Logistics

- Problem set 3 available later today
- 7Eleven-1 due tomorrow
- Midterm Exam, October 10th
  - Covers sequence alignment, models of sequence substitution
    - See Study Guide on Syllabus page
  - Lectures through Sept. 19 (and review on Sept 21)
  - Closed book
  - Two pages of notes

Two widely used families of Amino Acid Substitution Matrices Parameterized for evolutionary divergence (*N*)

- PAM matrices, Dayhoff *et al*, 1978
- BLOSUM (Block Sum) matrices, Hennikoff &Hennikoff, 1991

Amino Acid Substitution Matrices Parameterized for evolutionary divergence (*N*)

Overall strategy for both PAM and BLOSUM

1. Trusted amino acid alignments
2. Obtain amino acid pair counts $\left(A_{xy}^{N}\right)$with corrections for
   - Evolutionary divergence
   - Sample biases
3. Estimate substitution frequencies, q_{xy}^{N}, from pair counts, A_{xy}^{N}
4. Log odds substitution matrix: $S^N[x,y] = c \log \frac{q^N_{xy}}{p_x p_y}$

Log odds substitution matrices

Two sequences have *N* PAMs divergence , if, on average, *N* amino acid replacements per 100 residues occurred since their separation

$$S^N[x,y] = c \log \frac{q^N_{xy}}{p_x p_y}$$

Frequency of x aligned with y in sequences with divergence *N*

Frequency of x aligned with y in "random" sequences

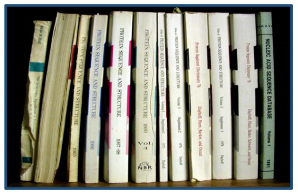
Scaling constant

Two widely used families of Amino Acid Substitution Matrices Parameterized for evolutionary divergence (*N*)

- PAM matrices, Dayhoff *et al*, 1978
- BLOSUM (Block Sum) matrices, Hennikoff &Hennikoff, 1991

PAM Matrices

Atlas of Protein Sequence & Structure
1965 - 1978



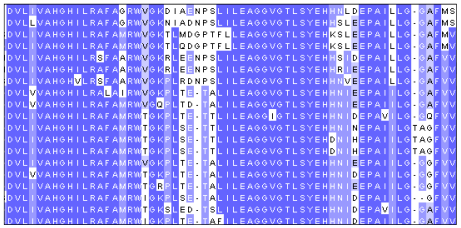
Examined 1572 changes in 71 groups of closely related proteins

Margaret Dayhoff
PhD in Chemistry, 47
Watson Computing Lab
Fellow 47 - 48

1. Trusted multiple sequence alignments

Examined 1572 changes in 71 groups of closely related proteins

At least 85% identical



2. Obtain amino acid pair counts (A_{xy}) with corrections for evolutionary divergence and sample biases

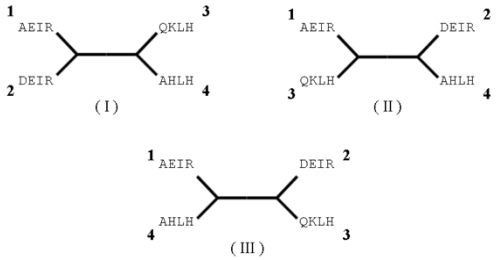
Counting amino acid pairs on a tree:

For each unrooted tree with *k* leaves

- Select the tree(s) that require the fewest substitutions to explain the data
- Count amino acid pairs on the branches of the tree

Suppose we have an alignment of four sequences. There are 3 hypotheses (i.e., 3 unrooted trees) for their evolutionary relationships

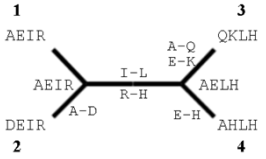
1. AEIR
2. DEIR
3. QKLH
4. AHLH



How to select the tree(s) that require the fewest substitutions to explain the data...

For a given a tree, assign labels to internal nodes that minimize the number of changes required to explain the data

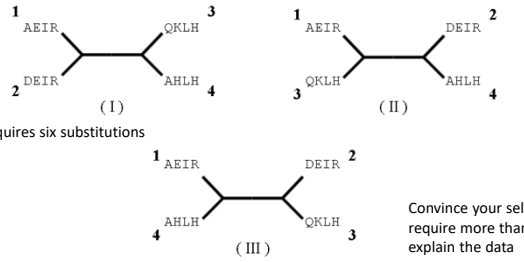
1. AEIR
2. DEIR
3. QKLH
4. AHLH



There may be more than one set of labels that satisfies this criterion

Tree I requires six substitutions

Select the most parsimonious tree; i.e., the tree that requires the fewest substitutions to explain the data.



