Overview

- Week 2: Comparative methods and concepts
  - Similarity *vs.* Homology
  - Global *vs.* Local Alignments
  - Scoring Matrices
  - BLAST
  - BLAT
- Week 3: 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

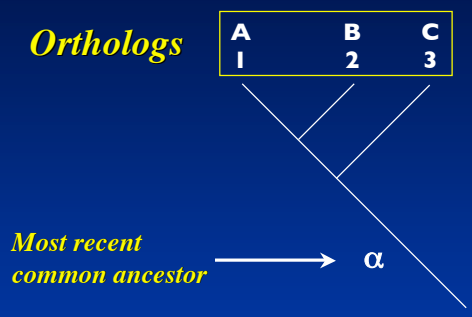
Defining the Terms

- 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*)

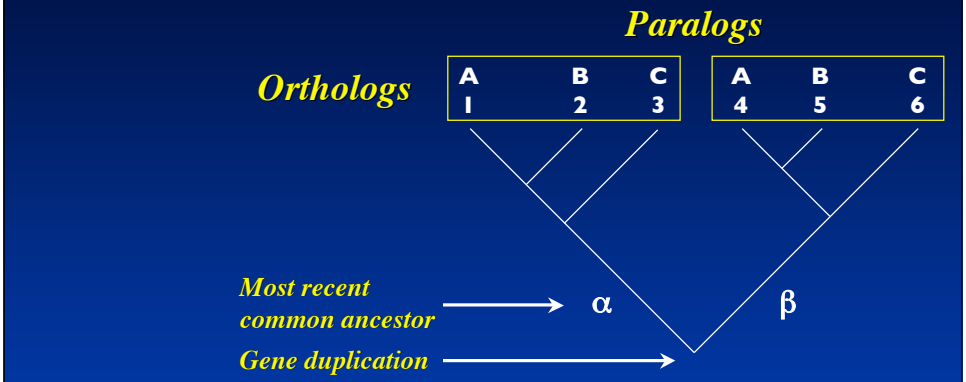
Defining the Terms

- 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



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: Comparative methods and concepts
  - Similarity *vs.* Homology
  - Global *vs.* Local Alignments
  - Scoring Matrices
  - BLAST
  - BLAT
- Week 3: 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 alignment could be generated for any two sequences being compared
- Best for sequences that share some similarity, or for sequences of different lengths

Overview

- Week 2: Comparative methods and concepts
  - Similarity *vs.* Homology
  - Global *vs.* Local Alignments
  - Scoring Matrices
  - BLAST
  - BLAT
- Week 3: Predictive methods and concepts
  - Profiles, patterns, motifs, and domains
  - Secondary structure prediction
  - Structures: VAST, Cn3D, and *de novo* prediction

Scoring Matrices

- Empirical weighting scheme representing physicochemical and biological characteristics of nucleotides and amino acids
  - Side chain structure and chemistry
  - Side chain function
- Amino acid-based examples:
  - Cys/Pro important for structure and function
  - Trp has bulky side chain
  - Lys/Arg have positively-charged side chains

Scoring Matrices

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

Scoring Matrices

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

Matrix Structure: Nucleotides

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

- *Simple match/mismatch scoring scheme:*
  - Match + 5
  - Mismatch – 4
- *Assumes each nucleotide occurs 25% of the time*

Matrix Structure: Proteins

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

BLOSUM62

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 of this analysis
  - 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, not accounting for conserved blocks or motifs
- 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

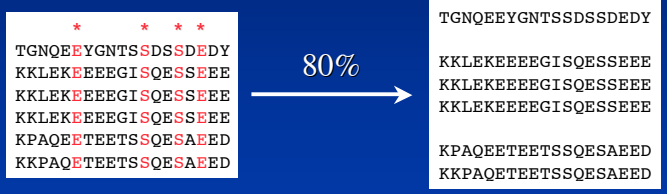
- Small, globular proteins of average composition used to derive matrices
- Errors in PAM 1 are magnified up to PAM 250 (only PAM 1 is based on direct observation)

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 2: Comparative methods and concepts
  - Similarity *vs.* Homology
  - Global *vs.* Local Alignments
  - Scoring Matrices
  - BLAST
  - BLAT
- Week 3: 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 with one another
  - when aligned, have maximal aggregate score (score cannot be improved by extension or trimming)
  - score must be above score threshhold *S*
  - gapped or ungapped
- Results not limited to the "best HSP" for any given sequence pair

BLAST Algorithms

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

Neighborhood Words

Query Word ($W = 3$)

Neighborhood Words

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

Neighborhood Score Threshold ($T = 13$)

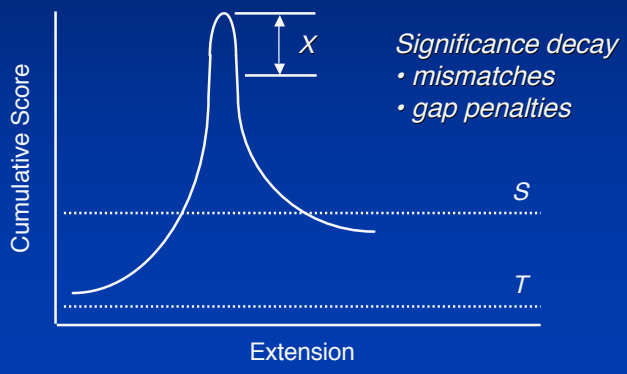
High-Scoring Segment Pairs

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

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

Extension

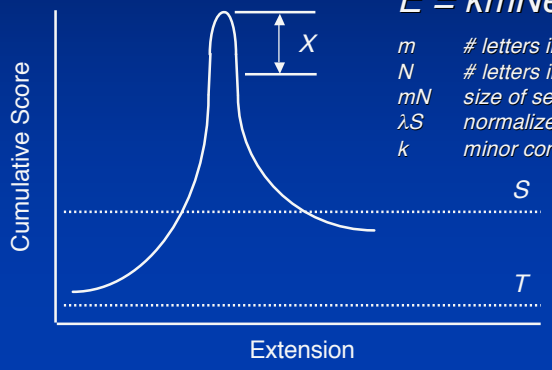
```
Query:   325  SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA  365
              +LA++L    TP+G R++ +W+ +P+ D   + ER   + A
Sbjct:   290  TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA  330
```



Significance decay
- *mismatches*
- *gap penalties*

Scores and Probabilities

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



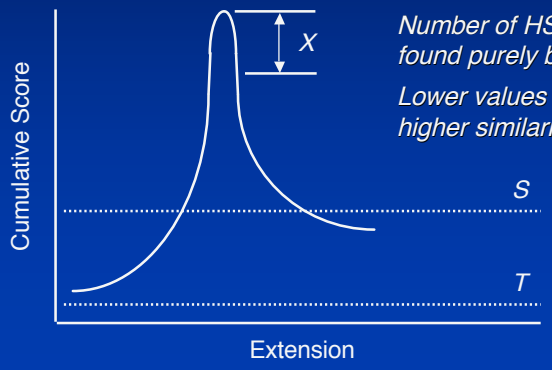
Karlin-Altschul Equation

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

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

Scores and Probabilities

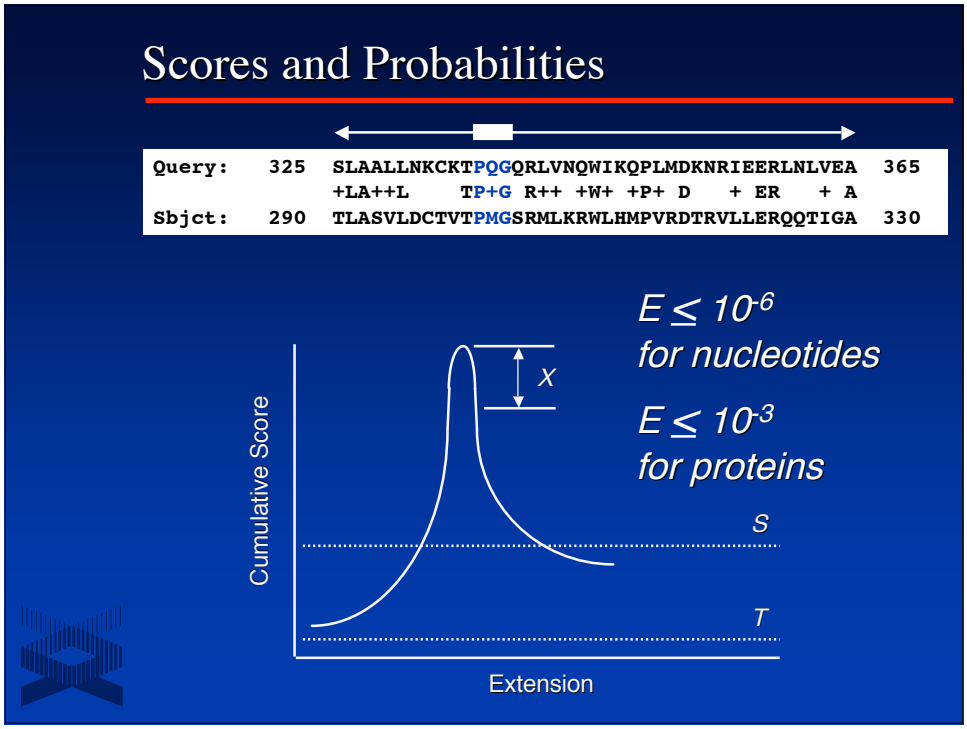
```
Query:   325  SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA  365
              +LA++L    TP+G R++ +W+ +P+ D   + ER   + A
Sbjct:   290  TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA  330
```

