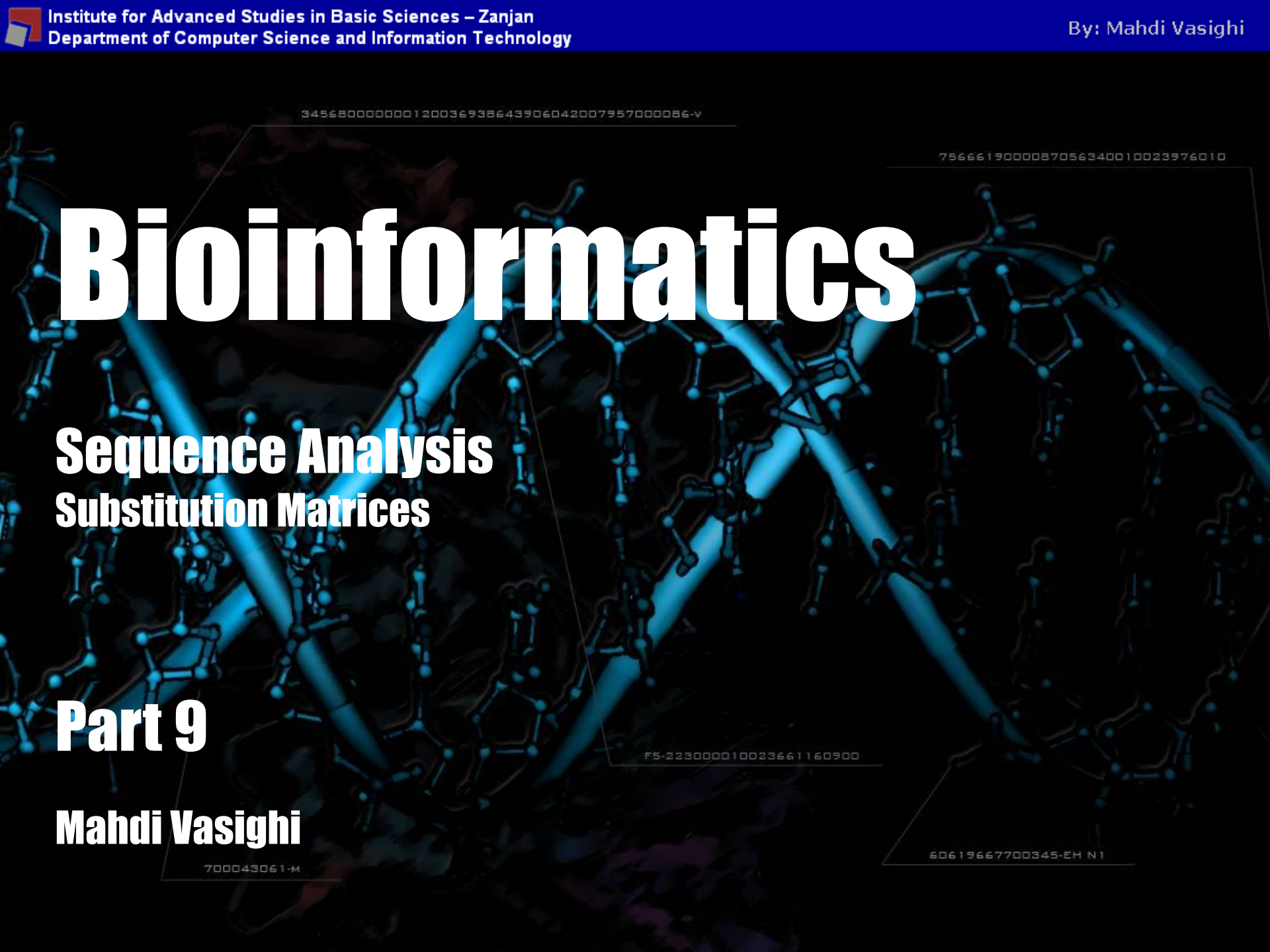


Bioinformatics

Sequence Analysis
Substitution Matrices

Part 9

Mahdi Vasighi

Pairwise Sequence Alignment

```
-GCGC-ATGGATTGAGCGA
TGCGCCATTGAT-GACC-A
```

```
------GCGCATGGATTGAGCGA
TGCGCC----ATTGATGACCA--
WHICH-ONE-IS-BETTER?---
```

