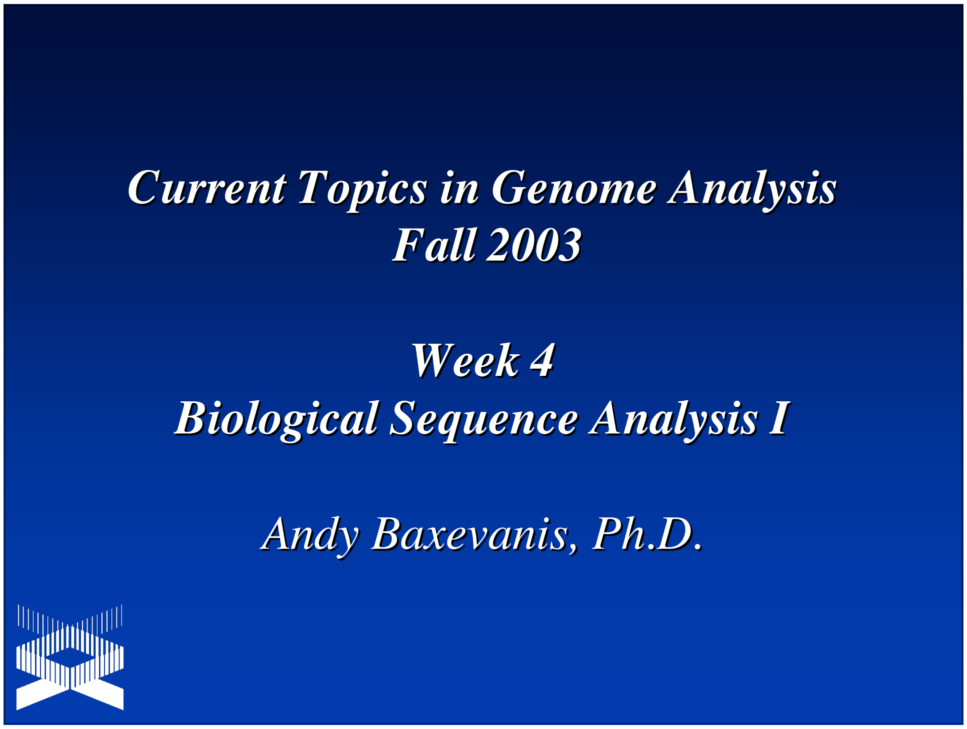
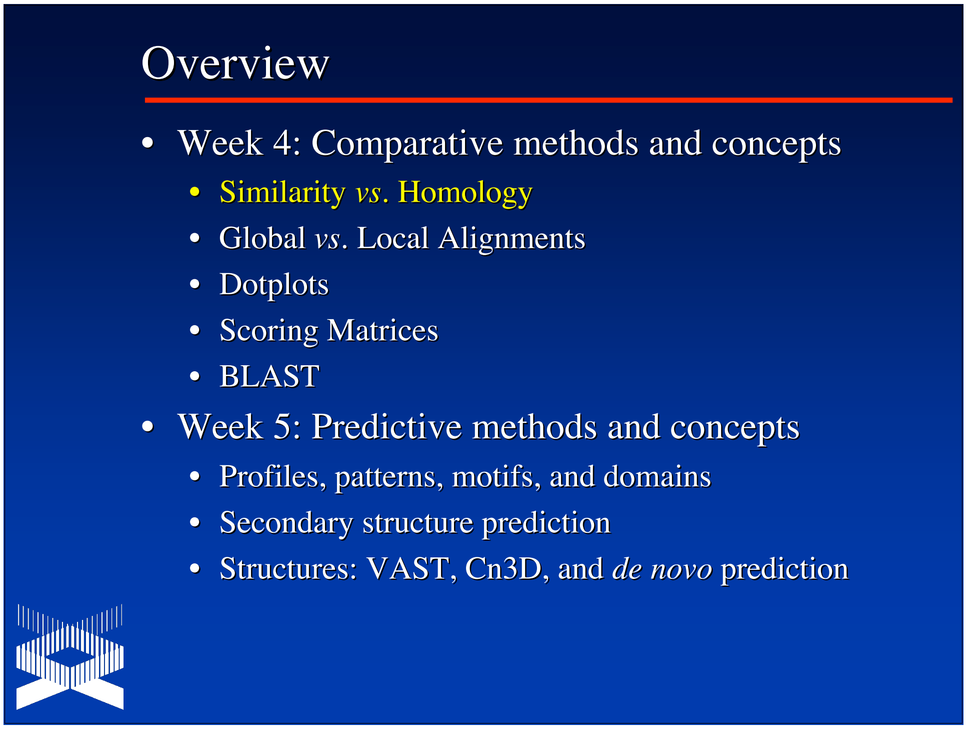


Current Topics in Genome Analysis Fall 2003

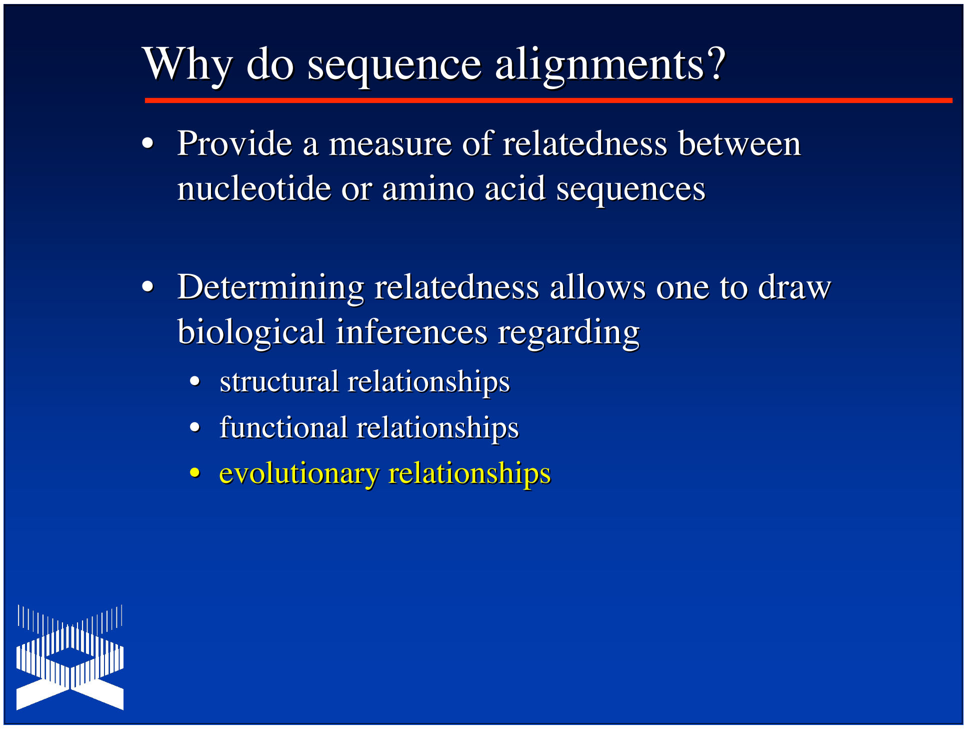
Week 4 Biological Sequence Analysis I

Andy Baxevanis, Ph.D.



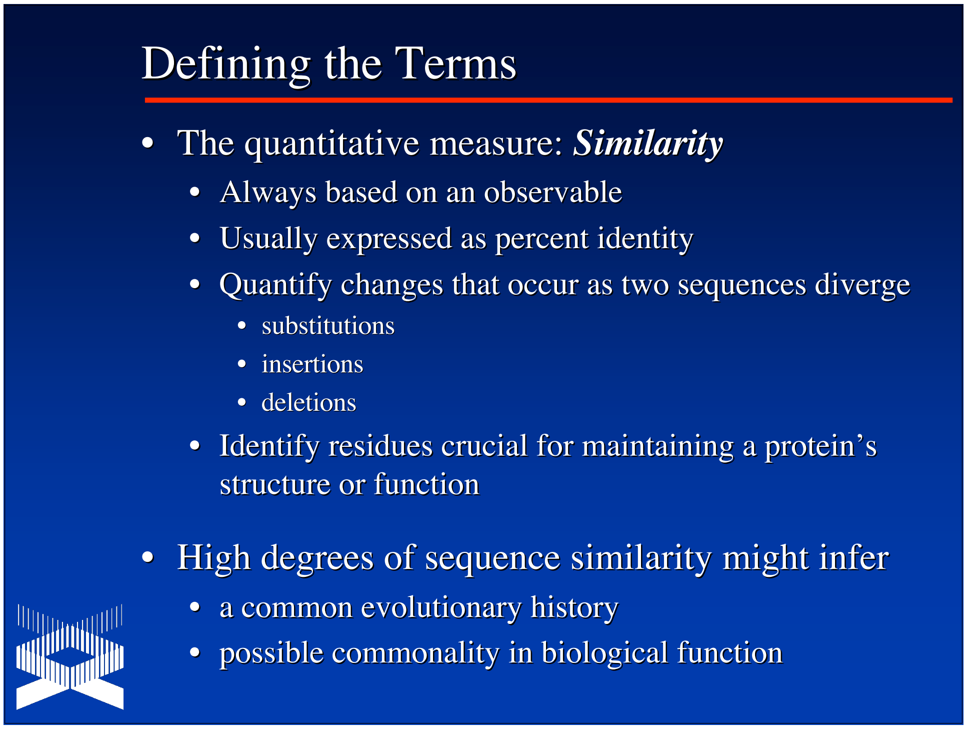
Overview

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



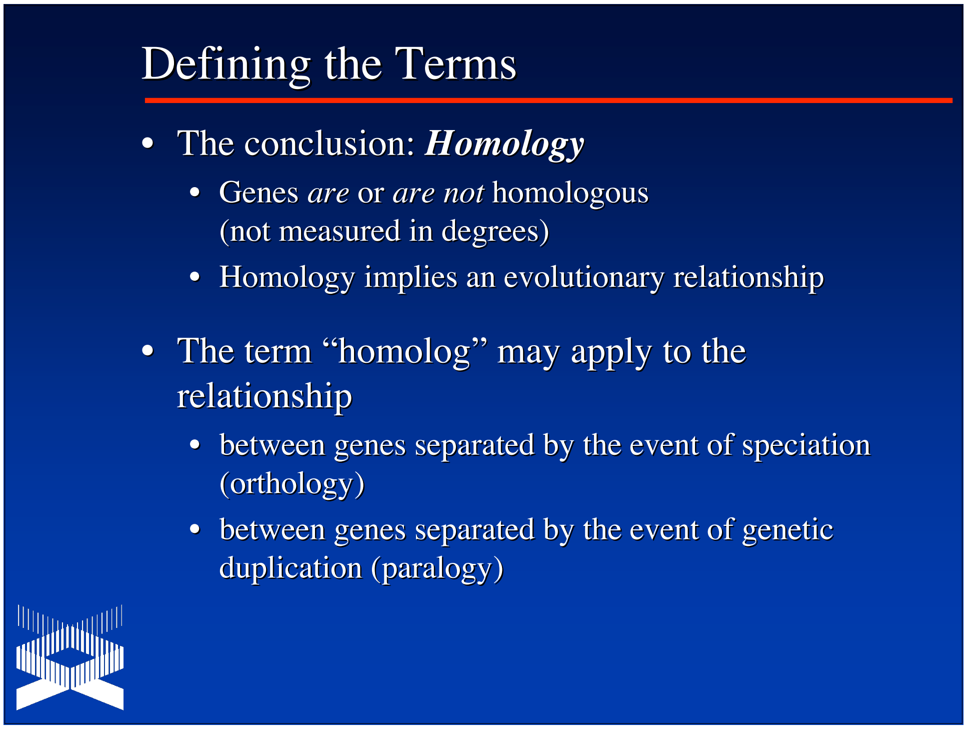
Why do sequence alignments?

- 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



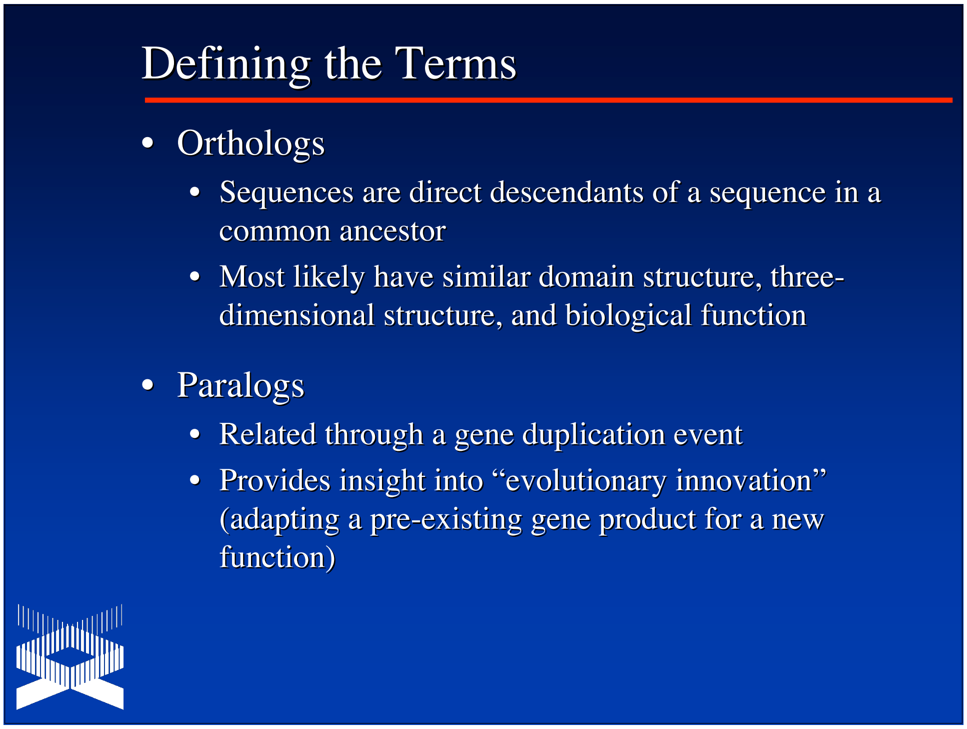
Defining the Terms

- 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 infer
  - a common evolutionary history
  - possible commonality in biological function



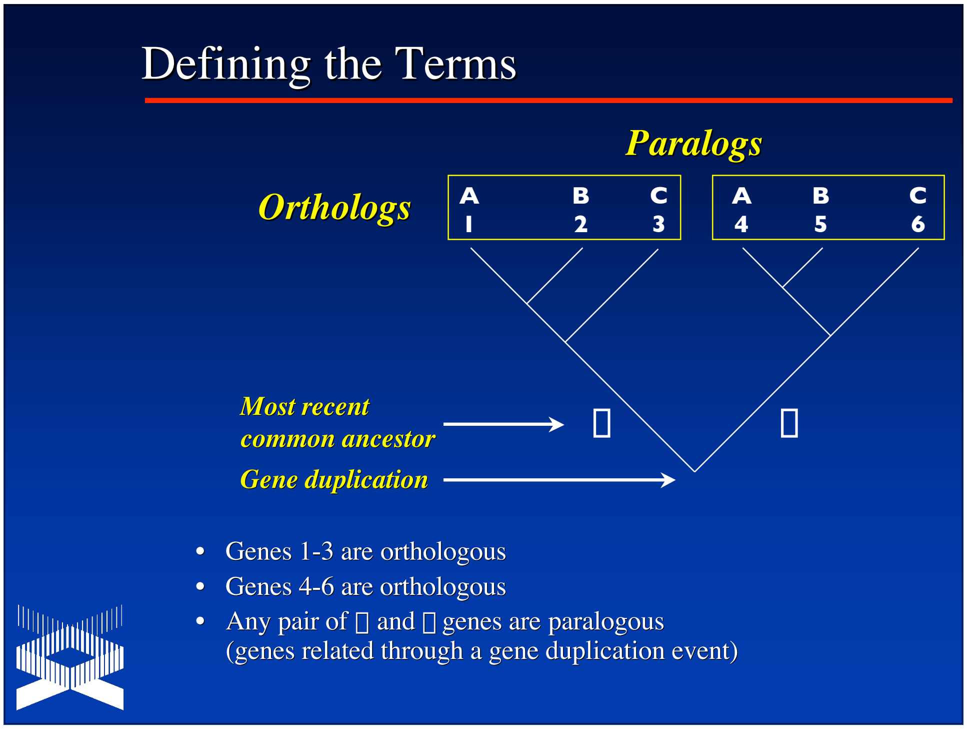
Defining the Terms

- 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

Paralogs

Orthologs

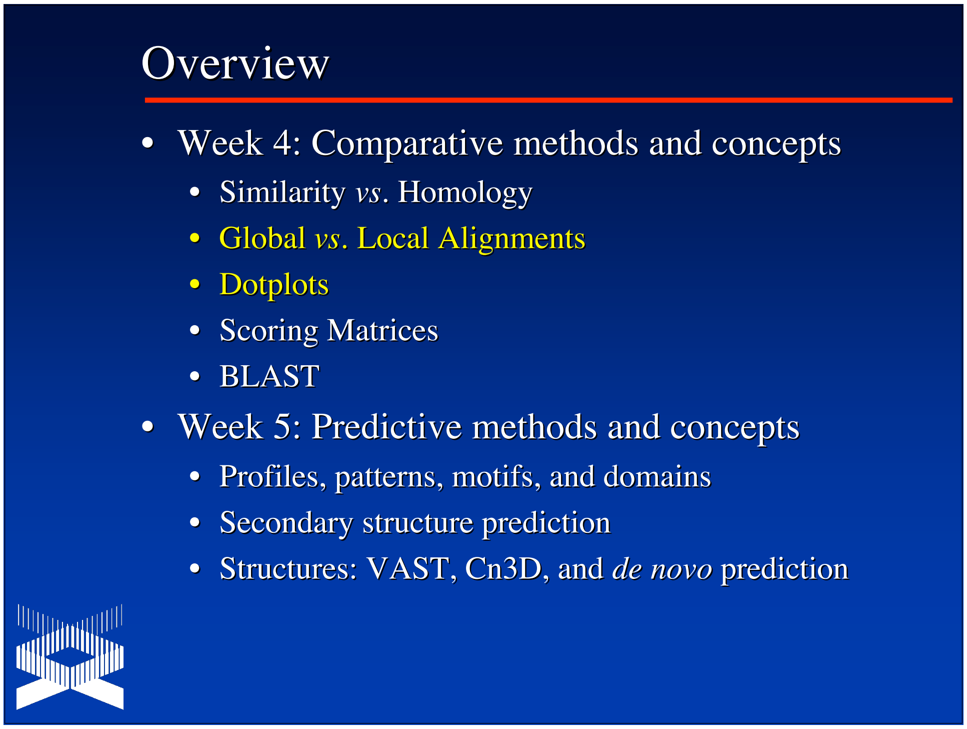
A 1 | B 2 | C 3

A 4 | B 5 | C 6

Most recent common ancestor → □ □

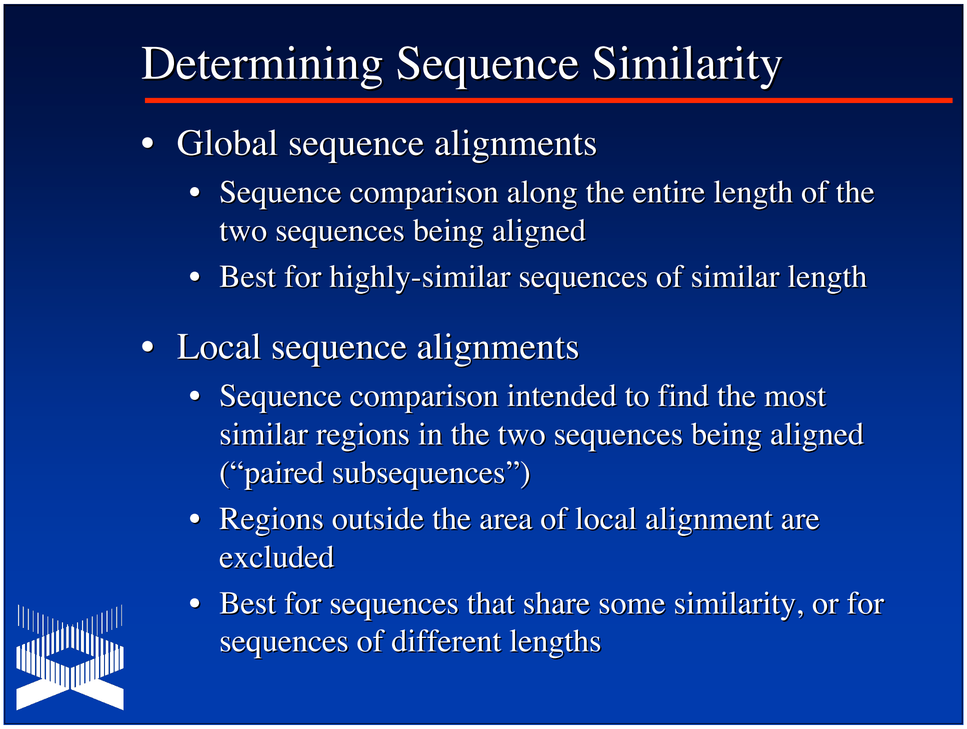
Gene duplication →

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



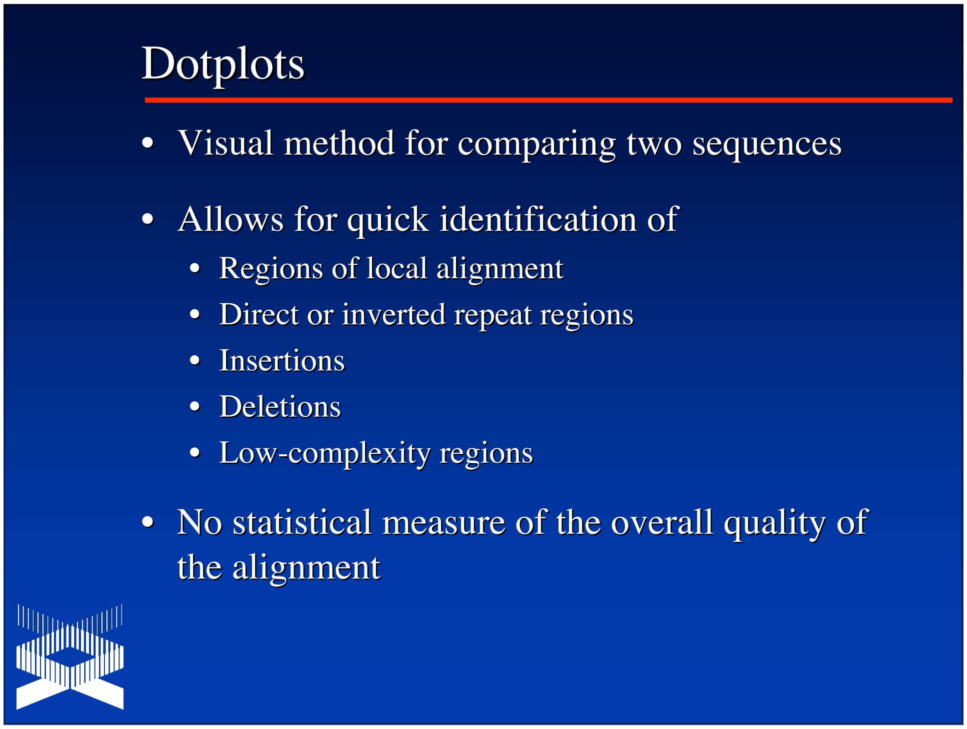
Overview

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



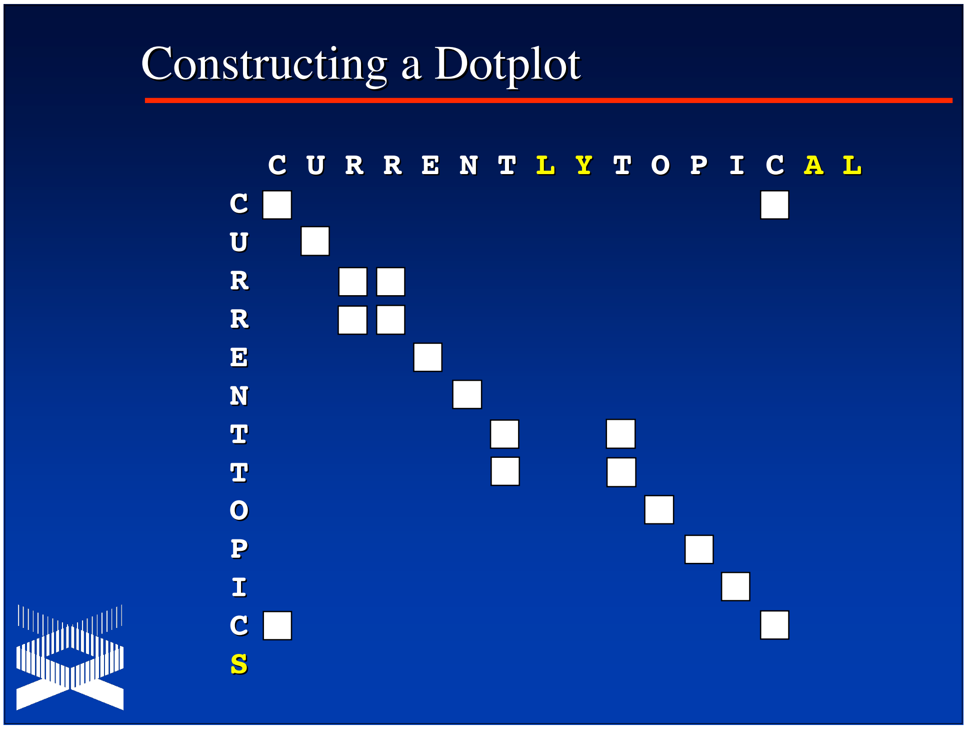
Determining Sequence Similarity

- Global sequence alignments
  - Sequence comparison along the entire length of the two sequences being aligned
  - Best for highly-similar sequences of similar length
- 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
  - Best for sequences that share some similarity, or for sequences of different lengths



