Sequence Alignment: Scoring Schemes

COMP 571
Luay Nakhleh, Rice University

Scoring Schemes

* Recall that an alignment score is aimed at providing a scale to measure the degree of similarity (or difference) between two sequences and thus make it possible to quickly distinguish among the many subtly different alignments that can be generated for any two sequences
* Scoring schemes contain two separate elements:
  * the first assigns a value to a pair of aligned residues
  * the second assigns penalties to gaps

Deriving a Substitution Matrix

* The alignment score attempts to measure the likelihood of a common evolutionary ancestor

* To achieve this mathematically, we consider the alignment of two residues from two sequences under two "competing" models: a **random** model, R, and a **match** (non-random, evolutionary) model, M

The Random Model (R)

* All sequences are assumed to be random selections from a given pool of residues, with every position in the sequence totally independent of every other

* Thus for a protein sequence, if the proportion of amino acid type a in the pool is p_a, this fraction will be reproduced in the amino acid composition of the protein

* In this model, the probability of residue a being aligned with residue b is simply $p_a p_b$

The Match Model (M)

* Sequences are related, due to an evolutionary process, and there is a high correlation between aligned residues
* The probability of occurrence of particular residues thus depends not on the pool of available residues, but on the residue at the equivalent position in the sequence of the common ancestor
* In this model, the probability of residue a being aligned with residue b is $q_{a,b}$, where the actual values of $q_{a,b}$ depend on the properties of the evolutionary process

The Odds Ratio

* So, we have $P(a,b|R)=p_a p_b$ and $P(a,b|M)=q_{a,b}$
* These two models can be compared by taking the odds ratio $q_{a,b}/p_a p_b$
* If this ratio is greater than 1, the match model is more likely to have produced the alignment of these residues

The Odds Ratio

* The odds ratio for the entire alignment is taken as the product of the odds ratios for the different positions

$$\prod_u \left(\frac{q_{a,b}}{p_a p_b} \right)_u$$

where u ranges over all positions in the alignment

The Log-odds Ratio

* It is frequently more practical to deal with sums rather than products, especially when small numbers are involved
* This can be achieved by taking logarithms of the odds ratio to give the log-odds ratio.
* This ratio can be summed over all positions of the alignment to give S, the score of the alignment:

$$S = \sum_u \log\left(\frac{q_{a,b}}{p_a p_b}\right)_u = \sum_u (s_{a,b})_u$$

where $s_{a,b}$ is the substitution matrix element associated with the alignment of residue types a and b

The Log-odds Ratio

* A positive value of $s_{a,b}$ means that the probability of those two residues being aligned is greater in the match model than in the random model
* The converse is true for negative $s_{a,b}$ values
* S is a measure of the relative likelihood of the whole alignment arising due to the match model as compared with the random model
* However, a positive S is not a sufficient test of the alignment's significance (more on the significance of scores later)

PAM Scoring Matrices

* It is strongly argued that the scoring matrices are best developed based on experimental data, thus reflecting the kind of relationships occurring in nature

* The first scoring matrices developed from known data were the PAM matrices

* Point accepted mutations matrix, derived by Dayhoff et al.

* Dayhoff et al. estimated the substitution probabilities by using known mutational histories (mutation here means substitution)

* 34 protein superfamilies were used, divided into 71 groups of near homologous sequences (>85% identity to reduce the number of superimposed mutations) and a phylogenetic tree was constructed for each group (including the inference of the most likely ancestral sequences at each internal node)

PAM Scoring Matrices

* Then, the **accepted point mutations** on each edge were estimated
  * A mutation is **accepted** if it is accepted by the species
  * This usually means that the new amino acid must have the same effect (must function in a similar way) as the old one, which usually requires strong physio-chemical similarity, dependent on how critical the position of the amino acid is

PAM Scoring Matrices

* Let τ be a time interval of evolution, measured in numbers of mutations per residue
* Dayhoff's procedure used the following steps:

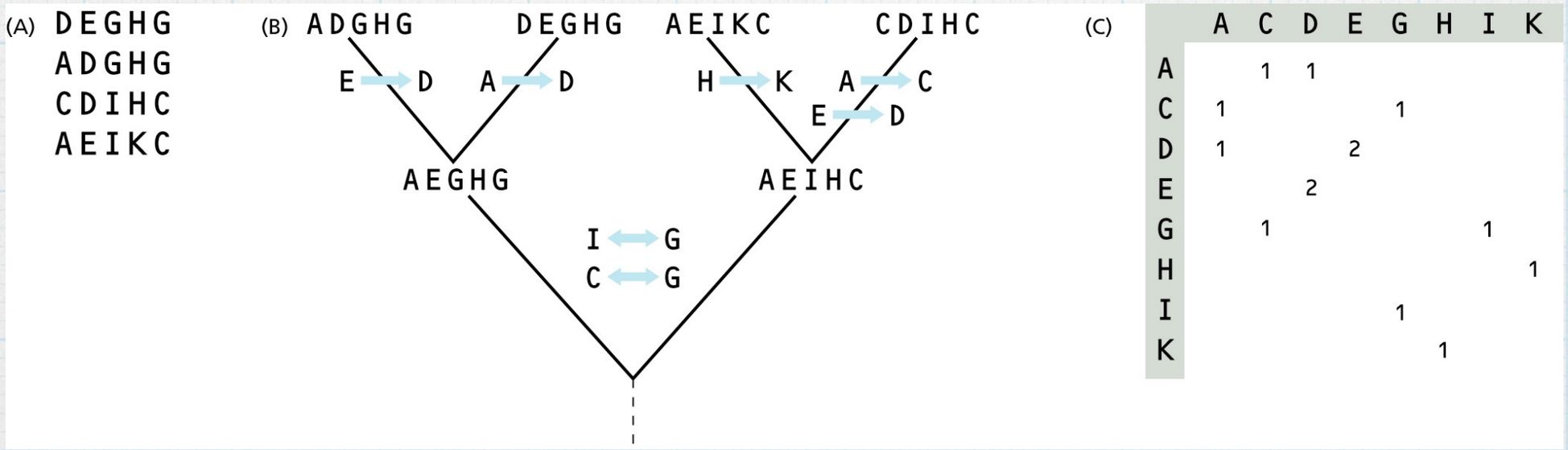
1. Divide the set of sequences into groups of similar sequences, and make a multiple alignment of each group
2. Construct phylogenetic trees for each group, and estimate the mutations on the edges
3. Define an evolutionary model to explain the evolution
4. Construct substitution matrices (the substitution matrix for an evolutionary interval τ given for each pair (a,b) of residues an estimate for the probability of a to mutate to b in a time interval τ)
5. Construct scoring matrices from the substitution matrices

The Evolutionary Model

* The evolutionary model used has the following assumption: the probability of a mutation in one position of a sequence is only dependent on which amino acid is in that position
* It is
  * independent of position and neighbor residues, and
  * independent of previous mutations in the position
* The biological clock is also assumed, which means that the rate of mutations is constant over time
* Hence, the time of evolution can be measured by the number of mutations observed in a certain number of residues
* This is measured in point accepted mutations (PAMs), and 1 PAM means one accepted mutation per 100 residues

Calculating the Substitution Matrix

* The substitution matrix is calculated by observing the number of accepted mutations in the constructed phylogenetic trees (1572 in the first experiment)



Calculating the Substitution Matrix

* The task is then to calculate a value for the relation between the amino acids a and b in terms of mutations
* This is done by first estimating the probability that a will be replaced by b in a certain evolutionary time τ, and denote this by M^{τ}_{ab}
* τ is measured in PAMs, and first we look at $\tau=1$ (M^{1}_{ab})
* When $\tau=1$, the time specification is often omitted, and the probability denoted by M_{ab}
* Note that M_{ab} need not be equal to M_{ba}
* M_{ab} depends on (1) the probability that a mutates and (2) the probability that a mutates to b given that a mutates

Calculating the Substitution Matrix

* The procedure can be described as follows

1. Find all accepted mutations in the data. From this calculate f_{ab} (the number of mutations from a to b or b to a), f_a (the total number of mutations that involve a), and f (the sum of f_a for all a)

