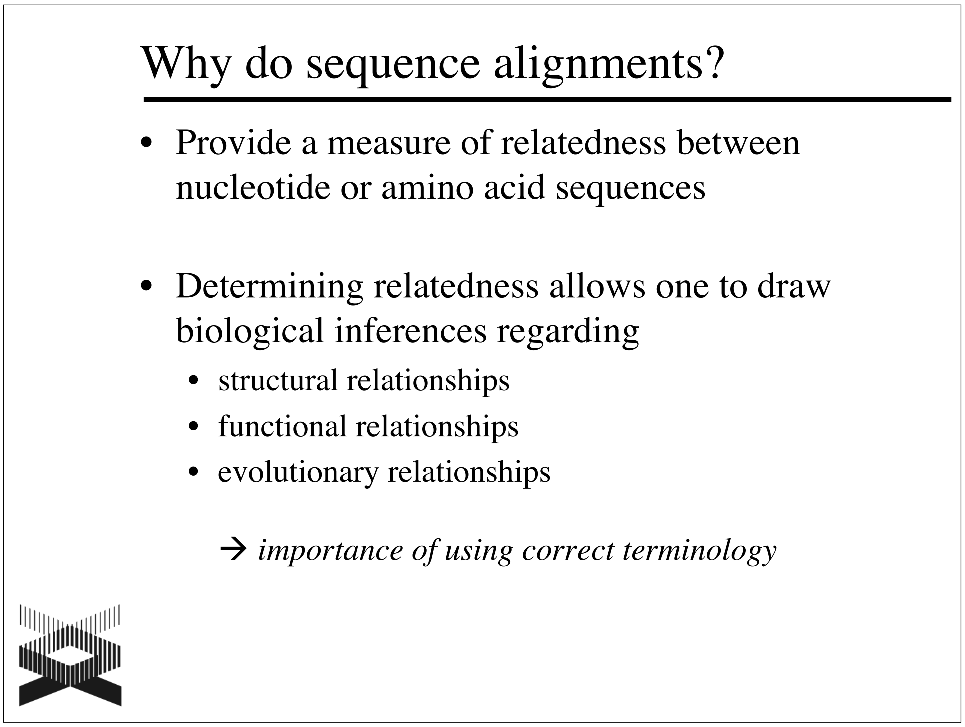
Current Topics in Genome Analysis Spring 2005

Week 4
Biological Sequence Analysis I

Andy Baxevanis, Ph.D.

Overview

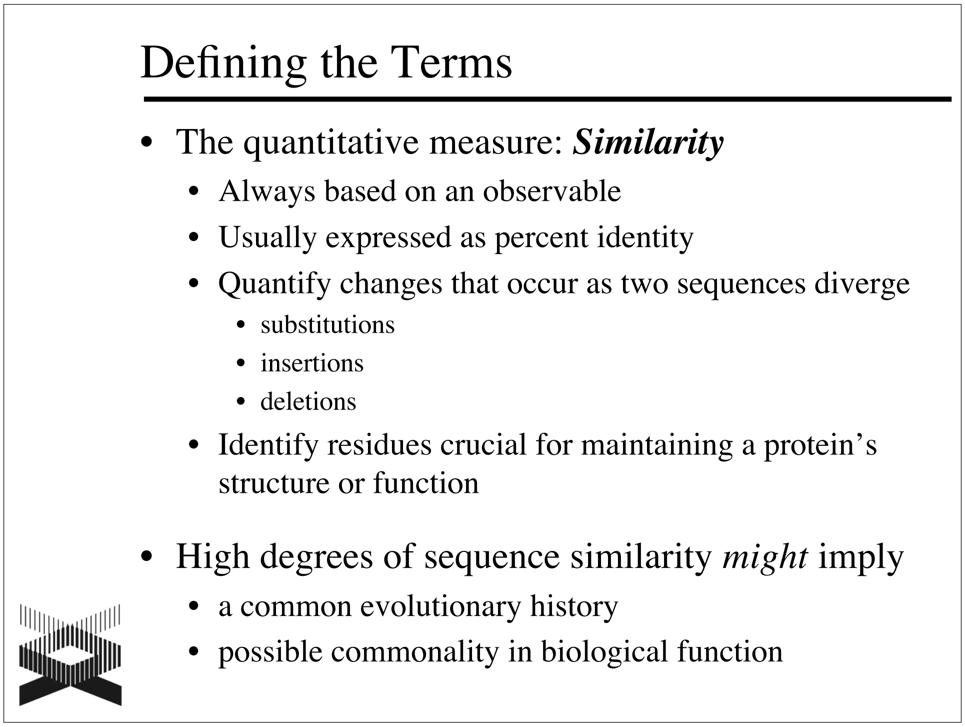
- Week 4: Comparative methods and concepts
  - Similarity *vs*. Homology
  - Global *vs*. Local Alignments
  - Scoring Matrices
  - BLAST
  - BLAT
- Week 5: Predictive methods and concepts
  - Profiles, patterns, motifs, and domains
  - Secondary structure prediction
  - Structures: VAST, Cn3D, and *de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



Why do sequence alignments?

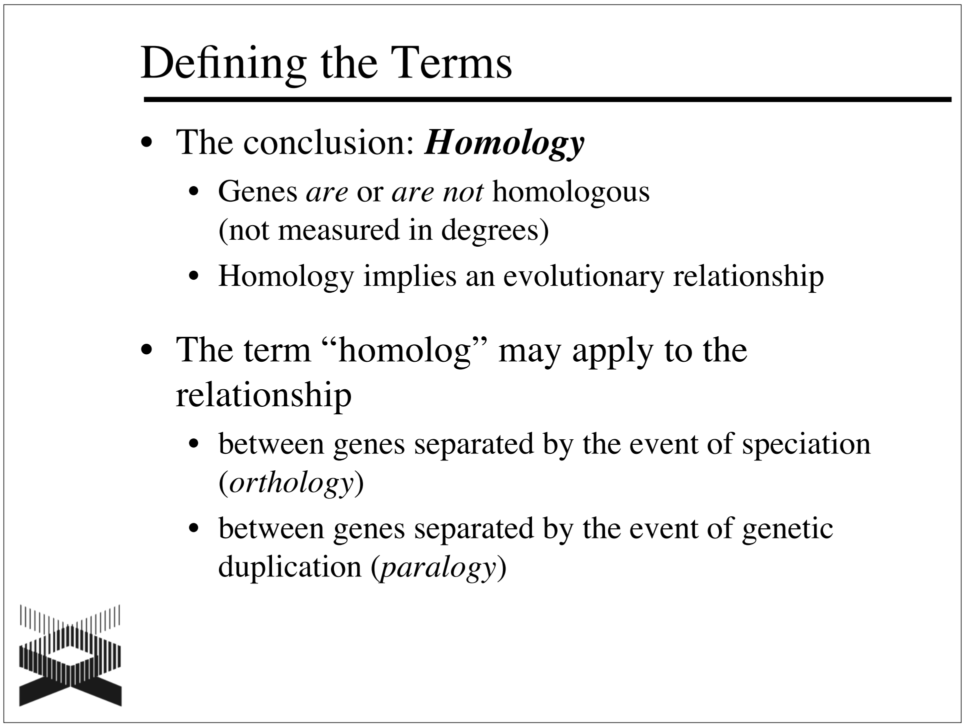
- Provide a measure of relatedness between nucleotide or amino acid sequences
- Determining relatedness allows one to draw biological inferences regarding
  - structural relationships
  - functional relationships
  - evolutionary relationships

→ *importance of using correct terminology*



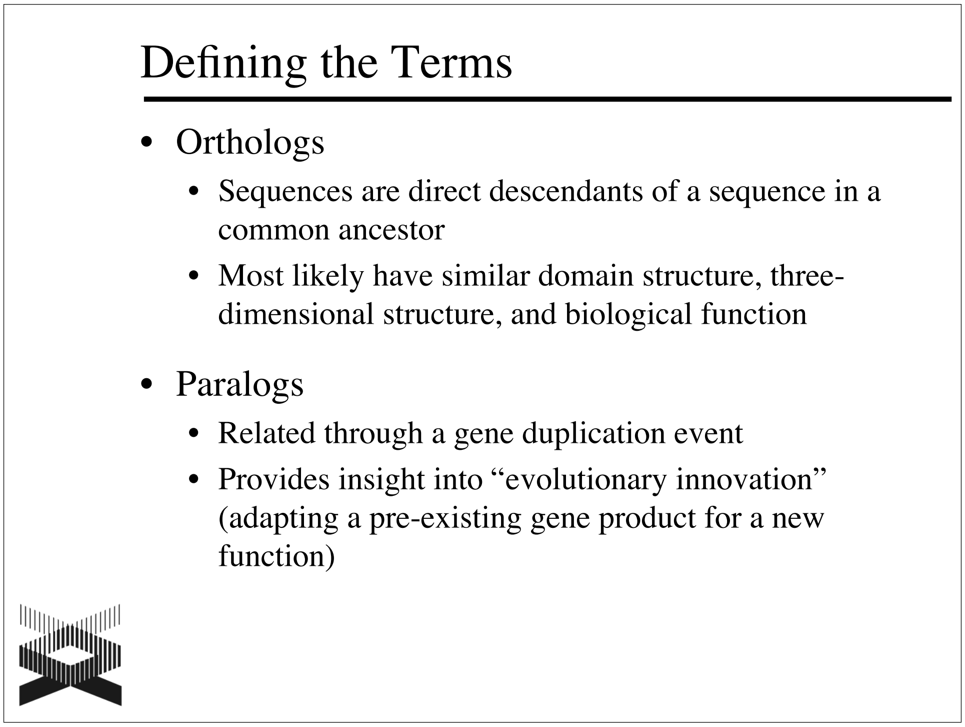
Defining the Terms

- 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*)

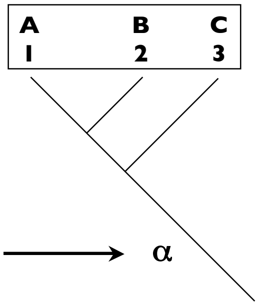


Defining the Terms

- 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

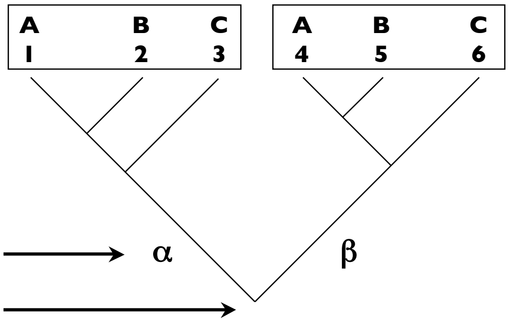


Most recent common ancestor →

Defining the Terms

Paralogs

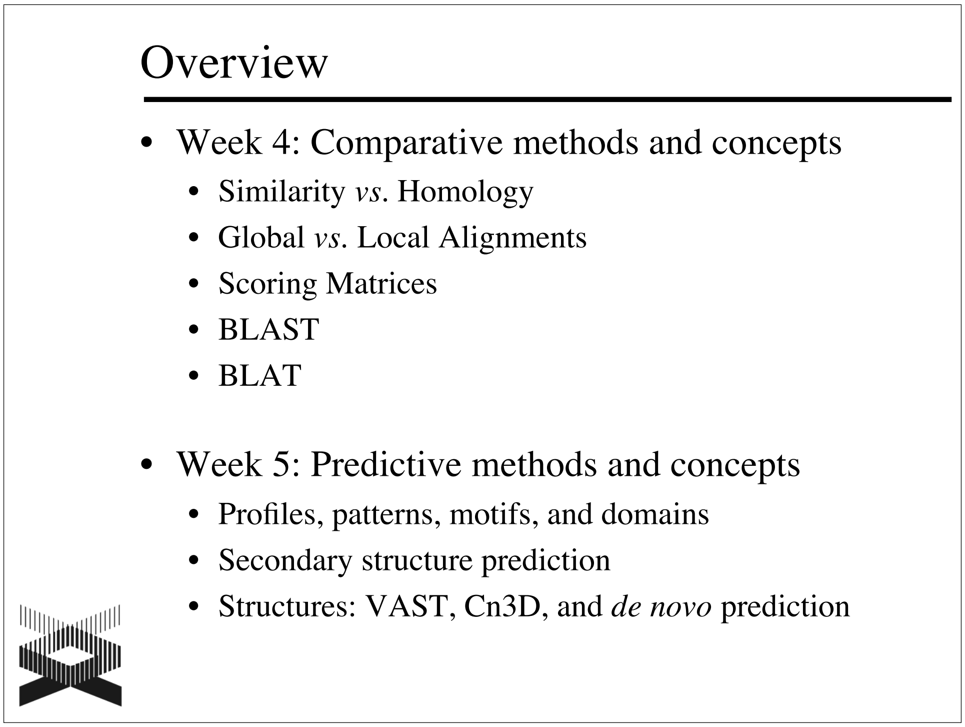
Orthologs



Most recent common ancestor →

Gene duplication →

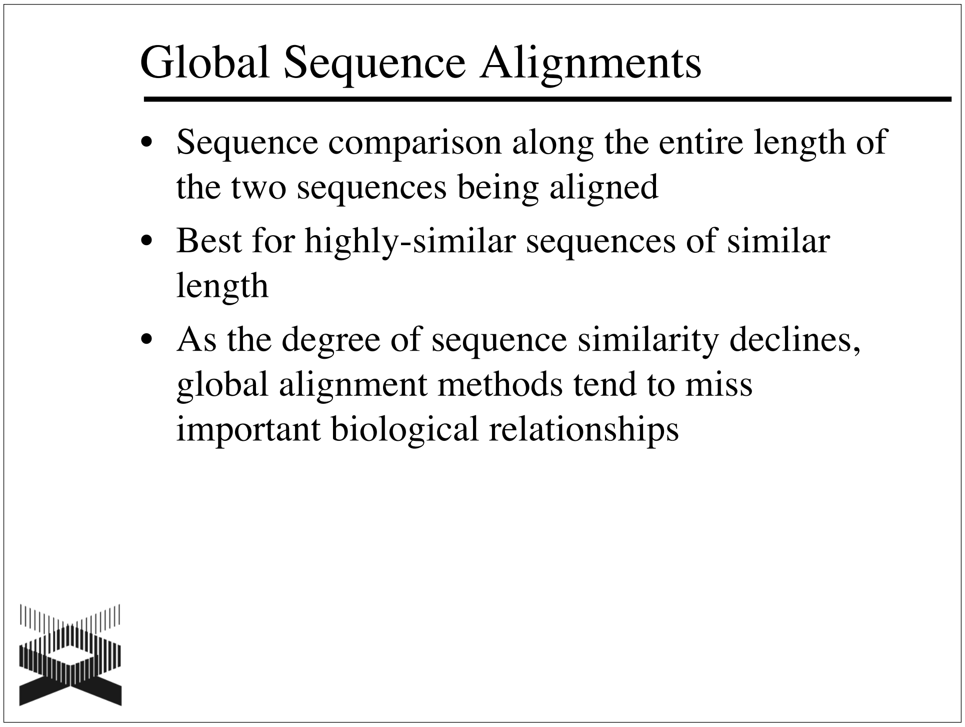
- Genes 1-3 are orthologous
- Genes 4-6 are orthologous
- Any pair of α and β genes are paralogous (genes related through a gene duplication event)



Overview

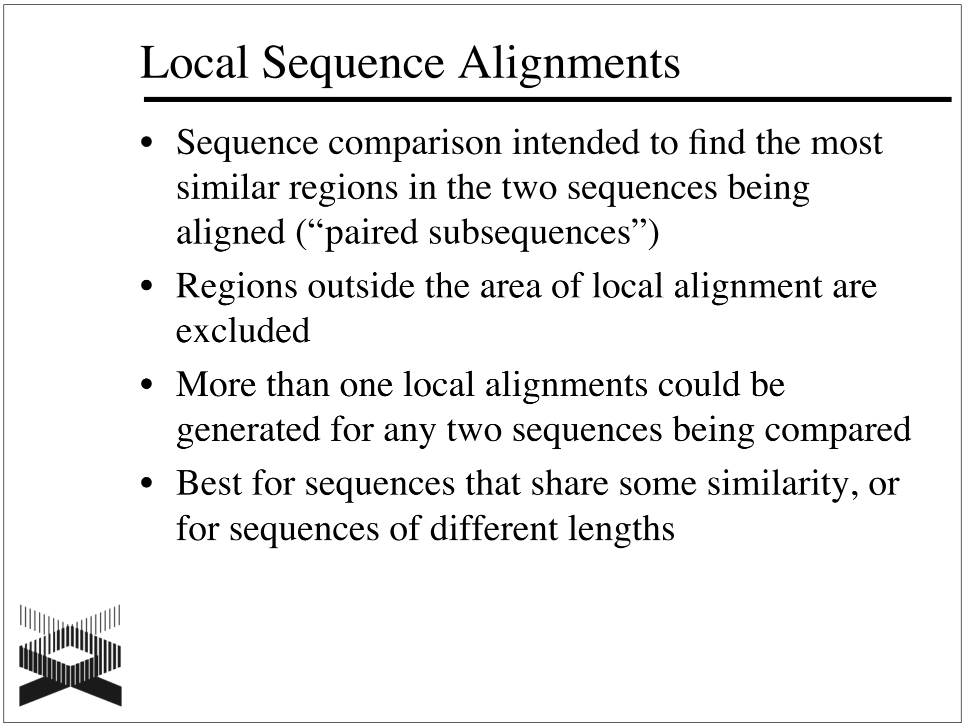
- Week 4: Comparative methods and concepts
  - Similarity *vs*. Homology
  - Global *vs*. Local Alignments
  - Scoring Matrices
  - BLAST
  - BLAT

- Week 5: Predictive methods and concepts
  - Profiles, patterns, motifs, and domains
  - Secondary structure prediction
  - Structures: VAST, Cn3D, and *de novo* prediction



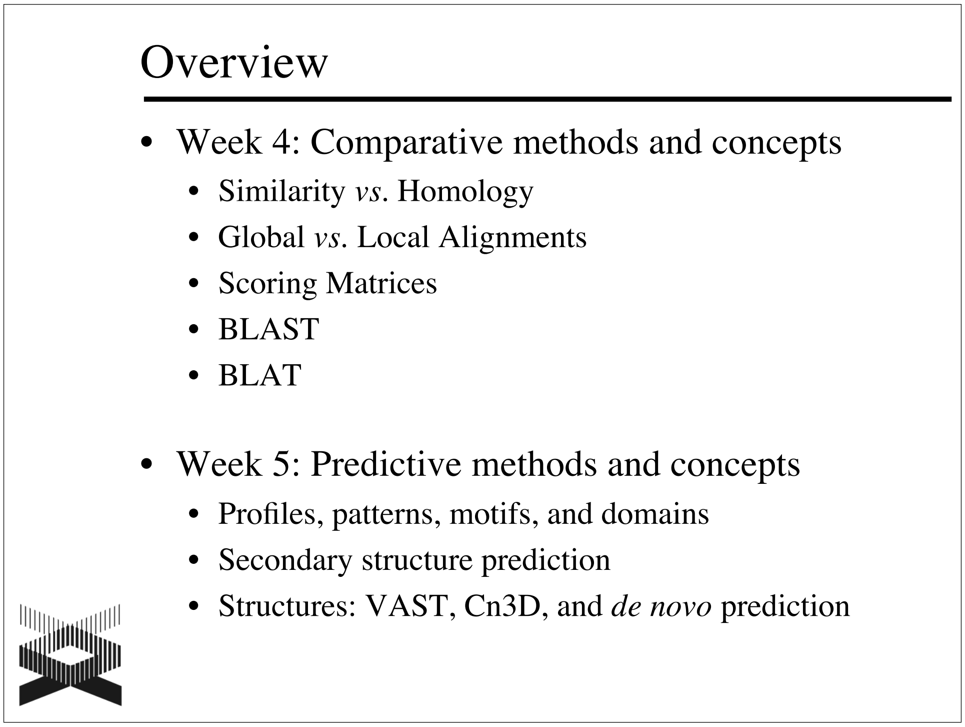
Global Sequence Alignments

- 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



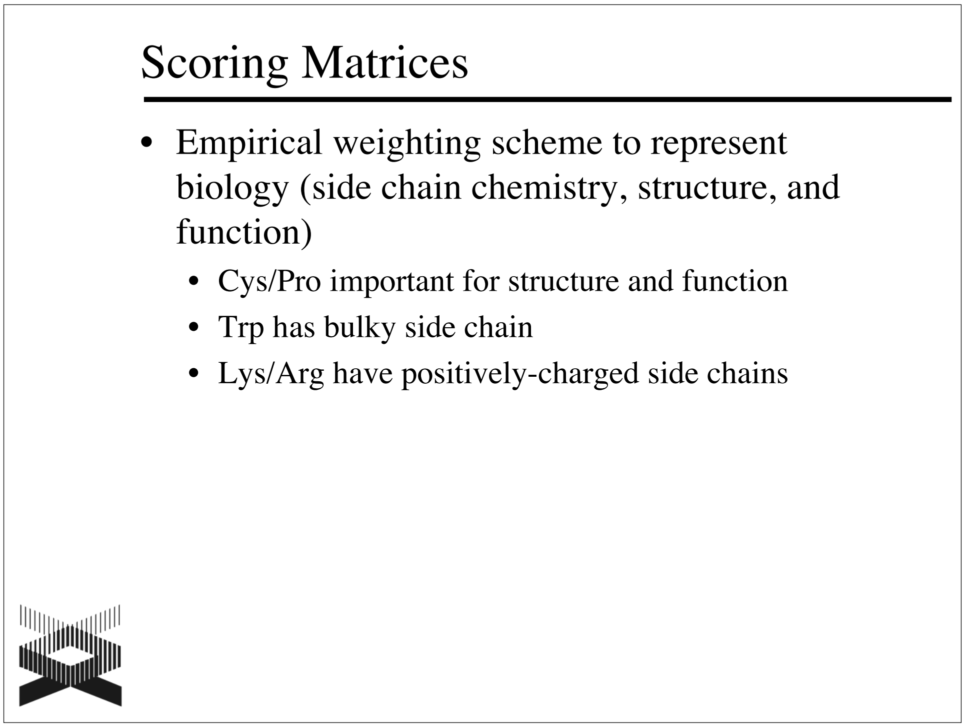
Local Sequence Alignments

- 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



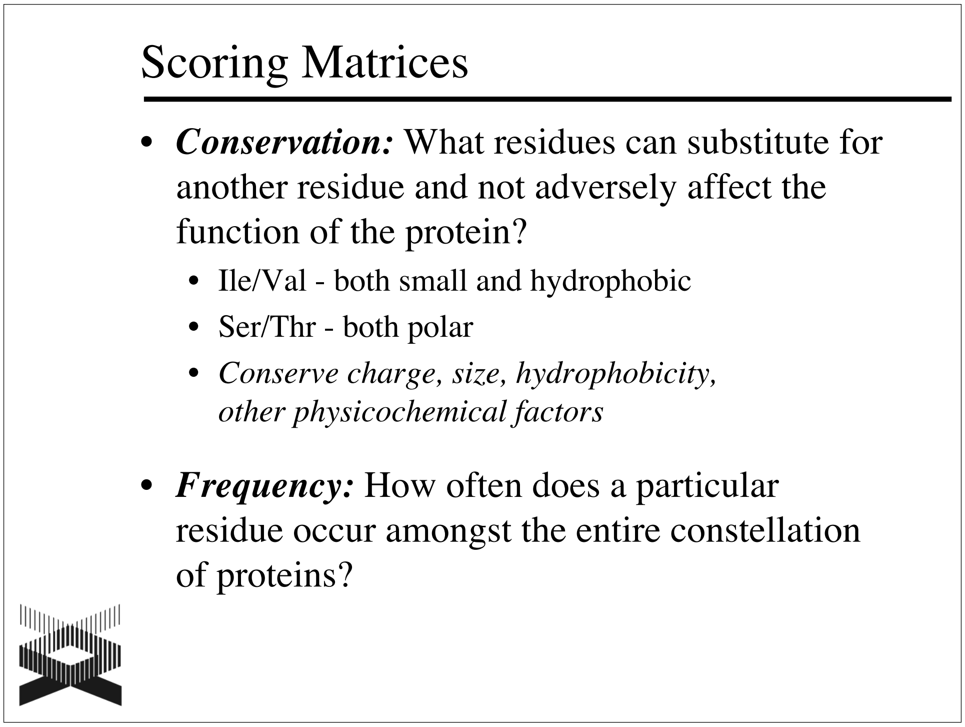
Overview

- 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

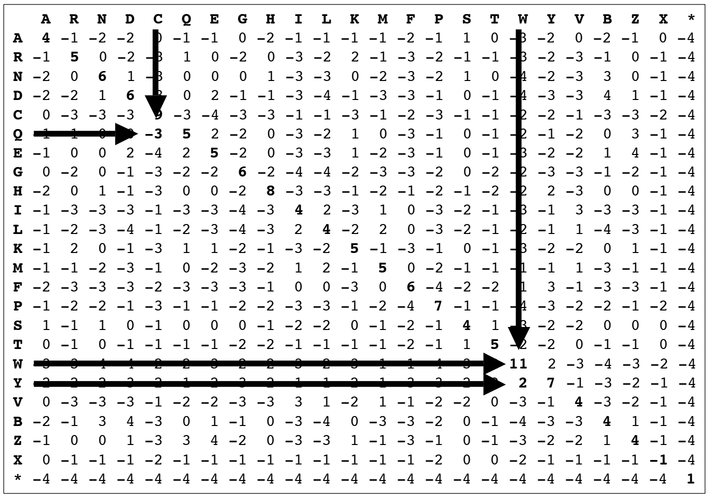
- Importance of understanding scoring matrices
  - Appear in all analyses involving sequence comparison
  - Implicitly represent particular evolutionary patterns
  - Choice of matrix can strongly influence outcomes

