Introduction to Bioinformatics

Substitution matrices

Jacques van Helden

Jacques.van-Helden@univ-amu.fr
Université d'Aix-Marseille, France
Lab. Technological Advances for Genomics and Clinics
(TAGC, INSERM Unit U1090)
http://tagc.univ-mrs.fr/

FORMER ADDRESS (1999-2011)
Université Libre de Bruxelles, Belgique
Bioinformatique des Génomes et des Réseaux (BiGRe lab)
http://www.bigre.ulb.ac.be/

Substitution matrix

- A ***substitution matrix*** indicates the score associated to each possible pair of aligned residues in an alignment.
  - Each row and each column represent one of the possible residues (4 for DNA, 20 for proteins).
  - The diagonal indicates identities.
  - The lower triangle indicates substitutions.
  - The upper triangle is symmetrical to the lower triangle, and does not need to be displayed.
  - Positive scores indicate that the aligned pair of residue is considered “beneficial” for the alignment. Note that some mismatches might have positive scores in protein alignments (see later).
  - Negative scores are considered as penalties associated to mismatches, and alignment algorithms will try to avoid aligning the pairs of residues.
- One could decide to give a lower cost to A-T substitutions, if we assume that these are more likely to occur in our sequences
- Example: the top matrix represents arbitrarily defined scores for DNA alignment
  - match 2
  - A-T mismatch -1
  - other mismatch -2
- The scoring scheme can be represented as a substitution matrix

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

		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Ala	A	4																			
Arg	R	-1	5																		
Asn	N	-2	0	6																	
Asp	D	-2	-2	1	6																
Cys	C	0	-3	-3	-3	9															
Gln	Q	-1	1	0	0	-3	5														
Glu	E	-1	0	0	2	-4	2	5													
Gly	G	0	-2	0	-1	-3	-2	-2	6												
His	H	-2	0	1	-1	-3	0	0	-2	8											
Ile	I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Scoring an alignment matrix with a substitution matrix

- Let us come back to our previous alignment matrix
- For each cell of the alignment matrix, we compare the residue in sequences A and B, and take the score for this pair of residues in the substitution matrix.

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

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

Substitution counts in 71 groups of aligned proteins (Dayhoff, 1978)

Ala	A																				
Arg	R	30																			
Asn	N	109	17																		
Asp	D	154	0	532																	
Cys	C	33	10	0	0																
Gln	Q	93	120	50	76	0															
Glu	E	266	0	94	831	0	422														
Gly	G	579	10	156	162	10	30	112													
His	H	21	103	226	43	10	243	23	10												
Ile	I	66	30	36	31	17	8	35	0	3											
Leu	L	95	17	37	0	0	75	15	17	40	253										
Lys	K	57	423	322	85	0	147	104	60	23	43	39									
Met	M	29	12	0	0	0	20	7	7	0	57	207	90								
Phe	F	20	7	7	0	0	0	0	1è	20	90	167	0	17							
Pro	P	345	67	20	10	10	93	40	49	50	7	43	43	4	7						
Ser	S	772	137	432	89	117	47	86	450	26	20	32	168	20	40	269					
Thr	T	590	20	169	57	10	37	31	50	14	129	52	200	28	10	73	696				
Trp	W	0	27	3	0	0	0	0	0	3	0	13	0	0	10	0	17	0			
Tyr	Y	20	3	36	0	30	0	10	0	40	13	23	10	0	260	0	22	23	6		
Val	V	365	20	13	17	33	27	37	97	30	661	303	17	77	10	50	43	18§	0	17	
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Hydrophobic		A				C	G				I	L		M		P					V
Aromatic										H					F				W	Y	
Polar				N			Q										S	T		Y	
Basic			R							H			K								
Acidic					D			E													

en rouge: chiffres illisibles

Substitution matrices for proteins

$$s_{i,j} = s_{j,i} = \log_2\left(\frac{f_{i,j}}{\left(f_i f_j\right)}\right)$$

	C	S	T	P	A	G	...
C	11.5						...
S	0.1	2.2					...
T	-0.5	1.5	2.5				...
P	-3.1	0.4	0.1	7.6			...
A	0.5	1.1	0.6	0.3	2.4		...
G	-2.0	0.4	-1.1	-1.6	0.5	1.6	...
...	...	...	...	...	...	...	...

- Margaret Dayhoff (1978) measured the rate of substitutions between each pair of amino acids, in a collection of aligned proteins.
- Scores are calculated as ***log-odds***
  - **Positive values** reflect frequent ("accepted") substitutions, i.e. substitutions that occur more frequently than expected by chance.
  - **Negative values** reflect rare ("unfavourable") mutations, i.e. substitutions that occur less frequently than expected by chance
- The diagonal reflect residue conservation

Reference: Dayhoff et al. (1978). A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*, vol. 5, suppl. 3, 345–352. National Biomedical Research Foundation, Silver Spring, MD, 1978.

