Spring 2025

BIOS 477/877

Bioinformatics and Molecular Evolution

Lecture 9

1

TODAY'S TOPICS

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

2

Substitution matrices based on empirical data

- PAM matrices
  - Dayhoff, Schwartz, and Orcutt (1978)
- BLOSUM matrices
  - Henikoff and Henikoff (1992)

Also see Eddy (2004) Nature Biotechnology 22: 1035-36

3

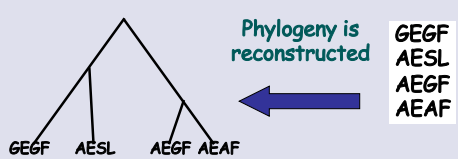
PAM matrices (Dayhoff *et al.* 1978)

- 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

PAM matrices

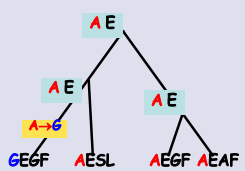
- Numbers of accepted point mutations: $f(a,b)$ are counted based on phylogenies
  - → Assumption: substitutions are equally likely in each direction (*e.g.,* G→A = A→G)



5

PAM matrices

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

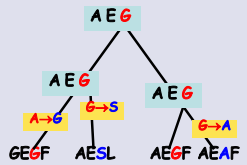


Using the maximum parsimony principle, ancestral sequences can be inferred

6

PAM matrices

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

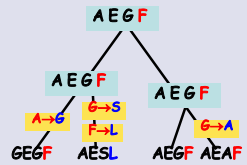


Using the maximum parsimony principle, ancestral sequences can be inferred

7

PAM matrices

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

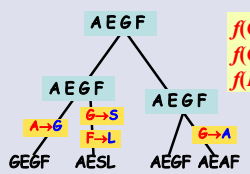


Substitutions can be identified along the phylogeny

8

PAM matrices

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



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

9

PAM matrices

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

Dayhoff *et al.* (1978)

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

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
A Ala																				
R Arg	30																			
N Asn	109	17																		
D Asp	154	0	532																	
C Cys	33	10	0	0																
Q Gln	93	120	50	76	0															
E Glu	266	0	94	831	0	422														
G Gly	579	10	156	162	10	30	112													
H His	21	103	226	43	10	243	23	10												
I Ile	66	30	36	13	17	8	35	0	3											
L Leu	95	17	37	0	0	75	15	17	40	253										
K Lys	57	477	322	85	0	147	104	60	23	43	39									
M Met	29	17	0	0	0	20	7	7	0	57	207	90								
F Phe	20	7	7	0	0	0	0	17	20	50	167	0	17							
P Pro	345	67	27	10	10	93	40	49	50	7	43	42	4	7						
S Ser	772	132	432	98	117	47	86	450	26	20	32	168	20	40	260					
T Thr	590	20	169	57	10	37	31	50	14	129	52	200	28	10	73	696				
W Trp	0	27	3	0	0	0	0	0	3	0	13	0	0	10	0	17	0			
Y Tyr	20	3	36	0	30	0	10	0	40	13	23	10	0	260	0	22	23	6		
V Val	365	20	13	17	33	27	37	97	30	661	203	17	77	10	50	43	186	0	17	

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

10

PAM matrices

- Relative mutability: $m(a)$

Probability that the amino acid a will change in a given small evolutionary interval

[from a pair]

GEGF
AEGG

Amino acid: a	A	E	F	G
Changes: $\Sigma f(i,a)$	1	0	1	2
Freq. of occurrence: $f(a)$ (Total # of the residue)	1	2	1	4
Relative mutability: $m(a)$	1	0	1	0.5

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

[combined from multiple trees]

Substitutions are collected from trees with different lengths

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

(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

PAM matrices

Relative mutability: $m(a)$

Dayhoff *et al.* (1978)

Table 21
Relative Mutabilities of the Amino Acids[a]

More mutable					Less 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		

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

12

PAM matrices

- **Mutation probability**

$M(a,b) = \lambda m(b) \text{ x } f(a,b) / \sum_a f(a,b)$, 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

PAM matrices

Mutation probability matrix: $M(a,b)$ Dayhoff *et al.* (1978)

ORIGINAL AMINO ACID

i \ j	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
A Ala	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
R Arg	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
N Asn	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
D Asp	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
C Cys	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Q Gln	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
E Glu	10	0	7	56	0	35	9865	4	2	3	3	4	1	0	3	4	2	0	1	2
G Gly	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
H His	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
I Ile	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
L Leu	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
K Lys	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
M Met	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
F Phe	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
P Pro	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
S Ser	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
T Thr	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
W Trp	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Y Tyr	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
V Val	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

REPLACEMENT AMINO ACID

Figure B2. 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 0.56% probability that Asp will be replaced by Glu. To simplify the appearance, the elements are shown multiplied by 10,000.

