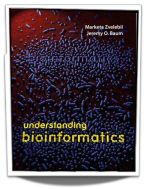
Introduction à la bioinformatique

4. la construction des matrices de substitution

Bibliographie



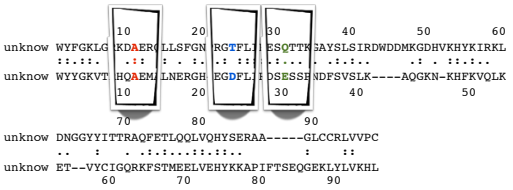
- Zvelebil et Baum, Understanding bioinformatics
- M. Dayhoff et al (1978) A model of evolutionary change in proteins. In Atlas of protein sequence and structure Vol 5, No suppl 3, p.345-351
- S. Henikoff et J.G. Henikoff (1991) Automated assembly of protein blocks for database searching Nucleic Acids Research 19(23):6565-6572
- S. Henikoff et J.G. Henikoff (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89:10915-10919

Objectifs

- Expliquer les principes d'évaluation d'un alignement
- Expliquer comment les matrices de substitution PAM et BLOSUM étaient construites
- Comprendre et être capable d'expliquer les différences entre les méthodes de construction

L'idée de base

Dans la troisième leçon, on a montré qu'un score de similarité entre les acides aminés est nécessaire pour l'évaluation d'un alignement



D'où vient le score pour l'alignement des résidus A–A (3), Q–E (2) et T–D (-1)?

Ce score représente la probabilité qu'un A ne change pas, qu'un Q change vers un E et qu'un T change vers un D ?

L'idée de base 2

- Deux mécanismes pourraient produire des différences entre des séquences de protéines
  - Un modèle aléatoire
  - Un modèle non-aléatoire (évolutif)
- Si on établit la probabilité de l'occurrence de l'alignement des résidus pour chaque modèle, on pourrait décider quel modèle est le plus probable de produire cet alignement

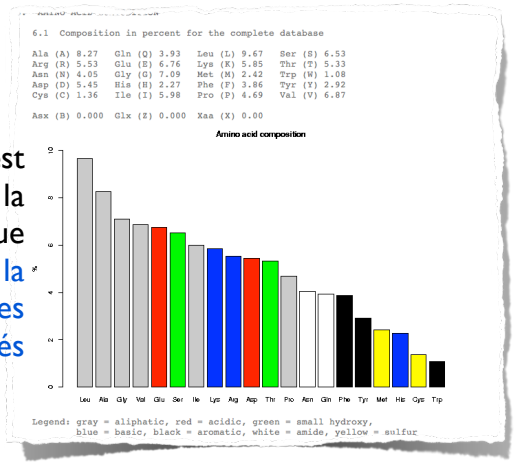
Modèle aléatoire

- Dans ce modèle, il n'y a pas de contraintes sur la composition de la séquence
- Le choix de résidu pour chaque position est indépendante des autres positions
  - la probabilité est dépendante sur la fréquence de chaque acides aminés dans la population des acides aminés
  - Example : $p_{tyr}=2.92\%$ et $p_{gly}=7.09\%$
- Si deux résidus a et b sont alignés, le modèle donne comme probabilité de l'occurrence $p_a\, p_b$ parce que les occurrences de ces résidus sont indépendants : $p_{tyr} * p_{gly} \sim 0.21\%$

la population est par exemple la fréquence des acides aminés dans les deux séquences ou ...

Modèle aléatoire

La probabilité est dépendante sur la fréquence de chaque acides aminés dans la population des acides aminés



6.1 Composition in percent for the complete database

Ala (A)	8.27	Gln (Q)	3.93	Leu (L)	9.67	Ser (S)	6.53
Arg (R)	5.53	Glu (E)	6.76	Lys (K)	5.85	Thr (T)	5.33
Asn (N)	4.05	Gly (G)	7.09	Met (M)	2.42	Trp (W)	1.08
Asp (D)	5.45	His (H)	2.27	Phe (F)	3.86	Tyr (Y)	2.92
Cys (C)	1.36	Ile (I)	5.98	Pro (P)	4.69	Val (V)	6.87

Asx (B) 0.000 Glx (Z) 0.000 Xaa (X) 0.00

Legend: gray = aliphatic, red = acidic, green = small hydroxy, blue = basic, black = aromatic, white = amide, yellow = sulfur

Modèle (évolutif) non-aléatoire

- Ce modèle propose que les séquences ont une association évolutives
  - c.-à.-d. Il existe des contraintes sur les combinaisons qu'on peut trouver
- La probabilité de l'occurrence dépend sur le résidu dans la même position dans l'ancêtre commun
  - la probabilité est $q_{a,b}$
    - La valeur de $q_{a,b}$ dépend sur la propriété du mécanisme évolutif
- Si deux résidus a et b sont alignés, le modèle donne comme probabilité de l'occurrence $q_{a,b}$ parce que les résidus sont corrélés

L'idée de base 2

- Le taux de chance $q_{a,b}/p_a p_b$ (*odds ratio*) montre quel modèle est le plus probable
  - Si $q_{a,b}/p_a p_b > 1$ le modèle non-aléatoir est plus probable pour produire cet alignement entre les deux résidus
- Le modèle doit expliquer l'alignement complet, c.-à-d. qu'on doit combiner les scores pour toutes les pairs de résidus dans l'alignement

$$\prod_u (q_{a,b}/p_a p_b)_u$$

u est la position dans l'alignement

L'idée de base 3

- Parfois, il est plus facile d'utiliser des sommes au lieu des produits.

$$\sum_u log(q_{a,b}/p_a p_b)_u$$

log-odds ratio

u est la position dans l'alignement

- Le taux de log-chance (*log-odds ratio*) est la valeur qu'on peut trouver dans une matrice de substitution

$$S = \sum_u (s_{a,b})_u$$

$s_{a,b} > 0$ signifie que la probabilité que les deux résidus sont alignés est plus grande dans le modèle non-aléatoire que dans le modèle aléatoire

