

Domaine SNV : Biologie, Agronomie, Science Alimentaire, Ecologie

Bioinformatique

www.facebook.com/DomaineSNV/

Alignement de séquences multiples

Jacques van Helden

Jacques.van-Helden@univ-amu.fr

Aix-Marseille Université (AMU), 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

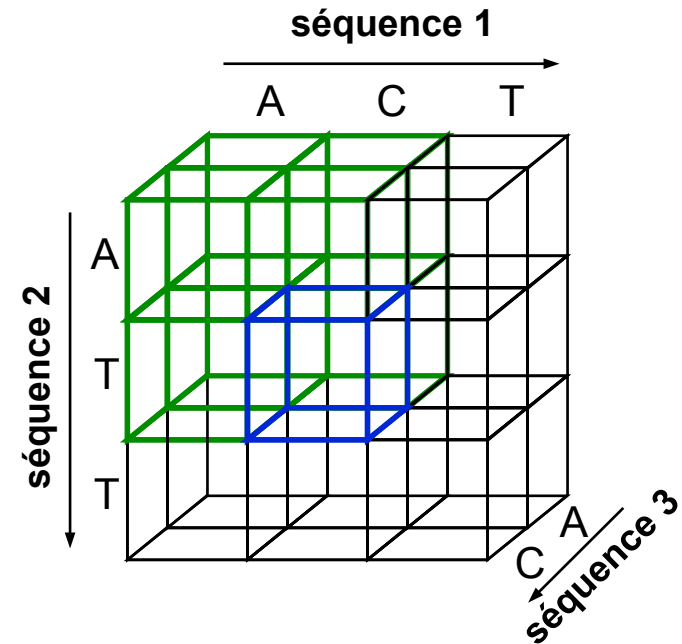
Bioinformatique des Génomes et des Réseaux (BiGRe lab)

<http://www.bigre.ulb.ac.be/>



Aignement multiple par programmation dynamique

- L'approche de programmation dynamique peut être étendue pour aligner 3 séquences.
 - Construction d'une matrice d'alignement tridimensionnelle.
 - Le meilleur score de chaque cellule est calculé sur base des cellules précédentes dans les 3 directions.
- On peut étendre le concept à N séquences en utilisant un hypercube à N dimensions.
- Problème : la taille de la matrice (mémoire occupée) et le temps d'exécution augmentent **exponentiellement** avec le nombre de séquences:
 - 2 séquences $L1 \times L2$
 - 3 séquences $L1 \times L2 \times L3$
 - 4 séquences $L1 \times L2 \times L3 \times L4$
 - n séquences $L1 \times L2 \times \dots \times L_n \sim L^n$
- Aligner N séquences en programmation dynamique requiert $O(L^n)$ opérations, ce qui devient très vite impraticable.
- L'efficacité peut être améliorée en ne considérant qu'un sous-espace de la matrice à N-dimension. Cependant, le nombre de séquences praticable reste très limité (~8 séquences max).

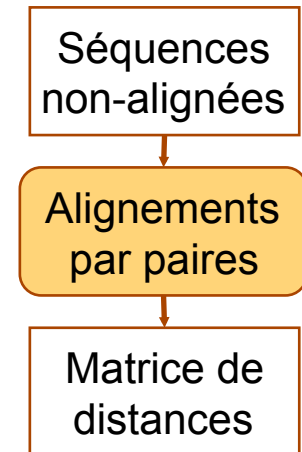


Alignement progressif

- Une approche alternative pour aligner des séquences multiples est de réaliser un **alignement progressif**.
- L'algorithme procède en plusieurs étapes:
 - Calculer une **matrice de distances**, qui indique la distance entre chaque paire de séquences.
 - Construire un **arbre guide** qui regroupe en premier lieu les séquences les plus proches, et remonte en regroupant progressivement les séquences les plus éloignées.
 - Utiliser ce arbre pour aligner progressivement les séquences.
- Il s'agit d'une approche **heuristique**
 - Cette approche est praticable pour un grand nombre de séquences, mais ne peut pas garantir de retourner l'alignement optimal.

Alignement progressif – 1^{ère} étape: construction de la matrice de distance

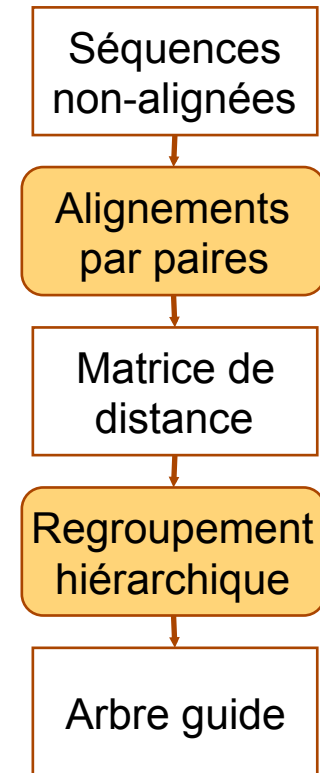
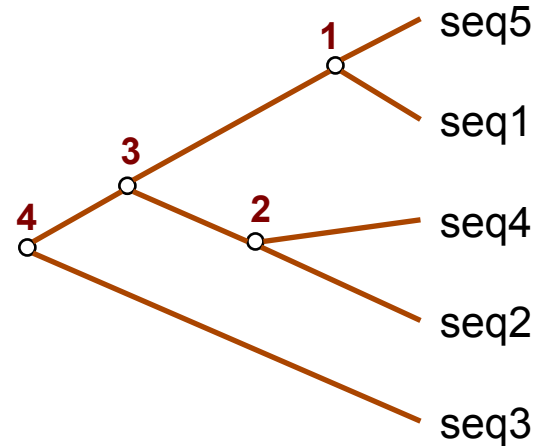
- On effectue un alignement par paires entre chaque paire de protéines
 - Alignement par programmation dynamique ou par BLAST.
 - Nombre d'alignements = $n * (n - 1) / 2$
- A partir de chaque alignement par paire, calculer la distance entre les deux séquences.
 - $d_{i,j} = s_{i,j} / L_{j,j}$
 - $d_{j,j}$ distance entre les séquences i and j
 - $L_{j,j}$ longueur de l'alignement
 - $s_{j,j}$ nombre de substitutions
- Remarques
 - Les gaps ne sont pas pris en compte dans la métrique de distance
 - La matrice est symétrique: $d_{i,j} = d_{j,i}$
 - Les éléments diagonaux sont nuls: $d_{i,i} = 0$



	seq 1	seq 2	...	seq n
seq 1	d1,1	d1,2	...	d1,n
seq 2	d2,1	d2,2	...	d2,n
...
seq n	dn,1	dn,2	...	dn,n

