Mesures de similarité

Mesures de similarité

- Les structures 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 devrait avoir un score de similarité plus élevé qu'une subs. à l'extérieur de cette classe.

Mesures de similarité

- Étant donnée une paire de séquences *(x,y)* alignées, on veut mesurer la vraisemblance que les séquences soient apparentées (versus non-apparentées).
- On a besoin d'un modèle qui assigne une probabilité à chaque substitution (ici on ignore les indels) d'un résidu par un autre (nucléotide ou AA).
  - Modèle le plus simple: Modèle *R* aléatoire. Chaque résidu a apparaît de façon indépendante à une certaine fréquence *f(a)*. Dans ce cas, la probabilité de l'alignement *(x,y)* est simplement:

$$P(x, y|R) = \prod_i f(x_i) \prod_j f(x_j)$$

  - Dans le cas général d'un modèle *M* avec une probabilité *p(a,b)* pour l'alignement de deux résidus *a* et *b*, la probabilité de l'alignement est:

$$P(x, y|M) = \prod_i p(x_i y_i)$$

Mesures de similarité

- La vraisemblance de l'alignement est le rapport des ratios:

$$\frac{P(x,y|M)}{P(x,y|R)} = \frac{\prod_i p(x_i y_i)}{\prod_i f(x_i) \prod_i f(y_i)} = \prod_i \frac{p(x_i y_i)}{f(x_i) f(y_i)}$$

- Afin d'avoir un score additif, on prend le log des ratios:

$$S = \sum_i s(x_i, y_i),$$

$$\text{avec } s(a,b) = \log\left(\frac{p(ab)}{f(a)\, f(b)}\right)$$

- Le score $s(a,b)$ peut être conservé dans une matrice.
- Pour les acides aminés, il s'agit d'une matrice 20 x 20 avec $s(a_i,a_j)$ à la position (i,j), où a_i est le i-ième et a_j le j-ième acide aminé dans un ordre donné.

Mesures de similarité

- Comment établir les probabilités de substitution?
- Méthode intuitive:
  - Prendre un ensemble de séquences alignées;
  - Considérer les fréquences d'apparition des résidus, et la fréquence d'alignement de chaque paire de résidus.
- Difficulté avec ça: Obtenir un bon échantillon d'alignements aléatoires; mais surtout, les alignements ne reflètent pas tous la même divergence. Certaines paires de séquences ont un ancêtre commun récent, donc divergent peu, alors que d'autres ont un ancêtre commun plus vieux, donc divergent plus.
- 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 séquences d'AA qui divergent d'un nombre spécifique d'unités PAM: 120 PAM, 250 PAM, etc.
- Signification: La case (i,j) d'une matrice t PAM contient la fréquence avec laquelle l'AA a_i est remplacé par l'AA a_j dans les séquences qui divergent de t unités PAM.
- Méthode idéale de construction d'une matrice t PAM:
  - Considérer un ensemble de seq. qui divergent de t unités PAM;
  - Aligner les séquences 2 à 2;
  - Compter le nbre d'alignements (a_i,a_j) pour chaque paire de résidus et diviser par le nbre total de colonnes → Fréquence $f(a_i,a_j)$.
  - Case (i,j) de la matrice contient $\log \frac{f(i,j)}{f(i)f(j)}$ où f(i), f(j) sont les fréquences de a_i, 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.
  - Plutôt que de calculer $p(ab)$ (probabilité de l'alignement ab), pour extrapoler à des temps de divergence plus grands, Dayhoff considère plutôt $P(b/a,t)$: Probabilité que a soit substitué en b en t unités de temps.

- $P(b/a,t) = p(ab)_t / f(a)$
- Obtenu en multipliant t fois la matrice 1 PAM (matrice M)
- $M^t(i,j)$: Probabilité que a_i se transforme en a_j en t unités de temps.
- Et donc, case (i,j) de la matrice t PAM:

$$\log \frac{p(a_i a_j)_t}{f(a_i) f(a_j)} = \log \frac{f(a_i) M^t(i,j)}{f(a_i) f(aj)} = \log \frac{M^t(i,j)}{f(aj)}$$

- Dans la pratique, on essaye plusieurs matrices PAM différentes. PAM 250 est la plus utilisée.

De PAM à BLOSUM

- Les matrices t 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 représentés 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 a_i, a_j se trouvent appariés, divisé par le nb de fois qu'ils seraient appariés dans des seq. aléatoires.
- Comme d'habitude, $s(a,b) = \log\left(\frac{p(ab)}{f(a)\,f(b)}\right)$

$$p(ab) = {}^{f(ab)}/_{\sum_{cd} f(cd)}$$

où *f(ab)* est la fréquence de l'appariement *(ab),* i.e. nbre d'appariements *(ab)* divisé par le produit de la taille des séquences appariés.

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.
- BLOSUM 62 est la plus utilisée pour les alignements sans indels, et BLOSUM 50 la plus utilisée pour les alignements avec indels.

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