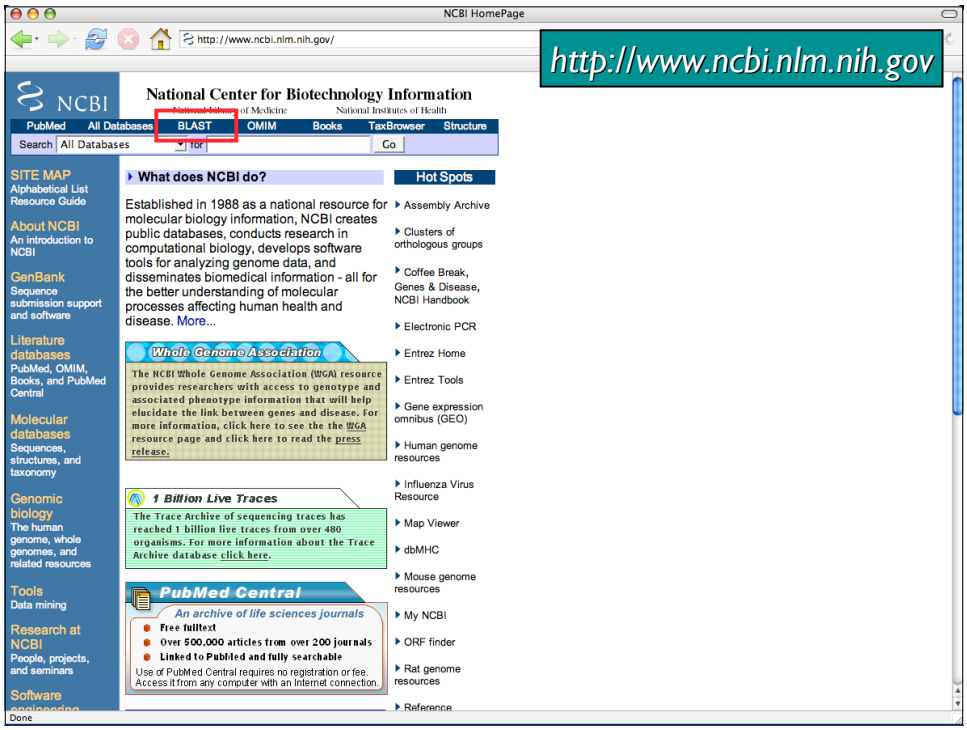


$E = kmNe^{-\lambda S}$

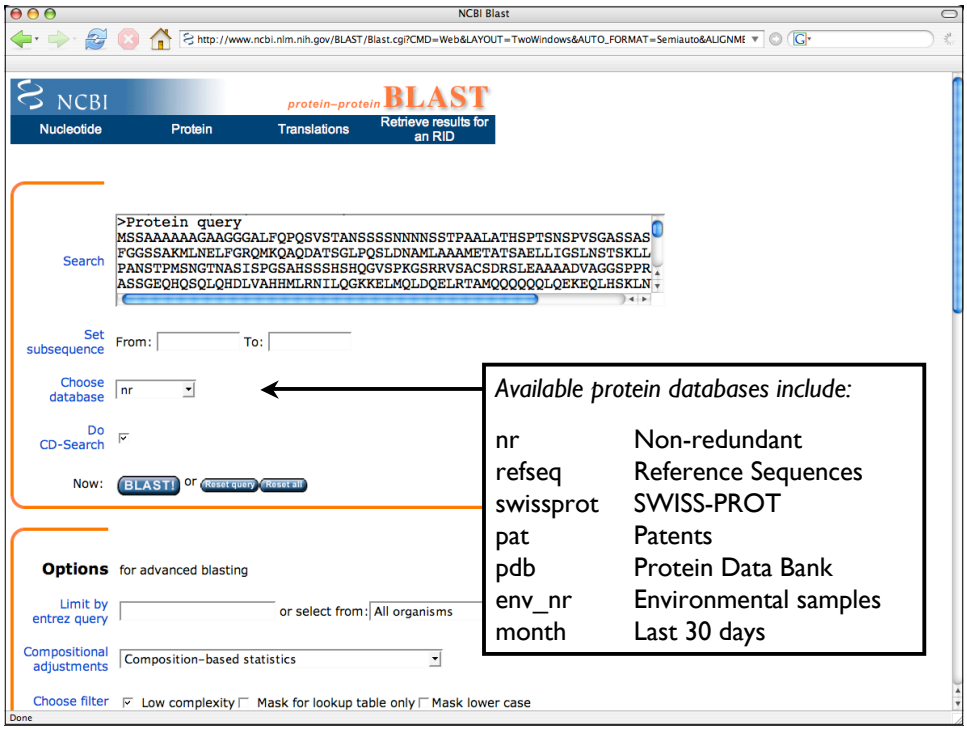
Number of HSPs found purely by chance

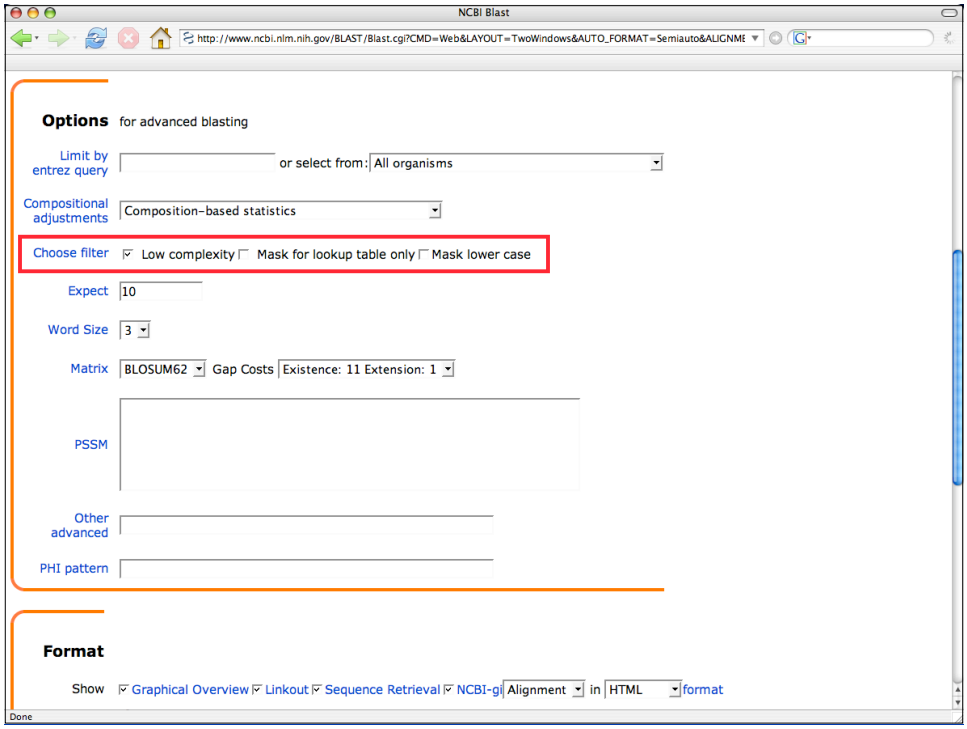
Lower values signify higher similarity

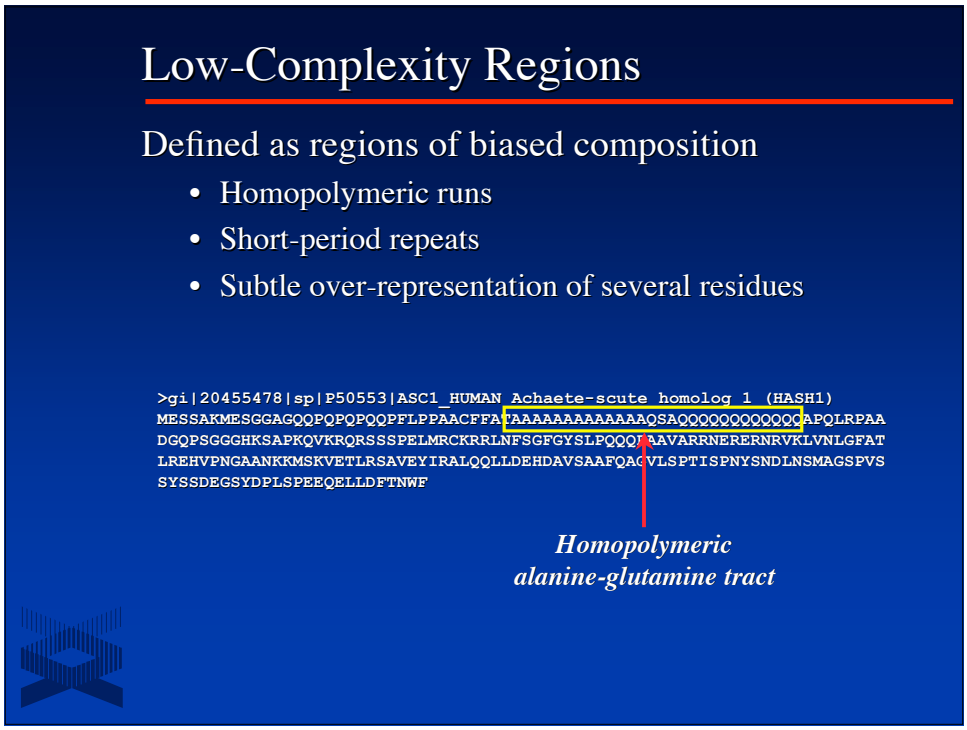

Scores and Probabilities
Query: 325 SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA 365
+LA++L TP+G R++ +W+ +P+ D + ER + A
Sbjct: 290 TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA 330
E ≤ 10-6
for nucleotides
E ≤ 10-3
for proteins
Cumulative Score
X
S
T
Extension


NCBI HomePage
http://www.ncbi.nlm.nih.gov/
http://www.ncbi.nlm.nih.gov
National Center for Biotechnology Information
PubMed All Databases BLAST OMIM Books TaxBrowser Structure
Search All Databases for Go
SITE MAP
Alphabetical List
Resource Guide
About NCBI
An introduction to NCBI
GenBank
Sequence submission support and software
Literature databases
PubMed, OMIM, Books, and PubMed Central
Molecular databases
Sequences, structures, and taxonomy
Genomic biology
The human genome, whole genomes, and related resources
Tools
Data mining
Research at NCBI
People, projects, and seminars
Software
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...
Whole Genome Association
The NCBI Whole Genome Association (WGA) resource provides researchers with access to genotype and associated phenotype information that will help elucidate the link between genes and disease. For more information, click here to see the the WGA resource page and click here to read the press release.
1 Billion Live Traces
The Trace Archive of sequencing traces has reached 1 billion live traces from over 480 organisms. For more information about the Trace Archive database click here.
PubMed Central
An archive of life sciences journals
Free fulltext
Over 500,000 articles from over 200 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.
Hot Spots
Assembly Archive
Clusters of orthologous groups
Coffee Break, Genes & Disease, NCBI Handbook
Electronic PCR
Entrez Home
Entrez Tools
Gene expression omnibus (GEO)
Human genome resources
Influenza Virus Resource
Map Viewer
dbMHC
Mouse genome resources
My NCBI
ORF finder
Rat genome resources
Reference
Done


NCBI BLAST
http://www.ncbi.nlm.nih.gov/BLAST/
http://www.ncbi.nlm.nih.gov/BLAST
NCBI
BLAST
About
Getting started
News
FAQs
More info
NAR 2004
NCBI Handbook
The Statistics of Sequence Similarity Scores
Software
Downloads
Developer info
Other resources
References
NCBI Contributors
Mailing list
Contact us
The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.
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)
Position-specific iterated and pattern-hit initiated BLAST (PSI- and PHI-BLAST)
Search for short, nearly exact matches
Search the conserved domain database (rpsblast)
Protein homology by domain architecture (cdart)
Translated
Translated query vs. protein database (blastx)
Protein query vs. translated database (tblastn)
Translated query vs. translated database (tblastx)
Genomes
Human, mouse, rat, chimp, cow, pig, dog, sheep, cat
Chicken, puffer fish, zebrafish
Fly, honey bee, other insects
Microbes, environmental samples
Plants, nematodes
Fungi, protozoa, other eukaryotes
Special
Search for gene expression data (GEO BLAST)
Align two sequences (bl2seq)
Screen for vector contamination (VecScreen)
Immunoglobin BLAST (IgBlast)
SNP BLAST
Meta
Retrieve results
Disclaimer
Privacy statement
Accessibility
This page is valid XHTML 1.0.
Done

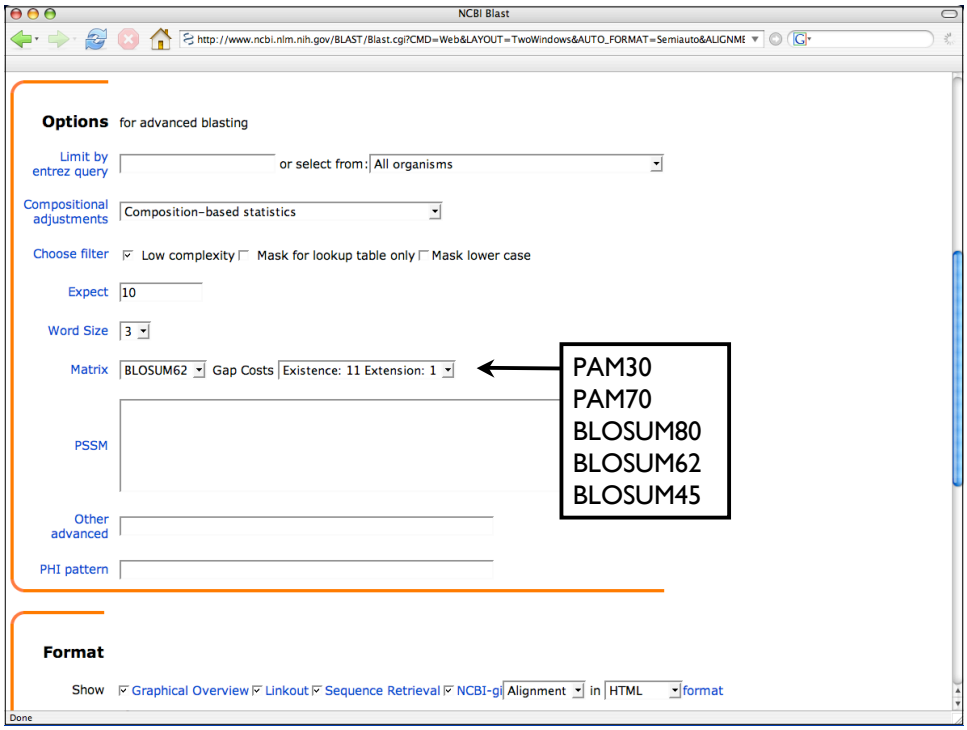

NCBI Blast
http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi?CMD=Web&LAYOUT=TwoWindows&AUTO_FORMAT=Semiauto&ALIGNM
NCBI
protein–protein BLAST
Nucleotide
Protein
Translations
Retrieve results for an RID
Search
>Protein query
MSSAAAAAAGAAGGGALFQPQSVSTANSSSSNNNNSSTPAALATHSPTSNSPVSGASSAS
FGGSSAKMLNELFGRQMKQAQDATSGLPQSLDNAMLAAAMETATSAELLIGSLNSTSKLL
PANSTPMSNGTNASISPGSAHSSSHSHQGVSPKGSRRVSACSDRSLEAAAADVAGGSPPR
ASSGEQHQSQLQHDLVAHHMLRNILQGKKELMQLDQELRTAMQQQQQQLQEKEQLHSKLN
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
Compositional adjustments
Composition-based statistics
Choose filter
Low complexity
Mask for lookup table only
Mask lower case
Done
Available protein databases include:
nr Non-redundant
refseq Reference Sequences
swissprot SWISS-PROT
pat Patents
pdb Protein Data Bank
env_nr Environmental samples
month Last 30 days

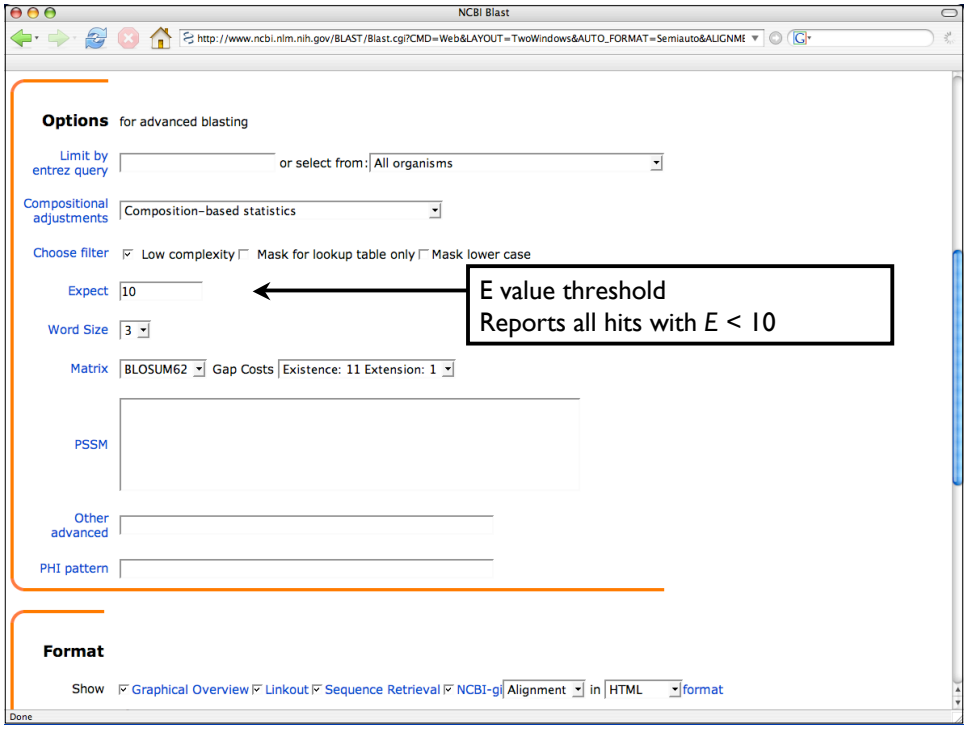

NCBI Blast
http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi?CMD=Web&LAYOUT=TwoWindows&AUTO_FORMAT=Semiauto&ALIGNME
Options for advanced blasting
Limit by entrez query
or select from: All organisms
Compositional adjustments
Composition-based statistics
Choose filter
Low complexity
Mask for lookup table only
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
Done

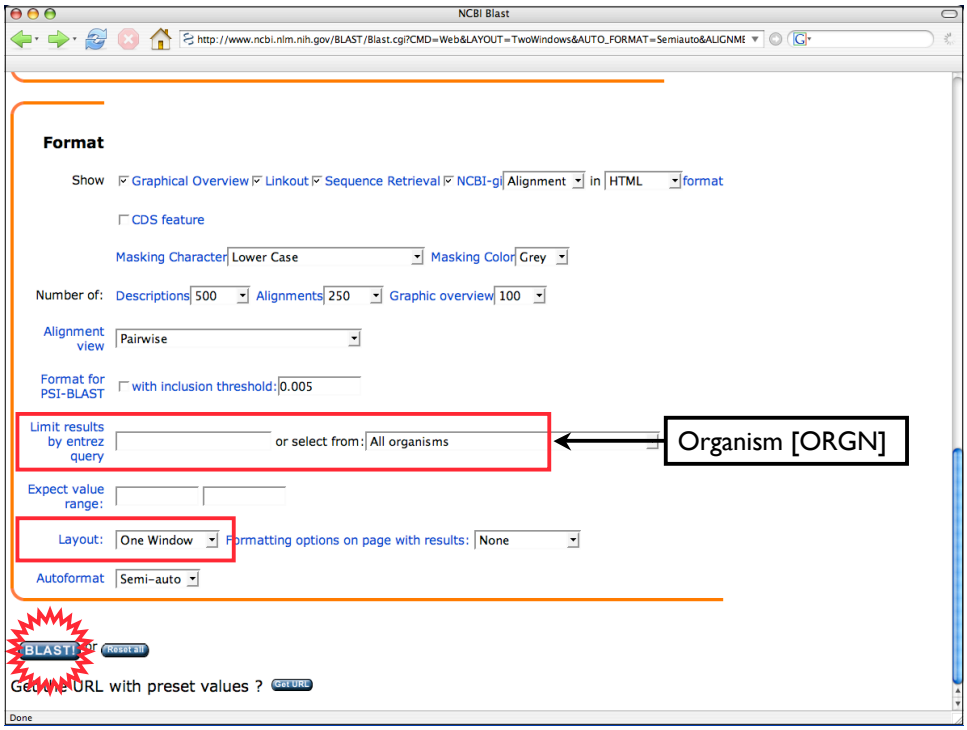

Low-Complexity Regions
Defined as regions of biased composition
Homopolymeric runs
Short-period repeats
Subtle over-representation of several residues
>gi|20455478|sp|P50553|ASC1_HUMAN Achaete-scute homolog 1 (HASH1)
MESSAKMESGGAGQQPQPQPQQPFLPPAACFFATAAAAAAAAAAAAAQSAQQQQQQQQQQQQAPQLRPAA
DGQPSGGGHKSAPKQVKRQRSSSPELMRCKRRLNFSGFGYSLPQQQLAAVARRNERERNRVKLVNLGFAT
LREHVPNGAANKKMSKVETLRSAVEYIRALQQLLDEHDAVSAAFQAGVLSPTISPNYSNDLNSMAGSPVS
SYSSDEGSYDPLSPEEQELLDFTNWF
Homopolymeric
alanine-glutamine tract