PAM scoring matrices

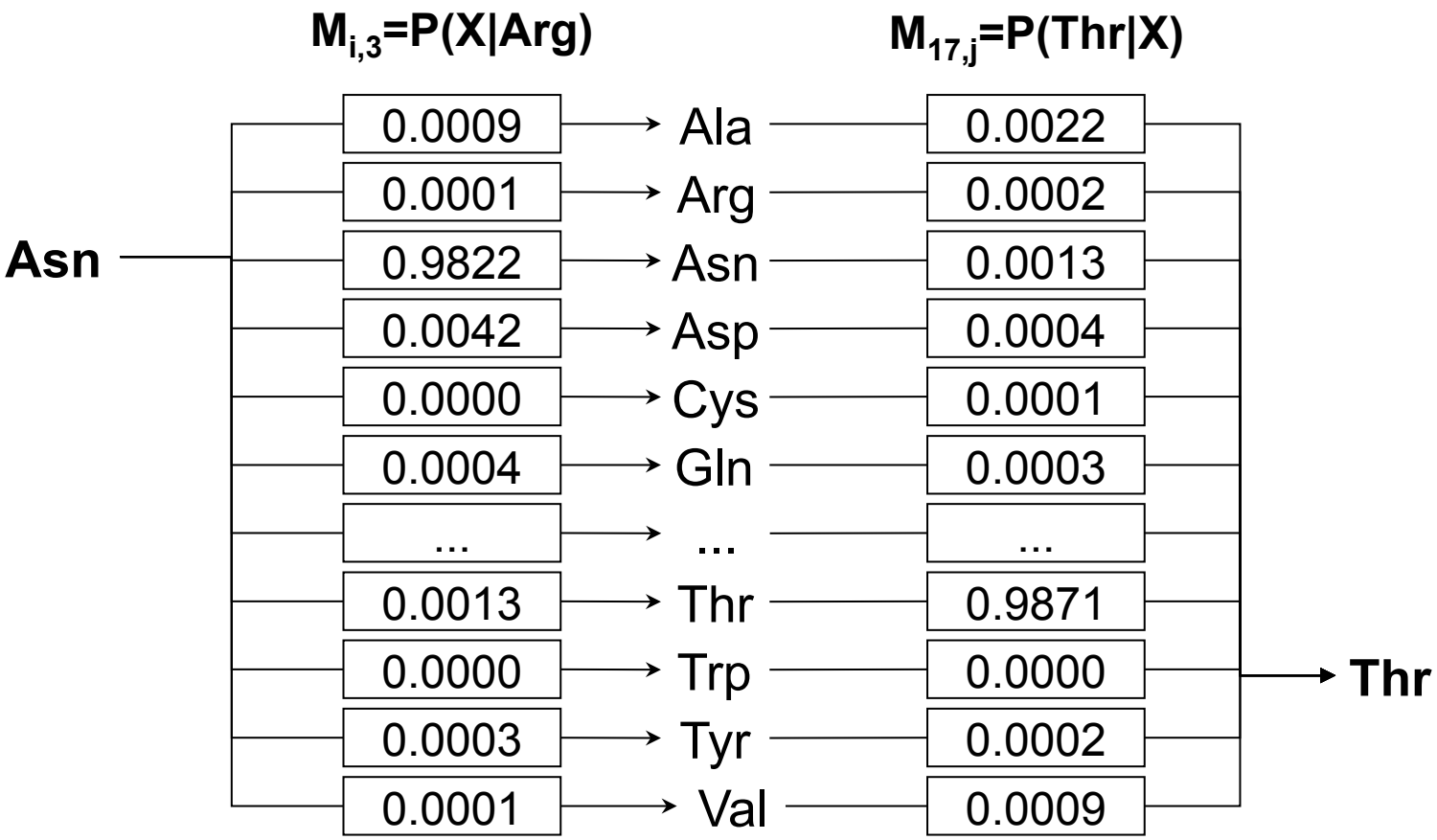
- The alignments used by Dayhoff had ~85% identity
- However, frequencies of substitutions are expected to depend on the rate of divergence between sequences: the number of substitutions increases with time.
- In order to take into account the divergence rate, Margaret Dayhoff calculated a **series of scoring matrices**, each reflecting a certain level of divergence

PAM001	rates of substitutions between amino-acid pairs expected for proteins with an average of 1% substitution per position
PAM050	rates of substitutions between amino-acid pairs expected for proteins with an average of 50% substitution per position
PAM250	250% mutations/position (**note:** a position could mutate several times)

- The substitution matrix must this be chosen according to the relatedness of the sequences to be aligned

Reference: Dayhoff et al. (1978). A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*, vol. 5, suppl. 3, 345–352. National Biomedical Research Foundation, Silver Spring, MD, 1978.

