Current Topics in Genome Analysis Spring 2005

Week 4
Biological Sequence Analysis I

Andy Baxevanis, Ph.D.

Overview

- Week 4: Comparative methods and concepts
  - Similarity *vs.* Homology
  - Global *vs.* Local Alignments
  - Scoring Matrices
  - BLAST
  - BLAT
- Week 5: Predictive methods and concepts
  - Profiles, patterns, motifs, and domains
  - Secondary structure prediction
  - Structures: VAST, Cn3D, and *de novo* prediction

Why do sequence alignments?

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

  → *importance of using correct terminology*

Defining the Terms

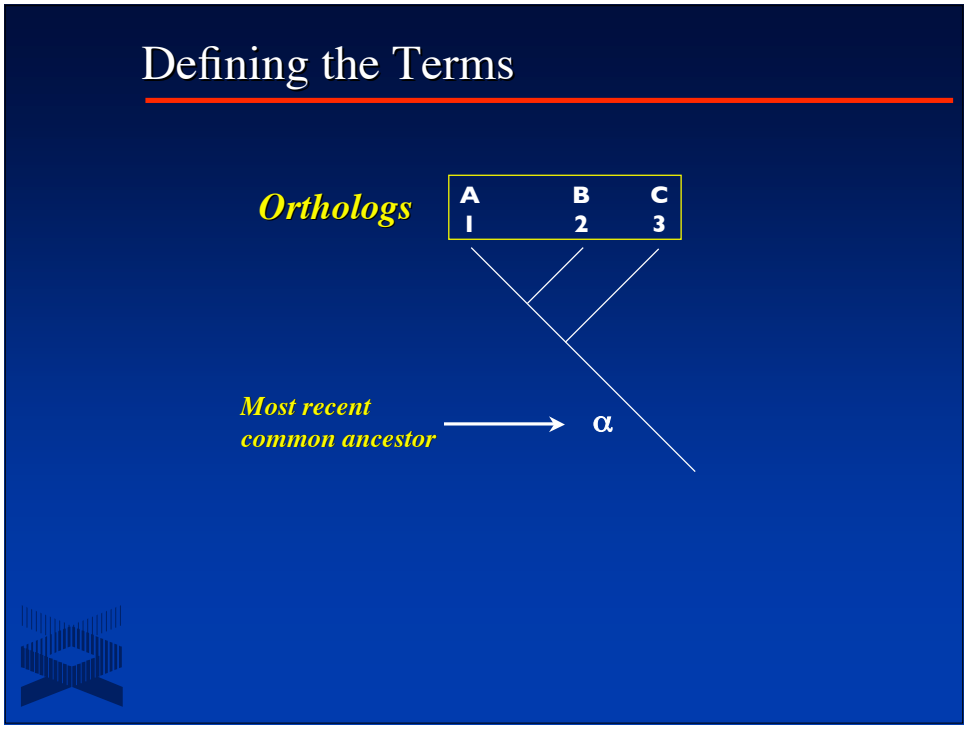
- The quantitative measure: ***Similarity***
  - Always based on an observable
  - Usually expressed as percent identity
  - Quantify changes that occur as two sequences diverge
    - substitutions
    - insertions
    - deletions
  - Identify residues crucial for maintaining a protein's structure or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

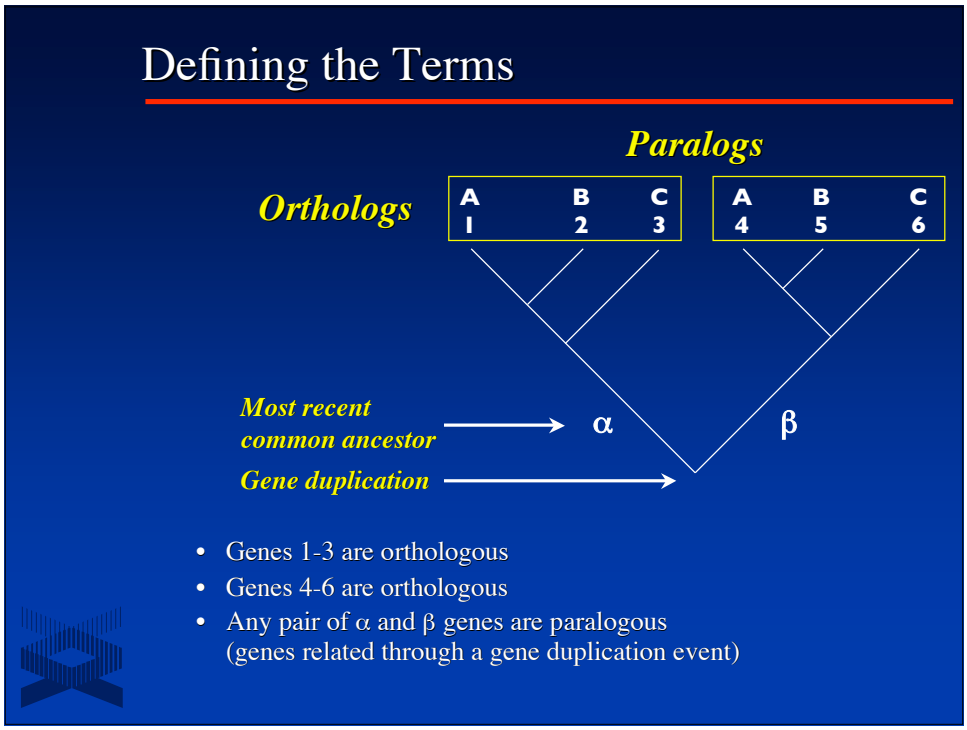
Defining the Terms

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

Defining the Terms

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


Defining the Terms
Orthologs
A 1
B 2
C 3
Most recent common ancestor
α


Defining the Terms
Paralogs
Orthologs
A 1
B 2
C 3
A 4
B 5
C 6
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)

Overview

- Week 4: Comparative methods and concepts
  - Similarity *vs.* Homology
  - Global *vs.* Local Alignments
  - Scoring Matrices
  - BLAST
  - BLAT
- Week 5: Predictive methods and concepts
  - Profiles, patterns, motifs, and domains
  - Secondary structure prediction
  - Structures: VAST, Cn3D, and *de novo* prediction

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 alignments could be generated for any two sequences being compared
- Best for sequences that share some similarity, or for sequences of different lengths

Overview

- Week 4: Comparative methods and concepts
  - Similarity *vs.* Homology
  - Global *vs.* Local Alignments
  - Scoring Matrices
  - BLAST
  - BLAT
- Week 5: Predictive methods and concepts
  - Profiles, patterns, motifs, and domains
  - Secondary structure prediction
  - Structures: VAST, Cn3D, and *de novo* prediction

Scoring Matrices

- Empirical weighting scheme to represent biology (side chain chemistry, structure, and function)
  - 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

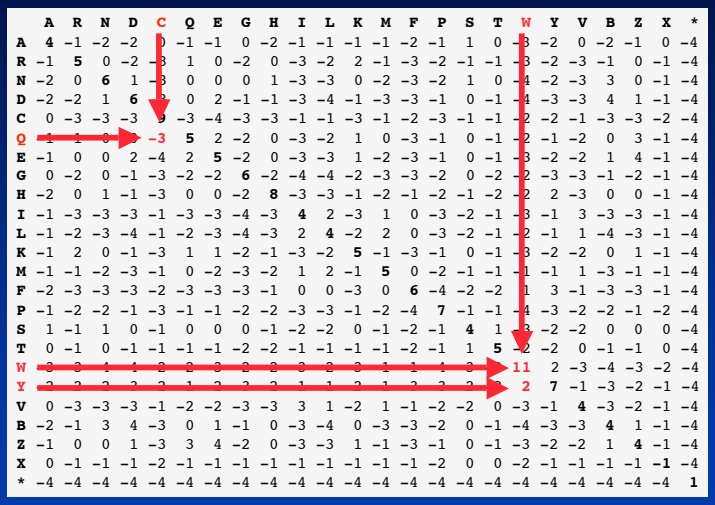
- Importance of understanding scoring matrices
  - Appear in all analyses involving sequence comparison
  - Implicitly represent particular evolutionary patterns
  - Choice of matrix can strongly influence outcomes

Matrix Structure: Nucleotides

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

- *Simple match/mismatch scoring scheme*
- *Assumes each nucleotide occurs 25% of the time*

Matrix Structure: Proteins



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

BLOSUM62

PAM Matrices

- Margaret Dayhoff and colleagues, 1978
  - Look at patterns of substitutions in highly related proteins (> 85% similar) within multiple sequence alignments
  - Analysis documented 1572 changes in 71 groups of proteins examined
  - Substitution tables constructed based on results
  - Given high degree of similarity within original sequence set, results represent substitution pattern that would be expected over short evolutionary distances

PAM Matrices

- Short evolutionary distance
  ∴ change in function unlikely
- Point Accepted Mutation (PAM)
  - The new side chain must function the same way as the old one ("acceptance")
  - On average, 1 PAM corresponds to 1 amino acid change per 100 residues
  - 1 PAM ~ 1% divergence
  - Extrapolate to predict patterns at longer evolutionary distances

PAM Matrices: Assumptions

- All sites assumed to be equally mutable
- Replacement of amino acids is independent of previous mutations at the same position
- Replacement is independent of surrounding residues
- Forces responsible for sequence evolution over shorter time spans are the same as those over longer time spans

PAM Matrices: Sources of Error

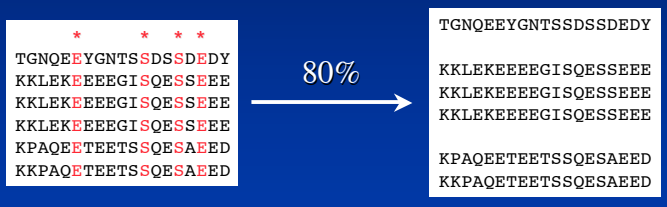
- Small, globular proteins of average composition used to derive matrices
- Errors in PAM 1 are magnified up to PAM 250 (only PAM 1 is based on direct observation)
- Does not account for conserved blocks or motifs

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

BLOSUM *n*

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

So many matrices...

Triple-PAM Strategy *(Altschul, 1991)*

PAM 40	Short alignments, highly similar	70-90%
PAM 160	Detecting known members of a protein family	50-60%
PAM 250	Longer, weaker local alignments	~ 30%

BLOSUM *(Henikoff, 1993)*

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

So many matrices...

- Matrix Equivalencies

PAM 250	~	BLOSUM 45
PAM 160	~	BLOSUM 62
PAM 120	~	BLOSUM 80

- Specialized matrices
  - Transmembrane proteins
  - Species-specific matrices

Wheeler, 2003

So many matrices...

No single matrix is the complete answer for all sequence comparisons

Gaps

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

Affine Gap Penalty

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

Deduction for a gap = $G + Ln$

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

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

Overview

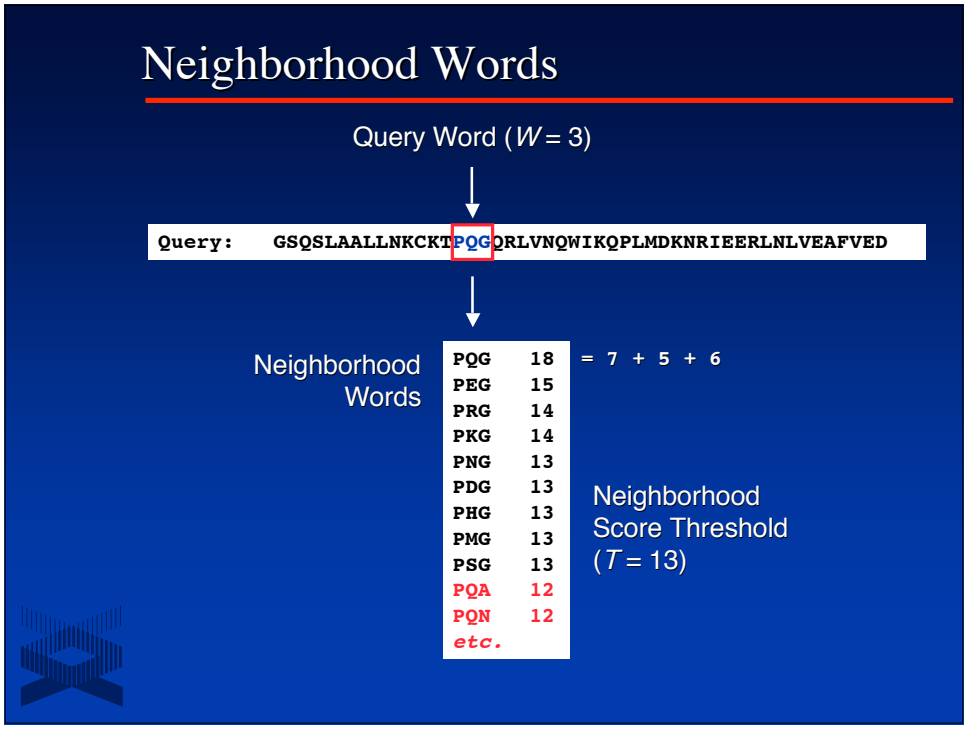
- Week 4: Comparative methods and concepts
  - Similarity *vs.* Homology
  - Global *vs.* Local Alignments
  - Scoring Matrices
  - BLAST
  - BLAT
- Week 5: Predictive methods and concepts
  - Profiles, patterns, motifs, and domains
  - Secondary structure prediction
  - Structures: VAST, Cn3D, and *de novo* prediction

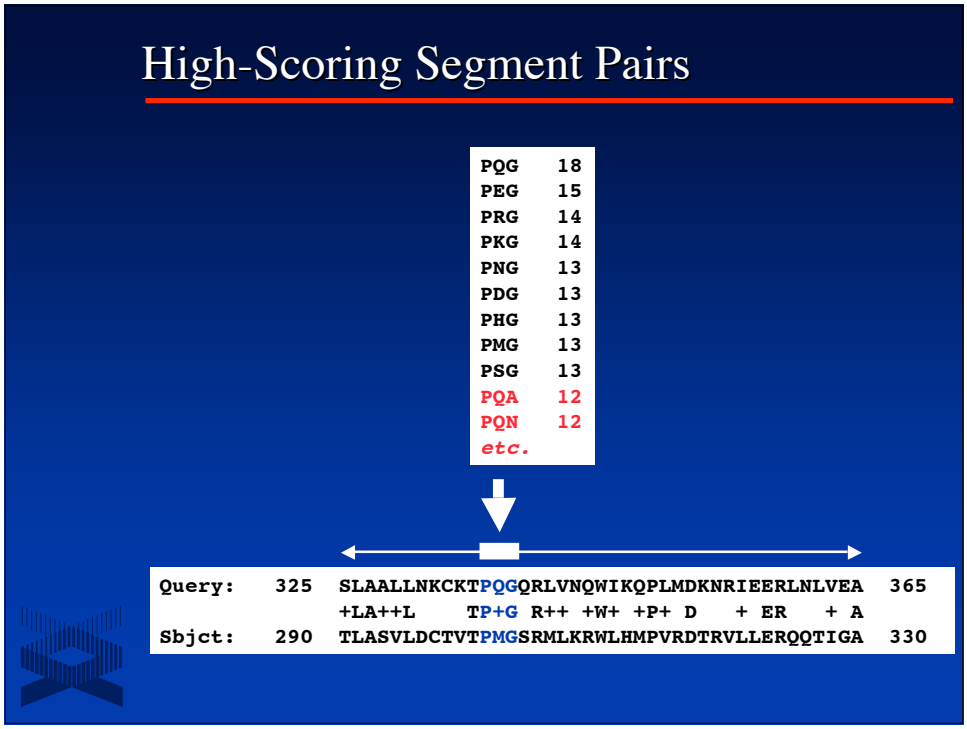
BLAST

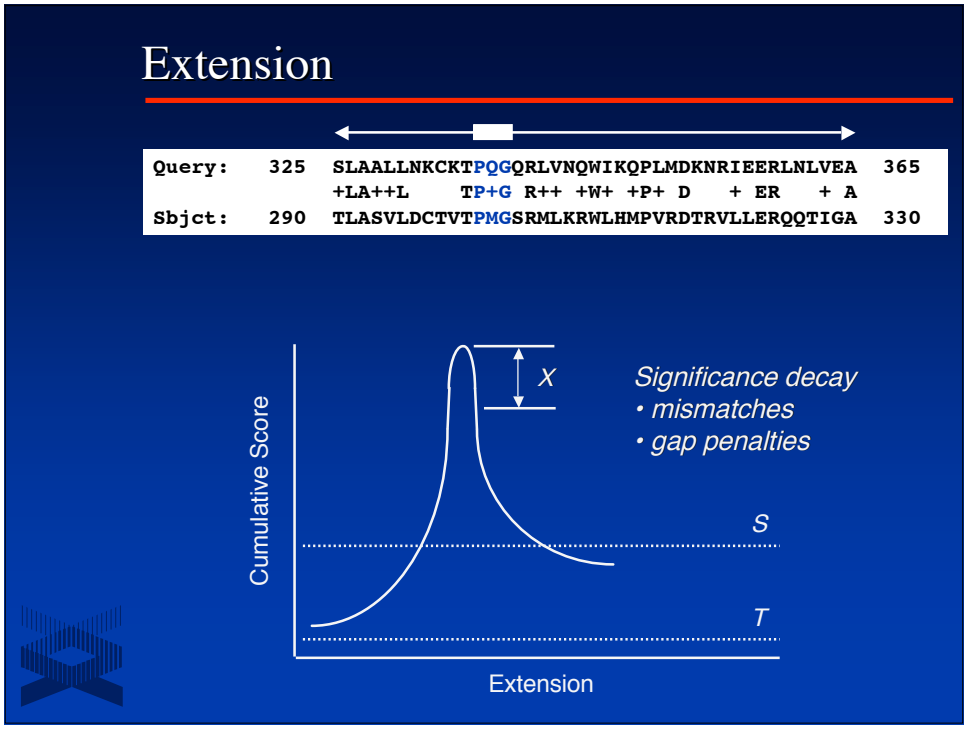
- Basic Local Alignment Search Tool
- Seeks high-scoring segment pairs (HSP)
  - pair of sequences that can be aligned without gaps
  - when aligned, have maximal aggregate score (score cannot be improved by extension or trimming)
  - score must be above score threshhold *S*
  - gapped or ungapped
- Results not limited to the "best HSP" for any given sequence pair

