Matrices de substitution

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://jacques.van-helden.perso.luminy.univ-amu.fr/

Mismatches et substitutions

- Quand on aligne deux ou plusieurs séquences, on observe souvent des résidus différents à la même position de l'alignement (« ***mismatches*** »), qui reflètent vraisemblablement qu'une substitution est survenue au sein de l'une des séquences ancestrales.
- On constate que certaines substitutions sont plus fréquentes que d'autres.
- Dans les séquences protéiques, les substitutions fréquentes correspondent généralement à des acides aminés qui partagent des propriétés chimiques (hydrophobie, polarité) ou stérique (encombrement du radical).
- Sur base de cette observation, on construit des ***matrices de substitutions*** qui serviront ensuite à pondérer les « *mismatches* » lors de l'alignement de nouvelles séquences.

Construction des matrices de substitutions

La série de matrices PAM (« point accepted mutation ») construite à partir d'alignements par paire

Exemple d'alignement par paires

```
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
# Length: 482
# Identity:     133/482 (27.6%)
# Similarity:   205/482 (42.5%)
# Gaps:          85/482 (17.6%)
# Score: 353.5
metL              16 KPGGSSLADVKCYLRVAGIMAEYSQPDDMMVVSAAGSTTWQLIWNLK-LS     64
                     ||||:|:|||.....|.|.:.....:. .::|:||:...|.|.|...: |.
lysC               8 KPGGTSVADFDAMNRSADIVLSDANV-RLVVLSASAGITNLLVALAEGLE     56

metL              65 QTDRLSAHQVQQTLRRYQCDLISGL----LPAEEADSLISAFVSDLERLA     110
                     ..:|.  :....:|..|..::..|    :..||.:.|:.. ::.|...|
lysC              57 PGERF---EKLDAIRNIQFAILERLRYPNVIREEIERLLEN-ITVLAEAA     102

metL             111 ALLDSGINDAVYAEVVGHGEVWSARLMSAVLNQQGLPAAWLDAREFLRA-     159
                     ||..|   .|:..|:|.|||:.|..|...:|.:.:.|.|.|.|:.:|.
lysC             103 ALATS---PALTDELVSHGELMSTLLFVEILRERDVQAQWFDVRKVMRTN     149

metL             160 ERAAQPQVDEGLSYPLLQQLLVQHPGKRLVVT-GFISRNNAGETVLLGRN     208
                     :|..:.:|.|......|....|:....:.||:| |||...|.|.|..|||.
lysC             150 DRFGRAEPDIAALAELAALQLLPRLNEGLVITQGFIGSENKGRTTTLGRG     199

metL             209 GSDYSATQIGALAGVSRVTIWSDVAGVYSADPRKVKDACLLPLLRLDEAS     258
                     ||||:|..:.......|||.|:||.|:|:.|||.|..|...:.:....||:
lysC             200 GSDYTAALLAEALHASRVDIWTDVPGIYTTDPRVVSAAKRIDEIAFAEAA     249

metL             259 ELARLAAPVLHARTLQPVSGSEIDLQLRCSYTPDQGSTRI---------E     299
                     |:|...|.|||..|.|...|:|.:.:..|..|..|.|.:          .
lysC             250 EMATFGAKVLHPATLLPAVRSDIPVFVGSSKDPRAGGTLVCNKTENPPLF     299

metL             300 RVLASGTGARIVTSHDDVCLIEPQVPASQDFKLAHKEIDQILKRAQVRPL     349
                     |.||.....::|.|       ..:..|:.| ||  |:..||.|
lysC             300 RALALRRNQTLLTLH------SLNMLHSRGF-LA--EVPGILAR------     334

metL             350 AVGVHNDRQLLQFCYTSEVA--------------DSAL--KILDEAGLPG     383
                        ||..  :....||||:              |:.|  .:|.|.....
lysC             335 ----HNIS--VDLITTSEVSVALTLDTTGSTSTGDTLLTQSLLMELSALC     378

metL             384 ELRLRQGLALVAMVGAGVTR------------NPLHCHRFWQQLKGQPVE     421
```

- La figure représente l'alignement de deux séquences peptidiques.
- Les *barres verticales* indiquent les *identités*.
- Les gaps sont marqués par des traits d'union.
- Les doubles points indiquent des substitutions qu'on retrouve souvent dans les alignements (« point accepted mutations).
- Les simples points indiquent les substitutions rares et celes qui ne sont pas spécialement fréquentes.

