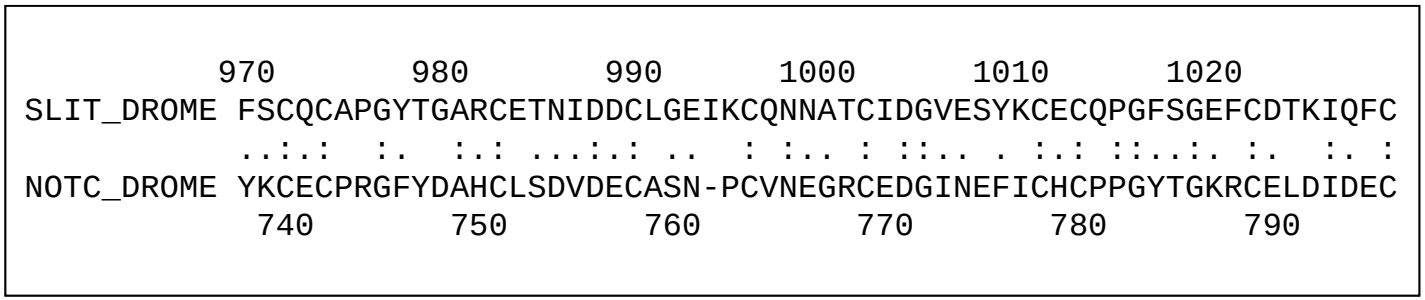
PAIRWISE SEQUENCE COMPARISON



- Comparing sequences for similarity
- Finding motifs
- Prediction of function of genes and proteins
- Construction of phylogeny

Main goals

- To understand the meaning of identity, similarity and homology
- To be able to compare sequences
  - Dot plot
  - Pairwise alignment
- To be able to explain substitution score matrices
- To be able to explain what the differences are between affine and linear gap functions
- To understand the differences between global and local alignments
- To be able to explain the concepts of optimal alignment and dynamic programming

Some definitions

Identity

Proportion of pairs of identical residues between two aligned sequences (generally expressed as a percentage)

This value strongly depends on how the two sequences are aligned.

Random sequences share more than 0% identity !!!

Similarity

Proportion of pairs of identical and similar residues between two aligned sequences. Similar residues are residues that can be substitute for one another without modifying the function

Similarity depends on the substitution matrix used and how the two sequences are aligned.

Distance

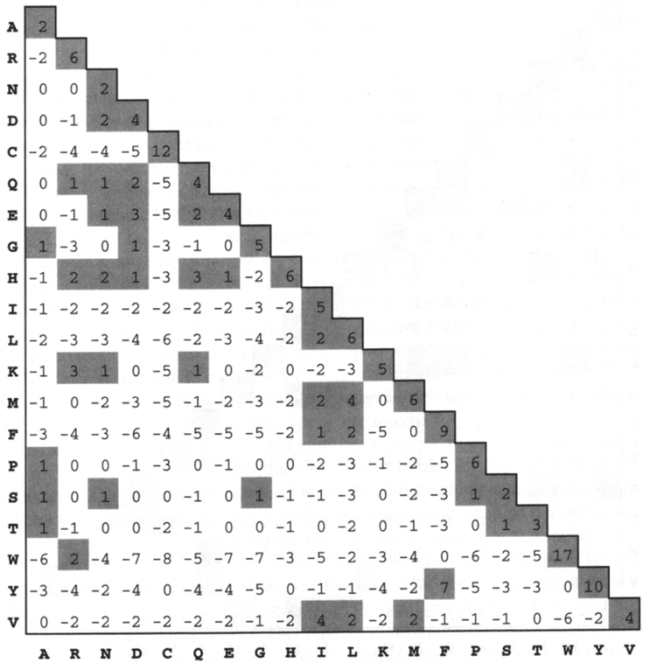
The number of observed changes in an optimal alignment of two sequences (generally expressed as %)

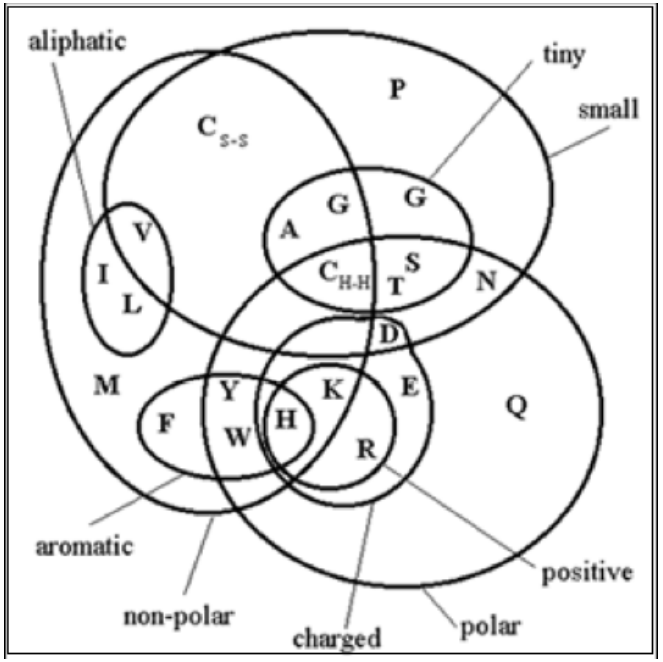
Value depends on sequence alignment method and model for multiple substitution

Length: 22
Identity: 17/22 (77.3%)
Similarity: 18/22 (81.8%)
Gaps: 1/22 (4.5%)
Score: 78.0
Matrix:Blosum62

```
1  GARFIELDTHELASTFA-TCAT   21
   ||||||||||||:...|| ||||
1  GARFIELDTHEVERYFASTCAT   22
```

Similarity reflects frequent amino-acid substitutions among homologous proteins and therefore equivalent physicochemical properties





Effect of the scoring matrix on similarity determination

```
# Matrix: PAM10
# Gap_penalty: 10.0
# Extend_penalty: 0.5
# Length: 25
# Identity: 17/25 (68.0%)
# Similarity: 17/25 (68.0%)
# Gaps: 7/25 (28.0%)
# Score: 109.5
```

```
1 GARFIELDTHE----LASTFATCAT     21
  ||||||||||||     .||   ||||
1 GARFIELDTHEVERYFAS---TCAT     22
```

Gaps are considered as additional residues !
The penalty values are linked to the scoring matrix used

Distance (the *p* distance)

- This distance is the proportion (p) of residue sites at which the two sequences to be compared are different.
- It is obtained by dividing the number of differences by the number of sites compared. Gaps are not considered !
- No correction for multiple substitutions at the same site

```
Sequence 1  1 GARFIELDTHELASTFA-TCAT   21
              ||||||||||||:...|| ||||
Sequence 2  1 GARFIELDTHEVERYFASTCAT   22
```

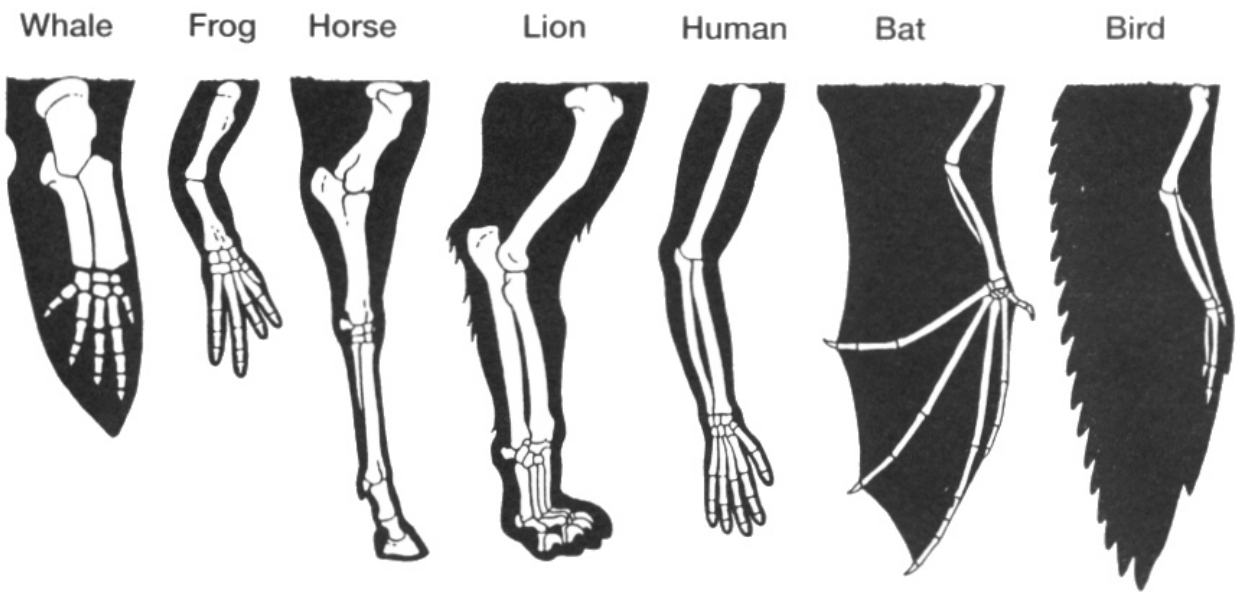
D = 4/21

Distance matrix (dismat)

1	2	
0.00	19.05	1
	0.00	2

D=1-S (the simplest model !)

