
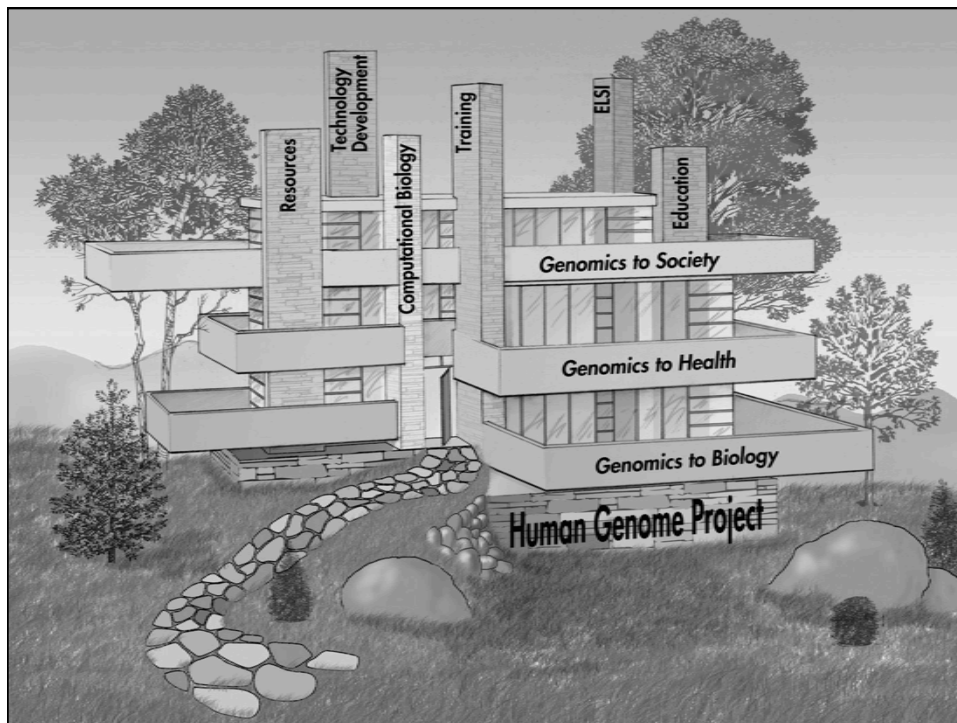


**Current Topics in Genome Analysis
Spring 2010**

Week 2: Biological Sequence Analysis

Andy Baxevanis, Ph.D.

 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



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 β

Gene duplication →

- 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	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0	-4	
R	-1	5	0	-2	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1	-4	
N	-2	0	6	1	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1	-4	
D	-2	-2	1	6	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	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-3	-2	-4	
Q	-1	0	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	-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	0	0	1	1	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	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

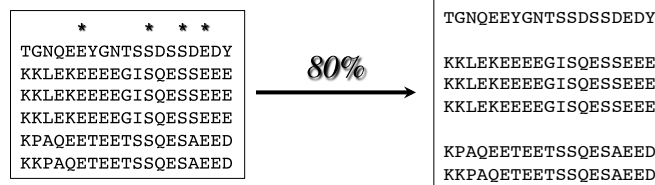
- Henikoff and Henikoff, 1992
- Blocks Substitution Matrix
 - Look only for differences in conserved, ungapped regions of a protein family (“blocks”)
 - Directly calculated, using no extrapolations
 - More sensitive to detecting structural or functional substitutions
 - Generally perform better than 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 sensitivity to the search, while statistical significance adds selectivity (see UNIT 4). Virtually every user chooses the default, typically PAM 250 or BLOSUM62. Despite the fact that the choice of matrix can strongly influence the outcome of the analysis, most users do not know why a particular matrix should be used. In general, scoring matrices implicitly represent a particular theory of protein sequence evolution. This unit provides guidance in the choice of a scoring matrix, 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 trademark acronym derived from Accepted Point Mutation (Dayhoff, 1978) is a probabilistic model for amino-acid replacement derived by computing 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 nonstationary 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 threonine (A to T) in some substituted side chain. The indicator is that if two aligned sequences manifest a higher than expected prevalence of these characteristic replacements, the sequences are related. An excellent discussion of the derivation and use of the PAM matrices is given in George et al. (1995).

PAM matrices are the result of computing the probability of one substitution per 100

amino acids, called the PAM 1 matrix. Higher PAM matrices are derived by multiplying the PAM 1 matrix by itself a defined number of times. Thus, a PAM 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 UNIT 4) regarding insertion and deletions. This second meaning, but considered that over evolutionary time, it is possible that an alanine was changed first to a glycine, then to a valine, and then back to an alanine. These silent substitutions are derived from observed amino acid frequency data in protein families and superfamilies.

Choosing a PAM Matrix

It is extremely important to note that PAM matrices are derived from protein sequence data available in the late 1960s and early 1970s. Most proteins known at that time were small, globular, and hydrophilic. If the researcher believes their protein contains substantial hydrophobic regions, such as membrane-spanning helices or sheets, the PAM matrices are less useful than others described in this unit. Dayhoff et al. (1978) were the first to define the nonprotein 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 some protein families. The user should be aware that while the terms "family" and "superfamily" are widely used and long ago, 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 detecting similarities 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

Further
Statistics and
Illustrations
3.5.1

Contributed by David Wheeler
Current Protocol in Bioinformatics (UNIT 3.5) 3.5.6
Copyright © 2005 by John Wiley & Sons, Inc.



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

↓

Neighborhood Words	PQG 18 PEG 15 PRG 14 PKG 14 PNG 13 PDG 13 PHG 13 PMG 13 PSG 13 PQA 12 PQN 12 etc.	= 7 + 5 + 6 Neighborhood Score Threshold ($T = 13$)
---------------------------	--	--