Occurrences de substitutions dans 71 groupes de protéines alignées (Dayhoff, 1978)

- En 1978, Margret Dayhoff réalise des alignements de séquences protéiques (71 groupes de protéines), et compte le nombre de substitutions et d'identités entre chaque paire d'acides aminés.
- Elle obtient les comptages représentés dans la matrice ci-dessous.
- Elle se sert ensuite de cette matrice de comptages pour dériver des matrices de « point accepted mutations » (PAM).

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													

Substitution matrices for proteins

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

	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) a mesuré les taux de substitutions entre chaque paire d'acides aminés, dans une collection de 71 alignements de paires de protéines.
- A partir des comptages bruts, elle dérive un score de ***log-odds***
  - f_i, f_j: *fréquences des résidus i et j, resp.*
  - $f_{i,j}$: *fréquence de la substitution i <-> j*
  - **Les valeurs positives** indiquent des substitutions fréquentes ("acceptées"), càd des substitutions observées plus fréquemment que ce à quoi l'on s'attendrait par hasard.
  - **Les valeurs négatives** indiquent les mutations rares, càd celles qu'on observe moins fréquemment que ce à quoi l'on s'attendrait par hasard. Ce taux inférieur est interprété comme un indice du contre-sélection, suggérant que ces mutations sont généralement défavorables pour la fonction de la protéine.
- La diagonale reflète le taux de conservation des résidus. Notons que certains résidus rares ont un score de conservation très important: le score de conservation n'est pas proportionnel à la fréquence.

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.

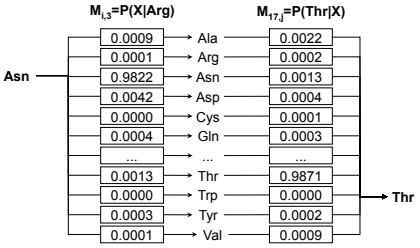
PAM scoring matrices

- Les alignements réalisés par Margret Dayhoff en 1987 comportaient un taux moyen d'identité de ~85%.
- Cependant, on s'attend à ce que les fréquences de substitutions dépendent du degré de divergence entre séquences, car leur nombre augmente avec le temps.
- Pour prendre en compte le taux de divergence, Margret Dayhoof a calculé une ***série de matrices de score***, reflétant chacune un certain taux de substitutions.
  - PAM001 — taux de substitutions entre acides aminés au terme d'un temps évolutif donnant lieu à ~1% de substitutions par position.
  - PAM050 — taux de substitutions entre acides aminés au terme d'un temps évolutif donnant lieu à ~50% de substitutions par position.
  - PAM250 — idem avec 250% mutations/position (**note:** une même position peut faire l'objet de plusieurs mutations successives)
- Quand on fait un alignement, on doit choisir l'une des matrices de cette série, en tenant compte du taux de différences entre les deux séquences qu'on veut aligner.

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 de la série de matrices PAM à partir de la PAM001

- Exemple: si l'on dispose de la matrice PAM001 (temps évolutif donnant ~1% de substitutions/position), on peut calculer la probabilité de substitution de l'Asn à la Thr en 2 unités temporelles (PAM002) en calculant la probabilité de chaque « trajet » de 2 substitutions.



