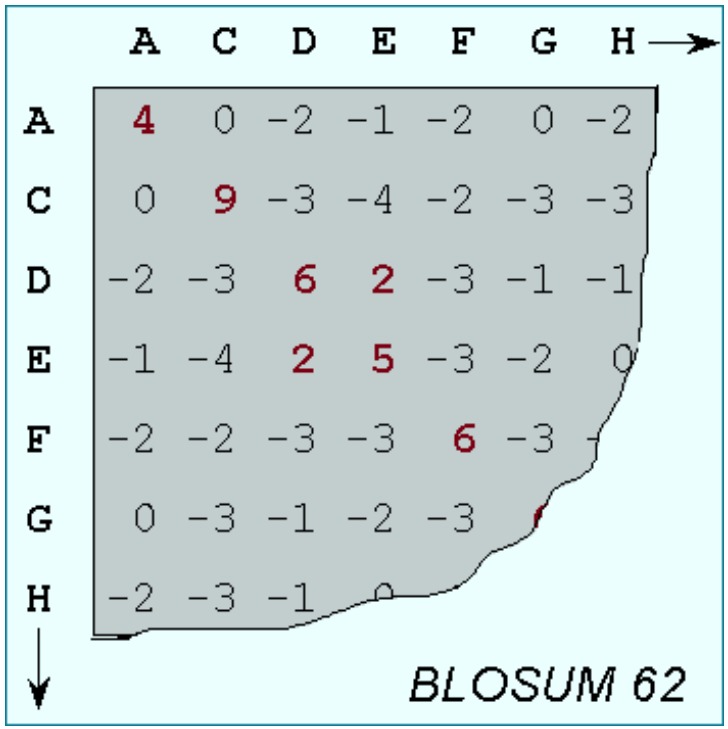
Virtual Bioinformatics Conference

- PLEASE Register! It's Free.
- http://www.ndsu.nodak.edu/virtual-genomics/conference_2002.htm
- September 24-26, 2002, Access Grid, Room ECS 212.
- You can be on TV!

BioPerl Modules

- **Bio::PreSeq**, module for reading, accessing, manipulating, analyzing single sequences.
- **Bio::UnivAln**, module for reading, parsing, writing, slicing, and manipulating multiple biosequences (sequence multisets and alignments).
- **Bio::Struct**, module for reading, writing, accessing, manipulating, and analyzing 3D structures.
- Support for invoking **BLAST** and other programs.
- Listing: bioperl-1.0.2::Bio & here.
- BioPerl Tutorial

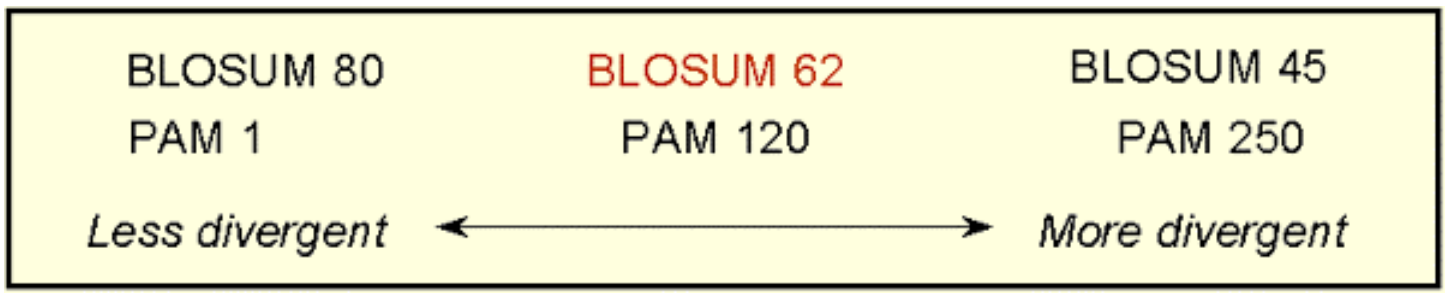
Substitution Matrices



	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	12																				C
S	0	2																			S
T	-2	1	3																		T
P	-3	1	0	6																	P
A	-2	1	1	1	2																A
G	-3	1	0	-1	1	5															G
N	-4	1	0	-1	0	0	2														N
D	-5	0	0	-1	0	1	2	4													D
E	-5	0	0	-1	0	0	1	3	4												E
Q	-5	-1	-1	0	0	-1	1	2	2	4											Q
H	-3	-1	-1	0	-1	-2	2	1	1	3	6										H
R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6									R
K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5								K
M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6							M
I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5						I
L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6					L
V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4				V
F	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9			F
Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10		Y
W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

PAM250

PAM vs BLOSUM



Which Substitution Matrix?

- BLOSUM-62 matrix best for detecting most weak protein similarities.
- For particularly long and weak alignments, BLOSUM-45 matrix may be superior.
-

Query Length	Substitution Matrix	Gap Costs
<35	PAM 30	(9,1)
35-50	PAM 70	(10,1)
50-85	BLOSUM 80	(10,1)
>85	BLOSUM 62	(11,1)

BLAST & FASTA

- FASTA

[Lipman Pearson ’85, ‘88]

- Basic Local Alignment Search Tool

[Altschul, Gish, Miller, Myers, Lipman ’90]

Search Strategy

FASTA Search Strategy

- Find "hot spots" of length k (exact match) for each length k word in query.
- Locate "runs" of "hot spots".
- Do detailed "Smith-Waterman" local alignment at these locations.

