Sujet de thèse : *Décomposition de graphes pour la reconstruction de réseaux phylogénétiques*

Pour représenter l'Évolution, le modèle classique est l'arbre phylogénétique, dont les feuilles représentent des espèces ou des gènes, et dont les noeuds internes correspondent aux ancêtres communs. L'étude du mécanisme de transfert latéral de gènes, mis en évidence à la fin des années 80 [1], a remis en cause l'idée d'Arbre de la Vie, privilégiant un modèle d'arbre augmenté de réticulations (correspondant aux échanges de gènes) ou plus généralement de réseau [2].

D'autres événements biologiques sont à l'origine de réticulations : les recombinaisons [3], qui correspondent à des échanges de matériel génétique au sein d'un même génome, très fréquents par exemple pour le virus VIH [4], et les hybridations, qui peuvent avoir lieu de façon naturelle en donnant des descendants fertiles, en particulier entre certaines variétés de végétaux. On remarque donc l'utilisation des réseaux phylogénétiques à la fois dans la construction d'une phylogénie des espèces [5], que la construction d'une phylogénie des gènes [6].

D'autre part, les réseaux phylogénétiques servent aussi de moyen de visualisation efficace pour mettre en relief des conflits entre arbres phylogénétiques [7].

Les approches bioinformatiques pour étudier ces problèmes sont donc généralement spécialisées, adaptées au phénomène évolutif étudié. Divers types de réseaux phylogénétiques ont été introduits, parmi lesquels les "splits graphs" [8], les réseaux réticulés [9], les "galled trees" [10], et les phylogénies hybrides [11], etc. L'étude de ces réseaux phylogénétiques fait l'objet d'une forte attention des bioinformaticiens, comme le montre la forte présence du thème aux deux conférences internationales majeures en bioinformatique WABI (deux sessions consacrées aux réseaux phylogénétiques à WABI06) et RECOMB (une session à RECOMB05).

Il existe plusieurs logiciels dédiés à la construction de réseaux phylogénétiques. Le plus utilisé est SplitsTree [12], cité dans plus de 300 articles de biologie et bioinformatique. Il permet la construction des diverses variantes de réseaux phylogénétiques à partir de données d'entrée variées, et en utilisant des plugins implémentant diverses méthodes de construction d'arbres ou réseaux phylogénétiques. Il devrait prochainement être disponible sous licence libre.

Nous nous proposons donc tout d'abord d'effectuer une synthèse des modèles existants, qui sont très spécifiques. Les "splits graphs" servent plutôt à visualiser des conflits entre plusieurs arbres phylogénétiques. Les "galled trees" sont une restriction des réseaux phylogénétiques où les cycles induits par les recombinaisons n'ont pas de noeud en commun, et tels que la construction en minimisant le nombre de recombinaisons est polynomiale. Les phylogénies hybrides sont utilisées dans le contexte de la minimisation du nombre d'hybridations. Il manque

une unification de ces travaux, la terminologie des diverses variantes de réseaux phylogénétiques n'ayant été précisée et fixée qu'en 2006 [12].

Cette synthèse permettra aussi d'extraire des propriétés communes à ces réseaux. Pour cela, nous utiliserons des méthodes classiques de décomposition de graphes (décompositon arborescentes, décomposition en branches, décomposition en coupe...) dans un but de simplification des réseaux. Nous relierons les propriétés dégagées à des invariants, paramètres de graphes classiques tels que par exemple la largeur arborescente [13]. En effet, de nombreux graphes ou réseaux issus de données réelles se sont avérés avoir une faible largeur arborescente. Nous partirons donc de données biologiques pour mesurer certains paramètres de ces réseaux et les mettre en relation avec des classes de graphes connues. Par exemple, les graphes triangulés sont connus pour jouer un rôle important pour le problème de la phylogénie parfaite [14].

Ceci nous permettra de mettre au point des algorithmes de complexité paramétrée (classe FPT) en identifiant dans les données du problème un paramètre borné. Le but est d'obtenir des algorithmes efficaces en pratique pour des problèmes dont la complexité théorique est importante dans le cas général.