Probabilities of AA_j replaced by AA_i

14

PAM matrices

Mutation probability matrix: $M(a,b)$ Dayhoff *et al.* (1978)

ORIGINAL AMINO ACID

i \ j	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
A Ala	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
R Arg	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
N Asn	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
D Asp	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
C Cys	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Q Gln	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
E Glu	10	0	7	56	0	35	9865	4	2	3	3	4	1	0	3	4	2	0	1	2
G Gly	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
H His	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
I Ile	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
L Leu	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
K Lys	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
M Met	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
F Phe	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
P Pro	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
S Ser	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
T Thr	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
W Trp	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Y Tyr	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
V Val	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

REPLACEMENT AMINO ACID

Ala

Gly

Ala will not be changed with 98.67% probability

Ala will be replaced by Gly with 0.21% probability

Figure B2. 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 0.56% probability that Asp will be replaced by Glu. To simplify the appearance, the elements are shown multiplied by 10,000.

Probabilities of AA_j replaced by AA_i

15

PAM matrices

Mutation probability matrix: $M(a,b)$ Dayhoff *et al.* (1978)

ORIGINAL AMINO ACID

i \ j	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
A Ala	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
R Arg	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
N Asn	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
D Asp	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
C Cys	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Q Gln	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
E Glu	10	0	7	56	0	35	9865	4	2	3	3	4	1	0	3	4	2	0	1	2
G Gly	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
H His	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
I Ile	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
L Leu	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
K Lys	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
M Met	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
F Phe	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
P Pro	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
S Ser	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
T Thr	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
W Trp	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Y Tyr	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
V Val	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

REPLACEMENT AMINO ACID

Mutation probability matrix is not symmetrical

Figure B2. 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 0.56% probability that Asp will be replaced by Glu. To simplify the appearance, the elements are shown multiplied by 10,000.

Probabilities of AA_j replaced by AA_i

16

PAM matrices

- **Relatedness odds score**

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

$M(a,b) = \lambda m(b) \text{ x } f(a,b) / \sum_a f(a,b)$

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)

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

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

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

$R(a,b) = R(b,a)$ ➜ Relatedness odds score matrix is symmetrical

17

PAM matrices

- **Relatedness odds score**

Odds ratio

$R(a,b) = M(a,b) / f(a)$ ← $\frac{P(a \text{ was derived from } b) \text{ [nonrandom]}}{P(\text{random occurrence of } a)}$

(>1 if nonrandom events happen more)

$M(a,b) = \lambda m(b) \text{ x } f(a,b) / \sum_a f(a,b)$

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)

- $S(a,b) = 10 \log_{10}\{M(a,b) / f(a)\}$ ← Log odds score

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

Log odds matrix: $S(a,b)$ Dayhoff *et al.* (1978)

+ score: substitutions happen more often than by chance
0 score: substitutions happen just by chance
- score: substitutions happen less often than by chance

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

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.

Log odds matrix is symmetrical: $S(a,b) = S(b,a)$

$10\ log_{10}\{R(a,b)\} = 10\ log_{10}\{R(b,a)\}$

19

PAM matrices

- **PAM1 matrix**

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

→ normalized (using λ) to represent an amount of evolution producing an average of one mutation per hundred amino acids
[Evolutionary interval of PAM1]

$100 \times \sum_b \{f(b)M(b,b)\} = 99$ where $M(b,b) = 1 - \lambda m(b)$
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)

e.g., M_{250}: Probability matrix after evolutionary interval of PAM250 (after 250 changes are produced in 100 aa)

20

PAM matrices

- **PAM1 matrix**
  → normalized (using λ) to represent an amount of evolution producing an average of one mutation per hundred amino acids
  [Evolutionary interval of PAM1]

  $100 \times \sum_b \{f(b)M(b,b)\} = 99$ where $M(b,b) = 1 - \lambda m(b)$
  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 probabilty matrix for PAMn
  $M_n = (M_1)^n$ (*e.g.*, PAM250 or $M_{250} = M_1^{250}$)