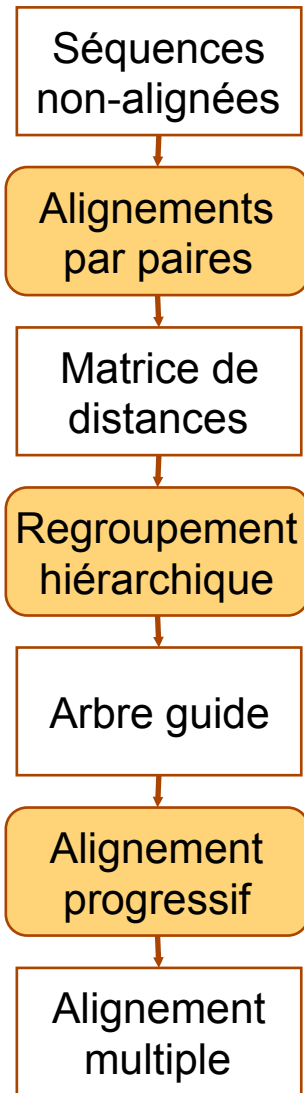
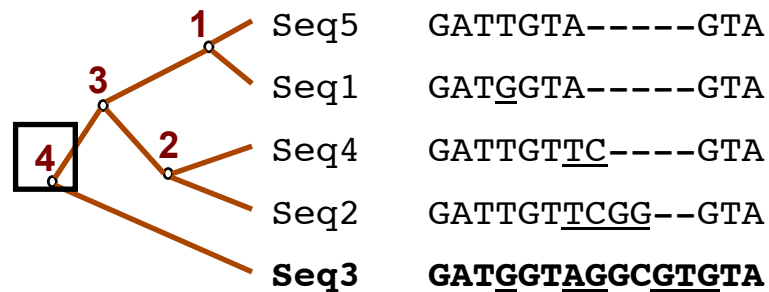
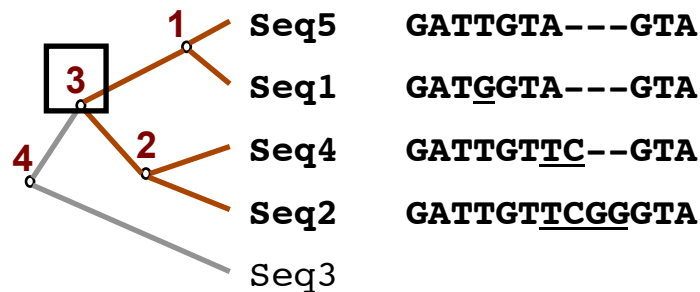
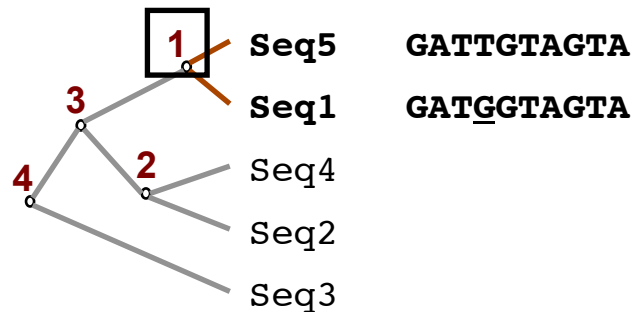
Alignement progressif – 2^{ème} étape : construire l'arbre guide

- On peut calculer un arbre à partir d'une matrice de distance par regroupement hiérarchique.
 - On commence par regrouper les deux séquences les plus proches (groupe **1**)
 - Regrouper ensuite les groupes les plus proches
 - Les deux séquences les plus proches (groupe **2**).
 - Un groupe avec un groupe (groupe **3**).
 - Une séquence avec un groupe précédent (groupe **4**).
- Cet arbre sera ensuite utilisé comme **guide** pour déterminer l'ordre d'incorporation des séquences dans l'alignement multiple.
- Attention ! Cet arbre ne doit pas être interprété comme un arbre phylogénétique.
 - Il sert uniquement à identifier les similarités les plus fortes entre séquences pour construire l'alignement multiple.



