


NATIONAL HUMAN GENOME RESEARCH INSTITUTE *Division of Intramural Research*

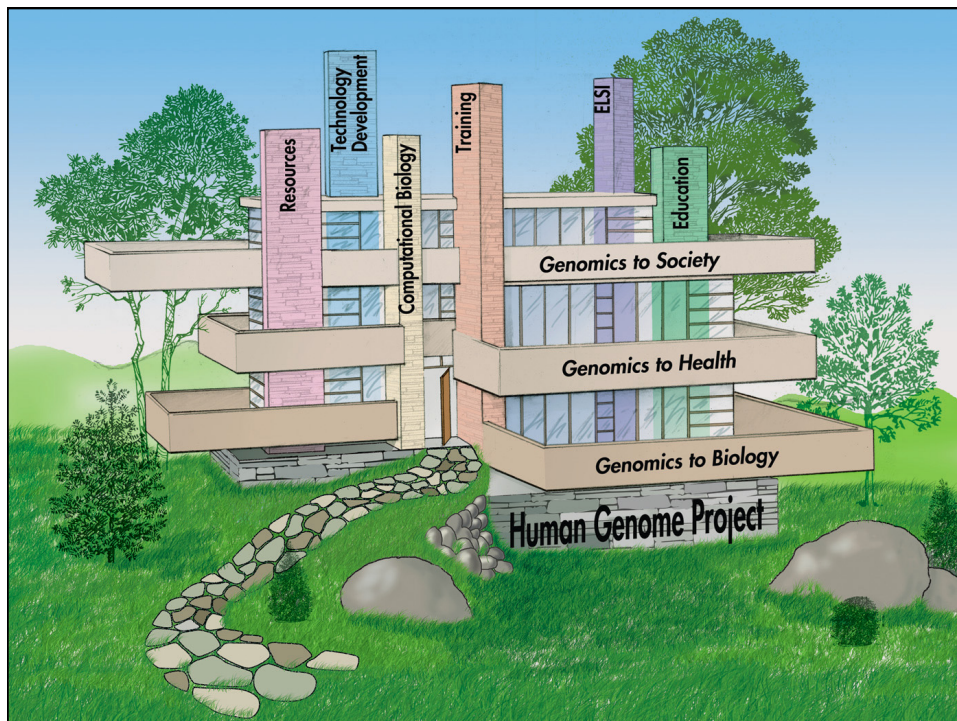



*Current Topics in Genome Analysis
Spring 2010*

Week 2: Biological Sequence Analysis

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, or deletions)
 - Identify residues crucial for maintaining a protein's structure or function
- High degrees of sequence similarity *might* imply
 - a common evolutionary history
 - possible commonality in biological function



Defining the Terms

- The conclusion: **Homology**
 - Genes *are* or *are not* homologous (not measured in degrees)
 - Homology implies an evolutionary relationship

It is worth repeating here that homology, like pregnancy, is indivisible⁸. You either are homologous (pregnant) or you are not. Thus, if what one means to assert is that 80% of the character states are identical one should speak of 80% identity, and not 80% homology.

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



Defining the Terms

- 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

Orthologs *Paralogs*

Most recent common ancestor → α

Gene duplication → 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)

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

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

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

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

- *Simple match/mismatch scoring scheme:*

Match +2

Mismatch -3

	A	T	G	C
A	2	-3	-3	-3
T	-3	2	-3	-3
G	-3	-3	2	-3
C	-3	-3	-3	2

- *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	-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	-3	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-3	-2	-4
Q	-1	0	0	-3	-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	-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	-3	-1	-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	1	1	2	2	2	2	2	2	2	3	1	1	1	1	3	2	1	1	-3	-2	-4	-4
Y	3	3	2	2	2	1	0	0	1	1	1	0	1	2	2	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	-4

BLOSUM62



NATIONAL HUMAN GENOME RESEARCH INSTITUTE
 Division of Intramural Research

BLOSUM Matrices

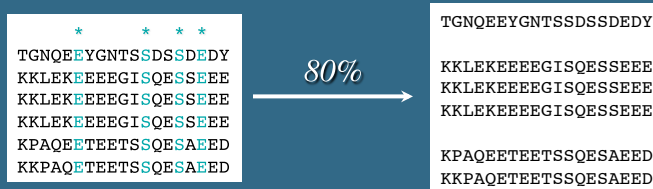
- Henikoff and Henikoff, 1992
- Blocks Substitution Matrix
 - Look only for differences in conserved, ungapped regions of a protein family (“blocks”)
 - Directly calculated, using no extrapolations
 - More sensitive to detecting structural or functional substitutions
 - Generally perform better than PAM matrices for local similarity searches (*Henikoff and Henikoff, 1993*)



NATIONAL HUMAN GENOME RESEARCH INSTITUTE
 Division of Intramural Research

BLOSUM n

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



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



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



So many matrices...

*No single matrix is
the complete answer for
all sequence comparisons*



Further Reading

Unit 3.5 Current Protocols in Bioinformatics

- PAM Matrices
- BLOSUM Matrices
- Specialized Scoring Matrices

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 variability to the search, while statistical significance adds selectivity (see *over 1*). Virtually every user chooses the default, typically PAM250 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, in understanding the assumptions underlying the PAM and BLOSUM scoring matrices and 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 (G to A) would have less of an effect on a protein's structure and function than the substitution of alanine for leucine (A to L) versus substitution of valine (V) for leucine (L). The reference is that if two aligned sequences contain a higher than expected number 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. (1995).

PAM matrices are the result of comparing 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 100 matrix is the result of performing 100 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 *over 1*) regarding insertion and deletions. This second meaning, but remember that over evolutionary time, it is possible that an alanine was changed for 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 superprotein 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 bioinformatics, most of the time the original definition of Dayhoff and colleagues is not being used (see below).

Leaving all potential candidates: 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 deriving identities in the 30% range (i.e., superfamilies), that is, when the two proteins are 70% different from each other (George et al., 1995). Another way to think about this is that the PAM 250

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

UNIT 3.5

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

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"

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Affine Gap Penalty

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

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

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



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



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 threshold S
 - gapped or ungapped
- Results not limited to the “best HSP” for any given sequence pair



BLAST Algorithms

<i>Program</i>	<i>Query Sequence</i>	<i>Target Sequence</i>
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: GSQSLAALLNKCKT **PQG** QRLVNQWIKQPLMDKNRIEERLNLVEAFVED

↓

<i>Neighborhood Words</i>	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.	<i>Neighborhood Score Threshold (T = 13)</i>
---------------------------	--	--

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.	

↓

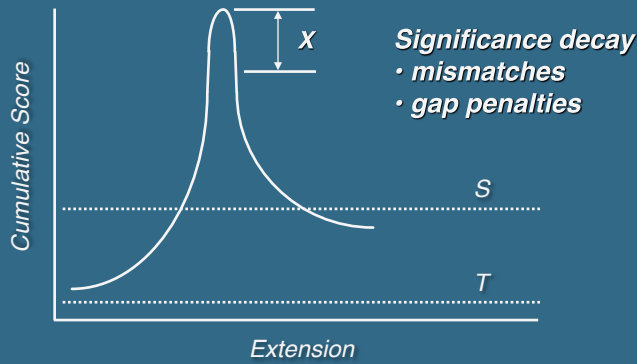
←		→
Query: 325	SLAALLNKCKT PQG QRLVNQWIKQPLMDKNRIEERLNLVEA	365
	+LA++L T R++ +W+ +P+ D + ER + A	
Sbjct: 290	TLASVLDCTVT PMG SRMLKRWLHMPVRDTRVLLERQQTIGA	330

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
 Division of Intramural Research