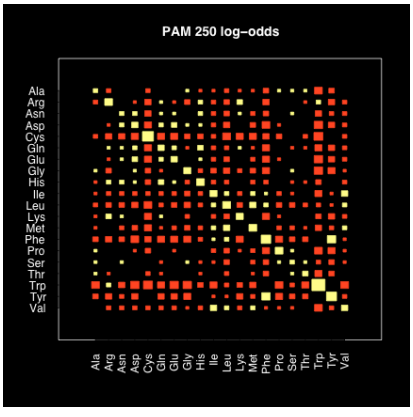
P(Asn -> Thr)= P(Asn -> Ala -> Thr) + P(Asn -> Arg -> Thr) + ... + P(Asn -> Val -> Thr)
= (0.0009)(0.0001) + (0.0001)(0.0002) + ... + (0.0001)(0.009)

PAM250 matrix

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

- La PAM250 est appropriée pour les alignements entre séquences très éloignées.
- Notes
  - La diagonale est constituée de scores positifs, qui reflètent la conservation.
  - les autres scores élevés correspondent souvent à des acides aminés partageant des propriétés physico-chimiques.

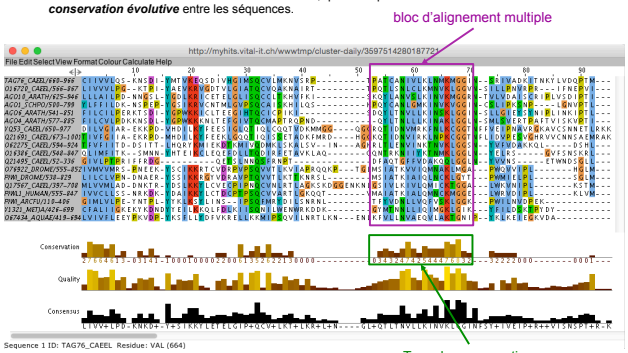
Hinton diagram of the PAM250 matrix



- 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.

La série BLOSUM matrices de substitutions construites à partir de blocs conservés

- Henikoff and Henikoff (1992) ont analysé les fréquences de substitutions dans des ***blocs d'alignements multiples*** générés à partir d'un grand nombre de familles de protéines (**blocks**).
- Ils en ont dérivé la série de matrices « ***BLOSUM*** », qui correspondent à des taux différents de ***conservation évolutive*** entre les séquences.



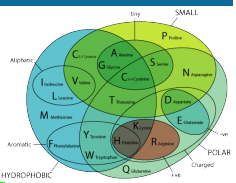
BLOSUM scoring matrices

- Henikoff and Henikoff (1992) ont analysé les fréquences de substitutions dans des blocs d'alignements multiples générés à partir d'un grand nombre de familles de protéines (**blocks**)
- Ils en ont dérivé la série de matrices « ***BLOSUM*** », qui correspondent à des taux différents de divergence évolutive entre les séquences.
- Exemples
  - La matrice *BLOSUM62* a été calculée sur des blocs de >=62% d'identité
  - La matrice *BLOSUM80* a été calculée sur des blocs de >=80% d'identité
- Quand on utilise les matrices BLOSUM pour aligner des séquences, on devrait systématiquement ***choisir la matrice la plus adéquate***, en fonction du pourcentage de similarité.
- Le problème est qu'avant de réaliser l'alignement, on connaît pas ce pourcentage. Comment résoudre cette circularité ?
  - On réalise un premier alignement avec une matrice « moyenne » (BLOSUM62).
  - On observe le % d'identité dans cet alignement.
  - On choisit alors la matrice dont l'indice est le plus proche de ce taux
  - On refait l'alignement avec la nouvelle matrice (sauf s'il s'agit de celle de départ).
- Exemples:
  - L'alignement présente 65.2% d'identité -> le premier alignement avec BLOSUM62 était correct.
  - L'alignement présente 28.4% d'identité -> on refait l'alignement avec BLOSUM30.
  - L'alignement présente 81.5% d'identité -> on refait l'alignement avec BLOSUM80.