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

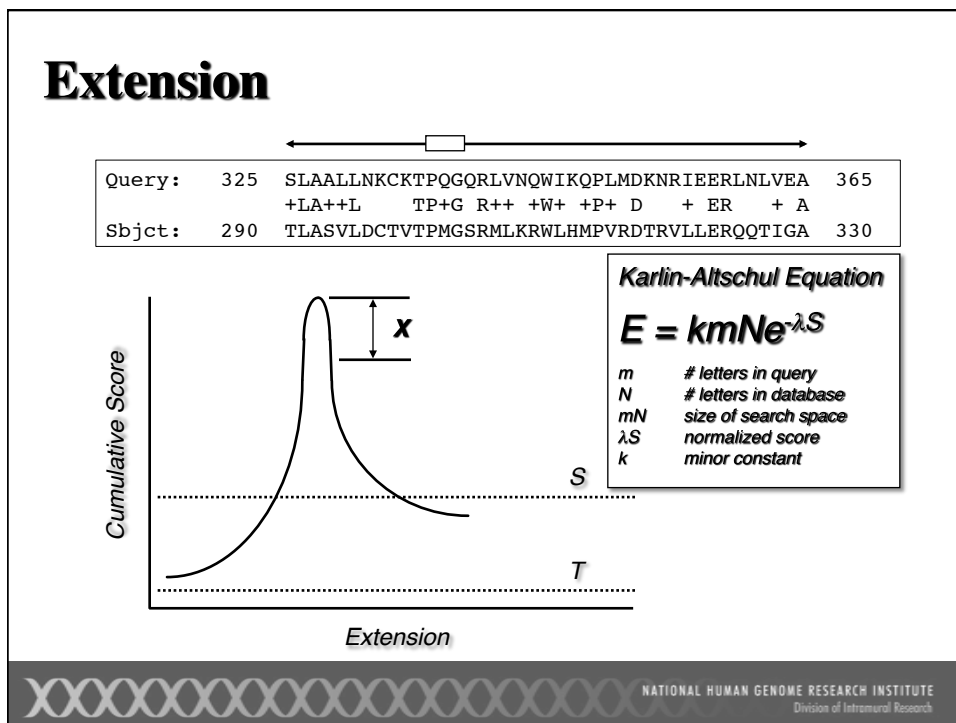
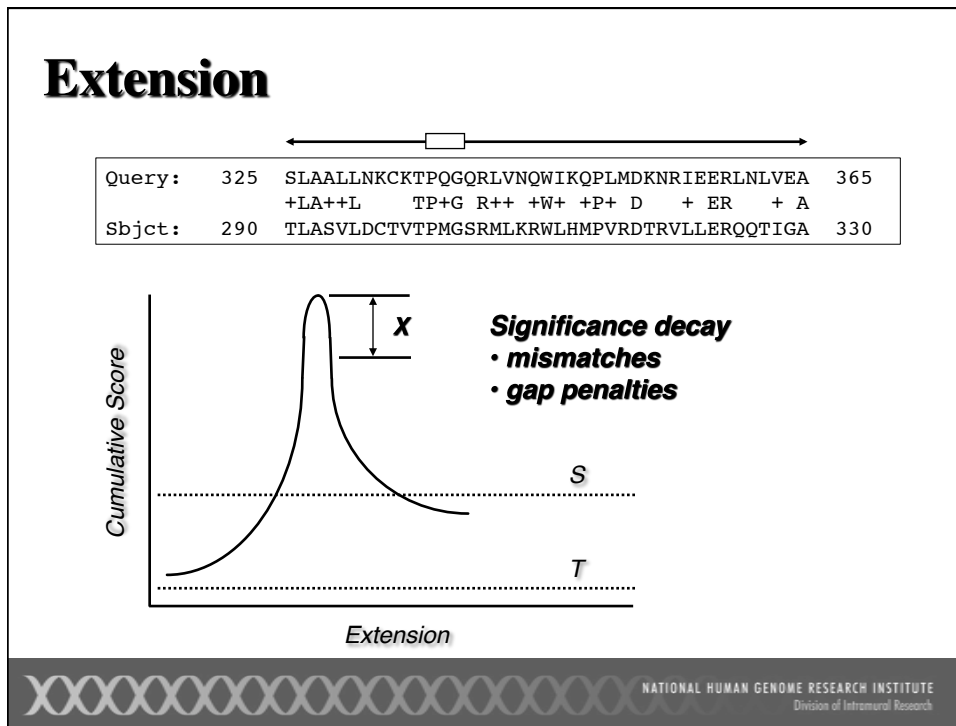
High-Scoring Segment Pairs

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

↓

	←	□	→	
Query:	325	SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA	365	
		+LA++L TP+G R++ +W+ +P+ D + ER + A		
Sbjct:	290	TLASVLDCTVTPMGSRLKRWLHMPVRDTRVLLERQQTIGA	330	

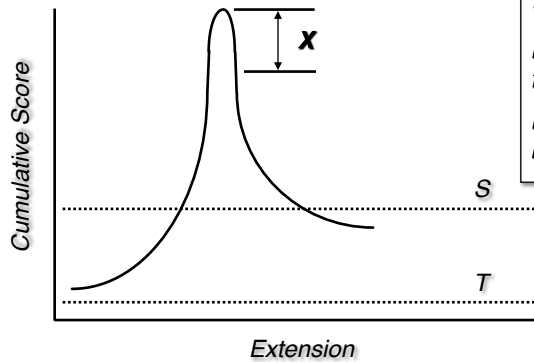
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



Scores and Probabilities

←—————|—————→

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



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

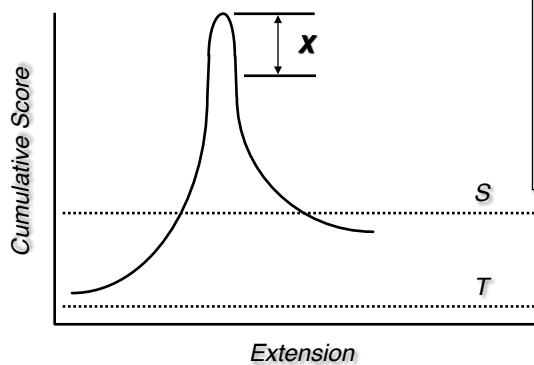
Number of HSPs
 found purely by chance
 Lower values signify
 higher similarity



Scores and Probabilities

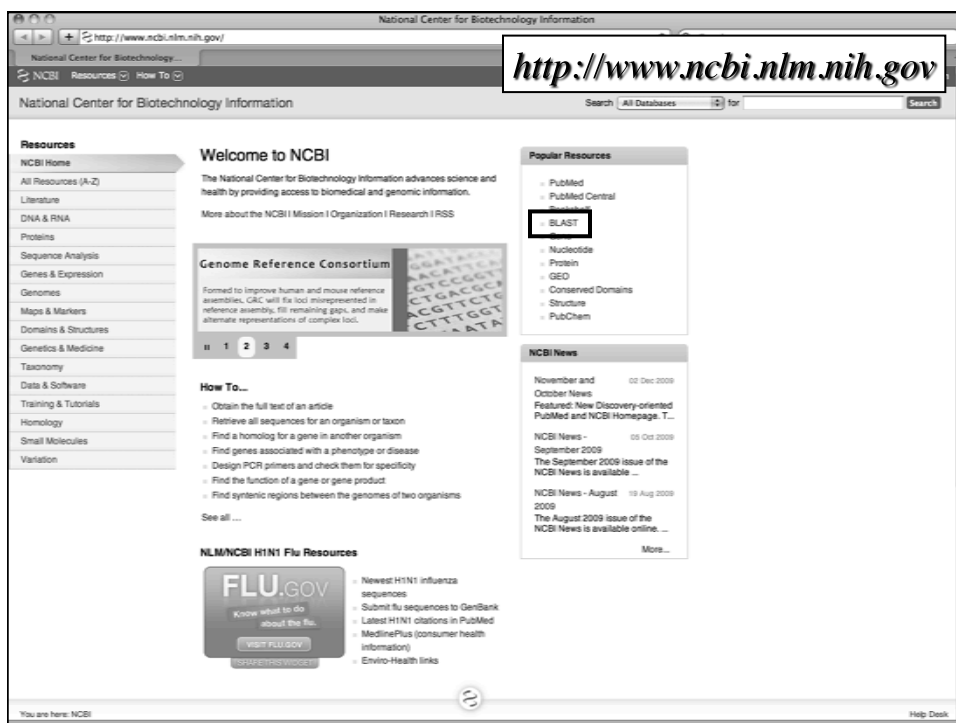
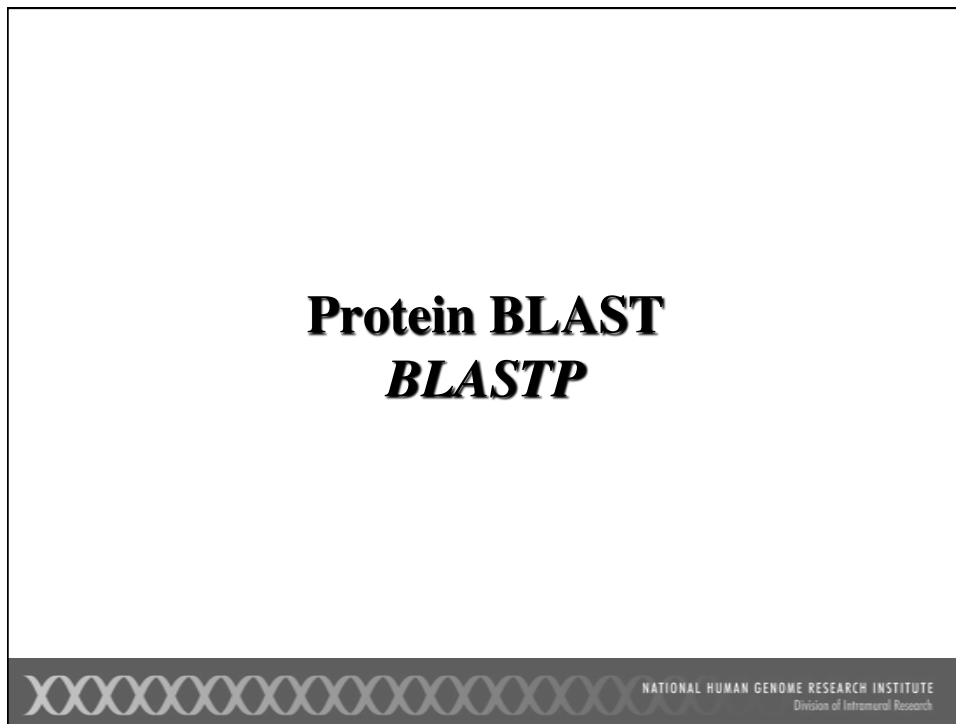
←—————|—————→

