Introduction to sequence Alignment II

Best aligned subsequences

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

6 matches: 6 * 5 = 30

1 mismatch: -4

1 gaps: 1 * -7 = -7

Total: 19

- Between any pair of sequences we have seen how an optimal local alignment can be found
- How good is the best?
- When you compare one sequence of interest to a database of sequences, you can get the optimal local alignment with any sequence in the database, which ones are “real”?

Extreme Value Distribution

- The probability distribution of the mean of a large number of iid random variables is Gaussian (Central Limit Theorem)

- The probability distribution of the maximum of a large number of iid random variables is the Extreme Value Distribution

- CDF:

$$F(X) = \exp(-e^{-(x-\mu)/\sigma})$$

For a pair of long sequences, we can find local segments that have high alignment scores. These are called high-scoring segment pairs (HSP). The E-value of a score S is given by

$$E = Kmne^{-\lambda S}$$

- E: Expected number of HSPs with score at least S
- M,N: lengths of the two sequences in comparison
- λ: scale parameter
- K: scale parameter
- λ and K depend on scoring system
- Normalizing raw S as

$$S' = \frac{\lambda S - \log(K)}{\log(2)}$$

We can rewrite

$$E = mn2^{-S'}$$

- The number of random HSPs with score ovre S is a Poisson r.v. with expected value E. So the probability of observing any HSP is

$$P(\text{Poisson}(E)=0)=1-e^{-E}$$

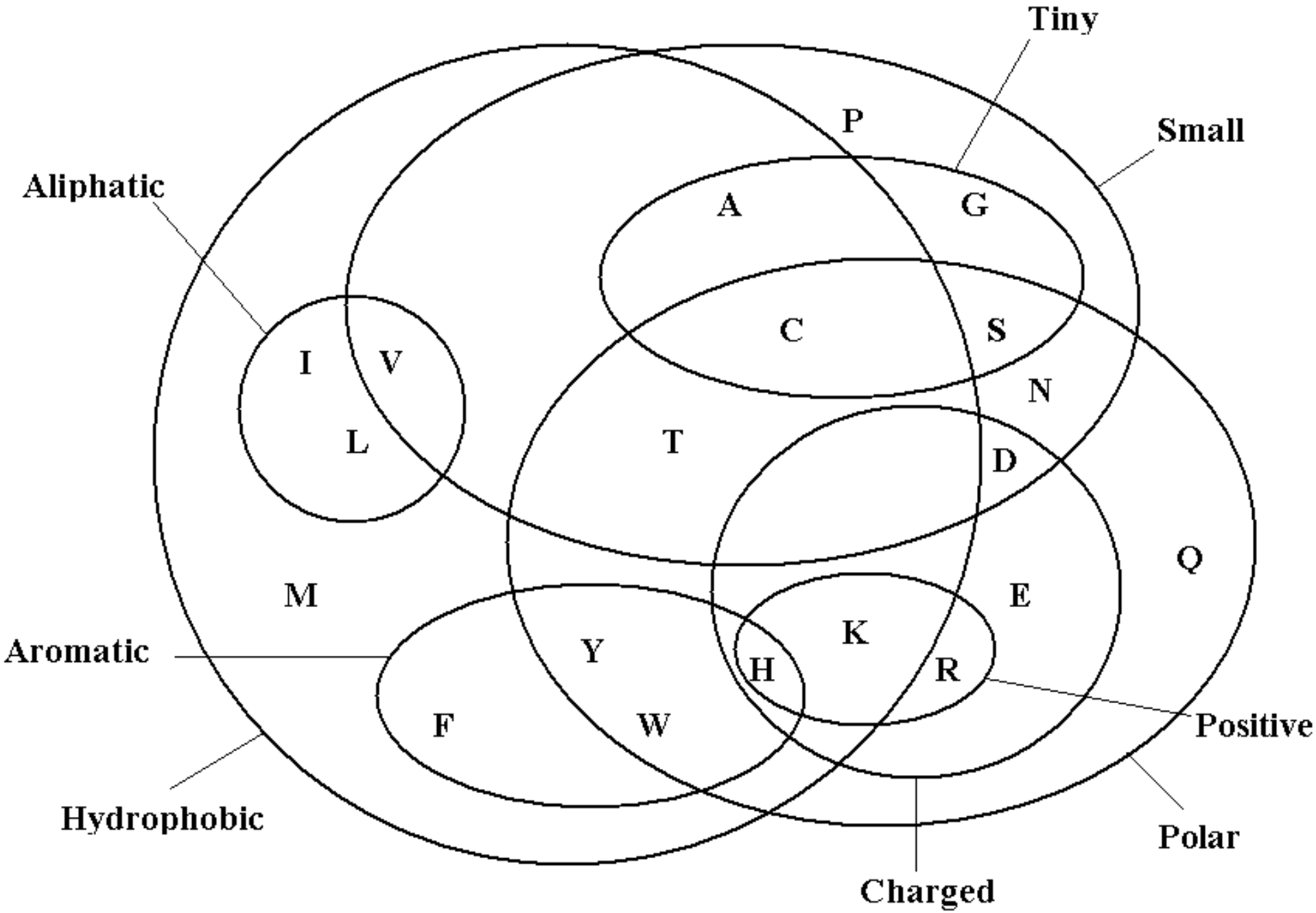
This is the p-value associated with the E-value of a score S.

