# Comment la reconstruction de génomes ancestraux peut aider à l'assemblage de génomes actuels

Yoann Anselmetti [*1,2], Vincent Berry[3,4], Cedric Chauve[5], Annie Chateau[3,4], Éric Tannier[2,6], Sèverine Bérard[1,3,4]

Session génomique des populations mercredi 29 11h30 Amphi Mérieux

[1] Institut des Sciences de l'Évolution - Montpellier (ISEM) – CNRS : UMR5554, Institut de recherche pour le développement [IRD] : UMR226, Université Montpellier II - Sciences et techniques – Place E. Bataillon CC 064, F-34 095 Montpellier Cedex 05, France
[2] Laboratoire de Biométrie et Biologie Évolutive (LBBE) – CNRS : UMR5558, Université Claude Bernard - Lyon I (UCBL), INRIA – 43 boulevard du 11 Novembre 1918, F-69 622 Villeurbanne Cedex, France
[3] Institut de Biologie Computationnelle (IBC) – CNRS : UMR5506, Université Montpellier II - Sciences et techniques – 95 rue de la Galéra, F-34 095 Montpellier, France
[4] Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier (LIRMM) – Université Montpellier II - Sciences et techniques, CNRS : UMR5506 – CC 477, 161 rue Ada, F-34 095 Montpellier Cedex 5, France
[5] SFU Discrete Mathematics Group (SFU-DMG) – Dept. Mathematics, SFU 8888 University Drive Burnaby, BC, V5A 1S6, Canada
[6] INRIA Rhône-Alpes (INRIA Grenoble Rhône-Alpes) – INRIA – ZIRST 655 avenue de l'Europe, Montbonnot, F-38 334 Saint Ismier Cedex, France

## Introduction

L'avènement des *NGS* et l'accès à un nombre croissant de génomes ces dernières années dans les bases de données permettent d'entreprendre des études sur l'histoire évolutive complexe de la structure des génomes. Pour cela, la plupart des méthodes d'analyse nécessitent des génomes complètement assemblés. Cependant, les génomes disponibles dans les bases de données sont souvent incomplètement assemblés et la plupart d'entre eux restent à l'état de *permanent draft genomes* [1]. Pour illustration, dans Ensembl (version 84) les génomes sont en moyenne constitués de 2 882 fragments, appelés contigs (écart-type=2977) et 58 des 69 espèces présentes dans la base de données ont leur génome composé de plus de 100 *contigs*. L'assemblage de génomes est un problème difficile qui consiste à ordonner et orienter des fragments d'ADN issus du séquençage afin de reconstruire les chromosomes composant le génome.

Il existe des méthodes de génomique comparative adaptées à la problématique d'assemblage afin d'améliorer l'ordre et l'orientation de marqueurs génomiques dans les génomes actuels à l'aide d'un ou plusieurs génomes de référence [2-6]. Parmi les méthodes dites multireference-guided assembly, certaines comme treecat [4], RACA [5] et celle d'Aganezov et al. [6] utilisent également les liens de parentés avec la structure des génomes de référence pour pondérer les apports de ces derniers dans l'assemblage du génome fragmenté (phylogeny-guided assembly). La méthode présentée dans cet article, ARt-DeCo (*Assembly Recovery through DeCo*) [7], s'inscrit dans cette catégorie de méthodes. Elle permet de réduire la fragmentation des génomes actuels composant un jeu de données, en prédisant des adjacences entre gènes situés aux extrémités de leurs contigs à l'aide de la composition des autres génomes du jeu de données. ARt-DeCo s'appuie sur la phylogénie des marqueurs génomiques utilisés mais n'a aucune contrainte quant à l'unicité et l'universalité de ceux-ci.

*. Intervenant

Dans un premier temps, nous présenterons succinctement le principe d'ARt-DeCo. Ensuite nous exposerons des simulations permettant d'évaluer la capacité d'ARt-DeCo à retrouver les adjacences de gènes. Enfin, nous détaillerons des applications d'ARt-DeCo sur des données réelles.