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

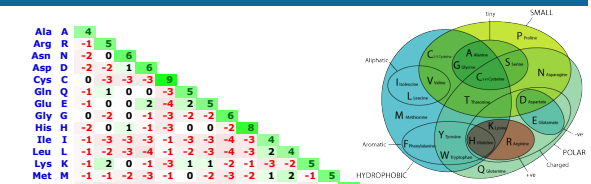
BLOSUM30

		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	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	-5	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	1	0	-3	-2	-1	-1	-2	1	2	2	6												
Phe	F	-2	-1	-1	-5	-3	-3	-4	-3	-3	0	2	-1	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	0	0	-1	0	-2	0	2	5								
Trp	W	-5	0	-7	-4	-2	-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	-4	0	-1	-1	-1	-2	-3	0	4			
Unkown	X	0	-1	0	-1	-2	0	-1	-1	-1	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																	



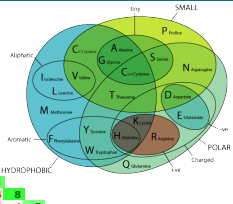
BLOSUM62

		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	-1	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	H	I	L		M	F	P					V				
Aromatic										H					F				W	Y					
Polar				N			Q										S	T		Y					
Basic			R							H			K												
Acidic					D			E																	



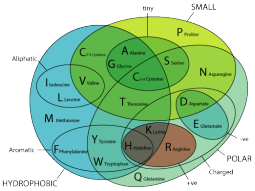
BLOSUM80

		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	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	-3	-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	-1	-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	-4	-4	-1	-3	-4	-2	0	-1	-5	-3	-4	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																	

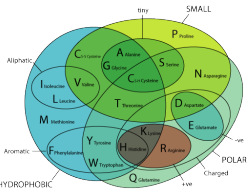


BLOSUM62 – Amino acid properties

		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	-1	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	H	I	L		M	F	P			W	Y	V				
Aromatic										H					F				W	Y					
Polar				N			Q										S	T		Y					
Basic			R							H			K												
Acidic					D			E																	



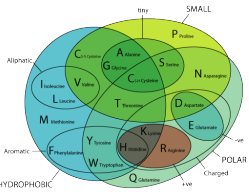
BLOSUM62 - substitutions between acidic residues



	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																							
R	-1	5																						
N	-2	0	6																					
D	-2	-2	1	6																				
C	0	-3	-3	-3	9																			
Q	-1	1	0	0	-3	5																		
E	-1	0	0	2	-4	2	5																	
G	0	-2	0	-1	-3	-2	-2	6																
H	-2	0	1	-1	-3	0	0	-2	8															
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4														
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4													
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5												
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	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4								
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5							
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-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	-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																	

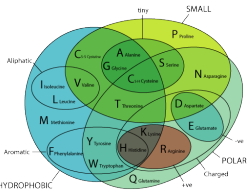
BLOSUM62- substitutions between basic residues



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

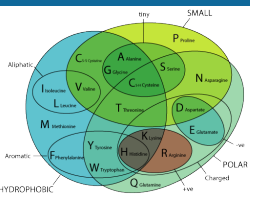
BLOSUM62 - substitutions between aromatic residues



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

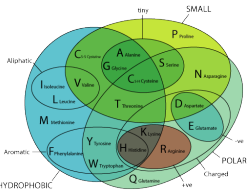
BLOSUM62 - substitutions between polar residues



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

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	*
A	4																							
R	-1	5																						
N	-2	0	6																					
D	-2	-2	1	6																				
C	0	-3	-3	-3	9																			
Q	-1	1	0	0	-3	5																		
E	-1	0	0	2	-4	2	5																	
G	0	-2	0	-1	-3	-2	-2	6																
H	-2	0	1	-1	-3	0	0	-2	8															
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4														
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4													
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5												
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	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4								
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5							
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-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	-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

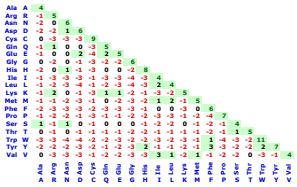
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
Hydrophobic	A				C			G		I	L		M		P					V				
Aromatic									H					F				W	Y					
Polar			N			Q										S	T		Y					
Basic		R							H			K												
Acidic				D			E																	

