Sequencing Alignment II

Lectures 17 – Nov 23, 2011
CSE 527 Computational Biology, Fall 2011

Instructor: Su-In Lee
TA: Christopher Miles

Monday & Wednesday 12:00-1:20
Johnson Hall (JHN) 022

1

Needleman-Wunsch Algorithm

- Key idea: build up an optimal alignment using previous solutions for optimal alignments of smaller subsequences.
- Optimal align of S[1], ..., S[i] vs T[1], ..., T[j]:

$$\begin{bmatrix} \boxed{\begin{matrix}\sim\sim\sim\sim \\ \sim\sim\sim\sim\end{matrix}} & \begin{matrix} S[i] \\ T[j] \end{matrix} \end{bmatrix}, \quad \begin{bmatrix} \boxed{\begin{matrix}\sim\sim\sim\sim \\ \sim\sim\sim\sim\end{matrix}} & \begin{matrix} S[i] \\ - \end{matrix} \end{bmatrix}, \text{ or } \begin{bmatrix} \boxed{\begin{matrix}\sim\sim\sim\sim \\ \sim\sim\sim\sim\end{matrix}} & \begin{matrix} - \\ T[j] \end{matrix} \end{bmatrix}$$

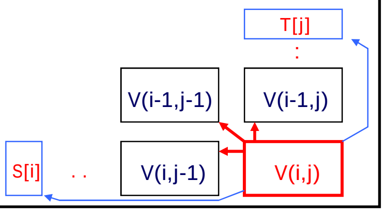
Opt align of $S_1 \ldots S_{i-1}$ & $T_1 \ldots T_{j-1}$
Value = V(i-1, j-1)

Opt align of $S_1 \ldots S_{i-1}$ & $T_1 \ldots T_j$
Value = V(i-1, j)

Opt align of $S_1 \ldots S_i$ & $T_1 \ldots T_{j-1}$
Value = V(i, j-1)

$$V(i,j) = \max \begin{Bmatrix} V(i\text{-}1,j\text{-}1) + \sigma(S[i],T[j]) \\ V(i\text{-}1,j) \quad + \sigma(S[i],\ -\ ) \\ V(i,j\text{-}1) \quad + \sigma(\ -\ ,\ T[j]) \end{Bmatrix},$$

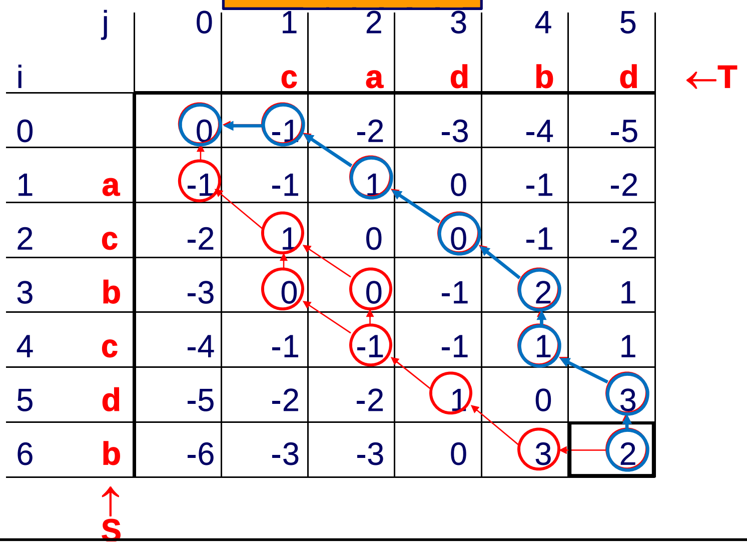
for all $1 \le i \le n,\ 1 \le j \le m.$



Align by Dynamic Programming

T': cadb-d-

S': -acbcdb



	j	0	1	2	3	4	5
i			c	a	d	b	d
0		0	-1	-2	-3	-4	-5
1	a	-1	-1	1	0	-1	-2
2	c	-2	1	0	0	-1	-2
3	b	-3	0	0	-1	2	1
4	c	-4	-1	-1	-1	1	1
5	d	-5	-2	-2	1	0	3
6	b	-6	-3	-3	0	3	2

←T

↑ S

Scoring Rules/Matrices

- How should σ be defined?
  - σ(A,G), σ(A,-), σ(A,-), etc?

$$V(i,j) = \max \left\{ \begin{array}{ll} V(i\text{-}1,j\text{-}1) & +\sigma(S[i],T[j]) \\ V(i\text{-}1,j) & +\sigma(S[i],\ -\ ) \\ V(i,j\text{-}1) & +\sigma(\ -\ ,\ T[j]) \end{array} \right\},$$

- Why are they important?
  - The choice of a scoring rule can strongly influence the outcome of sequence analysis

- What do they mean?
  - Scoring matrices implicitly represent a particular theory of evolution
  - Elements of the matrices specify the similarity of one residue to another

Refers to an amino acid

Outline: Scoring Alignments

- Probabilistic meaning
- Scoring matrices
  - PAM: scoring based on evolutionary statistics
  - BLOSUM: tuning to evolutionary conservation
- Gaps revisited

Probabilistic Interpretation

```
X: TCCAGGTG-GAT
   | || ||| | |
Y: TGCAAGTGCG-T
```

Chance or true homology?

