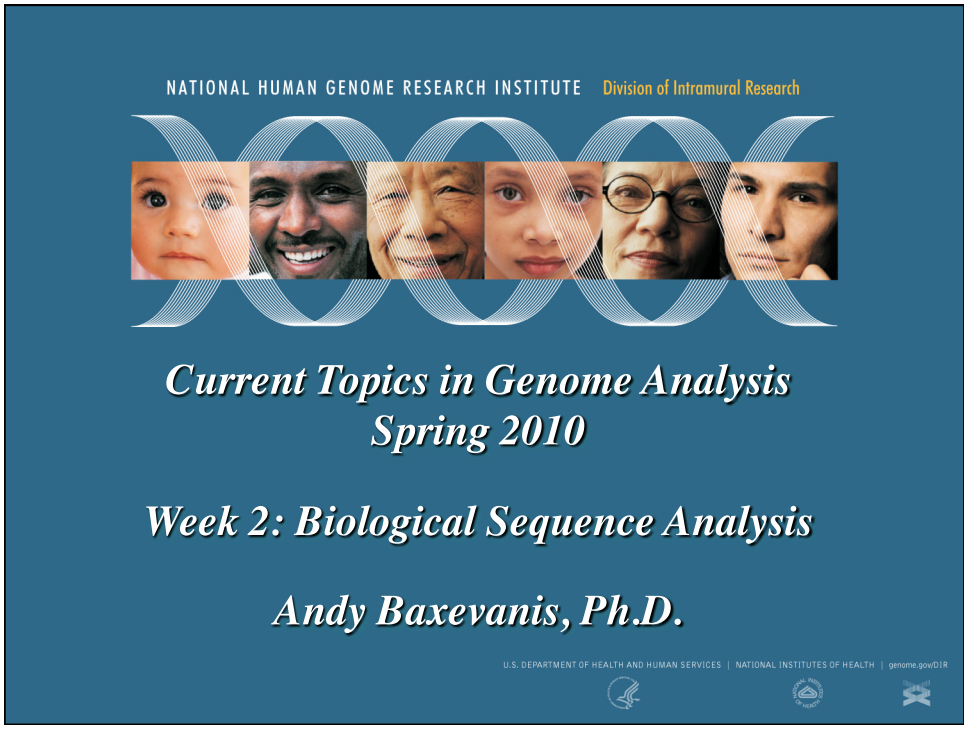
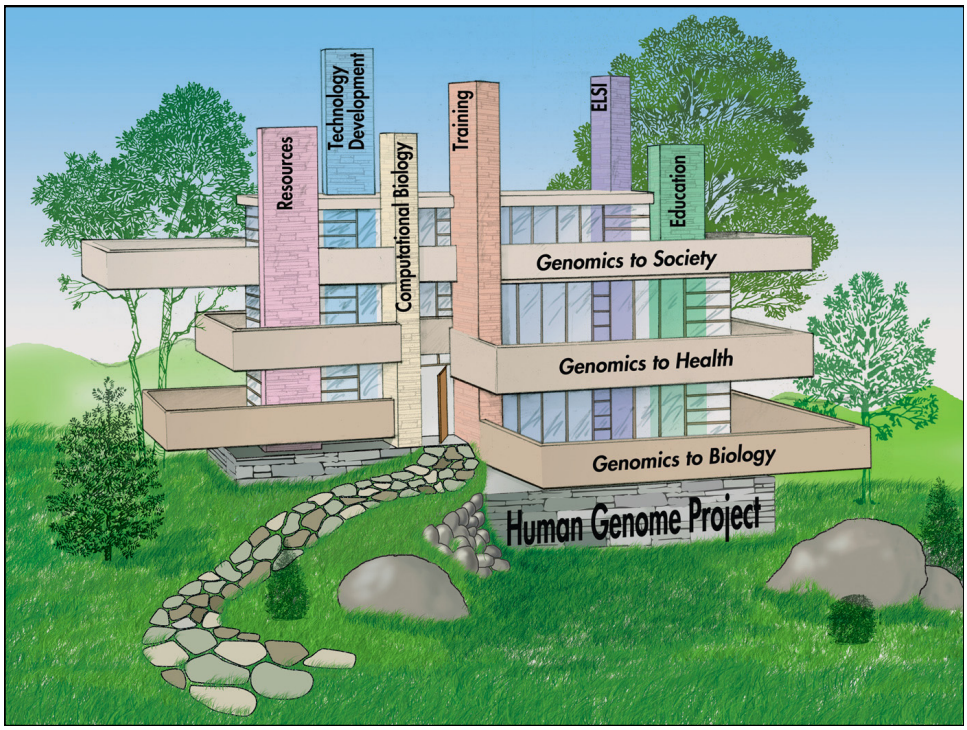

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research
Current Topics in Genome Analysis
Spring 2010
Week 2: Biological Sequence Analysis
Andy Baxevanis, Ph.D.
U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES | NATIONAL INSTITUTES OF HEALTH | genome.gov/DIR


Resources
Technology Development
Computational Biology
Training
ELSI
Education
Genomics to Society
Genomics to Health
Genomics to Biology
Human Genome Project

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

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*

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Defining the Terms

- The quantitative measure: ***Similarity***
  - Always based on an observable
  - Usually expressed as percent identity
  - Quantify changes that occur as two sequences diverge (substitutions, insertions, or deletions)
  - Identify residues crucial for maintaining a protein's structure or function
- High degrees of sequence similarity *might* imply
  - a common evolutionary history
  - possible commonality in biological function

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Defining the Terms

- The conclusion: ***Homology***
  - 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 indivisible[8]. You either are homologous (pregnant) or you are not. Thus, if what one means to assert is that 80% of the character states are identical one should speak of 80% identity, and not 80% homology.

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

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

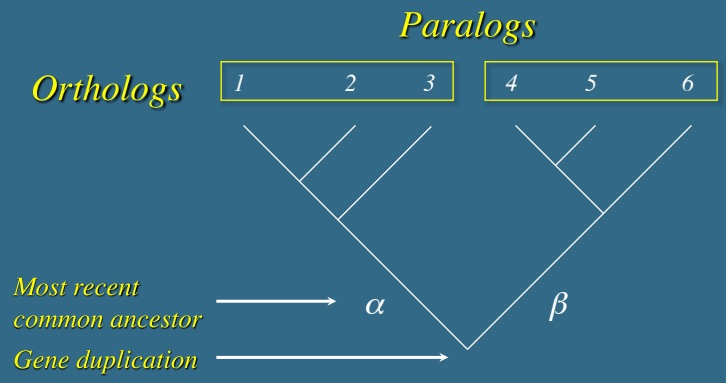
Defining the Terms

- 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



- 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)

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

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

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

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

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

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

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Scoring Matrices

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

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Scoring Matrices

- Why is understanding scoring matrices important?
  - Appear in all analyses involving sequence comparison
  - Implicitly represent particular evolutionary patterns
  - Choice of matrix can strongly influence outcomes of analyses

Matrix Structure: Nucleotides

- *Simple match/mismatch scoring scheme:*

  Match +2
  Mismatch – 3

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

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

Matrix Structure: Proteins

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

BLOSUM62

BLOSUM Matrices

- 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)*

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

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

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

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Which one to choose?

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

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

So many matrices...

No single matrix is the complete answer for all sequence comparisons

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

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 4.4). 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, as understanding the assumptions underlying the PAM and BLOSUM scoring matrices can aid in making the proper choice. The selection of PAM matrices is covered first, after which the selection of BLOSUM matrices is discussed, and finally a brief overview of the wide variety of specialized scoring matrices is provided.

PAM MATRICES

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

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

Biologically, the PAM 50 matrix means that in 100 amino acids there have been 50 substitutions, while the PAM 250 matrix means there have been 2.5 amino acid replacements at each site (see UNIT 5.1 regarding insertions and deletions). This sounds unusual, but remember that over evolutionary time, it is possible that an alanine was changed to a glycine, then to a valine, and then back to an alanine. These silent substitutions are derived from observed amino acid frequency data in protein families and superfamilies.

Choosing a PAM Matrix

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

Locating all potential similarities: PAM 250

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

Finding Similarities and Inferring Homologies

3.5.1

Contributed by David Wheeler
Current Protocols in Bioinformatics (2003) 3.5.1-3.5.6
Copyright © 2003 by John Wiley & Sons, Inc.

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

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	$G > L$			

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

BLAST

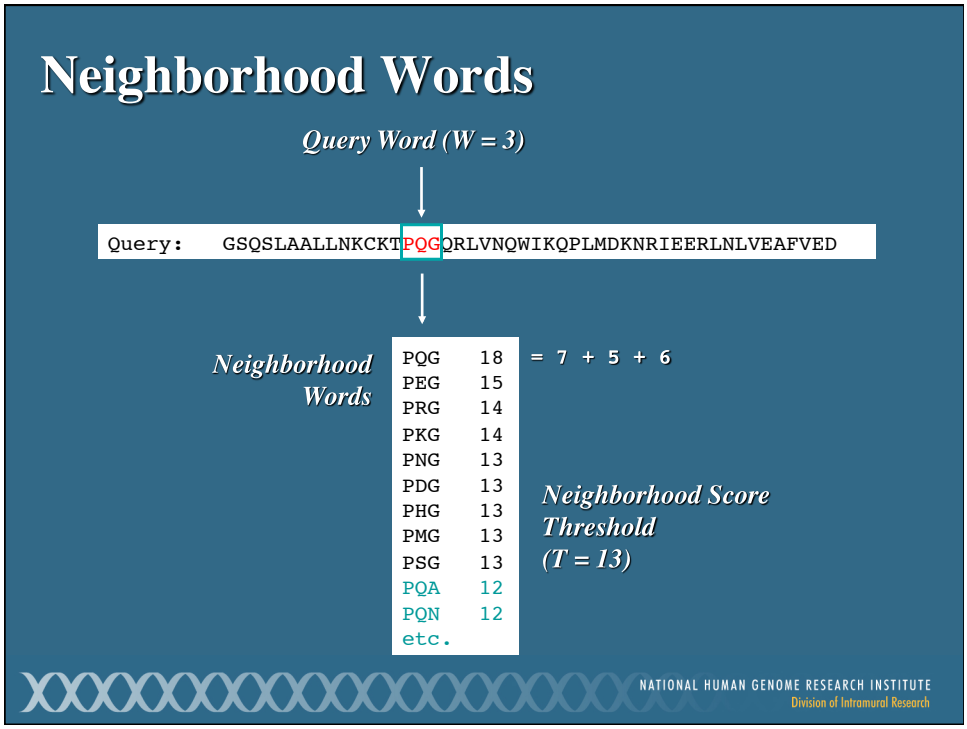
- Basic Local Alignment Search Tool
- Seeks high-scoring segment pairs (HSP)
  - pair of sequences that can be aligned with one another
  - when aligned, have maximal aggregate score (score cannot be improved by extension or trimming)
  - score must be above score threshhold *S*
  - gapped or ungapped
- Results not limited to the "best HSP" for any given sequence pair

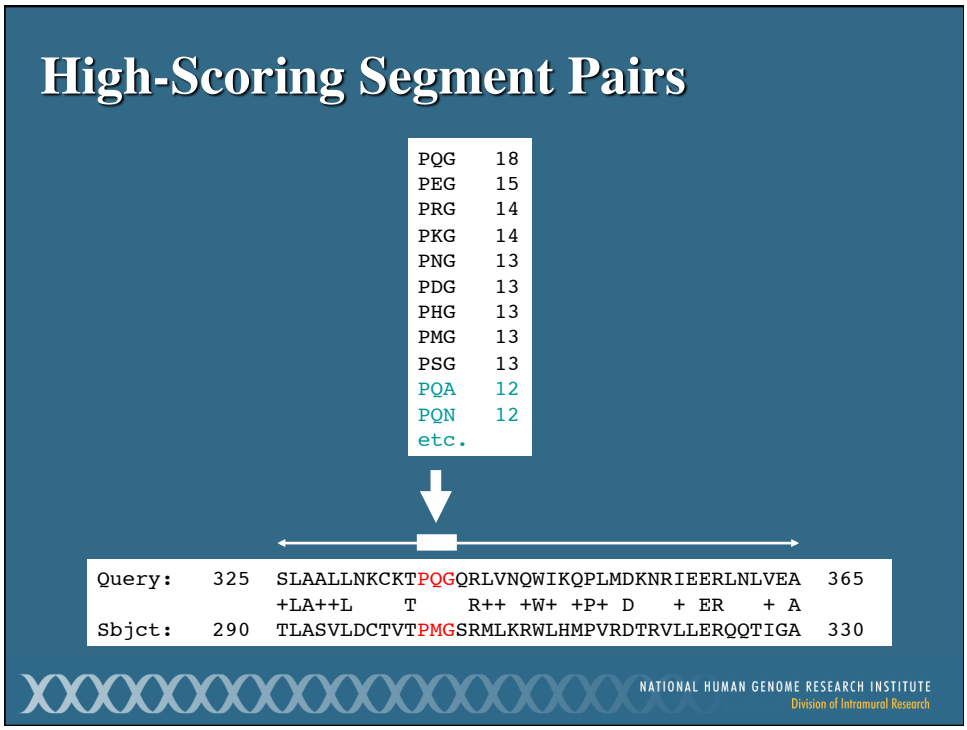
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