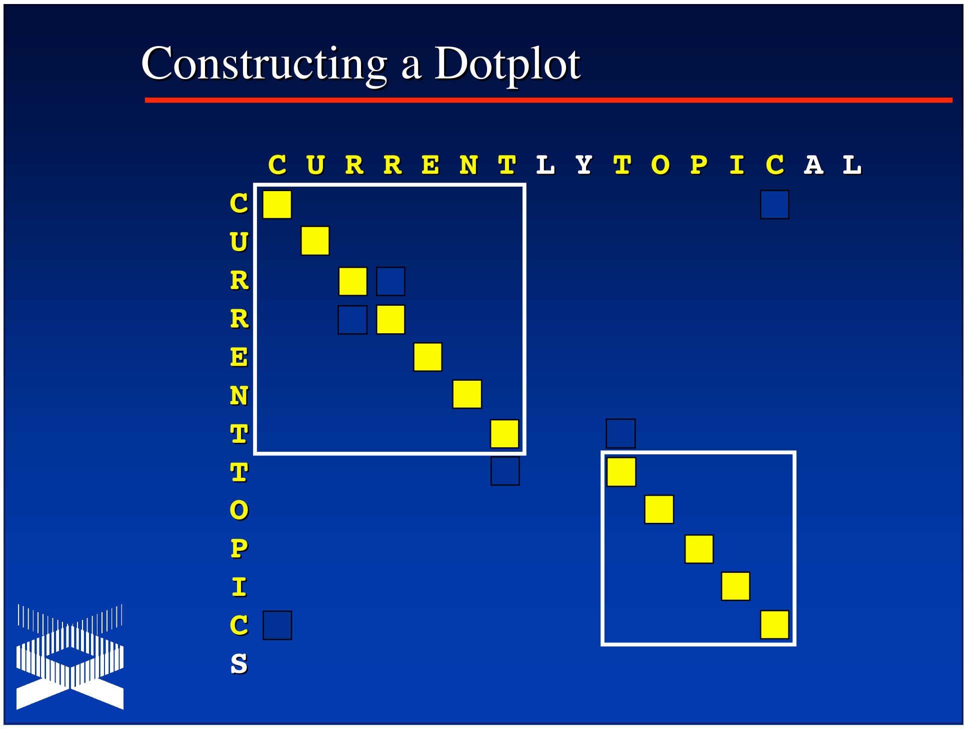
Dotplots

- Visual method for comparing two sequences
- Allows for quick identification of
  - Regions of local alignment
  - Direct or inverted repeat regions
  - Insertions
  - Deletions
  - Low-complexity regions
- No statistical measure of the overall quality of the alignment



Constructing a Dotplot

	C	U	R	R	E	N	T	L	Y	T	O	P	I	C	A	L
C	□													□		
U		□														
R			□	□												
R			□	□												
E					□											
N						□										
T							□			□						
T							□			□						
O											□					
P												□				
I													□			
C	□													□		
S																



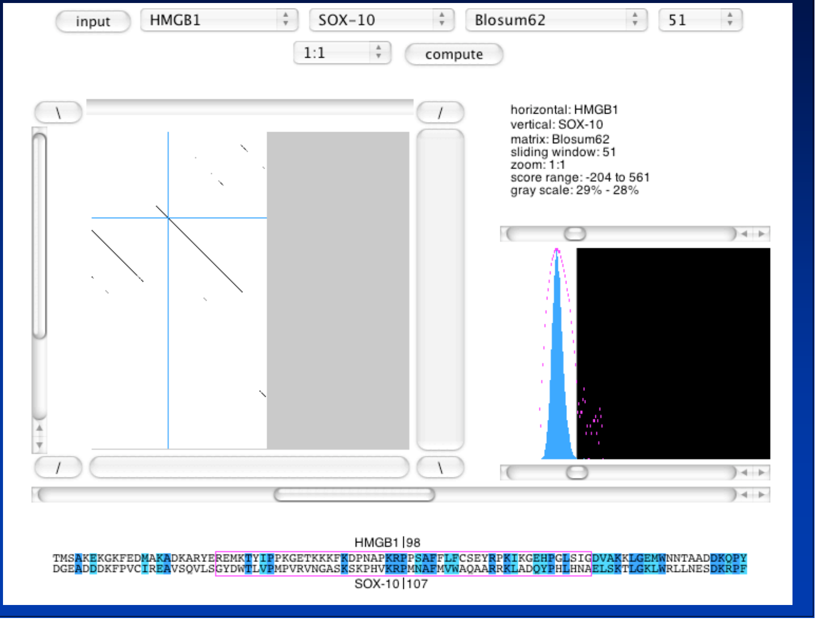
Constructing a Dotplot

	C	U	R	R	E	N	T	L	Y	T	O	P	I	C	A	L
C	■													□		
U		■														
R			■	□												
R			□	■												
E					■											
N						■										
T							■			□						
T							□			■						
O											■					
P												■				
I													■			
C	□													■		
S																

Tools for Constructing Dotplots

- Dotlet (Java applet)
  http://www.isrec.isb-sib.ch/java/dotlet/Dotlet.html
- Dotter
  http://www.cgr.ki.se/cgr/groups/sonhammer/Dotter.html
- Dottup (for complete genomes)
  http://www.emboss.org
- Dotplot subroutines also available through several software suites (GCG, DNA Strider)

Finding Regions of Local Alignment



horizontal: HMGB1
vertical: SOX-10
matrix: Blosum62
sliding window: 51
zoom: 1:1
score range: -204 to 561
gray scale: 29% - 28%

```
                                 HMGB1 |98
TMSAKEKGKFEDMAKADKARYEREMKTYIPPKGETKKKFKDPNAPKRPPSAFFLFCSEYRPKIKGEHPGLSIGDVAKKLGEMWNNTAADDKQPY
DGEADDDKFPVCIREAVSQVLSGYDWTLVPMPVRVNGASKSKPHVKRPMNAFMVWAQAARRKLADQYPHLHNAELSKTLGKLWRLLNESDKRPF
                                 SOX-10 |107
```

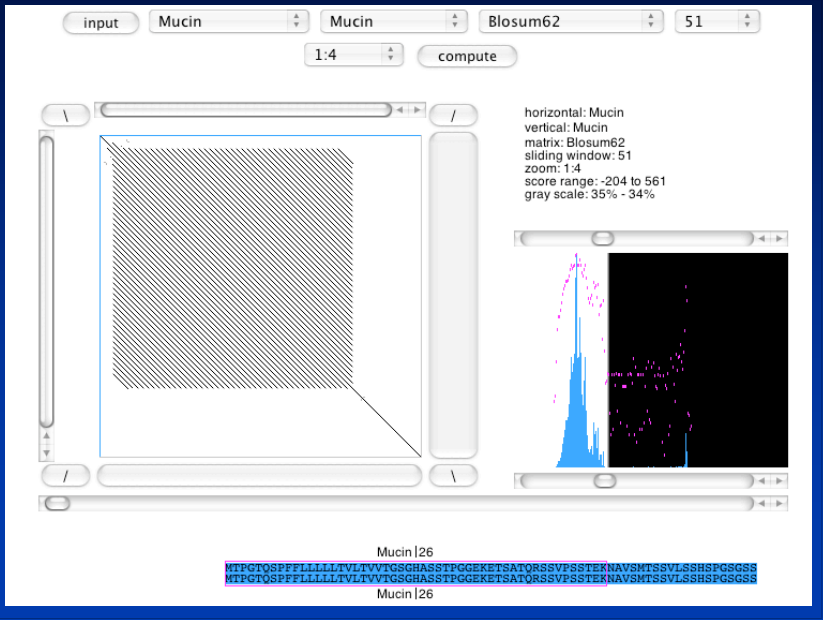
Identifying Repeats

```
>gi|189599|gb|AAA60019.1| mucin
MTPGTQSPFFLLLLLTVLTVVTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHSPGSGSSTTQGQDVTL
APATEPASGSAATWGQDVTSVPVTRPALGSTTPPAHDVTSAPDNKPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTS
APDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTS
APDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTS
APDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTS
APDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTS
APDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTS
APDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTS
APDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTS
APDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTS
APDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTS
APDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDNRPALGSTAPPVHNVTS
ASGSASGSASTLVHNGTSARATTTPASKSTPFSIPSHHSDTPTTLASHSTKTDASSTHHSSVPPLTSSNHSTSPQLSTGV
SFFFLSFHISNLQFNSSLEDPSTDYYQELQRDISEMFLQIYKQGGFLGLSNIKFRPGSVVVQLTLAFREGTINVHDVETQ
FNQYKTEAASRYNLTISDVSVSDVPFPFSAQSGAGVPGWGIALLVLVCVLVALAIVYLIALAVCQCRRKNYGQLDIFPAR
DTYHPMSEYPTYHTHGRYVPPSSTDRSPYEKVSAGNGGSSLSYTNPAVAAASANL
```

PAPGSTAPPAHGVTSAPDTR

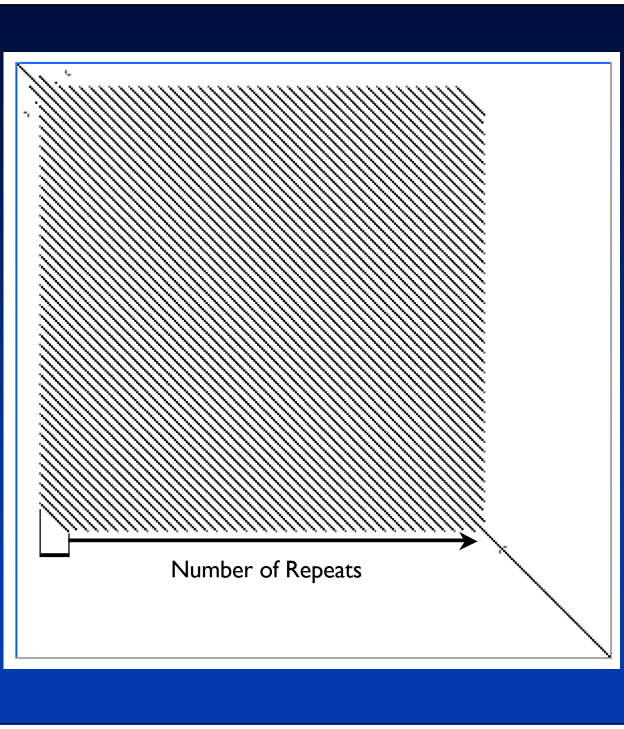
40 tandem repeats of 20 amino acids

Identifying Repeats



horizontal: Mucin
vertical: Mucin
matrix: Blosum62
sliding window: 51
zoom: 1:4
score range: -204 to 561
gray scale: 35% - 34%

Mucin|26
MTPGTQSPFFLLLLLTVLTVVTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHSPGSGSS
MTPGTQSPFFLLLLLTVLTVVTGSGHASSTPGGEKETSATQRSSVPSSTEKNAVSMTSSVLSSHSPGSGSS
Mucin|26



Shortest diagonal:
Beginning and
end points of
single repeat

Identifying Low-Complexity Regions

- Regions of biased composition
  - Homopolymeric runs
  - Short-period repeats
  - Subtle over-representation of several residues
- 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)

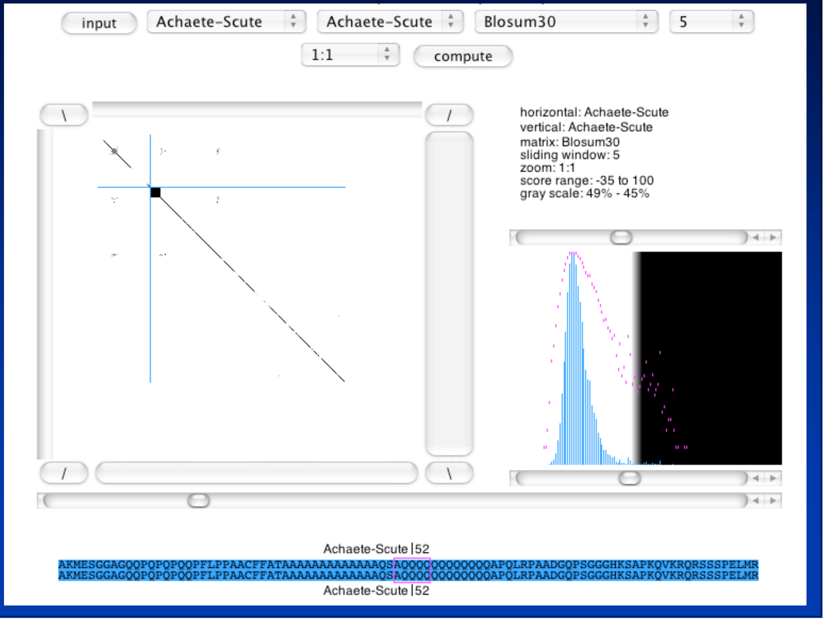
Identifying Low-Complexity Regions

Example: *Drosophila* achaete-scute

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

Homopolymeric alanine-glutamine tract

Identifying Low-Complexity Regions

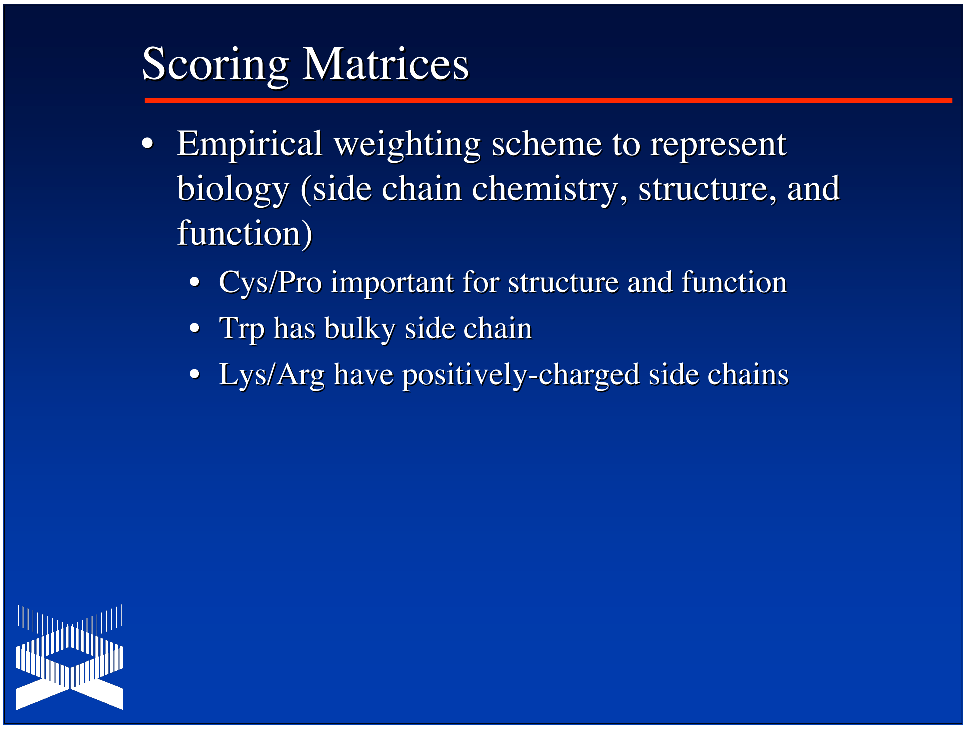


horizontal: Achaete-Scute
vertical: Achaete-Scute
matrix: Blosum30
sliding window: 5
zoom: 1:1
score range: -35 to 100
gray scale: 49% - 45%

Achaete-Scute|52

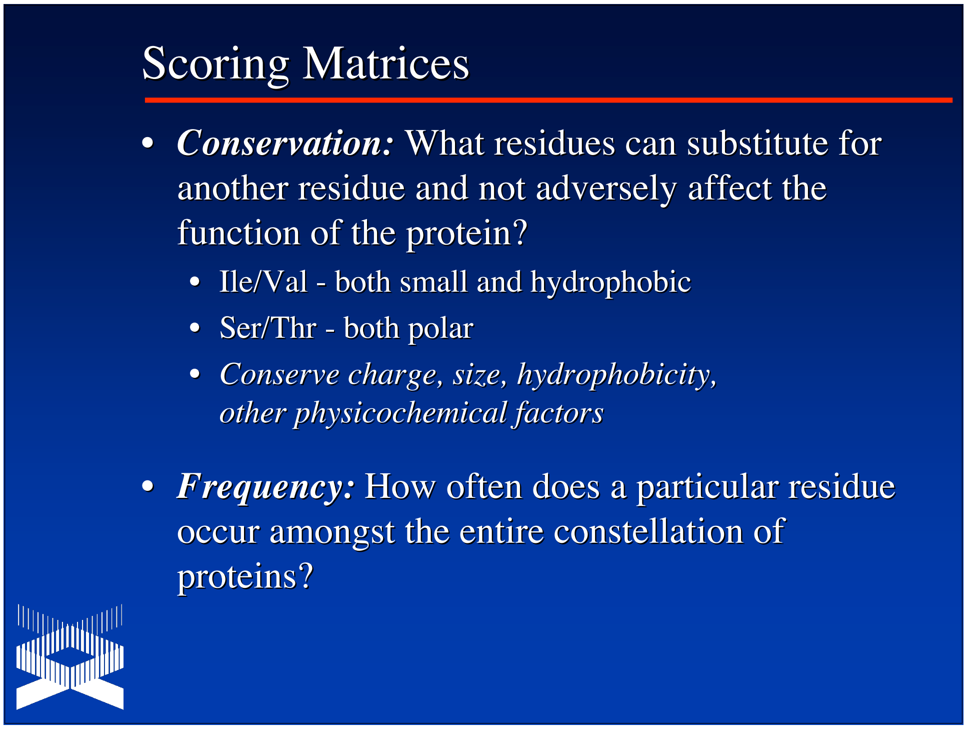
```
AKMESGGAGQQPQPQPQQPFLPPAACFFATAAAAAAAAAAAAAQSAQQQQQQQQQQQQQAPQLRPAADGQPSGGGHKSAPKQVKRQRSSSPELMR
AKMESGGAGQQPQPQPQQPFLPPAACFFATAAAAAAAAAAAAAQSAQQQQQQQQQQQQQAPQLRPAADGQPSGGGHKSAPKQVKRQRSSSPELMR
```

Achaete-Scute|52



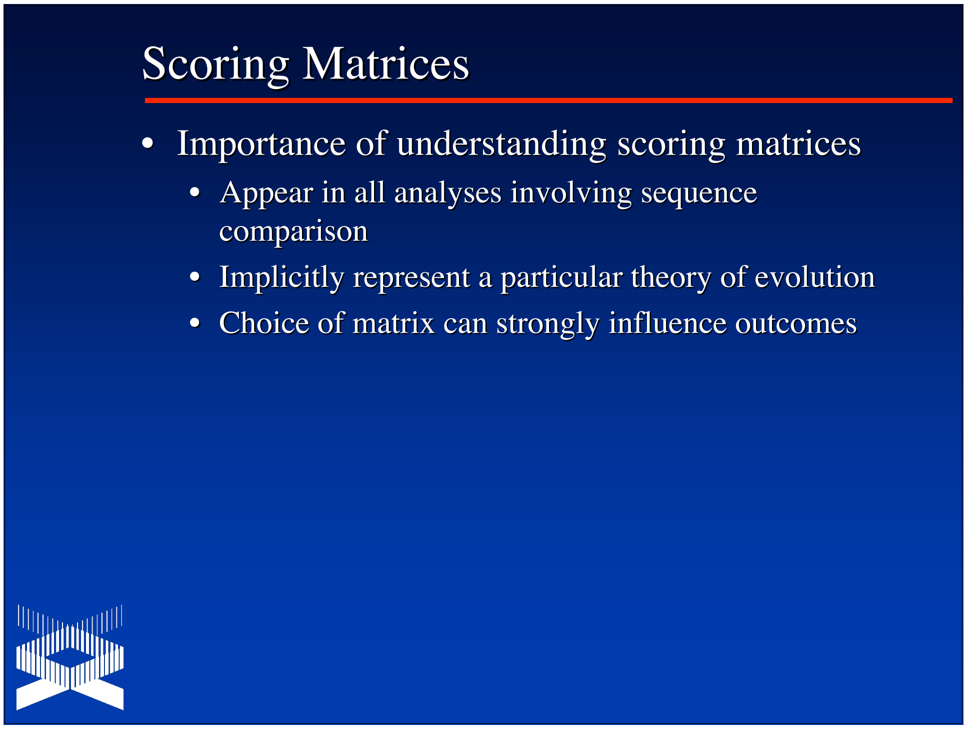
Scoring Matrices

- 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



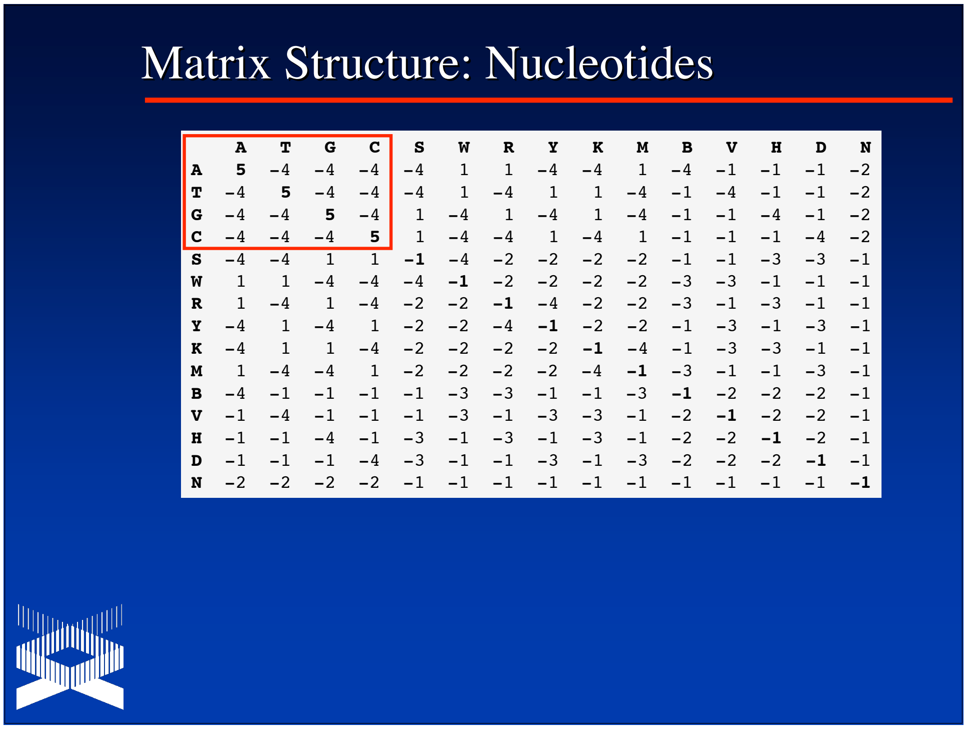
Scoring Matrices

- ***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

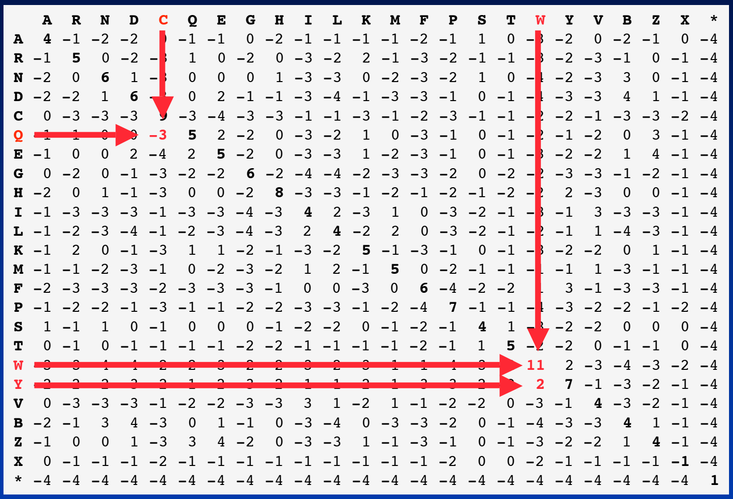
- Importance of understanding scoring matrices
  - Appear in all analyses involving sequence comparison
  - Implicitly represent a particular theory of evolution
  - Choice of matrix can strongly influence outcomes