Pairwise Sequence Alignment

Scoring Alignment

```
ACGTCTAG
| |
ACTCTAG-
```

2 matches
5 mismatches
1 not aligned

```
ACGTCTAG
    | | | | |
-ACTCTAG
```

5 matches
2 mismatches
1 not aligned

```
ACGTCTAG
| |  | | | | |
AC-TCTAG
```

7 matches
0 mismatches
1 not aligned

It depends on how you score the matches, mismatches and gaps. Since we are interested in the similarity of the two sequences, we would want to reward a match and penalize a mismatch/gap.

For example:

Status	Score
Match	+1
Mismatch	-1
Indel	-2

Pairwise Sequence Alignment

Scoring Alignment

Status	Score
Match	+1
Mismatch	-1
Indel	-2

```
-GCGC-ATGGATTGAGCGA
TGCGCCATTGAT-GACC-A
```

Score = (+1×13) + (-1×2) + (-2×4) = 3

```
-------GCGCATGGATTGAGCGA
TGCGCC-----ATTGATGACCA--
```

Score = (+1×5) + (-1×6) + (-2×12) = -25

Pairwise Sequence Alignment

Scoring Alignment

Status	Score
Match	+1
Mismatch	-1
Indel	-2

```
TGCG--ATGGATTGCGCGA
TGCGACATGGAT-GAC-CA
```

Score = (+1×12) + (-1×3) + (-2×4) = 1

?

```
TGCGA--TGGATTGCGCGA
TGCGACATGGATGACC--A
```

Score = (+1×12) + (-1×3) + (-2×4) = 1

Pairwise Sequence Alignment

Scoring Alignment

The goal is finding alignments that are evolutionarily likely.

Which one of the following alignments seems more likely?

```
GHGKKVADALVNAVDHVADSALSDLHAHKL
GHGKK-----V-A-D--A-SALSDLHAHKL
```

```
GHGKKVADALVNAVDHVADSALSDLHAHKL
GHGKKVADA----------SALSDLHAHKL
```

- We can achieve this kind of alignment by penalizing more for a new gap, than for extending an existing gap!

Pairwise Sequence Alignment

Scoring Alignment

Status	Score
Match	+1
Mismatch	0
Gap open	-2
Gap extension	-1

```
ACGTCTGATACGCCGTATAGTCTATCT
     ||||| |||   || ||||||||
----CTGATTCGC---ATCGTCTATCT
```

Matches: 18 × (+1)
Mismatches: 2 × 0
Open: 2 × (–2)
Extension: 5 × (–1)

Score = +9

Pairwise Sequence Alignment

Scoring Alignment

Any scoring scheme can be represented as a substitution matrix. For example:

	C	T	A	G
C	1	-1	-1	-1
T	-1	1	-1	-1
A	-1	-1	1	-1
G	-1	-1	-1	1

Can we used a similar miss-match score for A → C and A → G ?

Transversions should be penalized more than transitions

– transitions: replacement of a purine base with another purine or replacement of a pyrimidine with another pyrimidine (A <-> G, C <-> T)

– transversions: replacement of a purine with a pyrimidine or vice versa.

– Transition mutations are more common than transversions

Pairwise Sequence Alignment

Scoring Alignment

Any scoring scheme can be represented as a substitution matrix. For example:

	C	T	A	G
C	1	-1	-5	-5
T	-1	1	-5	-5
A	-5	-5	1	-1
G	-5	-5	-1	1

Transition-Transversion matrix

Can we used a similar miss-match score for A → C and A → G ?