Extrapolation of the PAM series from PAM001

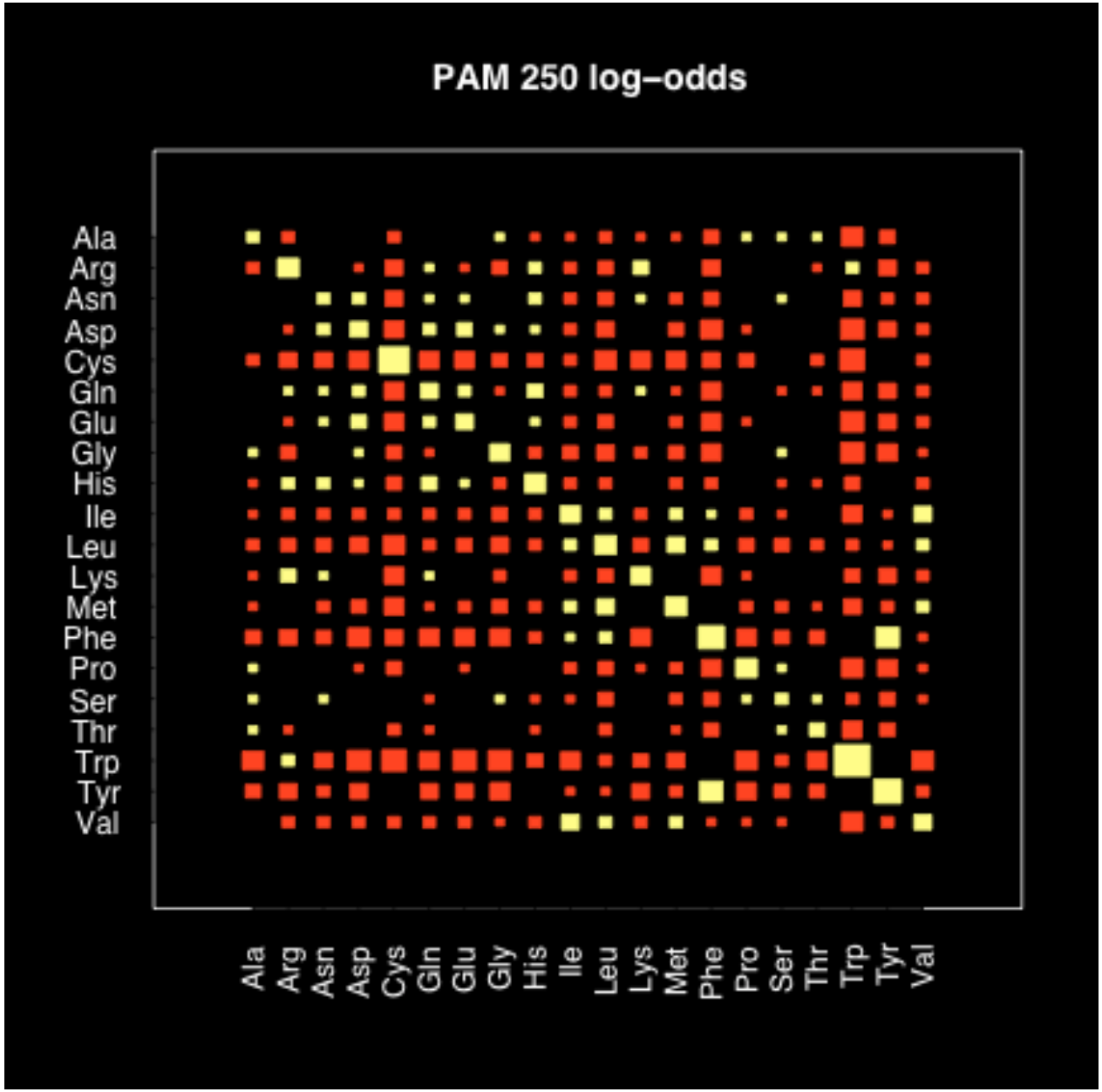


P(Asn -> Thr)= P(Asn -> Ala -> Thr) + P(Asn -> Arg -> Thr) + ... + P(Asn -> Val -> Thr)
= (0.0009)(0.0001) + (0.0001)(0.0002) + ... + (0.0001)(0.009)

PAM250 matrix

Cys	C	12																		
Ser	S	0	2																	
Thr	T	-2	1	3																
Pro	P	-1	1	0	6															
Ala	A	-2	1	1	1	2														
Gly	G	-3	1	0	-1	1	5													
Asn	N	-4	1	0	-1	0	0	2												
Asp	D	-5	0	0	-1	0	1	2	4											
Glu	E	-5	0	0	-1	0	0	1	3	4										
Gln	Q	-5	-1	-1	0	0	-1	1	2	2	4									
His	H	-3	-1	-1	0	-1	-2	2	1	1	3	6								
Arg	R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6							
Lys	K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5						
Met	M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6					
Ile	I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5				
Leu	L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6			
Val	V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4		
Phe	F	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9	
Tyr	Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10
Trp	W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17
		C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
		Cys	Ser	Thr	Pro	Ala	Gly	Asn	Asp	Glu	Gln	His	Arg	Lys	Met	Ile	Leu	Val	Phe	Tyr	Trp
Hydrophobic		C			P	A	G								M	I	L	V			
Aromatic												H							F	Y	W
Polar			S	T				N			Q									Y	
Basic												H	R	K							
Acidic									D	E											

Hinton diagram of the PAM250 matrix



- Yellow boxes indicate positive values (accepted mutations)
- Red boxes indicate negative values (avoided mutations).
- The area of each box is proportional to the absolute value of the log-odds score.

BLOSUM scoring matrices

- Henikoff and Henikoff (1992) analyzed substitution rates on the basis of aligned regions (**blocks**)
- They calculated scoring matrices from blocks with different percentages of protein divergence
- Example:
  - BLOSUM62calculated from blocks with ~62% identity
  - BLOSUM80calculated from blocks with ~80% identity
- When these substitution matrices are used to score sequence alignments, one should always choose the matrix appropriate to the expected percentage of similarity.