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**

Matrix Structure: Proteins

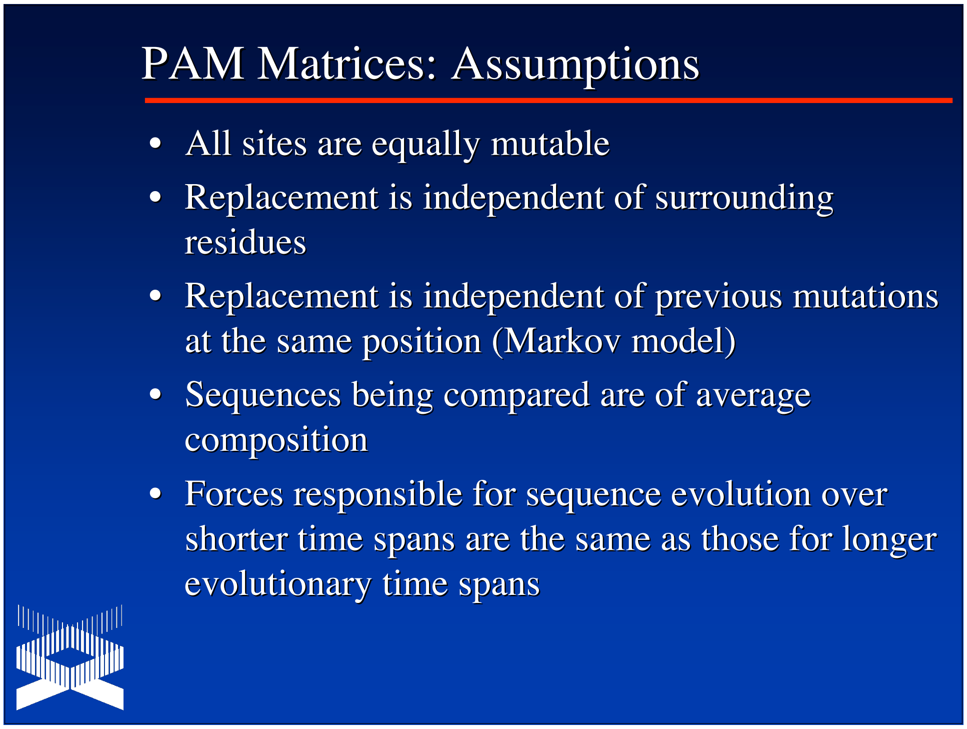


	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	**4**	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0	-4
R	-1	**5**	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1	-4
N	-2	0	**6**	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1	-4
D	-2	-2	1	**6**	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1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

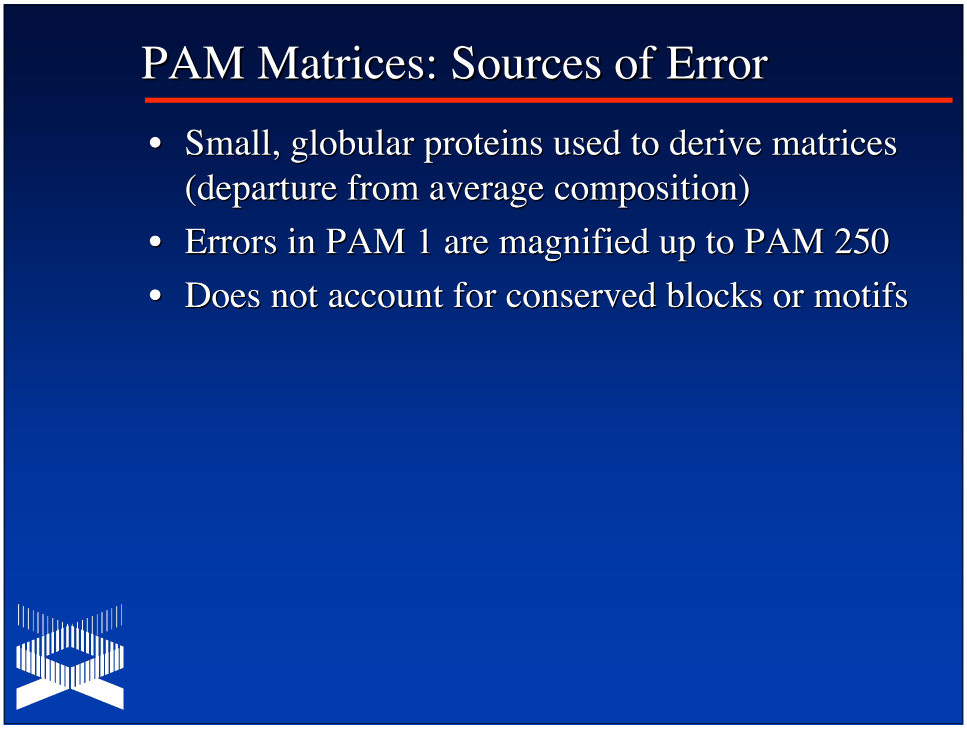
PAM Matrices

- Margaret Dayhoff, 1978
- Point Accepted Mutation (PAM)
  - Look at patterns of substitutions in highly related proteins (> 85% similar), based on multiple sequence alignments
  - 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 are equally mutable
- Replacement is independent of surrounding residues
- Replacement is independent of previous mutations at the same position (Markov model)
- Sequences being compared are of average composition
- Forces responsible for sequence evolution over shorter time spans are the same as those for longer evolutionary time spans



PAM Matrices: Sources of Error

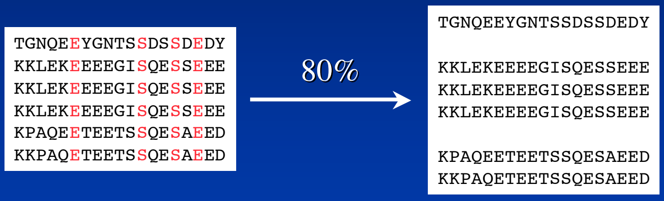
- Small, globular proteins used to derive matrices (departure from average composition)
- Errors in PAM 1 are magnified up to PAM 250
- Does not account for conserved blocks or motifs

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 structural or functional substitutions
  - Generally perform better than PAM matrices for local similarity searches *(Henikoff and Henikoff, 1993)*

BLOSUM *n*

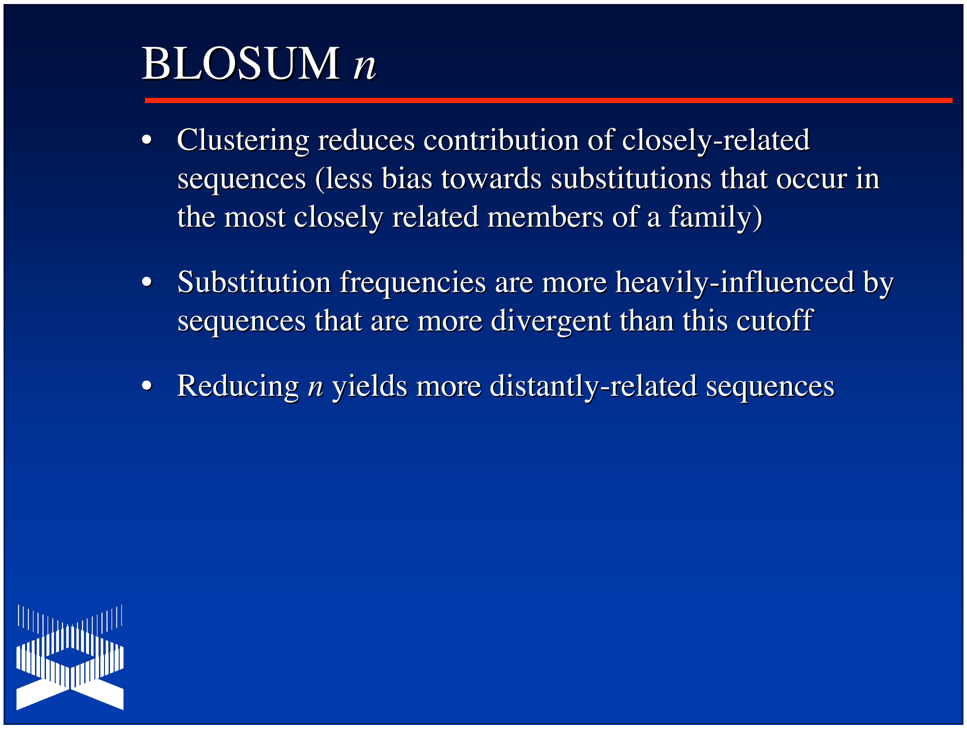
- Calculated from sequences sharing no more than *n*% identity
- Contribution of sequences > *n*% identical clustered and weighted to 1



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

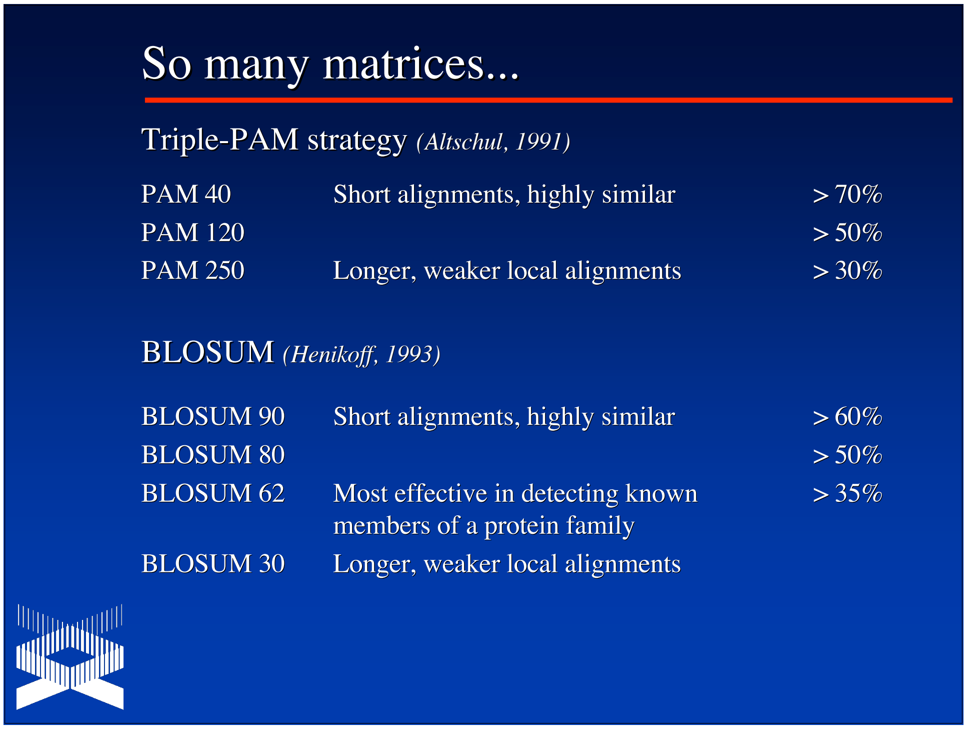
```
TGNQEEYGNTSSDSSDEDY

KKLEKEEEEGISQESSEEE
KKLEKEEEEGISQESSEEE
KKLEKEEEEGISQESSEEE

KPAQEETEETSSQESAEED
KKPAQETEETSSQESAEED
```

A+T Hook Domain (Block IPB000637B)



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



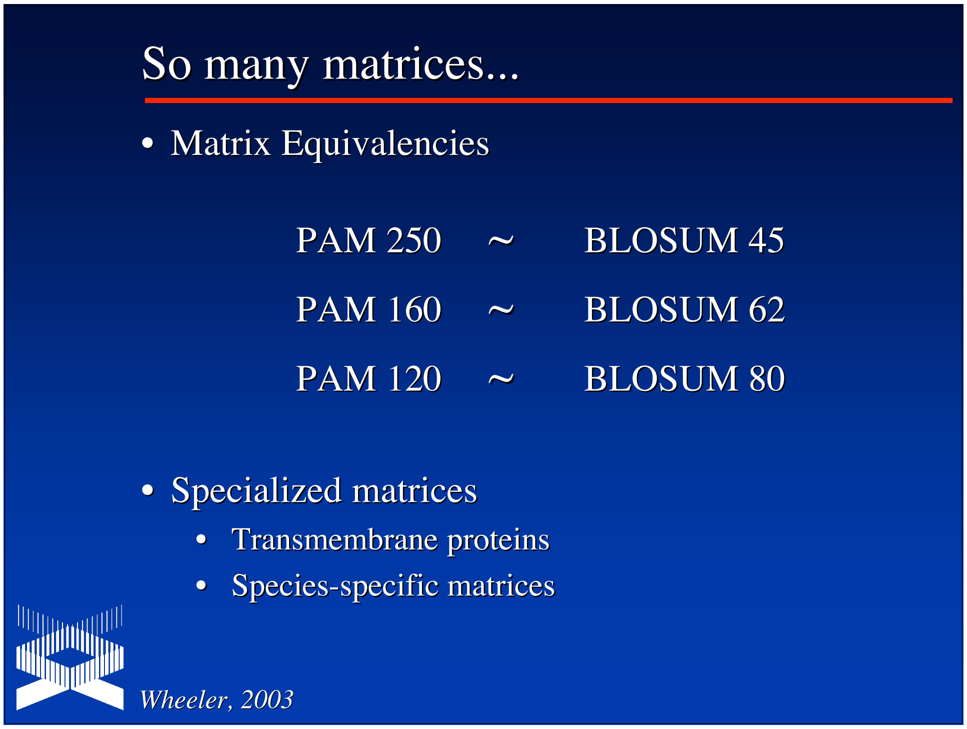
So many matrices...

Triple-PAM strategy *(Altschul, 1991)*

PAM 40	Short alignments, highly similar	> 70%
PAM 120		> 50%
PAM 250	Longer, weaker local alignments	> 30%

BLOSUM *(Henikoff, 1993)*

BLOSUM 90	Short alignments, highly similar	> 60%
BLOSUM 80		> 50%
BLOSUM 62	Most effective in detecting known members of a protein family	> 35%
BLOSUM 30	Longer, weaker local alignments	



So many matrices...

- Matrix Equivalencies

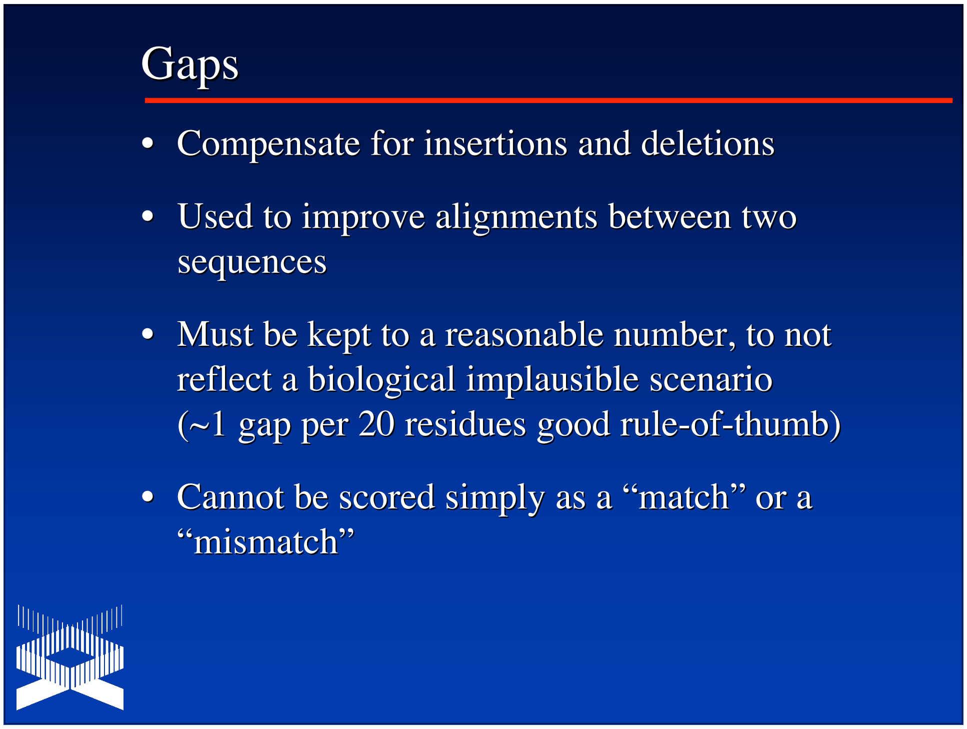
PAM 250	~	BLOSUM 45
PAM 160	~	BLOSUM 62
PAM 120	~	BLOSUM 80

- Specialized matrices
  - Transmembrane proteins
  - Species-specific matrices

Wheeler, 2003

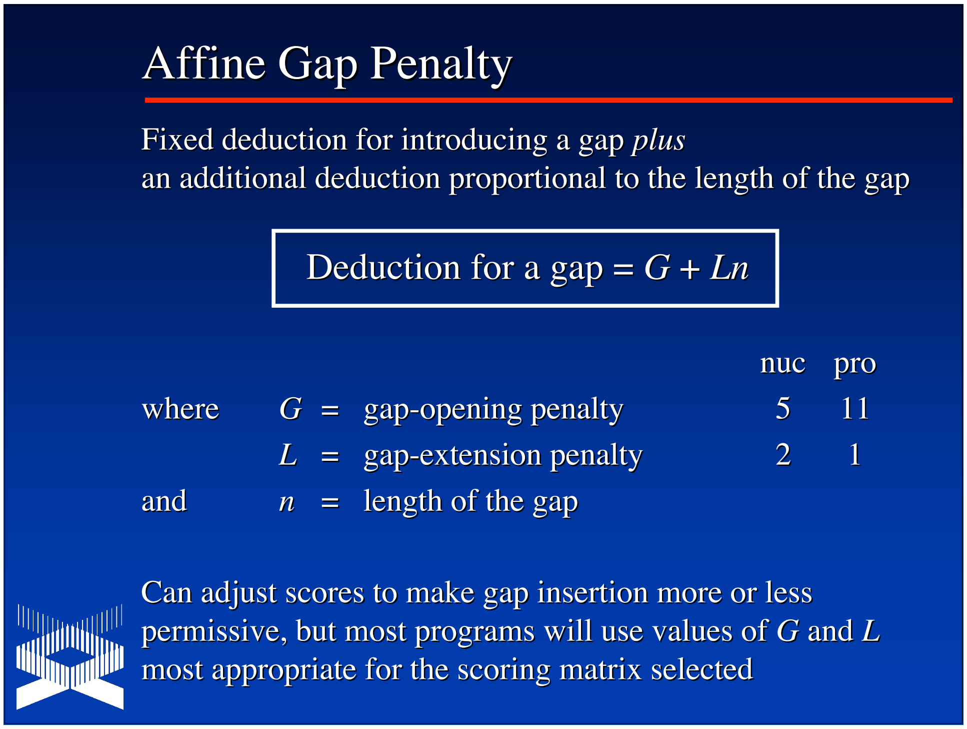
So many matrices...

