Vendredi 6 février 2004

Démo 4 – Alignement Multiple

Jean-Eudes Duchesne

0. Introduction

L'alignement multiple est l'extension naturelle à l'alignement de deux séquences.

Intérêts biologiques :

- Inférer des relations d'évolution entre les espèces.
- Déduire des similarités de fonction ou de structure.
- Trouver des caractéristiques communes à une famille de protéines ou de séquences d'ARN. Obtenir un modèle.

Intérêts par rapport à l'alignement de 2 séquences :

- Amplifie les différences et ressemblances entre les alignements.
- Alignement multiple capable de définir une famille tendis que l'alignement de deux séquences va plutôt servir à aligner une séquence à une famille.
- La comparaison multiple permet de déduire des contraintes de structure (positions très corrélées)
- Alignement multiple permet de définir un contexte phylogénique.

Il n'existe pas de méthode qui donne LE meilleur alignement. Pour obtenir un alignement biologiquement significatif, il faut combiner les méthodes automatiques avec certaines connaissances préalables sur les séquences. Par exemple, certaines régions sont indispensables à la protéine et doivent être alignées correctement. Il n'est pas improbable qu'un expert modifie un alignement obtenu automatiquement pour que celui-ci représente mieux le contexte biologique.

1. Score SP (Sum of Pairs)

Le score d'un alignement est la somme des scores des alignements induits pour toutes les paires de séquences.

Ex :

S_1	A	A	G	A	A	-	A
S_2	A	T	-	A	A	T	G
S_3	C	T	G	-	G	-	G

Si l'on considère la distance d'édition entre les séquences :
$S_1 – S_2 = 4$
$S_2 – S_3 = 5$
$S_1 – S_3 = 5$
Score SP = 4 + 5 + 5 = 14

Il n'existe pas de justification théorique pour le score SP, c'est seulement facile à calculer. Par contre trouver le score SP minimal pour un ensemble de séquences non alignées est un problème NP-complet. Donc, le score SP ne peut être calculé par programmation dynamique simple. Il faut donc utiliser une heuristique qui ne donne pas l'alignement optimal, mais néanmoins donne un « bon » alignement en temps polynomial.

2. Méthode de l'étoile centrale

La méthode de l'étoile centrale utilise le score SP et défini une méthode polynomiale pour obtenir un alignement. Supposons un ensemble $S = \{s_1, s_2, s_3, \ldots, s_n\}$ de séquences à aligner. L'algorithme se déroule comme suit :

1) Trouver une séquence dans S qui deviendra le centre de notre étoile (séquence de référence pour toutes les autres). La séquence s* doit être la séquence de l'ensemble S qui minimise la distance d'édition entre toutes les autres séquences.
2) Une fois s* trouvée, il faut aligner toutes les séquences restantes à s*. Pour ce faire, nous alignons séquentiellement toutes les séquences à s*. Si à une étape donnée un gap est ajouté dans s*, ce gap est propagé à toutes les séquences préalablement alignées à s*.
3) Le score est obtenu en calculant le score SP à partir de l'alignement obtenu.

Ex : Utilisons la méthode de l'étoile centrale pour aligner le séquences suivantes {ATGTTG, ATCGTTTG, GCCTGCG, GCCTTTGC, ACATTG}.

D'abord choisir s* en utilisant les valeurs obtenues de l'alignement entre les séquences :

	ATGTTG	ATCGTTG	GCCTGCG	GCCTTTGC	ACATTG	total
ATGTTG	0	2	5	5	2	14
ATCGTTTG	2	0	5	4	3	14
GCCTGCG	5	5	0	3	4	17
GCCTTTGC	5	4	3	0	4	16
ACATTG	**2**	**3**	**4**	**4**	**0**	**13**

Les résultats obtenus indiquent que s* = ACATTG.

Maintenant, reconstruisons séquentiellement l'alignement obtenu :

Étape 1 : Alignement de ACATTG et ATGTTG

A C A T T G
A T G T T G

Étape 2 : Ajout de ATCGTTTG dans l'alignement

A - C A - T T G
A - T G - T T G
A T C G T T T G

Étape 3 : Ajout de GCCTGCG dans l'alignement

A	-	C	A	-	T	T	-	G
A	-	T	G	-	T	T	-	G
A	T	C	G	T	T	T	-	G
G	-	C	C	-	T	G	C	G

Étape 4 : Ajout de GCCTTTGC et obtention de l'alignement final

A	-	C	A	-	T	T	-	G	-
A	-	T	G	-	T	T	-	G	-
A	T	C	G	T	T	T	-	G	-
G	-	C	C	-	T	G	C	G	-
G	-	C	C	T	T	T	-	G	C

Ce tableau nous permet de calculer le score SP de l'alignement, qui donne 38.

