BIOS 477/877
Bioinformatics and Molecular Evolution

Instructor: Etsuko Moriyama
(School of Biological Sciences)

Spring 2026 **Lecture 9**

1

Today's topics

- Amino Acid Substitution Matrix
  - Dayhoff's PAM Matrix
  - BLOSUM Matrix

2

Amino acid substitution matrices based on empirical data

- PAM matrices
  Dayhoff, Schwartz, and Orcutt (1978)
- BLOSUM matrices
  Henikoff and Henikoff (1992)
  - Also see Eddy (2004)

3

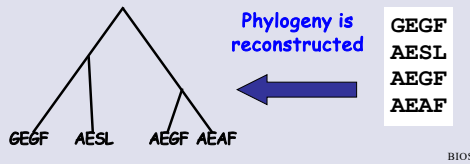
PAM matrices (Dayhoff et al. 1978)

- Accepted point mutations (point accepted mutations, percent accepted mutations)
  → accepted by selection: no (or very weak) deleterious effect, maintaining the function
- Based on 1,572 changes in 71 groups of closely related proteins (34 protein families)
  - at least 85% identical
  - no ambiguity in alignments, no gap
  - most likely observed substitutions do not affect protein functions (accepted by selection, close to neutral)
  - successive (multiple) substitutions at one site are minimal (no hidden substitution)

4

Accepted point mutations: $f(a,b)$

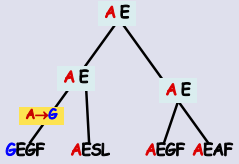
- Numbers of accepted point mutations: $f(a,b)$
  - Counted based on phylogenies
  - Assumption:
    Substitutions are equally likely in each direction (e.g., G→A = A→G)



5

Accepted point mutations: $f(a,b)$

- Numbers of accepted point mutations: $f(a,b)$
  - Counted based on phylogenies
  - Assumption:
    Substitutions are equally likely in each direction (e.g., G→A = A→G)

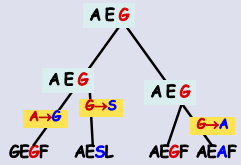


Using the maximum parsimony principle, ancestral sequences can be inferred

6

Accepted point mutations: $f(a,b)$

- Numbers of accepted point mutations: $f(a,b)$
  - Counted based on phylogenies
  - Assumption:
    Substitutions are equally likely in each direction (e.g., G→A = A→G)

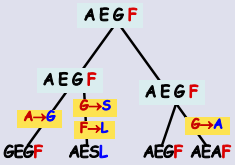


Using the maximum parsimony principle, ancestral sequences can be inferred

7

Accepted point mutations: $f(a,b)$

- Numbers of accepted point mutations: $f(a,b)$
  - Counted based on phylogenies
  - Assumption:
    Substitutions are equally likely in each direction (e.g., G→A = A→G)

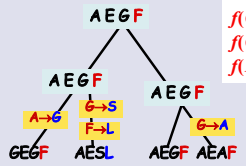


Substitutions can be identified along the phylogeny

8

Accepted point mutations: $f(a,b)$

- Numbers of accepted point mutations: $f(a,b)$
  - Counted based on phylogenies
  - Assumption:
    Substitutions are equally likely in each direction (e.g., G→A = A→G)



$f(G,A) = Freq(G \rightarrow A) + Freq(A \rightarrow G) = 2$
$f(G,S) = 1$
$f(F,L) = 1$

9

Accepted point mutations: $f(a,b)$

Numbers of accepted point mutations: $f(a,b)$

Based on 1,572 changes, but still missing 35 types of substitutions

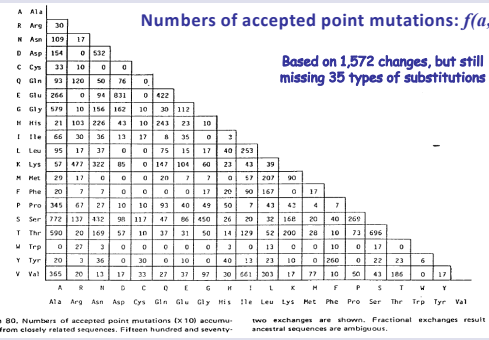


Figure 80. Numbers of accepted point mutations (X 10) accumulated from closely related sequences. Fifteen hundred and seventy-two exchanges are shown. Fractional exchanges result when ancestral sequences are ambiguous.

