Prédiction de synténies dans le génome ancestral des amniotes

Frédéric Boyer[1], Cedric Chauve[2] and Éric Tannier[3]

[1] Institut de Recherches en Technologies et Sciences pour le Vivant ; Laboratoire Biologie, Informatique et Mathématiques ; CEA Grenoble, F-38000 Grenoble, France
frederic.boyer@cea.fr

[2] Dept. Mathematics, Simon Fraser University, Burnaby (BC), Canada ; CGL and LaCIM, Université du Québec À Montréal, Montréal (QC), Canada.
cedric.chauve@sfu.ca

[3] INRIA Rhône-Alpes, France ; Université de Lyon, F-69000, Lyon ; Université Lyon 1 ; CNRS, UMR5558, Laboratoire de Biométrie et Biologie Évolutive, F-69622, Villeurbanne, France
eric.tannier@inria.fr

Abstract: *We describe a general methodological framework for reconstructing ancestral genome segments and chromosomes based on conserved synteny. This framework can be applied to various kinds of genomes, whether or not large duplications have occured in some species. It benefits (1) from a general principle of detecting homologies between chromosomal segments with large losses of similarity due to duplications and losses, and (2) from a combinatorial framework used to aggregate synteny in physical mapping studies. We illustrate the possibility of handling largely duplicated genomes by reconstructing synteny in the amniote ancestor, using mammalian, chicken and teleost genomes. We compare the contigs we obtain with other published propositions on the same genome, and discuss the convergences and divergences.*

Keywords: Ancestral chromosomes, ancestral synteny, genome rearrangements, amniote genomes, physical mapping, double synteny, whole genome duplication.

1 Introduction

La reconstruction de karyotypes ancestraux en utilisant des homologies entre segments chromosomiques d'espèces actuelles a été initiée dans les années 80 par des techniques de cytogénétique (bandes chromosomiques, hybridation in-situ) et appliquée à des génomes de mammifères (voir par exemple [20]). Au-delà de cette échelle évolutive, les homologies sont moins visibles, et il a fallu attendre des données de séquence génomiques et des méthodes bioinformatiques pour pouvoir prédire le passé des chromosomes. Les premières méthodes ont été appliquées sur les génomes de mammifères (voir par exemple [5,15], et une revue dans [17]), montrant plus ou moins de convergence avec les résultats cytogénétiques (voir des discussions sur les divergences [9,4]) ; leur application à des génomes ancestraux plus anciens se heurte toujours à l'impossibilité pour ces méthodes de prendre en compte des duplications comme évènements évolutifs. Or, les génomes des poissons téléostes ont subi une duplication globale [12] dans une branche ancienne de leur évolution, et sont les seuls génomes séquencés et assemblés exploitables pour reconstruire les chromosomes de l'ancêtre des amniotes, ou des tétrapodes. D'autres méthodes plus récentes et très différentes ont été mises au point pour reconstruire le génome ancestral des amniotes [18,14], et le manque d'un cadre méthodologique formel se retrouve dans ces premiers résultats, qui prédisent par exemple de 18 chromosomes [14] à 26

chromosmes [18] et des associations synténiques nettement divergentes, avec les mêmes données et le même but : reconstruire les chromosomes de l'ancêtre des vertébrés.

Nous avons développé un cadre méthodologique général pour la reconstruction de chromosomes ancestraux (pour une description plus complète, voir [7]). Il est inspiré par des approches développées pour la cartographie physique ou l'assemblage des chromosomes, et permet aussi bien de reconstruire des segments de chromosomes d'ancêtres de mammifères [7] que d'ancêtres plus éloignés nécessitant la prise en compte de larges duplications dans le génome. La méthode consiste à appliquer successivement deux phases :

- À partir de marqueurs génomiques et des relations d'homologie entre différentes espèces actuelles, nous construisons tout d'abord une collection de groupes de marqueurs qui sont probablement synténiques (ici, contigus) dans le génome ancestral : les "syntons" ancestraux.
- À partir de la collection de syntons ancestraux, nous construisons une structure combinatoire, le PQ-arbre ([3]) des syntons, qui contient toutes les possibilités de segments chromosomiques ancestraux dans lesquels les groupes synténiques sont contigus ; éventuellement certains syntons sont éliminés durant cette phase si leur présence est incompatible avec l'existence d'une solution.

La représentation du résultat sous forme de PQ-arbre permet d'éviter de choisir, par de l'optimisation ou un critère arbitraire, entre plusieurs solutions possibles, souvent équivalentes : le PQ-arbre fournit toutes les solutions de façon compacte.