Homology: an anatomy-based example

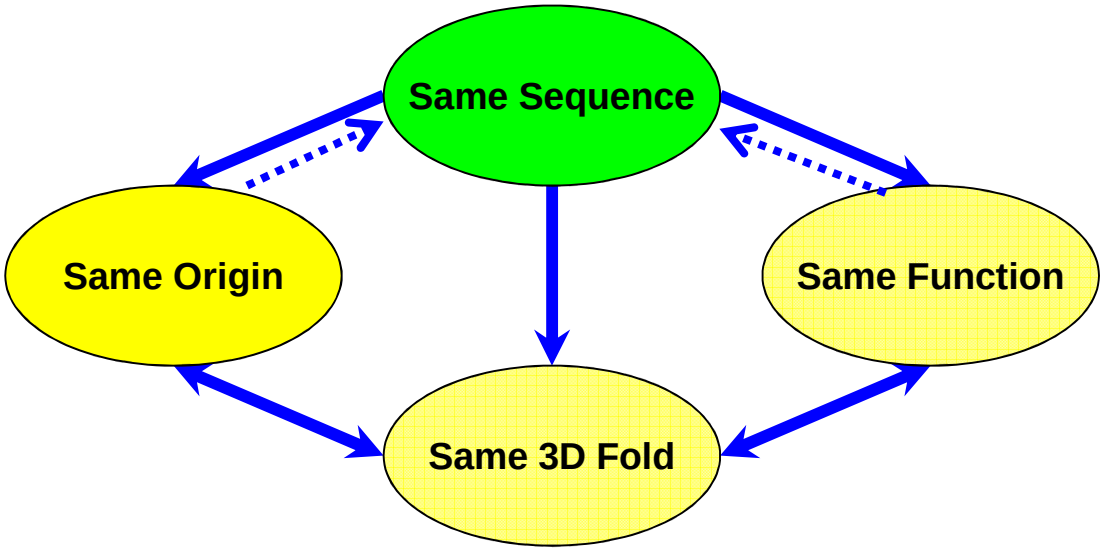


Whale clipper and human leg look different although the limb structures are similar .
Homology focuses on common (limb) ancestry and shows how closely related species have changed from their ancient ancestors

In contrast analogous structures are not necessarily evidence that the two species came from a common ancestor

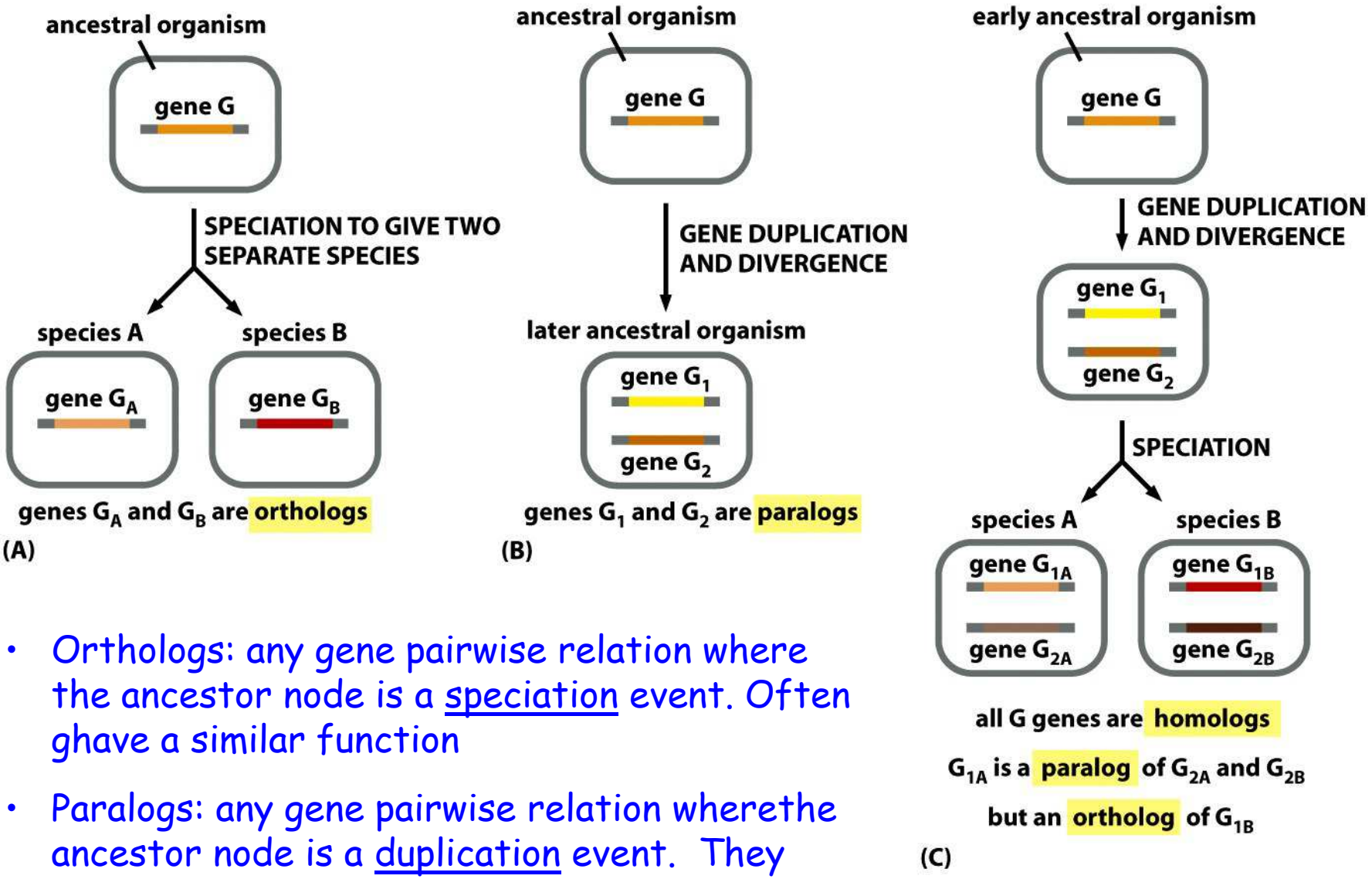
Sequence, function and 3-D fold should not be equally considered indicators of homology

- Homologous sequences do not necessarily serve the same function
- 3-D structure may be conserved while sequence is not



Evolutionary relationships between homologous sequences

Orthology versus paralogy

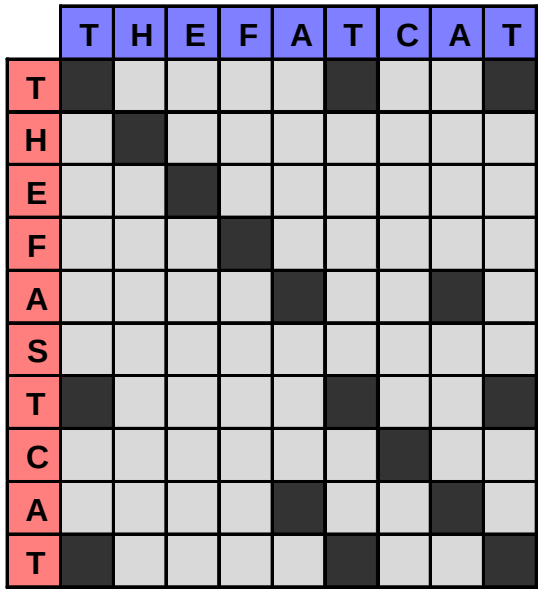


- Orthologs: any gene pairwise relation where the ancestor node is a speciation event. Often ghave a similar function
- Paralogs: any gene pairwise relation wherethe ancestor node is a duplication event. They tend to have different functions

1. Dotplot

A visual assesment of similarity based on identity

- The horizontal and vertical axes correspond to the sequences being compared
  - A dot is placed at each position where two residues match
  - A region of similarity stands out as a diagonal (graphical representation)




```
THEFA-TCAT
|||||  ||||
THEFASTCAT
```

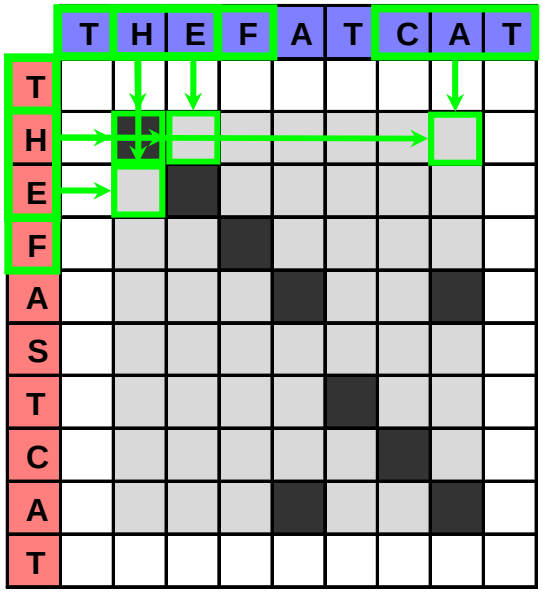
Note

The noise is reduced by calculating a score using a window

The score is then compared to a threshold or stringency

Window approach

- Each window of the first sequence is aligned (without gaps) to each window of the 2nd sequence
- A colour is set into a rectangular array according to the score of the aligned windows (threshold = 16)

