Intervalles conservés et Neighbor-Joining

Anne Bergeron, Cedric Chauve
LaCIM and Département d'Informatique
Université du Québec à Montréal

Mathieu Blanchette
McGill Center for Bioinformatics and School of Computer Science, McGill University

Proceedings of the Rencontres de la Société Francophone de Classification 2004, pages 111-114.

Intervalles conservés et Neighbor-Joining

Anne Bergeron[1], Mathieu Blanchette[2] et Cedric Chauve[1]

(1)LaCIM et Département d'Informatique, Université du Québec à Montréal, Canada.
`[anne,chauve]@lacim.uqam.ca`
(2)McGill Center for Bioinformatics et School of Computer Science, McGill University, Montréal, Canada.
`blanchem@mcb.mcgill.ca`

RÉSUMÉ. *Nous présentons dans cette note une application de la notion d'intervalles conservés (introduite par Bergeron et Stoye, [BER 03]) dans le cadre de la phylogénie, et plus particulièrement de l'algorithme Neighbor-Joining utilisé avec des données basées sur l'ordre des gènes. Nous proposons notamment une modification de cet algorithme permettant d'obtenir des informations sur de possibles génomes ancestraux. Nous étudions la pertinence de ces résultats dans le cadre d'un jeu de données constitué de génomes de mitochondries de Métazoaires, déjà considéré par Blanchette et al. [BLA 99] et par Bourque et Pevzner [BOU 02].*

MOTS-CLÉS : *Phylogénies, Ordre des gènes, Intervalles conservés, Neighbor-Joining, Mitochondries, Métazoaires.*

1. Introduction

Dans cet article, nous nous intéressons à la reconstruction de phylogénies basée sur l'étude de l'ordre des gènes. Cette approche, qui repose entre autres sur la modélisation de génomes par des permutations signées, a connu de nombreux développements récents [MOR 02, SAN 98, BOU 02] (voir aussi [SAN 00] pour une revue plus large des problèmes reliés à l'étude de l'ordre des gènes). Les algorithmes proposés dans ce cadre reposent sur le calcul du médian de trois génomes pour une distance donnée, basée sur la notion de réarrangements génomiques, (distance de breakpoints ou d'inversions, avec les méthodes BPAnalysis [SAN 98], GRAPPA [MOR 02], MGR-MEDIAN [BOU 02]). Ce problème du médian est NP-complet mais peut être résolu de manière non exacte à l'aide d'heuristiques efficaces, ce qui permet en particulier de proposer des génomes ancestraux des organismes étudiés, c'est-à-dire des génomes potentiels pour les nœuds internes de l'arbre phylogénétique reconstruit. Cette capacité à reconstruire un génome ancestral est un des points forts des méthodes basées sur l'ordre des gènes.

Récemment, Bergeron et Stoye [BER 03] ont introduit le concept d'*ensemble d'intervalles conservés* (Section 2) d'un ensemble de génomes représentés par des permutations signées, qui permet de décrire de manière compacte certains groupes de gènes qui sont consécutifs dans tous ces génomes, groupes de gènes qu'il est tentant de considérer comme des invariants, en termes d'ordre des gènes, provenant d'ancêtres communs. L'intérêt de cette notion repose entre autres sur le fait que, le nombre de scenarios d'évolution les plus parcimonieux, en termes de réarrangements génomiques, entre deux génomes pouvant être exponentiel en la taille de ces génomes [BER 02], le nombre de génomes ancestraux possibles est très grand, et le choix de l'un d'entre eux en particulier devient alors un choix arbitraire. De son côté, la notion d'intervalles conservés permet de représenter un ensemble de génomes ancestraux sans avoir à en exhiber un en particulier.

Le but de notre travail est d'effectuer une première exploration de l'utilisation de la notion d'intervalles conservés dans le cadre du calcul de phylogénies basées sur l'ordre des gènes. Plus précisément, nous proposons une adaptation de l'algorithme Neighbor-Joining, et en fait de tout algorithme de clustering hiérarchique, basée sur les intervalles conservés, permettant donc d'étudier des génomes décrits par l'ordre de leurs gènes et de déterminer des informations intéressantes sur leurs génomes ancestraux (Section 3). Nous utilisons notre méthode sur un jeu de données de génomes de mitochondries de Métazoaires étudié par Blanchette et al. [BLA 99], ainsi que par Bourque et Pevzner [BOU 02] (Section 4). Nous obtenons des résultats qui sont cohérents avec ceux de [BLA 99, BOU 02], notamment en termes des génomes ancestraux proposés, tout en permettant d'apporter certaines précisions et corrections sur la structure de ces génomes et des scenarios d'évolution proposés.