2. Calculate the frequency p_a for all a (this is the relative occurrence of amino acid a in the data)

3. Calculate the relative mutability m_a, which is a measure of the probability that a will mutate in the evolutionary time of interest. m_a depends on f_a (m_a should increase with increasing f_a) and p_a (m_a should decrease with increasing p_a). Hence, m_a can be defined as $m_a = K f_a / p_a$, where K is a constant (for the value of K, see the next slide)

4. For determining M_{ab} we can now use the facts that (a) the probability that a mutates (in time 1 PAM) is m_a, and (b) the probability that a mutates to b, given that a mutates, is f_{ab}/f_a. Therefore,

   - for $a \neq b$, $M_{ab} = m_a f_{ab}/f_a$
   - for $a = b$, $M_{aa} = 1 - m_a$

The Constant K

* The probability that an arbitrary mutation contains a is $f_a/(f/2)$
* The probability that it is from a is (since $f_{ab}=f_{ba}$) is $1/2\,(f_a/(f/2))= f_a/f$
* Among 100 residues there are $100p_a$ occurrences of a, hence the probability for any one of these to mutate is

$$m_a = \frac{1}{100p_a}\frac{f_a}{f} = \frac{1}{100f}\frac{f_a}{p_a}$$

- As a check, we can find expected number of mutations per 100 residues

$$\sum_a (100p_a)m_a = \sum_a 100p_a \frac{f_a}{f100p_a} = \frac{1}{f}\sum_a f_a = \frac{f}{f} = 1$$

ORIGINAL AMINO ACID

REPLACEMENT AMINO ACID

		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	1	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	2	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

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

Matrices for General Evolutionary Times

* Due to the independence properties of the model (Markov model), M^z, for an arbitrary evolutionary time z, can be computed as M raised to the power z (matrix M multiplied by itself z times)

ORIGINAL AMINO ACID

REPLACEMENT AMINO ACID	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	13	6	9	9	5	8	9	12	5	8	6	7	7	4	11	11	11	2	4	9
R Arg	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
N Asn	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
D Asp	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
C Cys	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
Q Gln	3	5	5	6	1	10	7	3	7	2	3	5	2	1	4	3	3	1	2	3
E Glu	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
G Gly	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
H His	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
I Ile	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
L Leu	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
K Lys	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
M Met	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
F Phe	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
P Pro	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
S Ser	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	5
T Thr	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
W Trp	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
Y Tyr	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
V Val	7	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	7	2	4	17

Figure 83. Mutation probability matrix for the evolutionary distance of 250 PAMs. To simplify the appearance, the elements are shown multiplied by 100. In comparing two sequences of average amino acid frequency at this evolutionary distance, there is a 13% probability that a position containing Ala in the first sequence will contain Ala in the second. There is a 3% chance that it will contain Arg, and so forth. The relationship of two sequences at a distance of 250 PAMs can be demonstrated by statistical methods.

Substitution Matrices

* These matrices tell how many mutations have been accepted, but not the percentage of residues that have mutated: some may have mutated more than once, others not at all

* Suppose two sequences q and d have evolutionary distance τ (τ mutations per 100 residues have occurred in the transition from the ancestral sequence, say q, to the derived one, d)