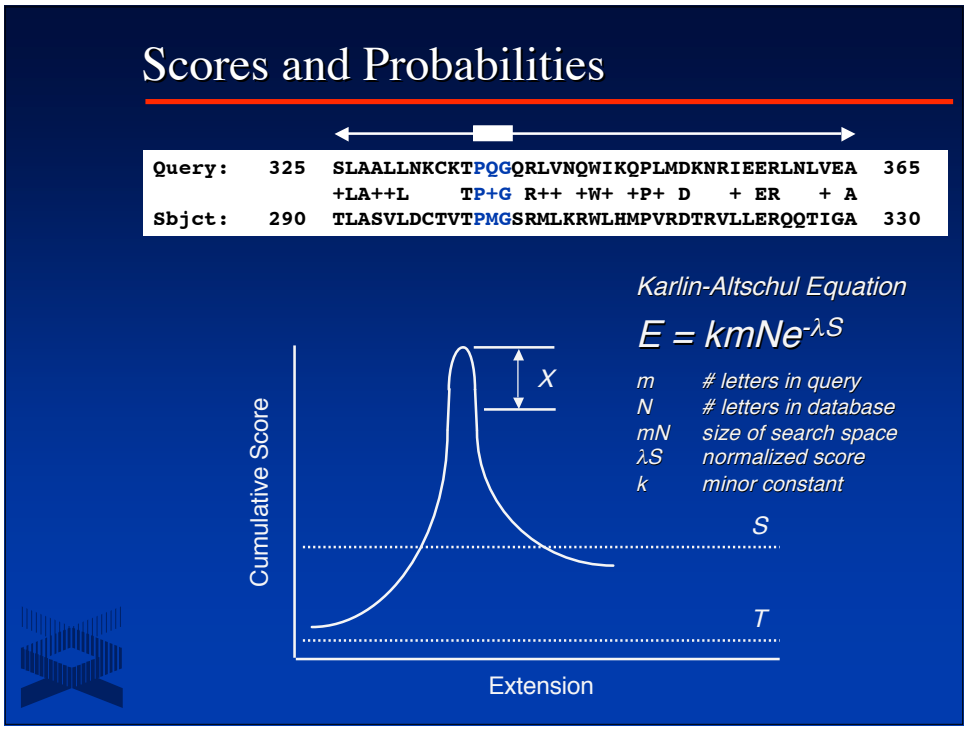
BLAST Algorithms

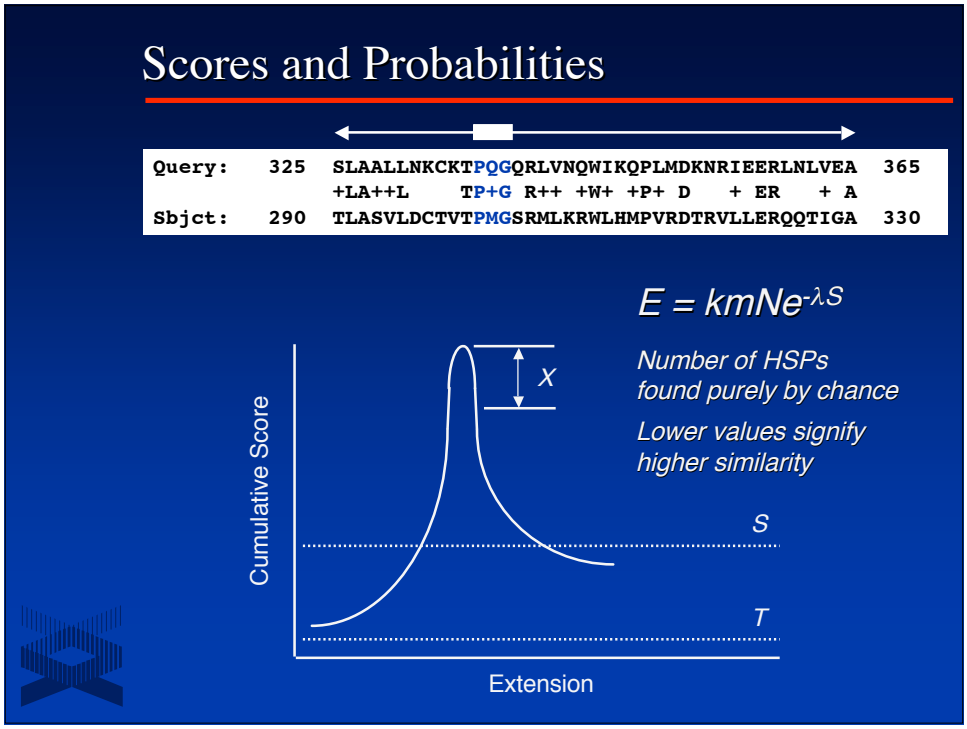
Program	*Query Sequence*	*Target Sequence*
BLASTN	**Nucleotide**	**Nucleotide**
BLASTP	**Protein**	**Protein**
BLASTX	**Nucleotide, six-frame translation**	**Protein**
TBLASTN	Protein	Nucleotide, six-frame translation
TBLASTX	Nucleotide, six-frame translation	Nucleotide, six-frame translation

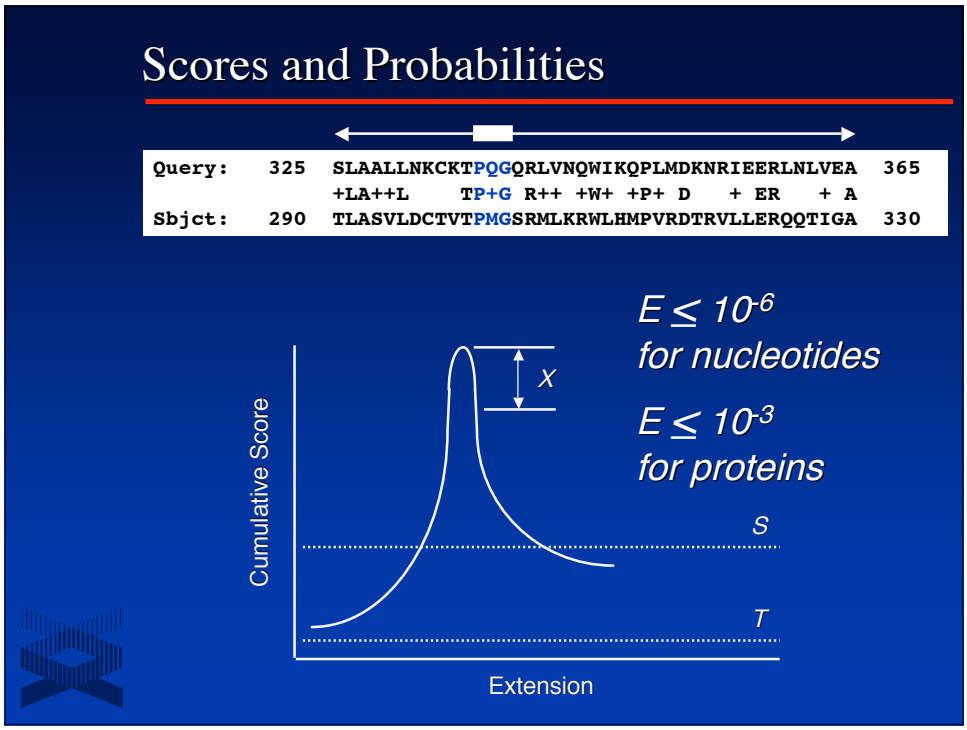

Neighborhood Words
Query Word (W = 3)
Query: GSQSLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEAFVED
Neighborhood Words
PQG 18
PEG 15
PRG 14
PKG 14
PNG 13
PDG 13
PHG 13
PMG 13
PSG 13
PQA 12
PQN 12
etc.
= 7 + 5 + 6
Neighborhood Score Threshold (T = 13)

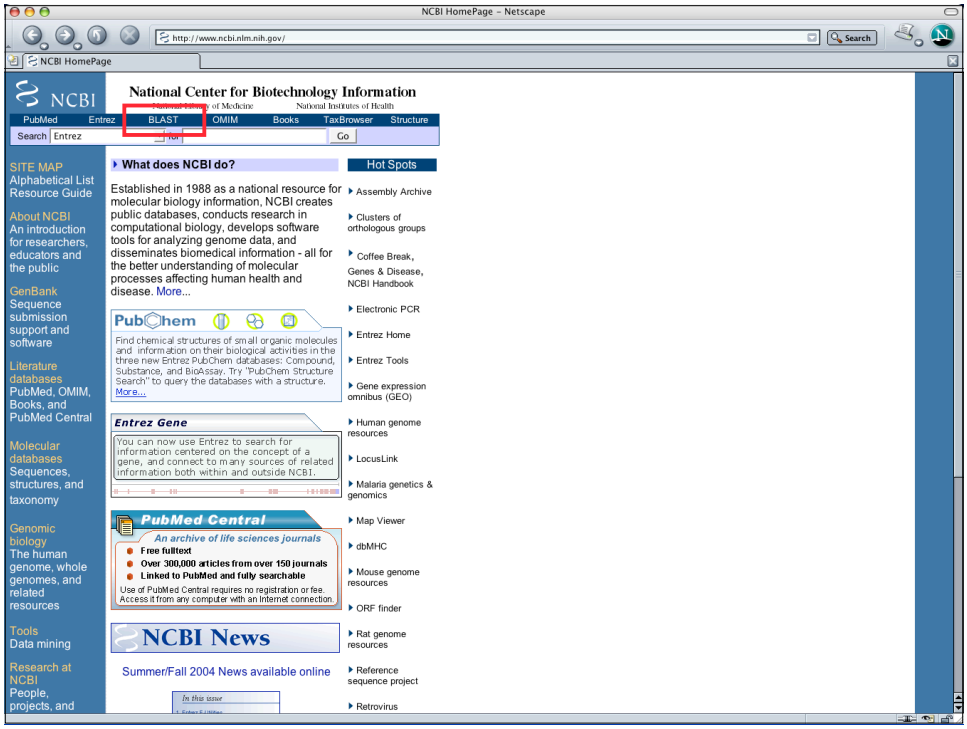

High-Scoring Segment Pairs
PQG 18
PEG 15
PRG 14
PKG 14
PNG 13
PDG 13
PHG 13
PMG 13
PSG 13
PQA 12
PQN 12
etc.
Query: 325 SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA 365
+LA++L TP+G R++ +W+ +P+ D + ER + A
Sbjct: 290 TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA 330

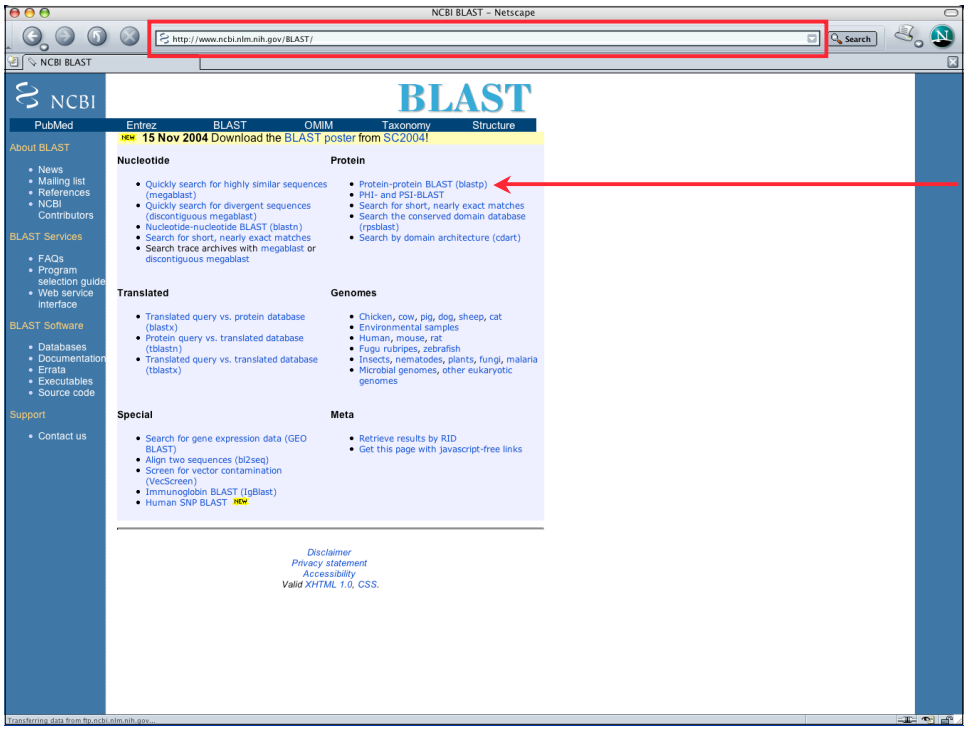

Extension
Query: 325 SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA 365
+LA++L TP+G R++ +W+ +P+ D + ER + A
Sbjct: 290 TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA 330
X
Significance decay
• mismatches
• gap penalties
S
T
Cumulative Score
Extension

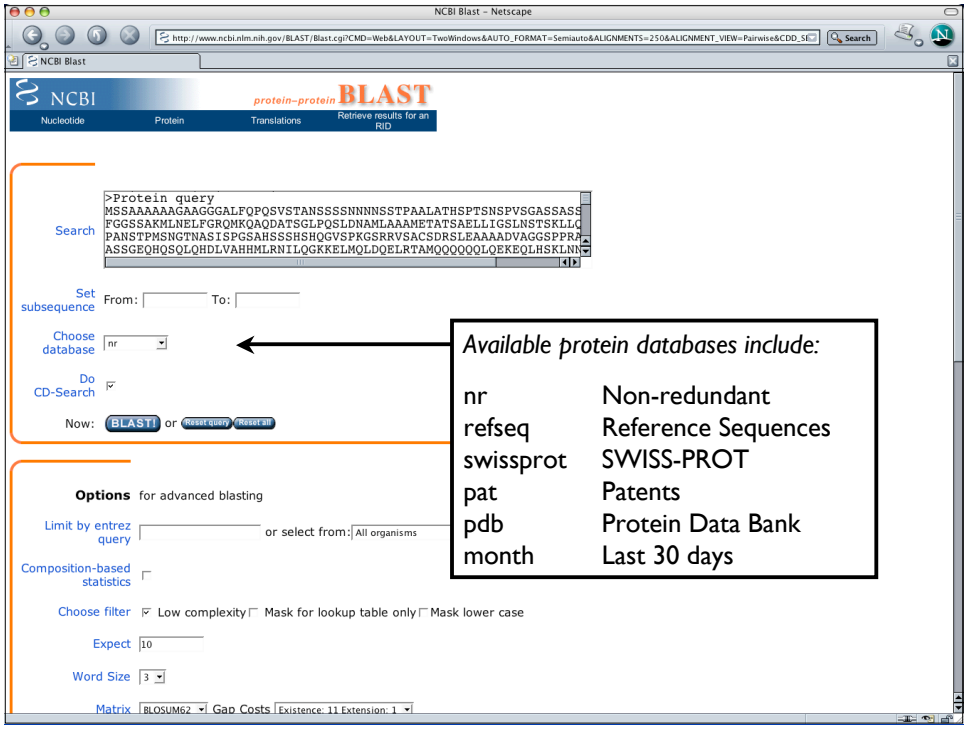

Scores and Probabilities
Query: 325 SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA 365
+LA++L TP+G R++ +W+ +P+ D + ER + A
Sbjct: 290 TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA 330
Karlin-Altschul Equation
E = kmNe-λS
m # letters in query
N # letters in database
mN size of search space
λS normalized score
k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

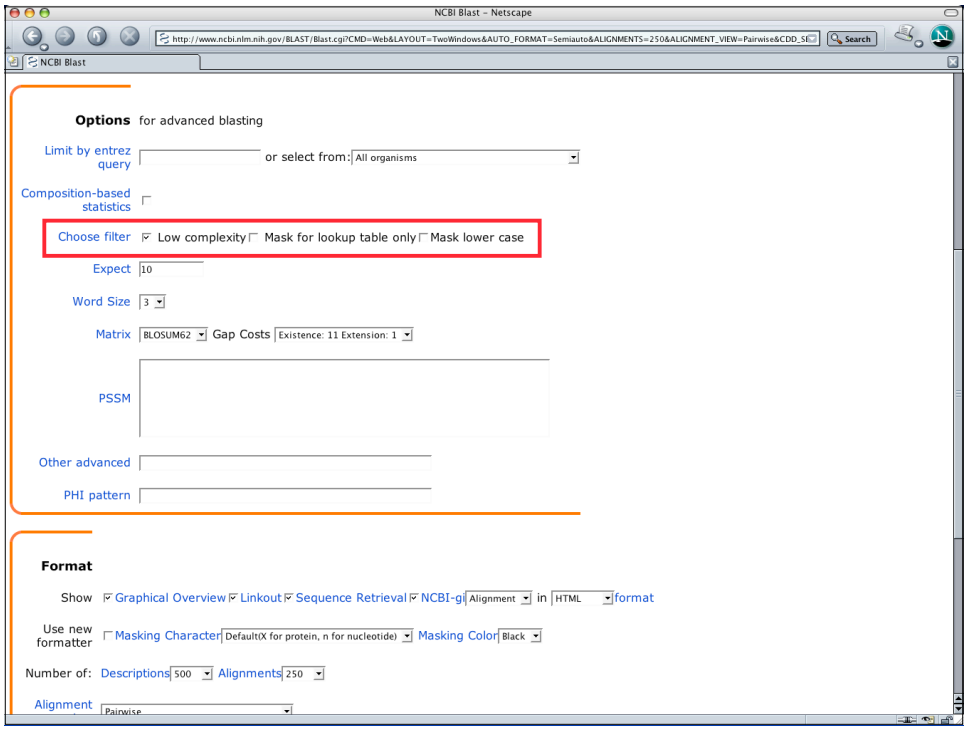

Scores and Probabilities
Query: 325 SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA 365
+LA++L TP+G R++ +W+ +P+ D + ER + A
Sbjct: 290 TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA 330
E = kmNe-λS
Number of HSPs found purely by chance
Lower values signify higher similarity
X
S
T
Cumulative Score
Extension


