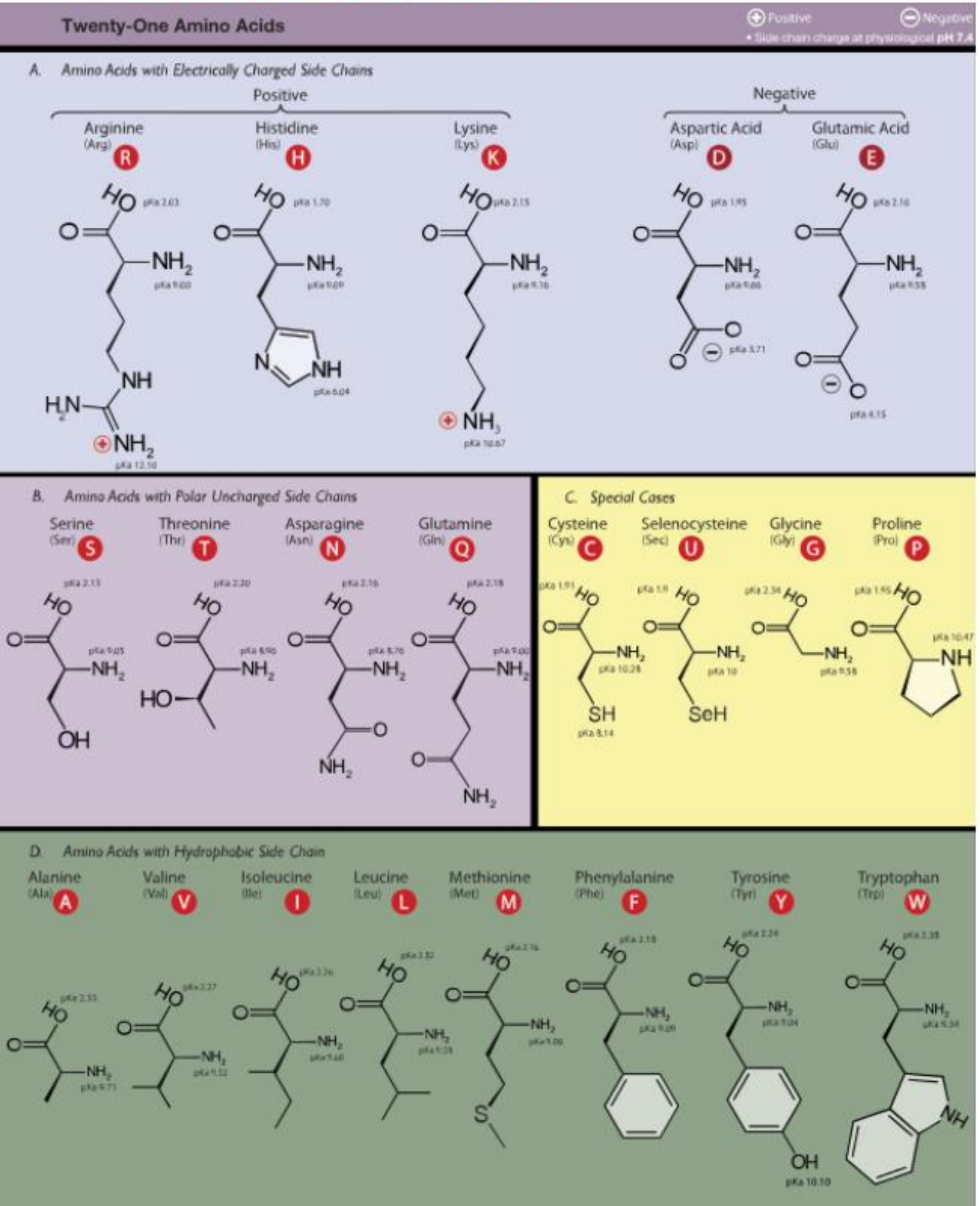
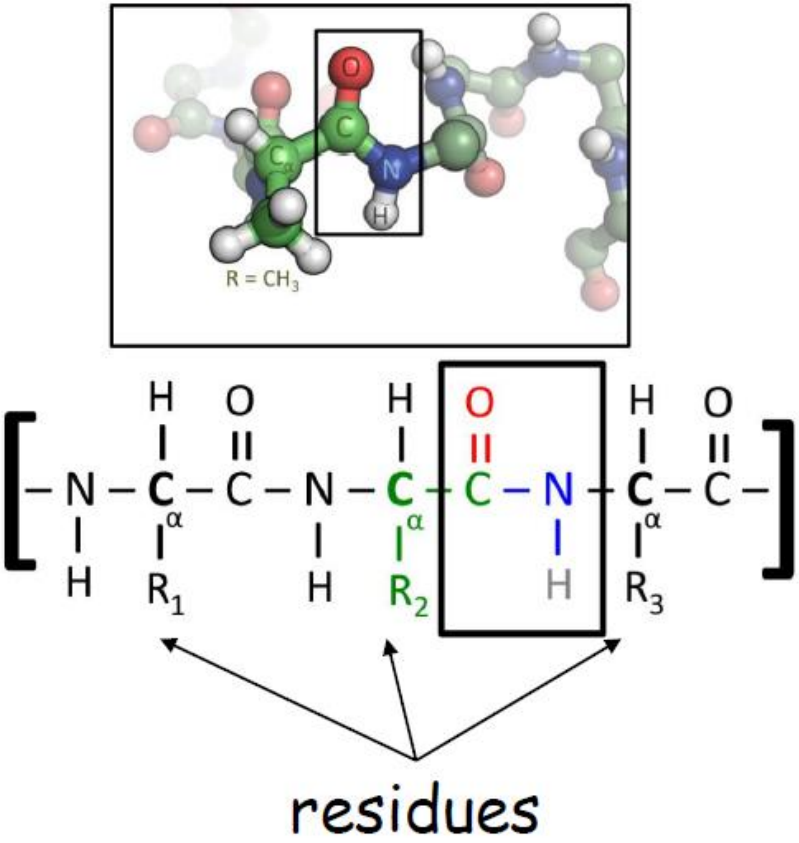
Protein comparison

Aminoacids

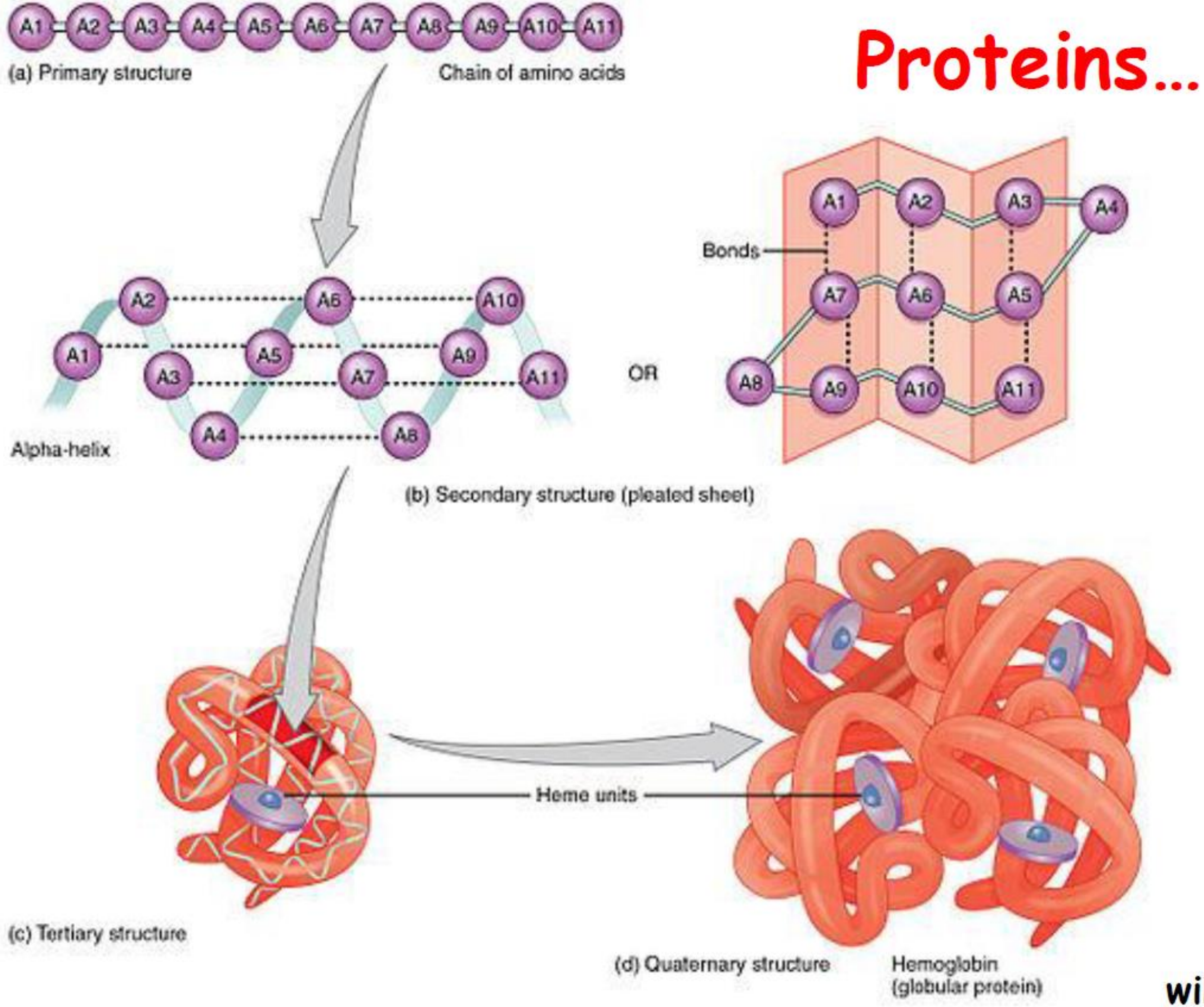


Pepdidic bond



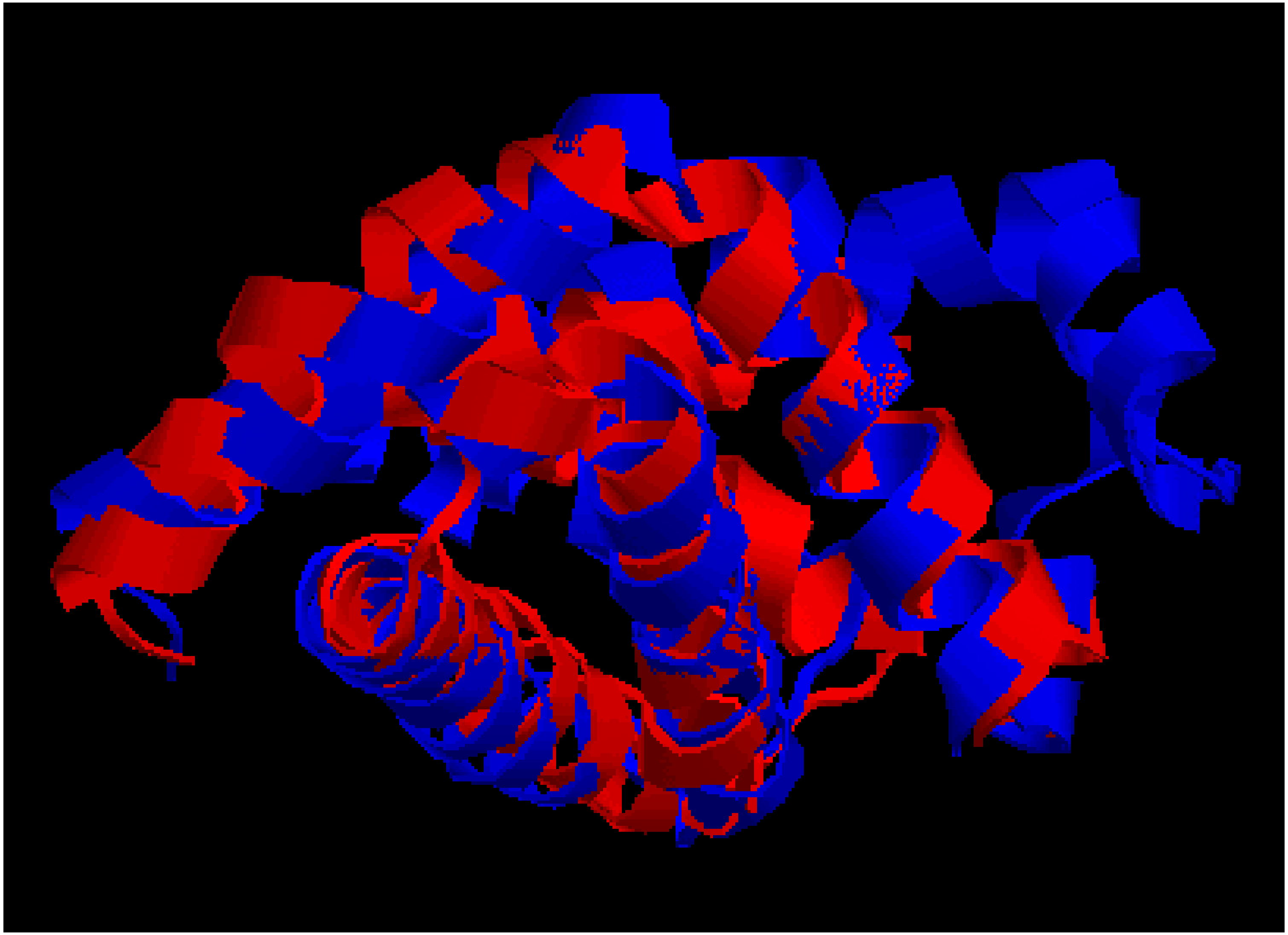
wikipedia

Proteins…



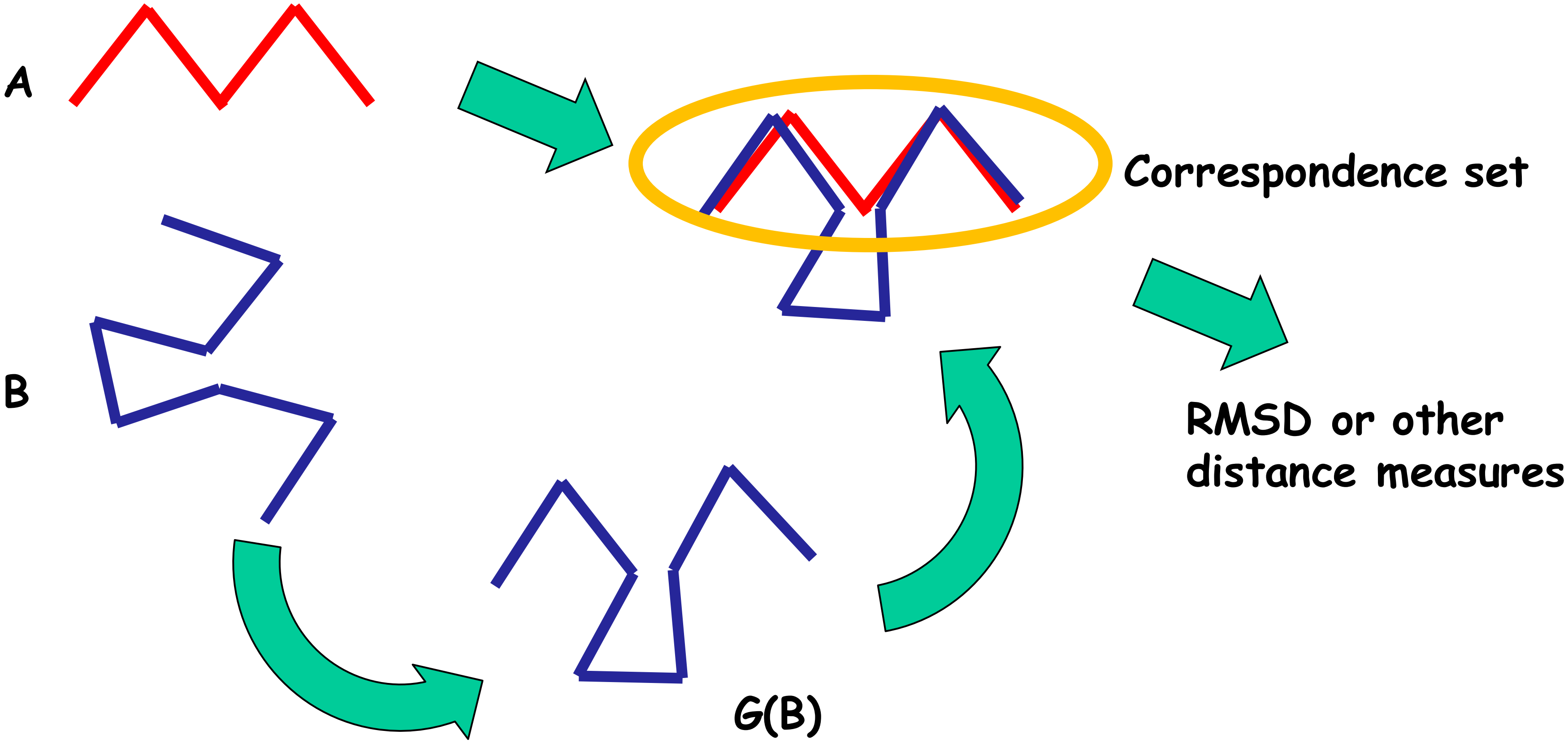
wikipedia

Structure superimposition



Sperm Whale Mioglobin vs bacterial Emoglobin

Structure superimposition problem



Formalizing the structure superimposition problem

Given two sets of points $A = (a_1, a_2, \ldots, a_n)$ and $B = (b_1, b_2, \ldots b_m)$ in Cartesian space, find the **optimal** subsets $A(P)$ and $B(Q)$ with $|A(P)| = |B(Q)|$, and find the **optimal** rigid body transformation G between the two subsets $A(P)$ and $B(Q)$ that minimizes a given distance metric D over all possible rigid body transformation G, i.e.

