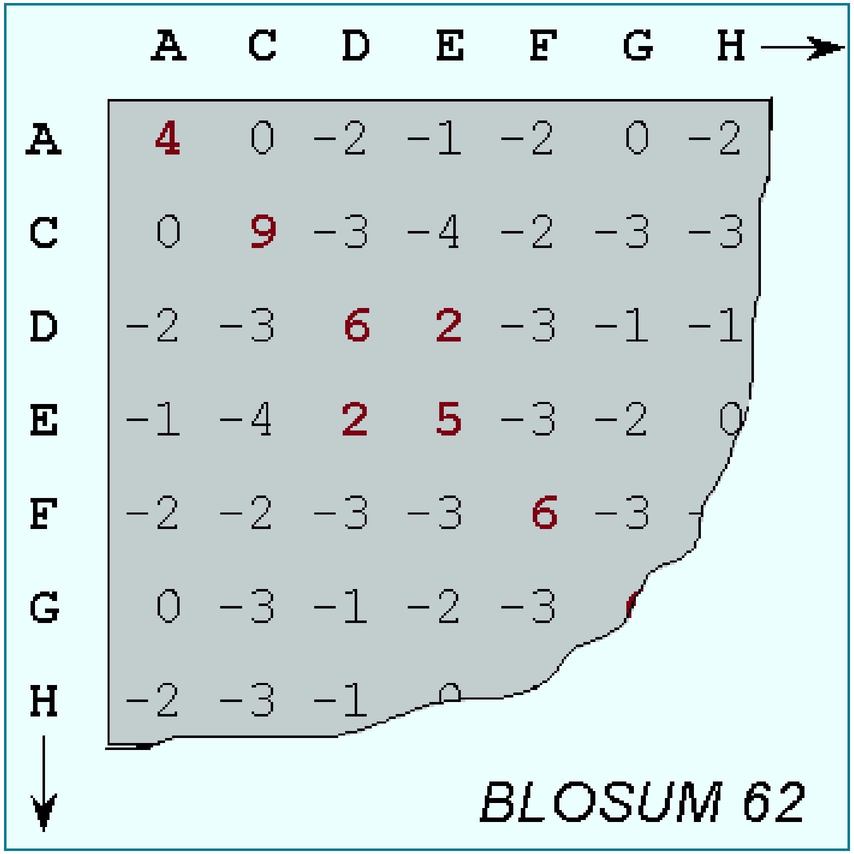
Alignment (Continued)

How to score mismatches?



	A	C	D	E	F	G	H
A	4	0	-2	-1	-2	0	-2
C	0	9	-3	-4	-2	-3	-3
D	-2	-3	6	2	-3	-1	-1
E	-1	-4	2	5	-3	-2	0
F	-2	-2	-3	-3	6	-3	-
G	0	-3	-1	-2	-3	6	
H	-2	-3	-1	0			