Reference: Henikoff, S. and Henikoff, J.G. (1992). Amino acid substitution matrices from protein blocks. PNAS 89:10915-10919.

BLOSUM30

Ala	A	4																							
Arg	R	-1	8																						
Asn	N	0	-2	8																					
Asp	D	0	-1	1	9																				
Cys	C	-3	-2	-1	-3	17																			
Gln	Q	1	3	-1	-1	-2	8																		
Glu	E	0	-1	-1	1	1	2	6																	
Gly	G	0	-2	0	-1	-4	-2	-2	8																
His	H	-2	-1	-1	-2	-5	0	0	-3	14															
Ile	I	0	-3	0	-4	-2	-2	-3	-1	-2	6														
Leu	L	-1	-2	-2	-1	0	-2	-1	-2	-1	2	4													
Lys	K	0	1	0	0	-3	0	2	-1	-2	-2	-2	4												
Met	M	1	0	0	-3	-2	-1	-1	-2	2	1	2	2	6											
Phe	F	-2	-1	-1	-5	-3	-3	-4	-3	-3	0	2	-1	-2	10										
Pro	P	-1	-1	-3	-1	-3	0	1	-1	1	-3	-3	1	-4	-4	11									
Ser	S	1	-1	0	0	-2	-1	0	0	-1	-1	-2	0	-2	-1	-1	4								
Thr	T	1	-3	1	-1	-2	0	-2	-2	-2	0	0	-1	0	-2	0	2	5							
Trp	W	-5	0	-7	-4	-2	-1	-1	1	-5	-3	-2	-2	-3	1	-3	-3	-5	20						
Tyr	Y	-4	0	-4	-1	-6	-1	-2	-3	0	-1	3	-1	-1	3	-2	-2	-1	5	9					
Val	V	1	-1	-2	-2	-2	-3	-3	-3	-3	4	1	-2	0	1	-4	-1	1	-3	1	5				
Asx	B	0	-2	4	5	-2	-1	0	0	-2	-2	-1	0	-2	-3	-2	0	0	-5	-3	-2	5			
Glx	Z	0	0	-1	0	0	4	5	-2	0	-3	-1	1	-1	-4	0	-1	-1	-1	-2	-3	0	4		
Unkown	X	0	-1	0	-1	-2	0	-1	-1	-1	0	0	0	0	-1	-1	0	0	-2	-1	0	-1	0	-1	
End	*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1
		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	Asx	Glx	Unk	End
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
Hydrophobic		A				C			G		I	L		M		P					V				
Aromatic										H					F				W	Y					
Polar				N			Q										S	T		Y					
Basic			R							H			K												
Acidic					D			E																	

BLOSUM62

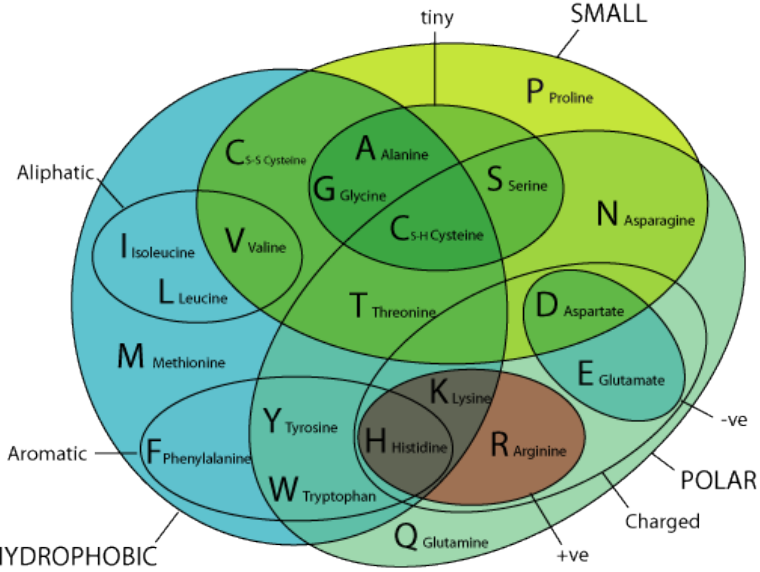
Ala	A	4																							
Arg	R	-1	5																						
Asn	N	-2	0	6																					
Asp	D	-2	-2	1	6																				
Cys	C	0	-3	-3	-3	9																			
Gln	Q	-1	1	0	0	-3	5																		
Glu	E	-1	0	0	2	-4	2	5																	
Gly	G	0	-2	0	-1	-3	-2	-2	6																
His	H	-2	0	1	-1	-3	0	0	-2	8															
Ile	I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4														
Leu	L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4													
Lys	K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5												
Met	M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5											
Phe	F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6										
Pro	P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7									
Ser	S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4								
Thr	T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5							
Trp	W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11						
Tyr	Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7					
Val	V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4				
Asx	B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4			
Glx	Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4		
Unkown	X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	
End	*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1
		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	Asx	Glx	Unk	End
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
Hydrophobic		A				C			G		I	L		M		P					V				
Aromatic										H					F				W	Y					
Polar				N			Q										S	T		Y					
Basic			R							H			K												
Acidic					D			E																	