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



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





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:

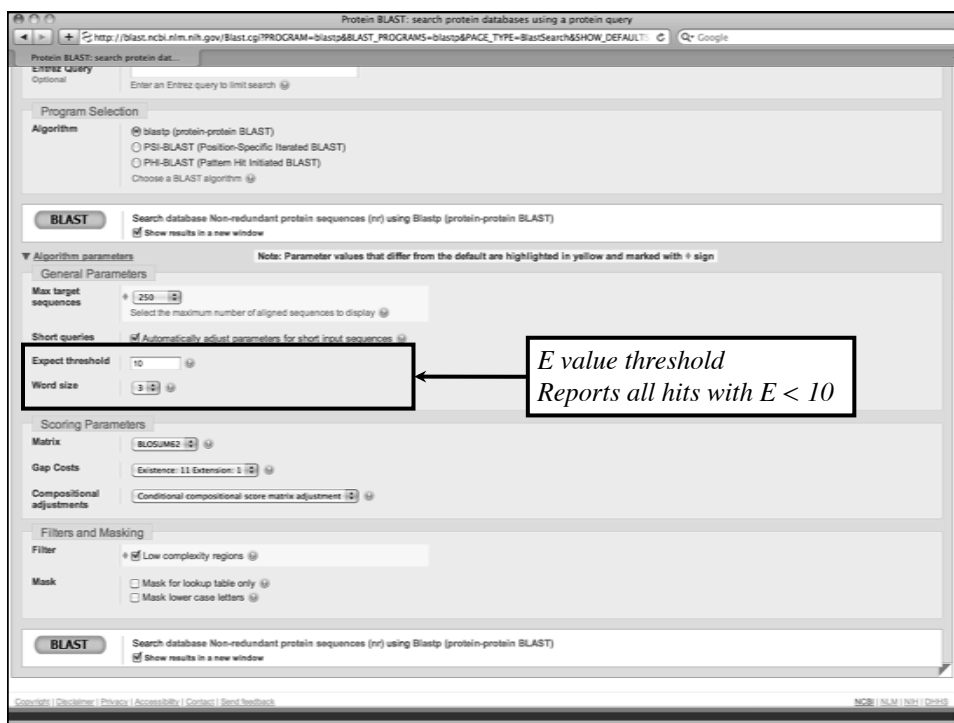
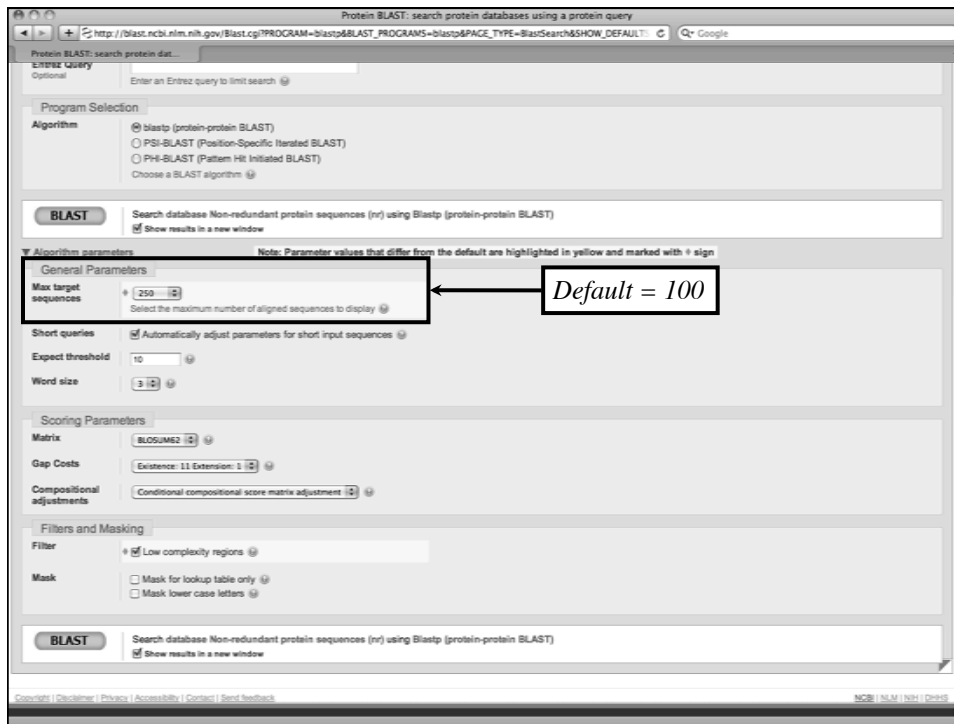
XM_123456	Model mRNA
XP_123456	Model proteins