La première phase est simplement une phase de détection de synténies, qui peut être implémentée de différentes façons. Nous utilisons un cadre très général pour détecter tous les blocs synténiques, en présence, ou non, de duplication et/ou pertes de matériel génomique. Par contre, ce cadre contraint à l'absence de marqueurs indifférenciés chez l'ancêtre reconstruit : il n'y aura qu'un exemplaire par marqueur génétique dans tout ancêtre reconstruit de cette façon. Nous décrivons plus précisément cette méthode dans la section 2. La seconde phase profite d'un cadre mathématique développé pour le problème combinatoire appelé "problème des uns consécutifs" [3], étudié notamment dans le cadre de la cartographie physique des chromosomes (voir par exemple [1]). Cette phase est décrite dans la section 3. En assemblant des synténies communes à des génomes de mammifères (humain, macaque, souris, rat, chien, vache, opossum), du poulet, et de certains poissons (tetraodon, medaka, poisson zèbre), nous obtenons finalement un jeu de segments chromosomiques qui sont probablement présents dans le génome ancestral des amniotes. Les résultats sont présentés et discutés dans la section 4.

2 Détection de synténies

Pour détecter un ensemble de marqueurs qui forment potentiellement un groupe contigu dans le génome de l'ancêtre de deux espèces actuelles, nous détectons les groupes de marqueurs contigus, et éventuellement avec un ordre différent, dans les deux espèces, et avec une souplesse réglée par un paramètre δ pour la contiguité (il peut y avoir des "trous", résultant d'insertions et/ou pertes, de taille δ dans le groupe) ; δ peut prendre une valeur différente pour les deux espèces. Formellement, les groupes sont des "équipes de gènes", dont les définitions et algorithmes de détections, fondés sur des méthodes de partitionnements de graphes, sont étudiés dans [10,2,6]. Sur cette base, la duplication globale de génome chez les poissons téléostes nécessite un traitement supplémentaire quand un poisson entre dans la comparaison.

2.1 En l'absence de duplications

Un synton ancestral en l'absence de duplications (dans la comparaison entre un mammifère et un oiseau par exemple) est simplement l'ensemble des marqueurs contigus dans la comparaison entre deux amniotes dont le dernier ancêtre commun est l'ancêtre de tous les amniotes (en pratique, le poulet et un mammifère). En présence de gènes dupliqués et pour permettre l'utilisation d'un paramètre δ différent pour les deux espèces, nous avons utilisé l'implémentation de [6].

2.2 En présence de duplications

Pour les comparaisons avec les poissons, la méthode décrite ci-dessus est plus difficile à appliquer, de par les nombreux réarrangements intra-chromosomiques qui ont bouleversé les syntons dans les chromosomes de poissons. Un paramètre δ beaucoup plus important est nécessaire pour capturer les synténies homologues, mais la spécificité pâtit de l'augmentation de ce paramètre. Pour pallier cette baisse de spécificité, nous utilisons le principe des "doubles synténies", pour retrouver des paralogies après de larges pertes de similarités.

En effet, des pertes massives de matériel génomique suivent habituellement une duplication, et deux segments paralogues ne présentent souvent plus de similarité après ces pertes. On peut par contre retrouver ces paralogies en comparant un génome dupliqué à un autre non dupliqué. C'est la méthode dite du "pivot", initiée par [13], [8], utilisée par [12,18] et systématisée par [19].

Ici, pour maintenir une unité dans les detections de segments ancestraux, nous proposons une méthode générale pour implémenter le principe du pivot, et de là retrouver des syntons ancestraux avec une bonne spécificité. Nous calculons donc les marqueurs contigus dans un génome dupliqué (poisson) et un non dupliqué (amniote) par la méthode de la section 2.1, avec un paramètre δ de l'ordre d'un chromosome entier pour le poisson, et plus restreint (du même ordre que pour la comparaison de deux amniotes) pour l'amniote. Pour chaque groupe de gène détecté comme synténique chez le poisson et co-localisés sur le génome amniote, son support est le segment chromosomique minimal de l'amniote qui contient tous les marqueurs du groupe. Si les supports de deux groupes s'intersectent, alors on définit un synton ancestral comme l'ensemble des marqueurs présents dans l'intersection des deux support.

Une illustration de cette méthode est présentée sur la Figure 1.