Scoring Rules vs. Scoring Matrices

- Nucleotide vs. Amino Acid Sequence
- The choice of a scoring rule can strongly influence the outcome of sequence analysis
- Scoring matrices implicitly represent a particular theory of evolution
- Elements of the matrices specify the similarity of one residue to another

Nucleotide sequence determines the amino acid sequence

1st position (5' end)	2nd position U	2nd position C	2nd position A	2nd position G	3rd position (3' end)
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G


Tiny
Small
Aliphatic
P
A
G
C
S
N
I
V
L
T
D
Q
M
E
K
Y
H
R
Aromatic
F
W
Positive
Hydrophobic
Polar
Charged

Log Likelihoods used as Scoring Matrices:

BLOSUM – Blocks Substitution Matrix:
2000 “blocks” from 500 families

PAM - % Accepted Mutations:
1500 changes in 71 groups w/ > 85% similarity

Log Likelihoods used as Scoring Matrices:

BLOSUM

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
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	-3	-4	-3	4										
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
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							
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

$$S_{ij} = 2 \cdot \log_2 \frac{p_{ij}}{p_i p_j}$$

Review: Likelihood Ratio of Aligning Two Sequences

$$\text{log lik ratio of alignment}$$

$$= \log \frac{\Pr(\text{alignment} \mid \text{common ancestry})}{\Pr(\text{alignment} \mid \text{by chance})}$$

$$= \log \frac{\prod \Pr(i \text{ aligned with } j \mid \text{common ancestry})}{\prod \Pr(i \text{ aligned with } j \mid \text{by chance})}$$

$$= \log \frac{\prod p_{ij}}{\prod p_i p_j} = \sum \log \frac{p_{ij}}{p_i p_j} = \sum_{i \in s,\, j \in t} S_{ij}$$

- PAM and BLOSUM matrices are all log likelihood matrices
- More specifically:
- An alignment that scores 6 means that the alignment by common ancestry is 2^6=64 times as likely as expected by chance.

$$S_{ij} = \log_2 \frac{p_{ij}}{p_i p_j}$$

- Notice different definitions of S: sometimes

$$S_{ij} = 2\log_2 \frac{p_{ij}}{p_i p_j} \qquad S_{ij} = 10\log_{10} \frac{p_{ij}}{p_i p_j}$$

