Analyse de séquences

Partie III : Alignement multiple

Sèverine Bérard

AMAP - Université Montpellier 2

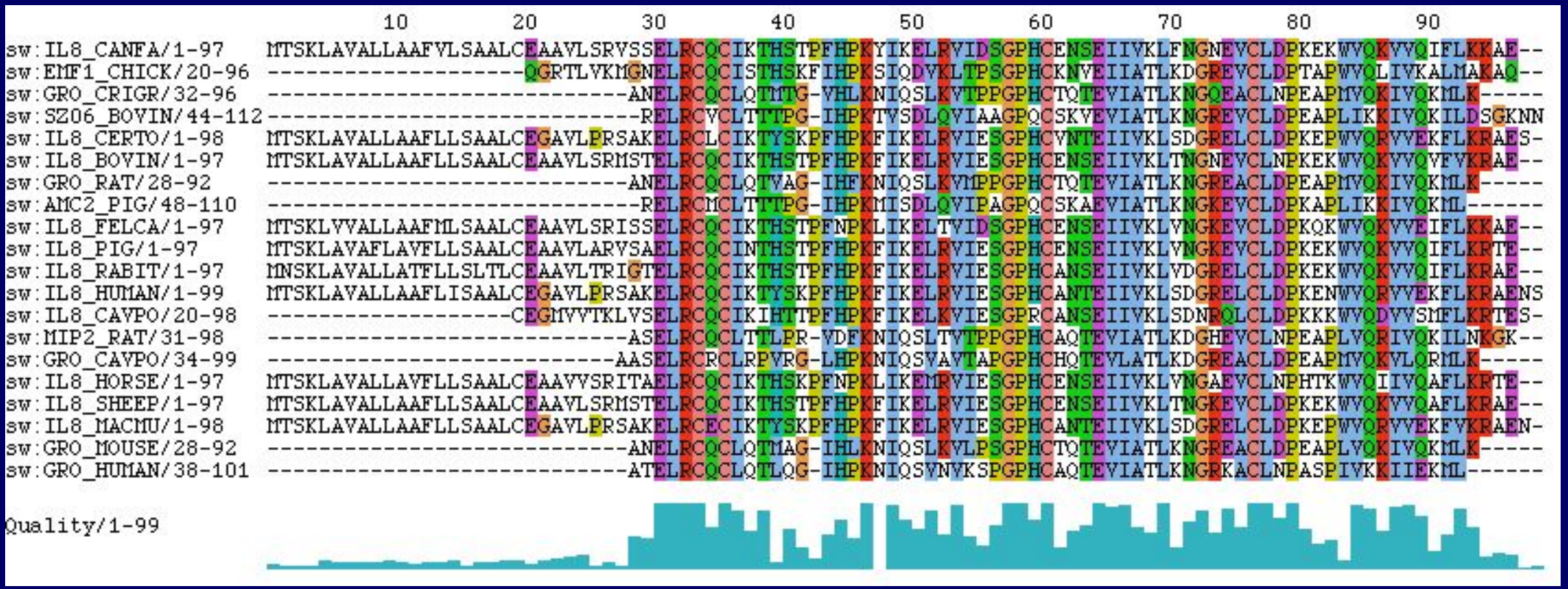
Alignement multiple

- Permet d'aligner plusieurs séquences simultanément
- En général pour les protéines
- Alignements faits à la main par les experts biologistes
- Généralisation naturelle de l'alignement 2 à 2 mais beaucoup plus complexe en terme de calcul

Problème : Étant données k séquences $s_1, s_2, \ldots, s_k$, trouver le meilleur alignement multiple pour ces séquences

- Ce pb est NP-complet ($\Rightarrow \nexists$ de solution exacte en tps raisonnable)

Exemples d'alignement



Intérêt des alignements multiples

- Recherche dans les banques $\Rightarrow$ plrs séquences similaires à la requête
  Il est naturel de vouloir aligner ces séquences entre elles

- MSA détectent les régions qui ont été conservées lors de l'évolution
  Très svt des domaines associés à une fonction clé de la molécule

- Plusieurs protéines de fonctions similaires dans différentes espèces
  $\rightarrow$ Quelles parties semblables ? $\Rightarrow$ Consensus/Profil
  $\rightarrow$ Quelles parties différentes ?

- Permet de trouver d'autres membres d'une famille de protéines

- Séquençage de génomes (assemblage, recouvrement EST)

- Point de départ pour les analyses phylogénétiques

Alignements multiples plus informatifs que les ali. de 2 séq.

Les 2 grands types de MSA

- Un MSA peut être global ou local (comme un alignement de 2 séquences)

- Global : l'alignement 2 à 2 est étendu pour inclure 3 séq. ou plus
  Des protéines de $\neq$ organismes peuvent être conservées sur toute la longueur si elles assurent une fonction biologique importante

  Logiciels : CLUSTALW, MULTALIN, T-COFFEE, DIALIGN, . . .

- Local : recherche de domaines/régions conservés
  Les domaines fonctionnels de protéines peuvent être conservés tandis que le reste de la séquence diverge

  Logiciels : BLOCKS Web site, eMOTIF, GIBBS, HMMER, . . .

Sensibilité et sélectivité

- Notions utiles pour comparer des méthodes
- Par exemple : on prend une séquence membre d'une famille de protéine et on l'utilise comme requête, les algorithmes retourne une liste de *hits* et on coupe à un certain seuil

	Séq. membres famille	Séq. non membres
Séq. trouvé au-dessus du seuil	Vrais positifs	Faux positifs
Séq. trouvé en dessous du seuil	Faux négatifs	Vrai négatifs

- $\text{Sensibilité} = \frac{\text{Vrais positifs}}{\text{Vrais positifs}+\text{Faux négatifs}}$
- $\text{Sélectivité} = \frac{\text{Vrais positifs}}{\text{Vrais positifs}+\text{Faux positifs}}$

Comment quantifier la qualité d'un alignement multiple ?

- Généraliser une fonction de score pour un alignement de k séq.
- Le score d'un MSA est la somme des scores de ses colonnes (ici les colonnes sont de hauteur k)
- Les colonnes sont considérées indépendantes
- Fonction à k paramètres ?
  Si alphabet de taille $\alpha \Rightarrow \alpha^k$ colonnes possibles ($21^5 > 4$ millions)
  On ne peut pas associer un coût à chaque colonne ! ? !
- Mesure raisonnable S, quelles propriétés ?

Propriétés d'une méthode de score

1. Même score pour les colonnes contenant les mêmes caractères (indépendamment de l'ordre)

$$S(I,-,I,V) = S(V,I,I,-) = S(V,I,-,I) = S(V,-,I,I) = \cdots$$

2. Récompense les colonnes avec beaucoup de résidus identiques ou similaires

3. Pénalise les colonnes avec des résidus différents et des espaces

- Plusieurs méthodes de score : méthode SP, méthode basée sur la phylogénie (arbre ou étoile), contenu en information, . . .

Méthode de score SP

- SP pour Sum of Pairs, c'est la méthode la plus utilisée

- Somme des scores de toutes les paires d'a.a. possibles dans une colonne ($p(-,-) = 0$), puis somme des scores des colonnes

$$SP(I,-,I,V) = p(I,-) + p(I,I) + p(I,V) + p(-,I) + p(-,V) + p(I,V)$$

- Exemple avec BLOSUM62 et indel=-2

1	N	N	N	I	V
2	N	N	N	I	V
3	N	N	N	–	V
4	N	N	C	I	V
5	N	C	C	I	V

- Il y a $\frac{k(k-1)}{2}$ paires/scores à calculer pour chaque colonne

Méthode de score SP

- SP pour Sum of Pairs, c'est la méthode la plus utilisée

- Somme des scores de toutes les paires d'a.a. possibles dans une colonne ($p(-,-) = 0$), puis somme des scores des colonnes

$$SP(I,-,I,V) = p(I,-) + p(I,I) + p(I,V) + p(-,I) + p(-,V) + p(I,V)$$

- Exemple avec BLOSUM62 et indel=-2

$p(N,N) = 6$

1	N	N	N	I	V
2	N	N	N	I	V
3	N	N	N	–	V
4	N	N	C	I	V
5	N	C	C	I	V
					60

- Il y a $\frac{k(k-1)}{2}$ paires/scores à calculer pour chaque colonne

Méthode de score SP

- SP pour Sum of Pairs, c'est la méthode la plus utilisée

- Somme des scores de toutes les paires d'a.a. possibles dans une colonne ($p(-,-) = 0$), puis somme des scores des colonnes

$$SP(I,-,I,V) = p(I,-) + p(I,I) + p(I,V) + p(-,I) + p(-,V) + p(I,V)$$

- Exemple avec BLOSUM62 et indel=-2

$p(N,N) = 6$

$p(N,C) = -3$

1	N	N	N	I	V
2	N	N	N	I	V
3	N	N	N	–	V
4	N	N	C	I	V
5	N	C	C	I	V
	60	24			

- Il y a $\frac{k(k-1)}{2}$ paires/scores à calculer pour chaque colonne

Méthode de score SP

- SP pour Sum of Pairs, c'est la méthode la plus utilisée

- Somme des scores de toutes les paires d'a.a. possibles dans une colonne ($p(-,-) = 0$), puis somme des scores des colonnes