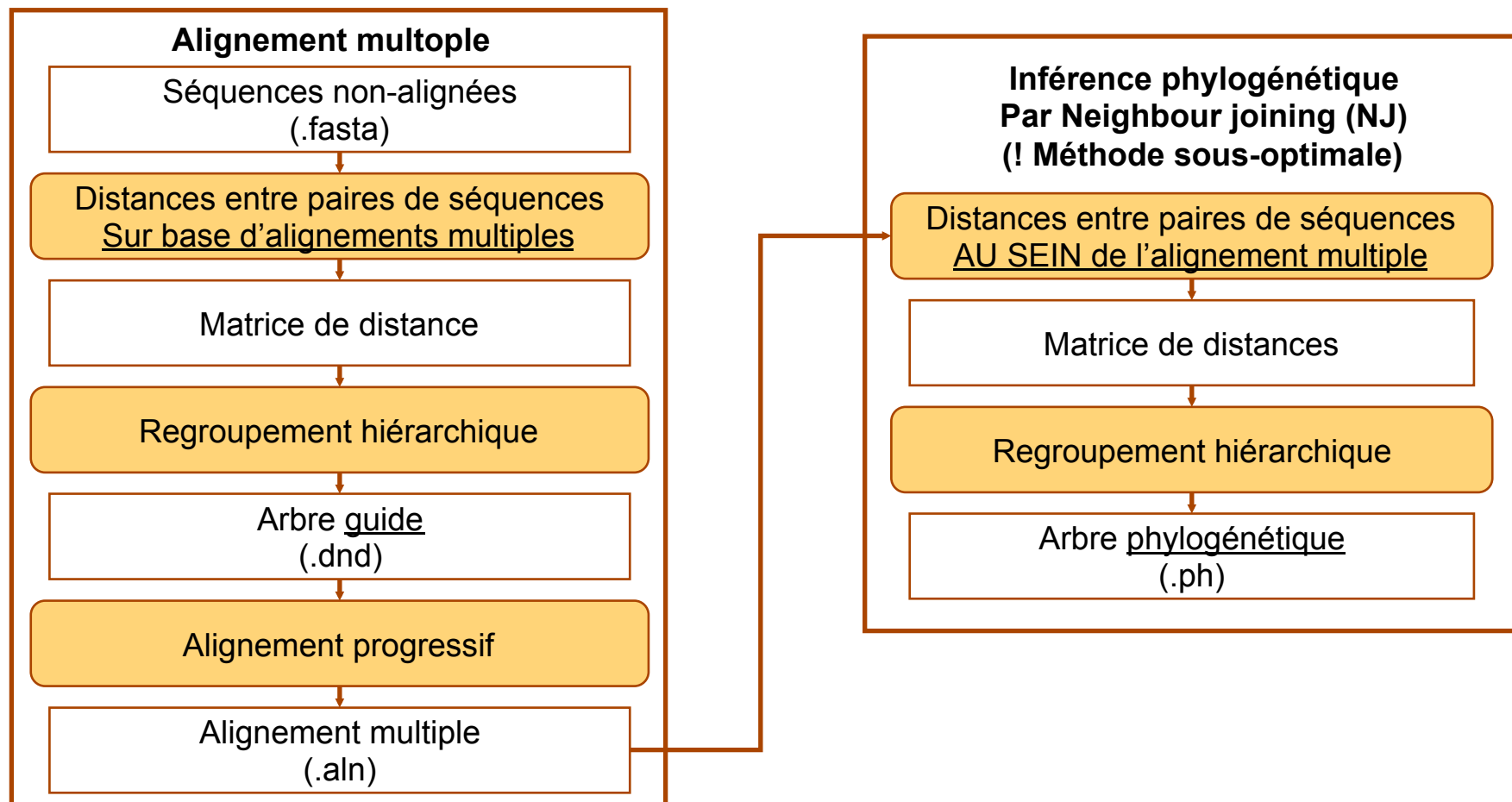
Alignement progressif – 3^{ème} étape: alignement multiple

- On construit un alignement multiple en incorporant progressivement les séquences selon leur ordre de regroupement dans l'arbre guide.

