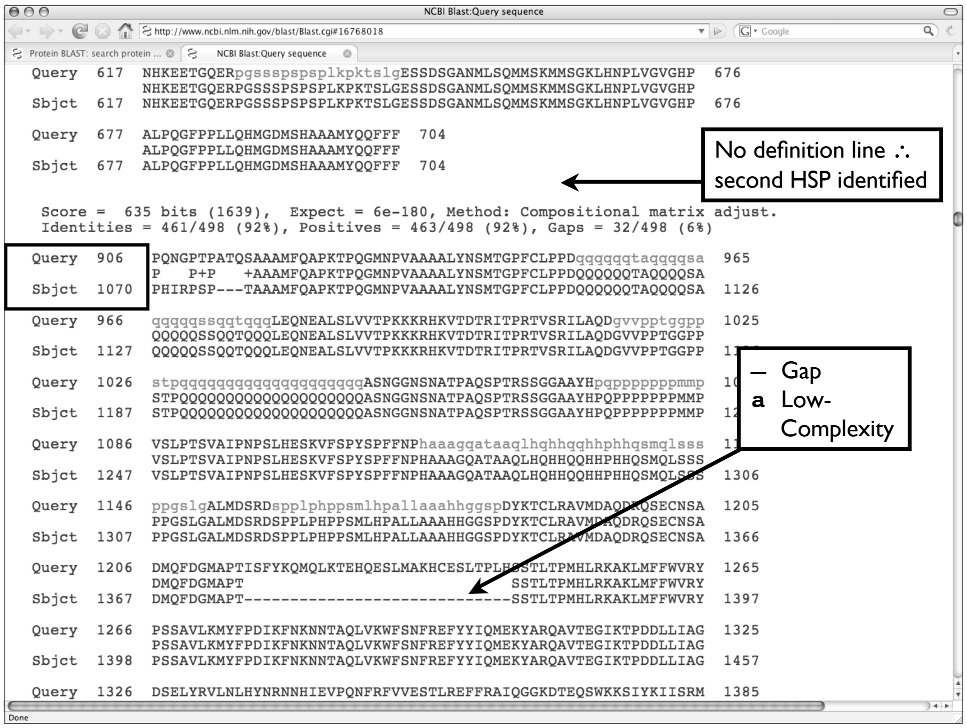
Reject


NCBI Blast:Query sequence
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi#16768018
Google
Protein BLAST: search protein ...
NCBI Blast:Query sequence
Done

```
>ref|NP_731565.2| UG prospero CG17228-PA, isoform A [Drosophila melanogaster]
 gb|AAN13501.2| G CG17228-PA, isoform A [Drosophila melanogaster]
Length=1535

 GENE ID: 41363 pros | prospero [Drosophila melanogaster]
(Over 100 PubMed links)

 Score =  938 bits (2425),  Expect = 0.0, Method: Compositional matrix ad
 Identities = 687/688 (99%), Positives = 687/688 (99%), Gaps = 0/688 (0%)

Query  17   LFQPQSVSTAnssssnnnnssTPAALATHsptsnspvsgassassllltaaFGNLFGGSSA  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  RAPEEPQLPTKKESVDDMLDEVELLGLHSRGSDMDSLASPSQSdmmlldkddvldedddd  436
            RAPEEPQLPTKKESVDDMLDEVELLGLHSRGSDMDSLASPS SDMMLLDKDDVLDEDDDD
Sbjct  377  RAPEEPQLPTKKESVDDMLDEVELLGLHSRGSDMDSLASPSHSDMMLLDKDDVLDEDDDD  436

Query  437  dCVEQKTSGSGCLKKPGMDLKRARVENIVSGMRCSPSSGLAQAGQLQVNGCKKRKLYQPQ  496
            DCVEQKTSGSGCLKKPGMDLKRARVENIVSGMRCSPSSGLAQAGQLQVNGCKKRKLYQPQ
Sbjct  437  DCVEQKTSGSGCLKKPGMDLKRARVENIVSGMRCSPSSGLAQAGQLQVNGCKKRKLYQPQ  496
```

≥ 25% for proteins
≥ 70% for nucleotides

— Gap
a Low-Complexity



NCBI Blast:Query sequence
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi#16768018
Google
Protein BLAST: search protein ...
NCBI Blast:Query sequence
Done

```
Query  617   NHKEETGQERpgssspspsplkpktslgESSDSGANMLSQMMSKMMSGKLHNPLVGVGHP  676
             NHKEETGQERPGSSSPSPSPLKPKTSLGESSDSGANMLSQMMSKMMSGKLHNPLVGVGHP
Sbjct  617   NHKEETGQERPGSSSPSPSPLKPKTSLGESSDSGANMLSQMMSKMMSGKLHNPLVGVGHP  676

Query  677   ALPQGFPPLLQHMGDMSHAAAMYQQFFF  704
             ALPQGFPPLLQHMGDMSHAAAMYQQFFF
Sbjct  677   ALPQGFPPLLQHMGDMSHAAAMYQQFFF  704


 Score =  635 bits (1639),  Expect = 6e-180, Method: Compositional matrix adjust.
 Identities = 461/498 (92%), Positives = 463/498 (92%), Gaps = 32/498 (6%)

Query  906   PQNGPTPATQSAAAMFQAPKTPQGMNPVAAAALYNSMTGPFCLPPDqqqqqqtaqqqqsa  965
             P   P+P   +AAAMFQAPKTPQGMNPVAAAALYNSMTGPFCLPPDQQQQQQTAQQQQSA
Sbjct  1070  PHIRPSP---TAAAMFQAPKTPQGMNPVAAAALYNSMTGPFCLPPDQQQQQQTAQQQQSA  1126

Query  966   qqqqqssqqtqqqLEQNEALSLVVTPKKKRHKVTDTRITPRTVSRILAQDgvvpptggpp  1025
             QQQQQSSQQTQQQLEQNEALSLVVTPKKKRHKVTDTRITPRTVSRILAQDGVVPPTGGPP
Sbjct  1127  QQQQQSSQQTQQQLEQNEALSLVVTPKKKRHKVTDTRITPRTVSRILAQDGVVPPTGGPP  11

Query  1026  stpqqqqqqqqqqqqqqqqqqqqASNGGNSNATPAQSPTRSSGGAAYHpqppppppppmmp  10
             STPQQQQQQQQQQQQQQQQQQQQASNGGNSNATPAQSPTRSSGGAAYHPQPPPPPPPPMMP
Sbjct  1187  STPQQQQQQQQQQQQQQQQQQQQASNGGNSNATPAQSPTRSSGGAAYHPQPPPPPPPPMMP  12

Query  1086  VSLPTSVAIPNPSLHESKVFSPYSPFFNPhaaagqataaqlhqhhqqhhphhqsmqlsss  11
             VSLPTSVAIPNPSLHESKVFSPYSPFFNPHAAAGQATAAQLHQHHQQHHPHHQSMQLSSS
Sbjct  1247  VSLPTSVAIPNPSLHESKVFSPYSPFFNPHAAAGQATAAQLHQHHQQHHPHHQSMQLSSS  1306

Query  1146  ppgslgALMDSRDspplphppsmlhpallaaahhggspDYKTCLRAVMDAQDRQSECNSA  1205
             PPGSLGALMDSRDSPPLPHPPSMLHPALLAAAHHGGSPDYKTCLRAVMDAQDRQSECNSA
Sbjct  1307  PPGSLGALMDSRDSPPLPHPPSMLHPALLAAAHHGGSPDYKTCLRAVMDAQDRQSECNSA  1366

Query  1206  DMQFDGMAPTISFYKQMQLKTEHQESLMAKHCESLTPLHSSTLTPMHLRKAKLMFFWVRY  1265
             DMQFDGMAPT                             SSTLTPMHLRKAKLMFFWVRY
Sbjct  1367  DMQFDGMAPT-----------------------------SSTLTPMHLRKAKLMFFWVRY  1397

Query  1266  PSSAVLKMYFPDIKFNKNNTAQLVKWFSNFREFYYIQMEKYARQAVTEGIKTPDDLLIAG  1325
             PSSAVLKMYFPDIKFNKNNTAQLVKWFSNFREFYYIQMEKYARQAVTEGIKTPDDLLIAG
Sbjct  1398  PSSAVLKMYFPDIKFNKNNTAQLVKWFSNFREFYYIQMEKYARQAVTEGIKTPDDLLIAG  1457

Query  1326  DSELYRVLNLHYNRNNHIEVPQNFRFVVESTLREFFRAIQGGKDTEQSWKKSIYKIISRM  1385
```