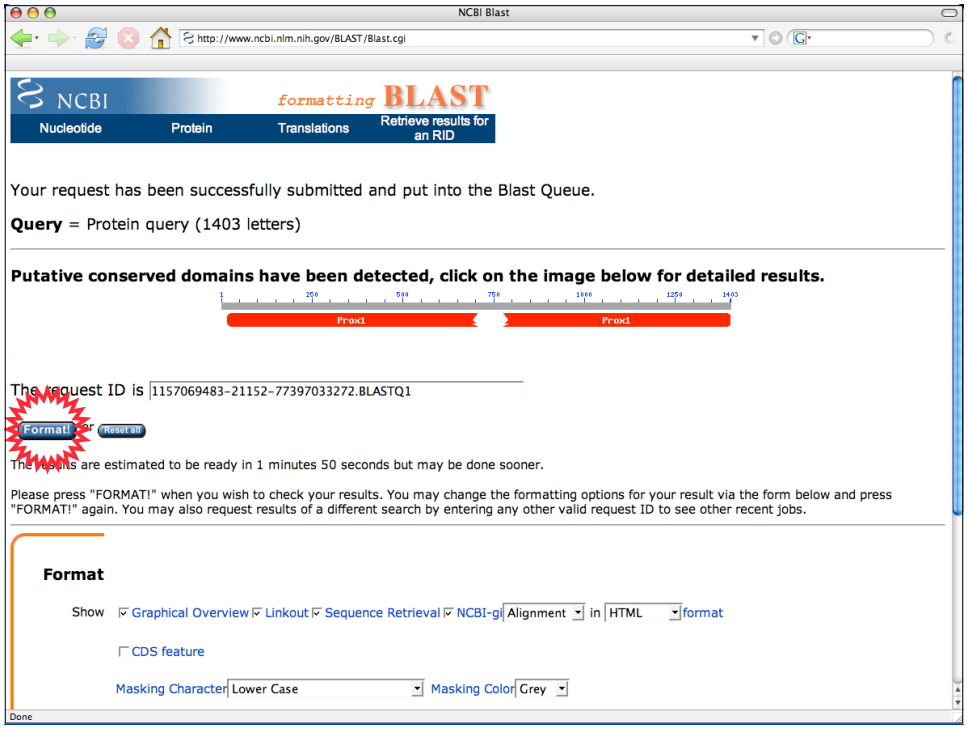
Identifying Low-Complexity Regions

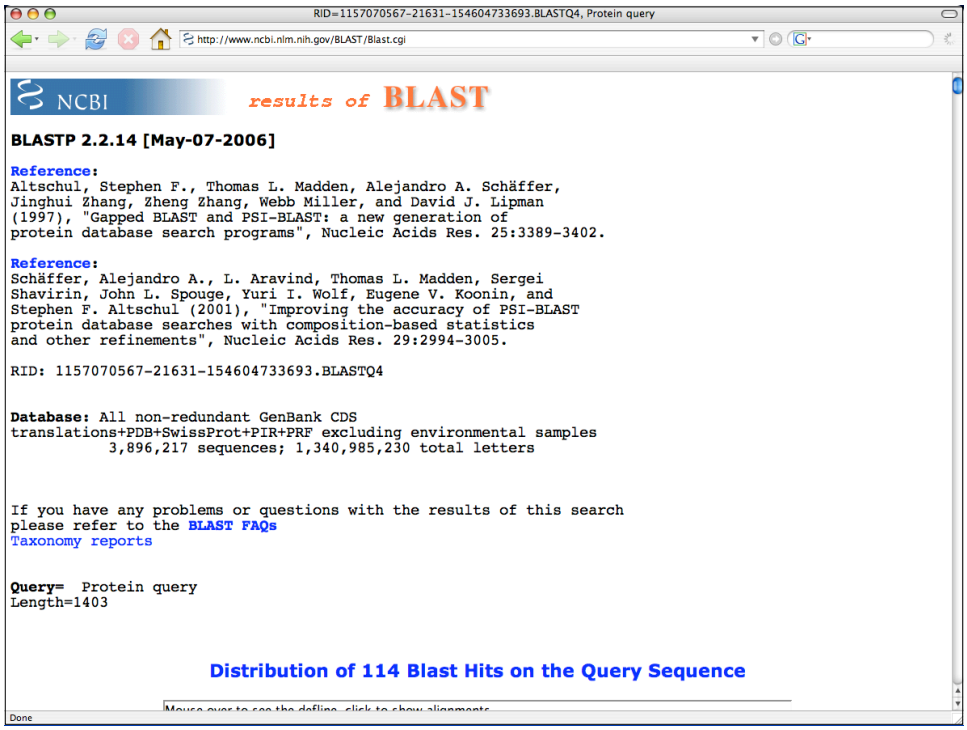
- Biological origins and role not well-understood
  - DNA replication errors (polymerase slippage)?
  - Unequal crossing-over?
- May confound sequence analysis
  - BLAST relies on uniformly-distributed amino acid frequencies
  - Often lead to false positives
  - Filtering is advised (and usually enabled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

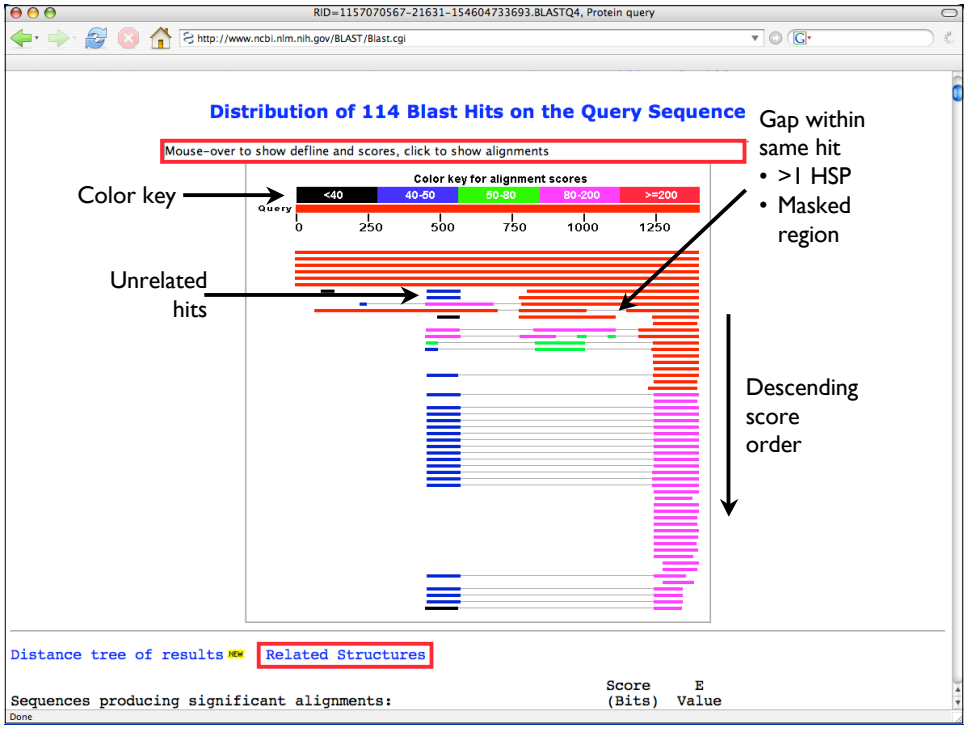


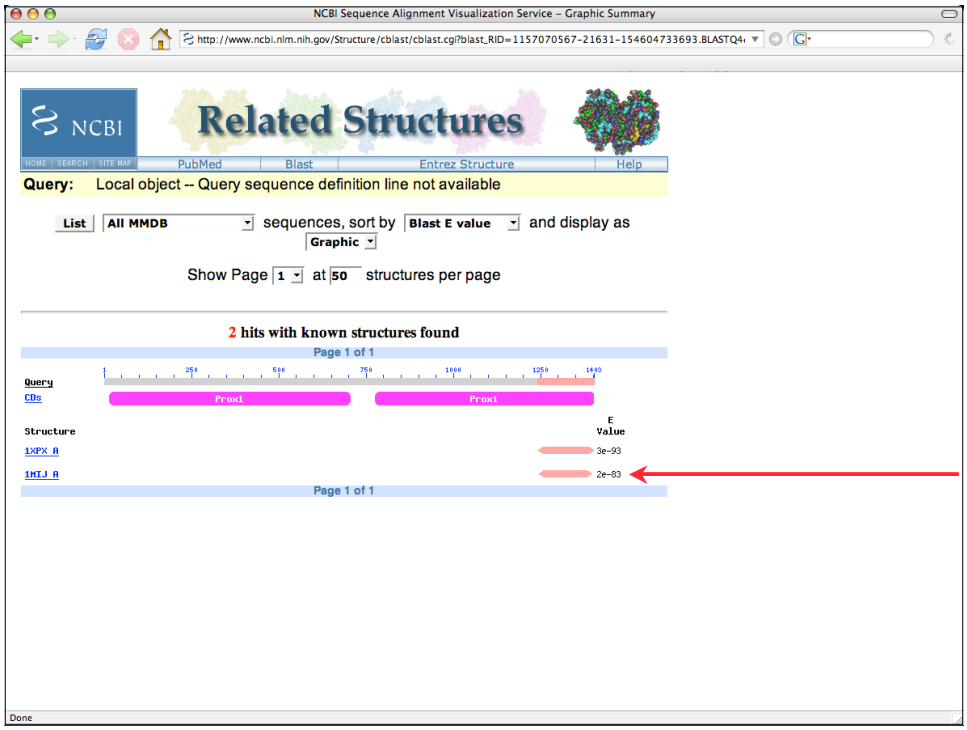

NCBI Blast
http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi?CMD=Web&LAYOUT=TwoWindows&AUTO_FORMAT=Semiauto&ALIGNME
Options for advanced blasting
Limit by entrez query
or select from: All organisms
Compositional adjustments
Composition-based statistics
Choose filter
Low complexity Mask for lookup table only 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
Done

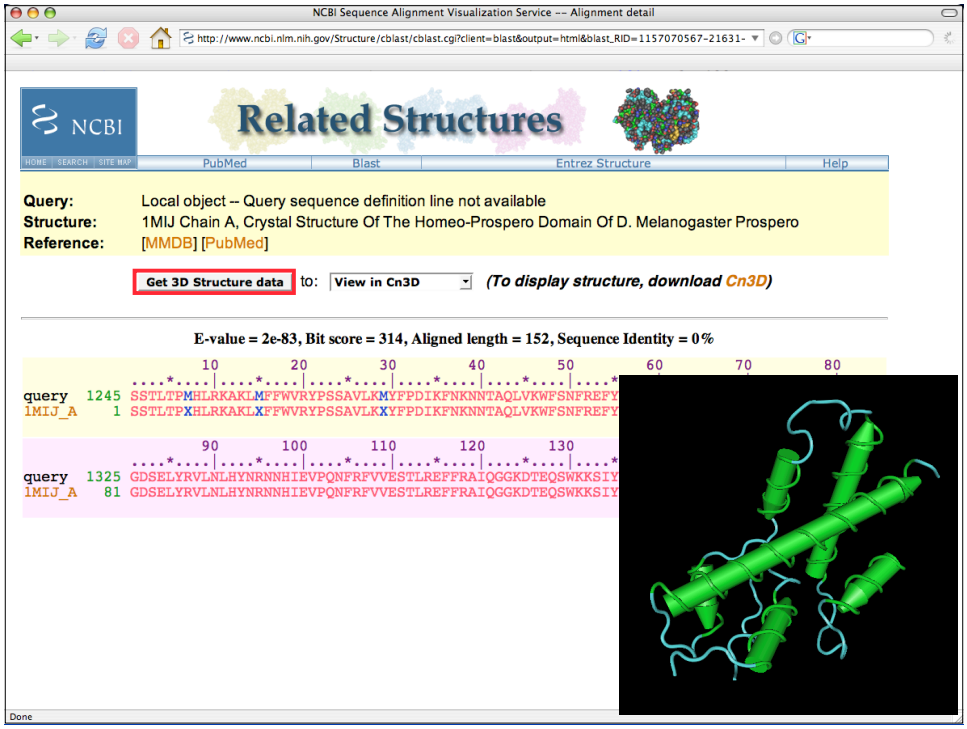

NCBI Blast
http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi?CMD=Web&LAYOUT=TwoWindows&AUTO_FORMAT=Semiauto&ALIGNME
Format
Show
Graphical Overview Linkout Sequence Retrieval NCBI-gi
Alignment
in
HTML
format
CDS feature
Masking Character
Lower Case
Masking Color
Grey
Number of:
Descriptions
500
Alignments
250
Graphic overview
100
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:
One Window
Formatting options on page with results:
None
Autoformat
Semi-auto
BLAST!
Reset all
Get URL with preset values ?
Get URL
Done

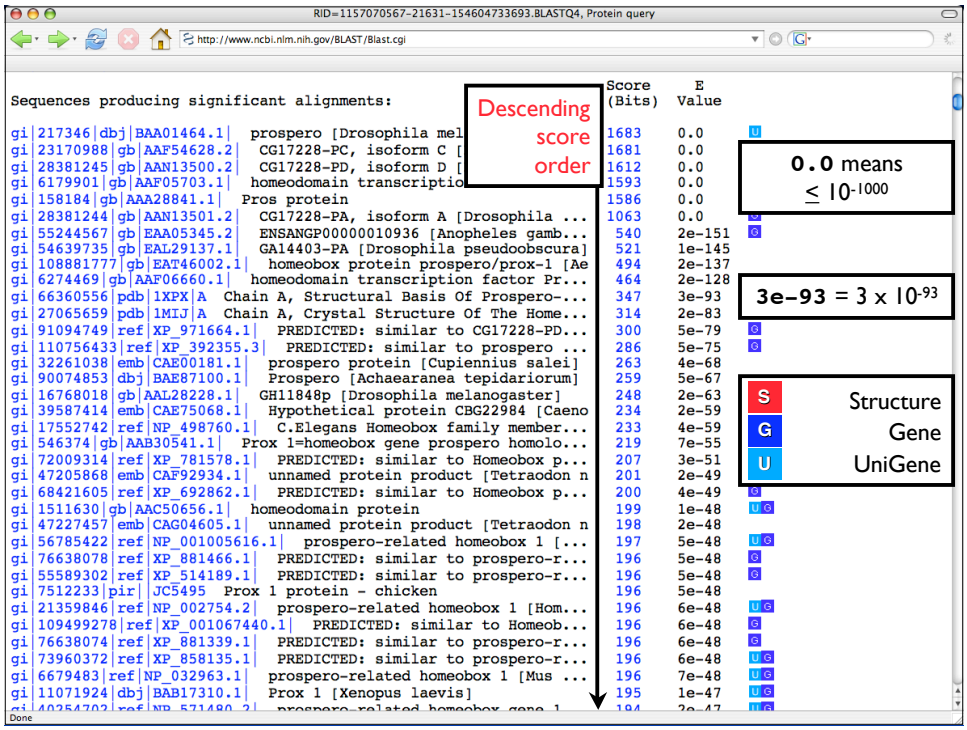

NCBI Blast
http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi
NCBI
formatting BLAST
Nucleotide
Protein
Translations
Retrieve results for an RID
Your request has been successfully submitted and put into the Blast Queue.
Query = Protein query (1403 letters)
Putative conserved domains have been detected, click on the image below for detailed results.
Prox1
Prox1
The request ID is 1157069483-21152-77397033272.BLASTQ1
Format!
Reset all
The results are estimated to be ready in 1 minutes 50 seconds but may be done sooner.
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
CDS feature
Masking Character
Lower Case
Masking Color
Grey
Done