Tree I requires six substitutions

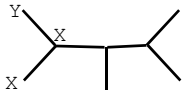
Convince your self that Trees II and III require more than six substitutions to explain the data

2. Obtain amino acid pair counts (A_{xy}) with corrections for evolutionary divergence and sample biases

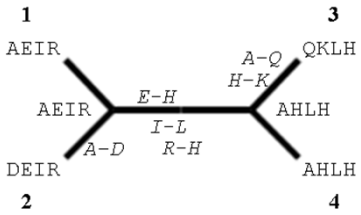
Counting amino acid pairs on a tree:

For each unrooted tree with *k* leaves

- Select the tree(s) that require the fewest substitutions to explain the data
- ➢ Count amino acid pairs on the branches of the tree
- For each branch,
  - if labeled x ——— y, $A^N_{xy} = A^N_{xy}+1$ and $A^N_{yx} = A^N_{yx}+1$
  - if labeled x ——— x, $A^N_{xx} = A^N_{xx}+2$



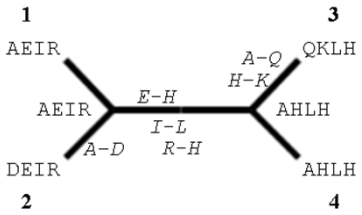
Impact of counting pairs on a tree: some examples



1. **AEIR**
2. **DEIR**
3. **QKLH**
4. **AHLH**

$A_{DQ} = 0$ $A_{DQ} = 1$

Impact of counting pairs on a tree: some examples

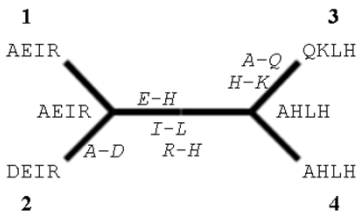


1. **AEIR**
2. **DEIR**
3. **QKLH**
4. **AHLH**

$A_{DQ} = 0$ $A_{DQ} = 1$

$A_{EK} = 0$ $A_{EK} = 1$

Impact of counting pairs on a tree: some examples



1. **AEIR**
2. **DEIR**
3. **QKLH**
4. **AHLH**

$A_{DQ} = 0$ $A_{DQ} = 1$

$A_{EK} = 0$ $A_{EK} = 1$

$A_{IL} = 1$ $A_{IL} = 4$

3. Estimate substitution frequencies, q_{xy}^N, from pair counts, A_{xy}

- Markov model with 20 states (A, C, D, E ... Y)
- Estimate 1 PAM transition matrix P^1 from A_{xy}
- N-PAM transition matrix: $P^1 = (P^1)^N$
- $q_{xy}^N = p_x P_{xy}^N$
- $S^N[x,y] = c \log \frac{q_{xy}^N}{p_x p_y}$

Is P_{xy}^N a symmetric matrix? No. (Check this algebraically).

Is $S^N[x,y]$ a symmetric matrix? Yes (Check this algebraically).

Two widely used families of Amino Acid Substitution Matrices Parameterized for evolutionary divergence (*N*)

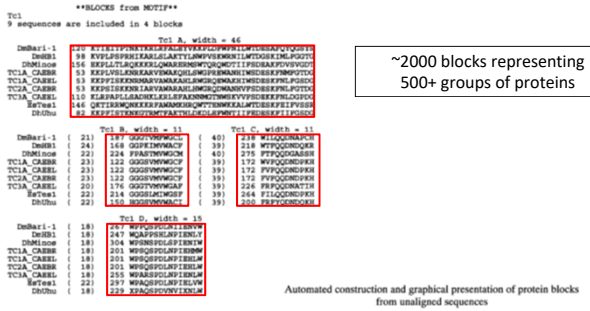
- PAM matrices, Dayhoff *et al*, 1978

➢BLOSUM (Block Sum) matrices, Hennikoff &Hennikoff, 1991

BLOSUM Matrices

- Trusted data
  - 2000 blocks of conserved regions in ~500 groups of proteins
- Count amino acid pairs: A_{xy}^N
  - **Parameterize by evolutionary distance, *N***
  - **Correct for sample bias**
- Calculate amino acid frequencies:
  - Related pairs: q_{xy}^N
  - Background pair frequencies calculated from blocks: E_{xy}
- Log likelihood scoring matrix
  - $S^N = 2\ \log_2 \frac{q_{xy}^N}{E_{xy}}$