## Principes d'ARt-DeCo

Pour simplifier cet exposé, considérons que les marqueurs génomiques utilisés sont des gènes. On définit deux gènes comme adjacents s'ils sont situés sur un même fragment génomique sans aucun autre gène entre eux. ARt-DeCo est basé sur l'algorithme DeCo (Detection of Coevolution) [8] qui permet la reconstruction de l'histoire évolutive des adjacences de gènes et donne ainsi accès à une estimation de la structure des génomes ancestraux.

La méthode prend en entrée un ensemble d'arbres phylogénétiques de gènes, les adjacences de gènes observées dans les génomes actuels ainsi que l'arbre phylogénétique des espèces. Elle applique un principe de parcimonie basé sur les coûts de cassure et de création d'adjacences pour calculer une histoire évolutive des adjacences de moindre coût. Une telle histoire de coût minimum est calculée par une méthode de programmation dynamique basée sur l'exploration des arbres phylogénétiques de gènes.

La méthode DeCo ne tient compte que des adjacences présentes dans les assemblages des bases de données. Or, on sait qu'une adjacence peut être réelle mais non observée situé à la jonction de deux contigs non ordonnés et orientés. ARt-DeCo a été conçu pour permettre d'inférer, les adjacences de génomes actuels non présentes dans les bases de données, en plus d'inférer les adjacences ancestrales (comme DeCo). Cette inférence prend en compte une probabilité pour deux gènes d'être adjacents en fonction du degré de fragmentation du génome auquel ils appartiennent. ARt-DeCo s'autorise à inférer l'existence de ces adjacences si elles contribuent à une histoire de coût minimum.

## Expérience et résultats biologiques

Nous présentons ici les résultats obtenus avec ARt-DeCo. Les deux premières expériences ont permis de valider l'approche générale. Ensuite, nous analysons de façon qualitative une adjacence prédite par ARt-DeCo, puis nous analysons les résultats préliminaires d'une variante de la méthode sur le jeu d'anophèles permettant l'apport de données de séquençage dans la reconstruction d'histoires évolutives.

### Simulations de fragmentation

Pour évaluer la capacité de la méthode à prédire des adjacences de gènes présentes dans les génomes mais non répertoriées dans les bases de données (adjacences qualifiées ici de « réelles »), nous avons aléatoirement occulté certaines adjacences et testé la capacité d'ARt-DeCo à les retrouver. Ce premier jeu de données est composé de 18 génomes d'anophèles récemment séquencés et assemblés [9], et de 11 534 arbres de gènes incluant 172 585 gènes disponibles sur la base de données VectorBase.

Les 18 espèces ont été fragmentées aléatoirement avec divers pourcentages d'adjacences occultées (0,1 %, 0,5 %, 1 %, 5 %, 10 %, 25 %, 50 % et 75 %) et chaque expérience répliquée 30 fois.

À partir des nouvelles adjacences proposées par ARt-DeCo, nous avons calculé le rappel et la précision de la méthode. Le rappel indique la proportion d'adjacences occultées qui ont été retrouvées. La précision correspond à la proportion d'adjacences correctes (i.e., occultées) parmi les adjacences prédites.

Les résultats (cf. Figure 1) montrent que la précision la plus faible constatée est de l'ordre de 80 %, obtenue pour le plus faible pourcentage d'adjacences occultées (0,1 %) et augmente graduellement jusqu'à atteindre un plateau avec une valeur fluctuant autour de 92 % à partir de 1

% d'adjacences occultées. Le rappel est de 69,75 % pour le plus faible pourcentage d'adjacences occultées (0,1%), et il décroît au fur et à mesure que plus d'adjacences sont occultées, jusqu'à atteindre 12,18 % pour le cas extrême où 75 % des adjacences sont occultées.

On observe donc que pour des génomes faiblement fragmentés (de 0,1 à 0,5 %), ARt-DeCo retrouve de l'ordre de 69 % des adjacences occultées mais infère également une proportion non négligeable d'adjacences non présentes dans les assemblages initiaux (jusqu'à _~20%). Pour des génomes plus fragmentés, ARt-DeCo retrouve une plus faible proportion d'adjacences mais avec une précision supérieure à 90 %.