Dayhoff *et al.* (1978)

10

Relative mutability: $m(a)$

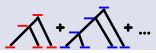
- Relative mutability: $m(a)$
  Probability that the amino acid a will change in a given small evolutionary interval

[from a pair]
GEGF
AEGF

$m(a) = \sum_i f(i,a) / f(a)$

Amino acid: a	A	E	F	G
Changes: $\sum_{f(i,a)}$	1	0	0	1
Freq. of occurrence: $f(a)$ (Total # of the residue)	1	2	2	3
Relative mutability: $m(a)$	1	0	0	0.3

[combined from multiple trees]



Substitutions are collected from trees with different lengths

$$m(a) = \frac{\text{Number of times amino acid } a \text{ is substituted by any other amino acid}}{\sum_{branch}\{(\text{Freq. of amino acid } a) \times (\text{Number of total substitutions}) \times 100\}}$$

(Number of occurrence of amino acid a) / (Total number of residues)

This denominator is called "the total exposure of the amino acid to mutation"

11

Relative mutability: $m(a)$

Table 21
Relative Mutabilities of the Amino Acids[a]

More mutable	Asn	134	His	66
	Ser	120	Arg	65
	Asp	106	Lys	56
	Glu	102	Pro	56
	Ala	100	Gly	49
	Thr	97	Tyr	41
	Ile	96	Phe	41
	Met	94	Leu	40
	Gln	93	Cys	20
	Val	74	Trp	18

Less mutable

[a] The value for Ala has been arbitrarily set at 100.

Dayhoff *et al.* (1978)

12

Mutation probability: $M(a,b)$

- **Mutation probability**

$M(a,b) = \lambda m(b) \times f(a,b) / \sum_a f(a,b)$, where $a \neq b$

$m(b)$: relative mutability of amino acid b

$f(a,b)$: frequency of accepted point mutations between amino acids a and b

$\sum_a f(a,b)$: number of times the amino acid b is substituted by any other amino acid

λ: proportionality constant (normalization factor)

→ The probability of the amino acid b being replaced by the amino acid a after a given evolutionary time

$M(b,b) = 1 - \lambda m(b)$

→ unchange probability (the diagonal elements)

13

Mutation probability matrix: M

Probabilities of AA_j replaced by AA_i

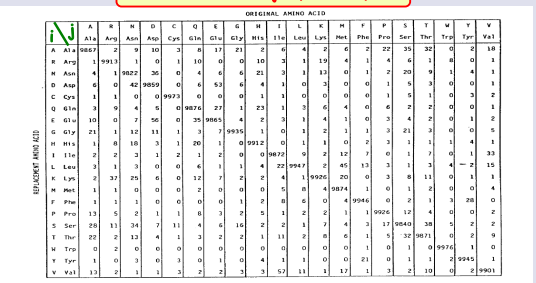


Figure 82. Mutation probability matrix for the evolutionary distance of 1 PAM. An element of this matrix, M_{ij}, gives the probability that the amino acid in column j will be replaced by the amino acid in row i after a given evolutionary interval, in this case 1 accepted point mutation per 100 amino acids. Thus, there is a 0.66% probability that Asp will be replaced by Glu. To simplify the appearance, the elements are shown multiplied by 10,000.

Dayhoff *et al.* (1978)

14

Mutation probability matrix: M

Probabilities of AA_j replaced by AA_i

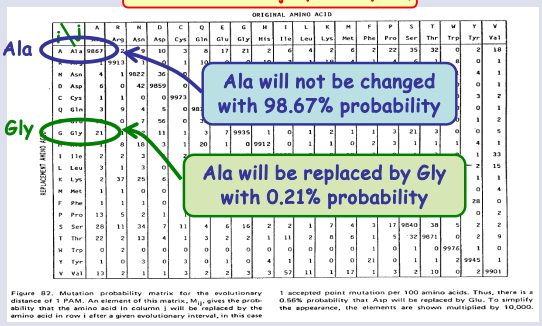


Figure 82. Mutation probability matrix for the evolutionary distance of 1 PAM. An element of this matrix, M_{ij}, gives the probability that the amino acid in column j will be replaced by the amino acid in row i after a given evolutionary interval, in this case 1 accepted point mutation per 100 amino acids. Thus, there is a 0.66% probability that Asp will be replaced by Glu. To simplify the appearance, the elements are shown multiplied by 10,000.

