

- 1) Pick a date for your presentation.  
Your writeup is due on the day of  
your presentation.
- 2) Candidate presentation dates are  
marked in the course outline  
(online)
- 3) Pick a paper or a project topic.
- 4) THANKS.

# Alignment scoring statistics and scoring matrices

Analysis of Biological Sequences 140.638

# Two sequences

ACTTCGCGCGAT and ACGTGGTTGATG

Possible alignments:

ACTTCGCGCGAT

| | | |

ACGTGGTTGATG

Which is the better alignment?

How can we tell?

ACTTCGCGCGAT-

| | | | | |

ACGTGGTT-GATG

We need a scoring function. For example, match = 2, mismatch = -1, gap = -2. Then aln #1 scores 0 and aln #2 scores 6.

# Different scoring rules -> different alignments

Dynamic programming: global alignment

Match=5, mismatch = -4, gap w = -7

```
G C T G G A A G G C A - T
| |   |   |   | | |   |
G C - A G - A - G C A C T
```

$$\text{Score: } 8*5 + 1*-4 + 4*-7 = 8$$

# Different scoring rules -> different alignments

Dynamic programming: global alignment

Match=5, mismatch = -4, gap w = -2

```
G C T G G A A G - G C A - T
| |           | |   | | |   |
G C - - - - A G A G C A C T
```

$$\text{Score} = 8 * 5 + 0 + 6 * -2 = 28$$

# Scoring rules/matrices

- Why are they important?
  - Choice of scoring rule can dramatically influence the sequence alignments obtained and, therefore, the analysis being done
  - Different scoring matrices have been developed for different situations; using the wrong one can make a big difference.
- What do they mean?
  - Scoring matrices implicitly represent a particular theory of evolution
  - Your goal is to figure out whether the two sequences have a common ancestor
  - Elements of the matrices specify relationships between amino acid residues or nucleotides

# Substitution Matrices

- We need scoring terms for each aligned residue pair
- Models: Random model (R): letter a occurs with frequency  $q_a$

Sequence x: x1 x2 x3 x4  
                  | | | |  
Sequence y: y1 y2 y3 y4

$$P(x,y|R) = \prod_i q_{x_i} q_{y_i}$$

# Substitution Matrices— random model

$$P(x,y|R) = \prod_i q_{x_i} q_{y_i}$$

$$x = \text{ACCTGCC} \quad p(A) = 0.2$$

$$y = \text{ACGTCCA} \quad p(T) = 0.2$$

$$\text{ACCTGCC} \quad p(C) = 0.3$$

$$\begin{array}{c} | | | | \\ \text{ACGTCCA} \end{array} \quad p(G) = 0.3$$

$$P(x,y|R) = p(A)^2 p(C)^2 p(C) p(G) p(T)^2 p(G) p(C) p(C)^2 p(C) p(A) = 6.29 * 10^{-9}$$

# Substitution Matrices— match model

- Models: Match model (M): aligned pairs of residues have joint probability  $p_{ab}$
- $p_{ab}$  = probability that a and b came from common ancestor residue

Sequence x: x1 x2 x3 x4  
                  | | | |  
Sequence y: y1 y2 y3 y4

$$P(\mathbf{x}, \mathbf{y} | M) = \prod_i p_{x_i y_i}$$

# Substitution Matrices

Odds ratio:

$$\frac{P(x, y|M)}{P(x, y|R)} = \frac{\prod_i p_{x_i y_i}}{\prod_i q_{x_i} q_{y_i}}$$
$$= \prod_i \left( \frac{p_{x_i y_i}}{q_{x_i} q_{y_i}} \right)$$

# Substitution Matrices

$$\prod_i \left( \frac{p_{x_i y_i}}{q_{x_i} q_{y_i}} \right)$$

Change to a sum by using logarithms . . .

$$\text{Score} = \sum_i \log \left( \frac{p_{x_i y_i}}{q_{x_i} q_{y_i}} \right) = \sum_i s(a, b)_i$$

Where  $s(a,b)$  is just the score of aligning a residue of type  $a$  to a residue of type  $b$