$$\boxed{\min_{G}\{D[A(P) - G(B(Q))]\}}$$

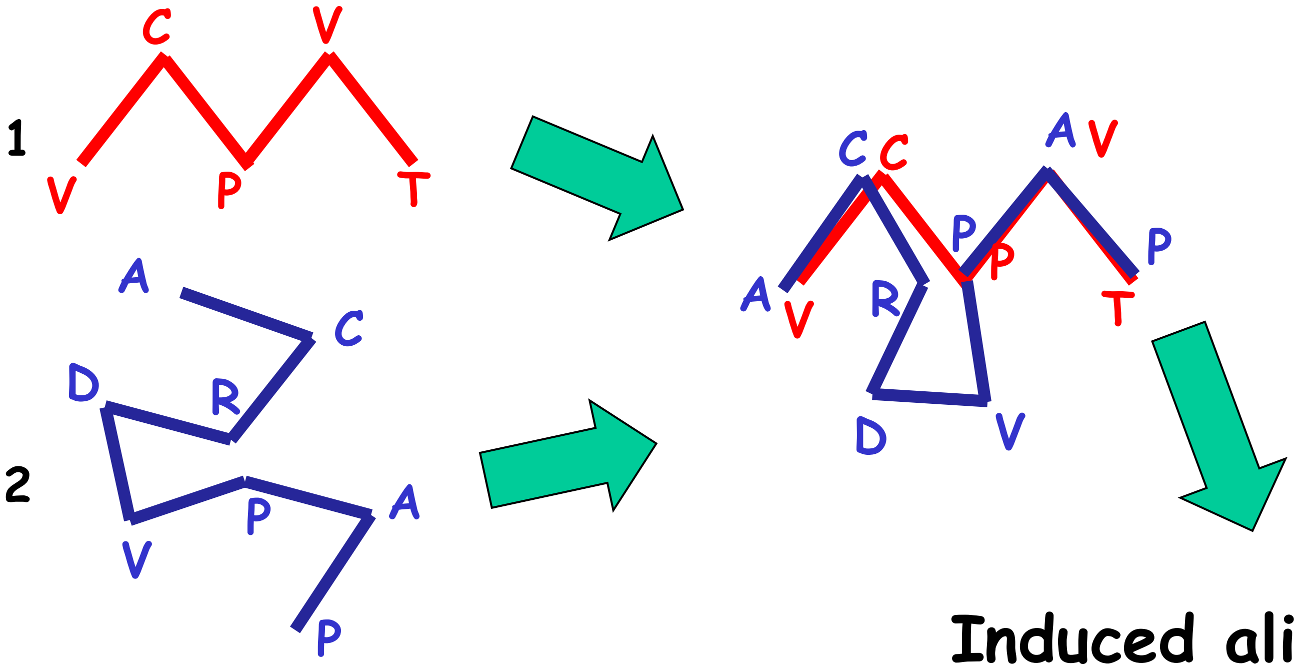
where, usually, D is $\mathbf{RMSD} = \sqrt{\dfrac{\overset{n}{\underset{i=1}{\text{å}}} (a_i - b_i)^2}{n}}$

The two subsets $A(P)$ and $B(Q)$ define a "**correspondence**", and $p = |A(P)| = |B(Q)|$ is called the correspondence length. Naturally, the correspondence length is maximal when $A(P)$ and $B(Q)$ are similar.

Therefore there are essentially two problems in structure alignment:
(*i.*) Find the correspondence set (which is *NP-hard*), and
(*ii.*) Find the alignment transform (which is $O(n)$).

Phil Bourne 2012

Structural superimposition induces a structural alignment beween the sequences

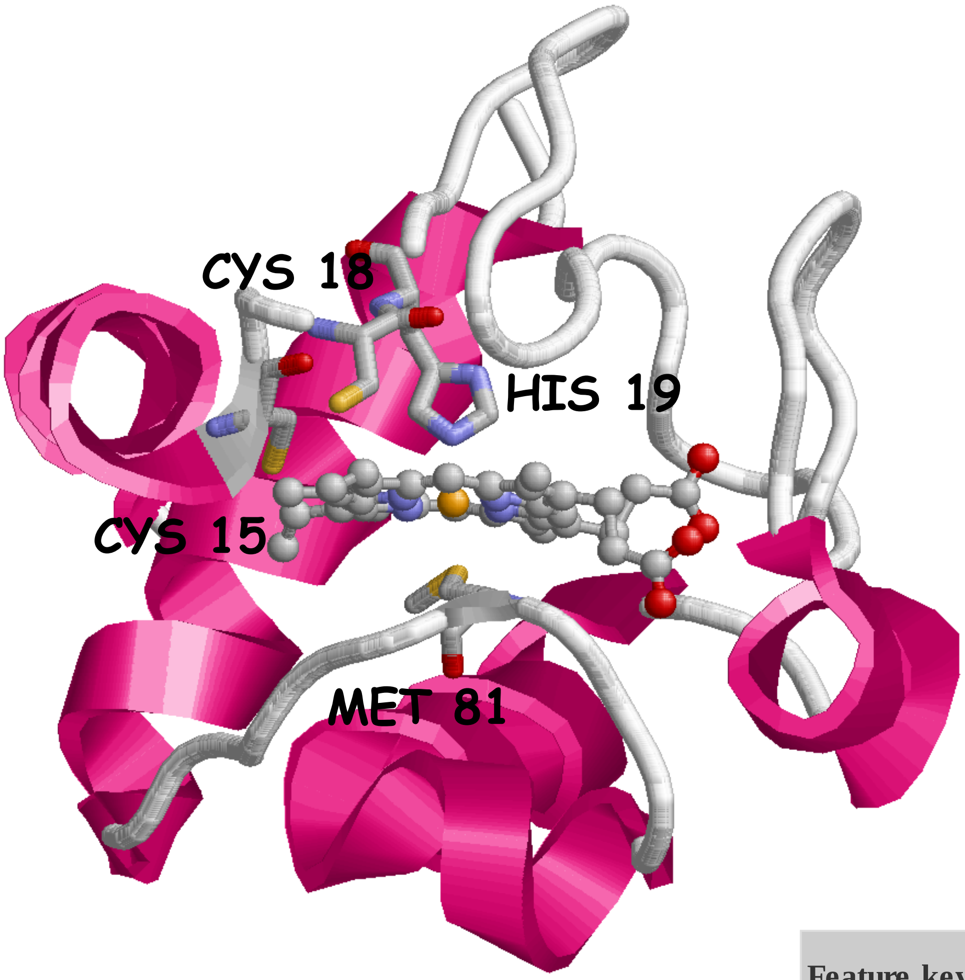


Induced alignment

```
VC---PVT
ACRDVPAP
.*...*..
```

Alignment lenght=8 res
Alignment identity = 2/8

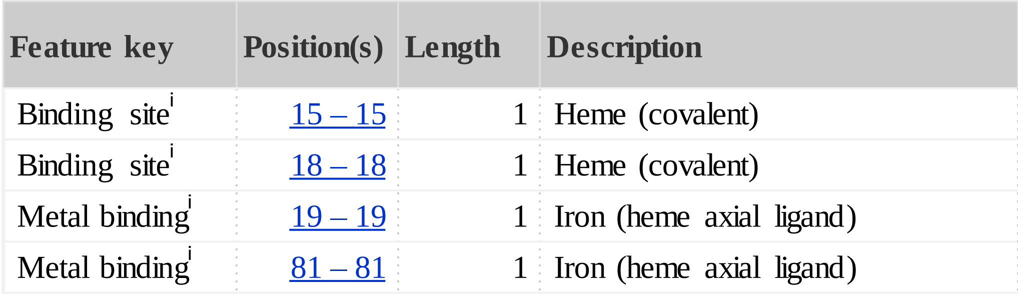
Sequence-to-structure relation: Cytochrome C



Electron carrier protein. The oxidized form of the cytochrome c heme group can accept an electron from the heme group of the cytochrome c1 subunit of cytochrome reductase. Cytochrome c then transfers this electron to the cytochrome oxidase complex, the final protein carrier in the mitochondrial electron-transport chain.

UniProt: P99999

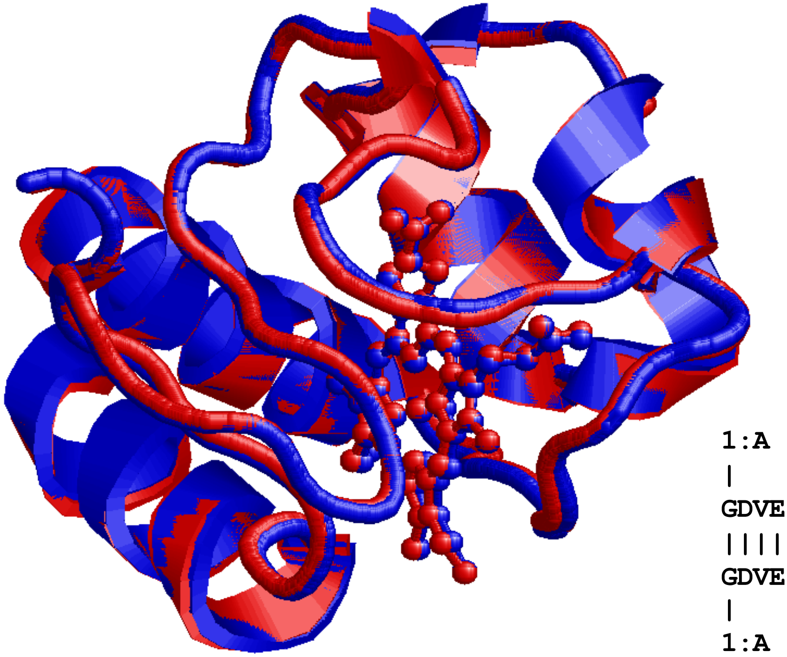
PDB: 3zcf:A



Feature key	Position(s)	Length	Description
Binding site	15 – 15	1	Heme (covalent)
Binding site	18 – 18	1	Heme (covalent)
Metal binding	19 – 19	1	Iron (heme axial ligand)
Metal binding	81 – 81	1	Iron (heme axial ligand)

Cytochrome C (*Homo* vs. *Horse*)

Human Cytochrome C - Uniprot:P99999. PDB: 3ZCF:A
Equine Cytochrome C – Uniprot: P00004. PDB 3O20:A



Structural alignment:
RMSD= 0,035 nm

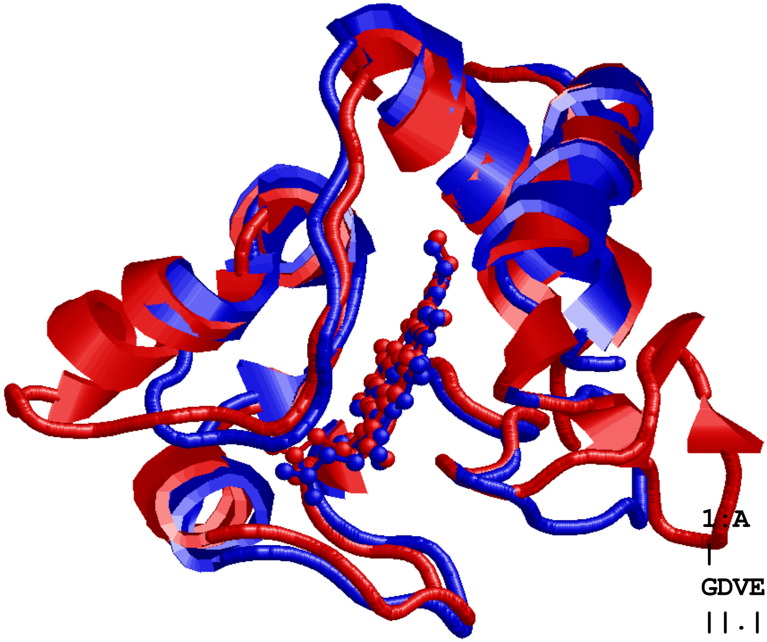
```
1:A                20:A                40:A                60:A
|        |    .    |    .    |    .    |    .    |    .    |    .    |
GDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNKGIIWGEDTLMEYLEN
||||||||||:.||.||||||||||||||||||||||||||||||:.||.|||||||.|.|:||||||||
GDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGFTYTDANKNKGITWKEETLMEYLEN
|        |    .    |    .    |    .    |    .    |    .    |    .    |
1:A                20:A                40:A                60:A
```

88% sequence identity

```
         80:A                100:A
    .    |    .    |    .    |
PKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE
||||||||||||.|||||.||.||||||||||||
PKKYIPGTKMIFAGIKKKTEREDLIAYLKKATNE
    .    |    .    |    .    |
         80:A                100:A
```

Cytochrome C (*Homo* vs. *Rhodobacter sphaeroides*)

Human Cytochrome C - Uniprot:P99999. PDB: 3ZCF:A
Cytochrome C2 *Rhodobacter Sph.* – Uniprot: P0C0X8. PDB 1CXC:A



Structural alignment:
RMSD= 0,18 nm

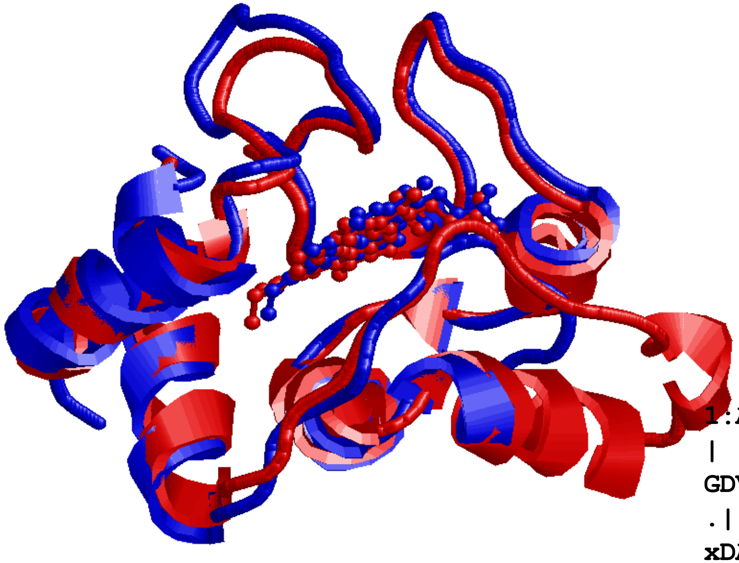
```
1:A                20:A                       40:A
|        |    .    |           .    |    .    |    .     |    .
GDVEKGKKIFIMKCSQCHTVEKGG-------KHKTGPNLHGLFGRKTGQAPGYS-YTAANKNKG---IIW
||.|.|.|.|. .|..||.:....       ..||||||:|..||..|....:. |....|..|   :.|
GDPEAGAKAFN-QCQTCHVIVDDSGTTIAGRNAKTGPNLYGVVGRTAGTQADFKGYGEGMKEAGAKGLAW
|      |     .    |    .    |    .    |    .    |    .    |    .    |
3:A               20:A                40:A                60:A
```

28% sequence identity

```
60:A                        80:A                100:A
|    .    |    .            |    .    |    .    |
GEDTLMEYLENPKKYI--------PGTKMIFVGIKKKEERADLIAYLKKATNE
.|:...:|.:.|.|::        ...||.| .:||..:...:.|||......
DEEHFVQYVQDPTKFLKEYTGDAKAKGKMTF-KLKKEADAHNIWAYLQQVAVR
   .    |    .    |    .    |     .    |    .    |
        80:A                100:A               120:A
```

Cytochrome C (*Homo* vs. *Rhodopseudomonas palustris*)

Human Cytochrome C - Uniprot:P99999. PDB: 3ZCF:A
Cytochrome C2 *Rhodopseudomons pal.* – Uniprot: P00091. PDB 1I8O:A



Structural alignment:

RMSD= 0,13 nm

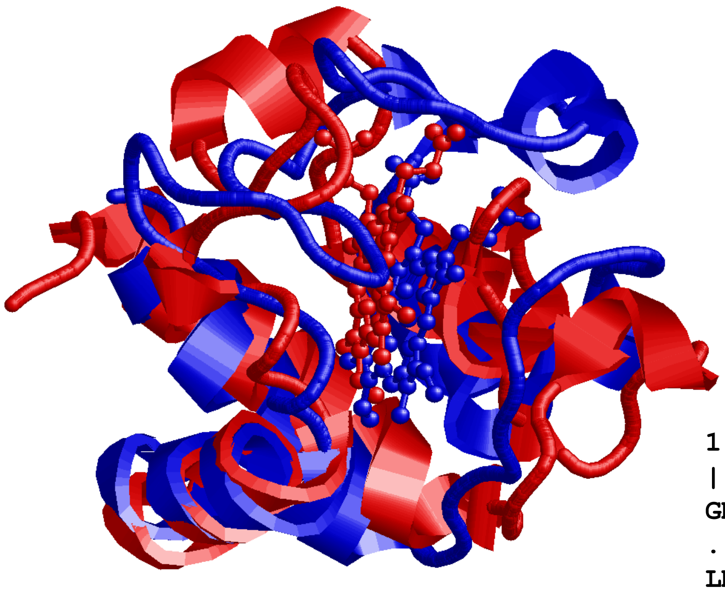
```
1:A                20:A                40:A                   60:A
|        |    .    |    .    |    .    |    .    |    .       |    .
GDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNKG---IIWGEDTLMEY
.|...|..:|. .|..||..   .|...||.|.|..|||.|.|.|:.|...|.|.|   ::|..|.:..|
xDAKAGEAVFK-QCMTCHRA---DKNMVGPALAGVVGRKAGTAAGFTYSPLNHNSGEAGLVWTADNIVPY
|        |     .       |    .    |    .    |    .    |    .    |    .
1:A                    20:A                40:A                60:A
```

29% sequence identity

```
                          80:A                100:A
  |    .                  |    .    |    .    |
LENPKKYIP--------------GTKMIFVGIKKKEERADLIAYLKKAT
|..|..::.              .|||.| .:...::|.|.:|||....
LADPNAFLKKFLTEKGKADQAVGVTKMTF-KLANEQQRKDVVAYLATLK
   |    .    |    .    |    .     |    .    |
             80:A                 100:A
```

Cytochrome C (*Homo vs. Arabidopsis thaliana*)

Human Cytochrome C - Uniprot:P99999. PDB: 3ZCF:A
Cytochrome C6A *Arabidopsis Thaliana* – Uniprot: Q93VA3. PDB 2CE0:A



Structural alignment:

RMSD= 0,35 nm

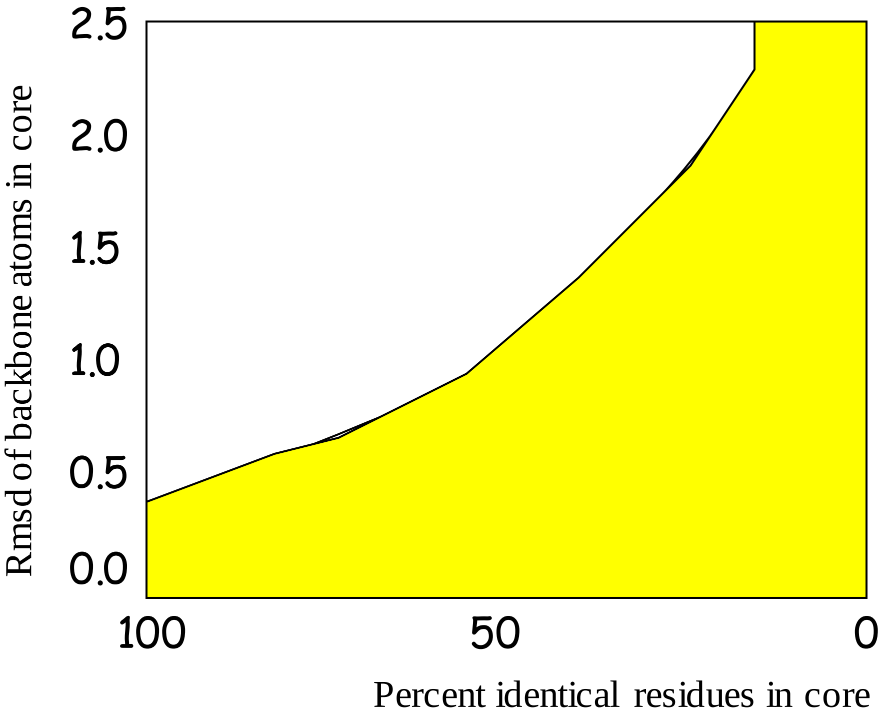
```
1:A                20:A                    40:A                60:A
|        |    .    |    .    |        .    |    .    |    .    |    .
GDVEKGKKIFIMKCSQCHTVEKGGKHKTGP---NLHG-LFGRKTGQAPGYSYTAANKNKGIIWGEDTLME
.|:::|..:|...|..||...  |...  .   .|.. ...|.                 .:..|:.:..
LDIQRGATLFNRACAACHDTG--GNII--QPGATLFTKDLERN-----------------GVDTEEEIYR
|      |    .    |      .      |    .    |                     .    |
3:A              20:A                    40:A
```

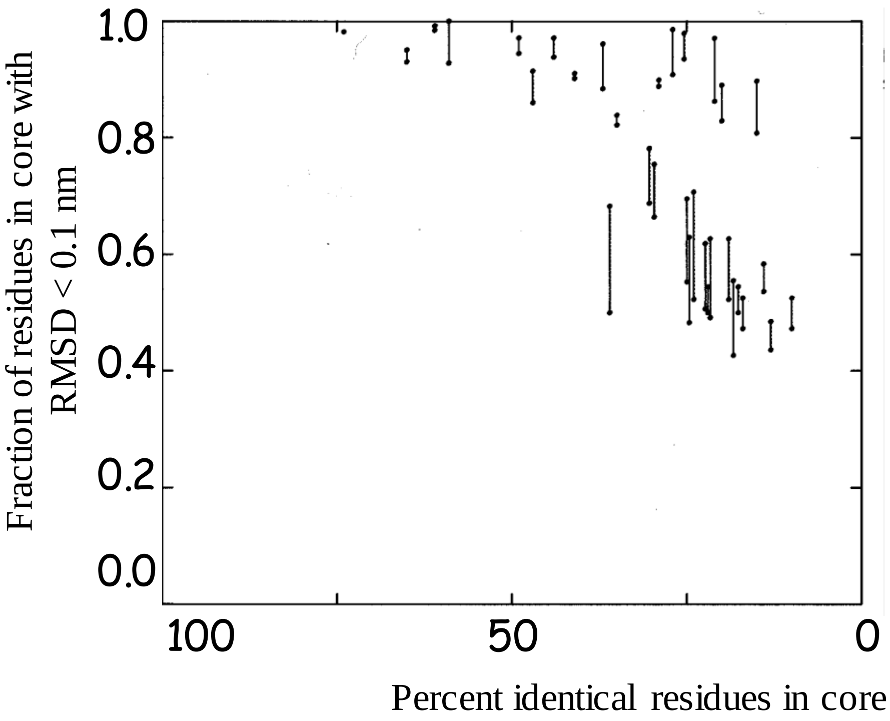
13% sequence identity

```
                      80:A                100:A
            |    .    |    .    |    .    |
YLE---------NPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE
...         ..|....|.......:... |...|..:.|....:
VTYFGKGRMPGFGEKCTPRGQCTFGPRLQDE-EIKLLAEFVKFQADQ
   .    |    .    |    .    |     .    |    .
        60:A                80:A
```

Sequence-to-structure relation

By structurally aligning a large set of strucures:



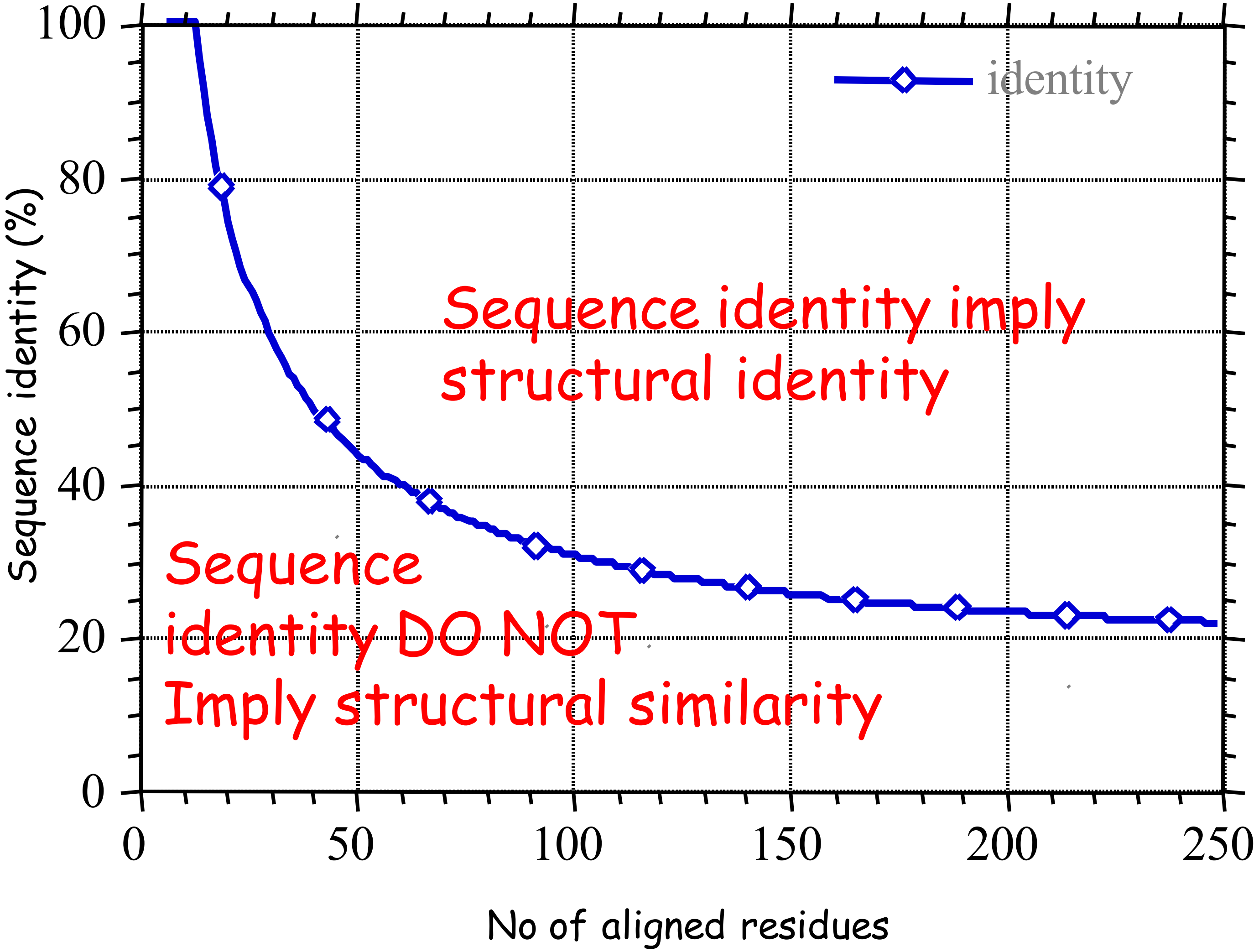


- Proteins more than 60% identical have more than 90% of residues that result less than 0.1 nm apart after superimposition

Chothia, C. & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823-826.

Sequence-to-structure relation

By structurally aligning a large set of strucures:



Rost B (1999). The twilight zone of protein alignments. *Protein Engineering* **12**, 85-94.

Sequence identity and structural similarity

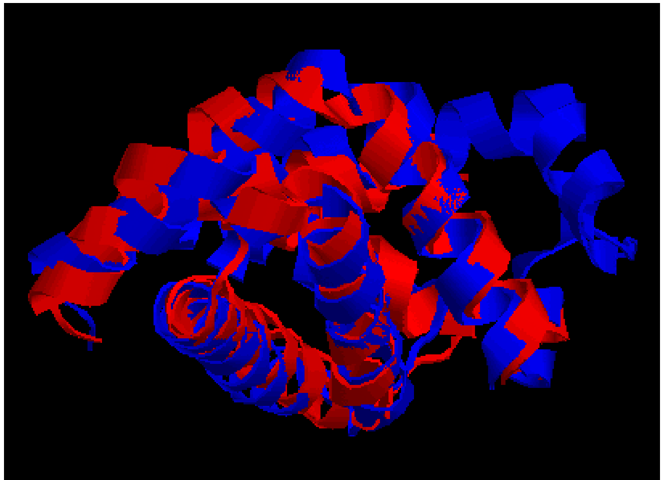
Sequences longer than 100 residues and sharing more the 30% of similar residues have similar structures

For shorter sequences the level of identity must be higher

This DO NOT mean that sequences sharing lower identity MUST have different structures

Example: Sperm Whale Mioglobin and bacterial Emoglobin

RMSD = 0.19 nm, Identity: 14%



Evolution did it (?)

Evolution: Variability and Natural Selection

Sequences of living organisms have evolved from ancestral sequences

Genomic sequences are continually changing at random

The environment operates a selection of the individuals on the basis of the fitness of their phenotypes

When the products of the modified gene (the proteins, the structural RNAs ....) fit worse with the environment than the original ones, the individual has a lower probability of surviving and the mutation has lower probability to be transmitted

NB. Are mutation always random? Not, at least when the mutation rate is taken into consideration (Radman polymerases)

Homology vs sequence similarity

Homology

Sequences are homologous when they derive from a common ancestor

Orthologous when they belong to different species
Paralogous when they are present into the same species (duplication)

Similarity

Sequences are similar if they share a large amount of residues along the sequence: it is a comparative criterion, not an evolutionary one.

Homology vs Sequence Similarity

Are homologous sequences always similar?
It depends on how much did they separated after the divergence.

Are similar sequences always homologous?
Different sequences could be evolved in a convergent way towards similar sequences. (Similarly to wings, independently evolved in insects, birds and bats)

In principle, homology and similarity are different concepts. However, sequences sharing high similarity are likely to be homologous.

Similarity can be measured as the degree of identity

Sequence alignment

Comparing sequences (without structures) can give information about the structural similarity among two proteins

Sequence comparison = sequence alignment

Pairwise Sequence Alignment

```
          EEELTKPRLLWALYFNMRDALSSG

VEKPRILYALYFNMRDSSDE
```

```
EEELTKPRLLWALYFNMRDALSSG-
---VEKPRILYALYFNMRD--SSDE
```

Alignment

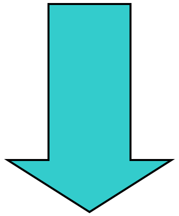
The process of lining up two or more sequences to achieve maximal levels of identity (and conservation, in the case of amino acid sequences) for the purpose of assessing the degree of similarity and the possibility of homology.