Matrices de substitution

- Deux approches différentes ont été construites pour déterminer les scores $S_{a,b}$
  - Toutes matrices de PAM ont été dérivées de la matrice PAM1 qui était construite en utilisant 71 groupes de protéines alignés sans espaces et les séquences dans chaque groupe ont une identité d'au moins 85%
  - Les matrices de BLOSUM ont été dérivées indépendamment en utilisant des groupes de sous-séquences qui chacune ont une identité de séquence spécifique
    - BLOSUM 62 a utilisé des groupes de sous-séquences alignées sans espaces qui ont une identité d'au moins 62%

Les matrices PAM

PAM = Point or percentage accepted mutations

22 A Model of Evolutionary Change in Proteins

M.O. Dayhoff, R.M. Schwartz, and B.C. Orcutt

In the eight years since we last examined the amino acid exchanges seen in closely related proteins,[1] the information has doubled in quantity and comes from a much wider variety of protein types. The matrices derived from these data that describe the amino acid replacement probabilities between two sequences at various evolutionary distances are more accurate and the scoring matrix that is derived is more sensitive in detecting distant relationships than the one that we previously derived.[2,3] The method used in this chapter is essentially the same as that described in the *Atlas*, Volume 3[4] and Volume 5.[1]

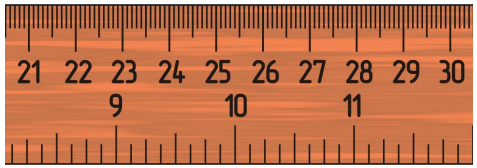
The matrix of accepted point mutations calculated from this tree is shown in Figure 79. We have assumed that the likelihood of amino acid X replacing Y is the same as that of Y replacing X, and hence 1 is entered in box YX as well as in box XY. This assumption is reasonable, because this likelihood should depend on the product of the frequencies of occurrence of the two amino acids and on their chemical and physical similarity. As a consequence of this assumption, no change in amino acid frequencies over evolutionary distance will be detected.

By comparing observed sequences with inferred ancestral sequences, rather than with each other, a sharper

Accepted Point Mutations

M. Dayhoff et al (1978) A model of evolutionary change in proteins. In Atlas of protein sequence and structure Vol 5, No suppl 3, p.345-351

Les Matrices PAM



L'idée de base est de construire **une unité** pour calculer la distance entre deux séquences

Quelle unité de base ? → PAM1 = une mutation acceptée pour 100 acide aminées

Les matrices PAM 2

Analyser les mutations ponctuelles acceptées (*accepted point mutations*) dans un grand nombre de groupes de séquences alignées (sans espaces)

Une mutation ponctuelle acceptée est la substitution d'un acide aminé par un autre qui était accepté par la sélection naturelle

Etant donné ces mutations ponctuelles acceptées, on pourrait calculer la probabilité qu'un acide aminé dans une position reste le même ou la probabilité qu'il change vers un autre acide aminé

Dayhoff et al. ont utilisé 71 groupes de protéines étroitement liés

Les matrices PAM 3

Comment a-t-on déterminé les mutations ponctuelles acceptées?

En analysant les séquences alignées (sans espaces) par groupe:

exemple imaginaire d'un groupe:

```
ACGH
DBGH
ADIJ
CBIJ
```

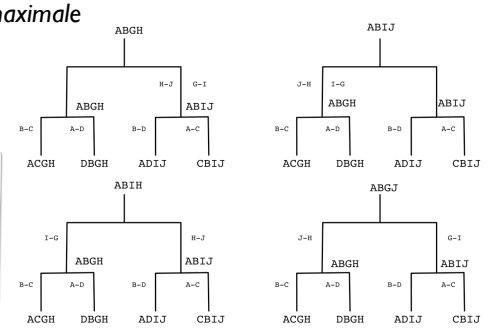

Les matrices PAM 4

Comme ces 4 séquences sont dérivées d'un ancêtre commun on peut reconstruire leur relation évolutive en construisant un arbre phylogénétique (leçon 10)

On peut dériver 4 arbres en utilisant la notion de la *parcimonie maximale*

le nombre minimal de substitutions

Chaque arbre produits 6 alignements :
(ACGH,ABGH) (DBGH,ABGH)
(ADIJ,ABIJ) (CBIJ,ABIJ)
(ABGH,ABGH) (ABIJ,ABGH)



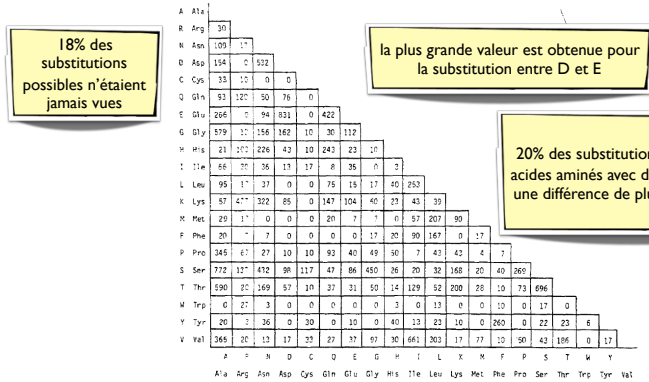
Les matrices PAM 5

Les substitutions annoté sur les branches des arbres donne les mutations ponctuelles acceptées

	A	B	C	D	G	H	I	J
A			4	4				
B			4	4				
C	4	4						
D	4	4						
G							4	
H								4
I					4			
J						4		

Les matrices PAM 6

Dayhoff et al ont créé des arbres phylogénétiques pour les 71 groupes de séquences et ont obtenu 1572 substitutions



18% des substitutions possibles n'étaient jamais vues

la plus grande valeur est obtenue pour la substitution entre D et E

20% des substitutions sont entre des acides aminés avec des codons qui ont une différence de plus d'un nucléotide