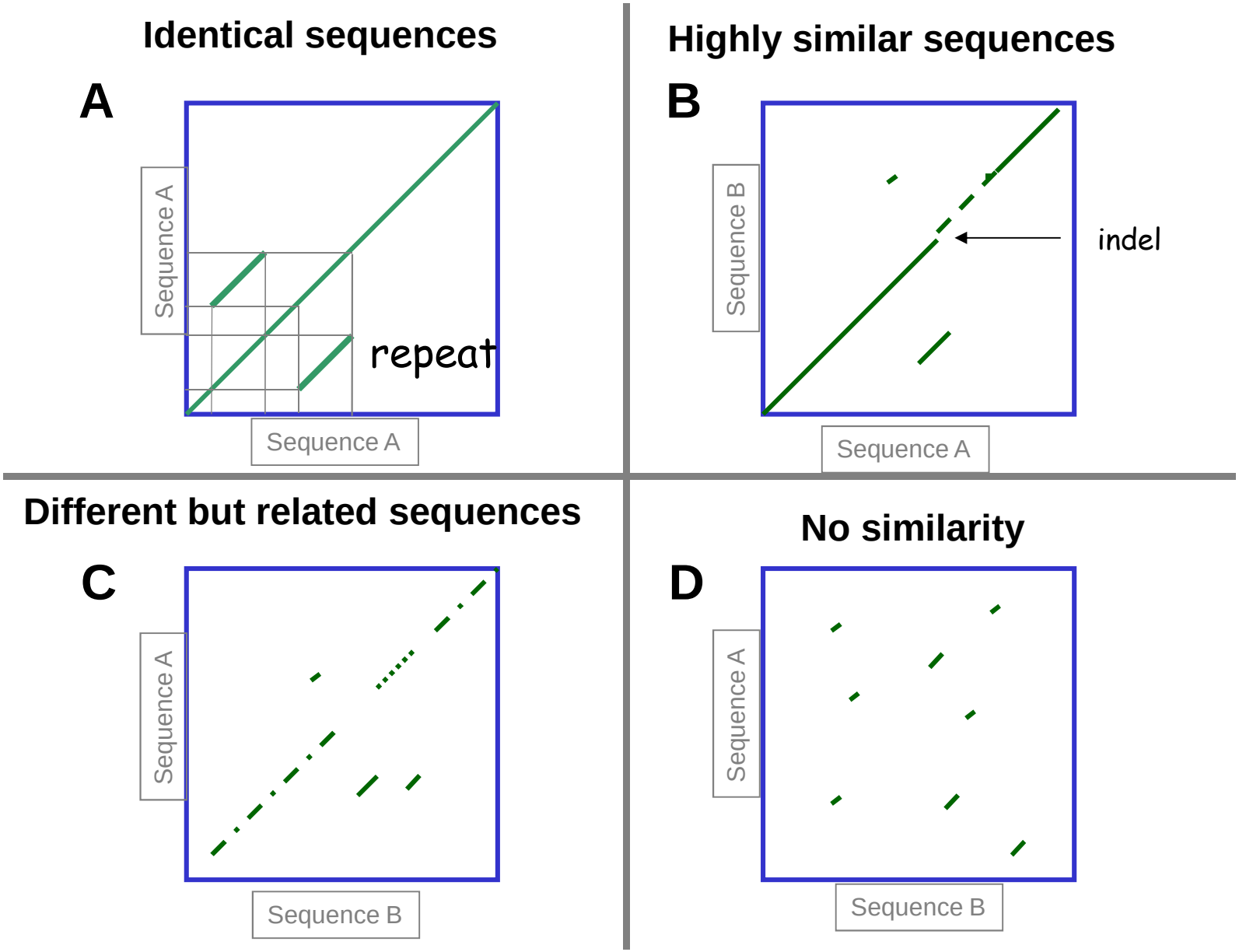


HEF

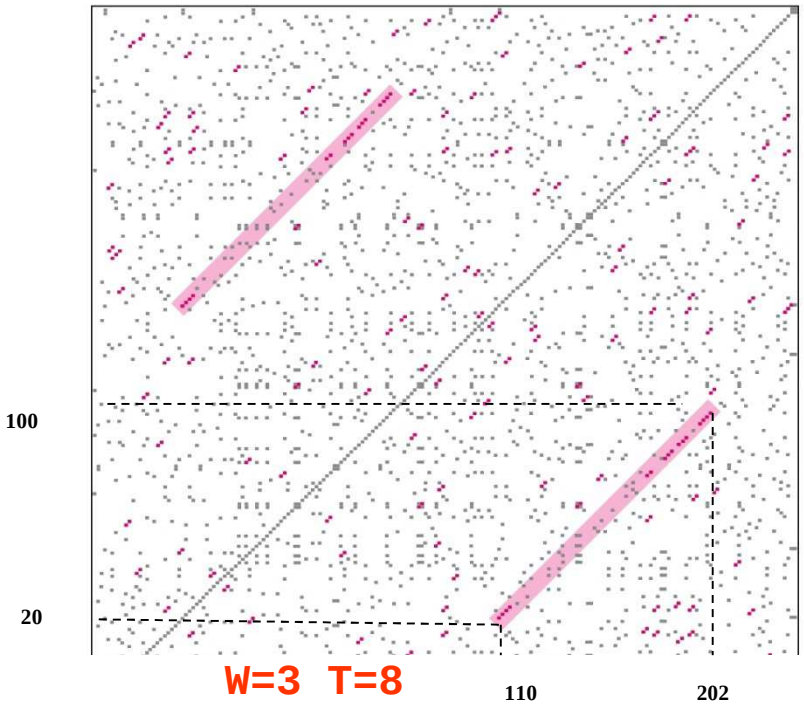
THE

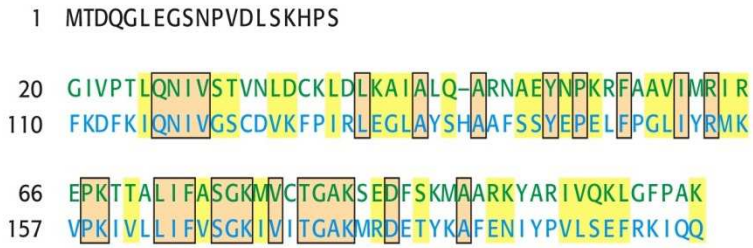
Score: 25

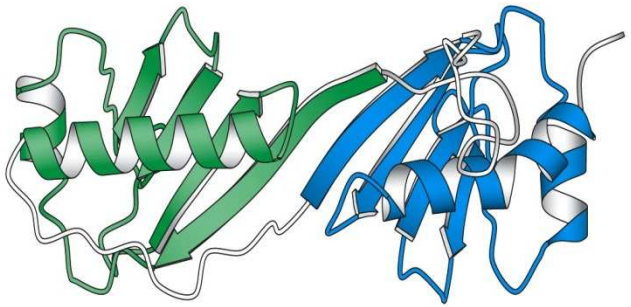
B. Dotplot utility



- Detection of repeats within the TATA-box binding protein TBP
  - direct repeat (same orientation) : protein domains or motifs
  - inverted repeat or palindrome (reverse orientation) : self-complementary sequences in mRNA



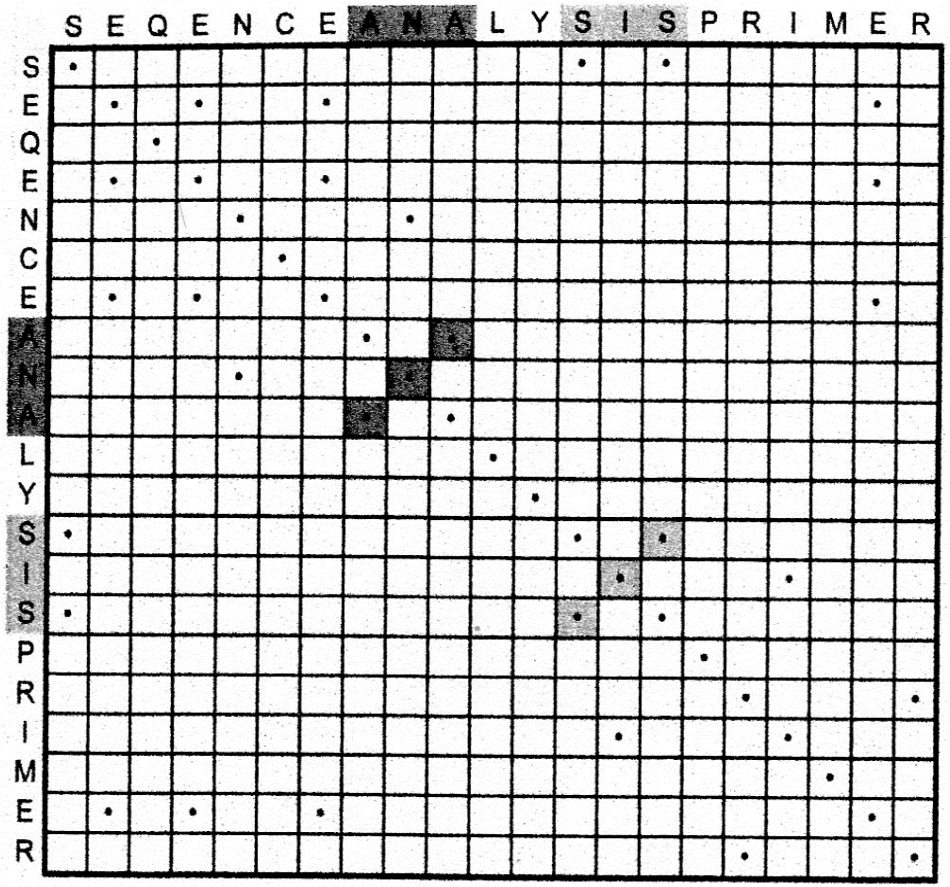




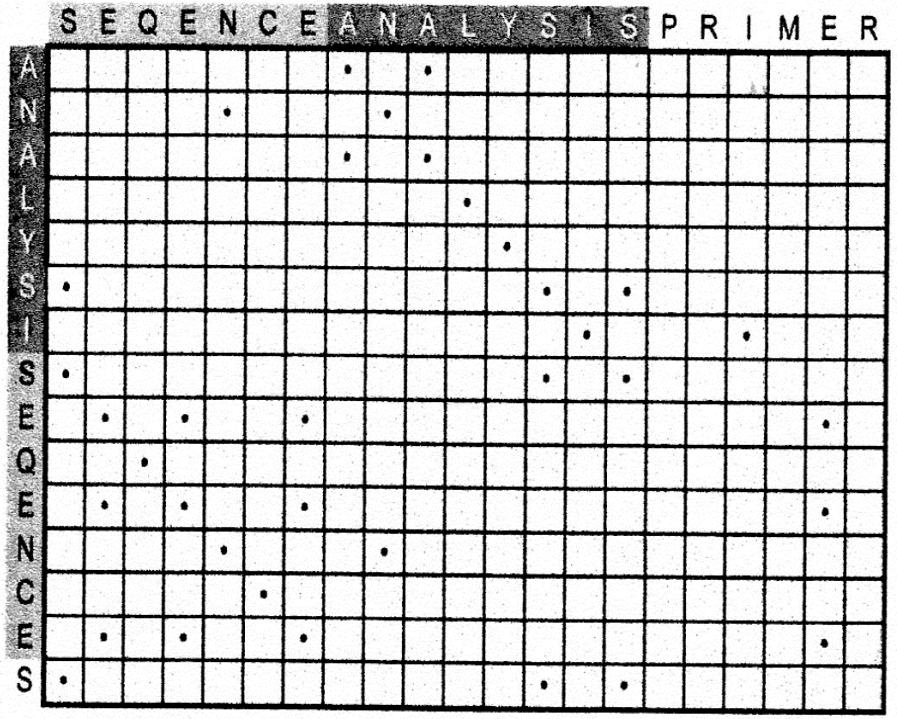
Detection of the repeats of the TATA-box-binding protein from the plant Arabidopsis

