

## The analysis of multiple DNA or protein sequences (III)-Weight Matrices for Sequence Similarity Scoring

### Importance of scoring matrices

- Scoring matrices appear in all analysis involving sequence comparison.
- The choice of matrix can strongly influence the outcome of the analysis.
- Scoring matrices implicitly represent a particular theory of evolution.
- Understanding theories underlying a given scoring matrix can aid in making proper choice.

When we consider scoring matrices, we encounter the convention that matrices have numeric indices corresponding to the rows and columns of the matrix. That is,  $M_{11}$  refers to the entry at the first row and the first column. In general,  $M_{ij}$  refers to the entry at the  $i$ th row and the  $j$ th column. To use this for sequence alignment, we simply associate a numeric value to each letter in the alphabet of the sequence. For example, if the alphabet is

$$\mathcal{A} = \{\mathbf{A}; \mathbf{C}; \mathbf{G}; \mathbf{T}\}$$

then  $A = 1$ ,  $C = 2$ , etc. Thus, one would find the score for a match between A and C at  $M_{12}$ .

### Some Examples

#### Nucleotide scoring:

##### 1. Identity matrix (similarity)

	A	T	C	G
A	1	0	0	0
T	0	1	0	0
C	0	0	1	0
G	0	0	0	1

For elements in row  $i$  by column  $j$ :  $S_{ij} = 1$  when  $i = j$ ;  $S_{ij} = 0$  when  $i \neq j$ .

##### 2. BLAST matrix (similarity)

	A	T	C	G
A	5	-4	-4	-4
T	-4	5	-4	-4
C	-4	-4	5	-4
G	-4	-4	-4	5

##### 3. Transition/Transversion Matrix

	A	T	C	G
A	0	5	5	1
T	5	0	1	5
C	5	1	0	5
G	1	5	5	0

Nucleotide bases fall into two categories depending on the ring structure of the base. Purines (Adenine and Guanine) are two ring bases, pyrimidines (Cytosine and Thymine) are single ring

bases. Mutations in DNA are changes in which one base is replaced by another. A mutation that conserves the ring number is called a transition (e.g., A → G or C → T). A mutation that changes the ring number is called a transversion (e.g. A → C or A → T and so on). Although there are more ways to create a transversion, the number of transitions observed to occur in nature (i.e., when comparing related DNA sequences) is much greater. Since the likelihood of transitions is greater, it is sometimes desirable to create a weight matrix which takes this propensity into account when comparing two DNA sequences. Use of a Transition/Transversion Matrix reduces noise in comparisons of distantly related sequences.

### Protein scoring

1. Identity matrix:  $S_{ij} = 1$  when  $i = j$ ,  $S_{ij} = 0$  when  $i \neq j$
2. Genetic Code Matrix

		A	S	G	L	K	V	T	P	E	D	N	I	Q	R	F	Y	C	H	M	W	Z	B	X
Ala	=	A	0	1	1	2	2	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2
Ser	=	S	1	0	1	1	2	2	1	1	2	2	1	1	2	1	1	1	1	2	2	1	2	2
Gly	=	G	1	1	0	2	2	1	2	2	1	1	2	2	2	1	2	2	1	2	2	1	2	2
Leu	=	L	2	1	2	0	2	1	2	1	2	2	1	1	1	1	2	2	1	1	1	2	2	
Lys	=	K	2	2	2	2	0	2	1	2	1	2	1	1	1	1	2	2	2	1	2	1	2	
Val	=	V	1	2	1	1	2	0	2	2	1	1	2	1	2	2	1	2	2	1	2	2	2	
Thr	=	T	1	1	2	2	1	2	0	1	2	2	1	1	2	1	2	2	2	1	2	2	2	
Pro	=	P	1	1	2	1	2	2	1	0	2	2	2	2	1	1	2	2	2	1	2	2	2	
Glu	=	E	1	2	1	2	1	1	2	2	0	1	2	2	1	2	2	2	2	2	2	1	2	
Asp	=	D	1	2	1	2	2	1	2	2	1	0	1	2	2	2	2	1	2	1	2	2	1	
Asn	=	N	2	1	2	2	1	2	1	2	2	1	0	1	2	2	2	1	2	1	2	2	1	
Ile	=	I	2	1	2	1	1	1	1	2	2	2	1	0	2	1	1	2	2	2	1	2	2	
Gln	=	Q	2	2	2	1	1	2	2	1	1	2	2	0	1	2	2	2	1	2	2	1	2	
Arg	=	R	2	1	1	1	1	2	1	1	2	2	2	1	1	0	2	2	1	1	1	2	2	
Phe	=	F	2	1	2	1	2	1	2	2	2	2	1	2	2	0	1	1	2	2	2	2	2	
Tyr	=	Y	2	1	2	2	2	2	2	2	1	1	2	2	2	1	0	1	1	3	2	2	1	
Cys	=	C	2	1	1	2	2	2	2	2	2	2	2	2	1	1	1	0	2	2	1	2	2	
His	=	H	2	2	2	1	2	2	2	1	2	1	1	2	1	1	2	1	0	2	2	2	1	
Met	=	M	2	2	2	1	1	1	1	2	2	2	2	1	2	1	2	3	2	2	0	2	2	
Trp	=	W	2	1	1	1	2	2	2	2	2	2	2	2	1	2	2	1	2	2	0	2	2	
Glx	=	Z	2	2	2	2	1	2	2	2	1	2	2	2	1	2	2	2	2	2	2	1	2	
Asx	=	B	2	2	2	2	2	2	2	2	1	1	2	2	2	2	1	2	1	2	2	2	1	
???	=	X	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	