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

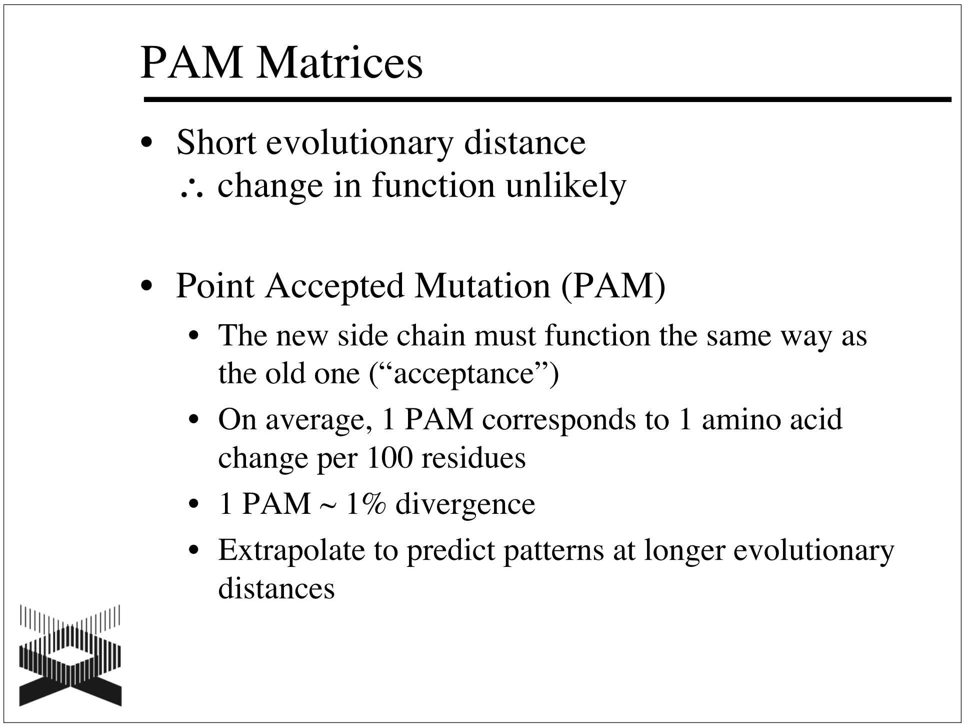


	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	**4**	−1	−2	−2	0	−1	−1	0	−2	−1	−1	−1	−1	−2	−1	1	0	−3	−2	0	−2	−1	0	−4
R	−1	**5**	0	−2	−3	1	0	−2	0	−3	−2	2	−1	−3	−2	−1	−1	−3	−2	−3	−1	0	−1	−4
N	−2	0	**6**	1	−3	0	0	0	1	−3	−3	0	−2	−3	−2	1	0	−4	−2	−3	3	0	−1	−4
D	−2	−2	1	**6**	−3	0	2	−1	−1	−3	−4	−1	−3	−3	−1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

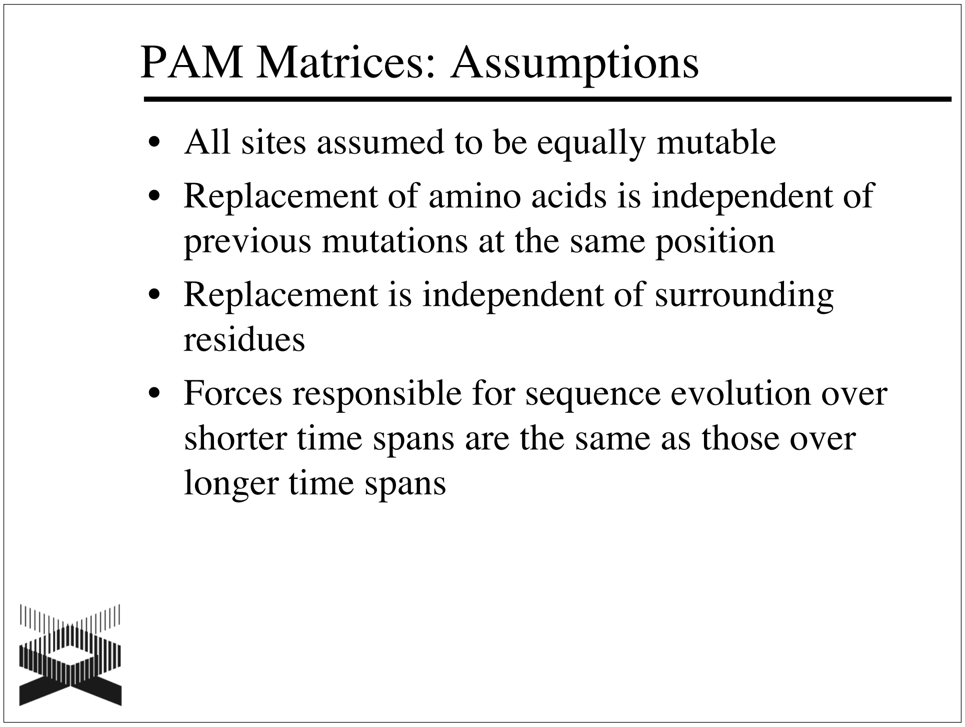
- Margaret Dayhoff and colleagues, 1978
  - Look at patterns of substitutions in highly related proteins (> 85% similar) within multiple sequence alignments
  - Analysis documented 1572 changes in 71 groups of proteins examined
  - Substitution tables constructed based on results
  - Given high degree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



PAM Matrices

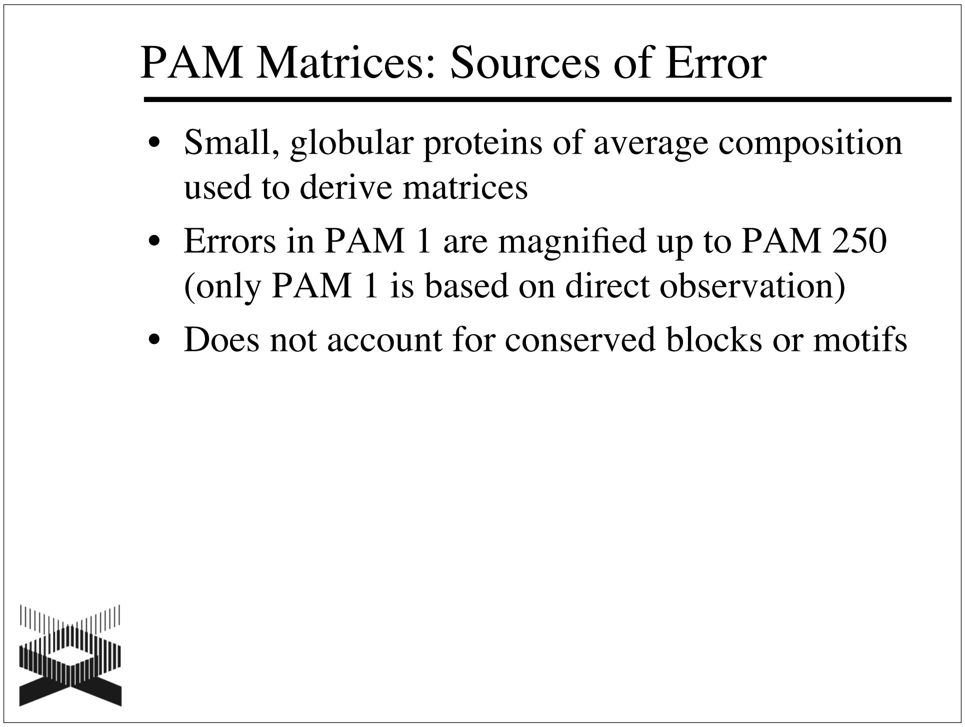
- Short evolutionary distance
  ∴ change in function unlikely

- 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



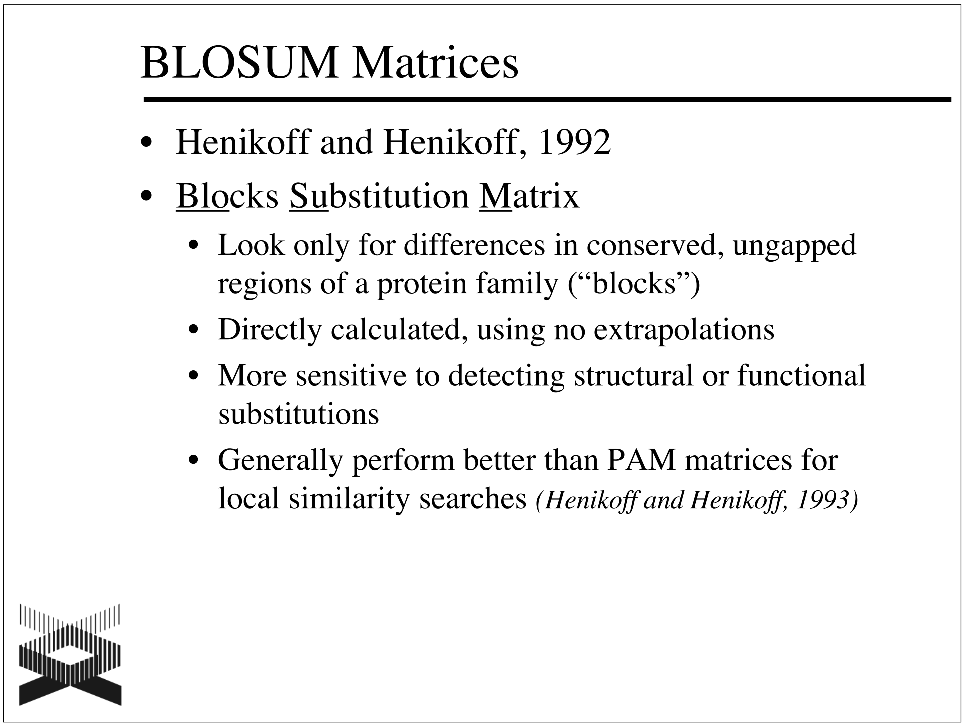
PAM Matrices: Assumptions

- 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

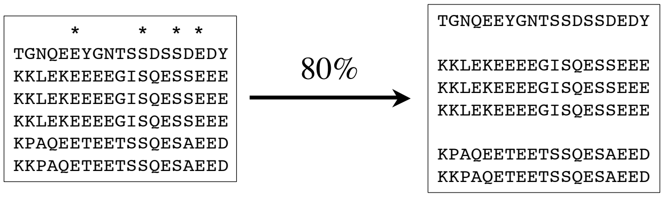


BLOSUM Matrices

- 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



```
     *     *  * *
TGNQEEYGNTSSDSSDEDY
KKLEKEEEEGISQESSEEE
KKLEKEEEEGISQESSEEE
KKLEKEEEEGISQESSEEE
KPAQEETEETSSQESAEED
KKPAQETEETSSQESAEED
```

80%

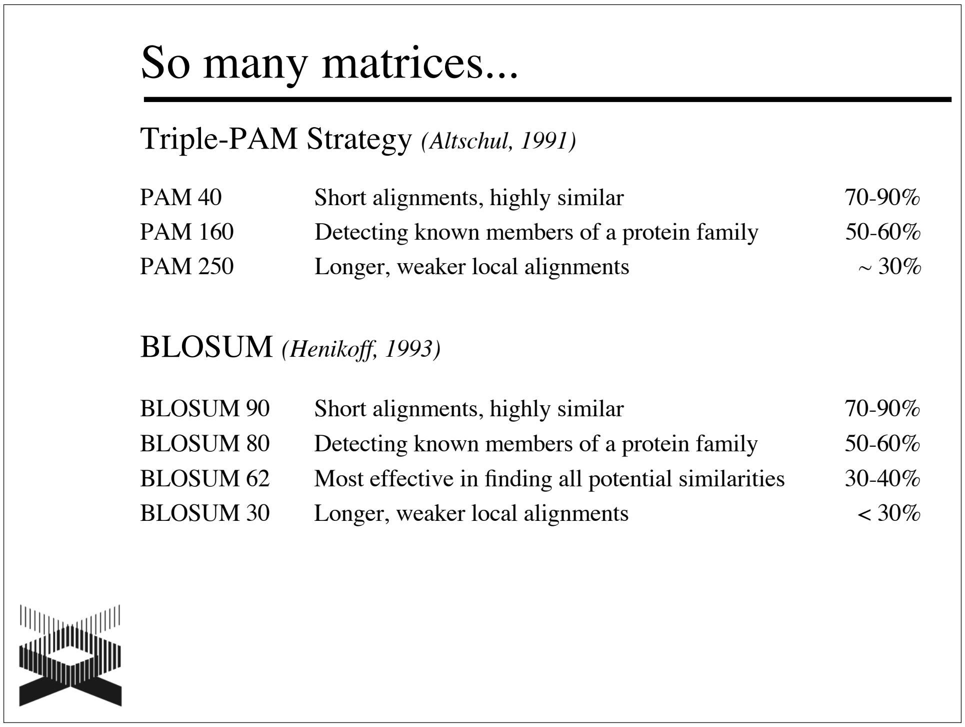
```
TGNQEEYGNTSSDSSDEDY

KKLEKEEEEGISQESSEEE
KKLEKEEEEGISQESSEEE
KKLEKEEEEGISQESSEEE

KPAQEETEETSSQESAEED
KKPAQETEETSSQESAEED
```

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



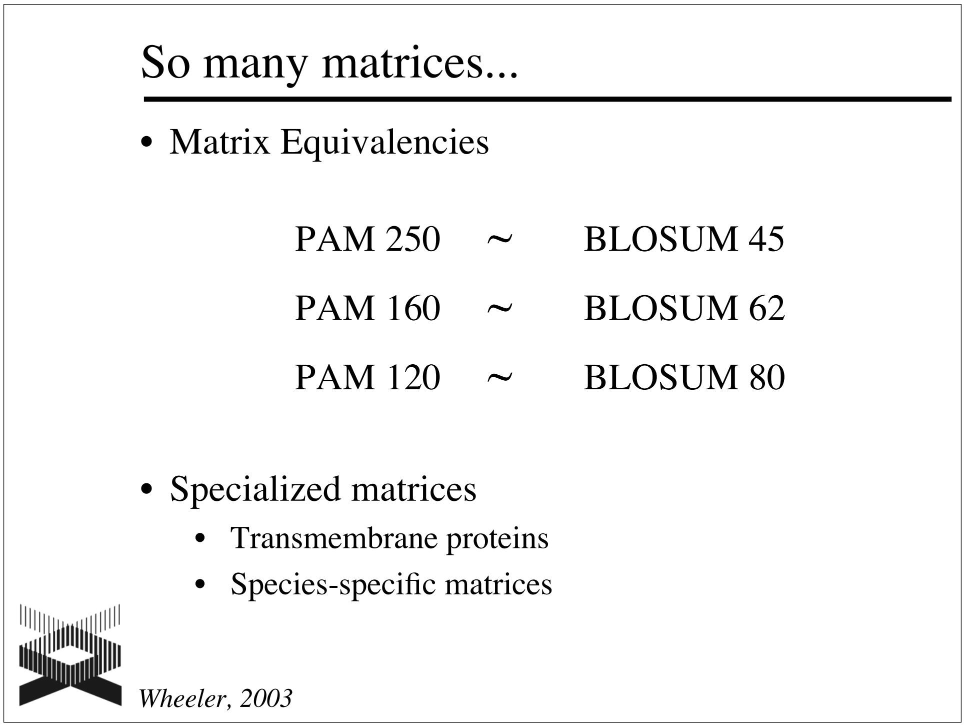
So many matrices...

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

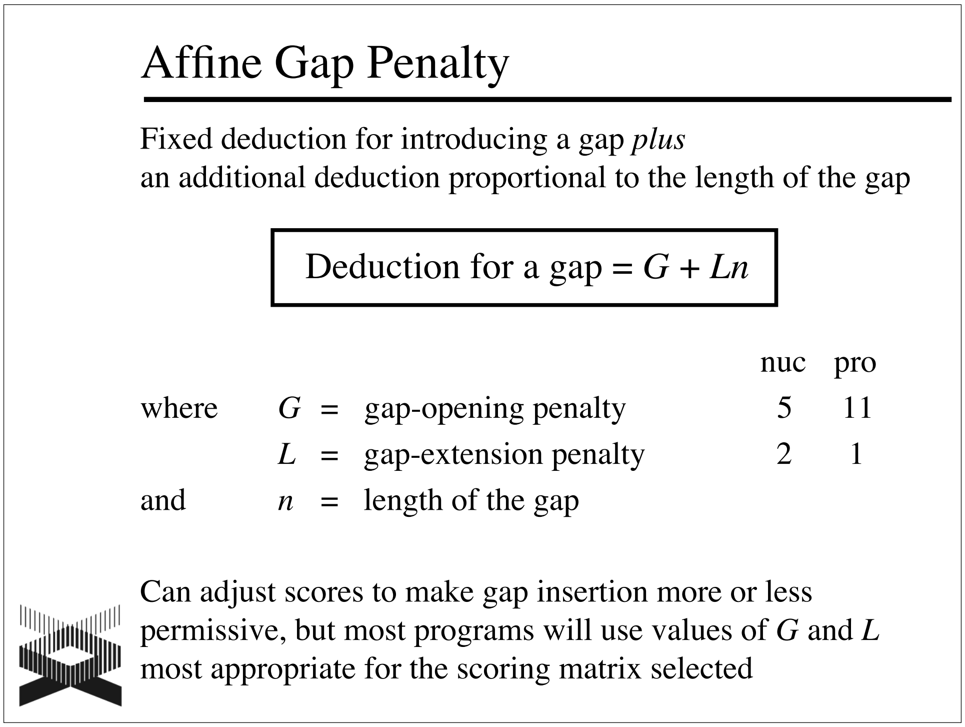
Wheeler, 2003

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”



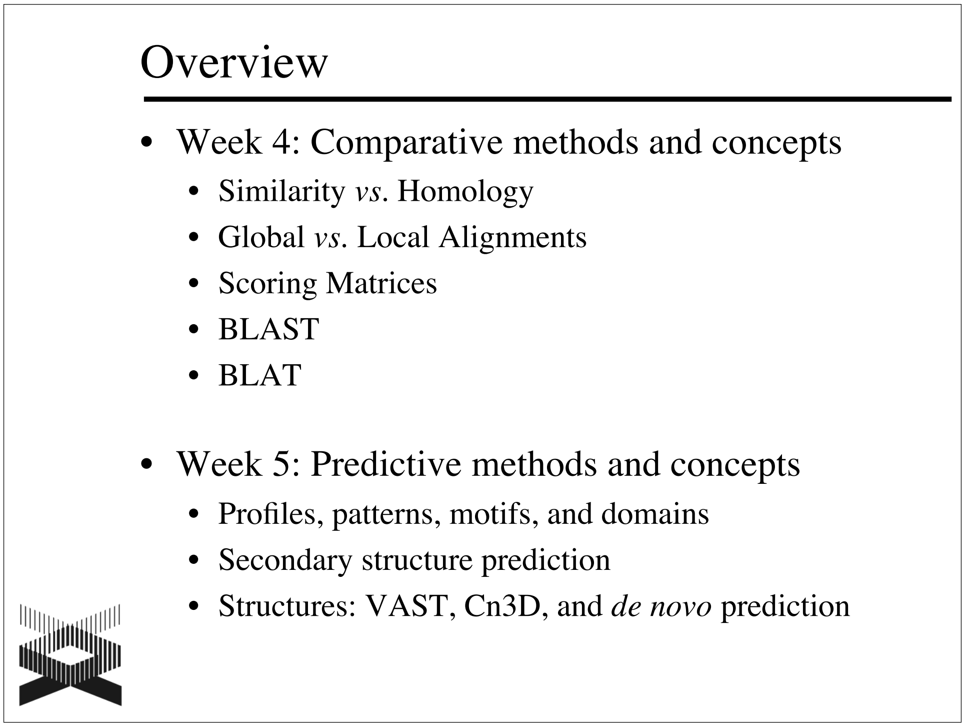
Affine Gap Penalty

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

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

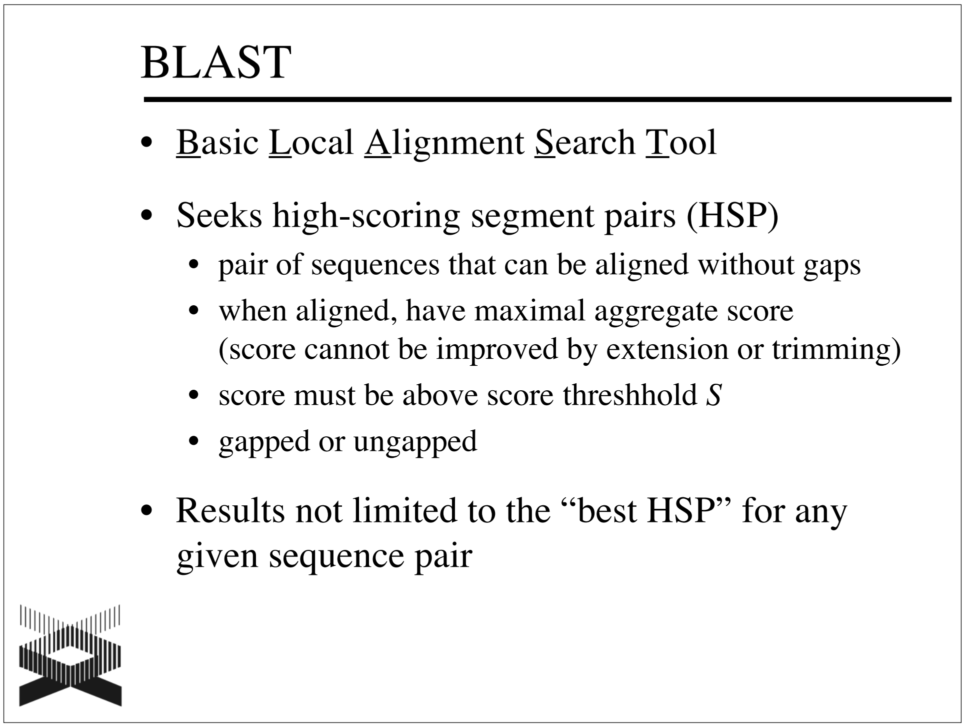
				nuc	pro
where	*G*	=	gap-opening penalty	5	11
	L	=	gap-extension penalty	2	1
and	*n*	=	length of the gap		