BLOSUM80

Ala	A	5																							
Arg	R	-2	6																						
Asn	N	-2	-1	6																					
Asp	D	-2	-2	1	6																				
Cys	C	-1	-4	-3	-4	9																			
Gln	Q	-1	1	0	-1	-4	6																		
Glu	E	-1	-1	-1	1	-5	2	6																	
Gly	G	0	-3	-1	-2	-4	-2	-3	6																
His	H	-2	0	0	-2	-4	1	0	-3	8															
Ile	I	-2	-3	-4	-4	-2	-3	-4	-5	-4	5														
Leu	L	-2	-3	-4	-5	-2	-3	-4	-4	-3	1	4													
Lys	K	-1	2	0	-1	-4	1	1	-2	-1	-3	-3	5												
Met	M	-1	-2	-3	-4	-2	0	-2	-4	-2	1	2	-2	6											
Phe	F	-3	-4	-4	-4	-3	-4	-4	-4	-2	-1	0	-4	0	6										
Pro	P	-1	-2	-3	-2	-4	-2	-2	-3	-3	-4	-3	-1	-3	-4	8									
Ser	S	1	-1	0	-1	-2	0	0	-1	-1	-3	-3	-1	-2	-3	-1	5								
Thr	T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-2	-1	-1	-2	-2	1	5							
Trp	W	-3	-4	-4	-6	-3	-3	-4	-4	-3	-3	-2	-4	-2	0	-5	-4	-4	11						
Tyr	Y	-2	-3	-3	-4	-3	-2	-3	-4	2	-2	-2	-3	-2	3	-4	-2	-2	2	7					
Val	V	0	-3	-4	-4	-1	-3	-3	-4	-4	3	1	-3	1	-1	-3	-2	0	-3	-2	4				
Asx	B	-2	-2	4	4	-4	0	1	-1	-1	-4	-4	-1	-3	-4	-2	0	-1	-5	-3	-4	4			
Glx	Z	-1	0	0	1	-4	3	4	-3	0	-4	-3	1	-2	-4	-2	0	-1	-4	-3	-3	0	4		
Unkown	X	-1	-1	-1	-2	-3	-1	-1	-2	-2	-2	-2	-1	-1	-2	-2	-1	-1	-3	-2	-1	-2	-1	-1	
End	*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1
		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	Asx	Glx	Unk	End
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
Hydrophobic		A				C			G		I	L		M		P					V				
Aromatic										H					F				W	Y					
Polar				N			Q										S	T		Y					
Basic			R							H			K												
Acidic					D			E																	

BLOSUM62 – Amino acid properties

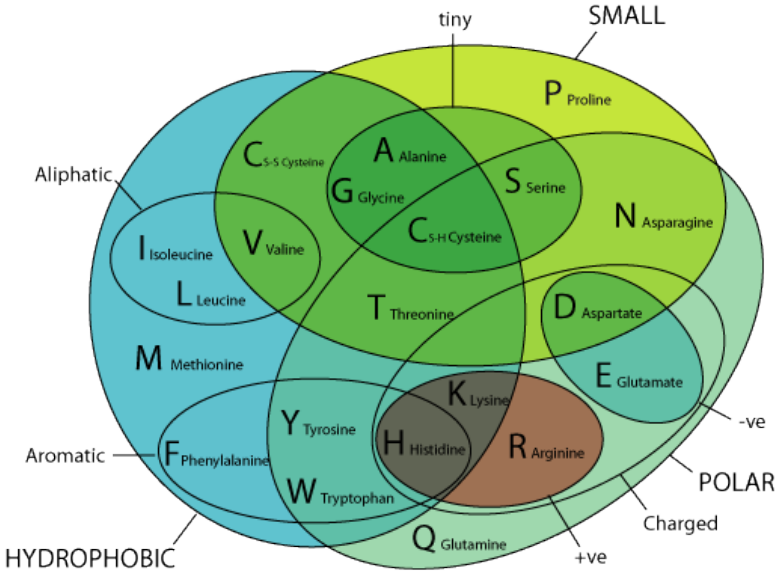
Ala	A	4																						
Arg	R	-1	5																					
Asn	N	-2	0	6																				
Asp	D	-2	-2	1	6																			
Cys	C	0	-3	-3	-3	9																		
Gln	Q	-1	1	0	0	-3	5																	
Glu	E	-1	0	0	2	-4	2	5																
Gly	G	0	-2	0	-1	-3	-2	-2	6															
His	H	-2	0	1	-1	-3	0	0	-2	8														
Ile	I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4													
Leu	L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4												
Lys	K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5											
Met	M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5										
Phe	F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6									
Pro	P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7								
Ser	S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4							
Thr	T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5						
Trp	W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11					
Tyr	Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7				
Val	V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4			
Asx	B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4		
Glx	Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4	
Unkown	X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1
End	*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1
		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	Asx	Glx	Unk	End
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
Hydrophobic		A				C			G		I	L		M		P					V				
Aromatic										H					F				W	Y					
Polar				N			Q										S	T		Y					
Basic			R							H			K												
Acidic					D			E																	