21

PAM matrices

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

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

PAM1

PAM250 ≈ 20% similarity

~PAM250

22

PAM matrices updated

- **JTT matrices**
  by Jones, Taylor, and Thornton (1992)
  → 59,190 accepted point mutations for 16,300 proteins
- **Gonnet matrices**
  by Gonnet, Cohen, 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
O61730             (  22) ISLAVLFFINVLVIGGNCLVIAAVF  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) AATALLLAIILVTIVGNSLVIISVF   3
A1AA_ORYLA|Q91175  (  28) VLGMVLGIFILFGVIGNILVILSVV   5
D2D1_XENLA|P24628  (  30) YYAMLLTLLVFVIVFGNVLVCIAVS   6
D2DR_CERAE|P52702  (  36) YYATLLTLLIAVIVFGNVLVCMAVS   5
O73810             (  33) YYAVLLTLLIFVIVFGNVLVCMAVS   5
OAR_BOMMO|Q17232   (  56) CTAIILTMIIISTVVGNILVILSVF   7
SSR5_MOUSE|O08858  (  39) LVPVLYLLVCTVGLGGNTLVIYVVL   4
```


25

BLOSUM matrices

Seq1 MCL
Seq2 GCV
Seq3 TCV
Seq4 MAI
Seq5 TCL

Observed amino acid pairs:
CC + CC + CA + CC

26

BLOSUM matrices

Seq1 MCL
Seq2 GCV
Seq3 TCV
Seq4 MAI
Seq5 TCL

Observed amino acid pairs:
CC + CC + CA + CC
+ CC + CA + CC

27

BLOSUM matrices

Seq1 MCL
Seq2 GCV
Seq3 TCV
Seq4 MAI
Seq5 TCL

Observed amino acid pairs:
CC + CC + CA + CC
+ CC + CA + CC
+ CA + CC

28

BLOSUM matrices

Seq1 MCL
Seq2 GCV
Seq3 TCV
Seq4 MAI
Seq5 TCL

Observed amino acid pairs:
CC + CC + CA + CC
+ CC + CA + CC
+ CA + CC
+ AC

29

BLOSUM matrices

Seq1 MCL
Seq2 GCV
Seq3 TCV
Seq4 MAI
Seq5 TCL

Observed amino acid pairs:
CC + CC + CA + CC
+ CC + CA + CC
+ CA + CC
+ AC

[3CA+AC]

= 6CC + 4CA

Each column has 10 pairs
→ total 30 pairs
for 3 columns

30

BLOSUM matrices

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

- Observed amino acid pairs: 6CC, 4CA (from 30 pairs total)
- Observed frequency of pairs in the alignment:
  $q_{CC} = 6/30 = 0.2$, $q_{CA} = 4/30 = 0.133$ (There are 20 other pairs, too)
- Observed frequency of each amino acid in the alignment:
  $p_i = q_{ii} + \Sigma_{i \neq j}\, q_{ij}/2$
  → $p_C = 0.2 + 0.133/2 = 0.267$ 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 matrices

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

- Observed and Expected frequencies of amino acid pairs are cumulatively counted from all columns of the BLOCKs
- Log odds scores are calculated for each amino acid pairs:

$$S_{ij} = \log (q_{ij}/e_{ij})$$

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)

32

BLOSUM matrices

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

- Observed and Expected frequencies of amino acid pairs are cumulatively counted from all columns of the BLOCKs
- Log odds scores are calculated for each amino acid pairs:

$$S_{ij} = 2\log_2 (q_{ij}/e_{ij})$$

From the example:

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

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

33

BLOSUM62 matrix

Log odds matrix (Henikoff and Henikoff 1992)

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

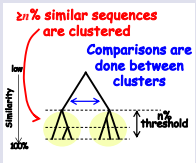
+ score: substitutions happen more often than by chance
0 score: substitutions happen just by chance
- score: substitutions happen less often than by chance

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

34

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., 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

35

BLOSUM and PAM matrices

- PAM matrices: based on mutational model of evolution
  → a transition probability matrix for a Markov process
  → any M_n matrix can be extrapolated based on PAM1 (M_1) matrix (*e.g.*, $M_{250} = M_1^{250}$)

Transition matrix

$AA_1 \xrightarrow{M_1} AA_2 \xrightarrow{M_1} AA_3 \xrightarrow{M_1}$

  → assume more distant changes are a reflection of the short-term changes
  → designed to track the evolutionary origins of proteins
- BLOSUM matrices: not based on explicit evolutionary model
  → based on local similarity
  → derived from all changes observed in the conserved blocks regardless of the overall degree of similarity
  → generated based on different similarity levels (BLOSUM50, BLOSUM62, *etc.*)
  → all BLOSUM matrices are generated based on observed data
  → designed to find conserved domains

36

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)

$\sum_j f(i,j)$: 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: $S(i,j) = S(j,i)$**

37

Log Odds Matrix (PAM250)

PAM matrix is symmetrical

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

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

Dayhoff *et al.* (1978)

38

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**

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

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

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 $e_{ij} = p_i p_j + p_j p_i = 2p_i p_j$

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

39

Log Odds Matrix (BLOSUM62)

BLOSUM matrix is symmetrical

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

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

Henikoff and Henikoff (1992)

40

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

- **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]

41

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.

42

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

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

Target frequency (q_{ii})

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

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

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

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

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

$$\text{Likelihood ratio (LR)} = \frac{\text{Likelihood of } H_1}{\text{Likelihood of } H_0}$$

$[0 < LR < +inf]$ $$= \frac{\text{Prob(an event}|H_1)}{\text{Prob(an event}|H_0)}$$

H_1: Hypothesis to be tested, H_0: Null hypothesis

44

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

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

$$\text{Log likelihood ratio} = \log\left\{\frac{\text{Likelihood of } H_1}{\text{Likelihood of } H_0}\right\}$$

$$= \log\{\text{Prob(an event}|H_1)\}-\log\{\text{Prob(an event}|H_0)\}$$

$[- < \log(LR) < +]$

H_1: Hypothesis to be tested, H_0: Null hypothesis

45

Log Odds Score and Target Frequencies

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

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

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

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

$$q_{ij} = p_ip_je^{\lambda S(i,j)}$$

Target frequency Expected (or background) frequency

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

λ can be estimated (matrix specific)

46