$$SP(I,-,I,V) = p(I,-) + p(I,I) + p(I,V) + p(-,I) + p(-,V) + p(I,V)$$

- Exemple avec BLOSUM62 et indel=-2

$p(N,N) = 6$

$p(N,C) = -3$

$p(C,C) = 9$

1	N	N	N	I	V
2	N	N	N	I	V
3	N	N	N	–	V
4	N	N	C	I	V
5	N	C	C	I	V
	60	24	9		

- Il y a $\frac{k(k-1)}{2}$ paires/scores à calculer pour chaque colonne

Méthode de score SP

- SP pour Sum of Pairs, c'est la méthode la plus utilisée

- Somme des scores de toutes les paires d'a.a. possibles dans une colonne ($p(-,-) = 0$), puis somme des scores des colonnes

$$SP(I,-,I,V) = p(I,-) + p(I,I) + p(I,V) + p(-,I) + p(-,V) + p(I,V)$$

- Exemple avec BLOSUM62 et indel=-2

$p(N,N) = 6$

$p(N,C) = -3$

$p(C,C) = 9$

$p(I,I) = 4$

$p(I,-) = -2$

1	N	N	N	I	V
2	N	N	N	I	V
3	N	N	N	–	V
4	N	N	C	I	V
5	N	C	C	I	V
	60	24	9	16	

- Il y a $\frac{k(k-1)}{2}$ paires/scores à calculer pour chaque colonne

Méthode de score SP

- SP pour Sum of Pairs, c'est la méthode la plus utilisée

- Somme des scores de toutes les paires d'a.a. possibles dans une colonne ($p(-,-) = 0$), puis somme des scores des colonnes

$$SP(I,-,I,V) = p(I,-) + p(I,I) + p(I,V) + p(-,I) + p(-,V) + p(I,V)$$

- Exemple avec BLOSUM62 et indel=-2

$p(N,N) = 6$
$p(N,C) = -3$
$p(C,C) = 9$
$p(I,I) = 4$
$p(I,-) = -2$
$p(V,V) = 4$

1	N	N	N	I	V
2	N	N	N	I	V
3	N	N	N	–	V
4	N	N	C	I	V
5	N	C	C	I	V
	60	24	9	16	40

- Il y a $\frac{k(k-1)}{2}$ paires/scores à calculer pour chaque colonne

Méthode de score SP

- SP pour Sum of Pairs, c'est la méthode la plus utilisée

- Somme des scores de toutes les paires d'a.a. possibles dans une colonne ($p(-,-) = 0$), puis somme des scores des colonnes

$$SP(I,-,I,V) = p(I,-) + p(I,I) + p(I,V) + p(-,I) + p(-,V) + p(I,V)$$

- Exemple avec BLOSUM62 et indel=-2

$p(N,N) = 6$
$p(N,C) = -3$
$p(C,C) = 9$
$p(I,I) = 4$
$p(I,-) = -2$
$p(V,V) = 4$

1	N	N	N	I	V	
2	N	N	N	I	V	
3	N	N	N	–	V	
4	N	N	C	I	V	
5	N	C	C	I	V	
	60	24	9	16	40	=

- Il y a $\frac{k(k-1)}{2}$ paires/scores à calculer pour chaque colonne

Méthode de score SP

- SP pour Sum of Pairs, c'est la méthode la plus utilisée

- Somme des scores de toutes les paires d'a.a. possibles dans une colonne ($p(-,-) = 0$), puis somme des scores des colonnes

$$SP(I,-,I,V) = p(I,-) + p(I,I) + p(I,V) + p(-,I) + p(-,V) + p(I,V)$$

- Exemple avec BLOSUM62 et indel=-2

$p(N,N) = 6$
$p(N,C) = -3$
$p(C,C) = 9$
$p(I,I) = 4$
$p(I,-) = -2$
$p(V,V) = 4$

1	N	N	N	I	V	
2	N	N	N	I	V	
3	N	N	N	–	V	
4	N	N	C	I	V	
5	N	C	C	I	V	
	60	24	9	16	40	= **149**

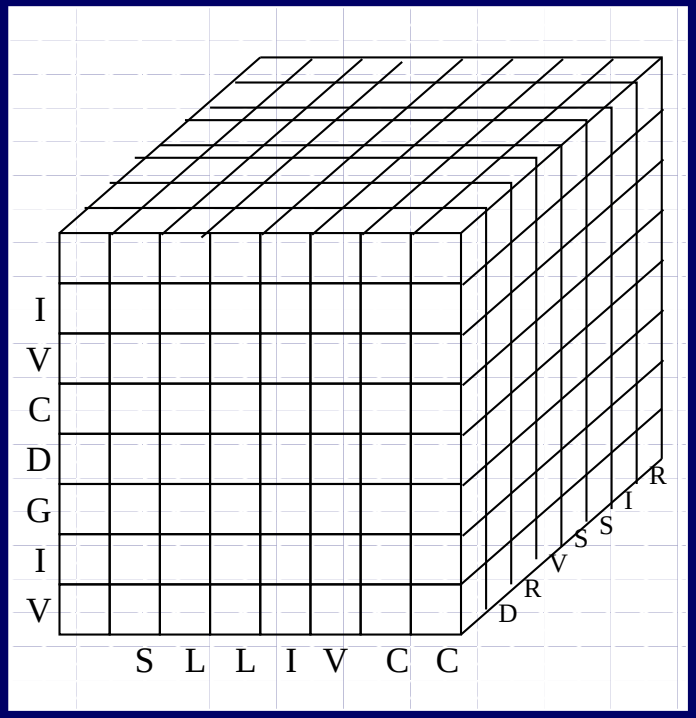
- Il y a $\frac{k(k-1)}{2}$ paires/scores à calculer pour chaque colonne

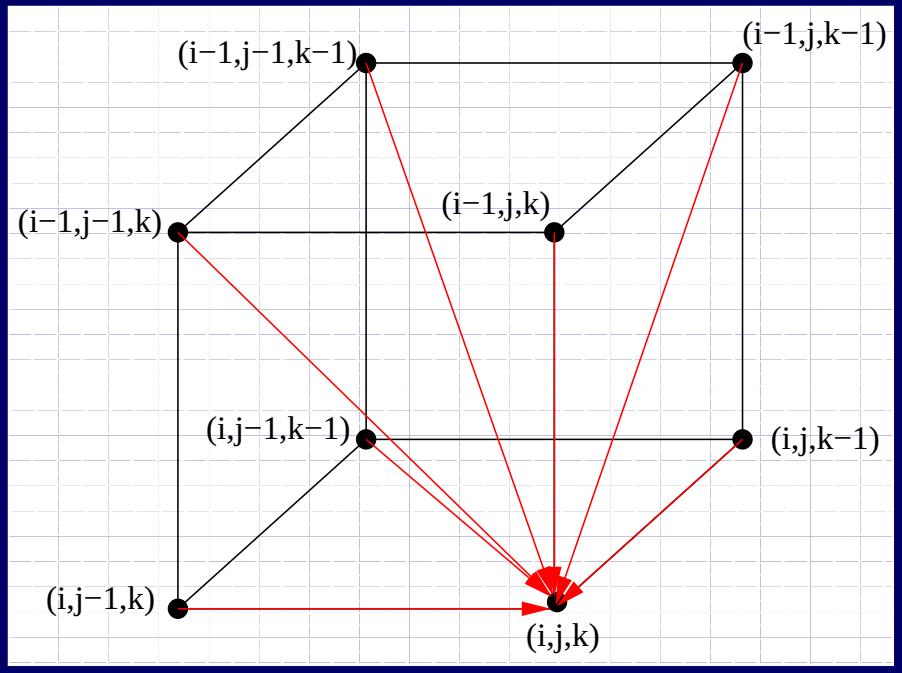
Méthode exacte

- Le principe de programmation dynamique est généralisable à k séquences de longueur n
- Matrice $\mathcal{A}$ de dimension n^k
- $\mathcal{A}(i_1, i_2, \ldots, i_k)$ contient le score d'alignement optimal entre les préfixes $s_1[1..i_1], s_2[1..i_2], \ldots, s_k[1..i_k]$
- Remplissage des $O(n^k)$ cases la table $\Rightarrow$ espace mémoire $O(n^k)$
- Chaque entrée dépend de $2^k - 1$ entrées déjà calculées
- Calculer le score SP requiert $O(k^2)$ car il y a $\frac{k(k-1)}{2}$ paires

Temps total d'exécution en $O(k^2 2^k n^k)$

Exemple avec 3 séquences





Exemple avec 3 séquences : la récurrence

- 3 séquences : r, s et t, méthode de score SP et g pénalité de gap
- $\mathcal{A}$ matrice de prog. dyn.

Initialisation : $\mathcal{A}(0,0,0) = 0$; $\mathcal{A}(i,0,0) = i \times 2g$;
$\mathcal{A}(0,j,0) = j \times 2g$; $\mathcal{A}(0,0,k) = k \times 2g$.

Remplissage :

$$\mathcal{A}(i,j,k) = \max \begin{cases} \mathcal{A}(i-1,j-1,k-1) & + & SP(r_i, s_j, t_k) \\ \mathcal{A}(i,j-1,k-1) & + & SP(-, s_j, t_k) \\ \mathcal{A}(i-1,j,k-1) & + & SP(r_i, -, t_k) \\ \mathcal{A}(i-1,j-1,k) & + & SP(r_i, s_j, -) \\ \mathcal{A}(i,j,k-1) & + & SP(-, -, t_k) \\ \mathcal{A}(i,j-1,k) & + & SP(-, s_j, -) \\ \mathcal{A}(i-1,j,k) & + & SP(r_i, -, -) \end{cases}$$

Méthode exacte : conclusion

- Méthode guère plus complexe pour k séquences que pour 2
- Facilement programmable
- Mais il est clair que le temps de calcul en $O(k^2 2^k n^k)$ devient prohibitif quand k augmente
- Illustration :
  - $\rightarrow$ 2 séquences de 100 a.a. $\rightarrow$ 1 sec.
  - $\rightarrow$ 3 séquences de 100 a.a. $\rightarrow$ 10 min
  - $\rightarrow$ 4 séquences de 100 a.a. $\rightarrow \sim$ 3 jours
  - $\rightarrow$ à partir de 9 séq. le tps de calcul dépasse l'âge de l'univers . . .