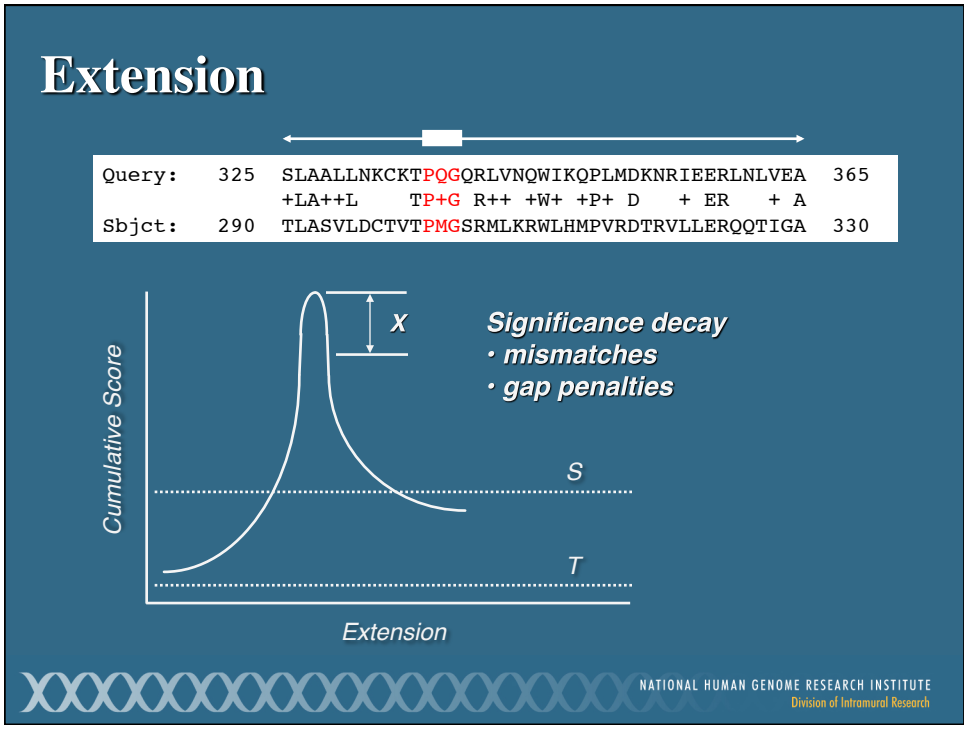
BLAST Algorithms

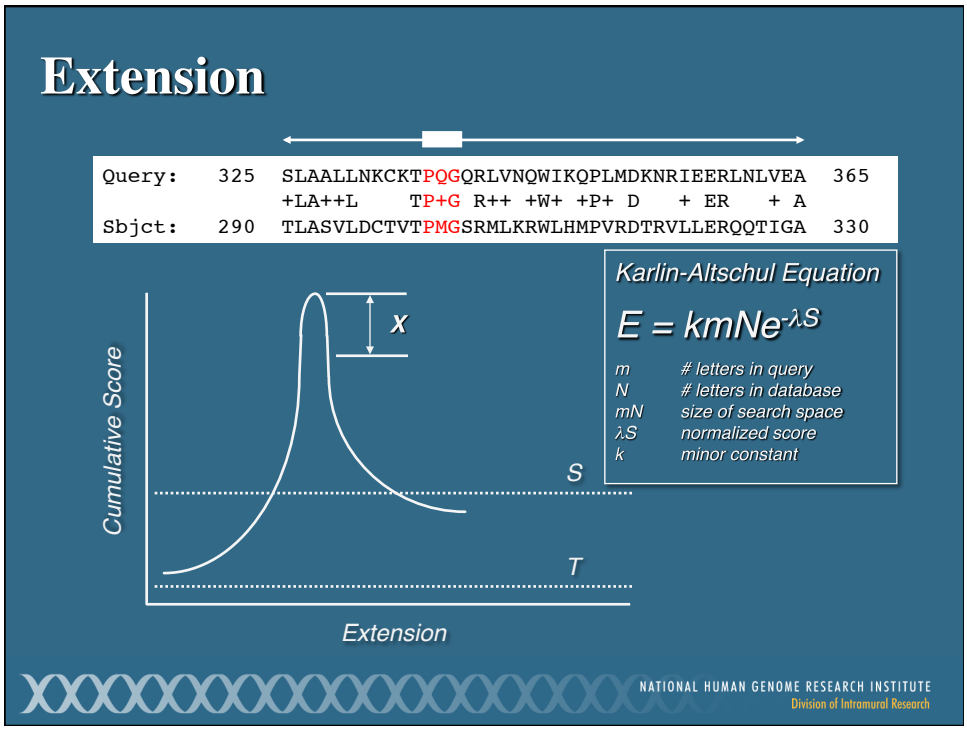
Program	*Query Sequence*	*Target Sequence*
BLASTN	Nucleotide	Nucleotide
BLASTP	Protein	Protein
BLASTX	Nucleotide, six-frame translation	Protein
TBLASTN	Protein	Nucleotide, six-frame translation
TBLASTX	Nucleotide, six-frame translation	Nucleotide, six-frame translation

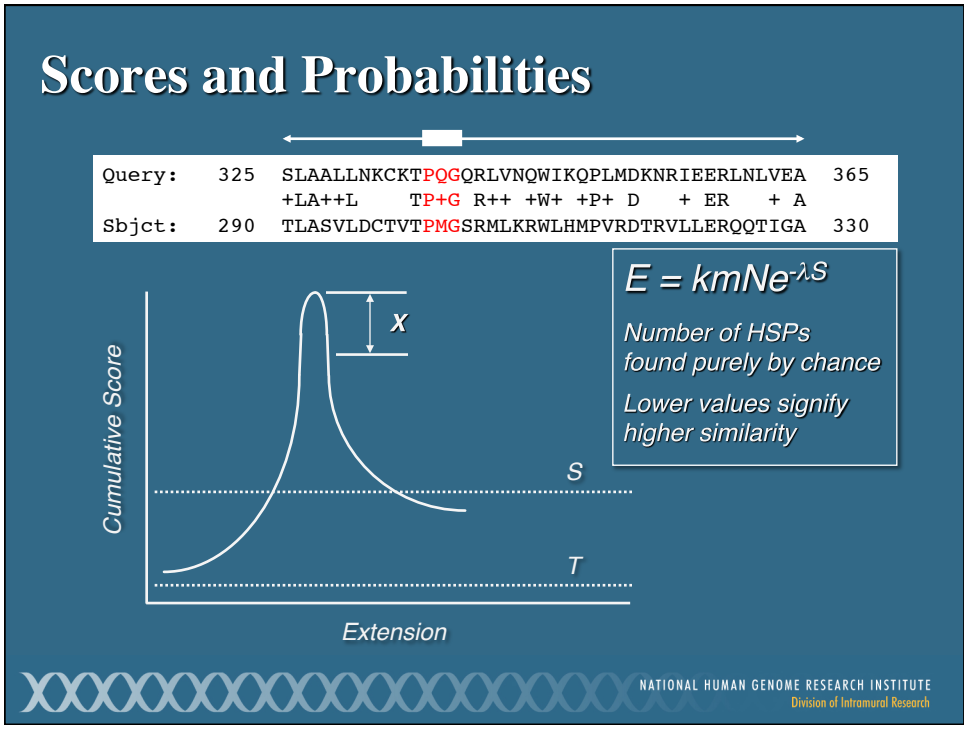
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

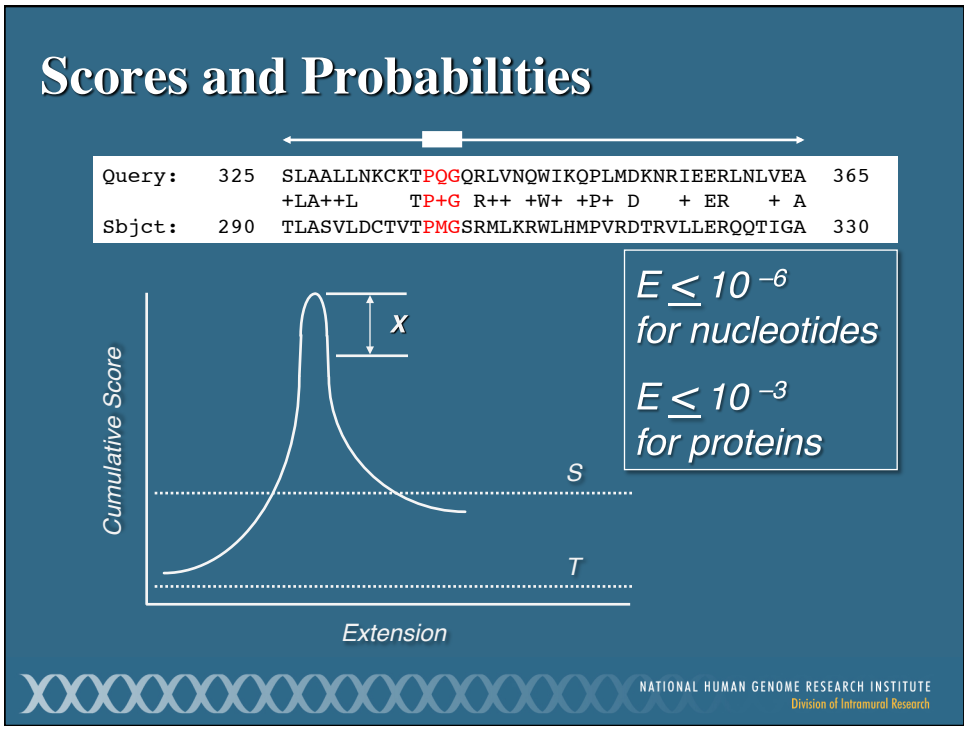

Neighborhood Words
Query Word (W = 3)
Query: GSQSLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEAFVED
Neighborhood Words
PQG 18
PEG 15
PRG 14
PKG 14
PNG 13
PDG 13
PHG 13
PMG 13
PSG 13
PQA 12
PQN 12
etc.
= 7 + 5 + 6
Neighborhood Score Threshold (T = 13)
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural 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 T R++ +W+ +P+ D + ER + A
Sbjct: 290 TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA 330
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research


Extension
Query: 325 SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA 365
+LA++L TP+G R++ +W+ +P+ D + ER + A
Sbjct: 290 TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA 330
X
Significance decay
• mismatches
• gap penalties
Cumulative Score
S
T
Extension
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

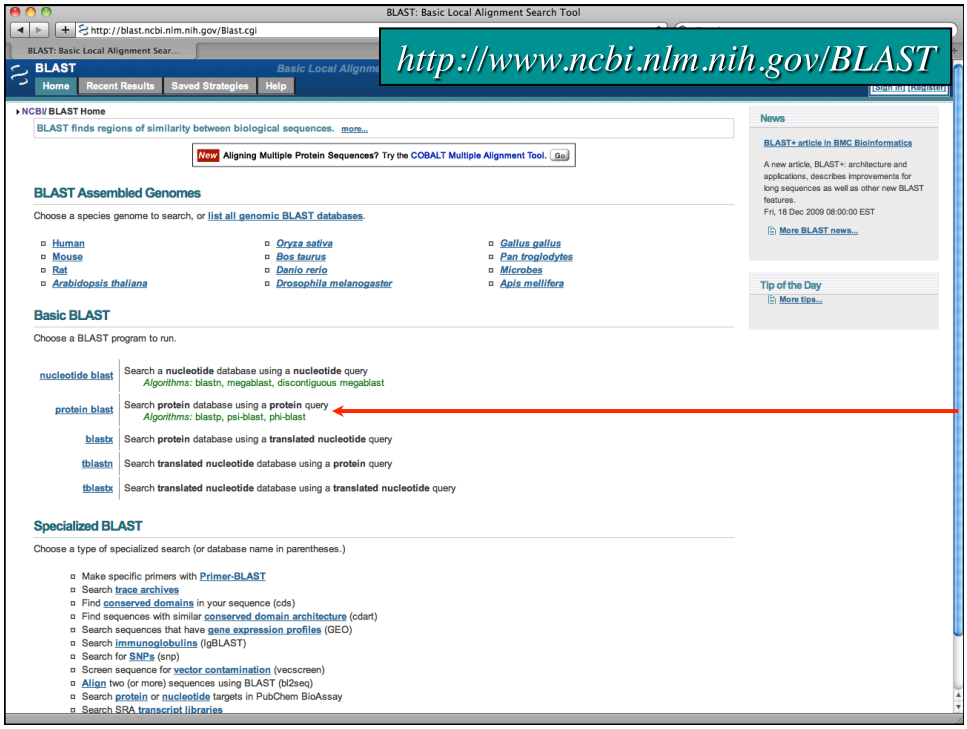

Extension
Query: 325 SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA 365
+LA++L TP+G R++ +W+ +P+ D + ER + A
Sbjct: 290 TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA 330
X
Karlin-Altschul Equation
E = kmNe-λS
m # letters in query
N # letters in database
mN size of search space
λS normalized score
k minor constant
Cumulative Score
S
T
Extension
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research