Slide by Vered Caspi, BGU. 21.12.2005

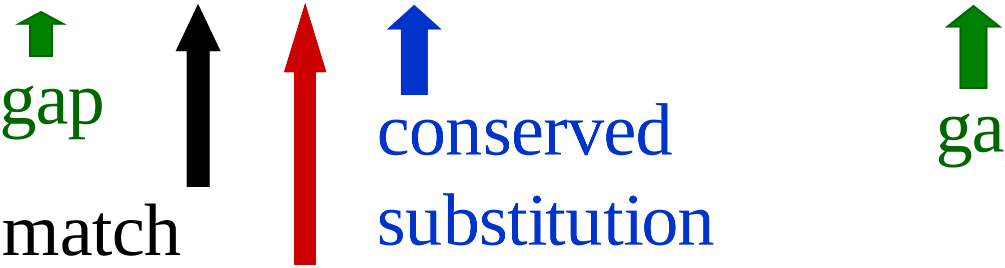
Pairwise Sequence Alignment

EEELTKPRLLWALYFNMRDALSSG

VEKPRILYALYFNMRDSSDE



```
EEELTKPRLLWALYFNMRDALSSG-
---VEKPRILYALYFNMRD--SSDE
```



end gap

mismatch

match

conserved substitution

gap

Slide by Vered Caspi, BGU. 21.12.2005

Sequence alignment

In order to define a sequence alignment procedures we must:

- To define a **score** (or a **distance)** between two aligned sequences

- To find an **algorithm** for finding the alignment with maximum score (or minimal distance)

- To **statistically evaluate** the significance of the alignment

Sequence alignment

In order to define a sequence alignment procedures we must:

- To define a **score** (or a **distance)** between two aligned sequences

- To find an **algorithm** for finding the alignment with maximum score (or minimal distance)

- To **statistically evaluate** the significance of the alignment

Distance between aligned sequences

Alignment without gaps

```
A:    ALASVLIRLITRLYP
B:    ASAVHLNRLITRLYP
```

The alignment consists of a sequence of paired residues

Defining a score for the substitution of residue *i* with residue *j*: *s(i,j)* [Substitution matrix] ,

the score of the two aligned sequences can be computed as the sum of substitution scores over the alignment length

$$Score(A, B, alignment) = \sum_{position\_k} s(A^k, B^k)$$

NB: it assume strong independence among the alignment positions

How to derive substitution scores?
1) Identity matrix

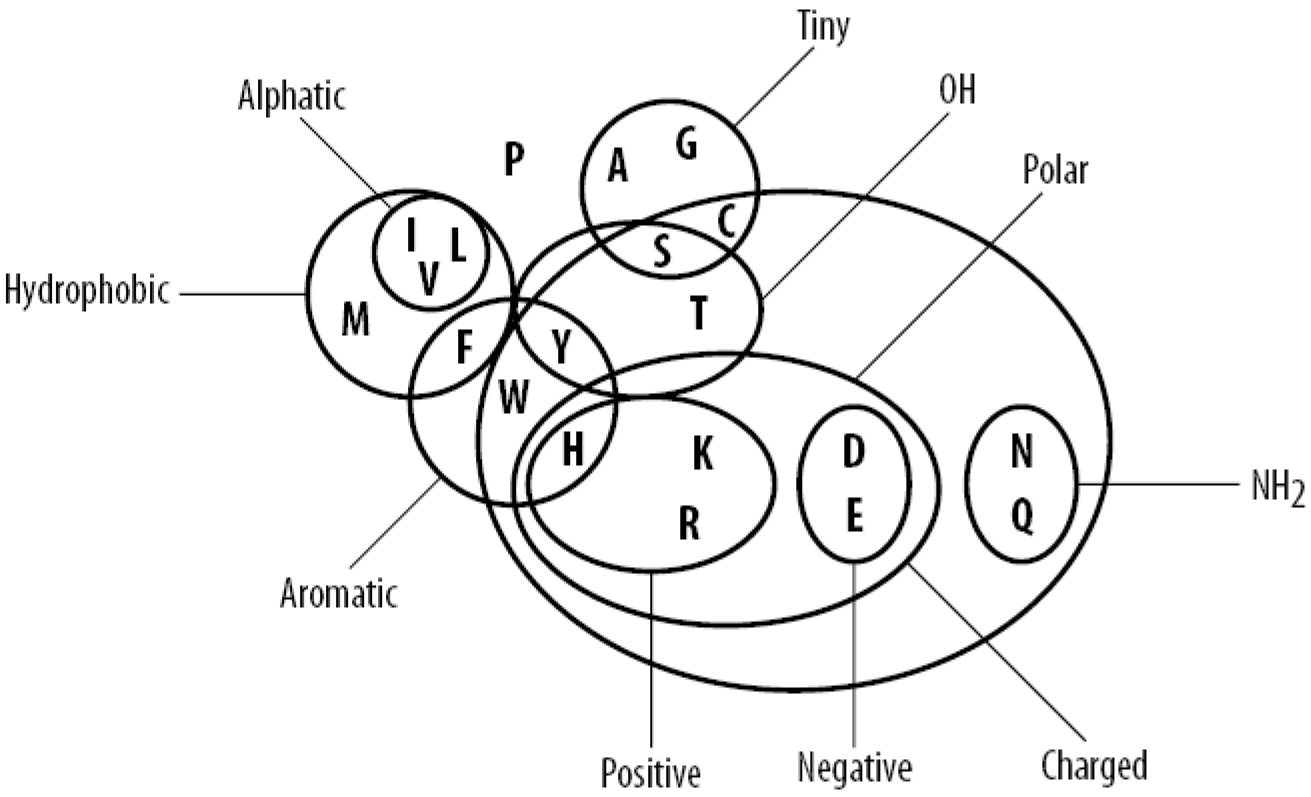
$s(i,j) = 1$ if $i=j$
$s(i,j) = 0$ if $i=j$

```
A:    ALASVLIRLITRLYP
B:    ASAVHLNRLITRLYP
      101001011111111
```

$$Score(A, B, alignment) = 11$$

All the mismatches are considered as equivalent
Is it realistic for proteins? For DNA?
Alternatives?

How to derive substitution scores?
1) Physical-chemical characteristics



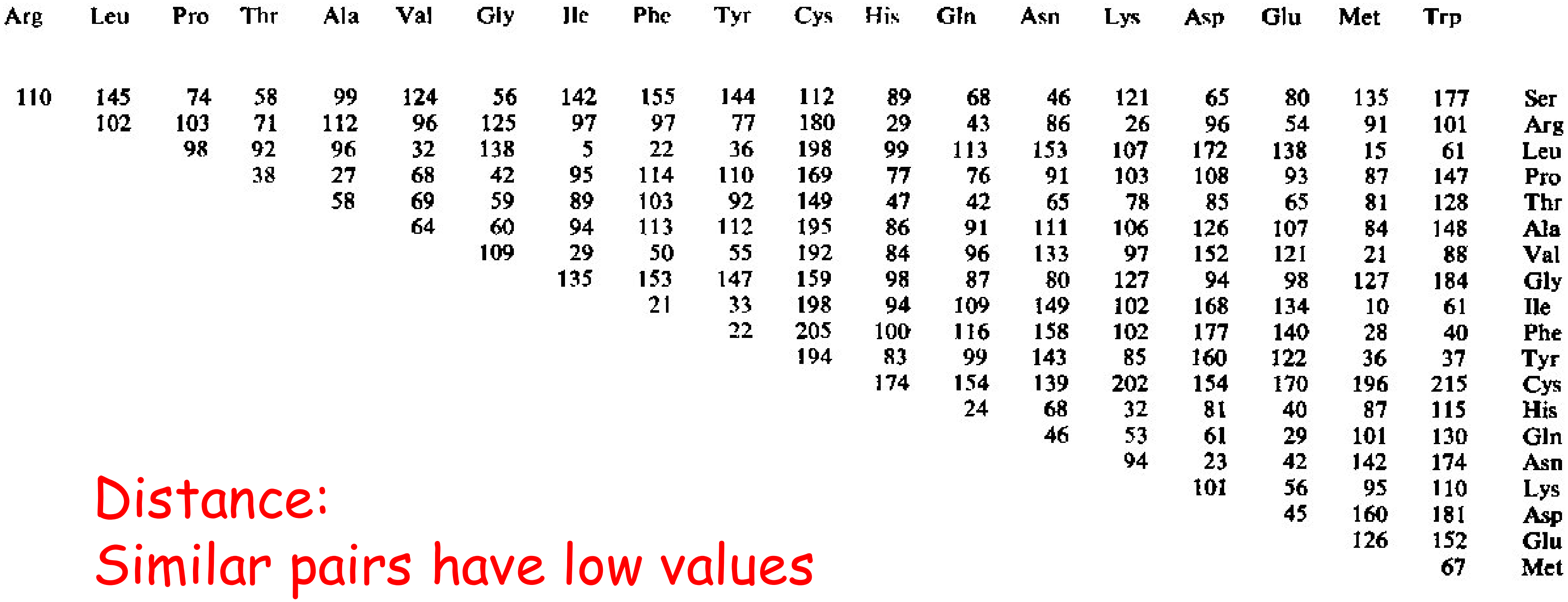
The similarity of pairs of amino acids (McLachlan, 1971)

Score:

Similar pairs have high values

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

Chemical distance (Grantham, 1974)



Arg	Leu	Pro	Thr	Ala	Val	Gly	Ile	Phe	Tyr	Cys	His	Gln	Asn	Lys	Asp	Glu	Met	Trp	
110	145	74	58	99	124	56	142	155	144	112	89	68	46	121	65	80	135	177	Ser
	102	103	71	112	96	125	97	97	77	180	29	43	86	26	96	54	91	101	Arg
		98	92	96	32	138	5	22	36	198	99	113	153	107	172	138	15	61	Leu
			38	27	68	42	95	114	110	169	77	76	91	103	108	93	87	147	Pro
				58	69	59	89	103	92	149	47	42	65	78	85	65	81	128	Thr
					64	60	94	113	112	195	86	91	111	106	126	107	84	148	Ala
						109	29	50	55	192	84	96	133	97	152	121	21	88	Val
							135	153	147	159	98	87	80	127	94	98	127	184	Gly
								21	33	198	94	109	149	102	168	134	10	61	Ile
									22	205	100	116	158	102	177	140	28	40	Phe
										194	83	99	143	85	160	122	36	37	Tyr
											174	154	139	202	154	170	196	215	Cys
												24	68	32	81	40	87	115	His
													46	53	61	29	101	130	Gln
														94	23	42	142	174	Asn
															101	56	95	110	Lys
																45	160	181	Asp
																	126	152	Glu
																		67	Met

Distance:
Similar pairs have low values

3) Substitution scores derived from known structural alignments among proteins

Given a set of good alignments it is possible to estimate the probability of the mutation between any pairs of residues

Given (many) pairs of aligned sequences, we estimate the frequency of substitution $i^A\text{->}j^B$ or $i^B\text{->}j^A$ (independently of the direction): P_{ij}

Ex:

```
A:    ALASVLIRAILRLYP
B:    ALAVLLNRLILRALP
```

$P(A,A)=$

3) Substitution scores derived from known structural alignments among proteins

Given a set of good alignments it is possible toestimate the probability of the mutation between any pairs of residues

Given (many) pairs of aligned sequences, we estimate the frequency of substitution i^A->j^B or i^B->j^A (independently of the direction): P_{ij}

Ex:

A: **A**L**A**SVLIR**A**ILR**L**YP

B: **A**L**A**VLLNR**L**ILR**A**LP

$P(A,A)= N(A^A,A^B)/N = 2/15$

$P(A,L)= P(L,A)=$

3) Substitution scores derived from known structural alignments among proteins

Given a set of good alignments it is possible toestimate the probability of the mutation between any pairs of residues

Given (many) pairs of aligned sequences, we estimate the frequency of substitution i^A->j^B or i^B->j^A (independently of the direction): P_{ij}

Ex:

A: **A**L**A**SVLIR**A**ILR**L**YP

B: **A**L**A**VLLNR**L**ILR**A**LP

$P(A,A)= N(A^A,A^B)/N = 2/15$

$P(A,L)= P(L,A)= [N(L^A,A^B)+N(A^A,L^B)]/N = 2/15$

How to estimate whether a substitution frequency is significant?

Which is the probability that the substitution i->j is random (and so not significant)?

1st set of known alignments

A: ALASVLIR**A**ILR**L**YP
B: ALAVLLNR**L**ILR**A**LP

2nd set of known alignments

A: L**L**LLAALL**L**ALLALL
B: L**A**LLAALL**A**ALLALL

$P(A,L) =$

How to estimate whether a substitution frequency is significant?

Which is the probability that the substitution i->j is random (and so not significant)?

1st set of known alignments

A: ALASVLIR**A**ILR**L**YP
B: ALAVLLNR**L**ILR**A**LP

2nd set of known alignments

A: L**L**LLAALL**L**ALLALL
B: L**A**LLAALL**A**ALLALL

$P(A,L) = 2/15$ in both the cases.

Are they equally significant?

The probability that the substitution is random depends on the frequency of the two substituted residues P_i e P_j

Comparison with the independence condition

Random substitution i^A -> j^B means that the two event: E_1 = (i in A) and E_2 = (j in B) are INDEPENDENT

The "non-randomness" degree is measured by comparing P_{ij} with the product P_iP_j

1st set of known alignments

A: ALASVLIR**A**ILR**L**YP

B: ALAVLLNR**L**ILR**A**LP

P(A)= 6/30, P(L) =10/30

P(A,L) = 2/15 > 1/15 = P(A)P(L): MORE FREQUENT THAN EXPECTED

2nd set of known alignments

A: L**L**LLAALL**L**ALLALL

B: L**A**LLAALL**A**ALLALL

P(A)= 10/30, P(L) =20/30

P(A,L) = 2/15 < 2/9 = P(A)P(L): LESS FREQUENT THAN EXPECTED

Substitution score

The ratio $r_{ij} = P_{ij}/P_iP_j$
determines whether the substitution *i* -> *j* is more or less frequent than expected by random.
Given an alignment between two sequences

A: SLDPIKHTYRALMNVDSLRTFPIL
B: SFGIKKHTKLAKLPVDTIKSWPIL

the probability of all the substitutions A->B is computed as the product of the ratios r_{ij} : $r_{SS}\, r_{LF}\, r_{DG}\, r_{PI}\, r_{IK}$ … (assuming the independence among the positions)

SCORE : $s(i,j) = \text{int}[K \log(P_{ij}/P_iP_j)]$ LOG-ODD SCORE

Thanks to the logarithm the scores can be added up

Minimal distance = Maximal score

Exercise

Compute the substitution score matrix starting from these alignments

```
ACAGGTGGACCT
ACTGGTCGACTT

CTATATGG
CCGGATCG
```

Substitution matrices: PAM

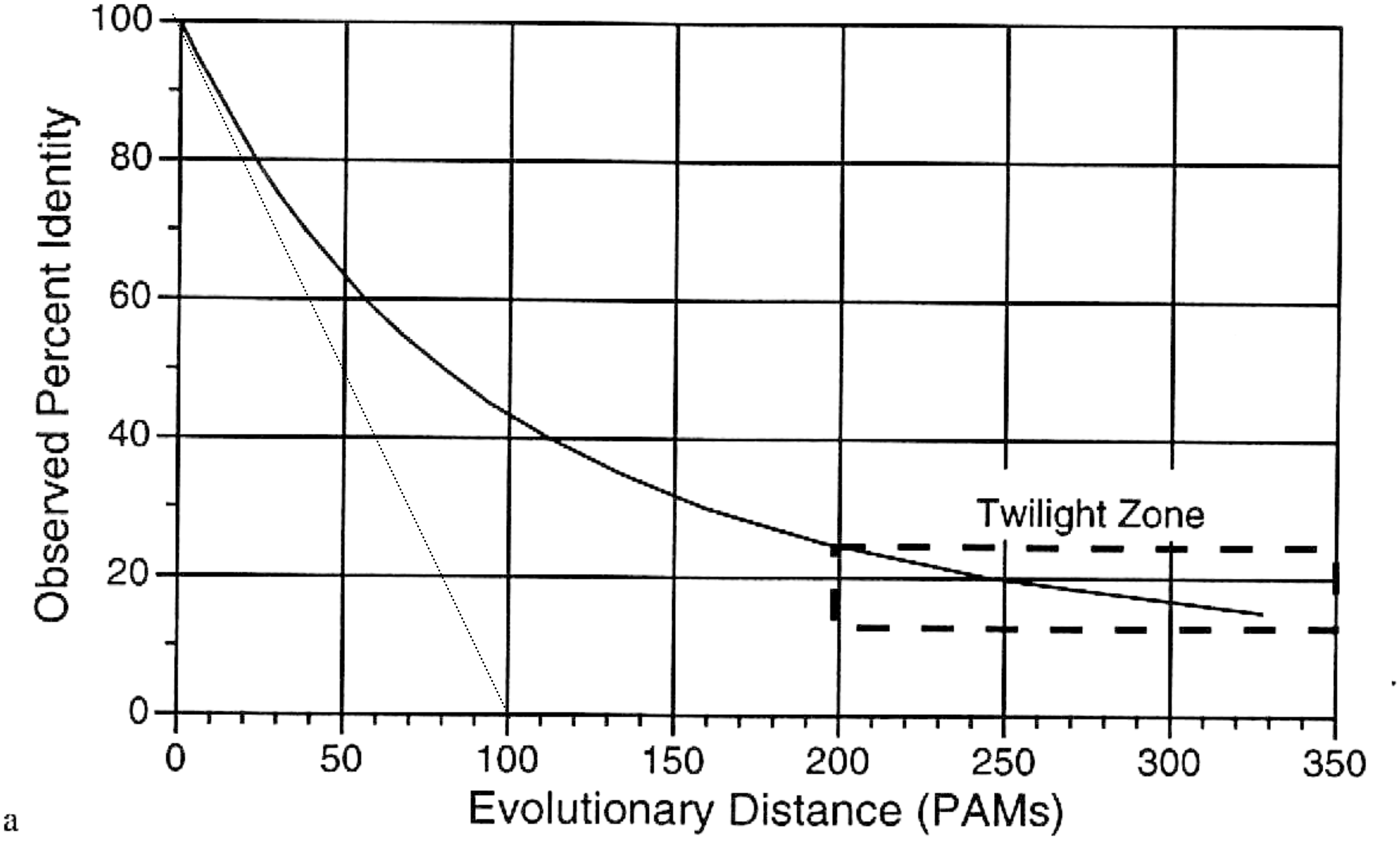
In this framework different matrices can be derived.
The fundamental difference resides in the sets of alignments adopted for building the matrices.

PAM: (Point Accepted Mutation) Margaret Oakley Dayhoff (1972 Atlas of protein sequences and structures)

A point accepted mutation is the replacement of a single amino acid in the primary structure of a protein with another single amino acid, which is accepted by the processes of natural selection.

PAMx: substitution matrix referring to the sequences undergoing x PAMs every 100 residues

Relationship between the PAM and the identity between two sequences



The number of mutational events (PAM) does NOT correspond to the number of different residues between two sequences, when mutations accumulates.