$\Rightarrow$ Mise au point d'algorithmes heuristiques performants et de bonne qualité (pb encore ouvert aujourd'hui)

Méthode progressive ou « alignement de profils »

- Difficile d'aligner simultanément k séquences $\Rightarrow$ on les aligne progressivement en commençant par les plus proches

$\rightarrow$ on aligne alors d'autres séquences avec ces premiers alignements

- Les alignement intermédiaires sont appelés « profils »

- À chaque étape on n'aligne que 2 éléments : séquence/séquence, séquence/profil ou profil/profil

- Utilise la prog. dyn. pour évaluer la similarité entre les séquences $\Rightarrow$ création d'un arbre évolutif qui sert de guide

- Propriété de cette méthode : toute création d'un trou est définitive « *once a gap, ever a gap* »

- Programme original CLUSTAL utilisé depuis 1988, régulièrement amélioré depuis [Higgins & Sharp, 88]
- CLUSTALW ([Thompson *et al*, 94]) est la version la plus récente, W pour *Weighting* ; les séquences et les paramètres sont pondérés
- Fonctionnalités : ajout d'une séq. ou d'un ali. à un ali. déjà fait, production d'un arbre phylogénétique, paramètre *slow/fast*, ...
- Méthode
  1. Construction de la matrice des $\frac{k(k-1)}{2}$ distances
  2. Logiciel de reconstruction phylogénétique $\Rightarrow$ arbre
  3. Alignement progressif suivant cet arbre
- Complexité pour k séquences de longueur n : $O(kn^2)$

CLUSTALW : le score

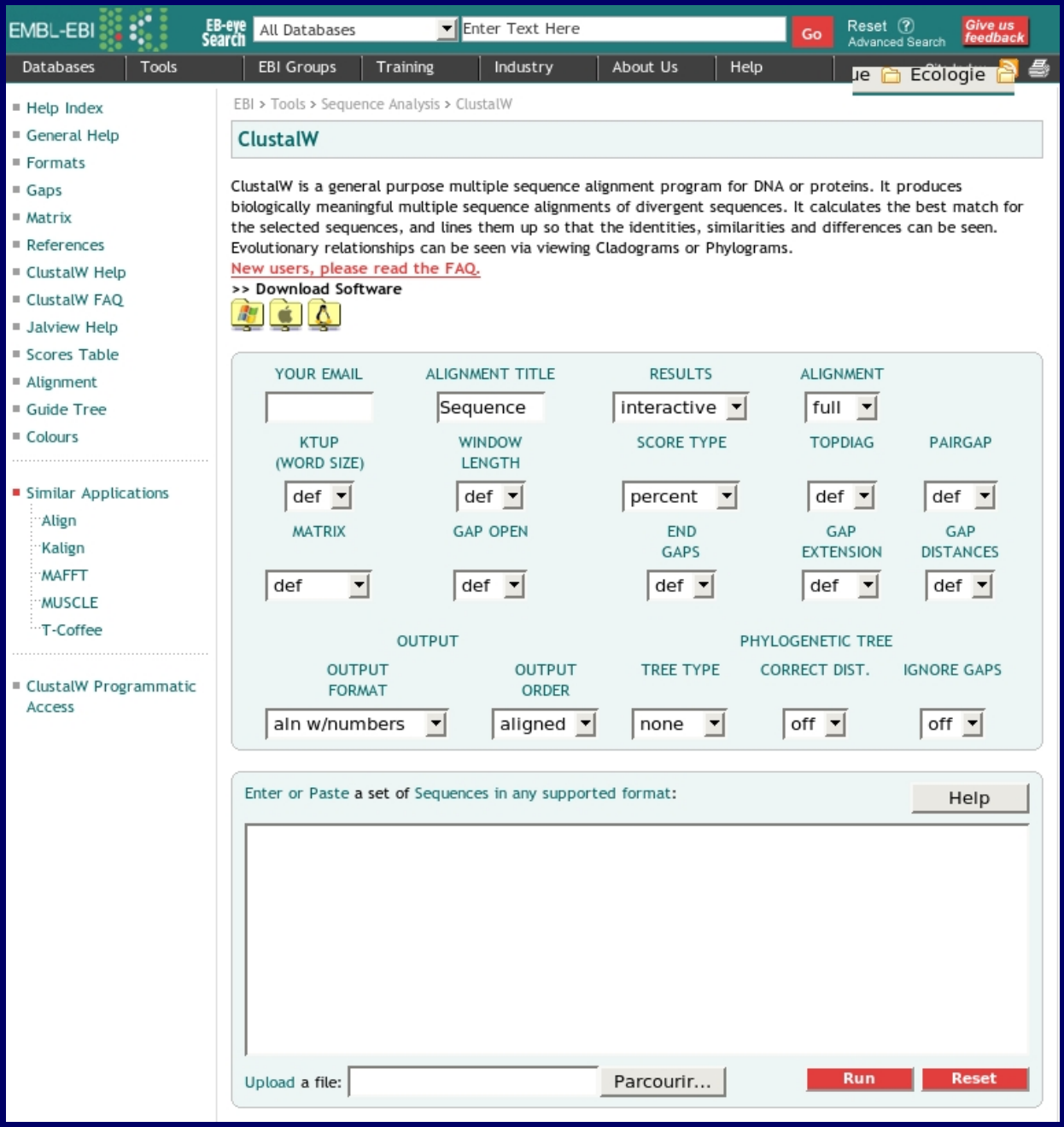
- Pondération des séquences :
  Une séquence similaire à d'autres dans le groupe a un poids faible, alors qu'une séquence moins proche des autres a un poids plus fort

  $\Rightarrow$ On diminue ainsi le poids des groupes de séquences similaires et privilégie les changements dans l'arbre évolutif

- Pénalité de gap :
  CLUSTALW pénalise les gaps de manière à les placer entre les domaines conservés (utilise une matrice spéciale et $\neq$ pénalités suivant les régions)

Méthode de score de CLUSTALW $\neq$ SP-score

CLUSTALW : Page d'accueil sur l'EBI



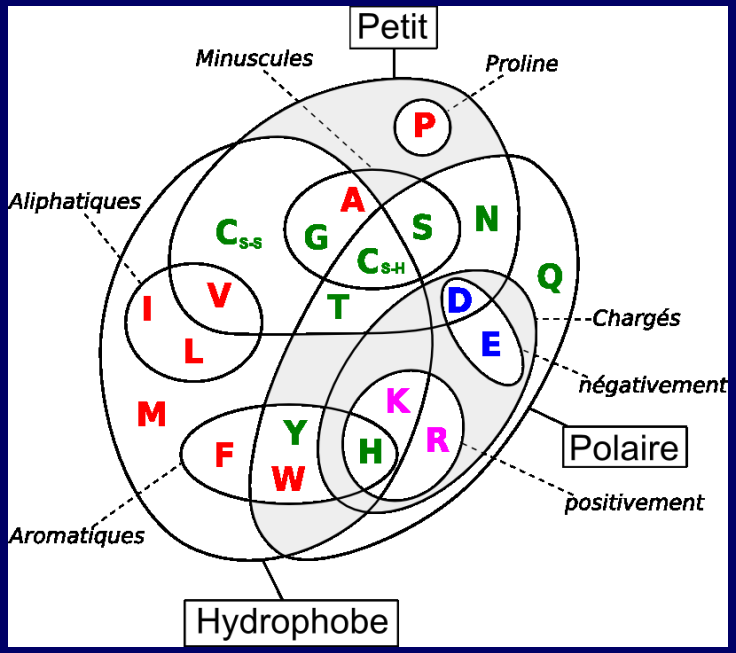
- CLUSTALW affiche par défaut les symboles suivants pour indiquer la degrés de conservation dans chaque colonne :

→ * caractère identique dans toute la colonne

→ : substitutions conservatives (suivant la table des couleurs)

→ . substitutions semi-conservatives

Code couleur de CLUSTALW



A.A.	Couleur	Description
AVFPMILW	Rouge	Petits (petits + hydrophobes (incl. aromatiques -Y))
DE	Bleu	Acides
RK	Magenta	Basiques - H
STYHCNGQ	Vert	Hydroxyl + sulfhydryl + amine + G
Autres	Gris	Unusual amino/imino acids . . .