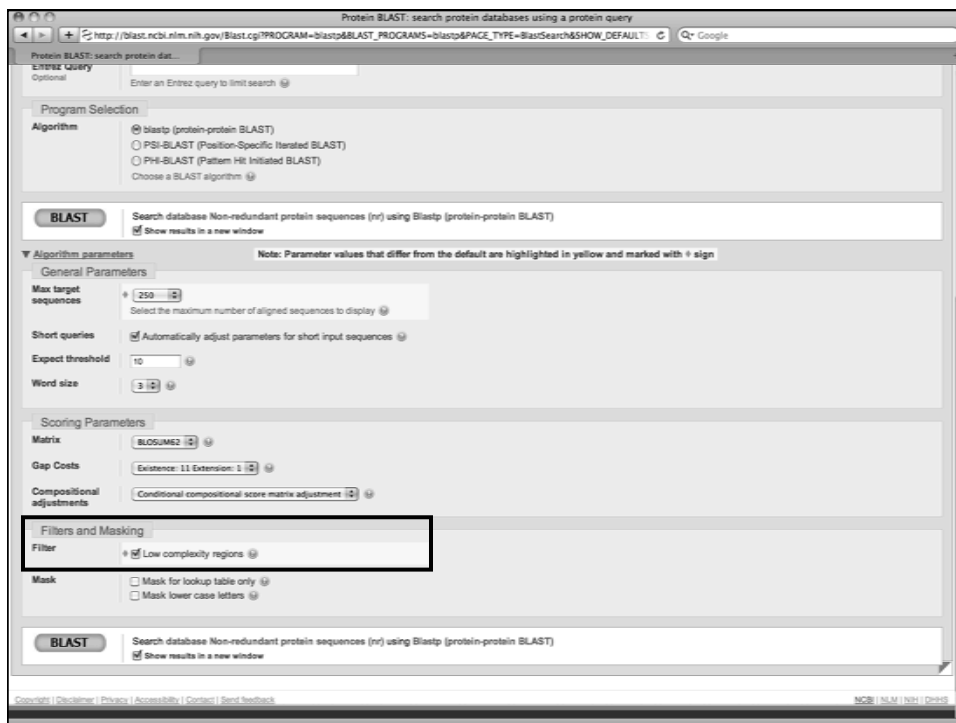
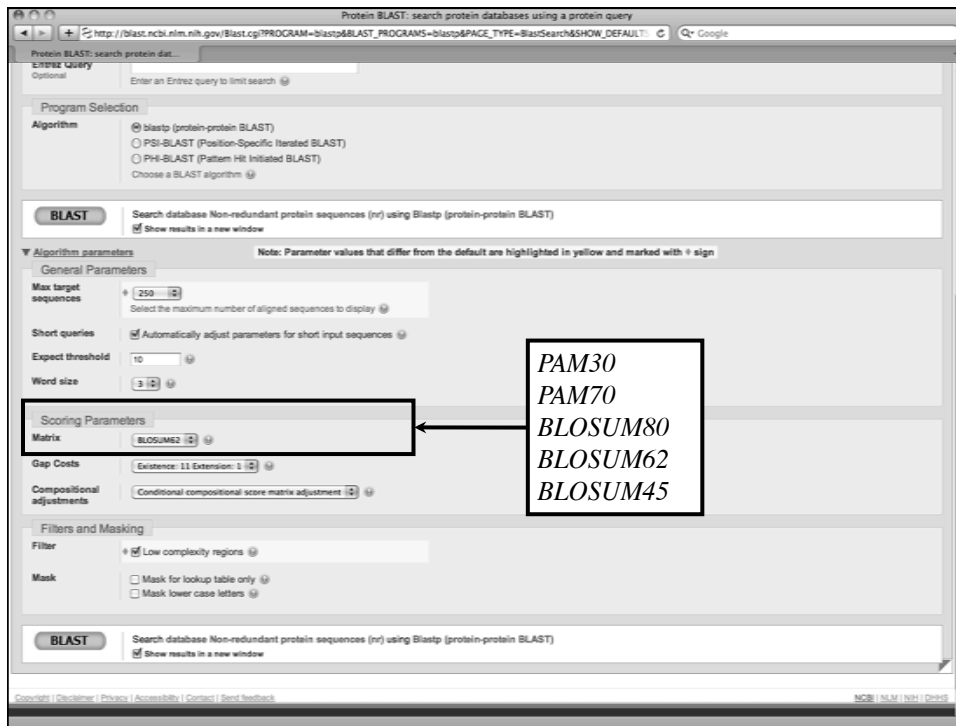
Complete key at

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



The screenshot shows the NCBI BLAST search interface. The 'Choose Search Set' section is highlighted with a box and an arrow pointing to the 'Organism' field. A text box next to the arrow contains the text 'Limit by organism or taxonomic group'. The interface includes fields for 'Enter Query Sequence', 'Choose Search Set', and 'Program Selection'. The 'Organism' field is currently empty, and the 'Exclude' checkbox is checked. The 'Program Selection' section shows 'blastp (protein-protein BLAST)' selected.