Scores and Probabilities
Query: 325 SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA 365
+LA++L TP+G R++ +W+ +P+ D + ER + A
Sbjct: 290 TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA 330
E ≤ 10-6
for nucleotides
E ≤ 10-3
for proteins
X
S
T
Cumulative Score
Extension


NCBI HomePage - Netscape
http://www.ncbi.nlm.nih.gov/
NCBI HomePage
NCBI
National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
PubMed
Entrez
BLAST
OMIM
Books
TaxBrowser
Structure
Search
Entrez
Go
SITE MAP
Alphabetical List
Resource Guide
About NCBI
An introduction for researchers, educators and the public
GenBank
Sequence submission support and software
Literature databases
PubMed, OMIM, Books, and PubMed Central
Molecular databases
Sequences, structures, and taxonomy
Genomic biology
The human genome, whole genomes, and related resources
Tools
Data mining
Research at NCBI
People, projects, and
What does NCBI do?
Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. More...
Hot Spots
Assembly Archive
Clusters of orthologous groups
Coffee Break, Genes & Disease, NCBI Handbook
Electronic PCR
Entrez Home
Entrez Tools
Gene expression omnibus (GEO)
Human genome resources
LocusLink
Malaria genetics & genomics
Map Viewer
dbMHC
Mouse genome resources
ORF finder
Rat genome resources
Reference sequence project
Retrovirus
PubChem
Find chemical structures of small organic molecules and information on their biological activities in the three new Entrez PubChem databases: Compound, Substance, and BioAssay. Try "PubChem Structure Search" to query the databases with a structure. More...
Entrez Gene
You can now use Entrez to search for information centered on the concept of a gene, and connect to many sources of related information both within and outside NCBI.
PubMed Central
An archive of life sciences journals
Free fulltext
Over 300,000 articles from over 150 journals
Linked to PubMed and fully searchable
Use of PubMed Central requires no registration or fee. Access it from any computer with an Internet connection.
NCBI News
Summer/Fall 2004 News available online
In this issue


NCBI BLAST - Netscape
http://www.ncbi.nlm.nih.gov/BLAST/
NCBI BLAST
NCBI
BLAST
PubMed
Entrez
BLAST
OMIM
Taxonomy
Structure
15 Nov 2004 Download the BLAST poster from SC2004!
About BLAST
News
Mailing list
References
NCBI Contributors
BLAST Services
FAQs
Program selection guide
Web service interface
BLAST Software
Databases
Documentation
Errata
Executables
Source code
Support
Contact us
Nucleotide
Quickly search for highly similar sequences (megablast)
Quickly search for divergent sequences (discontiguous megablast)
Nucleotide-nucleotide BLAST (blastn)
Search for short, nearly exact matches
Search trace archives with megablast or discontiguous megablast
Protein
Protein-protein BLAST (blastp)
PHI- and PSI-BLAST
Search for short, nearly exact matches
Search the conserved domain database (rpsblast)
Search by domain architecture (cdart)
Translated
Translated query vs. protein database (blastx)
Protein query vs. translated database (tblastn)
Translated query vs. translated database (tblastx)
Genomes
Chicken, cow, pig, dog, sheep, cat
Environmental samples
Human, mouse, rat
Fugu rubripes, zebrafish
Insects, nematodes, plants, fungi, malaria
Microbial genomes, other eukaryotic genomes
Special
Search for gene expression data (GEO BLAST)
Align two sequences (bl2seq)
Screen for vector contamination (VecScreen)
Immunoglobin BLAST (IgBlast)
Human SNP BLAST
Meta
Retrieve results by RID
Get this page with javascript-free links
Disclaimer
Privacy statement
Accessibility
Valid XHTML 1.0, CSS.


NCBI Blast – Netscape
NCBI
protein–protein BLAST
Nucleotide
Protein
Translations
Retrieve results for an RID
Search
>Protein query
Set subsequence
From:
To:
Choose database
nr
Do CD-Search
Now:
BLAST!
or
Reset query
Reset all
Options for advanced blasting
Limit by entrez query
or select from:
All organisms
Composition-based statistics
Choose filter
Low complexity
Mask for lookup table only
Mask lower case
Expect
10
Word Size
3
Matrix
BLOSUM62
Gap Costs
Existence: 11 Extension: 1
Available protein databases include:
nr Non-redundant
refseq Reference Sequences
swissprot SWISS-PROT
pat Patents
pdb Protein Data Bank
month Last 30 days


NCBI Blast – Netscape
Options for advanced blasting
Limit by entrez query
or select from:
All organisms
Composition-based statistics
Choose filter
Low complexity
Mask for lookup table only
Mask lower case
Expect
10
Word Size
3
Matrix
BLOSUM62
Gap Costs
Existence: 11 Extension: 1
PSSM
Other advanced
PHI pattern
Format
Show
Graphical Overview
Linkout
Sequence Retrieval
NCBI-gi
Alignment
in
HTML
format
Use new formatter
Masking Character
Masking Color
Black
Number of:
Descriptions
500
Alignments
250
Alignment
Pairwise

Low-Complexity Regions

Defined as regions of biased composition

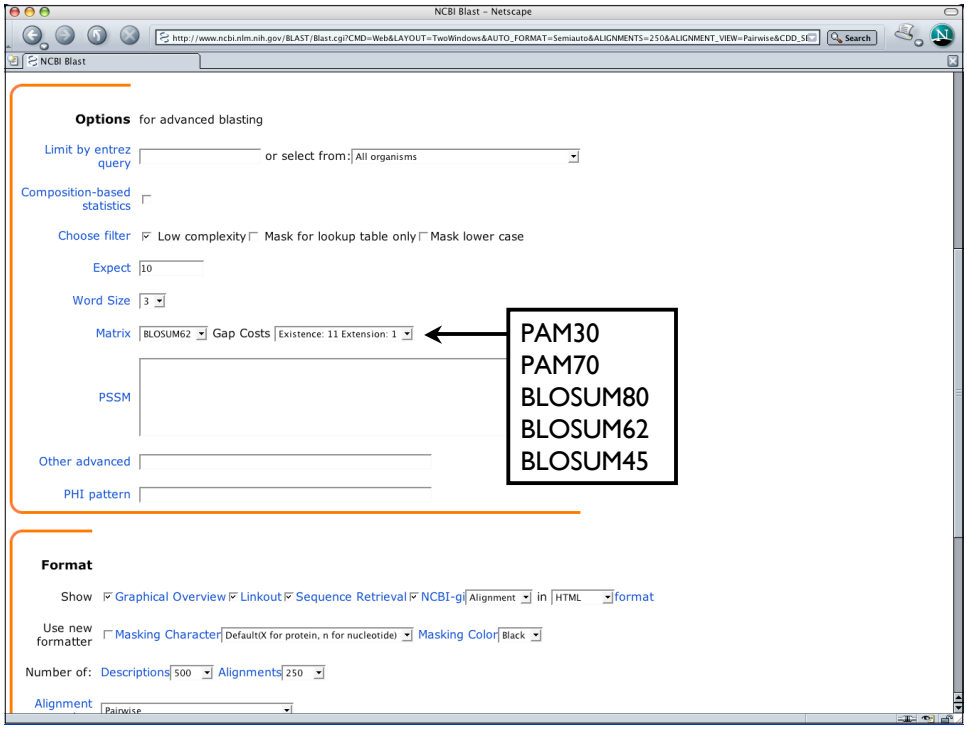
- Homopolymeric runs
- Short-period repeats
- Subtle over-representation of several residues