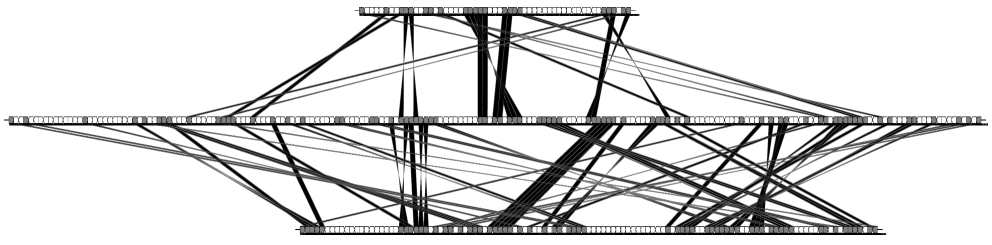

Fig. 1. Une double synténie : un segment de chromosome amniote (au centre) est homologue à deux segments de poissons (en haut et en bas), alors que la paralogie entre les deux segments de poissons n'est pas détectable (peu de gènes sont conservés en deux copies). On peut remarquer que les syntons sont très réarrangés, ce qui justifie la méthode de détection et la méthode de reconstruction de l'ancêtre, indépendantes de l'ordre des gènes dans un synton.

Dans un cas comme dans l'autre (duplications ou non), nous ne conservons que les syntons non conflictuels (ne chevauchant pas d'autres marqueurs dans les génomes amniotes), plus grand qu'une taille minimale, fixée en paramètre (ici, couvrant au moins 500Kb).

3 Assemblage des synténies

En sortie de la première phase, nous disposons d'une famille $\mathcal{F} = \{F_1, \ldots, F_m\}$ de syntons ancestraux, c'est-à-dire m sous-ensembles de l'ensemble des marqueurs $\{1, \ldots, n\}$. Chaque élément de $\mathcal{F}$ est un groupe de marqueurs prédits comme étant contigus dans le génome ancestral. Le problème est alors d'ordonner les marqueurs de façon à ce que chaque synton soit effectivement contigu.

C'est là qu'intervient la méthodologie utilisée en cartographie physique de génomes : on construit une matrice $\mathcal{M}$ de taille $m \times n$, dont les colonnes sont les marqueurs et les lignes les éléments de $\mathcal{F}$: $\mathcal{M}_{i,j} = 1$ si le gène j appartient au synton F_i, et $\mathcal{M}_{i,j} = 0$ sinon. Ordonner les marqueurs revient ici à trouver une permutation des colonnes de la matrice, de façon à ce que dans chaque ligne, les entrées "1" soient consécutives. C'est le problème classique des "uns consécutifs". Si une solution existe, un algorithme [16] donne la solution en temps linéaire. Sinon, certains groupes de marqueurs ne sont pas des syntons ancestraux, et il faut les enlever de l'instance. Enlever un nombre minimum de lignes pour que la matrice ait la propriété des uns consécutifs est un problème NP-complet, et nous avons programmé un algorithme de type brancher/élaguer (branch&bound) qui peut trouver la solution optimale si le nombre de syntons à enlever n'est pas trop élevé.

S'il exite un ordre des colonnes de $\mathcal{M}$ qui vérifie la propriété des uns consécutifs, il y a souvent un très grand nombre de solutions. Toutes les solutions peuvent être représentées dans une structure combinatoire, le PQ-arbre, introduit dans [3]. Le PQ-arbre est un arbre à deux types de noeuds, les noeuds de type Q, dont les descendants doivent être ordonnés selon une permutation donnée ou son inverse, et les noeuds de type P, dans lesquels les descendants peuvent être ordonnés de façon quelconque. Toutes les solutions peuvent être obtenues en choisissant un sens aux noeuds Q et une permutation pour les noeuds P. Par exemple, pour notre reconstruction de l'ancêtre des amniotes, plus de 10^{10} solutions sont représentées dans la figure 4.

4 Segments chromosomiques d'un génome ancestral des amniotes

Appliquée aux données de mammifères, oiseaux et poissons afin de trouver des synténies dans le génome proto-amniote, cette méthode reconstruit 34 segments chromosomiques après optimisation (toutes les solutions ont le même nombre de segments, et varient dans l'ordonnancement des marqueurs à l'intérieur des chromosomes). Ici, l'algorithme brancher/élaguer permet de trouver une solution optimale en très peu de temps (quelques secondes), et en enlevant très peu de syntons candidats. Ceci tend à prouver que la méthode de détection des syntons est suffisamment spécifique. Sa sensibilité n'est sans doute pas suffisante pour que nous puissions affirmer que les segments trouvés sont les chromosomes ancestraux, et annoncer un nombre de chromosomes pour le proto-amniote, d'autant plus que des données manquent sur les petits chromosomes du poulet, et que nous éliminons les plus petits syntons.

Ces morceaux de proto-chromosomes amniotes sont illustrés dans la figure 4, avec leur correspondance dans le génome du poulet, parce qu'il présente plus de similitudes avec le génome de poulet qu'avec un génome de mammifère.