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



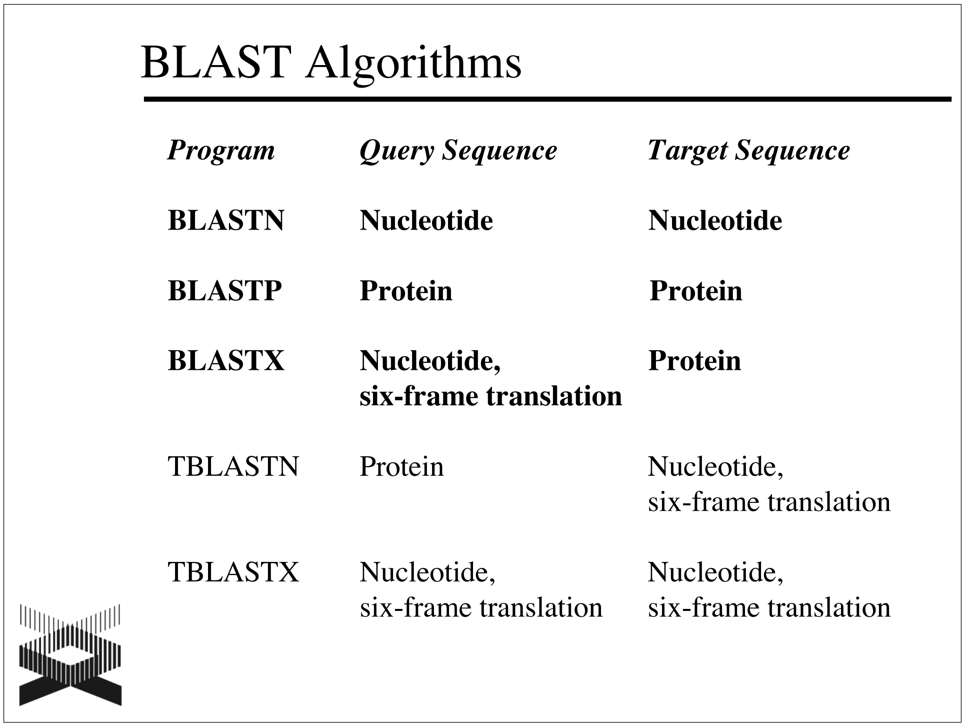
Overview

- 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

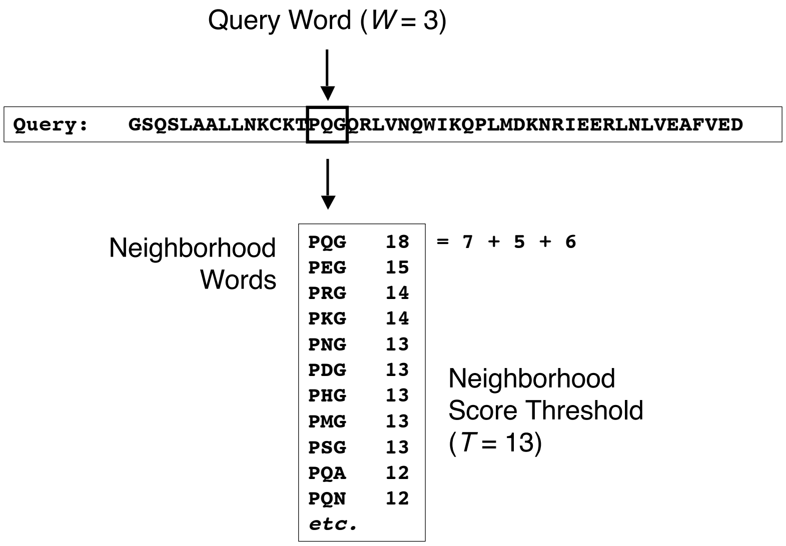
- Basic Local Alignment Search Tool
- Seeks high-scoring segment pairs (HSP)
  - pair of sequences that can be aligned without gaps
  - when aligned, have maximal aggregate score (score cannot be improved by extension or trimming)
  - score must be above score threshhold *S*
  - gapped or ungapped
- 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 ($W = 3$)

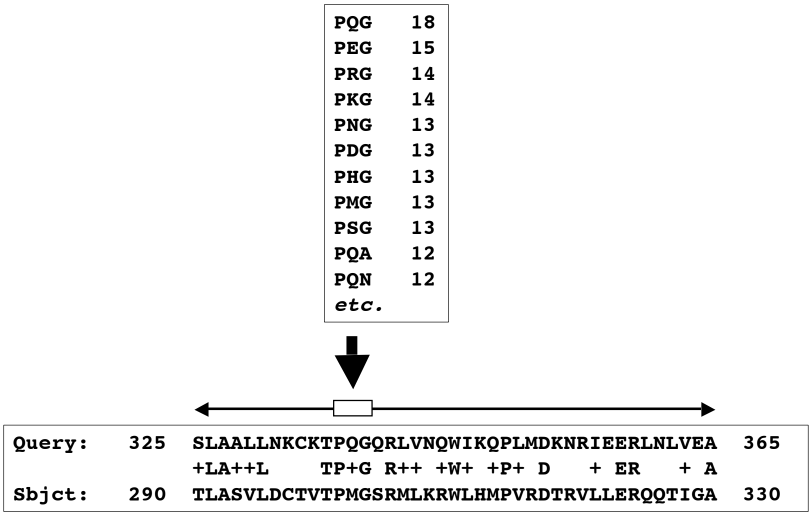
```
Query:   GSQSLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEAFVED
```

Neighborhood Words

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

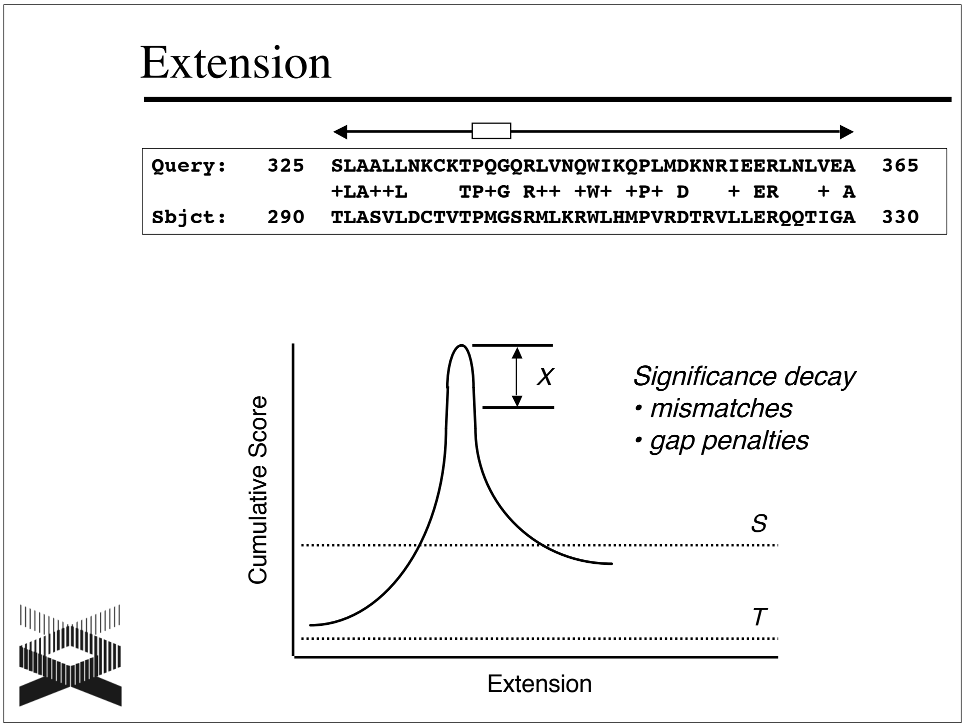
Neighborhood Score Threshold ($T = 13$)

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.	

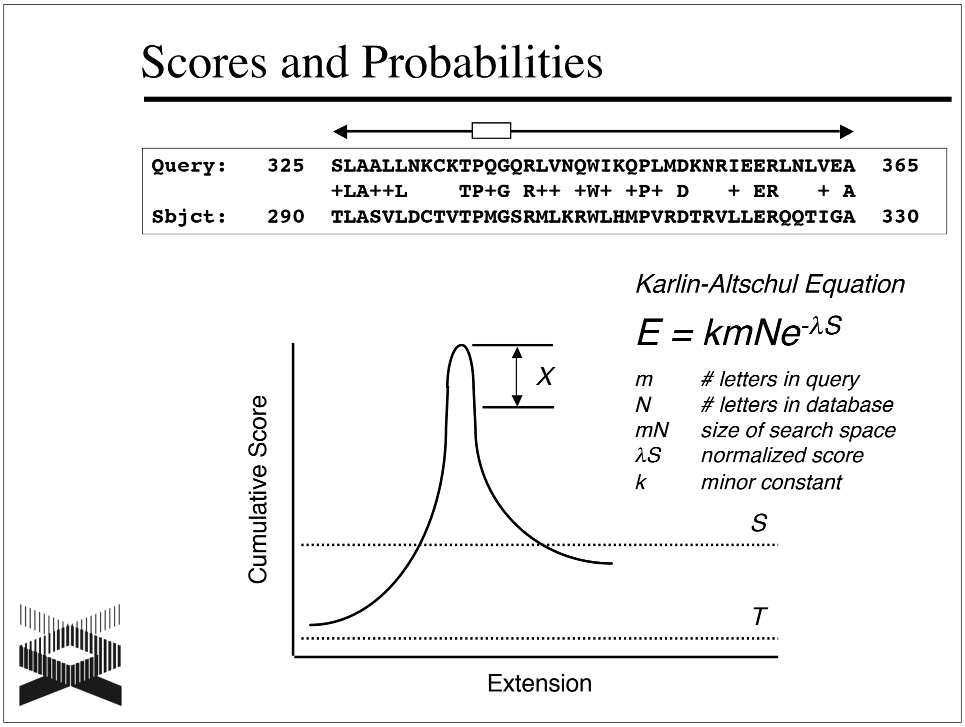
```
Query:   325  SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA  365
              +LA++L    TP+G R++ +W+ +P+ D   + ER   + A
Sbjct:   290  TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA  330
```



Extension

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

Significance decay
- *mismatches*
- *gap penalties*



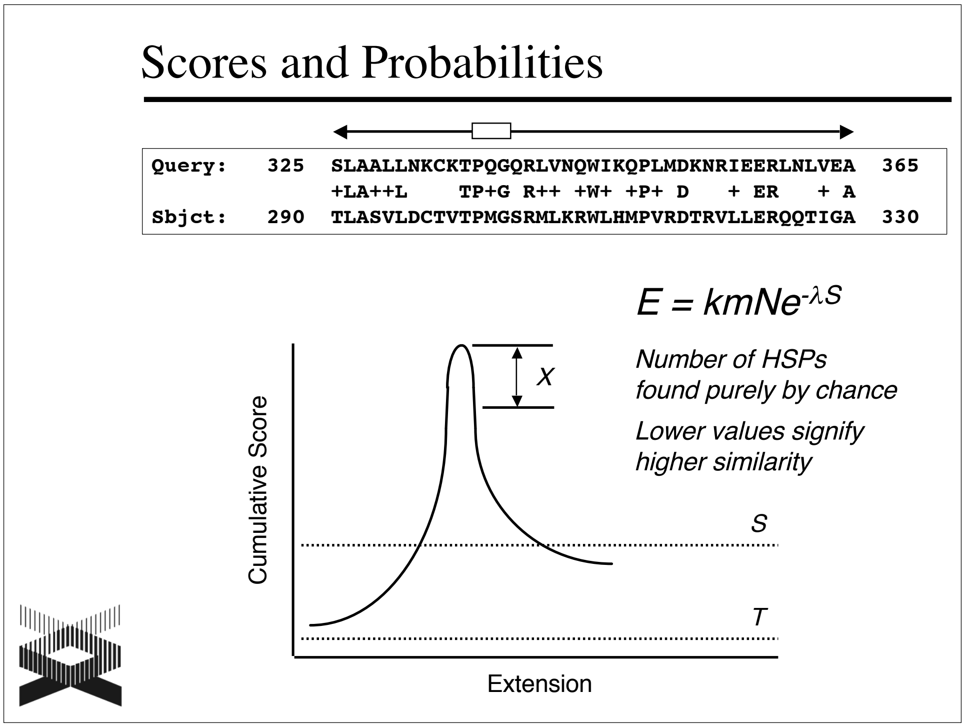
Scores and Probabilities

```
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*



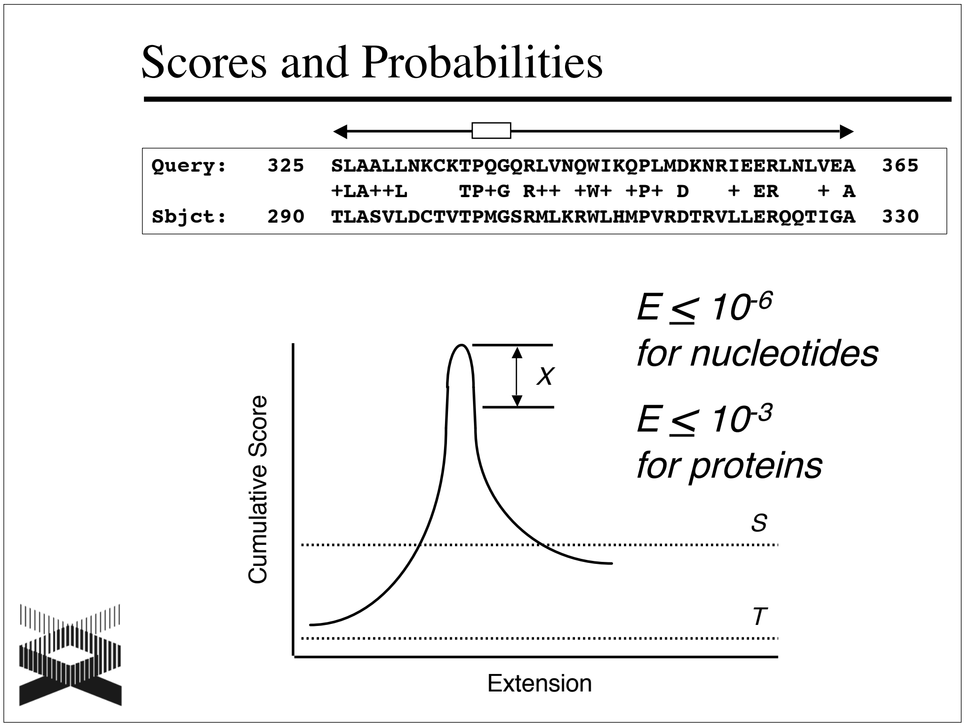
Scores and Probabilities

```
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

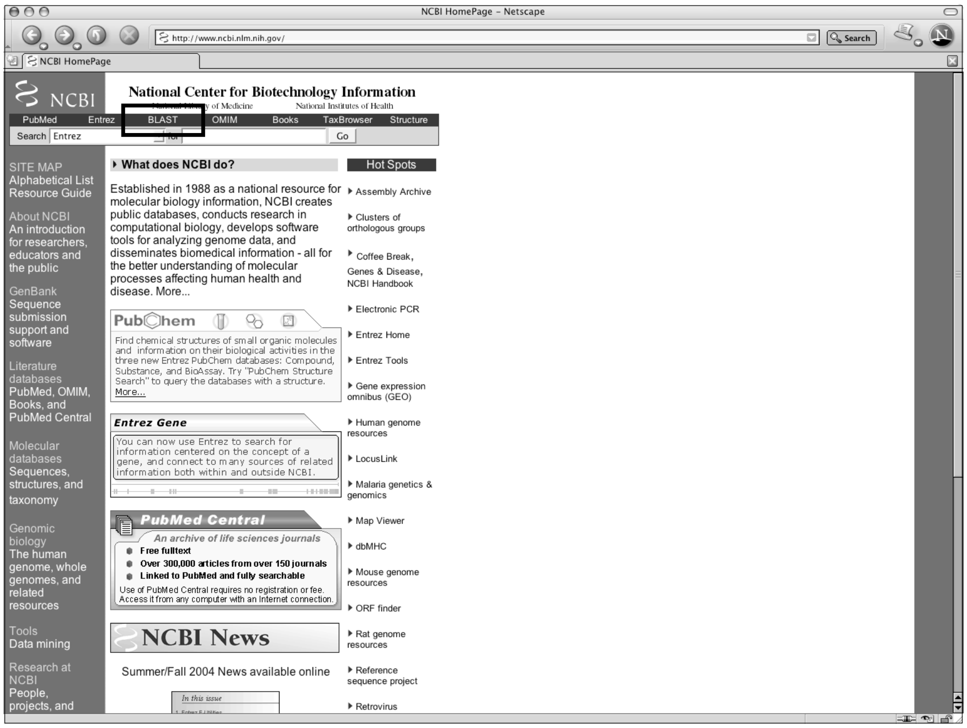


Scores and Probabilities

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

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

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



http://www.ncbi.nlm.nih.gov/

National Center for Biotechnology Information
National Library of Medicine National Institutes of Health

PubMed | Entrez | BLAST | OMIM | Books | TaxBrowser | Structure

Search Entrez for

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...

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.

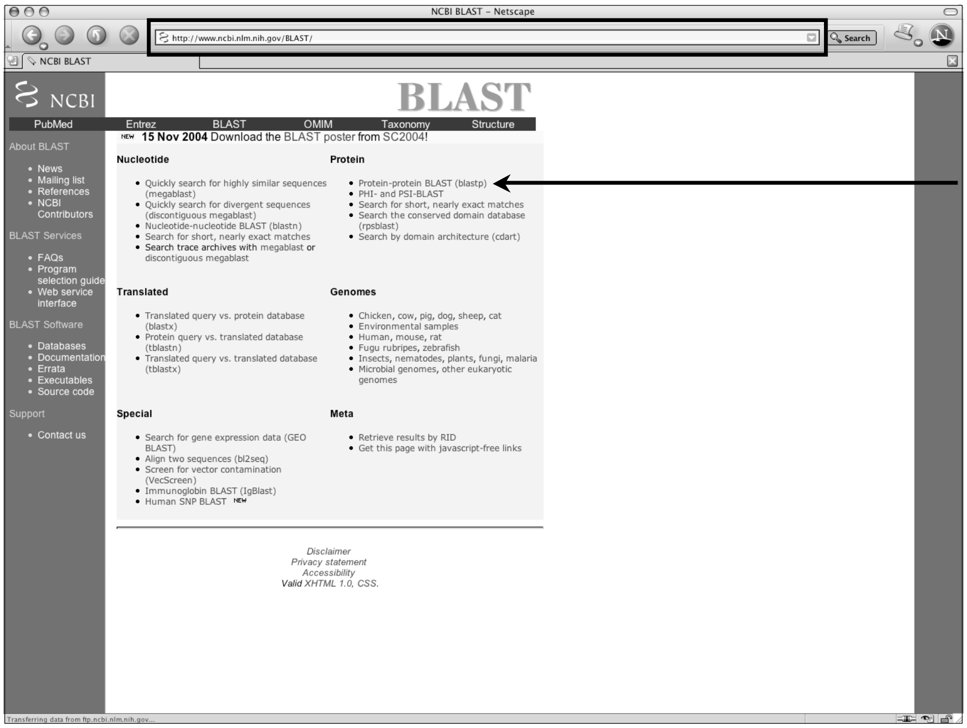
NCBI News

Summer/Fall 2004 News available online

In this issue

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



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

BLAST

PubMed | Entrez | BLAST | OMIM | Taxonomy | Structure

NEW **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

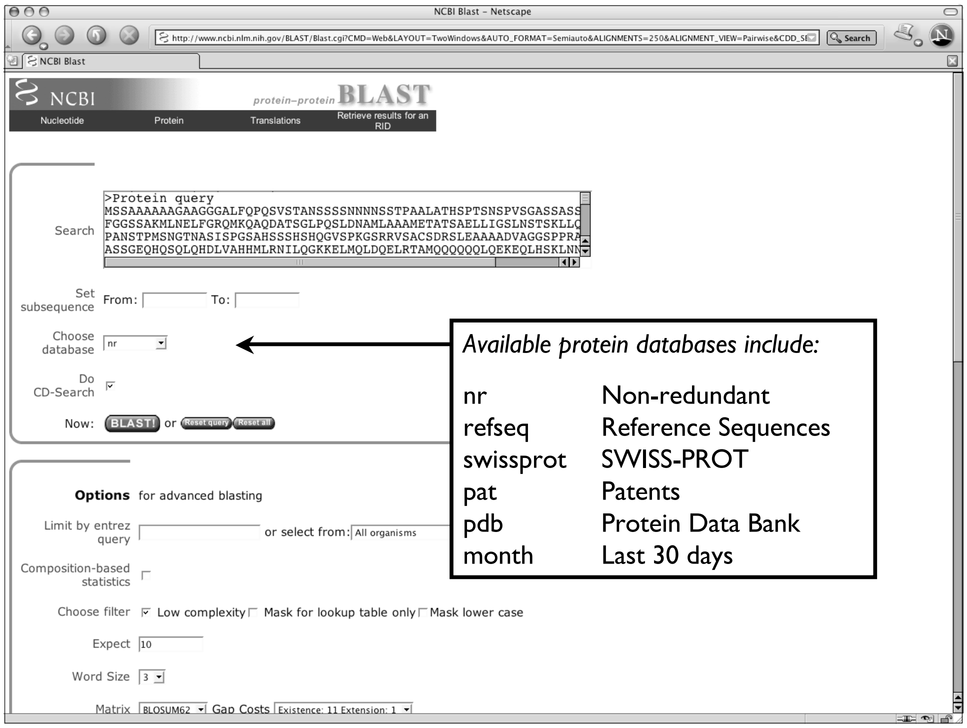
- Search for gene expression data (GEO BLAST)
- Align two sequences (bl2seq)
- Screen for vector contamination (VecScreen)
- Immunoglobin BLAST (IgBlast)
- Human SNP BLAST NEW

Meta

- Retrieve results by RID
- Get this page with javascript-free links

Disclaimer
Privacy statement
Accessibility
Valid XHTML 1.0, CSS.

Transferring data from ftp.ncbi.nlm.nih.gov...