Sharing a common ancestor

Likelihood Ratio

```
X: TCCAGGTG-GAT
   | || ||| | |
Y: TGCAAGTGCG-T
```

$$\frac{\text{Pr(Data|Homology)}}{\text{Pr(Data|Chance)}}$$

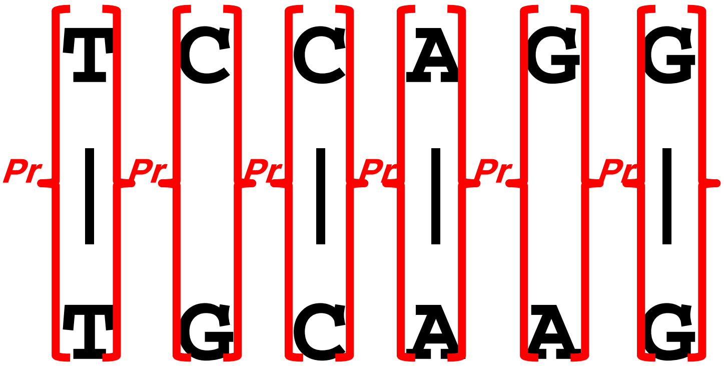
Pr(Data | Chance)

Given an alignment between TCCAGG and TGCAAG,

```
Pr(T)Pr(C)Pr(C)Pr(A)Pr(G)Pr(G)
   |        |     |        |
Pr(T)Pr(G)Pr(C)Pr(A)Pr(A)Pr(G)
```

Pr(Data | Homology)

Given an alignment between TCCAGG and TGCAAG,



Likelihood Ratio

X: T C C A G G

Y: T G C A A G

$$\frac{\text{Pr(Data | homology)}}{\text{Pr(Data | Chance)}}$$

$$= \prod_i \frac{\Pr(x_i y_i)}{\Pr(x_i)\Pr(y_i)}$$

Score: Log Likelihood Ratio

- The most commonly used alignment score of aligning two sequences is the log likelihood ratio of the alignment under two models
  - Common ancestry
  - By chance

$$Score = \log\left(\prod_i \frac{\Pr(x_i y_i)}{\Pr(x_i)\Pr(y_i)}\right) =$$

$$= \sum_i \log\left(\frac{\Pr(x_i y_i)}{\Pr(x_i)\Pr(y_i)}\right) = \sum_i s(x_i, y_i)$$

The *S* in a Scoring Matrix

(as log likelihood ratio)

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

Prob(x_i aligned with y_i at position i | common ancestry)

$$s(x_i, y_i) = \log\left(\frac{\Pr(x_i y_i)}{\Pr(x_i)\Pr(y_i)}\right)$$

Prob(x_i aligned with y_i at position i | by change)

How do we acquire the probabilities Pr(a), Pr(a,b)?

Making a Scoring Matrix

- Scoring matrices S are created based on biological evidence.
  - Alignments can be thought of as two sequences that differ due to mutations.
  - Some of these mutations have little effect on the protein’s function, therefore some penalties will be less harsh than others.

```
TCCAGGTG-GAT
| || ||| | |
TGCAAGTGCG-T
```

Scoring Matrix: Example

	A	R	N	K
A	5	-2	-1	-1
R	-	7	-1	3
N	-	-	7	0
K	-	-	-	6

```
AKRANR
KAAANK
```

-1 + (-1) + (-2) + 5 + 7 + 3 = 11

- Notice that although R (arginine) and K (Lysine) are different amino acids, they have a positive score.
- Why? They are both positively charged amino acids → will not greatly change function of protein.

Conservation

- Amino acid changes that tend to preserve the physical/ chemical properties of the original residue
  - Polar to polar
    - aspartate (D) → glutamate (E)
  - Nonpolar to nonpolar
    - alanine (A) → valine (V)
  - Similarly behaving residues
    - leucine (L) to isoleucine (I)

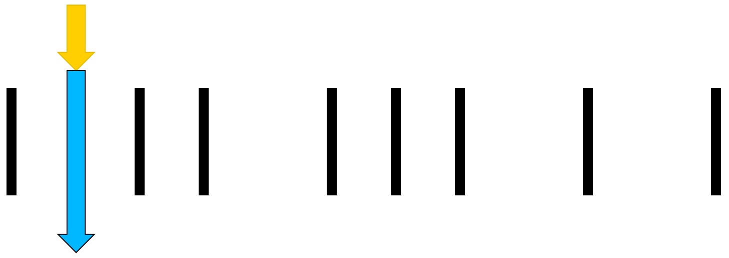
- More prone to mutate in the evolutionary process.