# Substitution matrices

Log-odds ratio  $\rightarrow$  log likelihood ratio that the pair (a,b) is related vs unrelated (depends on scoring matrices)

The alignment score is the log likelihood that the sequences have common ancestry

# Two major scoring matrices

- PAM = accepted point mutation
  - 71 trees with 1572 accepted mutations, sequences with >85% identity
  - “accepted” means new amino acid doesn’t disrupt the protein’s function too severely
  - PAM1 means average of 1% change over all amino acids
  - 1 PAM = 10my evolutionary distance
- BLOSUM = Blocks substitution matrices
  - Based on BLOCKS database (Henikoff & Henikoff, 1992) of over 2000 conserved amino acid patterns in over 500 proteins

# PAM matrices

- Each matrix describes changes expected for a given period of evolutionary time (measured by expected similarity of proteins)
- Count # of changes to each amino acid in the phylogenetic group and divide by the “exposure to mutation” of the residue
- Exposure to mutation = frequency of occurrence of amino acid \* #amino acid changes in the group/100 sites

# PAM matrices

- Amino acid changes are modeled by a Markov process, so each mutation is independent of previous mutations and of adjacent positions
  - This means that we can calculate the matrices for more distantly related proteins by multiplying matrices for closely related proteins (PAM 250 = PAM1 multiplied by itself 250 times)
  - PAM 250 = 250% change over 2500 my. ~20% similarity at this level; shown to be best for proteins of 14-27% similarity

# Matrix multiplication

$$\begin{pmatrix} X1 & X2 \\ Y1 & Y2 \end{pmatrix} \times \begin{pmatrix} X1 & X2 \\ Y1 & Y2 \end{pmatrix}$$

$$= \begin{pmatrix} X1*X1 + X2*Y1 & X1*X2 + X2*Y2 \\ Y1*X1 + Y2*Y1 & Y1*X2 + Y2*Y2 \end{pmatrix}$$

# PAM matrices

- PAM120: 40% similarity
- PAM80: 50% similarity
- PAM60: 60% similarity
  
- Simulations have confirmed these numbers
  
- Choosing the best PAM matrix: ungapped alignment score will be highest when the correct matrix is used.

# PAM matrices—assumptions

- $P(X \rightarrow Y) = P(Y \rightarrow X)$
- $P(X \rightarrow Z \rightarrow Y)$  is low in a single PAM period
- Markov model/independence
- All sequences have similar amino acid composition

# An example: obtaining the PAM250 score for Tyr $\leftrightarrow$ Phe (from Mount book)

- Original PAM data: 1572 observed amino acid changes, 260 were between Phe and Tyr
- $260/1572$  is multiplied by relative mutability of Phe and by the pair exchange frequency  $\rightarrow$  **mutation probability score**
- **Relative mutability** = chance that the amino acid will change (Dayhoff used # times observed to change)
- **Pair exchange frequency** = fraction of Phe  $\rightarrow$  Tyr/all Phe mutations
- Normalize to a sum of 1% probability of any amino acid change, then take log odds  $\rightarrow$  PAM1

## Example: Tyr<->Phe

260/1572 \* mutability \* pair exchange frequency THEN  
normalize to 1% change over whole matrix

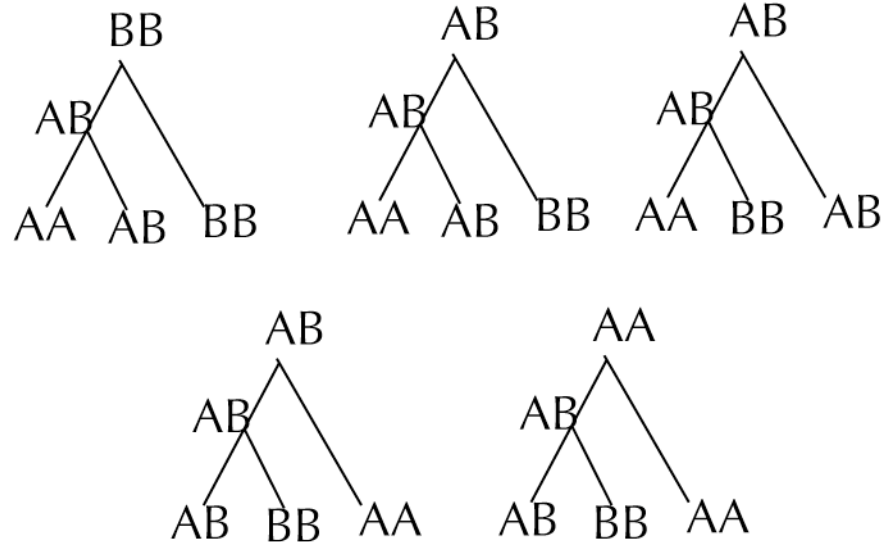
Normalized probability scores for some Phe mutations