BLOSUM 62

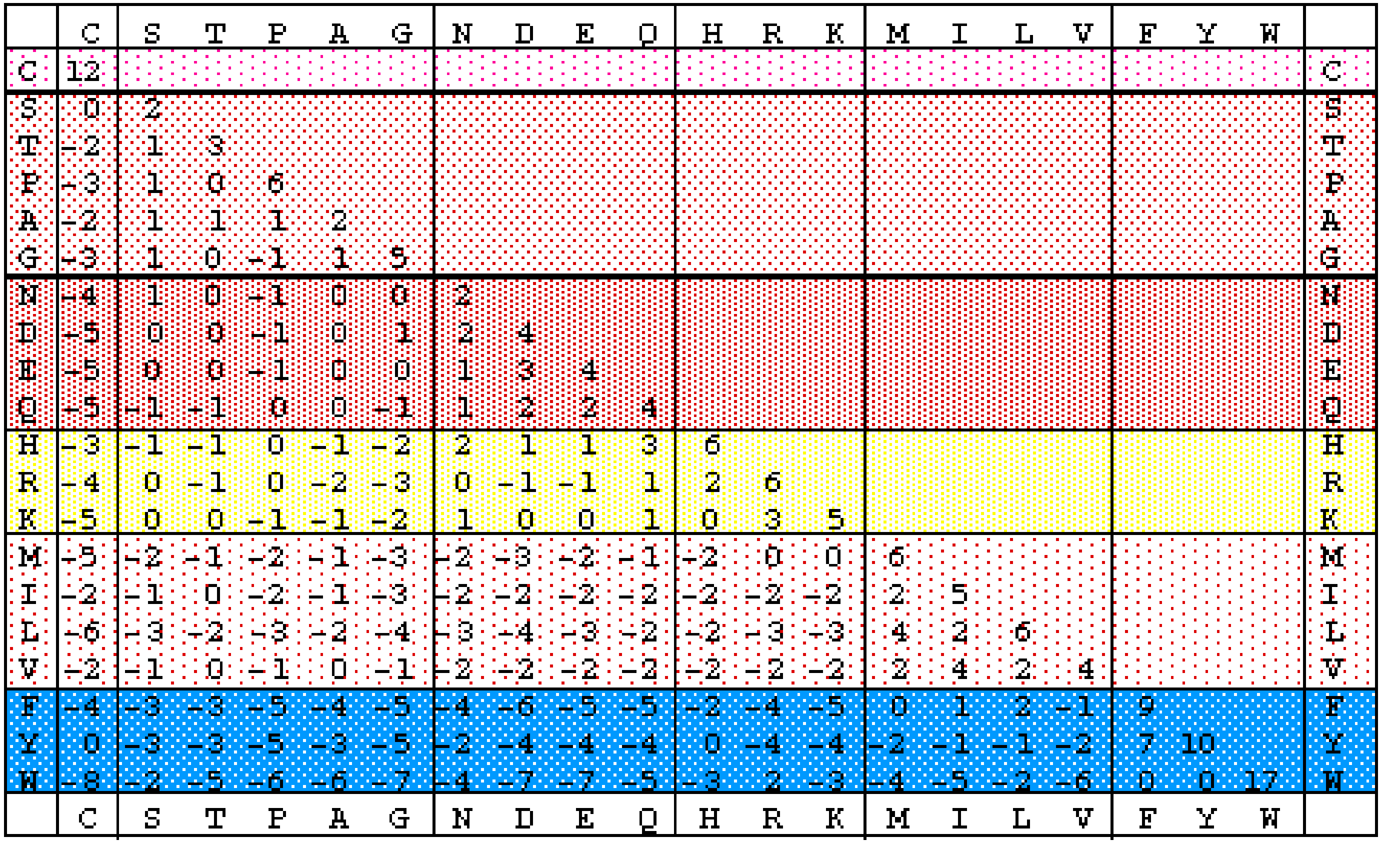
BLOSUM n Substitution Matrices

- For each amino acid pair a, b
  - For each BLOCK
    - Align all proteins in the BLOCK
    - Eliminate proteins that are more than n% identical
    - Count F(a), F(b), F(a,b)
    - Compute Log-odds Ratio

$$\log\left(\frac{F(a,b)}{F(a)F(b)}\right)$$

Alternative Substitution Matrices

PAM250



	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	

Point Accepted Mutations (PAM)

- PAM is a unit of evolutionary distance.
- Protein sequences A and B are 1 PAM unit apart if one is converted to the other with an average of 1 accepted point mutation per 100 amino acids.
- Point Mutation ⇔ Substitutions (No InDels)
- Accepted ⇔ incorporated into protein and passed onto progeny

True or False?

- If |A| = |B| = 400, and A and B are 1 PAM unit apart, then the expected number of differences between A and B is exactly 4.
- If |A| = |B|, and A and B are 100 PAM units apart, then they are expected to be different in every position.
- If A and B are 250 PAM units apart, then they are as distinct as a pair of random sequences.