Edit Operations Over Time

```
TCCAGGTG-GAT
| || ||| | |
TGCAAGTGCG-T
```

Edit Operations Over Time

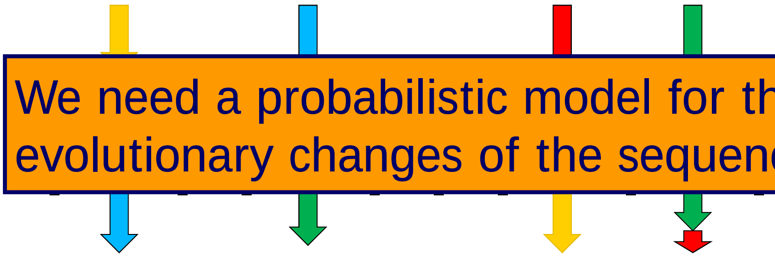
TCCAGGTG-GAT



TGCAAGTGCG-T

Edit Operations Over Time

TCCAGGTG-GAT



We need a probabilistic model for the evolutionary changes of the sequence

TGCAAGTGCG-T

Most Widely Used Scoring Matrices

- Amino acid substitution matrices
  - PAM
  - BLOSUM
- DNA substitution matrices
  - Warning: when the sequences of interest code for protein, it is almost always better to compare the protein translations than to compare the DNA sequences directly.
  - DNA is less conserved than protein sequences
    - After only a small amount of evolutionary change, the DNA sequences, when compared using simple nucleotide substitution scores, contain less information with which to deduce homology than do the encoded protein sequences
  - Less effective to compare coding regions at nucleotide level

PAM

- **P**oint **A**ccepted **M**utation*
- 1 PAM = PAM_1 = 1% average change of all amino acid positions
  - After 100 PAMs of evolution, not every residue will have changed
    - some residues may have mutated several times
    - some residues may have returned to their original state
    - some residues may not changed at all

* Dayhoff, M. O.; Schwartz, R. M.; Orcutt, B. C. (1978). "A model of evolutionary change in proteins". Atlas of Protein Sequence and Structure **5** (3): 345–352.

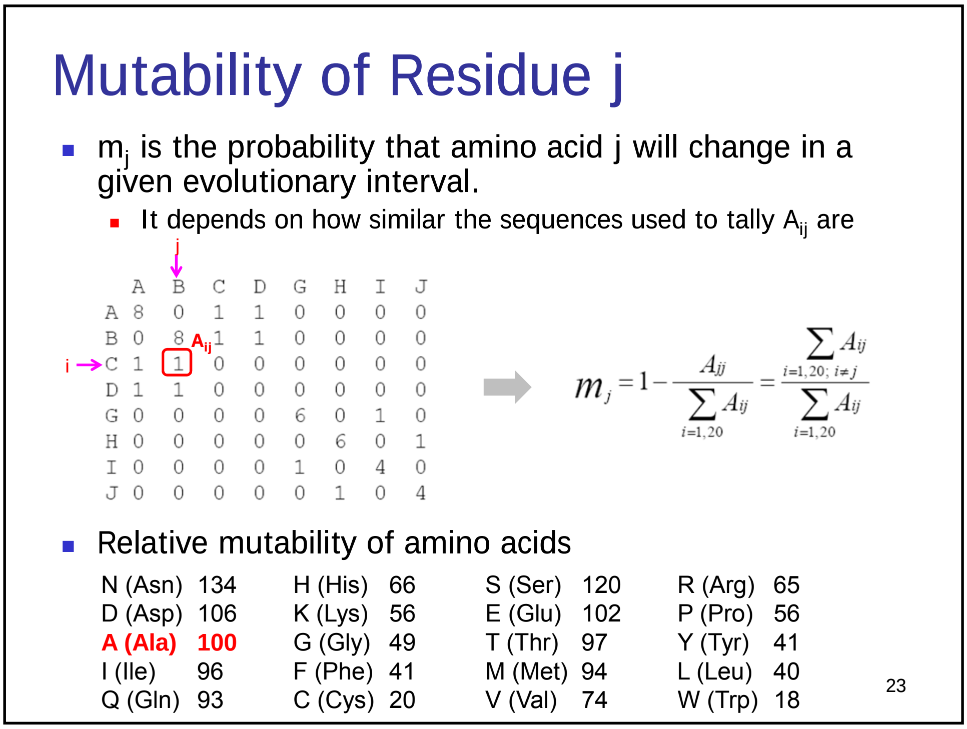
PAM Matrices: Training Data

- Take aligned set of closely related proteins
  - 71 groups of proteins that were at least 85% similar
- Each group of sequences were organized into a phylogenetic tree
  - Creates a model of the order in which substitutions occurred
- Count the number of changes of each amino acid into every other amino acid
  - Each substitution is considered to be an “accepted mutation” - an amino acid change “accepted” by natural selection

PAM: Point Accepted Mutation

- A_{ij}: number of times amino acid j mutates to amino acid i.
  - A mutation could go in both directions, therefore the tally of mutation i-j enters both A_{ij} and A_{ji} entries, while the tally of conservation i-i enters A_{ii} entry twice.