2. Génomes, permutations signées et intervalles conservés

Soient $\mathcal{A}$ un ensemble de n gènes, et $\mathcal{G}$ un ensemble de génomes sur $\mathcal{A}$ sans répétition de gène : dans chaque génome, chaque gène apparaît une et une seule fois, et est signé $+$ ou $-$ dépendant du brin d'ADN sur lequel il est exprimé. Nous utilisons la représentation classique d'un tel ensemble de génomes par un ensemble de permutations signées sur $\{1,\ldots,n\}$, chaque gène $\mathcal{A}$ correspondant à un élément de $\{1,\ldots,n\}$. Voici par exemple, un ensemble de 3 génomes sur 10 gènes.

$$\begin{array}{rcccccccccc} G_1 = & +1 & -2 & +3 & +4 & +5 & +6 & -7 & -8 & +9 & +10 \\ G_2 = & +1 & -5 & +2 & -4 & -3 & +6 & +7 & -8 & +9 & +10 \\ G_3 = & +1 & -5 & +2 & -4 & -3 & +6 & +8 & -7 & +9 & +10 \end{array}$$

La notion d'intervalles conservés d'un ensemble de permutations signées a été introduite récemment par Bergeron et Stoye [BER 03] comme une mesure de similarité entre un ensemble de génomes. Soit $\mathcal{G}$ un ensemble de permutations signées sur $\{1,\ldots,n\}$. Un intervalle $[a,b]$ (où a et b sont deux entiers signés sur $\{1,\ldots,n\}$) est un intervalle conservé de $\mathcal{G}$ si les conditions suivantes sont vérifiées :

1. dans toute permutation de $\mathcal{G}$, soit a précède b, soit $-b$ précède $-a$,
2. les éléments situés entre a et b (ou $-b$ et $-a$) sont les mêmes, au signe près, dans toute permutation de $\mathcal{G}$.

On note $IC(\mathcal{G})$ l'ensemble des intervalles conservés de $\mathcal{G}$: $IC(\{G_1,G_2,G_3\}) = \{[1,6],[3,4],[6,10],[6,9],[9,10],[1,10]\}$.

Intuitivement, on peut appréhender l'ensemble des intervalles conservés de $\mathcal{G}$ comme l'ensemble des groupes de gènes qui ne devraient pas être séparés par un scénario d'évolution plausible pour ce groupe de génomes, et qui devraient donc se retrouver ensembles dans les génomes des ancêtres communs aux organismes considérés. $IC(\mathcal{G})$ peut aussi être vu comme une représentation de l'ensemble des permutations signées respectant ces intervalles, c'est-à-dire l'ensemble des ancêtres potentiels possibles vis-à-vis d'un scenario d'évolution, pour les génomes de $\mathcal{G}$, ne cassant aucun de ces intervalles.

Cette mesure de similarité entre génomes induit une distance entre deux ensembles de permutations signées : la distance entre $\mathcal{G}_1$ et $\mathcal{G}_2$, notée $d(\mathcal{G}_1,\mathcal{G}_2)$, est donnée par $|IC(\mathcal{G}_1)|+|IC(\mathcal{G}_2)|-2|IC(\mathcal{G}_1\cup\mathcal{G}_2)|$. Bergeron et Stoye [BER 03] ont proposé un algorithme linéaire en temps permettant, étant donnés deux ensembles d'intervalles conservés $\mathcal{I}_1$ et $\mathcal{I}_2$, correspondant à deux ensembles de permutations $\mathcal{G}_1$ et $\mathcal{G}_2$, de calculer l'union de ces deux ensembles d'intervalles conservés (c'est-à-dire l'ensemble d'intervalles conservés de $\mathcal{G}_1\cup\mathcal{G}_2$).

3. Neighbor-Joining et intervalles conservés

Le principe général des algorithmes de clustering hiérarchiques basés sur l'analyse de matrices de distances est le suivant (voir [FEL 03] pour une description précise de plusieurs méthodes, incluant Neighbor-Joining) :

1. la matrice de distance est indexée par un ensemble de groupes de taxa, chaque groupe étant structuré en arbre phylogénétique, et au départ chaque groupe est réduit à un seul taxon,

2. deux groupes sont choisis, selon un critère numérique calculable à partir de la matrice de distances et variant selon les algorithmes, pour être fusionnés en un seul groupe,

3. puis les deux groupes choisis sont supprimés de la matrice et remplacés par le nouveau groupe créé, dont les distances aux autres groupes sont calculées, selon une méthode variant selon les algorithmes, et l'on retourne à l'étape 2 si la matrice comporte plus de deux groupes.