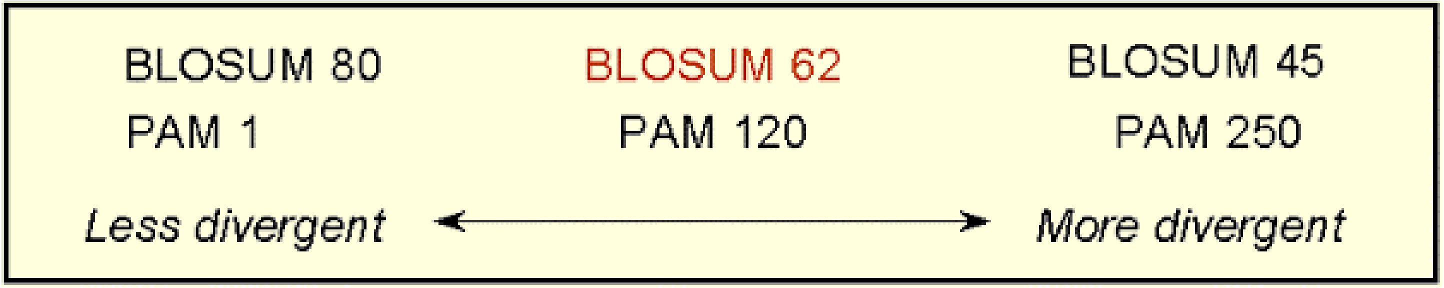
>15%

PAM Substitution Matrices

- Align very similar pairs of sequences (<15% difference).
- Identify and ignore InDels.
- For each amino acid pair (a,b) compute log-odds ratio:

$$\log\left(\frac{F(a,b)}{F(a)F(b)}\right)$$

PAM vs BLOSUM



Less divergent		More divergent
BLOSUM 80	BLOSUM 62	BLOSUM 45
PAM 1	PAM 120	PAM 250

Less divergent ⟷ *More divergent*

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 for “Bright Angel Trail”

- Bright
  - “Bright Futures” (health initiative), “Bright Lights Film Journal”, “The Bright Side” (crisis site), “The Armory of Bright Blades” (knife store), “Bright Ideas” (home improvement site), “Bright Angel Trail”, …
- Angel
  - “Angel of Fashion Award”, “Angel Island State Park”, “Recursive Angel” (poetry), “Angel Flight West” (free medical transportation), “Bright Angel Trail”, …
- Trail
  - “Appalachian Trail”, “Oregon Trail”, “Trail of Tears”, “Bright Angel Trail”, …

FASTA 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 Strategy

- 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 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)
- http://www.ncbi.nlm.nih.gov/BLAST/ (BLAST)

Perl: Practical Extraction & Report Language

- Created by Larry Wall, early 90s
- Portable, "glue" language for interfacing C/Fortran code, WWW/CGI, graphics, numerical analysis and much more
- Easy to use and extensible
- OOP support, simple databases, simple data structures.
- From interpreted to compiled
- high-level features, and relieves you from manual memory management, segmentation faults, bus errors, most portability problems, etc, etc.
- Competitors: Python, Tcl, Java

Perl Features

- Perl – many features
  - Bit Operations, Pattern Matching, Subroutines, Packages & Modules, Objects, Interprocess Communication, Threads, Compiling, Process control
- Competitors to Perl: Python, Tcl, Java

BioPerl

- Routines for handling biosequence and alignment data.
- Why? Human Genome Project: Same project, same data. different data formats! Different input formats. Different output formats for comparable utility programs.
- BioPerl was useful to interchange data and meaningfully exchange results. “Perl Saved the Human Genome Project”
- Many routine tasks automated using BioPerl.
- String manipulations (string operations: substring, match, etc.; handling string data: names, annotations, comments, bibliographical references; regular expression operations)
- Modular: modules in any language

Sequencing Project

- a trace editor to analyze, and display the short DNA read chromatograms from DNA sequencing machines.
- a read assembler, to find overlaps between the reads and assemble them together into long contiguous sections.
- an assembly editor, to view the assemblies and make changes in places where the assembler went wrong.
- a database to keep track of it all.

Managing a Large Project

- Devise a common data exchange format.
- Use modules that have already been developed.
- Write Perl scripts to convert to and from common data exchange format.
- Write Perl scripts to “glue” it all together.

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

Miscellaneous

- pTk – to enable building Perl-driven GUIs for X-Window systems.
- BioJava
- BioPython
- The BioCORBA Project provides an object-oriented, language neutral, platform-independent method for describing and solving bioinformatics problems.

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!