# Algorithmes incrémentaux pour la gestion de grandes familles de séquences homologues

## Implémentation dans le logiciel *Mentalign*

Jean-François Dufayard[1*], Guy Perrière[2] et Manolo Gouy[2]

[1] INRIA Rhône-Alpes, projet Helix, Montbonnot, 38334 Saint Ismier Cedex, France.
[2] UMR 5558, Biométrie, Biologie Évolutive, 69622 Villeurbanne Cedex, France.
[*]Présentateur de l'article. (tél : +33 4 76 61 55 05, fax : +33 4 76 61 54 08).

## Introduction

La gestion de familles de séquences homologues est un problème central de la génomique comparative dont la solution repose sur le calcul d'alignements multiples et la construction d'arbres phylogénétiques associés à ces familles, problèmes bioinformatiques connus pour leur difficulté. De nombreuses méthodes existent pour résoudre ces problèmes mais elles possèdent deux limitations importantes :

- Les familles contenant un petit nombre de séquences (c'est-à-dire moins de 1000) sont en général traitées facilement par les méthodes classiques. Cependant, au-delà d'un millier de séquences, il est assez difficile de construire un arbre phylogénétique et encore plus difficile de calculer un alignement multiple.
- Le séquençage étant devenue une activité intensive, il est très fréquent d'avoir à ajouter de nouvelles séquences à des familles de séquences homologues. Cependant, la plupart des méthodes classiques ne permettent pas d'ajouter des séquences à un alignement ou à un arbre phylogénétique sans les recalculer entièrement. Quant aux méthodes le permettant, soit elles s'avèrent trop lentes pour pouvoir être utilisée efficacement sur des alignements comportant plusieurs milliers de séquences (comme celle implémentée dans Clustal-w [5]), soit elles nécessitent des données qui ne sont pas forcément disponibles (comme les méthodes d'alignement utilisées dans le projet *RDP II* [1] qui utilisent des données relatives aux structures secondaires des ARN étudiés).

## Méthodes

Deux méthodes incrémentales ont donc été créées, résolvant les problèmes suivants :

- ajouter une séquence à un alignement, sans recalculer entièrement ce dernier.
- ajouter une séquence à un arbre phylogénétique, sans le reconstruire entièrement.

Elles sont basées sur les méthodes d'alignement progressif [2] et sur la méthode de reconstruction d'arbre appelée *Neighbor-Joining* [4]. Elles sont applicables à toute famille de séquences homologues, protéiques ou nucléiques.

Elles sont dépendantes l'une de l'autre, car l'arbre phylogénétique et l'alignement multiple sont traités en même temps. En effet, la localisation de la nouvelle séquence dans l'arbre est très dépendante de son alignement avec les autres ; et inversement, l'alignement de la nouvelle séquence avec les autres est très dépendant de sa localisation dans l'arbre. Aussi, ces méthodes incrémentales consistent en re-calculs partiels successifs des positions dans l'arbre et de l'alignement de la nouvelle séquence, jusqu'à stabilisation des données.

L'ajout d'une séquence dans un arbre et un alignement peut se décomposer en deux étapes successives :

- D'abord, la nouvelle séquence est ajoutée à l'arbre par re-calculs successifs de topologie par la méthode du *Neighbor-Joining*. Cette étape peut se formaliser comme suit :

**Soit** $S_{nouv}$ la nouvelle séquence, à ajouter à l'arbre A, selon l'alignement de ses séquences S :

**Supposer** un emplacement de $S_{nouv}$ dans l'arbre A.
**Itérer** :
**Aligner rapidement** $S_{nouv}$ avec S, selon son emplacement supposé dans A.
**Recalculer** la topologie de A autour de l'emplacement de $S_{nouv}$, incluant $S_{nouv}$.
**Jusqu'à** que la position de $S_{nouv}$ dans A soit stable.

- Une fois qu'on a défini l'emplacement définitif de la nouvelle séquence dans l'arbre, la seconde étape consiste à l'aligner avec les séquences de l'alignement initial. En s'inspirant des méthodes d'alignement progressif, et en considérant l'arbre phylogénétique comme guide de l'alignement, on recalcule les alignements d'ensembles à chaque nœud de l'arbre, depuis le nœud d'insertion de la nouvelle séquence et jusqu'à la racine.

Ces méthodes peuvent aussi être considérées comme des méthodes de reconstruction d'arbres et d'alignements à part entière. En effet, elles ont été conçues pour être utilisées dans les cas extrêmes où on ajoute successivement toutes les séquences d'une famille, en partant d'un alignement initial réduit à une seule séquence.

## Implémentation

Ces méthodes ont été implémentées dans un logiciel appelé *Mentalign* codé en Java. *Mentalign* peut visualiser arbres et alignements en mode graphique, calculer entièrement des arbres et des alignements, ou ajouter des séquences à des arbres et des alignements existants.

## Discussion et perspectives

*Mentalign* apporte des méthodes entièrement nouvelles permettant d'ajouter à un alignement et un arbre, même de très grandes tailles, de nouvelles séquences. Donc, par rapport aux deux méthodes existantes (celles implémentées dans Clustal-w et le projet *RDP II*), *Mentalign* améliore de façon importante les performances en terme de taille des familles traitables, et peut s'appliquer à n'importe quelle famille de séquences homologues.