- Comparison of genomic and cDNA copies

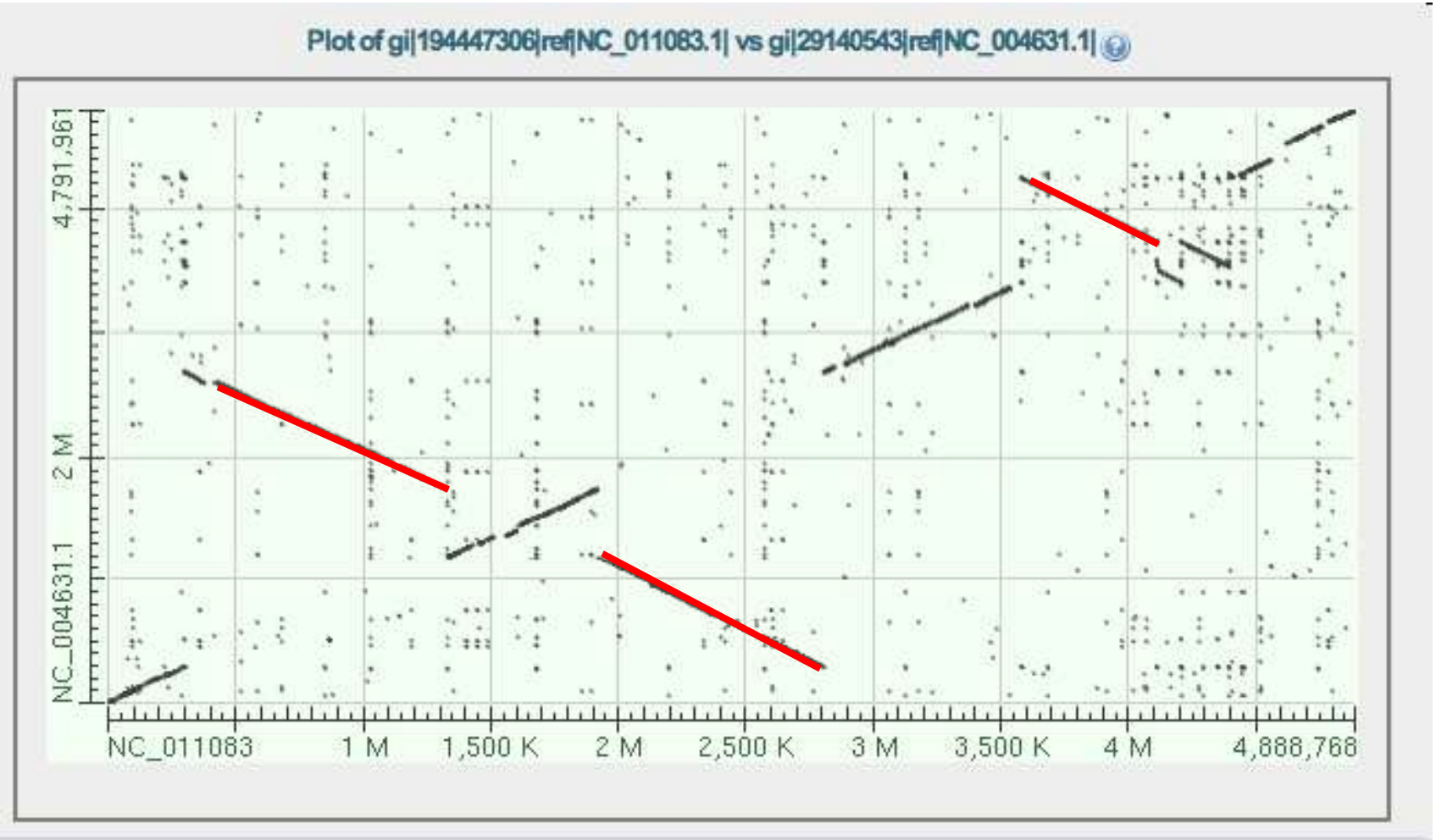
- Detection of palindromic sequences ANA and SIS



- Detection of inverted sequences



- Genome comparison revealed inversion of gene positions



Dot plot comparisons of the genome of two *S. enterica* strains

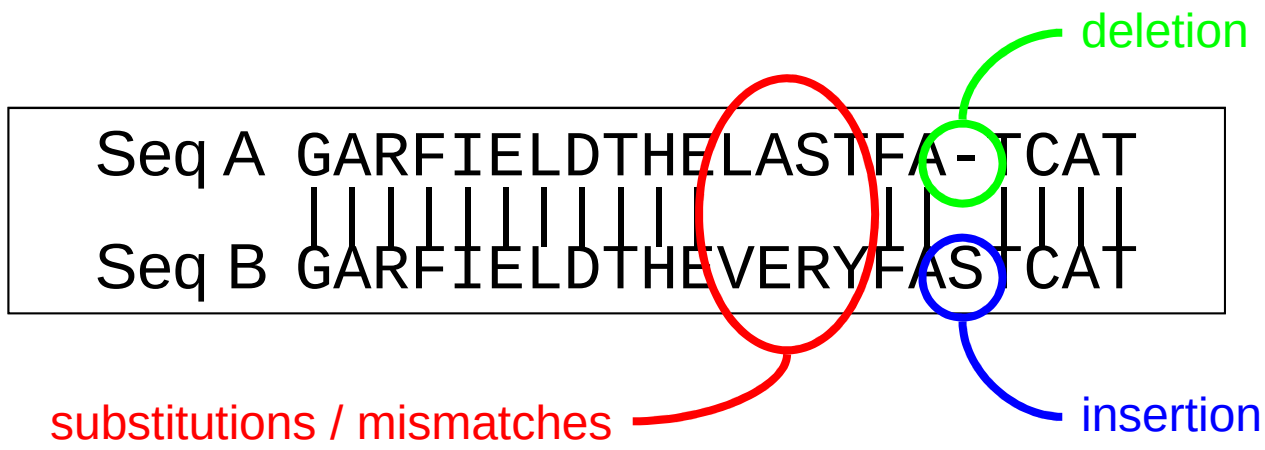
C. Dotplot limitations

- ⇨ It's a visual aid. The human eye can rapidly identify similar regions in sequences.
- ⇨ It's a good way to explore sequence organisation.
- ⇨ It does not provide an alignment.

2. Pairwise sequence alignments

A. Concept

Explicit residue matching between sequences



B. Alignment evaluation

Consider the sequence fragments below: a simple alignment show some conserved portions

```
CGATGCAGACGTCA
||||||||
CGATGCAAGACGTCA
```

but also:

```
 CGATGCAGACGTCA
        ||||||||
CGATGCAAGACGTCA
```

We need a way to evaluate the biological meaning of alignment

- Tolerant to errors: mismatches, insertion/deletions (indels)
- Evaluation of the alignment in a biological concept (significance)

B1. Scoring system

The following alignment :

```
CGAGGCACAACGTCA
|||  |||   ||||||
CGATGCAAGACGTCA
```

looks better than:

```
TTGGACAGCAATCAG
      ||   |
CGATGCAAGACGTCA
```

We can express this notion more rigorously, by using a scoring system and looking for the highest score

A simple way to score an alignment is to count 1 for each match and 0 for each mismatch (identity score)

```
CGAGGCACAACGTCA
|||  |||   ||||||
CGATGCAAGACGTCA
```

Score : 12

```
TTGGACAGCAATCAG
      ||   |
CGATGCAAGACGTCA
```

Score : 3

B2. Gaps

Insertions or deletions

- Proteins often contain regions where residues have been inserted or deleted during evolution
- There are constraints on where these insertions and deletions can happen (between structural or functional elements like: alpha helices, active site, etc.)

The alignment

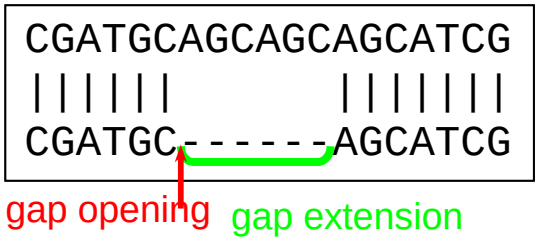
```
GCATGCATGCAACTGCAT
|||||||||
GCATGCATGGGCAACTGCAT
```

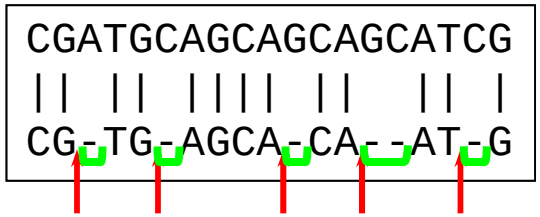
can be improved by inserting a gap

```
GCATGCATG--CAACTGCAT
|||||||||  |||||||||
GCATGCATGGGCAACTGCAT
```

Gap opening and extension penalties

- Homologous sequences may have different lengths. We need a model simulating the evolutionary mechanisms involved in gap occurrence
- Insertion tend to be several residues long rather than just a single residue long





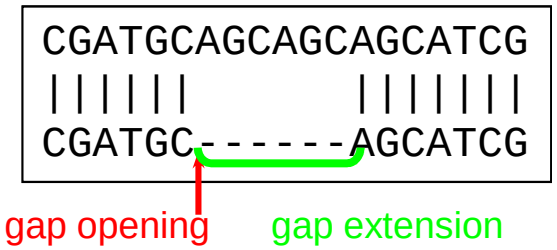
Two alignments with identical number of gaps but very different gap distribution.