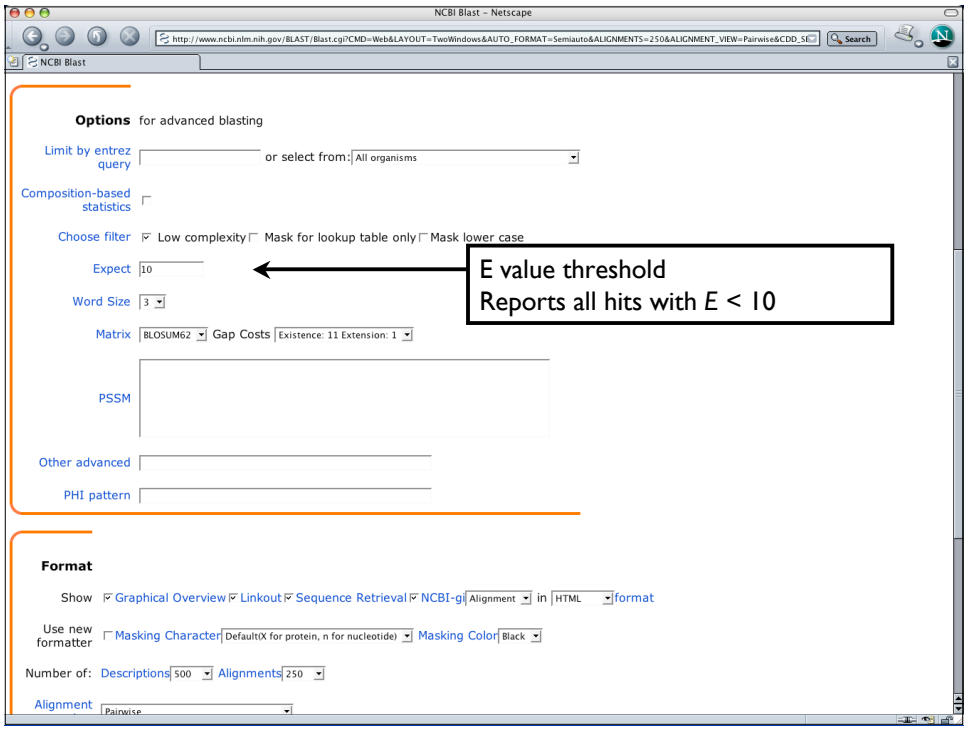
```
>gi|20455478|sp|P50553|ASC1_HUMAN Achaete-scute homolog 1 (HASH1)
MESSAKMESGGAGQQPQPQPQQPFLPPAACFFATAAAAAAAAAAAAAQSAQQQQQQQQQQQQAPQLRPAA
DGQPSGGGHKSAPKQVKRQRSSSPELMRCKRRLNFSGFGYSLPQQQPAAVARRNERERNRVKLVNLGFAT
LREHVPNGAANKKMSKVETLRSAVEYIRALQQLLDEHDAVSAAFQAGVLSPTISPNYSNDLNSMAGSPVS
SYSSDEGSYDPLSPEEQELLDFTNWF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

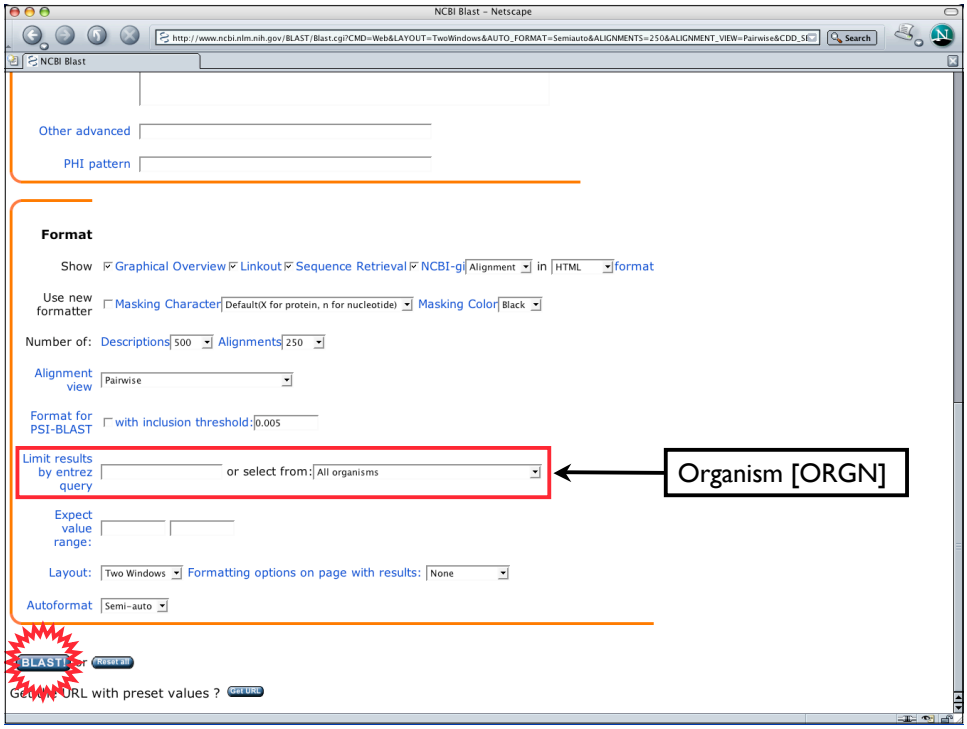
Homopolymeric alanine-glutamine tract

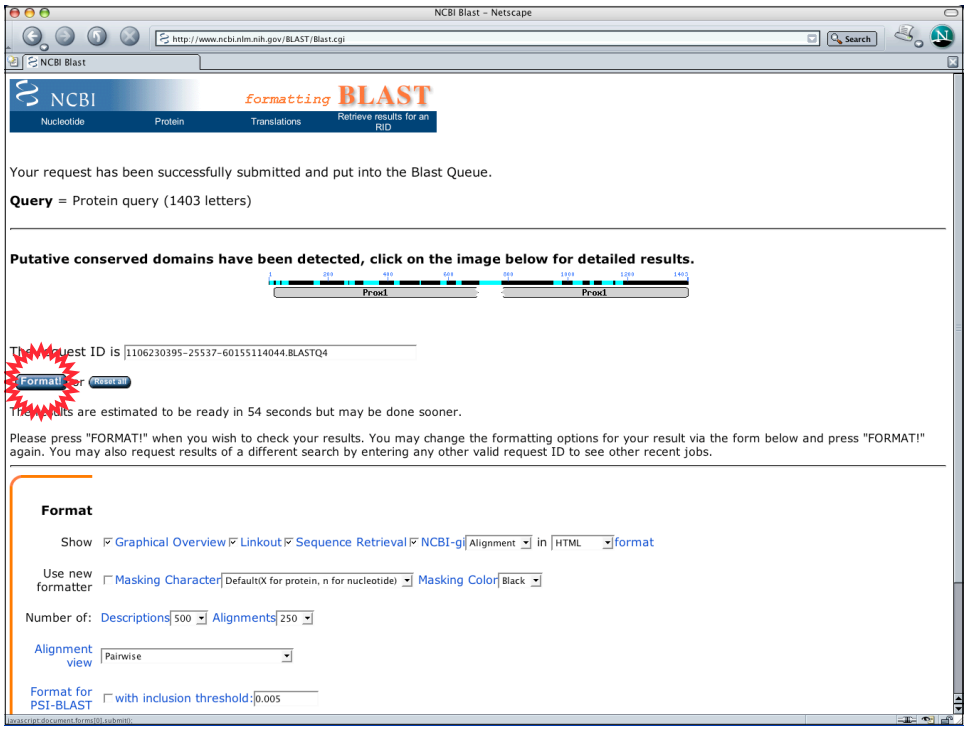
Identifying Low-Complexity Regions

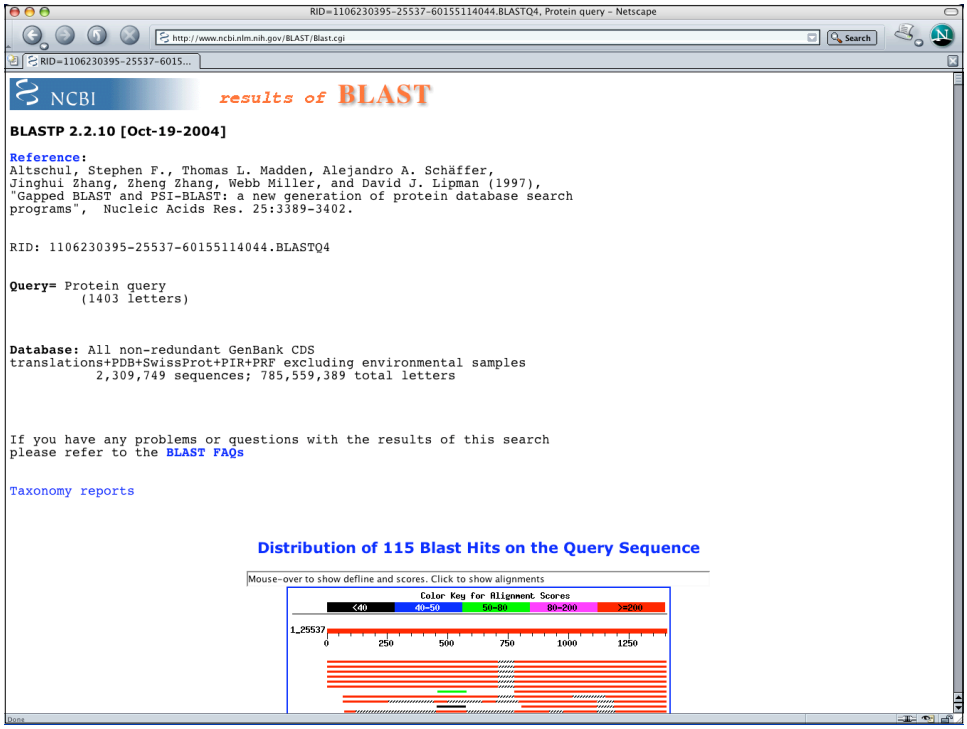
- Biological origins and role not well-understood
  - DNA replication errors (polymerase slippage)?
  - Unequal crossing-over?
- May confound sequence analysis
  - BLAST relies on uniformly-distributed amino acid frequencies
  - Often lead to false positives
  - Filtering is advised (and usually enabled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

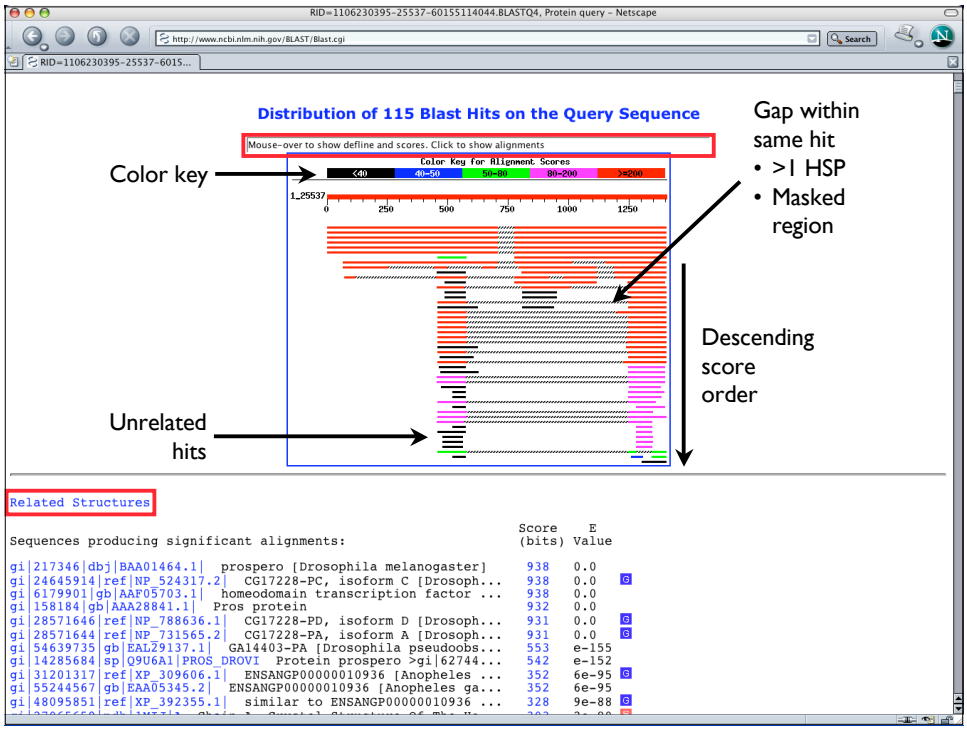

NCBI Blast – Netscape
NCBI Blast
Options for advanced blasting
Limit by entrez query
or select from: All organisms
Composition-based statistics
Choose filter
Low complexity
Mask for lookup table only
Mask lower case
Expect
10
Word Size
3
Matrix
BLOSUM62
Gap Costs
Existence: 11 Extension: 1
PSSM
Other advanced
PHI pattern
PAM30
PAM70
BLOSUM80
BLOSUM62
BLOSUM45
Format
Show
Graphical Overview
Linkout
Sequence Retrieval
NCBI-gi
Alignment
in
HTML
format
Use new formatter
Masking Character
Masking Color
Black
Number of:
Descriptions
500
Alignments
250
Alignment
Pairwise

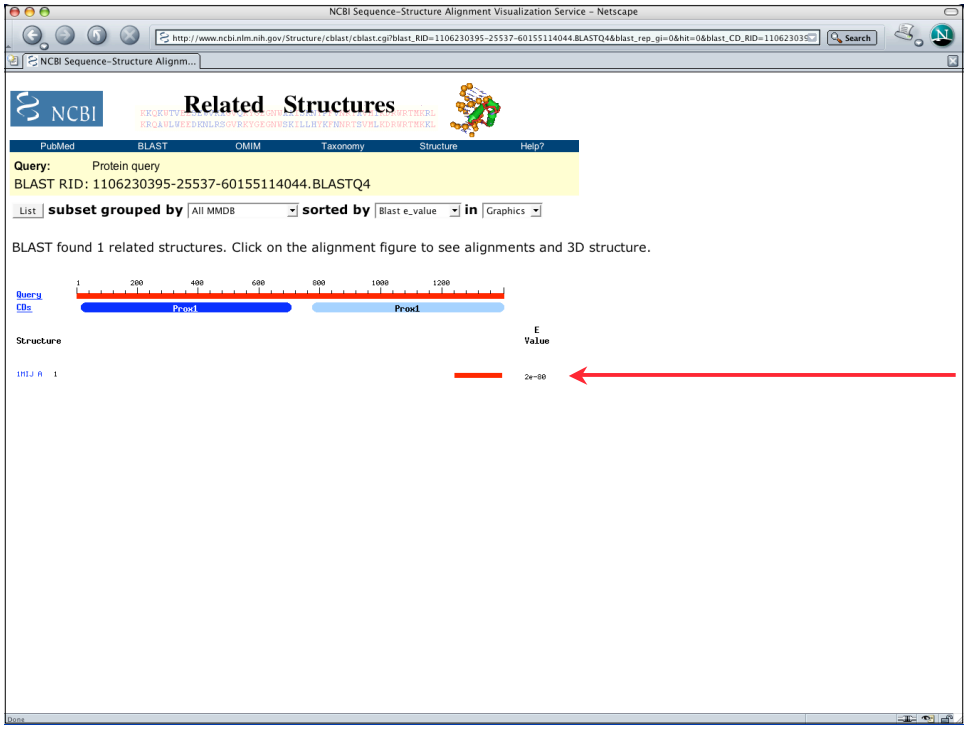

NCBI Blast – Netscape
NCBI Blast
Options for advanced blasting
Limit by entrez query
or select from: All organisms
Composition-based statistics
Choose filter
Low complexity
Mask for lookup table only
Mask lower case
Expect
10
E value threshold
Reports all hits with E < 10
Word Size
3
Matrix
BLOSUM62
Gap Costs
Existence: 11 Extension: 1
PSSM
Other advanced
PHI pattern
Format
Show
Graphical Overview
Linkout
Sequence Retrieval
NCBI-gi
Alignment
in
HTML
format
Use new formatter
Masking Character
Masking Color
Black
Number of:
Descriptions
500
Alignments
250
Alignment
Pairwise

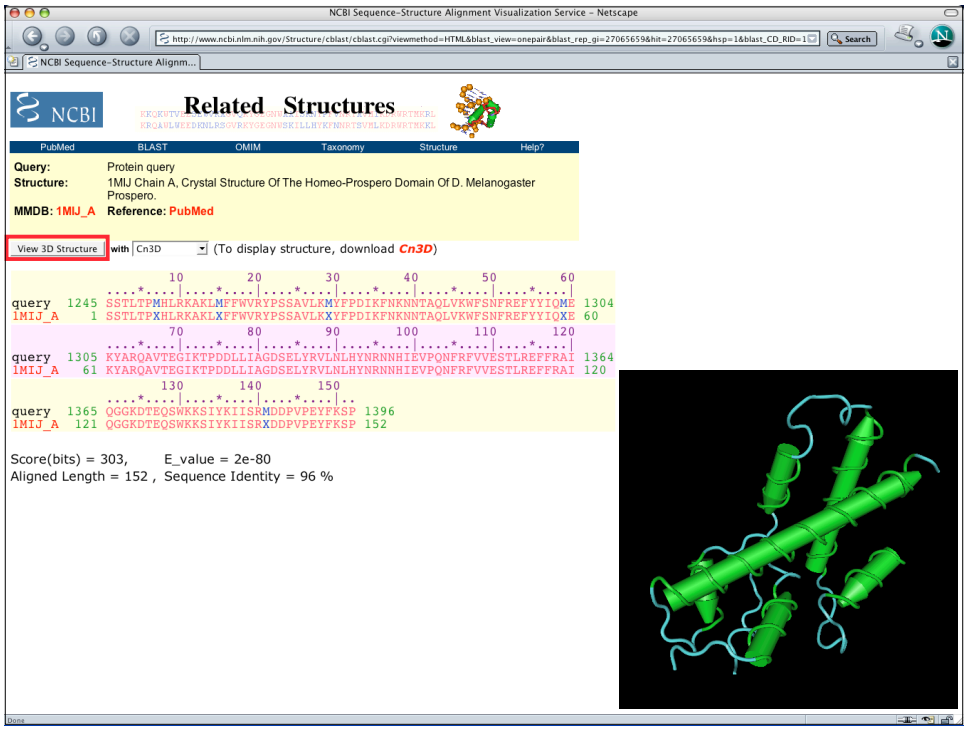

NCBI Blast - Netscape
NCBI Blast
Other advanced
PHI pattern
Format
Show Graphical Overview Linkout Sequence Retrieval NCBI-gi Alignment in HTML format
Use new formatter Masking Character Default(X for protein, n for nucleotide) Masking Color Black
Number of: Descriptions 500 Alignments 250
Alignment view Pairwise
Format for PSI-BLAST with inclusion threshold: 0.005
Limit results by entrez query or select from: All organisms
Organism [ORGN]
Expect value range:
Layout: Two Windows Formatting options on page with results: None
Autoformat Semi-auto
BLAST! or Reset all
Get URL with preset values ? Get URL

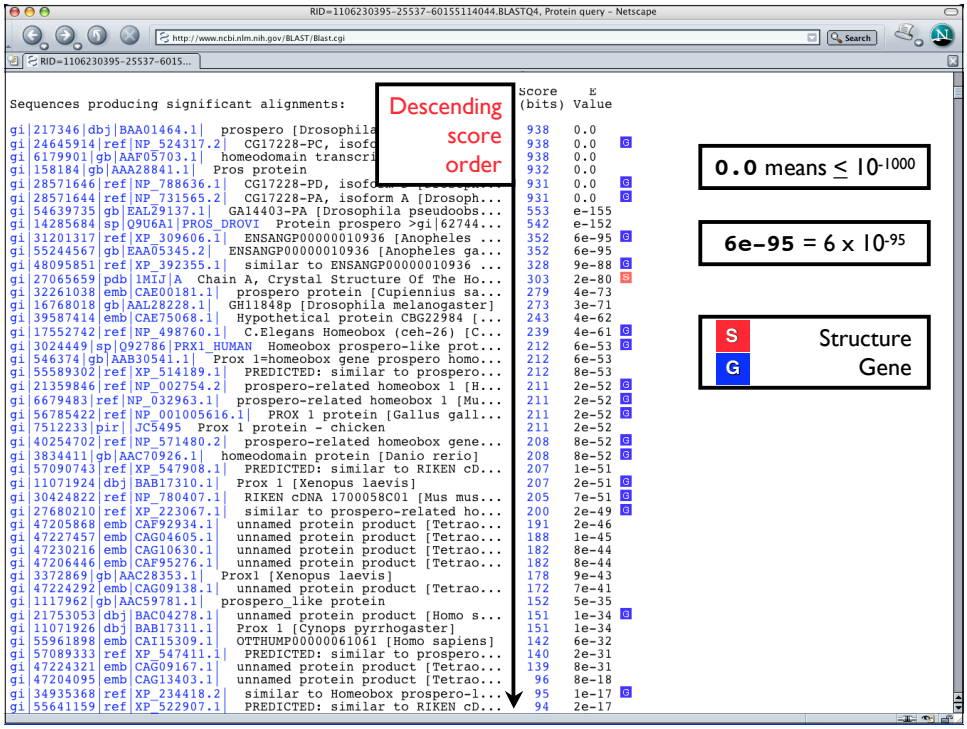

NCBI Blast - Netscape
NCBI Blast
NCBI formatting BLAST
Nucleotide Protein Translations Retrieve results for an RID
Your request has been successfully submitted and put into the Blast Queue.
Query = Protein query (1403 letters)
Putative conserved domains have been detected, click on the image below for detailed results.
ProcI
ProcI
Format! or Reset all
Please press "FORMAT!" when you wish to check your results. You may change the formatting options for your result via the form below and press "FORMAT!" again. You may also request results of a different search by entering any other valid request ID to see other recent jobs.
Format
Show Graphical Overview Linkout Sequence Retrieval NCBI-gi Alignment in HTML format
Use new formatter Masking Character Default(X for protein, n for nucleotide) Masking Color Black
Number of: Descriptions 500 Alignments 250
Alignment view Pairwise
Format for PSI-BLAST with inclusion threshold: 0.005

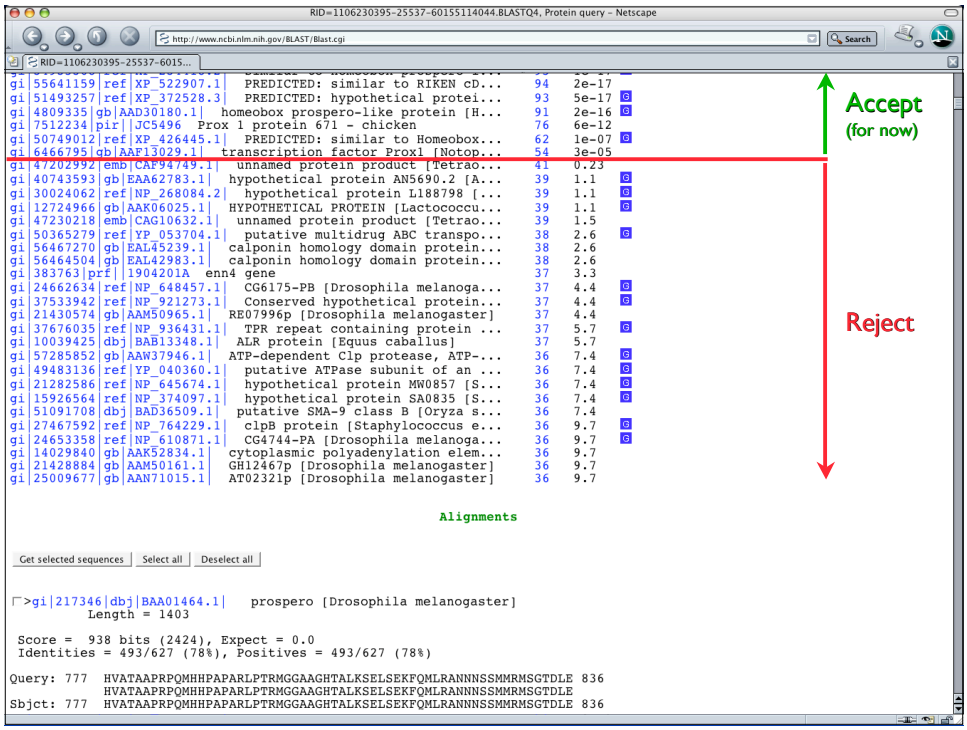

RID=1106230395-25537-60155114044.BLASTQ4, Protein query - Netscape
NCBI
results of BLAST
BLASTP 2.2.10 [Oct-19-2004]
Reference:
Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs", Nucleic Acids Res. 25:3389-3402.
RID: 1106230395-25537-60155114044.BLASTQ4
Query= Protein query
(1403 letters)
Database: All non-redundant GenBank CDS
translations+PDB+SwissProt+PIR+PRF excluding environmental samples
2,309,749 sequences; 785,559,389 total letters
If you have any problems or questions with the results of this search
please refer to the BLAST FAQs
Taxonomy reports
Distribution of 115 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
1_25537
0
250
500
750
1000
1250

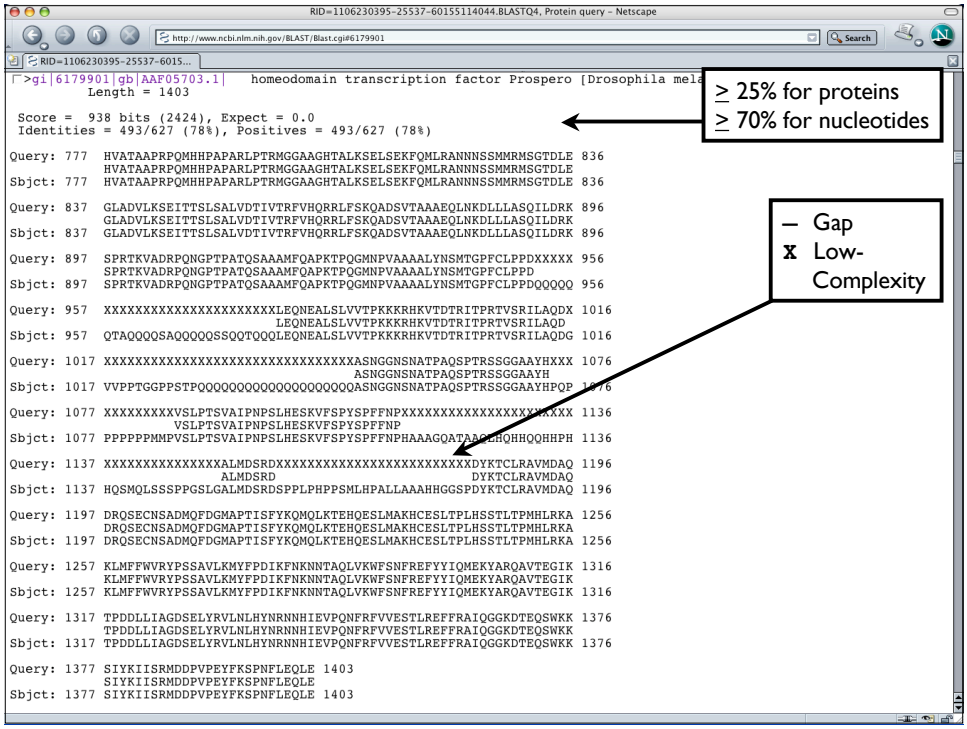

RID=1106230395-25537-60155114044.BLASTQ4, Protein query - Netscape
Distribution of 115 Blast Hits on the Query Sequence
Mouse-over to show defline and scores. Click to show alignments
Color key
Color Key for Alignment Scores
<40
40-50
50-80
80-200
>=200
1_25537
0
250
500
750
1000
1250
Gap within same hit
• >1 HSP
• Masked region
Descending score order
Unrelated hits
Related Structures
Sequences producing significant alignments:
Score (bits)
E Value
gi|217346|dbj|BAA01464.1| prospero [Drosophila melanogaster] 938 0.0
gi|24645914|ref|NP_524317.2| CG17228-PC, isoform C [Drosoph... 938 0.0
gi|6179901|gb|AAF05703.1| homeodomain transcription factor ... 938 0.0
gi|158184|gb|AAA28841.1| Pros protein 932 0.0
gi|28571646|ref|NP_788636.1| CG17228-PD, isoform D [Drosoph... 931 0.0
gi|28571644|ref|NP_731565.2| CG17228-PA, isoform A [Drosoph... 931 0.0
gi|54639735|gb|EAL29137.1| GA14403-PA [Drosophila pseudoobs... 553 e-155
gi|14285684|sp|Q9U6A1|PROS_DROVI Protein prospero >gi|62744... 542 e-152
gi|31201317|ref|XP_309606.1| ENSANGP00000010936 [Anopheles ... 352 6e-95