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


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

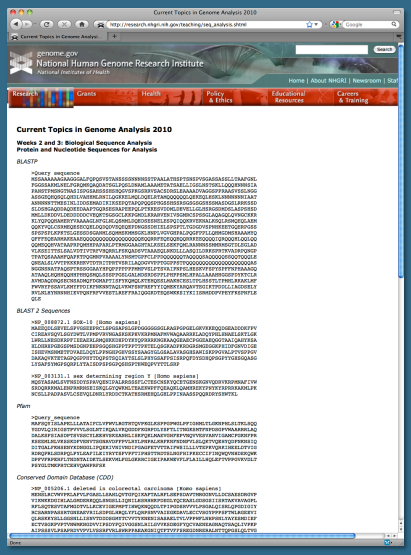
Protein BLAST
BLASTP

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

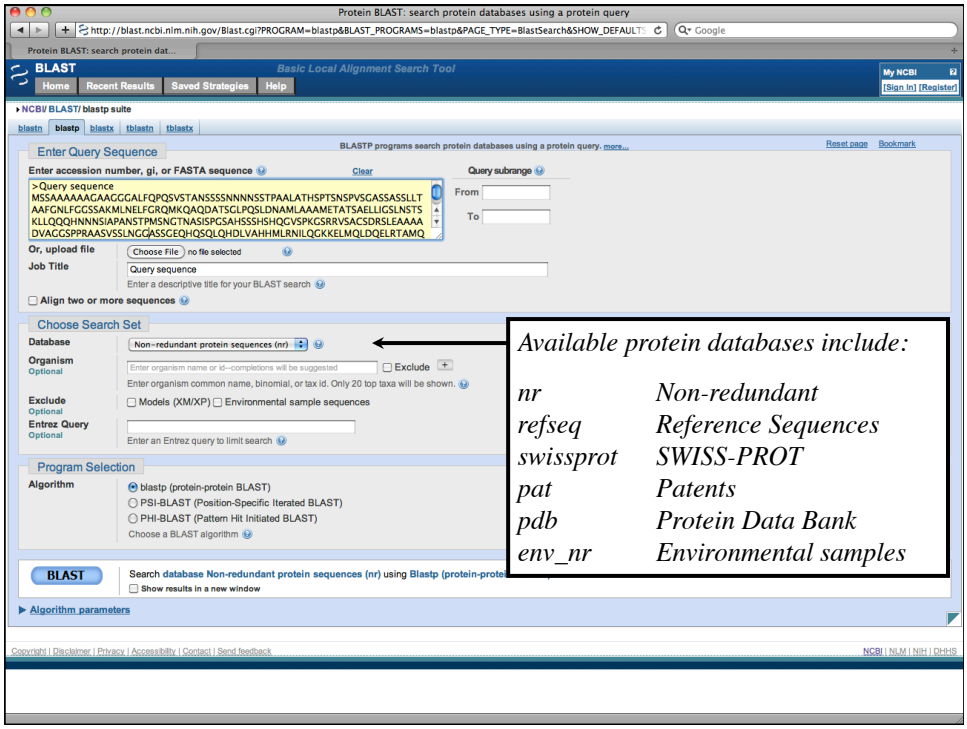


Sequences Used in Examples

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



NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



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

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

RefSeq Accesion Format

From curation of GenBank entries:

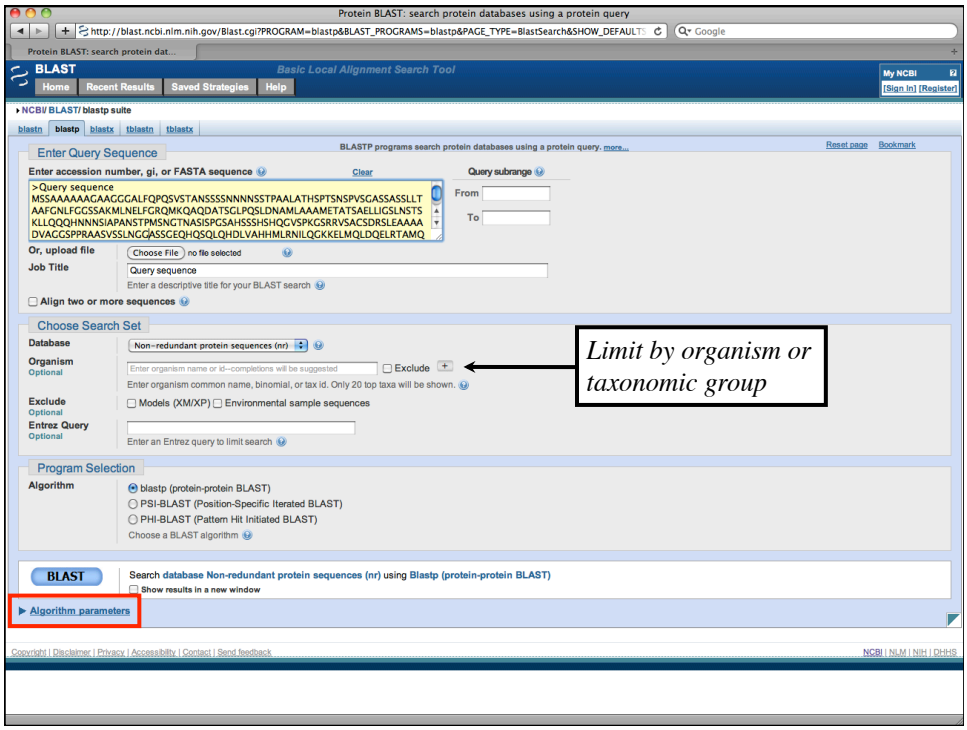
NT_123456	Genomic contigs
NM_123456	mRNAs
NP_123456	Proteins

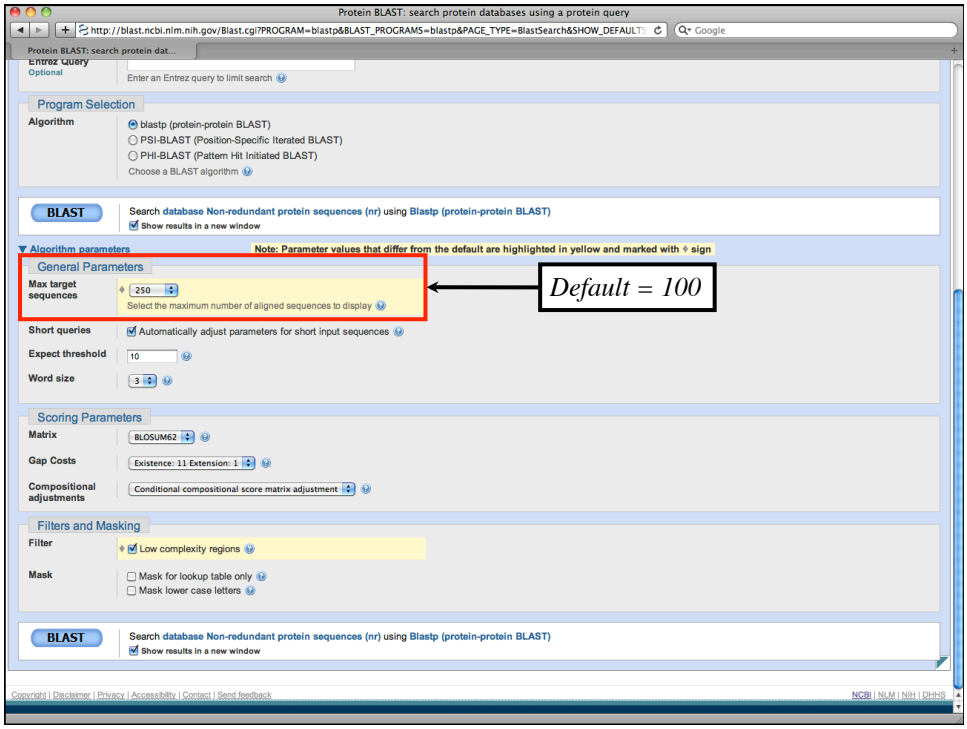
From genome annotation:

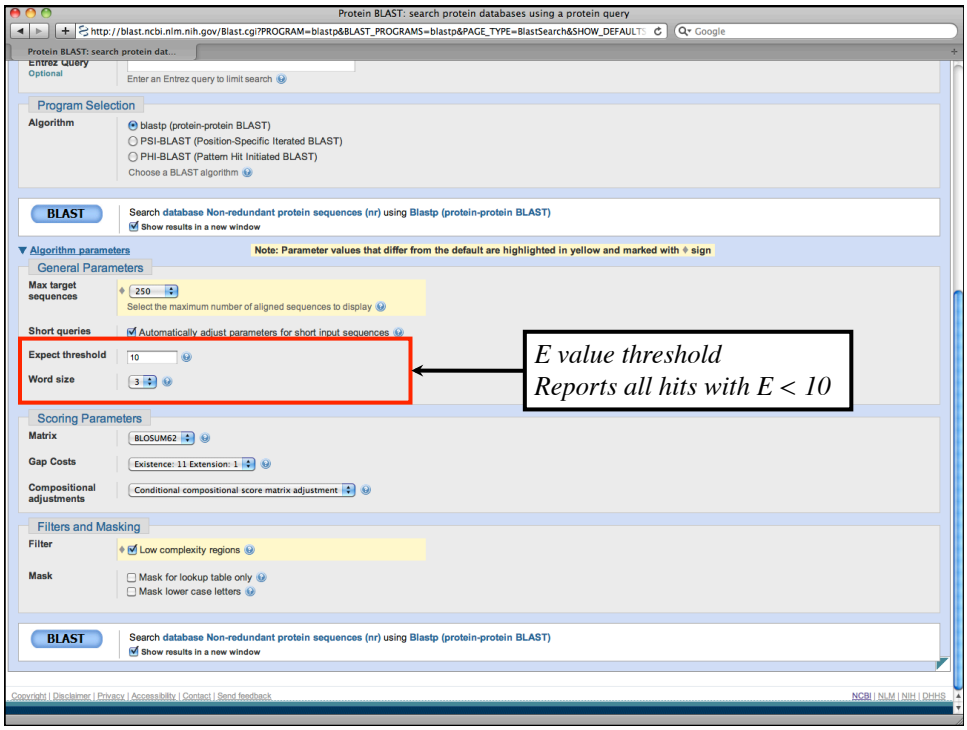
XM_123456	Model mRNA
XP_123456	Model proteins

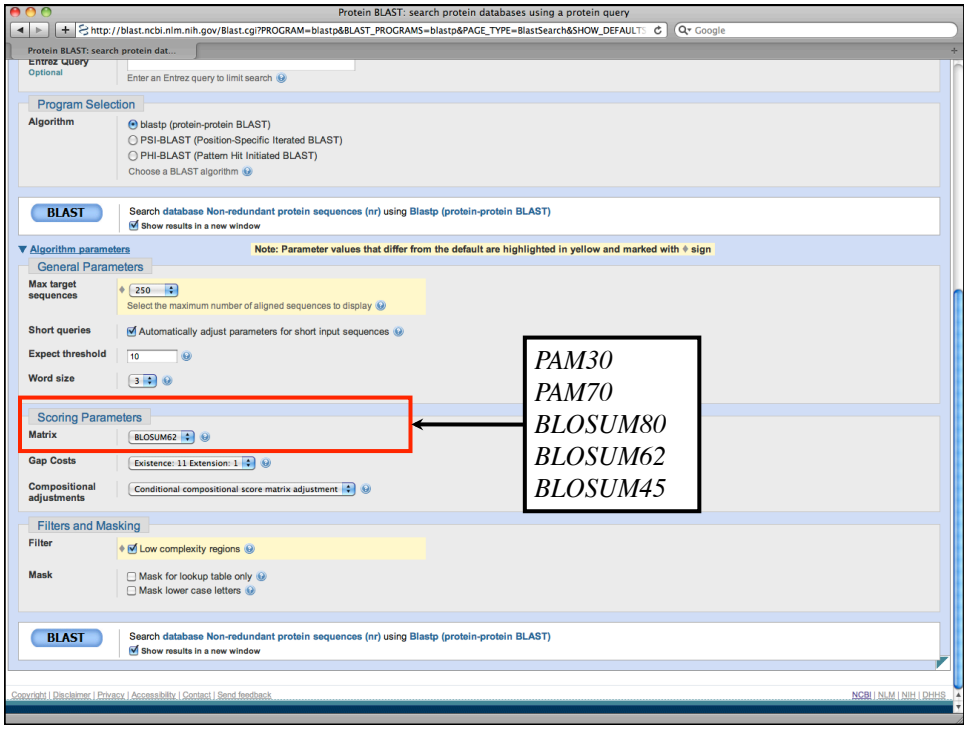
Complete key at
http://www.ncbi.nlm.nih.gov/RefSeq/key.html