aa change	PAM1	PAM250
F->A	0.0002	0.04
F->R	0.0001	0.01
F->F	0.9946	0.32
F->Y	0.0021	0.15

# PAM construction example

Phylogeny-based: Sequences AA, AB, BB

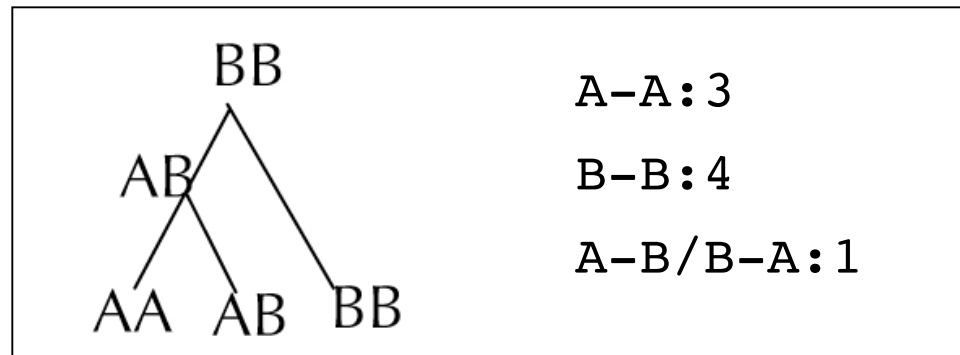
Construct all possible phylogenetic trees relating the three sequences



# PAM construction example

- Count up the times that each residue is aligned to itself or the other residue

A-A: 30  
B-B: 30  
A-B/B-A: 10



Divide by the number of trees (5) to get the matrix of counts

	A	B
A	6	
B	2	6

# PAM construction example

- Mutability:

Likelihood that residue  $j$  will change in the given evolutionary interval

	A	B
A	6	
B	2	6

$$m_j = 1 - \frac{A_{jj}}{\sum_{i=1}^{20} A_{ij}}$$

$$m_a = 1 - (6/8) = 1/4$$

$$m_b = 1 - (6/8) = 1/4$$

# PAM construction example

- $p_j$  is the probability that  $j$  occurs randomly in a sequence

$$P_j = \frac{\sum_{i=1}^{20} A_{ij}}{\sum_{i=1}^{20} \sum_{k=1}^{20} A_{ik}}$$

All the ways to mutate to  $j$

All mutations

$\sum p_j m_j$  is the total mutation rate of all amino acids

Then scale so that  $\lambda * \sum p_j m_j = 1\%$

# BLOSUM

- Henikoff & Henikoff used PROTOMAT program to create BLOCKS database from Prosite catalog
- PROTOMAT looks for **A1-d1-A2-d2-A3** where **A1**, **A2**, **A3** are conserved residues and  $d1, d2 < 25$  residue intervening sequence
- These initial patterns are consolidated into larger patterns by PROTOMAT
- Sequences above a similarity threshold are clustered into families (% threshold -> BLOSUM #)