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)  
MESSAKMESGGAGQQPQPQQPFLPPAACFFAIAAAAAAAAAAQAQQQQQQQQQQQAPQLRPAA  
DGQPSGGGHSAPKQVKRQRSSPELMRCKRRLNFSGFGYSLFQQQAAVARRNERERNRKLVLNLFAT  
LREHVPNGAANKKMSKVETLRSAYEYIRALQQLLDEHDAVSAAFQAVLSPTISPNYSNDLNSMAGSPVS  
SYSSDEGSYDPLSPPEQELDFTNWF
```

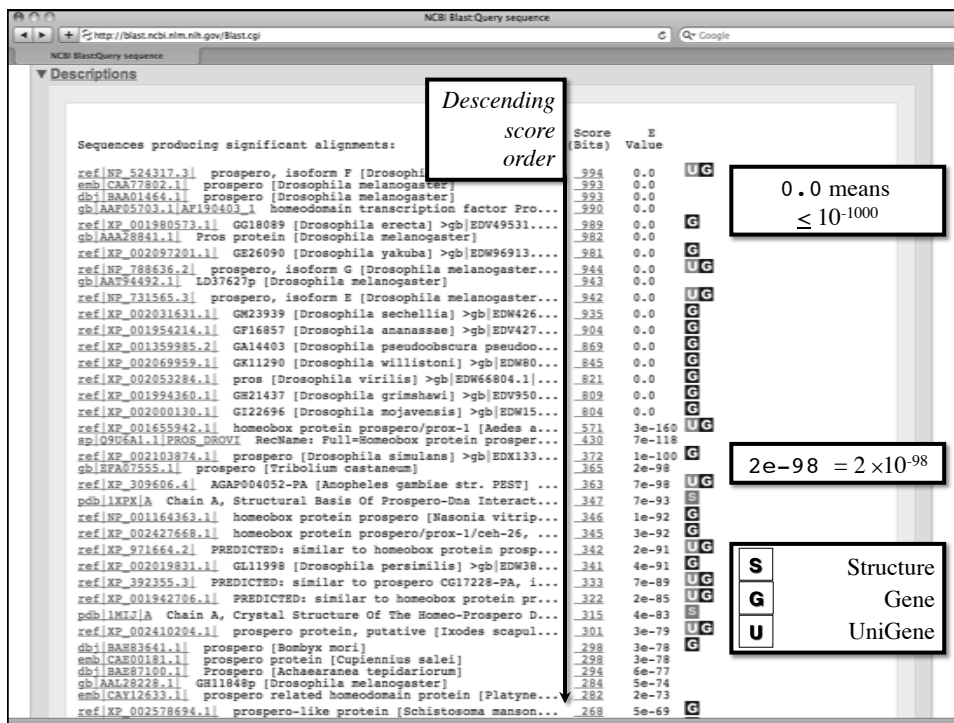
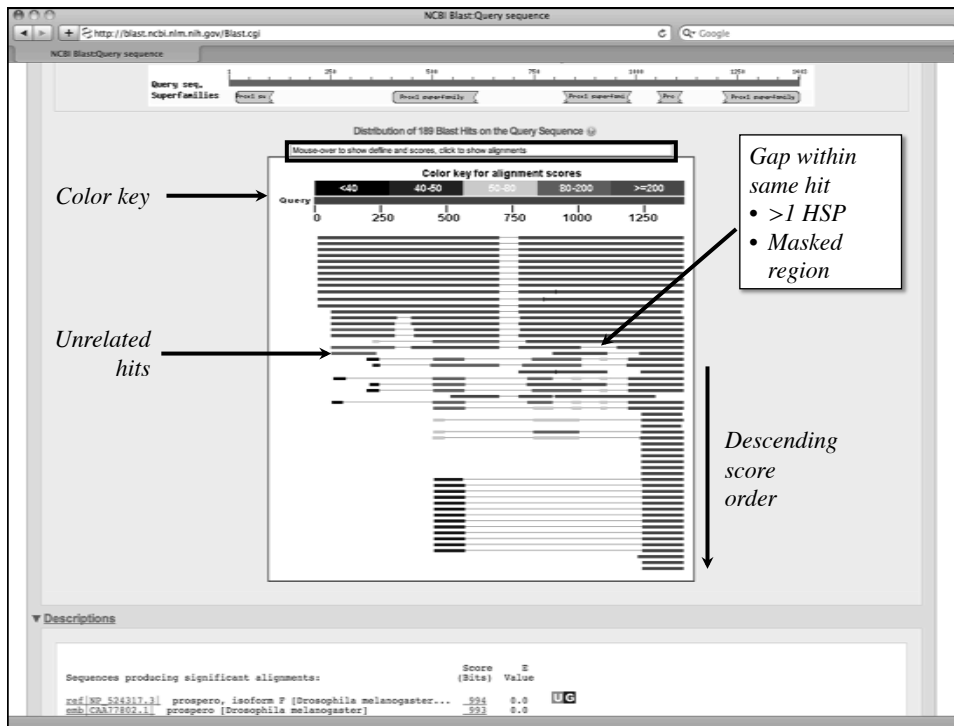
Homopolymeric
alanine-glutamine tract

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 for the protein query. Below it, the 'Program Selection' section has radio buttons for 'blastp (protein-protein BLAST)', 'PSI-BLAST (Position-Specific Iterated BLAST)', and 'PHI-BLAST (Pattern Hit Initiated BLAST)'. The 'blastp' option is selected. A 'BLAST' button is visible. The 'Algorithm parameters' section includes 'General Parameters' with 'Max target sequences' set to 250, 'Short queries' checked, 'Expect threshold' at 10, and 'Word size' at 3. 'Scoring Parameters' includes 'Matrix' set to BLOSUM62, 'Gap Costs' set to Existence: 11 Extension: 1, and 'Compositional adjustments' checked. 'Filters and Masking' includes 'Filter' checked for 'Low complexity regions' and 'Mask' options for 'Mask for lookup table only' and 'Mask lower case letters'. A 'BLAST' button is also present at the bottom left of the parameters section.

The screenshot shows the NCBI BlastQuery sequence results page. The 'Query sequence' section displays: Query ID: lc152739, Description: Query sequence, Molecule type: amino acid, Query Length: 1403, Database Name: nr, Description: All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects, Program: BLASTP 2.2.22+. Below this is a 'Graphic Summary' section with a 'Show Conserved Domains' button. A horizontal bar chart shows 'Putative conserved domains have been detected, click on the image below for detailed results.' Below that, a 'Distribution of 189 Blast Hits on the Query Sequence' is shown as a horizontal bar chart. A 'Color key for alignment scores' is provided: <table border='1'><tr><th><math><math><math><math></tr></table>



NCBI Blast Query 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 UG
 sp|Q388N5.2|PROX2_HUMAN RecName: Full=Prospero homeobox prote... 208 5e-51 UG
 ref|NP_010773877.1| 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 UG
 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 G
 ref|NP_001071961.1| transcription factor protein [Clona intes... 199 2e-48 UG
 emb|CAG04695.1| unnamed protein product [Tetraodon nigroviridis] 198 4e-48 G
 emb|CAF95276.1| 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|AAC28335.1| 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/ceb-26 [... 167 9e-39 UG
 db|BAB17311.1| Prox 1 [Cynops pyrrhogaster] 161 4e-37 G
 gb|EAW81198.1| hCG22353 [Homo sapiens] 158 4e-36 G
 db|BAC04278.1| unnamed protein product [Homo sapiens] 157 8e-36 G
 gb|AAC59781.1| prospero-like protein [Takifugu rubripes] 156 1e-35 G
 gb|EDL02848.1| RIKEN cDNA 1700058C01, isoform CRA_a [Mus musc... 154 7e-35 G
 emb|CAI15389.1| prospero homeobox 1 [Homo sapiens] 154 1e-34 UG
 ref|NP_849216.1| PREDICTED: similar to prospero-related homeo... 154 1e-34 G