$$\begin{array}{cccccccccc}
 & & & \overset{j}{\downarrow} & & & & & & \\
 & & A & B & C & D & G & H & I & J \\
 & A & 8 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
 & B & 0 & 8\ A_{ij} & 1 & 1 & 0 & 0 & 0 & 0 \\
i \rightarrow & C & 1 & \boxed{1} & 0 & 0 & 0 & 0 & 0 & 0 \\
 & D & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
 & G & 0 & 0 & 0 & 0 & 6 & 0 & 1 & 0 \\
 & H & 0 & 0 & 0 & 0 & 0 & 6 & 0 & 1 \\
 & I & 0 & 0 & 0 & 0 & 1 & 0 & 4 & 0 \\
 & J & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 4
\end{array}$$



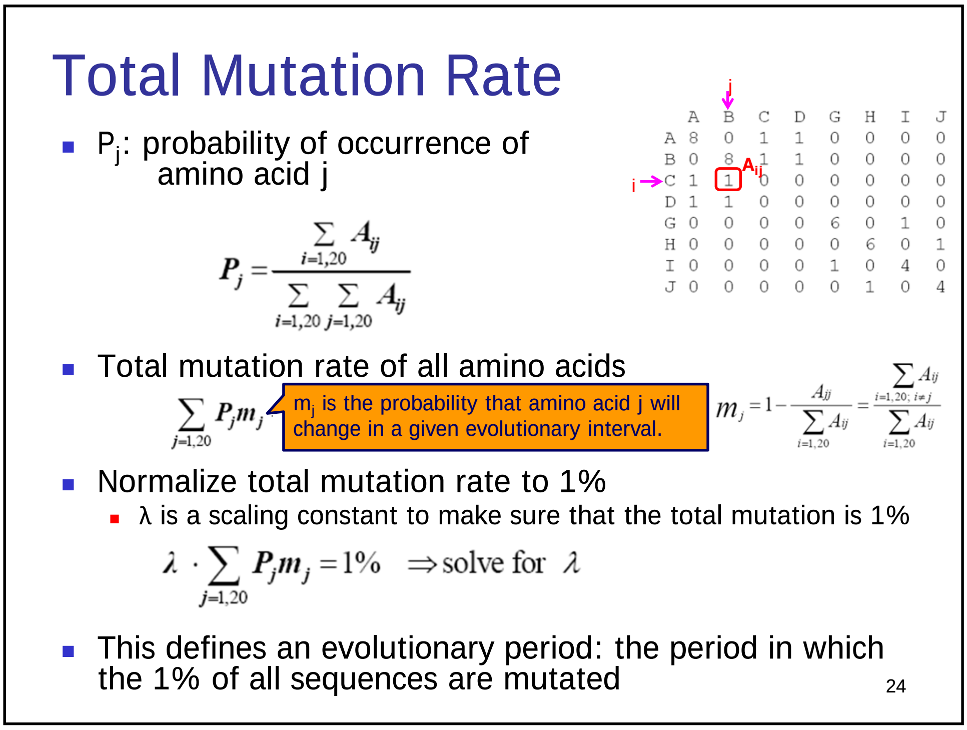
Mutability of Residue j

- m_j is the probability that amino acid j will change in a given evolutionary interval.
  - It depends on how similar the sequences used to tally A_{ij} are

$$\begin{array}{c|cccccccc} & A & \overset{j}{\downarrow}B & C & D & G & H & I & J \\ \hline A & 8 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ B & 0 & 8 & {}^{A_{ij}}1 & 1 & 0 & 0 & 0 & 0 \\ i \rightarrow C & 1 & \boxed{1} & 0 & 0 & 0 & 0 & 0 & 0 \\ D & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ G & 0 & 0 & 0 & 0 & 6 & 0 & 1 & 0 \\ H & 0 & 0 & 0 & 0 & 0 & 6 & 0 & 1 \\ I & 0 & 0 & 0 & 0 & 1 & 0 & 4 & 0 \\ J & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 4 \end{array} \quad \Rightarrow \quad m_j = 1 - \frac{A_{jj}}{\sum_{i=1,20} A_{ij}} = \frac{\sum_{i=1,20;\, i \neq j} A_{ij}}{\sum_{i=1,20} A_{ij}}$$

- Relative mutability of amino acids

N (Asn)	134	H (His)	66	S (Ser)	120	R (Arg)	65
D (Asp)	106	K (Lys)	56	E (Glu)	102	P (Pro)	56
A (Ala)	**100**	G (Gly)	49	T (Thr)	97	Y (Tyr)	41
I (Ile)	96	F (Phe)	41	M (Met)	94	L (Leu)	40
Q (Gln)	93	C (Cys)	20	V (Val)	74	W (Trp)	18