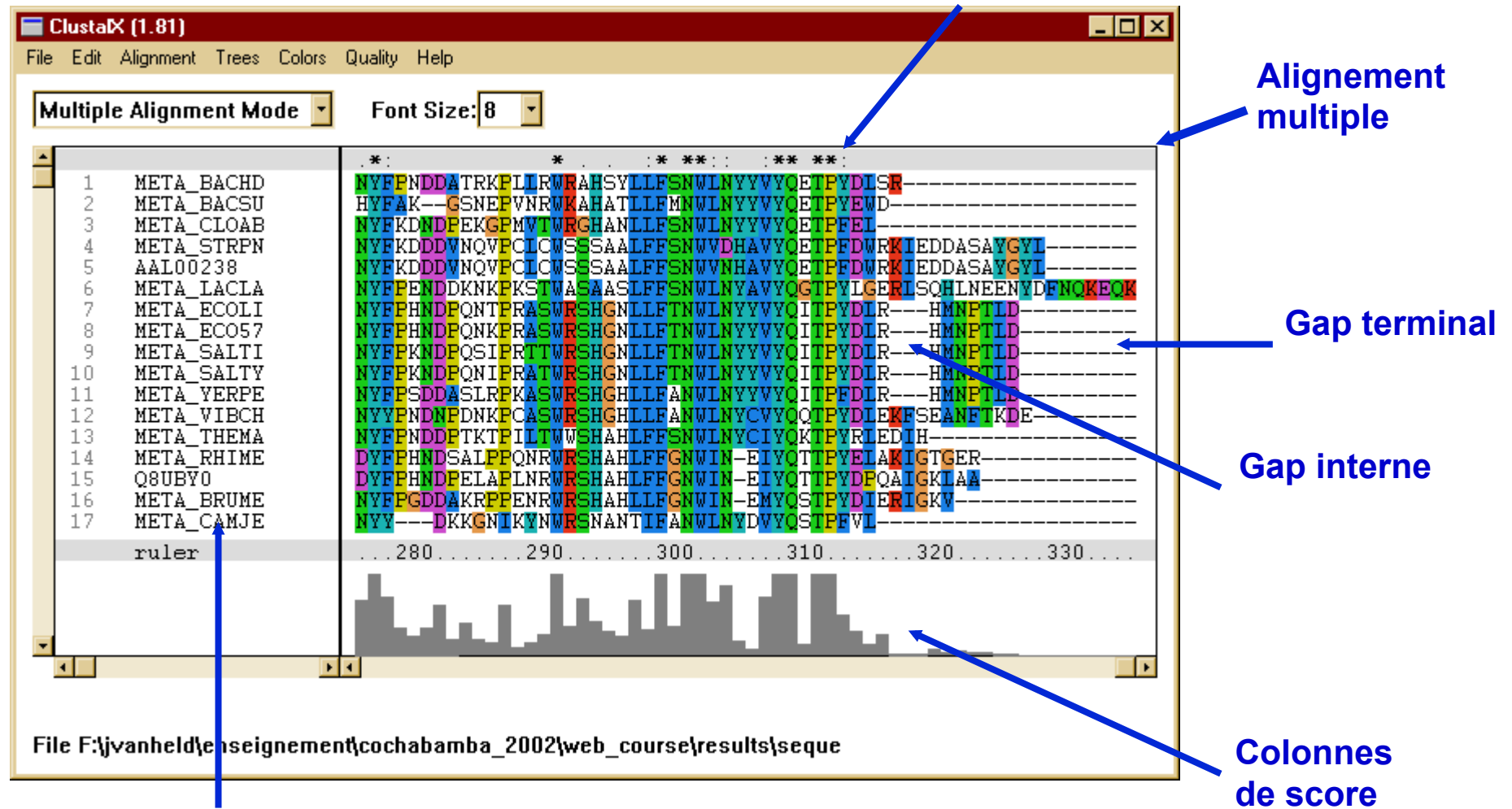


Alignement progressif et inférence phylogénétique par Neighbour joining

- Attention ! L'arbre guide n'est pas un arbre phylogénétique.
 - Son rôle se limite à proposer un ordre pour construire l'alignement multiple.
 - Il n'a pas pour vocation de prédire l'histoire évolutive des divergences entre séquences.
- On peut éventuellement, dans un second temps, inférer un arbre phylogénétique à partir de l'alignement multiple, par la méthode « **Neighbor joining** ».
 - Cette méthode est cependant sous-optimale pour l'inférence phylogénétique.



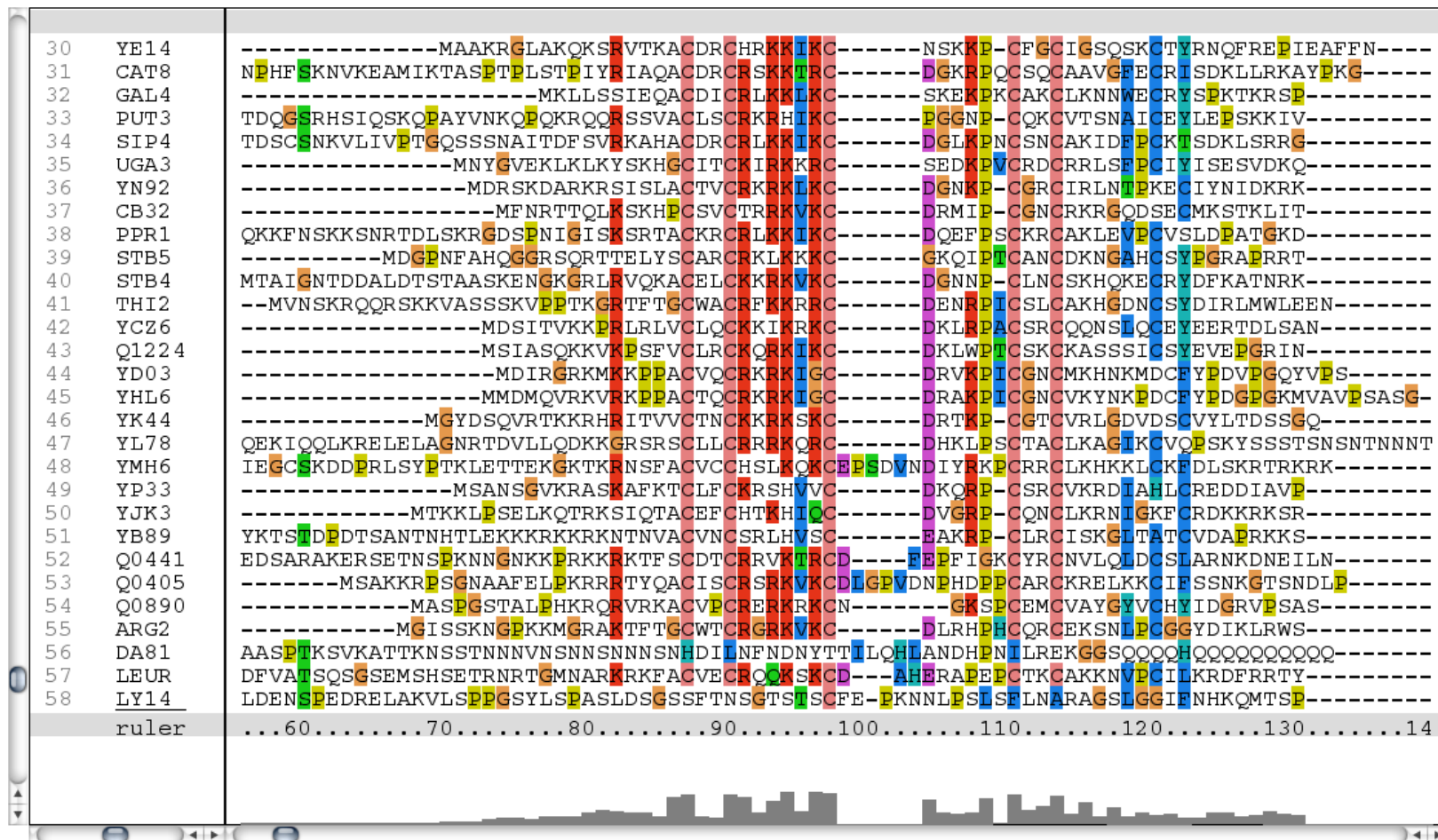
Alignement multiple global : Homoserine-O-dehydrogenase



Identifiants de séquences

Alignement des protéines à domaine Zinc cluster

- Un cas difficile: aligner les protéines contenant le domaine Zinc cluster Zn(2)Cys(6)
 - La région conservée est restreinte au comaind Zinc cluster
 - Ce domaine n'est pas composé de résidus contigus: il contient des positions variables et conservées interspersées.
 - L'alignement met en évidence 5 cystéines parmi les 6 qui donnent son nom au domaine.