BLOSUM matrices for Protein

- S. Henikoff and J. Henikoff (1992). “Amino acid substitution matrices from protein blocks”. *PNAS* 89: 10915-10919
- Training Data: ~2000 conserved blocks from BLOCKS database. Ungapped, aligned protein segments. Each block represents a conserved region of a protein family

Constructing BLOSUM Matrices of Specific Similarities

- Sets of sequences have widely varying similarity. Sequences are clustered into one group whenever percentage of identical residues exceeds L%
- Calculating A_{ab} as frequency of residue a in one cluster aligning with b in another cluster, weighted by $(1/n_1n_2)$
- If clustering threshold is 62%, final matrix is BLOSUM62

Constructing a BLOSUM matrix: A toy example

1. Counting mutations

VVAPV
AAAPA
PVAPV
PAAAV

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

$N_{AV} = N_{VA} = 1 + 2*2 + 0 + 0 + 3 = 8$
$N_{AP} = N_{PA} = 2 + 0 + 0 + 3 + 0 = 5$
$N_{PV} = N_{VP} = 2 + 0 + 0 + 0 + 0 = 2$

N_{VP} is the number of V-P pairs

2. Tallying mutation frequencies

q_{ij}=	A	V	P
A	14	8	5
V	8	8	2
P	5	2	8

q_{ij}: number of times amino acid j mutates to amino acid i.
A mutation could go in both directions, therefore the tally of A-V pair enters both q_{AV} and q_{VA} entries, while the tally of A-A pair enters q_{AA} entry twice.

3. Matrix of mutation probs.

p_{ij} is the probability that a mutation occurs between amino acid i and amino acid j

$$p_{ij} = \frac{q_{ij}}{\sum_{i=1,20} \sum_{j=1,20} q_{ij}}$$

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

4. Calculate abundance of each residue (Marginal prob)

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

$$p_i = \frac{\sum_{j=1,20} q_{ij}}{\sum_{i=1,20} \sum_{j=1,20} q_{ij}}$$

VVAPV
AAAPA
PVAPV
PAAAV

p_i =	A	V	P
	9/20	6/20	5/20

5. Obtaining a BLOSUM matrix

The BLOSUM log-likelihood matrix:

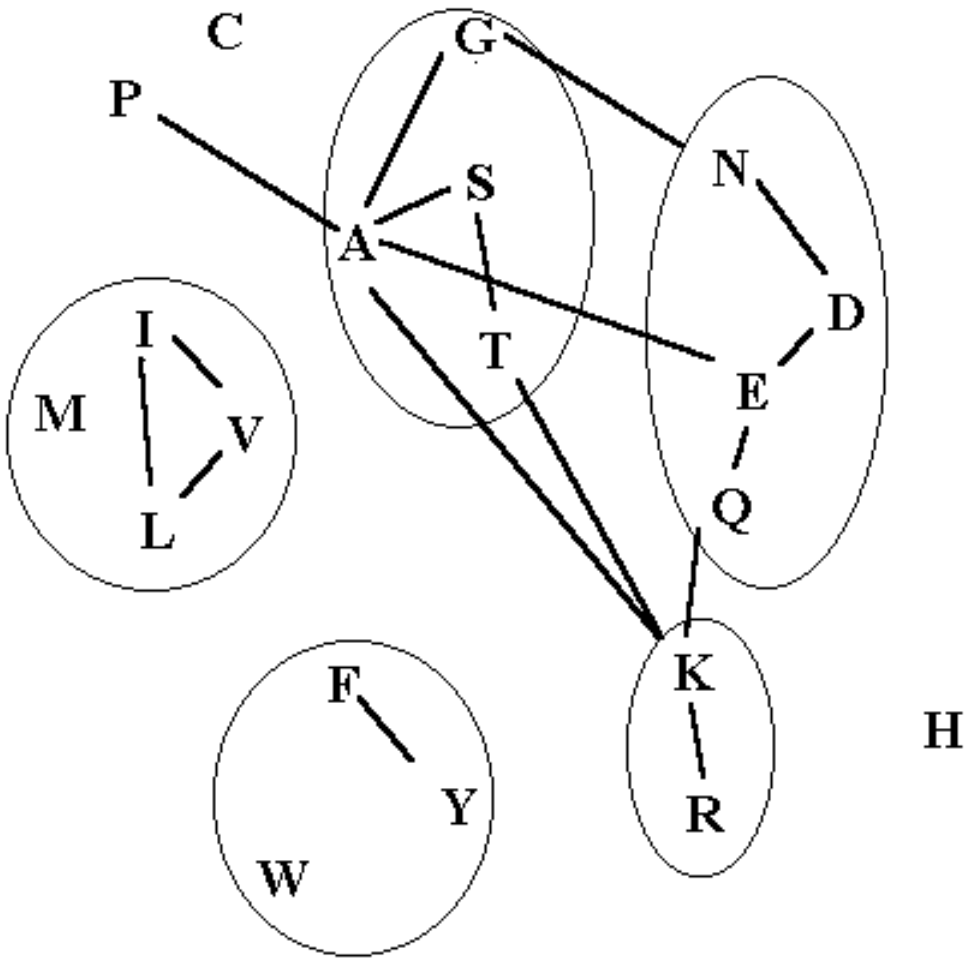
$$S_{ij} = 2\log_2 \frac{p_{ij}}{p_i p_j}$$