No definition line ∴ second HSP identified

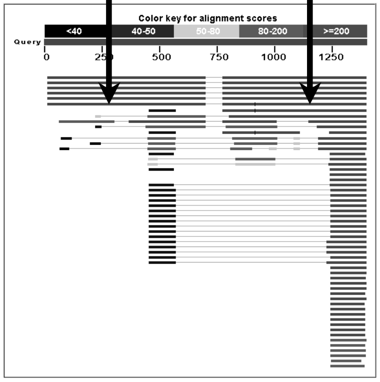
– Gap
a Low-Complexity

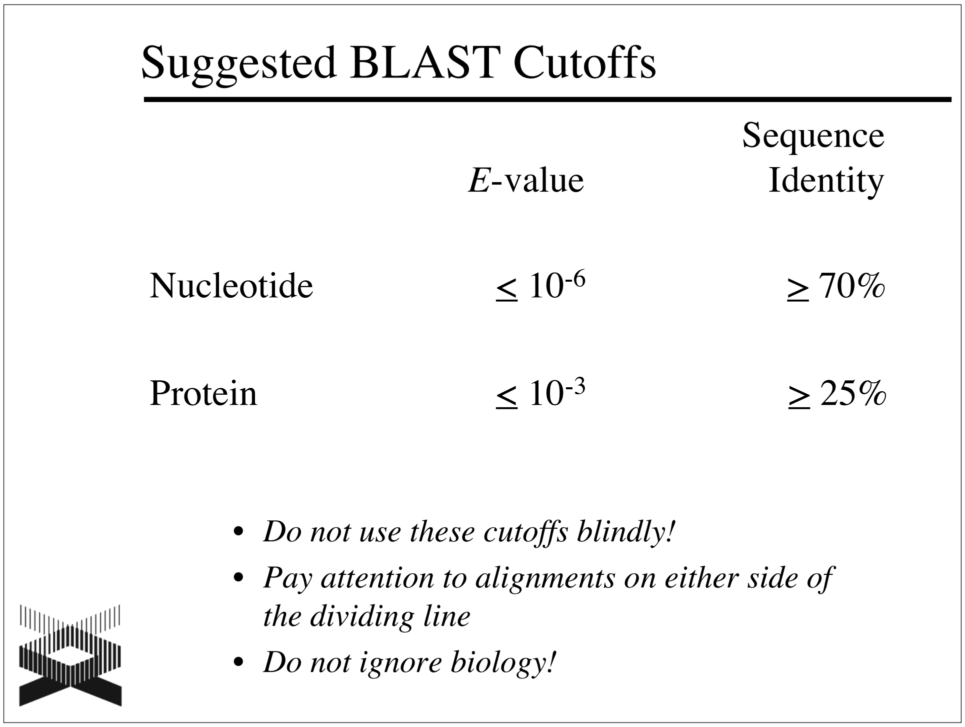
```
>ref|NP_731565.2| UG prospero CG17228-PA, isoform A [Drosophila melanogaster]
 gb|AAN13501.2| G CG17228-PA, isoform A [Drosophila melanogaster]
Length=1535

 Score =  938 bits (2425),  Expect = 0.0, Method: Compositional matrix adjust.
 Identities = 687/688 (99%), Positives = 687/688 (99%), Gaps = 0/688 (0%)


 Score =  635 bits (1639),  Expect = 6e-180, Method: Compositional matrix adjust.
 Identities = 461/498 (92%), Positives = 463/498 (92%), Gaps = 32/498 (6%)
```

✓ ✓ ✓ ✓

HSP 1
Q: 17-704
S: 17-704

HSP 2
Q: 906-1403
S: 1070-1535

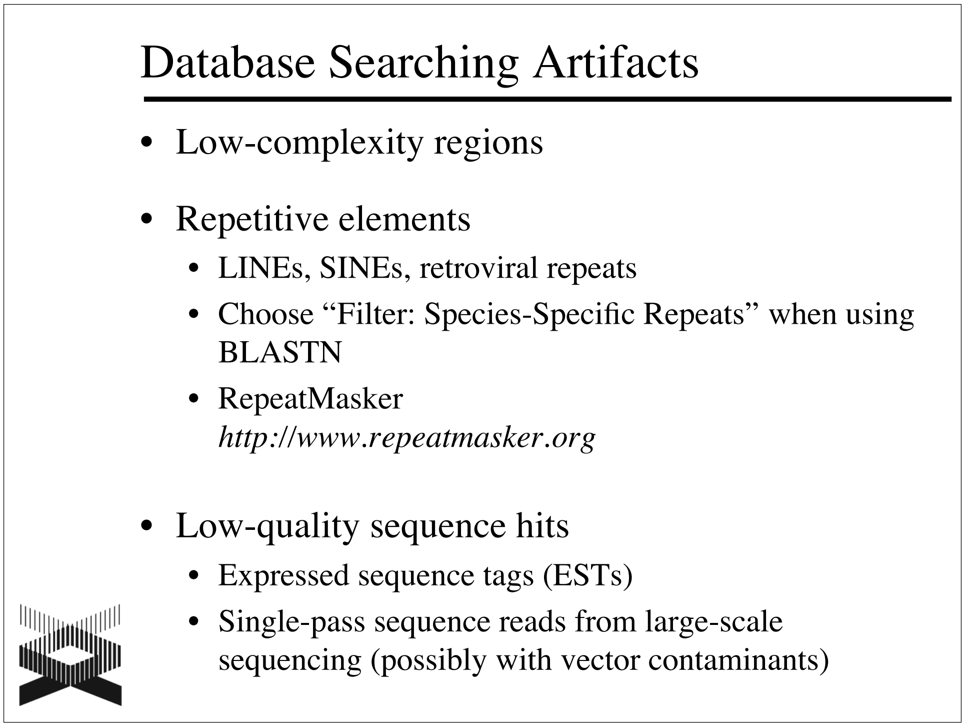




Suggested BLAST Cutoffs

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

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

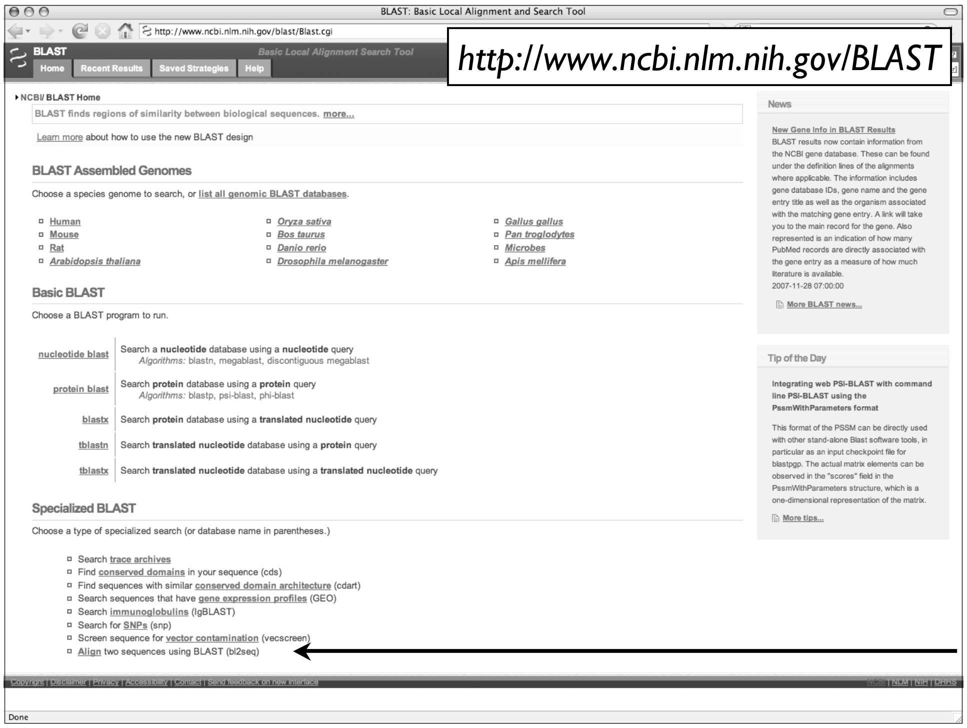


Database Searching Artifacts

- Low-complexity regions
- Repetitive elements
  - LINEs, SINEs, retroviral repeats
  - Choose “Filter: Species-Specific Repeats” when using BLASTN
  - RepeatMasker
    http://www.repeatmasker.org
- Low-quality sequence hits
  - Expressed sequence tags (ESTs)
  - Single-pass sequence reads from large-scale sequencing (possibly with vector contaminants)

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 Alignment and Search Tool
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi
BLAST
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback on new interface
NCBI | NLM | NIH | DHHS
Done

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

▸ NCBI/ BLAST Home

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

Learn more about how to use the new BLAST design

BLAST Assembled Genomes

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

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

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

Specialized BLAST

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

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

News

New Gene Info in BLAST Results
BLAST results now contain information from the NCBI gene database. These can be found under the definition lines of the alignments where applicable. The information includes gene database IDs, gene name and the gene entry title as well as the organism associated with the matching gene entry. A link will take you to the main record for the gene. Also represented is an indication of how many PubMed records are directly associated with the gene entry as a measure of how much literature is available.
2007-11-28 07:00:00

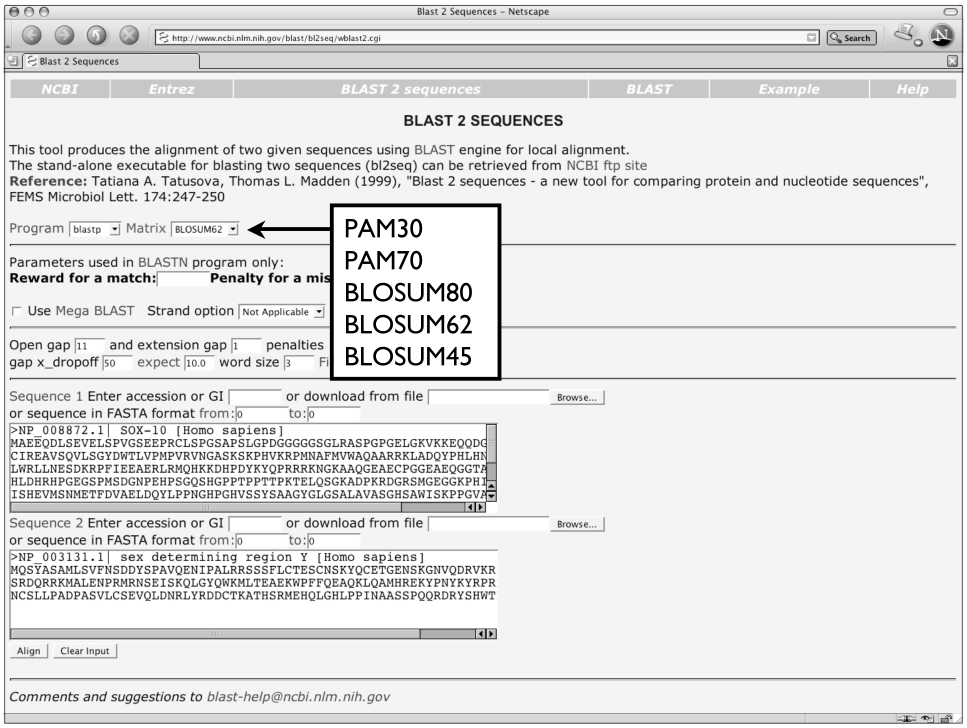
More BLAST news...

Tip of the Day

Integrating web PSI-BLAST with command line PSI-BLAST using the PssmWithParameters format

This format of the PSSM can be directly used with other stand-alone Blast software tools, in particular as an input checkpoint file for blastpgp. The actual matrix elements can be observed in the "scores" field in the PssmWithParameters structure, which is a one-dimensional representation of the matrix.

More tips...