Nous pouvons observer plusieurs synténies et les comparer avec d'autres méthodes : à notre connaissance, deux études antérieures [14,18] ont abouti à des propositions de proto-génomes d'am-

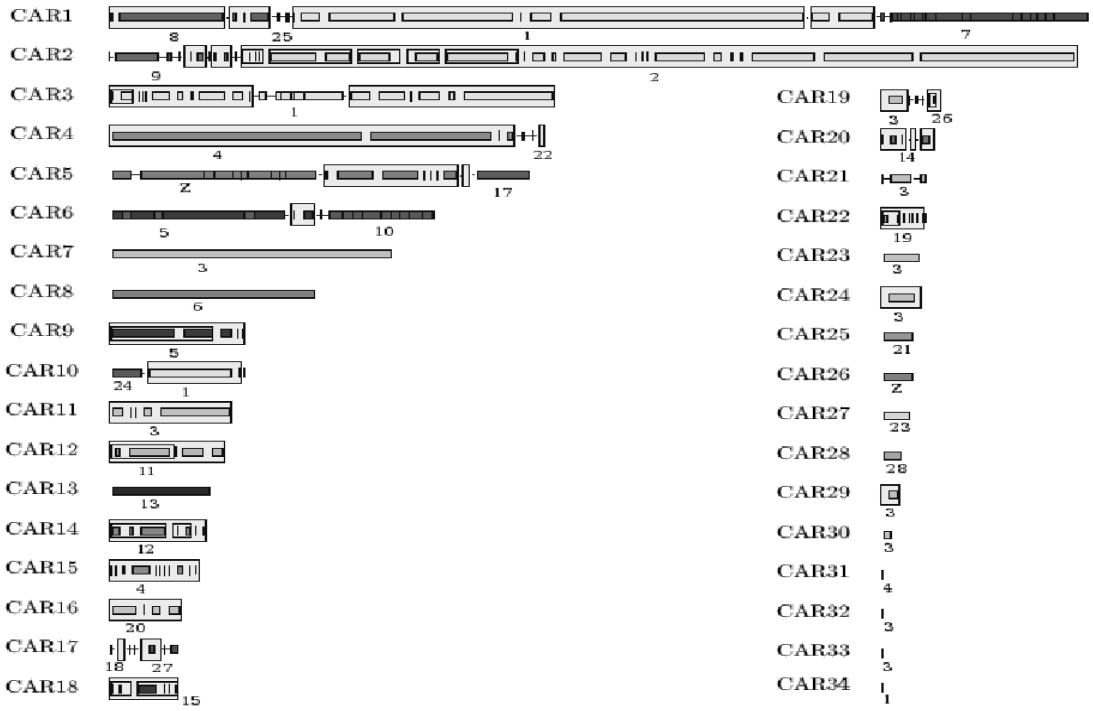

Fig. 2. Le PQ-arbre des segments proto-amniotes, avec les correspondances sur le génome du poulet. Chaque segment est un descendant direct de la racine du PQ-arbre. Un noeud linéaire est représenté par la suite de ses descendants reliés par des traits horizontaux, et un noeud premier est représenté par un cadre, à l'intérieur duquel sont ses descendants, dans un ordre quelconque. Tous les ancêtres possibles sont donc représentés ici, avec quelques parties chromosomiques où l'ordre des marqueurs n'est pas fixé.

niotes. Ces propositions sont très divergentes : non seulement les nombres de chromosomes varient entre 18 [14] et 26 [18], mais les synténies retrouvées ne sont pas toujours compatibles : sur 13 associations synténiques entre segments de chromosomes différents de poulets retrouvées par [14] et 6 retrouvées par [18], seulement deux sont communes. La raison est la divergence des méthodologies, et l'absence d'un cadre formel bien décrit, que nous espérons combler en partie ici. Nous donnons un résumé des différences, en incluant la comparaison avec nos résultats, dans le tableau 1. Nous retrouvons bien les deux synténies communes aux deux méthodes précédentes, plus 4 associations trouvées dans [14] et une dans [18]. Trois associations synténiques sont propres à notre étude.

Méthode Kohn *et al*	Méthode Nakatani *et al.*	Notre méthode
2-9-16, 1-24, 3-14, 5-10,	2-9, 1-14-18	2-9, 1-24, 5-10
17-Z, 4-22, 8-18, 18-27-19,	13-17-Z,	17-Z, 4-22, 18-27
21-23-26-32	1-7	8-25-1-7, 3-26

Tableau 1. Synténies retrouvées entre morceaux de chromosomes aviaires, pour trois méthodes différentes. Les numéros se rapportent à un morceau de chromosome, porté par le chromosome qui porte ce numéro. Les autres chromosomes prédits ne contiennent que des morceaux issus d'un même chromosome de poulet.