Dayhoff *et al.* (1978)

15

Mutation probability matrix: M

Probabilities of AA_j replaced by AA_i

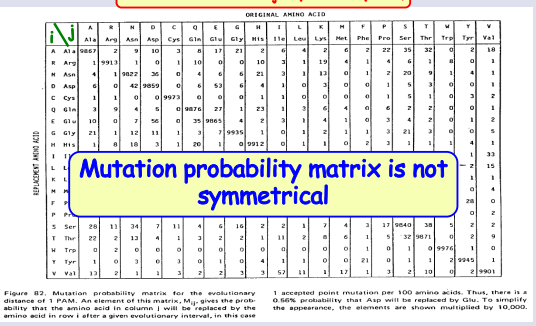


Figure 82. Mutation probability matrix for the evolutionary distance of 1 PAM. An element of this matrix, M_{ij}, gives the probability that the amino acid in column j will be replaced by the amino acid in row i after a given evolutionary interval, in this case 1 accepted point mutation per 100 amino acids. Thus, there is a 0.66% probability that Asp will be replaced by Glu. To simplify the appearance, the elements are shown multiplied by 10,000.

Dayhoff *et al.* (1978)

16

PAM scores: $R(a,b)$

- **Relatedness odds score**

$R(a,b) = M(a,b) / f(a)$

$M(a,b) = \lambda m(b) \times f(a,b) / \sum_a f(a,b)$

the probability that amino acid b will change to a in a related sequence in a given interval

$f(a)$: the chance of a random occurrence of amino acid a

= frequency of occurrence of amino acid a

= (number of occurrences of the residue a) / (total number of residues)

$R(a,b) = M(a,b) / f(a) = \{\lambda m(b) \times f(a,b) / \sum_a f(a,b)\} / f(a)$ $\quad m(b) = \sum_a f(a,b) / f(b)$

$= \{\lambda \sum_a f(a,b) \times f(a,b)\} / \{\sum_a f(a,b) \times f(a) \times f(b)\} = \lambda f(a,b) / \{f(a) \times f(b)\}$

$R(b,a) = M(b,a) / f(b) = \{\lambda m(a) \times f(b,a) / \sum_b f(b,a)\} / f(b)$

$= \lambda f(b,a) / \{f(b) \times f(a)\} = R(a,b)$

$R(a,b) = R(b,a)$

→ Relatedness odds score matrix is symmetrical

17

PAM scores: $S(a,b)$

- **Relatedness odds score**

$R(a,b) = M(a,b) / f(a)$

Odds ratio ⇓ [nonrandom]

$\frac{P(a \text{ was derived from } b)}{P(\text{random occurrence of } a)}$

→ >1 if nonrandom events happen more

- **Log odds score**

$S(a,b) = 10\ log_{10}\{M(a,b) / f(a)\}$

If $R(a,b) = M(a,b)/f(a) > 1$ or $S(a,b) > 0$

→ substitutions happen more often than by chance

18

PAM matrices: $S(a,b)$

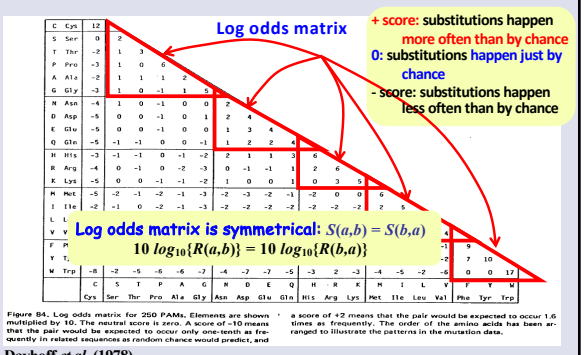


Figure 84. Log odds matrix for 250 PAMs. Elements are shown multiplied by 10. The neutral score is zero. A score of -10 means that the pair would be expected to occur only one-tenth as frequently in related sequences as random chance would predict, and a score of +2 means that the pair would be expected to occur 1.6 times as frequently. The order of the amino acids has been arranged to illustrate the patterns in the mutation data.