Transversions should be penalized more than transitions

– transitions: replacement of a purine base with another purine or replacement of a pyrimidine with another pyrimidine (A <-> G, C <-> T)

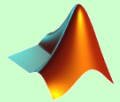
– transversions: replacement of a purine with a pyrimidine or vice versa.

– Transition mutations are more common than transversions

Pairwise Sequence Alignment

Scoring Alignment

nuc44



Purpose: Convert nucleotide sequence to amino acid sequence

Syntax: `SeqAA = nt2aa(SeqNT)`

Example

```
S = nuc44
```

```
S =
     5    -4    -4    -4     1    -4    -4     1    -4     1    -4    -1    -1    -1    -2
    -4     5    -4    -4    -4     1    -4     1     1    -4    -1    -4    -1    -1    -2
    -4    -4     5    -4     1    -4     1    -4     1    -4    -1    -1    -4    -1    -2
    -4    -4    -4     5    -4     1     1    -4    -4     1    -1    -1    -1    -4    -2
     1    -4     1    -4    -1    -4    -2    -2    -2    -2    -3    -1    -3    -1    -1
    -4     1    -4     1    -4    -1    -2    -2    -2    -2    -1    -3    -1    -3    -1
    -4    -4     1     1    -2    -2    -1    -4    -2    -2    -1    -1    -3    -3    -1
     1     1    -4    -4    -2    -2    -4    -1    -2    -2    -3    -3    -1    -1    -1
    -4     1     1    -4    -2    -2    -2    -2    -1    -4    -1    -3    -3    -1    -1
     1    -4    -4     1    -2    -2    -2    -2    -4    -1    -3    -1    -1    -3    -1
    -4    -1    -1    -1    -3    -1    -1    -3    -1    -3    -1    -2    -2    -2    -1
    -1    -4    -1    -1    -1    -3    -1    -3    -3    -1    -2    -1    -2    -2    -1
    -1    -1    -4    -1    -3    -1    -3    -1    -3    -1    -2    -2    -1    -2    -1
    -1    -1    -1    -4    -1    -3    -3    -1    -1    -3    -2    -2    -2    -1    -1
    -2    -2    -2    -2    -1    -1    -1    -1    -1    -1    -1    -1    -1    -1    -1
```