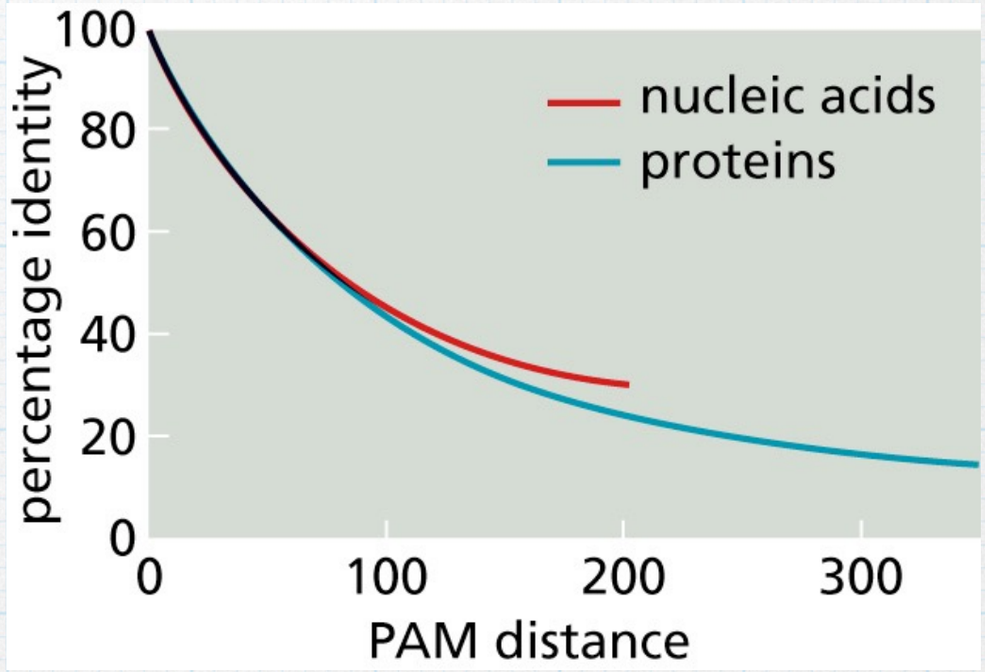
* With

$$100\left(1-\sum_c p_c M_{cc}^{\tau}\right)$$

we find how many residues on average are different per 100 residues

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
45	67
50	80
55	94
60	112
65	133
70	159
75	195
80	246
85	328



Obtaining a Scoring Matrix

* So far we have obtained a substitution matrix, but not a scoring matrix
* Using the log-odds ratio, we need to divide the probability under the match model (given by the substitution matrix) by the probability under the random model

$$S_{ab} = \log \frac{M_{ab}}{p_b}$$

PAM120

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

BLOSUM Scoring Matrices

* In the Dayhoff model, the scoring values are derived from protein sequences with at least 85% identity
* Alignments are, however, most often performed on sequences of less similarity, and the scoring matrices for use in these cases are calculated from the 1 PAM matrix
* Henikoff and Henikoff (1992) have therefore developed scoring matrices based on known alignments of more diverse sequences

BLOSUM Scoring Matrices

* They take a group of related proteins and produce a set of blocks representing this group, where a block is defined as an ungapped region of aligned amino acids
* An example of two blocks is

K	I	F	I	M	K
N	L	F	K	T	R
K	I	F	K	T	K
K	L	F	E	S	R
K	I	F	K	G	R

G	D	E	V	K
G	D	S	K	K
G	D	P	K	A
G	D	A	E	R
G	D	A	A	K

BLOSUM Scoring Matrices

* The Henikoffs used over 2000 blocks in order to derive their scoring matrices
* For each column in each block they counted the number of occurrences of each pair of amino acids, when all pairs of segments were used
* Then the frequency distribution of all 210 different pairs of amino acids were found
* A block of length w from an alignment of m sequences makes (wm(m-1))/2 pairs of amino acids

BLOSUM Scoring Matrices

* We define
  * h_{ab} as the number of occurrences of the amino acid pair (ab) (note that $h_{ab}=h_{ba}$)
  * T as the total number of pairs in the alignment

$$T = \sum_{c} \sum_{e \geq c} h_{ce}, \qquad c, e \in \mathcal{M}$$

    where ≥ is interpreted as a total ordering over the amino acids

- $f_{ab}=h_{ab}/T$ (the frequency of observed pairs)

BLOSUM Scoring Matrices

* Example

K	I	F	I	M	K
N	L	F	K	T	R
K	I	F	K	T	K
K	L	F	E	S	R
K	I	F	K	G	R

$h_{KR}=6$
$h_{KK}=1$
$h_{RR}=3$

G	D	E	V	K
G	D	S	K	K
G	D	P	K	A
G	D	A	E	R
G	D	A	A	K

For the two blocks: $h_{KR}=9$; there are 110 pairs
Hence: $f_{KR}=9/110$

Log-odds Matrix

* For each pair (ab), the expected probability that they are aligned by chance, e_{ab}, must be calculated
* Then
  * $f_{ab} > e_{ab}$, the observed frequency is higher than expected by chance, which indicates a biological relation between the amino acids a and b
  * $f_{ab} < e_{ab}$, the observed frequency is less than expected by chance, which indicates a biological ´aversion´ between the amino acids a and b
  * $f_{ab} = e_{ab}$, which indicates biological neutrality between the amino acids a and b