La caractéristique de base de notre algorithme consiste à utiliser la distance d'intervalles entre groupes de taxa (génomes).

La première innovation concerne l'étape 2 du schéma algorithmique décrit ci-dessus. Elle consiste à associer à chaque groupe de taxa $\mathcal{G}$ utilisé pour indexer la matrice de distance, c'est-à-dire à chaque nœud sommet de l'arbre phylogénétique calculé, l'ensemble des intervalles conservés des génomes de ce groupe, à savoir $IC(\mathcal{G})$. Ainsi, à la différence des versions classiques des algorithmes de Neighbor-Joining, chaque nœud interne N de l'arbre phylogénétique contient un codage compact de l'ensemble des génomes ancestraux possibles, pour les taxa $\mathcal{G}_N$ contenus dans le sous-arbre de racine N, compatibles avec les intervalles conservés des génomes de $\mathcal{G}_N$.

La seconde innovation consiste, lors de l'étape 3, à calculer la distance du nouveau groupe créé, notons-le $\mathcal{G}_N$, à tout groupe restant dans la matrice, par exemple un groupe $\mathcal{G}_M$, de manière non purement numérique, c'est-à-dire en calculant une moyenne de la distance à $\mathcal{G}_M$ des taxa présents dans $\mathcal{G}_N$, mais en calculant la distance d'intervalles entre les ensembles d'intervalles conservés $IC(\mathcal{G}_N)$ et $IC(\mathcal{G}_M)$.

Ces deux innovations induisent une augmentation, par rapport aux versions classiques de Neighbor-Joining, de la complexité en temps et en espace du calcul d'un arbre phylogénétique pour m taxa décrits par des génomes sur n gènes : l'étape 3 demande d'effectuer le calcul de l'union d'intervalles conservés, ce qui induit une augmentation du temps de calcul d'un facteur $O(n)$, et le stockage sur chaque nœud interne d'un ensemble d'intervalles conservés demande un accroissement de l'espace nécessaire au stockage d'un sommet en $O(n)$. Cependant, l'algorithme proposé reste polynomial en temps de calcul ($O(n^3m)$) et en espace ($O(m(mn))$).

4. Application aux génomes de mitochondries de Métazoaires

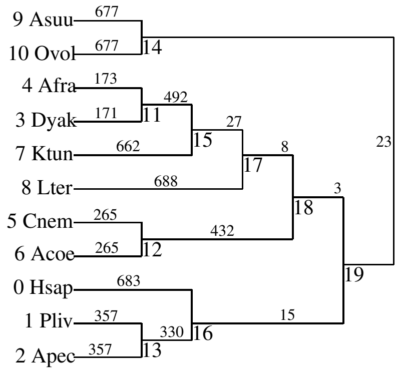
Nous avons appliqué notre méthode avec l'algorithme Neighbor-Joining (tel que décrit dans [FEL 03, p. 167]) à l'ensemble de génomes de mitochondries de Métazoaires étudié par Blanchette et al. [BLA 99] et par Bourque et Pevzner [BOU 02].

Ces données se composent des organismes suivants : *Homo sapiens*, (Hsap, groupe des Chordés), *Asterina pectinifera* et *Paracentrotus lividus* (Apec et Pliv, groupe des Echinodermes), *Drosophila yakuba* et *Artemia franciscana* (Dyak et Afra, groupe des Arthropodes), *Albinaria coerulea*, *Cepaea nemoralis* et *Katharina tunicata* (Acoe, Cnem et Ktun, groupe des Mollusques), *Lumbricus terrestris* (Lter, groupe des Annelida), *Ascaris suum* et *Onchocera volvulus* (Asuu et Ovol, groupe des Nematodes). Pour chacun de ces organismes, suivant [BOU 02], nous avons considéré leurs génomes mitochondriaux, dont l'ordre des gènes est disponible dans la base de données *Mitochondrial Gene Arrangement Guide 6.0* maintenue par Jeffrey L. Boore. Ces génomes partagent essentiellement le même ensemble de 36 gènes (les gènes `atp8` et `UNK` ont été supprimés des génomes les contenant car ils ne sont pas présents pas dans tous les génomes).