Total Mutation Rate

- P_j: probability of occurrence of amino acid j

$$\begin{array}{c|cccccccc} & A & \overset{j}{\downarrow}B & C & D & G & H & I & J \\ \hline A & 8 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ B & 0 & 8 & {}^{A_{ij}}1 & 1 & 0 & 0 & 0 & 0 \\ i \rightarrow C & 1 & \boxed{1} & 0 & 0 & 0 & 0 & 0 & 0 \\ D & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ G & 0 & 0 & 0 & 0 & 6 & 0 & 1 & 0 \\ H & 0 & 0 & 0 & 0 & 0 & 6 & 0 & 1 \\ I & 0 & 0 & 0 & 0 & 1 & 0 & 4 & 0 \\ J & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 4 \end{array}$$

$$P_j = \frac{\sum_{i=1,20} A_{ij}}{\sum_{i=1,20} \sum_{j=1,20} A_{ij}}$$

- Total mutation rate of all amino acids

$$\sum_{j=1,20} P_j m_j$$

m_j is the probability that amino acid j will change in a given evolutionary interval.

$$m_j = 1 - \frac{A_{jj}}{\sum_{i=1,20} A_{ij}} = \frac{\sum_{i=1,20;\, i \neq j} A_{ij}}{\sum_{i=1,20} A_{ij}}$$

- Normalize total mutation rate to 1%
  - λ is a scaling constant to make sure that the total mutation is 1%

$$\lambda \cdot \sum_{j=1,20} P_j m_j = 1\% \quad \Rightarrow \text{solve for } \lambda$$

- This defines an evolutionary period: the period in which the 1% of all sequences are mutated

Normalized Mutation Probability Matrix

- Normalize mutation probability matrix such that the total mutation rate is 1%

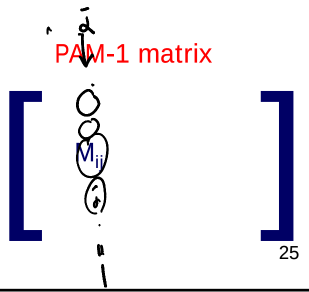
M_{ij} $(i \neq j)$: Probability of amino acid j changing into i in the evolutionary period

$$M_{ij} = \lambda \frac{A_{ij}}{\sum_{i=1,20} A_{ij}}$$

M_{jj}: Probability of amino acid j not changing in PAM-1

$$M_{jj} = 1 - \sum_{i=1,20; i \neq j} M_{ij} = 1 - \lambda m_j$$

$$A = \begin{array}{c|cccccccc} & A & B & C & D & G & H & I & J \\ \hline A & 8 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ B & 0 & 8 & 1 & 1 & 0 & 0 & 0 & 0 \\ C & 1 & \overset{A_{ij}}{\boxed{1}} & 0 & 0 & 0 & 0 & 0 & 0 \\ D & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ G & 0 & 0 & 0 & 0 & 6 & 0 & 1 & 0 \\ H & 0 & 0 & 0 & 0 & 0 & 6 & 0 & 1 \\ I & 0 & 0 & 0 & 0 & 1 & 0 & 4 & 0 \\ J & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 4 \end{array}$$



PAM-1 matrix

$$M = \begin{bmatrix} & M_{ij} & & \end{bmatrix}$$

Mutation Probability Matrix (transposed) M*10000

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
R	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
N	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
D	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
C	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Q	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
E	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
G	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
H	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
I	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
L	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
K	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
M	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
F	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
P	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
S	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
T	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
W	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Y	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
V	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

* Dayhoff, M. O.; Schwartz, R. M.; Orcutt, B. C. (1978). "A model of evolutionary change in proteins". Atlas of Protein Sequence and Structure **5** (3): 345–352.

In Two PAM1 Periods

- $M^{(1)}$: PAM-1 mutation probability matrix
- $M^{(2)}$: PAM-2 mutation probability matrix
  - Mutations that happen in twice the evolution period of that for a PAM1

- {A→R} = {A→A and A→R} or
  {A→N and N→R} or
  {A→D and D→R} or
  ... or
  {A→V and V→R} or

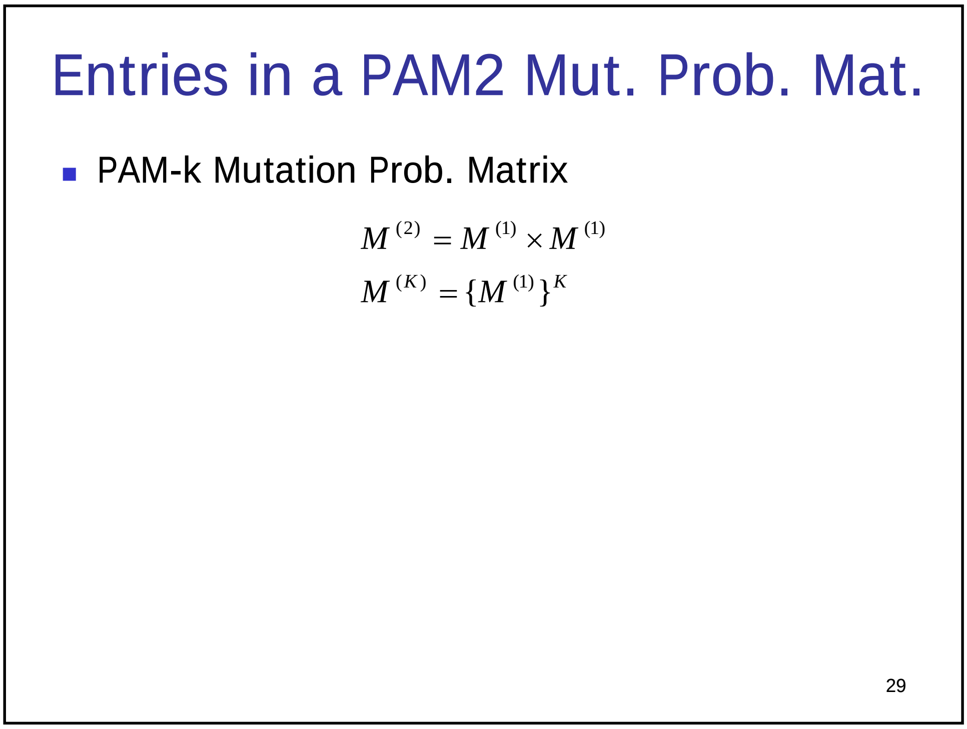
Entries in a PAM-2 Mut. Prob. Mat.