Pour analyser plus finement l'effet d'ARt-DeCo, nous avons effectué la même expérience mais en simulant des cassures chez une seule espèce à la fois. Pour cela, trois espèces placées à des positions différentes dans la phylogénie des anophèles ont été choisies :

— *Anopheles gambiae* localisée en profondeur dans l'arbre,
— *Anopheles minimus* située en profondeur dans l'arbre mais avec peu d'espèces proches,
— *Anopheles albimanus* en position d'outgroup par rapport aux autres espèces de l'arbre.

Pour les trois espèces (cf. Figure 2), la précision et le rappel décroissent légèrement lorsque la fragmentation artificielle augmente :

— Pour *A. gambiae* :
  — Précision : Max : 100 % | Min : 87,37 %
  — Rappel : Max : 70 % | Min : 68 %
— Pour *A. minimus* :
  — Précision : Max : 99,43 % | Min : 95,54 %
  — Rappel : Max : 54 % | Min : 53,06 %
— Pour *A. albimanus* :
  — Précision : Max : 99,49 % | Min : 93,68 %
  — Rappel : Max : 38,89 % | Min : 37,37 %

On observe que pour les trois espèces, on obtient une bonne précision fluctuant entre 100 et 87,37 %. Le rappel plus élevé chez *A. gambiae* que chez les autres espèces peut s'expliquer par le voisinage d'espèces proches dans la phylogénie. Comme le faible rappel de *A. albimanus* peut être expliqué par sa position d'*outgroup* dans l'arbre des espèces.

En conclusion, on observe que la performance d'ARt-DeCo à lier des contigs n'est pas la même suivant la proportion de génomes fragmentés dans le jeu de données et le degré de fragmentation de ces génomes. Pour le cas, où l'ensemble des génomes est fragmenté (cf. Figure 1), les résultats montrent que ARt-DeCo obtient un meilleur compromis rappel/précision pour des génomes moyennement fragmentés et retrouve 333 vraies adjacences sur 340 adjacences prédites lorsque 493 sont occultées. Tandis que dans le cas où une seule espèce est fragmentée (cf. Figure 2), ARt-DeCo est plus performant pour de faibles fragmentations, chez *A. gambiae* pour 9 adjacences occultées, ARt-DeCo en prédit 7 toutes valides.

**Passage à l'échelle**

Pour déterminer la capacité de l'algorithme ARt-DeCo à travailler sur de grands jeux de données, nous l'avons appliqué aux 69 espèces eucaryotes de la base de données Ensembl (version 79). Ce jeu est composé de 20 279 arbres de gènes contenant 1 222 543 gènes codant pour des protéines et 1 023 492 adjacences chez les génomes actuels. Une grande proportion des génomes actuels sont fortement fragmentés, dont le génome du wallaby (*Macropus eugenii*) composé de 12 704 *contigs*. L'algorithme prédit 36 445 nouvelles adjacences sur l'ensemble des espèces du jeu de données en $\approx$ 18h sur un ordinateur de bureau.

L'analyse des résultats montre que plus les génomes sont fragmentés plus l'assemblage est amélioré, c'est-à-dire que le ratio du nombre d'adjacences prédites sur nombre d'adjacences manquantes est plus élevé dans ces génomes là (cf. [7]).

### Analyse détaillée d'une adjacence prédite

L'algorithme ARt-DeCo peut également être combiné avec l'algorithme DeClone [10] qui permet d'explorer l'ensemble des solutions parcimonieuses co-optimales d'ARt-DeCo et ainsi fournir un score à une adjacence. Ce score correspond à la proportion de scénarios dans lesquels l'adjacence a été prédite par ARt-DeCo et mesure ainsi un support de confiance compris entre 0 et 1.

Sur le jeu de données que nous considérons maintenant, qui contient 39 placentaires, ARt-DeCo prédit 22 675 nouvelles adjacences dont 95 % ont un support > 0,9.