BLAST Improvements

- Lipman et al.: speeded up finding “runs” of “hot spots”.
- Eugene Myers ’94: “Sublinear algorithm for approximate keyword matching”.
- Karlin, Altschul, Dembo ’90, ’91: “Statistical Significance of Matches”

General Bioinformatics Resources

- PubMed (PubMed) at National Center for Biotechnology Information (NCBI) at the National Institutes of Health (NIH):
- http://www4.ncbi.nlm.nih.gov/entrez/query.fcgi
- Try Lambda Cro (73101), Ecoli Sigma-70 (1SIG), Ecoli Sigma factor (1072030), Bacteriorhodopsin (14194473), 1baza vs. 1myka (P-22 Arc repressors)
- http://www.ncbi.nlm.nih.gov/BLAST/ (BLAST)

BLAST Overview

- Program(s) to search all sequence databases
- Tremendous Speed/Less Sensitive
- Statistical Significance reported
- WWWBLAST, QBLAST (send now, retrieve results later), Standalone BLAST, BLASTcl3 (Client version, TCP/IP connection to NCBI server), BLAST URLAPI (to access QBLAST, no local client)

Homework!

- Run the BLAST Tutorials.

BLAST Cont'd

- **Nucleotide BLAST**
  - Standard
  - MEGABLAST (Compare large sets, Near-exact searches)
  - Short Sequences (higher E-value threshold, smaller word size, no low-complexity filtering)
- **Protein BLAST**
  - Standard
  - PSI-BLAST (Position Specific Iterated BLAST)
  - PHI-BLAST (Pattern Hit Initiated BLAST; reg expr. Or Motif search)
  - Short Sequences (higher E-value threshold, smaller word size, no low-complexity filtering, PAM-30)
- **Translating BLAST**
  - Blastx: Search nucleotide sequence in protein database (6 reading frames)
  - Tblastn: Search protein sequence in nucleotide dB
  - Tblastx: Search nucleotide seq (6 frames) in nucleotide DB (6 frames)

BLAST Cont'd

- **RPS BLAST**
  - Compare protein sequence against Conserved Domain DB; Helps in predicting rough structure and function
- **Pairwise BLAST**
  - blastp (2 Proteins), blastn (2 nucleotides), tblastn (protein-nucleotide w/ 6 frames), blastx (nucleotide-protein), tblastx (nucleotide w/6 frames-nucleotide w/ 6 frames)
- **Specialized BLAST**
  - Human & Other finished/unfinished genomes
  - P. falciparum: Search ESTs, STSs, GSSs, HTGs
  - VecScreen: screen for contamination while sequencing
  - IgBLAST: Immunoglobin sequence database

BLAST Credits

- Stephen Altschul
- Jonathan Epstein
- David Lipman
- Tom Madden
- Scott McGinnis
- Jim Ostell
- Alex Schaffer
- Sergei Shavirin
- Heidi Sofia
- Jinghui Zhang

Useful Terms

- **E value:** Expectation value. expected # of alignments with scores equivalent to or better than S to occur by chance. The lower the E value, the more significant the score.
- **P value:** The probability of an alignment occurring with the given score, S, or better. Calculated by relating the observed score, S, to the expected distribution of HSP scores from comparisons of random sequences of the same length and composition as the query to the database. The most highly significant P values will be those close to 0.
- **HSP:** High-scoring segment pair. Local alignments with no gaps that achieve high alignment scores
- **Identity (Similarity):** The extent to which two (nucleotide or amino acid) sequences are invariant (similar).

Databases used by BLAST

- **Protein**
  - nr (everything), swissprot, pdb, alu, individual genomes
- **Nucleotide**
  - nr, dbest, dbsts, htgs (unfinished genomic sequences), gss, pdb, vector, mito, alu, epd
- **Misc**

Rules of Thumb

- Most sequences with significant similarity over their entire lengths are homologous.
- Matches that are > 50% identical in a 20-40 aa region occur frequently by chance.
- Distantly related homologs may lack significant similarity. Homologous sequences may have few absolutely conserved residues.
- A homologous to B & B to C $\Rightarrow$ A homologous to C.
- Low complexity regions, transmembrane regions and coiled-coil regions frequently display significant similarity without homology.
- Greater evolutionary distance implies that length of a local alignment required to achieve a statistically significant score also increases.

Rules of Thumb

- Results of searches using different scoring systems may be compared directly using normalized scores.
- If S is the (raw) score for a local alignment, the **normalized** score S' (in bits) is given by

$$S' = \frac{\lambda - \ln(\mathrm{K})}{\ln(2)}$$

  The parameters depend on the scoring system.
- **Statistically significant normalized score**,

$$S' > \log\left(\frac{N}{E}\right)$$

  where E-value = E, and N = size of search space.

Homologs: Orthologs & Paralogs

- **Homology:** Similarity due to common ancestry.
- **Orthologs:** Homologous sequences in different species that arose from a common ancestral gene during speciation; may or may not be responsible for a similar function.
  - **Paralogs:** Homologous sequences within a single species that arose by gene duplication.