 gb|EFP18559.1| hypothetical protein FANDA 009835 [Ailuropoda ... 152 2e-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|AAD30189.1|AC006530.2 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
 db|J05454| Prox 1 protein 671 - chicken 80.1 2e-12 UG
 ref|NP_001100671.1| prospero homeobox 1 [Rattus norvegicus] >... 44.7 0.091 UG
 emb|CAF94749.1| unnamed protein product [Tetraodon nigroviridis] 43.5 0.17 G
 emb|CAF58279.1| Prox1 protein [Xenopus tropicalis] 42.0 0.64 G
 gb|AAF13029.1|AF070733.1 transcription factor Prox1 [Notopth... 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 Blast Query 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 LFQPQSVSTANSSSSNNNSSTPAALATHSPtenspvsqassasslltaeFCNLFQGS... 76
 LFPQPSVSTANSSSSNNNSSTPAALATHSPtenspvsqassasslltaeFCNLFQGS...
 LFPQPSVSTANSSSSNNNSSTPAALATHSPtenspvsqassasslltaeFCNLFQGS... 376

Query 77 KMLNELFGRQMKQAQDATSGLPQSLDNAMLAAMETATSAEILLIGSLNSTSKLLQQQHNN 136
 KMLNELFGRQMKQAQDATSGLPQSLDNAMLAAMETATSAEILLIGSLNSTSKLLQQQHNN
 KMLNELFGRQMKQAQDATSGLPQSLDNAMLAAMETATSAEILLIGSLNSTSKLLQQQHNN 436

Query 137 NSIAPANSTPMSNGTNASISPGSAHSSSHSHQGVSPKGSRRVSAACSDRSLEAAADVAGG 196
 NSIAPANSTPMSNGTNASISPGSAHSSSHSHQGVSPKGSRRVSAACSDRSLEAAADVAGG
 NSIAPANSTPMSNGTNASISPGSAHSSSHSHQGVSPKGSRRVSAACSDRSLEAAADVAGG 496

Query 197 SPPRAASVSLNGGASGSGEQHSQQLQHDLVAAHMLRNILQKKEIMQLDQELRTAMQQQQ 256
 SPPRAASVSLNGGASGSGEQHSQQLQHDLVAAHMLRNILQKKEIMQLDQELRTAMQQQQ
 SPPRAASVSLNGGASGSGEQHSQQLQHDLVAAHMLRNILQKKEIMQLDQELRTAMQQQQ 556

Query 257 qllqeqeqlHSLKLNnnnnnnaaTANNnnntTMSINLIDDSEMA DIKISEPQTAPQPQ 316
 QLLQEQQLHSLKLNnnnnnnaaTANNnnntTMSINLIDDSEMA DIKISEPQTAPQPQ
 QLLQEQQLHSLKLNnnnnnnaaTANNnnntTMSINLIDDSEMA DIKISEPQTAPQPQ 616

Query 317 QphgshsrsrgsgsgshssmasdgsLrrkssdsLdeHGagddaqdeedaPTGQRSES 376
 QSPHGS SHSRSGSGSGSHSSMASDGS LRRKSSDSDLDSHGAQDDAQDEEDAAPTQRSES
 QSPHGS SHSRSGSGSGSHSSMASDGS LRRKSSDSDLDSHGAQDDAQDEEDAAPTQRSES 676

Query 377 RAPEEPQLPTKKEVDMLDEVELLGLSHRSGSDMSLASPSHSmmlLdkddvldedddd 436
 RAPEEPQLPTKKEVDMLDEVELLGLSHRSGSDMSLASPSHSmmlLdkddvldedddd
 RAPEEPQLPTKKEVDMLDEVELLGLSHRSGSDMSLASPSHSmmlLdkddvldedddd 736

Query 437 dCVEQRTSGSGCLKPKGMDLKRARVENIVSGMRCSPSSGLAQAGLQVNGCKKRLYQFPQ 496
 DCVEQRTSGSGCLKPKGMDLKRARVENIVSGMRCSPSSGLAQAGLQVNGCKKRLYQFPQ
 DCVEQRTSGSGCLKPKGMDLKRARVENIVSGMRCSPSSGLAQAGLQVNGCKKRLYQFPQ 796

Query 497 QHAMERYVaaagLNFGLNLQSMMLDQDSESNLESFPQIQKRVKKNALKSLRSMQEQ 556
 QHAMERYVaaagLNFGLNLQSMMLDQDSESNLESFPQIQKRVKKNALKSLRSMQEQ
 QHAMERYVaaagLNFGLNLQSMMLDQDSESNLESFPQIQKRVKKNALKSLRSMQEQ 856

≥ 25% for proteins
 ≥ 70% for nucleotides

– Gap
 a Low-Complexity

NCBI Blast Query sequence

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

Annotations:

- No definition line → Second HSP identified
- Gap a Low-Complexity

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

<40	40-60	60-80	80-100	>=200
-----	-------	-------	--------	-------

Query

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...](#)

[Aligning Multiple Protein Sequences? Try the COBALT Multiple Alignment Tool.](#)

BLAST Assembled Genomes

Choose a species genome to search, or [list all genomic BLAST databases.](#)

<input type="checkbox"/> Human	<input type="checkbox"/> <i>Oriza sativa</i>	<input type="checkbox"/> <i>Gallus gallus</i>
<input type="checkbox"/> Mouse	<input type="checkbox"/> <i>Bos taurus</i>	<input type="checkbox"/> <i>Pan troglodytes</i>