Substitution matrices: PAM

PAMx: Margaret Oakley Dayhoff (1978)

1,572 changes in 71 groups of closely related proteins (85% min sequence identity)

Original formulation considers manually built phylogenetic trees where hypothetical ancestor sequences are inferred.

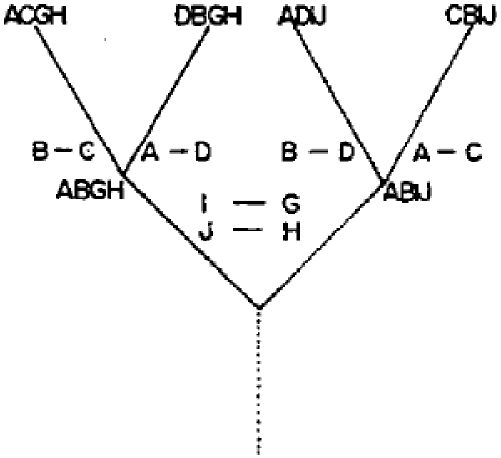
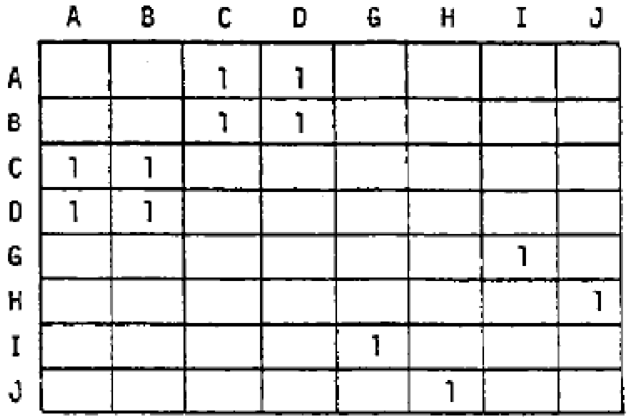


Figure 78. Simplified phylogenetic tree. Four "observed" proteins are shown at the top. Inferred ancestors are shown at the nodes. Amino acid exchanges are indicated along the branches.



	A	B	C	D	G	H	I	J
A			1	1				
B			1	1				
C	1	1						
D	1	1						
G							1	
H								1
I					1			
J						1		

Figure 79. Matrix of accepted point mutations derived from the tree of Figure 78.

Substitution matrices: PAM

PAMx: Margaret Oakley Dayhoff (1978)

1,572 changes in 71 groups of closely related proteins (85% min sequence identity)

General formulation can consider manually built sequence alignments without referring to phylogenesis.

With highly similar sequences, alignments are easily compiled.

Ideally, a conditional probability matrix could be computed using sequences with 1% of mutations

$A^1_{ij} = P(j|i) = N(i,j)/N(i)$ PAM1 probability matrix

Substitution matrices: PAM

PAMx: Margaret Oakley Dayhoff (1978)

1,572 changes in 71 groups of closely related proteins (85% min sequence identity)

It is possible to compute the probability matrix

$A_{ij} = P(j|i) = N(i,j)/N(i)$

If sequences are not 99% identical, matrix referring to 1% PAM is computed (iteratively) rescaling off-diagonal elements:

$$\sum_{i=1}^{20} P(i) \sum_{j \neq i} P(j|i) = 0.01$$

and then diagonal elements, imposing $\sum_{j=1}^{20} P(j|i) = 1$

Substitution matrices: PAM

PAMx: Margaret Oakley Dayhoff (1978)

1,572 changes in 71 groups of closely related proteins (85% min sequence identity)

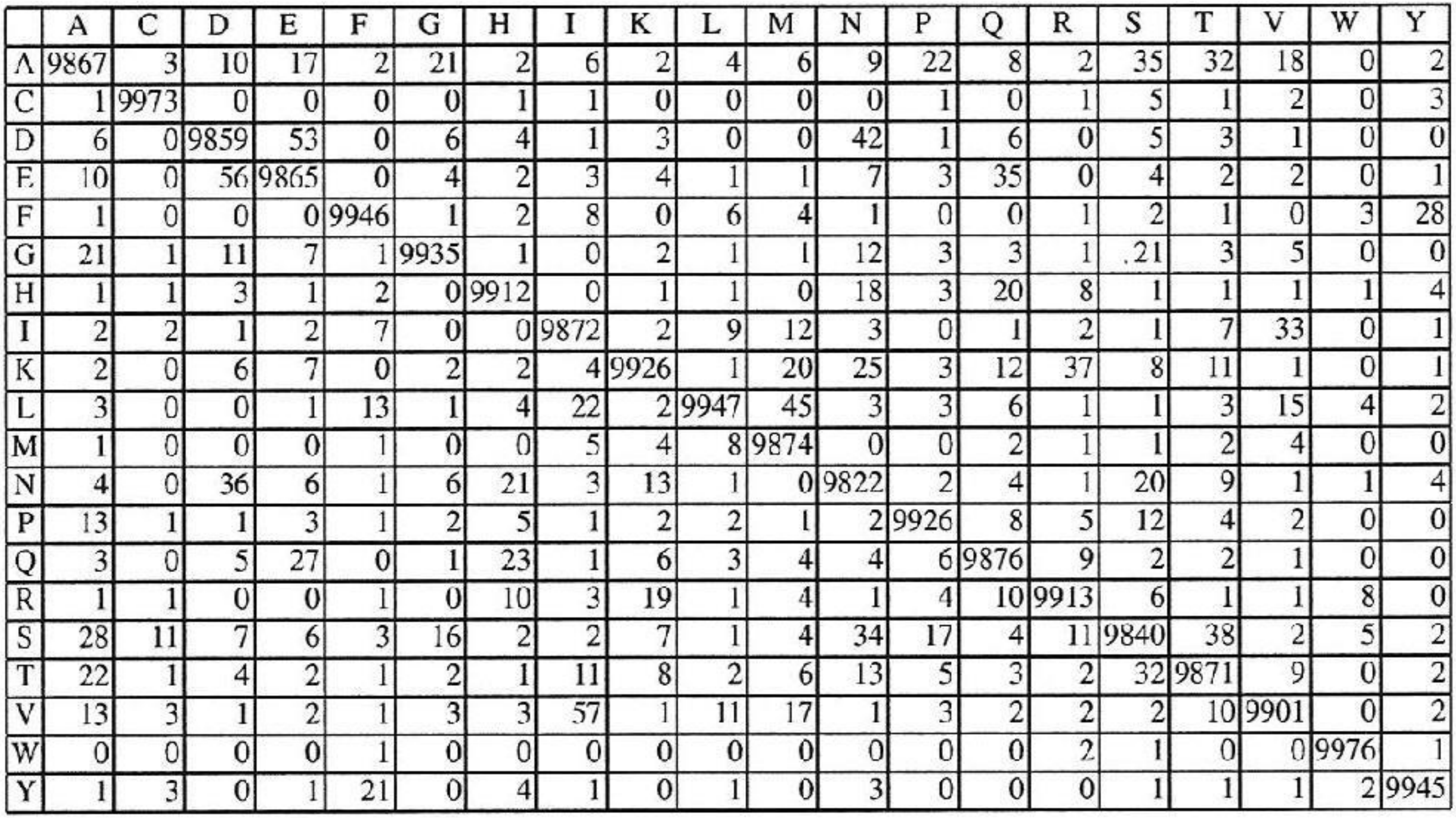
It is possible to compute the probability matrix

A^{1}_{ij} PAM1 probability matrix

$Score(PAM1)_{ij} = K \, Log(A^{1}_{ij} / P_i)$ PAM1 log-odd matrix

PAM 1 Probability matrix : P(j|i) (A^{1}_{ij} x10000)

Table 14.5.4. PAM1 matrix.



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	9867	3	10	17	2	21	2	6	2	4	6	9	22	8	2	35	32	18	0	2
C	1	9973	0	0	0	0	1	1	0	0	0	0	1	0	1	5	1	2	0	3
D	6	0	9859	53	0	6	4	1	3	0	0	42	1	6	0	5	3	1	0	0
E	10	0	56	9865	0	4	2	3	4	1	1	7	3	35	0	4	2	2	0	1
F	1	0	0	0	9946	1	2	8	0	6	4	1	0	0	1	2	1	0	3	28
G	21	1	11	7	1	9935	1	0	2	1	1	12	3	3	1	21	3	5	0	0
H	1	1	3	1	2	0	9912	0	1	1	0	18	3	20	8	1	1	1	1	4
I	2	2	1	2	7	0	0	9872	2	9	12	3	0	1	2	1	7	33	0	1
K	2	0	6	7	0	2	2	4	9926	1	20	25	3	12	37	8	11	1	0	1
L	3	0	0	1	13	1	4	22	2	9947	45	3	3	6	1	1	3	15	4	2
M	1	0	0	0	1	0	0	5	4	8	9874	0	0	2	1	1	2	4	0	0
N	4	0	36	6	1	6	21	3	13	1	0	9822	2	4	1	20	9	1	1	4
P	13	1	1	3	1	2	5	1	2	2	1	2	9926	8	5	12	4	2	0	0
Q	3	0	5	27	0	1	23	1	6	3	4	4	6	9876	9	2	2	1	0	0
R	1	1	0	0	1	0	10	3	19	1	4	1	4	10	9913	6	1	1	8	0
S	28	11	7	6	3	16	2	2	7	1	4	34	17	4	11	9840	38	2	5	2
T	22	1	4	2	1	2	1	11	8	2	6	13	5	3	2	32	9871	9	0	2
V	13	3	1	2	1	3	3	57	1	11	17	1	3	2	2	2	10	9901	0	2
W	0	0	0	0	1	0	0	0	0	0	0	0	0	0	2	1	0	0	9976	1
Y	1	3	0	1	21	0	4	1	0	1	0	3	0	0	0	1	1	1	2	9945

Very stringent matrix: very low values off diagonal

Substitution matrices: PAM

To derive a score matrix for sequences undergone to *n* mutational events every 100 residues:

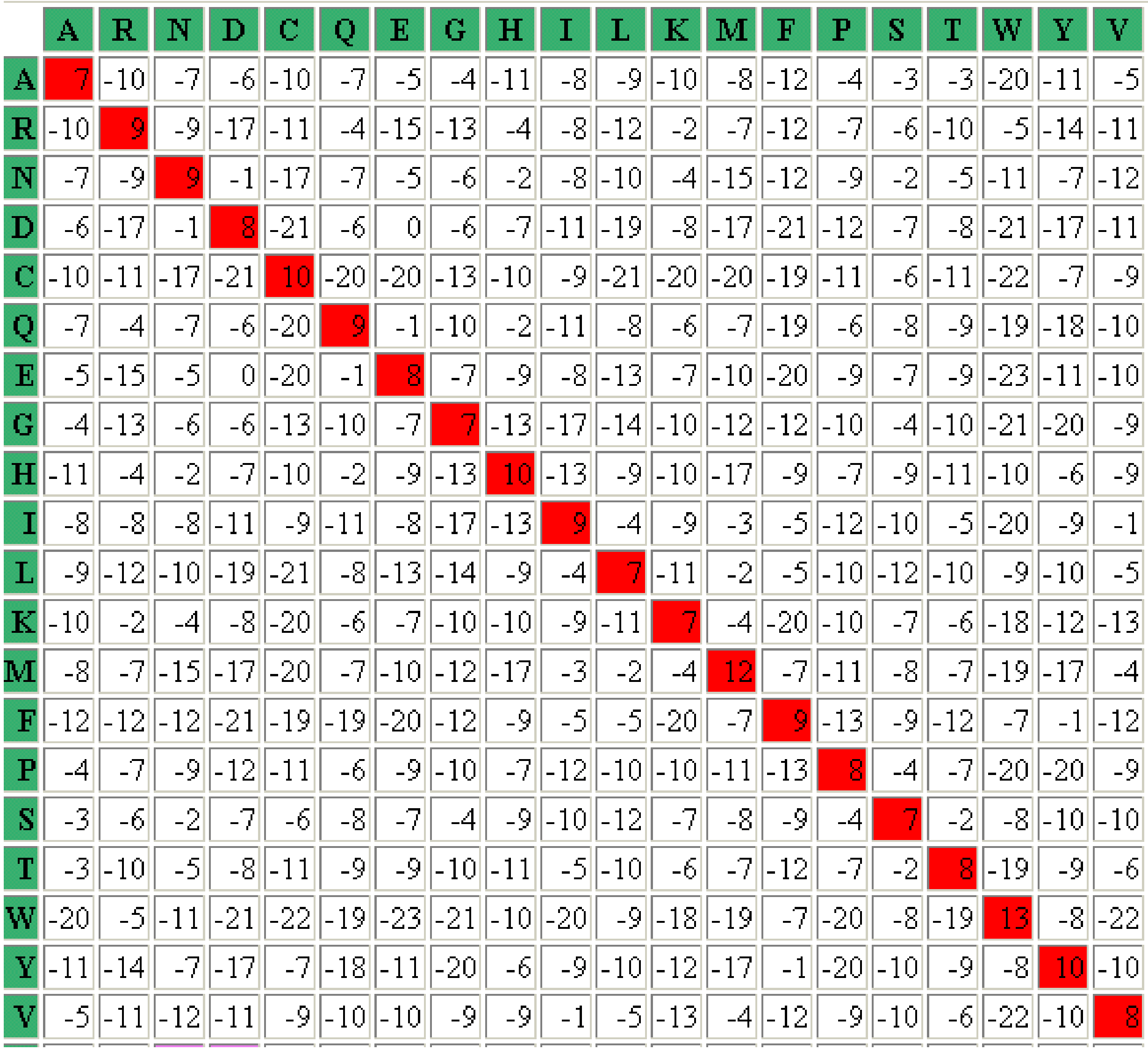
$$A^n_{ij}=(A^1_{ij})^n$$

n=2 $P(i|j) = \Pi_l P(i|l)\, P(l|j)$

The residue j can change into i via any intermediate l

$Score\ (PAMn)_{ij} = Log(A^n_{ij} / P_i)$

PAM10 log odd matrix



	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	7	-10	-7	-6	-10	-7	-5	-4	-11	-8	-9	-10	-8	-12	-4	-3	-3	-20	-11	-5
R	-10	9	-9	-17	-11	-4	-15	-13	-4	-8	-12	-2	-7	-12	-7	-6	-10	-5	-14	-11
N	-7	-9	9	-1	-17	-7	-5	-6	-2	-8	-10	-4	-15	-12	-9	-2	-5	-11	-7	-12
D	-6	-17	-1	8	-21	-6	0	-6	-7	-11	-19	-8	-17	-21	-12	-7	-8	-21	-17	-11
C	-10	-11	-17	-21	10	-20	-20	-13	-10	-9	-21	-20	-20	-19	-11	-6	-11	-22	-7	-9
Q	-7	-4	-7	-6	-20	9	-1	-10	-2	-11	-8	-6	-7	-19	-6	-8	-9	-19	-18	-10
E	-5	-15	-5	0	-20	-1	8	-7	-9	-8	-13	-7	-10	-20	-9	-7	-9	-23	-11	-10
G	-4	-13	-6	-6	-13	-10	-7	7	-13	-17	-14	-10	-12	-12	-10	-4	-10	-21	-20	-9
H	-11	-4	-2	-7	-10	-2	-9	-13	10	-13	-9	-10	-17	-9	-7	-9	-11	-10	-6	-9
I	-8	-8	-8	-11	-9	-11	-8	-17	-13	9	-4	-9	-3	-5	-12	-10	-5	-20	-9	-1
L	-9	-12	-10	-19	-21	-8	-13	-14	-9	-4	7	-11	-2	-5	-10	-12	-10	-9	-10	-5
K	-10	-2	-4	-8	-20	-6	-7	-10	-10	-9	-11	7	-4	-20	-10	-7	-6	-18	-12	-13
M	-8	-7	-15	-17	-20	-7	-10	-12	-17	-3	-2	-4	12	-7	-11	-8	-7	-19	-17	-4
F	-12	-12	-12	-21	-19	-19	-20	-12	-9	-5	-5	-20	-7	9	-13	-9	-12	-7	-1	-12
P	-4	-7	-9	-12	-11	-6	-9	-10	-7	-12	-10	-10	-11	-13	8	-4	-7	-20	-20	-9
S	-3	-6	-2	-7	-6	-8	-7	-4	-9	-10	-12	-7	-8	-9	-4	7	-2	-8	-10	-10
T	-3	-10	-5	-8	-11	-9	-9	-10	-11	-5	-10	-6	-7	-12	-7	-2	8	-19	-9	-6
W	-20	-5	-11	-21	-22	-19	-23	-21	-10	-20	-9	-18	-19	-7	-20	-8	-19	13	-8	-22
Y	-11	-14	-7	-17	-7	-18	-11	-20	-6	-9	-10	-12	-17	-1	-20	-10	-9	-8	10	-10
V	-5	-11	-12	-11	-9	-10	-10	-9	-9	-1	-5	-13	-4	-12	-9	-10	-6	-22	-10	8