Dayhoff *et al.* (1978)

19

PAM1 matrix

- **PAM1 matrix**
  - normalized to represent an amount of evolution producing an average of one mutation per hundred amino acids
    → Evolutionary interval of PAM1
  - $100 \times \sum_b \{f(b)\mathrm{M}(b,b)\} = 99$ where $\mathrm{M}(b,b) = 1 - \lambda m(b)$
    → within 100 amino acids 99 are unchanged (or only 1 changed)
- **M_1: PAM1 mutation probability matrix**
  → shows the probability of AA_j replaced by AA_i after the evolutionary interval of PAM1 (when one mutation per 100 aa is found)
- **PAM1**: from $S(a,b) = 10\ log_{10}\{M_1(a,b) / f(a)\}$,
  where $M(a,b) = \lambda m(b) \times f(a,b) / \sum_a f(a,b)$
  Normalizing (scaling) factor

20

PAM1 and PAMn

- **PAM1 matrix**
  - normalized to represent an amount of evolution producing an average of one mutation per hundred amino acids
    → Evolutionary interval of PAM1
  - $100 \times \sum_b \{f(b)\mathrm{M}(b,b)\} = 99$ where $\mathrm{M}(b,b) = 1 - \lambda m(b)$
    → within 100 amino acids 99 are unchanged (or only 1 changed)
- **M_1: PAM1 mutation probability matrix**
  → shows the probability of AA_j replaced by AA_i after the evolutionary interval of PAM1 (when one mutation per 100 aa is found)
- **M_n: mutation probability matrix for PAMn**
  $M_n = (M_1)^n$ *e.g.*, M_{250} (for PAM250) = M_1^{250}
  → Probability matrix after evolutionary interval of PAM250 (after 250 changes are produced in 100 aa)

21

PAM distance

Table 23
Correspondence between Observed Differences and the Evolutionary Distance

Observed Percent Difference	Evolutionary Distance in PAMs
1	1
5	5
10	11
15	17
20	23
25	30
30	38
35	47
40	56
65	133
70	159
80	246

PAM1

$100[1-S_b\{f(b)M_n(b,b)\}]$ for PAMn

PAM250 ≈ 20% similarity

~PAM250

Dayhoff *et al.* (1978)

22

PAM matrices updated

- **JTT matrices** (Jones, Taylor, and Thornton1992)
  - Based on 59,190 accepted point mutations for 16,300 proteins
- **Gonnet matrices** (Gonnet, Cohen, and Benner, 1992)
  - Based on exhaustive pairwise alignment from the protein database (~8,344,353 amino acids).

23

BLOSUM matrices (Henikoff and Henikoff 1992)

- **Blocks substitution matrix**
  - Based on ~2,000 conserved amino acid patterns (or ungapped blocks), representing more than 500 families.
  - Based on local, multiple alignment of all commonly-occurring motifs (blocks) in the protein sequence database.

→ The Blocks Database
(no longer available, but used to generate BLOSUM matrices)

24

BLOCK entry example

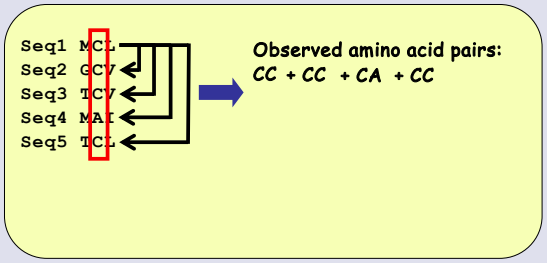
Blocks are multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins

```
Block PR00237A

ID   GPCRRHODOPSN; BLOCK
AC   PR00237A; distance from previous block=(5,490)
DE   Rhodopsin-like GPCR superfamily signature
BL   adapted;   width=25; seqs=739; 99.5%=1613; strength=1138
OAR1_LOCMI|Q25321   (  53) VTAVSLSLIILITIVGNVLVVLSVF   3
OAR2_LOCMI|Q25322   (  53) VTAVSLSLIILITIVGNVLVVLSVF   3
Q61730              (  22) ISLAVLEFINVLVIGGNCLVIAAVF  22
OAR_DROME|P22270    ( 111) LTALVLSVIIVLTIIGNILVILSVF   3
DOP2_DROME|Q24563   ( 110) GLLAFLFLFSFATVFGNSLVILAVI   5
OAR_HELVI|Q25188    (  54) CTAIVLTLIIISTIVGNILVILSVF   6
Q93128              (  35) ISLLALAFLNLMVVAGNLLVVMAVF   9
Q93126              (  35) ISLLALAFLNLMVVAGNLLVVMAVF   9
A1AB_MESAU|P18841   (  47) SVGLVLGAFILFAIVGNILVILSVA   5
A1AB_RAT|P15823     (  47) SVGLVLGAFILFAIVGNILVILSVA   5
A1AB_HUMAN|P35368   (  47) SVGLVLGAFILFAIVGNILVILSVA   5
Q93127              (  34) AATALLLAIILVLTIVGNSLVIISVF   3
A1AA_ORYLA|Q91175   (  28) VLGNVLGIFILFGVIGNILVILSVV   5
D2D1_XENLA|P24628   (  30) YYAMLLTLLVFVIVFGNVLVCIAVS   6
D2DR_CERAE|P52702   (  36) YYATLLTLLIAVIVFGNVLVCMAVS   5
Q73810              (  33) YYAVLLTLLIFVIVFGNVLVCMAVS   5
OAR_BOMMO|Q17232    (  56) CTAIILTMIIISTVVGNILVILSVF   7
SSR5_MOUSE|O08858   (  39) LVPVLYLLVCTVGLGGNTLVIYVVL   4
```