Local multiple alignment

Alignement progressif - résumé

- Temps de calcul pour N séquences à aligner
 - Construction de l'arbre-guide quadratique: proportionnel à $N(N-1)/2 = N^2/2 - N/2$
 - Alignement des séquences linéaire : proportionnel à N
- Méthode heuristique
 - Permet de traiter quelques dizaines de séquences en un temps raisonnable
 - Ne peut cependant pas garantir de retourner la réponse optimale (celle qui maximise le score d'alignement).
- Le programme clustalX
 - fournit une interface interactive à l'algorithme d'alignement progressif clustalW.
 - En outre, il présente des fonctionnalités additionnelles:
 - Marquage des segments de séquences mal alignés (« low-scoring segments »).
 - L'alignement peut être raffiné manuellement
 - Réalignement de quelques séquences sélectionnées par l'utilisateur
 - Réalignement de colonnes sélectionnées par l'utilisateur.

Bioinformatics

***La détection de motifs
dans les séquences biologiques***

Matrices de profil (matrices de scores spécifiques de la position) (=position-specific scoring matrices, PSSM)

- En partant d'un alignement multiple, on peut construire une matrice qui indique les résidus les plus représentatifs de chaque position: **matrice de scores spécifiques de la position** (en anglais: Position-Specific Scoring Matrix, PSSM).
 - Chaque colonne représente une position de l'alignement
 - Chaque ligne correspond à un résidu (20 lignes pour les motifs protéiques, 4 lignes pour les motifs nucléiques).
 - Les valeurs indiquent le nombre d'occurrences de chaque résidu à chaque position de l'alignement multiple.

Construction d'une matrice PSSM

Alignement multiple

W	S	K	T	N	V	T	S	T	L	H	I	C	W	G	A	Q	A	G	L
W	S	K	T	N	V	T	S	T	L	H	I	C	W	G	A	Q	A	G	L
W	T	Q	S	H	V	H	R	T	L	N	I	C	W	A	A	Q	A	A	V
F	L	K	Q	N	V	T	S	S	M	Y	I	C	W	G	A	M	A	A	L
W	S	V	T	N	V	T	S	T	I	H	I	C	W	G	A	Q	A	G	L
W	S	K	D	H	V	T	S	T	L	F	V	C	W	A	V	Q	A	A	L
W	S	K	S	H	V	Y	S	S	L	H	I	C	W	G	A	Q	A	A	L
W	T	T	T	N	V	H	S	T	L	N	V	C	W	G	G	M	A	A	V
W	A	K	D	H	V	T	S	T	L	F	V	C	W	A	V	Q	A	A	L
W	A	K	D	H	V	Y	S	T	L	H	I	C	W	G	A	Q	A	G	L
W	S	R	H	N	V	Y	S	T	M	F	I	C	W	A	A	Q	A	G	L
W	A	K	A	H	V	T	S	T	L	F	I	C	W	A	A	Q	A	G	L
W	A	K	E	H	V	T	S	T	L	Y	V	C	W	A	V	Q	A	A	L
W	T	Q	T	N	V	H	S	T	L	N	V	C	W	G	A	M	A	A	I
W	S	K	T	H	V	Y	S	T	L	H	I	C	W	G	A	Q	A	G	L

Matrice de scores spécifiques de la position (occurrences)

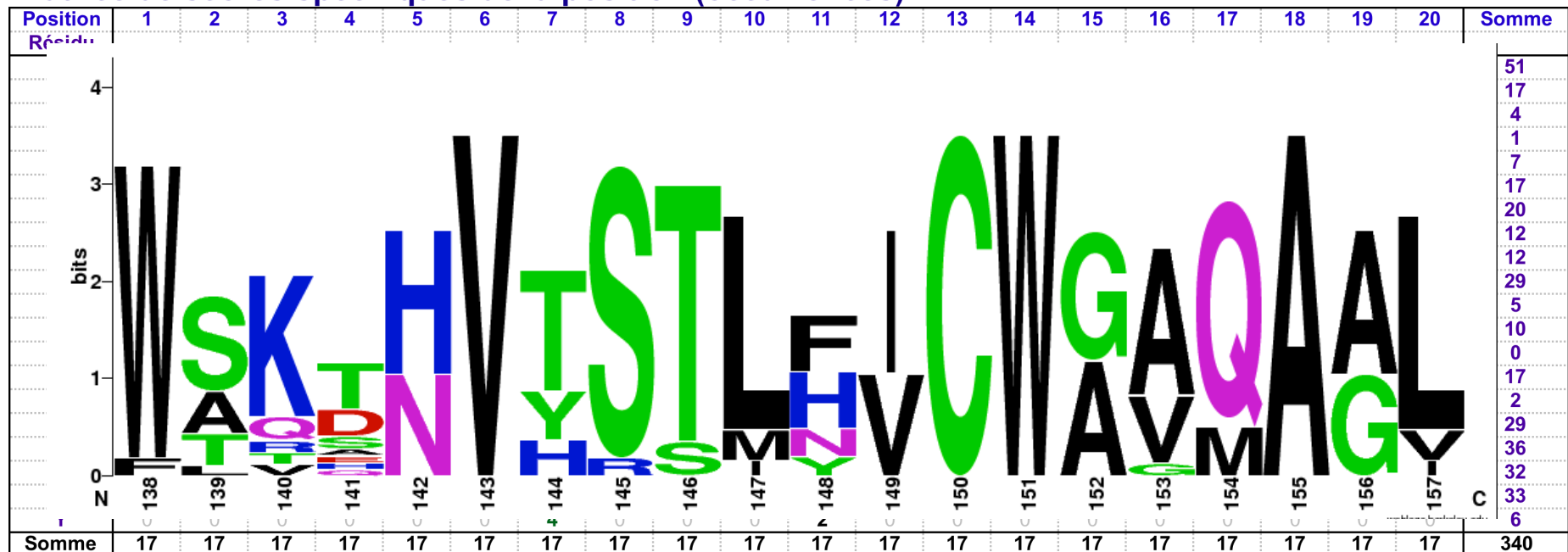
[illegible]