L'arbre obtenu est décrit en Figure 1. On peut notamment remarquer qu'à l'exception du taxon *Katharina tunicata*, dont la "mobilité" posait aussi problème dans [BLA 99, BOU 02], tous les clades importants ont été reconstruits, et notamment les deusteromes (Chordés et Echinodermes), réunion que les méthodes Neighbor-Joining et Fitch classiques basées sur les distances d'inversions ou de breakpoints ne retrouvaient pas [BLA 99]. Ce point important semble indiquer que la distance d'intervalles est une bonne mesure de distance d'évolution en termes de réarrangements génomiques. De plus, les branchements différents entre l'arbre que nous exhibons et les arbres produits par [BLA 99, BOU 02], notamment la position du groupe des Arthropodes, ne concernent que des branchements correspondant à de courtes branches, à la fois dans notre arbre et dans ceux de [BLA 99, BOU 02], ce qui incite à relativiser ces différences.

Concernant les ensembles d'intervalles conservés étiquetant les nœuds internes de cet arbres (les "génomes ancestraux"), par manque de place nous ne pouvons les inclure dans cette note, mais ils sont disponibles à l'url `www.lacim.uqam.ca/~chauve/SFC04/`. On peut faire les constations suivantes. Premièrement, les ensembles d'intervalles conservés assignés à chacun des ancêtres des groupes Echinodermes, deusteuromes (Echinodermes et Chordés), Arthropodes, Arthropodes+Ktun, Arthropodes+Mollusques+Annelidés, et Nematodes n'indiquent aucune contradiction avec les génomes ancestraux proposés dans [BLA 99, BOU 02], bien que nous considérions

FIG. 1. *Arbre obtenu par Neighbor-Joining et intervalles conservés, avec longueur de branches.*



dans notre étude l'ensemble des 36 gènes communs aux 11 génomes à la différence de [BLA 99] qui ne prenaient pas en compte les ARN de transfert. De plus, les structures d'intervalles conservés proposés permettent d'affiner l'information sur ces génomes ancestraux en précisant quels groupes de gènes ont évolués ensembles, par exemple le groupe (cox2, K, atp6, cox3) chez les deusteuromes. Deuxièmement, les intervalles conservés de l'ancêtre immédiat (sommet 12) des taxa Acoe et Cnem (Mollusques privé de Ktun) conservent l'ARN de transfert P dans un groupe contenant les gènes A (tRNA-Ala) et nad6 et situé en début de génome, alors que MGR-MEDIAN rejette ce gène P en fin du génome ancestral, ce qui semble en contradiction avec la structure des génomes de Acoe et Cnem, dans lesquels P est en début de génome, proche de A et nad6. La solution proposée par notre algorithme, qui ne repose pas sur l'optimisation d'un scénario en termes de réarrangements mais sur une mesure de similarité, semble alors plus logique du point de vue de la structure de génomes de Acoe et Cnem.

En conclusion, cette étude préliminaire semble indiquer que les notions d'intervalles conservés et de distance d'intervalles sont pertinentes dans le cadre de l'inférence de phylogénies basées sur l'ordre des gènes, notamment pour la reconstruction de génomes ancestraux, et son étude devrait être étendue à des méthodes autres que Neighbor-Joining.

5. Bibliographie

[SAN 98] Sankoff, D. et Blanchette, M., Multiple Genome Rearrangement and Breakpoint Phylogeny, *J. Comp. Biol.*, vol. 5, n^o 3, p. 555-570, 1998.

[BLA 99] Blanchette, M., Kunisawa, T. et Sankoff, D., Gene Order Breakpoint Evidence in Animal Mitochondrial Phylogeny, *J. Mol. Evol.*, vol. 49, n^o 2, p. 193-203, 1999.

[SAN 00] Sankoff, D. and Nadeau, J. (éditeurs), *Comparative Genomics. Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignement and the Evolution of Gene Families*, Kluwer Academic Publishers, 2000.

[BOU 02] Bourque, G. et Pevzner, P., Genome-Scale Evolution : Reconstructing Gene Orders in the Ancestral Species, *Genome Res.*, vol. 12, n^o 1, p. 26-36, 2002.

[BER 02] Bergeron, A., Chauve, C., Hartman, T. et StOnge, K., Properties of Sequences of Reversals that Sort a Signed Permutations, *Actes de JOBIM 2002*, p. 99-108, 2002.

[MOR 02] Moret, B., Tang, J., Wang, L., et Warnow, T., Steps toward accurate reconstruction of phylogenies from gene-order data, *J. Comput. Syst. Sci.*, vol. 65, n^o 3, p. 508-525, 2002.

[FEL 03] Felsenstein, J., *Inferring Phylogenies*, Sinauer Associates, Inc., 2003.

[BER 03] Bergeron, A. et Stoye, J., On the Similarity of Sets of Permutations and its Applications to Genome Comparison, *Actes de COCOON 2003* vol. 2697 de *Lecture Notes in Comput. Sci.*, p. 68-79, 2003.