Very stringent matrix: no positive value out of the diagonal

Substitution matrices: PAM

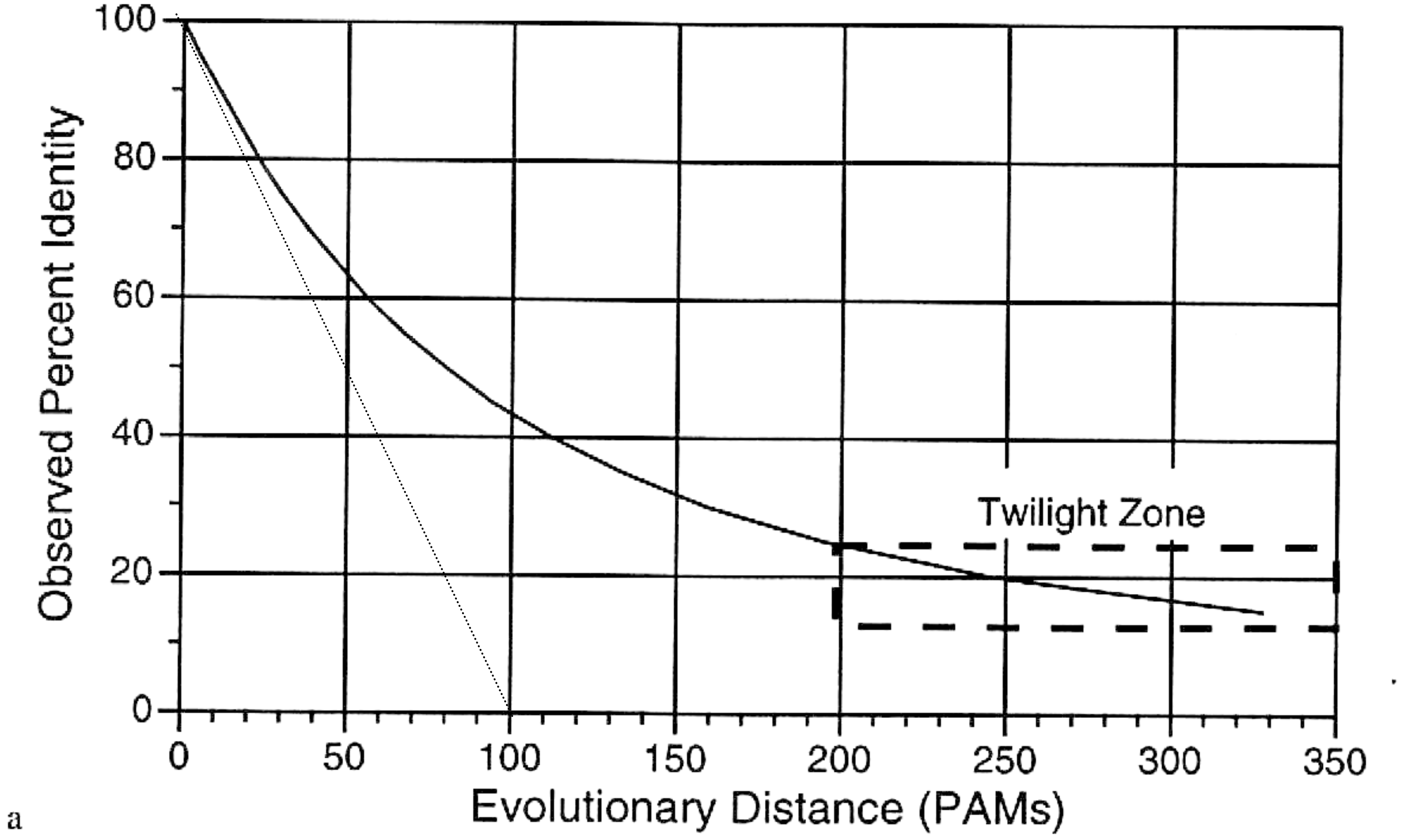
To derive a score matrix for sequences undergone to *n* mutational events every 100 residues:

n % mutational events does not mean that n out of 100 residue are different among the two sequences:

as the number of mutational events increases, different mutations can occur in the same position

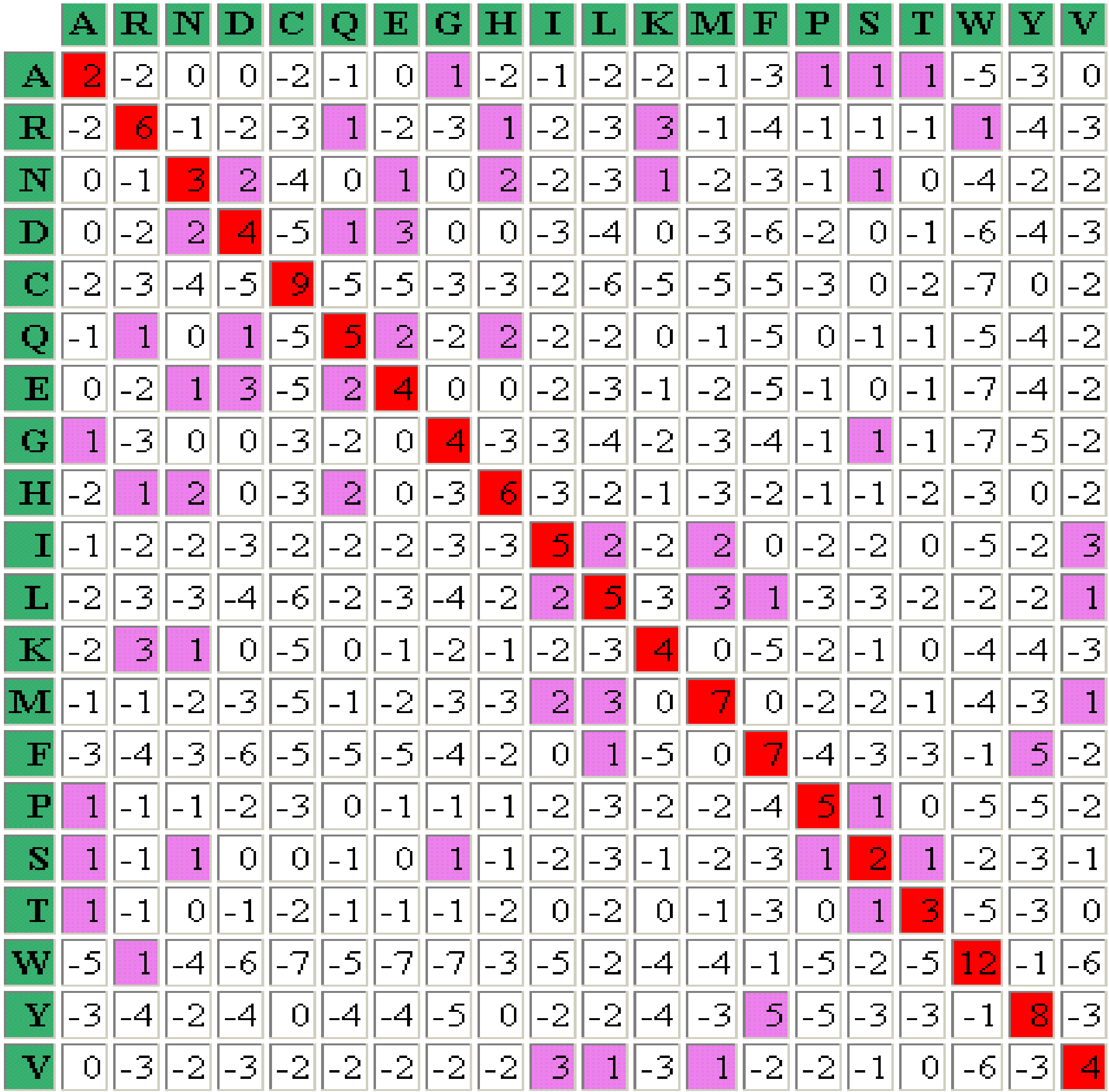
Then 100 mutational events in a 100-residue sequence leave some unvaried position

Relationship between the PAM and the identity between two sequences



The number of mutational events (PAM) does NOT correspond to the number of different residues between two sequences, when mutations accumulates.

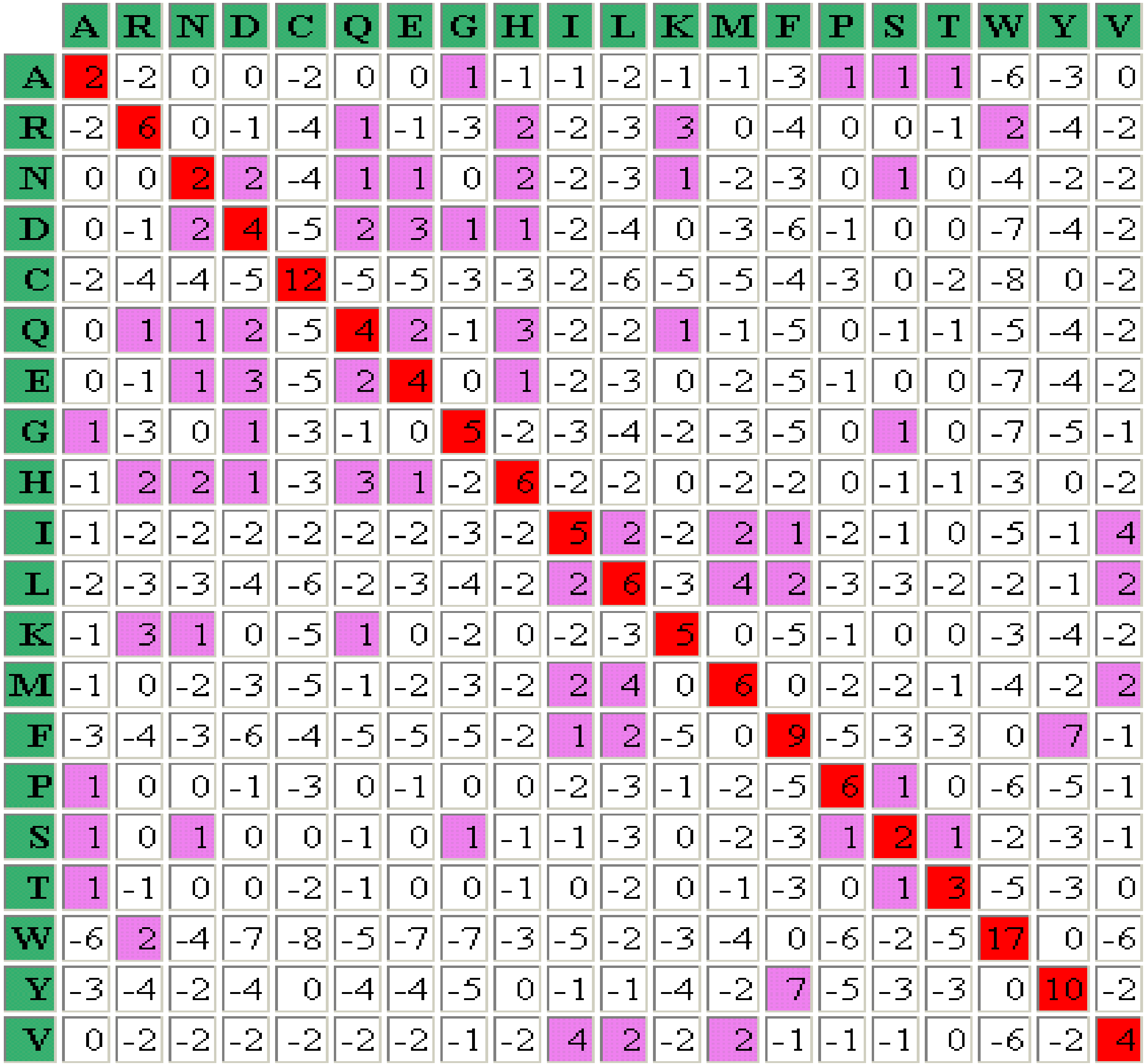
PAM160 log odd matrix



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

Some positive values out of the diagonal: residue pairs endowed with positive scores are SIMILAR

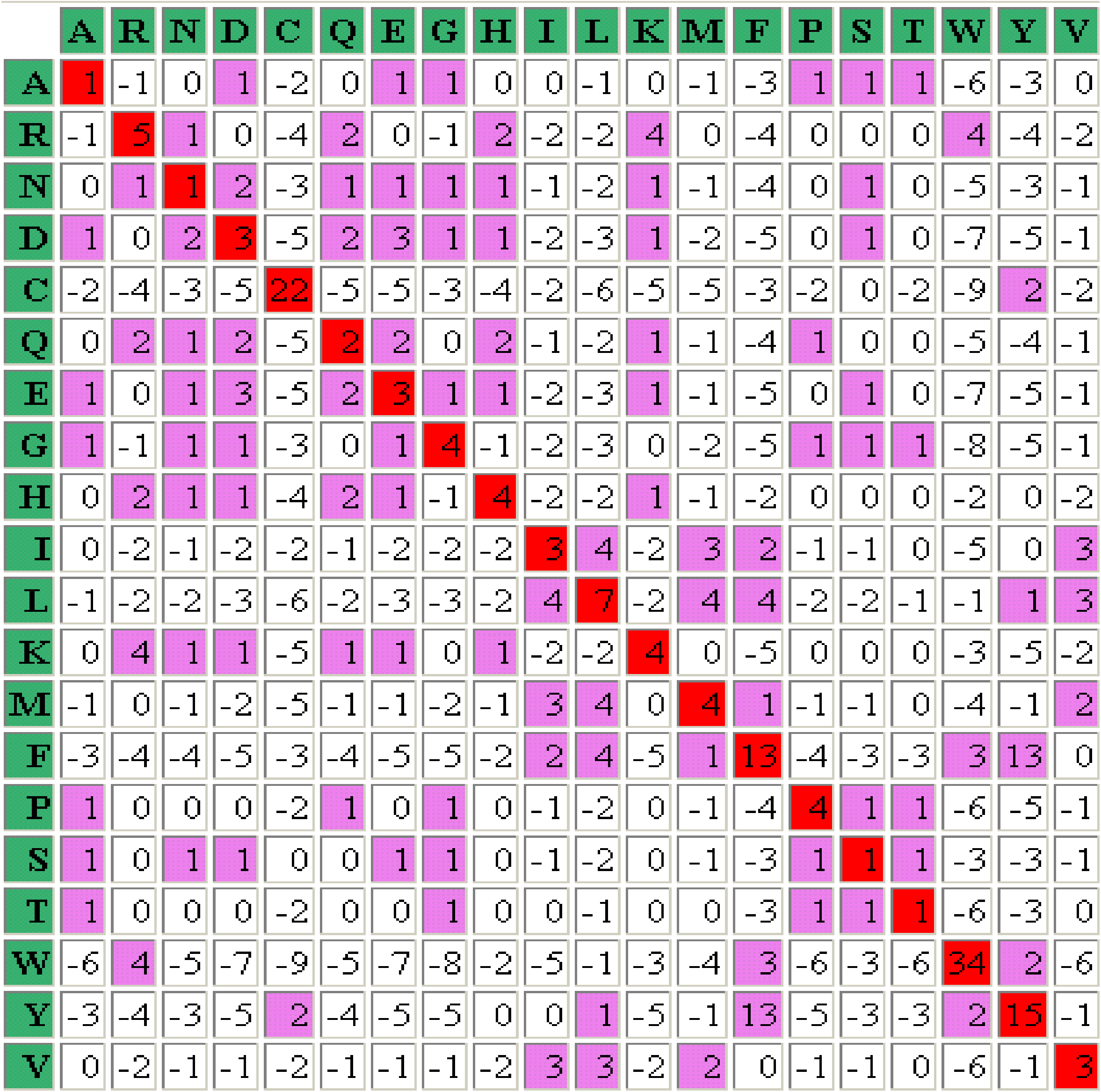
PAM250 log odd matrix



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

Often adopted

PAM500 log odd matrix



	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	1	-1	0	1	-2	0	1	1	0	0	-1	0	-1	-3	1	1	1	-6	-3	0
R	-1	5	1	0	-4	2	0	-1	2	-2	-2	4	0	-4	0	0	0	4	-4	-2
N	0	1	1	2	-3	1	1	1	1	-1	-2	1	-1	-4	0	1	0	-5	-3	-1
D	1	0	2	3	-5	2	3	1	1	-2	-3	1	-2	-5	0	1	0	-7	-5	-1
C	-2	-4	-3	-5	22	-5	-5	-3	-4	-2	-6	-5	-5	-3	-2	0	-2	-9	2	-2
Q	0	2	1	2	-5	2	2	0	2	-1	-2	1	-1	-4	1	0	0	-5	-4	-1
E	1	0	1	3	-5	2	3	1	1	-2	-3	1	-1	-5	0	1	0	-7	-5	-1
G	1	-1	1	1	-3	0	1	4	-1	-2	-3	0	-2	-5	1	1	1	-8	-5	-1
H	0	2	1	1	-4	2	1	-1	4	-2	-2	1	-1	-2	0	0	0	-2	0	-2
I	0	-2	-1	-2	-2	-1	-2	-2	-2	3	4	-2	3	2	-1	-1	0	-5	0	3
L	-1	-2	-2	-3	-6	-2	-3	-3	-2	4	7	-2	4	4	-2	-2	-1	-1	1	3
K	0	4	1	1	-5	1	1	0	1	-2	-2	4	0	-5	0	0	0	-3	-5	-2
M	-1	0	-1	-2	-5	-1	-1	-2	-1	3	4	0	4	1	-1	-1	0	-4	-1	2
F	-3	-4	-4	-5	-3	-4	-5	-5	-2	2	4	-5	1	13	-4	-3	-3	3	13	0
P	1	0	0	0	-2	1	0	1	0	-1	-2	0	-1	-4	4	1	1	-6	-5	-1
S	1	0	1	1	0	0	1	1	0	-1	-2	0	-1	-3	1	1	1	-3	-3	-1
T	1	0	0	0	-2	0	0	1	0	0	-1	0	0	-3	1	1	1	-6	-3	0
W	-6	4	-5	-7	-9	-5	-7	-8	-2	-5	-1	-3	-4	3	-6	-3	-6	34	2	-6
Y	-3	-4	-3	-5	2	-4	-5	-5	0	0	1	-5	-1	13	-5	-3	-3	2	15	-1
V	0	-2	-1	-1	-2	-1	-1	-1	-2	3	3	-2	2	0	-1	-1	0	-6	-1	3

