Mesures de similarité

Matrices de substitutions

- La structure des protéines déterminent leur fonction.
- Des séquences assez différentes peuvent se replier en la même structure, et donc assurer la même fonction.
- Des « substitutions » d'acides aminés qui préservent la même structure ne devraient pas être trop nuisibles à la fonction de la protéine.
- Par exemple, 6 aa sont hydrophobes. Ils préfèrent être à l'intérieur de la structure pour évider d'être en contact avec l'eau. Et donc une substitution d'un aa hydrophobe pour un autre dans la même classe est plus acceptable qu'une subs. à l'extérieur de cette classe.
- Il est important de définir des matrices de substitutions appropriées. Deux classes de matrices sont utilisées: PAM et BLOSUM.

Matrices PAM

- PAM: “Point Accepted Mutations”.
- Probabilité d’une substitution d’un AA en un autre.
- Ensemble de matrices utilisées pour évaluer un alignement de séquences de protéines.
- Introduites par Margaret Dayhoff en 1978.

PAM 250

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

Unité PAM

- Unité de mesure du taux de divergence entre 2 séquences d'AA. Représente une distance d'évolution.

- Définition: S_1, S_2 divergent d'1 unité PAM si la suite de substitutions qui a converti S_1 en S_2 est telle qu'en moyenne, une seule mutation est survenue tous les 100 AA.

  Exp.: S_1 diverge de 5 PAM de S_2

- Mutations acceptées: celles incorporées dans la protéine et transmises. Soit sans effet, soit bénéfique à l'organisme.
- Pas de correspondance absolue entre unités PAM et divergence de séquences. Plusieurs mut. peuvent être survenues à la même pos.

Divergence d'AA ≤ unités PAM

Exemple: Deux seq. qui divergent de 100 PAM ne sont pas différentes à chaque pos.

En fait, deux seq. qui divergent de 200 PAM sont susceptibles de contenir 25% d'identité de seq.

Matrices PAM

- Différentes matrices PAM pour comparer des seq. d'AA qui divergent d'un nb spécifique d'unités PAM: 120 PAM, 250 PAM…
- Signification: La case (i,j) d'une mat. *n* PAM contient la fréquence avec laquelle l'AA A_i est remplacé par l'AA A_j dans les seq. qui divergent de *n* unités PAM
- Méthode idéale de const. d'une mat. *n* PAM:
  - Considérer un ensemble de seq qui divergent de n unités PAM
  - Aligner les seq. 2 à 2
  - Compter le nb. d'alignements A_i,A_j, pour chaque A_i, A_j. Diviser par le nb total d'appariements - - - > f(i,j)
  - Case (i,j) de la mat. Contient *log [f(i,j)/ f(i)f(j)]* où f(i) fréquence de A_i et f(j) freq. de A_j

- Méthode précédente nécessite d'aligner correctement les séquences. Alignement pour avoir la matrice, et matrice pour avoir l'alignement???
- Méthode de Dayhoff (1979):
  - Pour des seq. très similaires (moins de 15% de différence), principalement la méthode idéale

  → M: Matrice 1 PAM.
  - Séquences plus divergentes: $M^n(i,j)$: probabilité que A_i se transforme en A_j en n unités PAM

    Case (i,j) de la matrice n PAM:

    $$\log [f(i)\, M^n(i,j) / f(i)f(j)] = \log [M^n(i,j) / f(j)]$$
- Dans la pratique, on essaye plusieurs matrices PAM différentes. PAM 250 est la plus utilisée.

De PAM à BLOSUM

- Les matrices N PAM sont obtenues par extrapolation de la matrice 1PAM obtenue pour des protéines très proches.
- Pas appropriées pour la comparaison de séquences de protéines très divergentes
- BLOSUM *(Heinikoff and Heinikoff 1992)*, *Block Substitution Matrix :* Basée sur des block, i.e. régions conservées d'alignements de protéines: substitutions observées.

PROSITE et BLOCKS

- PROSITE: Dictionnaire de sites de protéines. Lié à Swiss-Prot.

  Motifs représenté par une exp. reg. ou par une matrice consensus

  Exemple: *G[GN] [SGA]GxRx[SGA]Cx(2)[IV]*

- BLOCKS: Dérivé de PROSITE. Dictionnaire de séquences conservées.

  BLOCK: Petit intervalle très conservé d'un alignement. Similarité de séquence, mais pas nécessairement similarité de fonction.

II. Matrices BLOSUM

- Dérivées de BLOCKS. Ensemble de blocs de *n* colonnes et *k* lignes
- Matrice BLOSUM: Nb de fois que Ai, Aj se trouvent appariés, divisé par le nb de fois qu'ils seraient appariés dans des seq. aléatoires.

Pour tous A_i, A_j, *n(i,j)* nb d'appariements *(Ai,Aj);*

f(i): freq. de *Ai; f(j):* freq. de *Aj*

$e(i,j) = n \binom{k}{2} f(i)\, f(j)$

$$s(i,j) = \log [n(i,j) / e(i,j)]$$

II. BLOSUM (suite)

- Caractéristique: Élimine la redondance dans les blocs.
- Matrice BLOSUM x (généralement entre 50 et 80): Pour tout couple de lignes contenant plus de x% de similarité, en garder une seule.
- La plus utilisée est BLOSUM 62

Matrice BLOSUM 62

Score positif pour les identités, et négatif pour les mismatchs

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