Les matrices PAM 7

Dans la prochaine étape, Dayhoff et al ont calculé la probabilité qu'un acide aminé change dans un petit intervalle évolutif donné = la mutabilité relative (*relative mutability*)

$$m_j = \frac{nombre\ de\ changements\ de\ j}{nombre\ d'occurrences\ de\ j}$$

Acide aminé	A	B	C	D	G	H	I	J
Substitutions	8	8	8	8	4	4	4	4
Occurrences	40	40	8	8	24	24	24	24
m_j	0,2	0,2	1	1	0,167	0,167	0,167	0,167

La mutabilité relative montre qu'il y a une différence dans le taux de substitutions pour les acides aminés différents

Les matrices PAM 8

La mutabilité relative pour chaque acide aminé obtenu par Dayhoff et al étaient:

Table 21
Relative Mutabilities of the Amino Acids[a]

Asn	134	His	66
Ser	120	Arg	65
Asp	106	Lys	56
Glu	102	Pro	56
Ala	100	Gly	49
Thr	97	Tyr	41
Ile	96	Phe	41
Met	94	Leu	40
Gln	93	Cys	20
Val	74	Trp	18

[a]The value for Ala has been arbitrarily set at 100.

Trp et Cys sont moins mutable

et

Asn, Ser, Asp et Glu sont plus mutable

Les matrices PAM 9

Etant donné la matrice des mutations ponctuelles acceptées et la mutabilité relative, on peut calculer *la probabilité de mutation* pour chaque paire d'acides aminés

$$M_{ij} = \frac{\lambda\, m_j A_{ij}}{\sum_i A_{ij}}$$

m_j est la mutabilité relative et A_{ij} est un élément de la matrice de mutations ponctuelles acceptées

$$M_{ii} = 1 - \lambda m_i$$

M_{ij} est la probabilité que l'acide aminé en colonne j soit substitué par l'acide aminé dans la ligne i dans cette unité de temps évolutif

λ est une constante importante qui est utilisé pour ajuster les M_{ij} de telle manière qu'un nombre spécifique de substitutions se produisent dans chaque 100 résidus

Les matrices PAM 10

Pour calculer λ on a besoin des fréquences relatives de l'exposition à la mutation (*effective frequency*) f_i

$$f_i = k \sum_b q_i^{(b)} N^{(b)}$$

$q_i^{(b)}$ est la fréquence d'acides aminés i dans le block b (gliss. 13) et $N^{(b)}$ est le nombre de substitution dans tous les arbres phylogénétiques construits pour le block b et k est une constante pour normaliser la somme des f_i à 1.

	q_i	N	f_i
A	2/16=0.125	24	0,125
B	0,125	24	0,125
C	0,125	24	0,125
D	0,125	24	0,125
H	0,125	24	0,125
H	0,125	24	0,125
I	0,125	24	0,125
J	0,125	24	0,125

Pour notre exemple, k = 0.0417

Les matrices PAM 11

Les fréquences effectives obtenu par Dayhoff et al étaient

Table 22
Normalized Frequencies of the Amino Acids in the Accepted Point Mutation Data

Gly	0.089	Arg	0.041
Ala	0.087	Asn	0.040
Leu	0.085	Phe	0.040
Lys	0.081	Gln	0.038
Ser	0.070	Ile	0.037
Val	0.065	His	0.034
Thr	0.058	Cys	0.033
Pro	0.051	Tyr	0.030
Glu	0.050	Met	0.015
Asp	0.047	Trp	0.010

Les matrices PAM 12

La constante λ ajustera les M_{ij} de telle manière qu'un nombre spécifique de substitutions se produisent dans chaque 100 résidus

Le nombre prévu d'acides aminés qui ne changera pas dans chaque 100 résidus est donné par:

$$x = 100 \sum_i f_i M_{ii} = 100 \sum_i f_i (1 - \lambda m_i)$$

$$\lambda = \frac{(100 - x)}{100 \sum_i f_i m_i}$$

Si on veut que, seulement 1% des résidus soient acceptés, λ deviendra

$$\lambda = \frac{1}{100 \sum_i f_i m_i}$$

Les matrices PAM 13

Les probabilités de mutation pour notre exemple sont:

λ = 0.0261

	A	B	C	D	G	H	I	J
A	99,48	0	1,3	1,3	0	0	0	0
B	0	99,48	1,3	1,3	0	0	0	0
C	0,26	0,26	97,4	0	0	0	0	0
D	0,26	0,26	0	97,4	0	0	0	0
G	0	0	0	0	99,57	0	0,43	0
H	0	0	0	0	0	99,57	0	0,43
I	0	0	0	0	0,43	0	99,57	0
J	0	0	0	0	0	0,43	0	99,57

Les éléments sont multiplié par 100

Cette matrice montre les probabilités de mutation pour chaque paire d'acides aminés dans un intervalle d'évolution spécifique : 1 mutation acceptée par 100 résidus = 1 Percent of Accepted mutations = 1 PAM

1 PAM est une unité d'évolution/divergence

Les matrices PAM 14

La matrice de probabilité de mutation dérivée par Dayhoff et al.

Cette matrice est le PAM1 de Dayhoff et al (les éléments sont multipliés par 10000)

	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
A Ala	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
R Arg	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
N Asn	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
D Asp	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
C Cys	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Q Gln	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
E Glu	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
G Gly	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
H His	1	0	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
I Ile	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
L Leu	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
K Lys	2	27	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
M Met	1	1	0	0	0	2	0	0	0	5	8	4	9674	1	0	1	2	0	0	4
F Phe	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
P Pro	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
S Ser	28	11	34	7	11	4	5	16	2	2	1	7	4	3	17	9840	38	5	2	2
T Thr	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
W Trp	0	2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	9976	1	0
Y Tyr	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
V Val	13	2	1	1	1	3	2	2	3	3	57	11	3	17	1	3	2	10	0	2	9901

REPLACEMENT AMINO ACID

Il y a 0.56% de probabilité que Asp soit remplacé par Glu

Les matrices PAM 15

La matrice de probabilité de mutation pourrait être utilisée pour la création des séquences et la prédiction de la parenté entre des séquences