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"



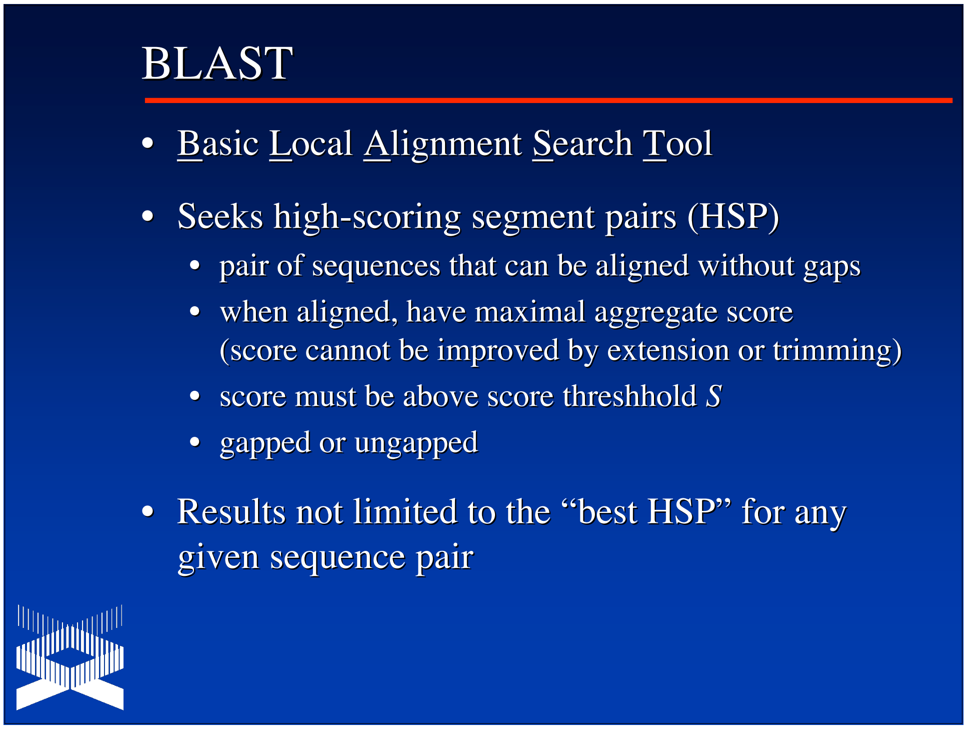
Affine Gap Penalty

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

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

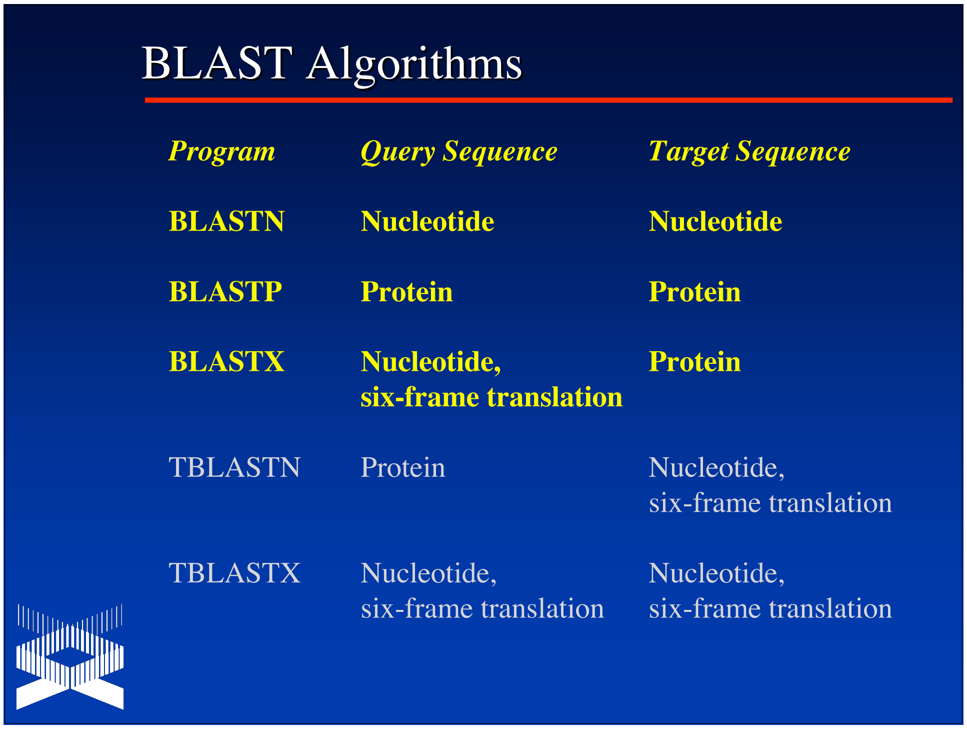
				nuc	pro
where	G	=	gap-opening penalty	5	11
	L	=	gap-extension penalty	2	1
and	n	=	length of the gap		

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



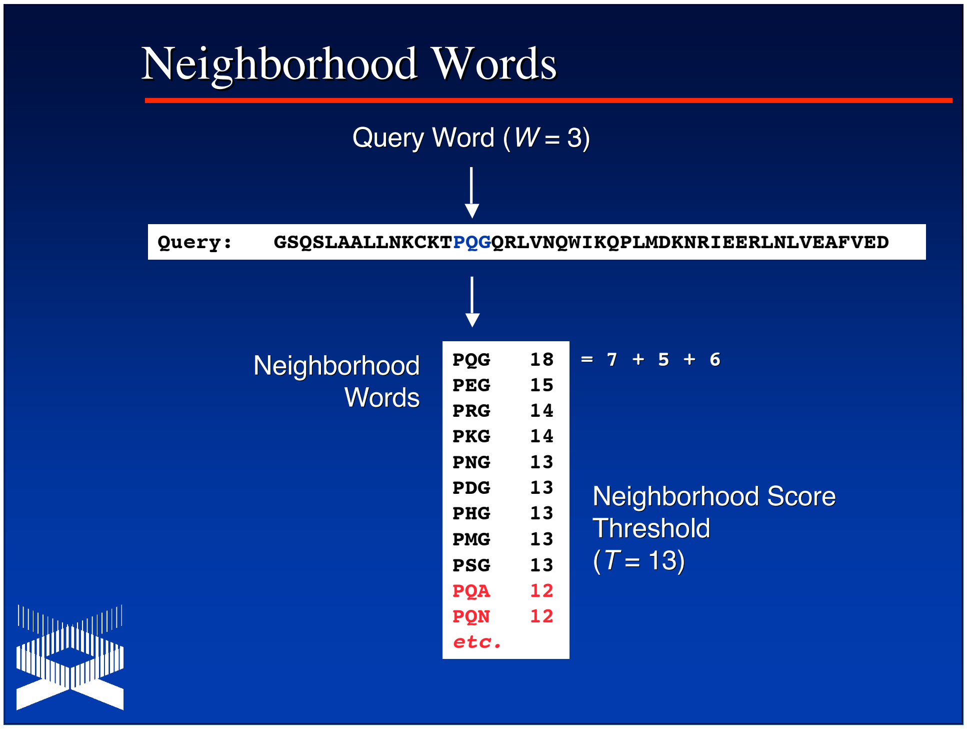
BLAST

- 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

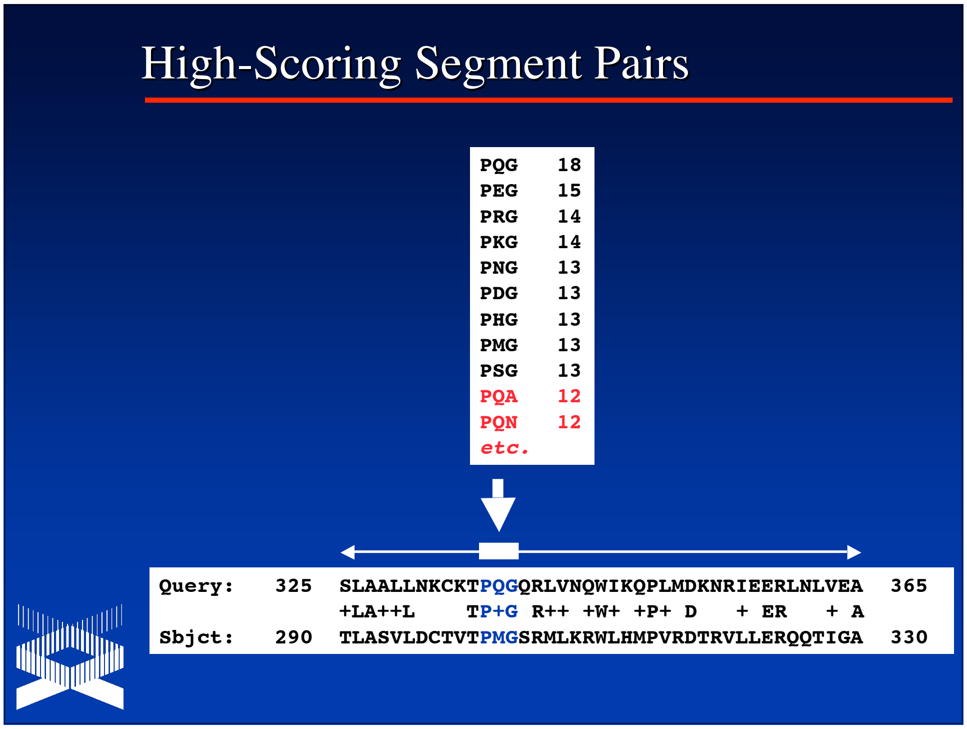
Query Word (*W* = 3)

```
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.		

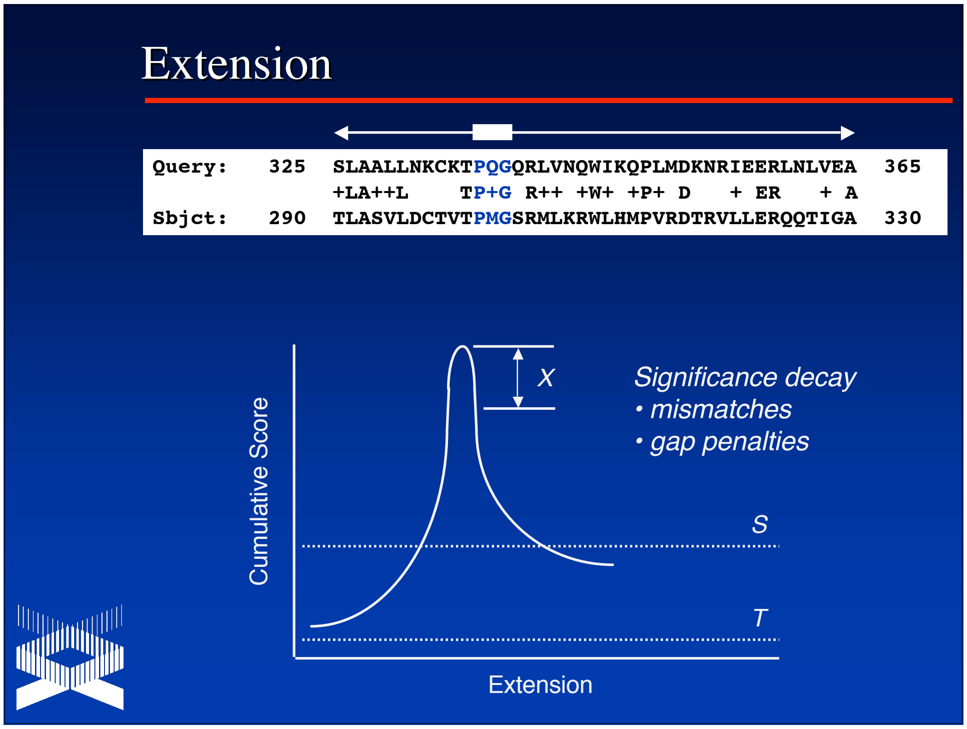
Neighborhood Score Threshold (*T* = 13)



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.	

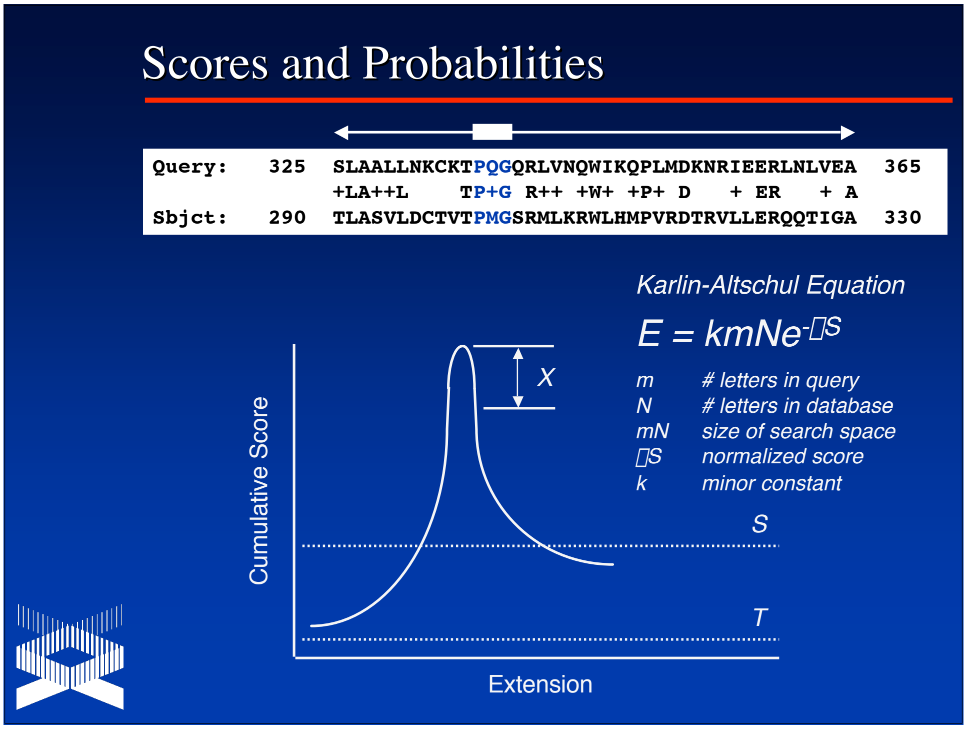
```
Query:  325  SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA  365
             +LA++L    TP+G R++ +W+ +P+ D   + ER   + A
Sbjct:  290  TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA  330
```



Extension

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

Significance decay
- *mismatches*
- *gap penalties*



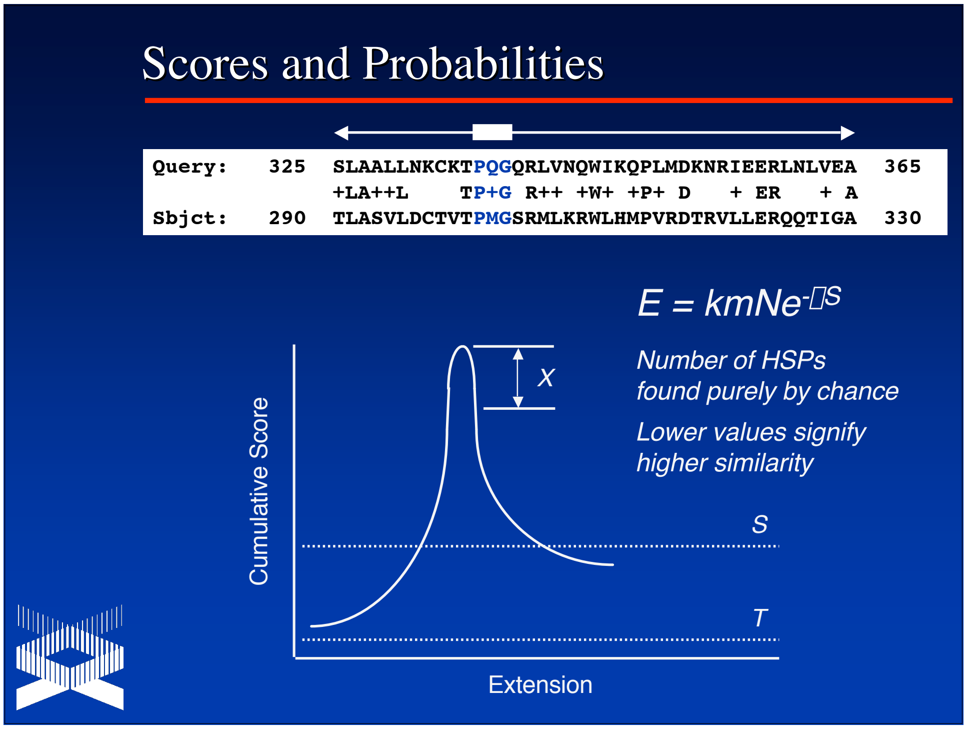
Scores and Probabilities

```
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*



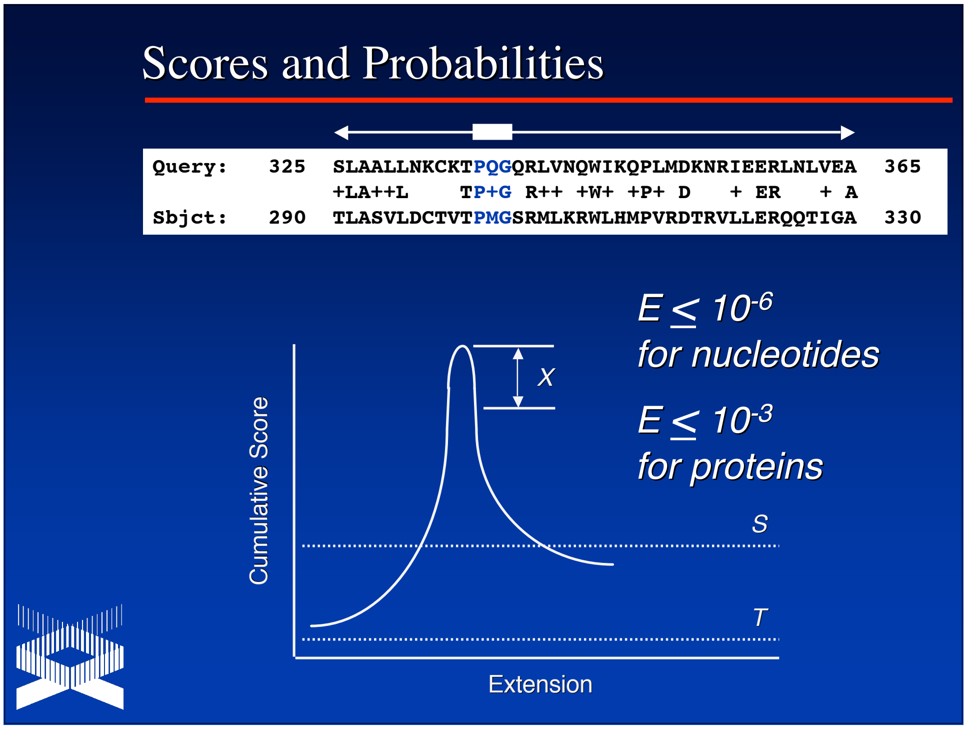
Scores and Probabilities

```
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

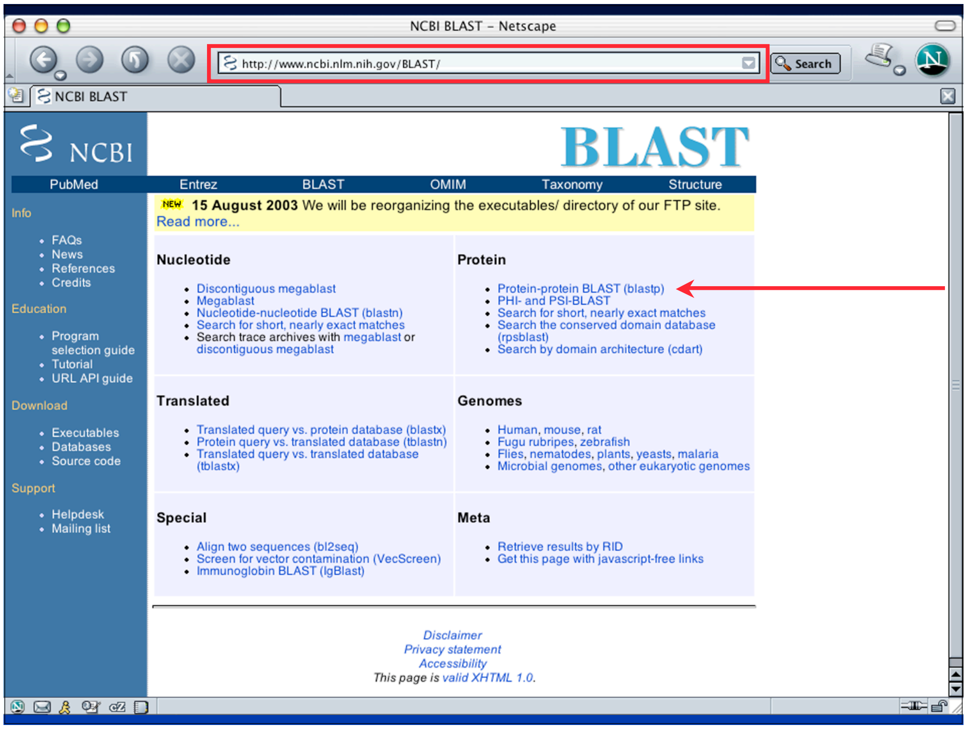


Scores and Probabilities

```
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



NEW **15 August 2003** We will be reorganizing the executables/ directory of our FTP site.
Read more...

Info

- FAQs
- News
- References
- Credits

Education

- Program selection guide
- Tutorial
- URL API guide

Download

- Executables
- Databases
- Source code

Support

- Helpdesk
- Mailing list

Nucleotide

- Discontiguous megablast
- 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

- Human, mouse, rat
- Fugu rubripes, zebrafish
- Flies, nematodes, plants, yeasts, malaria
- Microbial genomes, other eukaryotic genomes

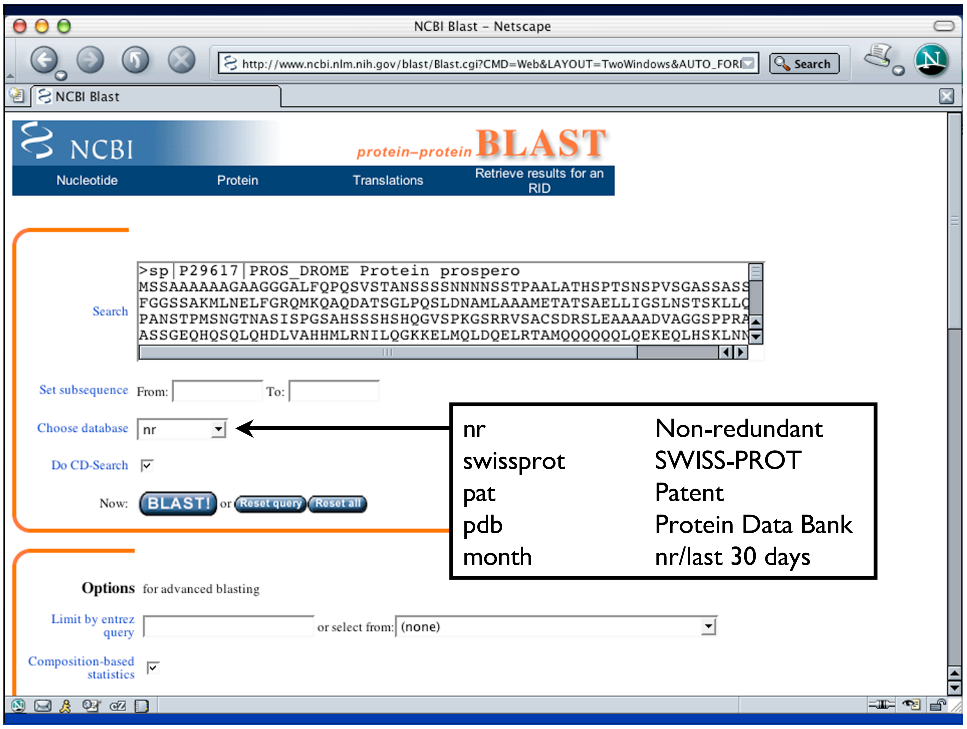
Special

- Align two sequences (bl2seq)
- Screen for vector contamination (VecScreen)
- Immunoglobin BLAST (IgBlast)

Meta

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

Disclaimer
Privacy statement
Accessibility
This page is valid XHTML 1.0.