BLOSUM62 - substitutions between acidic residues

		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	Asx	Glx	Unk	End
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
Ala	A	4																							
Arg	R	-1	5																						
Asn	N	-2	0	6																					
Asp	D	-2	-2	1	6																				
Cys	C	0	-3	-3	-3	9																			
Gln	Q	-1	1	0	0	-3	5																		
Glu	E	-1	0	0	2	-4	2	5																	
Gly	G	0	-2	0	-1	-3	-2	-2	6																
His	H	-2	0	1	-1	-3	0	0	-2	8															
Ile	I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4														
Leu	L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4													
Lys	K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5												
Met	M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5											
Phe	F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6										
Pro	P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7									
Ser	S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4								
Thr	T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5							
Trp	W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11						
Tyr	Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7					
Val	V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4				
Asx	B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4			
Glx	Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4		
Unkn	X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	
End	*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

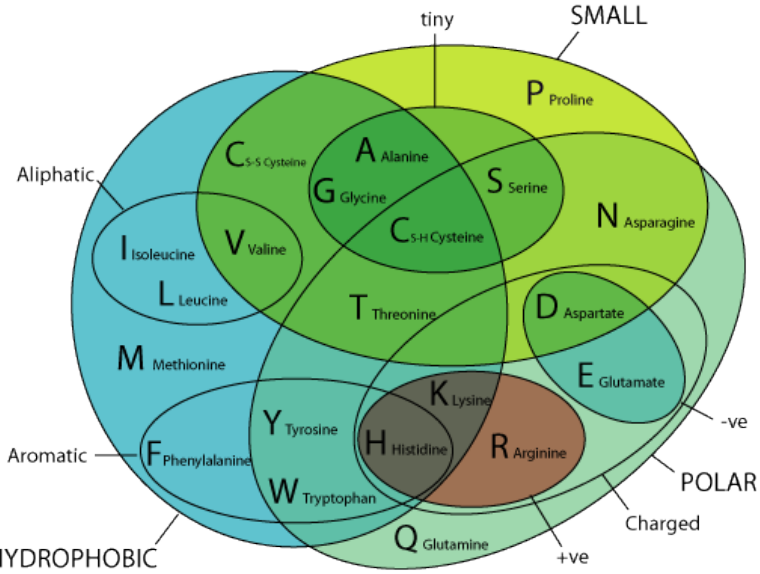
Hydrophobic	A				C			G		I	L		M		P					V				
Aromatic									H					F				W	Y					
Polar			N			Q										S	T		Y					
Basic		R							H			K												
Acidic				D			E																	



BLOSUM62- substitutions between basic residues

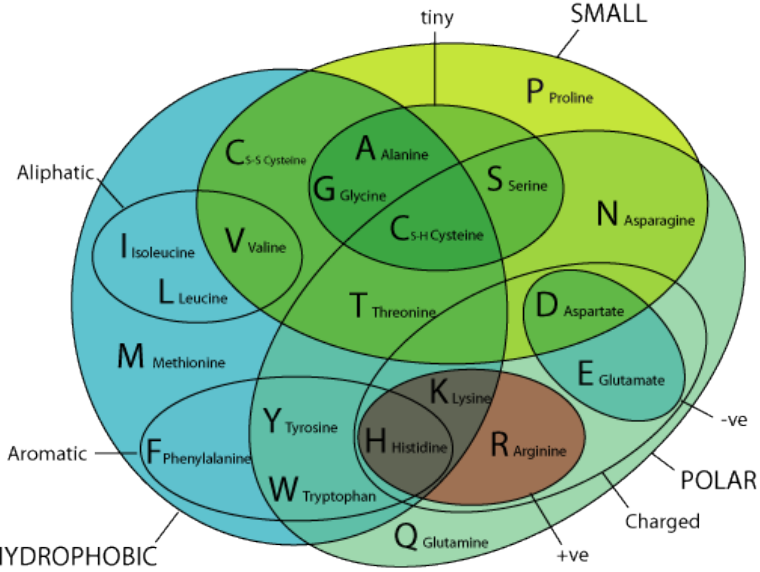
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
Ala	A	4																							
Arg	R	-1	5																						
Asn	N	-2	0	6																					
Asp	D	-2	-2	1	6																				
Cys	C	0	-3	-3	-3	9																			
Gln	Q	-1	1	0	0	-3	5																		
Glu	E	-1	0	0	2	-4	2	5																	
Gly	G	0	-2	0	-1	-3	-2	-2	6																
His	H	-2	0	1	-1	-3	0	0	-2	8															
Ile	I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4														
Leu	L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4													
Lys	K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5												
Met	M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5											
Phe	F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6										
Pro	P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7									
Ser	S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4								
Thr	T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5							
Trp	W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11						
Tyr	Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7					
Val	V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4				
	B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4			
	Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4		
	X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	
	*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