NCBI Blast – Netscape
http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi?CMD=Web&LAYOUT=TwoWindows&AUTO_FORMAT=Semiauto&ALIGNMENTS=250&ALIGNMENT_VIEW=Pairwise&CDD_SE
Search
NCBI Blast
NCBI
protein–protein BLAST
Nucleotide
Protein
Translations
Retrieve results for an RID

Search

```
>Protein query
MSSAAAAAAGAAGGGALFQPQSVSTANSSSSNNNNSSTPAALATHSPTSNSPVSGASSASS
FGGSSAKMLNELFGRQMKQAQDATSGLPQSLDNAMLAAAMETATSAELLIGSLNSTSKLLQ
PANSTPMSNGTNASISPGSAHSSSHSHQGVSPKGSRRVSACSDRSLEAAAADVAGGSPPRA
ASSGEQHQSQLQHDLVAHHMLRNILQGKKELMQLDQELRTAMQQQQQQLQEKEQLHSKLNN
```

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

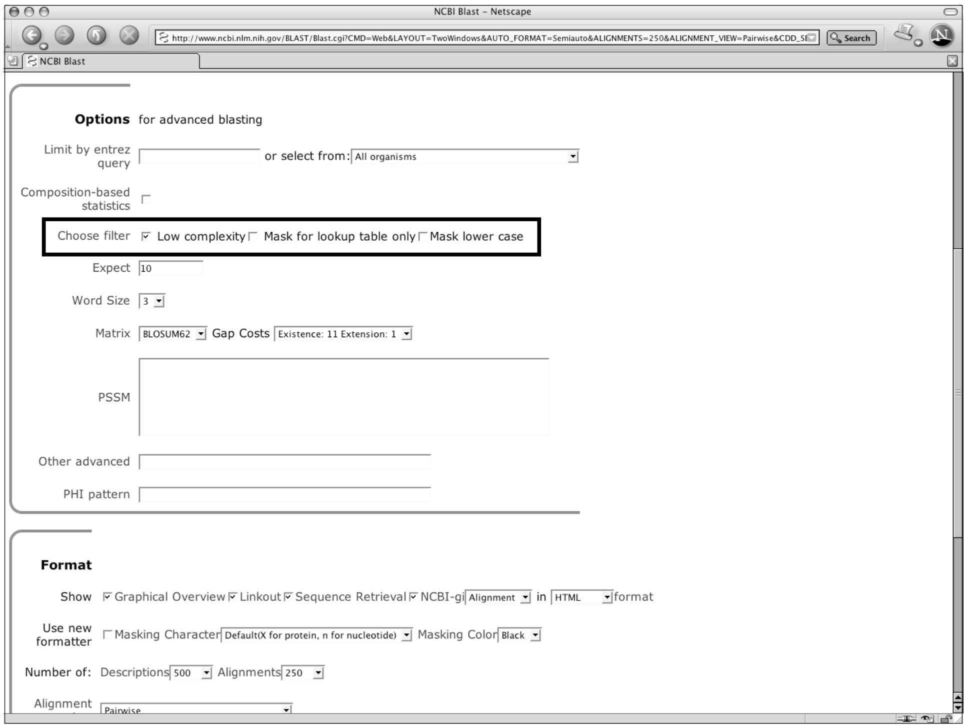
Expect 10

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
http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi?CMD=Web&LAYOUT=TwoWindows&AUTO_FORMAT=Semiauto&ALIGNMENTS=250&ALIGNMENT_VIEW=Pairwise&CDD_SE
Search
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

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

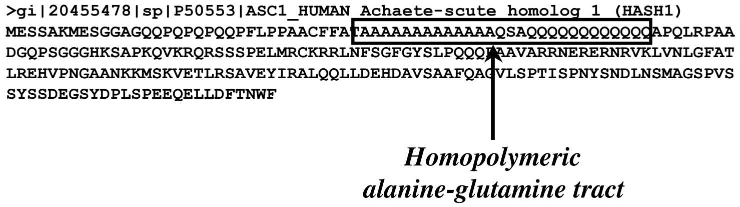
Number of: Descriptions 500 Alignments 250

Alignment Pairwise

Low-Complexity Regions

Defined as regions of biased composition

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

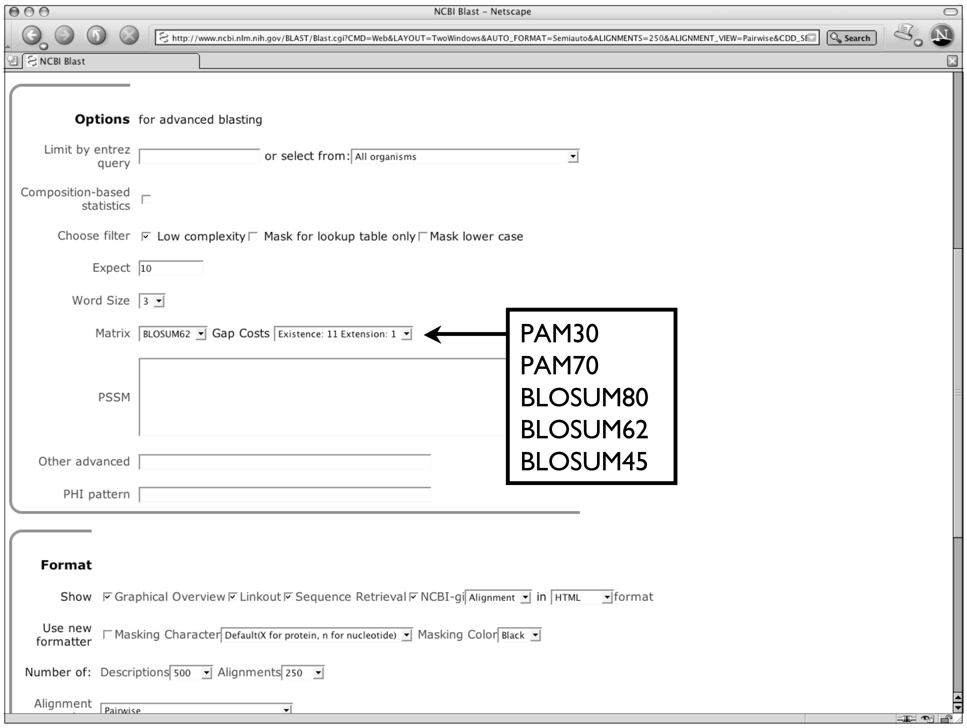


```
>gi|20455478|sp|P50553|ASC1_HUMAN Achaete-scute homolog 1 (HASH1)
MESSAKMESGGAGQQPQPQPQQPFLPPAACFFATAAAAAAAAAAAAAQSAQQQQQQQQQQQQAPQLRPAA
DGQPSGGGHKSAPKQVKRQRSSSPELMRCKRRLNFSGFGYSLPQQQPAAVARRNERERNRVKLVNLGFAT
LREHVPNGAANKKMSKVETLRSAVEYIRALQQLLDEHDAVSAAFQAGVLSPTISPNYSNDLNSMAGSPVS
SYSSDEGSYDPLSPEEQELLDFTNWF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

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
http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi?CMD=Web&LAYOUT=TwoWindows&AUTO_FORMAT=Semiauto&ALIGNMENTS=250&ALIGNMENT_VIEW=Pairwise&CDD_SE
Search
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

Format

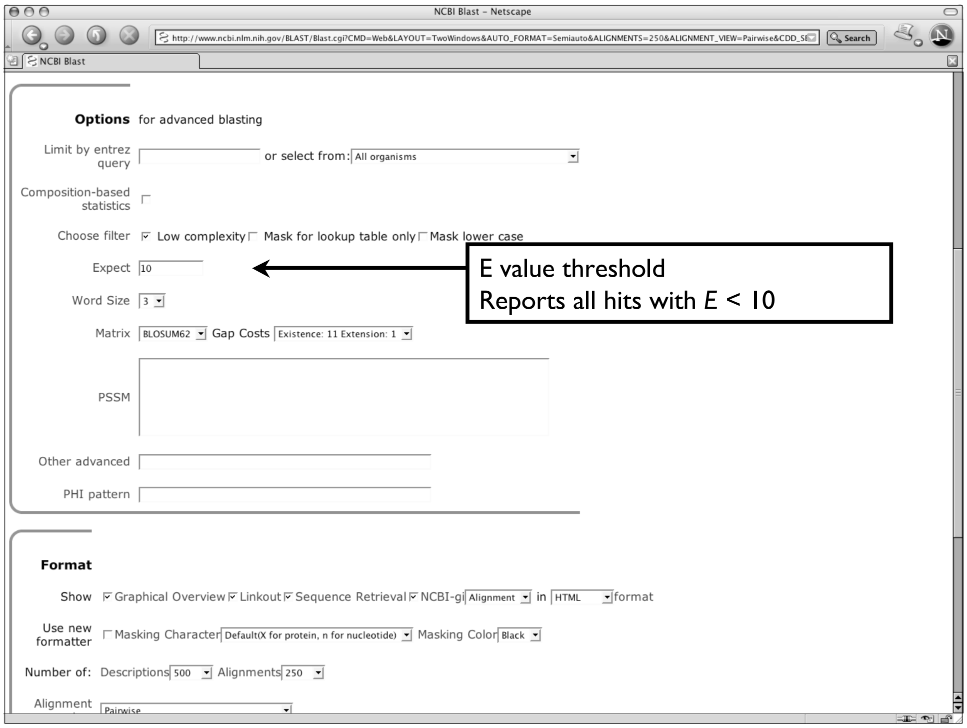
Show: ☑ Graphical Overview ☑ Linkout ☑ Sequence Retrieval ☑ NCBI-gi Alignment in HTML format

Use new formatter: □ Masking Character Default(X for protein, n for nucleotide) Masking Color Black

Number of: Descriptions 500 Alignments 250

Alignment: Pairwise

PAM30
PAM70
BLOSUM80
BLOSUM62
BLOSUM45


NCBI Blast – Netscape
http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi?CMD=Web&LAYOUT=TwoWindows&AUTO_FORMAT=Semiauto&ALIGNMENTS=250&ALIGNMENT_VIEW=Pairwise&CDD_SE
Search
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

Format

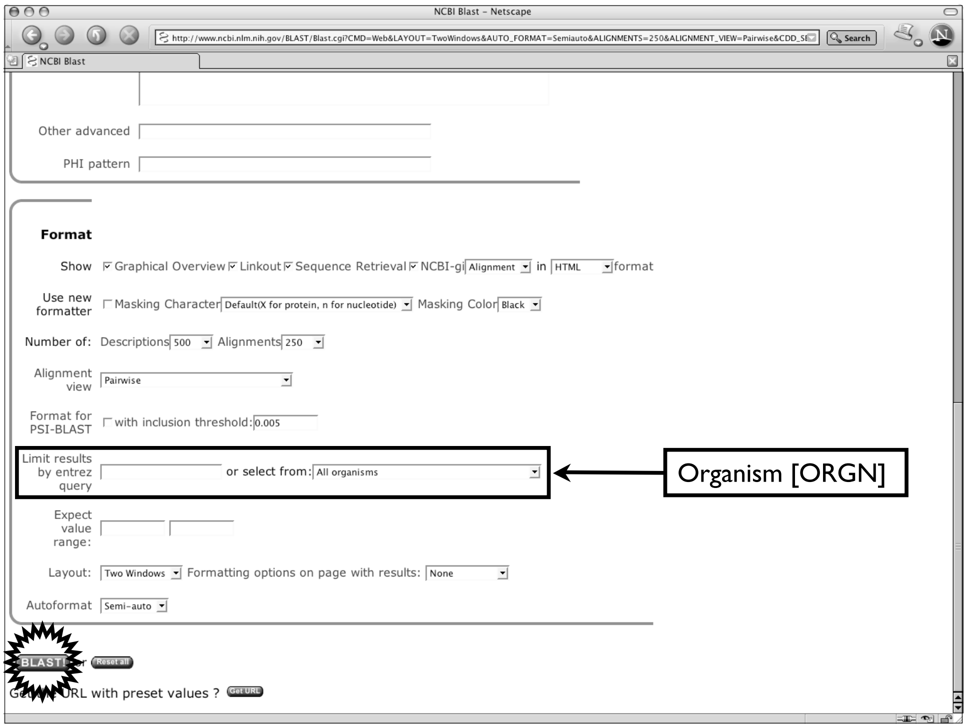
Show: ☑ Graphical Overview ☑ Linkout ☑ Sequence Retrieval ☑ NCBI-gi Alignment in HTML format

Use new formatter: □ Masking Character Default(X for protein, n for nucleotide) Masking Color Black

Number of: Descriptions 500 Alignments 250

Alignment: Pairwise

E value threshold
Reports all hits with $E < 10$



http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi?CMD=Web&LAYOUT=TwoWindows&AUTO_FORMAT=Semiauto&ALIGNMENTS=250&ALIGNMENT_VIEW=Pairwise&CDD_SE

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]

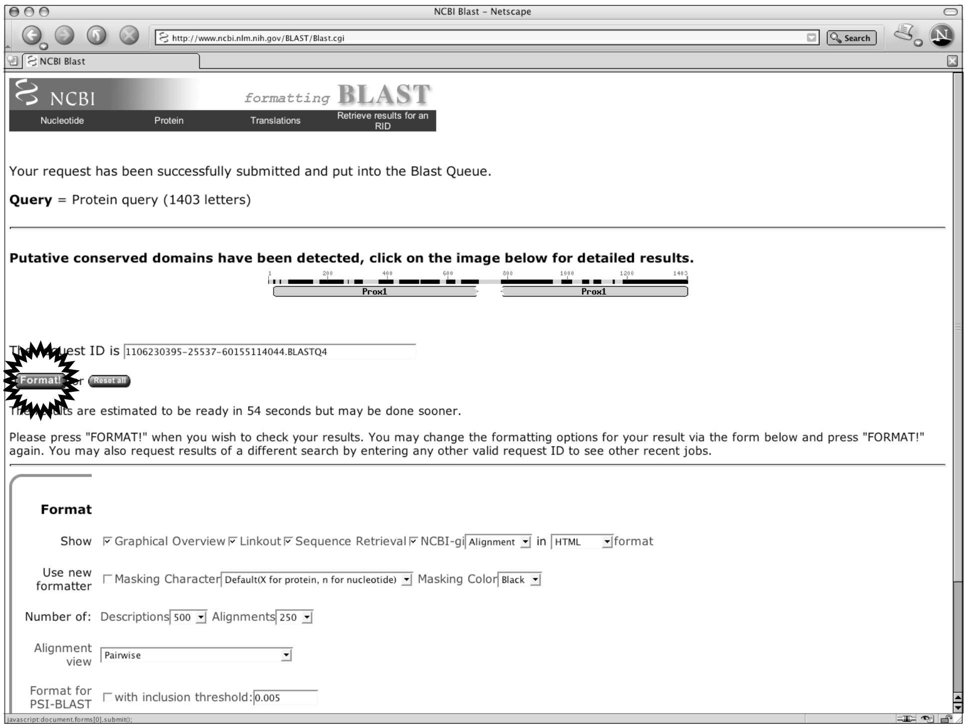
Expect value range:

Layout: Two Windows Formatting options on page with results: None

Autoformat Semi-auto

BLAST! or Reset all

Get a URL with preset values ? Get URL



http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi

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.

The request ID is 1106230395-25537-60155114044.BLASTQ4

Format! or Reset all

The results are estimated to be ready in 54 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

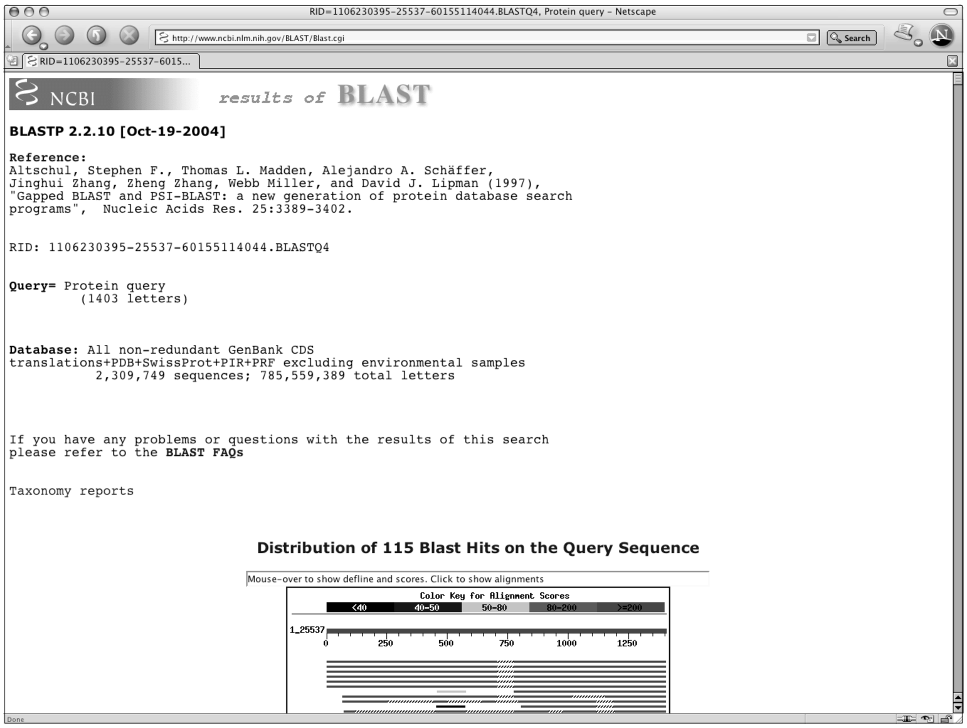
Show ☑ Graphical Overview ☑ Linkout ☑ Sequence Retrieval ☑ NCBI-gi Alignment in HTML format

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

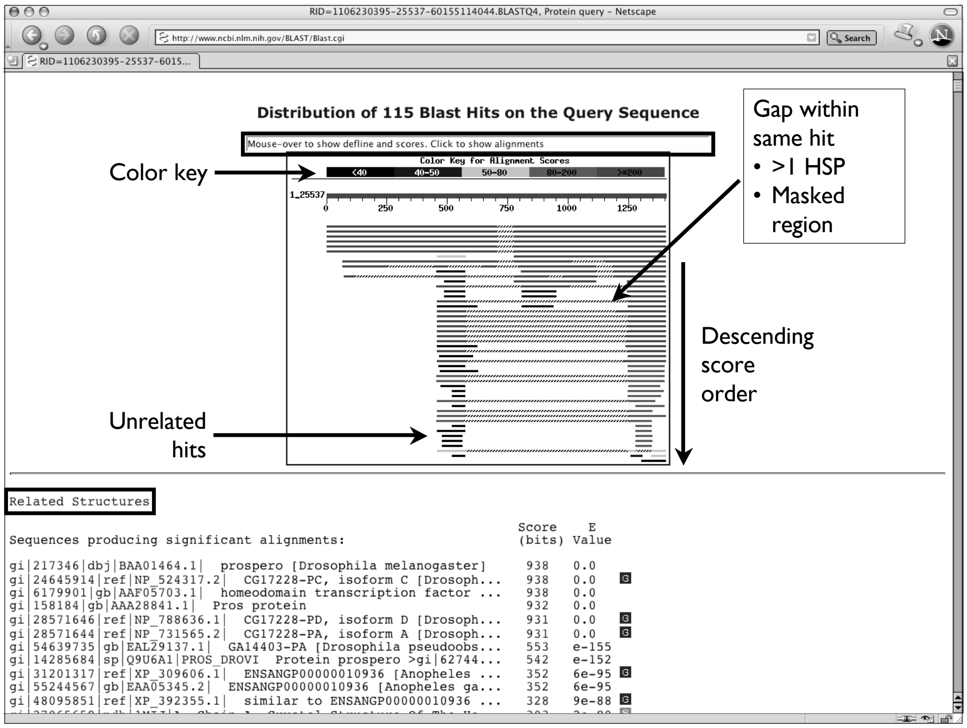

RID=1106230395-25537-60155114044.BLASTQ4, Protein query – Netscape
http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi
Search
RID=1106230395-25537-6015...
NCBI
results of BLAST
Color Key for Alignment Scores
<40
40-50
50-80
80-200
>=200
1_25537
0
250
500
750
1000
1250
Done

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


RID=1106230395-25537-60155114044.BLASTQ4, Protein query – Netscape
http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi
Search
RID=1106230395-25537-6015...
Color Key for Alignment Scores
<40
40-50
50-80
80-200
>=200
1_25537
0
250
500
750
1000
1250

Distribution of 115 Blast Hits on the Query Sequence

Mouse-over to show defline and scores. Click to show alignments

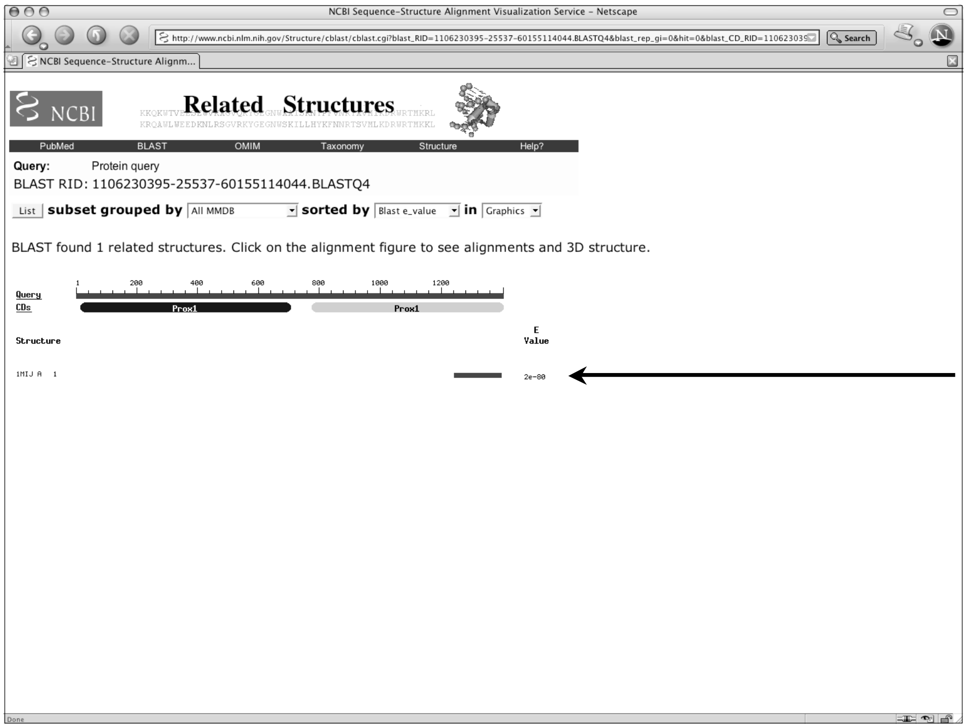
Color key

Gap within same hit
- >1 HSP
- Masked region

Descending score order

Unrelated hits

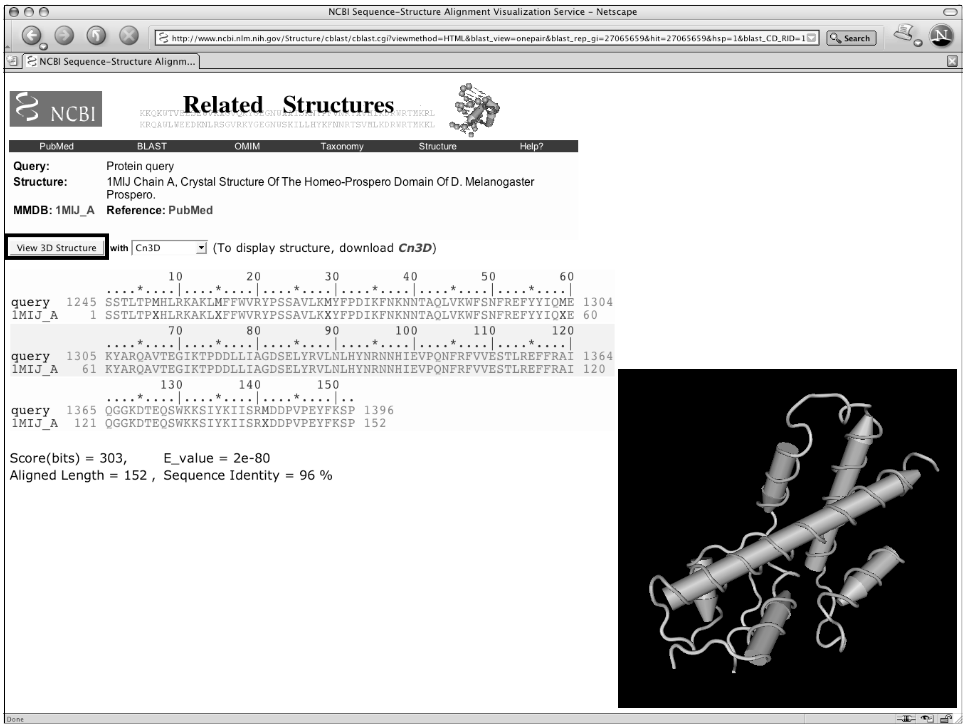
```
Related Structures