gi|55244567|gb|EAA05345.2| ENSANGP00000010936 [Anopheles ga... 352 6e-95
gi|48095851|ref|XP_392355.1| similar to ENSANGP00000010936 ... 328 9e-88

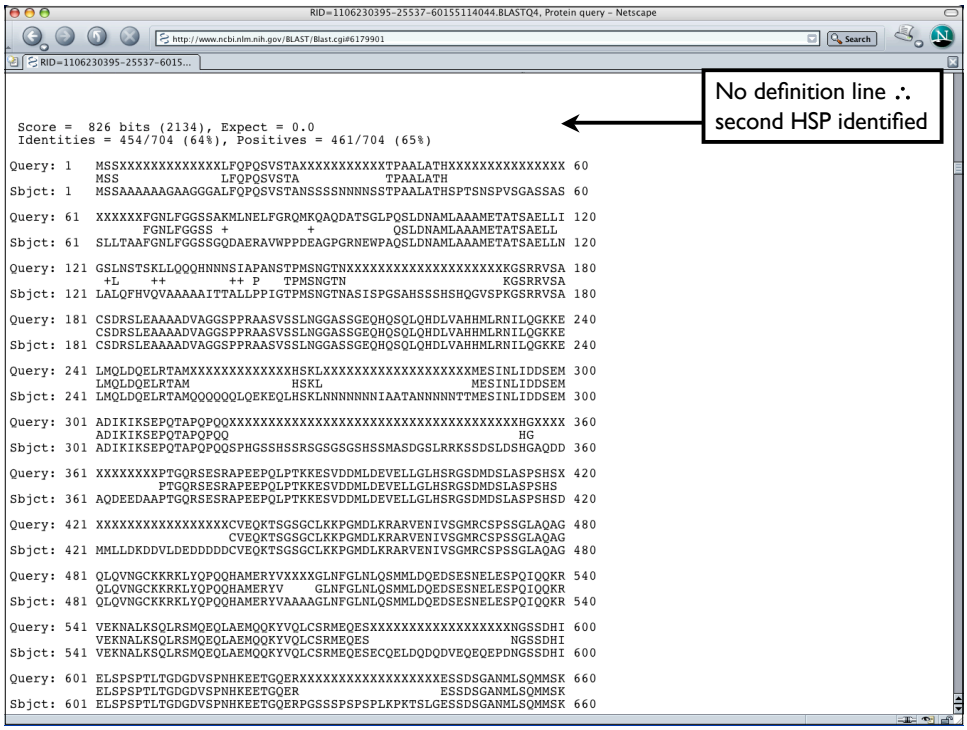

NCBI Sequence-Structure Alignment Visualization Service - Netscape
Related Structures
PubMed BLAST OMIM Taxonomy Structure Help?
Query: Protein query
BLAST RID: 1106230395-25537-60155114044.BLASTQ4
List subset grouped by All MMDB sorted by Blast e_value in Graphics
BLAST found 1 related structures. Click on the alignment figure to see alignments and 3D structure.
Query
CDs
Prod1
Prod1
Structure
E Value
1MIJ A 1
2e-80

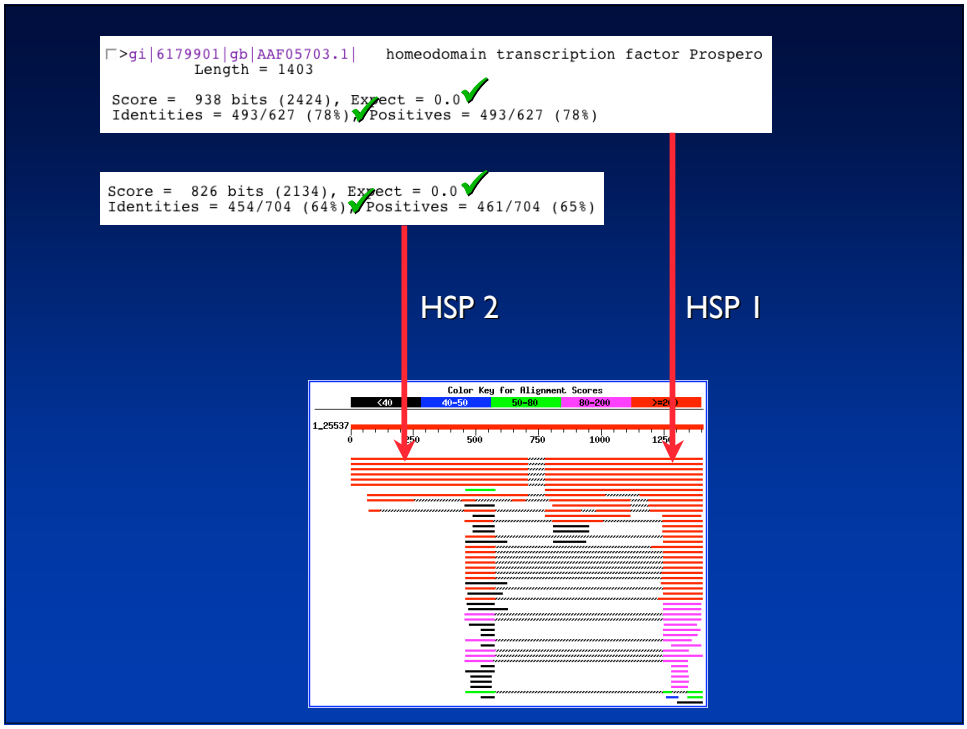

NCBI Sequence-Structure Alignment Visualization Service - Netscape
Related Structures
PubMed BLAST OMIM Taxonomy Structure Help?
Query: Protein query
Structure: 1MIJ Chain A, Crystal Structure Of The Homeo-Prospero Domain Of D. Melanogaster Prospero.
MMDB: 1MIJ_A Reference: PubMed
View 3D Structure with Cn3D (To display structure, download Cn3D)
query 1245 SSTLTPMHLRKAKLMFFWVRYPSSAVLKMYFPDIKFNKNNTAQLVKWFSNFREFYYIQME 1304
1MIJ_A 1 SSTLTPXHLRKAKLXFFWVRYPSSAVLKXYFPDIKFNKNNTAQLVKWFSNFREFYYIQXE 60
query 1305 KYARQAVTEGIKTPDDLLIAGDSELYRVLNLHYNRNNHIEVPQNFRFVVESTLREFFRAI 1364
1MIJ_A 61 KYARQAVTEGIKTPDDLLIAGDSELYRVLNLHYNRNNHIEVPQNFRFVVESTLREFFRAI 120
query 1365 QGGKDTEQSWKKSIYKIISRMDDPVPEYFKSP 1396
1MIJ_A 121 QGGKDTEQSWKKSIYKIISRXDDPVPEYFKSP 152
Score(bits) = 303, E_value = 2e-80
Aligned Length = 152 , Sequence Identity = 96 %


RID=1106230395-25537-60155114044.BLASTQ4, Protein query - Netscape
Descending score order
0.0 means ≤ 10-1000
6e-95 = 6 x 10-95
S Structure
G Gene


RID=1106230395-25537-60155114044.BLASTQ4, Protein query - Netscape
Accept (for now)
Reject
Alignments
Get selected sequences
Select all
Deselect all


RID=1106230395-25537-60155114044.BLASTQ4, Protein query - Netscape
≥ 25% for proteins
≥ 70% for nucleotides
– Gap
x Low-Complexity


RID=1106230395-25537-60155114044.BLASTQ4, Protein query - Netscape
No definition line ∴
second HSP identified



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!*

Database Searching Artifacts

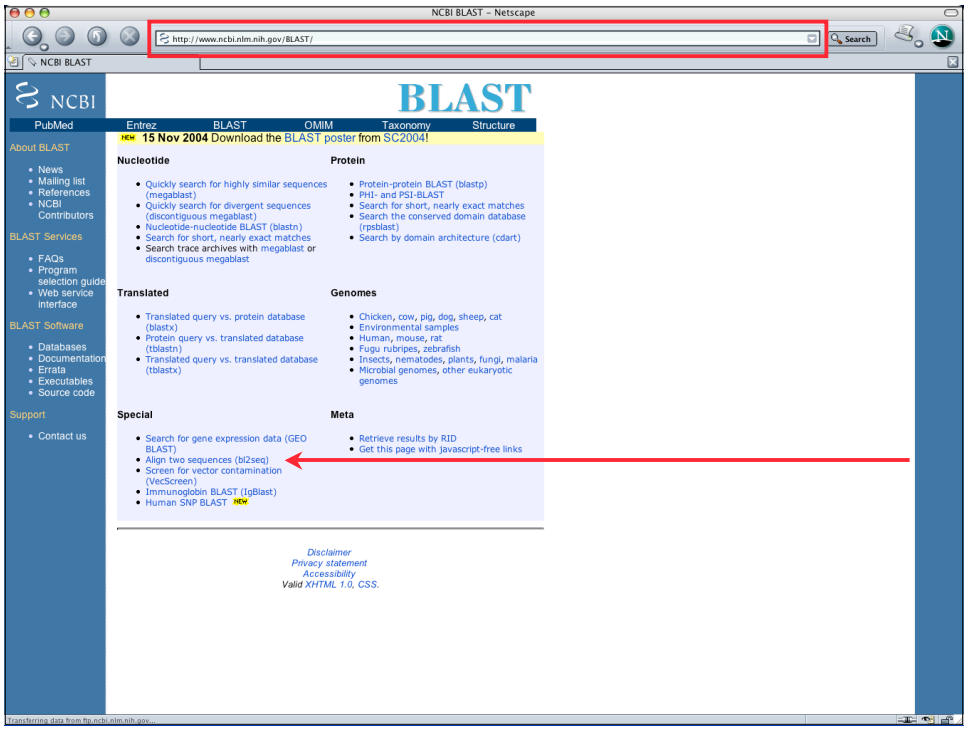
- Low-complexity regions
  - Nucleotide searches: removed with DUST (➔ N)
  - Protein searches: removed with SEG (➔ X)
- Repetitive elements
  - LINE, SINE, Alu
  - Automatic masking "experimental and still under development"
  - RepeatMasker
    http://www.repeatmasker.org

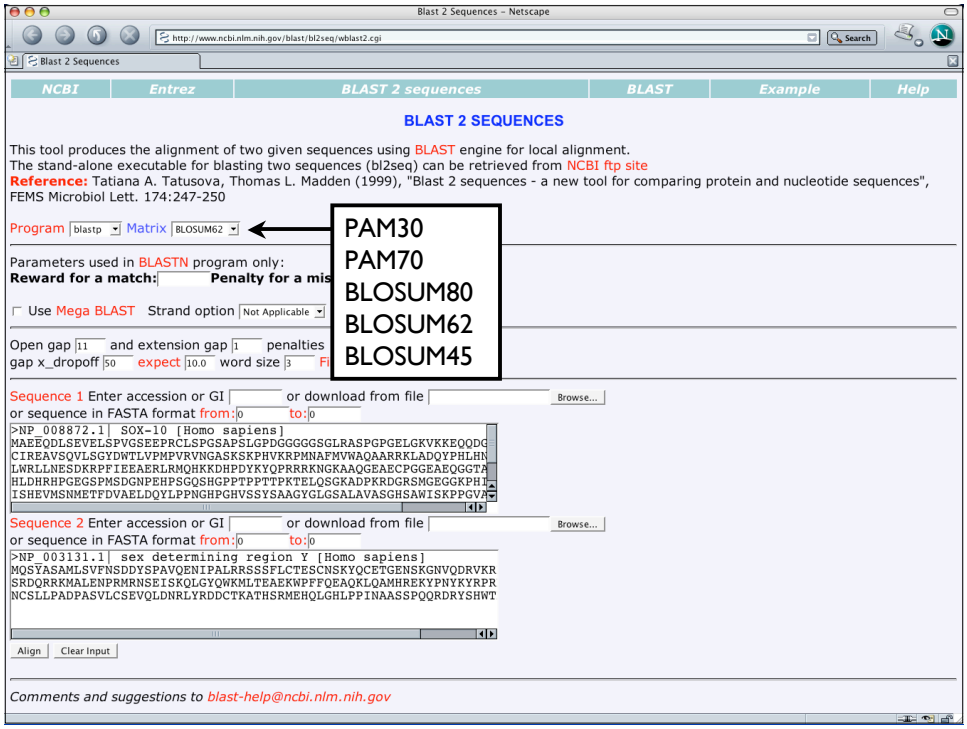
Database Searching Artifacts

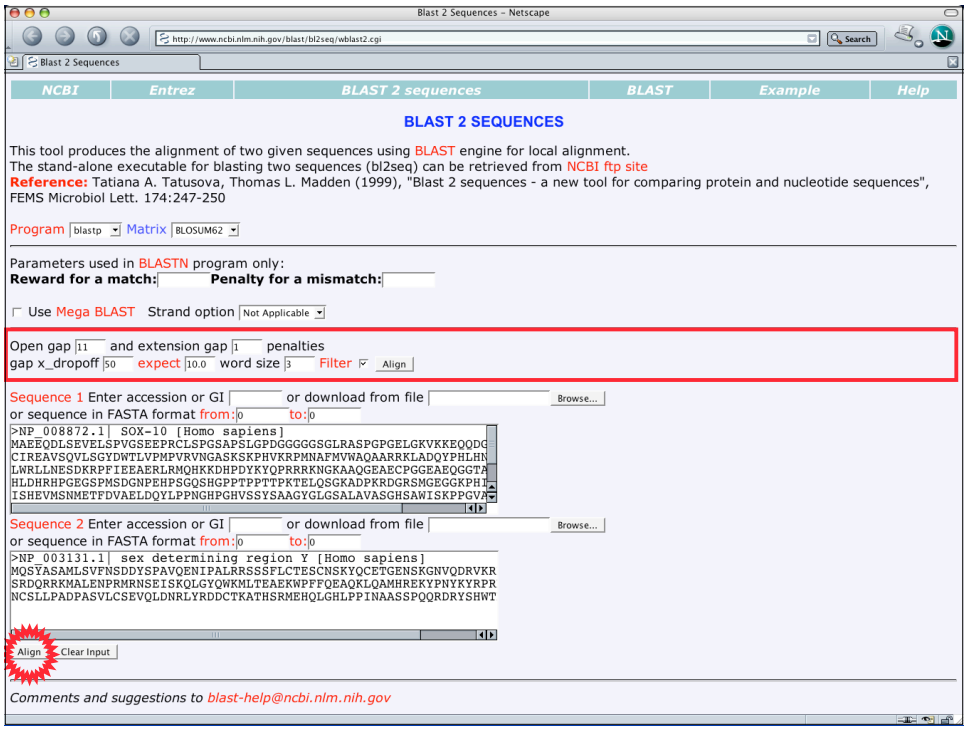
- Low-quality sequence hits
  - Expressed sequence tags (ESTs)
  - Single-pass sequence reads from large-scale sequencing (possibly with vector contaminants)

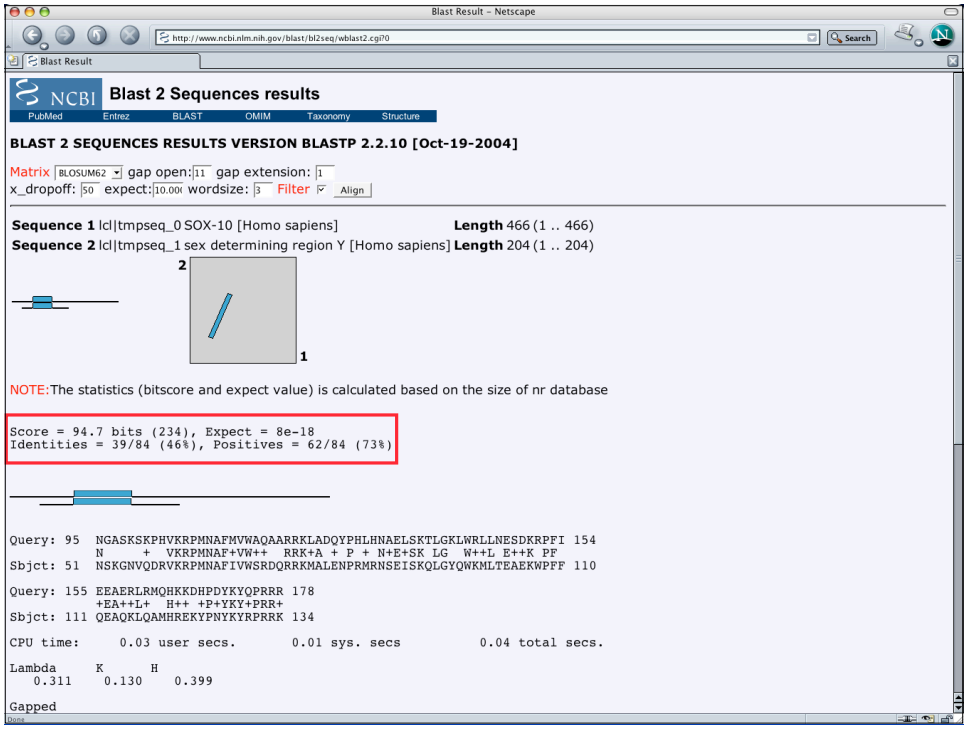
BLAST 2 Sequences

- 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
- Implementations
  - NCBI Web interface
  - bl2seq downloadable executable
    ftp://ncbi.nlm.nih.gov/blast/executables/




Blast 2 Sequences - Netscape
http://www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi
Search
Blast 2 Sequences
NCBI
Entrez
BLAST 2 sequences
BLAST
Example
Help
BLAST 2 SEQUENCES
This tool produces the alignment of two given sequences using BLAST engine for local alignment.
The stand-alone executable for blasting two sequences (bl2seq) can be retrieved from NCBI ftp site
Reference: Tatiana A. Tatusova, Thomas L. Madden (1999), "Blast 2 sequences - a new tool for comparing protein and nucleotide sequences",
FEMS Microbiol Lett. 174:247-250
Program blastp Matrix BLOSUM62
PAM30
PAM70
BLOSUM80
BLOSUM62
BLOSUM45
Parameters used in BLASTN program only:
Reward for a match:
Penalty for a mis
Use Mega BLAST
Strand option Not Applicable
Open gap 11 and extension gap 1 penalties
gap x_dropoff 50 expect 10.0 word size 3
Sequence 1 Enter accession or GI
or download from file
Browse...
or sequence in FASTA format from: 0 to: 0
>NP_008872.1| SOX-10 [Homo sapiens]
MAEEQDLSEVELSPVGSEEPRCLSPGSAPSLGPDGGGGGSGLRASPGPGELGKVKKEQQDG
CIREAVSQVLSGYDWTLVPMPVRVNGASKSKPHVKRPMNAFMVWAQAARRKLADQYPHLHN
LWRLLNESDKRPFIEEAERLRMQHKKDHPDYKYQPRRRKNGKAAQGEAECPGGEAEQGGTA
HLDHRHPGEGSPMSDGNPEHPSGQSHGPPTPPTTPKTELQSGKADPKRDGRSMGEGGKPHT
ISHEVMSNMETFDVAELDQYLPPNGHPGHVSSYSAAGYGLGSALAVASGHSAWISKPPGVA
Sequence 2 Enter accession or GI
or download from file
Browse...
or sequence in FASTA format from: 0 to: 0
>NP_003131.1| sex determining region Y [Homo sapiens]
MQSYASAMLSVFNSDDYSPAVQENIPALRRSSSFLCTESCNSKYQCETGENSKGNVQDRVKR
SRDQRRKMALENPRMRNSEISKQLGYQWKMLTEAEKWPFFQEAQKLQAMHREKYPNYKYRPR
NCSLLPADPASVLCSEVQLDNRLYRDDCTKATHSRMEHQLGHLPPINAASSPQQRDRYSHWT
Align
Clear Input
Comments and suggestions to blast-help@ncbi.nlm.nih.gov


Blast 2 Sequences - Netscape
http://www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi
Search
Blast 2 Sequences
NCBI
Entrez
BLAST 2 sequences
BLAST
Example
Help
BLAST 2 SEQUENCES
This tool produces the alignment of two given sequences using BLAST engine for local alignment.
The stand-alone executable for blasting two sequences (bl2seq) can be retrieved from NCBI ftp site
Reference: Tatiana A. Tatusova, Thomas L. Madden (1999), "Blast 2 sequences - a new tool for comparing protein and nucleotide sequences",
FEMS Microbiol Lett. 174:247-250
Program blastp Matrix BLOSUM62
Parameters used in BLASTN program only:
Reward for a match:
Penalty for a mismatch:
Use Mega BLAST
Strand option Not Applicable
Open gap 11 and extension gap 1 penalties
gap x_dropoff 50 expect 10.0 word size 3 Filter Align
Sequence 1 Enter accession or GI
or download from file
Browse...
or sequence in FASTA format from: 0 to: 0
>NP_008872.1| SOX-10 [Homo sapiens]