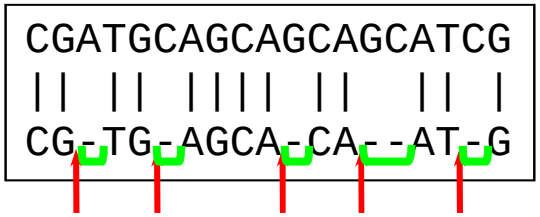
- An affine gap penalty function is preferred over a linear system
  - Smaller penalty on lengthening an existing gap (GEP) than introducing a new gap (GOP)

$W_{(k)} = a + bk$ with $b < a$ k: residu number

Example

- With a match score of 1 and a mismatch score of 0
- With an opening penalty of 10 (GOP) and extension penalty of 1 (GEP), we have the following score:





13 x 1 – (10 + 6 x 1) = -3

13 x 1 – [4 x 10 + (10+ 6 x 1)] = -43

$$S = x - (\Sigma w_k z_k)$$

S = similarity score; x = number of match; w_k, gap cost; and z_k = number of gaps with length k

Remark : depending on the program used, GOP may include the first residue

Effect of different GOP values on global alignment results


```
# Commandline: needle
#    -auto
#    -asequence uniprot:myg_human
#    -bsequence uniprot:lgb2_luplu
#    -brief
#    -outfile outfile
#    -aformat3 srspair
# Align_format: srspair
# Report_file: outfile
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: MYG_HUMAN
# 2: LGB2_LUPLU
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 173
# Identity:      39/173 (22.5%)
# Similarity:    69/173 (39.9%)
# Gaps:          38/173 (22.0%)
# Score: 67.5
#
#
#=======================================

MYG_HUMAN           1 MG-LSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKH     49
                      || |::.:..||.:.|.:..|:||.|....|.:.:..|...   |.|..
LGB2_LUPLU          1 MGALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAK---DLFSF     47

MYG_HUMAN          50 LKSEDEM-KASEDLKKHGATV----------LTALGGILKKKGHHEAEIK     88
                      ||...|: :.:.:|:.|...|          |...|.::.     :|.:|
LGB2_LUPLU         48 LKGTSEVPQNNPELQAHAGKVFKLVYEAAIQLQVTGVVVT-----DATLK     92

MYG_HUMAN          89 PLAQSHATK-----HKIPVKYLEFISECIIQVLQSKHPGDFGADAQGAMN    133
                      .|...|.:|     |...||  |.|.:.|.:|:.:|...:.:....|.:
LGB2_LUPLU         93 NLGSVHVSKGVADAHFPVVK--EAILKTIKEVVGAKWSEELNSAWTIAYD    140

MYG_HUMAN         134 KALELFRKDM--ASNYKELGFQG    154
                      :...:.:|:|  |:
LGB2 LUPLU        141 ELAIVIKKEMNDAA---------    154
```

GOP= 10

Score= 67.5

```
#    -auto
#    -asequence uniprot:myg_human
#    -bsequence uniprot:lgb2_luplu
#    -gapopen 5
#    -brief
#    -outfile outfile
#    -aformat3 srspair
# Align_format: srspair
# Report_file: outfile
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: MYG_HUMAN
# 2: LGB2_LUPLU
# Matrix: EBLOSUM62
# Gap_penalty: 5.0
# Extend_penalty: 0.5
#
# Length: 193
# Identity:      44/193 (22.8%)
# Similarity:    73/193 (37.8%)
# Gaps:          78/193 (40.4%)
# Score: 126.5
#
#
#=======================================

MYG_HUMAN           1 MG-LSDGEWQLVLNVWGKVEADIPGHGQE--VLIRLFKGHPETLE-----     42
                      || |::.:..||.:.|.:..|:||.|...  :|:        ||
LGB2_LUPLU          1 MGALTESQAALVKSSWEEFNANIPKHTHRFFILV---------LEIAPAA     41

MYG_HUMAN          43 KFDKFKHLKSEDEM-KASEDLKKH-G--------ATV-LTALGGILKKKG     81
                      | |.|..||...|: :.:.:|:.| |        |.: |...|.::.
LGB2_LUPLU         42 K-DLFSFLKGTSEVPQNNPELQAHAGKVFKLVYEAAIQLQVTGVVVT---     87

MYG_HUMAN          82 HHEAEIKPLAQSHATK-----HKIPVKYLEFISECIIQVLQSKHPGDFGA    126
                        :|.:|.|...|.:|     | .||     :.|.|::.::
LGB2_LUPLU         88 --DATLKNLGSVHVSKGVADAH-FPV-----VKEAILKTIK---------    120

MYG_HUMAN         127 DAQGA-----MNKAL-----EL---FRKDM--ASNYKELGFQG    154
                      :..||     :|.|.     ||   .:|:|  |:
LGB2_LUPLU        121 EVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA---------    154
```

GOP= 5

Score= 126.5

C. Use of scoring matrix for sequence alignment significance

- Nucleotide scoring matrix matrices are based on
  - Unitary matrix (1,0) (GCM)
  - FASTA matrix (5, -4) , NCBI BLAST (1,-2)
  - Transversion/transition matrix (1, -5, -1)

  ⇨Scoring system scales are arbitrary
- Protein scoring matrices give a score for substitution of one amino-acid by another
  - The genetic code (GCM)
  - Physico-chemical properties of amino-acids : hydrophobicity, acid / base, sterical properties, ...
  - Amino acid substitutions: PAM and BLOSUM

Hydrophobicity scoring matrix (x10)

Ala	-0.5
Cys	-1.0
Asp	2.5
Glu	2.5
Phe	-2.5
Gly	-1.5
His	-0.5
Ile	-1.8
Lys	3.0
Leu	-1.8
Met	-1.3
Asn	0.2
Pro	-1.4
Gln	0.2
Arg	3.0
Ser	0.3
Thr	-0.4
Val	-1.5
Trp	-3.4
Tyr	-2.3

Levit's scale of hydrophobicity

	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	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	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	5	5	4	4	4	3	3	3	2	1
Glu = E	9	9	10	10	8	8	7	6	6	6	5	5	5	5	5	4	4	4	3	3	3	2	1
Asx = B	8	8	8	8	10	10	8	8	8	8	7	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	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	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	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	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	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	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	8	7	7	5
His = H	5	5	5	5	7	7	9	9	9	9	10	10	10	10	9	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	9	8	8	8	7	7	5
Cys = C	4	4	5	5	6	6	8	8	8	8	9	9	9	9	10	10	9	9	9	9	8	8	5
Met = M	3	3	4	4	6	6	8	8	8	8	9	9	9	9	10	10	10	10	9	9	8	8	7
Pro = P	3	3	4	4	6	6	7	8	8	8	8	8	9	9	9	10	10	10	9	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	8	7
Leu = L	3	3	3	3	5	5	7	7	7	7	8	8	8	8	9	9	9	10	10	10	9	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	9	8
Tyr = Y	2	2	3	3	4	4	6	6	6	6	7	7	7	7	8	8	9	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	8	9	9	10	10	9
Trp = W	0	0	1	1	3	3	4	4	4	5	5	5	5	5	6	7	7	7	8	8	8	9	10

$S_{(S,R)} = 1 - (2.7)/6.4$

Probabilistic foundation for scoring matrices

- Differences in protein sequences might result from distinct mechanisms:
  - A random model
  - A non random (evolutionary) change model

- Determining occurrence probability for each model allows identification of the most likely mechanism

$$S_{ij} = \log\left(\frac{observed}{expected\,by\,chance}\right)$$

Random system

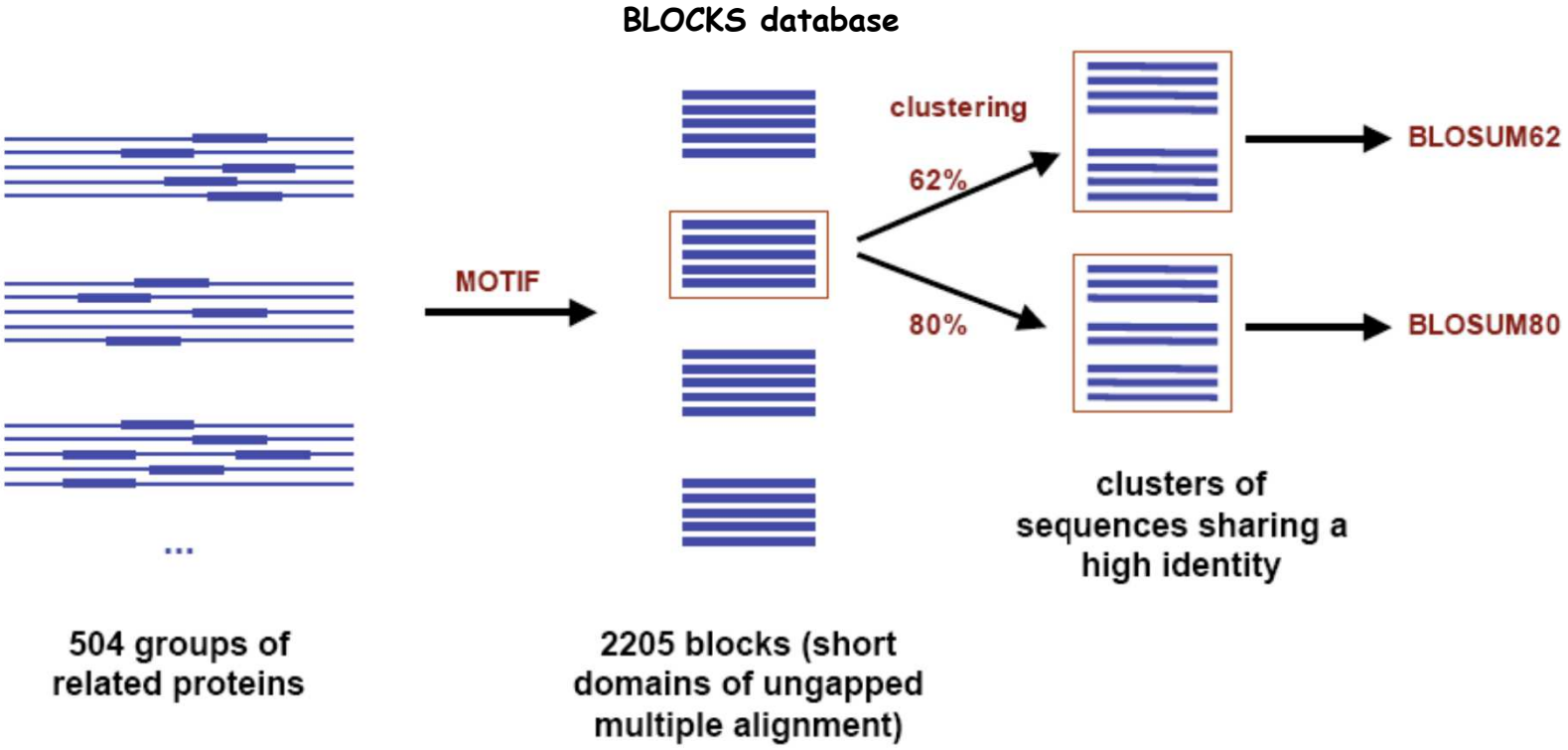
- No constraint on amino acid composition
- The nature of an amino-acid at each position is independent from other positions
  - Probability depens on amino-acid background frequency in the population , then $p_i = f_i$
- The probability of amino acid i and j matching (independent occurence) is $p_i.p_j$

Non random (evolutionary) system

- Some constraint on amino acid composition
- The probability of occurrence of a residu is determined by the residu at the same position in the ancestral sequence
- Probability of mutual substitution is $q_{i,j}$ (correlated residues and target frequency)
  - This value is dependent on the evolutionary mechanism