                                                                 Score    E
Sequences producing significant alignments:                      (bits) 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   G
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   G
gi|28571644|ref|NP_731565.2|  CG17228-PA, isoform A [Drosoph...   931   0.0   G
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 G
gi|55244567|gb|EAA05345.2|  ENSANGP00000010936 [Anopheles ga...   352   6e-95
gi|48095851|ref|XP_392355.1|  similar to ENSANGP00000010936 ...   328   9e-88 G
```


NCBI Sequence-Structure Alignment Visualization Service - Netscape
http://www.ncbi.nlm.nih.gov/Structure/cblast/cblast.cgi?blast_RID=1106230395-25537-60155114044.BLASTQ4&blast_rep_gi=0&hit=0&blast_CD_RID=11062303
Search
NCBI Sequence-Structure Alignm...
NCBI
Related Structures
PubMed
BLAST
OMIM
Taxonomy
Structure
Help?
List
All MMDB
Blast e_value
Graphics
1
200
400
600
800
1000
1200
Query
CDs
Prox1
Prox1
Structure
E Value
1MIJ A 1
2e-80
Done

Query: Protein query
BLAST RID: 1106230395-25537-60155114044.BLASTQ4

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.

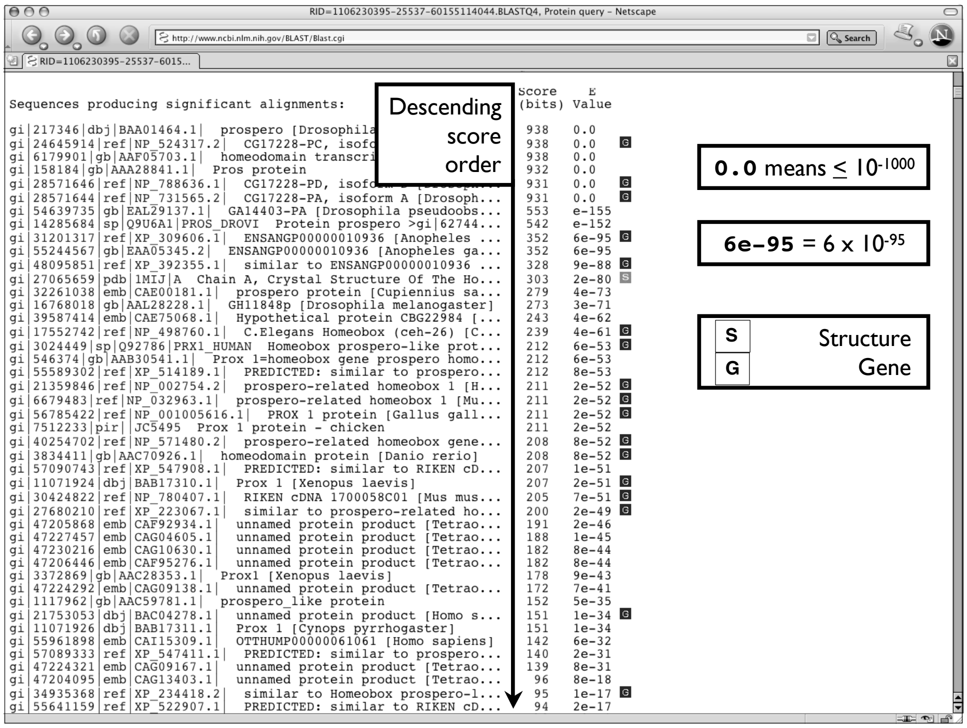

NCBI Sequence-Structure Alignment Visualization Service - Netscape
http://www.ncbi.nlm.nih.gov/Structure/cblast/cblast.cgi?viewmethod=HTML&blast_view=onepair&blast_rep_gi=27065659&hit=27065659&hsp=1&blast_CD_RID=1
Search
NCBI Sequence-Structure Alignm...
NCBI
Related Structures
PubMed
BLAST
OMIM
Taxonomy
Structure
Help?
Done

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***)

```
                    10        20        30        40        50        60
               ....*....|....*....|....*....|....*....|....*....|....*....|
query   1245  SSTLTPMHLRKAKLMFFWVRYPSSAVLKMYFPDIKFNKNNTAQLVKWFSNFREFYYIQME 1304
1MIJ_A     1  SSTLTPXHLRKAKLXFFWVRYPSSAVLKXYFPDIKFNKNNTAQLVKWFSNFREFYYIQXE 60
                    70        80        90       100       110       120
               ....*....|....*....|....*....|....*....|....*....|....*....|
query   1305  KYARQAVTEGIKTPDDLLIAGDSELYRVLNLHYNRNNHIEVPQNFRFVVESTLREFFRAI 1364
1MIJ_A    61  KYARQAVTEGIKTPDDLLIAGDSELYRVLNLHYNRNNHIEVPQNFRFVVESTLREFFRAI 120
                   130       140       150
               ....*....|....*....|....*....|..
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
http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi
Search
RID=1106230395-25537-6015...

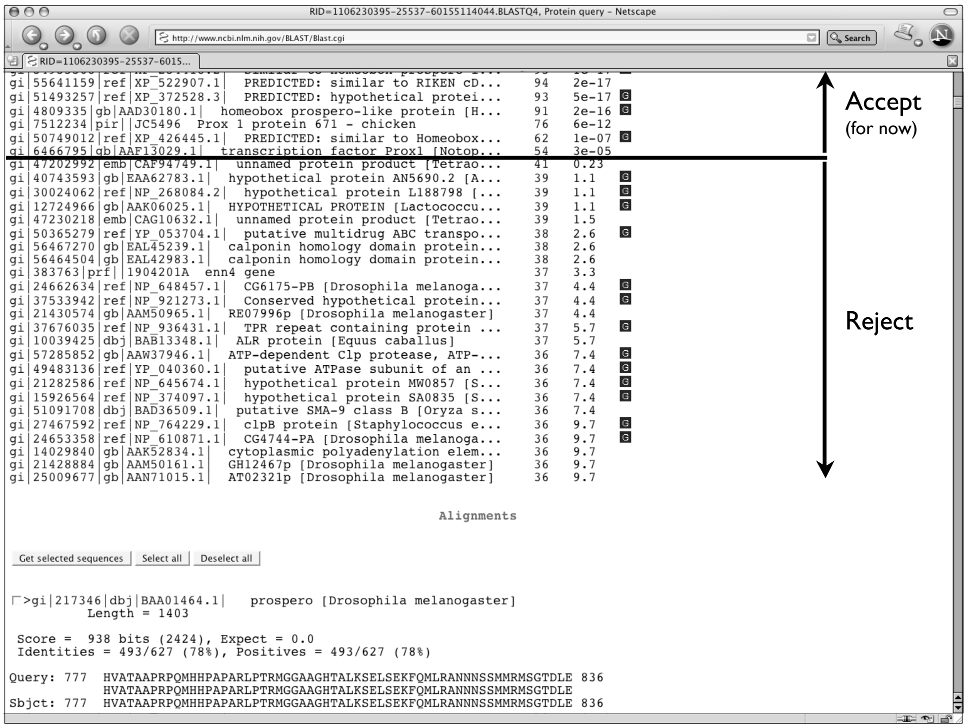
```
                                                                 Score    E
Sequences producing significant alignments:                      (bits) Value

gi|217346|dbj|BAA01464.1|  prospero [Drosophila                   938   0.0
gi|24645914|ref|NP_524317.2|  CG17228-PC, isofo                   938   0.0    G
gi|6179901|gb|AAF05703.1|  homeodomain transcri                   938   0.0
gi|158184|gb|AAA28841.1|  Pros protein                            932   0.0
gi|28571646|ref|NP_788636.1|  CG17228-PD, isofo                   931   0.0    G
gi|28571644|ref|NP_731565.2|  CG17228-PA, isoform A [Drosoph...   931   0.0    G
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  G
gi|55244567|gb|EAA05345.2|  ENSANGP00000010936 [Anopheles ga...   352   6e-95
gi|48095851|ref|XP_392355.1|  similar to ENSANGP00000010936 ...   328   9e-88  G
gi|27065659|pdb|1MIJ|A  Chain A, Crystal Structure Of The Ho...   303   2e-80  S
gi|32261038|emb|CAE00181.1|  prospero protein [Cupiennius sa...   279   4e-73
gi|16768018|gb|AAL28228.1|  GH11848p [Drosophila melanogaster]    273   3e-71
gi|39587414|emb|CAE75068.1|  Hypothetical protein CBG22984 [...   243   4e-62
gi|17552742|ref|NP_498760.1|  C.Elegans Homeobox (ceh-26) [C...   239   4e-61  G
gi|3024449|sp|Q92786|PRX1_HUMAN  Homeobox prospero-like prot...   212   6e-53  G
gi|546374|gb|AAB30541.1|  Prox 1=homeobox gene prospero homo...   212   6e-53
gi|55589302|ref|XP_514189.1|  PREDICTED: similar to prospero...   212   8e-53
gi|21359846|ref|NP_002754.2|  prospero-related homeobox 1 [H...   211   2e-52  G
gi|6679483|ref|NP_032963.1|  prospero-related homeobox 1 [Mu...   211   2e-52  G
gi|56785422|ref|NP_001005616.1|  PROX 1 protein [Gallus gall...   211   2e-52  G
gi|7512233|pir||JC5495  Prox 1 protein - chicken                  211   2e-52
gi|40254702|ref|NP_571480.2|  prospero-related homeobox gene...   208   8e-52  G
gi|3834411|gb|AAC70926.1|  homeodomain protein [Danio rerio]      208   8e-52  G
gi|57090743|ref|XP_547908.1|  PREDICTED: similar to RIKEN cD...   207   1e-51
gi|11071924|dbj|BAB17310.1|  Prox 1 [Xenopus laevis]              207   2e-51  G
gi|30424822|ref|NP_780407.1|  RIKEN cDNA 1700058C01 [Mus mus...   205   7e-51  G
gi|27680210|ref|XP_223067.1|  similar to prospero-related ho...   200   2e-49  G
gi|47205868|emb|CAF92934.1|  unnamed protein product [Tetrao...   191   2e-46
gi|47227457|emb|CAG04605.1|  unnamed protein product [Tetrao...   188   1e-45
gi|47230216|emb|CAG10630.1|  unnamed protein product [Tetrao...   182   8e-44
gi|47206446|emb|CAF95276.1|  unnamed protein product [Tetrao...   182   8e-44
gi|3372869|gb|AAC28353.1|  Prox1 [Xenopus laevis]                 178   9e-43
gi|47224292|emb|CAG09138.1|  unnamed protein product [Tetrao...   172   7e-41
gi|1117962|gb|AAC59781.1|  prospero_like protein                  152   5e-35
gi|21753053|dbj|BAC04278.1|  unnamed protein product [Homo s...   151   1e-34  G
gi|11071926|dbj|BAB17311.1|  Prox 1 [Cynops pyrrhogaster]         151   1e-34
gi|55961898|emb|CAI15309.1|  OTTHUMP00000061061 [Homo sapiens]    142   6e-32
gi|57089333|ref|XP_547411.1|  PREDICTED: similar to prospero...   140   2e-31
gi|47224321|emb|CAG09167.1|  unnamed protein product [Tetrao...   139   8e-31
gi|47204095|emb|CAG13403.1|  unnamed protein product [Tetrao...    96   8e-18
gi|34935368|ref|XP_234418.2|  similar to Homeobox prospero-l...    95   1e-17  G
gi|55641159|ref|XP_522907.1|  PREDICTED: similar to RIKEN cD...    94   2e-17
```

Descending score order

0.0 means ≤ 10^{-1000}

6e-95 = 6 x 10^{-95}

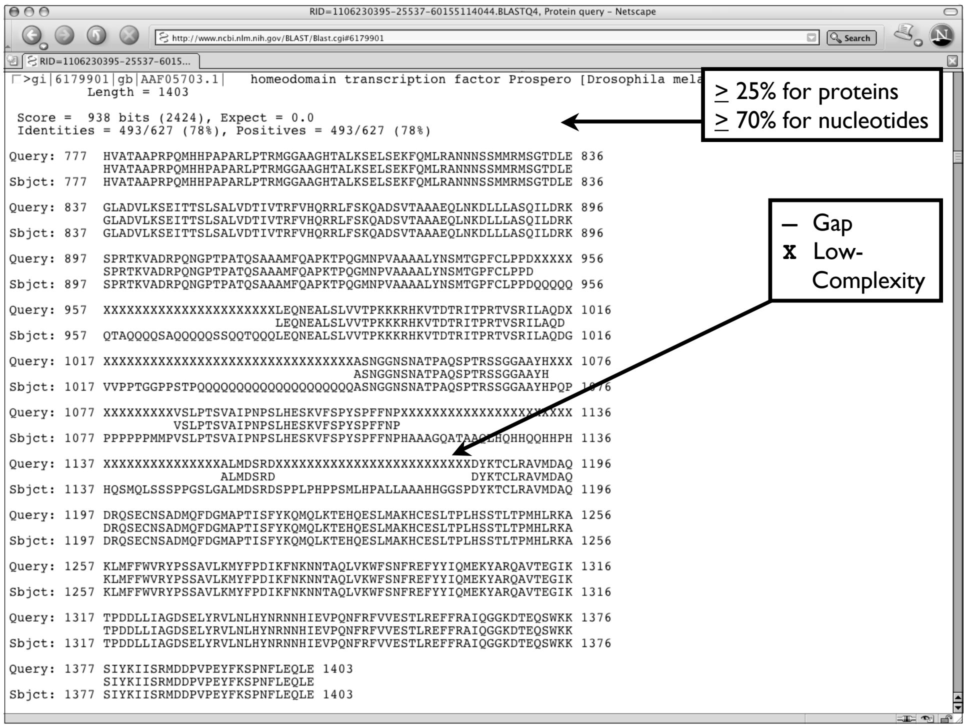
S	Structure
G	Gene


RID=1106230395-25537-60155114044.BLASTQ4, Protein query – Netscape
http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi
Search
RID=1106230395-25537-6015...
Get selected sequences
Select all
Deselect all

```
gi|55641159|ref|XP_522907.1|  PREDICTED: similar to RIKEN cD...    94   2e-17
gi|51493257|ref|XP_372528.3|  PREDICTED: hypothetical protei...    93   5e-17  G
gi|4809335|gb|AAD30180.1|  homeobox prospero-like protein [H...    91   2e-16  G
gi|7512234|pir||JC5496  Prox 1 protein 671 - chicken               76   6e-12
gi|50749012|ref|XP_426445.1|  PREDICTED: similar to Homeobox...    62   1e-07  G
gi|6466795|gb|AAF13029.1|  transcription factor Prox1 [Notop...    54   3e-05
gi|47202992|emb|CAF94749.1|  unnamed protein product [Tetrao...    41   0.23
gi|40743593|gb|EAA62783.1|  hypothetical protein AN5690.2 [A...    39   1.1    G
gi|30024062|ref|NP_268084.2|  hypothetical protein L188798 [...    39   1.1    G
gi|12724966|gb|AAK06025.1|  HYPOTHETICAL PROTEIN [Lactococcu...    39   1.1    G
gi|47230218|emb|CAG10632.1|  unnamed protein product [Tetrao...    39   1.5
gi|50365279|ref|YP_053704.1|  putative multidrug ABC transpo...    38   2.6    G
gi|56467270|gb|EAL45239.1|  calponin homology domain protein...    38   2.6
gi|56464504|gb|EAL42983.1|  calponin homology domain protein...    38   2.6
gi|383763|prf||1904201A  enn4 gene                                 37   3.3
gi|24662634|ref|NP_648457.1|  CG6175-PB [Drosophila melanoga...    37   4.4    G
gi|37533942|ref|NP_921273.1|  Conserved hypothetical protein...    37   4.4    G
gi|21430574|gb|AAM50965.1|  RE07996p [Drosophila melanogaster]     37   4.4
gi|37676035|ref|NP_936431.1|  TPR repeat containing protein ...    37   5.7    G
gi|10039425|dbj|BAB13348.1|  ALR protein [Equus caballus]          37   5.7
gi|57285852|gb|AAW37946.1|  ATP-dependent Clp protease, ATP-...    36   7.4    G
gi|49483136|ref|YP_040360.1|  putative ATPase subunit of an ...    36   7.4    G
gi|21282586|ref|NP_645674.1|  hypothetical protein MW0857 [S...    36   7.4    G
gi|15926564|ref|NP_374097.1|  hypothetical protein SA0835 [S...    36   7.4    G
gi|51091708|dbj|BAD36509.1|  putative SMA-9 class B [Oryza s...    36   7.4
gi|27467592|ref|NP_764229.1|  clpB protein [Staphylococcus e...    36   9.7    G
gi|24653358|ref|NP_610871.1|  CG4744-PA [Drosophila melanoga...    36   9.7    G
gi|14029840|gb|AAK52834.1|  cytoplasmic polyadenylation elem...    36   9.7
gi|21428884|gb|AAM50161.1|  GH12467p [Drosophila melanogaster]     36   9.7
gi|25009677|gb|AAN71015.1|  AT02321p [Drosophila melanogaster]     36   9.7


                                   Alignments


□>gi|217346|dbj|BAA01464.1|   prospero [Drosophila melanogaster]
          Length = 1403

 Score =  938 bits (2424), Expect = 0.0
 Identities = 493/627 (78%), Positives = 493/627 (78%)

Query: 777  HVATAAPRPQMHHPAPARLPTRMGGAAGHTALKSELSEKFQMLRANNNSSMMRMSGTDLE 836
            HVATAAPRPQMHHPAPARLPTRMGGAAGHTALKSELSEKFQMLRANNNSSMMRMSGTDLE
Sbjct: 777  HVATAAPRPQMHHPAPARLPTRMGGAAGHTALKSELSEKFQMLRANNNSSMMRMSGTDLE 836
```

Accept
(for now)

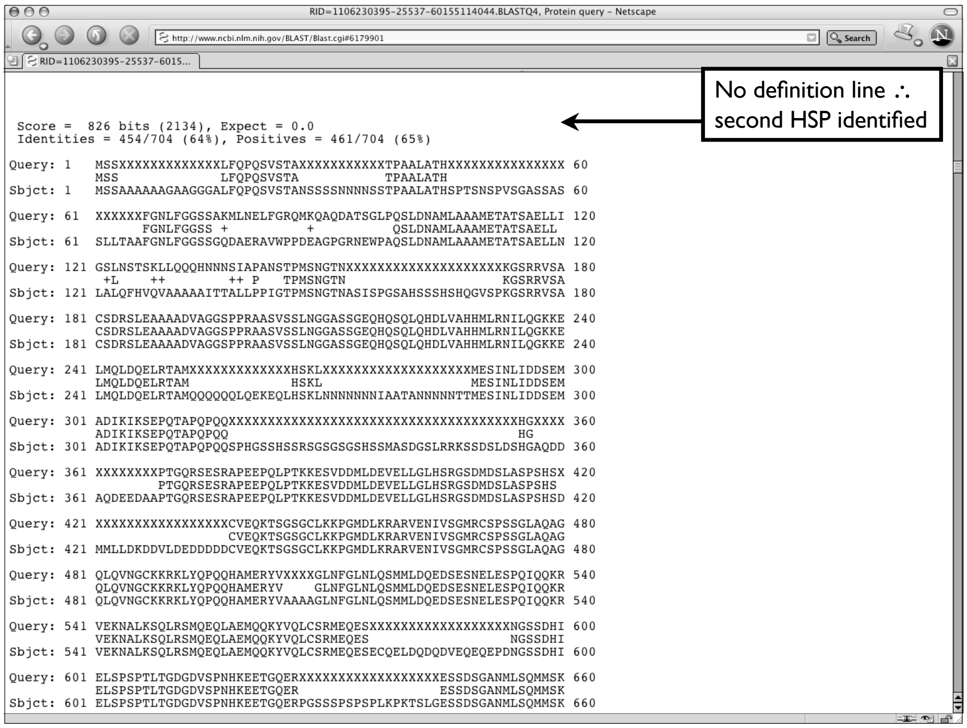
Reject


RID=1106230395-25537-60155114044.BLASTQ4, Protein query – Netscape
http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi#6179901
Search
RID=1106230395-25537-6015...

```
>gi|6179901|gb|AAF05703.1|    homeodomain transcription factor Prospero [Drosophila mela
          Length = 1403

 Score =  938 bits (2424), Expect = 0.0
 Identities = 493/627 (78%), Positives = 493/627 (78%)

Query: 777  HVATAAPRPQMHHPAPARLPTRMGGAAGHTALKSELSEKFQMLRANNNSSMMRMSGTDLE 836
            HVATAAPRPQMHHPAPARLPTRMGGAAGHTALKSELSEKFQMLRANNNSSMMRMSGTDLE
Sbjct: 777  HVATAAPRPQMHHPAPARLPTRMGGAAGHTALKSELSEKFQMLRANNNSSMMRMSGTDLE 836

Query: 837  GLADVLKSEITTSLSALVDTIVTRFVHQRRLFSKQADSVTAAAEQLNKDLLLASQILDRK 896
            GLADVLKSEITTSLSALVDTIVTRFVHQRRLFSKQADSVTAAAEQLNKDLLLASQILDRK
Sbjct: 837  GLADVLKSEITTSLSALVDTIVTRFVHQRRLFSKQADSVTAAAEQLNKDLLLASQILDRK 896

Query: 897  SPRTKVADRPQNGPTPATQSAAAMFQAPKTPQGMNPVAAAALYNSMTGPFCLPPDXXXXX 956
            SPRTKVADRPQNGPTPATQSAAAMFQAPKTPQGMNPVAAAALYNSMTGPFCLPPD
Sbjct: 897  SPRTKVADRPQNGPTPATQSAAAMFQAPKTPQGMNPVAAAALYNSMTGPFCLPPDQQQQQ 956

Query: 957  XXXXXXXXXXXXXXXXXXXXXXLEQNEALSLVVTPKKKRHKVTDTRITPRTVSRILAQDX 1016
                                  LEQNEALSLVVTPKKKRHKVTDTRITPRTVSRILAQD
Sbjct: 957  QTAQQQQSAQQQQQSSQQTQQQLEQNEALSLVVTPKKKRHKVTDTRITPRTVSRILAQDG 1016

Query: 1017 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXASNGGNSNATPAQSPTRSSGGAAYHXXX 1076
                                            ASNGGNSNATPAQSPTRSSGGAAYH
Sbjct: 1017 VVPPTGGPPSTPQQQQQQQQQQQQQQQQQQQQASNGGNSNATPAQSPTRSSGGAAYHPQP 1076

Query: 1077 XXXXXXXXXVSLPTSVAIPNPSLHESKVFSPYSPFFNPXXXXXXXXXXXXXXXXXXXXXX 1136
                     VSLPTSVAIPNPSLHESKVFSPYSPFFNP
Sbjct: 1077 PPPPPPMMPVSLPTSVAIPNPSLHESKVFSPYSPFFNPHAAAGQATAAQLHQHHQQHHPH 1136

Query: 1137 XXXXXXXXXXXXXXXALMDSRDXXXXXXXXXXXXXXXXXXXXXXXXXDYKTCLRAVMDAQ 1196
                           ALMDSRD                         DYKTCLRAVMDAQ
Sbjct: 1137 HQSMQLSSSPPGSLGALMDSRDSPPLPHPPSMLHPALLAAAHHGGSPDYKTCLRAVMDAQ 1196

Query: 1197 DRQSECNSADMQFDGMAPTISFYKQMQLKTEHQESLMAKHCESLTPLHSSTLTPMHLRKA 1256
            DRQSECNSADMQFDGMAPTISFYKQMQLKTEHQESLMAKHCESLTPLHSSTLTPMHLRKA
Sbjct: 1197 DRQSECNSADMQFDGMAPTISFYKQMQLKTEHQESLMAKHCESLTPLHSSTLTPMHLRKA 1256

Query: 1257 KLMFFWVRYPSSAVLKMYFPDIKFNKNNTAQLVKWFSNFREFYYIQMEKYARQAVTEGIK 1316
            KLMFFWVRYPSSAVLKMYFPDIKFNKNNTAQLVKWFSNFREFYYIQMEKYARQAVTEGIK
Sbjct: 1257 KLMFFWVRYPSSAVLKMYFPDIKFNKNNTAQLVKWFSNFREFYYIQMEKYARQAVTEGIK 1316

Query: 1317 TPDDLLIAGDSELYRVLNLHYNRNNHIEVPQNFRFVVESTLREFFRAIQGGKDTEQSWKK 1376
            TPDDLLIAGDSELYRVLNLHYNRNNHIEVPQNFRFVVESTLREFFRAIQGGKDTEQSWKK
Sbjct: 1317 TPDDLLIAGDSELYRVLNLHYNRNNHIEVPQNFRFVVESTLREFFRAIQGGKDTEQSWKK 1376

Query: 1377 SIYKIISRMDDPVPEYFKSPNFLEQLE 1403
            SIYKIISRMDDPVPEYFKSPNFLEQLE
Sbjct: 1377 SIYKIISRMDDPVPEYFKSPNFLEQLE 1403
```