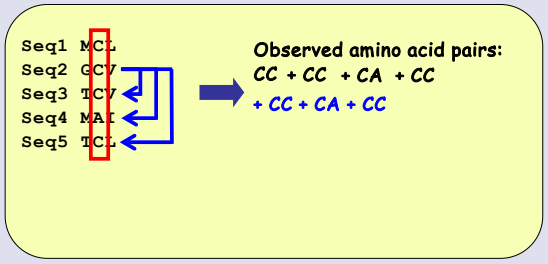

25

BLOSUM matrices: how to count



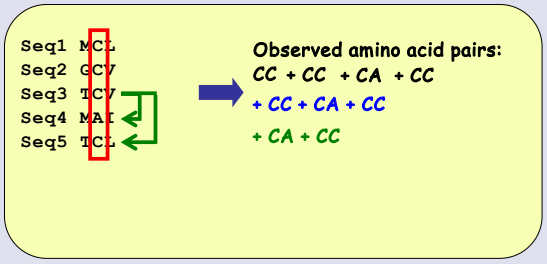
26

BLOSUM matrices: how to count



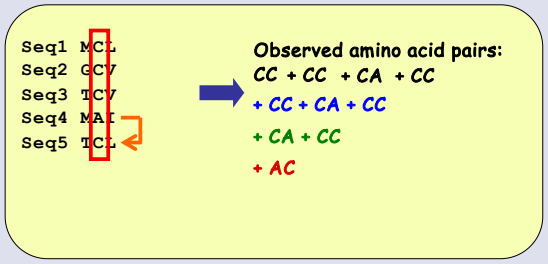
27

BLOSUM matrices: how to count



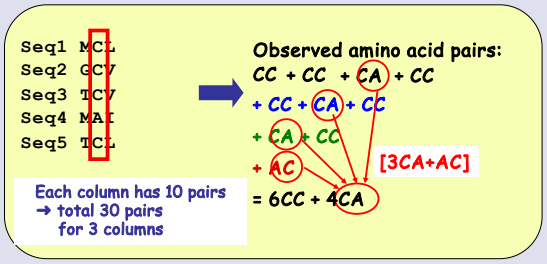
28

BLOSUM matrices: how to count



29

BLOSUM matrices: how to count



30

Observed vs. estimated AA pairs

```
Seq1 MCL
Seq2 GCV
Seq3 TCV
Seq4 MAL
Seq5 TCI
```

- Observed amino acid pairs in the alignment: 6CC, 4CA (from 30 pairs total)
- Observed frequency of pairs

  $q_{CC} = 6/30 = 0.2$, $q_{CA} = 4/30 = 0.133$ (There are 20 other pairs, too)
- Observed frequency of each amino acid:

  $p_i = q_{ii} + \Sigma_{i \neq j} q_{ij}/2$

  $p_C = 0.2 + 0.133/2 = 0.267$ and $p_A = 0.13/2 = 0.067$
- Expected frequencies of amino acid pairs:

  $e_{ii} = p_i p_i = p_i^2$ and $e_{ij} = p_i p_j + p_j p_i = 2 p_i p_j$

  $e_{CC} = 0.267^2 = 0.071$, $e_{AA} = 0.067^2 = 0.004$,

  $e_{CA} = 2 \times 0.267 \times 0.067 = 0.036$

31

BLOSUM scores

- Frequencies of amino acid pairs are cumulatively counted from all columns of the BLOCKs alignment
- Log odds score is calculated for each amino acid pair:
  - $S_{ij} = \log (q_{ij}/e_{ij})$
  - In bit units: $S_{ij} = \log_2 (q_{ij}/e_{ij})$
  - Usually in half-bit units: $S_{ij} = 2\log_2 (q_{ij}/e_{ij})$