Extension

← ————— →

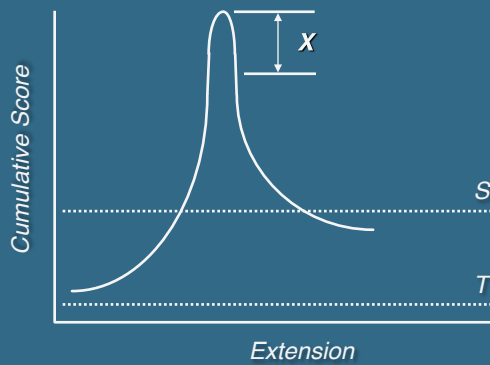
Query:	325	SLAALLNKCKT PQG QRLVNQWIKQPLMDKNRIEERLNLVEA	365
		+LA++L TP+G R++ +W+ +P+ D + ER + A	
Sbjct:	290	TLASVLDCTV TPMG SRMLKRWLHMPVRDTRVLLERQQTIGA	330



Extension

← ————— →

Query:	325	SLAALLNKCKT PQG QRLVNQWIKQPLMDKNRIEERLNLVEA	365
		+LA++L TP+G R++ +W+ +P+ D + ER + A	
Sbjct:	290	TLASVLDCTV TPMG SRMLKRWLHMPVRDTRVLLERQQTIGA	330



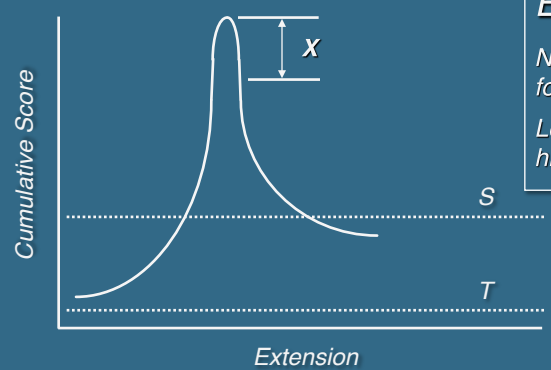
Karlin-Altschul Equation

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

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

Scores and Probabilities

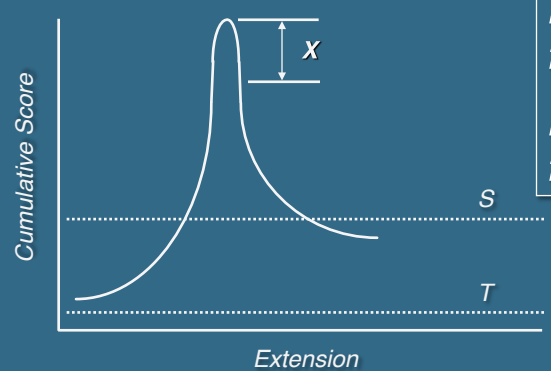
Query:	325	SLAALLNKCKT PQG QRLVNQWIKQPLMDKNRIEERLNLVEA	365
		+LA++L TP+G R++ +W+ +P+ D + ER + A	
Sbjct:	290	TLASVLDCTVTP MG SRMLKRWLHMPVRDTRVLLERQQTIGA	330



$E = kmNe^{-\lambda S}$
 Number of HSPs found purely by chance
 Lower values signify higher similarity

Scores and Probabilities

Query:	325	SLAALLNKCKT PQG QRLVNQWIKQPLMDKNRIEERLNLVEA	365
		+LA++L TP+G R++ +W+ +P+ D + ER + A	
Sbjct:	290	TLASVLDCTVTP MG SRMLKRWLHMPVRDTRVLLERQQTIGA	330



$E \leq 10^{-6}$
 for nucleotides
 $E \leq 10^{-3}$
 for proteins



Protein BLAST: search protein databases using a protein query

BLASTP programs search protein databases using a protein query, *mass...*

Enter Query Sequence

Enter accession number, gi, or FASTA sequence

Query sequence: MSSAAAAAGCGGALFQPSVSTANSSSSNNNSSTPAALATHSPTSNSPVSGASSASSLLT
 AAFGNLFGGSSAKMLNELFGRQMKQADATSGLPQSLDNMLAAAMETATSAELLIGSLNSTS
 KLLQQQHNNNSAPNSTPMSNGTNAQSPSCSAHSSSHHOGYSPKCSRRVYSCDSRSLLEAAA
 DVACCSPPRAASVSSLNGGASCEQHSQLHDLVAHHMLNLQKKELMQLDQELRTAMQ

Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism: Optional

Exclude: Optional

Program Selection

Algorithm: blastp (protein-protein BLAST)

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

Available protein databases include:

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

RefSeq

- *Goal:* Provide a single reference sequence for each molecule of the central dogma (DNA, mRNA, protein)
- Distinguishing Features
 - Non-redundancy
 - Updates to reflect the current knowledge of sequence data and biology
 - Ongoing curation by NCBI staff and collaborators, with review status indicated on each record

RefSeq Accession Format

From curation of GenBank entries:

NT_123456 Genomic contigs
NM_123456 mRNAs
NP_123456 Proteins

From genome annotation:

XM_123456 Model mRNA
XP_123456 Model proteins

Complete key at

<http://www.ncbi.nlm.nih.gov/RefSeq/key.html>



A screenshot of the NCBI BLAST search interface. The browser address bar shows "http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch&SHOW_DEFAULT...". The interface includes a search bar with a query sequence: "MSSAAAAAGCGGALFQPVSTANSSSSNNNSSTPAALATHSPTSNSPVSGASSLLT...". Below the search bar, there are sections for "Choose Search Set" and "Program Selection". In the "Choose Search Set" section, the "Database" is set to "Non-redundant protein sequences (nr)". The "Organism" field is empty, and the "Exclude" section has "Models (XM/XP)" and "Environmental sample sequences" checked. A callout box with an arrow points to the "Organism" field, containing the text "Limit by organism or taxonomic group". The "Program Selection" section has "blastp (protein-protein BLAST)" selected. At the bottom left, a red box highlights the "Algorithm parameters" link.

Protein BLAST: search protein databases using a protein query

Protein BLAST: search protein dat...
Entrez query
Optional Enter an Entrez query to limit search

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

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

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

General Parameters
Max target sequences **+ 250**
Select the maximum number of aligned sequences to display

Short queries Automatically adjust parameters for short input sequences

Expect threshold 10
Word size 3

Scoring Parameters
Matrix BLOSUM62
Gap Costs Existence: 11 Extension: 1
Compositional adjustments Conditional compositional score matrix adjustment

Filters and Masking
Filter **+ Low complexity regions**
Mask Mask for lookup table only
 Mask lower case letters

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

Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback

NCBI | NLM | NIH | CHHS

Protein BLAST: search protein databases using a protein query

Protein BLAST: search protein dat...
Entrez query
Optional Enter an Entrez query to limit search

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

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

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

General Parameters
Max target sequences **+ 250**
Select the maximum number of aligned sequences to display