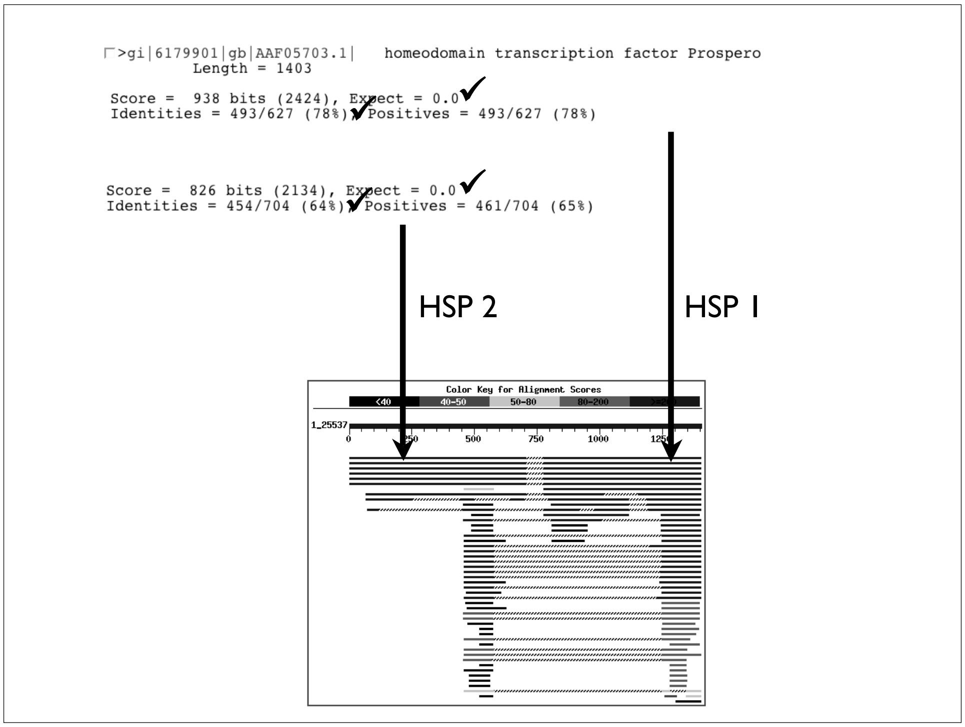
≥ 25% for proteins
≥ 70% for nucleotides

– Gap
x Low-Complexity


RID=1106230395-25537-60155114044.BLASTQ4, Protein query – Netscape
http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi#6179901
Search
RID=1106230395-25537-6015...

```
 Score =  826 bits (2134), Expect = 0.0
 Identities = 454/704 (64%), Positives = 461/704 (65%)

Query: 1    MSSXXXXXXXXXXXXXLFQPQSVSTAXXXXXXXXXXXTPAALATHXXXXXXXXXXXXXXX 60
            MSS             LFQPQSVSTA           TPAALATH
Sbjct: 1    MSSAAAAAAGAAGGGALFQPQSVSTANSSSSNNNNSSTPAALATHSPTSNSPVSGASSAS 60

Query: 61   XXXXXXFGNLFGGSSAKMLNELFGRQMKQAQDATSGLPQSLDNAMLAAAMETATSAELLI 120
                  FGNLFGGSS +          +          QSLDNAMLAAAMETATSAELL
Sbjct: 61   SLLTAAFGNLFGGSSGQDAERAVWPPDEAGPGRNEWPAQSLDNAMLAAAMETATSAELLN 120

Query: 121  GSLNSTSKLLQQQHNNNSIAPANSTPMSNGTNXXXXXXXXXXXXXXXXXXXXKGSRRVSA 180
             +L    ++        ++ P   TPMSNGTN                    KGSRRVSA
Sbjct: 121  LALQFHVQVAAAAAITTALLPPIGTPMSNGTNASISPGSAHSSSHSHQGVSPKGSRRVSA 180

Query: 181  CSDRSLEAAAADVAGGSPPRAASVSSLNGGASSGEQHQSQLQHDLVAHHMLRNILQGKKE 240
            CSDRSLEAAAADVAGGSPPRAASVSSLNGGASSGEQHQSQLQHDLVAHHMLRNILQGKKE
Sbjct: 181  CSDRSLEAAAADVAGGSPPRAASVSSLNGGASSGEQHQSQLQHDLVAHHMLRNILQGKKE 240

Query: 241  LMQLDQELRTAMXXXXXXXXXXXXXHSKLXXXXXXXXXXXXXXXXXXXMESINLIDDSEM 300
            LMQLDQELRTAM             HSKL                   MESINLIDDSEM
Sbjct: 241  LMQLDQELRTAMQQQQQQLQEKEQLHSKLNNNNNNNIAATANNNNNTTMESINLIDDSEM 300

Query: 301  ADIKIKSEPQTAPQPQQXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXHGXXXX 360
            ADIKIKSEPQTAPQPQQ                                     HG
Sbjct: 301  ADIKIKSEPQTAPQPQQSPHGSSHSSRSGSGSGSHSSMASDGSLRRKSSDSLDSHGAQDD 360

Query: 361  XXXXXXXXPTGQRSESRAPEEPQLPTKKESVDDMLDEVELLGLHSRGSDMDSLASPSHSX 420
                    PTGQRSESRAPEEPQLPTKKESVDDMLDEVELLGLHSRGSDMDSLASPSHS
Sbjct: 361  AQDEEDAAPTGQRSESRAPEEPQLPTKKESVDDMLDEVELLGLHSRGSDMDSLASPSHSD 420

Query: 421  XXXXXXXXXXXXXXXXXCVEQKTSGSGCLKKPGMDLKRARVENIVSGMRCSPSSGLAQAG 480
                             CVEQKTSGSGCLKKPGMDLKRARVENIVSGMRCSPSSGLAQAG
Sbjct: 421  MMLLDKDDVLDEDDDDDCVEQKTSGSGCLKKPGMDLKRARVENIVSGMRCSPSSGLAQAG 480

Query: 481  QLQVNGCKKRKLYQPQQHAMERYVXXXXGLNFGLNLQSMMLDQEDSESNELESPQIQQKR 540
            QLQVNGCKKRKLYQPQQHAMERYV    GLNFGLNLQSMMLDQEDSESNELESPQIQQKR
Sbjct: 481  QLQVNGCKKRKLYQPQQHAMERYVAAAAGLNFGLNLQSMMLDQEDSESNELESPQIQQKR 540

Query: 541  VEKNALKSQLRSMQEQLAEMQQKYVQLCSRMEQESXXXXXXXXXXXXXXXXXXNGSSDHI 600
            VEKNALKSQLRSMQEQLAEMQQKYVQLCSRMEQES                  NGSSDHI
Sbjct: 541  VEKNALKSQLRSMQEQLAEMQQKYVQLCSRMEQESECQELDQDQDVEQEQEPDNGSSDHI 600

Query: 601  ELSPSPTLTGDGDVSPNHKEETGQERXXXXXXXXXXXXXXXXXXESSDSGANMLSQMMSK 660
            ELSPSPTLTGDGDVSPNHKEETGQER                  ESSDSGANMLSQMMSK
Sbjct: 601  ELSPSPTLTGDGDVSPNHKEETGQERPGSSSPSPSPLKPKTSLGESSDSGANMLSQMMSK 660
```

No definition line ∴
second HSP identified

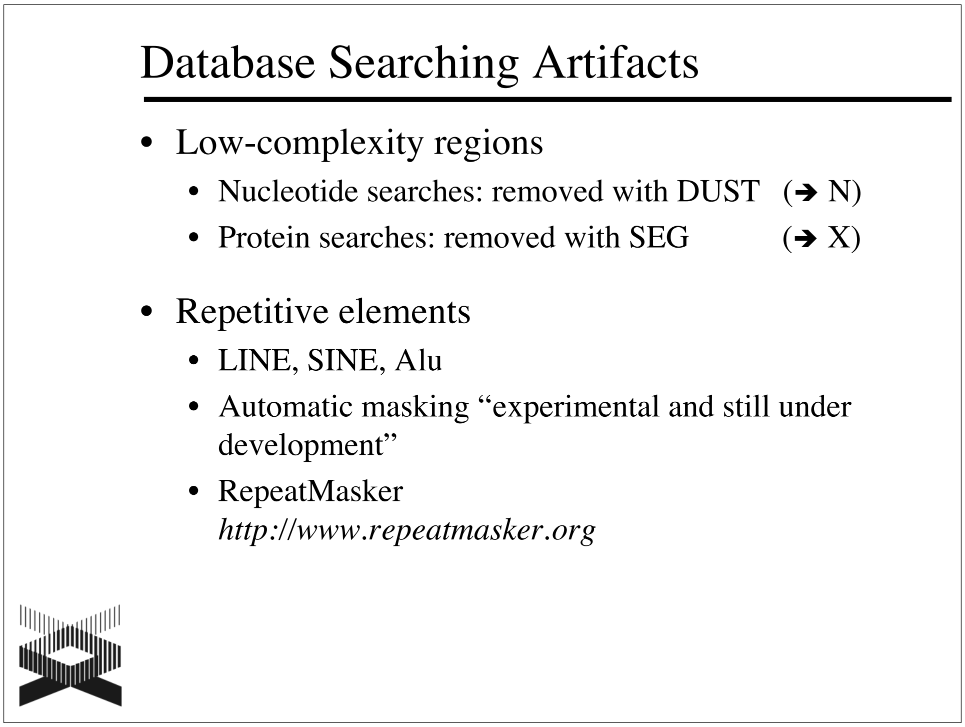


```
>gi|6179901|gb|AAF05703.1|   homeodomain transcription factor Prospero
        Length = 1403

 Score =  938 bits (2424), Expect = 0.0 ✓
 Identities = 493/627 (78%),✓ Positives = 493/627 (78%)


 Score =  826 bits (2134), Expect = 0.0 ✓
 Identities = 454/704 (64%),✓ Positives = 461/704 (65%)
```

Suggested BLAST Cutoffs

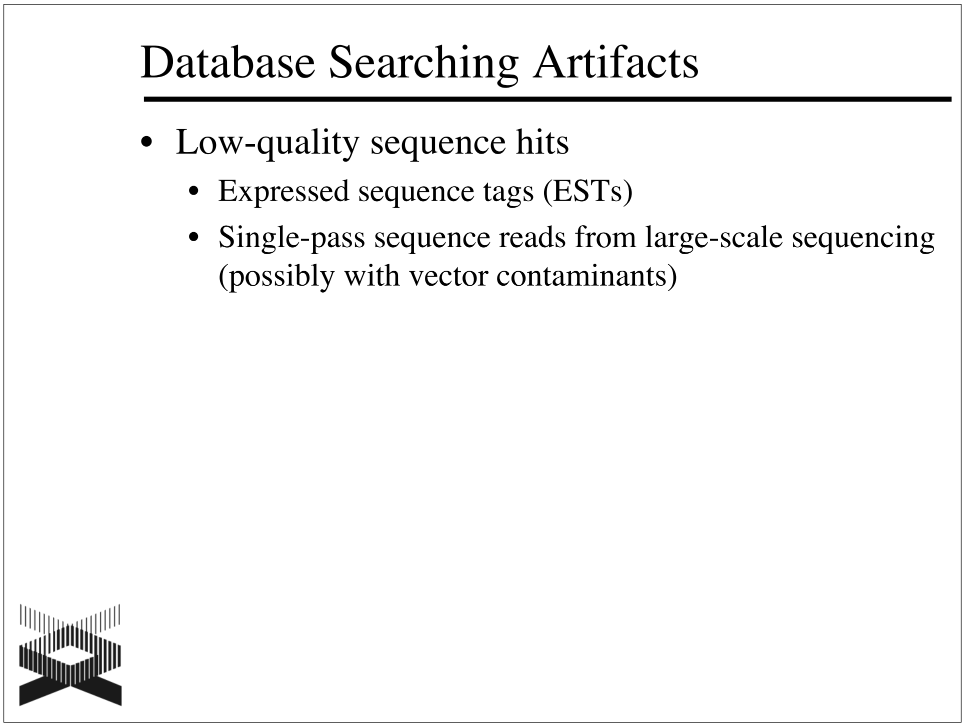
	E value	Sequence Identity
Nucleotide	$\leq 10^{-6}$	$\geq$ 70%
Protein	$\leq 10^{-3}$	$\geq$ 25%

- *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

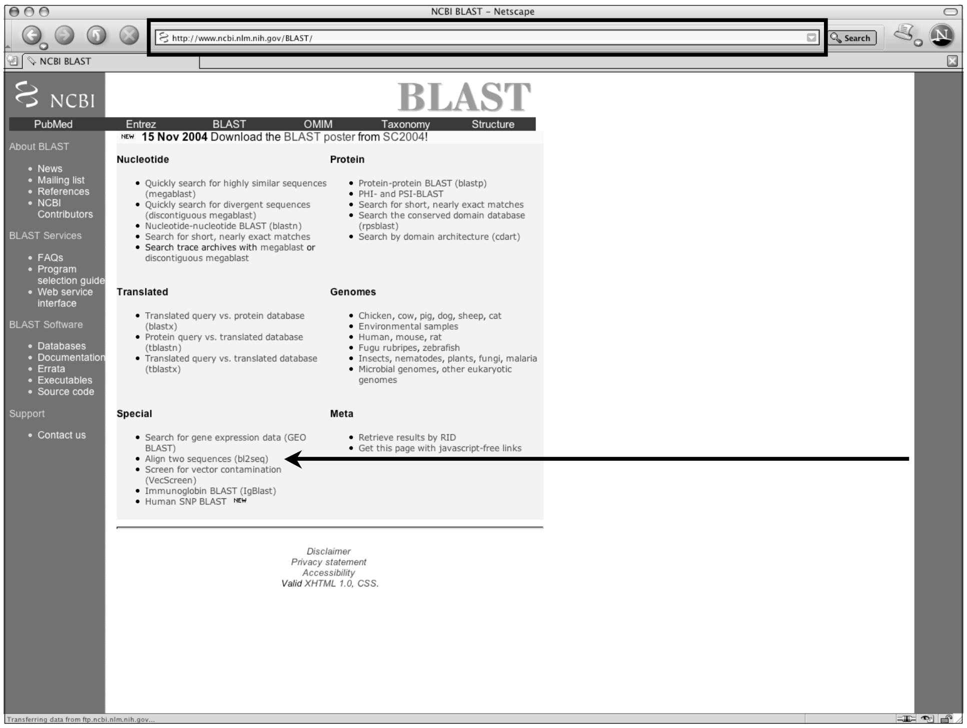


Database Searching Artifacts

- 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/*



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

NEW **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)

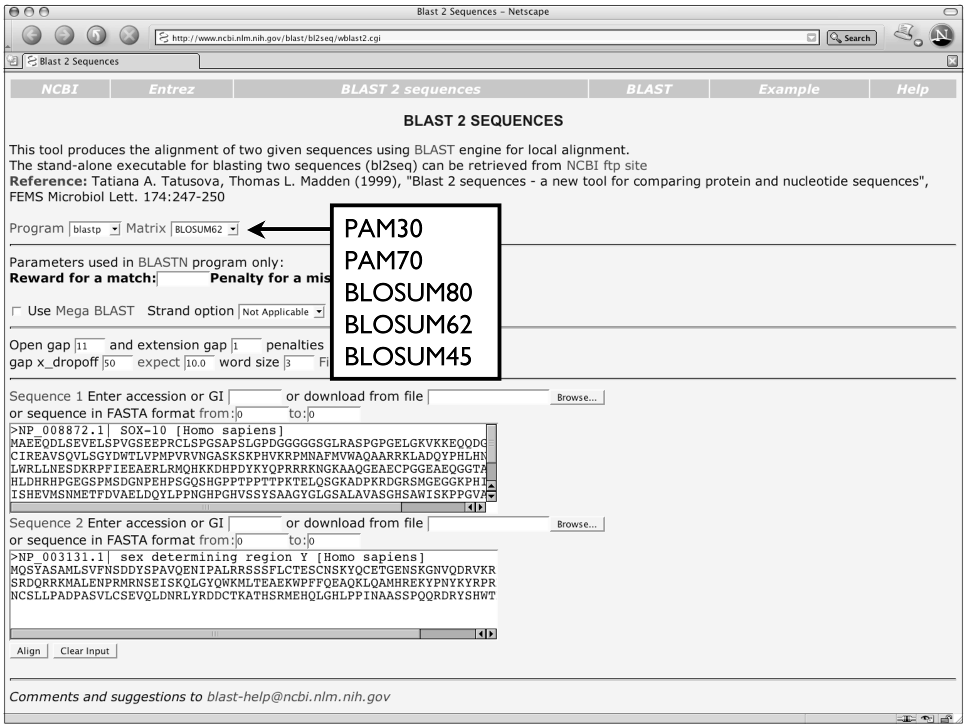
Genomes
- Chicken, cow, pig, dog, sheep, cat
- Environmental samples
- Human, mouse, rat
- Fugu rubripes, zebrafish
- Insects, nematodes, plants, fungi, malaria
- Microbial genomes, other eukaryotic genomes

Special
- Search for gene expression data (GEO BLAST)
- Align two sequences (bl2seq)
- Screen for vector contamination (VecScreen)
- Immunoglobin BLAST (IgBlast)
- Human SNP BLAST NEW

Meta
- Retrieve results by RID
- Get this page with javascript-free links

Disclaimer
Privacy statement
Accessibility
Valid *XHTML 1.0, CSS.*

Transferring data from ftp.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

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 Fi

Sequence 1 Enter accession or GI or download from file Browse...
or sequence in FASTA format from: 0 to: 0

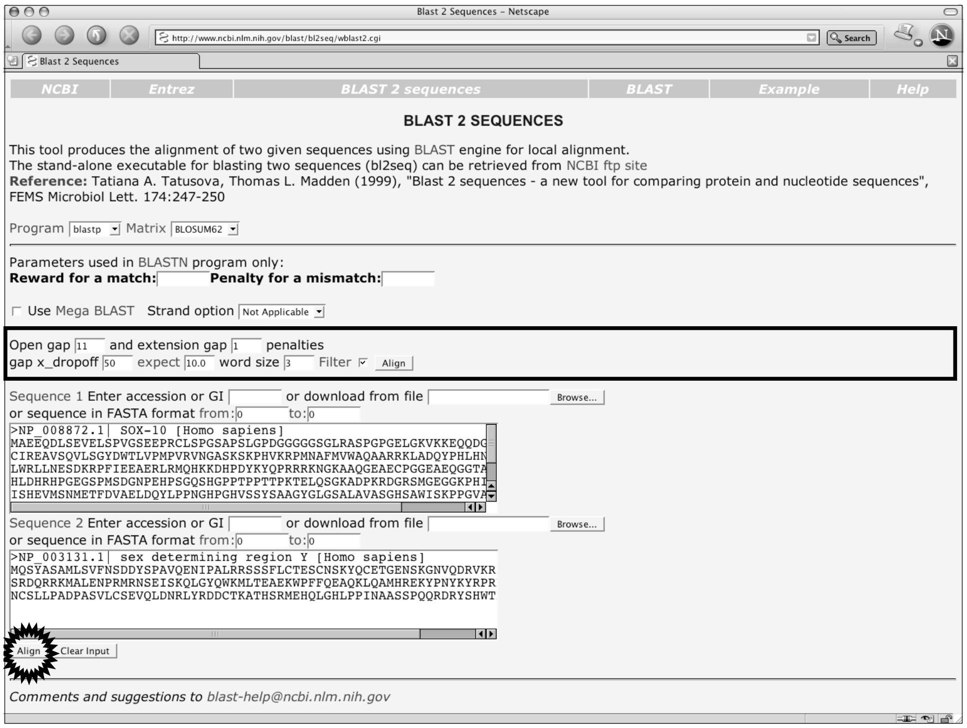
```
>NP_008872.1| SOX-10 [Homo sapiens]
MAEEQDLSEVELSPVGSEEPRCLSPGSAPSLGPDGGGGGSGLRASPGPGELGKVKKEQQDG
CIREAVSQVLSGYDWTLVPMPVRVNGASKSKPHVKRPMNAFMVWAQAARRKLADQYPHLHN
LWRLLNESDKRPFIEEAERLRMQHKKDHPDYKYQPRRRKNGKAAQGEAECPGGEAEQGGTA
HLDHRHPGEGSPMSDGNPEHPSGQSHGPPTPPTTPKTELQSGKADPKRDGRSMGEGGKPHI
ISHEVMSNMETFDVAELDQYLPPNGHPGHVSSYSAAGYGLGSALAVASGHSAWISKPPGVA
```

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
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