- De nouvelles séquences pourraient être créées qui se différencient d' 1,2 ou plusieurs PAM d'une séquence originale
- Ou des séquences avec un nombre spécifique de substitutions pourraient être créées en utilisant aussi les mutabilités relatives de tous les acides aminés (gliss. 18)

La distance entre des séquences est exprimée en PAM

GRKDAERQLLSFGNPRGTF

Les matrices PAM 15

La matrice PAM1 donne la probabilité de mutation pour un intervalle d'évolution dans lequel 1% des substitutions étaient acceptées

Les probabilités de mutation pour des intervalles d'évolution plus grands sont obtenues par la multiplication du PAM1 avec elle-même

PAM2 = PAM1xPAM1=PAM1^2 — intervalle d'évolution : 2 substitutions acceptées pour ~~chaque~~ 100 résidus

PAM120 = PAM1^{120} — intervalle d'évolution : 120 substitutions acceptées pour ~~chaque~~ 100 résidus

PAM250 = PAM1^{250} — intervalle d'évolution : 250 substitutions acceptées pour ~~chaque~~ 100 résidus

divergence

fréquences effectives

Les matrices PAM 16

Cette matrice est le PAM250 de Dayhoff et al (les éléments sont multipliés par 100)

ORIGINAL AMINO ACID

REPLACEMENT AMINO ACID	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
A Ala	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
R Arg	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
N Asn	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
D Asp	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
C Cys	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
Q Gln	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3
E Glu	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
G Gly	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
H His	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
I Ile	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
L Leu	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
K Lys	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
M Met	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
F Phe	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
P Pro	7	5	5	4	3	5	4	5	6	3	3	4	3	2	20	6	5	1	2	4
S Ser	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
T Thr	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
W Trp	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
Y Tyr	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
V Val	7	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	7	2	4	17

Quand on fait la comparaison entre deux séquences,

il y a toujours une probabilité de 52% qu'un Cys reste un Cys

il y a une probabilité de 20% qu'un Tyr est remplacé par un Phe

Les matrices PAM 17

Au début, on a dit qu'on avait besoin d'un score de similarité pour faire l'alignement des séquences et ce score est le taux log-chance entre la probabilité d'occurrence de deux acides aminés alignés dans un modèle d'évolution ($q_{a,b}$) et un modèle aléatoire ($p_a p_b$)

$$\sum_u log(q_{a,b}/p_a p_b)_u$$

Ici le score est le taux entre la probabilité de mutation entre les deux acides aminés (modèle d'évolution) et la probabilité qu'on ait un certain acide aminé dans la deuxième séquence (modèle aléatoire)

$$R_{a,b} = \sum_u log(M_{a,b}/f_a)_u$$

$R_{a,b} = R_{b,a}$

Les matrices PAM 18

Le PAM250 taux de log-chance de Dayhoff et al (les éléments ont été achevés et multipliés par 10)

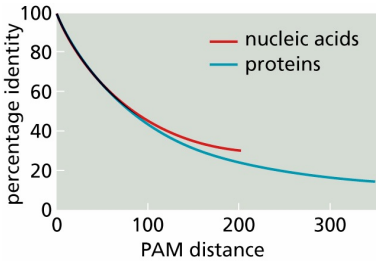
	C Cys	S Ser	T Thr	P Pro	A Ala	G Gly	N Asn	D Asp	E Glu	Q Gln	H His	R Arg	K Lys	M Met	I Ile	L Leu	V Val	F Phe	Y Tyr	W Trp
C Cys	12																			
S Ser	0	2																		
T Thr	-2	1	3																	
P Pro	-3	1	0	6																
A Ala	-2	1	1	1	2															
G Gly	-3	1	0	-1	1	5												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

Cette matrice est utile pour trouver des relations distantes entre des séquences

Les matrices PAM 18

Si deux séquences ont une différence d'identité de 40%, quelle matrice est-ce qu'on doit utiliser pour la production d'un bon alignement ?

$$d = 100(1 - \sum_i f_i M_{ii}^n)$$



identité (%)	différence d (%)	PAM index n
99	1	1
95	5	5
90	10	11
85	15	17
80	20	23
75	25	30
70	30	38
60	40	56
50	50	80
40	60	112
30	70	159
20	80	246
14	86	350

Les matrices BLOSUM

BLOSUM = Block Substitution matrix

Proc. Natl. Acad. Sci. USA
Vol. 89, pp. 10915–10919, November 1992
Biochemistry

Amino acid substitution matrices from protein blocks

(amino acid sequence/alignment algorithms/data base searching)

STEVEN HENIKOFF* AND JORJA G. HENIKOFF

Howard Hughes Medical Institute, Basic Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98104

Communicated by Walter Gilbert, August 28, 1992 (received for review July 13, 1992)

ABSTRACT **Methods for alignment of protein sequences typically measure similarity by using a substitution matrix with scores for all possible exchanges of one amino acid with another. The most widely used matrices are based on the Dayhoff model of evolutionary rates. Using a different approach, we have derived substitution matrices from about 2000 blocks of aligned sequence segments characterizing more than 500 groups of related proteins. This led to marked improvements in alignments and in searches using queries from each of the groups.**

Among the most useful computer-based tools in modern biology are those that involve sequence alignments of proteins, since these alignments often provide important insights into gene and protein function. There are several different

new sequence and every other sequence in the block. For example, if the residue of the new sequence that aligns with the first column of the first block is A and the column has 9 A residues and 1 S residue, then there are 9 AA matches and 1 AS mismatch. This procedure is repeated for all columns of all blocks with the summed results stored in a table. The new sequence is added to the group. For another new sequence, the same procedure is followed, summing these numbers with those already in the table. Notice that successive addition of each sequence to the group leads to a table consisting of counts of all possible amino acid pairs in a column. For example, in the column consisting of 9 A residues and 1 S residue, there are 8 + 7 + . . . 1 = 36 possible AA pairs, 9 AS or SA pairs, and no SS pairs. Counts of all possible pairs in each column of each block in the data base are summed.

