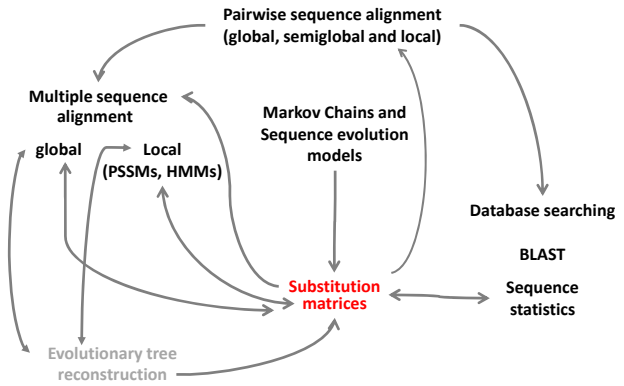
Logistics

- 7Eleven-1 due Fri, 11:59pm, Sep 27th
- Problem set 3 due Sat, 6pm, Sep 28th
- In-class Exam, October 1st
  - Covers lectures through Sept. 19
    - Sequence alignment
    - Models of sequence substitution
    - Log likelihood ratios
  - Closed book
  - Two pages of notes

Solution sets posted Saturday

No late assignments will be accepted once solution sets are posted

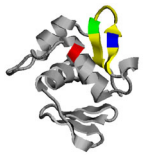


Two widely used families of Amino Acid Substitution Matrices Parameterized for evolutionary divergence (*N*)

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

Substitution Matrices do not model gaps
Why not treat gaps as a symbol in the alphabet?

Sequence evolution models and substitution matrices model events at one site



Problem:

A gap may be the result of the gain/loss of 10 residues in a single event.

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 $(A_{xy}^N)$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

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

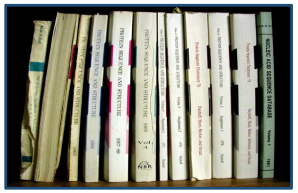
- q^N_{xy}: Frequency of x aligned with y in sequences with divergence *N*
- $p_x p_y$: Frequency of x aligned with y in "random" sequences
- c: Scaling constant

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

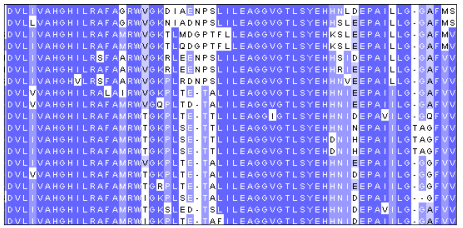
Evolutionary divergence (amino acids)

- PAM: Percent Accepted Mutation
  - *Accepted Mutations* are mutations that are retained and passed on to future generations
- We say the divergence between two sequences is *N* PAMs, if, on average, *N* amino acid replacements per 100 residues (including multiple substitutions) occurred since their separation.

1. Trusted multiple sequence alignments

Examined 1572 changes in 71 groups of closely related proteins

At least 85% identical



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** (A_{xy}^N) 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: $S^N[x,y] = c \log \frac{q_{xy}^N}{p_x p_y}$

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

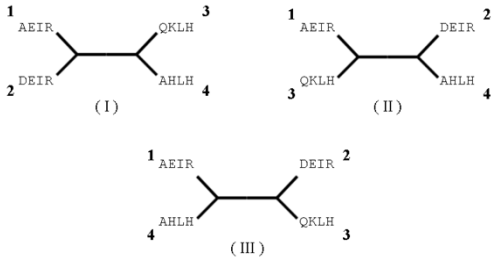
Counting amino acid pairs on a tree:

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

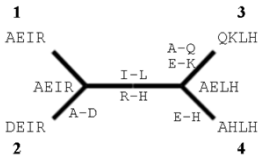
1. AEIR
2. DEIR
3. QKLH
4. AHLH



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



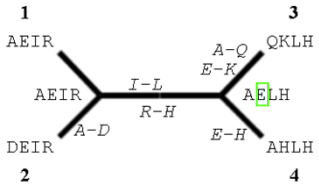
There may be more than one set of labels that satisfies this criterion

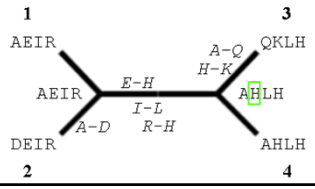
Tree I requires six substitutions

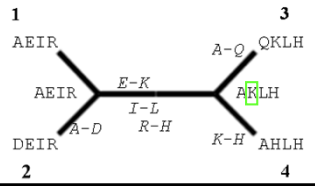
Three most parsimonious ways to assign internal labels to Tree (I)

In each case, six substitutions are required to explain the data.

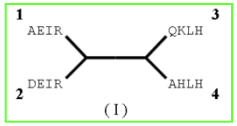
1. **AEIR**
2. **DEIR**
3. **QKLH**
4. **AHLH**



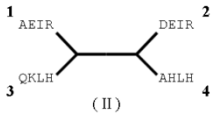


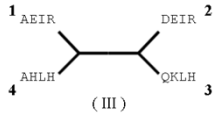


Select the most parsimonious tree; i.e., the tree that requires the fewest substitutions to explain the data.



Tree I requires six substitutions





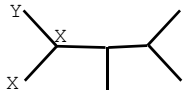
Convince your self that Trees II and III require more than six substitutions to explain the data

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

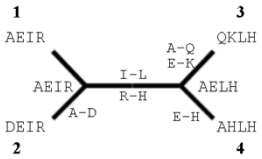
Counting amino acid pairs on a tree:

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**

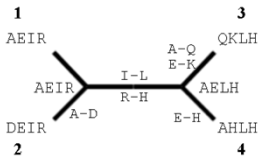
Counts in along tree branches

$A_{DQ} = 0$

$A_{DQ} = 1$

Counts in the multiple sequence alignment

Impact of counting pairs on a tree: some examples



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

Counts in along tree branches

$A_{DQ} = 0$

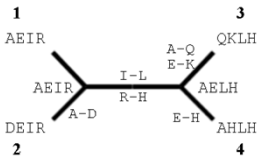
$A_{EK} = 0$

$A_{DQ} = 1$

$A_{EK} = 1$

Counts in the multiple sequence alignment

Impact of counting pairs on a tree: some examples



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

Counts in along tree branches

$A_{DQ} = 0$

$A_{EK} = 0$

$A_{IL} = 1$

$A_{DQ} = 1$

$A_{EK} = 1$

$A_{IL} = 4$

Counts in the multiple sequence alignment

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

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: $P^1 = (P^1)^N$
- $q_{xy}^{N} = p_x P_{xy}^{N}$
- $S^N[x,y] = c \log \frac{q^N_{xy}}{p_x p_y}$

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

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