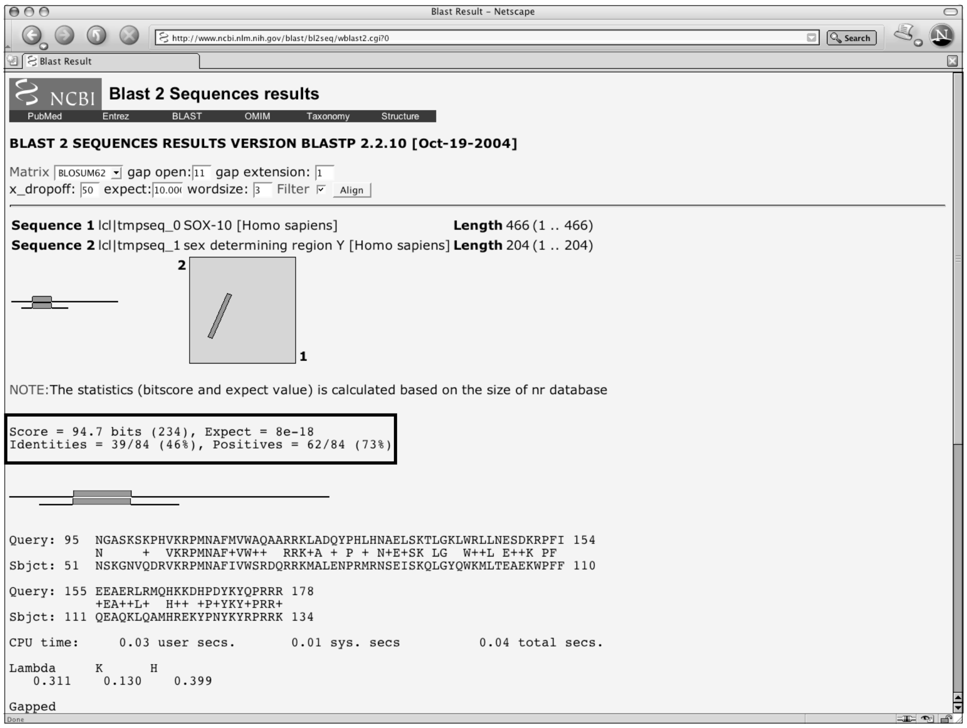
```
>NP_008872.1| SOX-10 [Homo sapiens]
MAEEQDLSEVELSPVGSEEPRCLSPGSAPSLGPDGGGGGSGLRASPGPGELGKVKKEQQDG
CIREAVSQVLSGYDWTLVPMPVRVNGASKSKPHVKRPMNAFMVWAQAARRKLADQYPHLHN
LWRLLNESDKRPFIEEAERLRMQHKKDHPDYKYQPRRRKNGKAAQGEAECPGGEAEQGGTA
HLDHRHPGEGSPMSDGNPEHPSGQSHGPPTPPTTPKTELQSGKADPKRDGRSMGEGGKPHI
ISHEVMSNMETFDVAELDQYLPPNGHPGHVSSYSAAGYGLGSALAVASGHSAWISKPPGVA
```

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 results

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 ☑

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)

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

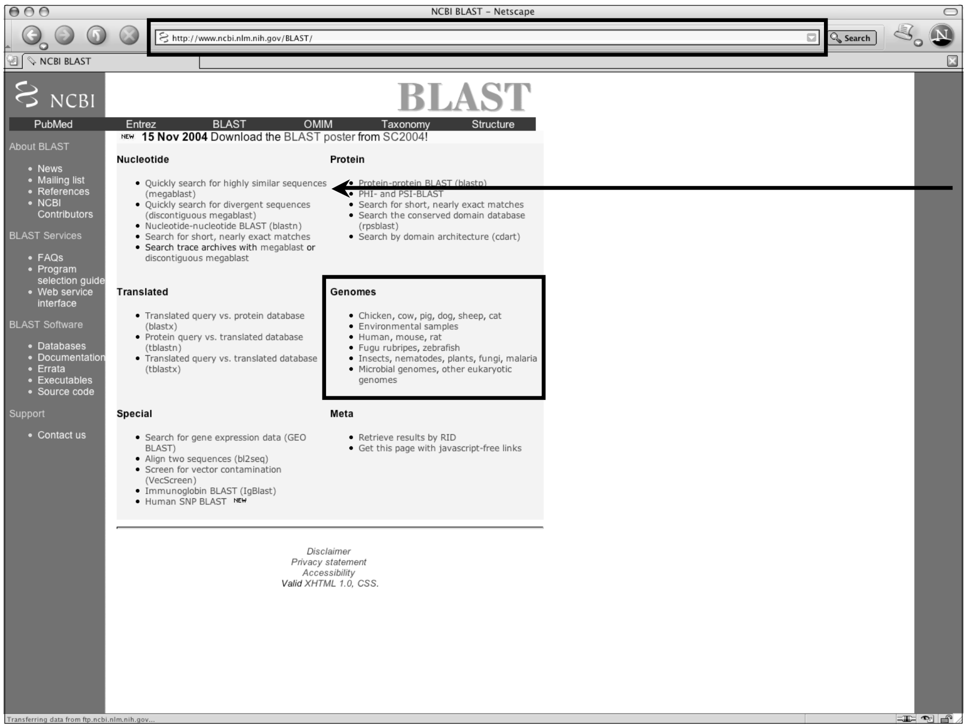
BLASTN *vs*. MegaBLAST

- Word size
  - BLASTN default = 11
  - MegaBLAST default = 28

- *Non-affine* gap penalties

$$\boxed{\text{Deduction for a gap} = r/2 - q}$$

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



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

PubMed | Entrez | BLAST | OMIM | Taxonomy | Structure

NEW **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

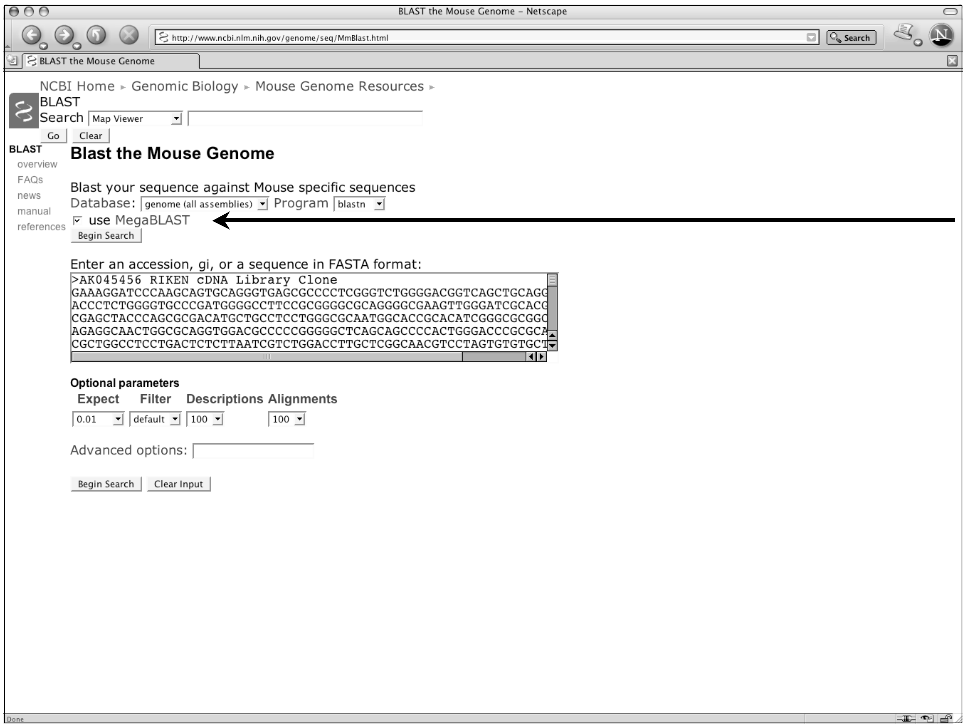
- Search for gene expression data (GEO BLAST)
- Align two sequences (bl2seq)
- Screen for vector contamination (VecScreen)
- Immunoglobin BLAST (IgBlast)
- Human SNP BLAST NEW

Meta

- Retrieve results by RID
- Get this page with javascript-free links

Disclaimer
Privacy statement
Accessibility
Valid *XHTML 1.0, CSS.*

Transferring data from ftp.ncbi.nlm.nih.gov...


BLAST the Mouse Genome – Netscape
http://www.ncbi.nlm.nih.gov/genome/seq/MmBlast.html
Search
BLAST the Mouse Genome
NCBI Home ▸ Genomic Biology ▸ Mouse Genome Resources ▸
BLAST
Search
Map Viewer
Go
Clear
BLAST
overview
FAQs
news
manual
references
Begin Search
Begin Search
Clear Input
Done

Blast the Mouse Genome

Blast your sequence against Mouse specific sequences

Database: genome (all assemblies) Program blastn

☑ use MegaBLAST

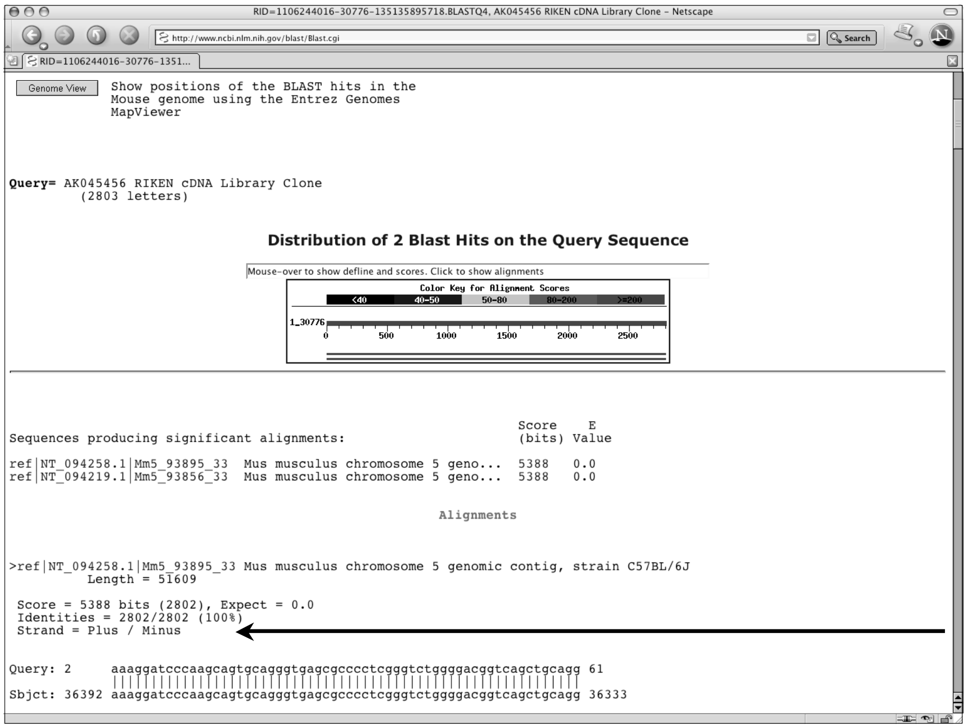
Enter an accession, gi, or a sequence in FASTA format:

```
>AK045456 RIKEN cDNA Library Clone
GAAAGGATCCCAAGCAGTGCAGGGTGAGCGCCCCTCGGGTCTGGGGACGGTCAGCTGCAGG
ACCCTCTGGGGTGCCCGATGGGGCCTTCCGCGGGGCGCAGGGGCGAAGTTGGGATCGCACG
CGAGCTACCCAGCGCGACATGCTGCCTCCTGGGCGCAATGGCACCGCACATCGGGCGCGGG
AGAGGCAACTGGCGCAGGTGGACGCCCCCGGGGGCTCAGCAGCCCCACTGGGACCCGCGCA
CGCTGGCCTCCTGACTCTCTTAATCGTCTGGACCTTGCTCGGCAACGTCCTAGTGTGTGCT
```

Optional parameters

Expect	Filter	Descriptions	Alignments
0.01	default	100	100

Advanced options:


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
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

Show positions of the BLAST hits in the
Mouse genome using the Entrez Genomes
MapViewer

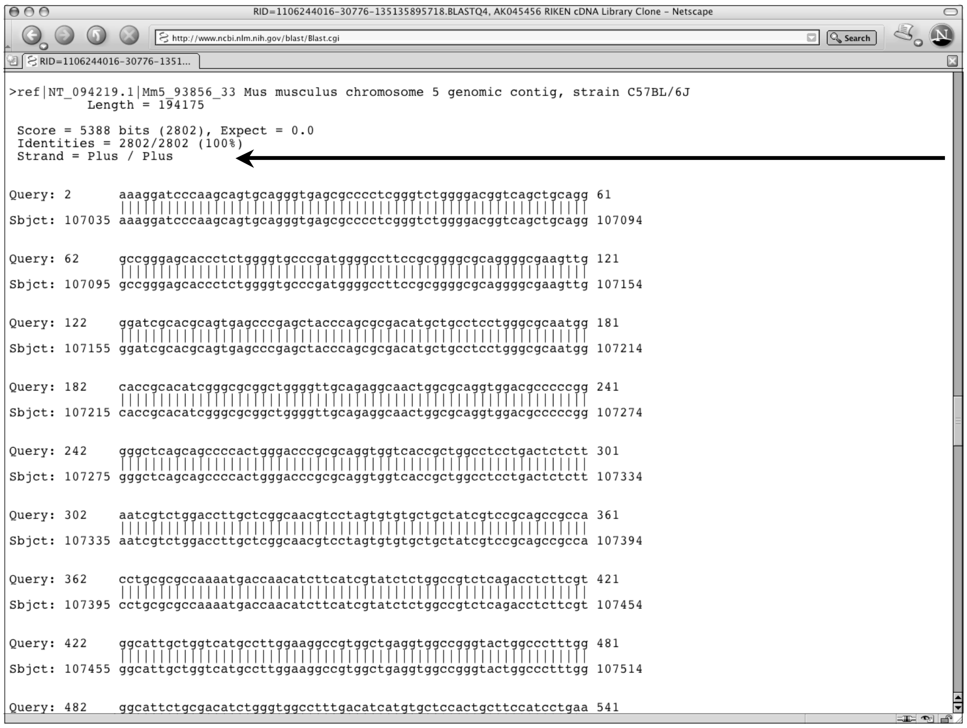
```
Query= AK045456 RIKEN cDNA Library Clone
         (2803 letters)
```

Distribution of 2 Blast Hits on the Query Sequence

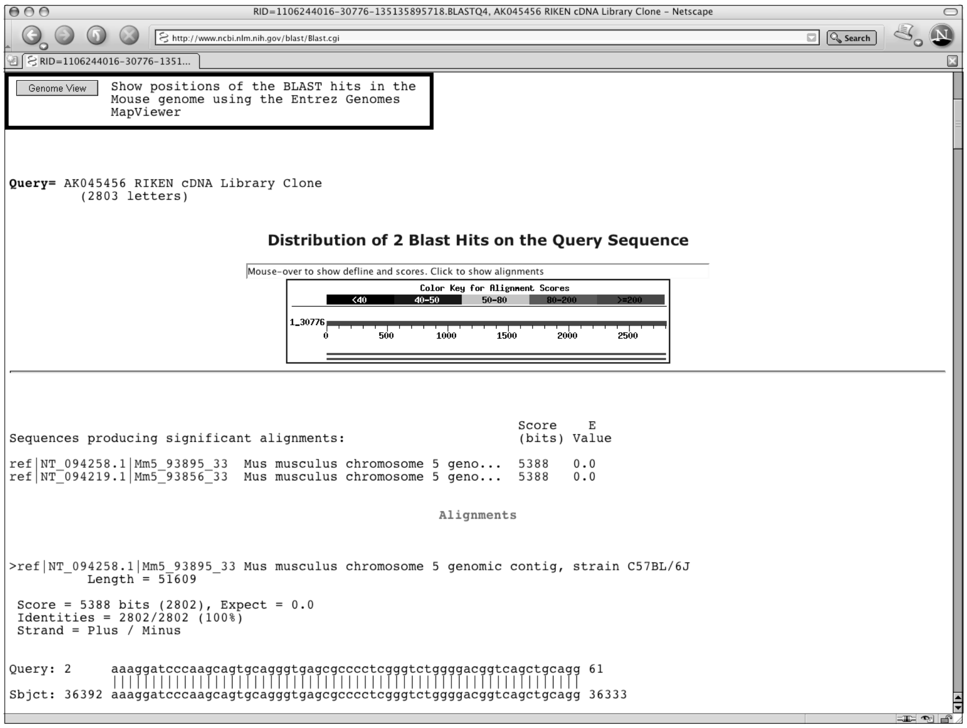
```
                                                                Score    E
Sequences producing significant alignments:                     (bits) Value

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
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

Show positions of the BLAST hits in the Mouse genome using the Entrez Genomes MapViewer

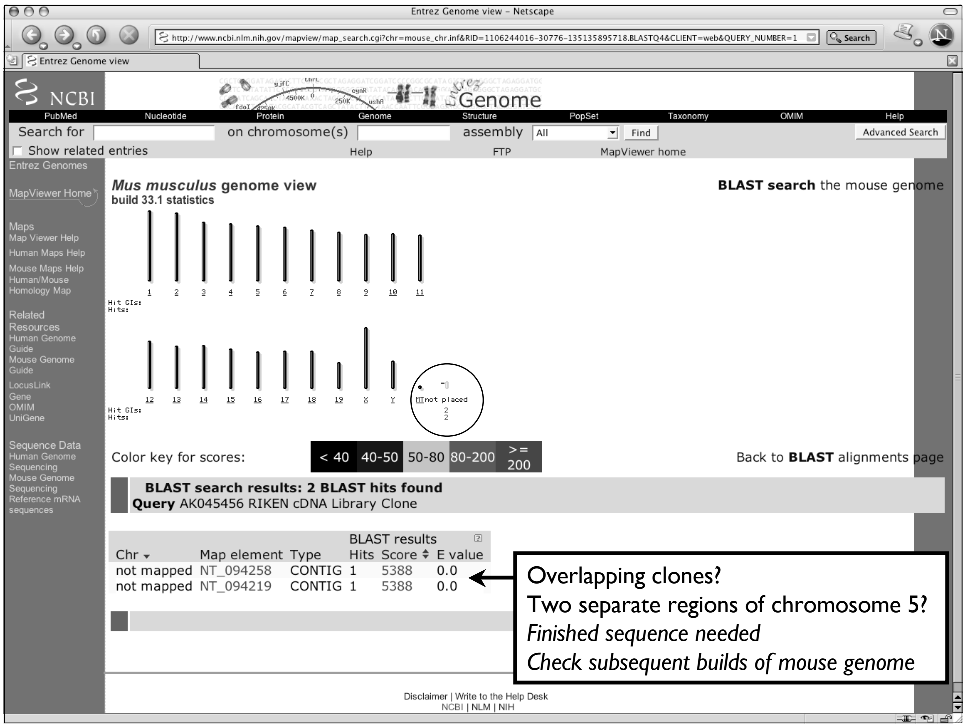
Query= AK045456 RIKEN cDNA Library Clone
(2803 letters)

Distribution of 2 Blast Hits on the Query Sequence