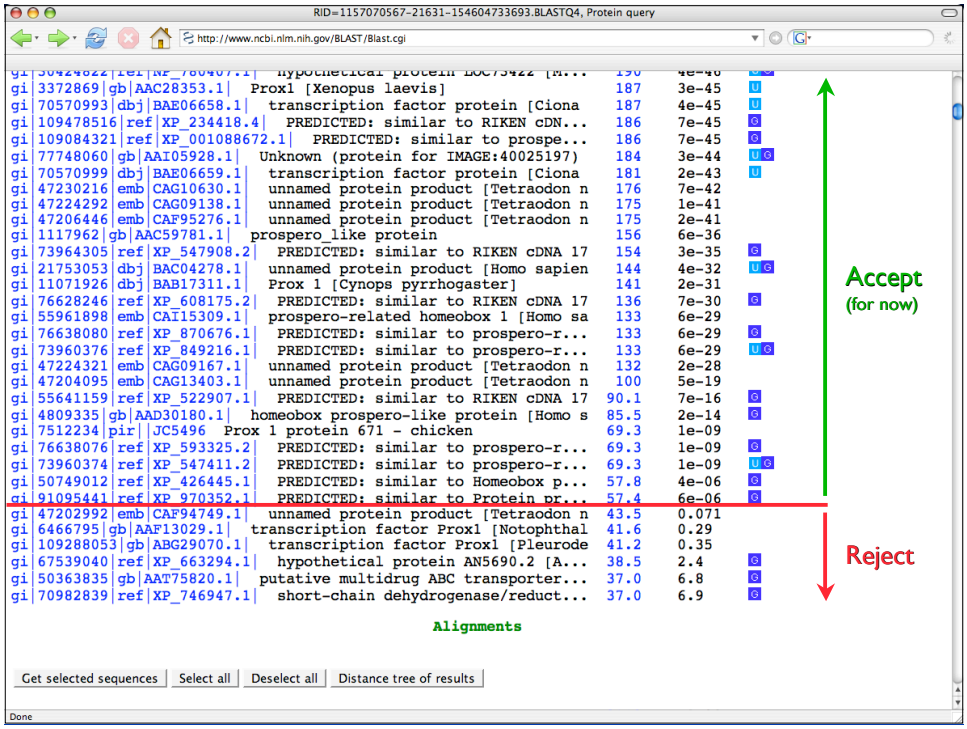

RID=1157070567-21631-154604733693.BLASTQ4, Protein query
http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi
NCBI
results of BLAST
BLASTP 2.2.14 [May-07-2006]
Reference:
Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman
(1997), "Gapped BLAST and PSI-BLAST: a new generation of
protein database search programs", Nucleic Acids Res. 25:3389-3402.
Reference:
Schäffer, Alejandro A., L. Aravind, Thomas L. Madden, Sergei
Shavirin, John L. Spouge, Yuri I. Wolf, Eugene V. Koonin, and
Stephen F. Altschul (2001), "Improving the accuracy of PSI-BLAST
protein database searches with composition-based statistics
and other refinements", Nucleic Acids Res. 29:2994-3005.
RID: 1157070567-21631-154604733693.BLASTQ4
Database: All non-redundant GenBank CDS
translations+PDB+SwissProt+PIR+PRF excluding environmental samples
3,896,217 sequences; 1,340,985,230 total letters
If you have any problems or questions with the results of this search
please refer to the BLAST FAQs
Taxonomy reports
Query= Protein query
Length=1403
Distribution of 114 Blast Hits on the Query Sequence
Done

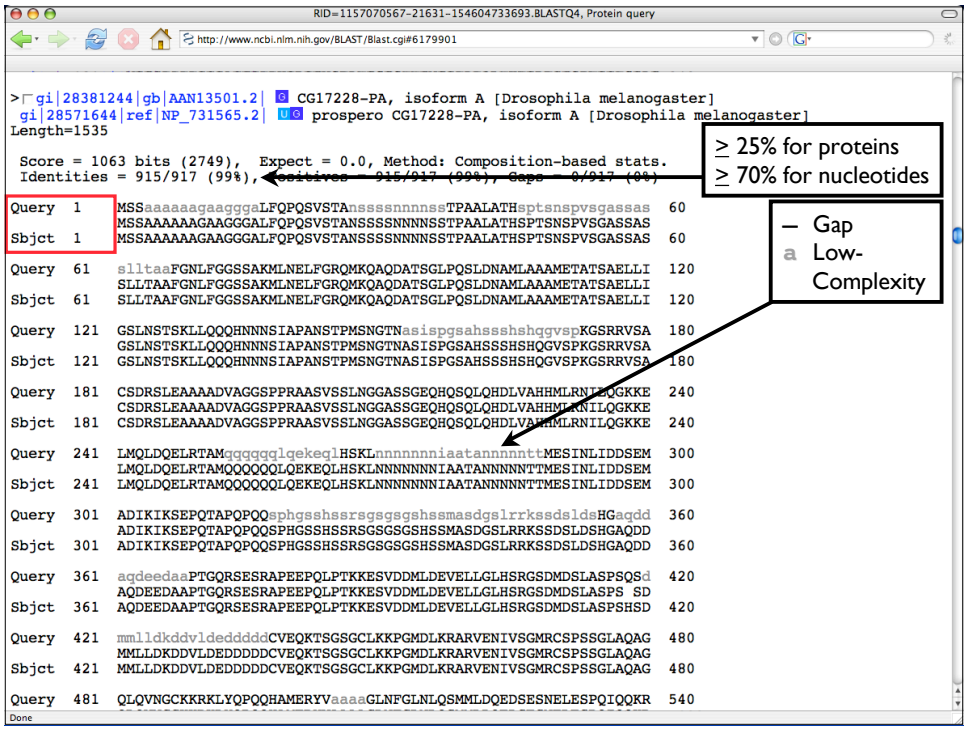

RID=1157070567-21631-154604733693.BLASTQ4, Protein query
http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi
Distribution of 114 Blast Hits on the Query Sequence
Mouse-over to show defline and scores, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
0
250
500
750
1000
1250
Color key
Unrelated hits
Gap within same hit
• >1 HSP
• Masked region
Descending score order
Distance tree of results NEW
Related Structures
Sequences producing significant alignments:
Score (Bits)
E Value
Done

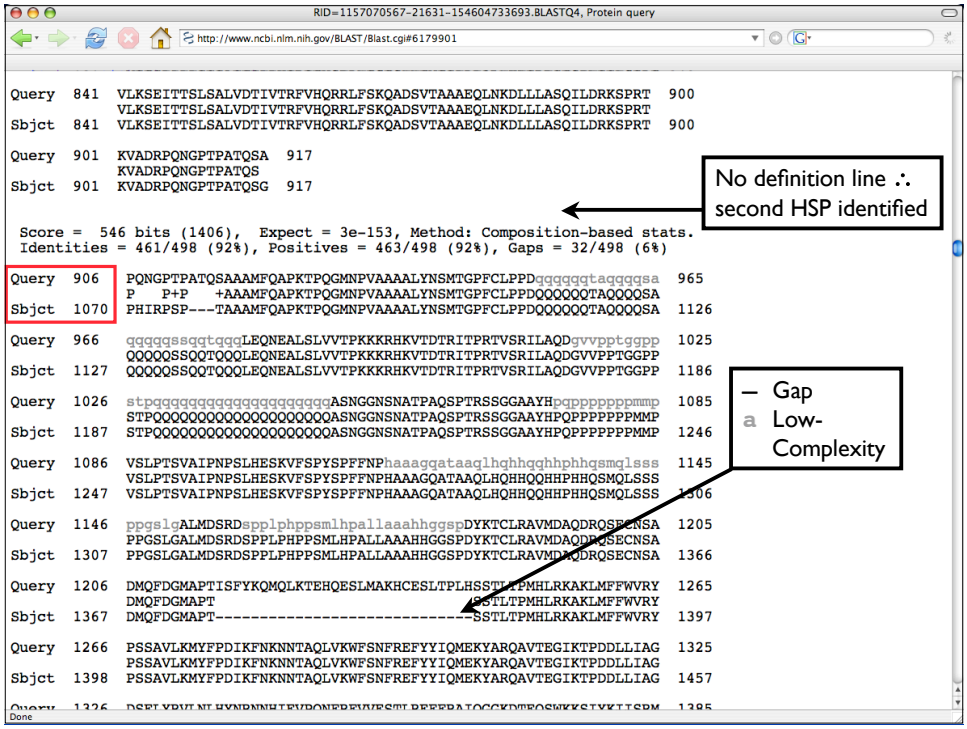

NCBI Sequence Alignment Visualization Service - Graphic Summary
http://www.ncbi.nlm.nih.gov/Structure/cblast/cblast.cgi?blast_RID=1157070567-21631-154604733693.BLASTQ4
NCBI
Related Structures
HOME | SEARCH | SITE MAP
PubMed
Blast
Entrez Structure
Help
Query: Local object -- Query sequence definition line not available
List
All MMDB
sequences, sort by
Blast E value
and display as
Graphic
Show Page 1 at 50 structures per page
2 hits with known structures found
Page 1 of 1
Query
CDs
Prox1
Prox1
Structure
E Value
1XPX A
3e-93
1MIJ A
2e-83
Page 1 of 1
Done

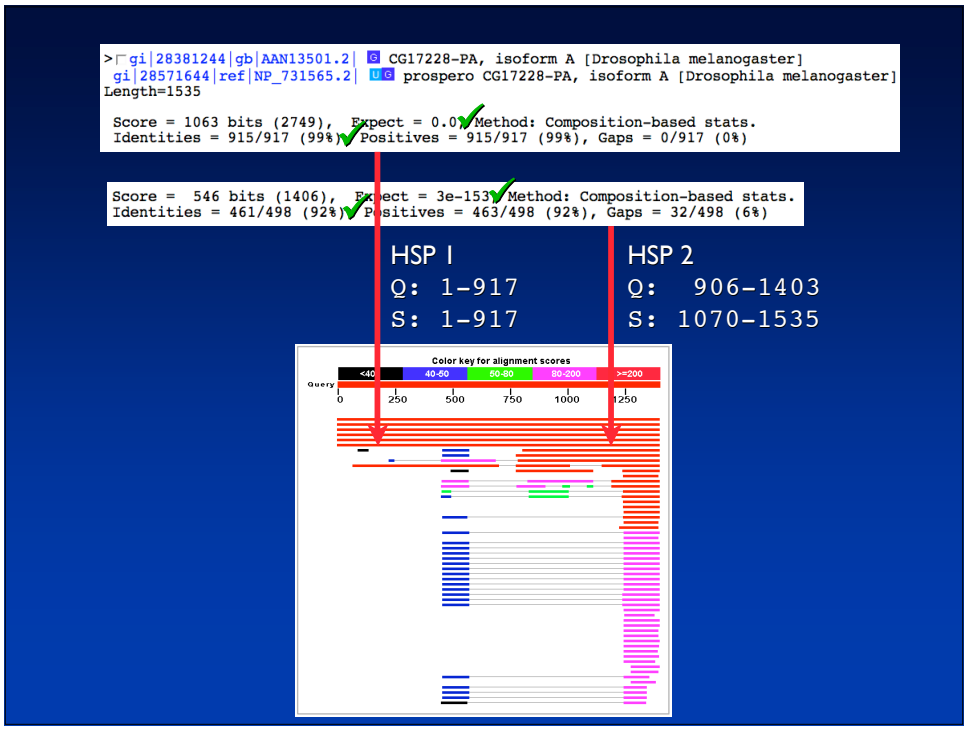

NCBI Sequence Alignment Visualization Service -- Alignment detail
Related Structures
PubMed
Blast
Entrez Structure
Help
Query: Local object -- Query sequence definition line not available
Structure: 1MIJ Chain A, Crystal Structure Of The Homeo-Prospero Domain Of D. Melanogaster Prospero
Reference: [MMDB] [PubMed]
Get 3D Structure data
to:
View in Cn3D
(To display structure, download Cn3D)
E-value = 2e-83, Bit score = 314, Aligned length = 152, Sequence Identity = 0%
query 1245 SSTLTPMHLRKAKLMFFWVRYPSSAVLKMYFPDIKFNKNNTAQLVKWFSNFREFY
1MIJ_A 1 SSTLTPXHLRKAKLXFFWVRYPSSAVLKXYFPDIKFNKNNTAQLVKWFSNFREFY
query 1325 GDSELYRVLNLHYNRNNHIEVPQNFRFVVESTLREFFRAIQGGKDTEQSWKKSIY
1MIJ_A 81 GDSELYRVLNLHYNRNNHIEVPQNFRFVVESTLREFFRAIQGGKDTEQSWKKSIY
Done


RID=1157070567-21631-154604733693.BLASTQ4, Protein query
http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi
Sequences producing significant alignments:
Score (Bits)
E Value
Descending score order
0.0 means ≤ 10-1000
3e-93 = 3 x 10-93
S Structure
G Gene
U UniGene
gi|217346|dbj|BAA01464.1| prospero [Drosophila mel 1683 0.0
gi|23170988|gb|AAF54628.2| CG17228-PC, isoform C [1681 0.0
gi|28381245|gb|AAN13500.2| CG17228-PD, isoform D [1612 0.0
gi|6179901|gb|AAF05703.1| homeodomain transcriptio 1593 0.0
gi|158184|gb|AAA28841.1| Pros protein 1586 0.0
gi|28381244|gb|AAN13501.2| CG17228-PA, isoform A [Drosophila ... 1063 0.0
gi|55244567|gb|EAA05345.2| ENSANGP00000010936 [Anopheles gamb... 540 2e-151
gi|54639735|gb|EAL29137.1| GA14403-PA [Drosophila pseudoobscura] 521 1e-145
gi|108881777|gb|EAT46002.1| homeobox protein prospero/prox-1 [Ae 494 2e-137
gi|6274469|gb|AAF06660.1| homeodomain transcription factor Pr... 464 2e-128
gi|66360556|pdb|1XPX|A Chain A, Structural Basis Of Prospero-... 347 3e-93
gi|27065659|pdb|1MIJ|A Chain A, Crystal Structure Of The Home... 314 2e-83
gi|91094749|ref|XP_971664.1| PREDICTED: similar to CG17228-PD... 300 5e-79
gi|110756433|ref|XP_392355.3| PREDICTED: similar to prospero ... 286 5e-75
gi|32261038|emb|CAE00181.1| prospero protein [Cupiennius salei] 263 4e-68
gi|90074853|dbj|BAE87100.1| Prospero [Achaearanea tepidariorum] 259 5e-67
gi|16768018|gb|AAL28228.1| GH11848p [Drosophila melanogaster] 248 2e-63
gi|39587414|emb|CAE75068.1| Hypothetical protein CBG22984 [Caeno 234 2e-59
gi|17552742|ref|NP_498760.1| C.Elegans Homeobox family member... 233 4e-59
gi|546374|gb|AAB30541.1| Prox 1=homeobox gene prospero homolo... 219 7e-55
gi|72009314|ref|XP_781578.1| PREDICTED: similar to Homeobox p... 207 3e-51
gi|47205868|emb|CAF92934.1| unnamed protein product [Tetraodon n 201 2e-49
gi|68421605|ref|XP_692862.1| PREDICTED: similar to Homeobox p... 200 4e-49
gi|1511630|gb|AAC50656.1| homeodomain protein 199 1e-48
gi|47227457|emb|CAG04605.1| unnamed protein product [Tetraodon n 198 2e-48
gi|56785422|ref|NP_001005616.1| prospero-related homeobox 1 [... 197 5e-48
gi|76638078|ref|XP_881466.1| PREDICTED: similar to prospero-r... 196 5e-48
gi|55589302|ref|XP_514189.1| PREDICTED: similar to prospero-r... 196 5e-48
gi|7512233|pir||JC5495 Prox 1 protein - chicken 196 5e-48
gi|21359846|ref|NP_002754.2| prospero-related homeobox 1 [Hom... 196 6e-48
gi|109499278|ref|XP_001067440.1| PREDICTED: similar to Homeob... 196 6e-48
gi|76638074|ref|XP_881339.1| PREDICTED: similar to prospero-r... 196 6e-48
gi|73960372|ref|XP_858135.1| PREDICTED: similar to prospero-r... 196 6e-48
gi|6679483|ref|NP_032963.1| prospero-related homeobox 1 [Mus ... 196 7e-48
gi|11071924|dbj|BAB17310.1| Prox 1 [Xenopus laevis] 195 1e-47
Done


