Algorithmique pour l'évolution des interactions géniques

7 novembre 2013

Directeur (30 %)	Vincent Berry (Professeur, LIRMM, UM2)
Co-encadrant (50 %)	Sèverine Bérard (Maître de conférences, ISEM, UM2)
Co-encadrant (20 %)	Annie Chateau (Maître de conférences, LIRMM, UM2)

Mots-clés Algorithmique, graphe, combinatoire, programmation dynamique, arbres phylogénétiques, évolution.

Contexte La connaissance de l'histoire évolutive des êtres vivants est une des clefs de la compréhension du fonctionnement des organismes que nous connaissons aujourd'hui. Ce sujet de thèse aborde la reconstruction de cette histoire à partir des espèces actuelles, sur la base des gènes qui composent leur génome.

Les espèces évoluent suivant un processus arboré – appelé arbre phylogénétique : une unique espèce ancestrale (racine de l'arbre), a donné naissance à deux espèces (noeuds internes), qui se sont à leur tour divisées, et ainsi de suite jusqu'à aboutir aux espèces actuelles (feuilles de l'arbre). Les gènes portés par les espèces évoluent eux aussi au cours du temps, selon des histoires qui leur sont parfois propres : naissance, mort, duplication ou transfert d'une espèce à une autre. Si on réduit un génome à un ensemble de gènes évoluant indépendamment les uns des autres, on connaît des méthodes (modèles, algorithmes) efficaces pour proposer une histoire évolutive possible de ces gènes [Doyon *et al*, 2010]. Mais c'est négliger que les gènes interagissent dans l'organisme et échangent de l'information.

Ces relations/interactions entre gènes évoluent elles aussi, mais les méthodes qui retracent leur évolution ne sont pas aussi développées que celles retraçant l'évolution des gènes. Il y a une théorie algorithmique à construire pour en rendre compte, où l'évolution des gènes, regroupés en familles, est modélisée par des *arbres*, et les interactions entre gènes par des *graphes*.

Un algorithme permettant de retracer l'histoire évolutive d'une relation entre deux familles de gènes a été développé par une encadrante de cette thèse [Bérard *et al*, 2012]. Ce travail a donné lieu à un logiciel, `DeCo` [`DeCo`, 2012], qui permet d'avoir une estimation des positions relatives des gènes dans les génomes d'espèces disparues à partir des génomes contemporains. Il a été appliqué sur des jeux de donnée de mammifères ($\sim$ 5000 arbres, $\sim$ 107000 gènes) et de plantes ($\sim$ 50000 arbres, $\sim$ 615000 gènes). C'est cependant un travail préliminaire ne prenant en compte qu'un petit sous-ensemble des évènements évolutifs possibles. En particulier, la non prise en compte des transferts empêche l'utilisation de cette méthode pour l'étude des bactéries, pour lesquelles nous disposons de grands jeux de données ($\sim$ 46000 arbres) constitués lors de précédents travaux [Penel et al., 2009, `Phylariane`, 2009-2012].

Plan de la thèse La thèse pourra commencer en se basant sur l'algorithme `DeCo` que l'on complétera pour intégrer à la fonction objectif des évènements importants tels que les duplications segmentales et les transferts. Ce problème n'est pas trivial puisque l'intégration de

nouveaux évènements risque de faire disparaître la propriété d'indépendance entre les évènements permettant l'utilisation des techniques de programmation dynamique (complexité polynomiale). Il est néanmoins important d'ajouter ces évènements dans la fonction objectif, car ils pourront permettre de comprendre l'apparition de nouvelles fonctions biologiques, de pathogénicité ou de résistance aux antibiotiques (pour les bactéries), phénomènes souvent portés par plusieurs gènes et liés aux duplications et aux transferts. Un des enjeux est aussi de conserver une complexité basse pour permettre l'exécution de la méthode sur de gros jeux de données, comme celui des plantes. Une première approche, dévelloppé dans une extension de `DeCo`, `DeCoLT` [Patterson et al., 2013], permet d'intégrer les transferts dans une certaine mesure. C'est un point de départ possible pour la prise en compte des autres événements évolutifs importants à considérer.

Un des points à améliorer dans la méthode `DeCo` concerne le respect de la contrainte de linéarité du génome. Dans sa version actuelle, `DeCo` autorise en effet un gène à être adjacent à plus de 2 autres gènes, ce qui est physiquement impossible (un génome étant une séquence linéaire). Ce choix avait été fait pour des raisons de temps de calcul et de complexité algorithmique, mais le respect de la contrainte de linéarité est sans aucun doute une des pistes les plus intéressantes pour l'amélioration des prédictions de la méthode.

On pourra aussi converger vers la linéarité des génomes ancestraux par le biais d'aller-retours entre génomes reconstruits et arbres de gènes. Il faut pour cela envisager une mesure de linéarité de ces génomes fondée sur les relations d'adjacences entre couples de gènes. Possiblement cette problématique de linéarité des génomes ancestraux amènera à considérer l'histoire évolutive conjointe non plus de deux familles de gènes mais d'un très grand nombre de familles de gènes.

Ce sujet de thèse est proposé dans le cadre de l'ANR Ancestrome (LBBE, ISEM), qui pourra financer les missions et équipements de l'étudiant(e).

Références

[Bérard *et al*, 2012] Sèverine Bérard, Coralie Gallien, Bastien Boussau, Gergely J. Szöllősi, Vincent Daubin et Eric Tannier. **Evolution of gene neighborhoods within reconciled phylogenies**. Bioinformatics, 28 (18) : i382-i388.

[`DeCo`, 2012] Logiciel `DeCo` programmé d'après l'algorithme de [Bérard *et al*, 2012]. Auteur : Sèverine Bérard. Disponible à l'adresse http://pbil.univ-lyon1.fr/software/DeCo/

[Doyon *et al*, 2010] Jean-Philippe Doyon, Celine Scornavacca, K. Yu. Gorbunov, Gergely J. Szöllősi, Vincent Ranwez et Vincent Berry. **An Efficient Algorithm for Gene/Species Trees Parsimonious Reconciliation with Losses, Duplications and Transfers**. RECOMB-CG p93-108.

[Patterson et al., 2013] Murray Patterson, Gergely J. Szöllősi, Vincent Daubin et Eric Tannier. **Lateral gene transfer, rearrangement, reconciliation**, BMC Bioinformatics, 14 (Suppl 15) :S4.

[Penel et al., 2009] Penel S, Arigon AM, Dufayard JF, Sertier AS, Daubin V, Duret L, Gouy M et Perrière G, **Databases of homologous gene families for comparative genomics**, BMC Bioinformatics, 10 (Suppl 6) :S3, http://pbil.univ-lyon1.fr/databases/hogenom

[`Phylariane`, 2009-2012] Projet ANR impliquant l'équipe MAB (LIRMM) et le LBBE (Lyon).

[Romiguier *et al*, 2010] J. Romiguier, V. Ranwez, E.J.P. Douzery et N. Galtier. **Contrasting GC-content dynamics across 33 mammalian genomes : Relationship with life-history traits and chromosome sizes**. Genome Research (2010) 20 : 1001-1009