3. CLUSTALW

CLUSTALW est un programme d'alignement multiple qui approche encore une fois la réponse optimale. Il est à noter que les versions courantes du programme n'utilise pas le score SP à l'étape 1 mais bien une matrice de pondération telle que vue aux démos précédentes.

Étapes :

1. Calcul de scores deux à deux entre les séquences.
2. Construire un arbre guide reflétant la similarité entre les séquences à partir de la matrice des scores en utilisant une méthode de regroupement.
3. Aligner les séquences en suivant l'arbre.

Ex : Alignement progressif avec CLUSTALW des 5 séquences suivantes :
>Seq1
ATCTCGAGA
>Seq2
ATCCGAGA
>Seq3
ATGTCGACGA
>Seq4
ATGTCGACAGA
>Seq5
ATTCAACGA

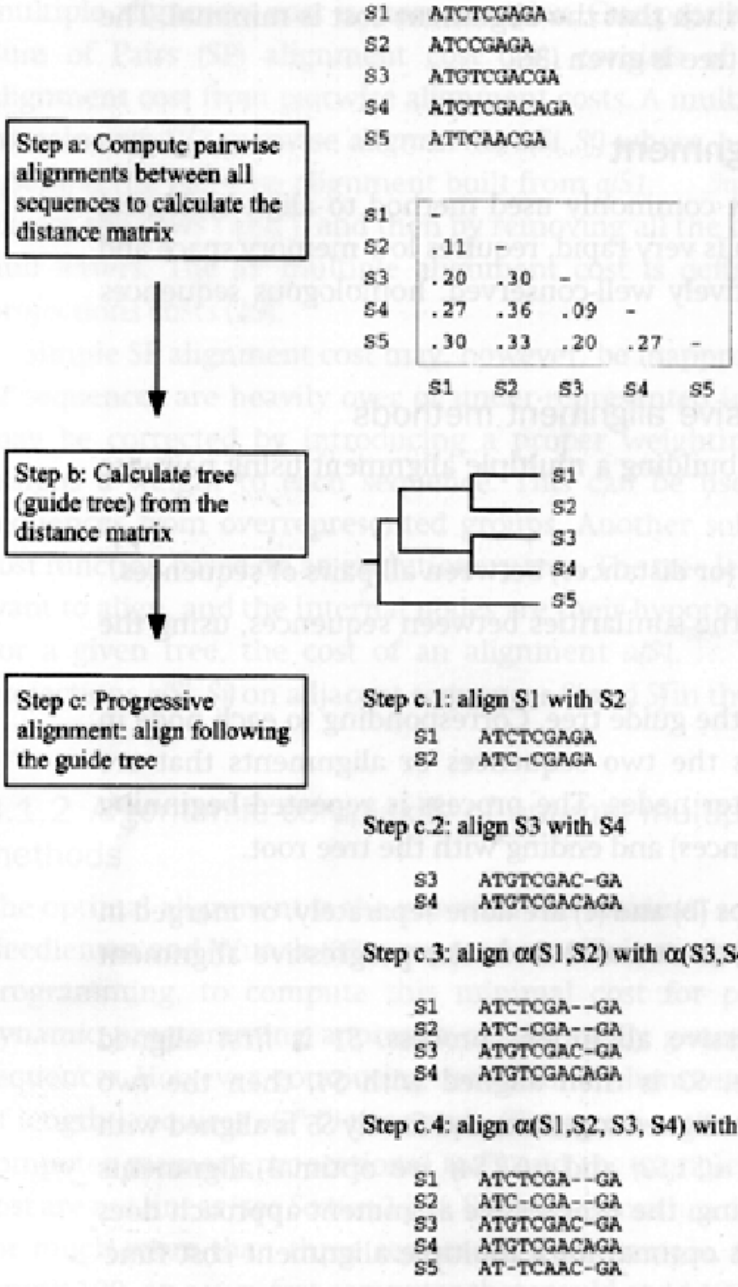

Figure 2 Progressive alignment process. (a) All sequences are compared to each other S_2. (b) A guide tree is calculated from the pairwise distance matrix. (c) Sequences are progressively aligned following the guide tree.