*Mentalign* a été testé sur de nombreuses familles de séquences, tant nucléiques que protéiques, dont certaines comportent plus de 10000 membres (comme par exemple la famille des cytochromes B, issue de la base HOVERGEN). Il semble envisageable d'appliquer ces algorithmes à des familles plus importantes (50000 ou 100000 séquences). Pour l'instant, les résultats d'alignements ont été testés en faisant divers calculs de *Dot Plot* sur des séquences éloignées dans l'arbre, afin de détecter d'éventuelles grosses anomalies. D'une manière générale, les arbres et les alignements obtenus ont été étudiés « à la main », et ont tous donné des résultats satisfaisants (c'est-à-dire ne présentant aucune anomalie importante visible). Cependant, des méthodes objectives de comparaison avec des alignements multiples et des arbres phylogénétiques obtenus par d'autres méthodes sont encore à l'étude.

Plusieurs améliorations peuvent être apportées à *Mentalign* :

- Pour l'instant, les alignements deux à deux d'ensembles de séquences sont assurés par l'algorithme de Needleman et Wunsch [4]. Cependant, il est envisageable d'utiliser d'autres méthodes plus performantes comme par exemple des techniques d'alignements par blocs.
- Il est souhaitable de développer un logiciel plus élaboré, incluant une gestion plus avancée des familles de séquences homologues. Par exemple, il serait pertinent d'ajouter à l'interface un éditeur d'alignement et d'arbre permettant de retoucher à la main certains résultats de *Mentalign*. L'utilisation par l'expert biologiste en serait alors facilitée.
- Enfin, il est nécessaire de choisir ou de définir une méthode pertinente permettant de comparer les résultats de *Mentalign* avec d'autres méthodes d'alignement et de construction d'arbre.

## Références


[1] Cole, J. R., B. Chai, et al. (2003). "*The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy.*" Nucleid Acids Research 1(31): 442-443.

[2] Feng, D. F. and R. F. Doolittle (1987). "*Progressive sequence alignment as a prerequisite to correct phylogenetic trees.*" Journal of molecular biology 25(4): 351-360.

[3] Needleman, S. B. and C. D. Wunsch (1970). "*A general method applicable to the search for similarities in the amino acid sequence of two proteins.*" Journal of molecular biology 48(1): 443-453.

[4] Saitou, N. and M. Nei (1987). "*The neighbor-joining method: a new method for reconstructing phylogenetic trees.*" Molecular Biology and Evolution 4(4): 406-425.

[5] Thompson, J. D., D. G. Higgins, et al. (1994). "*Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice.*" Nucleid Acids Research 22: 4673-4680.


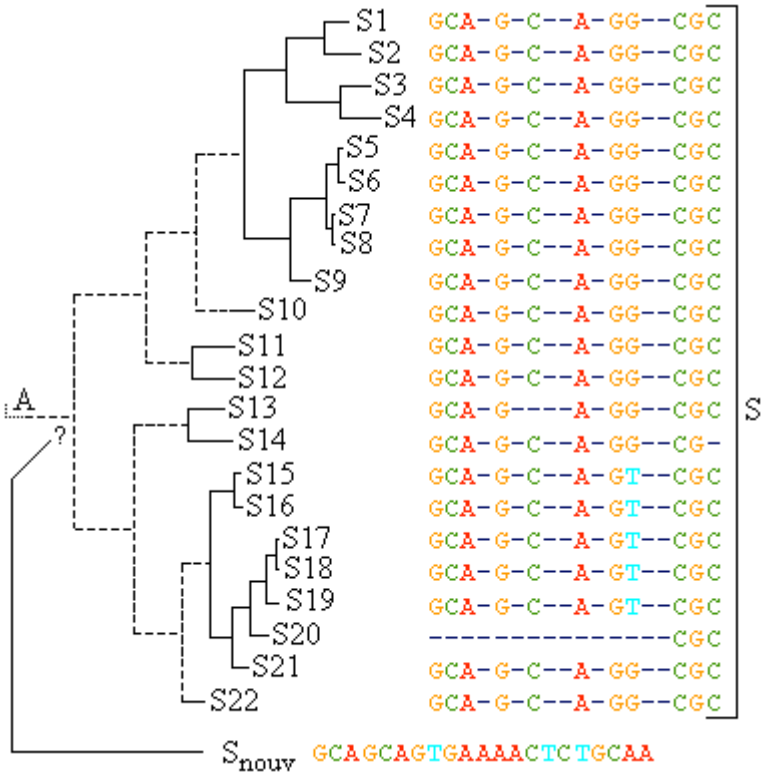


**Figure 1 – Étape d'insertion d'une nouvelle séquence dans un arbre, selon l'alignement multiple de ses séquences.** La nouvelle séquence $S_{nouv}$ (en bas) n'est pas alignée avec les séquences S de l'alignement (à droite). Supposant une position de $S_{nouv}$ dans l'arbre A de S (à gauche), on aligne rapidement $S_{nouv}$ et ses séquences voisines dans A, et on recalcule une partie de la topologie de A autour la position de $S_{nouv}$. Cette procédure est itérée jusqu'à que la position de $S_{nouv}$ se stabilise dans A.