Pairwise Sequence Alignment

Scoring Alignment

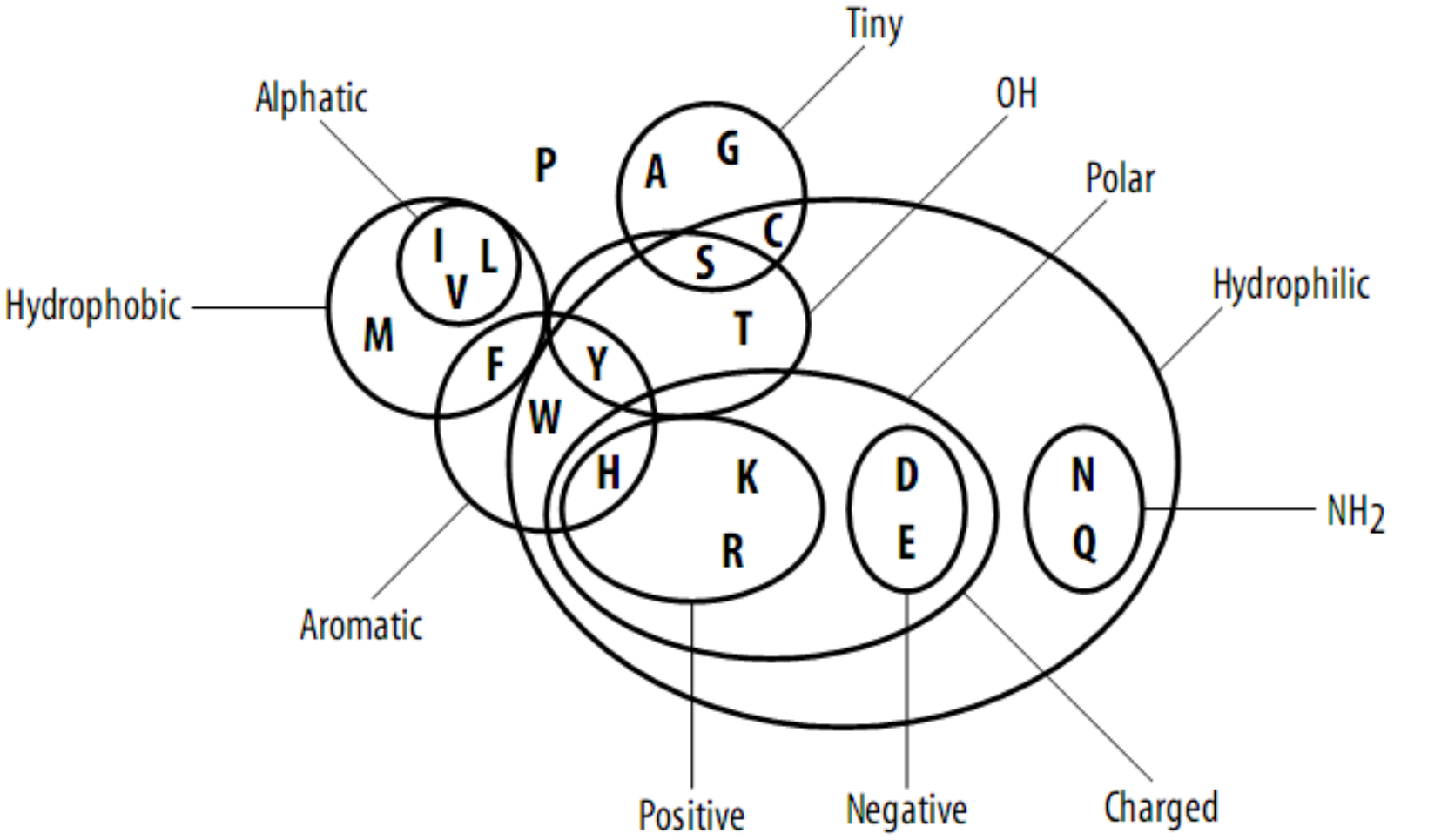
What about substituation matrix for protein sequence alignment?

- The same residues in a position give the score value 1, and different residues give 0.
- The same residues give a score 1, similar residues (for example: Tyr/Phe, or Ile/Leu) give 0.5, and all others 0.
- One may calculate, using well established sequence alignments, the frequencies (**probabilities**) that a particular residue in a position is exchanged for another.

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

Pairwise Sequence Alignment

Scoring Alignment



Pairwise Sequence Alignment

Scoring Alignment

Types of substitution matrices:

- DNA matrices
- PAM (Point Accepted Mutation)
- BLOSUM (BLOcks of amino acid SUbstitution Matrix)

Pairwise Sequence Alignment

Scoring Alignment

PAM (Point Accepted Mutations)

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

PAM is 20 × 20 matrix which elements are the probability of amino acid *a* changing into amino acid *b*.

PAM scoring was derived by Dayhoff (1970) based on 1572 observed mutations in 71 families of closely related proteins.

Pairwise Sequence Alignment

Scoring Alignment

PAM (Point Accepted Mutations)

- Starts with a multiple sequence alignment of very similar (>85% identity) proteins. Assumed to be homologous
- Compute the *relative mutability*, *m*i, of each amino acid (e.g. U_A = how many times was alanine(A) substituted with anything else?)

$$U_i = \frac{\text{\# times a.a. is observed to change}}{\text{\# times a.a. occurs in aligned sequences}}$$

Relative mutabilities

ACGCTAFKI
GCGCTAFKI
ACGCTAFKL
GCGCTGFKI
GCGCTLFKI
ASGCTAFKL
ACACTAFKL

- Across *all pairs* of sequences, there are 28 A $\rightarrow$ X substitutions
- There are 10 ALA residues, so U_A = 2.8

Pairwise Sequence Alignment

Scoring Alignment

aa	m_i	aa	m_i
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

Values according Dayhoff (1978) The value for Ala has been arbitrarily set at 100.