$$\begin{aligned}
&\Pr(\mathrm{A} \to \mathrm{R\ in\ 2\ periods}) = \\
&\quad \Pr(\mathrm{A} \to \mathrm{A\ in\ 1st\ period}) \times \Pr(\mathrm{A} \to \mathrm{R\ in\ 2nd\ period}) + \\
&\quad \Pr(\mathrm{A} \to \mathrm{N\ in\ 1st\ period}) \times \Pr(\mathrm{N} \to \mathrm{R\ in\ 2nd\ period}) + \\
&\quad \Pr(\mathrm{A} \to \mathrm{D\ in\ 1st\ period}) \times \Pr(\mathrm{D} \to \mathrm{R\ in\ 2nd\ period}) + \\
&\ldots
\end{aligned}$$

$$P_{AR}^{(2)} = P_{AA} \cdot P_{AR} + P_{AN} \cdot P_{NR} + P_{AD} \cdot P_{DR} + \ldots$$

$$\begin{bmatrix} P_{AA} & P_{AR} & P_{AN} & \cdots \\ P_{RA} & P_{RR} & P_{RN} & \cdots \\ P_{NA} & P_{NR} & P_{NN} & \cdots \\ & \vdots & & \end{bmatrix} \times \begin{bmatrix} P_{AA} & P_{AR} & P_{AN} & \cdots \\ P_{RA} & P_{RR} & P_{RN} & \cdots \\ P_{NA} & P_{NR} & P_{NN} & \cdots \\ & \vdots & & \end{bmatrix}$$

PAM-1 matrix PAM-1 matrix

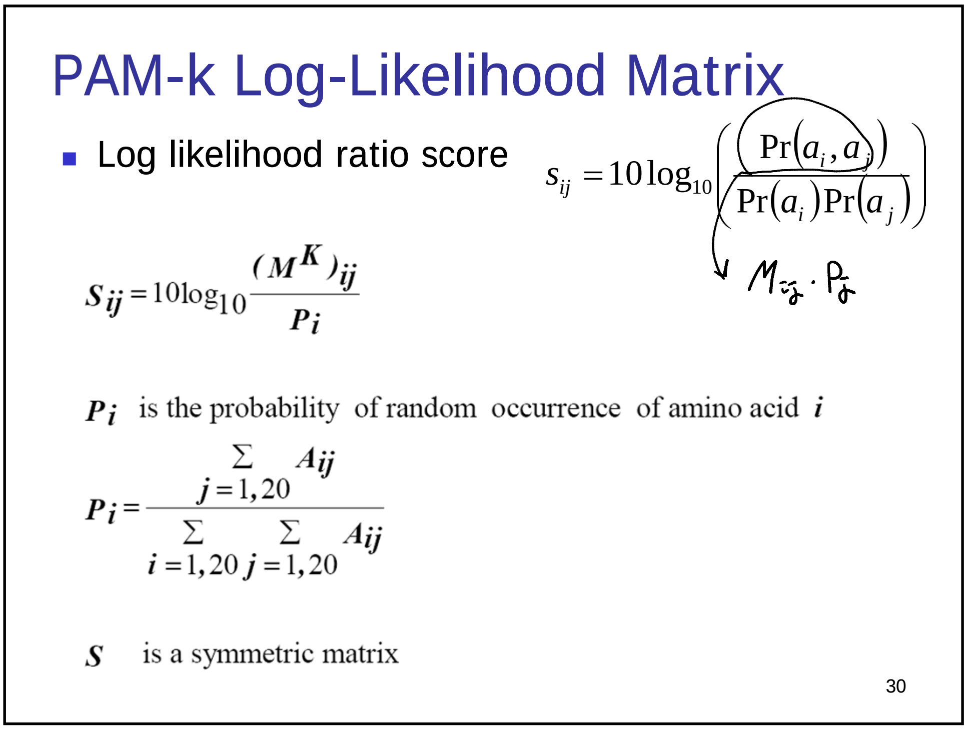


Entries in a PAM2 Mut. Prob. Mat.