Substitution matrices

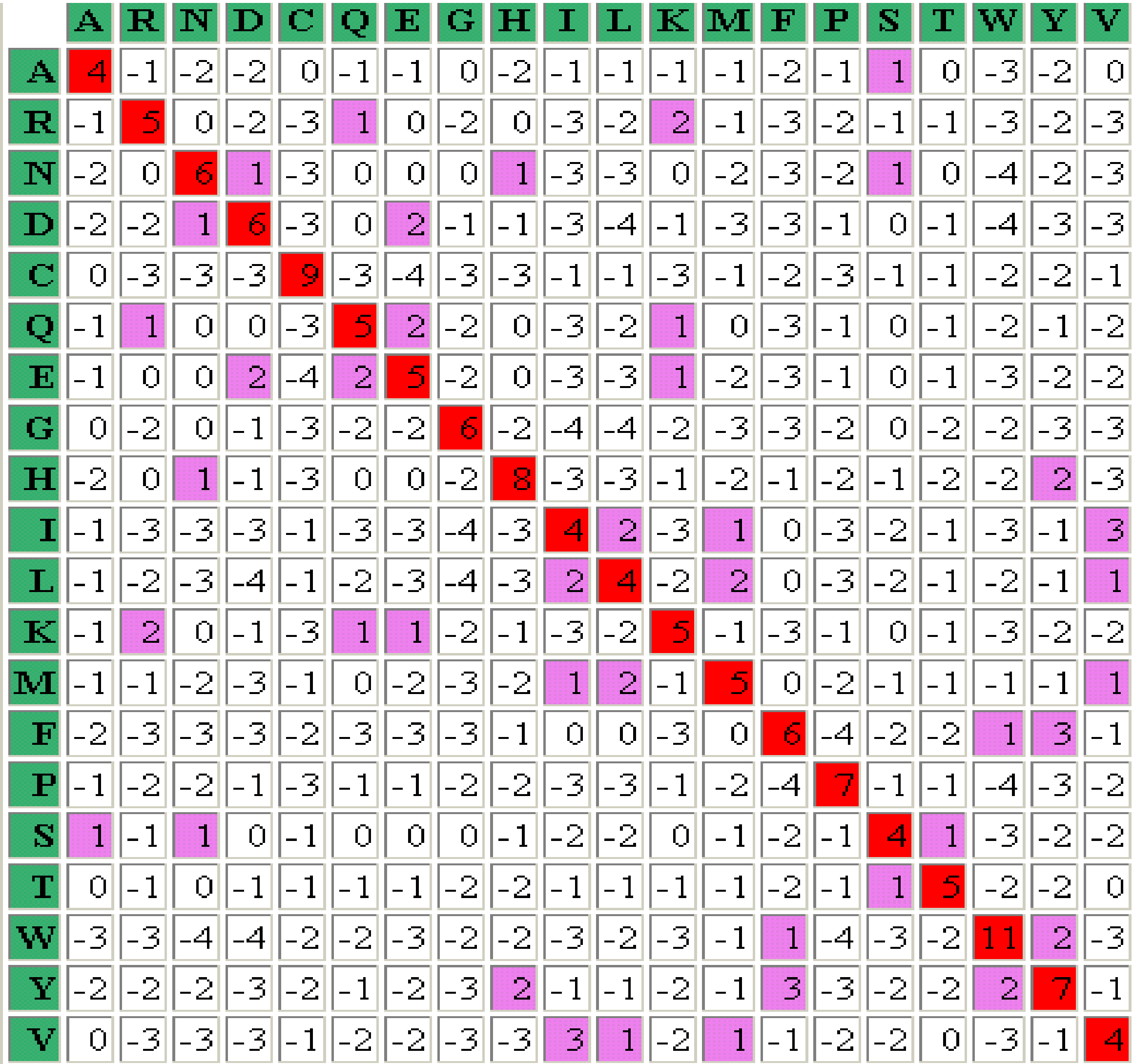
PAM matrices are computed under the hypothesis that substitution scores for distant sequences can be derived from the rate of mutation observed in pairs of very similar sequences.

BLOSUMx: BLOck Substitution Matrix (Henikoff and Henikoff (1992)

Family of matrices computed directly starting from curated alignments of sequences with at most x% of identical residues

For highly similar sequences low PAMs or high BLOSUMs have to used. The contrary, for distant sequences

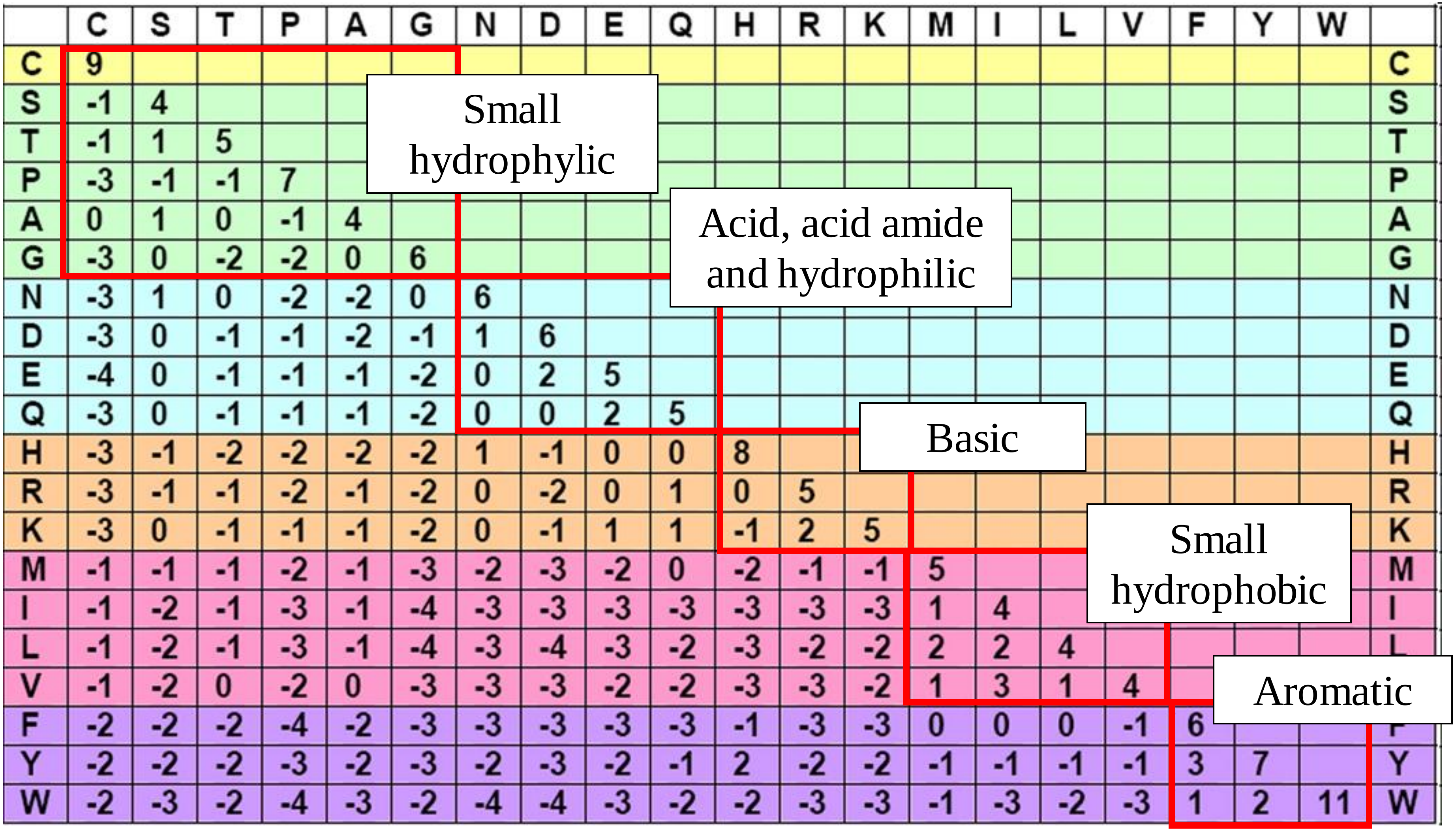
BLOSUM62



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

Often adopted

BLOSUM62 Matrix



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

Small hydrophylic

Acid, acid amide and hydrophilic

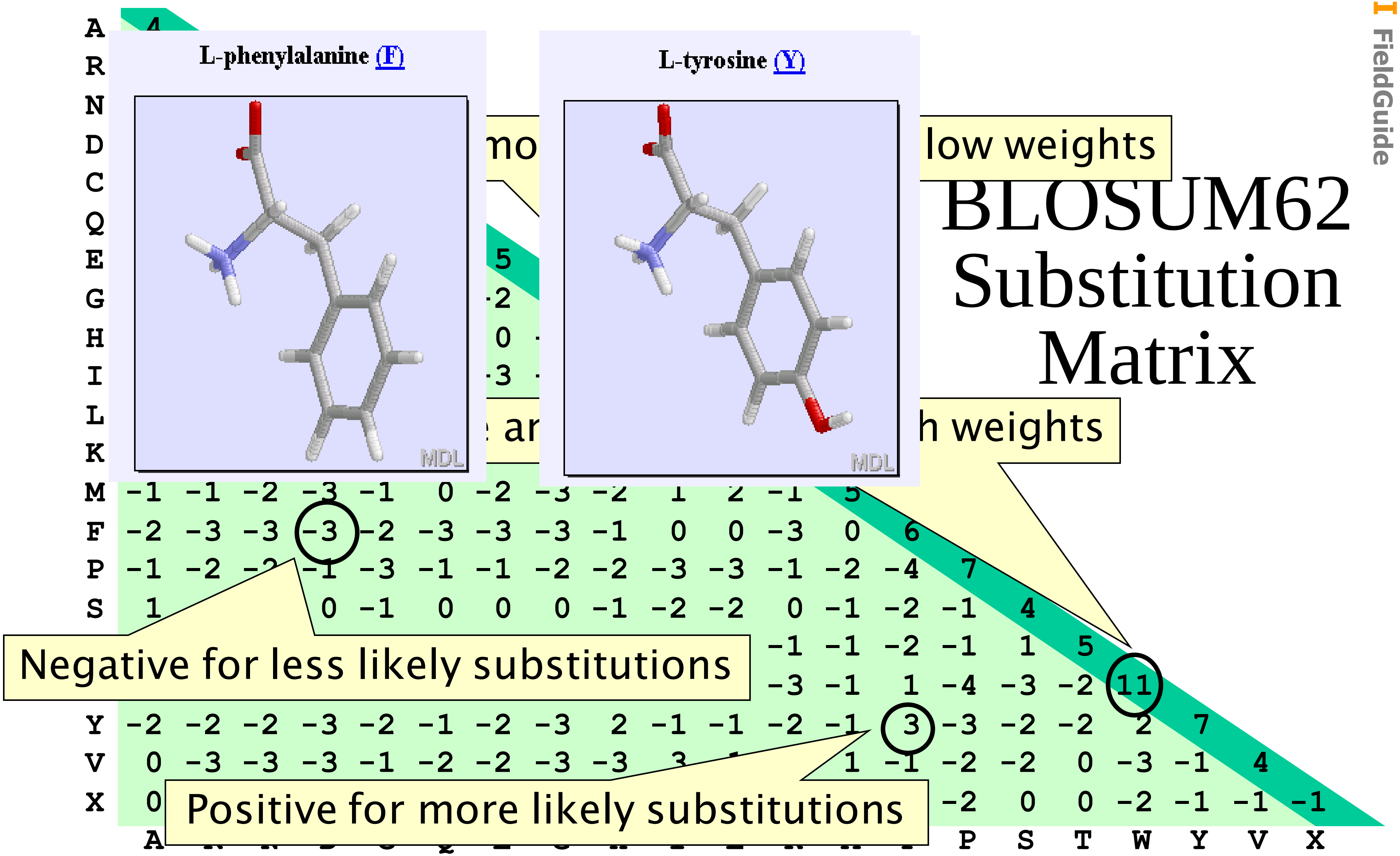
Basic

Small hydrophobic

Aromatic

Scoring Systems - Proteins

BLOSUM62 Substitution Matrix

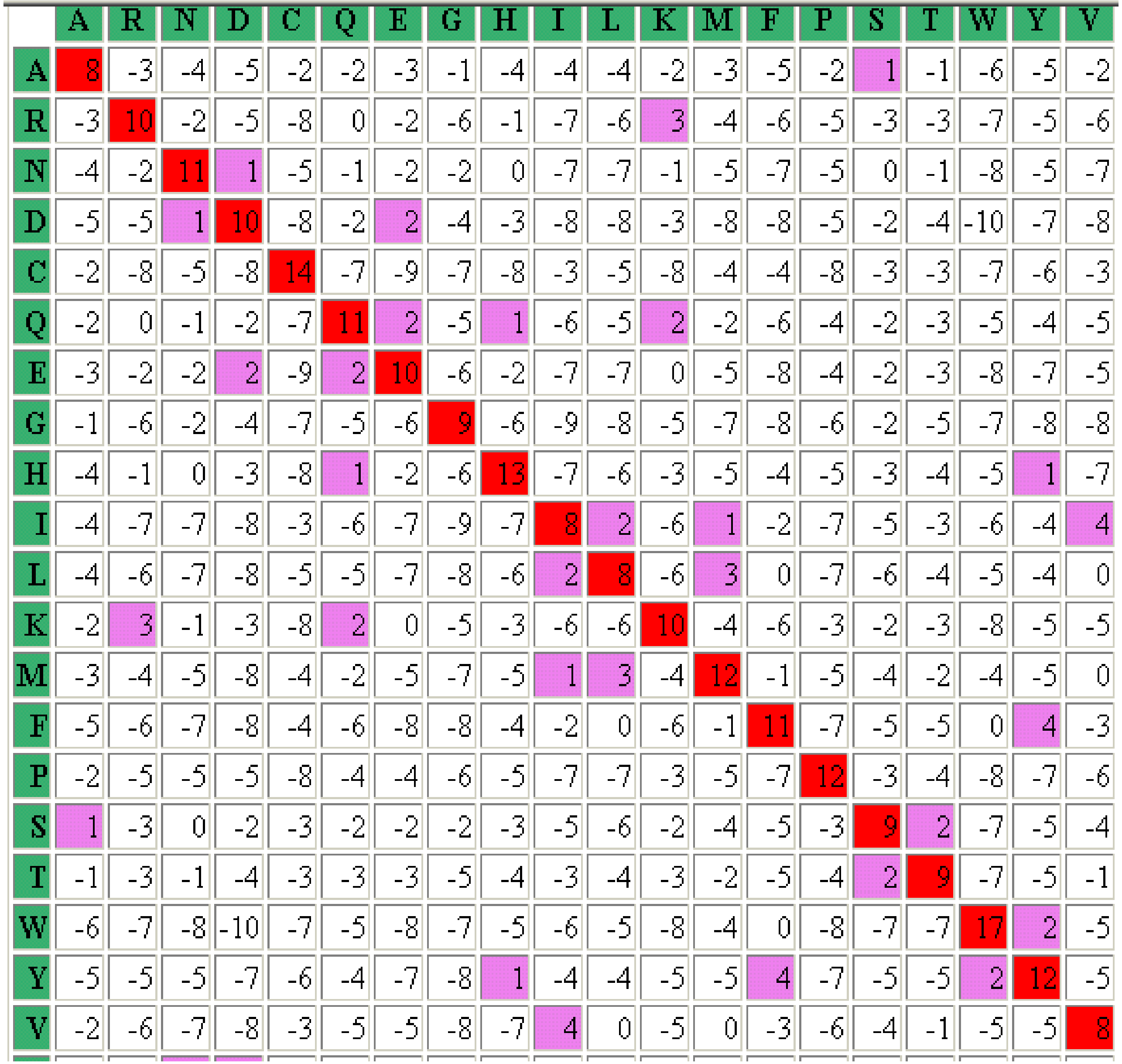

L-phenylalanine (F)
MDL
L-tyrosine (Y)
MDL
mo
low weights
e ar
h weights
A R N D C Q E G H I L K M F P S T W Y V X
M −1 −1 −2 −3 −1 0 −2 −3 −2 1 2 −1 5
F −2 −3 −3 −3 −2 −3 −3 −3 −1 0 0 −3 0 6
P −1 −2 −2 −1 −3 −1 −1 −2 −2 −3 −3 −1 −2 −4 7
S 1 0 −1 0 0 0 −1 −2 −2 0 −1 −2 −1 4
−1 −1 −2 −1 1 5
−3 −1 1 −4 −3 −2 11
Y −2 −2 −2 −3 −2 −1 −2 −3 2 −1 −1 −2 −1 3 −3 −2 −2 2 7
V 0 −3 −3 −3 −1 −2 −2 −3 −3 3 1 −1 −2 −2 0 −3 −1 4
X 0 −2 0 0 −2 −1 −1 −1
P S T W Y V X