*bit = binary digit (0 or 1)

```
Seq1 MCL
Seq2 GCV
Seq3 TCV
Seq4 MAL
Seq5 TCI
```

[From the example]

$q_{CC} = 0.2$, $e_{CC} = 0.071$ ➜ $S_{CC}=2\log_2(0.2/0.071)=2.99$

$q_{CA} = 0.133$, $e_{AC} = 0.036$ ➜ $S_{CA}=2\log_2(0.133/0.036)=3.77$

32

BLOSUM62 matrix

```
C  9
S -1  4
T -1  1  5
P -3 -1 -1  7
A  0  1  0 -1  4
G -3  0 -2 -2  0  6
N -3  1  0 -2 -2  0  6
D -3  0 -1 -1 -2 -1  1  6
E -4  0 -1 -1 -1 -2  0  2  5
Q -3  0 -1 -1 -1 -2  0  0  2  5
H -3 -1 -2 -2 -2 -2  1 -1  0  0  8
R -3 -1 -1 -2 -1 -2  0 -2  0  1  0  5
K -3  0 -1 -1 -1 -2  0 -1  1  1 -1  2  5
M -1 -1 -1 -2 -1 -3 -2 -3 -2  0 -2 -1 -1  5
I -1 -2 -1 -3 -1 -4 -3 -3 -3 -3 -3 -3 -3  1  4
L -1 -2 -1 -3 -1 -4 -3 -4 -3 -2 -3 -2 -2  2  2  4
V -1 -2  0 -2  0 -3 -3 -3 -2 -2 -3 -3 -2  1  3  1  4
F -2 -2 -2 -4 -2 -3 -3 -3 -3 -1 -3 -3  0  0  0 -1  6
Y -2 -2 -2 -3 -2 -3 -2 -3 -2 -1  2 -2 -2 -1 -1 -1 -1  3  7
W -2 -3 -2 -4 -3 -2 -4 -4 -3 -2 -2 -3 -3 -1 -3 -2 -3  1  2 11
   C  S  T  P  A  G  N  D  E  Q  H  R  K  M  I  L  V  F  Y  W
```

+ score: substitutions happen more often than by chance

0: substitutions happen just by chance

- score: substitutions happen less often than by chance

Note: Usually BLOSUM matrices are shown in half-bit units.

33

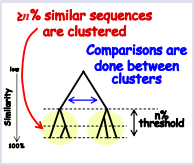
BLOSUM matrices

- BLOSUMn: n represents the similarity threshold

  (*e.g.*, BLOSUM62, BLOSUM45, BLOSUM80)

  ➜ for any n, the corresponding BLOSUM matrix is generated mainly comparing sequences that are less than n% identical

  e.g., For BLOSUM62:
  - Sequences with ≥62% identity are clustered and treated as a single sequence for counting (3 sequences in one cluster are counted as 1/3 each, instead of 1 each for counting).



- All BLOSUM matrices are based on observed alignments

  ➜ they are not extrapolated from comparisons of closely related proteins

34

Log odds matrix

	AA_1	AA_2
AA_1	S_{11}	S_{21}
AA_2	S_{12}	S_{22}

- Log odds (Lod) score: PAM matrix

  $S(i,j) = 10\log_{10}\{M(i,j)/f(i)\} = 10\log_{10}[f(i,j)/\{f(i) \times f(j)\}]$

  $M(i,j) = m(j) \times f(i,j)/\sum_i f(i,j)$: Mutation probability $AA_j \rightarrow AA_i$

  $f(i,j)$: Frequency of accepted point mutations $AA_i \leftrightarrow AA_j$

  $f(i)$, $f(j)$ = (frequency of AA_i or AA_j) / (total no. residues)

  $\sum_i f(i,j)$: Frequency of AA_j substituted by any other AA

  $m(j) = \sum_i f(i,j) / f(j)$: Relative mutability of AA_j

$S(i,j) = 10\log_{10}\{m(j) \times f(i,j)/\sum_i f(i,j)/f(i)\}$

$= 10\log_{10}[\{\sum_i f(i,j) \times f(i,j)\}/\{\sum_i f(i,j) \times f(i) \times f(j)\}]$