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

Constructing the real BLOSUM62 Matrix
Same procedure using ~2000 blocks

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
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	-3	-4	-3	4										
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
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							
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

$$S_{ij} = 2 \cdot \log_2 \frac{p_{ij}}{p_i p_j}$$


C
G
P
S
N
A
D
I
T
M
E
V
Q
L
K
F
H
Y
R
W

PAM Matrices
(Point Accepted Mutations)

Mutations accepted by natural selection

Definition 1 *An* accepted mutation *is a mutation that occurred and was positively selected by the environment; that is, it did not cause the demise of the particular organism where it occurred.*

PAM Matrices

- Atlas of Protein Sequence and Structure, Suppl 3, 1978, **M.O. Dayhoff**. ed. National Biomedical Research Foundation, 1
- Necessary ingredients to build PAM matrix
  - A list of accepted mutations
  - Probability of occurrence of each amino acid

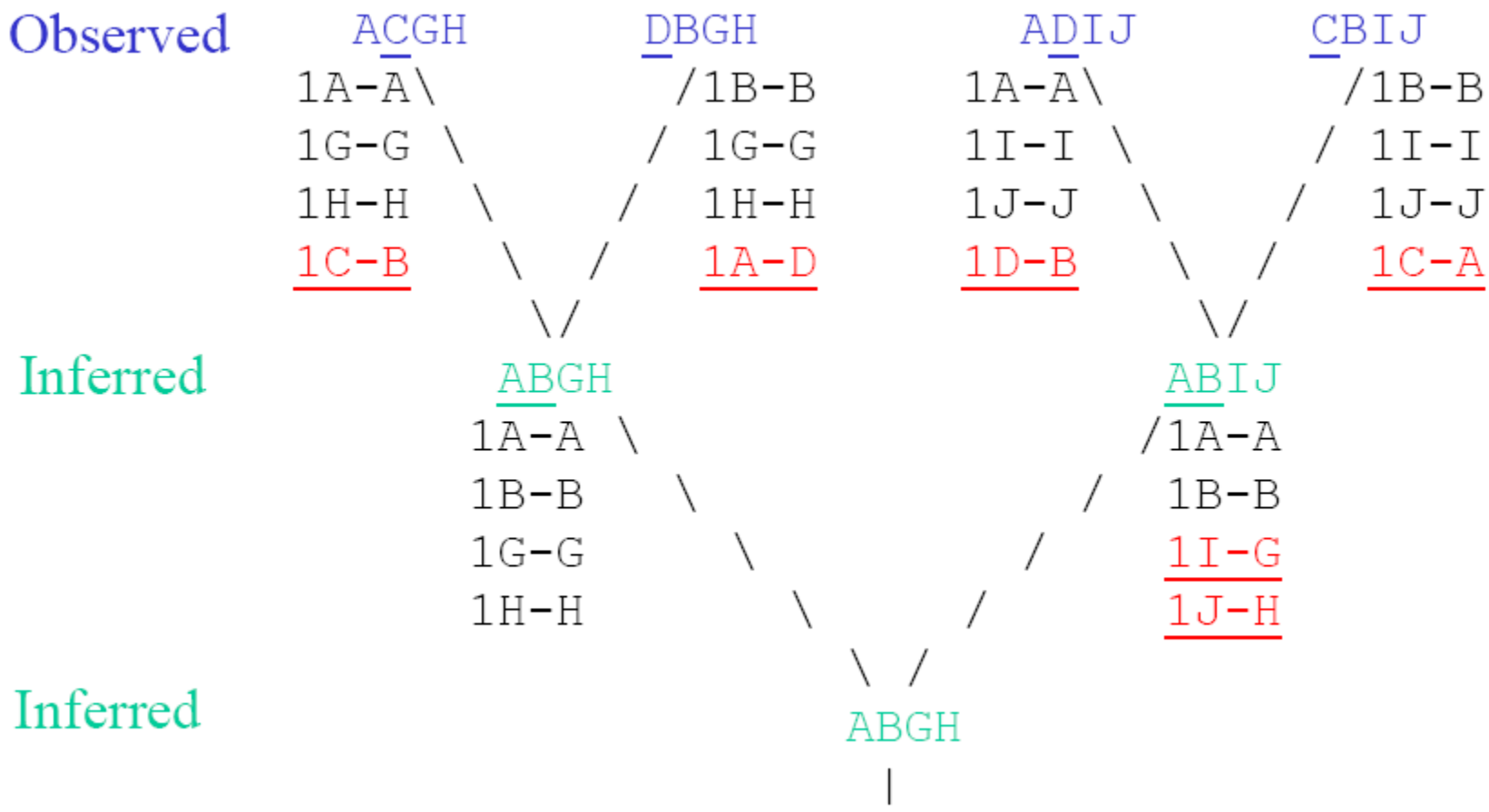
Constructing PAM Matrix: Training Data

- Align sequences that are more than 85% identical.
  minimize ambiguity in alignments.