RID=1157070567-21631-154604733693.BLASTQ4, Protein query
http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi
gi|3372869|gb|AAC28353.1| Prox1 [Xenopus laevis] 187 3e-45
gi|70570993|dbj|BAE06658.1| transcription factor protein [Ciona 187 4e-45
gi|109478516|ref|XP_234418.4| PREDICTED: similar to RIKEN cDN... 186 7e-45
gi|109084321|ref|XP_001088672.1| PREDICTED: similar to prospe... 186 7e-45
gi|77748060|gb|AAI05928.1| Unknown (protein for IMAGE:40025197) 184 3e-44
gi|70570999|dbj|BAE06659.1| transcription factor protein [Ciona 181 2e-43
gi|47230216|emb|CAG10630.1| unnamed protein product [Tetraodon n 176 7e-42
gi|47224292|emb|CAG09138.1| unnamed protein product [Tetraodon n 175 1e-41
gi|47206446|emb|CAF95276.1| unnamed protein product [Tetraodon n 175 2e-41
gi|1117962|gb|AAC59781.1| prospero_like protein 156 6e-36
gi|73964305|ref|XP_547908.2| PREDICTED: similar to RIKEN cDNA 17 154 3e-35
gi|21753053|dbj|BAC04278.1| unnamed protein product [Homo sapien 144 4e-32
gi|11071926|dbj|BAB17311.1| Prox 1 [Cynops pyrrhogaster] 141 2e-31
gi|76628246|ref|XP_608175.2| PREDICTED: similar to RIKEN cDNA 17 136 7e-30
gi|55961898|emb|CAI15309.1| prospero-related homeobox 1 [Homo sa 133 6e-29
gi|76638080|ref|XP_870676.1| PREDICTED: similar to prospero-r... 133 6e-29
gi|73960376|ref|XP_849216.1| PREDICTED: similar to prospero-r... 133 6e-29
gi|47224321|emb|CAG09167.1| unnamed protein product [Tetraodon n 132 2e-28
gi|47204095|emb|CAG13403.1| unnamed protein product [Tetraodon n 100 5e-19
gi|55641159|ref|XP_522907.1| PREDICTED: similar to RIKEN cDNA 17 90.1 7e-16
gi|4809335|gb|AAD30180.1| homeobox prospero-like protein [Homo s 85.5 2e-14
gi|7512234|pir||JC5496 Prox 1 protein 671 - chicken 69.3 1e-09
gi|76638076|ref|XP_593325.2| PREDICTED: similar to prospero-r... 69.3 1e-09
gi|73960374|ref|XP_547411.2| PREDICTED: similar to prospero-r... 69.3 1e-09
gi|50749012|ref|XP_426445.1| PREDICTED: similar to Homeobox p... 57.8 4e-06
gi|91095441|ref|XP_970352.1| PREDICTED: similar to Protein pr... 57.4 6e-06
gi|47202992|emb|CAF94749.1| unnamed protein product [Tetraodon n 43.5 0.071
gi|6466795|gb|AAF13029.1| transcription factor Prox1 [Notophthal 41.6 0.29
gi|109288053|gb|ABG29070.1| transcription factor Prox1 [Pleurode 41.2 0.35
gi|67539040|ref|XP_663294.1| hypothetical protein AN5690.2 [A... 38.5 2.4
gi|50363835|gb|AAT75820.1| putative multidrug ABC transporter... 37.0 6.8
gi|70982839|ref|XP_746947.1| short-chain dehydrogenase/reduct... 37.0 6.9
Accept
(for now)
Reject
Alignments
Get selected sequences
Select all
Deselect all
Distance tree of results
Done


RID=1157070567-21631-154604733693.BLASTQ4, Protein query
http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi#6179901
>gi|28381244|gb|AAN13501.2| CG17228-PA, isoform A [Drosophila melanogaster]
gi|28571644|ref|NP_731565.2| prospero CG17228-PA, isoform A [Drosophila melanogaster]
Length=1535
Score = 1063 bits (2749), Expect = 0.0, Method: Composition-based stats.
Identities = 915/917 (99%), Positives = 915/917 (99%), Gaps = 0/917 (0%)
Query 1 MSSaaaaaagaagggaLFQPQSVSTAnssssnnnnssTPAALATHsptsnspvsgassas 60
MSSAAAAAAGAAGGGALFQPQSVSTANSSSSNNNNSSTPAALATHSPTSNSPVSGASSAS
Sbjct 1 MSSAAAAAAGAAGGGALFQPQSVSTANSSSSNNNNSSTPAALATHSPTSNSPVSGASSAS 60
Query 61 slltaaFGNLFGGSSAKMLNELFGRQMKQAQDATSGLPQSLDNAMLAAAMETATSAELLI 120
SLLTAAFGNLFGGSSAKMLNELFGRQMKQAQDATSGLPQSLDNAMLAAAMETATSAELLI
Sbjct 61 SLLTAAFGNLFGGSSAKMLNELFGRQMKQAQDATSGLPQSLDNAMLAAAMETATSAELLI 120
Query 121 GSLNSTSKLLQQQHNNNSIAPANSTPMSNGTNasispgsahssshshqgvspKGSRRVSA 180
GSLNSTSKLLQQQHNNNSIAPANSTPMSNGTNASISPGSAHSSSHSHQGVSPKGSRRVSA
Sbjct 121 GSLNSTSKLLQQQHNNNSIAPANSTPMSNGTNASISPGSAHSSSHSHQGVSPKGSRRVSA 180
Query 181 CSDRSLEAAAADVAGGSPPRAASVSSLNGGASSGEQHQSQLQHDLVAHHMLRNILQGKKE 240
CSDRSLEAAAADVAGGSPPRAASVSSLNGGASSGEQHQSQLQHDLVAHHMLRNILQGKKE
Sbjct 181 CSDRSLEAAAADVAGGSPPRAASVSSLNGGASSGEQHQSQLQHDLVAHHMLRNILQGKKE 240
Query 241 LMQLDQELRTAMqqqqqqlqekeqlHSKLnnnnnnniaatannnnnttMESINLIDDSEM 300
LMQLDQELRTAMQQQQQQLQEKEQLHSKLNNNNNNNIAATANNNNNTTMESINLIDDSEM
Sbjct 241 LMQLDQELRTAMQQQQQQLQEKEQLHSKLNNNNNNNIAATANNNNNTTMESINLIDDSEM 300
Query 301 ADIKIKSEPQTAPQPQQsphgsshssrsgsgsgshssmasdgslrrksasdsldsHGaqdd 360
ADIKIKSEPQTAPQPQQSPHGSSHSSRSGSGSGSHSSMASDGSLRRKSSDSLDSHGAQDD
Sbjct 301 ADIKIKSEPQTAPQPQQSPHGSSHSSRSGSGSGSHSSMASDGSLRRKSSDSLDSHGAQDD 360
Query 361 aqdeedaaPTGQRSESRAPEEPQLPTKKESVDDMLDEVELLGLHSRGSDMDSLASPSQSd 420
AQDEEDAAPTGQRSESRAPEEPQLPTKKESVDDMLDEVELLGLHSRGSDMDSLASPS SD
Sbjct 361 AQDEEDAAPTGQRSESRAPEEPQLPTKKESVDDMLDEVELLGLHSRGSDMDSLASPSHSD 420
Query 421 mmlldkddvldeddddCVEQKTSGSGCLKKPGMDLKRARVENIVSGMRCSPSSGLAQAG 480
MMLLDKDDVLDEDDDDDCVEQKTSGSGCLKKPGMDLKRARVENIVSGMRCSPSSGLAQAG
Sbjct 421 MMLLDKDDVLDEDDDDDCVEQKTSGSGCLKKPGMDLKRARVENIVSGMRCSPSSGLAQAG 480
Query 481 QLQVNGCKKRKLYQPQQHAMERYVaaaaGLNFGLNLQSMMLDQEDSESNELESPQIQQKR 540
≥ 25% for proteins
≥ 70% for nucleotides
– Gap
a Low-Complexity
Done


RID=1157070567-21631-154604733693.BLASTQ4, Protein query
http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi#6179901
No definition line ∴
second HSP identified
– Gap
a Low-
Complexity
Done


>gi|28381244|gb|AAN13501.2| CG17228-PA, isoform A [Drosophila melanogaster]
gi|28571644|ref|NP_731565.2| prospero CG17228-PA, isoform A [Drosophila melanogaster]
Length=1535
Score = 1063 bits (2749), Expect = 0.0, Method: Composition-based stats.
Identities = 915/917 (99%), Positives = 915/917 (99%), Gaps = 0/917 (0%)
Score = 546 bits (1406), Expect = 3e-153, Method: Composition-based stats.
Identities = 461/498 (92%), Positives = 463/498 (92%), Gaps = 32/498 (6%)
HSP 1
Q: 1-917
S: 1-917
HSP 2
Q: 906-1403
S: 1070-1535
Color key for alignment scores
<40
40-60
60-80
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!*

Database Searching Artifacts

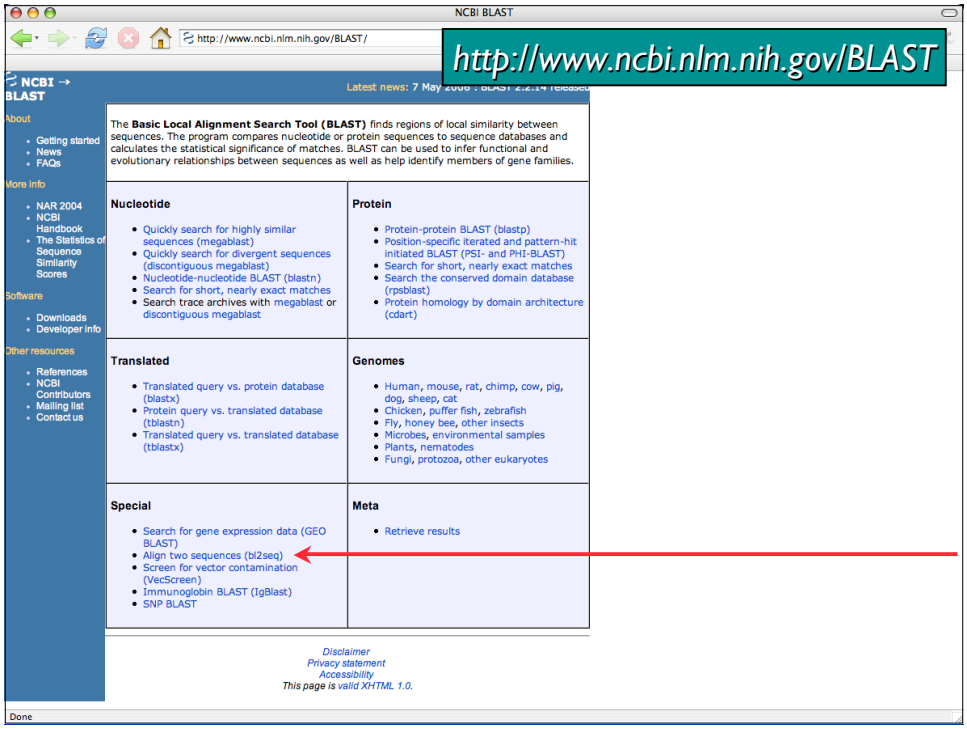
- Low-complexity regions
  - Nucleotide searches: removed with DUST (➔ N)
  - Protein searches: removed with SEG (➔ X)
- Repetitive elements
  - LINEs, SINEs, retroviral repeats
  - Choose "Filter: Human Repeats" when using BLASTN
  - RepeatMasker *http://www.repeatmasker.org*