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



BLOSUM62 - substitutions between aromatic residues

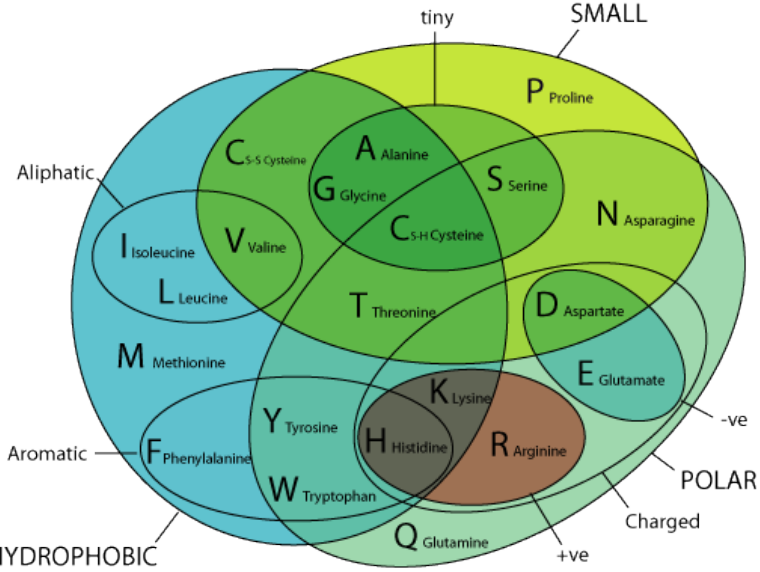
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
Ala	A	4																							
Arg	R	-1	5																						
Asn	N	-2	0	6																					
Asp	D	-2	-2	1	6																				
Cys	C	0	-3	-3	-3	9																			
Gln	Q	-1	1	0	0	-3	5																		
Glu	E	-1	0	0	2	-4	2	5																	
Gly	G	0	-2	0	-1	-3	-2	-2	6																
His	H	-2	0	1	-1	-3	0	0	-2	8															
Ile	I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4														
Leu	L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4													
Lys	K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5												
Met	M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5											
Phe	F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6										
Pro	P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7									
Ser	S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4								
Thr	T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5							
Trp	W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11						
Tyr	Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7					
Val	V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4				
	B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4			
	Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4		
	X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	
	*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1
		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val				
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
Hydrophobic		A				C			G		I	L		M		P					V				
Aromatic										H					F				W	Y					
Polar				N			Q										S	T		Y					
Basic			R							H			K												
Acidic					D			E																	



BLOSUM62 - substitutions between polar residues

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
Ala	A	4																							
Arg	R	-1	5																						
Asn	N	-2	0	6																					
Asp	D	-2	-2	1	6																				
Cys	C	0	-3	-3	-3	9																			
Gln	Q	-1	1	0	0	-3	5																		
Glu	E	-1	0	0	2	-4	2	5																	
Gly	G	0	-2	0	-1	-3	-2	-2	6																
His	H	-2	0	1	-1	-3	0	0	-2	8															
Ile	I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4														
Leu	L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4													
Lys	K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5												
Met	M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5											
Phe	F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6										
Pro	P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7									
Ser	S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4								
Thr	T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5							
Trp	W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11						
Tyr	Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7					
Val	V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4				
	B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4			
	Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4		
	X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	
	*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

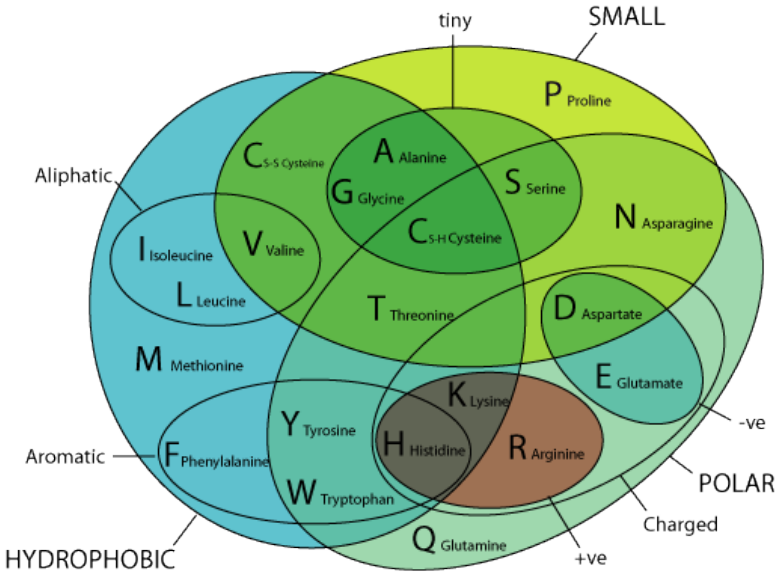
	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	B	Z	X	*
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
Hydrophobic	A				C			G		I	L		M		P					V				
Aromatic									H					F				W	Y					
Polar			N			Q										S	T		Y					
Basic		R							H			K												
Acidic				D			E																	



BLOSUM62 - substitutions between hydrophobic residues

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
Ala	A	4																							
Arg	R	-1	5																						
Asn	N	-2	0	6																					
Asp	D	-2	-2	1	6																				
Cys	C	0	-3	-3	-3	9																			
Gln	Q	-1	1	0	0	-3	5																		
Glu	E	-1	0	0	2	-4	2	5																	
Gly	G	0	-2	0	-1	-3	-2	-2	6																
His	H	-2	0	1	-1	-3	0	0	-2	8															
Ile	I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4														
Leu	L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4													
Lys	K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5												
Met	M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5											
Phe	F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6										
Pro	P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7									
Ser	S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4								
Thr	T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5							
Trp	W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11						
Tyr	Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7					
Val	V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4				
	B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4			
	Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4		
	X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	
	*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
Hydrophobic	A				C			G		I	L		M		P					V				
Aromatic									H					F				W	Y					
Polar			N			Q										S	T		Y					
Basic		R							H			K												
Acidic				D			E																	