Alignement étoile

- Un autre type d'alignement progressif : l'alignement étoile

- Principe : alignement 2 à 2 entre une séquence fixée (le centre de l'étoile) et toutes les autres séquences

- Méthode :
  1. Choisir la séquence centre s_c
  2. Alignement optimal entre toutes les s_i pour $i \neq c$ et s_c
  3. Agrégation des alignements avec le technique « *Once a gap, always a gap* », s_c est utilisée comme guide

- Procédure en $O(kn^2 + k^2l)$

- Le pb majeur avec l'alignement progressif est qu'une erreur faite au début de l'alignement ne peut être corrigée par la suite

- Les méthodes itératives tentent de pallier ce pb en réalignant des sous-groupes de séquences de manière répétée puis en alignant ces sous-groupes dans l'alignement global de toutes les séquences

- L'objectif est d'améliorer le score d'alignement global

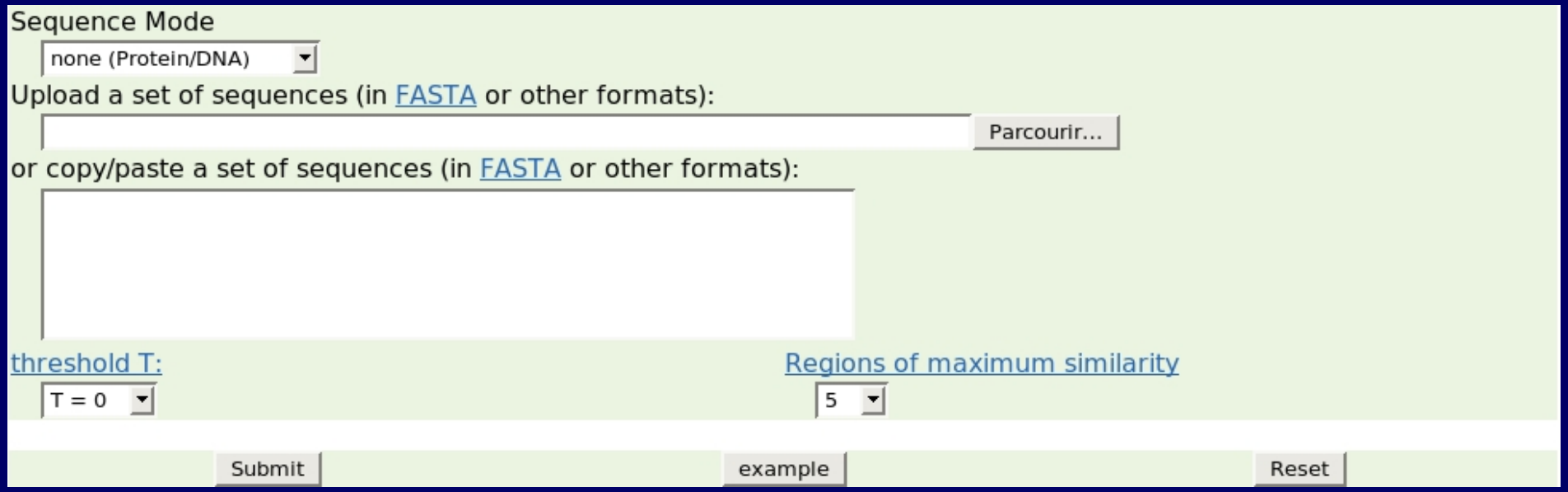
- La sélection de ces sous-groupes peut de faire sur la base d'un arbre phylogénétique (prédit de manière similaire aux méth. prog.), ou la séparation d'une à deux séquences du reste, ou de manière aléatoire

- DIALIGN 2.2.1 (oct. 2007) [Morgenstern, 2004]

- Spécialité : se base sur des similarités locales pour aligner des séquences très divergentes ou de longueur différentes

- Méthode :
  1. Repérer les régions alignées sans gaps dans les alignements 2 à 2 ($\sim$ diagonales continues dans un dotplot)
  2. Cherche un ensemble compatibles de diagonales pondérées pouvant produire un ali. et maximisant la somme des poids
  3. DIALIGN produit un alignement à partir de ces diagonales

- Disponible sur http://bibiserv.techfak.uni-bielefeld.de/dialign/



- Schéma de score dans DIALIGN : score de l'ali = somme des scores des diagonales qui le composent ⇒ pas de pénalités de gap
- Paramètres particuliers :
  - Seuil T → permet de ne considérer que les diagonales de score supérieur à T (nécessaire dans la V1)
  - Régions de similarité maximale → donne le nb max de caractères '*' utilisés pour représenter la similarité locale sous chaque col.

```
uniprot:MYH1    1898   SKFRRIQHEL EEAEERADIA ESQVNKLRVK SREVHTKIIS EE----
uniprot:MYH2    1900   AKFRKLQHEL EEAEERADIA ESQVNKLRVK SREVHTKVIS EE----
uniprot:MYH3    1899   SKFRKIQHEL EEAEERADIA ESQVNKLRAK SREIGKKAES EE----
uniprot:MYH3    1895   TKFRKAQHEL EEAEERADIA ESQVNKLRAK TRDFTSSRMV VHESEE
uniprot:MYH3     149   TKFRKAQHEL E--------- ---------- ---------- ------
uniprot:MYH3    1895   TKFRKAQHEL EEAEERADIA ESQVNKLRAK TRDFTSSRMV VHESEE
uniprot:MYH4    1898   AKFRKLQHEL EEAEERADIA ESQVNKLRVK SREVHTKVIS EE----
uniprot:MYH4    1897   SKFRKLQHEL EEAEERADIA ESQVNKLRVK SREVHTKVIS EE----
uniprot:MYH6    1896   SKFRKVQHEL DEAEERADIA ESQVNKLRAK SRDIGAKQKM HDEE--
uniprot:MYH6    1896   SKFRKVQHEL DEAEERADIA ESQVNKLRAK SRDIGAKQKM HDEE--
uniprot:MYH6    1896   SKFRKVQHEL DEAEERADIA ESQVNKLRAK SRDIGAK-KM HDEE--
uniprot:MYH6    1895   SKFRKVQHEL DEAEERADIA ESQVNKLRAK SRDIGAKQKM HDEE--
uniprot:MYH7    1894   SKFRKVQHEL DEAEERADIA ESQVNKLRAK SRDIGTKGLN EE----
uniprot:MYH7    1893   SKFRKVQHEL DEAEERADIA ESQVNKLRAK SRDIGAKGLN EE----

                       ********** ********** ********** ********** ******
                       ********** ********** ********** *******
                       ********** ********** ********** ****
                       ********** ********** ********** ***
                       ********** *
```

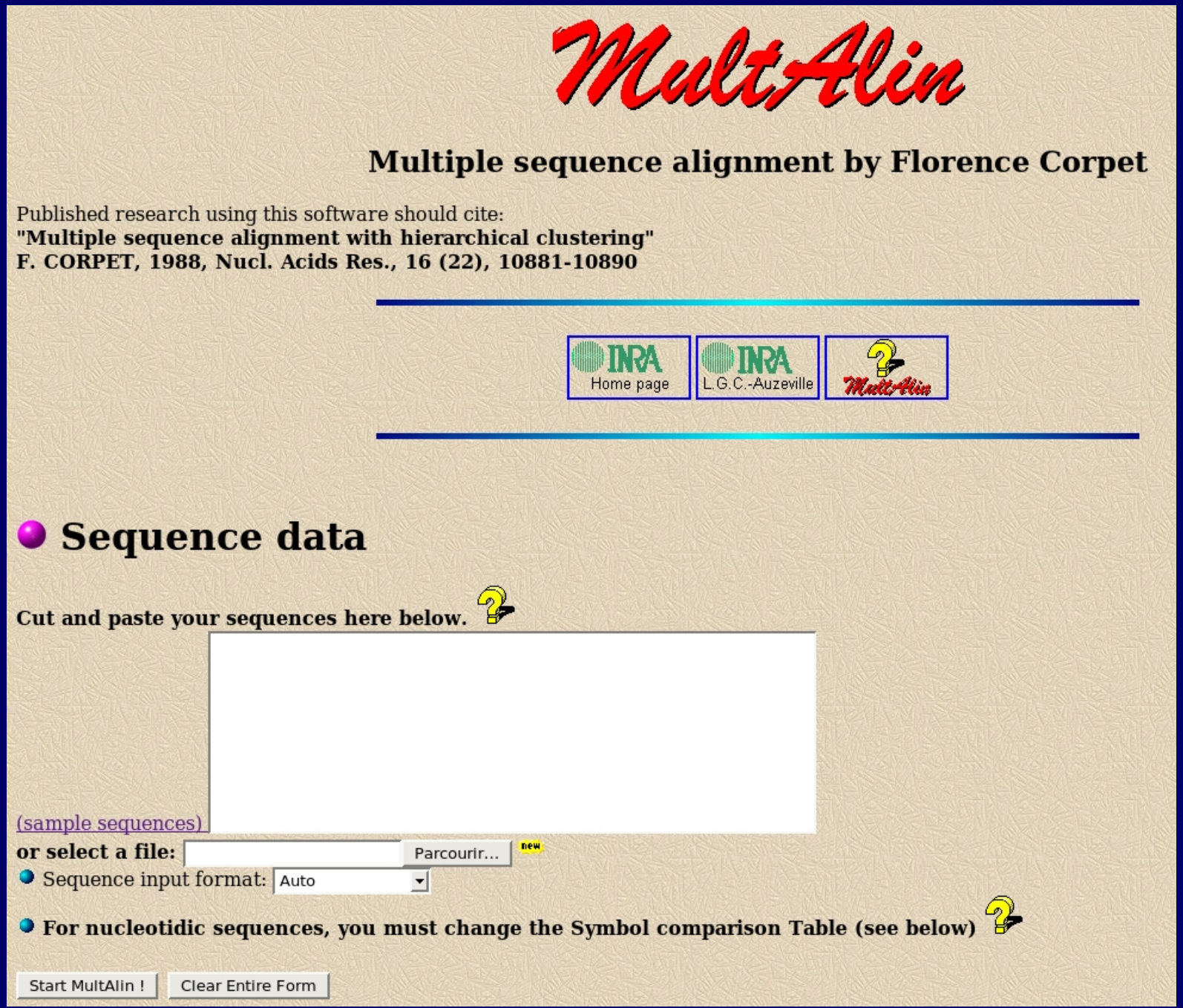
- 3 formats de sortie possibles : FASTA, MSF et DIALIGN ci-dessus pour l'alignement multiple
- + un arbre au format PHYLIP
- DIALIGN peut aussi produire des alignements multiples locaux