S. Henikoff et J.G. Henikoff (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89:10915-10919

Les matrices BLOSUM 2

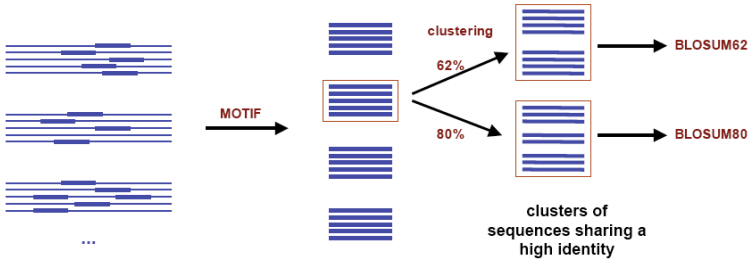
L'objectif était de créer des bonnes matrices de substitution pour trouver des régions conservées entre des séquences de protéines

Les matrices BLOSUM ne sont pas basées sur la distance d'évolution comme les matrices PAM

Les matrices BLOSUM étaient construites en utilisant la base de données BLOCKS (http://blocks.fhcrc.org/)

La base de données BLOCKS est composée d'une collection de blocs qui représente des familles protéiques qui pourrait être utilisées pour détecter des relations entre des séquences ADN ou des séquences protéiques avec ces familles connues.

Les matrices BLOSUM 3



S. Henikoff et J.G. Henikoff (1991) Automated assembly of protein blocks for database searching Nucleic Acids Research 19(23):6565-6572

Les matrices BLOSUM 3

Les séquences sont mises dans le même groupe quand ils ont **une identité** de C% ou plus que C%.

Un exemple d'un bloc de 7 séquences alignées dans lequel on a trouvé un motif conservé : une cystéine

Exemple : C= 50%

groupe 1:	groupe 2:	groupe 3:
ATCKQ		
SSCRN		
ATCRN		
TECRQ		
ASCKN		
SECEN		
SDCEQ		

Les matrices BLOSUM 3

Les séquences sont mises dans le même groupe quand ils ont **une identité** de C% ou plus que C%.

Un exemple d'un bloc de 7 séquences alignées dans lequel on a trouvé un motif conservé : une cystéine

Exemple : C= 50%

groupe 1: groupe 2: groupe 3:

ATCKQ
SSCRN
ATCRN
TECRQ
ASCKN
SECEN
SDCEQ

Les matrices BLOSUM 3

Les séquences sont mises dans le même groupe quand ils ont **une identité** de C% ou plus que C%.

Un exemple d'un bloc de 7 séquences alignées dans lequel on a trouvé un motif conservé : une cystéine

Exemple : C= 50%

groupe 1: groupe 2: groupe 3:

ATCKQ ATCKQ
SSCRN → SSCRN
ATCRN
TECRQ
ASCKN
SECEN
SDCEQ

Les matrices BLOSUM 3

Les séquences sont mises dans le même groupe quand ils ont **une identité** de C% ou plus que C%.

Un exemple d'un bloc de 7 séquences alignées dans lequel on a trouvé un motif conservé : une cystéine

Exemple : C= 50%

groupe 1: groupe 2: groupe 3:

ATCKQ ATCKQ
SSCRN → SSCRN
ATCRN → ATCRN
TECRQ
ASCKN
SECEN
SDCEQ

Les matrices BLOSUM 3

Les séquences sont mises dans le même groupe quand ils ont **une identité** de C% ou plus que C%.

Un exemple d'un bloc de 7 séquences alignées dans lequel on a trouvé un motif conservé : une cystéine

Exemple : C= 50%

groupe 1: groupe 2: groupe 3:

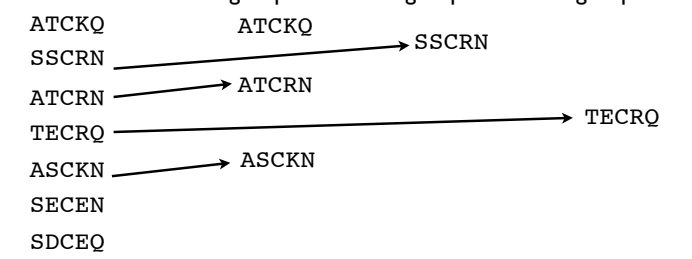
ATCKQ ATCKQ
SSCRN → SSCRN
ATCRN → ATCRN
TECRQ → TECRQ
ASCKN
SECEN
SDCEQ

Les matrices BLOSUM 3

Les séquences sont mises dans le même groupe quand ils ont **une identité** de C% ou plus que C%.

Exemple : C= 50%

Un exemple d'un bloc de 7 séquences alignées dans lequel on a trouvé un motif conservé : une cystéine

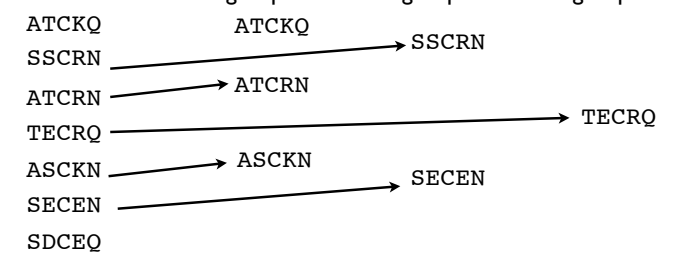


Les matrices BLOSUM 3

Les séquences sont mises dans le même groupe quand ils ont **une identité** de C% ou plus que C%.