- **Trp** and **Cys** are less mutable
- **Asn**, **Ser**, **Asp** and **Glu** are most mutable

Pairwise Sequence Alignment

Scoring Alignment

PAM (Point Accepted Mutations)

- **Mutation probability matrix:** It gives probability that a substitution will occur in in specified unit of evolutionary time

To prepare the Dayhoff PAM matrices, amino acid substitutions that occur in a group of evolving proteins were estimated using 1572 changes in 71 groups of protein sequences that were at least 85% similar.

$$M(i,j) = u_i \times \frac{\#\ \text{times a.a. } i \text{ changes to a.a. } j}{\text{total number of changes in a.a. } i}$$

Pairwise Sequence Alignment

Scoring Alignment

	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

1 PAM (×10⁴)

$$\left[\text{PAM 1}\right]^{2} = \text{PAM 2}$$

Other PAM matrices are extrapolated from PAM1.

- PAM-1 : one substitution per 100 residues (a PAM unit of time)
- When we multiply the PAM matrices many times, the error is magnified
- The diagonal represents the probability to still observe the same residue.

Pairwise Sequence Alignment

Scoring Alignment

PAM (Point Accepted Mutations)

It expresses scores as **log-odds scores** values:

Mutation probability matrix number

Occurrence probability of a

$$S(a \rightarrow b) = 10\ log_{10}(M_{ab}/P_a)$$

$$S(b \rightarrow a) = 10\ log_{10}(M_{ba}/P_b)$$

$$S(a,b) = \frac{\left[S(a \rightarrow b) + S(b \rightarrow a)\right]}{2}$$

Pairwise Sequence Alignment

Scoring Alignment

S^{250}

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

So, a PAM score of -8 indicates that (in related sequences) the mutation would be expected to occur $10^{-0.8}$ times more frequently than random.

$-8 = 10 \log_{10}(M_{CL}/P_L)$

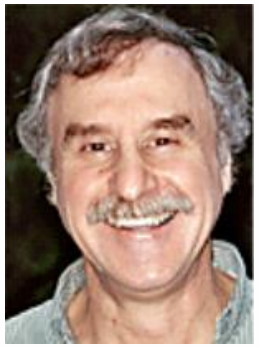
So rare mutation in evaluation

PAM substitution matrices are based on the rate of divergence between sequences

Pairwise Sequence Alignment

Scoring Alignment

Blocks Amino Acid **Su**bstitution **M**atrices (BLOSUM)



Steven Henikoff

- Based on database of ungapped local alignments (BLOCKS)
- Alignments have lower similarity than PAM alignments.
- BLOSUM index indicates the percent identity level of sequences in the alignment
- BLOSUM62 → 62% similarity

```
AABCDA---BBCDA
DABCDA-A-BBCBB
BBBCDABA-BCCA-
-AACDAC-DCBCDB
-CBADAB-DBBDC-
AAACAA---BBCCC
```

Pairwise Sequence Alignment

Scoring Alignment

Just as with the PAM matrix, we will compute the BLOSUM score as the (log) ratio of the observed probability of substitution of one amino acid by another divided by the probability expected purely due to chance.

A	A	I
S	A	L
T	A	L
T	A	V
A	A	L

Log odds ratio: $s_{ij} = \log_2 \frac{q_{ij}}{e_{ij}}$

$$e_{ii} = p_i^2$$

$$e_{ij} = 2p_i p_j \qquad (i \neq j)$$

Value stored for BLOSUM = $2 \times \boldsymbol{S}_{ij}$ is rounded to nearest integer

Pairwise Sequence Alignment

Scoring Alignment

Step 1: Count pair frequencies for each pair of amino acids *i* and *j*, for each column *k* of each block:

	A	I	L	S	T	V
A	10+1					
I		0				
L		3	3			
S	2			0		
T	4			2	1	
V		1	3			0

A	A	I
S	A	L
T	A	L
T	A	V
A	A	L

Pairwise Sequence Alignment

Scoring Alignment

Step 2: Normalize results to obtain probabilities:

A	A	I
S	A	L
T	A	L
T	A	V
A	A	L

q_{ij}

	A	I	L	S	T	V
A	11/30					
I		0				
L		3/30	3/30			
S	2/30			0		
T	4/30			2/30	1/30	
V		1/30	3/30			0

Pairwise Sequence Alignment

Scoring Alignment

Step 2: compute frequency of occurrence of i and j:

A A I
S A L
T A L
T A V
A A L

e_{ij}

	A	I	L	S	T	V
A	$(7/15)^2$					
I	…	0				
L	…	…	$(3/15)^2$			
S	…	…	…	$(1/15)^2$		
T	…	…	…	…	$(2/15)^2$	
V	…	…	…	…	…	0

Pairwise Sequence Alignment

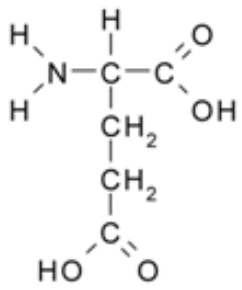
Scoring Alignment

The BLOSUM62 amino acid substitution matr

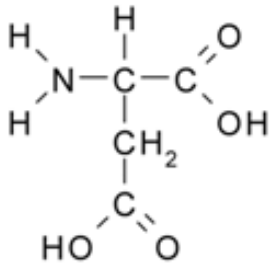
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L
C	9															
S	−1	4														
T	−1	1	5													
P	−3	−1	−1	7												
A	0	1	0	−1	4											
G	−3	0	−2	−2	0	6										
N	−3	1	0	−2	−2	0	6									
D	−3	0	−1	−1	−2	−1	1	6								
E	−4	0	−1	−1	−1	−2	0	2	5							
Q	−3	0	−1	−1	−1	−2	0	0	2	5						
H	−3	−1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

BLOSUM score of 2 indicates that the mutation would be expected to occur 2 times more frequently than random.

Score = 2 $\log_2(q_{EP}/p_E p_Q)$



Glutamic acid



Aspartic acid

BLOSUM substitution matrices are based on the conservation of domains in proteins

Pairwise Sequence Alignment

Scoring Alignment

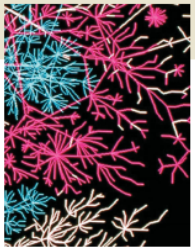
How to Select a Scoring Matrix?

For general alignment purpose, **BLOSUM62** or **PAM120** are recommended.
For **PAM**x matrices, higher x detects more divergent sequences;
For **BLOSUM**n matrices, lower n detects more divergent sequences.
PAM120 matrix is the most appropriate for database searches
PAM200 matrix is the most appropriate for comparing two specific proteins with suspected homology

Pairwise Sequence Alignment

Scoring Alignment

Nature Biotechnology 22, 1035 - 1036 (2004)



_computational BIOLOGY

PRIMER

Nature Publishing Group http://www.nature.com/naturebiotechnology

Where did the BLOSUM62 alignment score matrix come from?

Sean R Eddy

Many sequence alignment programs use the BLOSUM62 score matrix to score pairs of aligned residues. Where did BLOSUM62 come from?

Back in the good old days, so many things were easier to understand. I once disassembled the engine of my 1972 MG just to see how it worked, but now I won't touch the squirrel's nest of technology that's inside my modern Honda Civic. Likewise, in the early days of sequence comparison, alignment scores were straightforward stuff that anybody could tweak. The first sequence comparisons just assigned –1 per mismatch and –1 per insertion/deletion, and if you

ment score is the sum of individual log-odds scores for each aligned residue pair. Those individual scores make up a 20 × 20 score matrix. The equation for calculating a score $s(a,b)$ for aligning two residues a and b is:

$$s(a,b) = \frac{1}{\lambda} \log \frac{p_{ab}}{f_a f_b}$$

The numerator (p_{ab}) is the likelihood of the hypothesis we want to test: that these

chance ($p_{ab} > f_a f_b$), then the odds ratio is greater than one and the score is positive. Operationally, we say that positive scores mean conservative substitutions, and negative scores indicate nonconservative substitutions. This definition of 'conservative substitution' in a score matrix is purely statistical. It has nothing directly to do with amino acid structure or biochemistry.