- Commonly-used scoring matrices are based on odds ratios, $q_{i,j}/p_i p_j$
  - If the ratio is >1, the non-random model is more likely to explain the alignment between the residues
- Two main families of substitution matrices:
  - PAM : likelihood of changes in homologous protein sequences during evolution (Dayhoff, Jones *et al*)
  - BLOSUM : substitutions in conserved blocks from protein families (Hennikoff and Hennikoff)
- Log-odds score $s_{ij} = log \frac{q_{ij}}{p_i p_j}$

1. BLOcks SUbstitution Matrix (BLOSUM, Henikoff & Henikoff 1992)



BLOSUM matrices are based on residue substitutions seen in conserved regions from the BLOCKS database

They are not based on evolutionary distances

Clustering according to identity percentage and weighting prevent a bias due to similar sequences overepresented in databases

Derivation steps for the BLOSUM scoring matrix

- Original data

```
A:  T L K K V Q K T
B:  T L K K V Q K T
C:  T L K K I Q K Q
D:  I I T K L Q K Q
E:  T I T K L Q K Q
F:  T L T K I Q K Q
G:  T L T Q I Q K Q
```

Multiple sequence alignment

¤	I¤	K¤	L¤	Q¤	T¤	V¤
I¤	8¤	·¤	16¤	··¤	6¤	6¤
K¤	¤	78¤	¤	6¤	12¤	¤
L¤	16¤	¤	22¤	¤	¤	4¤
Q¤	¤	6¤	¤	62¤	10¤	¤
T¤	6¤	12¤	¤	10¤	44¤	¤
V¤	6¤	¤	4¤	¤	¤	2¤

Residue pairs frequency matrix

- $q_{ij} = f_{ij} / \sum_{i} \sum_{j=1}^{i} f_{ij}$ with f_{ij} = frequency of the aligned *i,j* pairs
- $R_{ij} = q_{ij} / p_i\, p_j$
- $S_{(i,j)} = 2 \log_2 R_{ij}$

On the basis of overall identity, blocks are splitted in clusters having different weights

Example for the BLOSUM 50 (50 % identity)

weight	Block
1/3	ATCKQ ATCRN ASCKN
1/3	SDCDN SDCEQ SECEN
1	TECRQ

Weighted frequency matrix

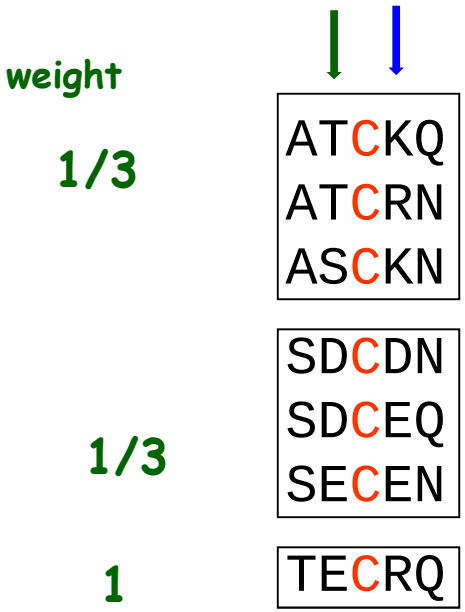
	A	C	D	E	K	N	Q	R	S	T
A										
C										
D										
E										
K										
N						4/9	16/9			
Q						16/9	7/9			
R										
S										
T										

$f_{Q,N}$ = (1/3 x 2/3)+(2/3 x 1/3)+(2/3 x 1)+(2/3 x1) = 16/9

$f_{Q,Q}$ = (1/3 x 1/3)+(1/3 x 1)+(1/3 x 1) = 7/9

$f_{N,N}$ = (2/3 x 2/3) = 4/9

Example continued



$f_{K,R}$ = (1/3 x 2) x 1

$f_{K,E}$ = (2/3 x 2/3)

$f_{E,E}$ = (1/3 x 1)

$f_{R,E}$ = (1/3 x 2/3)+(2/3 x 1)

$f_{R,R}$ = (1/3 x 1)

f_{KK}= 0

Completely filled matrix

	A	C	D	E	K	N	Q	R	S	T
A	0	0	0	0	0	0	0	0	1	1
C	0	3	0	0	0	0	0	0	0	0
D	0	0	0	2/3	2/9	0	0	4/9	2/9	4/9
E	0	0	2/3	3/9	4/9	0	0	8/9	4/9	8/9
K	0	0	2/9	4/9	0	0	0	2/3	0	0
N	0	0	0	0	0	4/9	16/9	0	0	0
Q	0	0	0	0	0	16/9	7/9	0	0	0
R	0	0	4/9	8/9	2/3	0	0	1/3	0	0
S	1	0	2/9	4/9	0	0	0	0	0	1
T	1	0	4/9	8/9	0	0	0	0	1	0

Example continued

Observed frequency of occurrence of a pair (i,j)

$$q_{ij} = f_{ij} / \sum_{1 \le i \le j} \sum f_{ij}$$

	A	C	D	E	K	N	Q	R	S	T
A	0	0	0	0	0	0	0	0	1	1
C	0	3	0	0	0	0	0	0	0	0
D	0	0	0	2/3	2/9	0	0	4/9	2/9	4/9
E	0	0	2/3	3/9	4/9	0	0	8/9	4/9	8/9
K	0	0	2/9	4/9	0	0	0	2/3	0	0
N	0	0	0	0	0	4/9	16/9	0	0	0
Q	0	0	0	0	0	16/9	7/9	0	0	0
R	0	0	4/9	8/9	2/3	0	0	1/3	0	0
S	1	0	2/9	4/9	0	0	0	0	0	1
T	1	0	4/9	8/9	0	0	0	0	1	0

	A	C	D	E	K	N	Q	R	S	T
A	0	0	0	0	0	0	0	0	0.07	0.07
C	0	0.20	0	0	0	0	0	0	0	0
D	0	0	0	0.04	0.01	0	0	0.03	0.01	0.03
E	0	0	0.04	0.02	0.03	0	0	0.06	0.03	0.06
K	0	0	0.01	0.03	0	0	0	0.04	0	0
N	0	0	0	0	0	0.03	0.12	0	0	0
Q	0	0	0	0	0	0.12	0.05	0	0	0
R	0	0	0.03	0.06	0.04	0	0	0.02	0	0
S	0.07	0	0.01	0.03	0	0	0	0	0	0.07
T	0.07	0	0.03	0.06	0	0	0	0	0.07	0

Frequency Matrix used for BLOSUM50

$q_{N,N} = 4/9 : 135/9 = 0.025$

Example continued

Determination of expected frequency (random model)

The probability of occurrence of residue Q :

$$p_Q = q_{Q,Q} + (1/2)\,\Sigma_{b \neq Q}\, q_{Q,b} \qquad (q_{Qb} = p_{q \rightarrow b} + p_{b \rightarrow Q})$$

$$p_Q = 0.052 + (0.119/2) = 0.112$$

and the probability of occurrence of residue N :

$$p_N = 0.030 + (0.119/2) = 0.090$$

The expected frequency of occurrence of (Q,Q) pairs

$$e_{QQ} = p_Q \times p_Q = 0.112 \times 0.112 = 0.013$$

and that of (Q,N) pairs

$$e_{QN} = 2p_Q \times p_N = 2 \times 0.112 \times 0.090 = 0.020$$

Example continued

Calculation of R_{ij} entry

For QQ : $q_{QQ}/e_{QQ} = 0.052/0.013 = 3.99$

For QN : $q_{QN}/e_{QN} = 0.119/0.020 = 5.93$

Calculation of S_{ij} entry

For QQ: $S_{Q,Q} = 2 \times \log_2 3.99 = 4$

For QN, $S_{Q,N} = 2 \times \log_2 5.93 = 5.1$

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

BLOSUM62

(Henikoff and Henikoff, on the basis of 2205 blocks of proteins)

$S_{(w,w)} = 11$

$R_{ww} = 2^{11/2} = 45.2$

Note: The original BLOSUM 62 matrix is quite different from the matrix that should have been calculated using the correct algorithm. The funny thing is that it works better than the revised matrix.

2. PAM scoring matrices (Dayhoff et al. 1978)

PAM = Percent (or Point) Accepted Mutation

- based on a model of evolutionary change in proteins
- unit of evolution = period of time required for 1 change /100 amino acids
- series of scoring matrices, each reflecting a certain level of divergence

PAM1 proteins with an evolutionary distance of 1% mutation per position

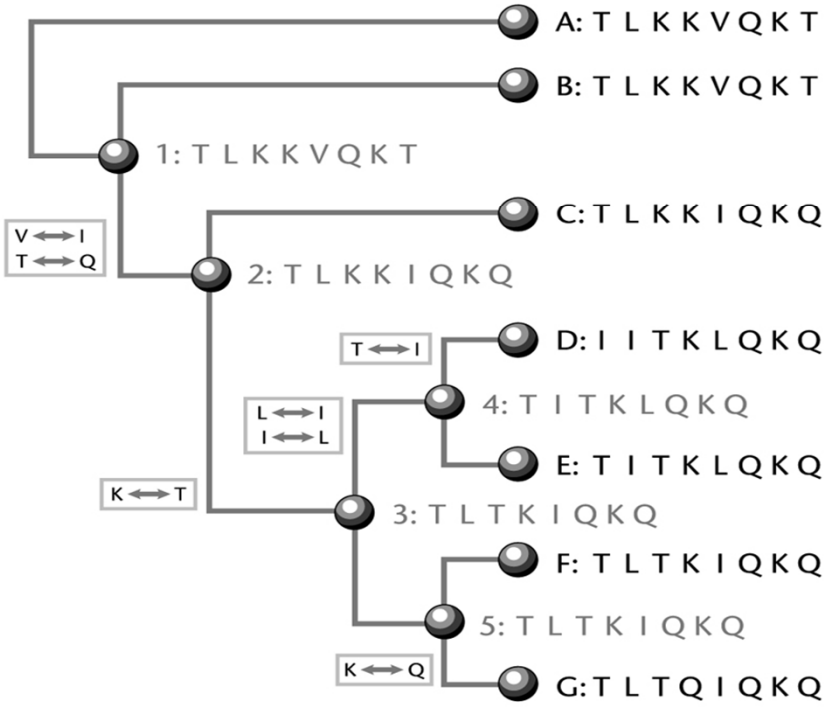
PAM50 = proteins with an evolutionary distance of 50% mutation per position

PAM250 = proteins with an evolutionary distance of 250% mutation per position (a position could mutate several times)