Exemple : C= 50%

Un exemple d'un bloc de 7 séquences alignées dans lequel on a trouvé un motif conservé : une cystéine

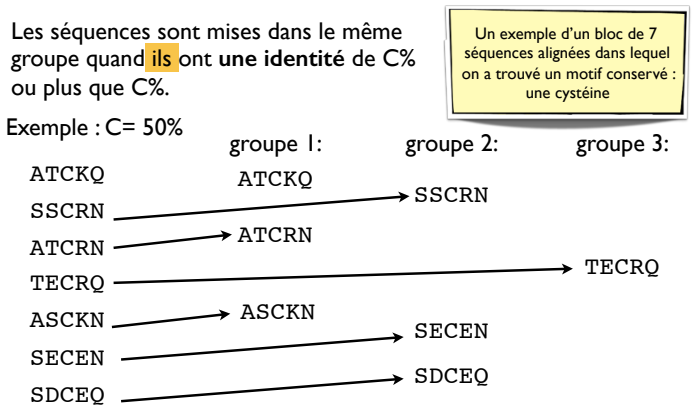


Les matrices BLOSUM 3

Les séquences sont mises dans le même groupe quand ils ont **une identité** de C% ou plus que C%.

Exemple : C= 50%

Un exemple d'un bloc de 7 séquences alignées dans lequel on a trouvé un motif conservé : une cystéine

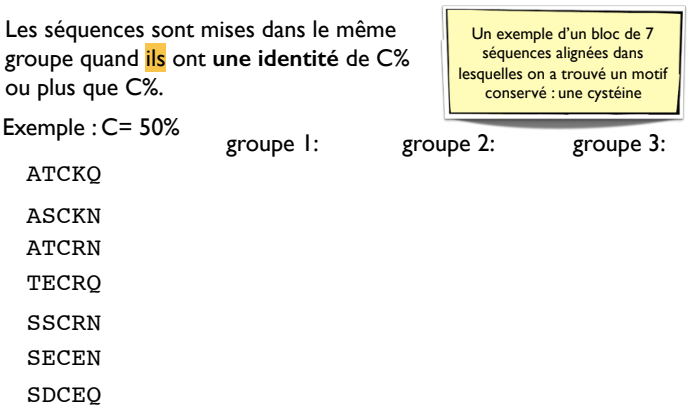


Les matrices BLOSUM 3

Les séquences sont mises dans le même groupe quand ils ont **une identité** de C% ou plus que C%.

Exemple : C= 50%

Un exemple d'un bloc de 7 séquences alignées dans lesquelles on a trouvé un motif conservé : une cystéine



Les matrices BLOSUM 3

Les séquences sont mises dans le même groupe quand ils ont **une identité** de C% ou plus que C%.

Un exemple d'un bloc de 7 séquences alignées dans lesquelles on a trouvé un motif conservé : une cystéine

Exemple : C= 50%

	groupe 1:	groupe 2:	groupe 3:
ATCKQ	ATCKQ		
ASCKN			
ATCRN			
TECRQ			
SSCRN			
SECEN			
SDCEQ			

37-2

Les matrices BLOSUM 3

Les séquences sont mises dans le même groupe quand ils ont **une identité** de C% ou plus que C%.

Un exemple d'un bloc de 7 séquences alignées dans lesquelles on a trouvé un motif conservé : une cystéine

Exemple : C= 50%

	groupe 1:	groupe 2:	groupe 3:
ATCKQ	ATCKQ		
ASCKN →	ASCKN		
ATCRN			
TECRQ			
SSCRN			
SECEN			
SDCEQ			

37-3

Les matrices BLOSUM 3

Les séquences sont mises dans le même groupe quand ils ont **une identité** de C% ou plus que C%.

Un exemple d'un bloc de 7 séquences alignées dans lesquelles on a trouvé un motif conservé : une cystéine

Exemple : C= 50%

	groupe 1:	groupe 2:	groupe 3:
ATCKQ	ATCKQ		
ASCKN →	ASCKN		
ATCRN →	ATCRN		
TECRQ			
SSCRN			
SECEN			
SDCEQ			

37-4

Les matrices BLOSUM 3

Les séquences sont mises dans le même groupe quand ils ont **une identité** de C% ou plus que C%.

Un exemple d'un bloc de 7 séquences alignées dans lesquelles on a trouvé un motif conservé : une cystéine

Exemple : C= 50%

	groupe 1:	groupe 2:	groupe 3:
ATCKQ	ATCKQ		
ASCKN →	ASCKN		
ATCRN →	ATCRN		
TECRQ →		TECRQ	
SSCRN			
SECEN			
SDCEQ			

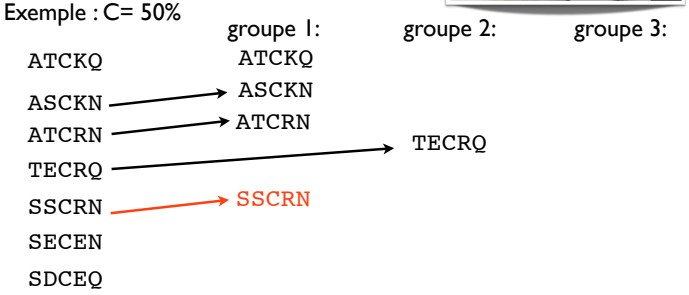
37-5

Les matrices BLOSUM 3

Les séquences sont mises dans le même groupe quand ils ont **une identité** de C% ou plus que C%.

Un exemple d'un bloc de 7 séquences alignées dans lesquelles on a trouvé un motif conservé : une cystéine

Exemple : C= 50%



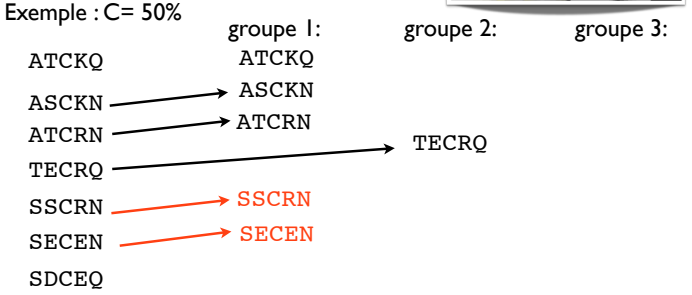
37-6

Les matrices BLOSUM 3

Les séquences sont mises dans le même groupe quand ils ont **une identité** de C% ou plus que C%.