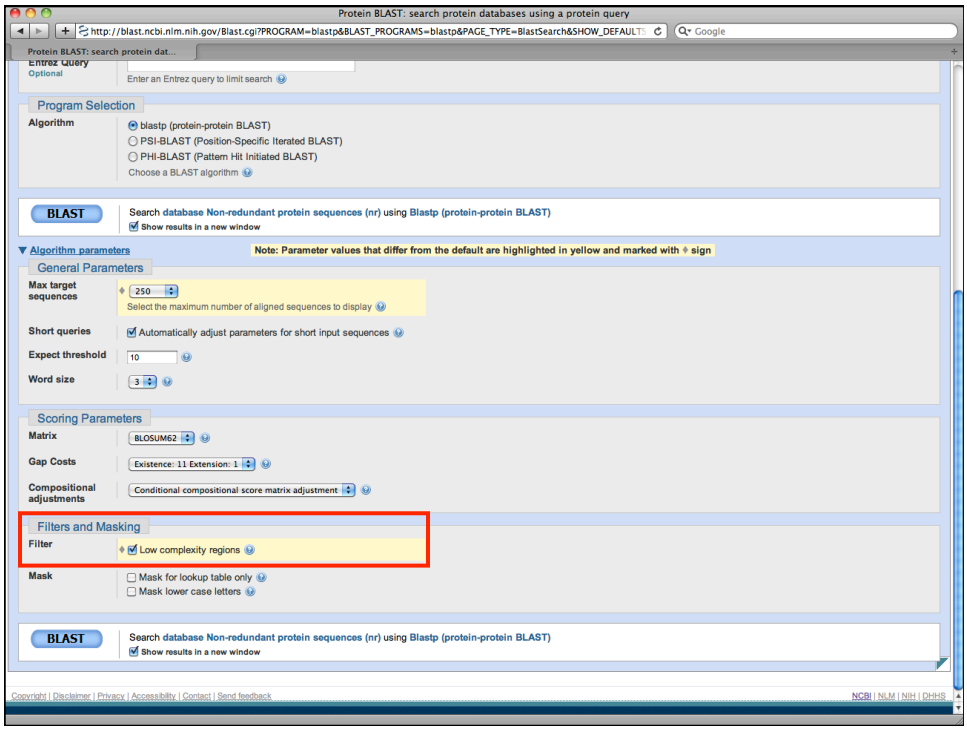
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research




Protein BLAST: search protein databases using a protein query
Program Selection
Algorithm
blastp (protein-protein BLAST)
PSI-BLAST (Position-Specific Iterated BLAST)
PHI-BLAST (Pattern Hit Initiated BLAST)
BLAST
Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)
Show results in a new window
Algorithm parameters
Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign
General Parameters
Max target sequences
250
Select the maximum number of aligned sequences to display
Default = 100
Short queries
Automatically adjust parameters for short input sequences
Expect threshold
10
Word size
3
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


Protein BLAST: search protein databases using a protein query
Program Selection
Algorithm
blastp (protein-protein BLAST)
PSI-BLAST (Position-Specific Iterated BLAST)
PHI-BLAST (Pattern Hit Initiated BLAST)
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
E value threshold
Reports all hits with E < 10
Scoring Parameters
Matrix
BLOSUM62
Gap Costs
Existence: 11 Extension: 1
Compositional adjustments
Conditional compositional score matrix adjustment
Filters and Masking
Filter
Low complexity regions
Mask
Mask for lookup table only
Mask lower case letters


Protein BLAST: search protein databases using a protein query
Program Selection
Algorithm
blastp (protein-protein BLAST)
PSI-BLAST (Position-Specific Iterated BLAST)
PHI-BLAST (Pattern Hit Initiated BLAST)
BLAST
Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)
Algorithm parameters
General Parameters
Scoring Parameters
Matrix
BLOSUM62
Gap Costs
Compositional adjustments
Filters and Masking
Filter
Low complexity regions
Mask
PAM30
PAM70
BLOSUM80
BLOSUM62
BLOSUM45


Protein BLAST: search protein databases using a protein query
Program Selection
Algorithm
blastp (protein-protein BLAST)
PSI-BLAST (Position-Specific Iterated BLAST)
PHI-BLAST (Pattern Hit Initiated BLAST)
BLAST
Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)
Algorithm parameters
General Parameters
Scoring Parameters
Matrix
BLOSUM62
Gap Costs
Compositional adjustments
Filters and Masking
Filter
Low complexity regions
Mask