Short queries Automatically adjust parameters for short input sequences

Expect threshold **10**
Word size 3

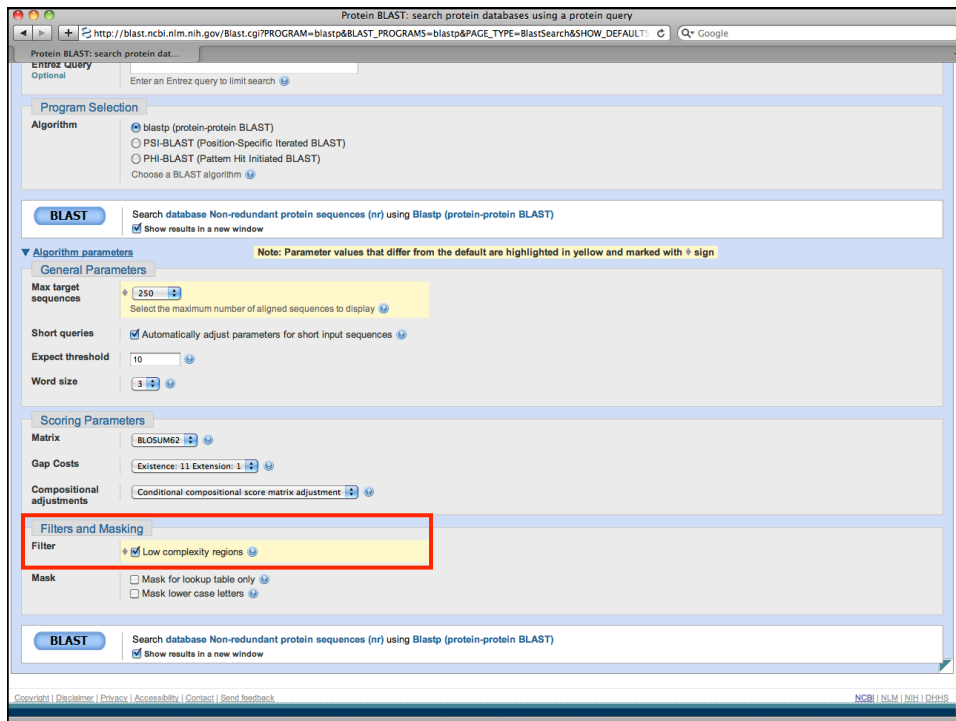
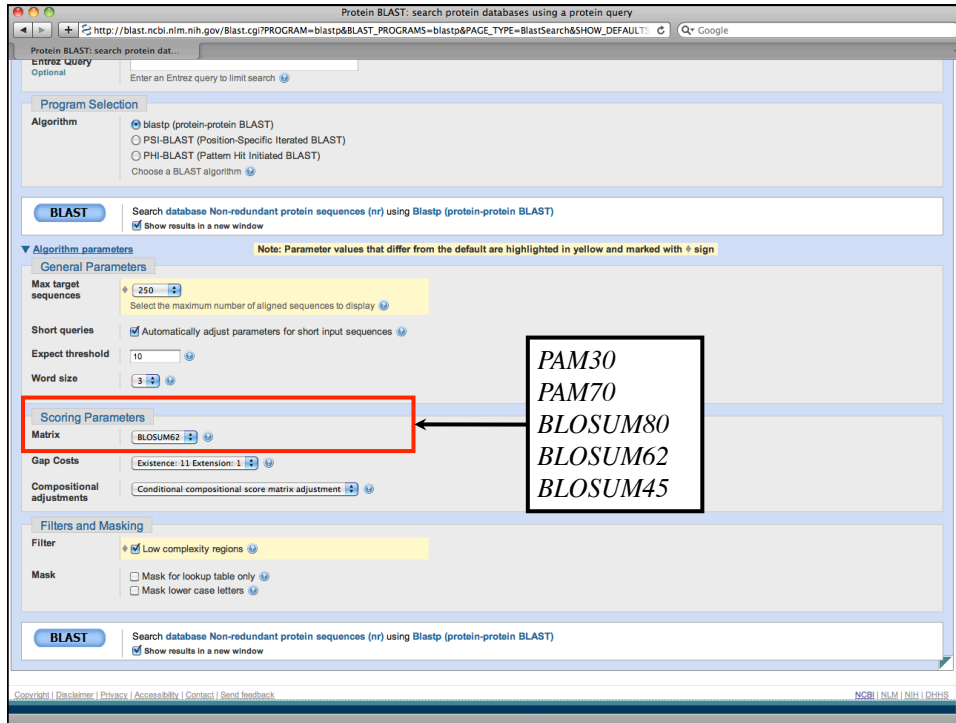
Scoring Parameters
Matrix BLOSUM62
Gap Costs Existence: 11 Extension: 1
Compositional adjustments Conditional compositional score matrix adjustment

Filters and Masking
Filter **+ Low complexity regions**
Mask Mask for lookup table only
 Mask lower case letters