Derivation of PAM matrices

- Analysis of point or percentage accepted mutations in homologous sequences during evolution

A:	T L K K V Q K T
B:	T L K K V Q K T
C:	T L K K I Q K Q
D:	I I T K L Q K Q
E:	T I T K L Q K Q
F:	T L T K I Q K Q
G:	T L T Q I Q K Q

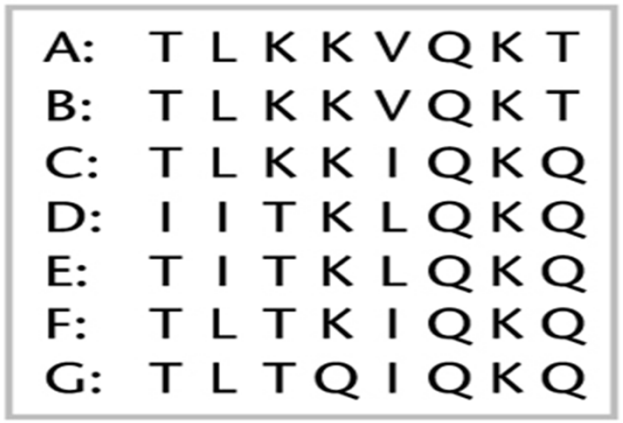


Sequence alignment

- Based on 1572 changes from 71 families, with identity > 85%
- Phylogenetic tree using parsimony

¤	I¤	K¤	L¤	Q¤	T¤	V¤	C
I¤	¤	¤	2¤	¤	1¤	1¤	C
K¤	¤	¤	¤	1¤	1¤	¤	C
L¤	2¤	¤	¤	¤	¤	¤	C
Q¤	¤	1¤	¤	¤	1¤	¤	C
T¤	1¤	1¤	¤	1¤	¤	¤	C
V¤	1¤	¤	¤	¤	¤	¤	C

·¶



PAM

¤	I¤	K¤	L¤	Q¤	T¤	V¤
I¤	¤	¤	2¤	¤	1¤	1¤
K¤	¤	¤	¤	1¤	1¤	¤
L¤	2¤	¤	¤	¤	¤	¤
Q¤	¤	1¤	¤	¤	1¤	¤
T¤	1¤	1¤	¤	1¤	¤	¤
V¤	1¤	¤	¤	¤	¤	¤

BLOSUM

¤	I¤	K¤	L¤	Q¤	T¤	V¤
I¤	8¤	·¤	16¤	··¤	6¤	6¤
K¤	¤	78¤	¤	6¤	12¤	¤
L¤	16¤	¤	22¤	¤	¤	4¤
Q¤	¤	6¤	¤	62¤	10¤	¤
T¤	6¤	12¤	¤	10¤	44¤	¤
V¤	6¤	¤	4¤	¤	¤	2¤

Difference between the BLOSUM and PAM models for counting residue substitutions

Steps followed to derive the PAM scoring matrix

Remember: we need log odds score $S_{ij} = log \frac{q_{ij}}{p_i p_j}$

- Compute relative mutability, m_j

$$m_j = \frac{Number\ of\ changes\ of\ j}{Exposure\ of\ j\ to\ mutation}$$

- Compute mutabilition probability

$$M_{ij} = \lambda m_j A_{ij} / \Sigma_i A_{ij}$$

- Compute log relatedness odds

$$R_{ij} = M_{ij} / f_i$$

Shown to be identical to $q_{ij}/p_i p_j$

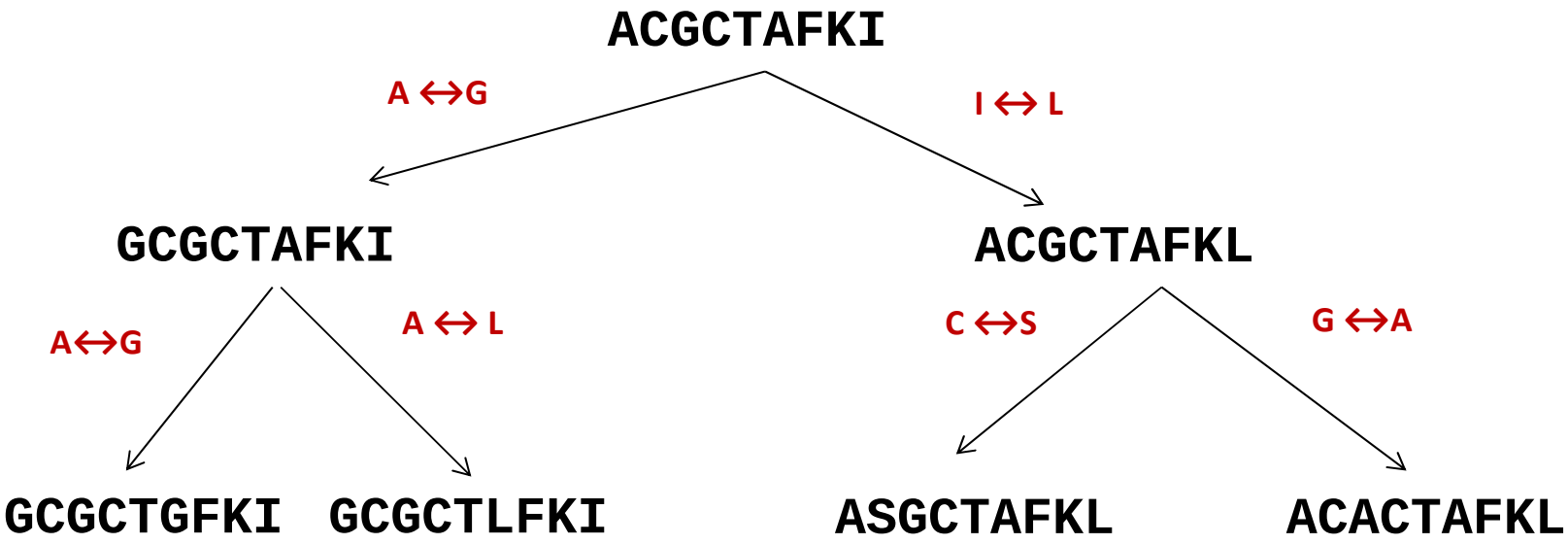
$$S_{ij} = log\ R_{ij}$$

Example illustrating the PAM model of protein evolution

- Consider the following alignment

```
ACACTAFKL   ACGCTAFKI
GCGCTGFKI   GCGCTAFKI
GCGCTLFKI   ACGCTAFKL
ASGCTAFKL
```

- Phylogenetic tree to determine which substitution occured during sequence evolution (maximum parcimony method)



Matrix of accepted point mutation count *A*

i \ *j*	A	C	F	G	I	K	L	T	S
A				3			1		
C									1
F									
G	3								
I							1		
K									
L	1				1				
T									
S		1							

A_{total} 4 1 0 3 1 0 2 0 1 = 12

For each pair of different residues (*i*,*j*), the total number A_{ij} of substitutions from *j* to *i* along the edges of the tree is calculated

A total of 12 changes is observed

- Relative mutability m_j

m_j is the number of times that amino acid *j* divided by the total number of mutations that could have affected the residue

There are a total of 12 changes , 4 of which affect the A residue. On the basis of its frequency (10/63), 1.9 was expected $\Rightarrow$ mA = 4/1.9 = 2.1

residue	A	C	F	G	I	K	L	S	T
changes	4	1	0	3	1	0	2	1	0
expected	1.90	2.48	1.33	1.90	0.76	1.33	0.76	0.19	1.33
m_j	2.11	0.40	0	1.58	1.31	0	2.63	5.25	0

Original Dayhoff's data

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

Note that the value of m_{Ala} has been set arbitrarily to 100 and the values of all other amino acids scaled accordingly.

• Mutational probability matrix (M)

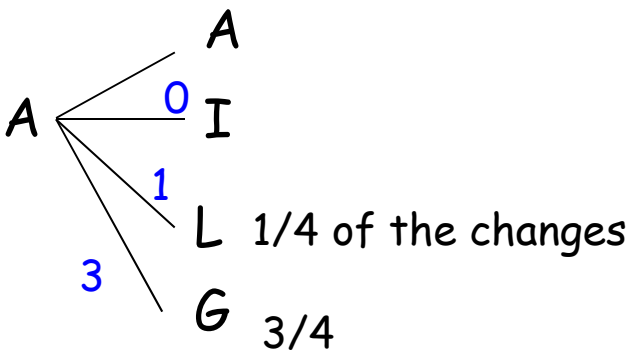
Let's define M_{ij} the probability of the amino acid in column *j* having been substituted by an amino acid in row *i* over a given evolutionary time unit.

$$M_{ij} = \lambda m_j A_{ij} / \Sigma_i A_{ij}$$

$$M_{jj} = 1-\lambda m_j$$

The constant λ represents a degree of freedom that could be used to connect the matrix M with an evolutionary time scale (number of mutations per 100 residues)

In our example



If A is mutated, the probability that it is mutated in G is $\frac{3}{4}$. Thus the probability that A is mutated in G is:

$$M_{GA} = \frac{3}{4} \times 0.021 = 0.01575 \quad (\lambda = 0.01)$$

The probability that A is mutated in L is:

$$M_{GA} = 1/4 \times 0.021 = 0.00525$$

and the probability that A is not mutated is

$$M_{AA} = 1 - 0.01575 - 0.00525 = 0.079$$

Mutational probability matrix

	A	C	F	G	I	K	L	S	T
A	0.9790			0.0157			0.0131		
C		0.9960						0.0525	
F			1						
G	0.0157			0.9843					
I					0.9869		0.0131		
K						1			
L	0.0053				0.0131		0.9738		
S		0.0040						0.9475	
T									1
	1	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**

A scaling factor makes the substitution probabilities over 1 PAM evolutionary time

$$M_{ij} = \lambda m_j A_{ij} / \Sigma_i A_{ij}$$

- ✓ Substitution probabilities are ajusted to account for the evolutionary distance between sequences
- ✓ λ was calculated to ensure that 1 substitution would occur on average per hundred residues (unit of time corresponding to 1 substitution).
- ✓ The expected number of amino acids remaining unchanged in a 100-residue protein is given by:

$$100 \sum f_j M_{jj}$$

$$100 \sum f_j (1-\lambda m_j)$$

- ✓ If only one substitution per residue is allowed, then $\lambda = 1/(100 \sum f_j m_j)$

In our example, $\sum f_j m_j = 1$ (2.11 x 0.158 + 0.4 x 0.082 + ... + 5.25 x 0.01587)

- The relatedness odds R_{ij} matrix

$$R_{ij} = M_{ij}/f_i$$

In our example, the relative frequencies of exposure to mutation (also called the effective frequencies) f_j are proportional to the average composition N_j/N multiplied by the number of mutations in the tree.