Log-odds Matrix

* To calculate the expected number of occurrences of the amino acid pairs, assume that the observed frequencies are equal to the frequencies in the actual population
* From this the expected probability that a specific amino acid a is in a pair can be calculated:
  * the number of residues in the considered data is $2T$
  * amino acid a occurs $2h_{aa}+\Sigma_{e\neq a}h_{ae}$ times
  * amino acid a occurs with a frequency of

$$p_a = \frac{2h_{aa} + \sum_{e\neq a} h_{ae}}{2T} = f_{aa} + \sum_{e\neq a} \frac{f_{ae}}{2}$$

- Suppose now that all pairs are separated, and that new pairs are drawn according to the observed frequencies; we get $e_{aa}=p_ap_a$ and $e_{ab}=2p_ap_b$ $(a\neq b)$
- In order to obtain the log-odds matrix we need to calculate the ratio between the observed and the expected frequencies for each amino acid pair
- This is simply f_{ab}/e_{ab}, and working with the logarithm of the odds, we take

$$R_{ab} = \log_2 \frac{f_{ab}}{e_{ab}}$$

Developing Scoring Matrices for Different Evolutionary Distances

* When comparing two sequences q and d with an evolutionary distance X, one should use segment pairs corresponding to this distance for constructing an appropriate scoring matrix

* For developing a matrix for an X% identity (if we take identity to reflect evolutionary distance), similar blocks with X or higher percentage identity are grouped into one group, and treated as one segment

Developing Scoring Matrices for Different Evolutionary Distances

* The procedure for developing a BLOSUM X matrix

  1. Collect a set of multiple alignments
  2. Find the blocks
  3. Group the segments with an X% identity
  4. Count the occurrences of all pairs of amino acids
  5. Develop the matrix, as explained before

- BLOSUM-62 is often used as the standard for ungapped alignments
- For gapped alignments, BLOSUM-50 is more often used

BLOSUM-62

	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

Comparing BLOSUM and PAM Matrices

* The basis for constructing the two sets of matrices is different

* BLOSUM matrices with a low percentage correspond to PAM matrices for large evolutionary distances

* By use of relative entropy, it can be found that PAM250 corresponds to BLOSUM-45 and PAM160 corresponds to BLOSUM-62, and PAM120 corresponds to BLOSUM-80

Comparing BLOSUM and PAM Matrices

* When comparing sequences it is always a question of which PAM or BLOSUM matrix to use, especially when the evolutionary distance between the sequences is unknown

* Different studies have concluded that for the PAM matrices it is generally best to try PAM40, PAM120, and PAM250

* When used for local alignments, lower PAM matrices find short local alignments, but higher PAM matrices find longer but weaker local alignments

Comparing BLOSUM and PAM Matrices

* Often a quick alignment is done first (using, for example, the identity scoring matrix), the evolutionary distance estimated, and the corresponding scoring matrix used
* However, several different matrices should be used, and the alignment that is judged to be evolutionarily the most accurate should be chosen

Scoring Matrices for Nucleotide Sequences

* The same techniques as those just described can be applied to nucleotides, although often simple scoring schemes such as +5 for a match and -4 for a mismatch are used

Gap Penalty Models

* A scoring scheme is required for insertions and deletions in alignments, as they are common evolutionary events

* The simplest method is to assign a gap penalty g on aligning any residue with a gap; that is, g=-E for a positive number E

* If the gap is n_{gap} residues long, then this **linear gap penalty** is defined as $g(n_{gap}) = -n_{gap}E$

Gap Penalty Models

* The observed preference for fewer and longer gaps can be modeled by using a higher penalty to initiate a gap (the **gap opening penalty**, or GOP, designated I) and then a lower penalty to extend an existing gap (the gap **extension penalty**, or GEP, designated E)