1. Trusted multiple sequence alignments


BLOCKS from MOTIF

Tc1
9 sequences are included in 4 blocks

Tc1 A, width = 46

Tc1 B, width = 11 Tc1 C, width = 11

Tc1 D, width = 15

Automated construction and graphical presentation of protein blocks from unaligned sequences

Steven Henikoff [a,b,*], Jorja G. Henikoff [a,1], William J. Alford [a,2], Shmuel Pietrokovski [a,3]

~2000 blocks representing 500+ groups of proteins

2. Count amino acid pairs: A_{xy}^N

Parameterize by evolutionary distance, *N*

Correct for sample bias

- Cluster sequences such that if *s1* and *s2* are in different clusters, then *identity(s1, s2) < N%*
- Count amino acid pairs in *s1* aligned with *s2* *only* if *s1* and *s2* are in different clusters
- Normalize for cluster size

An example...

BLOSUM clustering example

1: KKRK
2: KKKK
3: KNRN
4: NRNR
5: KNKN
6: KRNR

Percent Sequence Identity

	[2]	[3]	[4]	[5]	[6]
[1]	0.75	0.5	0	0.25	0.25
[2]		0.25	0	0.5	0.25
[3]			0	0.75	0.25
[4]				0	0.75
[5]					0.25

Unclustered sequences: Every sequence is at least 25% identical

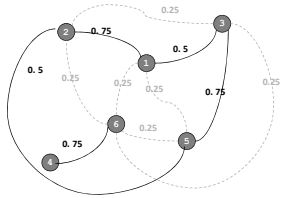
Percent Sequence Identity

	[2]	[3]	[4]	[5]	[6]
[1]	**0.75**	**0.5**	0	0.25	0.25
[2]		0.25	0	**0.5**	0.25
[3]			0	**0.75**	0.25
[4]				0	**0.75**
[5]					0.25

1: KKRK
2: KKKK
3: KNRN
5: KNKN

< 45% identical

4: NRNR
6: KRNR



< 65% id

1: KKRK
2: KKKK

< 65% id

3: KNRN
5: KNKN

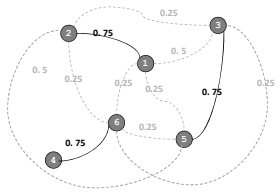
< 65% id

4: NRNR
6: KRNR

< 65% id

Percent Sequence Identity

	[2]	[3]	[4]	[5]	[6]
[1]	**0.75**	0.5	0	0.25	0.25
[2]		0.25	0	0.5	0.25
[3]			0	**0.75**	0.25
[4]				0	**0.75**
[5]					0.25



Two widely used families of Amino Acid Substitution Matrices Parameterized for evolutionary divergence (*N*)

- PAM matrices, Dayhoff *et al*, 1978
- BLOSUM (Block Sum) matrices, Hennikoff &Hennikoff, 1991

Similarities and differences between PAM and BLOSUM

	PAM	BLOSUM
Evolutionary model	Explicit evolutionary model	None
Data	Full length MSAs of closely related sequences.	Conserved blocks. i.e., ungapped local MSAs
Bias correction	Trees	Clustering
Multiple substitutions	Markov model: $P^n = (P^1)^n$	Implicitly represented in data (clustering)
Evolutionary distance	Markov model: $P^n = (P^1)^n$	Clustering
Matrices	Transition and log odds scoring matrices	Log odds scoring matrix only.
Parameter *n*	Distance increases with *n*	Distance decreases with *n*
Biophysical properties	Derived indirectly from data	Derived indirectly from data

Comparing PAM and BLOSUM matrices

Less divergent

PAM	Sequence identity	BLOSUM
20	83%	
30		
60	63%	
70		
100	43%	90
120	38%	80
160	30%	60
200	25%	50
250	20%	45

More divergent