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

Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback

NCBI | NLM | NIH | CHHS



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)  
MESSAKMESGGAGQQPQPQPQQPFLPPAACFFATAAAAAAAAAAAAAQAQQQQQQQQQQQAPQLRPAA  
DQOPSGGGHKSAPKQVKRORSSPELMRCKRRLNFSGFGYSLPQQQIAAVARRNERERNRVRLVNLGFAT  
LREHVPNGAANKKMSKVETLRSAVEYIRALQQLLDEHDAVSAAFQAVLSPTISPNYSNDLNSMAGSPVS  
SYSSDEGSYDPLSPPEEQELLDFTNWF
```

*Homopolymeric
alanine-glutamine tract*



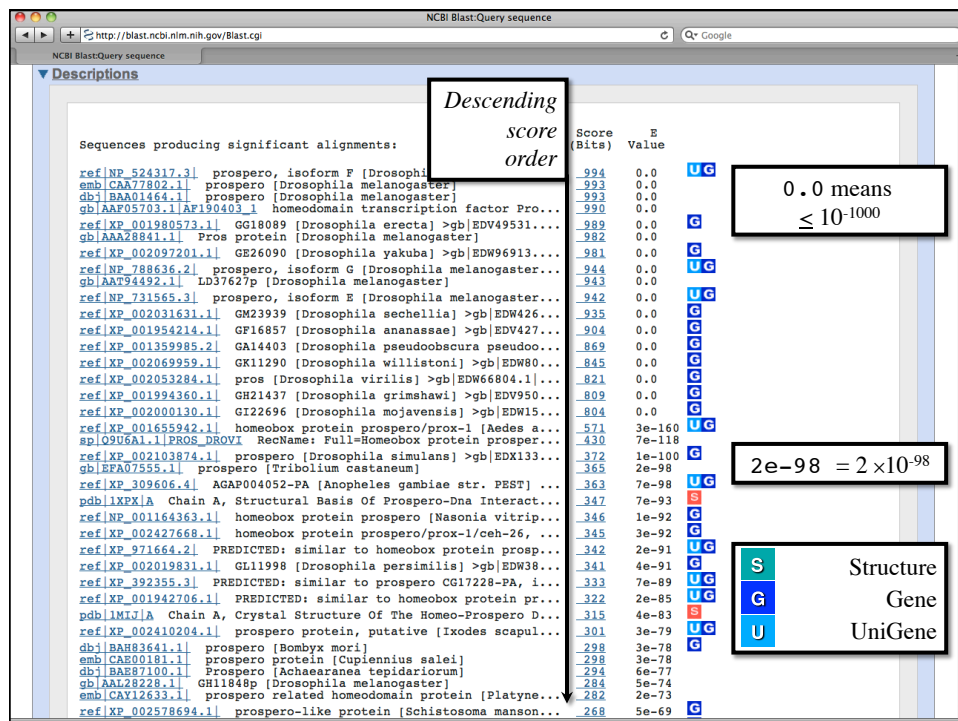
Identifying Low-Complexity Regions

- 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)



The screenshot shows the Protein BLAST search interface. At the top, there is a search bar and a "BLAST" button. Below this, the "Algorithm" section is set to "blastp (protein-protein BLAST)". The "General Parameters" section includes "Max target sequences" set to 250, "Short queries" checked, "Expect threshold" set to 10, and "Word size" set to 3. The "Scoring Parameters" section shows the "Matrix" set to "BLOSUM62". The "Filters and Masking" section has "Filter" set to "Low complexity regions". A red starburst graphic highlights the "BLAST" button at the bottom left. The URL in the browser is http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BLASTSearch&SHOW_DEFAULTS=.

The screenshot shows the NCBI Blast-Query sequence results page. The query sequence is "SuperFam11Loc" with a length of 1403 amino acids. The database is "nr" (All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects). The results are displayed as a "Graphic Summary" showing the distribution of 189 Blast Hits on the query sequence. A color key for alignment scores is provided: <math><40</math> (black), 40-60 (blue), 60-80 (green), 80-200 (magenta), and >=200 (red). The query sequence is shown as a horizontal bar with red boxes indicating conserved domains. The alignment results are shown as a grid of colored bars representing the distribution of hits across the query sequence.



NCBI BlastQuery sequence

ref|XP_001517520.1| PREDICTED: hypothetical protein [Ornithor... 210 1e-51 UG
 ref|XP_522907.2| PREDICTED: hypothetical protein [Pan troglod... 209 1e-51 UG
 ref|XP_001845683.1| homeobox protein prospero/prox-1 [Culex q... 209 2e-51 UG
 ref|XP_001088672.1| PREDICTED: similar to prospero-related ho... 209 2e-51 G
 sp|Q3B8N5.2|PROX2_HUMAN RecName: Full=Prospero homeobox prote... 208 5e-51 G
 ref|NP_001073877.1| prospero homeobox 2 [Homo sapiens] 207 7e-51 UG
 gb|AAI05928.1| PROX2 protein [Homo sapiens] >gb|AAI05721.1| P... 207 7e-51 G
 gb|ACT78708.1| prospero-like protein Prox3 [Danio rerio] 204 5e-50 G
 ref|XP_001845682.1| prospero [Culex quinquefasciatus] >gb|EDS... 204 5e-50 UG
 ref|XP_692862.3| PREDICTED: similar to Homeobox prospero-like... 204 6e-50 UG
 ref|XP_001919536.1| PREDICTED: similar to prox1-like protein ... 204 9e-50 UG
 ref|XP_002199957.1| PREDICTED: similar to prospero homeobox 2... 203 2e-49 UG
 emb|CAF92934.1| unnamed protein product [Tetraodon nigroviridis] 202 4e-49 UG
 ref|NP_001071961.1| transcription factor protein [Clona intes... 199 2e-48 UG
 emb|CAG04605.1| unnamed protein product [Tetraodon nigroviridis] 198 4e-48 UG
 emb|CAG95276.1| unnamed protein product [Tetraodon nigroviridis] 196 2e-47 UG
 emb|CAG10630.1| unnamed protein product [Tetraodon nigroviridis] 195 4e-47 G
 ref|XP_002019832.1| GL11997 [Drosophila persimilis] >gb|EDW38... 189 3e-45 G
 gb|AAC28353.1| Prox1 [Xenopus laevis] 187 8e-45 UG
 emb|CAG09138.1| unnamed protein product [Tetraodon nigroviridis] 175 3e-41 UG
 ref|XP_5471908.2| PREDICTED: similar to RIKEN cDNA 1700058C01 ... 168 4e-39 UG
 ref|XP_002575867.1| homeobox protein prospero/prox-1/ceh-26 [... 167 9e-39 UG
 dbj|BAB17311.1| Prox 1 [Cynops pyrrhogaster] 161 4e-37 G