MAEEQDLSEVELSPVGSEEPRCLSPGSAPSLGPDGGGGGSGLRASPGPGELGKVKKEQQDG
CIREAVSQVLSGYDWTLVPMPVRVNGASKSKPHVKRPMNAFMVWAQAARRKLADQYPHLHN
LWRLLNESDKRPFIEEAERLRMQHKKDHPDYKYQPRRRKNGKAAQGEAECPGGEAEQGGTA
HLDHRHPGEGSPMSDGNPEHPSGQSHGPPTPPTTPKTELQSGKADPKRDGRSMGEGGKPHT
ISHEVMSNMETFDVAELDQYLPPNGHPGHVSSYSAAGYGLGSALAVASGHSAWISKPPGVA
Sequence 2 Enter accession or GI
or download from file
Browse...
or sequence in FASTA format from: 0 to: 0
>NP_003131.1| sex determining region Y [Homo sapiens]
MQSYASAMLSVFNSDDYSPAVQENIPALRRSSSFLCTESCNSKYQCETGENSKGNVQDRVKR
SRDQRRKMALENPRMRNSEISKQLGYQWKMLTEAEKWPFFQEAQKLQAMHREKYPNYKYRPR
NCSLLPADPASVLCSEVQLDNRLYRDDCTKATHSRMEHQLGHLPPINAASSPQQRDRYSHWT
Align
Clear Input
Comments and suggestions to blast-help@ncbi.nlm.nih.gov



MegaBLAST

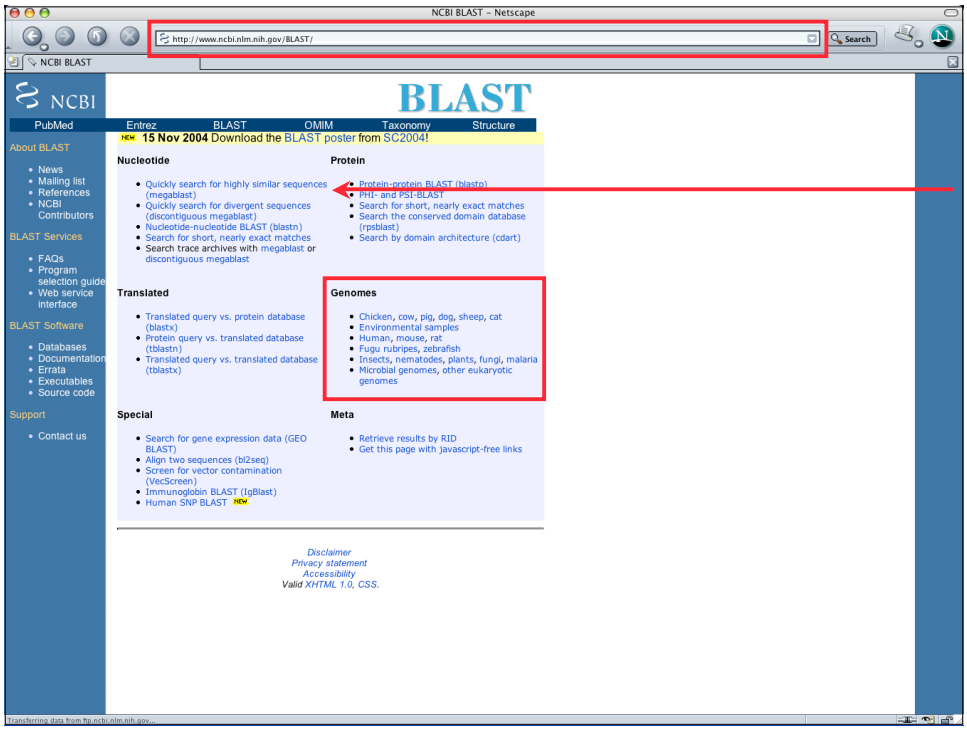
- Optimized for aligning very long and/or highly-similar sequences
- Good for batch nucleotide searches
- Search targets include
  - Entire eukaryotic genomes
  - Complete chromosomes and contigs from RefSeq
- Run speeds approximately 10 times faster than BLASTN
  - Adjusted word size
  - Different gap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

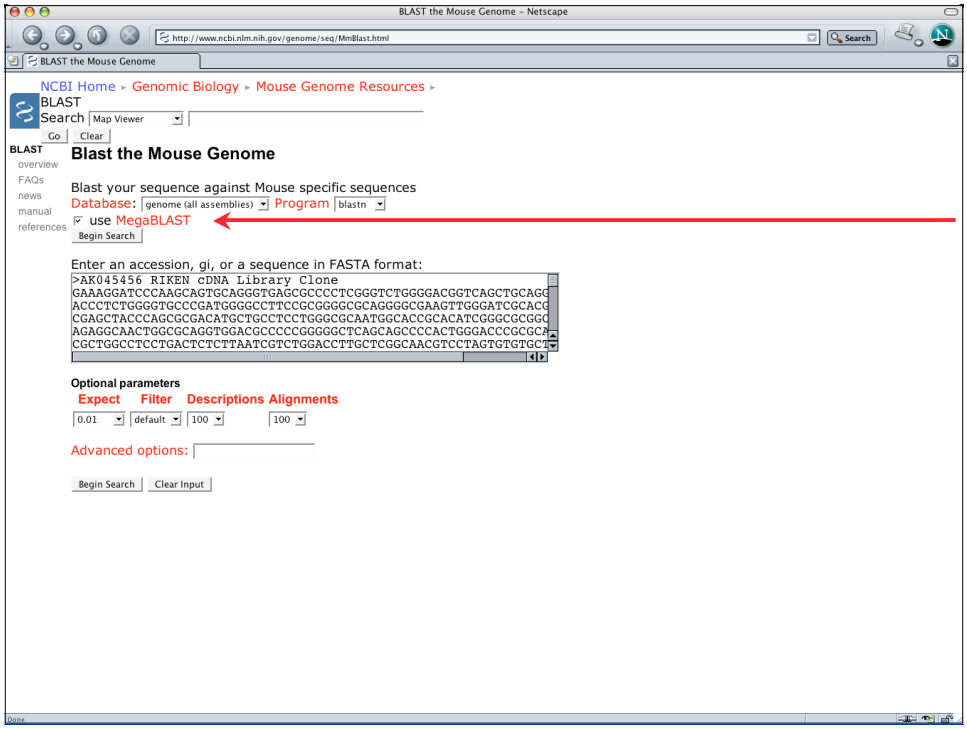
BLASTN *vs.* MegaBLAST

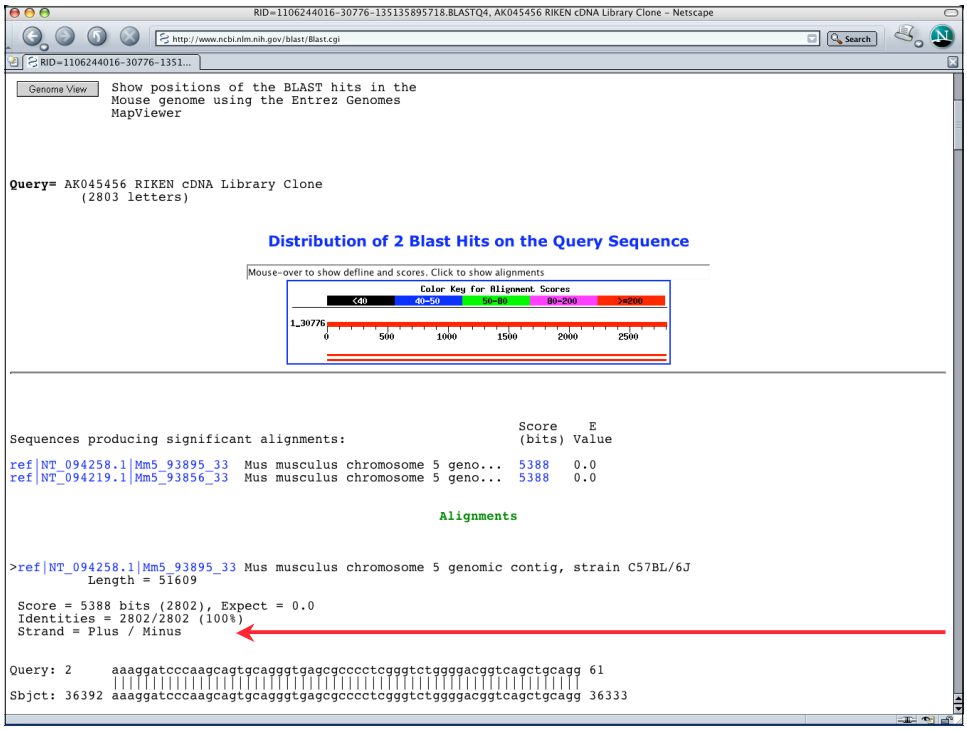
- Word size
  - BLASTN default = 11
  - MegaBLAST default = 28
- *Non-affine* gap penalties

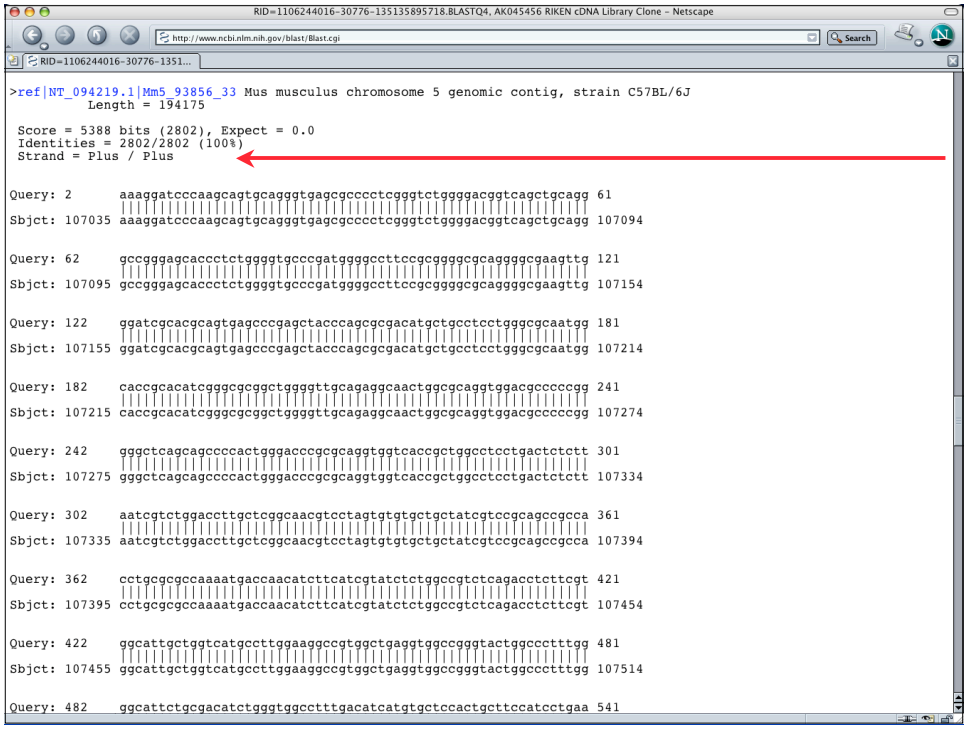
Deduction for a gap = $r/2 - q$

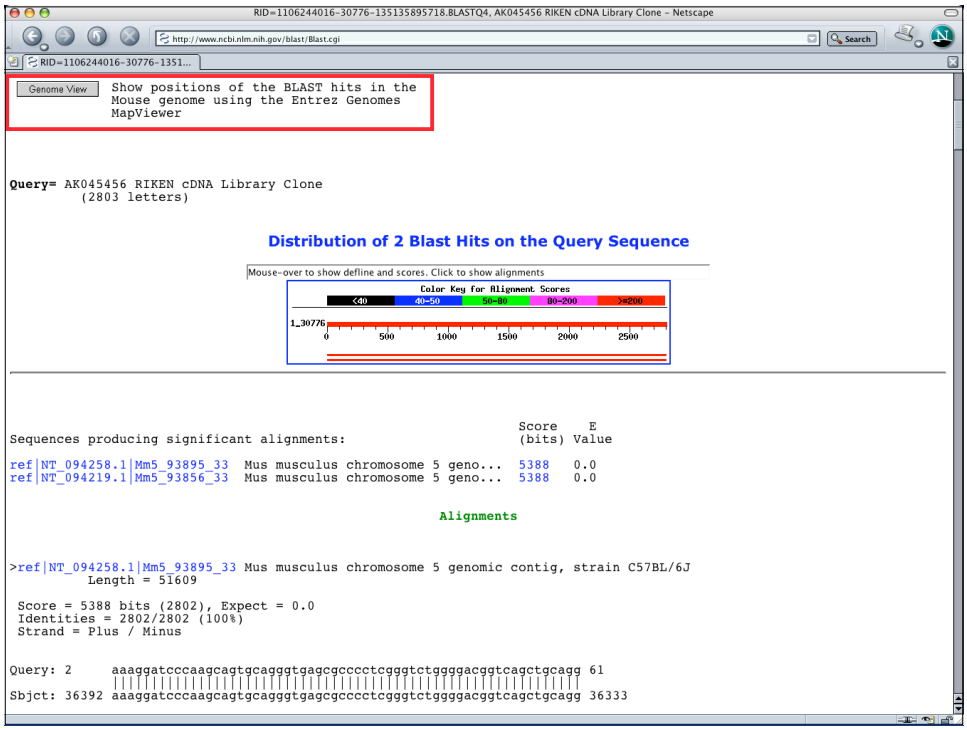
where r = match reward (default 1)
q = mismatch penalty (default -2)
and **no penalty for opening the gap**

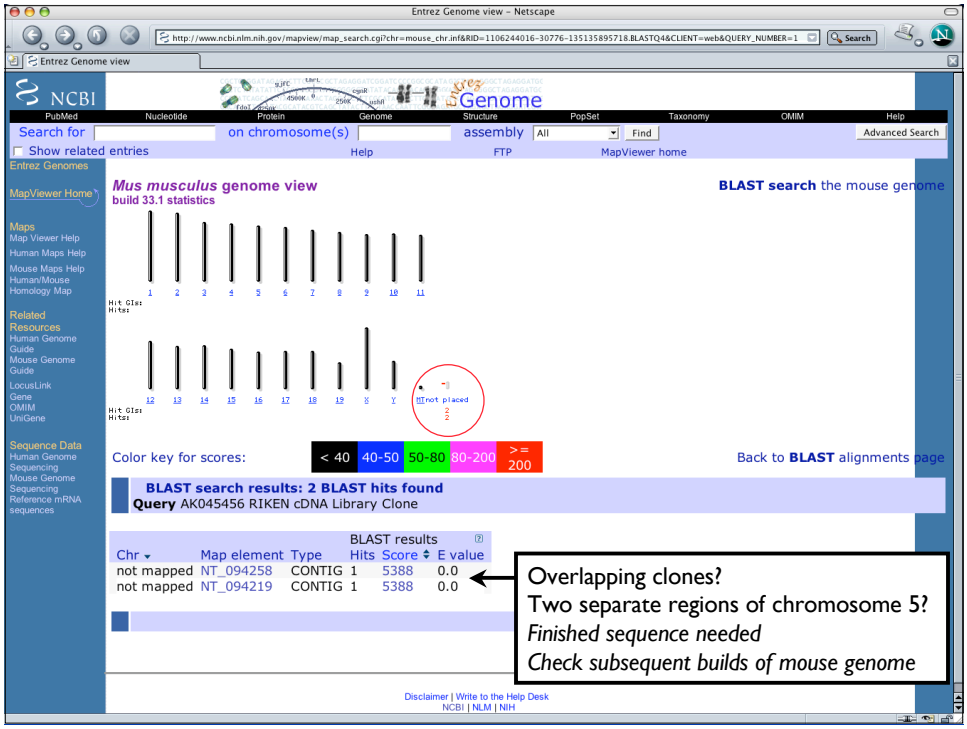



BLAST the Mouse Genome - Netscape
http://www.ncbi.nlm.nih.gov/genome/seq/MmBlast.html
BLAST the Mouse Genome
NCBI Home ▸ Genomic Biology ▸ Mouse Genome Resources ▸
BLAST Search Map Viewer
Go Clear
BLAST
overview
FAQs
news
manual
references
Blast the Mouse Genome
Blast your sequence against Mouse specific sequences
Database: genome (all assemblies) Program blastn
use MegaBLAST
Begin Search
Enter an accession, gi, or a sequence in FASTA format:
>AK045456 RIKEN cDNA Library Clone
GAAAGGATCCCAAGCAGTGCAGGGTGAGCGCCCCTCGGGTCTGGGGACGGTCAGCTGCAGG
ACCCTCTGGGGTGCCCGATGGGGCCTTCCGCGGGGCGCAGGGGCGAAGTTGGGATCGCACC
CGAGCTACCCAGCGCGACATGCTGCCTCCTGGGCGCAATGGCACCGCACATCGGGCGCGGG
AGAGGCAACTGGCGCAGGTGGACGCCCCCGGGGGCTCAGCAGCCCCACTGGGACCCGCGCA
CGCTGGCCTCCTGACTCTCTTAATCGTCTGGACCTTGCTCGGCAACGTCCTAGTGTGTGCT
Optional parameters
Expect Filter Descriptions Alignments
0.01 default 100 100
Advanced options:
Begin Search Clear Input


RID=1106244016-30776-135135895718.BLASTQ4, AK045456 RIKEN cDNA Library Clone - Netscape
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi
RID=1106244016-30776-1351...
Genome View
Show positions of the BLAST hits in the Mouse genome using the Entrez Genomes MapViewer
Query= AK045456 RIKEN cDNA Library Clone
(2803 letters)
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
1_30776
0 500 1000 1500 2000 2500
Sequences producing significant alignments: Score (bits) E Value
ref|NT_094258.1|Mm5_93895_33 Mus musculus chromosome 5 geno... 5388 0.0
ref|NT_094219.1|Mm5_93856_33 Mus musculus chromosome 5 geno... 5388 0.0
Alignments
>ref|NT_094258.1|Mm5_93895_33 Mus musculus chromosome 5 genomic contig, strain C57BL/6J
Length = 51609
Score = 5388 bits (2802), Expect = 0.0
Identities = 2802/2802 (100%)
Strand = Plus / Minus
Query: 2 aaaggatcccaagcagtgcagggtgagcgcccctcgggtctggggacggtcagctgcagg 61
Sbjct: 36392 aaaggatcccaagcagtgcagggtgagcgcccctcgggtctggggacggtcagctgcagg 36333


RID=1106244016-30776-135135895718.BLASTQ4, AK045456 RIKEN cDNA Library Clone - Netscape
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi
Search
RID=1106244016-30776-1351...
>ref|NT_094219.1|Mm5_93856_33 Mus musculus chromosome 5 genomic contig, strain C57BL/6J
Length = 194175
Score = 5388 bits (2802), Expect = 0.0
Identities = 2802/2802 (100%)
Strand = Plus / Plus
Query: 2 aaaggatcccaagcagtgcagggtgagcgcccctcgggtctggggacggtcagctgcagg 61
Sbjct: 107035 aaaggatcccaagcagtgcagggtgagcgcccctcgggtctggggacggtcagctgcagg 107094
Query: 62 gccgggagcaccctctggggtgcccgatggggccttccgcggggcgcaggggcgaagttg 121
Sbjct: 107095 gccgggagcaccctctggggtgcccgatggggccttccgcggggcgcaggggcgaagttg 107154
Query: 122 ggatcgcacgcagtgagcccgagctacccagcgcgacatgctgcctcctgggcgcaatgg 181
Sbjct: 107155 ggatcgcacgcagtgagcccgagctacccagcgcgacatgctgcctcctgggcgcaatgg 107214
Query: 182 caccgcacatcgggcgcggctggggttgcagaggcaactggcgcaggtggacgcccccgg 241
Sbjct: 107215 caccgcacatcgggcgcggctggggttgcagaggcaactggcgcaggtggacgcccccgg 107274
Query: 242 gggctcagcagccccactgggacccgcgcaggtggtcaccgctggcctcctgactctctt 301
Sbjct: 107275 gggctcagcagccccactgggacccgcgcaggtggtcaccgctggcctcctgactctctt 107334
Query: 302 aatcgtctggaccttgctcggcaacgtcctagtgtgtgctgctatcgtccgcagccgcca 361
Sbjct: 107335 aatcgtctggaccttgctcggcaacgtcctagtgtgtgctgctatcgtccgcagccgcca 107394
Query: 362 cctgcgcgccaaaatgaccaacatcttcatcgtatctctggccgtctcagacctcttcgt 421
Sbjct: 107395 cctgcgcgccaaaatgaccaacatcttcatcgtatctctggccgtctcagacctcttcgt 107454
Query: 422 ggcattgctggtcatgccttggaaggccgtggctgaggtggccgggtactggccctttgg 481
Sbjct: 107455 ggcattgctggtcatgccttggaaggccgtggctgaggtggccgggtactggccctttgg 107514
Query: 482 ggcattctgcgacatctgggtggcctttgacatcatgtgctccactgcttccatcctgaa 541


RID=1106244016-30776-135135895718.BLASTQ4, AK045456 RIKEN cDNA Library Clone - Netscape
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi
Search
RID=1106244016-30776-1351...
Genome View
Show positions of the BLAST hits in the Mouse genome using the Entrez Genomes MapViewer
Query= AK045456 RIKEN cDNA Library Clone
(2803 letters)
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
1_30776
0
500
1000