5 Données

Nous avons utilisé comme marqueurs les gènes orthologues entre génomes humain (hg18), macaque (rheMac2), souris (mm9), rat (rn4), chien (canFam2), vache (bosTau3), opossum (monDom5),

poulet (galGal4), tetraodon (TETRAODON 7), medaka (HdrR) et poisson zèbre (Zv7), tous fournis par la base Compara de Ensembl [11].

Acknowledgements

Eric Tannier est financé par l'Agence Nationale de la Recherche (GIP ANR JC05_49162 et NT05-3_45205) et par le Centre National de la Recherche Scientifique. Cedric Chauve est financé par un "NSERC Discovery Grant" et un "SFU Startup Grant".

Références

[1] F. Alizadeh, R.M. Karp, D.K. Weisser et G. Zweig, Physical mapping of chromosomes using unique probes, *Journal of Computational Biology*, 2 :159-184, 1995.

[2] M.-P. Beal, A. Bergeron, S. Corteel et M. Raffinot, An Algorithmic View of Gene Teams, *Theoretical Computer Science*, 320 :395-418, 2004.

[3] K.S. Booth et G.S. Lueker, Testing for the Consecutive Ones Property, Interval Graphs, and Graph Planarity Using PQ-tree Algorithms, *Journal of Computer and System Science*, 13 :335-379, 1976.

[4] G. Bourque, G. Tesler et P.A. Pevzner, The convergence of cytogenetics and rearrangement-based models for ancestral genome reconstruction, *Genome Research*, 16 :311-313, 2006.

[5] G. Bourque et P.A. Pevzner, Genome-Scale Evolution : reconstructing Gene Orders in the ancestral Species, *Genome Research*, 12 :26-36, 2002.

[6] F. Boyer, A. Morgat, L. Labarre, J. Pothier et A. Viari, Syntons, metabolons and interactons : an exact graph-theoretical approach for exploring neighbourhood between genomic and functional data, *Bioinformatics* 21 :4209-4215, 2005.

[7] C. Chauve et E. Tannier, A methodological framework for the reconstruction of contiguous regions of ancestral genomes and its application to mammalian genomes, Rapport INRIA RR-6494, soumis.

[8] F.S. Dietrich *et al.* The *Ashbya gossypii* Genome as a Tool for Mapping the Ancient *Saccharomyces cerevisiae* Genome, *Science*, 304 :304-307, 2004.

[9] L. Froenicke *et al.*, Are molecular cytogenetics and bioinformatics suggesting diverging models of ancestral mammalian genomes ?, *Genome Research*, 16 :306-310, 2006.

[10] A.-T. Gai, M. Habib, C. Paul et M. Raffinot, Identifying Common Connected Components of Graphs. *Rapport* RR-LIRMM 03-016, LIRMM, Université de Montpellier 2, 2003.

[11] T. J. P. Hubbard *et al*, Ensembl 2007 *Nucleic Acids Res.* 35 :Database issue :D610-D617, 2007.

[12] O. Jaillon *et al*, Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype, *Nature*, 431(7011) :946-957, 2004.

[13] M. Kellis, B.W. Birren, E. S. Lander, Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*, *Nature*, 428 :617-624, 2004.

[14] M. Kohn, J. Högel, W. Vogel, P. Minich, H. Kehrer-Sawatzki, J.A. Graves et H. Hameister, Reconstruction of a 450-My-old ancestral vertebrate protokaryotype, *Trends Genet.* 22 :203-210, 2006.

[15] J. Ma *et al.*, Reconstructing contiguous regions of an ancestral genome, *Genome Research*, 16 :1557-1565, 2006.

[16] R.M. McConnell, A certifying algorithm for the consecutive-ones property, *Proceedings of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, 761-770, 2004.

[17] M. Muffato et H. Roest Crollius, Paleogenomics, or the recovery of lost genomes from the mist of times, *BioEssays*, 30 :122-134, 2008.

[18] Y. Nakatani, H. Takeda, Y. Kohara et S. Morishita, Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates, *Genome Research*, 17 :1254-1265, 2007.

[19] Y. Van de Peer, Computational approaches to unveiling ancient genome duplications, *Nature Reviews*, 5 :752-763, 2004.

[20] J. Wienberg, The evolution of eutherian chromosomes, *Current Opinion in Genetics and Development*, 14 :657-666, 2004.