 gb|EAW81198.1| hCG22353 [Homo sapiens] 158 4e-36 G
 dbj|BAC04278.1| unnamed protein product [Homo sapiens] 157 8e-36 G
 gb|AAC95978.1| prospero-like protein [Takifugu rubripes] 156 1e-35 G
 gb|EDL02840.1| RIKEN cDNA 1700058C01, isoform CRA_a [Mus musc... 154 7e-35 G
 emb|CAI15309.1| prospero homeobox 1 [Homo sapiens] 154 1e-34 UG
 ref|XP_849216.1| PREDICTED: similar to prospero-related homeo... 154 1e-34 UG
 gb|EFP8550.1| hypothetical protein PANDA 009835 [Aluicropoda ... 153 3e-34 G
 emb|CAG09167.1| unnamed protein product [Tetraodon nigroviridis] 150 1e-33 UG
 emb|CAG13403.1| unnamed protein product [Tetraodon nigroviridis] 100 1e-18 G
 gb|AAD30180.1|ACO06530_2 homeobox prospero-like protein [Homo... 97.4 1e-17 UG
 ref|XP_547411.2| PREDICTED: similar to prospero-related homeo... 80.1 2e-12 UG
 pir|JC5496 Prox 1 protein 671 - chicken 80.1 2e-12 UG
 ref|NP_001108671.1| prospero homeobox 1 [Rattus norvegicus] >... 44.7 0.091 UG
 emb|CAF94749.1| unnamed protein product [Tetraodon nigroviridis] 43.5 0.17 G
 emb|CAP58279.1| Prox1 protein [Xenopus tropicalis] 42.0 0.64 G
 gb|AAF13029.1|AF070733_1 transcription factor Prox1 [Notophth... 40.4 1.8 G
 gb|ABG29070.1| transcription factor Prox1 [Pleurodeles waltl] 38.9 5.3 G

▼ Alignments Select All Get selected sequences Distance tree of results Multiple alignment

NCBI BlastQuery sequence

>ref|NP_731565.3| UG prospero, isoform E [Drosophila melanogaster]
 gb|AAI13501.3| G prospero, isoform E [Drosophila melanogaster]
 Length=1835

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

Score = 942 bits (2435), Expect = 0.0, Method: Compositional matrix adjust.
 Identities = 688/688 (100%), Positives = 688/688 (100%), Gaps = 0/688 (0%)

Query 17 LFQPQSVSTANSSSSNNNSSTPAALATHSPTNSPVSGASSASSLLTAAFGNLFQGGSSA 76
 LFQPQSVSTANSSSSNNNSSTPAALATHSPTNSPVSGASSASSLLTAAFGNLFQGGSSA
 Sbjct 317 LFQPQSVSTANSSSSNNNSSTPAALATHSPTNSPVSGASSASSLLTAAFGNLFQGGSSA 376

Query 77 KMLNELFGRQMKQAQDATSLPQLSDNAMLAAAMETATSSELLIGSLNSTSKLLQQQHNN 136
 KMLNELFGRQMKQAQDATSLPQLSDNAMLAAAMETATSSELLIGSLNSTSKLLQQQHNN
 Sbjct 377 KMLNELFGRQMKQAQDATSLPQLSDNAMLAAAMETATSSELLIGSLNSTSKLLQQQHNN 436

Query 137 NSIAPANSTPMSNGTNASIPGSAHSSSHSHGQVSPKGSRRVSACSDRSLEAAADVAGG 196
 NSIAPANSTPMSNGTNASIPGSAHSSSHSHGQVSPKGSRRVSACSDRSLEAAADVAGG
 Sbjct 437 NSIAPANSTPMSNGTNASIPGSAHSSSHSHGQVSPKGSRRVSACSDRSLEAAADVAGG 496

Query 197 SPPRAASVSSLNGGASSGEQHQSLQDHLVAHMLRLNLOKQKELMQLDQELRTAMQQQQ 256
 SPPRAASVSSLNGGASSGEQHQSLQDHLVAHMLRLNLOKQKELMQLDQELRTAMQQQQ
 Sbjct 497 SPPRAASVSSLNGGASSGEQHQSLQDHLVAHMLRLNLOKQKELMQLDQELRTAMQQQQ 556

Query 257 qqlqekeqLHSLKLNNNNNNIAATANNNNNTMESINLIDDSEMA DIKIKSEPTAPQPQ 316
 QQLQEKEQLHSLKLNNNNNNIAATANNNNNTMESINLIDDSEMA DIKIKSEPTAPQPQ
 Sbjct 557 QQLQEKEQLHSLKLNNNNNNIAATANNNNNTMESINLIDDSEMA DIKIKSEPTAPQPQ 616

Query 317 QSPHGSSSRSRSGSGSGSHSSMASDGLRRKSSDLSLDSHAQDDAQDEEDAAPTQORSSES 376
 QSPHGSSSRSRSGSGSGSHSSMASDGLRRKSSDLSLDSHAQDDAQDEEDAAPTQORSSES
 Sbjct 617 QSPHGSSSRSRSGSGSGSHSSMASDGLRRKSSDLSLDSHAQDDAQDEEDAAPTQORSSES 676

Query 377 RAPEEPQLPTKKEVDMLDEVELLGLHSRGSMDLSLSPSHSdmlldkddvldedddd 436
 RAPEEPQLPTKKEVDMLDEVELLGLHSRGSMDLSLSPSHSdmlldkddvldedddd
 Sbjct 677 RAPEEPQLPTKKEVDMLDEVELLGLHSRGSMDLSLSPSHSdmlldkddvldedddd 736

Query 437 dCVEQKTSGSGCLKFKPGMDLKRARVENIVSGMRCSPSSGLAQAGLQVNGCKRRKLYQFP 496
 DCVEQKTSGSGCLKFKPGMDLKRARVENIVSGMRCSPSSGLAQAGLQVNGCKRRKLYQFP
 Sbjct 737 DCVEQKTSGSGCLKFKPGMDLKRARVENIVSGMRCSPSSGLAQAGLQVNGCKRRKLYQFP 796

Query 497 QHAMERYVAAAAGLNFGLNLQSMMLDQEDSESENELESFQIQKRVKKNALKSOLRSMQEQ 556
 QHAMERYVAAAAGLNFGLNLQSMMLDQEDSESENELESFQIQKRVKKNALKSOLRSMQEQ
 Sbjct 797 QHAMERYVAAAAGLNFGLNLQSMMLDQEDSESENELESFQIQKRVKKNALKSOLRSMQEQ 856

≥ 25% for proteins
 ≥ 70% for nucleotides

Gap
 Low-Complexity

NCBI BlastQuery sequence

http://blast.ncbi.nlm.nih.gov/Blast.cgi#221378762