1500
2000
2500
Sequences producing significant alignments:
Score (bits)
E Value
ref|NT_094258.1|Mm5_93895_33 Mus musculus chromosome 5 geno... 5388 0.0
ref|NT_094219.1|Mm5_93856_33 Mus musculus chromosome 5 geno... 5388 0.0
Alignments
>ref|NT_094258.1|Mm5_93895_33 Mus musculus chromosome 5 genomic contig, strain C57BL/6J
Length = 51609
Score = 5388 bits (2802), Expect = 0.0
Identities = 2802/2802 (100%)
Strand = Plus / Minus
Query: 2 aaaggatcccaagcagtgcagggtgagcgcccctcgggtctggggacggtcagctgcagg 61
Sbjct: 36392 aaaggatcccaagcagtgcagggtgagcgcccctcgggtctggggacggtcagctgcagg 36333



Overview

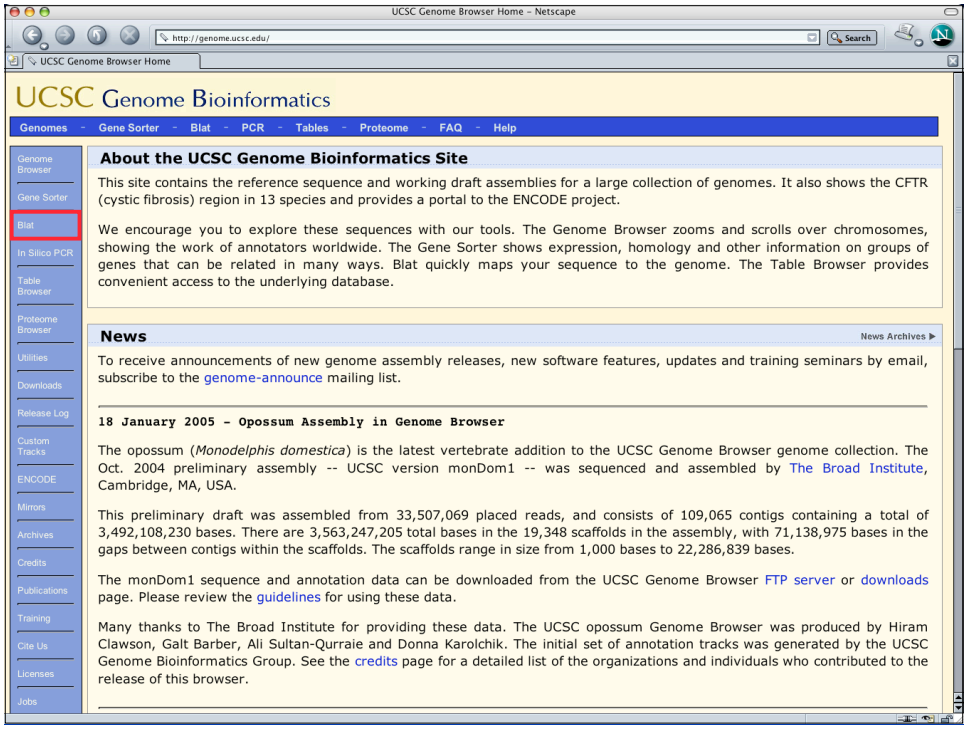
- Week 4: Comparative methods and concepts
  - Similarity *vs.* Homology
  - Global *vs.* Local Alignments
  - Scoring Matrices
  - BLAST
  - BLAT
- Week 5: Predictive methods and concepts
  - Profiles, patterns, motifs, and domains
  - Secondary structure prediction
  - Structures: VAST, Cn3D, and *de novo* prediction

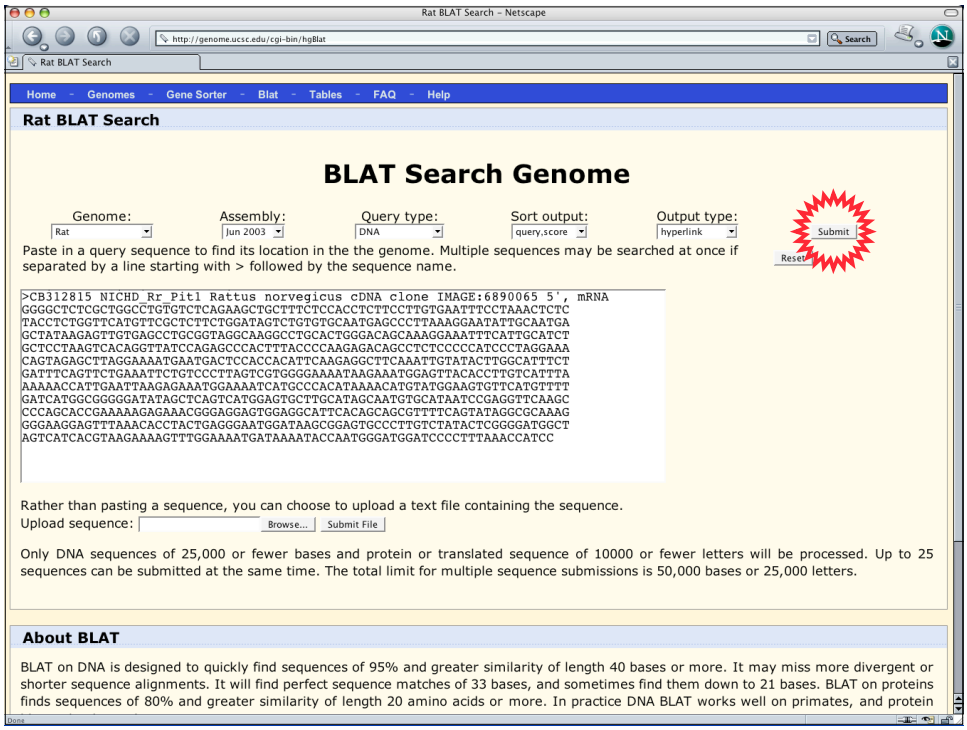
BLAT

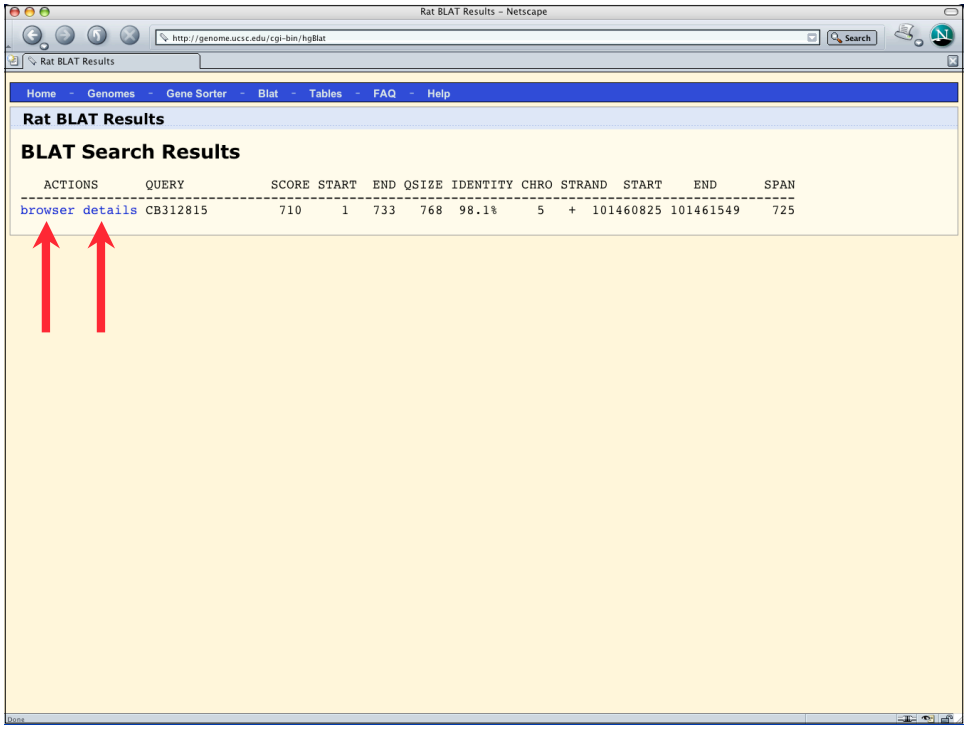
- "BLAST-Like Alignment Tool"
- Designed to rapidly-align longer nucleotide sequences ($L \geq 40$) having > 95% sequence similarity
- Can find exact matches reliably down to $L = 33$
- Method of choice when looking for exact matches in nucleotide databases
- 500 times faster for mRNA/DNA searches
-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

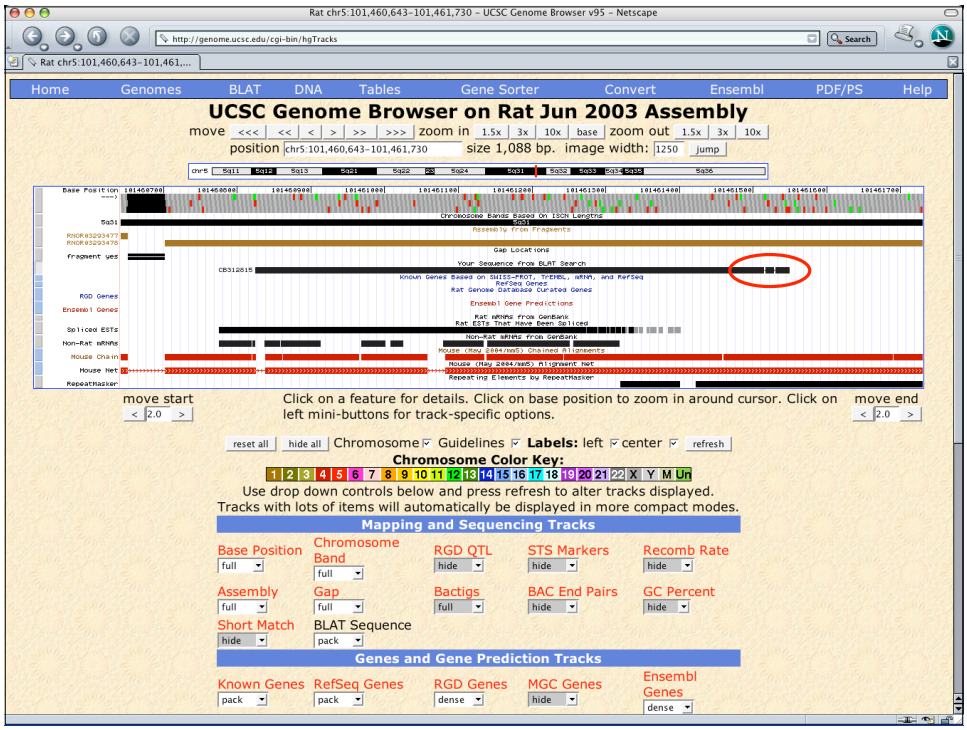
When to Use BLAT

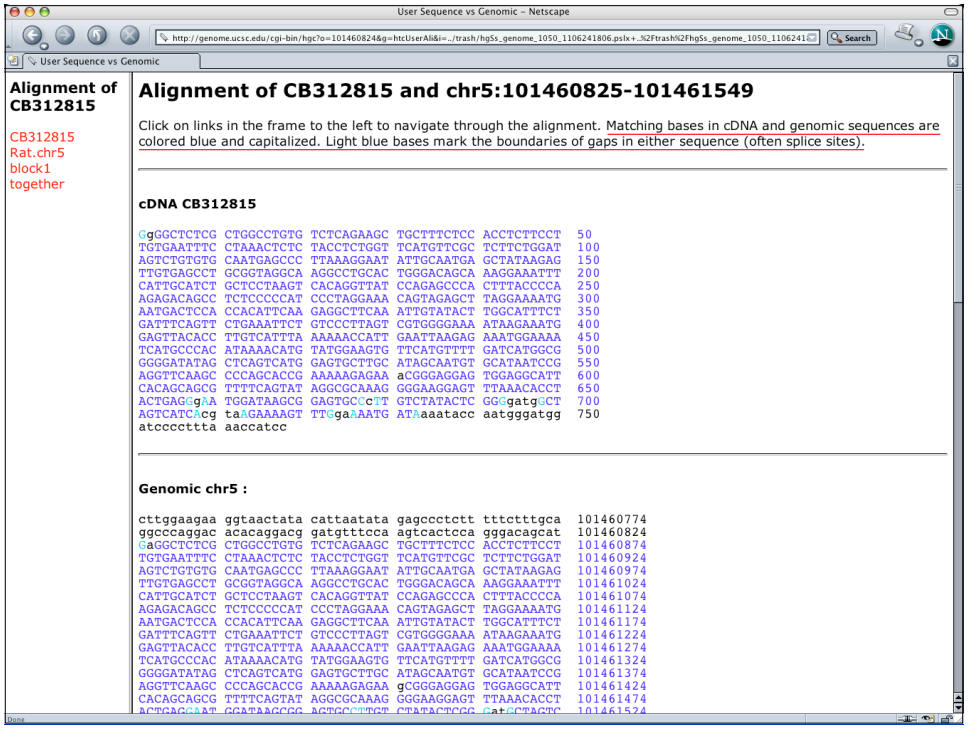
- To characterize an unknown gene or sequence fragment
  - Find its genomic coordinates
  - Determine gene structure (the presence and position of exons)
  - Identify markers of interest in the vicinity of a sequence
- To find highly-similar sequences
  - Identify gene family members
  - Identify putative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

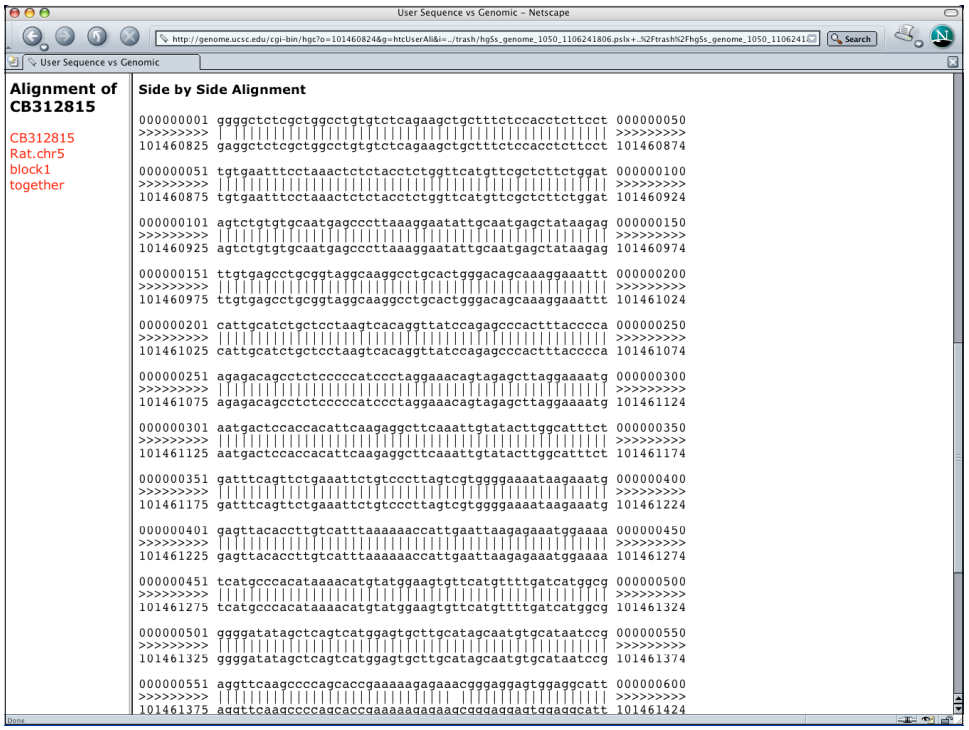

UCSC Genome Browser Home – Netscape
http://genome.ucsc.edu/
Search
UCSC Genome Browser Home
UCSC Genome Bioinformatics
Genomes - Gene Sorter - Blat - PCR - Tables - Proteome - FAQ - Help
Genome Browser
Gene Sorter
Blat
In Silico PCR
Table Browser
Proteome Browser
Utilities
Downloads
Release Log
Custom Tracks
ENCODE
Mirrors
Archives
Credits
Publications
Training
Cite Us
Licenses
Jobs
About the UCSC Genome Bioinformatics Site
This site contains the reference sequence and working draft assemblies for a large collection of genomes. It also shows the CFTR (cystic fibrosis) region in 13 species and 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.
News
News Archives
To receive announcements of new genome assembly releases, new software features, updates and training seminars by email, subscribe to the genome-announce mailing list.
18 January 2005 - Opossum Assembly in Genome Browser
The opossum (Monodelphis domestica) is the latest vertebrate addition to the UCSC Genome Browser genome collection. The Oct. 2004 preliminary assembly -- UCSC version monDom1 -- was sequenced and assembled by The Broad Institute, Cambridge, MA, USA.
This preliminary draft was assembled from 33,507,069 placed reads, and consists of 109,065 contigs containing a total of 3,492,108,230 bases. There are 3,563,247,205 total bases in the 19,348 scaffolds in the assembly, with 71,138,975 bases in the gaps between contigs within the scaffolds. The scaffolds range in size from 1,000 bases to 22,286,839 bases.
The monDom1 sequence and annotation data can be downloaded from the UCSC Genome Browser FTP server or downloads page. Please review the guidelines for using these data.
Many thanks to The Broad Institute for providing these data. The UCSC opossum Genome Browser was produced by Hiram Clawson, Galt Barber, Ali Sultan-Qurraie and Donna Karolchik. The initial set of annotation tracks was generated by the UCSC Genome Bioinformatics Group. See the credits page for a detailed list of the organizations and individuals who contributed to the release of this browser.


Rat BLAT Search – Netscape
http://genome.ucsc.edu/cgi-bin/hgBlat
Search
Rat BLAT Search
Home - Genomes - Gene Sorter - Blat - Tables - FAQ - Help
Rat BLAT Search
BLAT Search Genome
Genome:
Rat
Assembly:
Jun 2003
Query type:
DNA
Sort output:
query,score
Output type:
hyperlink
Submit
Reset
Paste in a query sequence to find its location in the the genome. Multiple sequences may be searched at once if separated by a line starting with > followed by the sequence name.
>CB312815 NICHD_Rr_Pit1 Rattus norvegicus cDNA clone IMAGE:6890065 5', mRNA
GGGGCTCTCGCTGGCCTGTGTCTCAGAAGCTGCTTTCTCCACCTCTTCCTTGTGAATTTCCTAAACTCTC
TACCTCTGGTTCATGTTCGCTCTTCTGGATAGTCTGTGTGCAATGAGCCCTTAAAGGAATATTGCAATGA
GCTATAAGAGTTGTGAGCCTGCGGTAGGCAAGGCCTGCACTGGGACAGCAAAGGAAATTTCATTGCATCT
GCTCCTAAGTCACAGGTTATCCAGAGCCCACTTTACCCCAAGAGACAGCCTCTCCCCCATCCCTAGGAAA
CAGTAGAGCTTAGGAAAATGAATGACTCCACCACATTCAAGAGGCTTCAAATTGTATACTTGGCATTTCT
GATTTCAGTTCTGAAATTCTGTCCCTTAGTCGTGGGGAAAATAAGAAATGGAGTTACACCTTGTCATTTA
AAAAACCATTGAATTAAGAGAAATGGAAAATCATGCCCACATAAAACATGTATGGAAGTGTTCATGTTTT
GATCATGGCGGGGGATATAGCTCAGTCATGGAGTGCTTGCATAGCAATGTGCATAATCCGAGGTTCAAGC
CCCAGCACCGAAAAAGAGAAACGGGAGGAGTGGAGGCATTCACAGCAGCGTTTTCAGTATAGGCGCAAAG
GGGAAGGAGTTTAAACACCTACTGAGGGAATGGATAAGCGGAGTGCCCTTGTCTATACTCGGGGATGGCT
AGTCATCACGTAAGAAAAGTTTGGAAAATGATAAAATACCAATGGGATGGATCCCCTTTAAACCATCC
Rather than pasting a sequence, you can choose to upload a text file containing the sequence.
Upload sequence:
Browse...
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.
About BLAT