Blast 2 Sequences – Netscape
http://www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi
Search
Blast 2 Sequences
NCBI
Entrez
BLAST 2 sequences
BLAST
Example
Help

BLAST 2 SEQUENCES

This tool produces the alignment of two given sequences using BLAST engine for local alignment.
The stand-alone executable for blasting two sequences (bl2seq) can be retrieved from NCBI ftp site
Reference: Tatiana A. Tatusova, Thomas L. Madden (1999), "Blast 2 sequences - a new tool for comparing protein and nucleotide sequences", FEMS Microbiol Lett. 174:247-250

Program blastp Matrix BLOSUM62

PAM30
PAM70
BLOSUM80
BLOSUM62
BLOSUM45

Parameters used in BLASTN program only:
Reward for a match: **Penalty for a mis**

□ Use Mega BLAST Strand option Not Applicable

Open gap 11 and extension gap 1 penalties
gap x_dropoff 50 expect 10.0 word size 3 Fi

Sequence 1 Enter accession or GI or download from file Browse...
or sequence in FASTA format from: 0 to: 0

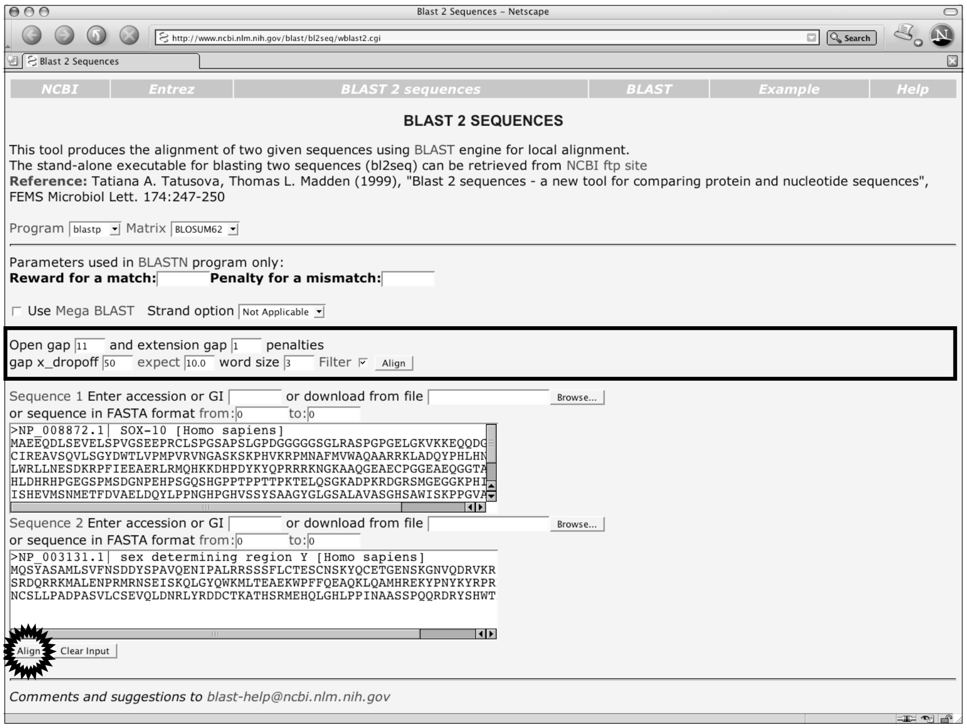
```
>NP_008872.1| SOX-10 [Homo sapiens]
MAEEQDLSEVELSPVGSEEPRCLSPGSAPSLGPDGGGGGSGLRASPGPGELGKVKKEQQDG
CIREAVSQVLSGYDWTLVPMPVRVNGASKSKPHVKRPMNAFMVWAQAARRKLADQYPHLHN
LWRLLNESDKRPFIEEAERLRMQHKKDHPDYKYQPRRRKNGKAAQGEAECPGGEAEQGGTA
HLDHRHPGEGSPMSDGNPEHPSGQSHGPPTPPTTPKTELQSGKADPKRDGRSMGEGGKPHI
ISHEVMSNMETFDVAELDQYLPPNGHPGHVSSYSAAGYGLGSALAVASGHSAWISKPPGVA
```

Sequence 2 Enter accession or GI or download from file Browse...
or sequence in FASTA format from: 0 to: 0

```
>NP_003131.1| sex determining region Y [Homo sapiens]
MQSYASAMLSVFNSDDYSPAVQENIPALRRSSSFLCTESCNSKYQCETGENSKGNVQDRVKR
SRDQRRKMALENPRMRNSEISKQLGYQWKMLTEAEKWPFFQEAQKLQAMHREKYPNYKYRPR
NCSLLPADPASVLCSEVQLDNRLYRDDCTKATHSRMEHQLGHLPPINAASSPQQRDRYSHWT
```

Align Clear Input

Comments and suggestions to blast-help@ncbi.nlm.nih.gov


Blast 2 Sequences – Netscape
http://www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi
Search
Blast 2 Sequences
NCBI
Entrez
BLAST 2 sequences
BLAST
Example
Help

BLAST 2 SEQUENCES

This tool produces the alignment of two given sequences using BLAST engine for local alignment.
The stand-alone executable for blasting two sequences (bl2seq) can be retrieved from NCBI ftp site
Reference: Tatiana A. Tatusova, Thomas L. Madden (1999), "Blast 2 sequences - a new tool for comparing protein and nucleotide sequences", FEMS Microbiol Lett. 174:247-250

Program blastp Matrix BLOSUM62

Parameters used in BLASTN program only:
Reward for a match: **Penalty for a mismatch:**

□ Use Mega BLAST Strand option Not Applicable

Open gap 11 and extension gap 1 penalties
gap x_dropoff 50 expect 10.0 word size 3 Filter ☑ Align

Sequence 1 Enter accession or GI or download from file Browse...
or sequence in FASTA format from: 0 to: 0

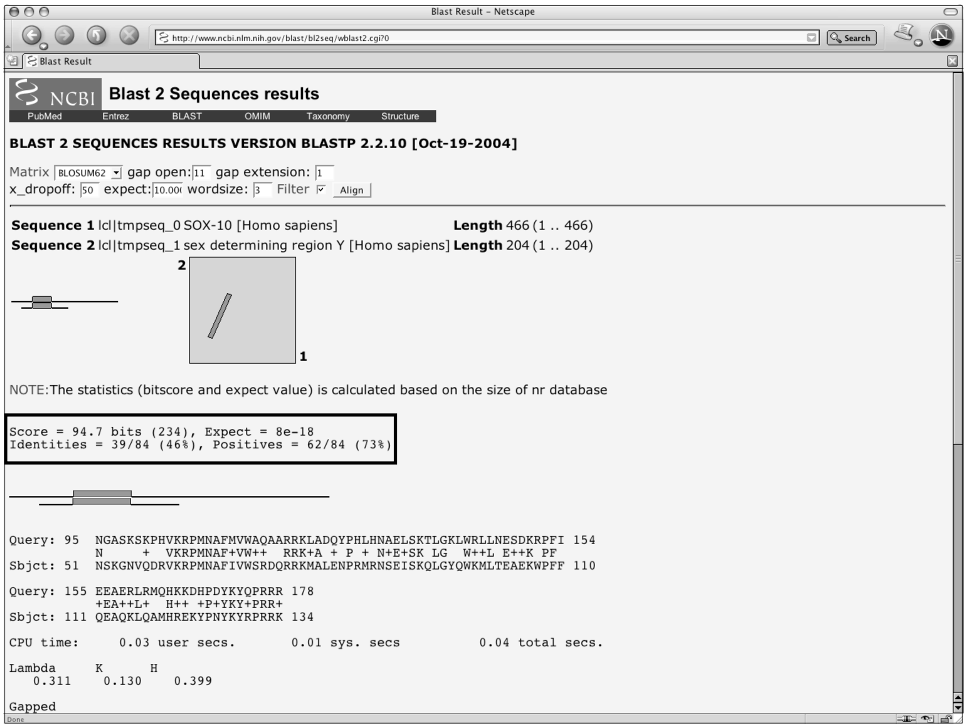
```
>NP_008872.1| SOX-10 [Homo sapiens]
MAEEQDLSEVELSPVGSEEPRCLSPGSAPSLGPDGGGGGSGLRASPGPGELGKVKKEQQDG
CIREAVSQVLSGYDWTLVPMPVRVNGASKSKPHVKRPMNAFMVWAQAARRKLADQYPHLHN
LWRLLNESDKRPFIEEAERLRMQHKKDHPDYKYQPRRRKNGKAAQGEAECPGGEAEQGGTA
HLDHRHPGEGSPMSDGNPEHPSGQSHGPPTPPTTPKTELQSGKADPKRDGRSMGEGGKPHI
ISHEVMSNMETFDVAELDQYLPPNGHPGHVSSYSAAGYGLGSALAVASGHSAWISKPPGVA
```

Sequence 2 Enter accession or GI or download from file Browse...
or sequence in FASTA format from: 0 to: 0

```
>NP_003131.1| sex determining region Y [Homo sapiens]
MQSYASAMLSVFNSDDYSPAVQENIPALRRSSSFLCTESCNSKYQCETGENSKGNVQDRVKR
SRDQRRKMALENPRMRNSEISKQLGYQWKMLTEAEKWPFFQEAQKLQAMHREKYPNYKYRPR
NCSLLPADPASVLCSEVQLDNRLYRDDCTKATHSRMEHQLGHLPPINAASSPQQRDRYSHWT
```

Align Clear Input

Comments and suggestions to blast-help@ncbi.nlm.nih.gov



Blast 2 Sequences results

BLAST 2 SEQUENCES RESULTS VERSION BLASTP 2.2.10 [Oct-19-2004]

Matrix BLOSUM62 gap open: 11 gap extension: 1
x_dropoff: 50 expect: 10.000 wordsize: 3 Filter ☑ Align

Sequence 1 lcl|tmpseq_0 SOX-10 [Homo sapiens] **Length** 466 (1 .. 466)
Sequence 2 lcl|tmpseq_1 sex determining region Y [Homo sapiens] **Length** 204 (1 .. 204)

NOTE:The statistics (bitscore and expect value) is calculated based on the size of nr database

```
Score = 94.7 bits (234), Expect = 8e-18
Identities = 39/84 (46%), Positives = 62/84 (73%)


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

CPU time:     0.03 user secs.       0.01 sys. secs          0.04 total secs.

Lambda     K      H
   0.311    0.130    0.399

Gapped
```