* This leads to the **affine gap penalty** forumla

$$g(n_{gap})=-I-(n_{gap}-1)E$$

* Typical ranges of the parameters for protein alignment are 7-15 for I and 0.5-2 for E

Gap Penalty Models

(A)

```
Bovine PI-3Kinase p110a        LNWENPDIMSELLFQNNEIIFKNGDDLRQDMLTLQIIRIMENIWQNQGLDLRMLPYGCLSIGDCVGLIEVVRNSHTIMQIQCKGGLKGAL
cAMP-dependent protein kinase --WENPAQNTAHLDQFERIKTLGTGSFGRVMLVKHMETGNHYAMKILDKQKVVKLKQIEHTLNEKRILQAVNFPFLVKLEFSFKDNSNLY

Bovine PI-3Kinase p110a        QFNSHTLHQWLKDKNKGEIYDAAIDLFTRSCAGYCVATFILGIGDRHNSNIMVKDDGQLFHIDFGHFLDHKKKKFGYKRERVPFVLTQDF
cAMP-dependent protein kinase MVMEYVPGGEMFSHLRRIGRFSEPHARFYAAQIVLTFEYLHSLDLIYRDLKPENLLIDQQGYIQVTDFGFAKRVKGRTWXLCGTPEYLAP

Bovine PI-3Kinase p110a        LIVISKGAQECTKTREFERFQEMCYKAYLAIRQHANLFINLFSMMLGSGMPELQSFDDIAYIRKTLALDKTEQEALEYFMKQMNDAHHGG
cAMP-dependent protein kinase EIILSKGYNKAVDWWALGVLIYEMAAGYPPFFADQPIQIYEKIVSGKVRFPSHFSSDLKDLLRNLLQVDLTKRFGNLKNGVNDIKNHKWF

Bovine PI-3Kinase p110a        WTTKMDWIFHTIKQHALN-----------------------------------
cAMP-dependent protein kinase ATTDWIAIYQRKVEAPFIPKFKGPGDTSNFDDYEEEEIRVXINEKCGKEFSEF
```

Very high gap penalty results in gaps only at beginning and end, and 10% sequence identity

(B)

```
Bovine PI-3Kinase p110a        LNWENPDIMSELLFQNNEIIFKNGDDLRQDMLTLQIIRIMENIWQNQGLDLRMLPYGCLSIGDCVGLIEVVRNSHTIMQIQCKGGLKGAL
cAMP-dependent protein kinase ?-WENPAQNTAHLDQFERIKTLGTGSFGRVMLVKHM--ETGNHYAMKILDKQKV-VKLKQIEHTLNEKRILQAVNFPFLVKLEFSFKDN-

Bovine PI-3Kinase p110a        QFNSHTLHQWLKDKNKGEIYDAAIDLFTRSCAGYCVATFILGIGDRHNSNIMVKD-DGQLFHIDFGHFLDHKKKKFGYKRERVPFVL--T
cAMP-dependent protein kinase -SNLYMVMEYVPGGEMFSHLRR-IGRFSEPHARFYAAQIVLTFEYLHSLDLIYRDLKPENLLIDQQGYIQVTDFGFAKRVKGRTWXLCGT

Bovine PI-3Kinase p110a        QDFL---IVISKGAQECTKTREFERF-QEMC--YKAYLAIRQHANLFINLFSMMLGSGMPELQSFDDIAYIRKTLALDKTEQEALEYFMK
cAMP-dependent protein kinase PEYLAPEIILSKGYNKAVDWWALGVLIYEMAAGYPPFFA-DQPIQIYEKIVSGKVRF--PSHFSSDLKDLLRNLLQVDLTKR--FGNLKN

Bovine PI-3Kinase p110a        QMNDAHHGGWTTKMDWI----------------------FHTIKQHAL----N----------
cAMP-dependent protein kinase GVNDIKNHKWFATTDWIAIYQRKVEAPFIPKFKGPGDTSNFDDYEEEEIRVXINEKCGKEFSEF
```

Very low gap penalty results in many more gaps, and 18% sequence identity