- PAM-k Mutation Prob. Matrix

$$M^{(2)} = M^{(1)} \times M^{(1)}$$

$$M^{(K)} = \{M^{(1)}\}^{K}$$



PAM-k Log-Likelihood Matrix

- Log likelihood ratio score

$$s_{ij} = 10\log_{10}\left(\frac{\boxed{\Pr(a_i, a_j)}}{\Pr(a_i)\Pr(a_j)}\right)$$

$\downarrow$ $M_{ij} \cdot P_j$

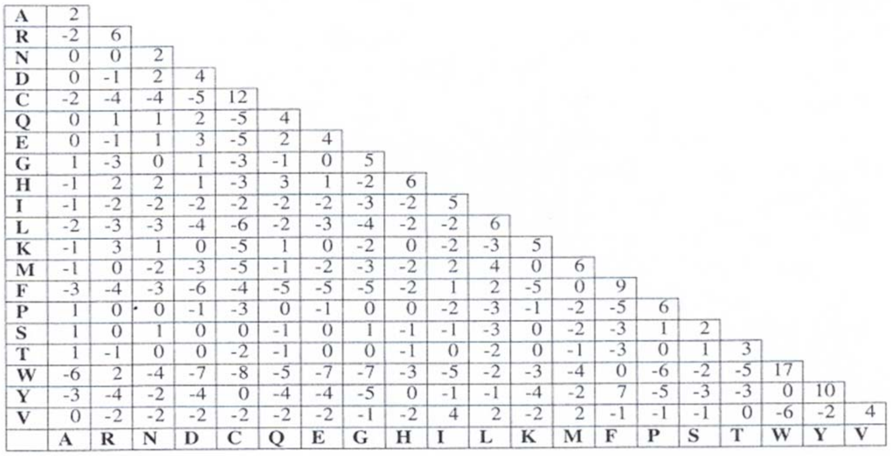
$$S_{ij} = 10\log_{10}\frac{(M^K)_{ij}}{P_i}$$

P_i is the probability of random occurrence of amino acid i

$$P_i = \frac{\sum\limits_{j=1,20} A_{ij}}{\sum\limits_{i=1,20}\sum\limits_{j=1,20} A_{ij}}$$

S is a symmetric matrix

PAM Score Matrix*



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

Log likelihood ratio matrix for PAM-250

* Dayhoff, M. O.; Schwartz, R. M.; Orcutt, B. C. (1978). "A model of evolutionary change in proteins". Atlas of Protein Sequence and Structure **5** (3): 345–352.

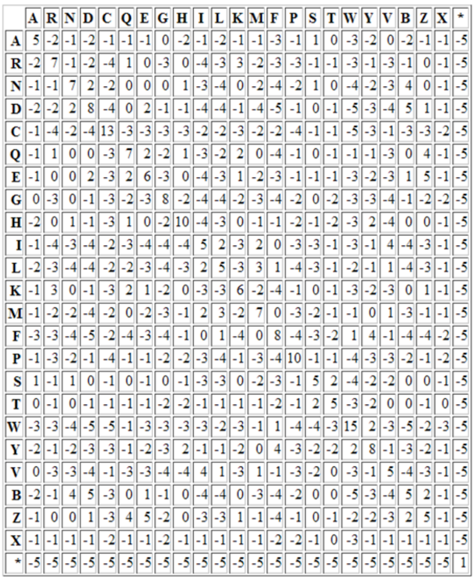
BLOSUM: Henikoff & Henikoff 92

- BLOSUM: Block Substitution Matrices
- Motivation: PAM use of matrix power can result in large errors
- Key idea: consider conserved patterns (blocks) of a large sample of proteins
  - Classify protein families (over 500 families)
  - Family has characteristic patterns (signatures) that are conserved
  - The probabilities used in the matrix calculation are computed by looking at "blocks" of conserved sequences found in multiple protein alignments.
- P(a,b) = probability of (a,b) substitution; P(a) = probability of “a”

```
 Bpi Bovine   npGivaRItqkgLdyacqqgvltlQkele
 Bpi Human    npGvvvRIsqkgLdyasqqgtaalQkelk
Cept Human    eaGivcRItkpaLlvlnhetakviQtafq
 Lbp Human    npGlvaRItdkgLqyaaqegllalQsell
 Lbp Rabbit   npGlitRItdkgLeyaareglllalQrkll
```

Scoring Matrices (e.g., BLOSUM)

- BLOSUMx=based on patterns that are x% similar
- The level of x% can provide different performance in identifying similarity
- BLOSUM62 provides good scoring (used as default)