NCBI Blast – Netscape
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?CMD=Web&LAYOUT=TwoWindows&AUTO_FORI
Search
NCBI Blast
NCBI
protein–protein BLAST
Nucleotide
Protein
Translations
Retrieve results for an RID

Search
```
>sp|P29617|PROS_DROME Protein prospero
MSSAAAAAAGAAGGGALFQPQSVSTANSSSSNNNNSSTPAALATHSPTSNSPVSGASSASS
FGGSSAKMLNELFGRQMKQAQDATSGLPQSLDNAMLAAAMETATSAELLIGSLNSTSKLLQ
PANSTPMSNGTNASISPGSAHSSSHSHQGVSPKGSRRVSACSDRSLEAAAADVAGGSPPRA
ASSGEQHQSQLQHDLVAHHMLRNILQGKKELMQLDQELRTAMQQQQQQLQEKEQLHSKLNN
```

Set subsequence From: To:

Choose database nr

Do CD-Search ☑

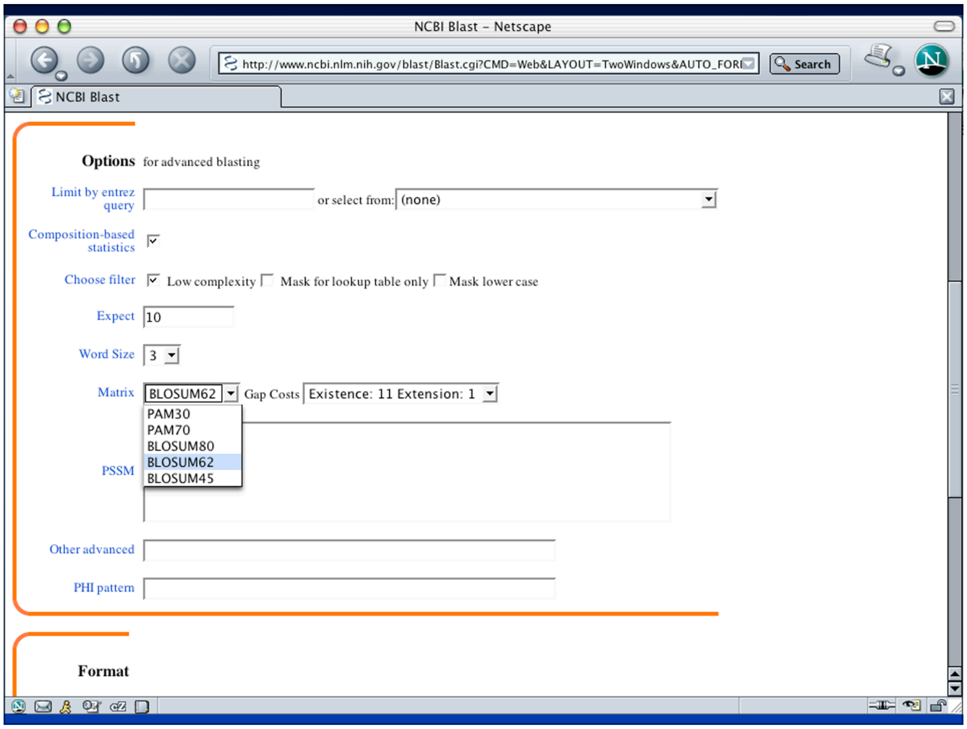
Now: BLAST! or Reset query Reset all

nr	Non-redundant
swissprot	SWISS-PROT
pat	Patent
pdb	Protein Data Bank
month	nr/last 30 days

Options for advanced blasting

Limit by entrez query or select from: (none)

Composition-based statistics ☑



NCBI Blast – Netscape
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?CMD=Web&LAYOUT=TwoWindows&AUTO_FORI
Search
NCBI Blast

Options for advanced blasting

Limit by entrez query: ______ or select from: (none)

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

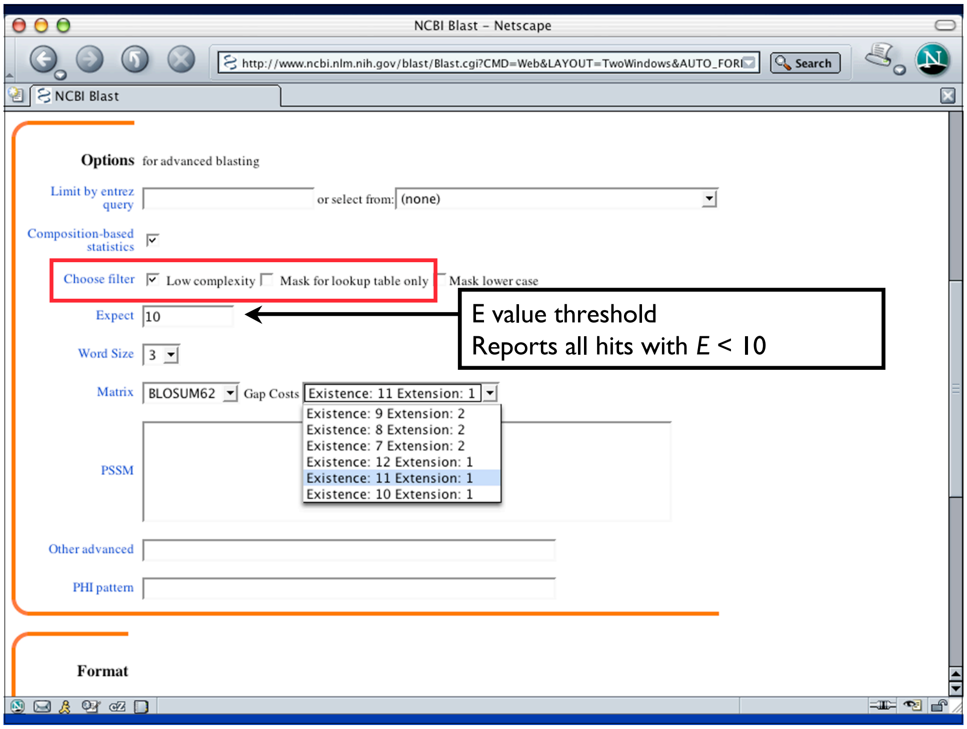
PAM30
PAM70
BLOSUM80
BLOSUM62
BLOSUM45

PSSM:

Other advanced:

PHI pattern:

Format



NCBI Blast – Netscape
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?CMD=Web&LAYOUT=TwoWindows&AUTO_FORI
Search
NCBI Blast

Options for advanced blasting

Limit by entrez query: ______ or select from: (none)

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

Existence: 9 Extension: 2
Existence: 8 Extension: 2
Existence: 7 Extension: 2
Existence: 12 Extension: 1
Existence: 11 Extension: 1
Existence: 10 Extension: 1

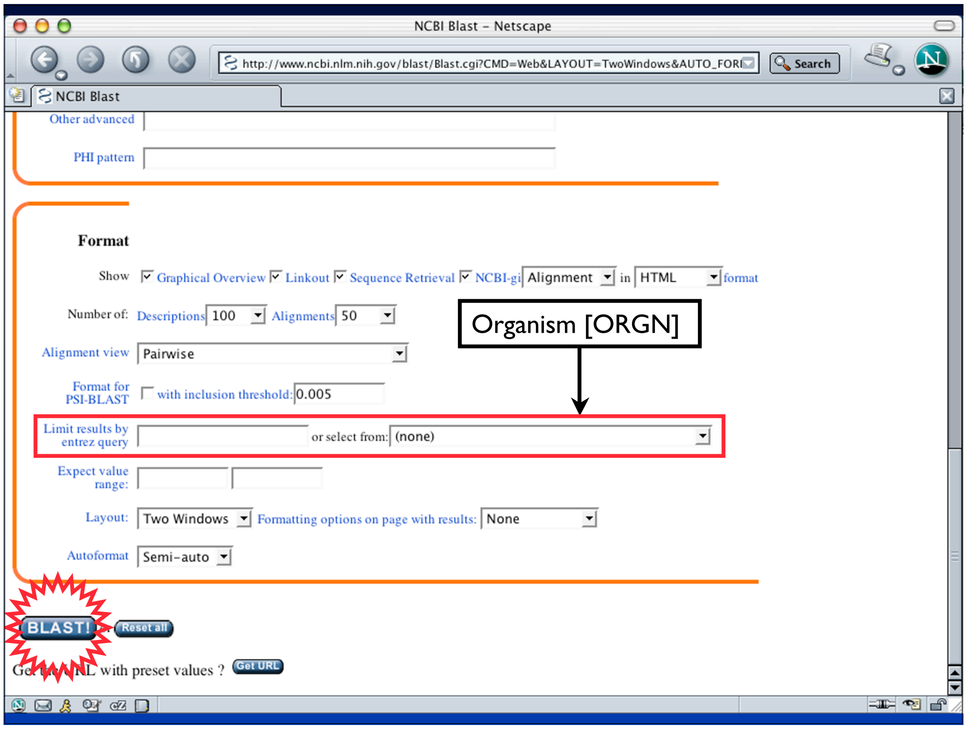
PSSM:

Other advanced:

PHI pattern:

Format

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


NCBI Blast – Netscape
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?CMD=Web&LAYOUT=TwoWindows&AUTO_FOR
Search
NCBI Blast
BLAST!
Reset all
Get URL
Organism [ORGN]

Other advanced

PHI pattern

Format

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

Number of: Descriptions 100 Alignments 50

Alignment view Pairwise

Format for PSI-BLAST □ with inclusion threshold: 0.005

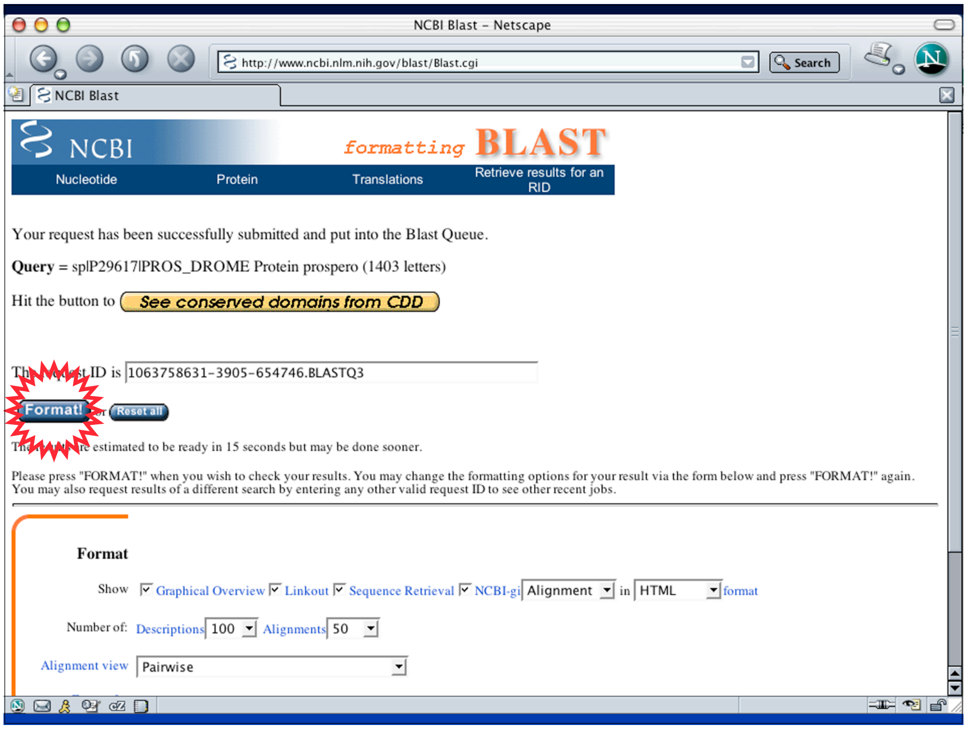
Limit results by entrez query or select from: (none)

Expect value range:

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

Autoformat Semi-auto

Get the URL with preset values ?



Your request has been successfully submitted and put into the Blast Queue.

Query = sp|P29617|PROS_DROME Protein prospero (1403 letters)

Hit the button to See conserved domains from CDD

The request ID is 1063758631-3905-654746.BLASTQ3

The results are estimated to be ready in 15 seconds but may be done sooner.

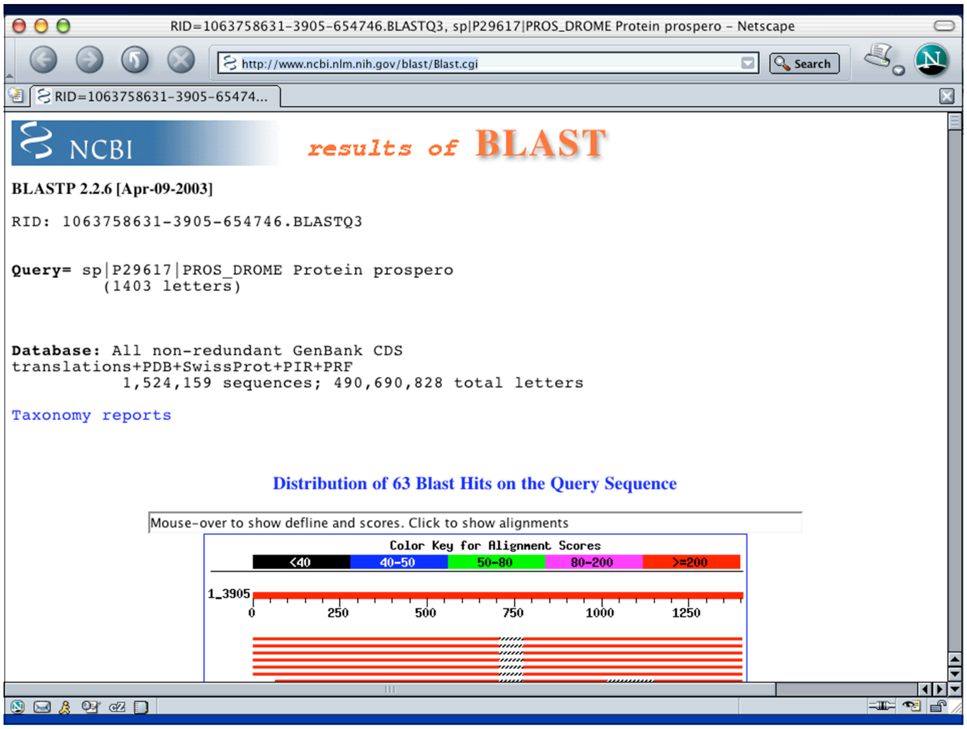
Please press "FORMAT!" when you wish to check your results. You may change the formatting options for your result via the form below and press "FORMAT!" again. You may also request results of a different search by entering any other valid request ID to see other recent jobs.

Format

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

Number of: Descriptions 100 Alignments 50

Alignment view Pairwise



RID=1063758631-3905-654746.BLASTQ3, sp|P29617|PROS_DROME Protein prospero - Netscape

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

RID=1063758631-3905-65474...

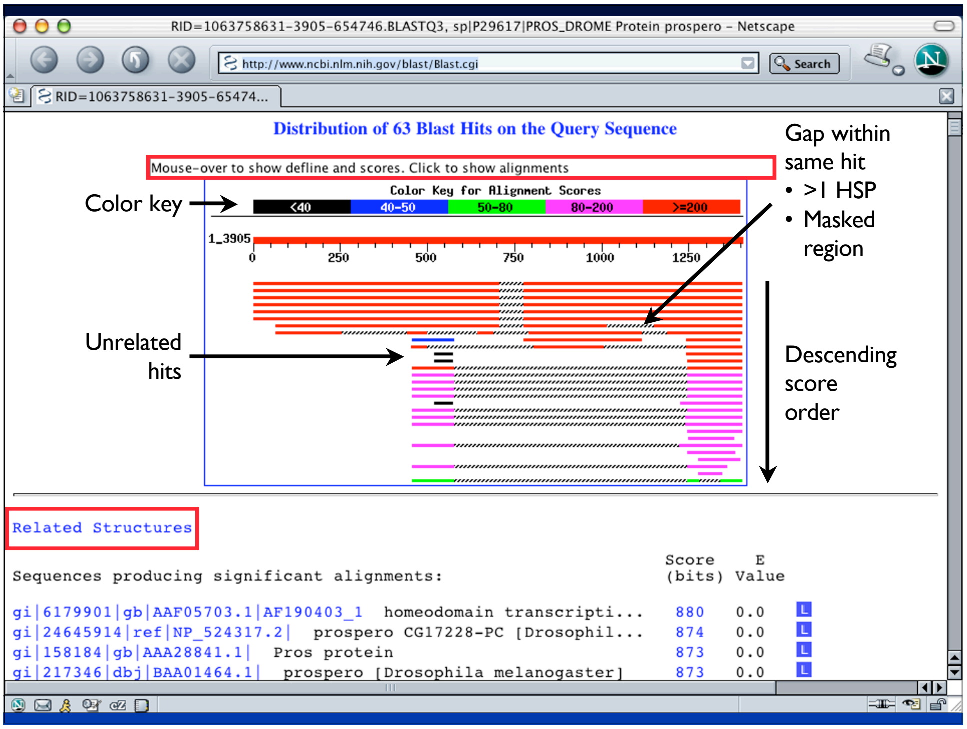
NCBI results of BLAST

BLASTP 2.2.6 [Apr-09-2003]

```
RID: 1063758631-3905-654746.BLASTQ3


Query= sp|P29617|PROS_DROME Protein prospero
         (1403 letters)


Database: All non-redundant GenBank CDS
translations+PDB+SwissProt+PIR+PRF
           1,524,159 sequences; 490,690,828 total letters

Taxonomy reports
```

Distribution of 63 Blast Hits on the Query Sequence

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



RID=1063758631-3905-654746.BLASTQ3, sp|P29617|PROS_DROME Protein prospero - Netscape

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

RID=1063758631-3905-65474...

Distribution of 63 Blast Hits on the Query Sequence

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

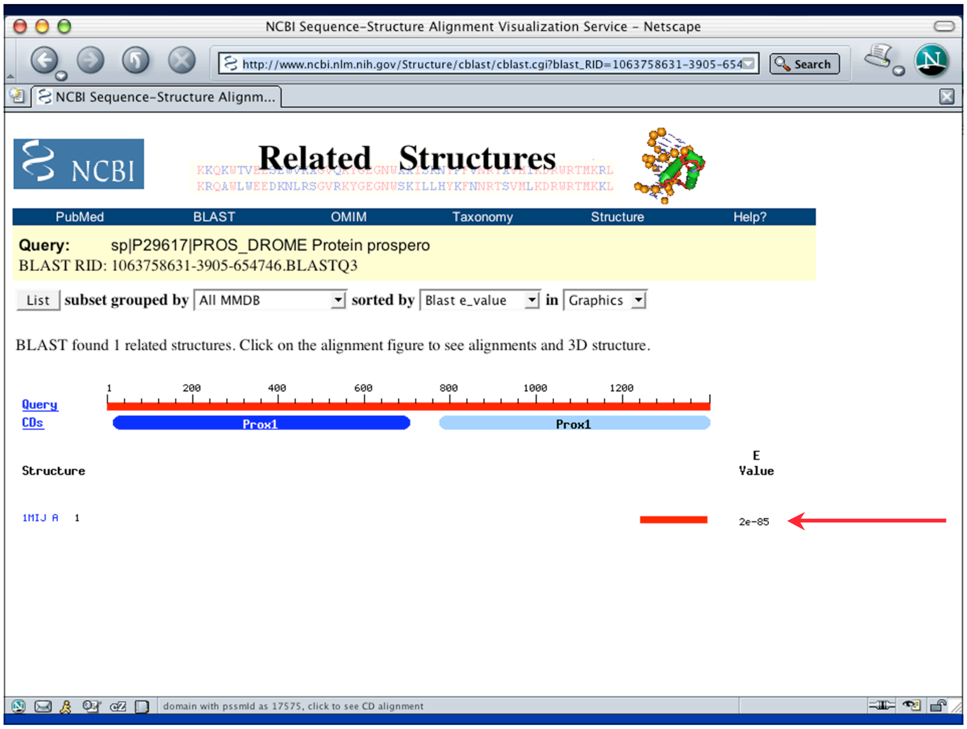
Color key

Unrelated hits

Gap within same hit
- >1 HSP
- Masked region

Descending score order

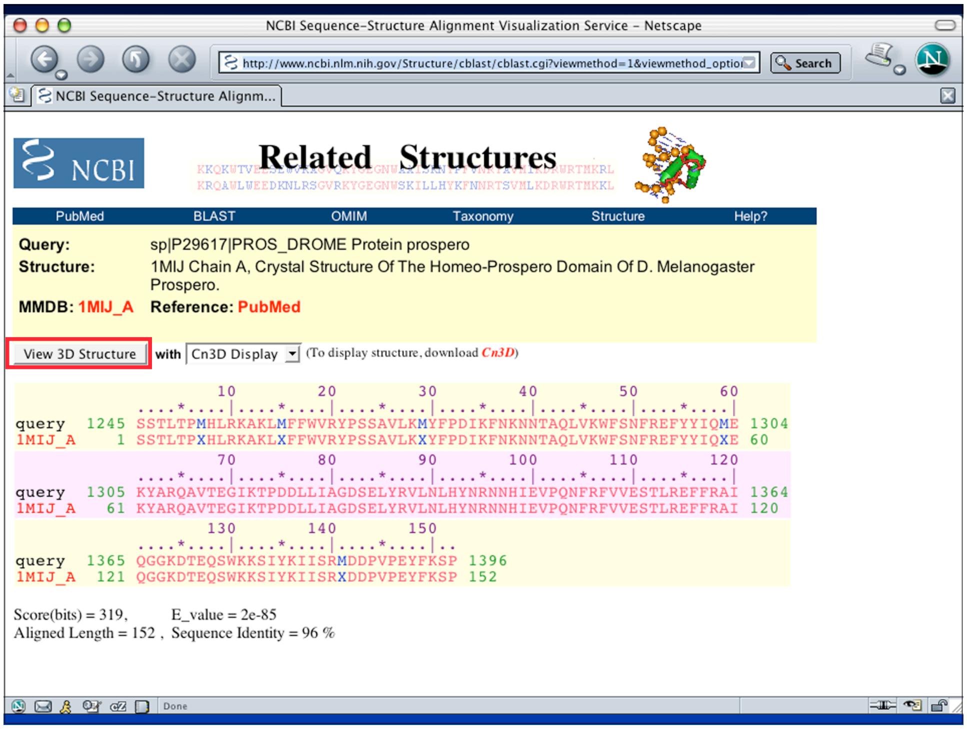
```
Related Structures

                                                        Score    E
Sequences producing significant alignments:             (bits) Value

gi|6179901|gb|AAF05703.1|AF190403_1  homeodomain transcripti...   880    0.0  L
gi|24645914|ref|NP_524317.2|  prospero CG17228-PC [Drosophil...   874    0.0  L
gi|158184|gb|AAA28841.1|  Pros protein                            873    0.0  L
gi|217346|dbj|BAA01464.1|  prospero [Drosophila melanogaster]     873    0.0  L
```


NCBI Sequence-Structure Alignment Visualization Service - Netscape
http://www.ncbi.nlm.nih.gov/Structure/cblast/cblast.cgi?blast_RID=1063758631-3905-654
Search
NCBI Sequence-Structure Alignm...
NCBI
Related Structures
PubMed
BLAST
OMIM
Taxonomy
Structure
Help?
1
200
400
600
800
1000
1200
Query
CDs
Prox1
Prox1
Structure
E Value
1MIJ A
1
2e-85
domain with pssmId as 17575, click to see CD alignment

Query: sp|P29617|PROS_DROME Protein prospero
BLAST RID: 1063758631-3905-654746.BLASTQ3

List **subset grouped by** All MMDB **sorted by** Blast e_value **in** Graphics

BLAST found 1 related structures. Click on the alignment figure to see alignments and 3D structure.