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)
MESSAKMESGGAGQQPQPQPQQPFLPPAACFFATAAAAAAAAAAAAAQSAQQQQQQQQQQQQQAPQLRPAA
DGQPSGGGHKSAPKQVKRQRSSSPELMRCKRRLNFSGFGYSLPQQQLAAVARRNERERNRVKLVNLGFAT
LREHVPNGAANKKMSKVETLRSAVEYIRALQQLLDEHDAVSAAFQAGVLSPTISPNYSNDLNSMAGSPVS
SYSSDEGSYDPLSPEEQELLDFTNWF
```

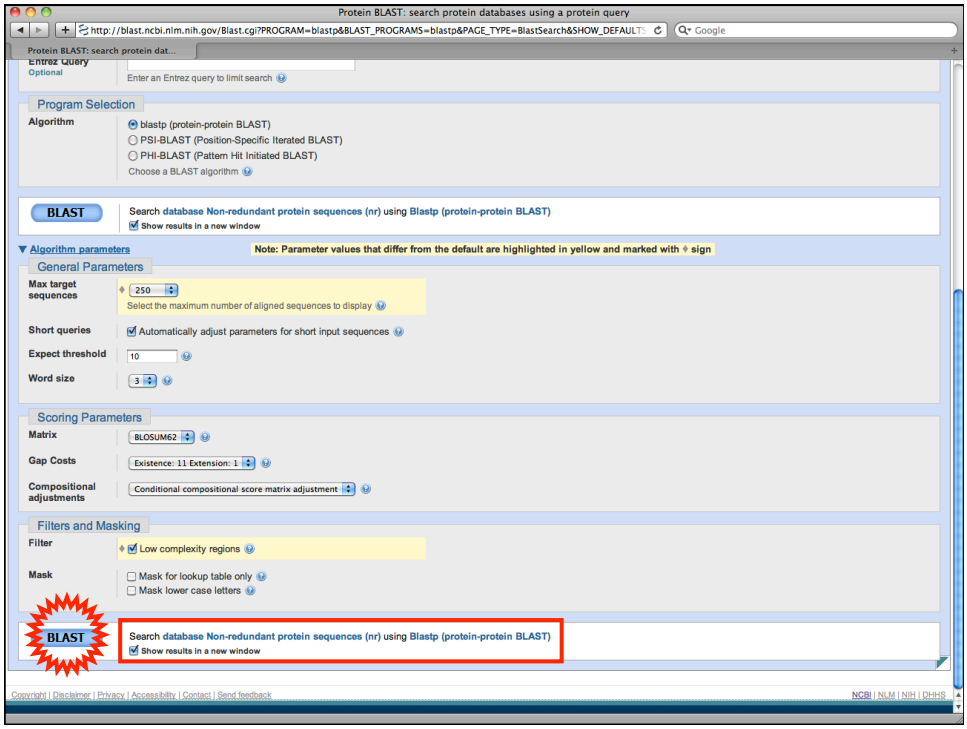
Homopolymeric alanine-glutamine tract

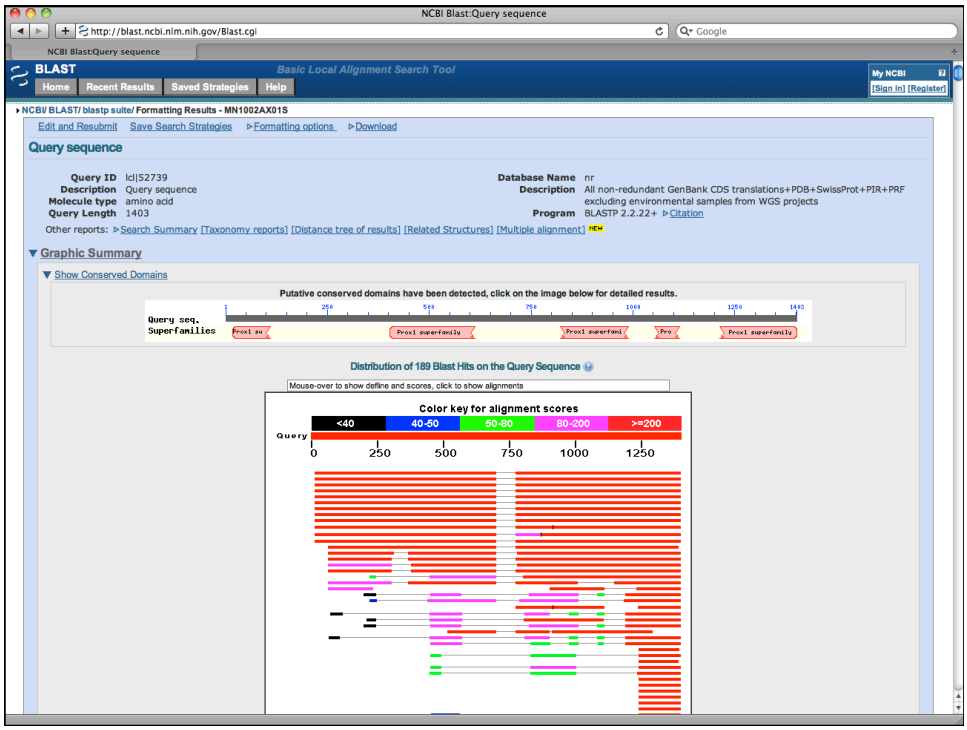
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Identifying Low-Complexity Regions

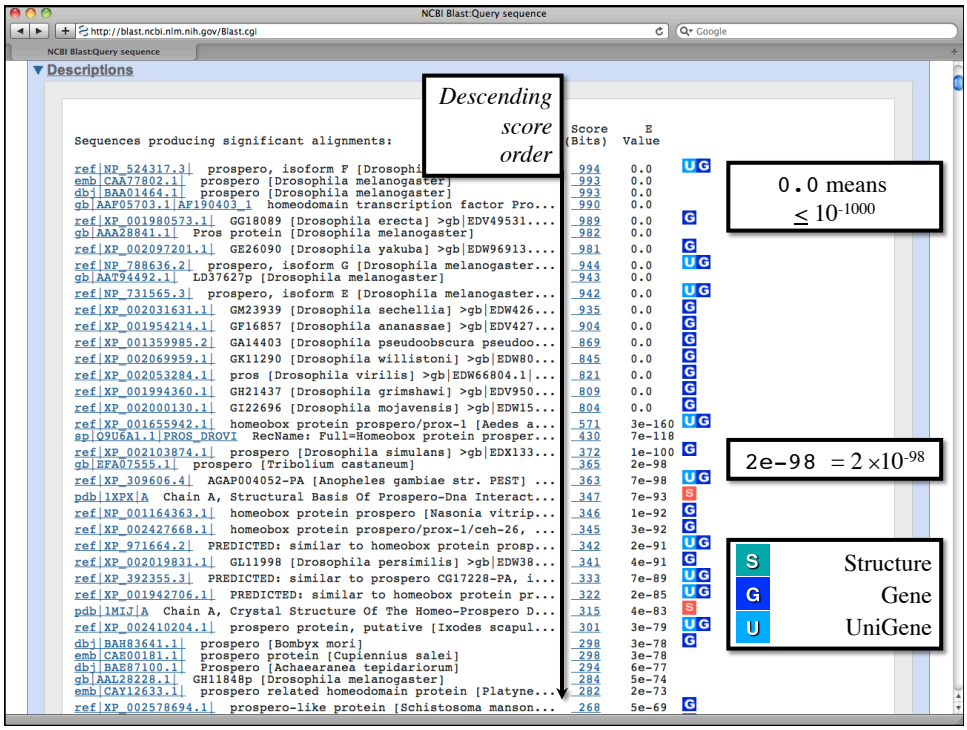
- Biological origins and role not well-understood
  - DNA replication errors (polymerase slippage)?
  - Unequal crossing-over?
- May confound sequence analysis
  - BLAST relies on uniformly-distributed amino acid frequencies
  - Often lead to false positives
  - Filtering is advised (but *not* enabled by default)

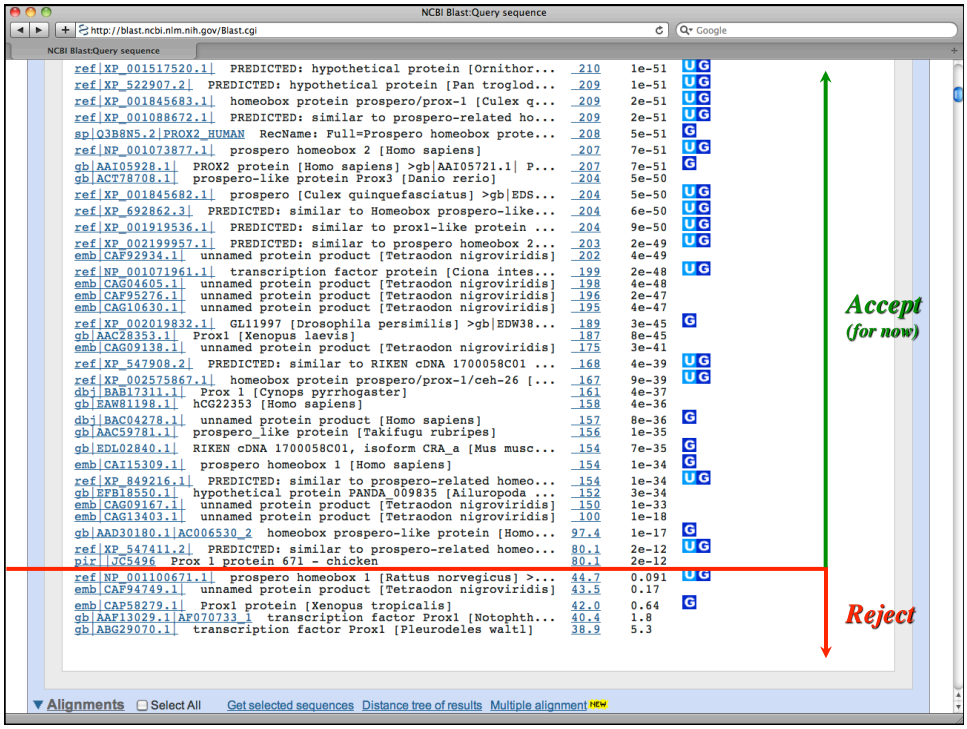
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research


Protein BLAST: search protein databases using a protein query
Program Selection
Algorithm
blastp (protein-protein BLAST)
PSI-BLAST (Position-Specific Iterated BLAST)
PHI-BLAST (Pattern Hit Initiated BLAST)
BLAST
Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)
Show results in a new window
General Parameters
Max target sequences
Short queries
Expect threshold
Word size
Scoring Parameters
Matrix
Gap Costs
Compositional adjustments
Filters and Masking
Filter
Mask
Mask for lookup table only
Mask lower case letters

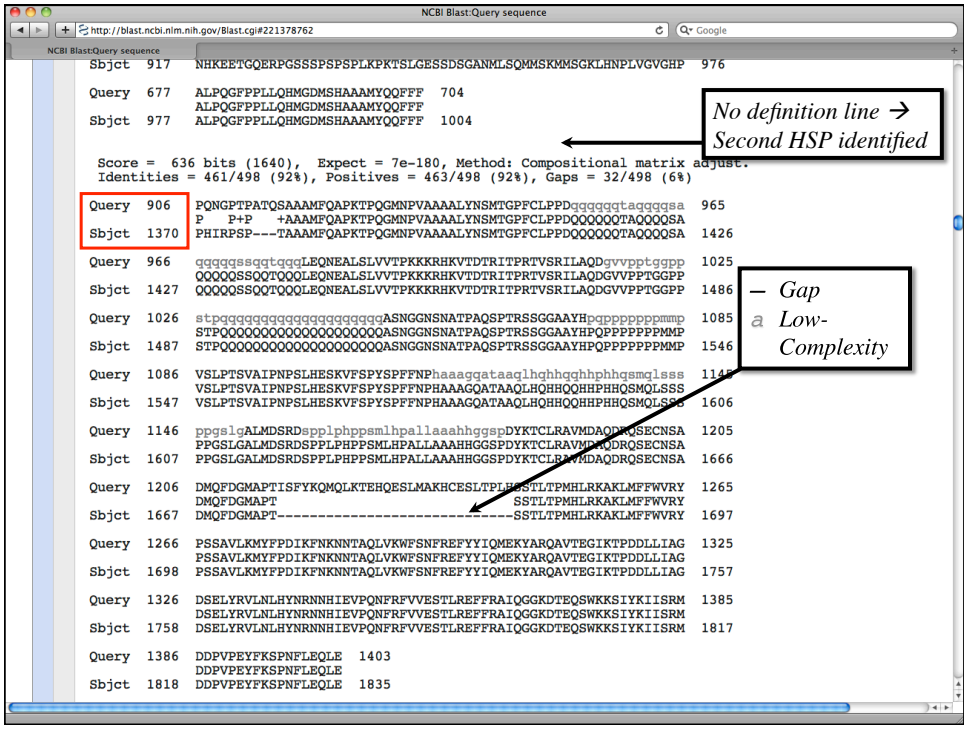

NCBI Blast:Query sequence
BLAST
Basic Local Alignment Search Tool
Query sequence
Query ID
Description
Molecule type
Query Length
Database Name
Program
Graphic Summary
Show Conserved Domains
Putative conserved domains have been detected, click on the image below for detailed results.
Distribution of 189 Blast Hits on the Query Sequence
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query

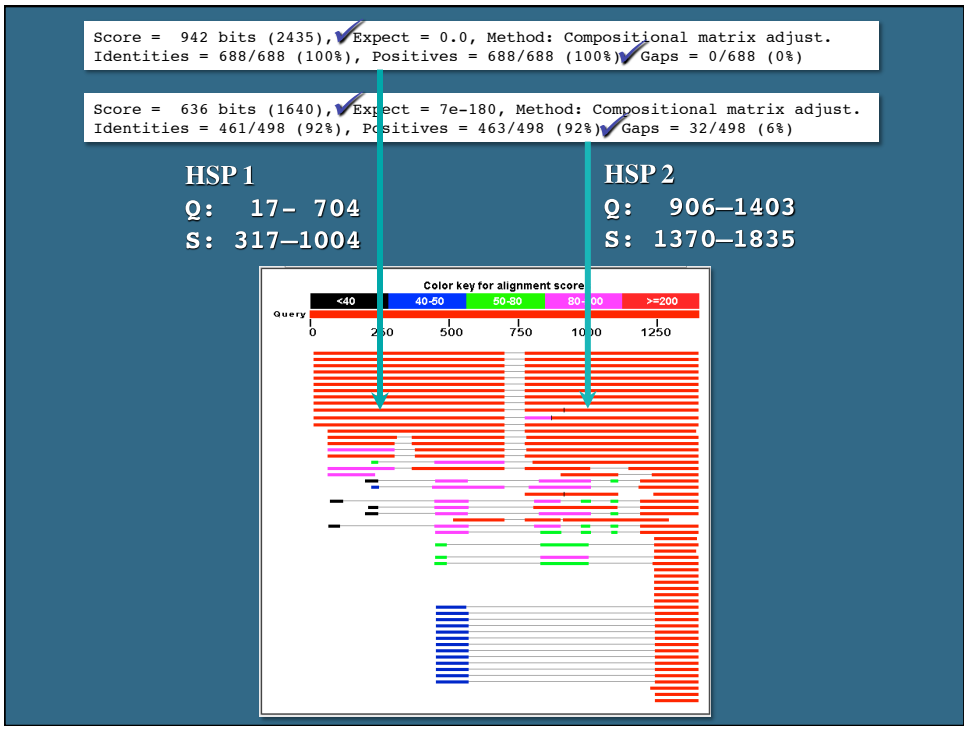

NCBI Blast:Query sequence
Distribution of 189 Blast Hits on the Query Sequence
Mouse-over to show define and scores, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
0
250
500
750
1000
1250
Color key
Unrelated hits
Gap within same hit
• >1 HSP
• Masked region
Descending score order
Descriptions
Sequences producing significant alignments:
Score (Bits)
E Value


NCBI Blast:Query sequence
Descriptions
Descending score order
Sequences producing significant alignments:
Score (Bits)
E Value
0.0 means ≤ 10-1000
2e-98 = 2 ×10-98
S Structure
G Gene
U UniGene


NCBI Blast:Query sequence
http://blast.ncbi.nlm.nih.gov/Blast.cgi
Accept
(for now)
Reject
Alignments
Select All
Get selected sequences
Distance tree of results
Multiple alignment


NCBI Blast:Query sequence
http://blast.ncbi.nlm.nih.gov/Blast.cgi#221378762
ref|NP_731565.3| prospero, isoform E [Drosophila melanogaster]
gb|AAN13501.3| 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%)
≥ 25% for proteins
≥ 70% for nucleotides
– Gap
a Low-Complexity


NCBI Blast:Query sequence
http://blast.ncbi.nlm.nih.gov/Blast.cgi#221378762
Google
NCBI Blast:Query sequence
Sbjct 917 NHKEETGQERPGSSSPSPSPSLKPKTSLGESSDSGANMLSQMMSKMMSGKLHNPLVGVGHP 976
Query 677 ALPQGFPPLLQHMGDMSHAAAMYQQFFF 704
ALPQGFPPLLQHMGDMSHAAAMYQQFFF
Sbjct 977 ALPQGFPPLLQHMGDMSHAAAMYQQFFF 1004
No definition line →
Second HSP identified
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 PQNGPTPATQSAAAMFQAPKTPQGMNPVAAAALYNSMTGPFCLPPDqqqqqqtaqqqqsa 965
P P+P +AAAMFQAPKTPQGMNPVAAAALYNSMTGPFCLPPDQQQQQQTAQQQQSA