MegaBLAST

- Optimized for aligning very long and/or highly-similar sequences
- Good for batch nucleotide searches
- Search targets include
  - Entire eukaryotic genomes
  - Complete chromosomes and contigs from RefSeq
- Run speeds approximately 10 times faster than BLASTN
  - Adjusted word size
  - Different gap scoring scheme

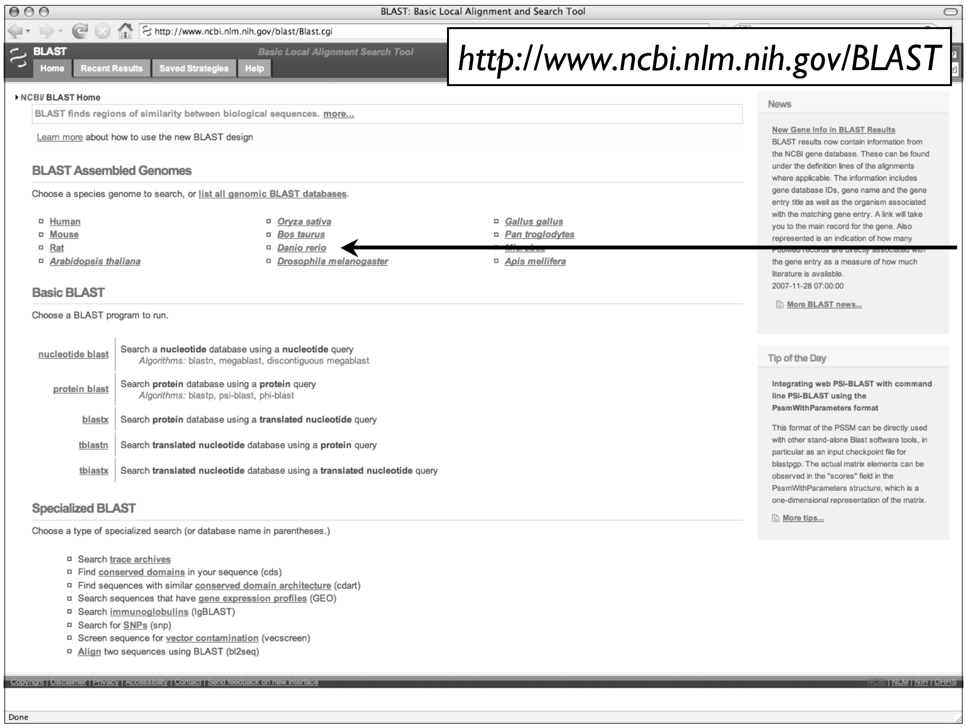
BLASTN *vs*. MegaBLAST

- Word size
  - BLASTN default = 11
  - MegaBLAST default = 28

- *Non-affine* gap penalties

$$\boxed{\text{Deduction for a gap} = r/2 - q}$$

where	r = match reward	(default = 1)
	q = mismatch penalty	(default = –2)
and	**no penalty for opening the gap**	



BLAST: Basic Local Alignment and Search Tool

http://www.ncbi.nlm.nih.gov/blast/Blast.cgi

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

NCBI/ BLAST Home

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

Learn more about how to use the new BLAST design

BLAST Assembled Genomes

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

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

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

Specialized BLAST

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

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

News

New Gene Info in BLAST Results
BLAST results now contain information from the NCBI gene database. These can be found under the definition lines of the alignments where applicable. The information includes gene database IDs, gene name and the gene entry title as well as the organism associated with the matching gene entry. A link will take you to the main record for the gene. Also represented is an indication of how many PubMed records are directly associated with the gene entry as a measure of how much literature is available.
2007-11-28 07:00:00

More BLAST news...

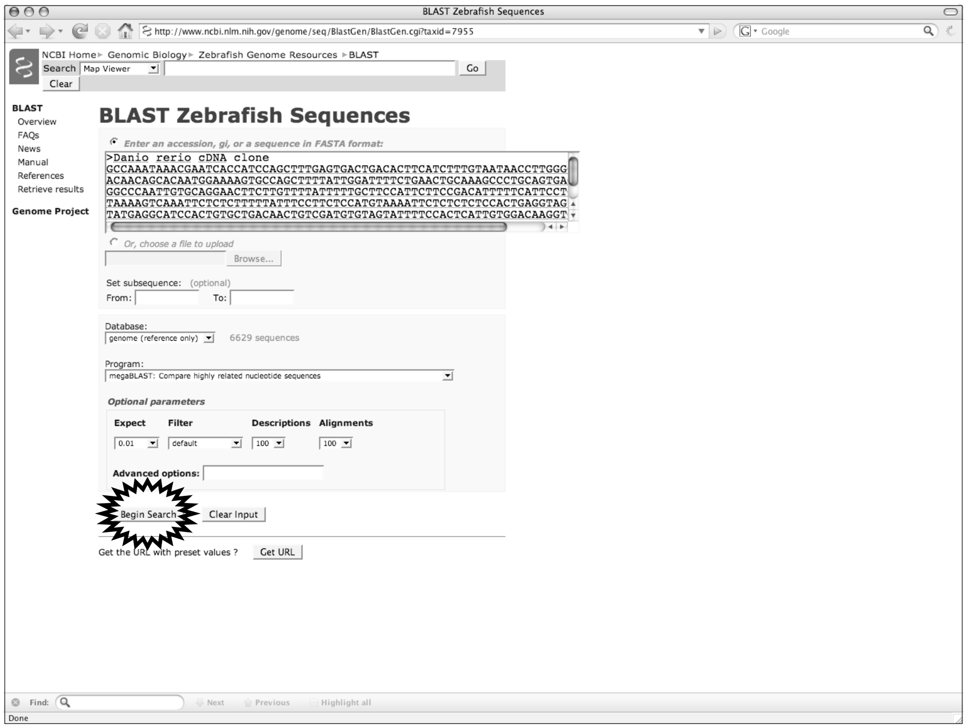
Tip of the Day

Integrating web PSI-BLAST with command line PSI-BLAST using the PssmWithParameters format

This format of the PSSM can be directly used with other stand-alone Blast software tools, in particular as an input checkpoint file for blastpgp. The actual matrix elements can be observed in the "scores" field in the PssmWithParameters structure, which is a one-dimensional representation of the matrix.

More tips...

Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback on new interface
NCBI | NLM | NIH | DHHS
Done



http://www.ncbi.nlm.nih.gov/genome/seq/BlastGen/BlastGen.cgi?taxid=7955

BLAST Zebrafish Sequences

● Enter an accession, gi, or a sequence in FASTA format:

```
>Danio rerio cDNA clone
GCCAAATAAACGAATCACCATCCAGCTTTGAGTGACTGACACTTCATCTTTGTAATAACCTTGGG
ACAACAGCACAATGGAAAAGTGCCAGCTTTTATTGGATTTTCTGAACTGCAAAGCCCTGCAGTGA
GGCCCAATTGTGCAGGAACTTCTTGTTTTATTTTTGCTTCCATTCTTCCGACATTTTTCATTCCT
TAAAAGTCAAATTCTCTCTTTTTATTTCCTTCTCCATGTAAAATTCTCTCTCTCCACTGAGGTAG
TATGAGGCATCCACTGTGCTGACAACTGTCGATGTGTAGTATTTTCCACTCATTGTGGACAAGGT
```

○ Or, choose a file to upload

Set subsequence: (optional)
From: To:

Database: genome (reference only) 6629 sequences

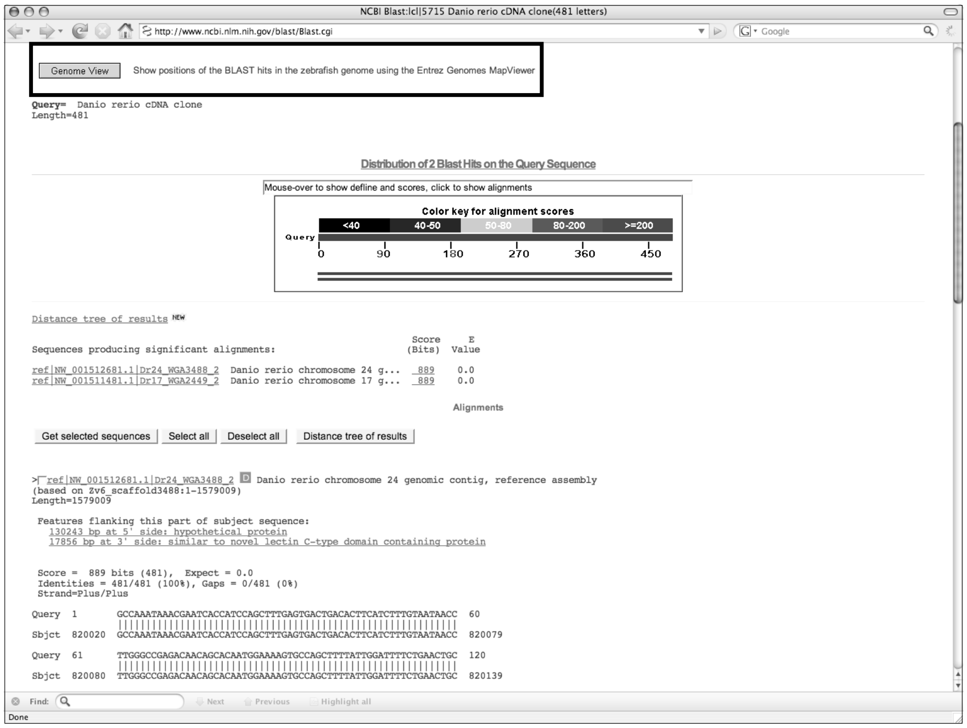
Program: megaBLAST: Compare highly related nucleotide sequences

Optional parameters

Expect	Filter	Descriptions	Alignments
0.01	default	100	100

Advanced options:

Get the URL with preset values ?



http://www.ncbi.nlm.nih.gov/blast/Blast.cgi

Show positions of the BLAST hits in the zebrafish genome using the Entrez Genomes MapViewer

```
Query=  Danio rerio cDNA clone
Length=481
```