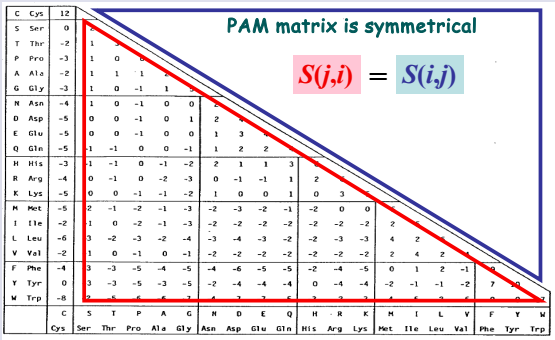
$= 10\log_{10}[f(i,j)/\{f(i) \times f(j)\}]$

$S(j,i) = 10\log_{10}[f(j,i)/\{f(j) \times f(i)\}] = S(i,j)$

➜ PAM matrix is symmetrical: $S(i,j) = S(j,i)$

35

Log odds matrix (PAM250)



Dayhoff *et al.* (1978)

36

Log odds matrix

	AA_1	AA_2
AA_1	S_{11}	S_{21}
AA_2	S_{12}	S_{22}

- **Log odds (Lod) score: BLOSUM matrix**

  $S(i,j) = 2\log_2(q_{ij}/e_{ij})$

  q_{ij}: Observed frequency of (AA_i, AA_j) pairs

  e_{ij}: Expected frequencies of (AA_i, AA_j) pairs

  $e_{ii} = p_i p_i = p_i^2$ and $e_{ij} = p_i p_j + p_j p_i = 2p_i p_j$

  $AA_i \rightarrow AA_j$, $AA_j \rightarrow AA_i$

  p_i: Observed frequency of AA_i in the pairs

  $p_i = q_{ii} + \Sigma_{i \neq j} q_{ij} / 2$

  → BLOSUM matrix is symmetrical: S(i,j) = S(j,i)

37

Log odds matrix (BLOSUM62)

BLOSUM matrix is symmetrical

$S(j,i) = S(i,j)$

```
C  9
S -1  4
T -1  1  5
P -3 -1 -1  7
A  0  1  0 -1  4
G -3  0 -2 -2  0  6
N -3  1  0 -2 -2  0  6
D -3  0 -1 -1 -2 -1  1  6
E -4  0 -1 -1 -1 -2  0  2  5
Q -3  0 -1 -1 -1 -2  0  0  2  5
H -3 -1 -2 -2 -2 -2  1 -1  0  0  8
R -3 -1 -1 -2 -1 -2  0 -2  0  1  0  5
K -3  0 -1 -1 -1 -2  0 -1  1  1 -1  2  5
M -1 -1 -1 -2 -1 -3 -2 -3 -2  0 -2 -1 -1  5
I -1 -2 -1 -3 -1 -4 -3 -3 -3 -3 -3 -3 -3  1  4
L -1 -2 -1 -3 -1 -4 -3 -4 -3 -2 -3 -2 -2  2  2  4
V -1 -2  0 -2  0 -3 -3 -3 -2 -2 -3 -3 -2  1  3  1  4
F -2 -2 -4 -2 -3 -3 -3 -3 -3 -1 -3 -3  0  0  0 -1  6  7
Y -2 -2 -2 -3 -2 -3 -2 -3 -2 -1  2 -2 -2 -1 -1 -1 -1  3  7
W -2 -3 -2 -4 -3 -2 -4 -4 -3 -2 -2 -3 -3 -1 -3 -2 -3  1  2 11
  C  S  T  P  A  G  N  D  E  Q  H  R  K  M  I  L  V  F  Y  W
```

Henikoff and Henikoff (1992)

38

Log odds matrix

- **PAM matrix**

  $S(i,j) = 10\log_{10}\{M(i,j)/f(i)\}$

  M(i,j): Mutation probability from AA_j to AA_i

  f(i): Frequency of AA_i (number of AA_i / total number of residues)

  →Probability to find AA_i by chance

- **BLOSUM matrix**

  $S(i,j) = 2\log_2(q_{ij}/e_{ij})$

  q_{ij}: Observed frequency of AA_i, AA_j pairs

  e_{ij}: Expected frequencies of AA_i, AA_j pairs [pairing by chance]

- **General form**

  $S(i,j) = 1/\lambda \log_2(q_{ij}/p_i p_j)$ [in bit unit]