Sbjct 1370 PHIRPSP---TAAAMFQAPKTPQGMNPVAAAALYNSMTGPFCLPPDQQQQQQTAQQQQSA 1426
Query 966 qqqqqssqqtqqqLEQNEALSLVVTPKKKRHKVTDTRITPRTVSRILAQDgvvpptggpp 1025
QQQQQSSQQTQQQLEQNEALSLVVTPKKKRHKVTDTRITPRTVSRILAQDGVVPPTGGPP
Sbjct 1427 QQQQQSSQQTQQQLEQNEALSLVVTPKKKRHKVTDTRITPRTVSRILAQDGVVPPTGGPP 1486
Query 1026 stpqqqqqqqqqqqqqqqqqqqqqASNGGNSNATPAQSPTRSSGGAAYHpqpppppppmmp 1085
STPQQQQQQQQQQQQQQQQQQQQASNGGNSNATPAQSPTRSSGGAAYHPQPPPPPPPMMP
Sbjct 1487 STPQQQQQQQQQQQQQQQQQQQQASNGGNSNATPAQSPTRSSGGAAYHPQPPPPPPPMMP 1546
Query 1086 VSLPTSVAIPNPSLHESKVFSPYSPFFNPhaaaggataaqlhqhhqqhhphhqsmqlsss 1145
VSLPTSVAIPNPSLHESKVFSPYSPFFNPHAAAGQATAAQLHQHHQQHHPHHQSMQLSSS
Sbjct 1547 VSLPTSVAIPNPSLHESKVFSPYSPFFNPHAAAGQATAAQLHQHHQQHHPHHQSMQLSSS 1606
Query 1146 ppgslgALMDSRDspplphppsmlhpallaaahhggspDYKTCLRAVMDAQDRQSECNSA 1205
PPGSLGALMDSRDSPPLPHPPSMLHPALLAAAHHGGSPDYKTCLRAVMDAQDRQSECNSA
Sbjct 1607 PPGSLGALMDSRDSPPLPHPPSMLHPALLAAAHHGGSPDYKTCLRAVMDAQDRQSECNSA 1666
Query 1206 DMQFDGMAPTISFYKQMQLKTEHQESLMAKHCESLTPLHSSTLTPMHLRKAKLMFFWVRY 1265
DMQFDGMAPT SSTLTPMHLRKAKLMFFWVRY
Sbjct 1667 DMQFDGMAPT----------SSTLTPMHLRKAKLMFFWVRY 1697
Query 1266 PSSAVLKMYFPDIKFNKNNTAQLVKWFSNFREFYYIQMEKYARQAVTEGIKTPDDLLIAG 1325
PSSAVLKMYFPDIKFNKNNTAQLVKWFSNFREFYYIQMEKYARQAVTEGIKTPDDLLIAG
Sbjct 1698 PSSAVLKMYFPDIKFNKNNTAQLVKWFSNFREFYYIQMEKYARQAVTEGIKTPDDLLIAG 1757
Query 1326 DSELYRVLNLHYNRNNHIEVPQNFRFVVESTLREFFRAIQGGKDTEQSWKKSIYKIISRM 1385
DSELYRVLNLHYNRNNHIEVPQNFRFVVESTLREFFRAIQGGKDTEQSWKKSIYKIISRM
Sbjct 1758 DSELYRVLNLHYNRNNHIEVPQNFRFVVESTLREFFRAIQGGKDTEQSWKKSIYKIISRM 1817
Query 1386 DDPVPEYFKSPNFLEQLE 1403
DDPVPEYFKSPNFLEQLE
Sbjct 1818 DDPVPEYFKSPNFLEQLE 1835
– 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
40-50
50-80
80-200
>=200
Query
0
250
500
750
1000
1250

Suggested BLAST Cutoffs

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

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

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Database Searching Artifacts

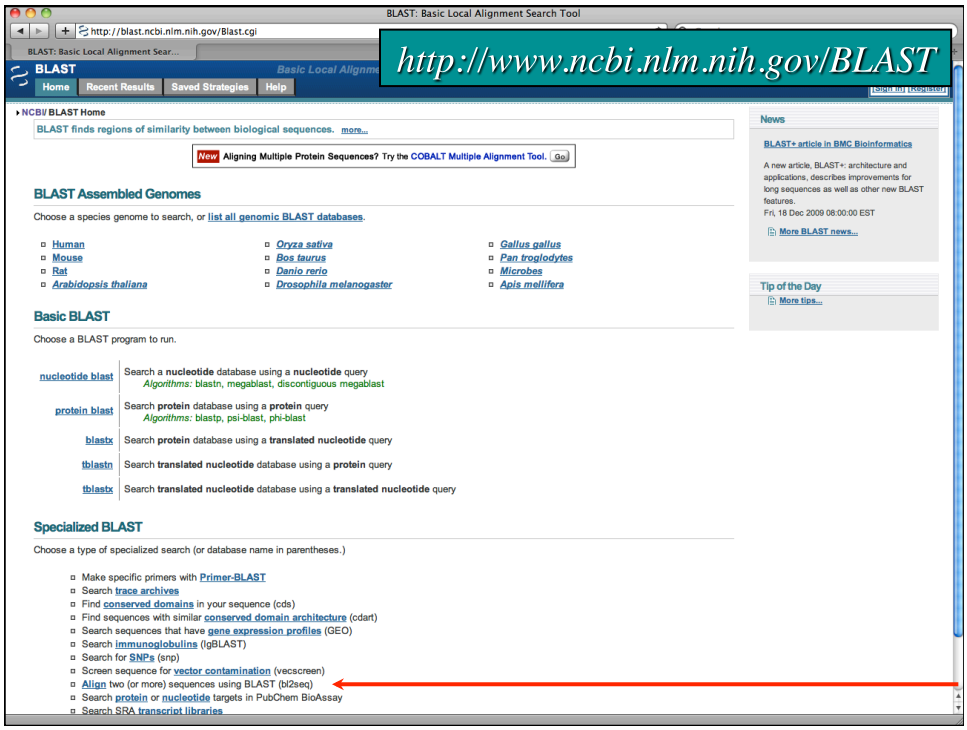
- Low-complexity regions
- Repetitive elements
  - LINEs, SINEs, retroviral repeats
  - Choose "Filter: Species-Specific Repeats" when using BLASTN
  - RepeatMasker
    http://www.repeatmasker.org
- Low-quality sequence hits
  - Expressed sequence tags (ESTs)
  - Single-pass sequence reads from large-scale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

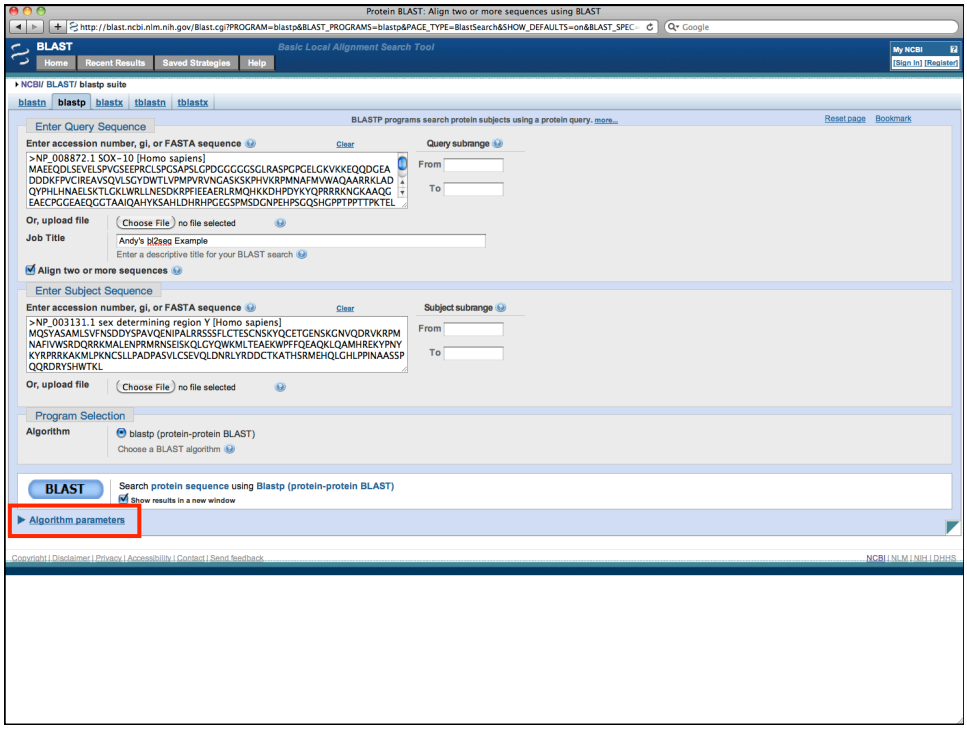
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

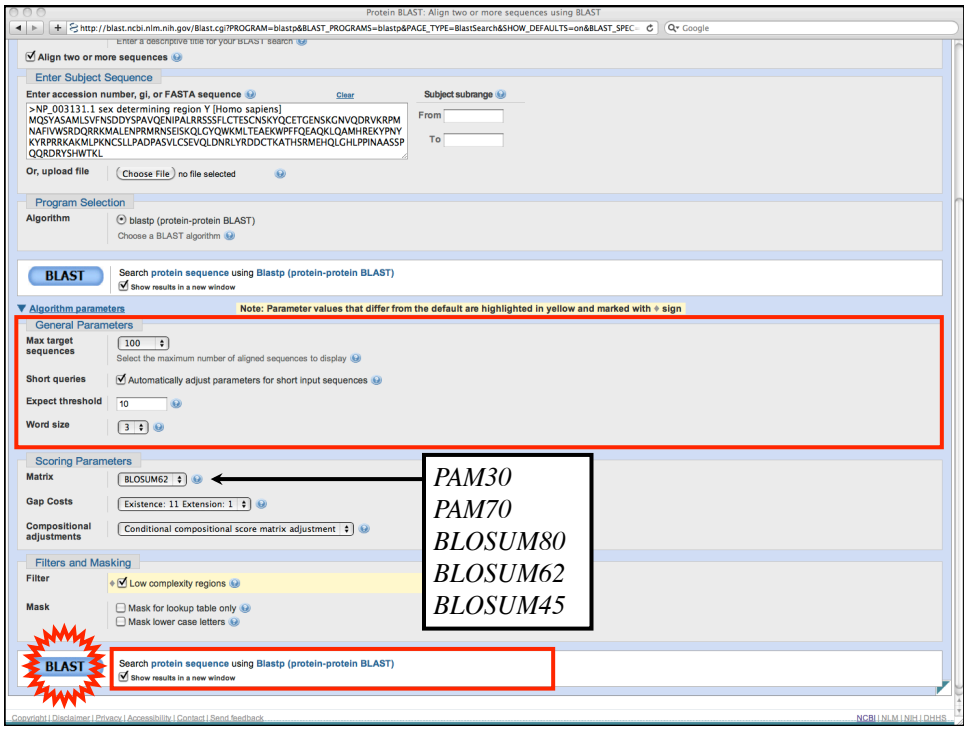
BLAST 2 Sequences

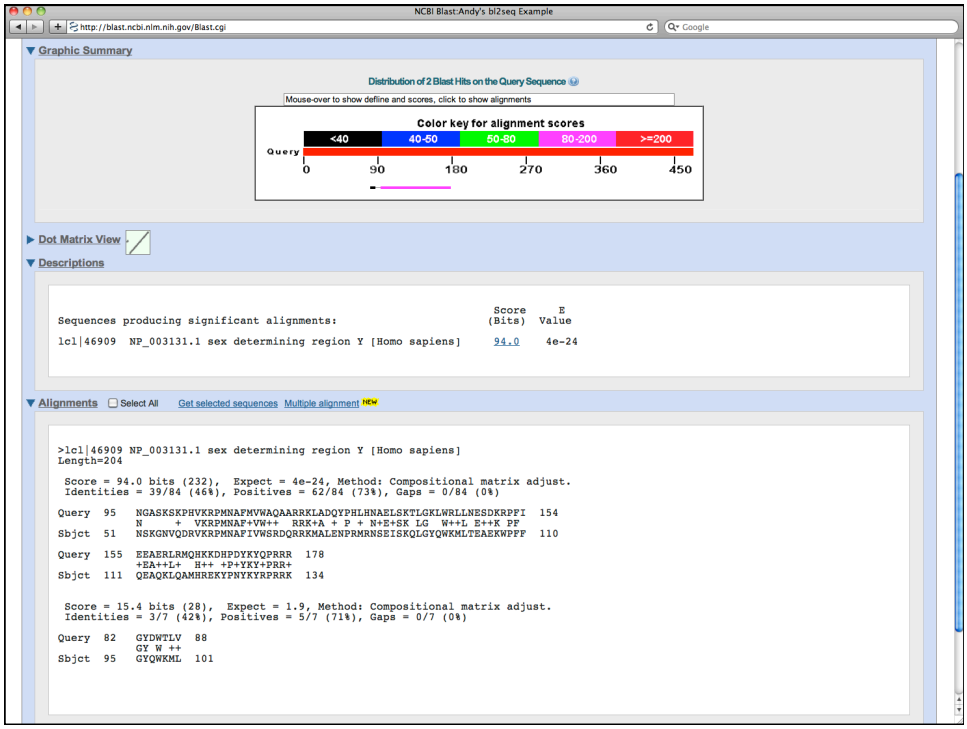
- Finds local alignments between two protein or nucleotide sequences of interest
  - All BLAST programs available
  - Select BLOSUM and PAM matrices available for protein comparisons
  - Same affine gap costs (adjustable)
  - Input sequences can be masked

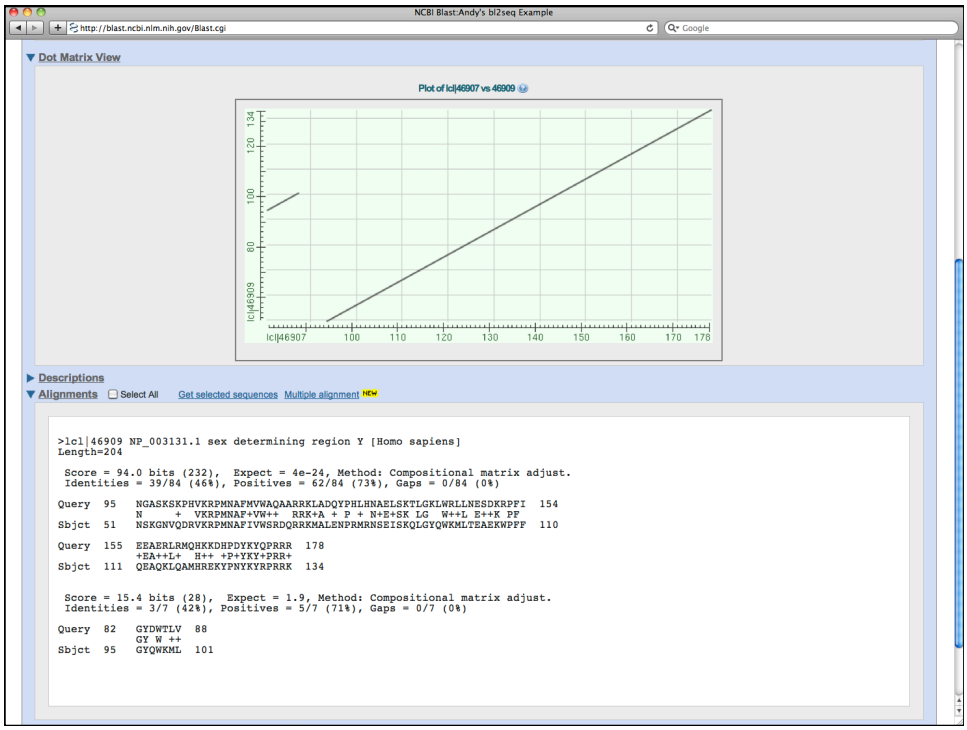
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research




Protein BLAST: Align two or more sequences using BLAST
BLAST
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
blastn
blastp
blastx
tblastn
tblastx
Enter Query Sequence
Enter accession number, gi, or FASTA sequence
Query subrange
Or, upload file
Job Title
Align two or more sequences
Enter Subject Sequence
Subject subrange
Program Selection
Algorithm
blastp (protein-protein BLAST)
BLAST
Search protein sequence using Blastp (protein-protein BLAST)
Algorithm parameters

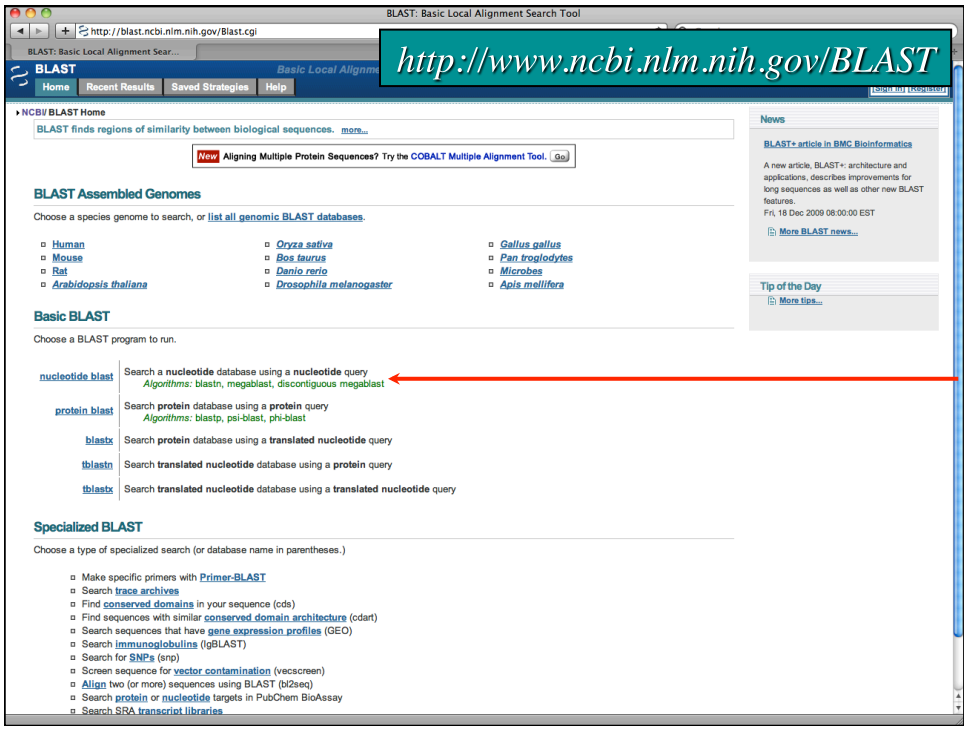

Algorithm parameters
Note: Parameter values that differ from the default are highlighted in yellow and marked with + sign