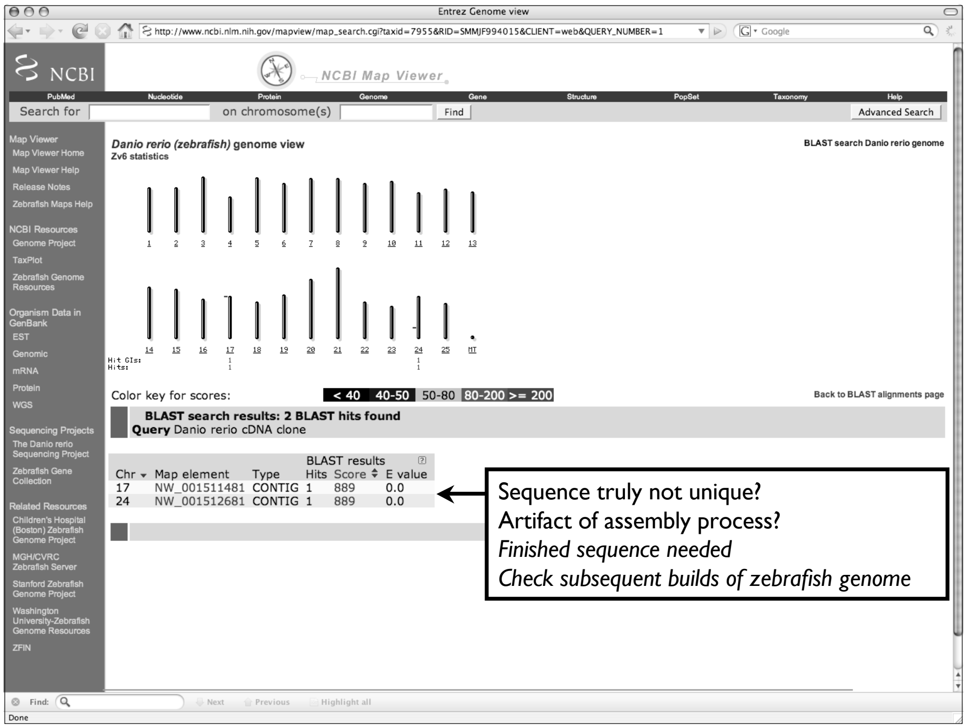
Distribution of 2 Blast Hits on the Query Sequence

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

```
Distance tree of results NEW

                                                                 Score     E
Sequences producing significant alignments:                      (Bits)  Value

ref|NW_001512681.1|Dr24_WGA3488_2  Danio rerio chromosome 24 g...   889    0.0
ref|NW_001511481.1|Dr17_WGA2449_2  Danio rerio chromosome 17 g...   889    0.0
```

Alignments

```
>ref|NW_001512681.1|Dr24_WGA3488_2 D Danio rerio chromosome 24 genomic contig, reference assembly
(based on Zv6_scaffold3488:1-1579009)
Length=1579009

 Features flanking this part of subject sequence:
   130243 bp at 5' side: hypothetical protein
   17856 bp at 3' side: similar to novel lectin C-type domain containing protein


 Score =  889 bits (481),  Expect = 0.0
 Identities = 481/481 (100%), Gaps = 0/481 (0%)
 Strand=Plus/Plus

Query  1       GCCAAATAAACGAATCACCATCCAGCTTTGAGTGACTGACACTTCATCTTTGTAATAACC  60
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  820020  GCCAAATAAACGAATCACCATCCAGCTTTGAGTGACTGACACTTCATCTTTGTAATAACC  820079

Query  61      TTGGGCCGAGACAACAGCACAATGGAAAAGTGCCAGCTTTTATTGGATTTTCTGAACTGC  120
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  820080  TTGGGCCGAGACAACAGCACAATGGAAAAGTGCCAGCTTTTATTGGATTTTCTGAACTGC  820139
```



Danio rerio (zebrafish) genome view
Zv6 statistics

BLAST search results: 2 BLAST hits found
Query Danio rerio cDNA clone

			BLAST results		
Chr	Map element	Type	Hits	Score	E value
17	NW_001511481	CONTIG	1	889	0.0
24	NW_001512681	CONTIG	1	889	0.0

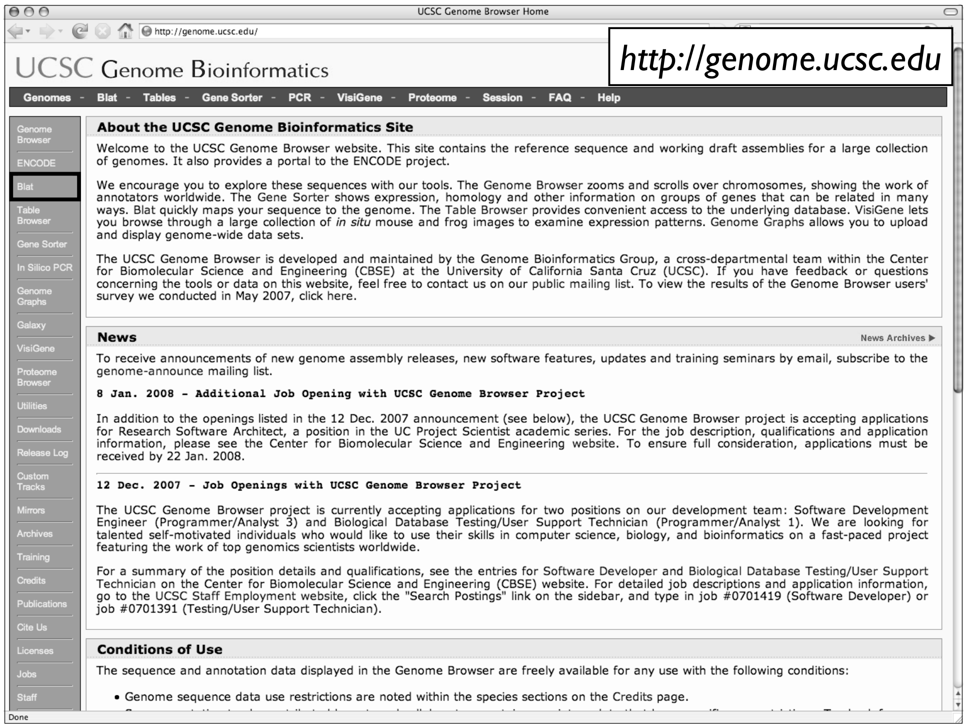
Sequence truly not unique?
Artifact of assembly process?
Finished sequence needed
Check subsequent builds of zebrafish genome

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 for mRNA/DNA searches
- May miss divergent or shorter sequence alignments
- Can be used on protein sequences

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 sequences
  - Identify gene family members
  - Identify putative homologs
- To display a specific sequence as a separate track



http://genome.ucsc.edu

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 a portal to the ENCODE project.

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 Center for Biomolecular Science and Engineering (CBSE) 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. To view the results of the Genome Browser users' survey we conducted in May 2007, click here.

News

To receive announcements of new genome assembly releases, new software features, updates and training seminars by email, subscribe to the genome-announce mailing list.

8 Jan. 2008 - Additional Job Opening with UCSC Genome Browser Project

In addition to the openings listed in the 12 Dec. 2007 announcement (see below), the UCSC Genome Browser project is accepting applications for Research Software Architect, a position in the UC Project Scientist academic series. For the job description, qualifications and application information, please see the Center for Biomolecular Science and Engineering website. To ensure full consideration, applications must be received by 22 Jan. 2008.

12 Dec. 2007 - Job Openings with UCSC Genome Browser Project

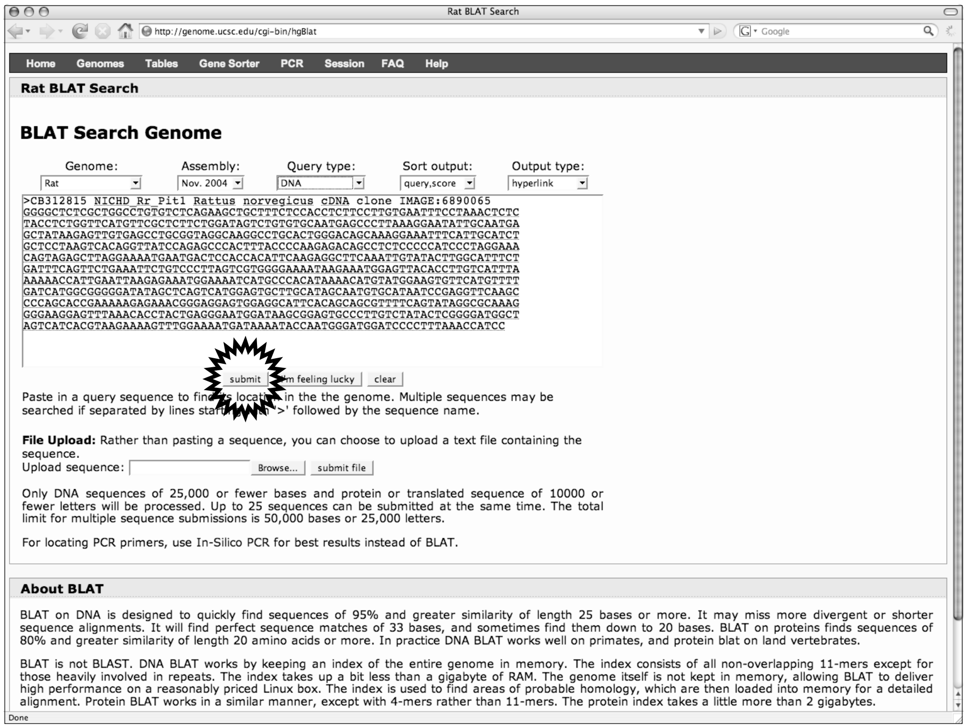
The UCSC Genome Browser project is currently accepting applications for two positions on our development team: Software Development Engineer (Programmer/Analyst 3) and Biological Database Testing/User Support Technician (Programmer/Analyst 1). We are looking for talented self-motivated individuals who would like to use their skills in computer science, biology, and bioinformatics on a fast-paced project featuring the work of top genomics scientists worldwide.