Utilisation des matrices de substitution pour mesurer la qualité d'un alignement

Matrices de substitutions

- Une ***matrice de substitution*** associe un score à chaque paire de résidus qu'on peut trouver dans un alignement.
  - Chaque ligne et chaque colonne représente l'un des résidus (4 nucléotides, 20 acide aminés).
  - La diagonale correspond aux identités.
  - Le triangle inférieur correspond à des substitutions.
  - Le triangle supérieur est symétrique au triangle inférieur, il n'est pas nécessaire d'indiquer les nombres.
- Les ***scores négatifs*** sont considérés comme des pénalités associées à certaines substitutions qu'on n'observe que rarement dans les alignements. Les algorithmes d'alignements tenteront donc d'éviter ces substitutions.
- Les ***scores positifs*** correspondent à des substitutions qu'on observe plus souvent que prévu, dans les alignements d'un grand nombre de séquences. Ceci suggère que ces substitutions particulières sont moins dommageable que d'autres, et on les qualifie donc de « ***substitutions conservatives*** » ou encore de « ***mutations ponctuelles acceptées*** » (***PAM***).
- Au sein d'un alignement, le terme ***similarité*** désigne les positions où se superposent des résidus ayant un score positif dans la matrice de substitution (identité ou substitution conservative).

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



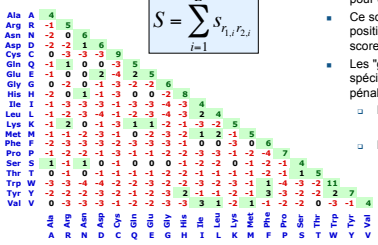
Matrices de substitutions nucléotidiques

- Pour les séquences nucléotidiques, on utilise généralement une pénalité identique pour toute les substitutions.
- Cependant, on pourrait décider d'assigner un coût plus faible à certaines substitutions (par exemple A<—>T) si l'on considère qu'elles ont plus de chance d'être observées dans des alignements (dans certains génomes, les résidus A et T sont deux fois plus fréquents que les C et G).
- Exemple: la matrice ci-jointe représente des scores définis de façon arbitraire
  - Identité +2
  - Substitution A-T -1
  - Autres substit. -2

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

Utilisation d'une matrice de substitution pour calculer le score d'un alignement

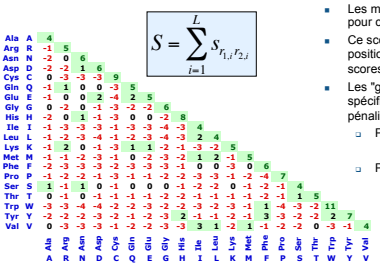
$$S = \sum_{i=1}^{L} s_{r_{1,i} r_{2,i}}$$



- Les matrices de substitution sont utilisées pour calculer le score d'un alignement.
- Ce score est la somme, pour toutes les positions de l'alignement (i *de 1 à* L) , des scores des paires de résidus ($r_{1,i}$ et $r_{2,i}$).
- Les "gaps" sont traités par une règle spécifique reposant sur deux paramètres de pénalité:
  - Pénalité d'ouverture de gap (***go***)
    - Valeurs typiques: entre -10 et -15
  - Pénalité d'extension de gap (***ge***)
    - Valeurs typiques: entre -0.5 et -2

Utilisation d'une matrice de substitution pour calculer le score d'un alignement

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



- Les matrices de substitution sont utilisées pour calculer le score d'un alignement.
- Ce score est la somme, pour toutes les positions de l'alignement (i *de 1 à* L) , des scores des paires de résidus ($r_{1,i}$ et $r_{2,i}$).
- Les "gaps" sont traités par une règle spécifique reposant sur deux paramètres de pénalité:
  - Pénalité d'ouverture de gap (***go***)
    - Valeurs typiques: entre -10 et -15
  - Pénalité d'extension de gap (***ge***)
    - Valeurs typiques: entre -0.5 et -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
```



```
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
```


Construction d'une matrice d'alignement à partir d'une matrice de substitutions

- Revenons à l'exemple précédent.
  - Haut: une matrice de scores (arbitraires) de substitution pour séquences d'ADN (un score pour chaque paire de nucléotides).
  - Bas: matrice d'alignement pour deux petites séquences d'ADN.
- Dans chaque cellule de la matrice d'alignement, on insère le score de la paire de résidus correspondants, extrait de la matrice de substitutions.

Matrice de substitutions

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

Matrice d'alignement

	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 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

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.