General Parameters
Max target sequences
Short queries
Expect threshold
Word size
Scoring Parameters
Matrix
Gap Costs
Compositional adjustments
Filters and Masking
Filter
Mask
BLAST
Search protein sequence using Blastp (protein-protein BLAST)
PAM30
PAM70
BLOSUM80
BLOSUM62
BLOSUM45

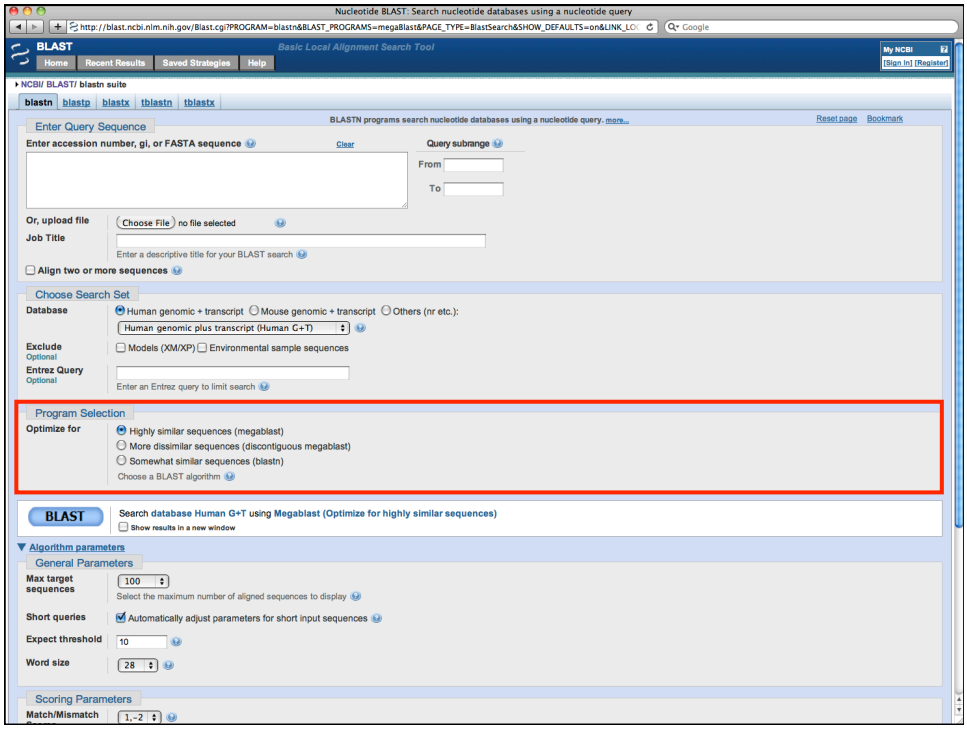

NCBI Blast:Andy's bl2seq Example
http://blast.ncbi.nlm.nih.gov/Blast.cgi
Google
Graphic Summary
Distribution of 2 Blast Hits on the Query Sequence
Mouse-over to show defline and scores, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
0
90
180
270
360
450
Dot Matrix View
Descriptions
Sequences producing significant alignments: Score (Bits) E Value
lcl|46909 NP_003131.1 sex determining region Y [Homo sapiens] 94.0 4e-24
Alignments
Select All
Get selected sequences
Multiple alignment
NEW
>lcl|46909 NP_003131.1 sex determining region Y [Homo sapiens]
Length=204
Score = 94.0 bits (232), Expect = 4e-24, Method: Compositional matrix adjust.
Identities = 39/84 (46%), Positives = 62/84 (73%), Gaps = 0/84 (0%)
Query 95 NGASKSKPHVKRPMNAFMVWAQAARRKLADQYPHLHNAELSKTLGKLWRLLNESDKRPFI 154
N + VKRPMNAF+VW++ RRK+A + P + N+E+SK LG W++L E++K PF
Sbjct 51 NSKGNVQDRVKRPMNAFIVWSRDQRRKMALENPRMRNSEISKQLGYQWKMLTEAEKWPFF 110
Query 155 EEAERLRMQHKKDHPDYKYQPRRR 178
+EA++L+ H++ +P+YKY+PRR+
Sbjct 111 QEAQKLQAMHREKYPNYKYRPRRK 134
Score = 15.4 bits (28), Expect = 1.9, Method: Compositional matrix adjust.
Identities = 3/7 (42%), Positives = 5/7 (71%), Gaps = 0/7 (0%)
Query 82 GYDWTLV 88
GY W ++
Sbjct 95 GYQWKML 101


NCBI Blast:Andy's bl2seq Example
http://blast.ncbi.nlm.nih.gov/Blast.cgi
Google
Dot Matrix View
Plot of lcl|46907 vs 46909
lcl|46909
134
120
100
80
lcl|46907
100
110
120
130
140
150
160
170
178
Descriptions
Alignments
Select All
Get selected sequences
Multiple alignment
NEW
>lcl|46909 NP_003131.1 sex determining region Y [Homo sapiens]
Length=204
Score = 94.0 bits (232), Expect = 4e-24, Method: Compositional matrix adjust.
Identities = 39/84 (46%), Positives = 62/84 (73%), Gaps = 0/84 (0%)
Query 95 NGASKSKPHVKRPMNAFMVWAQAARRKLADQYPHLHNAELSKTLGKLWRLLNESDKRPFI 154
N + VKRPMNAF+VW++ RRK+A + P + N+E+SK LG W++L E++K PF
Sbjct 51 NSKGNVQDRVKRPMNAFIVWSRDQRRKMALENPRMRNSEISKQLGYQWKMLTEAEKWPFF 110
Query 155 EEAERLRMQHKKDHPDYKYQPRRR 178
+EA++L+ H++ +P+YKY+PRR+
Sbjct 111 QEAQKLQAMHREKYPNYKYRPRRK 134
Score = 15.4 bits (28), Expect = 1.9, Method: Compositional matrix adjust.
Identities = 3/7 (42%), Positives = 5/7 (71%), Gaps = 0/7 (0%)
Query 82 GYDWTLV 88
GY W ++
Sbjct 95 GYQWKML 101

Nucleotide BLAST
MegaBLAST and BLASTN

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research





Nucleotide-Based BLAST Algorithms

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

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

BLAT

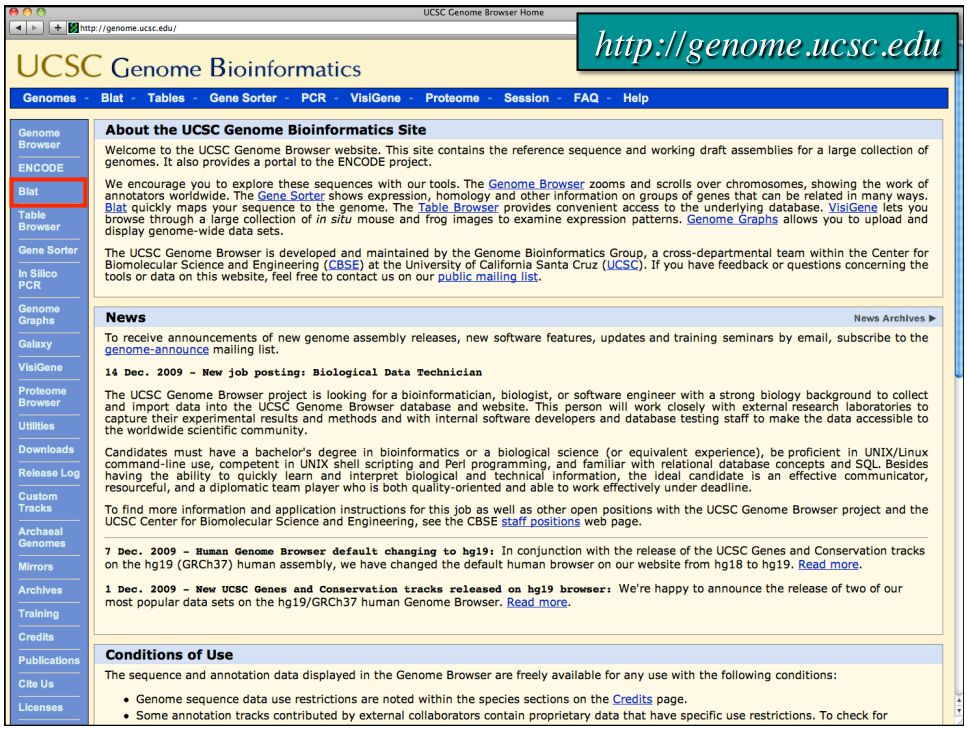
- "BLAST-Like Alignment Tool"
- Designed to rapidly-align longer nucleotide sequences ($L \geq 40$) having > 95% sequence similarity
- Can find exact matches reliably down to $L = 33$
- Method of choice when looking for exact matches in nucleotide databases
- 500 times faster for mRNA/DNA searches
-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

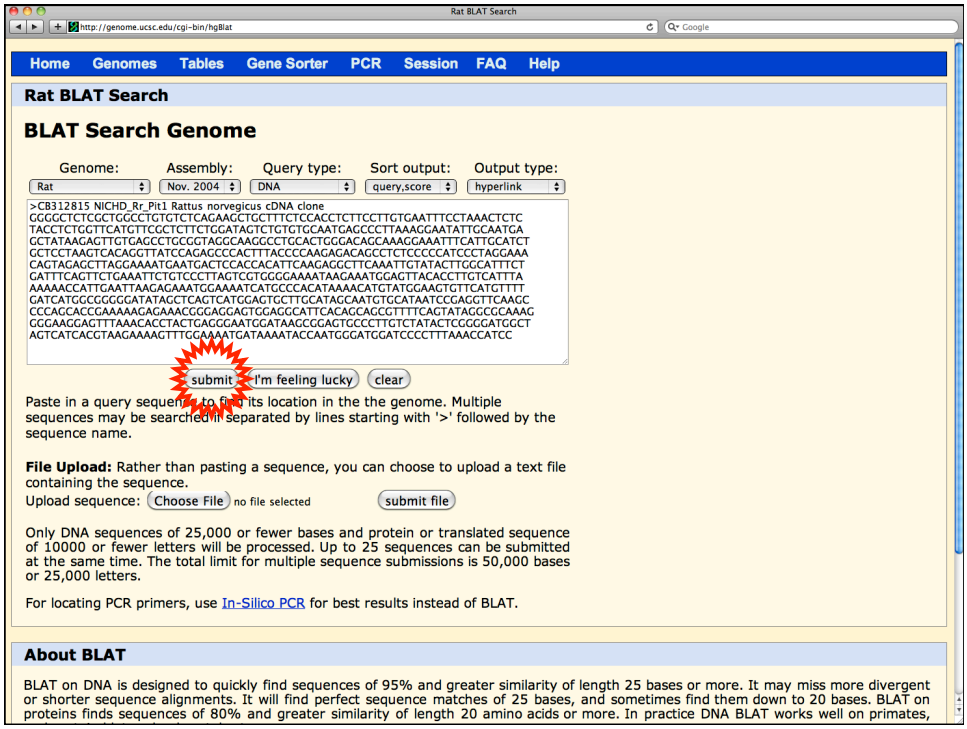
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

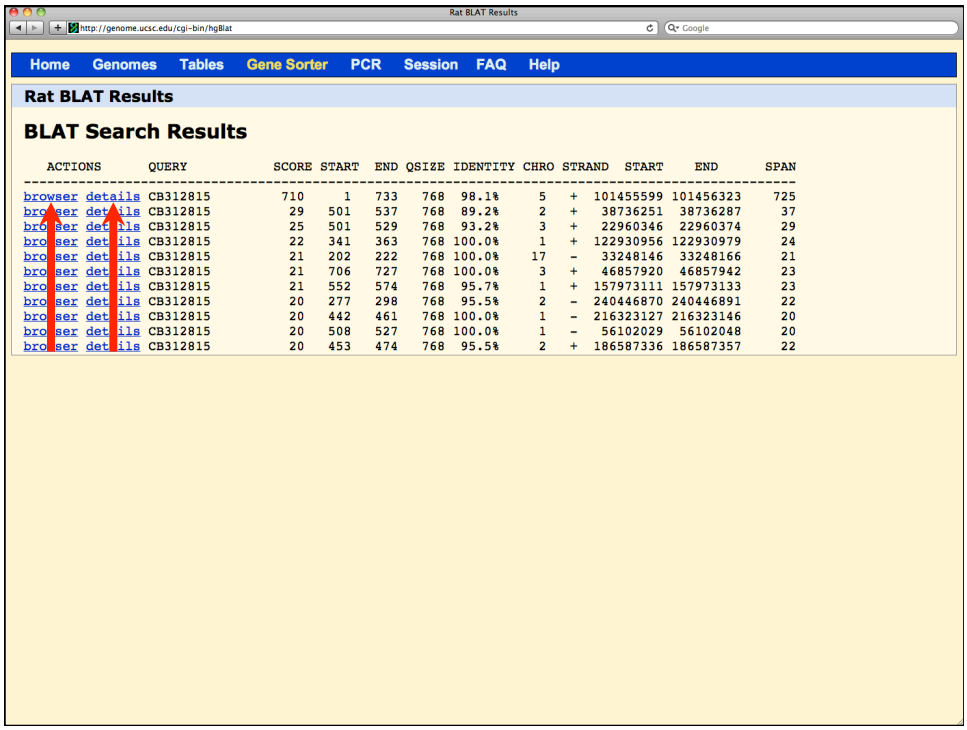
When to Use BLAT

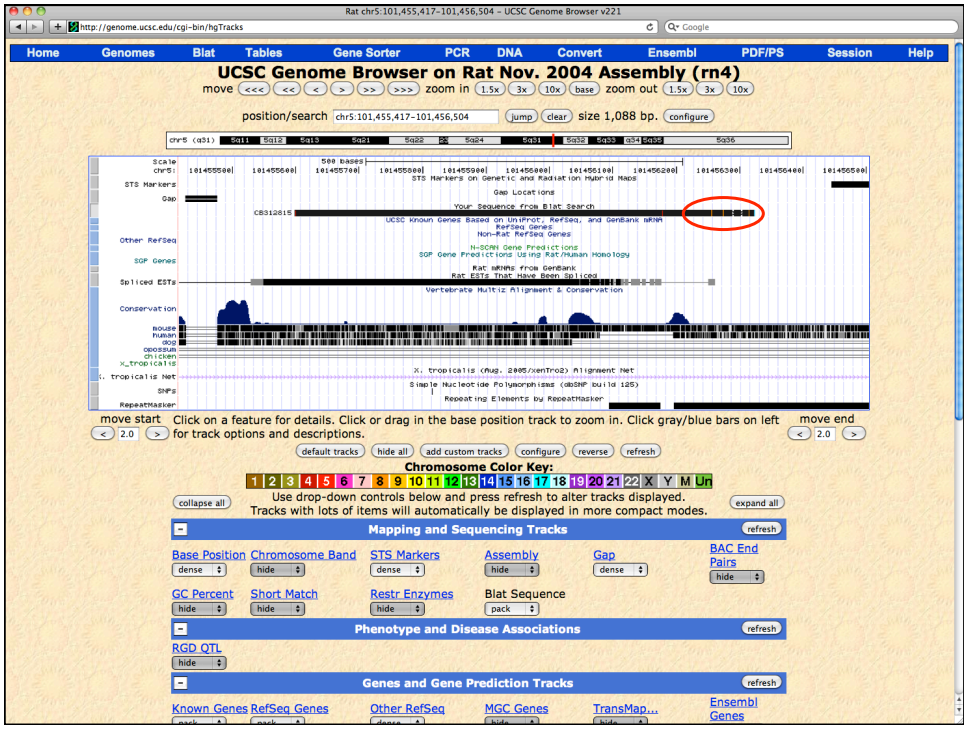
- To characterize an unknown gene or sequence fragment
  - Find its genomic coordinates