For a summary of the position details and qualifications, see the entries for Software Developer and Biological Database Testing/User Support Technician on the Center for Biomolecular Science and Engineering (CBSE) website. For detailed job descriptions and application information, go to the UCSC Staff Employment website, click the "Search Postings" link on the sidebar, and type in job #0701419 (Software Developer) or job #0701391 (Testing/User Support Technician).

Conditions of Use

The sequence and annotation data displayed in the Genome Browser are freely available for any use with the following conditions:

- Genome sequence data use restrictions are noted within the species sections on the Credits page.



Rat BLAT Search

BLAT Search Genome

Genome: Rat
Assembly: Nov. 2004
Query type: DNA
Sort output: query,score
Output type: hyperlink

```
>CB312815 NICHD_Rr_Pit1 Rattus norvegicus cDNA clone IMAGE:6890065
GGGGCTCTCGCTGGCCTGTGTCTCAGAAGCTGCTTTCTCCACCTCTTCCTTGTGAATTTCCTAAACTCTC
TACCTCTGGTTCATGTTCGCTCTTCTGGATAGTCTGTGTGCAATGAGCCCTTAAAGGAATATTGCAATGA
GCTATAAGAGTTGTGAGCCTGCGGTAGGCAAGGCCTGCACTGGGACAGCAAAGGAAATTTCATTGCATCT
GCTCCTAAGTCACAGGTTATCCAGAGCCCACTTTACCCCAAGAGACAGCCTCTCCCCCATCCCTAGGAAA
CAGTAGAGCTTAGGAAAATGAATGACTCCACCACATTCAAGAGGCTTCAAATTGTATACTTGGCATTTCT
GATTTCAGTTCTGAAATTCTGTCCCTTAGTCGTGGGGAAAATAAGAAATGGAGTTACACCTTGTCATTTA
AAAAACCATTGAATTAAGAGAAATGGAAAATCATGCCCACATAAAACATGTATGGAAGTGTTCATGTTTT
GATCATGGCGGGGGATATAGCTCAGTCATGGAGTGCTTGCATAGCAATGTGCATAATCCGAGGTTCAAGC
CCCAGCACCGAAAAAGAGAAACGGGAGGAGTGGAGGCATTCACAGCAGCGTTTTCAGTATAGGCGCAAAG
GGGAAGGAGTTTAAACACCTACTGAGGGAATGGATAAGCGGAGTGCCCTTGTCTATACTCGGGGATGGCT
AGTCATCACGTAAGAAAAGTTTGGAAAATGATAAAATACCAATGGGATGGATCCCCTTTAAACCATCC
```

Paste in a query sequence to find its location in the the genome. Multiple sequences may be searched if separated by lines starting with '>' followed by the sequence name.

File Upload: Rather than pasting a sequence, you can choose to upload a text file containing the sequence.
Upload sequence:

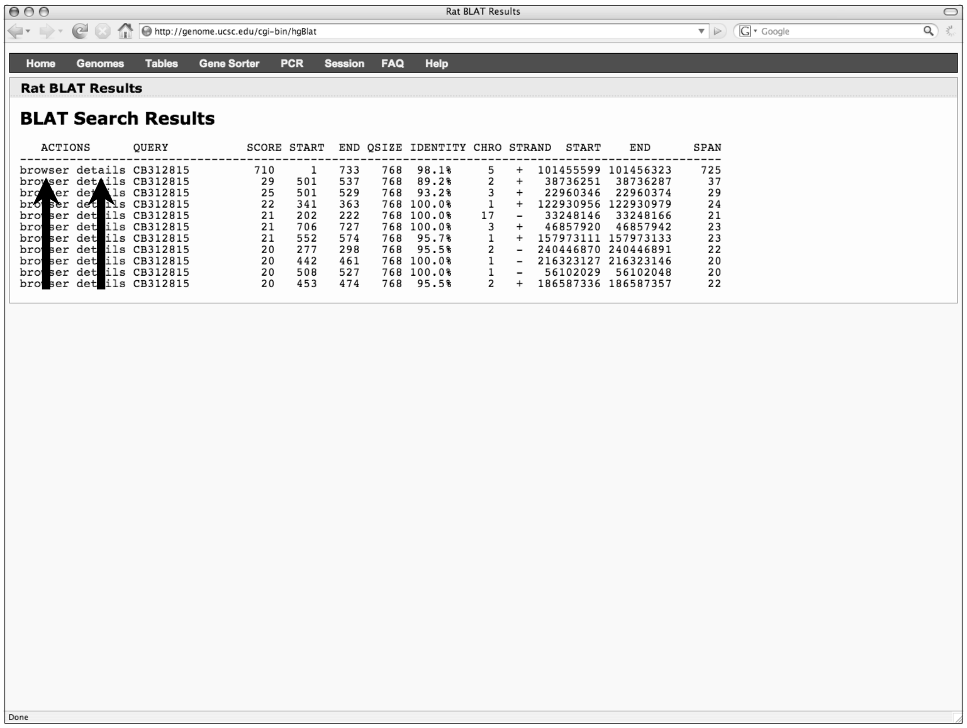
Only DNA sequences of 25,000 or fewer bases and protein or translated sequence of 10000 or fewer letters will be processed. Up to 25 sequences can be submitted at the same time. The total limit for multiple sequence submissions is 50,000 bases or 25,000 letters.

For locating PCR primers, use In-Silico PCR for best results instead of BLAT.

About BLAT

BLAT on DNA is designed to quickly find sequences of 95% and greater similarity of length 25 bases or more. It may miss more divergent or shorter sequence alignments. It will find perfect sequence matches of 33 bases, and sometimes find them down to 20 bases. BLAT on proteins finds sequences of 80% and greater similarity of length 20 amino acids or more. In practice DNA BLAT works well on primates, and protein blat on land vertebrates.

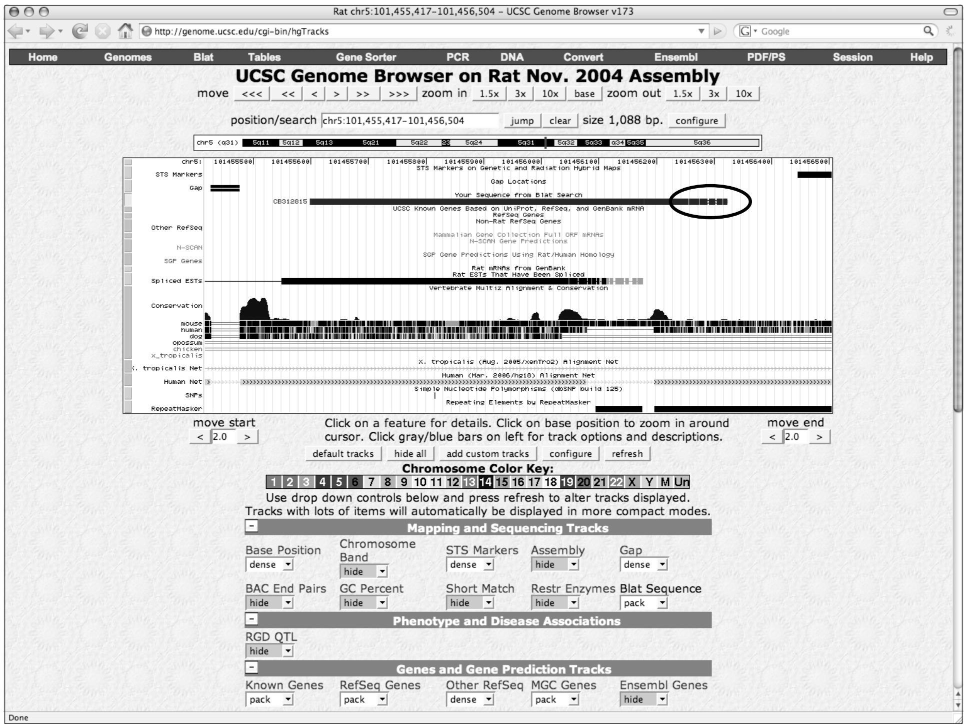
BLAT is not BLAST. DNA BLAT works by keeping an index of the entire genome in memory. The index consists of all non-overlapping 11-mers except for those heavily involved in repeats. The index takes up a bit less than a gigabyte of RAM. The genome itself is not kept in memory, allowing BLAT to deliver high performance on a reasonably priced Linux box. The index is used to find areas of probable homology, which are then loaded into memory for a detailed alignment. Protein BLAT works in a similar manner, except with 4-mers rather than 11-mers. The protein index takes a little more than 2 gigabytes.



Rat BLAT Results

BLAT Search Results

```
   ACTIONS      QUERY           SCORE START  END QSIZE IDENTITY CHRO STRAND  START    END      SPAN
---------------------------------------------------------------------------------------------------
browser details CB312815          710     1   733   768  98.1%     5   +   101455599 101456323   725
browser details CB312815           29   501   537   768  89.2%     2   +    38736251  38736287    37
browser details CB312815           25   501   529   768  93.2%     3   +    22960346  22960374    29
browser details CB312815           22   341   363   768 100.0%     1   +   122930956 122930979    24
browser details CB312815           21   202   222   768 100.0%    17   -    33248146  33248166    21
browser details CB312815           21   706   727   768 100.0%     3   +    46857920  46857942    23
browser details CB312815           21   552   574   768  95.7%     1   +   157973111 157973133    23
browser details CB312815           20   277   298   768  95.5%     2   -   240446870 240446891    22
browser details CB312815           20   442   461   768 100.0%     1   -   216323127 216323146    20
browser details CB312815           20   508   527   768 100.0%     1   -    56102029  56102048    20
browser details CB312815           20   453   474   768  95.5%     2   +   186587336 186587357    22
```


Rat chr5:101,455,417–101,456,504 – UCSC Genome Browser v173
http://genome.ucsc.edu/cgi-bin/hgTracks
Google
Home
Genomes
Blat
Tables
Gene Sorter
PCR
DNA
Convert
Ensembl
PDF/PS
Session
Help
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
jump
clear
configure
chr5 (q31) 5q11 5q12 5q13 5q21 5q22 23 5q24 5q31 5q32 5q33 q34 5q35 5q36
chr5: 101455500 101455600 101455700 101455800 101455900 101456000 101456100 101456200 101456300 101456400 101456500
STS Markers
STS Markers on Genetic and Radiation Hybrid Maps
Gap
Gap Locations
Your Sequence from Blat Search
CB312815
UCSC Known Genes Based on UniProt, RefSeq, and GenBank mRNA
RefSeq Genes
Non-Rat RefSeq Genes
Other RefSeq
Mammalian Gene Collection Full ORF mRNAs
N-SCAN Gene Predictions
N-SCAN
SGP Genes
SGP Gene Predictions Using Rat/Human Homology
Rat mRNAs from GenBank
Rat ESTs That Have Been Spliced
Spliced ESTs
Vertebrate Multiz Alignment & Conservation
Conservation
mouse
human
dog
opossum
chicken
X_tropicalis
X. tropicalis (Aug. 2005/xenTro2) Alignment Net
X. tropicalis Net
Human (Mar. 2006/hg18) Alignment Net
Human Net
Simple Nucleotide Polymorphisms (dbSNP build 125)
SNPs
Repeating Elements by RepeatMasker
RepeatMasker
move start
< 2.0 >
move end
< 2.0 >
default tracks
hide all
add custom tracks
configure
refresh
Chromosome Color Key:
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y M Un
Done

UCSC Genome Browser on Rat Nov. 2004 Assembly

position/search chr5:101,455,417-101,456,504 size 1,088 bp.