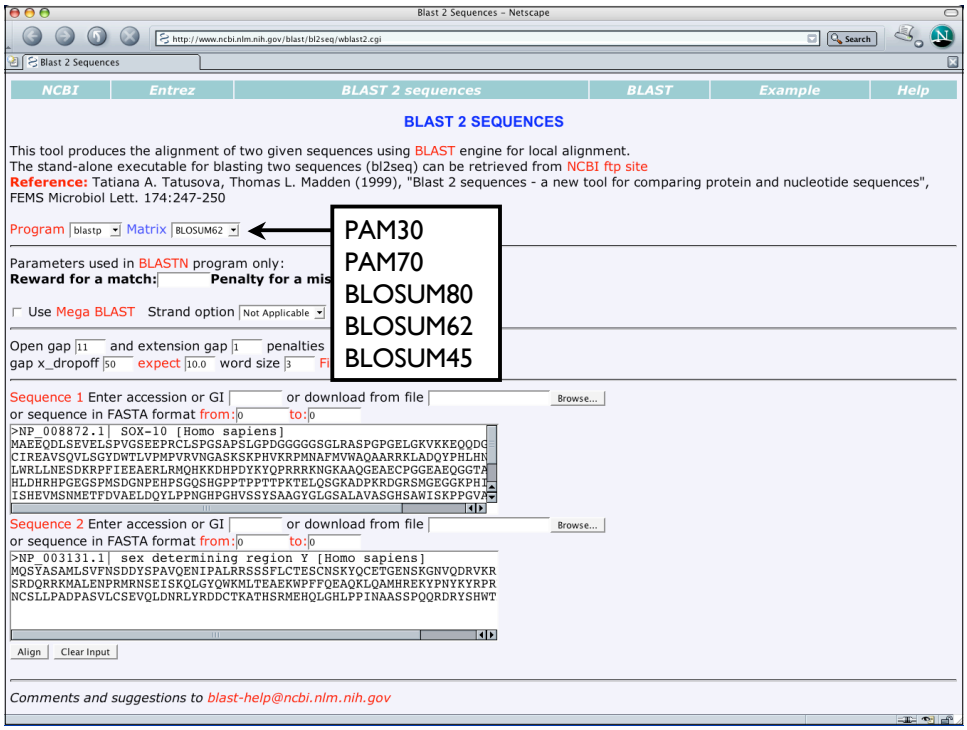
Database Searching Artifacts

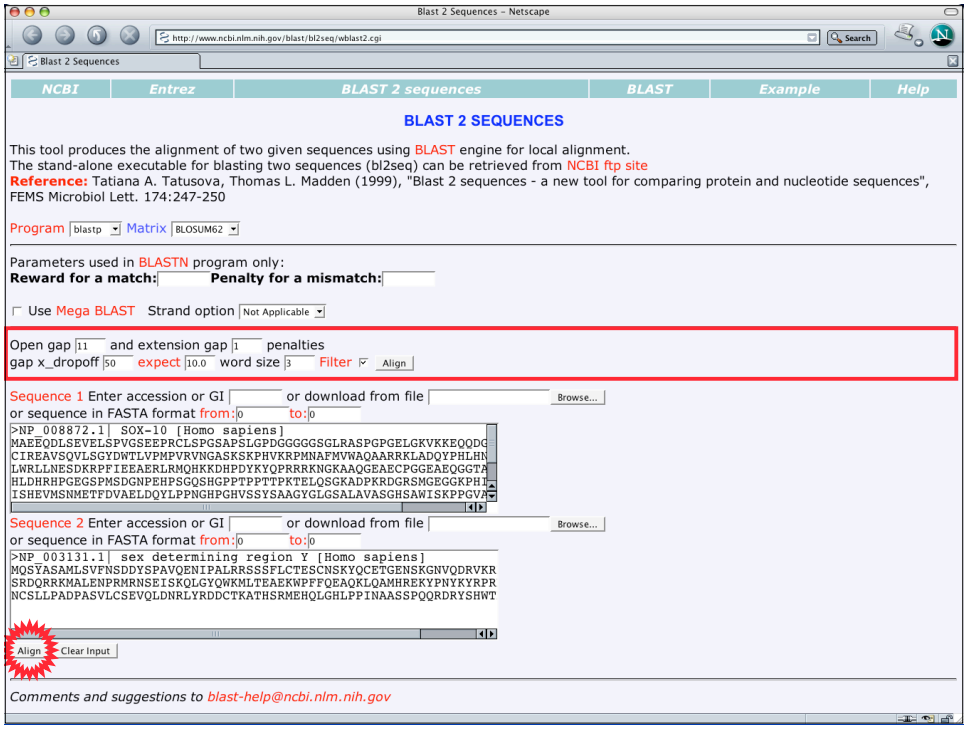
- Low-quality sequence hits
  - Expressed sequence tags (ESTs)
  - Single-pass sequence reads from large-scale sequencing (possibly with vector contaminants)

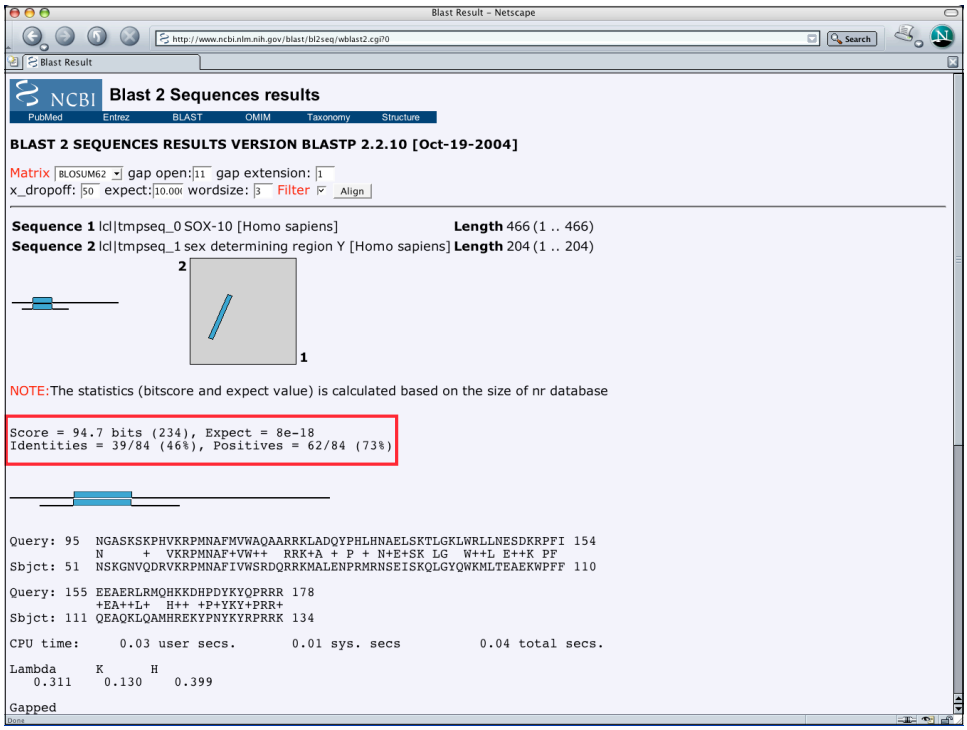
BLAST 2 Sequences

- 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


NCBI BLAST
http://www.ncbi.nlm.nih.gov/BLAST/
http://www.ncbi.nlm.nih.gov/BLAST
NCBI
BLAST
Latest news: 7 May 2006 : BLAST 2.2.14 released
About
Getting started
News
FAQs
More info
NAR 2004
NCBI Handbook
The Statistics of Sequence Similarity Scores
Software
Downloads
Developer info
Other resources
References
NCBI Contributors
Mailing list
Contact us
The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.
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)
Position-specific iterated and pattern-hit initiated BLAST (PSI- and PHI-BLAST)
Search for short, nearly exact matches
Search the conserved domain database (rpsblast)
Protein homology by domain architecture (cdart)
Translated
Translated query vs. protein database (blastx)
Protein query vs. translated database (tblastn)
Translated query vs. translated database (tblastx)
Genomes
Human, mouse, rat, chimp, cow, pig, dog, sheep, cat
Chicken, puffer fish, zebrafish
Fly, honey bee, other insects
Microbes, environmental samples
Plants, nematodes
Fungi, protozoa, other eukaryotes
Special
Search for gene expression data (GEO BLAST)
Align two sequences (bl2seq)
Screen for vector contamination (VecScreen)
Immunoglobin BLAST (IgBlast)
SNP BLAST
Meta
Retrieve results
Disclaimer
Privacy statement
Accessibility
This page is valid XHTML 1.0.
Done


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
HLDHRHPGEGSPMSDGNPEHPSGQSHGPPTPPTTPKTELQSGKADPKRDGRSMGEGGKPHI
ISHEVMSNMETFDVAELDQYLPPNGHPGHVSSYSAAGYGLGSALAVASGHSAWISKPPGVA
Sequence 2 Enter accession or GI or download from file Browse...
or sequence in FASTA format from: 0 to: 0
>NP_003131.1| sex determining region Y [Homo sapiens]
MQSYASAMLSVFNSDDYSPAVQENIPALRRSSSFLCTESCNSKYQCETGENSKGNVQDRVKR
SRDQRRKMALENPRMRNSEISKQLGYQWKMLTEAEKWPFFQEAQKLQAMHREKYPNYKYRPR
NCSLLPADPASVLCSEVQLDNRLYRDDCTKATHSRMEHQLGHLPPINAASSPQQRDRYSHWT
Align
Clear Input
Comments and suggestions to blast-help@ncbi.nlm.nih.gov


Blast 2 Sequences - Netscape
http://www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi
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


Blast Result - Netscape
http://www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi?0
Blast Result
NCBI
Blast 2 Sequences results
PubMed Entrez BLAST OMIM Taxonomy Structure
BLAST 2 SEQUENCES RESULTS VERSION BLASTP 2.2.10 [Oct-19-2004]
Matrix BLOSUM62 gap open:11 gap extension: 1
x_dropoff: 50 expect:10.00 wordsize: 3 Filter Align
Sequence 1 lcl|tmpseq_0 SOX-10 [Homo sapiens] Length 466 (1 .. 466)
Sequence 2 lcl|tmpseq_1 sex determining region Y [Homo sapiens] Length 204 (1 .. 204)
2
1
NOTE:The statistics (bitscore and expect value) is calculated based on the size of nr database
Score = 94.7 bits (234), Expect = 8e-18
Identities = 39/84 (46%), Positives = 62/84 (73%)
Query: 95 NGASKSKPHVKRPMNAFMVWAQAARRKLADQYPHLHNAELSKTLGKLWRLLNESDKRPFI 154
N + VKRPMNAF+VW++ RRK+A + P + N+E+SK LG W++L E++K PF
Sbjct: 51 NSKGNVQDRVKRPMNAFIVWSRDQRRKMALENPRMRNSEISKQLGYQWKMLTEAEKWPFF 110
Query: 155 EEAERLRMQHKKDHPDYKYQPRRR 178
+EA++L+ H++ +P+YKY+PRR+
Sbjct: 111 QEAQKLQAMHREKYPNYKYRPRRK 134
CPU time: 0.03 user secs. 0.01 sys. secs 0.04 total secs.
Lambda K H
0.311 0.130 0.399
Gapped

MegaBLAST

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

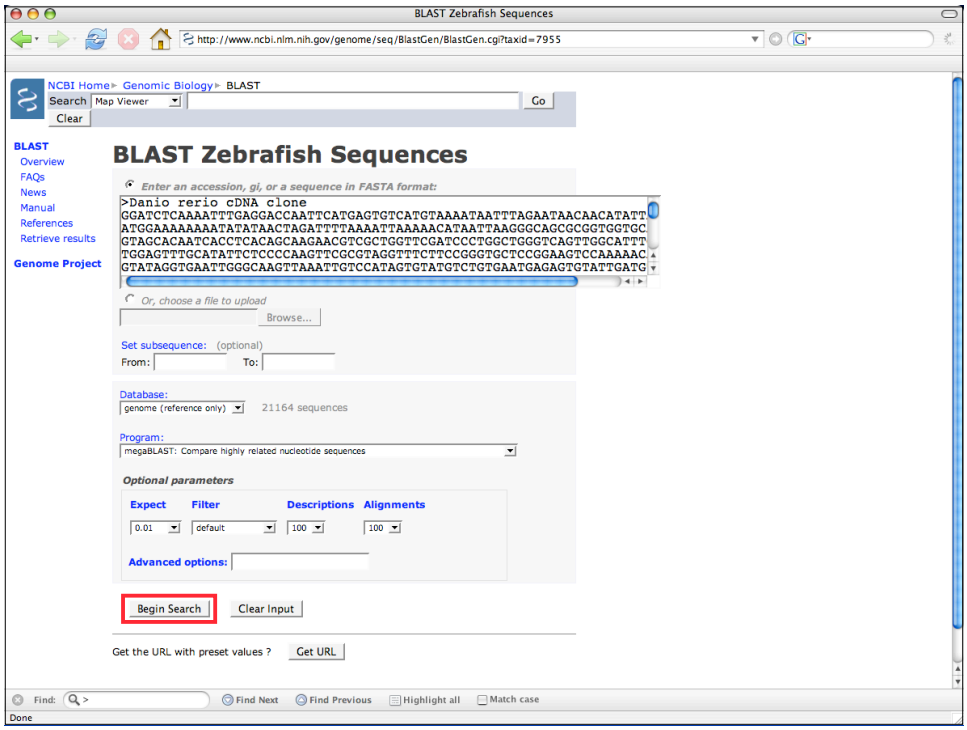
BLASTN *vs.* MegaBLAST

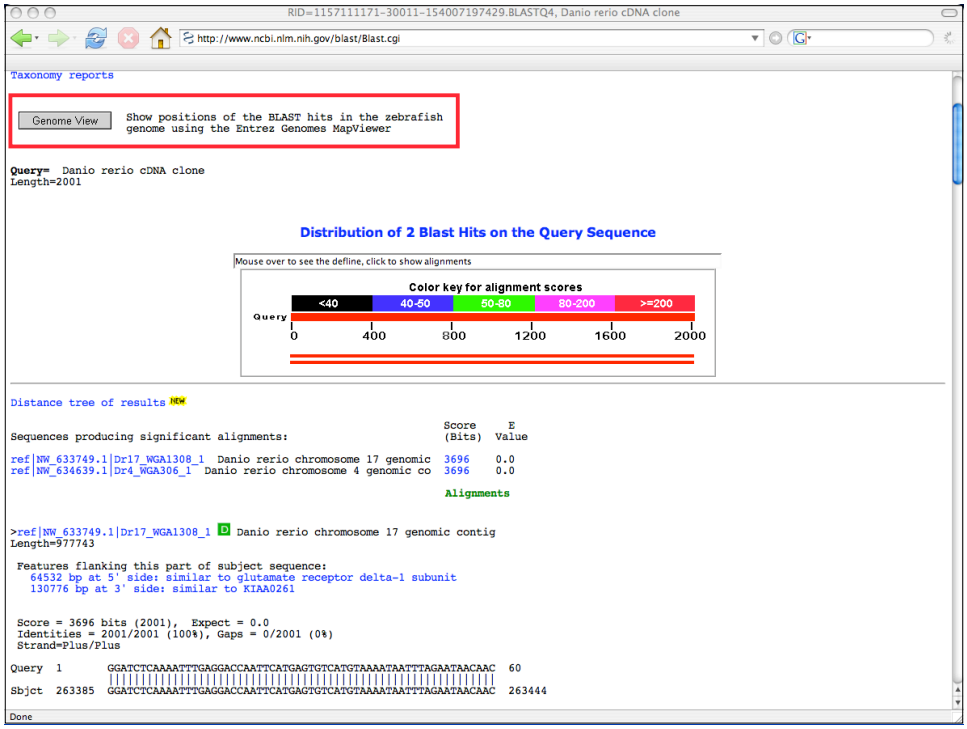
- Word size
  - BLASTN default = 11
  - MegaBLAST default = 28
- *Non-affine* gap penalties