# BLOSUM construction

## 1. Count mutations

VVAPV

AAAPA

PVAPV

PAAAV

$$N_{AA} = 0 + 1 + (4*3)/2 + 0 + 0 = 7$$

$$N_{VV} = 0 + 1 + 0 + 0 + (3*2)/2 = 4$$

$$N_{PP} = 1 + 0 + 0 + (3*2)/2 + 0 = 4$$

$$N_{AV} = N_{VA} = 1 + 2*2 + 0 + 0 + 3 = 8$$

$$N_{AP} = N_{PA} = 2 + 0 + 0 + 3 + 0 = 5$$

$$N_{PV} = N_{VP} = 2 + 0 + 0 + 0 + 0 = 2$$

# BLOSUM construction

## 2. Tallying mutation frequencies

$q_{ij}$  = # times amino acid  $j$  mutates to amino acid  $i$

Since we don't know ancestry, each mutation gets entered twice

VVAPV  
AAAPA  
PVAPV  
PAAAV

	A	V	P
A	14	8	5
V	8	8	2
P	5	2	8

# BLOSUM construction

## 3. Matrix of mutation probabilities

- Create probabilities from mutation frequencies

VVAPV  
 AAAPA  
 PVAPV  
 PAAAV

	A	V	P
A	14	8	5
V	8	8	2
P	5	2	8

$p_{ij}$	A	V	P
A	14/60	8/60	5/60
V	8/60	8/60	2/60
P	5/60	2/60	8/60

# BLOSUM construction

## 4. Calculate abundance of each residue (Marginal probability)

$p_i$  is the marginal probability, meaning the expected probability of occurrence of amino acid  $i$

	A	V	P	$P_i$
A	14/60	8/60	5/60	$27/60 = 9/20$
V	8/60	8/60	2/60	$18/60 = 6/20$
P	5/60	2/60	8/60	$15/60 = 5/20$
	9/20	6/20	5/20	1

# BLOSUM construction

## 5. Obtaining a BLOSUM matrix

- BLOSUM is a log-likelihood matrix:

$$S_{ij} = 2\log_2(p_{ij}/(p_i p_j))$$

$S_{ij}$	A	V	P
A	0.409		
V	-0.036	1.134	
P	-0.866	-2.34	2.19

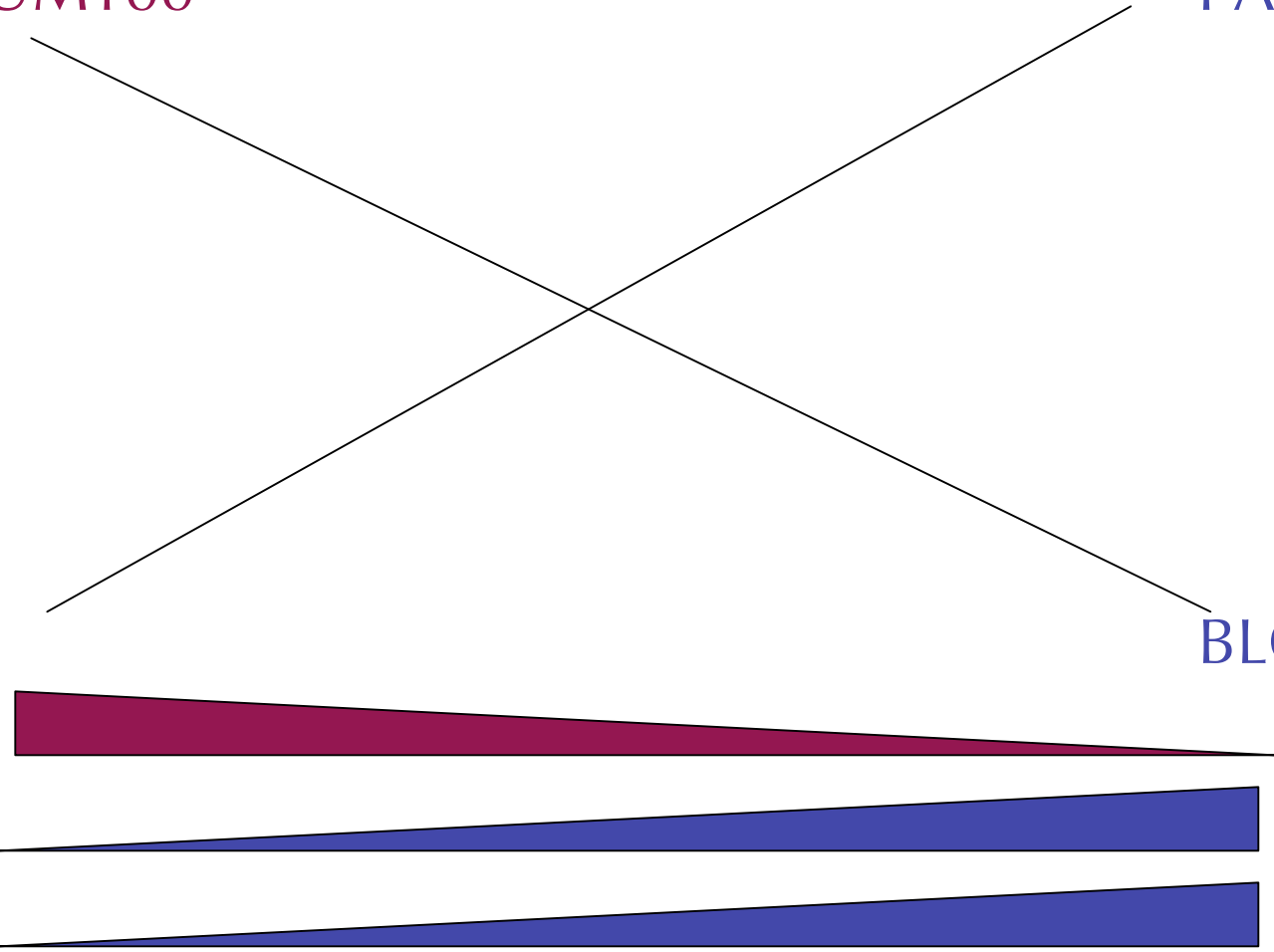
# Choice of matrix

BLOSUM100

PAM250

PAM1

BLOSUM30



Similarity  
Ev.distance  
Seq.length

# BLOSUM vs PAM

- BLOSUM: based on short conserved sequences (blocks)
  - Based on a range of evolutionary periods
  - Each matrix constructed separately
  - Indirectly accounts for interdependence of residues
  - Range of sequences, range of replacements
  - Overcounts related mutations
  - SCORING MATRIX
- PAM: evolutionary model
  - Based on extrapolation from a short evolutionary period
  - Errors in PAM1 are magnified through PAM250
  - Assumes Markov process
  - Many sequences depart from average composition
  - Rare replacements too infrequent to be represented accurately
  - SUBSTITUTION MATRIX

# Issues

- Both BLOSUM and PAM matrices are derived from small sets of sequences from biased databases
- Both types of matrices require aligned sequences for their construction
- Both types of matrices depend on global, ungapped alignments

# Other matrix types

- Simple identity
- Scored according to codon tables
- Chemical properties of amino acids
- 3D structural alignments
- Dipeptide-based
- Family-specific (kinase, transmembrane etc)
- DNA

# Determining correct alignment

- Can also use information theory/relative entropy to score alignments
- Average mutual information content per pair:

$$H = - \sum_i q_a q_b s(a, b)_i$$

Doesn't work for gapped alignments though.

# Gap penalties

- Until now we have talked about ungapped alignments and their properties
- Inclusion of gaps and gap penalties is necessary to obtain the best alignment

$$w_x = g + rx$$

g = opening penalty  
r = penalty for extending gap  
x = length of the gap

This is the affine gap penalty.

# Gap penalties

$$w_x = g + rx$$

g = opening penalty

r = penalty for extending gap

x = length of the gap

Why? Basic idea is that once two sequences have even a single insertion or deletion at a site, that site does not diverge significantly more with more indels.