Click on a feature for details. Click on base position to zoom in around cursor. Click gray/blue bars on left for track options and descriptions.

Use drop down controls below and press refresh to alter tracks displayed.
Tracks with lots of items will automatically be displayed in more compact modes.

Mapping and Sequencing Tracks

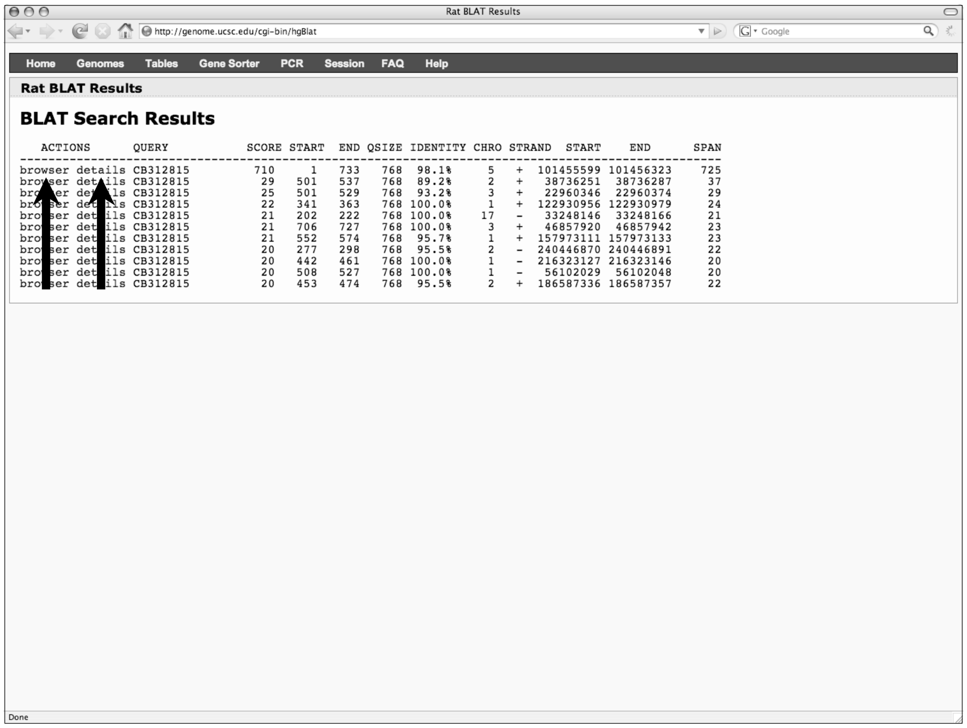
Base Position	Chromosome Band	STS Markers	Assembly	Gap
dense	hide	dense	hide	dense
BAC End Pairs	GC Percent	Short Match	Restr Enzymes	Blat Sequence
hide	hide	hide	hide	pack

Phenotype and Disease Associations

RGD QTL: hide

Genes and Gene Prediction Tracks

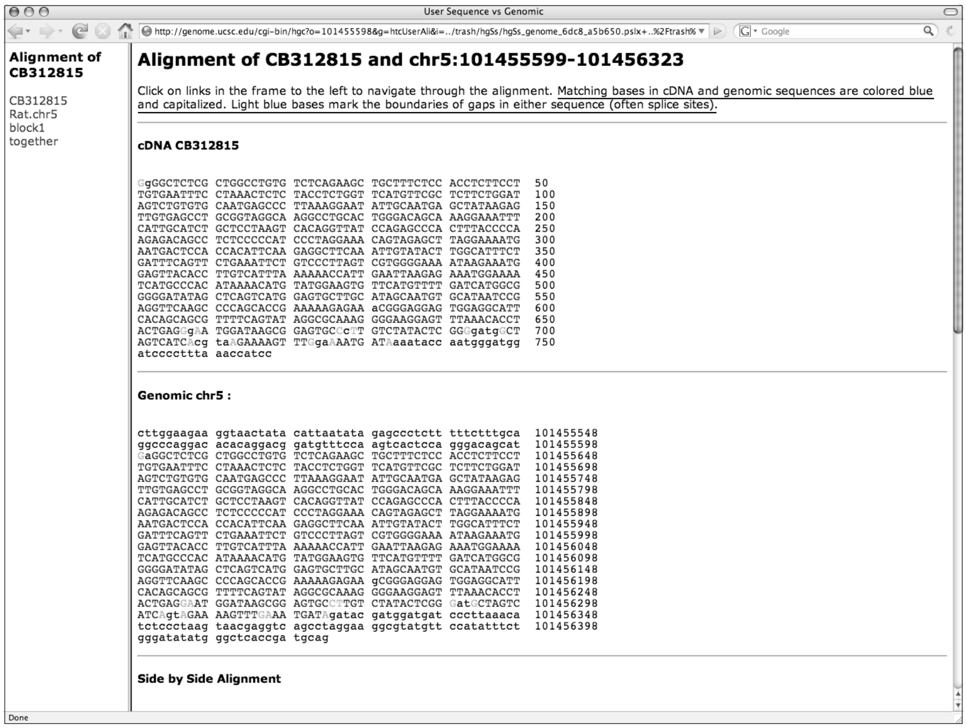
Known Genes	RefSeq Genes	Other RefSeq	MGC Genes	Ensembl Genes
pack	pack	dense	pack	hide



Rat BLAT Results

BLAT Search Results

```
   ACTIONS      QUERY           SCORE START  END QSIZE IDENTITY CHRO STRAND  START      END      SPAN
---------------------------------------------------------------------------------------------------
browser details CB312815          710     1   733   768  98.1%     5   +   101455599 101456323   725
browser details CB312815           29   501   537   768  89.2%     2   +    38736251  38736287    37
browser details CB312815           25   501   529   768  93.2%     3   +    22960346  22960374    29
browser details CB312815           22   341   363   768 100.0%     1   +   122930956 122930979    24
browser details CB312815           21   202   222   768 100.0%    17   -    33248146  33248166    21
browser details CB312815           21   706   727   768 100.0%     3   +    46857920  46857942    23
browser details CB312815           21   552   574   768  95.7%     1   +   157973111 157973133    23
browser details CB312815           20   277   298   768  95.5%     2   -   240446870 240446891    22
browser details CB312815           20   442   461   768 100.0%     1   -   216323127 216323146    20
browser details CB312815           20   508   527   768 100.0%     1   -    56102029  56102048    20
browser details CB312815           20   453   474   768  95.5%     2   +   186587336 186587357    22
```


User Sequence vs Genomic
http://genome.ucsc.edu/cgi-bin/hgc?o=101455598&g=htcUserAli&i=../trash/hgSs/hgSs_genome_6dc8_a5b650.pslx+..%2Ftrash%
Google
Done

Alignment of CB312815

CB312815
Rat.chr5
block1
together

Alignment of CB312815 and chr5:101455599-101456323

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).

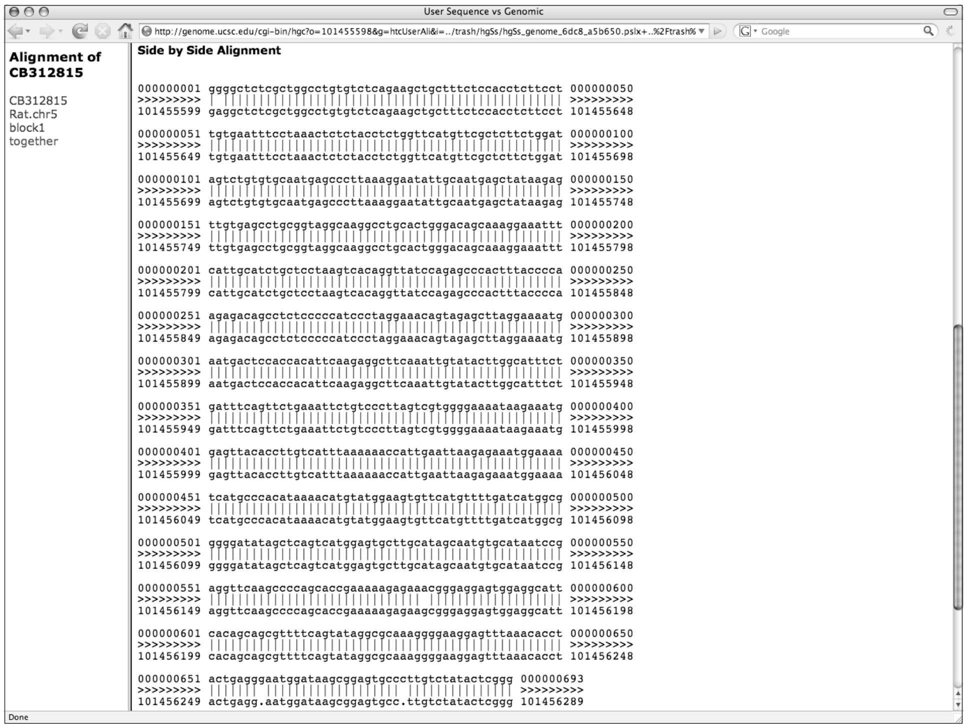
cDNA CB312815

```
GgGGCTCTCG CTGGCCTGTG TCTCAGAAGC TGCTTTCTCC ACCTCTTCCT  50
TGTGAATTTC CTAAACTCTC TACCTCTGGT TCATGTTCGC TCTTCTGGAT  100
AGTCTGTGTG CAATGAGCCC TTAAAGGAAT ATTGCAATGA GCTATAAGAG  150
TTGTGAGCCT GCGGTAGGCA AGGCCTGCAC TGGGACAGCA AAGGAAATTT  200
CATTGCATCT GCTCCTAAGT CACAGGTTAT CCAGAGCCCA CTTTACCCCA  250
AGAGACAGCC TCTCCCCCAT CCCTAGGAAA CAGTAGAGCT TAGGAAAATG  300
AATGACTCCA CCACATTCAA GAGGCTTCAA ATTGTATACT TGGCATTTCT  350
GATTTCAGTT CTGAAATTCT GTCCCTTAGT CGTGGGGAAA ATAAGAAATG  400
GAGTTACACC TTGTCATTTA AAAAACCATT GAATTAAGAG AAATGGAAAA  450
TCATGCCCAC ATAAAACATG TATGGAAGTG TTCATGTTTT GATCATGGCG  500
GGGGATATAG CTCAGTCATG GAGTGCTTGC ATAGCAATGT GCATAATCCG  550
AGGTTCAAGC CCCAGCACCG AAAAAGAGAA aCGGGAGGAG TGGAGGCATT  600
CACAGCAGCG TTTTCAGTAT AGGCGCAAAG GGGAAGGAGT TTAAACACCT  650
ACTGAGGgAA TGGATAAGCG GAGTGCCcTT GTCTATACTC GGGgatgGCT  700
AGTCATCAcg taAGAAAAGT TTGgaAAATG ATAaaatacc aatgggatgg  750
atcccctta aaccatcc
```