This explains some details in BLOSUM62 that may seem counterintuitive at first

Pairwise Sequence Alignment

Scoring Alignment

```
pam(50,'extended','false')

ans =

     5    -5    -2    -2    -5    -3    -1    -1    -5    -3    -5    -5    -4    -7     0     0     0   -11    -6    -1
    -5     8    -4    -7    -6     0    -7    -7     0    -4    -7     1    -3    -8    -3    -2    -5    -1    -8    -6
    -2    -4     7     2    -8    -2    -1    -2     1    -4    -6     0    -6    -7    -4     1    -1    -7    -3    -6
    -2    -7     2     7   -11    -1     3    -2    -2    -6   -10    -3    -8   -12    -6    -2    -3   -12    -9    -6
    -5    -6    -8   -11     9   -11   -11    -7    -6    -5   -12   -11   -11   -10    -6    -2    -6   -13    -3    -5
    -3     0    -2    -1   -11     8     2    -5     2    -6    -4    -2    -3   -10    -2    -4    -4   -10    -9    -5
    -1    -7    -1     3   -11     2     7    -3    -3    -4    -7    -3    -5   -11    -4    -3    -4   -13    -7    -5
    -1    -7    -2    -2    -7    -5    -3     6    -7    -8    -9    -6    -7    -8    -4    -1    -4   -12   -11    -4
    -5     0     1    -2    -6     2    -3    -7     9    -7    -5    -4    -8    -5    -3    -4    -5    -6    -2    -5
    -3    -4    -4    -6    -5    -6    -4    -8    -7     8     0    -5     0    -1    -7    -5    -1   -11    -5     3
    -5    -7    -6   -10   -12    -4    -7    -9    -5     0     6    -6     2    -1    -6    -7    -5    -5    -5    -1
    -5     1     0    -3   -11    -2    -3    -6    -4    -5    -6     6    -1   -11    -5    -3    -2    -9    -8    -7
    -4    -3    -6    -8   -11    -3    -5    -7    -8     0     2    -1    10    -3    -6    -4    -3   -10    -8     0
    -7    -8    -7   -12   -10   -10   -11    -8    -5    -1    -1   -11    -3     9    -8    -5    -7    -3     3    -6
     0    -3    -4    -6    -6    -2    -4    -4    -3    -7    -6    -5    -6    -8     8    -1    -3   -11   -11    -4
     0    -2     1    -2    -2    -4    -3    -1    -4    -5    -7    -3    -4    -5    -1     6     1    -4    -5    -4
     0    -5    -1    -3    -6    -4    -4    -4    -5    -1    -5    -2    -3    -7    -3     1     6   -10    -5    -2
   -11    -1    -7   -12   -13   -10   -13   -12    -6   -11    -5    -9   -10    -3   -11    -4   -10    13    -4   -12
    -6    -8    -3    -9    -3    -9    -7   -11    -2    -5    -5    -8    -8     3   -11    -5    -5    -4     9    -6
    -1    -6    -6    -6    -5    -5    -5    -4    -5     3    -1    -7     0    -6    -4    -4    -2   -12    -6     7
```

Example

```
Smat = pam(50)
Smat = pam(50,'extended','false')
```

Pairwise Sequence Alignment

Scoring Alignment

Summary:

Substitution matrices allow to detect similarities between more distant proteins than what would be detected with the simple identity of residues.

- Different substitution scoring matrices have been established

Limitations of the substitution scoring matrices

- They assumed independence between successive residues
- They have been derived from manually aligned sequences
- They have been built from a limited data set