Alignement structurel

- Lorsqu'on dispose de structures 3D, on peut comparer des protéines en les superposant
- Mais il faut connaître quels sont les a.a. qui se correspondent
- C'est le but des méthodes d'alignement multiple structurel
  → Alignement structurel des séquences
  → Superposition 3D de leur structure
- Logiciels, pour 2 ou plusieurs protéines : SSAP, DALI, STAMP, ...
- Méthode de STAMP
  1. Alignement structurel de 2 séquences
  2. Les structures sont superposées selon cet alignement
  3. Calcul d'une matrice de scores de similarité structurelle
  4. Prog. dyn. sur cette matrice pour trouver meilleur score & ali.
  5. Étant donné l'alignement on recommence jusqu'à convergence

  Si plusieurs séquences, toutes les paires de structures sont comparées et un arbre est calculé, et on le suit comme pour un alignement progressif

- Programme de Florence Corpet de l'INRA Toulouse [Corpet, 88]
- Alignement multiple avec clustering hiérarchique
- Principe similaire à CLUSTAL mais pas d'évolution depuis 2000
- Méthode :
  1. Initialisation : tous les ali. 2 à 2 sont faits et on garde leur score
  2. Clustering hiérarchique des séquences à partir de ces scores
  3. Alignements selon cet arbre hiérarchique $\rightarrow$ ali. complet
  4. L'alignement est montré et on calcule son score SP (et donc le score de chaque paire de séquences alignées)
  5. Nouveau clustering hiérarchique avec ces nouveaux scores
  6. Si le nouveau clustering est différent du 1er, on peut recommencer (pas 3.) jusqu'à ce que ce soit stable

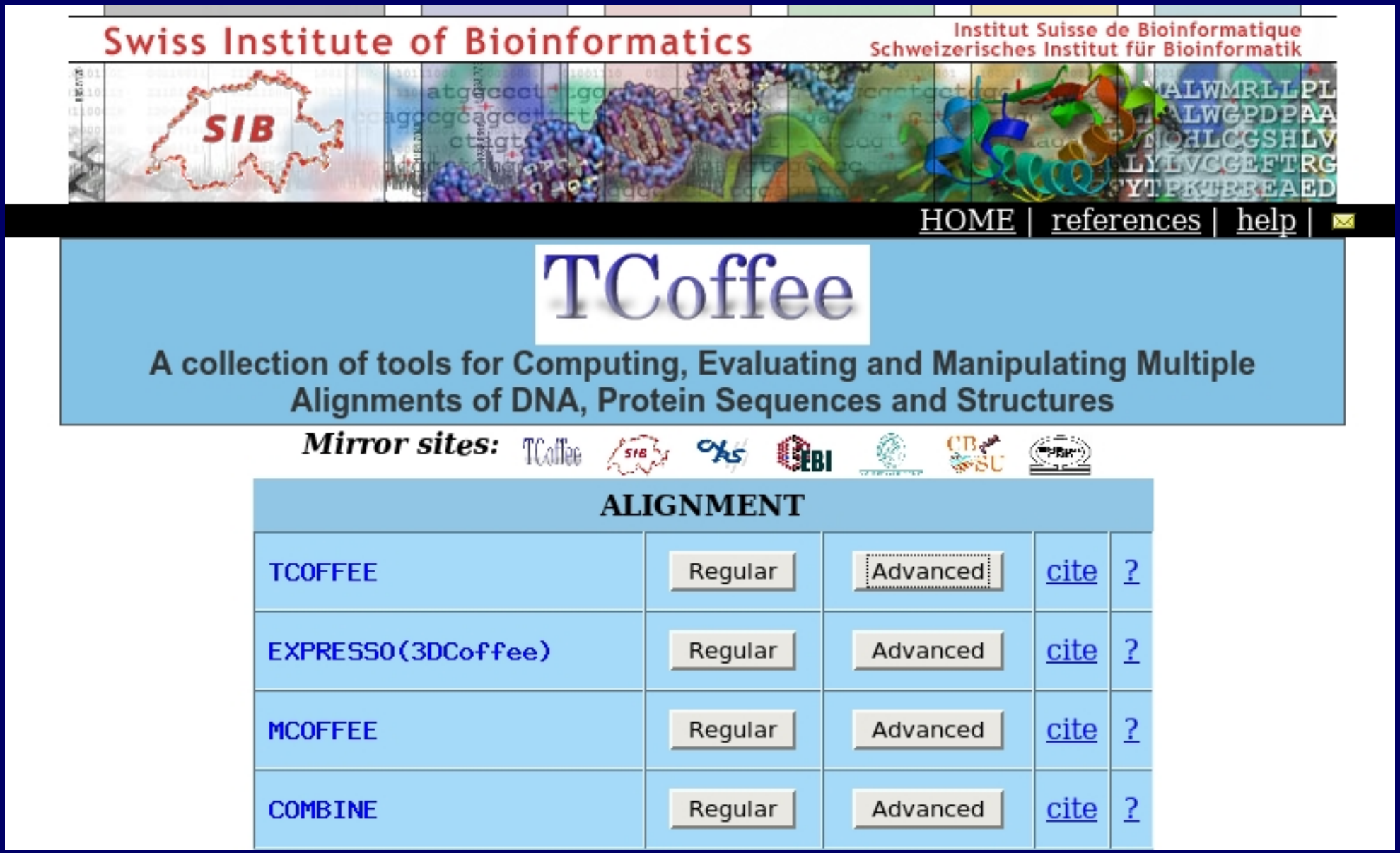

MultAlin
Multiple sequence alignment by Florence Corpet
Published research using this software should cite:
"Multiple sequence alignment with hierarchical clustering"
F. CORPET, 1988, Nucl. Acids Res., 16 (22), 10881-10890
INRA Home page
INRA L.G.C.-Auzeville
MultAlin
Sequence data
Cut and paste your sequences here below.
(sample sequences)
or select a file:
Parcourir...
new
Sequence input format:
Auto
For nucleotidic sequences, you must change the Symbol comparison Table (see below)
Start MultAlin !
Clear Entire Form

Résultats graphiques de MULTALIN

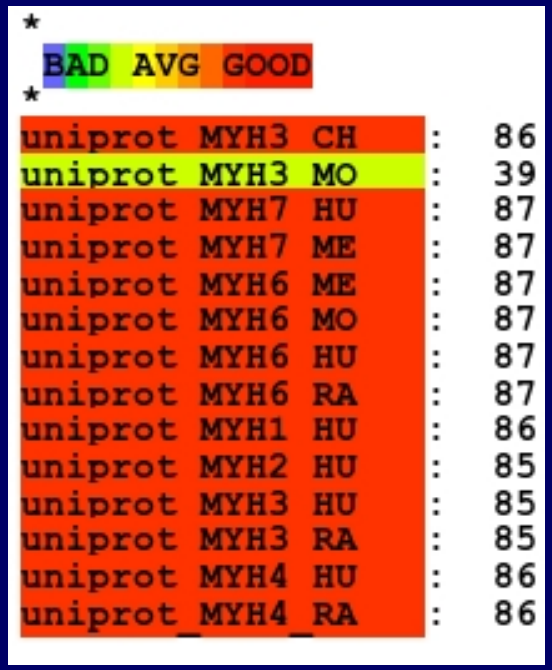
- Plusieurs formats d'entrée et de sortie disponibles
- Code couleur suivant le degrés de consensus de la colonne :
  - Rouge → fort consensus, lettre majuscule
  - Bleu → faible consensus, lettre minuscule
  - Noir → neutre, pas de caractère reporté
- Possibilité de modification des paramètres de la méthode et en particulier de ceux d'affichage

- T-COFFEE V5.05 (2007), [Notredame et al, 2000]
- Méthode :
  1. Construction d'une bibliothèque d'alignement 2 à 2, globaux et locaux, pour chaque paire de séquences du jeu d'entrée
  2. Phase de "scorage" pour donner un poids aux alignements puis aux caractères alignés (« *library extension* »)
  3. Cette bibliothèque est ensuite utilisée pour guider une phase d'alignement progressif pour trouver un MSA préservant la consistance des alignements 2 à 2
- Produit de meilleurs alignements mais encore trop lent pour de gros jeux de séquences
- Originalité : permet de combiner des résultats de CLUSTALW, DIALIGN et d'un alignement structural par exemple

T-COFFEE : Page d'accueil



T-COFFEE : Résultats graphiques



- Les parties oranges et rouges sont considérées comme bien alignées
- Plusieurs formats de sortie : .aln (CLUSTALW), FASTA, PHYLIP et T-COFFEE ci-dessus
- + 2 formats d'arbres

- Les MSA peuvent servir à faire des recherches plus sensibles dans les banques de données qu'en utilisant 1 seule séq. requête
- Rappel du fonctionnement de PSI-BLAST
  1. On cherche des séquences similaires à la requête
  2. Construction d'un profil ou PSSM
  3. (en boucle) On cherche avec ce profil et on y intègre les nouvelles séquences trouvées
  4. Fin qd l'ensemble de séq. est stable ou le nb max d'itérations est dépassée
- PSI-BLAST fournit des alignement multiples différents de ceux obtenus avec les outils classiques

→ Normalement les MSA sont plus long que les séquences qu'ils contiennent mais PSI-BLAST supprime les régions qui nécessiteraient d'introduire des gap dans la séq. requête

- Les méthodes de MSA décrites précédemment produisent un alignement global des séquences