```
                                                                 Score    E
Sequences producing significant alignments:                      (bits) 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
```



***Mus musculus* genome view**
build 33.1 statistics

BLAST search the mouse genome

Color key for scores: < 40 | 40-50 | 50-80 | 80-200 | >= 200

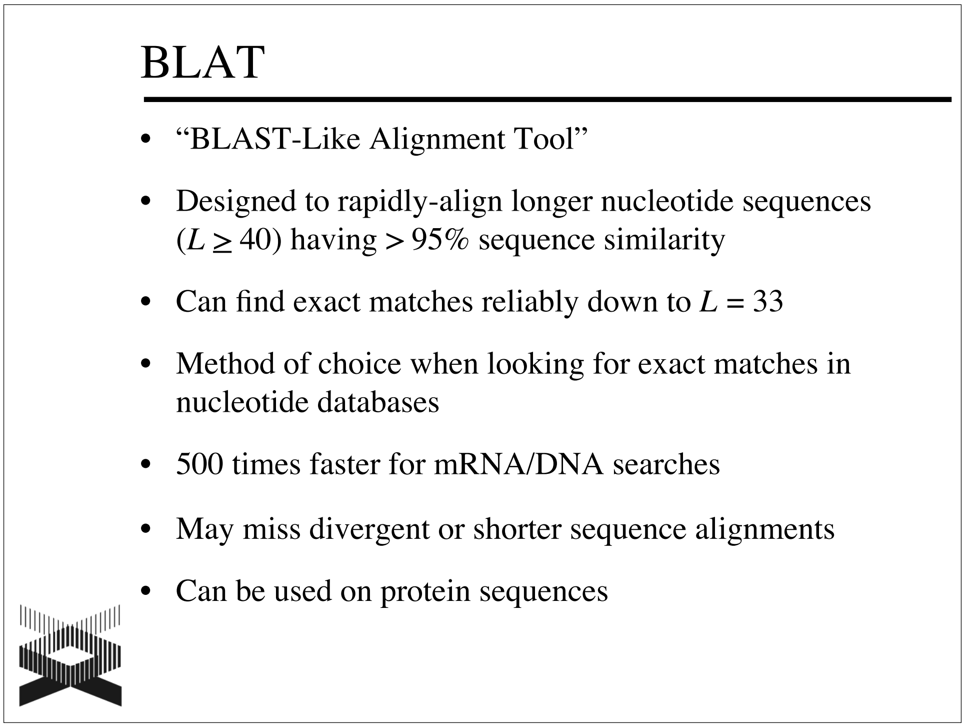
BLAST search results: 2 BLAST hits found
Query AK045456 RIKEN cDNA Library Clone

			BLAST results		
Chr	Map element	Type	Hits	Score	E value
not mapped	NT_094258	CONTIG	1	5388	0.0
not mapped	NT_094219	CONTIG	1	5388	0.0

Overlapping clones?
Two separate regions of chromosome 5?
Finished sequence needed
Check subsequent builds of mouse genome

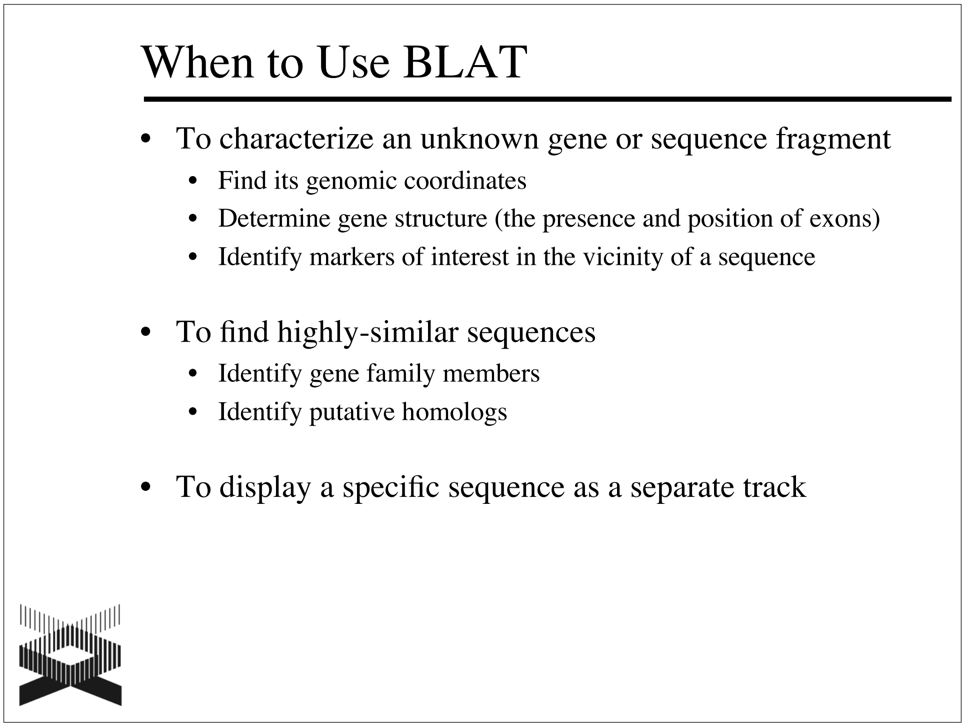
Overview

- 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



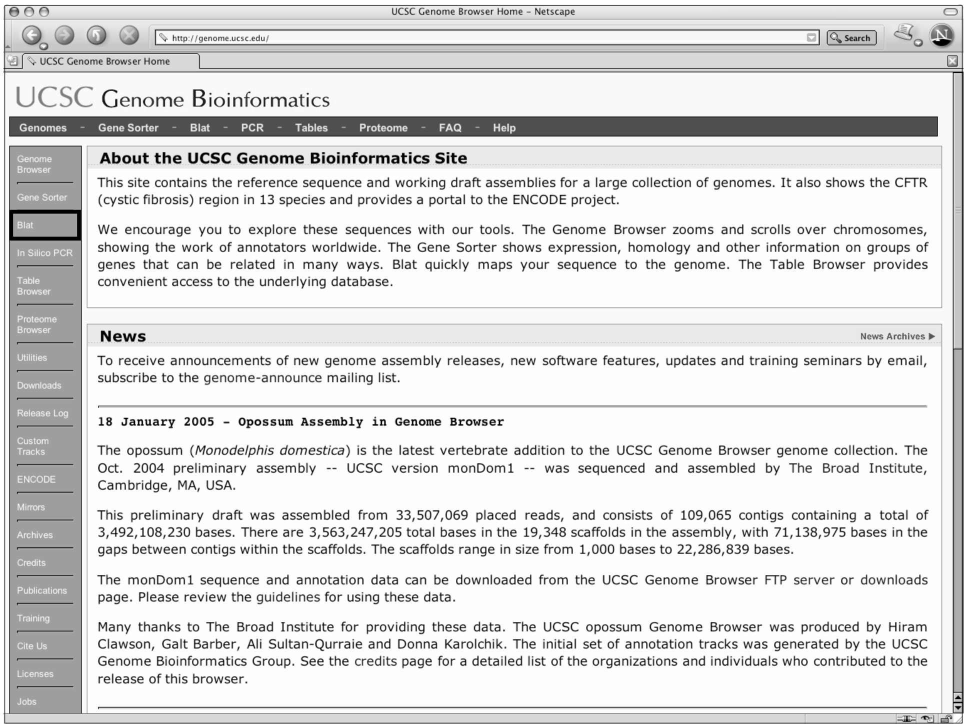
BLAT

- "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



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

To receive announcements of new genome assembly releases, new software features, updates and training seminars by email, subscribe to the genome-announce mailing list.

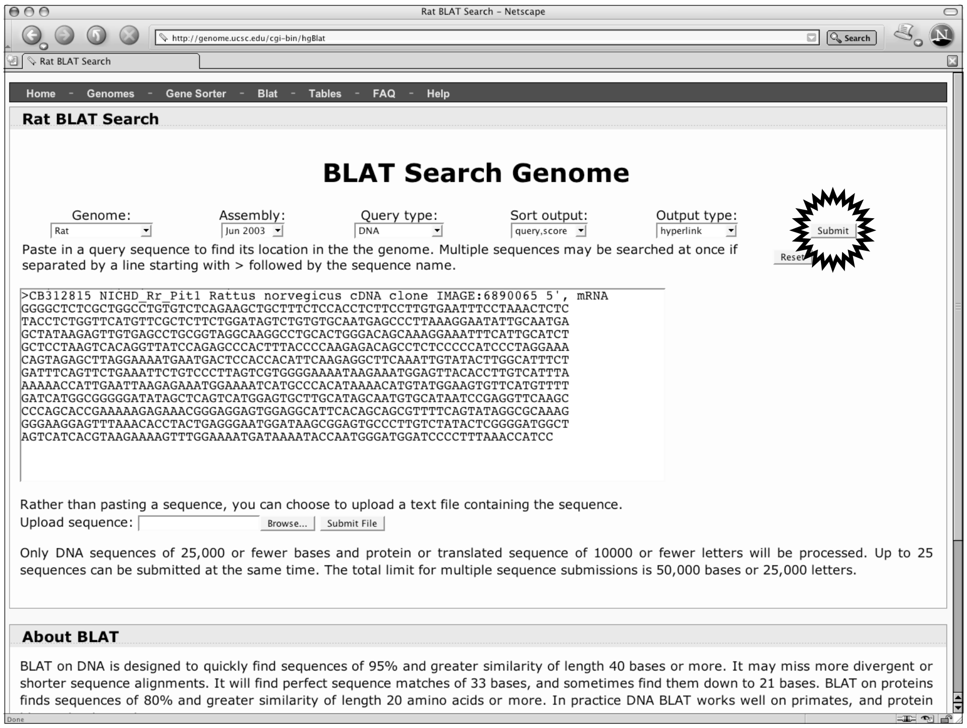
18 January 2005 - Opossum Assembly in Genome Browser

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
Submit
Reset
Browse...
Submit File
Done

Rat BLAT Search

BLAT Search Genome

Genome: Rat
Assembly: Jun 2003
Query type: DNA
Sort output: query,score
Output type: hyperlink

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.

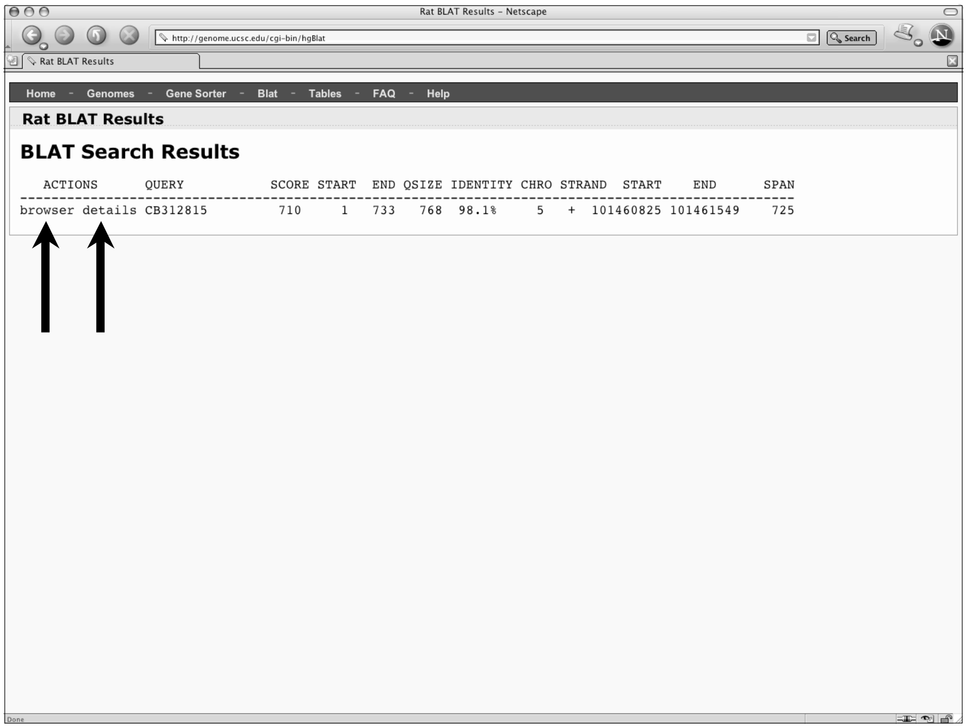
```
>CB312815 NICHD_Rr_Pit1 Rattus norvegicus cDNA clone IMAGE:6890065 5', mRNA
GGGGCTCTCGCTGGCCTGTGTCTCAGAAGCTGCTTTCTCCACCTCTTCCTTGTGAATTTCCTAAACTCTC
TACCTCTGGTTCATGTTCGCTCTTCTGGATAGTCTGTGTGCAATGAGCCCTTAAAGGAATATTGCAATGA
GCTATAAGAGTTGTGAGCCTGCGGTAGGCAAGGCCTGCACTGGGACAGCAAAGGAAATTTCATTGCATCT
GCTCCTAAGTCACAGGTTATCCAGAGCCCACTTTACCCCAAGAGACAGCCTCTCCCCCATCCCTAGGAAA
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:

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

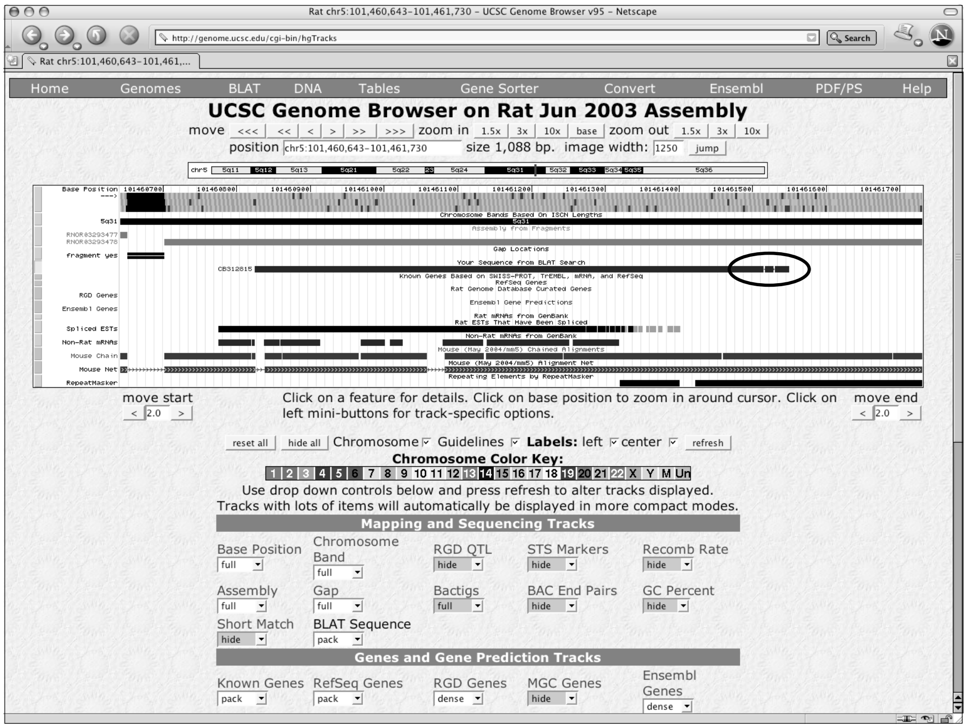
BLAT on DNA is designed to quickly find sequences of 95% and greater similarity of length 40 bases or more. It may miss more divergent or shorter sequence alignments. It will find perfect sequence matches of 33 bases, and sometimes find them down to 21 bases. BLAT on proteins finds sequences of 80% and 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
Done

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
```



UCSC Genome Browser on Rat Jun 2003 Assembly

position chr5:101,460,643-101,461,730 size 1,088 bp. image width: 1250

move start 2.0

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 2.0

Chromosome ☑ Guidelines ☑ Labels: left ☑ center ☑

Chromosome Color Key:

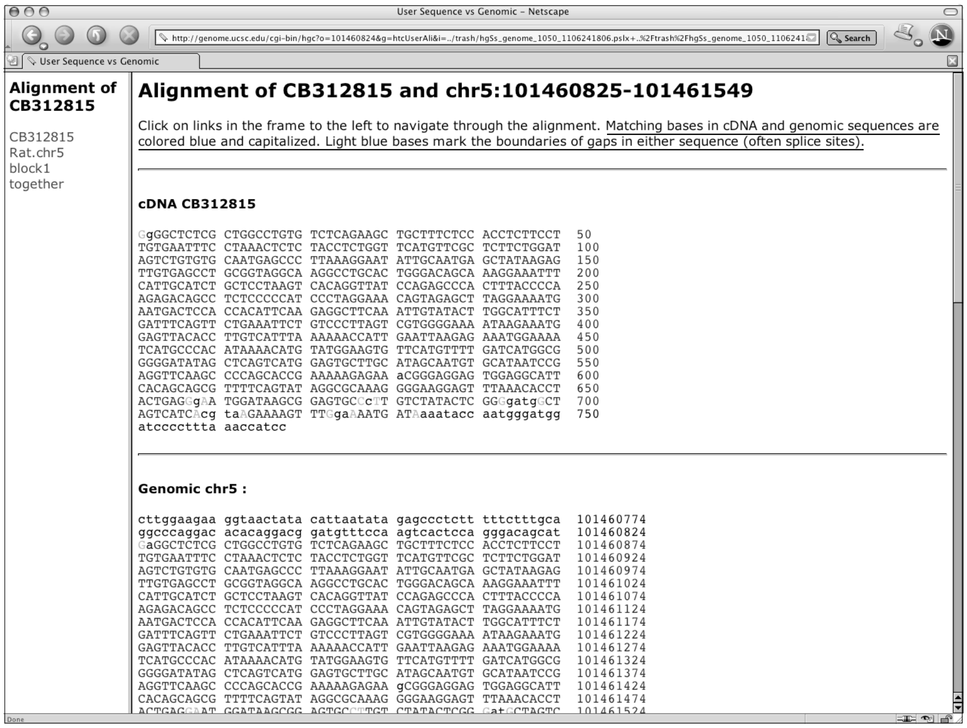
Use drop down controls below and press refresh to alter tracks displayed.
Tracks with lots of items will automatically be displayed in more compact modes.

Mapping and Sequencing Tracks

Base Position full	Chromosome Band full	RGD QTL hide	STS Markers hide	Recomb Rate hide
Assembly full	Gap full	Bactigs full	BAC End Pairs hide	GC Percent hide
Short Match hide	BLAT Sequence pack			

Genes and Gene Prediction Tracks

Known Genes pack	RefSeq Genes pack	RGD Genes dense	MGC Genes hide	Ensembl Genes dense


User Sequence vs Genomic – Netscape
http://genome.ucsc.edu/cgi-bin/hgc?o=101460824&g=htcUserAli&i=../trash/hgSs_genome_1050_1106241806.pslx+..%2Ftrash%2FhgSs_genome_1050_1106241
Search
User Sequence vs Genomic
Done

Alignment of CB312815

CB312815
Rat.chr5
block1
together

Alignment of CB312815 and chr5:101460825-101461549

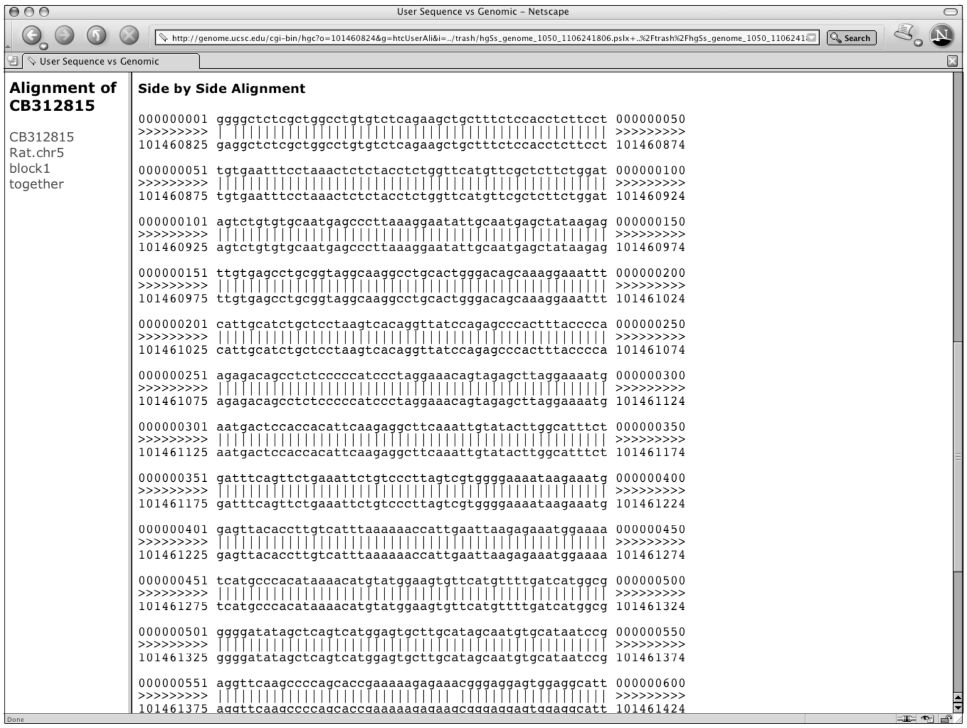
Click on links in the frame to the left to navigate through the alignment. Matching bases in cDNA and genomic sequences are colored blue and capitalized. Light blue bases mark the boundaries of gaps in either sequence (often splice sites).

cDNA CB312815

```
GgGGCTCTCG 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
```

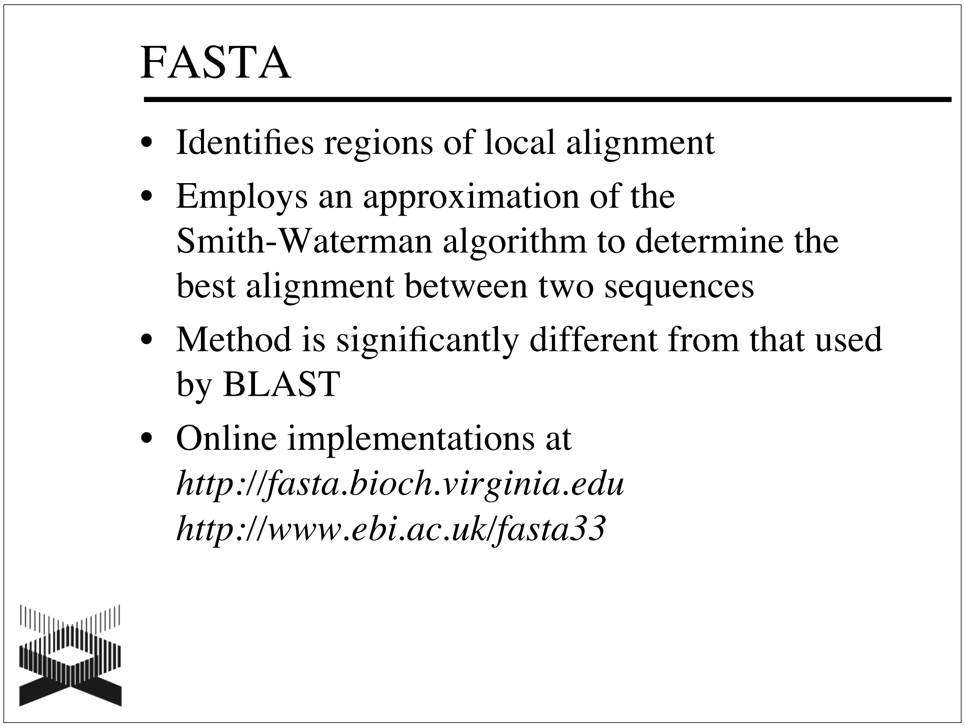

User Sequence vs Genomic – Netscape
http://genome.ucsc.edu/cgi-bin/hgc?o=101460824&g=htcUserAli&i=../trash/hgSs_genome_1050_1106241806.pslx+..%2Ftrash%2FhgSs_genome_1050_1106241
Search
User Sequence vs Genomic
Done

Alignment of CB312815

CB312815
Rat.chr5
block1
together

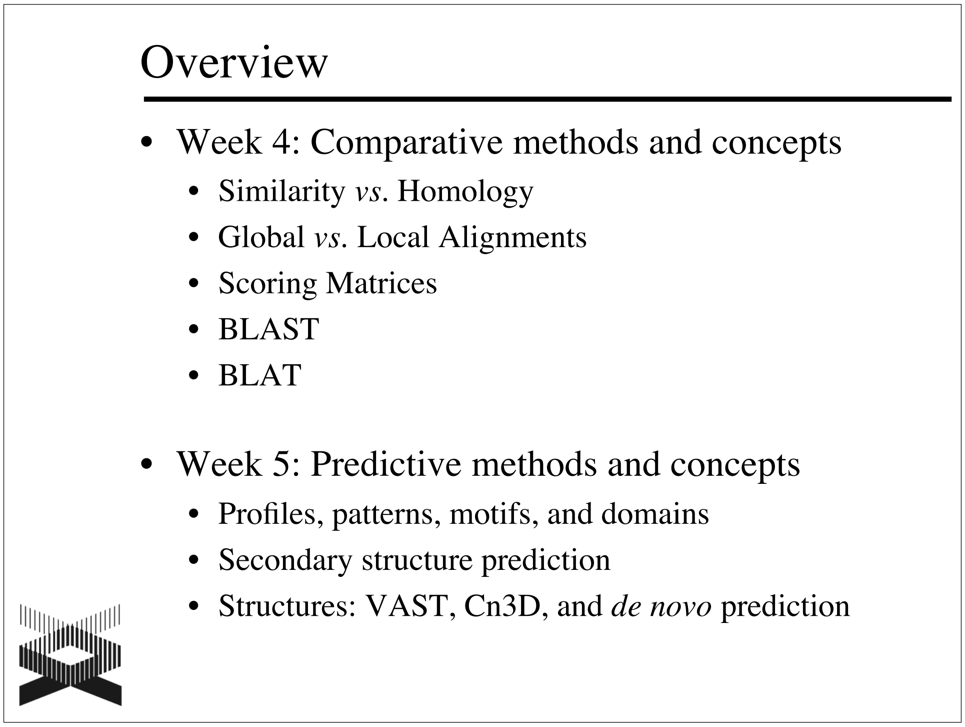
Side by Side Alignment

```
000000001 ggggctctcgctggcctgtgtctcagaagctgctttctccacctcttcct 000000050
>>>>>>>>> | |||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>
101460825 gaggctctcgctggcctgtgtctcagaagctgctttctccacctcttcct 101460874

000000051 tgtgaatttcctaaactctctacctctggttcatgttcgctcttctggat 000000100
>>>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>
101460875 tgtgaatttcctaaactctctacctctggttcatgttcgctcttctggat 101460924

000000101 agtctgtgtgcaatgagcccttaaaggaatattgcaatgagctataagag 000000150
>>>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>
101460925 agtctgtgtgcaatgagcccttaaaggaatattgcaatgagctataagag 101460974

000000151 ttgtgagcctgcggtaggcaaggcctgcactgggacagcaaaggaaattt 000000200
>>>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>
101460975 ttgtgagcctgcggtaggcaaggcctgcactgggacagcaaaggaaattt 101461024

000000201 cattgcatctgctcctaagtcacaggttatccagagcccactttacccca 000000250
>>>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>
101461025 cattgcatctgctcctaagtcacaggttatccagagcccactttacccca 101461074

000000251 agagacagcctctcccccatccctaggaaacagtagagcttaggaaaatg 000000300
>>>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>
101461075 agagacagcctctcccccatccctaggaaacagtagagcttaggaaaatg 101461124

000000301 aatgactccaccacattcaagaggcttcaaattgtatacttggcatttct 000000350
>>>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>
101461125 aatgactccaccacattcaagaggcttcaaattgtatacttggcatttct 101461174

000000351 gatttcagttctgaaattctgtcccttagtcgtggggaaaataagaaatg 000000400
>>>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>
101461175 gatttcagttctgaaattctgtcccttagtcgtggggaaaataagaaatg 101461224

000000401 gagttacaccttgtcatttaaaaaaccattgaattaagagaaatggaaaa 000000450
>>>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>
101461225 gagttacaccttgtcatttaaaaaaccattgaattaagagaaatggaaaa 101461274

000000451 tcatgcccacataaaacatgtatggaagtgttcatgttttgatcatggcg 000000500
>>>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>
101461275 tcatgcccacataaaacatgtatggaagtgttcatgttttgatcatggcg 101461324

000000501 ggggatatagctcagtcatggagtgcttgcatagcaatgtgcataatccg 000000550
>>>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>
101461325 ggggatatagctcagtcatggagtgcttgcatagcaatgtgcataatccg 101461374

000000551 aggttcaagccccagcaccgaaaaagagaaacgggaggagtggaggcatt 000000600
>>>>>>>>> |||||||||||||||||||||||||||||| ||||||||||||||||||| >>>>>>>>>
101461375 aggttcaagccccagcaccgaaaaagagaagcgggaggagtggaggcatt 101461424
```



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.