La représentation « logo » d'une matrice de profil

Alignement multiple

W	S	K	T	N	V	T	S	T	L	H	I	C	W	G	A	Q	A	G	L
W	S	K	T	N	V	T	S	T	L	H	I	C	W	G	A	Q	A	G	L
W	T	Q	S	N	V	T	S	T	M	N	I	C	W	A	A	M	A	V	L
F	L	K	Q	N	V	T	S	T	I	Y	I	C	W	G	A	Q	A	A	L
W	S	V	T	N	V	T	S	T	L	H	I	C	W	G	A	Q	A	A	L
W	S	K	D	H	V	T	S	T	L	F	V	C	W	A	V	Q	A	A	L
W	S	K	D	H	V	T	S	T	L	H	I	C	W	G	A	Q	A	A	L
W	T	T	S	N	V	H	S	T	L	N	V	C	W	A	G	M	A	A	V
W	A	K	D	H	V	T	S	T	L	F	V	C	W	A	V	Q	A	A	L
W	A	K	D	H	V	T	S	T	L	F	I	C	W	A	V	Q	A	A	L
W	S	K	T	H	V	Y	S	T	M	F	I	C	W	A	A	Q	A	G	L
W	A	R	H	N	V	T	S	T	L	Y	I	C	W	A	A	Q	A	G	L
W	A	K	E	H	V	T	S	T	L	F	V	C	W	A	V	Q	A	A	L
W	T	Q	T	N	V	H	S	T	L	N	V	C	W	G	A	M	A	A	I
W	S	K	T	H	V	Y	S	T	L	H	I	C	W	G	A	Q	A	G	L

Matrice de scores spécifiques de la position (occurrences)



Conversion de la matrice d'occurrences en matrice position-poids (position-weight matrix)

Matrice de scores spécifiques de la position (occurrences)

Position Résidu	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Somme	Fréq
A	0	4	0	1	0	0	0	0	0	0	0	0	0	0	8	11	0	17	10	0	51.00	0.15
C	0	0	0	0	0	0	0	0	0	0	0	0	17	0	0	0	0	0	0	0	17.00	0.05
D	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4.00	0.01
E	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.00	0.00
F	1	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	7.00	0.02
G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	1	0	0	7	0	17.00	0.05
H	0	0	0	1	10	0	3	0	0	0	6	0	0	0	0	0	0	0	0	0	20.00	0.06
I	0	0	0	0	0	0	0	0	0	1	0	10	0	0	0	0	0	0	0	1	12.00	0.04
K	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12.00	0.04
L	0	1	0	0	0	0	0	0	0	14	0	0	0	0	0	0	0	0	0	14	29.00	0.09
M	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	3	0	0	0	5.00	0.01
N	0	0	0	0	7	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	10.00	0.03
P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00	0.00
Q	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0	0	14	0	0	0	17.00	0.05
R	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	2.00	0.01
S	0	9	0	2	0	0	0	16	2	0	0	0	0	0	0	0	0	0	0	0	29.00	0.09
T	0	3	1	7	0	0	10	0	15	0	0	0	0	0	0	0	0	0	0	0	36.00	0.11
V	0	0	1	0	0	17	0	0	0	0	0	7	0	0	5	0	0	0	0	2	32.00	0.09
W	16	0	0	0	0	0	0	0	0	0	0	0	0	17	0	0	0	0	0	0	33.00	0.10
Y	0	0	0	0	0	0	4	0	0	0	2	0	0	0	0	0	0	0	0	0	6.00	0.02
Somme	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	340	1.00

Matrice de poids

Position Residue	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
A	-1.72	0.19	-1.72	-0.39	-1.72	-1.72	-1.72	-1.72	-1.72	-1.72	-1.72	-1.72	-1.72	-1.72	0.49	0.63	-1.72	0.82	0.59	-1.72
C	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	1.28	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26
D	-0.70	-0.70	-0.70	1.21	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70
E	-0.30	-0.30	-0.30	1.02	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30
F	0.42	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	1.18	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90
G	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	1.00	0.07	-1.26	-1.26	0.89	-1.26
H	-1.32	-1.32	-1.32	0.00	0.98	-1.32	0.46	-1.32	-1.32	-1.32	0.76	-1.32	-1.32	-1.32	-1.32	-1.32	-1.32	-1.32	-1.32	-1.32
I	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	0.21	-1.11	1.19	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	0.21
K	-1.11	-1.11	1.27	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11
L	-1.48	-0.15	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	0.97	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	0.97
M	-0.78	-0.78	-0.78	-0.78	-0.78	-0.78	-0.78	-0.78	-0.78	0.83	-0.78	-0.78	-0.78	-0.78	-0.78	-0.78	1.01	-0.78	-0.78	-0.78
N	-1.04	-1.04	-1.04	-1.04	1.11	-1.04	-1.04	-1.04	-1.04	-1.04	0.74	-1.04	-1.04	-1.04	-1.04	-1.04	-1.04	-1.04	-1.04	-1.04
P	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q	-1.26	-1.26	0.36	0.07	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	1.19	-1.26	-1.26	-1.26
R	-0.48	-0.48	0.85	-0.48	-0.48	-0.48	-0.48	0.85	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48
S	-1.48	0.78	-1.48	0.14	-1.48	-1.48	-1.48	1.03	0.14	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48
T	-1.57	0.22	-0.25	0.58	-1.57	-1.57	0.73	-1.57	0.91	-1.57	-1.57	-1.57	-1.57	-1.57	-1.57	-1.57	-1.57	-1.57	-1.57	-1.57
V	-1.52	-1.52	-0.20	-1.52	-1.52	1.01	-1.52	-1.52	-1.52	-1.52	-1.52	0.63	-1.52	-1.52	-1.52	0.49	-1.52	-1.52	-1.52	0.09
W	0.98	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53	1.00	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53
Y	-0.85	-0.85	-0.85	-0.85	-0.85	-0.85	1.06	-0.85	-0.85	-0.85	0.77	-0.85	-0.85	-0.85	-0.85	-0.85	-0.85	-0.85	-0.85	-0.85
Somme	-17.8	-14.4	-13.7	-10.7	-17.2	-19.1	-15.7	-17.8	-17.6	-16.3	-14.1	-17.2	-19.1	-19.1	-17.2	-16	-17.4	-19.1	-17.2	-16.3