<input type="checkbox"/> Rat	<input type="checkbox"/> <i>Danio rerio</i>	<input type="checkbox"/> Microbes
<input type="checkbox"/> <i>Arabidopsis thaliana</i>	<input type="checkbox"/> <i>Drosophila melanogaster</i>	<input type="checkbox"/> <i>Apis mellifera</i>

Basic BLAST

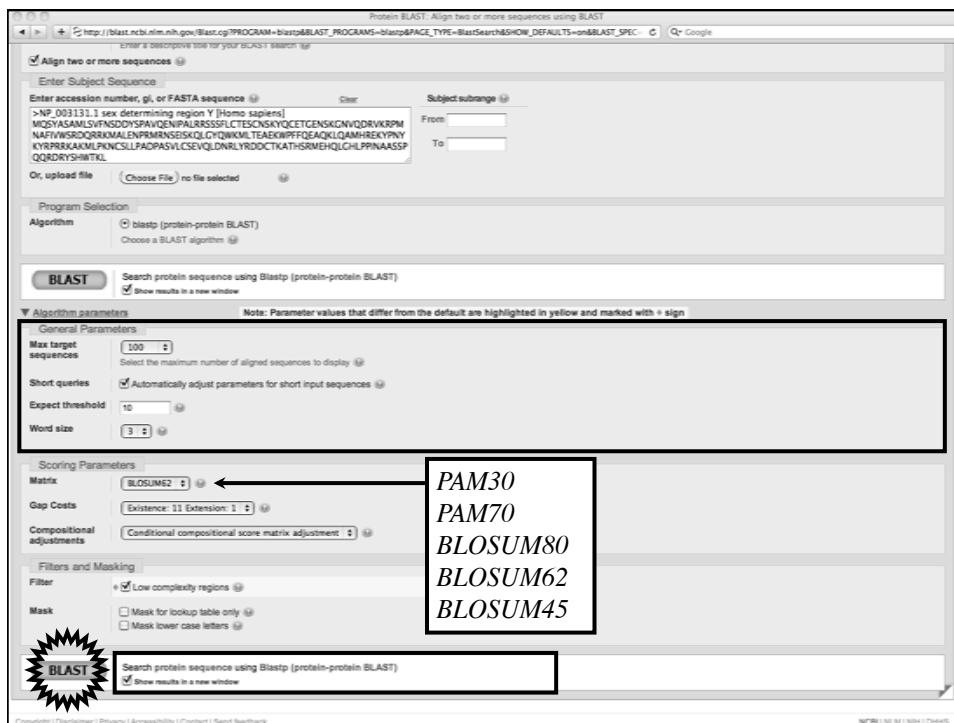
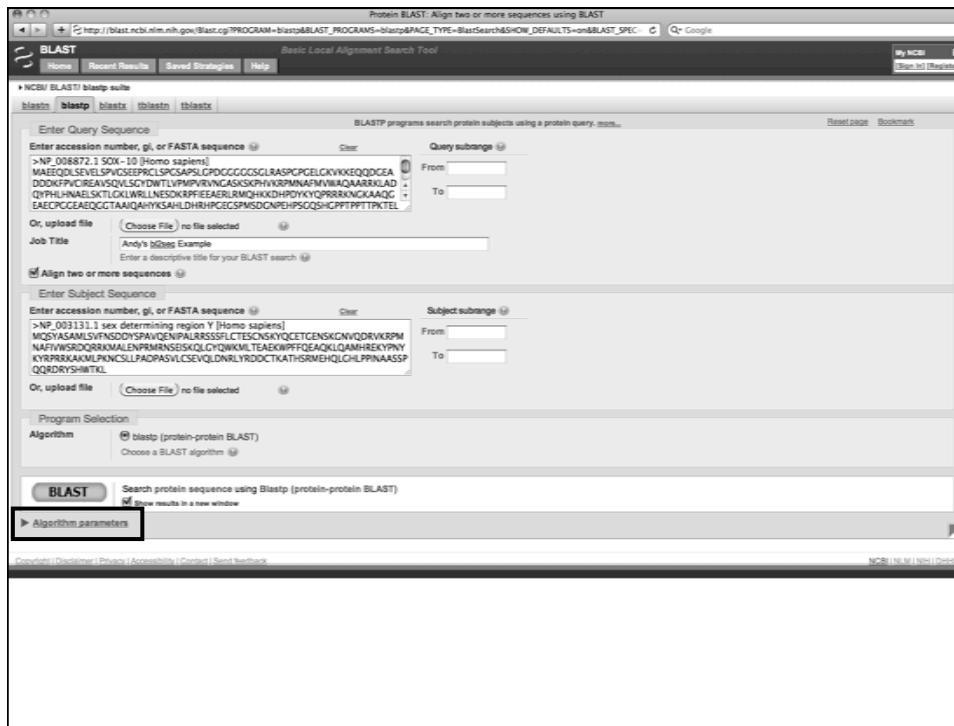
Choose a BLAST program to run.

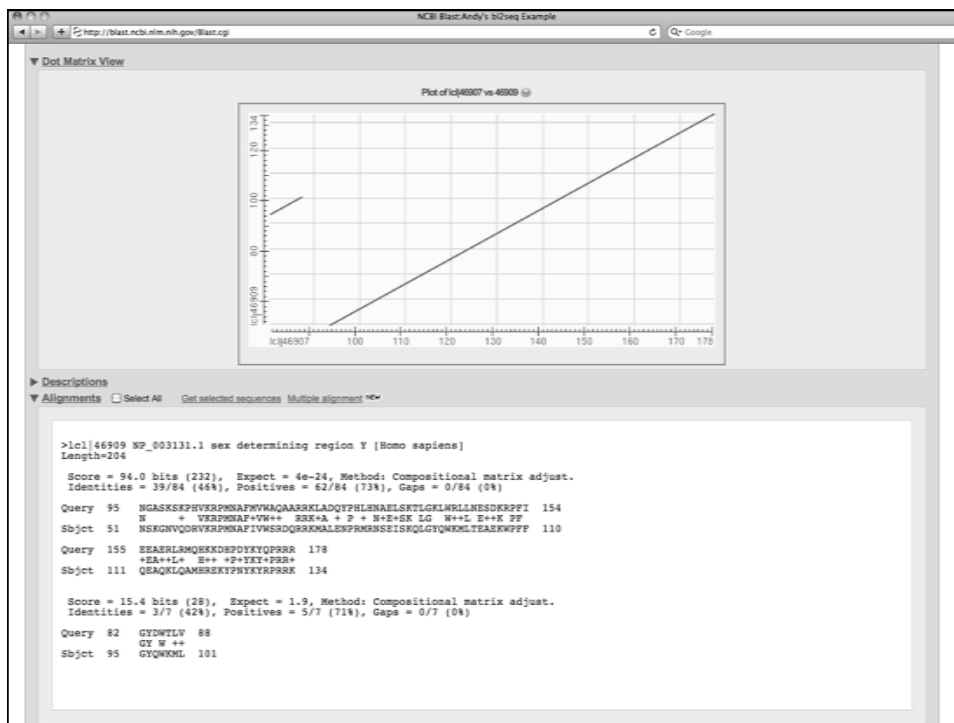
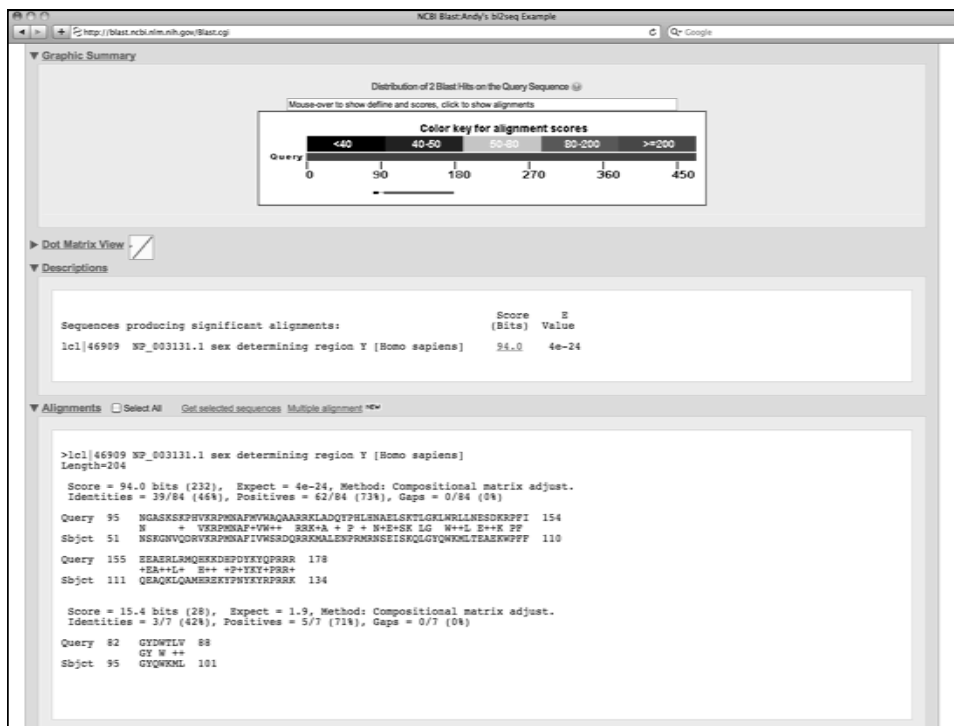
nucleotide blast	Search a nucleotide database using a nucleotide query Algorithms: blastn, megablast, discontinuous megablast
protein blast	Search protein database using a protein query Algorithms: blastp, psi-blast, phi-blast
tblastx	Search protein database using a translated nucleotide query
tblastn	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](#) (vectorscreen)
- [Align two \(or more\) sequences](#) using BLAST (zlibseq)
- Search protein or nucleotide targets in [PubChem BioAssay](#)
- Search [SRA transcript libraries](#)

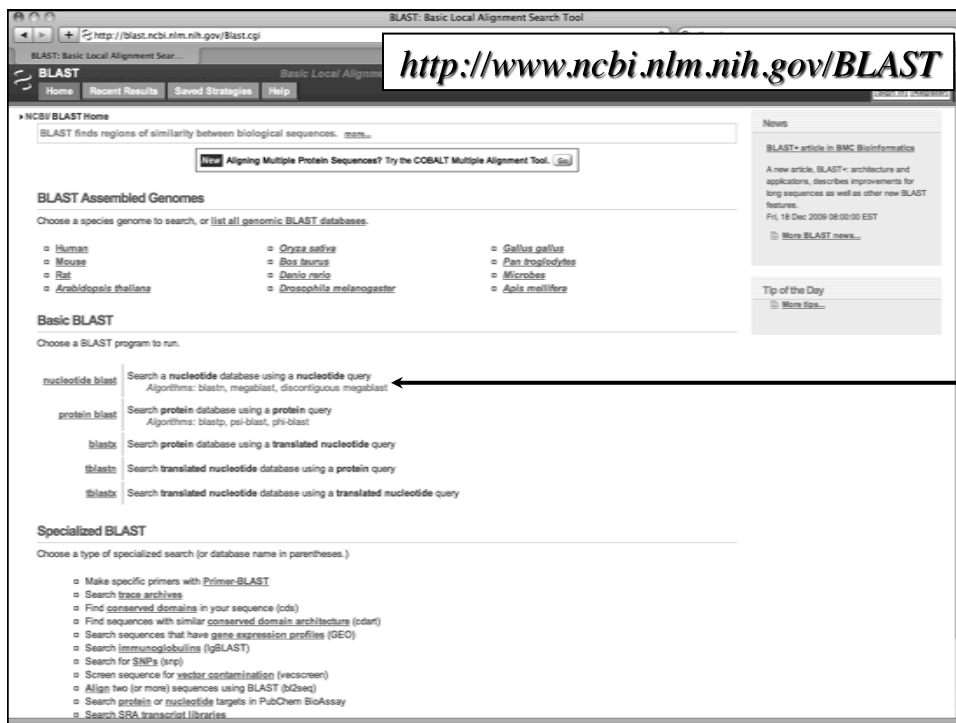




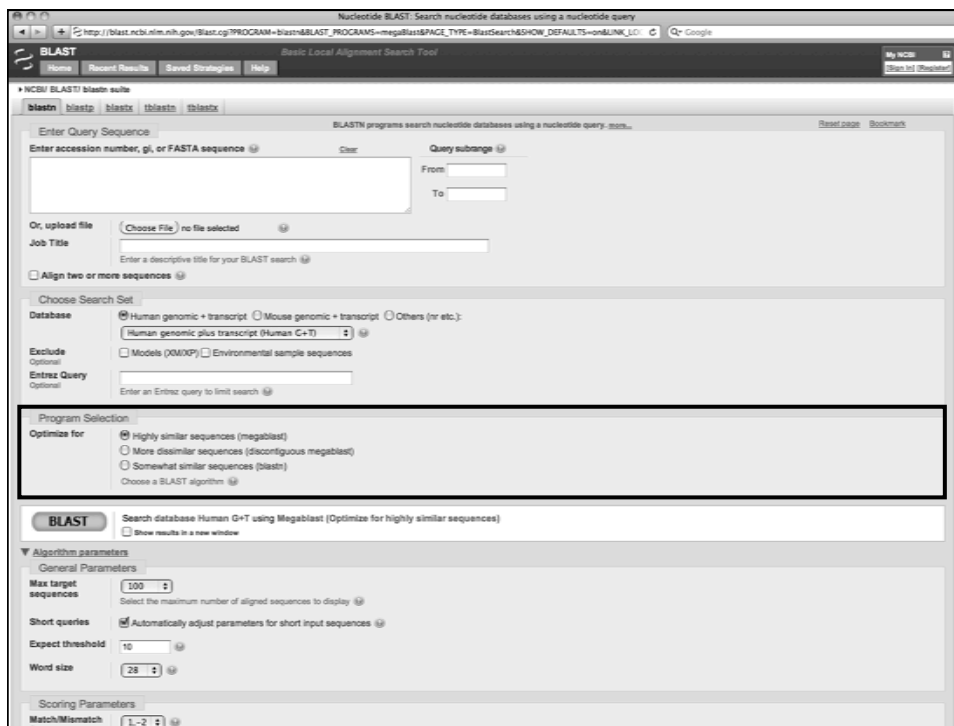
Nucleotide BLAST *MegaBLAST and BLASTN*



NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



The screenshot shows the NCBI BLAST website interface. At the top, the URL <http://www.ncbi.nlm.nih.gov/BLAST> is displayed. The main content area is titled "NCBI BLAST Home" and includes a search bar with the text "BLAST finds regions of similarity between biological sequences." Below this, there are several sections: "BLAST Assembled Genomes" with a list of species (Human, Mouse, Rat, Arabidopsis thaliana, Oriza sativa, Bos taurus, Danio rerio, Drosophila melanogaster, Gallus gallus, Pan troglodytes, Microbes, Apis mellifera); "Basic BLAST" with a list of search programs (nucleotide_blast, protein_blast, blastx, tblastx); and "Specialized BLAST" with a list of specialized search options (Primer-BLAST, trace_archives, conserved_domains, conserved_domain_architecture, gene_expression_profiles, immunoglobulins, SNPs, vector_contamination, pairwise, protein_or_nucleotide_targets, SRA_transcript_libraries). A red arrow points to the "nucleotide_blast" option in the Basic BLAST section.



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



<http://genome.ucsc.edu>

UCSC Genome Bioinformatics

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

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 NICHD_Rr_Px1 Rattus norvegicus cDNA clone
GGGCTCTCGCTGGCTGTGCTCAGAACTGGCTTCTCCACCTCTTCTGTGAATTCCTAAACTCTC
TACCTCTGGTTCATGCTCCCTCTTCTGGATGCTGTGTGCAATGACCCCTTAAAGGAATTTGCAATGA
CCTATAAGACTTGTGACCTCGCGTAGCCAGGCTTGCACCTGGCAGCAGAAAGAAATTCATTGGCATC
CTCTCTAAGTCAAGGTTATCCAGAGCCACTTTACCCGAAGAGAGAGCTTCCCCCATCCCTAGGAAA
CAGTAGACCTTACGAAATGAATGACTCCACCACATTCAGAGGCTTCAAATGTACTGGCATTCT
GATTCAGTCTCAAATTCGTCCCTAGTCTGGGAAATAGAAATGGAGTTACACCTTGTGATTTA
AAAAACATTAATTAAGACAAATGCAAAATCATGCCACATAAAACATGTATGAAGTGTTCATGTTT
CATCATGGCCCGGATATAGCTGACTGCTGACTGCTTGCATAGCAATTTGCATATCCAGCTTCAAGC
CCGACCCGAAAGAGAAACGGGAGGAGTGGACGATTCACAGCACCGTTTTCAGTATAGGCCCAAG
GGAAAGAGTAAACACCTACTCAGGGAATGATAAGCCGAGTGCCTTCTCTATACTCGGGGATGGCT
AGTCATCAGCTAAGAAAGTTCCATATGATAAATACCAATGGATGGATCCCCCTTAAACCATCC
```