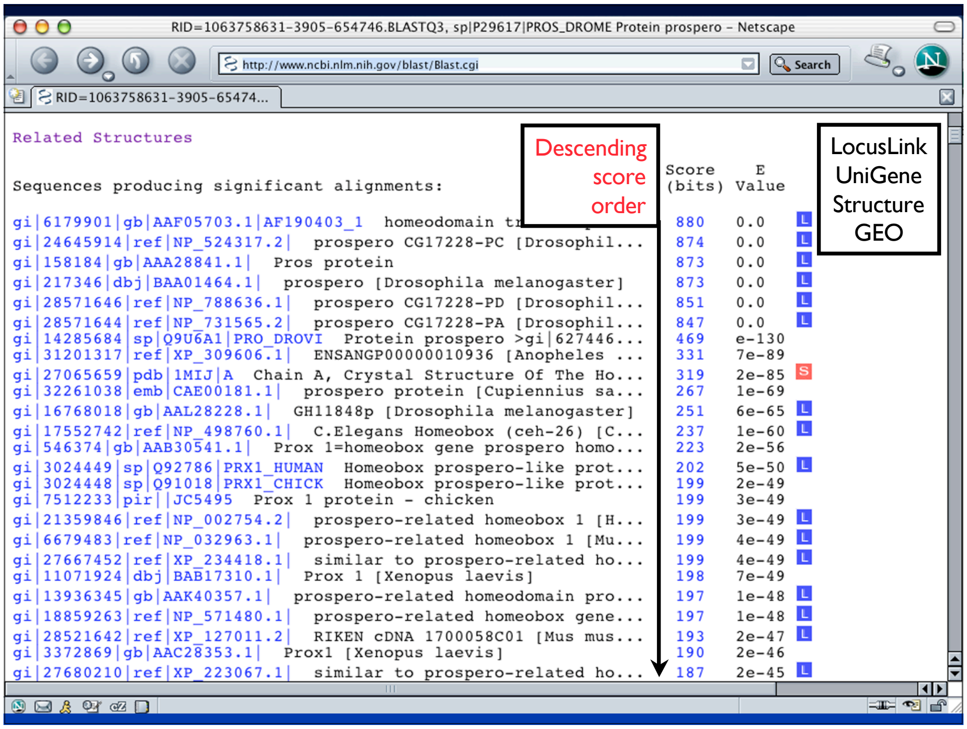

NCBI Sequence-Structure Alignment Visualization Service - Netscape
http://www.ncbi.nlm.nih.gov/Structure/cblast/cblast.cgi?viewmethod=1&viewmethod_optio
Search
NCBI Sequence-Structure Alignm...
NCBI
Related Structures
PubMed
BLAST
OMIM
Taxonomy
Structure
Help?
Done

Query: sp|P29617|PROS_DROME Protein prospero
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 Display (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
```

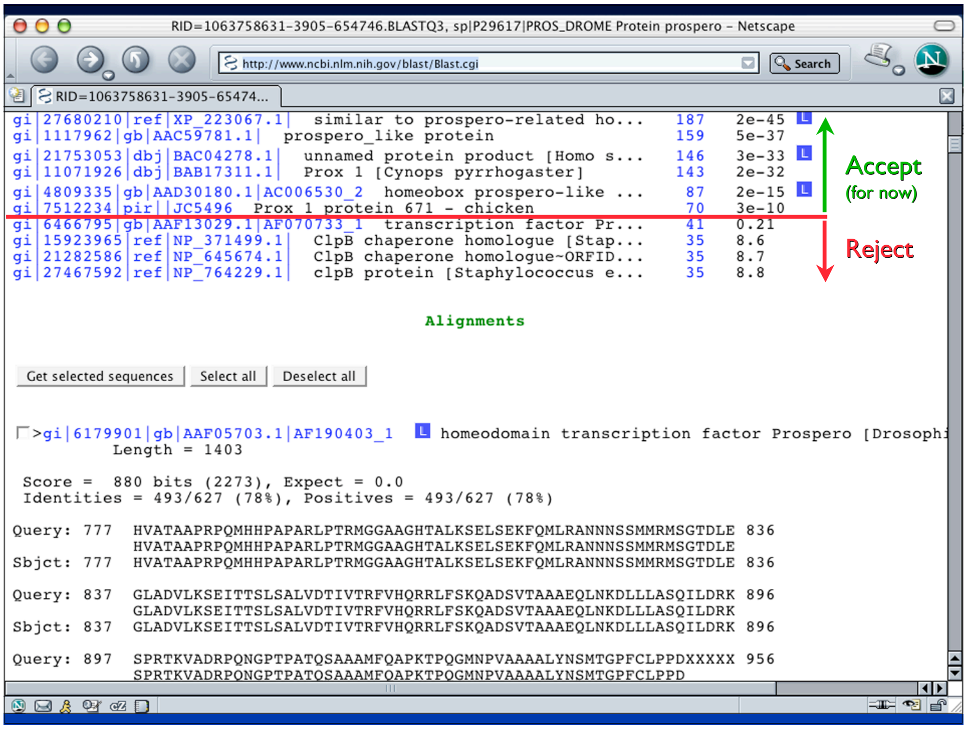
Score(bits) = 319, E_value = 2e-85
Aligned Length = 152 , Sequence Identity = 96 %


RID=1063758631-3905-654746.BLASTQ3, sp|P29617|PROS_DROME Protein prospero - Netscape
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi
Search
RID=1063758631-3905-65474...

```
Related Structures

                                                                      Score    E
Sequences producing significant alignments:                          (bits) Value

gi|6179901|gb|AAF05703.1|AF190403_1  homeodomain tr                   880   0.0    L
gi|24645914|ref|NP_524317.2|  prospero CG17228-PC [Drosophil...       874   0.0    L
gi|158184|gb|AAA28841.1|  Pros protein                                873   0.0    L
gi|217346|dbj|BAA01464.1|  prospero [Drosophila melanogaster]         873   0.0    L
gi|28571646|ref|NP_788636.1|  prospero CG17228-PD [Drosophil...       851   0.0    L
gi|28571644|ref|NP_731565.2|  prospero CG17228-PA [Drosophil...       847   0.0    L
gi|14285684|sp|Q9U6A1|PRO_DROVI  Protein prospero >gi|627446...       469   e-130
gi|31201317|ref|XP_309606.1|  ENSANGP00000010936 [Anopheles ...       331   7e-89
gi|27065659|pdb|1MIJ|A  Chain A, Crystal Structure Of The Ho...       319   2e-85  S
gi|32261038|emb|CAE00181.1|  prospero protein [Cupiennius sa...       267   1e-69
gi|16768018|gb|AAL28228.1|  GH11848p [Drosophila melanogaster]        251   6e-65  L
gi|17552742|ref|NP_498760.1|  C.Elegans Homeobox (ceh-26) [C...       237   1e-60  L
gi|546374|gb|AAB30541.1|  Prox 1=homeobox gene prospero homo...       223   2e-56
gi|3024449|sp|Q92786|PRX1_HUMAN  Homeobox prospero-like prot...       202   5e-50  L
gi|3024448|sp|Q91018|PRX1_CHICK  Homeobox prospero-like prot...       199   2e-49
gi|7512233|pir||JC5495  Prox 1 protein - chicken                      199   3e-49
gi|21359846|ref|NP_002754.2|  prospero-related homeobox 1 [H...       199   3e-49  L
gi|6679483|ref|NP_032963.1|  prospero-related homeobox 1 [Mu...       199   4e-49  L
gi|27667452|ref|XP_234418.1|  similar to prospero-related ho...       199   4e-49  L
gi|11071924|dbj|BAB17310.1|  Prox 1 [Xenopus laevis]                  198   7e-49
gi|13936345|gb|AAK40357.1|  prospero-related homeodomain pro...       197   1e-48  L
gi|18859263|ref|NP_571480.1|  prospero-related homeobox gene...       197   1e-48  L
gi|28521642|ref|XP_127011.2|  RIKEN cDNA 1700058C01 [Mus mus...       193   2e-47  L
gi|3372869|gb|AAC28353.1|  Prox1 [Xenopus laevis]                     190   2e-46
gi|27680210|ref|XP_223067.1|  similar to prospero-related ho...       187   2e-45  L
```

Descending score order

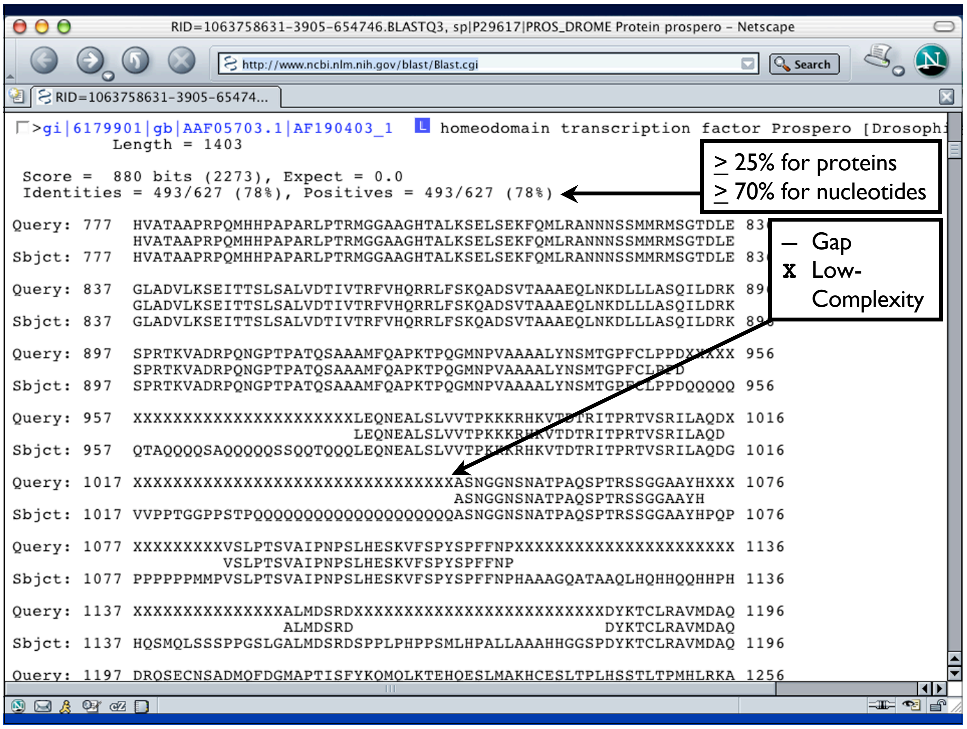
LocusLink
UniGene
Structure
GEO


RID=1063758631-3905-654746.BLASTQ3, sp|P29617|PROS_DROME Protein prospero - Netscape
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi
Search
RID=1063758631-3905-65474...
Get selected sequences
Select all
Deselect all

```
gi|27680210|ref|XP_223067.1|  similar to prospero-related ho...       187   2e-45  L
gi|1117962|gb|AAC59781.1|  prospero_like protein                      159   5e-37
gi|21753053|dbj|BAC04278.1|  unnamed protein product [Homo s...       146   3e-33  L
gi|11071926|dbj|BAB17311.1|  Prox 1 [Cynops pyrrhogaster]             143   2e-32
gi|4809335|gb|AAD30180.1|AC006530_2  homeobox prospero-like ...        87   2e-15  L
gi|7512234|pir||JC5496  Prox 1 protein 671 - chicken                   70   3e-10
gi|6466795|gb|AAF13029.1|AF070733_1  transcription factor Pr...        41   0.21
gi|15923965|ref|NP_371499.1|  ClpB chaperone homologue [Stap...        35   8.6
gi|21282586|ref|NP_645674.1|  ClpB chaperone homologue~ORFID...        35   8.7
gi|27467592|ref|NP_764229.1|  clpB protein [Staphylococcus e...        35   8.8

                                Alignments

□>gi|6179901|gb|AAF05703.1|AF190403_1  L homeodomain transcription factor Prospero [Drosophi
          Length = 1403

 Score =  880 bits (2273), 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
```

Accept
(for now)

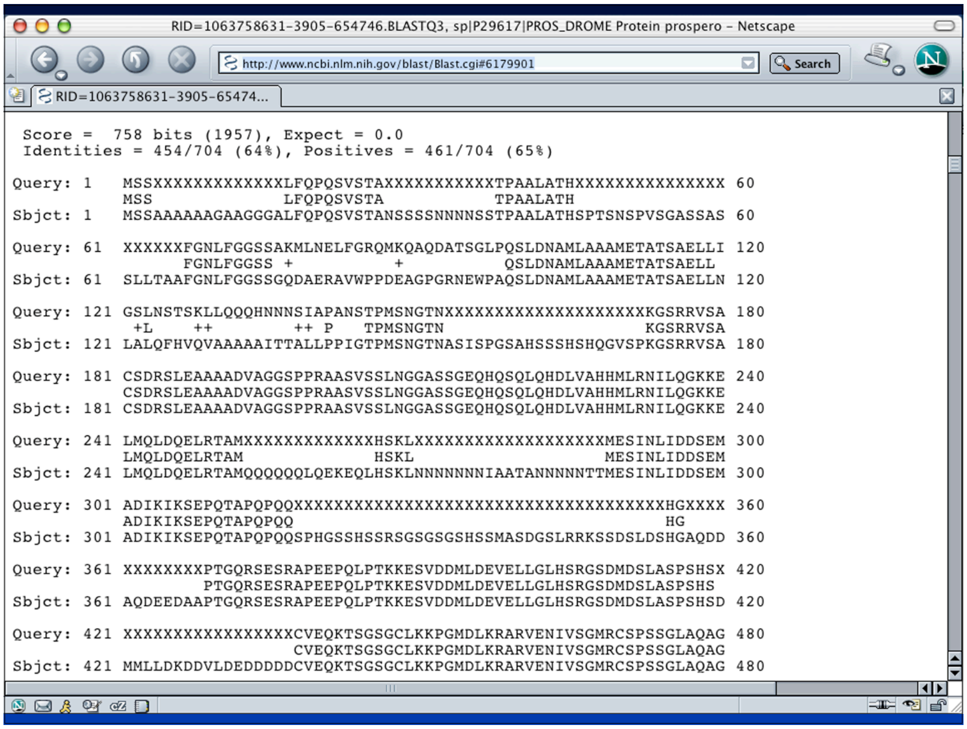
Reject


RID=1063758631-3905-654746.BLASTQ3, sp|P29617|PROS_DROME Protein prospero – Netscape
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi
Search
RID=1063758631-3905-65474...

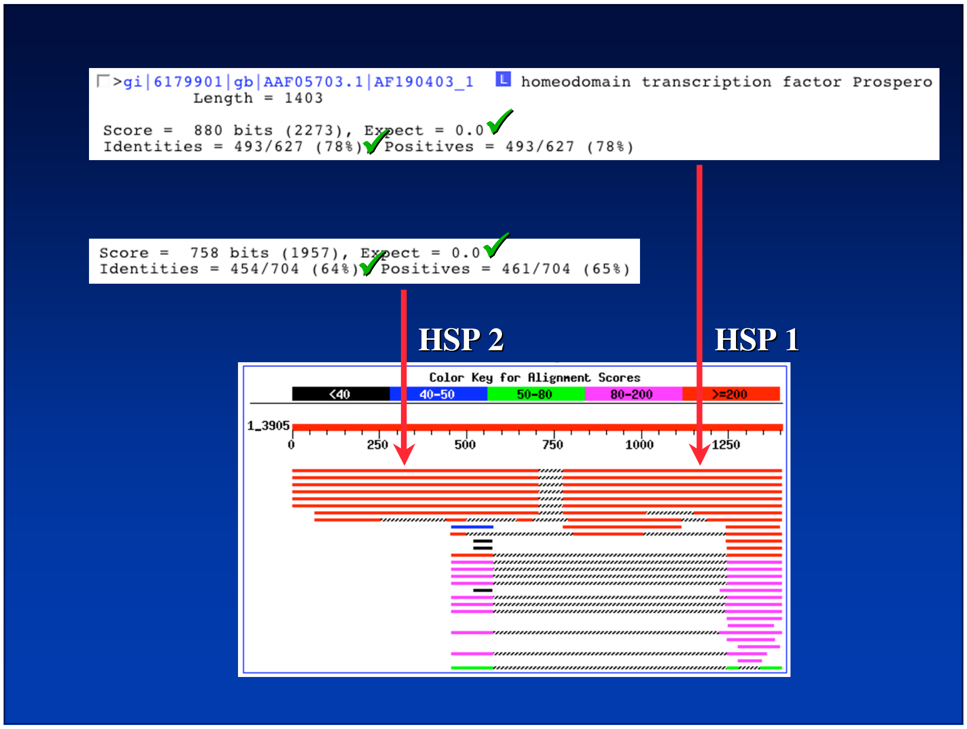
```
>gi|6179901|gb|AAF05703.1|AF190403_1  L homeodomain transcription factor Prospero [Drosophi
          Length = 1403

 Score =  880 bits (2273), 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
```

≥ 25% for proteins
≥ 70% for nucleotides

– Gap
x Low-Complexity


RID=1063758631-3905-654746.BLASTQ3, sp|P29617|PROS_DROME Protein prospero – Netscape
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi#6179901
Search
RID=1063758631-3905-65474...

```
 Score =  758 bits (1957), 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
```



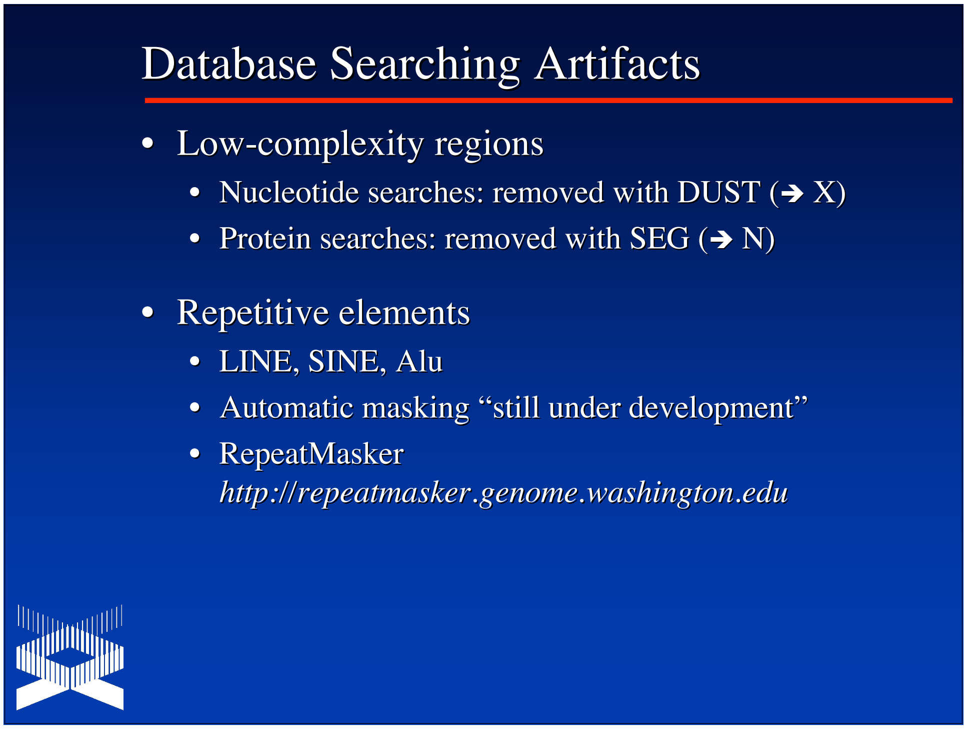
```
>gi|6179901|gb|AAF05703.1|AF190403_1  L  homeodomain transcription factor Prospero
          Length = 1403

 Score =  880 bits (2273), Expect = 0.0 ✓
 Identities = 493/627 (78%) ✓ Positives = 493/627 (78%)
```