  $S(i,j) = 1/\lambda \log_e(q_{ij}/p_i p_j)$ [in nat unit]

39

Log odds matrix

	AA_1	AA_2
AA_1	S_{11}	S_{21}
AA_2	S_{12}	S_{22}

- **Log odds (Lod) score: general**

  also called log odds ratio or log likelihood ratio

  $S(i,j) = 1/\lambda \log_2(q_{ij}/p_i p_j)$ [in bit unit]

  $S(i,j) = 1/\lambda \log_e(q_{ij}/p_i p_j)$ [in nat unit]

  If λ=1/2, $S(i,j) = 2\log_2(q_{ij}/p_i p_j)$ [a half-bit unit]

  q_{ij}: the frequency of the amino acid pair, AA_i and AA_j

  p_i, p_j: the individual frequency of AA_i or AA_j

  λ : a scaling factor ($1/\lambda$=2 is used with BLOSUM)

$$S(i,j) = {}^{1}/_{\lambda} \log \left(\frac{\text{Observed freq. of amino acid pair } i \text{ and } j}{\text{Expected freq. of amino acid pair } i \text{ and } j} \right)$$

* In the general format, substitutions does not have to be symmetrical. $S_{12} = S_{21}$ is not assumed.

40

Log odds matrix

	AA_1	AA_2
AA_1	S_{11}	S_{21}
AA_2	S_{12}	S_{22}

- **Log odds (Lod) score: general**

  also called log odds ratio or log likelihood ratio

Target frequency (q_{ii})

$$S(i,j) = {}^{1}/_{\lambda} \log \left(\frac{\text{Observed freq. of amino acid pair } i \text{ and } j}{\text{Expected freq. of amino acid pair } i \text{ and } j} \right)$$

Background frequency ($p_i p_j$)

$[- < S(i,j) < +]$

41

Relation to hypothesis testing

- **Hypothesis testing**
  - Test a hypothesis (H_1) against a null hypothesis (H_0)

    H_1: also called as "alternative hypothesis"
  - Likelihood ratio test: likelihood (probability of an event given a hypothesis) of each hypothesis is compared

$$\text{Likelihood ratio (LR)} = \frac{\text{Likelihood of } H_1}{\text{Likelihood of } H_0} = \frac{\text{Prob(an event}|H_1)}{\text{Prob(an event}|H_0)}$$

[0 < LR < +inf]

42

Relation to hypothesis testing

- Hypothesis testing
  - Test a hypothesis (H_1) against a null hypothesis (H_0)
    H_1: also called as "alternative hypothesis"
  - Likelihood ratio test: likelihood (probability of an event given a hypothesis) of each hypothesis is compared

$$\text{Log likelihood ratio} = \log\{\frac{\text{Likelihood of } H_1}{\text{Likelihood of } H_0}\}$$
$$= \log\{\text{Prob(an event}|H_1)\} - \log\{\text{Prob(an event}|H_0)\}$$
$$[- < \log(LR) < +]$$

43

Log odds matrix

	AA_1	AA_2	
AA_1	S_{11}	S_{21}	
AA_2	S_{12}	S_{22}	

- Log odds (Lod) score: general
  also called log odds ratio or log likelihood ratio

Target frequency (q_{ij})

$$S(i,j) = {}^{1}/_{\lambda}\, log\left(\frac{\text{Observed freq. of amino acid pair } i \text{ and } j}{\text{Expected freq. of amino acid pair } i \text{ and } j}\right)$$

Background frequency ($p_i p_j$)

$[- < S(i,j) < +]$

H_1: Homologous hypothesis (residues *i* and *j* are related)
H_0: Random hypothesis (residues *i* and *j* are unrelated)

44

Log odds score and target frequencies

$$S(i,j) = 1/\lambda \log_e(q_{ij}/p_i p_j)$$

[or $S(i,j) = 1/\lambda \log_2(q_{ij}/p_i p_j)$ for BLOSUM]

$$\lambda S(i,j) = \log_e(q_{ij}/p_i p_j)$$
$$e^{\lambda S(i,j)} = q_{ij}/p_i p_j$$
$$q_{ij} = p_i p_j e^{\lambda S(i,j)}$$

Target frequency — Expected (or background) frequency

$$\Sigma_i \Sigma_j\, q_{ij} = \Sigma_i \Sigma_{j\,(i<j)}\, p_i p_j\, e^{\lambda S(i,j)} = 1$$

λ can be estimated (matrix specific)

45