Parmi elles, une nouvelle adjacence proposée chez le panda (*Ailuropoda melanoleuca*) a été analysée en détail. L'analyse chez les espèces proches du voisinage plus large des gènes impliqués dans cette adjacence a montré de fortes similarités avec la situation dans le génome du panda (cf. Figure 3). L'analyse du voisinage des gènes concernés conforte donc la forte confiance dans l'existence de l'adjacence prédite. Cette confiance est renforcée par l'inférence par ARt-DeCo d'autres adjacences homologues à l'adjacence du panda chez trois autres espèces (*Ochotona princeps*, *Tupaia belangeri* & *Dipodomys ordii*) toutes avec des supports > 0,99.

Une analyse systématique reste à mener pour évaluer l'ensemble des prédictions.

### Intégration des données de séquençage

Récemment, l'algorithme ARt-DeCo a été amélioré pour permettre l'intégration des liens de *scaffolding* dans son calcul. Ces liens de scaffolding prédits par des logiciels comme BESST [11], apportent des informations de structure non présentes dans les bases de données. Combinés aux adjacences connues, ils permettent une reconstruction d'histoires évolutives d'adjacences plus complètes augmentant ainsi le pouvoir prédictif d'ARt-DeCo.

Des données de séquençage *paired-end* et *mate-pair* sont disponibles pour le jeu de données des 18 anophèles. Les 34 542 liens de *scaffolding* calculés par BESST sur ce jeu de données ont permis à ARt-DeCo d'inférer 5 894 nouvelles adjacences représentant ≈25 % des prédictions.

## Conclusion

ARt-DeCo est une méthode de prédiction d'adjacences de gènes basée sur la phylogénie qui s'attache à réduire la fragmentation des assemblages de génomes actuels. Intuitivement, ces génomes peuvent se corriger mutuellement si les zones génomiques fragmentées sont différentes d'un génome à l'autre. ARt-DeCo s'appuie sur un ensemble de marqueurs pour lequel les histoires évolutives sont connues. Il est capable de gérer de larges jeux de données, même si ceux-ci contiennent des gènes dupliqués. ARt-DeCo est un pas en avant dans la réduction de la fragmentation des génomes séquencés. Les simulations et l'analyse de cas réels montrent que les adjacences proposées sont fiables.

## Références

[1] GOLD database : `https://gold.jgi.doe.gov/statistics/`

[2] Lu et al., (2014). CAR : contig assembly of prokaryotic draft genomes using rearrangements. *BMC Bioinformatics*, 15:381–390.

[3] Bosi et al., (2015). MEDUSA : a multi-draft based scaffolder. *Bioinformatics*, 31(15):2443–51.

[4] Husemann & Stoye (2010). Phylogenetic comparative assembly. *Algorithms for Molecular Biology*, 5(1):3–14.

[5] Kim et al., (2013). Reference-assisted chromosome assembly. *Proceedings of the National Academy of Sciences (PNAS)*, 110(5):1785–90.

[6] Aganezov et al., (2015). Scaffold assembly based on genome rearrangement analysis. *Computational Biology and Chemistry*, 57(August):46–53.

[7] Anselmetti et al. (2015) Ancestral gene synteny reconstruction improves extant species scaffolding, *BMC genomics*. 16(Suppl 10):S11.

[8] Bérard et al., (2012) Evolution of gene neighborhoods within reconciled phylogenies, *Bioinformatics*. 28:i382–i388.

[9] Neafsey et al., (2015). Highly evolvable malaria vectors: The genomes of 16 Anopheles mosquitoes. *Science*, 347(6217):1258522–1 – 1258522–8.

[10] Chauve et al.(2014). Evolution of genes neighborhood within reconciled phylogenies: an ensemble approach. In *Lecture Notes in Computer Science (Advances in Bioinformatics and Computational Biology)* (Vol. 8826 LNBI, pp. 49–56).

[11] Sahlin et al., (2014). BESST - Efficient scaffolding of large fragmented assemblies. *BMC Bioinformatics*, 15(1):281

**Mots clefs :** reconstruction de génomes ancestraux, assemblage de génomes, adjacences de gènes, évolution, parcimonie, programmation dynamique