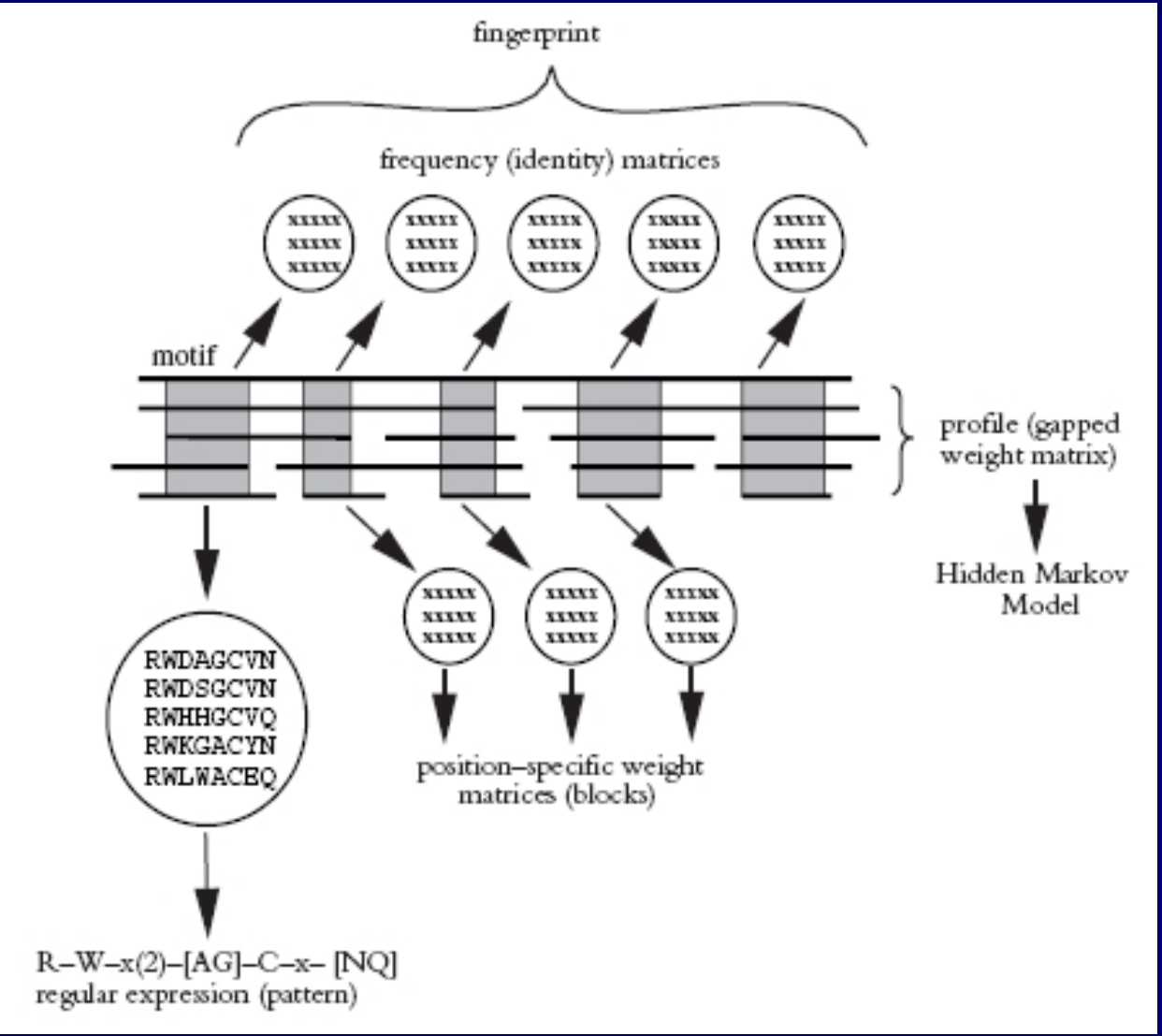
→ Bon point de départ pour les analyses phylogénétiques

- Les méthodes de MSA locaux alignent les régions les plus similaires des séquences, ignorant les régions dissimilaires

- La production d'un MSA local peut se faire soit
  1. en utilisant des méthodes d'alignements multiple local
  2. en partant d'un MSA global et en sélectionnant les régions les plus conservées

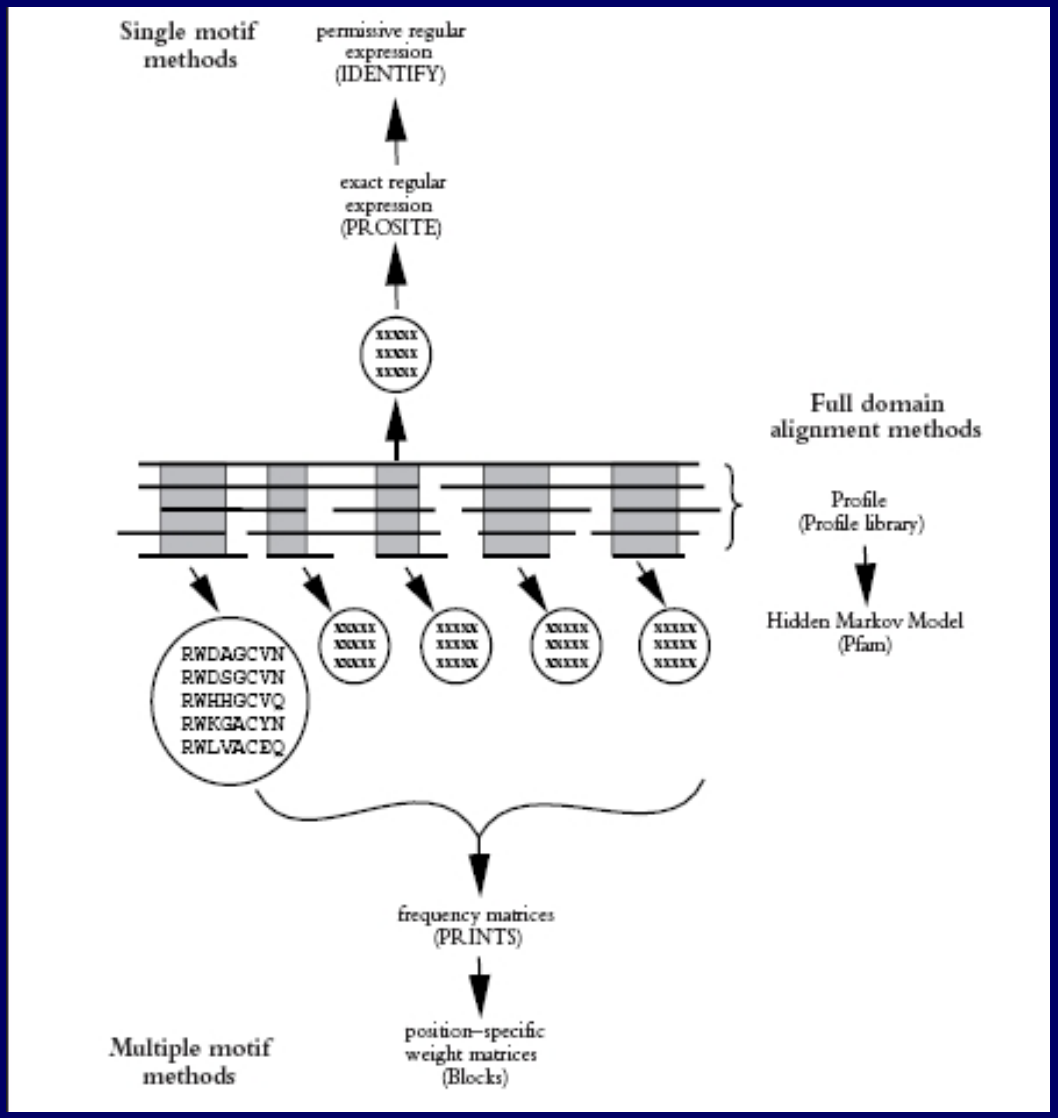
- Il existe de nombreuses banques de familles de séquences protéiques basées sur le partage d'une caractéristique commune

Extraction d'un MSA global



[Attwood, 2000]

Création de banques de familles protéiques



[Attwood, 2000]

Régions conservées dans les protéines

- La nature réutilise souvent un même élément de séquence pour réaliser des fonctions analogues

- Ces éléments conservés se structurent généralement en domaines indépendants dans les protéines et prennent en charge une des tâches de la protéine : liaison à l'ADN ou à l'ARN, fixation d'un ligand, activité enzymatique, . . .