Score based on minimum number of base changes required to convert one amino acid into another.

3. Physical/chemical characteristics

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

Attempt to quantify some physical or chemical attribute of the residues and arbitrarily assign weights based on similarities of the residues in this chosen property.

---

Log odds matrices:  $S_{ij} = \log \frac{q_{ij}}{p_i p_j}$

S is the log odds ratio of two probabilities: the probability that two residues,  $i$  and  $j$ , are aligned by evolutionary descent and the probability that they are aligned by chance.

- $q_{ij}$  are the frequencies that residue  $i$  and  $j$  are observed to align in sequences known to be related. They are derived from a “transition probability matrix.”
- $p_i$  and  $p_j$  are frequencies of occurrences of residue  $i$  and  $j$  in the set of sequences.
- e.g., PAM250, BLOSUM62 et al.

---

## PAM and BLOSUM matrices (OPTIONAL)

### PAM Matrix

Summary of steps:

1. Align sequences that are at least 85% identical.
  - Minimize ambiguity in alignments.
  - Minimize the number of coincident mutations.
2. Reconstruct phylogenetic trees and infer ancestral sequences. 71 trees containing 1,572 exchanges were used.
3. Tally replacements “accepted” by natural selection, in all pair-wise comparisons (each  $A_{ij}$  is the number of times amino acid  $j$  was replaced by amino acid  $i$  in all comparisons).
4. Compute amino acid mutability,  $m_j$ , i.e., the propensity of a given amino acid,  $j$ , to be replaced.
5. Combine data from 3 & 4 to produce a *Mutation Probability Matrix* for one PAM of evolutionary distance, according to the following formulae:

$$M_{ij} = \frac{m_j A_{ij}}{\sum_i A_{ij}}, \quad M_{jj} = 1 - m_j$$

6. Calculate *Log Odds Matrix* for similarity scoring: Divide each element of the Mutation Data Matrix,  $M$ , by the frequency of occurrence of each residue:

$$R_{ij} = \frac{M_{ij}}{f_i}, \quad R \text{ is a } \textit{Relatedness Odds Matrix}, f_i \text{ is the frequency of residue } i.$$

The Log Odds Matrix,  $S_{ij}$ , is calculated from the relatedness odds matrix,  $R_{ij}$ , simply by taking the log of each  $R_{ij}$ .

7. Different protein families manifest different PAM rates.

### Assumptions in PAM model:

1. Replacement (mismatch) at any site depends only on the amino acid at that site and the probability given by the table (Markov model).
2. Sequences that are being compared have average amino acid composition.

### Sources of error in PAM model

1. Many sequences depart from average composition.

2. Rare replacements were observed too infrequently to resolve relative probabilities accurately (for 36 pairs no replacements were observed!).
3. Errors in 1PAM are magnified in the extrapolation to 250 PAM.
4. The Markov process is an imperfect representation of evolution: Distantly related sequences usually have islands (blocks) of conserved residues. This implies that replacement is not equally probable over entire sequence.

### **BLOSUM (Blocks Substitution Matrix) Matrix**

Steven Henikoff and Jorja G. Henikoff (1992). Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. 89: 10915-10919.

1. Starting data is conserved blocks:
  - Aligned, un-gapped sequences
  - Widely varying similarity, but measures are taken to avoid biasing the sample with frequently occurring highly related sequences.
2. Tallies of replacements are made by straight forward tallying of all pairs of aligned residues  $f_{ij}$ .
  - The observed frequency of each pair is:  $q_{ij} = f_{ij}/(\text{total number of residue pairs})$ .
  - This includes cases of  $i=j$  (i.e. no replacement observed).
  - The expected frequency of each pair is essentially the product of the frequencies of each residue in the data set.
3. Similar sequences in a block, above a threshold percent similarity are clustered and members of the cluster count fractionally toward the final tally.
  - Reduces the number of identical pairs (AA, SS, TT, etc., matches) in the final tallies.
  - Somewhat analogous to increasing the PAM distance.
  - If clustering threshold is 80%, final matrix is BLOSUM 80.
  - Clustering at 62% reduces the number of blocks contributing to the table by 25%-still  $1.25 \times 10^6$  pairs contributed!
  - Least frequent amino acid pair replacement was observed 2369 times!