Un exemple d'un bloc de 7 séquences alignées dans lesquelles on a trouvé un motif conservé : une cystéine

Exemple : C= 50%



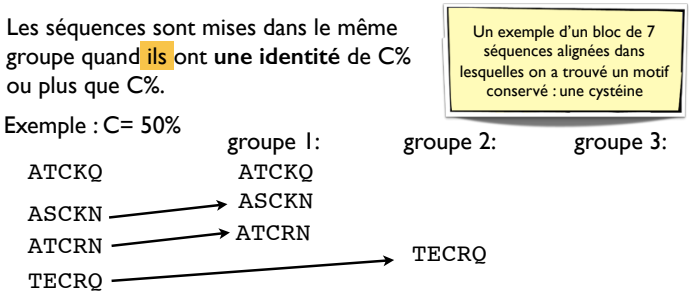
37-7

Les matrices BLOSUM 3

Les séquences sont mises dans le même groupe quand ils ont **une identité** de C% ou plus que C%.

Un exemple d'un bloc de 7 séquences alignées dans lesquelles on a trouvé un motif conservé : une cystéine

Exemple : C= 50%



37-8

Les matrices BLOSUM 4

Dans les slides suivants, la dérivation des matrices de BLOSUM sera démontré

Ici l'identité est 62%

Comme les séquences dans chaque groupe ont une identité de ≥62%, on appelle la matrice qui sera dérivée de ces données le BLOSUM62

ATCKQ |
ATCRN |
ASCKN |
SSCRN |
SDCEQ |
SECEN |
TECRQ |

Le grand problème est que les bases de données de séquences sont fortement décentrées vers certaines espèces et protéines

Pour réduire ce décentrement les séquences sont groupées en utilisant leur identité.

Leur importance est pondérée en utilisant un poids basé sur le nombre de séquences dans le groupe

38

Les matrices BLOSUM 5

Dans les glissières suivantes, la dérivation de les matrices de BLOSUM sera démontré

Ici l'identité est 50%

parce que les séquences dans chaque groupe ont une identité de ≥50%, on appelle la matrice qui sera dérivée de ces données le BLOSUM50

ATCKQ
ATCRN 1/3
ASCKN

SSCRN
SDCEQ 1/3
SECEN

TECRQ 1

Le grand problème est que les bases de données de séquences sont fortement décentrées vers certaines espèces et protéines

Pour réduire ce décentrement les séquences sont groupés sur leur identité et leur importance est pondéré en utilisant un poids basé sur le nombre de séquences dans le groupe

Les matrices BLOSUM 6

Utilisant ces poids, on calcul une matrice de fréquences pondérées

Comme ici l'identité est 62%, chaque séquence a le même poids, c.a.d. 1

ATCKQ
ATCRN
ASCKN
SSCRN
SDCEQ
SECEN
TECRQ

$f_{Q,N}=12+0+0+0+0$
$f_{Q,Q}=3$
$f_{N,N}=6$

	A	C	D	E	K	N	Q	R	S	T
A										
C										
D										
E										
K										
N						6	12			
Q						12	3			
R										
S										
T										

Le nombre total de paire de groupes est $c*(n*(n-1)/2)$, n est le nombre de groupes et c le nombre de colonnes

Les matrices BLOSUM 7

Utilisant ces poids, on calcul une matrice de fréquences pondérées

Comme ici l'identité est 50%, les séquences ont des poids différents

1/4 ATCKQ
ATCRN
ASCKN
SSCRN
1/2 SDCEQ
SECEN
1 TECRQ

	A	C	D	E	K	N	Q	R	S	T
A										
C										
D										
E										
K										
N						3/8	14/8			
Q						14/8	7/8			
R										
S										
T										

$f_{Q,N}=(1/4*1/2)+(3/4*1/2)+(3/4*1)+(1/2*1)$ $f_{Q,Q}=(1/4*1/2)+(1/2*1)+(1/4*1)$ $f_{N,N}=(3/4*1/2)$

Les matrices BLOSUM 8

Utilisant ces poids, on calcul une matrice de fréquences pondérées

1/4 ATCKQ
ATCRN
ASCKN
SSCRN
1/2 SDCEQ
SECEN
1 TECRQ

$f_{K,R}=(2/4*1)$
$f_{K,E}=(2/4*1)$
$f_{R,E}=(2/4*1)+(1*1)$
$f_{R,R}=(2/4*1)$
$f_{K,K}=0$
$f_{E,E}=0+(1/2*1)$

	A	C	D	E	K	N	Q	R	S	T
A										
C										
D										
E				1/2	1/2			6/4		
K				1/2	0			1/2		
N						3/8	14/8			
Q						14/8	7/8			
R				6/4	1/2			1/2		
S										
T										

On tient compte de toutes les colonnes pour le calcul de $f_{a,b}$

Les matrices BLOSUM 9

La matrice de fréquences pondérées pour BLOSUM50 est :

		A	C	D	E	K	N	Q	R	S	T
	ATCKQ										
1/4	ATCRN										
	ASCKN										
	SSCRN										
	SDCEQ										
1/2	SECEN										
1	TECRQ										
	A	0	0	0	0	0	0	0	0	3/4	3/4
	C	0	3	0	0	0	0	0	0	0	0
	D	0	0	0	1/2	0	0	0	0	1/4	1/4
	E	0	0	1/2	1/2	1/2	0	0	6/4	3/4	3/4
	K	0	0	0	1/2	0	0	0	1/2	0	0
	N	0	0	0	0	0	3/8	14/8	0	0	0
	Q	0	0	0	0	0	14/8	7/8	0	0	0
	R	0	0	0	6/4	1/2	0	0	1/2	0	0
	S	3/4	0	1/4	3/4	0	0	0	0	1/4	5/4
	T	3/4	0	1/4	3/4	0	0	0	0	5/4	0