  - Determine gene structure (the presence and position of exons)
  - Identify markers of interest in the vicinity of a sequence
- To find highly-similar sequences
  - Identify gene family members
  - Identify putative homologs
- To display a specific sequence as a separate track

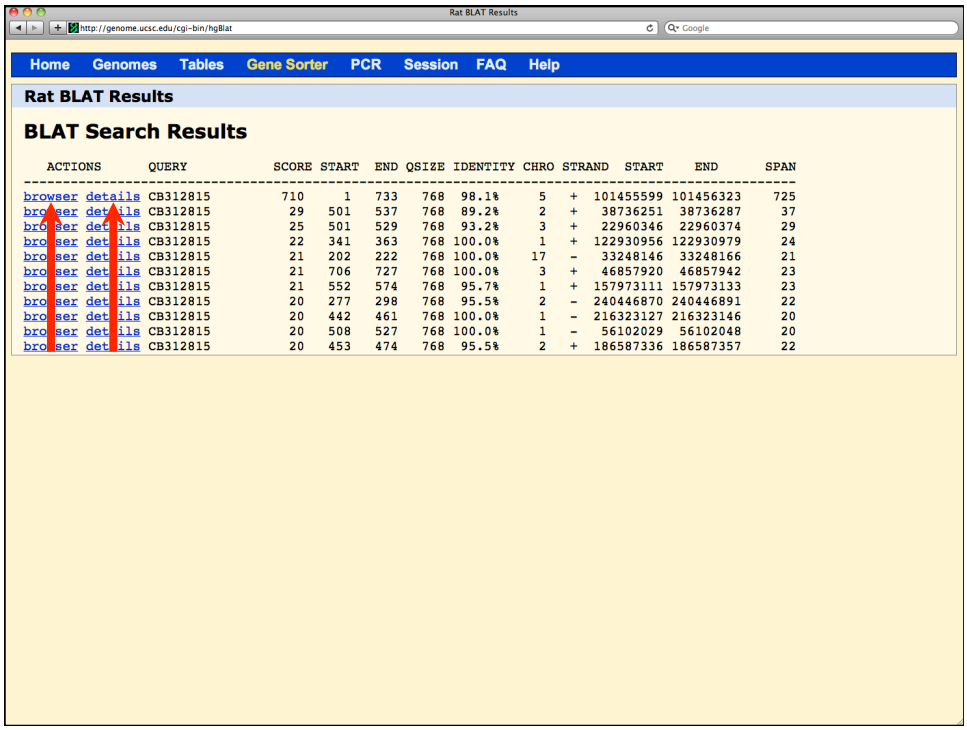
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



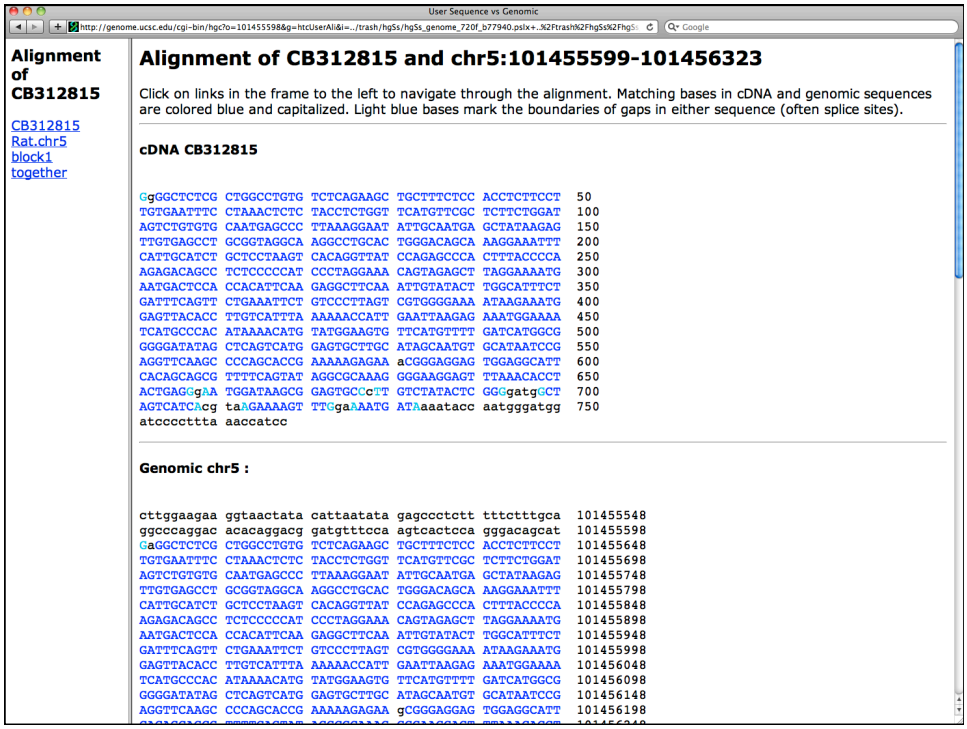

Rat BLAT Search
http://genome.ucsc.edu/cgi-bin/hgBlat
Home Genomes Tables Gene Sorter PCR Session FAQ Help
Rat BLAT Search
BLAT Search Genome
Genome: Assembly: Query type: Sort output: Output type:
Rat Nov. 2004 DNA query,score hyperlink
>CB312815 NICHD_Rr_Pit1 Rattus norvegicus cDNA clone
GGGGCTCTCGCTGGCCTGTGTCTCAGAAGCTGCTTTCTCCACCTCTTCCTTGTGAATTTCCTAAACTCTC
TACCTCTGGTTCATGTTCGCTCTTCTGGATAGTCTCTCTGCAATGAGCCCTTAAAGGAATATTGCAATGA
GCTATAAGAGTTGTGAGCCTGCGGTAGGCAAGGCCTGCACTGGGACAGCAAAGGAAATTTCATTGCATCT
GCTCCTAAGTCACAGGTTATCCAGAGCCCACTTTACCCCAAGAGACAGCCTCTCCCCCATCCCTAGGAAA
CAGTAGAGCTTAGGAAAATGAATGACTCCACCACATTCAAGAGGCTTCAAATTGTATACTTGGCATTTCT
GATTTCAGTTCTGAAATTCTGTCCCTTAGTCGTGGGGAAAATAAGAAATGGAGTTACACCTTGTCATTTA
AAAAACCATTGAATTAAGAGAAATGGAAAATCATGCCCACATAAAACATGTATGGAAGTGTTCATGTTTT
GATCATGGCGGGCGATATAGCTCAGTCATGGAGTGCTTGCATAGCAATGTGCATAATCCGAGGTTCAAGC
CCCAGCACCGAAAAACAGAAACGGAGGAGTGGAGGCATTCACAGCAGCGTTTTCAGTATAGGCCGCAAAG
GGGAAGGAGTTTAAACACCTACTGAGGGAATGGATAAGCCGAGTGCCCTTGTCTATACTCGGGGATGGCT
AGTCATCACGTAAGAAAAGTTTGGAAAATGATAAAATACCAATGGGATGGATCCCCTTTAAACCATCC
submit I'm feeling lucky clear
Paste in a query sequence to find its location in the the genome. Multiple sequences may be searched if separated by lines starting with '>' followed by the sequence name.
File Upload: Rather than pasting a sequence, you can choose to upload a text file containing the sequence.
Upload sequence: Choose File no file selected submit file
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,

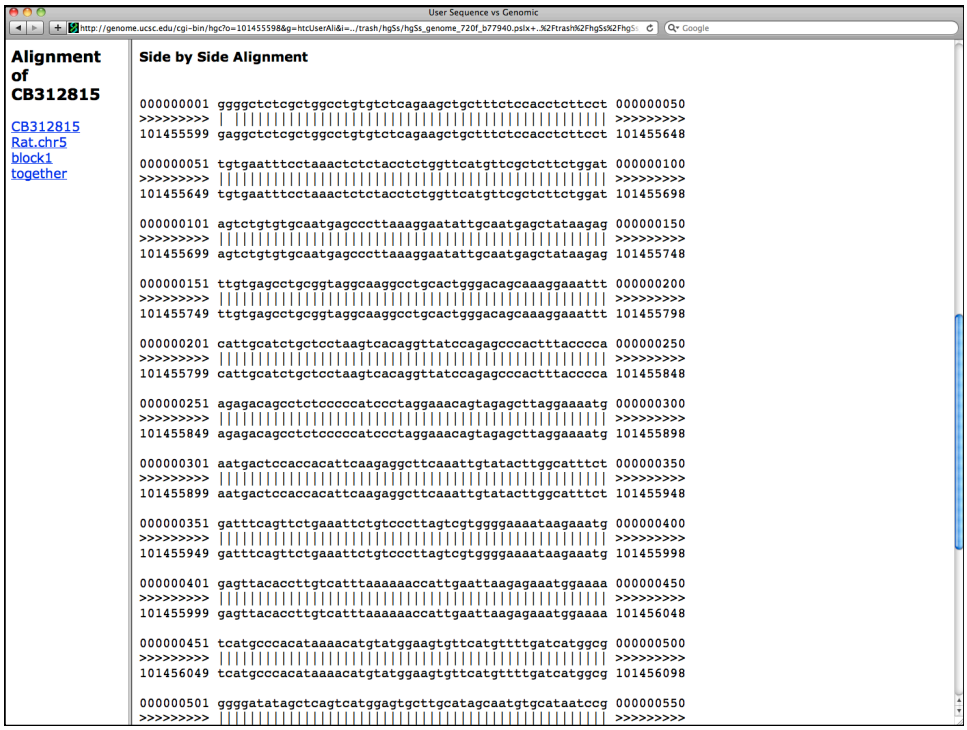

Rat BLAT Results
http://genome.ucsc.edu/cgi-bin/hgBlat
Home Genomes Tables Gene Sorter PCR Session FAQ Help
Rat BLAT Results
BLAT Search Results
ACTIONS QUERY SCORE START END QSIZE IDENTITY CHRO STRAND START END SPAN
browser details CB312815 710 1 733 768 98.1% 5 + 101455599 101456323 725
browser details 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





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


User Sequence vs Genomic
Alignment of CB312815
CB312815
Rat.chr5
block1
together
Alignment of CB312815 and chr5:101455599-101456323
Click on links in the frame to the left to navigate through the alignment. Matching bases in cDNA and genomic sequences are colored blue and capitalized. Light blue bases mark the boundaries of gaps in either sequence (often splice sites).
cDNA CB312815
Genomic chr5 :

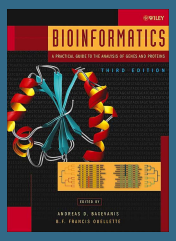

User Sequence vs Genomic
Alignment of CB312815
CB312815
Rat.chr5
block1
together
Side by Side Alignment

FASTA

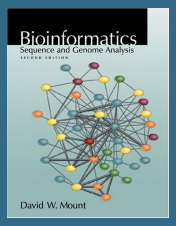
- Identifies regions of local alignment
- Employs an approximation of the Smith-Waterman algorithm to determine the best alignment between two sequences
- Method is significantly different from that used by BLAST
- Online implementations at
  http://fasta.bioch.virginia.edu
  http://www.ebi.ac.uk/fasta33

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Further Reading



Chapter 11
Assessing Pairwise Sequence Similarity: BLAST and FASTA



Chapter 6
Sequence Database Searching for Similar Sequences

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research