R_{SC} = 0.004/0.01587 = 0.252

R_{CS} = 0.052/0.206 = 0.252

Odd-score matrix R

	A	C	F	G	I	K	L	S	T
A	6.167			0.0957			0.0829		
C		4.827						0.252	
F			9						
G	0.099			6.202					
I					15.542		0.206		
K						9			
L	0.083				0.206		15.34		
S		0.252						59.7	
T									9

- PAM-1 matrix = 10x log odd-score matrix S

	A	C	F	G	I	K	L	S	T
A	7.9			-10.2			-10.8		
C		6.8						-5.99	
F			9.54						
G	-10.04			7.9					
I					11.9		-6.86		
K						9.54			
L	-10.8				-6.86		11.8		
S		-5.99						17.7	
T									9.54

• Dayhoff's matrix of replacement A_{ij} (x 10)

	ala	arg	asn	asp	cys	gln	glu	gly	his	ile	leu	lys	met	phe	pro	ser	thr	trp	tyr	val
A																				
R	30																			
N	109	17																		
D	154	0	532																	
C	33	10	0	0																
Q	93	120	50	76	0															
E	266	0	94	831	0	422														
G	579	10	156	162	10	30	112													
H	21	103	226	43	10	243	23	10												
I	66	30	36	13	17	8	35	0	3											
L	95	17	37	0	0	75	15	17	40	253										
K	57	477	322	85	0	147	104	60	23	43	39									
M	29	17	0	0	0	20	7	7	0	57	207	90								
F	20	7	7	0	0	0	0	17	20	90	167	0	17							
P	345	67	27	10	10	93	40	49	50	7	43	43	4	7						
S	772	137	432	98	117	47	86	450	26	20	32	168	20	40	269					
T	590	20	169	57	10	37	31	50	14	129	52	200	28	10	73	696				
W	0	27	3	0	0	0	0	0	3	0	13	0	0	10	0	17	0			
Y	20	3	36	0	30	0	10	0	40	13	23	10	0	260	0	22	23	6		
V	365	20	13	17	33	27	37	97	30	661	303	17	77	10	50	43	186	0	17	

Based on 1572 changes from 71 families, with identity > 85%

18% of possible substitutions not seen

$A_{DE} = 83,1$

• Mutation probability matrix at 1 PAM (x 10,000)

$$M_{ij} = \lambda\, m_j A_{ij} / \Sigma_i A_{ij}$$

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

$$\Sigma\ f_j \times M_{jj} = 0.99$$

PAM250 = PAM1^{250}

This matrix gives the probability of a substitution for a period of time allowing 2,5 changes per residue

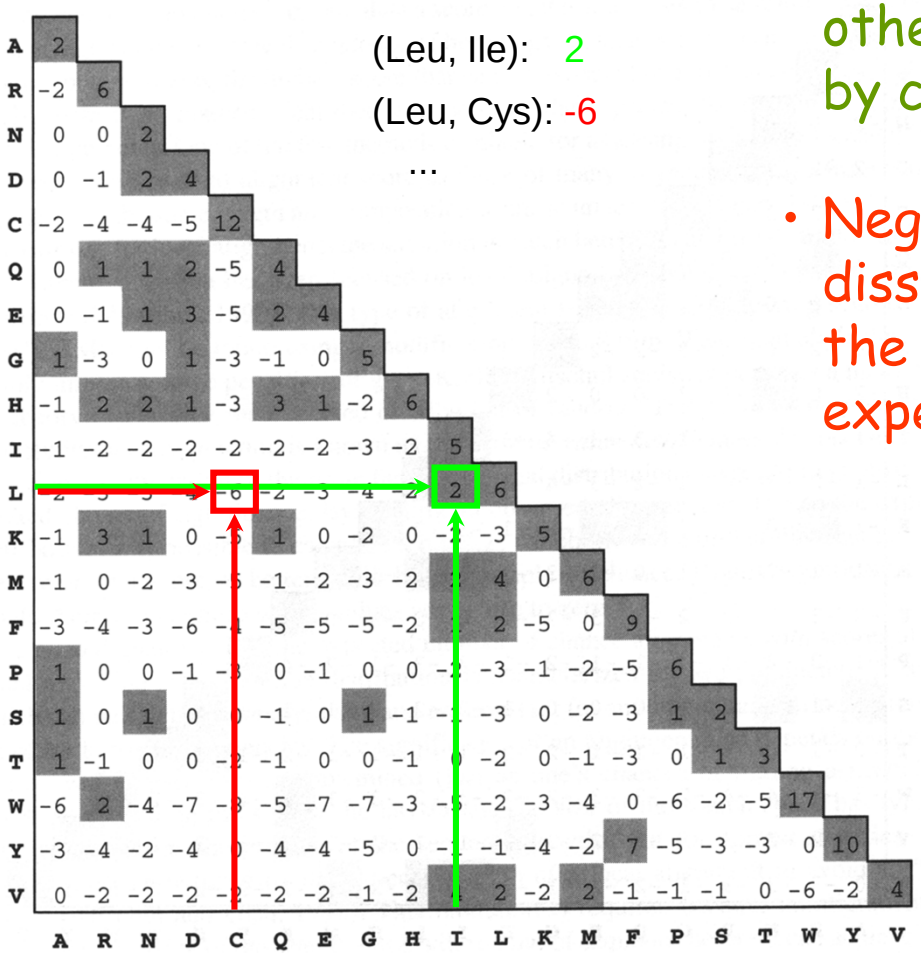
PAM250 Mutation Matrix

250 PAM evolutionary distance X100

ORIGINAL AMINO ACID

	Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V
Ala A	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
Arg R	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
Asn N	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
Asp D	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
Cys C	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
Gln Q	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3
Glu E	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
Gly G	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
His H	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
Ile I	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
Leu L	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
Lys K	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
Met M	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
Phe F	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
Pro P	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
Ser S	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
Thr T	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
Trp W	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
Tyr Y	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
Val V	7	4	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	72	4	17

- Log odds form of PAM 250 (X10)

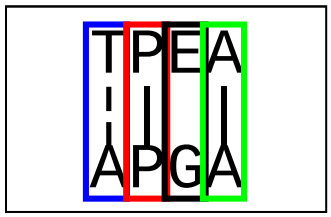


- Positive score: the amino acids are similar, mutations from one into the other occur more often then expected by chance during evolution
- Negative score: the amino acids are dissimilar, the mutation from one into the other occurs less often then expected by chance during evolution

The (Leu, Ile) match is 1,6 times more likely to occur among homologous sequences than by chance

Calculation of a raw alignment score

Raw score of an alignment



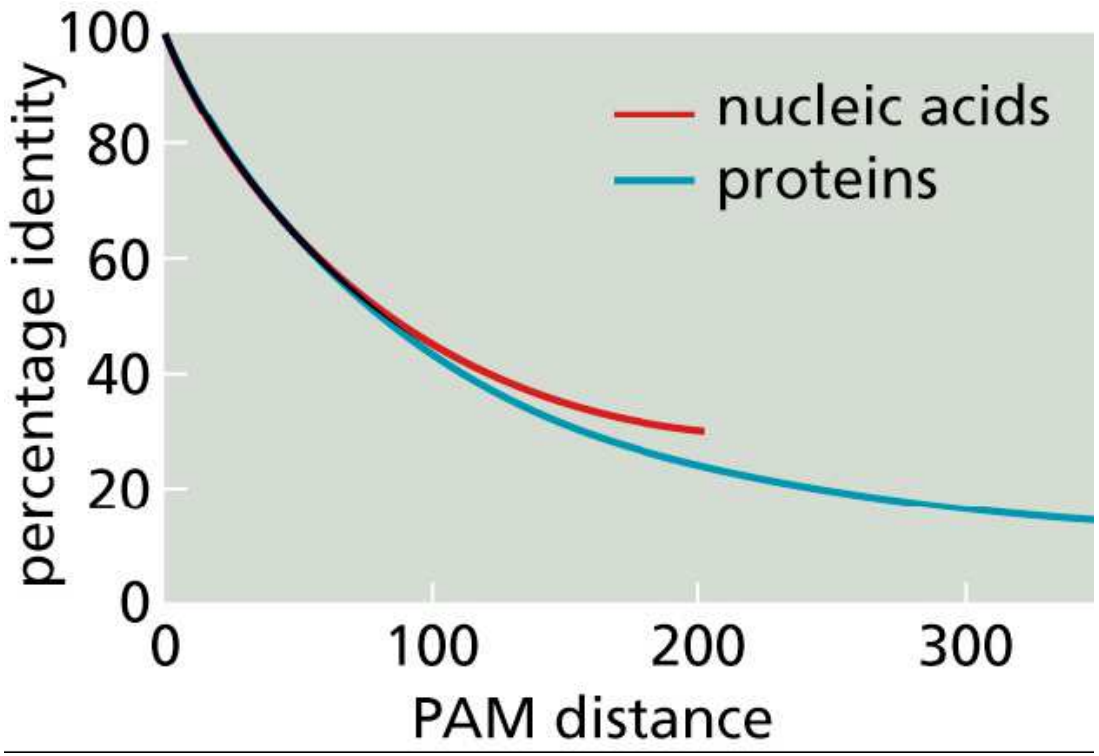
Score = 1 + 6 + 0 + 2 = 9

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

PAM250 matrix

Which PAM matrix to use?

If aligned sequences share 20% identity, the best matrix to use is PAM250



Observed difference (%)	Evolutionary distance (PAMs)
1	1
10	11
20	23
30	39
40	58
50	83
60	117
70	170
80	260

PAM	400	350	280	240	200	180	160	130	120
BLOSUM	30	35	40	45	50	55	62	75	80

On the basis of their information content

Concluding remarks

- Substitution matrices and gap penalties introduce biological information into the alignment algorithms.

- It is not because two sequences can be aligned that they share a common biological history. The relevance of the alignment must be assessed with a statistical score.

- There are many ways to align two sequences.
  Do not blindly trust your alignment to be the only truth.
  Especially gapped regions may be quite variable.