  minimize the number of coincident and mediated mutations.

- Reconstruct phylogenetic trees and infer ancestral sequences.
  71 trees containing 1,572 “accepted mutations” were used.

- Tally "accepted mutations": A_{ij} is the number of times the mutation i $\Leftrightarrow$ j was observed to occur.

PAM: Phylogenetic Tree



PAM: Accepted Point Mutation

	A	B	C	D	G	H	I	J
A	8	0	1	1	0	0	0	0
B	0	8	1	1	0	0	0	0
C	1	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

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.

Mutability

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

m_j is the probability that amino acid j will change in a given evolutionary interval. The absolute values of m_j depend on how similar the sequences used to tally A_{ij} are

Ser	149	Met	122	Asn	111	Ile	110	Glu	102
Ala	100	Gln	98	Asp	90	Thr	90	Trp	22
Val	80	Lys	57	Pro	56	His	50	Gly	48
Phe	45	Arg	44	Leu	38	Tyr	34	Cys	27
Gap	84								

The value of Ala (m_{Ala}) has been set arbitrarily to 100 and the values of all other amino acids scaled accordingly. (Adapted from Table 21. Atlas of Protein Sequence and Structure, Suppl 3, 1978, M.O. Dayhoff, ed. National Biomedical Research Foundation, 1979.)