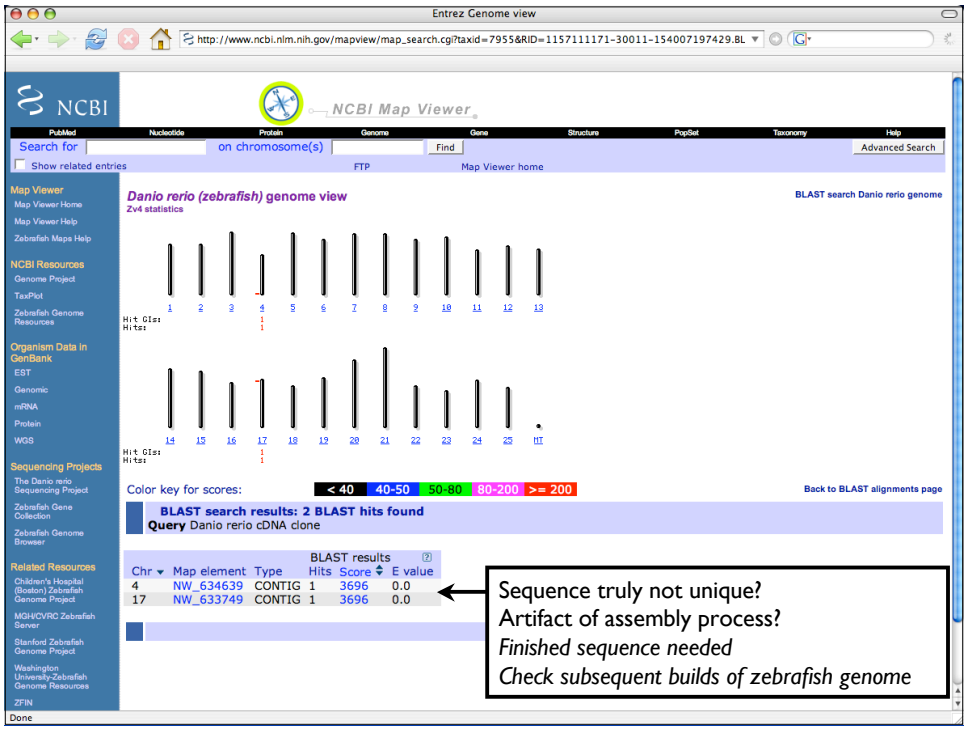
Deduction for a gap = $r/2 - q$

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


NCBI BLAST
http://www.ncbi.nlm.nih.gov/BLAST/
http://www.ncbi.nlm.nih.gov/BLAST
NCBI → BLAST
About
Getting started
News
FAQs
More info
NAR 2004
NCBI Handbook
The Statistics of Sequence Similarity Scores
Software
Downloads
Developer info
Other resources
References
NCBI Contributors
Mailing list
Contact us
The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.
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)
Position-specific iterated and pattern-hit initiated BLAST (PSI- and PHI-BLAST)
Search for short, nearly exact matches
Search the conserved domain database (rpsblast)
Protein homology by domain architecture (cdart)
Translated
Translated query vs. protein database (blastx)
Protein query vs. translated database (tblastn)
Translated query vs. translated database (tblastx)
Genomes
Human, mouse, rat, chimp, cow, pig, dog, sheep, cat
Chicken, puffer fish, zebrafish
Fly, honey bee, other insects
Microbes, environmental samples
Plants, nematodes
Fungi, protozoa, other eukaryotes
Special
Search for gene expression data (GEO BLAST)
Align two sequences (bl2seq)
Screen for vector contamination (VecScreen)
Immunoglobin BLAST (IgBlast)
SNP BLAST
Meta
Retrieve results
Disclaimer
Privacy statement
Accessibility
This page is valid XHTML 1.0.
Done


BLAST Zebrafish Sequences
http://www.ncbi.nlm.nih.gov/genome/seq/BlastGen/BlastGen.cgi?taxid=7955
NCBI Home ▸ Genomic Biology ▸ BLAST
Search Map Viewer
Go
Clear
BLAST
Overview
FAQs
News
Manual
References
Retrieve results
Genome Project
BLAST Zebrafish Sequences
Enter an accession, gi, or a sequence in FASTA format:
>Danio rerio cDNA clone
GGATCTCAAAATTTGAGGACCAATTCATGAGTGTCATGTAAAATAATTTAGAATAACAACATATT
ATGGAAAAAAAATATATAACTAGATTTTAAAATTAAAAACATAATTAAGGGCAGCGCGGTGGTGC
GTAGCACAATCACCTCACAGCAAGAACGTCGCTGGTTCGATCCCTGGCTGGGTCAGTTGGCATTT
TGGAGTTTGCATATTCTCCCCAAGTTCGCGTAGGTTTCTTCCGGGTGCTCCGGAAGTCCAAAAAC
GTATAGGTGAATTGGGCAAGTTAAATTGTCCATAGTGTATGTCTGTGAATGAGAGTGTATTGATG
Or, choose a file to upload
Browse...
Set subsequence: (optional)
From:
To:
Database:
genome (reference only)
21164 sequences
Program:
megaBLAST: Compare highly related nucleotide sequences
Optional parameters
Expect
Filter
Descriptions
Alignments
0.01
default
100
100
Advanced options:
Begin Search
Clear Input
Get the URL with preset values ?
Get URL
Find:
Find Next
Find Previous
Highlight all
Match case
Done


RID=1157111171-30011-154007197429.BLASTQ4, Danio rerio cDNA clone
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi
Taxonomy reports
Genome View
Show positions of the BLAST hits in the zebrafish genome using the Entrez Genomes MapViewer
Query= Danio rerio cDNA clone
Length=2001
Distribution of 2 Blast Hits on the Query Sequence
Mouse over to see the defline, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
0
400
800
1200
1600
2000
Distance tree of results
Sequences producing significant alignments:
Score (Bits)
E Value
ref|NW_633749.1|Dr17_WGA1308_1 Danio rerio chromosome 17 genomic 3696 0.0
ref|NW_634639.1|Dr4_WGA306_1 Danio rerio chromosome 4 genomic co 3696 0.0
Alignments
>ref|NW_633749.1|Dr17_WGA1308_1 Danio rerio chromosome 17 genomic contig
Length=977743
Features flanking this part of subject sequence:
64532 bp at 5' side: similar to glutamate receptor delta-1 subunit
130776 bp at 3' side: similar to KIAA0261
Score = 3696 bits (2001), Expect = 0.0
Identities = 2001/2001 (100%), Gaps = 0/2001 (0%)
Strand=Plus/Plus
Query 1 GGATCTCAAAATTTGAGGACCAATTCATGAGTGTCATGTAAAATAATTTAGAATAACAAC 60
Sbjct 263385 GGATCTCAAAATTTGAGGACCAATTCATGAGTGTCATGTAAAATAATTTAGAATAACAAC 263444
Done


Entrez Genome view
http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=7955&RID=1157111171-30011-154007197429.BL
NCBI
NCBI Map Viewer
PubMed Nucleotide Protein Genome Gene Structure PopSet Taxonomy Help
Search for
on chromosome(s)
Find
Advanced Search
Show related entries
FTP
Map Viewer home
Map Viewer
Map Viewer Home
Map Viewer Help
Zebrafish Maps Help
NCBI Resources
Genome Project
TaxPlot
Zebrafish Genome Resources
Organism Data in GenBank
EST
Genomic
mRNA
Protein
WGS
Sequencing Projects
The Danio rerio Sequencing Project
Zebrafish Gene Collection
Zebrafish Genome Browser
Related Resources
Children's Hospital (Boston) Zebrafish Genome Project
MGH/CVRC Zebrafish Server
Stanford Zebrafish Genome Project
Washington University-Zebrafish Genome Resources
ZFIN
Danio rerio (zebrafish) genome view
Zv4 statistics
BLAST search Danio rerio genome
Hit GIs:
Hits:
Color key for scores:
< 40
40-50
50-80
80-200
>= 200
Back to BLAST alignments page
BLAST search results: 2 BLAST hits found
Query Danio rerio cDNA clone
BLAST results
Chr
Map element
Type
Hits
Score
E value
4 NW_634639 CONTIG 1 3696 0.0
17 NW_633749 CONTIG 1 3696 0.0
Sequence truly not unique?
Artifact of assembly process?
Finished sequence needed
Check subsequent builds of zebrafish genome
Done

Overview

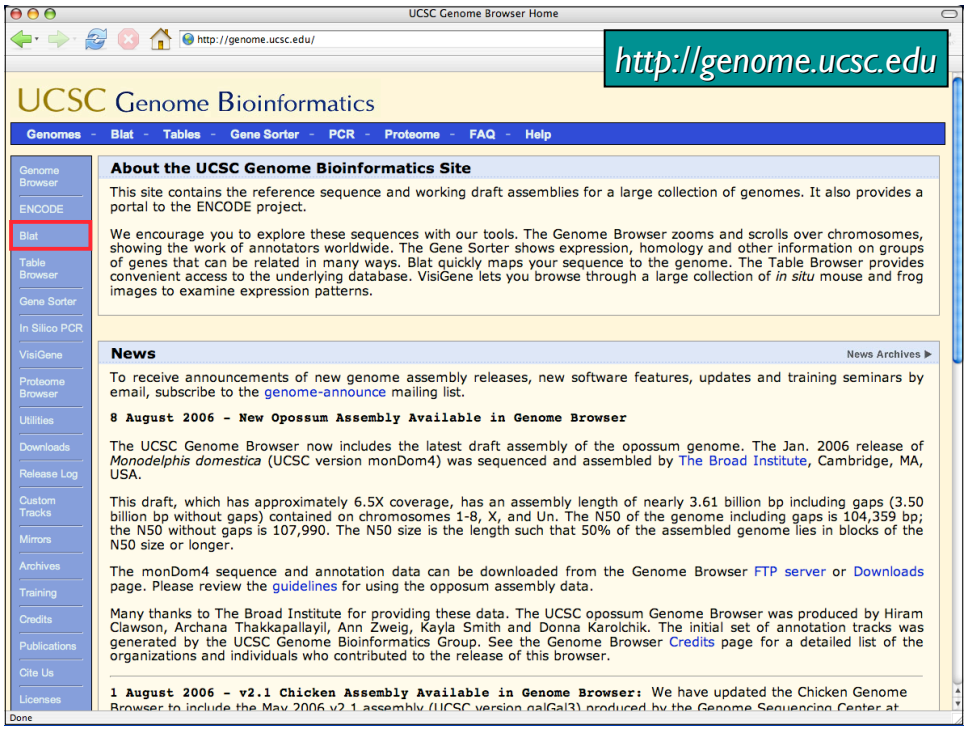
- Week 2: Comparative methods and concepts
  - Similarity *vs.* Homology
  - Global *vs.* Local Alignments
  - Scoring Matrices
  - BLAST
  - BLAT
- Week 3: Predictive methods and concepts
  - Profiles, patterns, motifs, and domains
  - Secondary structure prediction
  - Structures: VAST, Cn3D, and *de novo* prediction

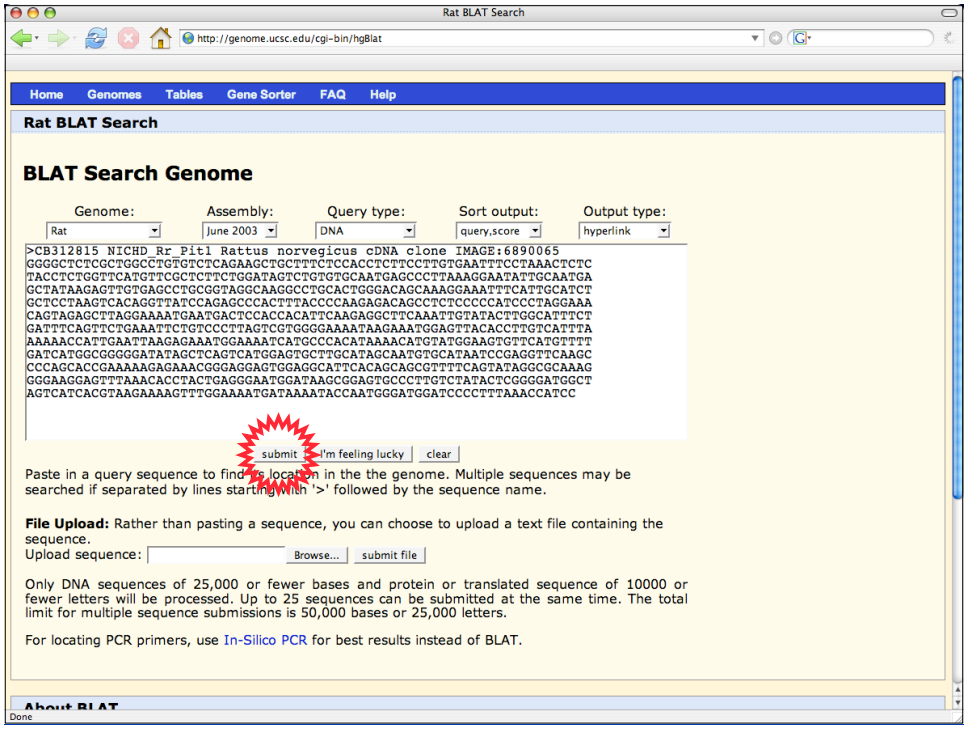
BLAT

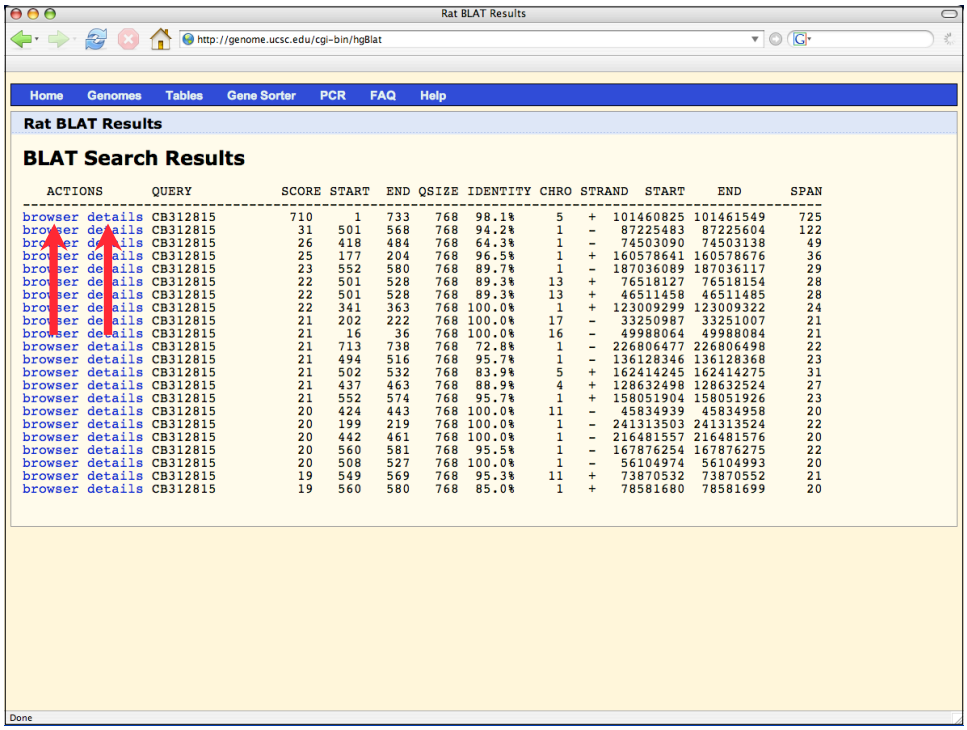
- "BLAST-Like Alignment Tool"
- Designed to rapidly-align longer nucleotide sequences ($L \geq 40$) having > 95% sequence similarity
- Can find exact matches reliably down to $L = 33$
- Method of choice when looking for exact matches in nucleotide databases
- 500 times faster for mRNA/DNA searches
-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