	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	5	-2	-1	-2	-1	-1	-1	0	-2	-1	-2	-1	-1	-3	-1	1	0	-3	-2	0	-2	-1	-1	-5
R	-2	7	-1	-2	-4	1	0	-3	0	-4	-3	3	-2	-3	-3	-1	-1	-3	-1	-3	-1	0	-1	-5
N	-1	-1	7	2	-2	0	0	0	1	-3	-4	0	-2	-4	-2	1	0	-4	-2	-3	4	0	-1	-5
D	-2	-2	2	8	-4	0	2	-1	-1	-4	-4	-1	-4	-5	-1	0	-1	-5	-3	-4	5	1	-1	-5
C	-1	-4	-2	-4	13	-3	-3	-3	-3	-2	-2	-3	-2	-2	-4	-1	-1	-5	-3	-1	-3	-3	-2	-5
Q	-1	1	0	0	-3	7	2	-2	1	-3	-2	2	0	-4	-1	0	-1	-1	-1	-3	0	4	-1	-5
E	-1	0	0	2	-3	2	6	-3	0	-4	-3	1	-2	-3	-1	-1	-1	-3	-2	-3	1	5	-1	-5
G	0	-3	0	-1	-3	-2	-3	8	-2	-4	-4	-2	-3	-4	-2	0	-2	-3	-3	-4	-1	-2	-2	-5
H	-2	0	1	-1	-3	1	0	-2	10	-4	-3	0	-1	-1	-2	-1	-2	-3	2	-4	0	0	-1	-5
I	-1	-4	-3	-4	-2	-3	-4	-4	-4	5	2	-3	2	0	-3	-3	-1	-3	-1	4	-4	-3	-1	-5
L	-2	-3	-4	-4	-2	-2	-3	-4	-3	2	5	-3	3	1	-4	-3	-1	-2	-1	1	-4	-3	-1	-5
K	-1	3	0	-1	-3	2	1	-2	0	-3	-3	6	-2	-4	-1	0	-1	-3	-2	-3	0	1	-1	-5
M	-1	-2	-2	-4	-2	0	-2	-3	-1	2	3	-2	7	0	-3	-2	-1	-1	0	1	-3	-1	-1	-5
F	-3	-3	-4	-5	-2	-4	-3	-4	-1	0	1	-4	0	8	-4	-3	-2	1	4	-1	-4	-4	-2	-5
P	-1	-3	-2	-1	-4	-1	-1	-2	-2	-3	-4	-1	-3	-4	10	-1	-1	-4	-3	-3	-2	-1	-2	-5
S	1	-1	1	0	-1	0	-1	0	-1	-3	-3	0	-2	-3	-1	5	2	-4	-2	-2	0	0	-1	-5
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	2	5	-3	-2	0	0	-1	0	-5
W	-3	-3	-4	-5	-5	-1	-3	-3	-3	-3	-2	-3	-1	1	-4	-4	-3	15	2	-3	-5	-2	-3	-5
Y	-2	-1	-2	-3	-3	-1	-2	-3	2	-1	-1	-2	0	4	-3	-2	-2	2	8	-1	-3	-2	-1	-5
V	0	-3	-3	-4	-1	-3	-3	-4	-4	4	1	-3	1	-1	-3	-2	0	-3	-1	5	-4	-3	-1	-5
B	-2	-1	4	5	-3	0	1	-1	0	-4	-4	0	-3	-4	-2	0	0	-5	-3	-4	5	2	-1	-5
Z	-1	0	0	1	-3	4	5	-2	0	-3	-3	1	-1	-4	-1	0	-1	-2	-2	-3	2	5	-1	-5
X	-1	-1	-1	-1	-2	-1	-1	-2	-1	-1	-1	-1	-1	-2	-2	-1	0	-3	-1	-1	-1	-1	-1	-5
*	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	1

Constructing BLOSUM *r*

- To avoid bias in favor of a certain protein, first eliminate sequences that are more than *r*% identical
- The elimination is done by either
  - removing sequences from the block, or
  - finding a cluster of similar sequences and replacing it by a new sequence that represents the cluster.
- BLOSUM *r* is the matrix built from blocks with no more the *r*% of similarity
  - E.g., BLOSUM62 is the matrix built using sequences with no more than 62% similarity.
  - Note: BLOSUM 62 is the default matrix for protein BLAST

Collecting substitution statistics

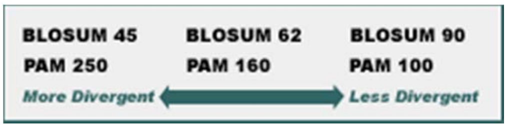
1. Count amino acids pairs in each column; e.g.,
   - 6 AA pairs, 4 AB pairs, 4 AC, 1 BC, 0 BB, 0 CC.
   - Total = 6+4+4+1=15
2. Normalize results to obtain probabilities (p_X's and p_{XY}'s)
3. Compute log likelihood ratio score matrix from probabilities:

   $s(X,Y) = \log (p_{XY} / (p_X\, p_y))$

A
A
B
A
C
A

Comparison

- PAM is based on an evolutionary model using phylogenetic trees
- BLOSUM assumes no evolutionary model, but rather conserved "blocks" of proteins



BLOSUM 45	BLOSUM 62	BLOSUM 90
PAM 250	PAM 160	PAM 100

More Divergent ⟷ *Less Divergent*