$$f'_{i,j} = \frac{n_{i,j} + p_i k}{\sum_{r=1}^A n_{r,j} + k}$$

$$W_{i,j} = \ln \left(\frac{f'_{i,j}}{p_i} \right)$$

Assignment d'un score à une séquence avec une matrice position-poids

Weight matrix

Position Residue	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
A	-1.72	0.19	-1.72	-0.39	-1.72	-1.72	-1.72	-1.72	-1.72	-1.72	-1.72	-1.72	-1.72	-1.72	0.49	0.63	-1.72	0.82	0.59	-1.72
C	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	1.28	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26
D	-0.70	-0.70	-0.70	1.21	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70
E	-0.30	-0.30	-0.30	1.02	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30
F	0.42	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	1.18	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90
G	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	1.00	0.07	-1.26	-1.26	0.89	-1.26
H	-1.32	-1.32	-1.32	0.00	0.98	-1.32	0.46	-1.32	-1.32	-1.32	0.76	-1.32	-1.32	-1.32	-1.32	-1.32	-1.32	-1.32	-1.32	-1.32
I	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	0.21	-1.11	1.19	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	0.21
K	-1.11	-1.11	1.27	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11
L	-1.48	-0.15	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	0.97	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	0.97
M	-0.78	-0.78	-0.78	-0.78	-0.78	-0.78	-0.78	-0.78	-0.78	0.83	-0.78	-0.78	-0.78	-0.78	-0.78	-0.78	1.01	-0.78	-0.78	-0.78
N	-1.04	-1.04	-1.04	-1.04	1.11	-1.04	-1.04	-1.04	-1.04	-1.04	0.74	-1.04	-1.04	-1.04	-1.04	-1.04	-1.04	-1.04	-1.04	-1.04
P	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q	-1.26	-1.26	0.36	0.07	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	1.19	-1.26	-1.26	-1.26
R	-0.48	-0.48	0.85	-0.48	-0.48	-0.48	-0.48	0.85	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48
S	-1.48	0.78	-1.48	0.14	-1.48	-1.48	-1.48	1.03	0.14	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48
T	-1.57	0.22	-0.25	0.58	-1.57	-1.57	0.73	-1.57	0.91	-1.57	-1.57	-1.57	-1.57	-1.57	-1.57	-1.57	-1.57	-1.57	-1.57	-1.57
V	-1.52	-1.52	-0.20	-1.52	-1.52	1.01	-1.52	-1.52	-1.52	-1.52	-1.52	0.63	-1.52	-1.52	-1.52	0.49	-1.52	-1.52	-1.52	0.09
W	0.98	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53	1.00	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53
Y	-0.85	-0.85	-0.85	-0.85	-0.85	-0.85	1.06	-0.85	-0.85	-0.85	0.77	-0.85	-0.85	-0.85	-0.85	-0.85	-0.85	-0.85	-0.85	-0.85
sum	-17.8	-14.4	-13.7	-10.7	-17.2	-19.1	-15.7	-17.8	-17.6	-16.3	-14.1	-17.2	-19.1	-19.1	-17.2	-16	-17.4	-19.1	-17.2	-16.3

Sequence	L	W	A	K	D	H	V	T	S	T	M	F	V	C	W	A	V	M	A	A	SUM
Score	-1.48	-1.53	-1.72	-1.11	-0.7	-1.32	-1.52	-1.57	0.136	-1.57	-0.78	-0.9	-1.52	-1.26	-1.53	0.628	-1.52	-0.78	0.587	-1.72	-21.1626

Assignment d'un score à une séquence avec une matrice position-poids

Weight matrix

Position Residue	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
A	-1.72	0.19	-1.72	-0.39	-1.72	-1.72	-1.72	-1.72	-1.72	-1.72	-1.72	-1.72	-1.72	-1.72	0.49	0.63	-1.72	0.82	0.59	-1.72
C	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	1.28	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26
D	-0.70	-0.70	-0.70	1.21	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70
E	-0.30	-0.30	-0.30	1.02	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30
F	0.42	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	1.18	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90
G	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	1.00	0.07	-1.26	-1.26	0.89	-1.26
H	-1.32	-1.32	-1.32	0.00	0.98	-1.32	0.46	-1.32	-1.32	-1.32	0.76	-1.32	-1.32	-1.32	-1.32	-1.32	-1.32	-1.32	-1.32	-1.32
I	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	0.21	-1.11	1.19	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	0.21
K	-1.11	-1.11	1.27	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11
L	-1.48	-0.15	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	0.97	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	0.97
M	-0.78	-0.78	-0.78	-0.78	-0.78	-0.78	-0.78	-0.78	-0.78	0.83	-0.78	-0.78	-0.78	-0.78	-0.78	-0.78	1.01	-0.78	-0.78	-0.78
N	-1.04	-1.04	-1.04	-1.04	1.11	-1.04	-1.04	-1.04	-1.04	-1.04	0.74	-1.04	-1.04	-1.04	-1.04	-1.04	-1.04	-1.04	-1.04	-1.04
P	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q	-1.26	-1.26	0.36	0.07	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	1.19	-1.26	-1.26	-1.26
R	-0.48	-0.48	0.85	-0.48	-0.48	-0.48	-0.48	0.85	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48
S	-1.48	0.78	-1.48	0.14	-1.48	-1.48	-1.48	1.03	0.14	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48
T	-1.57	0.22	-0.25	0.58	-1.57	-1.57	0.73	-1.57	0.91	-1.57	-1.57	-1.57	-1.57	-1.57	-1.57	-1.57	-1.57	-1.57	-1.57	-1.57
V	-1.52	-1.52	-0.20	-1.52	-1.52	1.01	-1.52	-1.52	-1.52	-1.52	-1.52	0.63	-1.52	-1.52	-1.52	0.49	-1.52	-1.52	-1.52	0.09
W	0.98	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53	1.00	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53
Y	-0.85	-0.85	-0.85	-0.85	-0.85	-0.85	1.06	-0.85	-0.85	-0.85	0.77	-0.85	-0.85	-0.85	-0.85	-0.85	-0.85	-0.85	-0.85	-0.85
sum	-17.8	-14.4	-13.7	-10.7	-17.2	-19.1	-15.7	-17.8	-17.6	-16.3	-14.1	-17.2	-19.1	-19.1	-17.2	-16	-17.4	-19.1	-17.2	-16.3