Les matrices BLOSUM 10

Ces fréquences pondérées sont utilisées pour le calcul des probabilités d'occurrence dans le modèle d'évolution ($q_{a,b}$)

$$q_{a,b}=\frac{f_{a,b}}{\sum_{1\le b\le a} f_{a,b}}$$

	A	C	D	E	K	N	Q	R	S	T
A	0	0	0	0	0	0	0	0	3/4	3/4
C	0	3	0	0	0	0	0	0	0	0
D	0	0	0	1/2	0	0	0	0	1/4	1/4
E	0	0	1/2	1/2	1/2	0	0	6/4	3/4	3/4
K	0	0	0	1/2	0	0	0	1/2	0	0
N	0	0	0	0	0	3/8	14/8	0	0	0
Q	0	0	0	0	0	14/8	7/8	0	0	0
R	0	0	0	6/4	1/2	0	0	1/2	0	0
S	3/4	0	1/4	3/4	0	0	0	0	1/4	5/4
T	3/4	0	1/4	3/4	0	0	0	0	5/4	0

	A	C	D	E	K	N	Q	R	S	T
A	0	0	0	0	0	0	0	0	0,05	0,05
C	0	0,2	0	0	0	0	0	0	0	0
D	0	0	0	0,033	0	0	0	0	0,0167	0,0167
E	0	0	0,033	0,033	0,033	0	0	0,1	0,05	0,05
K	0	0	0	0,033	0	0	0	0,033	0	0
N	0	0	0	0	0	0,025	0,1167	0	0	0
Q	0	0	0	0	0	0,1167	0,0583	0	0	0
R	0	0	0	0,1	0,033	0	0	0,033	0	0
S	0,05	0	0,0167	0,05	0	0	0	0	0,0167	0,0833
T	0,05	0	0,017	0,05	0	0	0	0	0,083	0

Les matrices BLOSUM 11

Ces fréquences pondérées sont utilisées pour le calcul des probabilités d'occurrence dans le modèle d'évolution ($q_{a,b}$)

$$q_{a,b}=\frac{f_{a,b}}{\sum_{1\le b\le a} f_{a,b}}$$

Si on avait utilisé les fréquences pondérées de BLOSUM62

	N	Q
N	0,057	0,114
Q	0,114	0,029

Les matrices BLOSUM 12

Pour le calcul des taux de log-chance on aura aussi besoin de la probabilité d'occurrence de cet alignement de résidu a et b dans le modèle aléatoire

La fréquence prévue pour l'alignement

$e_{aa}=p_a^2$ pour des résidus identiques

$e_{ab}=2p_ap_b$ pour des résidus différents

La fréquence prévue par résidu

$$p_a=q_{a,a}+\frac{1}{2}\sum_{a\neq b} q_{a,b}$$

Les matrices BLOSUM 13

Pour le calcul des taux de log-chance on aura aussi besoin de la probabilité d'occurrence de cet alignement de résidu a et b dans le modèle aléatoire

$p_Q = q_{Q,Q} + (1/2)\sum_{b \neq Q} q_{Q,b}$

$p_Q = 0.058 + (0.117/2) = 0.117$

$p_N = q_{N,N} + (1/2)\sum_{b \neq N} q_{N,b}$

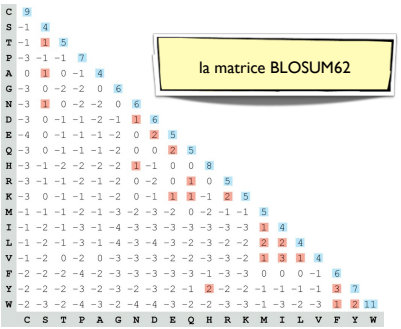
$p_Q = 0.025 + (0.117/2) = 0.0834$

	A	C	D	E	K	N	Q	R	S	T
A	0	0	0	0	0	0	0	0	0,05	0,05
C	0	5	0	0	0	0	0	0	0	0
D	0	0	0	0,033	0	0	0	0	0,0167	0,0167
E	0	0	0,033	0,033	0,033	0	0	0,1	0,05	0,05
K	0	0	0	0,033	0	0	0	0,033	0	0
N	0	0	0	0	0	0,025	0,1167	0	0	0
Q	0	0	0	0	0	0,1167	0,0583	0	0	0
R	0	0	0	0,1	0,033	0	0	0,033	0	0
S	0,05	0	0,0167	0,05	0	0	0	0	0,0167	0,0833
T	0,05	0	0,017	0,05	0	0	0	0	0,083	0

Les matrices BLOSUM 14

Comme dans les matrices PAM, les matrices BLOSUM contiennent le taux de log-chance entre la probabilité d'occurrence du pair dans le modèle d'évolution et le modèle aléatoire

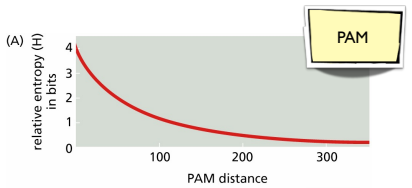
$$s_{a,b} = 2\log_2\left(\frac{q_{a,b}}{e_{a,b}}\right)$$



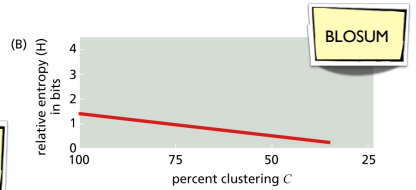
Quelles PAM ou BLOSUM?

En utilisant l'entropie relatif (H) on peut déterminer si la matrice est capable de distinguer entre l'évolution et la chance:

$$H = \sum_{a,b} q_{a,b} s_{a,b} = \sum_{a,b} q_{a,b} \log\left(\frac{q_{a,b}}{p_a p_b}\right)$$



La Longueur d'alignement (sans espaces) qui est nécessaire pour distinguer entre le modèle augmente inversement avec H



S.F.Altschul (1991) Amino acid subsitution matrices from an information theoretic perspective J Mol Biol 219:555-565