Substitution matrices - summary

- Different substitution scoring matrices have been established
  - Residue categories (Phylip)
  - PAM (Dayhoff, 1979).
    - PAM means “Percent Accepted Mutations”
  - BLOSUM (Henikoff & Henikoff, 1992).
    - BLOSUM means “Block sum”.
- Substitution matrices allow to detect similarities between more distant proteins than what would be detected with the simple identity of residues.
- The matrix must be chosen carefully, depending on the expected rate of conservation between the sequences to be aligned.
- Beware
  - With **PAM** matrices
    - the score indicates the percentage of substitution per position
      -> **higher numbers** are appropriate for **more distant** proteins
  - With **BLOSUM** matrices
    - the score indicates the percentage of conservation
      -> **higher numbers** are appropriate for **more conserved** proteins

Scoring an alignment with a substitution matrix

Ala	A	4																		
Arg	R	-1	5																	
Asn	N	-2	0	6																
Asp	D	-2	-2	1	6															
Cys	C	0	-3	-3	-3	9														
Gln	Q	-1	1	0	0	-3	5													
Glu	E	-1	0	0	2	-4	2	5												
Gly	G	0	-2	0	-1	-3	-2	-2	6											
His	H	-2	0	1	-1	-3	0	0	-2	8										
Ile	I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4									
Leu	L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4								
Lys	K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5							
Met	M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5						
Phe	F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6					
Pro	P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7				
Ser	S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4			
Thr	T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5		
Trp	W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	
Tyr	Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7
Val	V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4
		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

$$S = \sum_{i=1}^{L} s_{r_{1,i}\, r_{2,i}}$$

- The substitution matrix can be used to assign a score to a pair-wise alignment.
- The score of the alignment is the sum, over all the aligned positions (i *from 1 to* L), of the scores of the pairs of residues ($r_{1,I}$ and $r_{2,I}$).
- Gaps are treated by subtracting a penalty, with two parameters:
  - Gap opening (***go***) penalty
    - Typical values : between -10 and -15
  - Gap extension (***ge***) penalty
    - Typical values: between -0.5 and -2

i	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***13***	***14***	***15***	***16***	***17***	***18***	***19***	***20***	***21***
	R	**L**	**A**	**S**	**V**	**E**	**T**	**D**	**M**	**P**	**–**	**–**	**–**	**–**	**–**	**L**	**T**	**L**	**R**	**Q**	**H**
	T	**L**	**T**	**S**	**L**	**Q**	**T**	**T**	**L**	**K**	**N**	**L**	**K**	**E**	**M**	**A**	**H**	**L**	**G**	**T**	**H**

S

Scoring an alignment with a substitution matrix

Ala	A	4																		
Arg	R	-1	5																	
Asn	N	-2	0	6																
Asp	D	-2	-2	1	6															
Cys	C	0	-3	-3	-3	9														
Gln	Q	-1	1	0	0	-3	5													
Glu	E	-1	0	0	2	-4	2	5												
Gly	G	0	-2	0	-1	-3	-2	-2	6											
His	H	-2	0	1	-1	-3	0	0	-2	8										
Ile	I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4									
Leu	L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4								
Lys	K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5							
Met	M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5						
Phe	F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6					
Pro	P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7				
Ser	S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4			
Thr	T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5		
Trp	W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	
Tyr	Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7
Val	V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4
		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

$$S = \sum_{i=1}^{L} s_{r_{1,i}\, r_{2,i}}$$

- The substitution matrix can be used to assign a score to a pair-wise alignment.
- The score of the alignment is the sum, over all the aligned positions (i *from 1 to* L), of the scores of the pairs of residues ($r_{1,I}$ and $r_{2,I}$).
- Gaps are treated by subtracting a penalty, with two parameters:
  - Gap opening (***go***) penalty
    - Typical values : between -10 and -15
  - Gap extension (***ge***) penalty
    - Typical values: between -0.5 and -2

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	
	R	L	A	S	V	E	T	D	M	P	–	–	–	–	–	L	T	L	R	Q	H	
	.	|	.	|	:	:	|	.	:	.	go	ge	ge	ge	ge	.	.	|	.	.	|	
	T	L	T	S	L	Q	T	T	L	K	N	L	K	E	M	A	H	L	G	T	H	
S	–1	+4	+0	+4	+1	+2	+5	–1	+2	–1	–10	–1	–1	–1	–1	–1	–2	+4	–2	–1	+8	= 7

Bibliography

- Substitution matrices
  - PAM series
    - Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. (1978). A model of evolutionary change in proteins. Atlas of Protein Sequence and Structure 5, 345--352.
  - BLOSUM substitution matrices
    - Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 89, 10915-9.
  - Gonnet matrices, built by an iterative procedure
    - Gonnet, G. H., Cohen, M. A. & Benner, S. A. (1992). Exhaustive matching of the entire protein sequence database. Science 256, 1443-5. 1.