- Un gd nb de domaines sont répertoriés dans des bq spécialisées

→ Ex : PRODOM, INRA Toulouse, ~ 50000 domaines communs à au moins 2 protéines, lg moy 100 $a.a.$

- Alignement systématique de ces domaines $\Rightarrow$ identification de motifs caractéristiques

Motif conservés / Expressions régulières

- Il existe des positions (d'un MSA) où l'on trouve toujours un même a.a. ou bien des a.a. similaires (F, Y aromatiques), elles constituent une signature caractéristique de la famille de domaines homologues

- On peut donc vouloir chercher des motifs défini de la manière suivante :
  F ou Y, suivi d'un a.a. qcq, suivi de C, suivi de 5 a.a. qcq, suivi de C

- Même motif mais en syntaxe PROSITE : `[FY]-X-C-X(5)-C`

- On appelle ce formalisme expressions régulières, *regex* en anglais

Construire un motif PROSITE

- Utilisation du code standard des protéines (IUPAC)
- La lettre `X` est utilisée pour identifier une position qui peut accepter n'importe quel résidu
- Les positions qui peuvent accepter plrs résidus sont indiquées par des `[]` qui renferment les résidus acceptables, ex : `[GTS]`
- Les résidus polymorphes sont indiqués par des $\{\}$ qui indiquent quels résidus sont impossibles
- Chaque indication de contenu est séparée de la suivante par un `-`, ex : `A-T-X-T-X-X-[GTS]`
- Les répétitions de positions ayant un contenu identique sont indiquées de la manière suivante :
  `A-A-A`$\rightarrow$`A(3)` ; `T-T`$\rightarrow$`T(2)` ; `[AT](2)`$\rightarrow$`A-A` ou `A-T` ou `T-A` ou `T-T`
  `X(1,3)` signifie que les alternatives `X`, `X-X` ou `X-X-X` sont possibles

Un mot sur PROSITE

- PROSITE est une base de données où chaque entrée décrit des domaines de protéines, des familles ou des sites fonctionnels ainsi que les motifs et profils associés
- PROSITE contient plus de 1400 motifs documentés
- Consensus PROSITE pour "Zinc finger C2H2-type"

```
ABC3G_LAGLA/285-305     Cfs..CaekVaeflqenpHvnl..H
ABRU_DROME/546-567      Cpk..CgkiYrsahtlrtHledk.H
ACE1_TRIRE/402-424      CrepgCtkeFkrpcdltkHekt..H
ACE2_SCHPO/445-467      ClyngCnkrIarkynvesHiqt..H
ACE2_SCHPO/475-495      Cdl..CkagFvrhhdlkrHlri..H
ACE2_YEAST/605-627      ClypnCnkvFkrrynirsHiqt..H
ACE2_YEAST/635-657      CdfpgCtkaFvrnhdlirHkis..H
ADNP_HUMAN/514-535      Cpy..CrstFndvekmaaHmrmv.H
ADNP_MOUSE/233-254      Cpy..CrstFndvekmaaHmrmv.H
...
```

C - X(2,4) - C - X(3) - [LIVMFYWC] - X(8) - H - X(3,5) - H

Chercher un motif dans une séquence

- Avantage :

→ Des algorithmes de recherche de motifs très puissants existent (domaine *pattern matching* en informatique)

→ Les expressions régulières permettent un codage simple du motif

→ Les méthodes qui cherchent des motifs pointent vers des courtes régions conservées qui peuvent représenter des fonctions biologiques

- Inconvénients :

→ Manque de souplesse : « *matche* ou *ne matche pas* », pas de mesure de similarité

→ N'autorisent pas les gaps

→ Souvent il n'y a pas d'analyse de la significativité du motif

→ Si la famille de protéines a plrs régions conservées, laquelle choisir ?

Expressions régulières floues

- *Fuzzy regex* en anglais
- Permet de relâcher les contraintes en autorisant les substitutions d'a.a. de même groupe
- Ces groupes sont définis sur la base des partages de propriétés physico-chimiques (cf. diagramme de Venn p. 22)

Petits	A G
Petits hydroxyle	S T
Basiques	K R
Aromatiques	F Y W
Basiques	H K R
Petits hydrophobes	V L I
Moyens hydrophobes	V L I M
Acides/Amides	D E N Q
Petits/polaires	A C G S T P

- Ex : motif `[AS]-D-[IVL]-G-X(4)-PG-C-[DE]-R-[FY](2)-Q`

→ Motif flou dérivé `[ASCGPT]-D-[IVLM]-G-X(4)-PG-C-[DE]-R-[FYW](2)-Q`

- Permet d'identifier des prot. assez éloignées mais augmente le bruit

- Banque de donnée eMOTIF

Analyse de profil

- Les profils sont construits à partir d'un MSA global des séquences, en extrayant les régions les plus conservées, donnant ainsi un MSA de plus petite taille

- Une matrice de scores pour ce MSA est alors calculée, on l'appelle profil (*profile* en anglais)

- Un profil est composé de colonnes contenant les scores pour les *matches*, les *mismatches* et les pénalités de gap

- Une fois calculé, le profil peut être utilisé pour chercher une séquence cible

Plus précisément

- Les profils sont similaires aux matrices de scores (PSSM)
- Un profil est une matrice de 23 colonnes : une pour chaque a.a., plus une colonne pour un a.a. inconnu 'Z' et 2 colonnes pour les pénalités d'ouverture et d'extension de gap
- Le profil a autant de lignes que le MSA a de colonnes
- Une séquence consensus est indiquée verticalement à la gauche du tableau, elle représente les a.a. les plus fréquents à chaque position
- Les scores reflètent le nb d'occurrences de chaque a.a. dans les séquences alignées

Exemple de profil


```
BIRC6_HUMAN  RRLAQEAVTLST.....S.........LPLSSSSSVFVRCde.......eRLDIMKVLITGP...ADTPY
COP10_ARATH  KRIQREMAELNI.....D.........PPPDCSAGPKGD-.........NLYHWIATIIGP...SGTPY
FTS1_HUMAN   YSLLAEFTLVVK.....Q.........KLPGVYVQPSYRS.........ALMWFGVIFI--...RHGLY
FTS1_MOUSE   YSLLAEFTLVVK.....Q.........KLPGVYVQPSYRS.........ALVWFGVIFI--...RHGLY
MMS2_SCHPO   FKLLEELEKGEKgl..gE.........SSCSYGLTNADDI.........TLSDWNATILGP...AHSVH
MMS2_YEAST   FRLLEELEKGEK.....Gf.........GPESCSYGLADSDd.......iTMTKWNGTILGP...PHSNH
...

      A   B   C   D   E   F   G   H   I   K   L   M   N   P   Q   R   S   T   V   W
'R'; -10, -6,-25, -9, -2,-13,-19, -7,-16, 16,-14, -6, -1,-17,  1, 18, -6, -4,-13,-20
'R'; -15,-12,-26,-13, -4,-13,-20, -5,-21, 16,-10, -6, -4,-20,  3, 47, -9, -6,-14,-21
'L';  -9,-28,-22,-30,-20,  4,-32,-22, 25,-26, 35, 17,-26,-26,-19,-20,-24, -9, 16,-21
'M';  -7,-14,-20,-19, -8, -8,-23, -7,  0, -5,  6, 14,-13,-19,  5, -3,-13, -7, -4,-19
'K';  -7, -2,-28, -1, 13,-27,-19, -7,-27, 33,-24,-10, -2,-11, 11, 26, -7, -9,-20,-20
'E'; -12, 15,-30, 26, 50,-32,-18,  1,-31,  8,-22,-20,  3, -2, 21, -1,  0,-10,-30,-31
'L'; -10,-25,-21,-28,-20, 17,-28,-16, 10,-19, 22,  9,-22,-27,-19,-12,-20, -7,  7,-12
...
```

Profil : les désavantages

- Les profils sont très liés aux MSA globaux dont ils sont extraits (reflet de la variation dans les séquences)

- Si il y a beaucoup de séquences similaires, le MSA et le profil déduit seront biaisés en faveur de ces séquences

- Diverses corrections dont pondération des séquences

- Autre problème : des a.a. peuvent de pas être représentés dans certaines colonnes à cause d'un trop petit nb de séq. incluses

Construction de profil

- Deux méthodes sont principalement utilisées pour construire des profils

→ Méthode moyenne : les scores du profil sont pondérés par la proportion de chaque a.a. dans chaque colonne (ex MEME)

→ Méthode évolutive : basée sur le modèle d'évolution de protéine de Dayoff, elle suppose différentes vitesses d'évolution pour chaque colonne du MSA

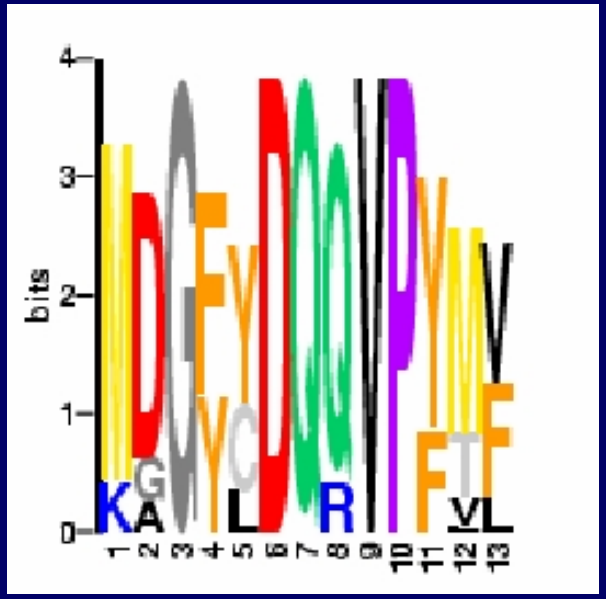
- Évaluation de profils : le "ROC plot" mesure la précision d'un profil pour discriminer entre les membres ou non-membres d'une famille de protéines lors d'une recherche de similarité dans les banques

Les blocs

- Comme les profils, les blocs représentent les régions conservées dans un MSA global
- Les blocs produits de cette manière sont aussi bon (*ou moins bon*) que le MSA dont ils sont issus
- Les blocs sont différents des profils car ils n'autorisent pas les indels
- Les blocs peuvent être constitués de plusieurs régions similaires mais non adjacentes
- Le motif trouvé dans chaque séquence est de même lg, et qd les séq. sont alignées, les caractères sont dans les mêmes colonnes
- Le serveur BLOCKS fournit des outils pour extraire des blocs de 10 à 55 a.a. provenant de MSA jusqu'à 400 séquences

Exemple de bloc


```
Block IPB006715A
ETV5_HUMAN|P41161  (   1) MDGFYDQQVPFMV  35
ETV1_HUMAN|P50549  (   1) MDGFYDQQVPYMV  33
ETV1_MOUSE|P41164  (   1) MDGFYDQQVPYVV  37
ETV4_HUMAN|P43268  (  73) KAGYLDQQVPYTF  77
ETV4_MOUSE|P28322  (  81) KGGYLDQRVPYTF 100
PEA3_BRARE|Q9PUQ1  (   5) MDGYLDQQVPYTL  50
...
```



Les empreintes

- Les empreintes ou *fingerprints* sont une autre représentation de plusieurs domaines conservés dans le MSA
- Les empreintes sont des matrices de comptage d'a.a. non pondérées
- Ces matrices sont donc creuses : la plupart des cases contenant 0
- L'utilisation de ces matrices brutes est plutôt rare
- On construit donc des empreintes par itérations successives sur les banques Swiss-Prot et TrEMBL pour qu'elles contiennent plus de variabilité
- Banque de données PRINTS (Version 38.0) contient 1900 empreintes, encodant 11170 motifs

Un mot sur les méthodes statistiques

- Les méthodes que l'on vient de voir utilisent un MSA global comme point de départ et tentent de localiser les régions conservées

- Une alternative est de chercher les motifs conservés en faisant des alignements "tests" et ensuite d'améliorer ces alignements en utilisant des méthodes statistiques

- Plusieurs méthodes regroupées sous ce terme : algorithmes EM (*Expectation Maximisation*), Gibbs Sampler, les modèles de Markov cachés (HMM), ...