Sbjct 917 NHKEETGQERPGSSSSPSPSLKPKTSLGESSDSGANLMSQMMKMMSGKLNHPVGVGHP 976

Query 677 ALPQGFPPLLQHMGMDSHAAMYYQFFF 704

ALPQGFPPLLQHMGMDSHAAMYYQFFF

Sbjct 977 ALPQGFPPLLQHMGMDSHAAMYYQFFF 1004

Score = 636 bits (1640), Expect = 7e-180, Method: Compositional matrix adjust.
 Identities = 461/498 (92%), Positives = 463/498 (92%), Gaps = 32/498 (6%)

Query 906 PONGPTPATQSAAMFQAPKTPQGMNPVAAAALYNSMTGPFCLPPDgggggtaggggsa 965

Sbjct 1370 P P+P +AAAMFQAPKTPQGMNPVAAAALYNSMTGPFCLPPDQQQQQTAQQQSA 1426

PHIRPSP---TAAAMFQAPKTPQGMNPVAAAALYNSMTGPFCLPPDQQQQQTAQQQSA

Query 966 gggggsgggtqqqLEQNEALSLVVTPKKRRHKVTDTRITPRTVSRILAQDgvpvptggpp 1025

Sbjct 1427 QQQQSSQQTQQLEQNEALSLVVTPKKRRHKVTDTRITPRTVSRILAQDGVVPTGGPP 1486

Query 1026 stpqqggggggggggggggggggASNGGNSNATPAQSPTRSSGGAAYHppppppppmmp 1085

Sbjct 1487 STPQQQQQQQQQQQQQQQQQQQASNGGNSNATPAQSPTRSSGGAAYHPQPPPPPPMMP 1546

Query 1086 VSLPTSVAIPNPSLHESKVFSPYSFPFNPhaaagqataaqlhghqghhphhggmslss 1145

Sbjct 1547 VSLPTSVAIPNPSLHESKVFSPYSFPFNPhaaagqataaqlhghqghhphhggmslss 1606

Query 1146 ppgslgALMDSRDppplphppsmhlpallaaahggspDYKTCRAVMDAQRQSECNESA 1205

Sbjct 1607 PPGSLGALMDSRDSPPPLPHPPSMHLPALLAAAHGGSPDYKTCRAVMDAQRQSECNESA 1666

Query 1206 DMQFDGMAPTISFYKQMLKTEHQESLMAKHCESLTPIHSTLTPMHLRKAklmffwvry 1265

Sbjct 1667 DMQFDGMAPT-----SSTLTPMHLRKAklmffwvry 1697

Query 1266 PSSAVLKMYPFDIKFNKNNTAQLVKWFSNFRFYIOMEKYARQAVTEGIKTPDDLLIAG 1325

Sbjct 1698 PSSAVLKMYPFDIKFNKNNTAQLVKWFSNFRFYIOMEKYARQAVTEGIKTPDDLLIAG 1757

Query 1326 DSELYRVNLHYNRNHHIEVPQNFREVVESTLREFFRAIQGKDEQSWKSIYKIISRM 1385

Sbjct 1758 DSELYRVNLHYNRNHHIEVPQNFREVVESTLREFFRAIQGKDEQSWKSIYKIISRM 1817

Query 1386 DDPVPEYFKSPNFLEQLE 1403

Sbjct 1818 DDPVPEYFKSPNFLEQLE 1835

Score = 942 bits (2435), Expect = 0.0, Method: Compositional matrix adjust.
 Identities = 688/688 (100%), Positives = 688/688 (100%), Gaps = 0/688 (0%)

Score = 636 bits (1640), Expect = 7e-180, Method: Compositional matrix adjust.
 Identities = 461/498 (92%), Positives = 463/498 (92%), Gaps = 32/498 (6%)

HSP 1
 Q: 17- 704
 S: 317-1004

HSP 2
 Q: 906-1403
 S: 1370-1835

Color key for alignment score

Query 0 250 500 750 1000 1250

<40 40-50 50-80 80-100 >=200

Suggested BLAST Cutoffs

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

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



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 Search Tool

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

BLAST finds regions of similarity between biological sequences. [more...](#)

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

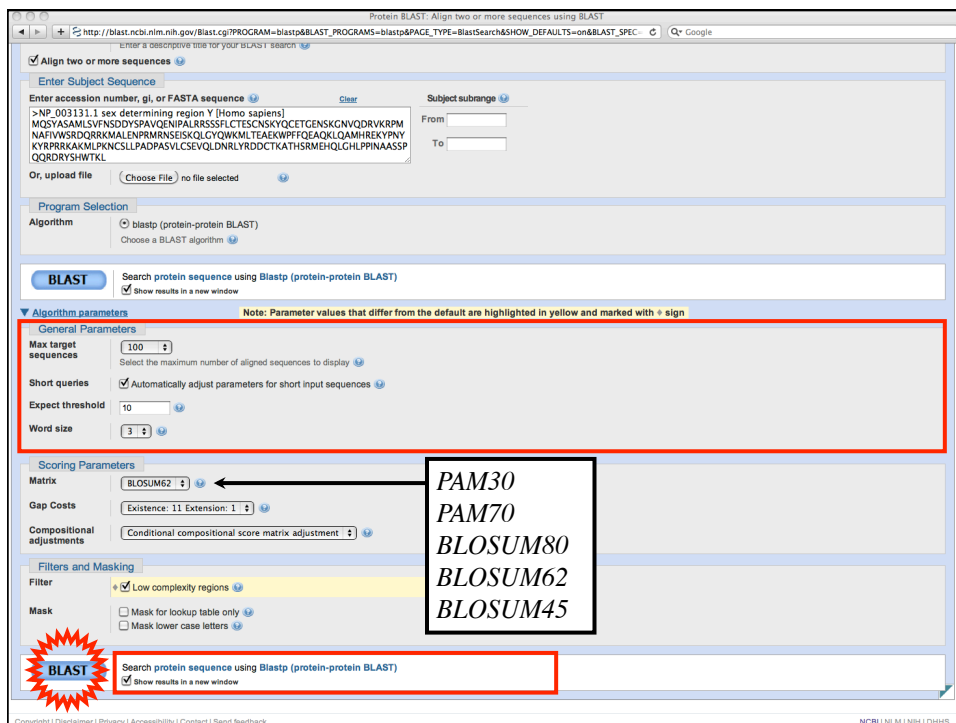
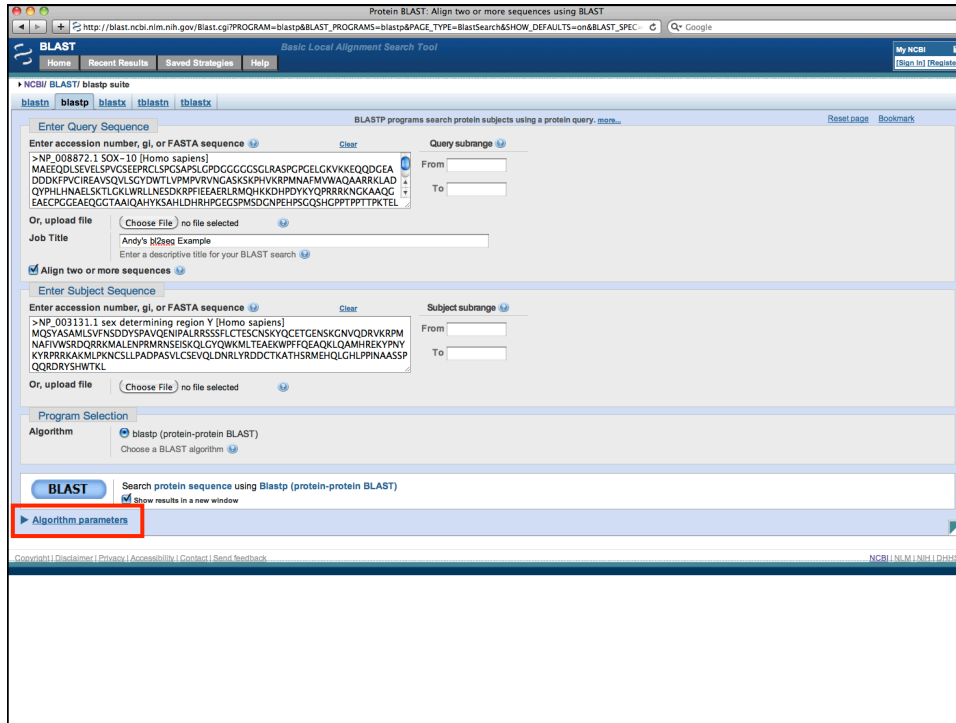
Choose a BLAST program to run.

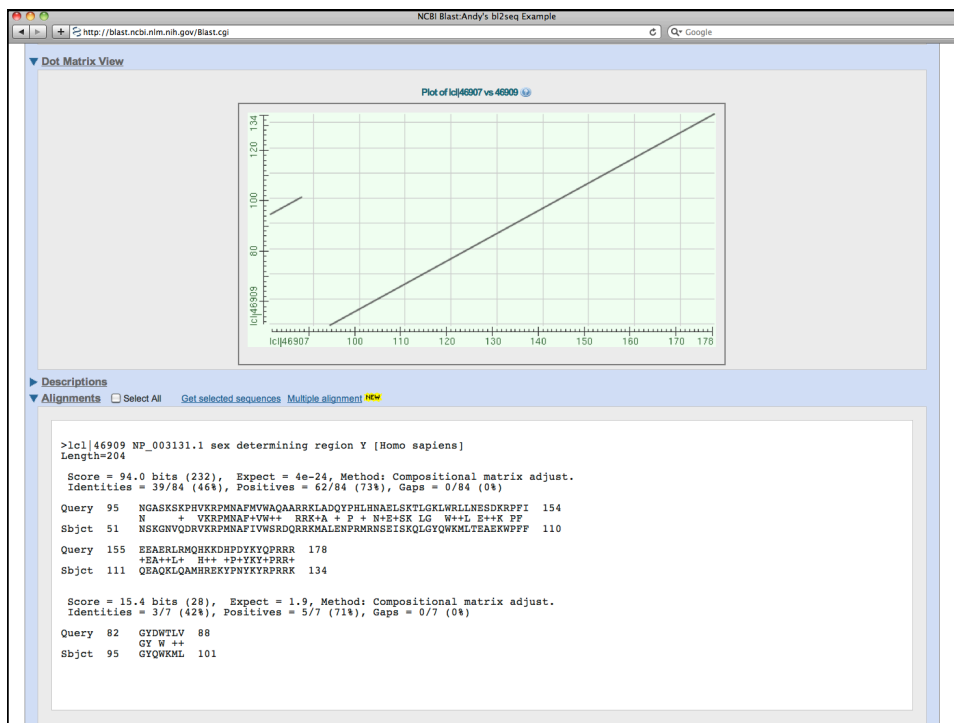
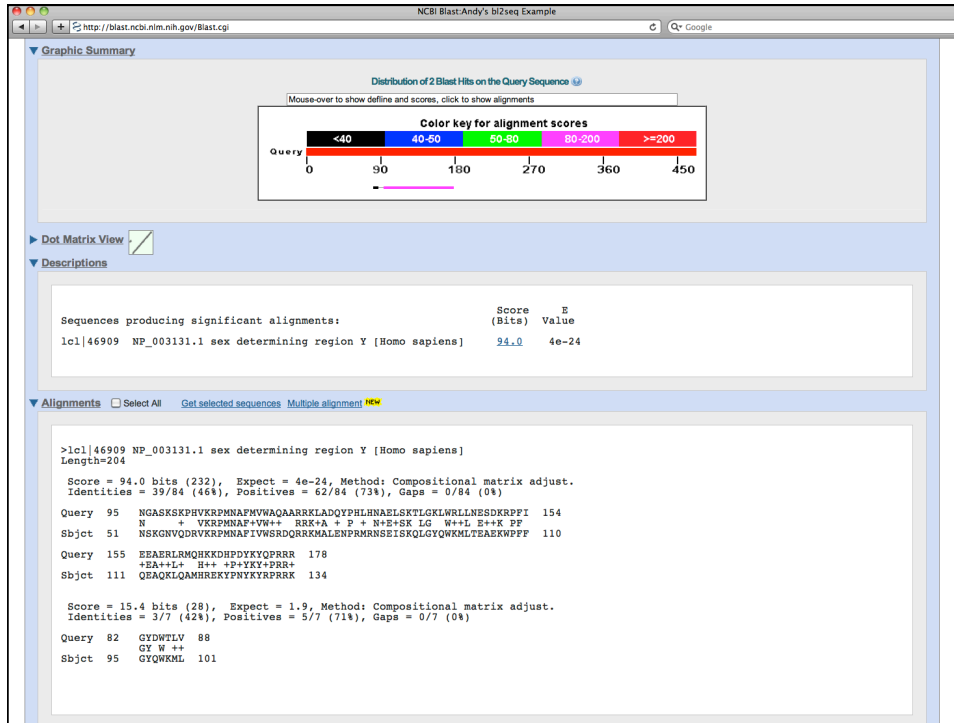
- nucleotide blast** Search a nucleotide database using a nucleotide query
Algorithms: blastn, megablast, discontinuous megablast
- protein blast** Search protein database using a protein query
Algorithms: blastp, psi-blast, phi-blast
- blastx** Search protein database using a translated nucleotide query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

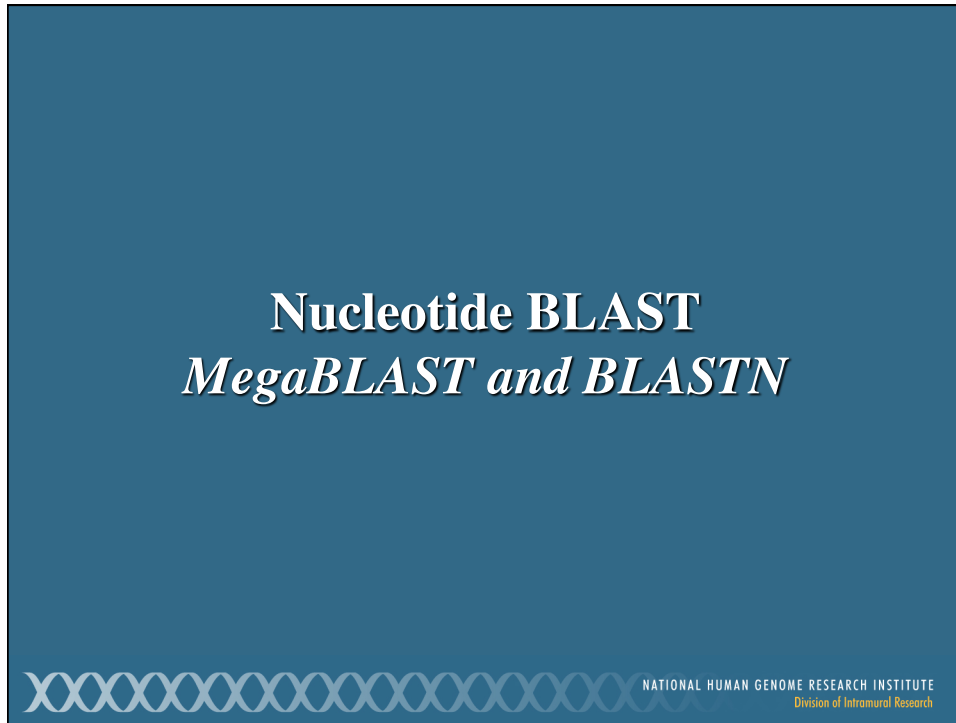
Specialized BLAST

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

- Make specific primers with [Primer-BLAST](#)
- 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 (or more) sequences using BLAST (tblastx)**
- Search [protein or nucleotide targets](#) in PubChem BioAssay
- Search [SRA transcript libraries](#)

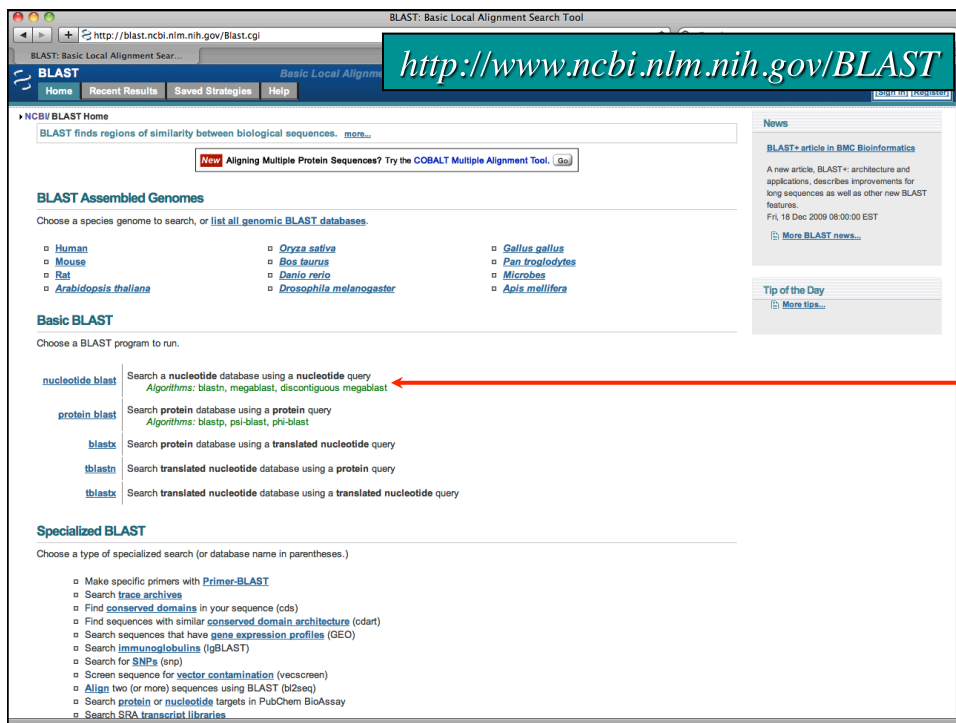






Nucleotide BLAST
MegaBLAST and BLASTN

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



BLAST: Basic Local Alignment Search Tool

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

BLAST finds regions of similarity between biological sequences. [more...](#)