submit I'm feeling lucky clear

Paste in a query sequence and its location in the genome. Multiple sequences may be searched, 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 10,000 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,

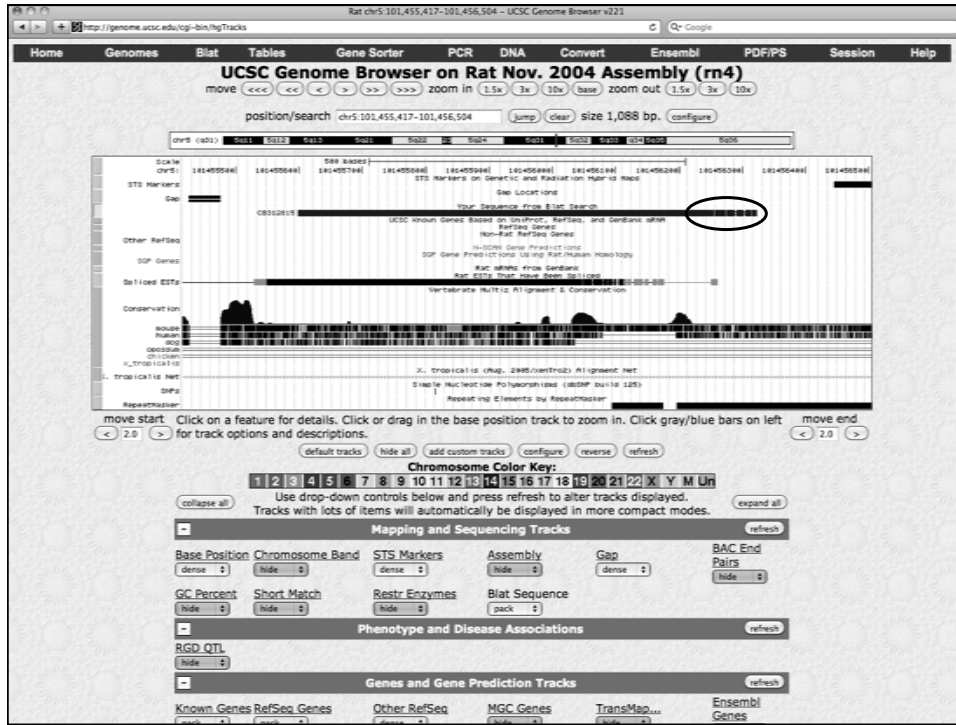
Rat BLAT Results

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

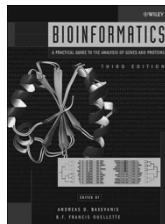
ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
browser details	CB312815	710	1	733	768	98.1%	5	+	101455599	101456323	725
browser details	CB312815	29	501	537	768	89.2%	2	+	38736251	38736287	37
browser details	CB312815	25	501	529	768	93.2%	3	+	22960346	22960374	29
browser details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

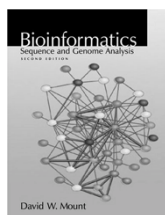


Further Reading



Chapter 11

*Assessing Pairwise Sequence Similarity:
BLAST and FASTA*



Chapter 6

*Sequence Database Searching for
Similar Sequences*