Negative for less likely substitutions

Positive for more likely substitutions

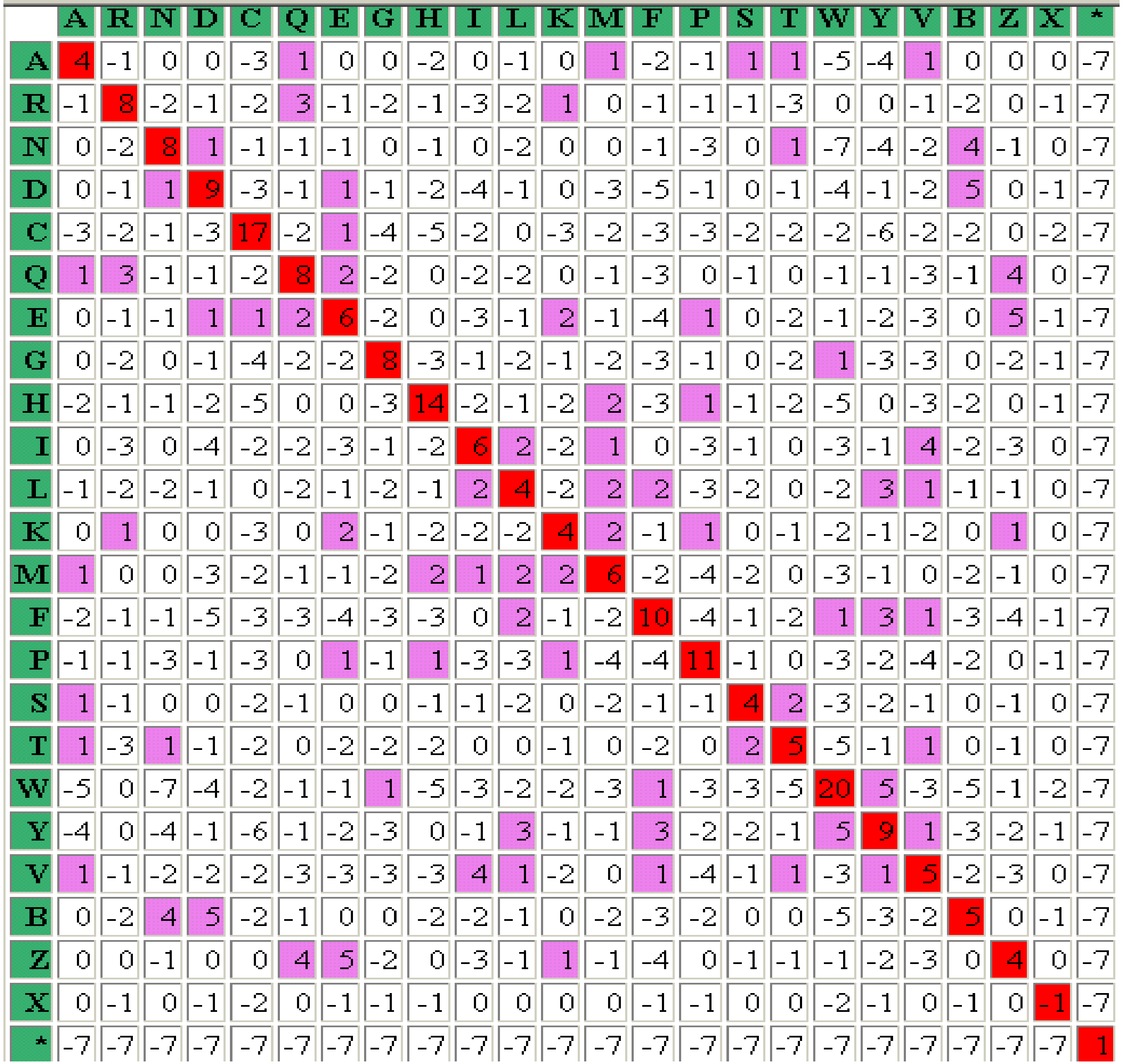
From NCBI field guides

BLOSUM90



	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	8	-3	-4	-5	-2	-2	-3	-1	-4	-4	-4	-2	-3	-5	-2	1	-1	-6	-5	-2
R	-3	10	-2	-5	-8	0	-2	-6	-1	-7	-6	3	-4	-6	-5	-3	-3	-7	-5	-6
N	-4	-2	11	1	-5	-1	-2	-2	0	-7	-7	-1	-5	-7	-5	0	-1	-8	-5	-7
D	-5	-5	1	10	-8	-2	2	-4	-3	-8	-8	-3	-8	-8	-5	-2	-4	-10	-7	-8
C	-2	-8	-5	-8	14	-7	-9	-7	-8	-3	-5	-8	-4	-4	-8	-3	-3	-7	-6	-3
Q	-2	0	-1	-2	-7	11	2	-5	1	-6	-5	2	-2	-6	-4	-2	-3	-5	-4	-5
E	-3	-2	-2	2	-9	2	10	-6	-2	-7	-7	0	-5	-8	-4	-2	-3	-8	-7	-5
G	-1	-6	-2	-4	-7	-5	-6	9	-6	-9	-8	-5	-7	-8	-6	-2	-5	-7	-8	-8
H	-4	-1	0	-3	-8	1	-2	-6	13	-7	-6	-3	-5	-4	-5	-3	-4	-5	1	-7
I	-4	-7	-7	-8	-3	-6	-7	-9	-7	8	2	-6	1	-2	-7	-5	-3	-6	-4	4
L	-4	-6	-7	-8	-5	-5	-7	-8	-6	2	8	-6	3	0	-7	-6	-4	-5	-4	0
K	-2	3	-1	-3	-8	2	0	-5	-3	-6	-6	10	-4	-6	-3	-2	-3	-8	-5	-5
M	-3	-4	-5	-8	-4	-2	-5	-7	-5	1	3	-4	12	-1	-5	-4	-2	-4	-5	0
F	-5	-6	-7	-8	-4	-6	-8	-8	-4	-2	0	-6	-1	11	-7	-5	-5	0	4	-3
P	-2	-5	-5	-5	-8	-4	-4	-6	-5	-7	-7	-3	-5	-7	12	-3	-4	-8	-7	-6
S	1	-3	0	-2	-3	-2	-2	-2	-3	-5	-6	-2	-4	-5	-3	9	2	-7	-5	-4
T	-1	-3	-1	-4	-3	-3	-3	-5	-4	-3	-4	-3	-2	-5	-4	2	9	-7	-5	-1
W	-6	-7	-8	-10	-7	-5	-8	-7	-5	-6	-5	-8	-4	0	-8	-7	-7	17	2	-5
Y	-5	-5	-5	-7	-6	-4	-7	-8	1	-4	-4	-5	-5	4	-7	-5	-5	2	12	-5
V	-2	-6	-7	-8	-3	-5	-5	-8	-7	4	0	-5	0	-3	-6	-4	-1	-5	-5	8

BLOSUM30



	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	4	-1	0	0	-3	1	0	0	-2	0	-1	0	1	-2	-1	1	1	-5	-4	1	0	0	0	-7
R	-1	8	-2	-1	-2	3	-1	-2	-1	-3	-2	1	0	-1	-1	-1	-3	0	0	-1	-2	0	-1	-7
N	0	-2	8	1	-1	-1	-1	0	-1	0	-2	0	0	-1	-3	0	1	-7	-4	-2	4	-1	0	-7
D	0	-1	1	9	-3	-1	1	-1	-2	-4	-1	0	-3	-5	-1	0	-1	-4	-1	-2	5	0	-1	-7
C	-3	-2	-1	-3	17	-2	1	-4	-5	-2	0	-3	-2	-3	-3	-2	-2	-2	-6	-2	-2	0	-2	-7
Q	1	3	-1	-1	-2	8	2	-2	0	-2	-2	0	-1	-3	0	-1	0	-1	-1	-3	-1	4	0	-7
E	0	-1	-1	1	1	2	6	-2	0	-3	-1	2	-1	-4	1	0	-2	-1	-2	-3	0	5	-1	-7
G	0	-2	0	-1	-4	-2	-2	8	-3	-1	-2	-1	-2	-3	-1	0	-2	1	-3	-3	0	-2	-1	-7
H	-2	-1	-1	-2	-5	0	0	-3	14	-2	-1	-2	2	-3	1	-1	-2	-5	0	-3	-2	0	-1	-7
I	0	-3	0	-4	-2	-2	-3	-1	-2	6	2	-2	1	0	-3	-1	0	-3	-1	4	-2	-3	0	-7
L	-1	-2	-2	-1	0	-2	-1	-2	-1	2	4	-2	2	2	-3	-2	0	-2	3	1	-1	-1	0	-7
K	0	1	0	0	-3	0	2	-1	-2	-2	-2	4	2	-1	1	0	-1	-2	-1	-2	0	1	0	-7
M	1	0	0	-3	-2	-1	-1	-2	2	1	2	2	6	-2	-4	-2	0	-3	-1	0	-2	-1	0	-7
F	-2	-1	-1	-5	-3	-3	-4	-3	-3	0	2	-1	-2	10	-4	-1	-2	1	3	1	-3	-4	-1	-7
P	-1	-1	-3	-1	-3	0	1	-1	1	-3	-3	1	-4	-4	11	-1	0	-3	-2	-4	-2	0	-1	-7
S	1	-1	0	0	-2	-1	0	0	-1	-1	-2	0	-2	-1	-1	4	2	-3	-2	-1	0	-1	0	-7
T	1	-3	1	-1	-2	0	-2	-2	-2	0	0	-1	0	-2	0	2	5	-5	-1	1	0	-1	0	-7
W	-5	0	-7	-4	-2	-1	-1	1	-5	-3	-2	-2	-3	1	-3	-3	-5	20	5	-3	-5	-1	-2	-7
Y	-4	0	-4	-1	-6	-1	-2	-3	0	-1	3	-1	-1	3	-2	-2	-1	5	9	1	-3	-2	-1	-7
V	1	-1	-2	-2	-2	-3	-3	-3	-3	4	1	-2	0	1	-4	-1	1	-3	1	5	-2	-3	0	-7
B	0	-2	4	5	-2	-1	0	0	-2	-2	-1	0	-2	-3	-2	0	0	-5	-3	-2	5	0	-1	-7
Z	0	0	-1	0	0	4	5	-2	0	-3	-1	1	-1	-4	0	-1	-1	-1	-2	-3	0	4	0	-7
X	0	-1	0	-1	-2	0	-1	-1	-1	0	0	0	0	-1	-1	0	0	-2	-1	0	-1	0	-1	-7
*	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	1

PAM Versus BLOSUM

- PAM is based on an evolutionary model.
- BLOSUM is based on protein families.
- PAM is based on global alignment.
- BLOSUM is based on local alignment.

BLOSUM 80	BLOSUM 62	BLOSUM 45
PAM 1	PAM 120	PAM 250
Less divergent	↔	*More divergent*

Distance between aligned sequences

Alignment without gaps

```
A:    ALASVLIRLITRLYP
B:    ASAVHLNRLITRLYP
```

The alignment consists of a sequence of paired residues

Defining a score for the substitution of residue *i* with residue *j*: *s(i,j)* [Substitution matrix] ,

the score of the two aligned sequences can be computed as the sum of substitution scores over the alignment length

$$Score(A, B, alignment) = \sum_{position\_k} s(A^k, B^k)$$

NB: it assume strong independence among the alignment positions

Distance between aligned sequences

Alignments with gaps

```
A:     ALASVLIRLIT--YP
B:     ASAVHL---ITRLYP
```

Deletion and Insertion

Some residues can be inserted or deleted

$$Score(A, B, aligments) = \sum_{nonGapPositions\_k} s(A^k, B^k) + \sigma(3) + \sigma(2)$$

Distance between aligned sequences

$$Score(A,B,aligments)=\sum_{nonGapPositions\_k} s(A^k,B^k)+\sigma(3)+\sigma(2)$$

The gap score is always negative and depends only on its length

Two main possibilities:

LINEAR

$\sigma(n) = -nd$ (each gapped position is equivalent)

AFFINE

$\sigma(n) = -d - (n-1)e$ (*d*: opening, *e*: extension with *d*>*e*)

N.B. All the scores are independent of the position along the sequence

Sequence alignment

Given two sequences, what is the maximal scoring alignment ?

Naïf solution: try all the possible alignments and chose the best scoring

The score of any alignment can be computed with as

$$Score(A,B)=\sum_{i} s(A^{i},B^{i})+\sum_{gap} \sigma(n_{gap})$$

How many possible alignments between to sequences?

Write ALL the possible ungapped alignments between the two sequences

A: tca
B: ga

Score the alignments using the following matrix and the linear gap penalty (d=2)

	A	C	T	G
A	2	-1	-1	0
C		2	0	-1
T			2	-1
G				2

How many possible alignments between two sequences?

Ungapped

Identical to the first

--tca	-tca	tca	tca	tca-	~~tca--~~
ga---	ga--	ga-	-ga	--ga	~~---ga~~
-10	**-7**	**-4**	**-1**	**-6**	

Given two sequences with lengths *m* and *n*, the number of shifts is *m* +*n*

How many possible alignments between to sequences?

Write ALL the possible gapped alignments between the two sequences

A: tca
B: ga

Score the alignments using the following matrix and the linear gap penalty (d=2)

	A	**C**	**T**	**G**
A	2	-1	-1	0
C		2	0	-1
T			2	-1
G				2

How many possible alignments between two sequences?

Gapped

```
--tca   -tca    -tca    -tca    t-ca
ga---   ga--    g-a-    g--a    ga--
gatca   gtaca   gtcaa   gtcaa   tgaca
22111   21211   21121   21112   12211

tca     tca     tc-a    tca     tca-
ga-     g-a     -ga-    -ga     --ga
tgcaa   tgcaa   tcgaa   tcgaa   tcaga
12121   12112   11221   11212   11122
```

The number of possible alignments is equal to the possible ways to intercalatevtwo sequences, preserving the order

Given two sequences with lengths m and n, the number of possible alignments is $(m+n)!/n!m!$

If n=m=80 there are $9 \cdot 10^{42}$ possible alignments !!!!!!!

Solution: to adopt dynamic programming strategies

Needleman-Wunsch
Smith-Waterman

Basic idea of dynamic programming

The complete computation of the alignment scores for all the possible alignments leads to compute the same things many times

ALSKLASPALSAKDLDSPALS
ALSKIADSLAPIKDLSPASLT

ALSKLASPALSAKDLDSPAL-S
ALSKIADSLAPIKDLSPASLT-

The two alignments are equal for most of the length

Scores are summed along the alignment: naif method computes the score for the first part of the alignment is computet two times: BETTER TO STORE AND REUSE IT

Basic idea of dynamic programming

Build the alignment step by step, storing the optimal alignment between substrings

Given the two sequences

ALSKLASPALSAKDLDSPALS, ALSKIADSLAPIKDLSPASLT

the best alignment between the substrings

$$\left\{\begin{array}{l}\mathtt{ALSKLASPA}\\ \mathtt{ALSKIAD}\end{array}\right.$$

is for sure deriving from one of the following possibilities:

$$\left\{\begin{array}{l|}\mathtt{ALSKLASP}\\ \mathtt{ALSKIA}\end{array}\right. + \begin{array}{c}\mathtt{A}\\ \mathtt{D}\end{array} \qquad \left\{\begin{array}{l|}\mathtt{ALSKLASP}\\ \mathtt{ALSKIAD}\end{array}\right. + \begin{array}{c}\mathtt{A}\\ -\end{array} \qquad \left\{\begin{array}{l|}\mathtt{ALSKLASPA}\\ \mathtt{ALSKIA}\end{array}\right. + \begin{array}{c}-\\ \mathtt{D}\end{array}$$

It is the highest scoring one