BLAT on DNA is designed to quickly find sequences of 95% and greater similarity of length 40 bases or more. It may miss more divergent or shorter sequence alignments. It will find perfect sequence matches of 33 bases, and sometimes find them down to 21 bases. BLAT on proteins finds sequences of 80% and greater similarity of length 20 amino acids or more. In practice DNA BLAT works well on primates, and protein


Rat BLAT Results - Netscape
http://genome.ucsc.edu/cgi-bin/hgBlat
Search
Rat BLAT Results
Home - Genomes - Gene Sorter - Blat - Tables - 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 + 101460825 101461549 725


Rat chr5:101,460,643-101,461,730 - UCSC Genome Browser v95 - Netscape
http://genome.ucsc.edu/cgi-bin/hgTracks
Search
Home Genomes BLAT DNA Tables Gene Sorter Convert Ensembl PDF/PS Help
UCSC Genome Browser on Rat Jun 2003 Assembly
move zoom in zoom out
position chr5:101,460,643-101,461,730 size 1,088 bp. image width: 1250 jump
Click on a feature for details. Click on base position to zoom in around cursor. Click on left mini-buttons for track-specific options.
move start move end
reset all hide all Chromosome Guidelines Labels: left center refresh
Chromosome Color Key:
Use drop down controls below and press refresh to alter tracks displayed.
Tracks with lots of items will automatically be displayed in more compact modes.
Mapping and Sequencing Tracks
Base Position
Chromosome Band
RGD QTL
STS Markers
Recomb Rate
Assembly
Gap
Bactigs
BAC End Pairs
GC Percent
Short Match
BLAT Sequence
Genes and Gene Prediction Tracks
Known Genes
RefSeq Genes
RGD Genes
MGC Genes
Ensembl Genes


User Sequence vs Genomic – Netscape
User Sequence vs Genomic
Alignment of CB312815
CB312815
Rat.chr5
block1
together
Alignment of CB312815 and chr5:101460825-101461549
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 – Netscape
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

Overview

- Week 4: Comparative methods and concepts
  - Similarity *vs.* Homology
  - Global *vs.* Local Alignments
  - Scoring Matrices
  - BLAST
  - BLAT
- Week 5: Predictive methods and concepts
  - Profiles, patterns, motifs, and domains
  - Secondary structure prediction
  - Structures: VAST, Cn3D, and *de novo* prediction

Further Reading

Altschul, S.F., Boguski, M.S., Gish, W., and Wootton, J.C. 1994. Issues in searching molecular sequence databases. *Nat. Genet.* 6: 119-129. A review of the issues that are of importance in using sequence similarity search programs, including potential pitfalls.

Baxevanis, A.D. Assessing pairwise sequence similarity: BLAST and FASTA. In *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*, third edition (Baxevanis, A.D. and Ouellette, B.F.F., eds.), John Wiley and Sons, 2005. An overview of the methods used to generate pairwise sequence alignments and assess the biological significance of results.

Henikoff, S. and Henikoff, J.G. 2000. Amino acid substitution matrices. *Adv. Protein Chem.* 54: 73-97. A comprehensive review covering the factors critical to the construction of protein scoring matrices.

Korf, I., Yandell, M., and Bedell, J. BLAST. O'Reilly and Associates, 2003. An in-depth treatment of the BLAST algorithm, its applications, as well as installation, hardware, and software considerations. The book provides "documentation" that is not easily found elsewhere.

Pearson, W.R. Finding protein and nucleotide similarities with FASTA. 2003. *Current Protocols in Bioinformatics* 3.9.1-3.9.23. An in-depth discussion of the FASTA algorithm, including worked examples and additional information regarding run options and use scenarios.

Wheeler, D.G. Selecting the right protein scoring matrix. 2003. *Current Protocols in Bioinformatics* 3.5.1-3.5.6. A discussion of PAM, BLOSUM, and specialized scoring matrices, with guidance regarding the proper choice of matrices for particular types of protein-based analyses.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1991. Basic local alignment search tool. *J. Mol. Biol.* 215: 403-410.

Altschul, S.F., Madden T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389-3402.

Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C. 1978. A model of evolutionary change in proteins. *In* Atlas of Protein Sequence and Structure, M.O. Dayhoff, ed., National Biomedical Research Foundation, Washington, 5: 345-352.

Henikoff, S. and Henikoff, J.G. 1991. Automated assembly of protein blocks for database searching. *Nucleic Acids Res.* 19: 6565-6572.

Henikoff, S. and Henikoff, J.G. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* 89: 10915-10919.

Henikoff, S. and Henikoff, J.G. 1993. Performance evaluation of amino acid substitution matrices. *Proteins Struct. Funct. Genet.* 17: 49-61.

Henikoff, S. and Henikoff, J.G. 2000. Amino acid substitution matrices. *Adv. Protein Chem.* 54: 73-97.

Karlin, S. and Altschul, S.F. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA* 87: 2264-2268.

Kent, W.J. 2002. BLAT: the BLAST-like alignment tool. *Genome Res.* 12: 656-664.

Pearson, W.R. 1995. Comparison of methods for searching protein sequence databases. *Protein Sci.* 4: 1145-1160.

Pearson, W.R. 2000. Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.* 132: 185-219.

Pearson, W.R. Finding protein and nucleotide similarities with FASTA. 2003. *Current Protocols in Bioinformatics* 3.9.1-3.9.23.

Pearson, W.R. and Lipman, D.J. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* 85: 2444-2448.

Smith, T.F. and Waterman, M.S. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* 147: 195-197.

Tatusova, T.A. and Madden, T.L. 1999. BLAST2Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbio. Lett.* 174: 247-250.

Wootton, J.C. and Federhen, S. 1993. Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.* 17: 149-163.