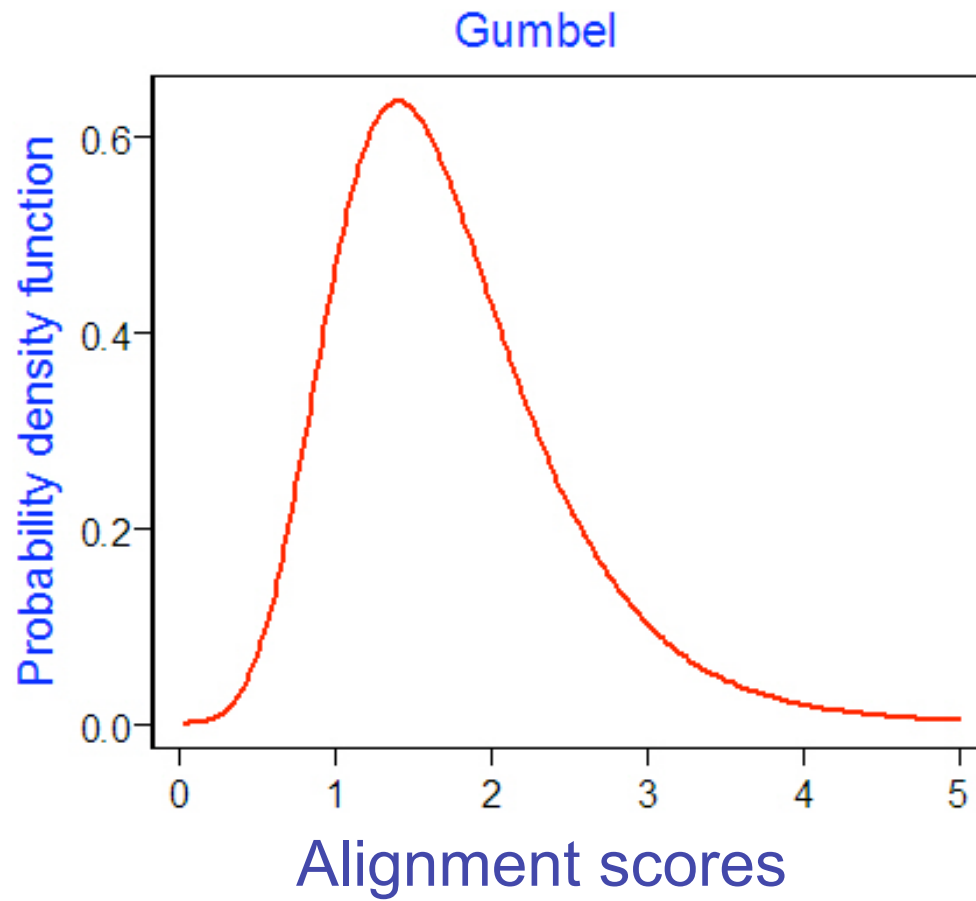
```
Query: RYALAIFFDD--RD-----ELG 114
       RYALAIFF + +D          ELG
Sbjct: RYALAIFFLNRYRKDGNFGLQELG 126

Query: RYALAIFFDDR-----DELG 114
       RYALAIFF +          ELG
Sbjct: RYALAIFFLNRYRKDGNFGLQELG 126
```

# Assessing the significance of sequence alignments

- Local alignments are rarely produced between random sequences
- Global alignments are readily produced between random sequences -> harder to assess the significance of a global alignment
- Can approximate a score by shuffling the sequences and realigning but this can be misleading
- Alignment scores follow the extreme value Gumbel distribution, not a normal distribution

# Assessing the significance of sequence alignments



# The Bayesian Approach

$P(M|x,y)$  = probability that the sequences are related  
(the model is correct, given the sequences)

$P(R) = 1 - P(M)$  where  $P(M)$  is the prior probability that the sequences are related and  $P(R)$  is the prior probability that the random model is correct (prior = before we see the data)

# The Bayesian Approach

$$\begin{aligned} P(M|x,y) &= \frac{P(x,y|M)P(M)}{P(x,y)} \\ &= \frac{P(x,y|M)P(M)}{P(x,y|M)P(M) + P(x,y|R)P(R)} \\ &= \frac{P(x,y|M)P(M)/P(x,y|R)P(R)}{1 + P(x,y|M)P(M)/P(x,y|R)P(R)} \end{aligned}$$

This is the probability that the match model is correct.

# The Bayesian Approach

- Need a specific mutational model a priori
- Computationally much more tractable
- BALSAs (Bayesian Algorithm for Local Sequence Alignment): creates pairwise ungapped alignments by sliding sequences to find optimal blocks
- BALSAs is also used for database searches