- Les matrices de scores produites par ces méthodes peuvent être utilisées pour la recherche de séquences présentant le même motif

- Une fois qu'on a obtenu un MSA global, il peut être intéressant d'éditer les séquences manuellement pour obtenir un *meilleur* alignement

- Un "bon" éditeur de séquences devrait avoir au moins les caractéristiques suivantes :

  → Prendre en compte l'affichage coloré du MSA où les couleurs aident à la compréhension

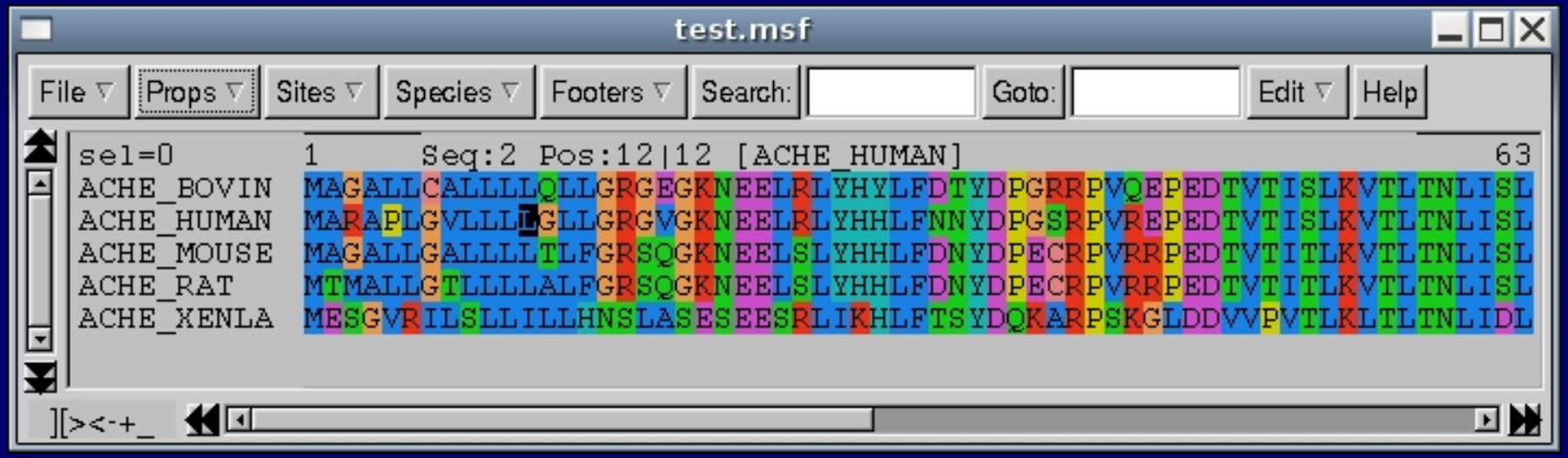
  → Reconnaissance des formats de MSA en sortie des logiciels d'alignement et maintient de ce format après l'édition

  → Interface graphique ergonomique, utilisation de la souris, . . .

Éditeurs

- Éditeurs : CINEMA (Colour INteractive Editor for Multiple Alignment), GDE (Genetic Data Environment), GeneDoc (pour Windows), SEAVIEW, . . .

- La plupart d'entre eux doivent être installés sur machines



- Il existe un grand nombre de formats de MSA
- Les deux plus fréquents sont le format MSF (Genetic's Computer Group) et le format ALN (CLUSTALW)
- On peut convertir l'un en l'autre facilement
- Exemple de format ALN :

```
CLUSTAL W (1.82) multiple sequence alignment


FOSB_MOUSE      MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA 60
FOSB_HUMAN      MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA 60
                ************************************************************

FOSB_MOUSE      ITTSQDLQWLVQPTLISSMAQSQGQPLASQPPAVDPYDMPGTSYSTPGLSAYSTGGASGS 120
FOSB_HUMAN      ITTSQDLQWLVQPTLISSMAQSQGQPLASQPPVVDPYDMPGTSYSTPGMSGYSSGGASGS 120
                ********************************.***************:*.**:******
```

Format ALN/ClustalW

- Le format ALN est le format d'origine du programme ClustalW. Le fichier commence par le mot clé "CLUSTAL" puis diverses informations concernant la version de CLUSTAL utilisée (ex : "CLUSTAL W (1.82) multiple sequence alignment")

- L'alignement est écrit en blocs de taille 60 résidus

- Chaque bloc commence avec le nom de la séquence et termine par la position de fin du bloc

- Des informations concernant les mises en correspondance des résidus est donnée en dessous de chaque bloc : "*" même résidu dans toute la colonne, ":" substitutions conservatives, "." substitutions semi-conservatives.

Format GCG/MSF

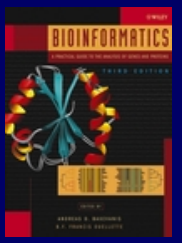
- Le fichier commence par autant de lignes de description que nécessaire
- Les commentaires sont terminés par une ligne commençant avec 2 barres obliques
- La 1re ligne reconnue est celle qui contient "MSF :", elle contient aussi la lg de la séquence, le type, etc
- Ensuite une ligne blanche
- Puis une ligne par séquence avec leur description (nom, lg, poids, . . .), ce bloc est aussi terminé par 2 barres obliques
- Encore une ligne blanche
- Enfin l'alignement proprement dit

Format GCG/MSF (2)


```
Les commentaires ....
//
MSF:  510  Type: P    Check:  7736   ..

Name: ACHE_BOVIN oo  Len:  510  Check:  7842  Weight:  16.0
Name: ACHE_HUMAN oo  Len:  510  Check:  8553  Weight:  17.8
Name: ACHE_MOUSE oo  Len:  510  Check:   229  Weight:  12.5
Name: ACHE_RAT oo  Len:  510  Check:  8410  Weight:  14.2
Name: ACHE_XENLA oo  Len:  510  Check:  2702  Weight:  39.2
//

ACHE_BOVIN      MAGALLCALL LLQLLGRGEG KNEELRLYHY LFDTYDPGRR PVQEPEDTVT
ACHE_HUMAN      MARAPLGVLL LLGLLGRGVG KNEELRLYHH LFNNYDPGSR PVREPEDTVT
ACHE_MOUSE      MAGALLGALL LLTLFGRSQG KNEELSLYHH LFDNYDPECR PVRRPEDTVT
ACHE_RAT        MTMALLGTLL LLALFGRSQG KNEELSLYHH LFDNYDPECR PVRRPEDTVT
ACHE_XENLA      MESGVRILSL LILLHNSLAS ESEESRLIKH LFTSYDQKAR PSKGLDDVVP

ACHE_BOVIN      ISLKVTLTNL ISLNEKEETL TTSVWIGIDW QDYRLNYSKG DFGGVETLRV
ACHE_HUMAN      ISLKVTLTNL ISLNEKEETL TTSVWIGIDW QDYRLNYSKD DFGGIETLRV
ACHE_MOUSE      ITLKVTLTNL ISLNEKEETL TTSVWIGIDW HDYRLNYSKD DFAGVGILRV
ACHE_RAT        ITLKVTLTNL ISLNEKEETL TTSVWIGIEW QDYRLNFSKD DFAGVEILRV
```

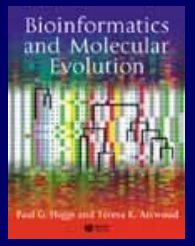
Références

- Chap 8 et 13 dans “Bioinformatics : A Practical Guide to the Analysis of Genes and Proteins, 3rd Edition”
  [Baxevanis & Ouellette,2005]



- Chap 2 et 4 dans “Bioinformatique Génomique et post-génomique”
  [Dardel & Képès, 2002]

- Chap 5, 7 et 9 dans “Bioinformatics and Molecular Evolution”
  [Higgs & Attwood, 2005]



- Illustrations :

  → Transp. 22 : Diagramme de Venn des propriétés des acides aminés. Wikipedia, entrée ’Acide aminé’ (http://fr.wikipedia.org/wiki/Acides_aminés)

  → Transp. 40 et 41 : Ces 2 schémas proviennent de l’article « The role of pattern databases in sequence analysis » de Terri K. Attwood, publié dans *Briefings in Bioinfomatics* en 2000