```
 Score =  758 bits (1957), Expect = 0.0 ✓
 Identities = 454/704 (64%) ✓ Positives = 461/704 (65%)
```

Suggested BLAST Cutoffs

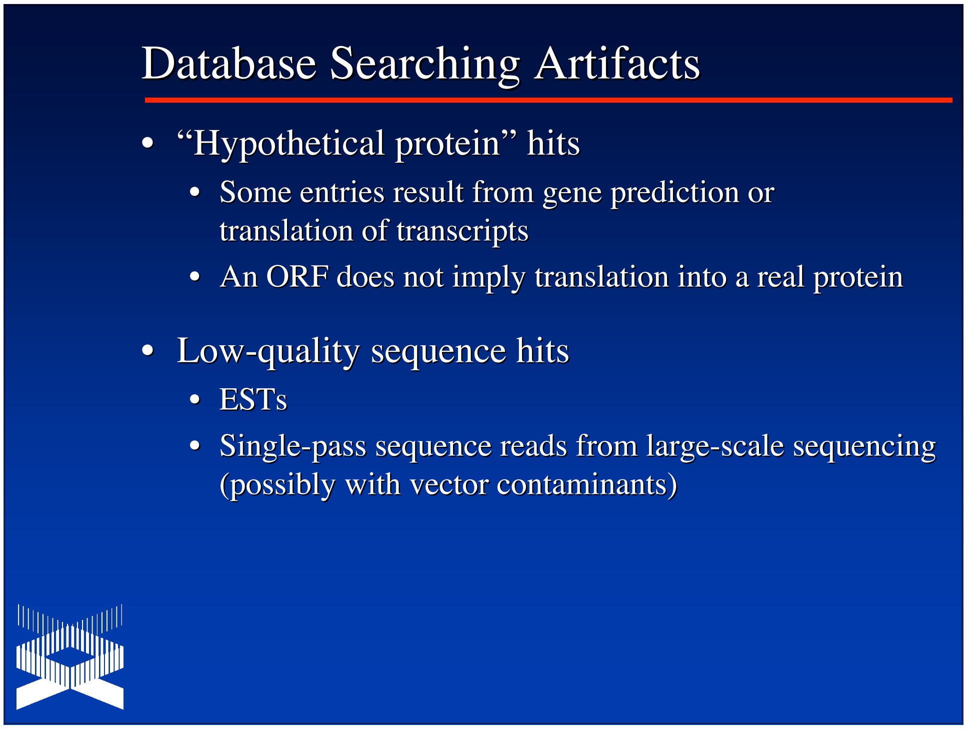
	E value	Sequence Identity
Nucleotide	$\leq 10^{-6}$	$\geq 70\%$
Protein	$\leq 10^{-3}$	$\geq 25\%$

PICK THE RIGHT MATRIX AND ALWAYS LOOK AT THE ALIGNMENTS!!!



Database Searching Artifacts

- Low-complexity regions
  - Nucleotide searches: removed with DUST (➔ X)
  - Protein searches: removed with SEG (➔ N)
- Repetitive elements
  - LINE, SINE, Alu
  - Automatic masking “still under development”
  - RepeatMasker
    http://repeatmasker.genome.washington.edu

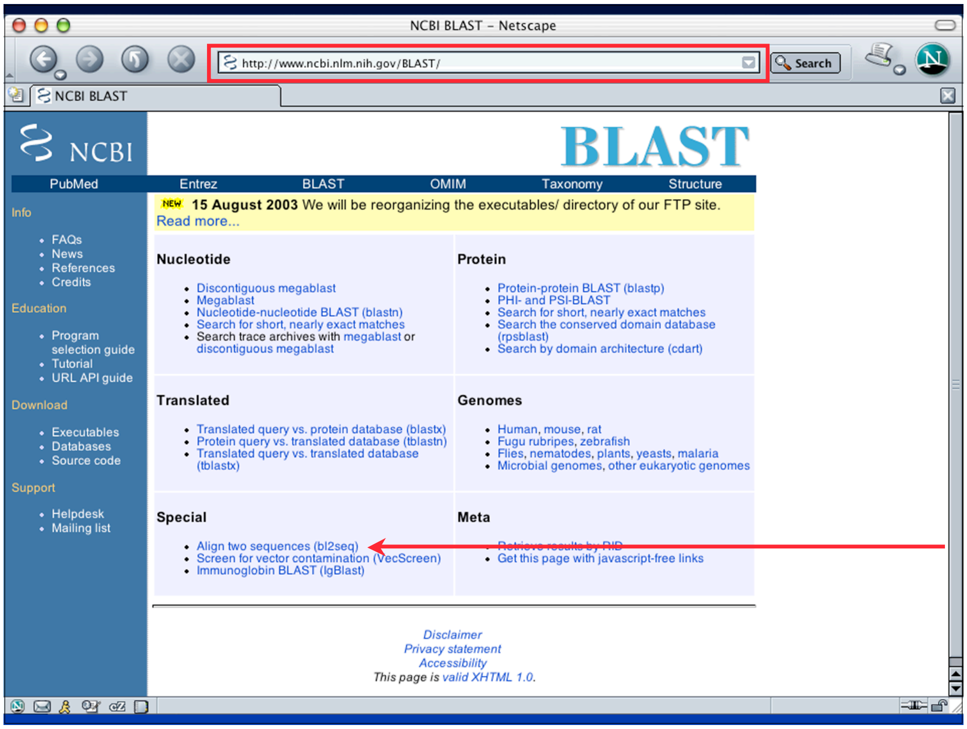


Database Searching Artifacts

- “Hypothetical protein” hits
  - Some entries result from gene prediction or translation of transcripts
  - An ORF does not imply translation into a real protein
- Low-quality sequence hits
  - ESTs
  - Single-pass sequence reads from large-scale sequencing (possibly with vector contaminants)

BLAST2SEQUENCES

- 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/*



NEW **15 August 2003** We will be reorganizing the executables/ directory of our FTP site. Read more...

Info
- FAQs
- News
- References
- Credits

Education
- Program selection guide
- Tutorial
- URL API guide

Download
- Executables
- Databases
- Source code

Support
- Helpdesk
- Mailing list

Nucleotide
- Discontiguous megablast
- 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)

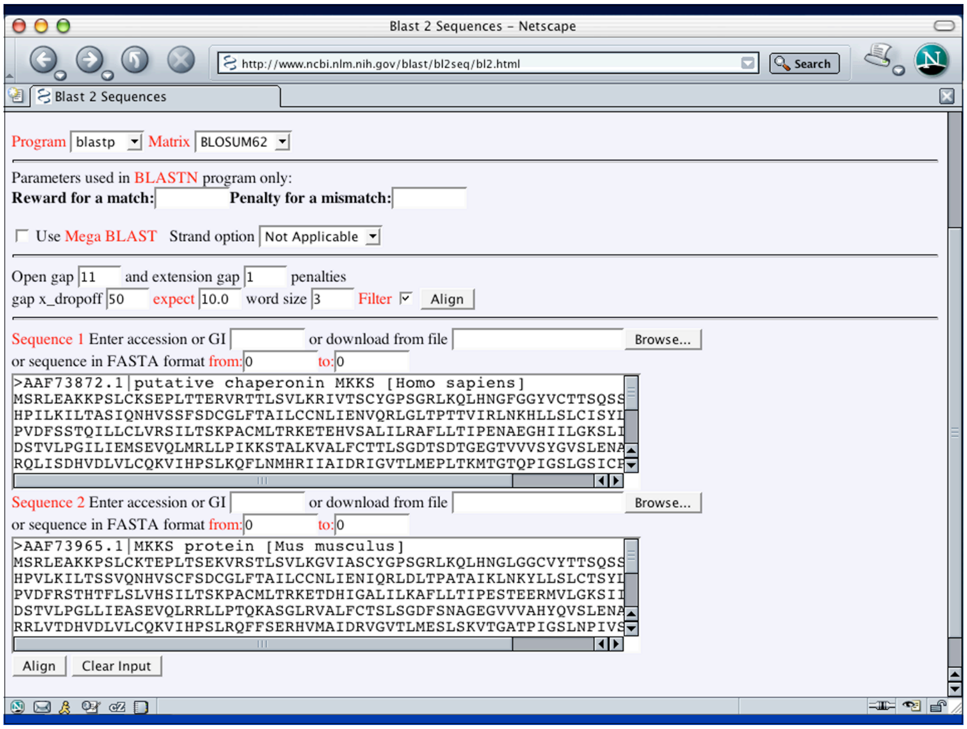
Translated
- Translated query vs. protein database (blastx)
- Protein query vs. translated database (tblastn)
- Translated query vs. translated database (tblastx)

Genomes
- Human, mouse, rat
- Fugu rubripes, zebrafish
- Flies, nematodes, plants, yeasts, malaria
- Microbial genomes, other eukaryotic genomes

Special
- Align two sequences (bl2seq)
- Screen for vector contamination (VecScreen)
- Immunoglobin BLAST (IgBlast)

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

Disclaimer
Privacy statement
Accessibility
This page is valid XHTML 1.0.



Blast 2 Sequences - Netscape
http://www.ncbi.nlm.nih.gov/blast/bl2seq/bl2.html
Search
Blast 2 Sequences

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

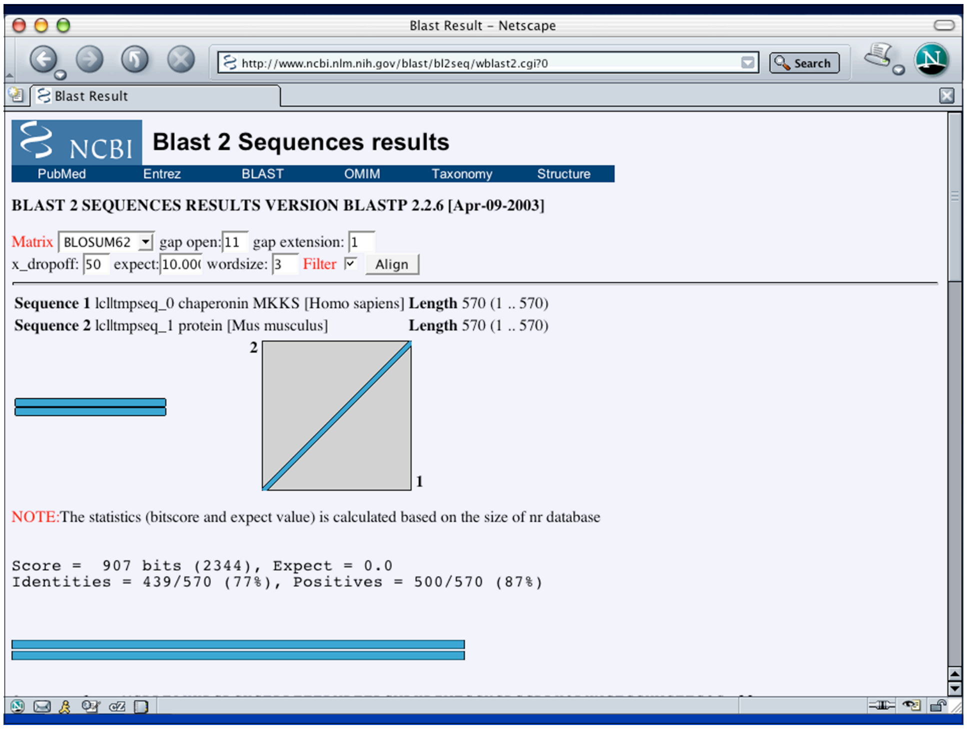
Sequence 1 Enter accession or GI or download from file Browse...
or sequence in FASTA format from: 0 to: 0

```
>AAF73872.1 putative chaperonin MKKS [Homo sapiens]
MSRLEAKKPSLCKSEPLTTERVRTTLSVLKRIVTSCYGPSGRLKQLHNGFGGYVCTTSQSS
HPILKILTASIQNHVSSFSDCGLFTAILCCNLIENVQRLGLTPTTVIRLNKHLLSLCISYI
PVDFSSTQILLCLVRSILTSKPACMLTRKETEHVSALILRAFLLTIPENAEGHIILGKSLI
DSTVLPGILIEMSEVQLMRLLPIKKSTALKVALFCTTLSGDTSDTGEGTVVVSYGVSLENA
RQLISDHVDLVLCQKVIHPSLKQFLNMHRIIAIDRIGVTLMEPLTKMTGTQPIGSLGSICP
```

Sequence 2 Enter accession or GI or download from file Browse...
or sequence in FASTA format from: 0 to: 0

```
>AAF73965.1 MKKS protein [Mus musculus]
MSRLEAKKPSLCKTEPLTSEKVRSTLSVLKGVIASCYGPSGRLKQLHNGLGGCVYTTSQSS
HPVLKILTSSVQNHVSCFSDCGLFTAILCCNLIENIQRLDLTPATAIKLNKYLLSLCTSYI
PVDFRSTHTFLSLVHSILTSKPACMLTRKETDHIGALILKAFLLTIPESTEERMVLGKSII
DSTVLPGLLIEASEVQLRRLLPTQKASGLRVALFCTSLSGDFSNAGEGVVVAHYQVSLENA
RRLVTDHVDLVLCQKVIHPSLRQFFSERHVMAIDRVGVTLMESLSKVTGATPIGSLNPIVS
```

Align Clear Input



Blast 2 Sequences results

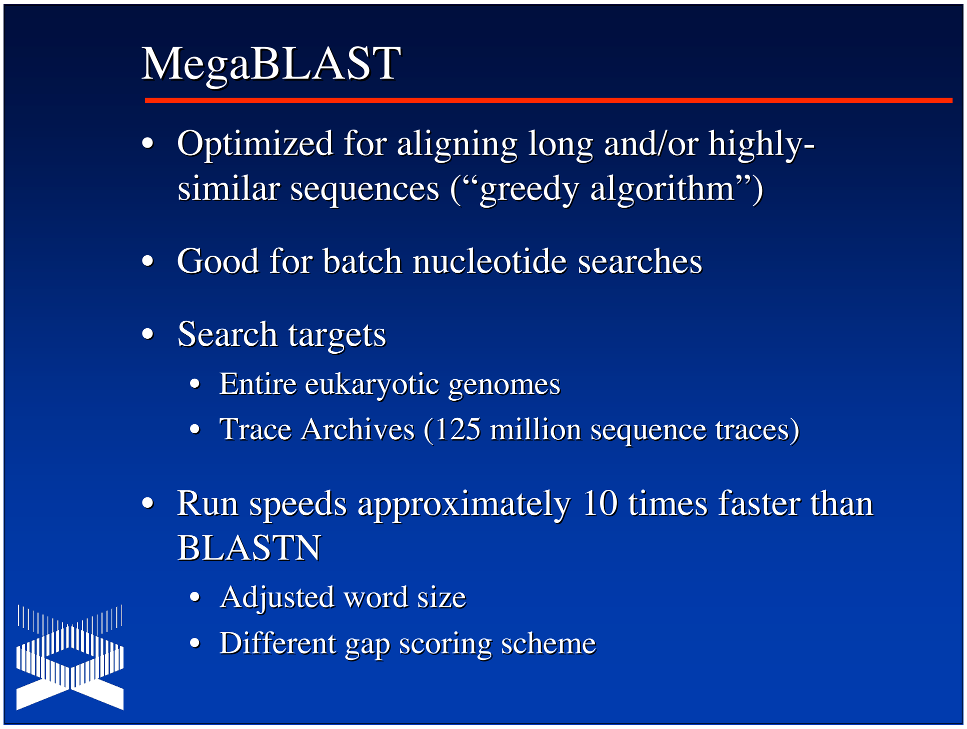
BLAST 2 SEQUENCES RESULTS VERSION BLASTP 2.2.6 [Apr-09-2003]

Matrix BLOSUM62 gap open: 11 gap extension: 1
x_dropoff: 50 expect: 10.00(wordsize: 3 Filter ☑ Align

Sequence 1 lcl|tmpseq_0 chaperonin MKKS [Homo sapiens] **Length** 570 (1 .. 570)
Sequence 2 lcl|tmpseq_1 protein [Mus musculus] **Length** 570 (1 .. 570)

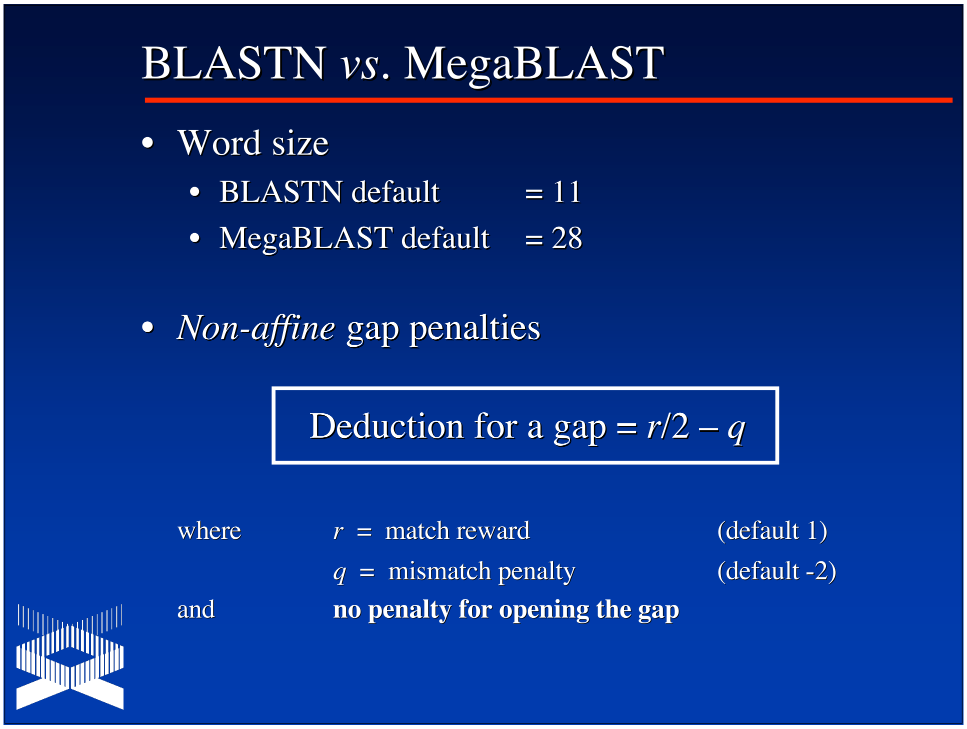
NOTE:The statistics (bitscore and expect value) is calculated based on the size of nr database

```
Score =  907 bits (2344), Expect = 0.0
Identities = 439/570 (77%), Positives = 500/570 (87%)
```



MegaBLAST

- Optimized for aligning long and/or highly-similar sequences (“greedy algorithm”)
- Good for batch nucleotide searches
- Search targets
  - Entire eukaryotic genomes
  - Trace Archives (125 million sequence traces)
- Run speeds approximately 10 times faster than BLASTN
  - Adjusted word size
  - Different gap scoring scheme



BLASTN *vs.* MegaBLAST

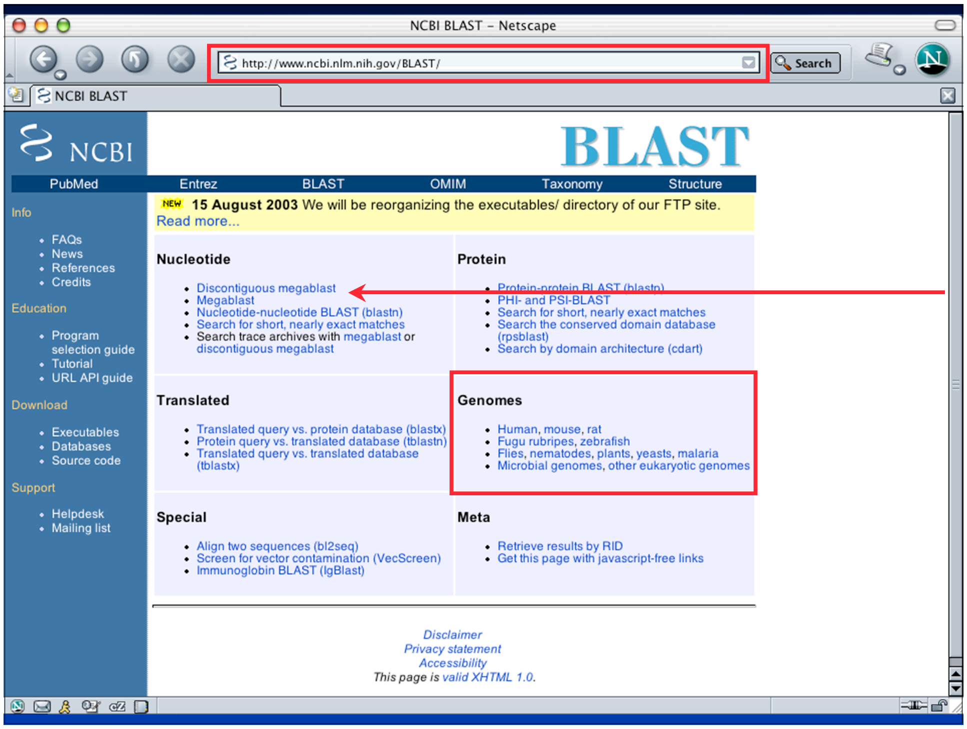
- Word size
  - BLASTN default = 11
  - MegaBLAST default = 28
- *Non-affine* gap penalties

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

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

Discontiguous MegaBLAST

- Designed specifically for the comparison of diverged sequences, particularly from different organisms
- Since these types of comparison may yield low degrees of identity, this variant performs better than the original MegaBLAST, which is optimized for sequences that are highly similar



NEW **15 August 2003** We will be reorganizing the executables/ directory of our FTP site. Read more...

Info

- FAQs
- News
- References
- Credits

Education

- Program selection guide
- Tutorial
- URL API guide

Download

- Executables
- Databases
- Source code

Support

- Helpdesk
- Mailing list

Nucleotide

- Discontiguous megablast
- 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

- Human, mouse, rat
- Fugu rubripes, zebrafish
- Flies, nematodes, plants, yeasts, malaria
- Microbial genomes, other eukaryotic genomes

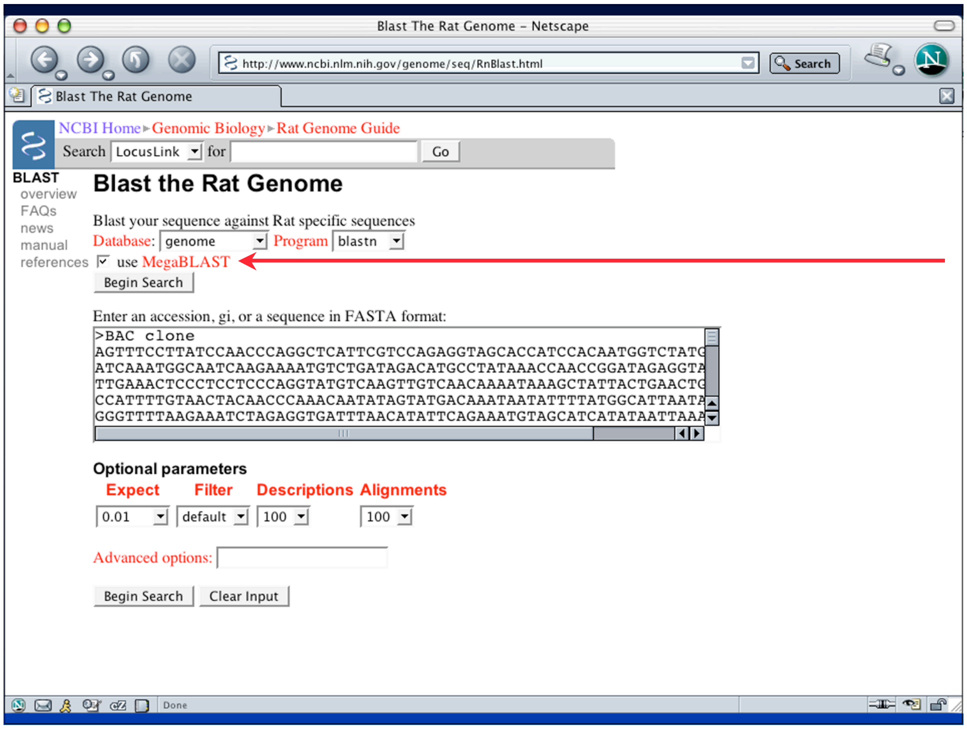
Special

- Align two sequences (bl2seq)
- Screen for vector contamination (VecScreen)
- Immunoglobin BLAST (IgBlast)

Meta

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

Disclaimer
Privacy statement
Accessibility
This page is valid XHTML 1.0.



Blast The Rat Genome – Netscape
http://www.ncbi.nlm.nih.gov/genome/seq/RnBlast.html
Search
Blast The Rat Genome
NCBI Home ▸ Genomic Biology ▸ Rat Genome Guide
Search
for
Go
BLAST
overview
FAQs
news
manual
references
Begin Search
Clear Input
Done

Blast the Rat Genome

Blast your sequence against Rat specific sequences

Database: genome Program blastn

☑ use MegaBLAST

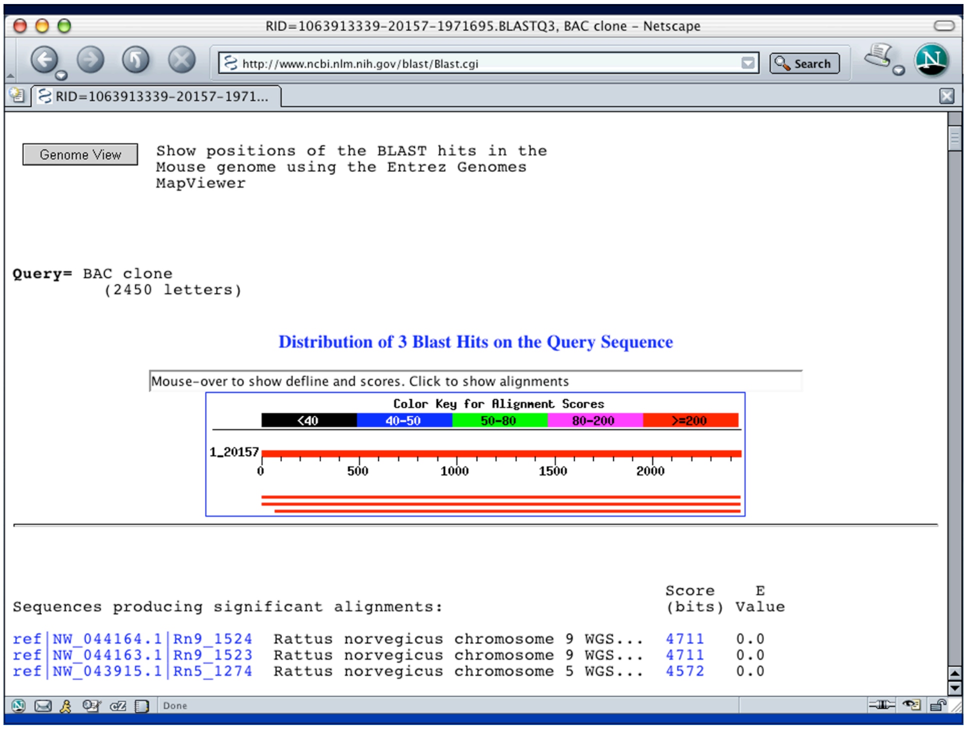
Enter an accession, gi, or a sequence in FASTA format:

```
>BAC clone
AGTTTCCTTATCCAACCCAGGCTCATTCGTCCAGAGGTAGCACCATCCACAATGGTCTATG
ATCAAATGGCAATCAAGAAAATGTCTGATAGACATGCCTATAAACCAACCGGATAGAGGTA
TTGAAACTCCCTCCTCCCAGGTATGTCAAGTTGTCAACAAAATAAAGCTATTACTGAACTG
CCATTTTGTAACTACAACCCAAACAATATAGTATGACAAATAATATTTTATGGCATTAATA
GGGTTTTAAGAAATCTAGAGGTGATTTAACATATTCAGAAATGTAGCATCATATAATTAAA
```

Optional parameters

Expect	Filter	Descriptions	Alignments
0.01	default	100	100

Advanced options:

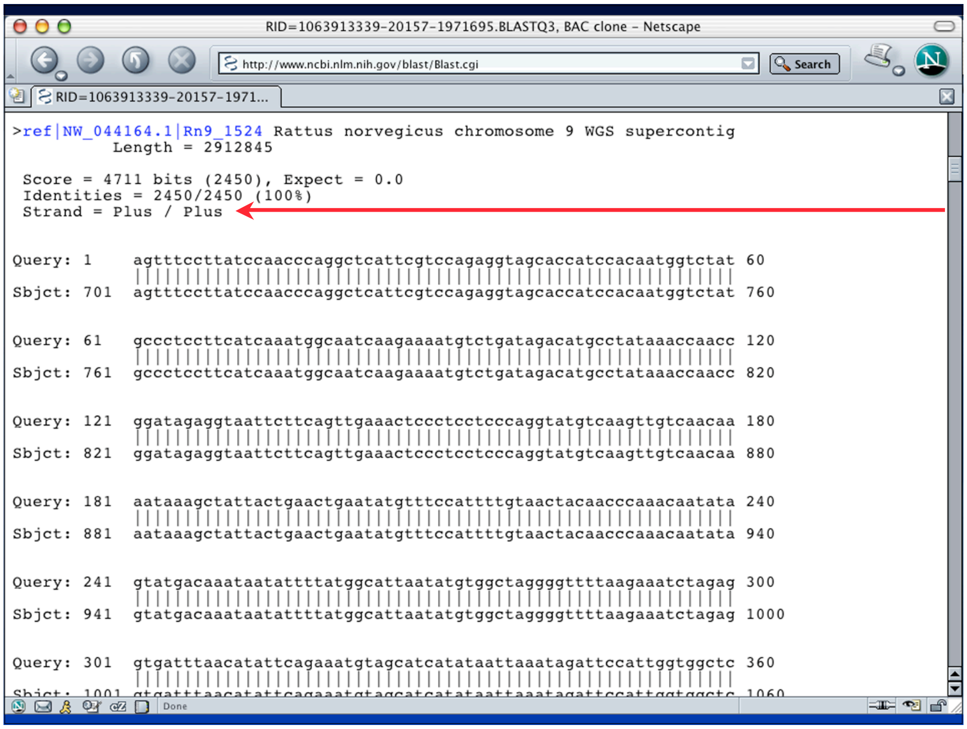


RID=1063913339-20157-1971695.BLASTQ3, BAC clone – Netscape
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi
Search
RID=1063913339-20157-1971...
Genome View
Mouse-over to show defline and scores. Click to show alignments
Color Key for Alignment Scores
<40
40-50
50-80
80-200
>=200
1_20157
0
500
1000
1500
2000
Done

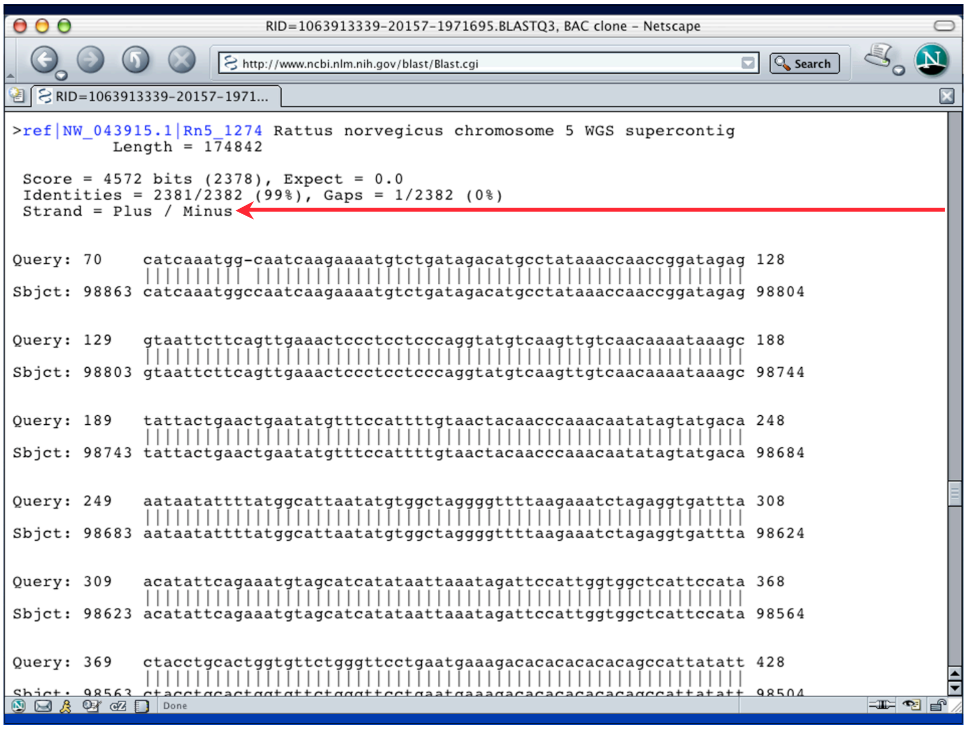
```
Show positions of the BLAST hits in the
Mouse genome using the Entrez Genomes
MapViewer

Query= BAC clone
         (2450 letters)
```

Distribution of 3 Blast Hits on the Query Sequence

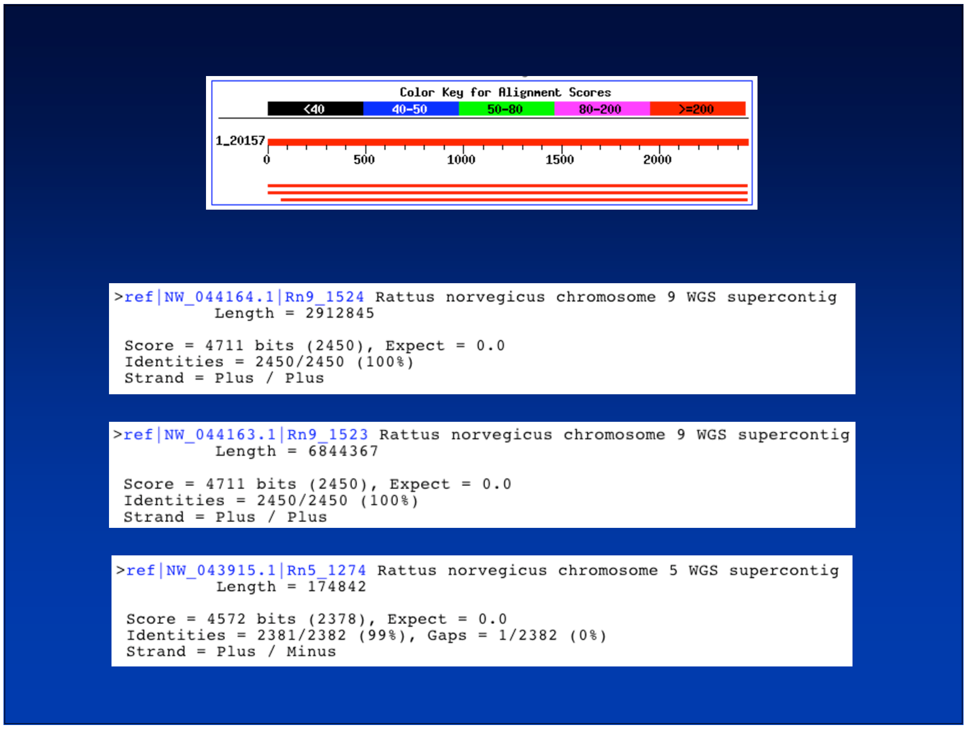
```
                                                                  Score    E
Sequences producing significant alignments:                       (bits) Value

ref|NW_044164.1|Rn9_1524  Rattus norvegicus chromosome 9 WGS...  4711    0.0
ref|NW_044163.1|Rn9_1523  Rattus norvegicus chromosome 9 WGS...  4711    0.0
ref|NW_043915.1|Rn5_1274  Rattus norvegicus chromosome 5 WGS...  4572    0.0
```


RID=1063913339-20157-1971695.BLASTQ3, BAC clone - Netscape
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi
Search
RID=1063913339-20157-1971...
Done

```
>ref|NW_044164.1|Rn9_1524 Rattus norvegicus chromosome 9 WGS supercontig
          Length = 2912845

 Score = 4711 bits (2450), Expect = 0.0
 Identities = 2450/2450 (100%)
 Strand = Plus / Plus


Query: 1    agtttccttatccaacccaggctcattcgtccagaggtagcaccatccacaatggtctat 60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 701  agtttccttatccaacccaggctcattcgtccagaggtagcaccatccacaatggtctat 760


Query: 61   gccctccttcatcaaatggcaatcaagaaaatgtctgatagacatgcctataaaccaacc 120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 761  gccctccttcatcaaatggcaatcaagaaaatgtctgatagacatgcctataaaccaacc 820


Query: 121  ggatagaggtaattcttcagttgaaactccctcctcccaggtatgtcaagttgtcaacaa 180
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 821  ggatagaggtaattcttcagttgaaactccctcctcccaggtatgtcaagttgtcaacaa 880


Query: 181  aataaagctattactgaactgaatatgtttccattttgtaactacaacccaaacaatata 240
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 881  aataaagctattactgaactgaatatgtttccattttgtaactacaacccaaacaatata 940


Query: 241  gtatgacaaataatattttatggcattaatatgtggctaggggttttaagaaatctagag 300
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 941  gtatgacaaataatattttatggcattaatatgtggctaggggttttaagaaatctagag 1000


Query: 301  gtgatttaacatattcagaaatgtagcatcatataattaaatagattccattggtggctc 360
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 1001 gtgatttaacatattcagaaatgtagcatcatataattaaatagattccattggtggctc 1060
```


RID=1063913339-20157-1971695.BLASTQ3, BAC clone - Netscape
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi
Search
RID=1063913339-20157-1971...
Done

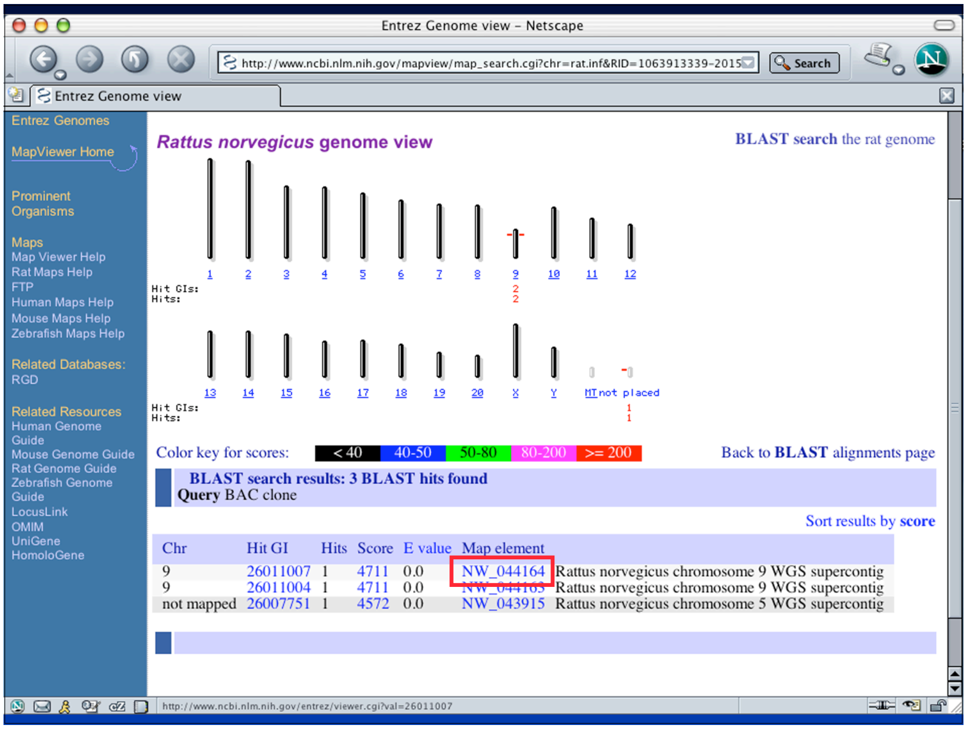
```
>ref|NW_043915.1|Rn5_1274 Rattus norvegicus chromosome 5 WGS supercontig
          Length = 174842

 Score = 4572 bits (2378), Expect = 0.0
 Identities = 2381/2382 (99%), Gaps = 1/2382 (0%)
 Strand = Plus / Minus


Query: 70    catcaaatgg-caatcaagaaaatgtctgatagacatgcctataaaccaaccggatagag 128
             |||||||||| |||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 98863 catcaaatggccaatcaagaaaatgtctgatagacatgcctataaaccaaccggatagag 98804


Query: 129   gtaattcttcagttgaaactccctcctcccaggtatgtcaagttgtcaacaaaataaagc 188
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 98803 gtaattcttcagttgaaactccctcctcccaggtatgtcaagttgtcaacaaaataaagc 98744


Query: 189   tattactgaactgaatatgtttccattttgtaactacaacccaaacaatatagtatgaca 248
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 98743 tattactgaactgaatatgtttccattttgtaactacaacccaaacaatatagtatgaca 98684


Query: 249   aataatattttatggcattaatatgtggctaggggttttaagaaatctagaggtgattta 308
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 98683 aataatattttatggcattaatatgtggctaggggttttaagaaatctagaggtgattta 98624


Query: 309   acatattcagaaatgtagcatcatataattaaatagattccattggtggctcattccata 368
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 98623 acatattcagaaatgtagcatcatataattaaatagattccattggtggctcattccata 98564


Query: 369   ctacctgcactggtgttctgggttcctgaatgaaagacacacacacacagccattatatt 428
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 98563 ctacctgcactggtgttctgggttcctgaatgaaagacacacacacacagccattatatt 98504
```



```
>ref|NW_044164.1|Rn9_1524 Rattus norvegicus chromosome 9 WGS supercontig
          Length = 2912845

 Score = 4711 bits (2450), Expect = 0.0
 Identities = 2450/2450 (100%)
 Strand = Plus / Plus
```

```
>ref|NW_044163.1|Rn9_1523 Rattus norvegicus chromosome 9 WGS supercontig
          Length = 6844367

 Score = 4711 bits (2450), Expect = 0.0
 Identities = 2450/2450 (100%)
 Strand = Plus / Plus
```

```
>ref|NW_043915.1|Rn5_1274 Rattus norvegicus chromosome 5 WGS supercontig
          Length = 174842

 Score = 4572 bits (2378), Expect = 0.0
 Identities = 2381/2382 (99%), Gaps = 1/2382 (0%)
 Strand = Plus / Minus
```



***Rattus norvegicus* genome view**

BLAST search the rat genome

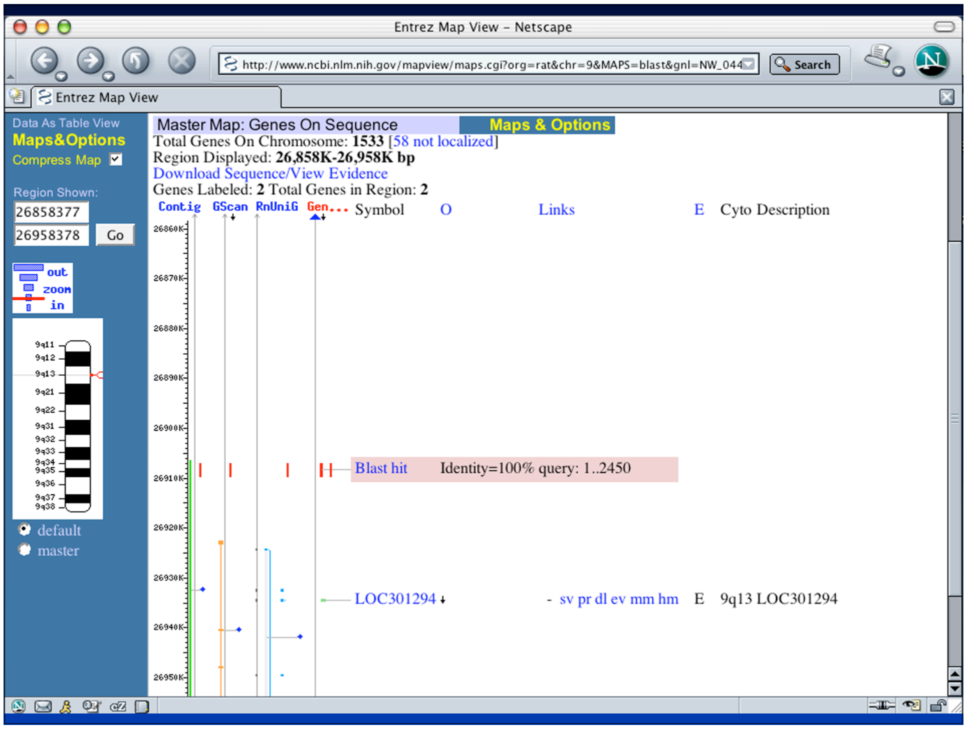
Color key for scores: ⬛ < 40 🟦 40-50 🟩 50-80 🟪 80-200 🟥 >= 200

Back to **BLAST** alignments page

BLAST search results: 3 BLAST hits found
Query BAC clone

Sort results by **score**

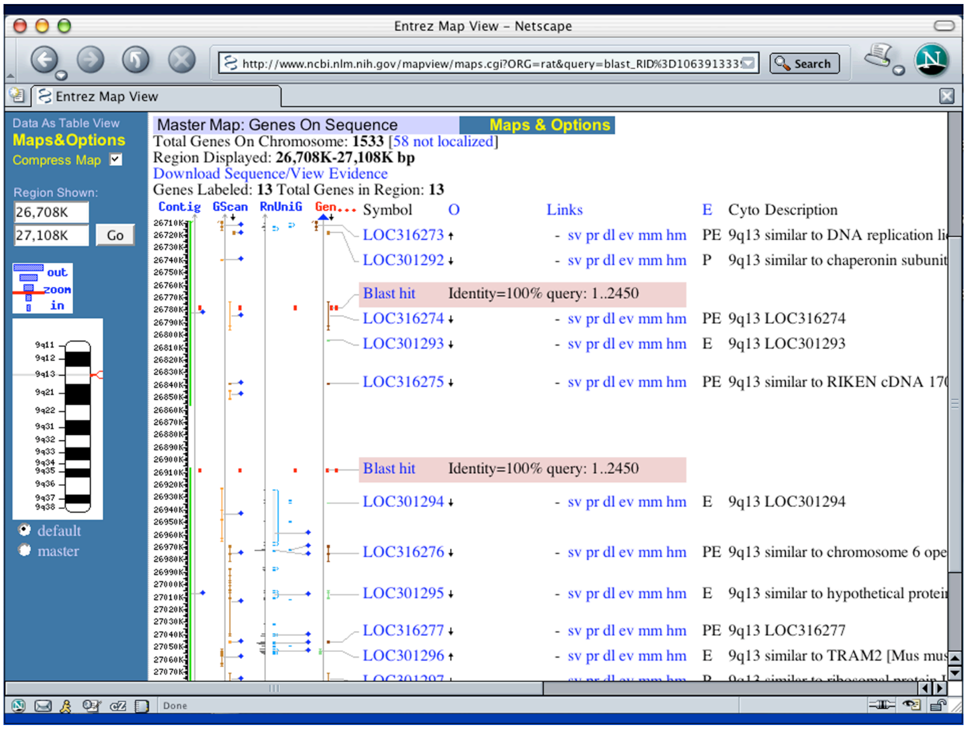
Chr	Hit GI	Hits	Score	E value	Map element	
9	26011007	1	4711	0.0	NW_044164	Rattus norvegicus chromosome 9 WGS supercontig
9	26011004	1	4711	0.0	NW_044163	Rattus norvegicus chromosome 9 WGS supercontig
not mapped	26007751	1	4572	0.0	NW_043915	Rattus norvegicus chromosome 5 WGS supercontig



Master Map: Genes On Sequence
Maps & Options

Total Genes On Chromosome: **1533** [58 not localized]
Region Displayed: **26,858K-26,958K bp**
Download Sequence/View Evidence
Genes Labeled: **2** Total Genes in Region: **2**

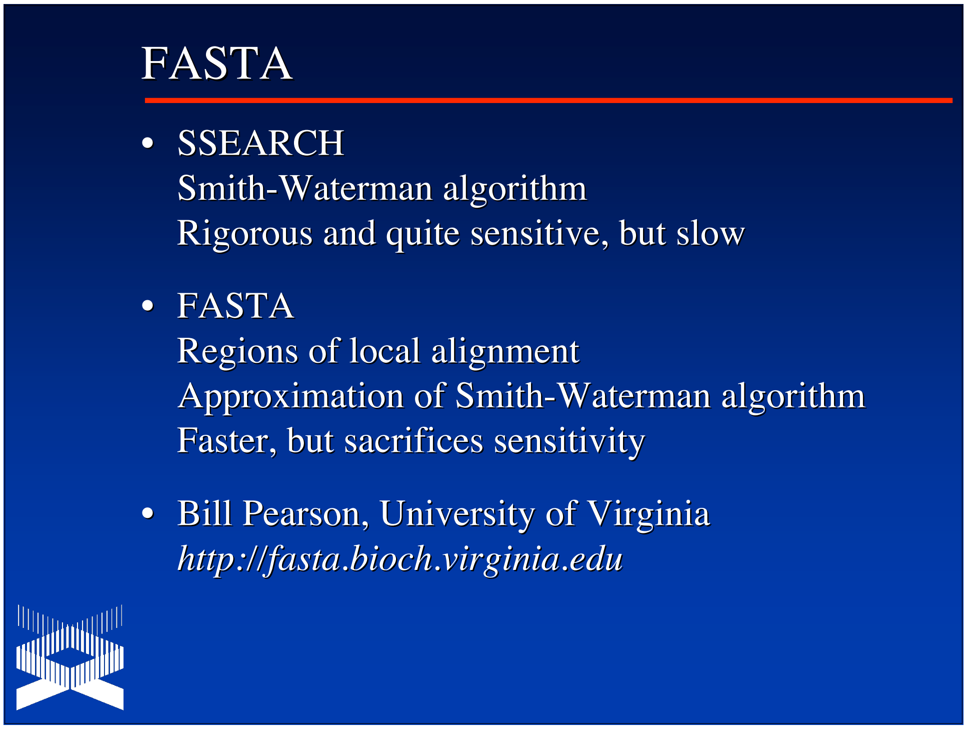
Symbol	O	Links	E	Cyto Description
Blast hit	Identity=100% query: 1..2450			
LOC301294 ↓	-	sv pr dl ev mm hm	E	9q13 LOC301294



Master Map: Genes On Sequence
Maps & Options

Total Genes On Chromosome: **1533** [58 not localized]
Region Displayed: **26,708K-27,108K bp**
Download Sequence/View Evidence
Genes Labeled: **13** Total Genes in Region: **13**

Symbol	O	Links	E	Cyto Description
LOC316273 ↑	-	sv pr dl ev mm hm	PE	9q13 similar to DNA replication li
LOC301292 ↓	-	sv pr dl ev mm hm	P	9q13 similar to chaperonin subunit
Blast hit	Identity=100% query: 1..2450			
LOC316274 ↓	-	sv pr dl ev mm hm	PE	9q13 LOC316274
LOC301293 ↓	-	sv pr dl ev mm hm	E	9q13 LOC301293
LOC316275 ↓	-	sv pr dl ev mm hm	PE	9q13 similar to RIKEN cDNA 17
Blast hit	Identity=100% query: 1..2450			
LOC301294 ↓	-	sv pr dl ev mm hm	E	9q13 LOC301294
LOC316276 ↓	-	sv pr dl ev mm hm	PE	9q13 similar to chromosome 6 ope
LOC301295 ↓	-	sv pr dl ev mm hm	E	9q13 similar to hypothetical protei
LOC316277 ↓	-	sv pr dl ev mm hm	PE	9q13 LOC316277
LOC301296 ↑	-	sv pr dl ev mm hm	E	9q13 similar to TRAM2 [Mus mus
LOC301297 ↓	-	sv pr dl ev mm hm	P	9q13 similar to ribosomal protein L



FASTA

- SSEARCH
  Smith-Waterman algorithm
  Rigorous and quite sensitive, but slow
- FASTA
  Regions of local alignment
  Approximation of Smith-Waterman algorithm
  Faster, but sacrifices sensitivity
- Bill Pearson, University of Virginia
  http://fasta.bioch.virginia.edu