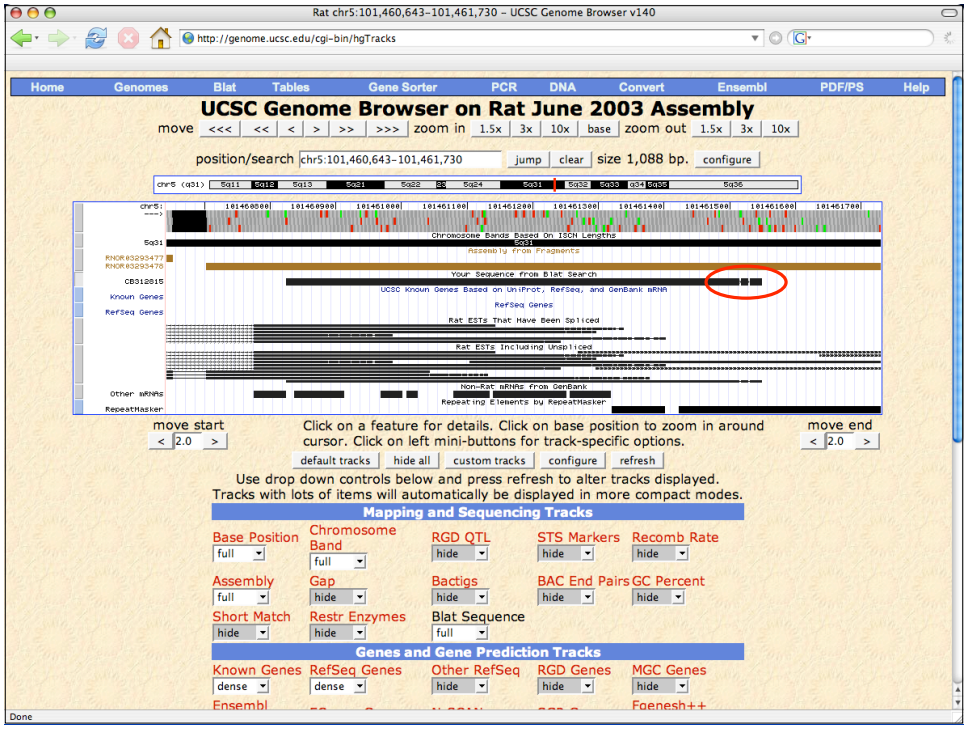
When to Use BLAT

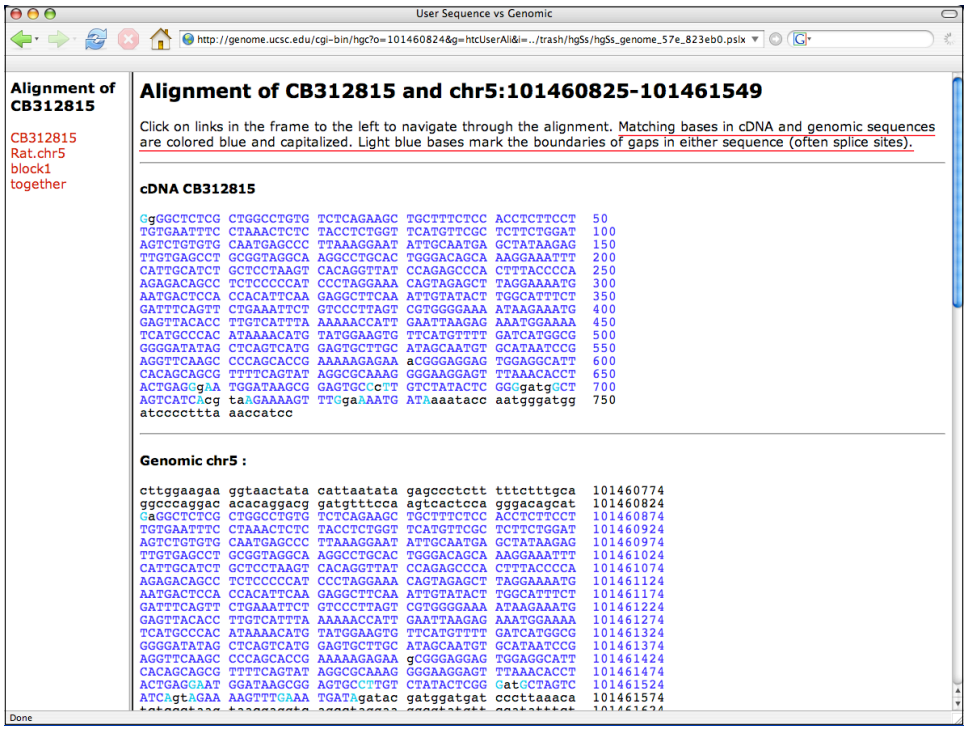
- To characterize an unknown gene or sequence fragment
  - Find its genomic coordinates
  - Determine gene structure (the presence and position of exons)
  - Identify markers of interest in the vicinity of a sequence
- To find highly-similar sequences
  - Identify gene family members
  - Identify putative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

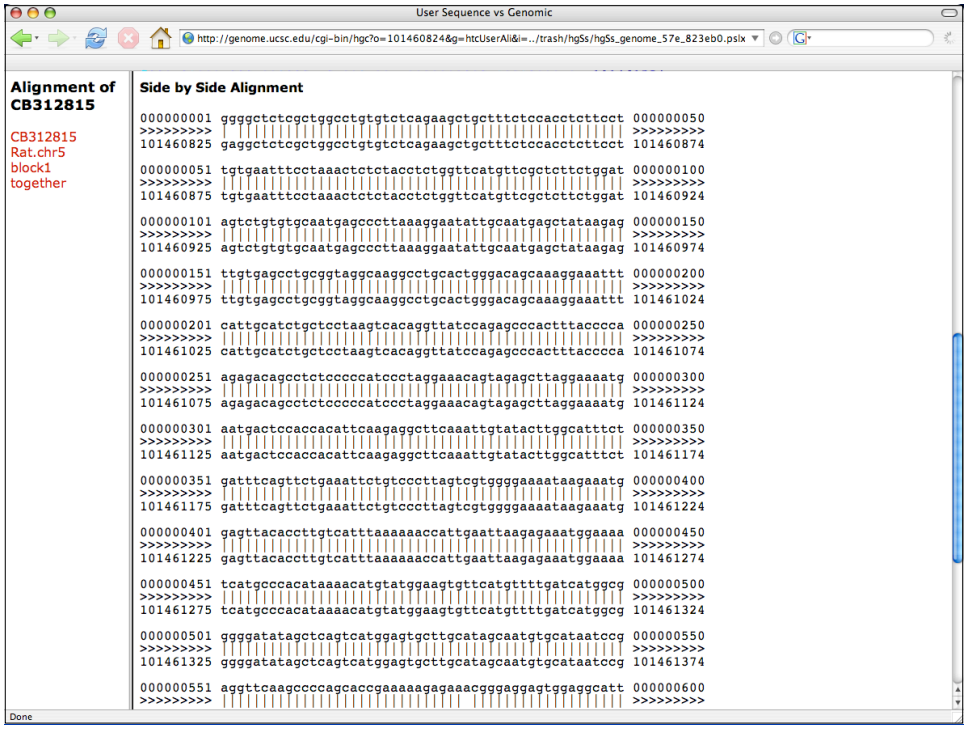



Rat BLAT Search
http://genome.ucsc.edu/cgi-bin/hgBlat
Home Genomes Tables Gene Sorter FAQ Help
Rat BLAT Search
BLAT Search Genome
Genome: Assembly: Query type: Sort output: Output type:
Rat June 2003 DNA query,score hyperlink
>CB312815 NICHD_Rr_Pit1 Rattus norvegicus cDNA clone IMAGE:6890065
GGGGCTCTCGCTGGCCTGTGTCTCAGAAGCTGCTTTCTCCACCTCTTCCTTGTGAATTTCCTAAACTCTC
TACCTCTGGTTCATGTTCGCTCTTCTGGATAGTCTGTGTGCAATGAGCCCTTAAAGGAATATTGCAATGA
GCTATAAGAGTTGTGAGCCTGCGGTAGGCAAGGCCTGCACTGGGACAGCAAAGGAAATTTCATTGCATCT
GCTCCTAAGTCACAGGTTATCCAGAGCCCACTTTACCCCAAGAGACAGCCTCTCCCCCATCCCTAGGAAA
CAGTAGAGCTTAGGAAAATGAATGACTCCACCACATTCAAGAGGCTTCAAATTGTATACTTGGCATTTCT
GATTTCAGTTCTGAAATTCTGTCCCTTAGTCGTGGGGAAAATAAGAAATGGAGTTACACCTTGTCATTTA
AAAACCATTGAATTAAGAGAAATGGAAAATCATGCCCACATAAAACATGTATCGAAGTGTTCATGTTTT
GATCATGGCGGGGGATATAGCTCAGTCATGGAGTGCTTGCATAGCAATGTGCATAATCCGAGGTTCAAGC
CCCAGCACCGAAAAAGAGAAACGGGAGGAGTGGAGGCATTCACAGCAGCGTTTTCAGTATAGGCGCAAAG
GGGAAGGAGTTTAAACACCTACTGAGGGAATGGATAAGCGGAGTGCCCTTGTCTATACTCGGGGATGGCT
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: 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.
For locating PCR primers, use In-Silico PCR for best results instead of BLAT.
About BLAT
Done


Rat BLAT Results
http://genome.ucsc.edu/cgi-bin/hgBlat
Home Genomes Tables Gene Sorter PCR 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
browser details CB312815 31 501 568 768 94.2% 1 - 87225483 87225604 122
browser details CB312815 26 418 484 768 64.3% 1 - 74503090 74503138 49
browser details CB312815 25 177 204 768 96.5% 1 + 160578641 160578676 36
browser details CB312815 23 552 580 768 89.7% 1 - 187036089 187036117 29
browser details CB312815 22 501 528 768 89.3% 13 + 76518127 76518154 28
browser details CB312815 22 501 528 768 89.3% 13 + 46511458 46511485 28
browser details CB312815 22 341 363 768 100.0% 1 + 123009299 123009322 24
browser details CB312815 21 202 222 768 100.0% 17 - 33250987 33251007 21
browser details CB312815 21 16 36 768 100.0% 16 - 49988064 49988084 21
browser details CB312815 21 713 738 768 72.8% 1 - 226806477 226806498 22
browser details CB312815 21 494 516 768 95.7% 1 - 136128346 136128368 23
browser details CB312815 21 502 532 768 83.9% 5 + 162414245 162414275 31
browser details CB312815 21 437 463 768 88.9% 4 + 128632498 128632524 27
browser details CB312815 21 552 574 768 95.7% 1 + 158051904 158051926 23
browser details CB312815 20 424 443 768 100.0% 11 - 45834939 45834958 20
browser details CB312815 20 199 219 768 100.0% 1 - 241313503 241313524 22
browser details CB312815 20 442 461 768 100.0% 1 - 216481557 216481576 20
browser details CB312815 20 560 581 768 95.5% 1 - 167876254 167876275 22
browser details CB312815 20 508 527 768 100.0% 1 - 56104974 56104993 20
browser details CB312815 19 549 569 768 95.3% 11 + 73870532 73870552 21
browser details CB312815 19 560 580 768 85.0% 1 + 78581680 78581699 20
Done


Rat chr5:101,460,643-101,461,730 - UCSC Genome Browser v140
http://genome.ucsc.edu/cgi-bin/hgTracks
Home Genomes Blat Tables Gene Sorter PCR DNA Convert Ensembl PDF/PS Help
UCSC Genome Browser on Rat June 2003 Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
position/search chr5:101,460,643-101,461,730 jump clear size 1,088 bp. configure
Chromosome Bands Based On ISCN Lengths
Assembly from Fragments
Your Sequence from Blat Search
UCSC Known Genes Based on UniProt, RefSeq, and GenBank mRNA
RefSeq Genes
Rat ESTs That Have Been Spliced
Rat ESTs Including Unspliced
Non-Rat mRNAs from GenBank
Repeating Elements by RepeatMasker
move start
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 end
default tracks hide all custom tracks configure refresh
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 Restr Enzymes Blat Sequence
Genes and Gene Prediction Tracks
Known Genes RefSeq Genes Other RefSeq RGD Genes MGC Genes
Ensembl
Done


User Sequence vs Genomic
http://genome.ucsc.edu/cgi-bin/hgc?o=101460824&g=htcUserAli&i=../trash/hgSs/hgSs_genome_57e_823eb0.psix
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
gGGCTCTCG CTGGCCTGTG TCTCAGAAGC TGCTTTCTCC ACCTCTTCCT 50
TGTGAATTTC CTAAACTCTC TACCTCTGGT TCATGTTCGC TCTTCTGGAT 100
AGTCTGTGTG CAATGAGCCC TTAAAGGAAT ATTGCAATGA GCTATAAGAG 150
TTGTGAGCCT GCGGTAGGCA AGGCCTGCAC TGGGACAGCA AAGGAAATTT 200
CATTGCATCT GCTCCTAAGT CACAGGTTAT CCAGAGCCCA CTTTACCCCA 250
AGAGACAGCC TCTCCCCCAT CCCTAGGAAA CAGTAGAGCT TAGGAAAATG 300
AATGACTCCA CCACATTCAA GAGGCTTCAA ATTGTATACT TGGCATTTCT 350
GATTTCAGTT CTGAAATTCT GTCCCTTAGT CGTGGGGAAA ATAAGAAATG 400
GAGTTACACC TTGTCATTTA AAAAACCATT GAATTAAGAG AAATGGAAAA 450
TCATGCCCAC ATAAAACATG TATGGAAGTG TTCATGTTTT GATCATGGCG 500
GGGGATATAG CTCAGTCATG GAGTGCTTGC ATAGCAATGT GCATAATCCG 550
AGGTTCAAGC CCCAGCACCG AAAAAGAGAA aCGGGAGGAG TGGAGGCATT 600
CACAGCAGCG TTTTCAGTAT AGGCGCAAAG GGGAAGGAGT TTAAACACCT 650
ACTGAGGgAA TGGATAAGCG GAGTGCCcTT GTCTATACTC GGGgatgGCT 700
AGTCATCAcg taAGAAAAGT TTGgaAAATG ATAaaatacc aatgggatgg 750
atccccttta aaccatcc
Genomic chr5 :
cttggaagaa ggtaactata cattaatata gagccctctt tttctttgca 101460774
ggcccaggac acacaggacg gatgtttcca agtcactcca gggacagcat 101460824
GaGGCTCTCG CTGGCCTGTG TCTCAGAAGC TGCTTTCTCC ACCTCTTCCT 101460874
TGTGAATTTC CTAAACTCTC TACCTCTGGT TCATGTTCGC TCTTCTGGAT 101460924
AGTCTGTGTG CAATGAGCCC TTAAAGGAAT ATTGCAATGA GCTATAAGAG 101460974
TTGTGAGCCT GCGGTAGGCA AGGCCTGCAC TGGGACAGCA AAGGAAATTT 101461024
CATTGCATCT GCTCCTAAGT CACAGGTTAT CCAGAGCCCA CTTTACCCCA 101461074
AGAGACAGCC TCTCCCCCAT CCCTAGGAAA CAGTAGAGCT TAGGAAAATG 101461124
AATGACTCCA CCACATTCAA GAGGCTTCAA ATTGTATACT TGGCATTTCT 101461174
GATTTCAGTT CTGAAATTCT GTCCCTTAGT CGTGGGGAAA ATAAGAAATG 101461224
GAGTTACACC TTGTCATTTA AAAAACCATT GAATTAAGAG AAATGGAAAA 101461274
TCATGCCCAC ATAAAACATG TATGGAAGTG TTCATGTTTT GATCATGGCG 101461324
GGGGATATAG CTCAGTCATG GAGTGCTTGC ATAGCAATGT GCATAATCCG 101461374
AGGTTCAAGC CCCAGCACCG AAAAAGAGAA gCGGGAGGAG TGGAGGCATT 101461424
CACAGCAGCG TTTTCAGTAT AGGCGCAAAG GGGAAGGAGT TTAAACACCT 101461474
ACTGAGGAAT GGATAAGCGG AGTGCCTTGT CTATACTCGG GatGCTAGTC 101461524
ATCAgtAGAA AAGTTTGAAA TGATAgatac gatggatgat ccctaaaca 101461574
Done



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