Sequence	W	A	K	D	H	V	T	S	T	M	F	V	C	W	A	V	M	A	A	L	SUM
Score	0.975	0.192	1.268	1.21	0.981	1.014	0.735	1.029	0.91	0.835	1.18	0.631	1.277	1.001	0.491	0.486	1.007	0.817	0.587	0.972	17.59818

Assignment d'un score à une séquence avec une matrice position-poids

Weight matrix

Position Residue	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
A	-1.72	0.19	-1.72	-0.39	-1.72	-1.72	-1.72	-1.72	-1.72	-1.72	-1.72	-1.72	-1.72	-1.72	0.49	0.63	-1.72	0.82	0.59	-1.72
C	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	1.28	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26
D	-0.70	-0.70	-0.70	1.21	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70	-0.70
E	-0.30	-0.30	-0.30	1.02	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30	-0.30
F	0.42	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	1.18	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90
G	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	1.00	0.07	-1.26	-1.26	0.89	-1.26
H	-1.32	-1.32	-1.32	0.00	0.98	-1.32	0.46	-1.32	-1.32	-1.32	0.76	-1.32	-1.32	-1.32	-1.32	-1.32	-1.32	-1.32	-1.32	-1.32
I	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	0.21	-1.11	1.19	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	0.21
K	-1.11	-1.11	1.27	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11	-1.11
L	-1.48	-0.15	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	0.97	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	0.97
M	-0.78	-0.78	-0.78	-0.78	-0.78	-0.78	-0.78	-0.78	-0.78	0.83	-0.78	-0.78	-0.78	-0.78	-0.78	-0.78	1.01	-0.78	-0.78	-0.78
N	-1.04	-1.04	-1.04	-1.04	1.11	-1.04	-1.04	-1.04	-1.04	-1.04	0.74	-1.04	-1.04	-1.04	-1.04	-1.04	-1.04	-1.04	-1.04	-1.04
P	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q	-1.26	-1.26	0.36	0.07	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	-1.26	1.19	-1.26	-1.26	-1.26
R	-0.48	-0.48	0.85	-0.48	-0.48	-0.48	-0.48	0.85	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48	-0.48
S	-1.48	0.78	-1.48	0.14	-1.48	-1.48	-1.48	1.03	0.14	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48	-1.48
T	-1.57	0.22	-0.25	0.58	-1.57	-1.57	0.73	-1.57	0.91	-1.57	-1.57	-1.57	-1.57	-1.57	-1.57	-1.57	-1.57	-1.57	-1.57	-1.57
V	-1.52	-1.52	-0.20	-1.52	-1.52	1.01	-1.52	-1.52	-1.52	-1.52	-1.52	0.63	-1.52	-1.52	-1.52	0.49	-1.52	-1.52	-1.52	0.09
W	0.98	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53	1.00	-1.53	-1.53	-1.53	-1.53	-1.53	-1.53
Y	-0.85	-0.85	-0.85	-0.85	-0.85	-0.85	1.06	-0.85	-0.85	-0.85	0.77	-0.85	-0.85	-0.85	-0.85	-0.85	-0.85	-0.85	-0.85	-0.85
sum	-17.8	-14.4	-13.7	-10.7	-17.2	-19.1	-15.7	-17.8	-17.6	-16.3	-14.1	-17.2	-19.1	-19.1	-17.2	-16	-17.4	-19.1	-17.2	-16.3

Sequence	A	K	D	H	V	T	S	T	M	F	V	C	W	A	V	M	A	A	L	V	SUM
Score	-1.72	-1.11	-0.7	1E-16	-1.52	-1.57	-1.48	-1.57	-0.78	-0.9	-1.52	-1.26	-1.53	-1.72	-1.52	-0.78	-1.72	0.817	-1.48	0.094	-21.9422

- PSI-blast signifie « Position-specific iterated BLAST » (Altschul et al., 1997)
- Principe : à partir d'une séquence requête, on collecte les séquences similaires sur base d'une matrice position-poids.
- Etapes
 1. Collecte de protéines similaires à la séquence requête par simple BLAST
 2. Alignement multiple des séquences collectées
 3. Construction d'une matrice de score spécifique de la position (PSSM) à partir de cet alignement
 4. Utilisation de la matrice pour scanner la base de données, et récolter une nouvelle série de séquences similaires
 5. Itérations à partir de l'étape 2
- La recherche par PSSM augmente la **sensibilité** de la recherche, et offre un **meilleur pouvoir de généralisation**.
 - Au départ d'une seule séquence, on collecte les séquences de la même famille, et on procède ensuite à une recherche sur base d'un motif (matrice PSSM) qui représente l'ensemble de cette famille.
 - Le motif est lui-même progressivement

Références

■ Matrices de substitutions

- 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.

■ Algorithmes d'alignement de séquences

- Needleman-Wunsch (pairwise, global)
 - Needleman, S. B. & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48, 443-53.
- Smith-Waterman (pairwise, local)
 - Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. J Mol Biol 147, 195-7.
- FastA (database searches, pairwise, local)
 - W. R. Pearson and D. J. Lipman. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA, 85:2444--2448, 1988.
- BLAST (database searches, pairwise, local)
 - S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. A basic local alignment search tool. J. Mol. Biol., 215:403--410, 1990.
 - S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs Nucleic Acids Res., 25:3389--3402, 1997.
- Clustal (multiple, global)
 - Higgins, D. G. & Sharp, P. M. (1988). CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene 73, 237-44.
 - Higgins, D. G., Thompson, J. D. & Gibson, T. J. (1996). Using CLUSTAL for multiple sequence alignments. Methods Enzymol 266, 383-402.
- Dialign (multiple, local)
 - Morgenstern, B., Frech, K., Dress, A. & Werner, T. (1998). DIALIGN: finding local similarities by multiple sequence alignment. Bioinformatics 14, 290-4.
- MUSCLE (multiple local)