Détection de régions génomiques homologues par un algorithme de flot avec coûts

Eric Audemard, Thomas Faraut, Thomas Schiex

INRA Toulouse ; Chemin de Borde Rouge BP 52627, 31326 Castanet Tolosan cedex , France
{eric.audemard,thomas.faraut,thomas.schiex}@toulouse.inra.fr

Mots-Clés : *bioinformatique, évolution des génomes, génomique comparative, flots, réseaux de transport.*

L'identification des régions génomiques homologues, résultant des évènements évolutifs de spéciation ou de duplication, est une étape importante dans l'étude des génomes. Qu'elles soient issues de duplications segmentales, à l'intérieur d'un génome, ou qu'elles définissent des blocs de synténie entre deux espèces, les régions homologues identifiées sont essentielles pour inférer la fonction des gènes (détection de gènes paralogues ou orthologues), pour la reconstruction de génomes ancestraux ou encore la détection d'éléments fonctionnels tels que les éléments de régulation. Les duplications peuvent également être à l'origine de caractères spécifiques d'espèces et pourraient expliquer une part importante de la variabilité génomique entre individus.

Peu après la séparation de deux espèces, la conservation des génomes est telle qu'il est facile de les comparer (c'est à dire de les aligner) et d'identifier les longs segments chromosomiques conservés. Lorsque la spéciation est plus ancienne, l'identification des régions conservées est rendue difficile par la poursuite des différents mécanismes évolutifs (mutations ponctuelles, réarrangements...). Ces modifications peuvent mener à la disparition locale de la similarité, à des changements d'orientation ou d'ordre entre éléments inclus dans la région (gènes, régulateurs...). Les seuls indices aisément identifiables permettant de retrouver ces régions homologues sont de courtes régions suffisamment similaires pour pouvoir être alignées (aussi appelées ancres), traces potentielles de l'homologie entre les deux régions. La densité de ces ancres dans une région, la succession d'une quantité importante d'ancres dans un ordre et/ou une orientation identique sont autant d'éléments permettant de suggérer l'existence d'une relation d'homologie.

L'identification de segments homologues entre deux génomes est une extension naturelle du problème d'alignement entre deux séquences. L'une des formalisations du problème d'alignement, identifie un segment à un chemin dans un graphe. Dans le cas de la recherche de segments homologues, les sommets représentent les ancres et les arcs la possibilité que deux ancres appartiennent à un même segment. Cette formalisation est utilisée implicitement ou explicitement par de nombreuses méthodes [2, 3] dédiées à la détection de segments homologues. Une fois le graphe créé, elles utilisent une approche gloutonne, utilisant un algorithme de "plus court chemin", pour sélectionner les chemins un par un. Il existe d'autres méthodes [4] avec des stratégies différentes, mais la sélection des chaînes reste gloutonne. Cependant les régions homologues ne sont pas indépendantes, elles sont le fruit d'une histoire évolutive cohérente composée d'évènements de spéciation et/ou de duplication. Les méthodes gloutonnes créent les chemins un par un sans tenir compte de la cohérence globale. Elles ne garantissent pas que le score de l'ensemble des chemins sera optimum.

La principale nouveauté de notre approche réside dans la proposition de formaliser le problème de

la recherche de segments homologues, sous la forme d'une reconstruction simultanée d'un ensemble de chemins de score optimum. Pour cela nous utilisons un algorithme de "flot de coût mininimum" [1], de complexité en $O(CVE)$ avec V (resp. E) le nombre de sommets (resp. d'arcs) du graphe et $C \leq V$ le nombre de chemins créés. De plus, la construction d'un "flot" permet de prendre en compte *a priori* des contraintes s'appliquant de façon globale à l'ensemble des chemins (par exemple une ancre ne doit appartenir qu'à un seul chemin). Ce qui est impossible avec les méthodes gloutonnes, qui génèrent des chemins indépendants. De plus la prise en compte de contraintes supplémentaires permet d'adapter simplement l'algorithme à l'identification de segments homologues au sein d'un même génome (appelé duplications segmentales). Nous avons implémenté notre méthode dans le logiciel ReD (**Re**gions **D**upliquées).

Nous avons évalué notre algorithme dans le cas de la recherche de duplications segmentales et de régions homologues à partir de séquences protéiques. Dans ces conditions les ancres sont définies par les similarités locales entre gènes. Nous avons commencé par comparer une méthode classique de "plus court chemin" avec la méthode de "flot". Les résultats montrent que l'approche gloutonne fournit des scores inférieurs et que cette différence augmente avec le nombre de chaînes construites. D'autre part l'approche par flot de coût minimum permet de construire des chemins plus longs et avec une meilleure linéarité des segments : les choix locaux réalisés par l'approche gloutonne pénalisent la qualité globale des chemins.

Ensuite, nous avons comparé ReD aux logiciels DAGchainer [2] et OSfinder [3]. Comme ces logiciels ne peuvent pas intégrer les contraintes globales, ils créent des ensembles de chemins qui possèdent entre 10% et 50% de chemins incohérents. Il est possible de filtrer ces chemins, mais le nombre de régions détectées diminue d'autant. Une alternative consiste à trouver un sous-ensemble de chemins cohérents mais ce n'est pas un problème simple, les chemins incompatibles s'excluant mutuellement. ReD règle ce problème en amont, dans son formalisme, et dans l'algorithme de flot associé, ce qui a l'avantage de créer un ensemble de chemins cohérent.

Cependant, la prise en compte de certaines contraintes au sein de son formalisme n'est pas encore idéale et peut parfois mener à une fragmentation excessive des chemins, dans les génomes fortement dupliqués. Enfin et surtout, il reste à évaluer les performances de ReD dans un contexte plus difficile et encore peu exploré : l'exploitation d'ancres détectées directement au niveau de l'ADN qui est largement plus bruitée qu'au niveau des protéines (utilisé traditionnellement pour détecter les ancres). ReD pourra alors être utilisé avant le processus d'annotation des génomes (détection des régions correspondant aux protéines) et ses sorties pourront alors alimenter ce processus de prédiction.

Références

[1] R. G. Busacker and P. J. Gowen. A procedure for determining minimal-cost network flow patterns. In *ORO Technical Report 15*, 1961.

[2] B. J. Haas, A. L. Delcher, J. R. Wortman, and S. L. Salzberg. Dagchainer : a tool for mining segmental genome duplications and synteny. *Bioinformatics*, 20(18) :3643–3646, December 2004.

[3] K. Popendorf T. Hachiya, Y. Osana and Y. Sakakibara. Accurate identification of orthologous segments among multiple genomes. *Bioinformatics*, 25, February 2009.

[4] K. Vandepoele, Y. Saeys, C. Simillion, J. Raes, and Y. Van De Peer. The automatic detection of homologous regions (adhore) and its application to microcolinearity between arabidopsis and rice. *Genome Res*, 12(11) :1792–1801, November 2002.