Genomic chr5 :

```
cttggaagaa ggtaactata cattaatata gagccctctt tttctttgca  101455548
ggcccaggac acacaggacg gatgtttcca agtcactcca gggacagcat  101455598
GaGGCTCTCG CTGGCCTGTG TCTCAGAAGC TGCTTTCTCC ACCTCTTCCT  101455648
TGTGAATTTC CTAAACTCTC TACCTCTGGT TCATGTTCGC TCTTCTGGAT  101455698
AGTCTGTGTG CAATGAGCCC TTAAAGGAAT ATTGCAATGA GCTATAAGAG  101455748
TTGTGAGCCT GCGGTAGGCA AGGCCTGCAC TGGGACAGCA AAGGAAATTT  101455798
CATTGCATCT GCTCCTAAGT CACAGGTTAT CCAGAGCCCA CTTTACCCCA  101455848
AGAGACAGCC TCTCCCCCAT CCCTAGGAAA CAGTAGAGCT TAGGAAAATG  101455898
AATGACTCCA CCACATTCAA GAGGCTTCAA ATTGTATACT TGGCATTTCT  101455948
GATTTCAGTT CTGAAATTCT GTCCCTTAGT CGTGGGGAAA ATAAGAAATG  101455998
GAGTTACACC TTGTCATTTA AAAAACCATT GAATTAAGAG AAATGGAAAA  101456048
TCATGCCCAC ATAAAACATG TATGGAAGTG TTCATGTTTT GATCATGGCG  101456098
GGGGATATAG CTCAGTCATG GAGTGCTTGC ATAGCAATGT GCATAATCCG  101456148
AGGTTCAAGC CCCAGCACCG AAAAAGAGAA gCGGGAGGAG TGGAGGCATT  101456198
CACAGCAGCG TTTTCAGTAT AGGCGCAAAG GGGAAGGAGT TTAAACACCT  101456248
ACTGAGGAAT GGATAAGCGG AGTGCCTTGT CTATACTCGG GatGCTAGTC  101456298
ATCAgtAGAA AAGTTTGAAA TGATAgatac gatggatgat cccttaaaca  101456348
tctccctaag taacgaggtc agcctaggaa ggcgtatgtt ccatatttct  101456398
gggatatatg ggctcaccga tgcag
```

Side by Side Alignment

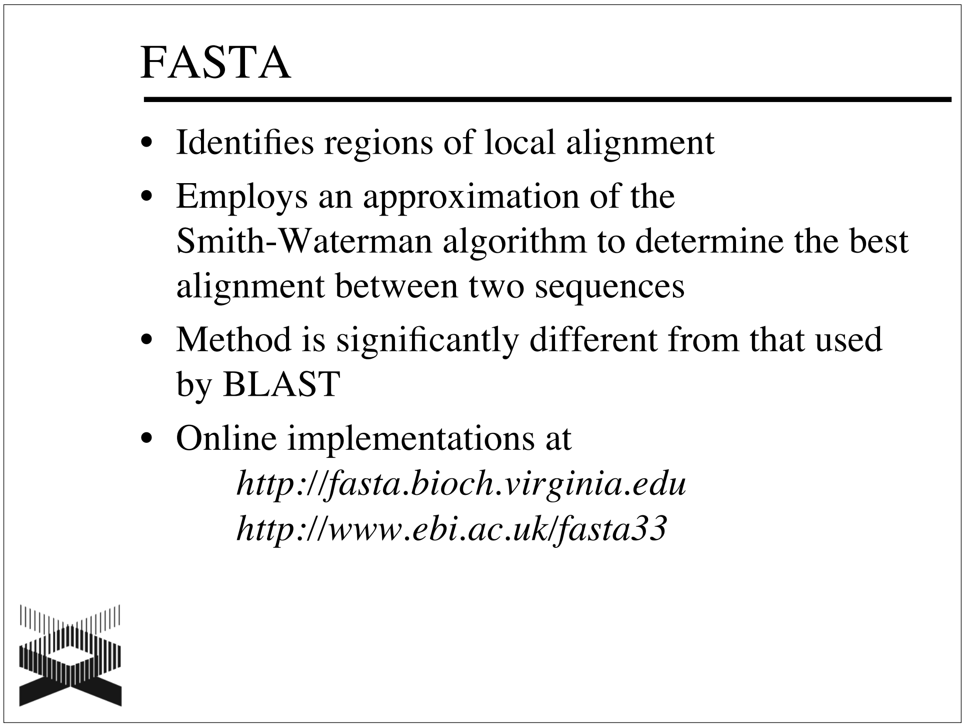

User Sequence vs Genomic
http://genome.ucsc.edu/cgi-bin/hgc?o=101455598&g=htcUserAli&i=../trash/hgSs/hgSs_genome_6dc8_a5b650.pslx+..%2Ftrash%
Google
Done

Alignment of CB312815

CB312815
Rat.chr5
block1
together

Side by Side Alignment

```
000000001 ggggctctcgctggcctgtgtctcagaagctgctttctccacctcttcct 000000050
>>>>>>>>> | |||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>
101455599 gaggctctcgctggcctgtgtctcagaagctgctttctccacctcttcct 101455648

000000051 tgtgaatttcctaaactctctacctctggttcatgttcgctcttctggat 000000100
>>>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>
101455649 tgtgaatttcctaaactctctacctctggttcatgttcgctcttctggat 101455698

000000101 agtctgtgtgcaatgagcccttaaaggaatattgcaatgagctataagag 000000150
>>>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>
101455699 agtctgtgtgcaatgagcccttaaaggaatattgcaatgagctataagag 101455748

000000151 ttgtgagcctgcggtaggcaaggcctgcactgggacagcaaaggaaattt 000000200
>>>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>
101455749 ttgtgagcctgcggtaggcaaggcctgcactgggacagcaaaggaaattt 101455798

000000201 cattgcatctgctcctaagtcacaggttatccagagcccactttacccca 000000250
>>>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>
101455799 cattgcatctgctcctaagtcacaggttatccagagcccactttacccca 101455848

000000251 agagacagcctctcccccatccctaggaaacagtagagcttaggaaaatg 000000300
>>>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>
101455849 agagacagcctctcccccatccctaggaaacagtagagcttaggaaaatg 101455898

000000301 aatgactccaccacattcaagaggcttcaaattgtatacttggcatttct 000000350
>>>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>
101455899 aatgactccaccacattcaagaggcttcaaattgtatacttggcatttct 101455948

000000351 gatttcagttctgaaattctgtcccttagtcgtggggaaaataagaaatg 000000400
>>>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>
101455949 gatttcagttctgaaattctgtcccttagtcgtggggaaaataagaaatg 101455998

000000401 gagttacaccttgtcatttaaaaaaccattgaattaagagaaatggaaaa 000000450
>>>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>
101455999 gagttacaccttgtcatttaaaaaaccattgaattaagagaaatggaaaa 101456048

000000451 tcatgcccacataaaacatgtatggaagtgttcatgttttgatcatggcg 000000500
>>>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>
101456049 tcatgcccacataaaacatgtatggaagtgttcatgttttgatcatggcg 101456098

000000501 ggggatatagctcagtcatggagtgcttgcatagcaatgtgcataatccg 000000550
>>>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>
101456099 ggggatatagctcagtcatggagtgcttgcatagcaatgtgcataatccg 101456148

000000551 aggttcaagccccagcaccgaaaaagagaaacgggaggagtggaggcatt 000000600
>>>>>>>>> |||||||||||||||||||||||||||||| ||||||||||||||||||| >>>>>>>>>
101456149 aggttcaagccccagcaccgaaaaagagaagcgggaggagtggaggcatt 101456198

000000601 cacagcagcgttttcagtataggcgcaaaggggaaggagtttaaacacct 000000650
>>>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>
101456199 cacagcagcgttttcagtataggcgcaaaggggaaggagtttaaacacct 101456248

000000651 actgagggaatggataagcggagtgcccttgtctatactcggg 000000693
>>>>>>>>> ||||||| ||||||||||||||||||| ||||||||||||||| >>>>>>>>>
101456249 actgagg.aatggataagcggagtgcc.ttgtctatactcggg 101456289
```



FASTA

- Identifies regions of local alignment
- Employs an approximation of the Smith-Waterman algorithm to determine the best alignment between two sequences
- Method is significantly different from that used by BLAST
- Online implementations at
  http://fasta.bioch.virginia.edu
  http://www.ebi.ac.uk/fasta33