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

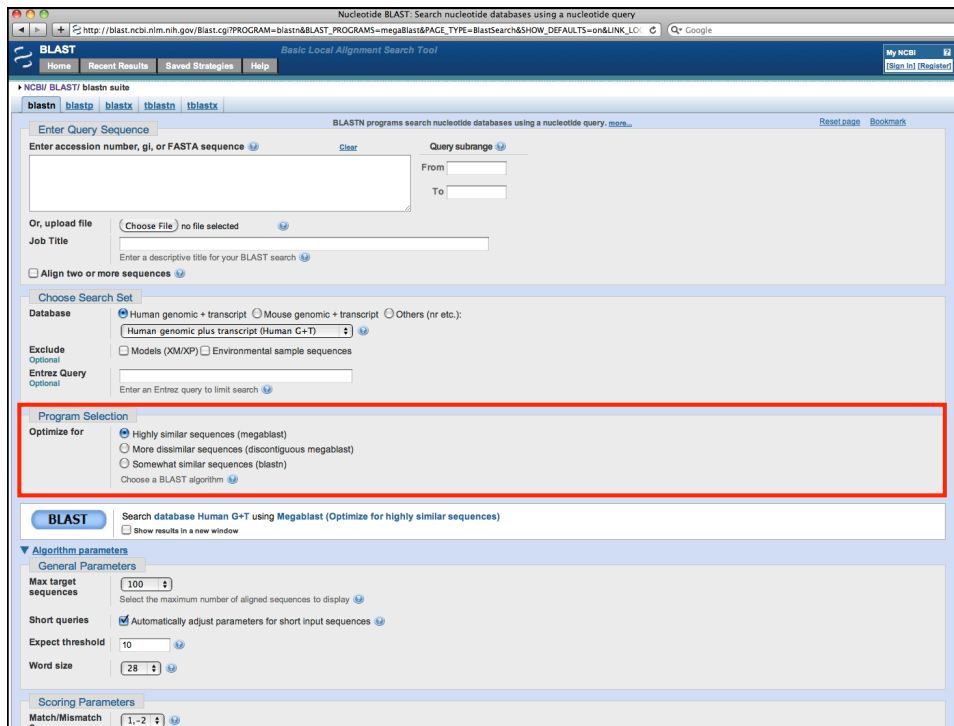
Choose a BLAST program to run.

- nucleotide blast** Search a nucleotide database using a nucleotide query
Algorithms: blastn, megablast, discontinuous 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.)

- Make specific primers with [Primer-BLAST](#)
- 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 (or more) sequences using BLAST (tblastn)
- Search [protein or nucleotide targets](#) in PubChem BioAssay
- Search [SRA transcript libraries](#)



Nucleotide-Based BLAST Algorithms

	<i>W</i>	<i>+/-</i>	<i>Gaps</i>
<i>Optimized for aligning very long and/or highly similar sequences (> 95%)</i>			
MegaBLAST (<i>default</i>)	28	1, -2	Linear
<i>Better for diverged sequences and/or cross-species comparisons (< 80%)</i>			
Discontiguous MegaBLAST	11	2, -3	Affine
BLASTN	11	2, -3	Affine
<i>Finding short, nearly exact matches (< 20 bases)</i>			
BLASTN <i>E = 1000, all filtering off</i>	7	2, -3	Affine

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



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



UCSC Genome Bioinformatics <http://genome.ucsc.edu>

Genomes - Blat - Tables - Gene Sorter - PCR - VisiGene - Proteome - Session - FAQ - Help

Genome Browser
ENCODE
Blat
Table Browser
Gene Sorter
In Silico PCR
Genome Graphs
Galaxy
VisiGene
Proteome Browser
Utilities
Downloads
Release Log
Custom Tracks
Archaeal Genomes
Mirrors
Archives
Training
Credits
Publications
Cite Us
Licenses

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](#).

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.

14 Dec. 2009 - New job posting: Biological Data Technician

The UCSC Genome Browser project is looking for a bioinformatician, biologist, or software engineer with a strong biology background to collect and import data into the UCSC Genome Browser database and website. This person will work closely with external research laboratories to capture their experimental results and methods and with internal software developers and database testing staff to make the data accessible to the worldwide scientific community.

Candidates must have a bachelor's degree in bioinformatics or a biological science (or equivalent experience), be proficient in UNIX/Linux command-line use, competent in UNIX shell scripting and Perl programming, and familiar with relational database concepts and SQL. Besides having the ability to quickly learn and interpret biological and technical information, the ideal candidate is an effective communicator, resourceful, and a diplomatic team player who is both quality-oriented and able to work effectively under deadline.

To find more information and application instructions for this job as well as other open positions with the UCSC Genome Browser project and the UCSC Center for Biomolecular Science and Engineering, see the CBSE [staff positions](#) web page.

7 Dec. 2009 - Human Genome Browser default changing to hg19: In conjunction with the release of the UCSC Genes and Conservation tracks on the hg19 (GRCh37) human assembly, we have changed the default human browser on our website from hg18 to hg19. [Read more.](#)

1 Dec. 2009 - New UCSC Genes and Conservation tracks released on hg19 browser: We're happy to announce the release of two of our most popular data sets on the hg19/GRCh37 human Genome Browser. [Read more.](#)

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.
- Some annotation tracks contributed by external collaborators contain proprietary data that have specific use restrictions. To check for

Rat BLAT Search

Home Genomes Tables Gene Sorter PCR Session FAQ Help

Rat BLAT Search

BLAT Search Genome

Genome: Assembly: Query type: Sort output: Output type:

```
>CB312815 NICH0_Rr.Pt1 Rattus norvegicus cDNA clone
GGGCTCTCGCTGGCTGTGCTCAGAACTGCTTCTCCACCTCTTCTGTGAATTCCTAAACTCTC
TACCTCTGGTTCATGCTCCCTCTTCTGGATAGTCTGTGCAATGAGCCCTTAAAGGAATTTGCAATGA
GCTATAAGAGTTGTGAGCCTCCGATAGCCAGGCTGCACTGGGACAGCAAGAAATTCATTGCATCT
GCTCTTAAGTCAAGGTTATCCAGAGCCCACTTACCCGAAGAGAGAGCTTCCCCCATCCCTAGGAAA
CAGTAGAGCTTAGGAAATGAATGACTCCACCACATCAAGAGGCTCAAATGTATACTGGCATTCT
GATTTGAGTCTGAAATTCGTCCCTAGTCTGGGAAATTAAGAAATGGAGTTACACCTTGCTATTTA
AAAAACCAATGAATTAAGCAAAATGGAAATCATCCACATAAAACATGTATGAACTGTTTCATGTTT
GATCATGGCCGGGATATAGCTCAGTCACTGACTGCTTGCATAGCAATGTGCATATCCGAGCTCAAGC
CCGACAGCAAAAGAGAAACGGGAGGATGACGACATTCACAGCAGGCTTTCAGTATAGCCGCAAG
GGGAGAGGTTTAAACACTACTAGGGAATGATAAGCCGAGTGCCTTCTATATACTGGGGATGGCT
AGTCATCACGTAAGAAAAGTTGGAAATGATAAAATCCAATGGATGGATCCCTTTAAACATCC
```

submit clear

Paste in a query sequence to find its location in the the genome. Multiple sequences may be searched in 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: no file selected

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 25 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,

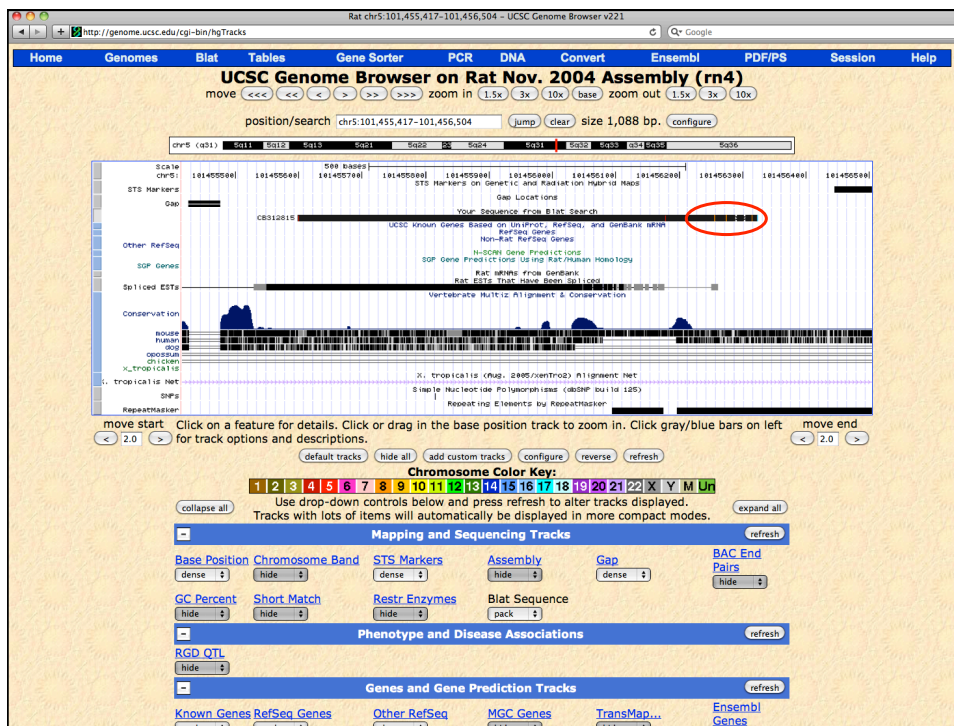
Rat BLAT Results

Home Genomes Tables Gene Sorter PCR Session FAQ Help

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 BLAT Results

BLAT Search Results

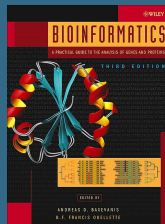
ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
browser details	CB312815	710	1	733	768	98.1%	5	+	101455599	101456323	725
browser details	CB312815	29	501	537	768	89.2%	2	+	38736251	38736287	37
browser details	CB312815	25	501	529	768	93.2%	3	+	22960346	22960374	29
browser details	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

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>

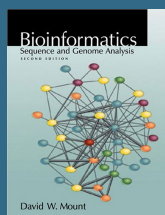


Further Reading



Chapter 11

*Assessing Pairwise Sequence Similarity:
BLAST and FASTA*



Chapter 6

*Sequence Database Searching for
Similar Sequences*