Total Mutation Rate

P_j is the probability of random occurrence of amino acid j

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

$\sum_{j=1,20} P_j m_j$ is the total mutation rate of all amino acids

Normalize Total Mutation Rate

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

$$\lambda \cdot \sum_{j=1,20} \boldsymbol{P}_j \boldsymbol{m}_j = 1\% \quad \Rightarrow \text{solve for } \lambda$$

Mutation Probability Matrix Normalized 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

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

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

$M^{(1)}$-- PAM1 mutation prob. matr.

$M^{(2)}$-- PAM2 Mutation Probability Matrix?

-- Mutations that happen in twice the evolution period of that for a PAM1

PAM Matrix: Assumptions

- The likelihood of amino acid Y replacing X is the same as that of X replacing Y:
  $\boldsymbol{P}(\mathrm{X} \rightarrow \mathrm{Y}) = \boldsymbol{P}(\mathrm{Y} \rightarrow \mathrm{X})$

- Very closely related proteins are used to decrease the mediated mutations such as $\mathrm{X} \rightarrow \mathrm{Z} \rightarrow \mathrm{Y}$

- Replacement at any site depends only on the amino acid at that site and the probability given by a Markov Model; all positions in a protein are equally mutable

- All sequences have average amino acid composition

In two PAM1 periods:

- {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}

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

$$\Pr(\text{A} \to \text{R in 2 periods}) =$$

$$\Pr(\text{A} \to \text{A in 1st period}) \times \Pr(\text{A} \to \text{R in 2nd period}) +$$

$$\Pr(\text{A} \to \text{N in 1st period}) \times \Pr(\text{N} \to \text{R in 2nd period}) +$$

$$\Pr(\text{A} \to \text{D in 1st period}) \times \Pr(\text{D} \to \text{R in 2nd period}) +$$

$\dots$

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

PAM-k Mutation Prob. Matrix

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

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

PAM-1 log likelihood matrix

$$S_{ij} = 10 \times \log_{10} \frac{P_{ij}^{(1)}}{p_i \cdot p_j}$$

PAM-k log likelihood matrix

$$S^{(k)}{}_{ij} = 10 \times \log_{10} \frac{P_{ij}^{(k)}}{p_i \cdot p_j}$$

PAM-250

Cys	C	12																		
Ser	S	0	2																	
Thr	T	-2	1	3																
Pro	P	-3	1	0	6															
Ala	A	-2	1	1	1	2														
Gly	G	-3	1	0	-1	1	5													
Asn	N	-4	1	0	-1	0	0	2												
Asp	D	-5	0	0	-1	0	1	2	4											
Glu	E	-5	0	0	-1	0	0	1	3	4										
Gln	Q	-5	-1	-1	0	0	-1	1	2	2	4									
His	H	-3	-1	-1	0	-1	-2	2	1	1	3	6								
Arg	R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	8							
Lys	K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5						
Met	M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6					
Ile	I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5				
Leu	L	-8	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	8			
Val	V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4		
Phe	F	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9	
Tyr	Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10
Trp	W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17
		C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W

- PAM60—60%, PAM80—50%,
- PAM120—40%
- PAM-250 matrix provides a better scoring alignment than lower-numbered PAM matrices for proteins of 14-27% similarity

Sources of Error in PAM

- Many sequences depart from average composition.

- Rare replacements were too infrequent to resolve relative probabilities accurately (for 36 pairs no replacements were observed).

- Errors in PAM-1 are magnified in the extrapolation to PAM-250.

- The Markov Process is an imperfect representation of evolution: Distantly related sequences usually have blocks of conserved residues. This implies that replacement is no equally probable over entire sequence.

Comparing Scoring Matrix

BLOSUM

- Based on a range of evol. Periods
- Conserved blocks
- Find conserved domains

PAM

- Based on extrapolation of a small evol. Period
- Track evolutionary origins
- Homologous seq.s during evolution

Choice of Scoring Matrix