Les algorithmes obtenus seront implémentés dans le logiciel SplitsTree. Les applications se situeront à la fois dans le domaine médical (inférences d'haplotypes impliqués dans des maladies génétiques [15]) et dans le domaine de l'Évolution (analyses de la diversité phylogénétique au sein d'une unité géographique ou d'un domaine du Vivant, afin de prévenir par exemple des risques d'extinction d'espèces, ou des risques épidémiologiques [16]).

Christophe Paul est chargé de recherche CNRS au LIRMM dans l'équipe Visualisation et Algorithmes de Graphes, et travaille sur l'algorithmique des graphes en général, et les décompositions de graphes en particulier, et leurs applications aux réseaux ou à la bioinformatique. Il est responsable du projet blanc ANR GRAAL (2006-2009) sur les décompositions de graphes. Vincent Berry est maître de conférences à l'Université de Montpellier II et ses activités de recherche dans l'équipe Méthodes et Algorithmes en Bioinformatique sont consacrées à la phylogénie. Tous deux ont déjà travaillé ensemble dans ce domaine [17].

Philippe Gambette est déjà familier des deux thèmes de cette thèse. Il a fait son stage de licence dans l'équipe Méthodes et Algorithmes pour la Bioinformatique du LIRMM en 2003 sous la direction d'Olivier Gascuel et Denis Bertrand, puis son stage de master 1 à l'université de Tübingen avec le professeur Daniel Huson, sur l'amélioration de la représentation des réseaux phylogénétiques. L'algorithme proposé a fait l'objet d'une publication d'un article de revue [18] et il l'a implémenté dans le logiciel SplitsTree. Il a réalisé son stage de master 2 dans le domaine de la théorie des graphes avec Michel Habib au LIAFA pour se familiariser avec les décompositions de graphes [19].

Références

[1] Johann Peter Gogarten, Henrik Kibak, Peter Dittrich, Lincoln Taiz, Emma Jean Bowman, Barry J. Bowman, Morris F. Manolson, Ronald J. Poole, Takayasu Date, Tairo Oshima, Jin Konishi, Kimitoshi Denda et Masasuke Yoshida : *Evolution of the Vacuolar H^+-ATPase : Implications for the Ori-*

gin of Eukaryotes, Proceedings of the National Academy of Sciences 86(17) p.6661-6665, 1989.

[2] W.F. Doolittle : *Phylogenetic classification and the universal tree*, Science 284(5423) p.2124-2129, 1999.

[3] Lusheng Wang, Kaizhong Zhang et Louxin Zhang : *Perfect Phylogenetic Networks with Recombination*, Journal of Computational Biology 8(1) p.69-78, 2001.

[4] David L. Robertson, Paul M. Sharp, Francine E. McCutchan et Beatrice H. Hahn : *Recombination in HIV-1*, Nature 374 p.124-126, 2002.

[5] Luay Nakhleh, Tandy Warnow et C.R. Linder : *Reconstructing reticulate evolution in species : theory and practice*, dans P.E. Bourne et Dan Gusfield, Research in Computational Biology p.337-346, 2004.

[6] Saara Finnilä, Mervi S. Lehtonen et Kari Majamaa : *Phylogenetic Network for European mtDNA*, The American Journal of Human Genetics 68 p.1475-1484, 2001.

[7] David Bryant et Vincent Moulton : *Neighbor-Net : An Agglomerative Method for the Construction of Phylogenetic Networks*, Molecular Biology and Evolution 21(2) p.255-265, 2004.

[8] Hans-Jürgen Bandelt et Andreas Dress : *Split decomposition : a new and useful approach to phylogenetic analysis of distance data*, Molecular Phylogenetics and Evolution 1 p.242-252, 1992.

[9] Vladimir Makarenkov et Pierre Legendre : *From a phylogenetic tree to a reticulated network*, Journal of Computational Biology 11(1) p.195-212, 2004.

[10] Dan Gusfield : *From a phylogenetic tree to a reticulated network*, Journal of Computer and System Sciences, Special issue on bioinformatics II p.381-398, 2005.

[11] Mihaela Baroni, Charles Semple et Mike Steel : *A Framework for Representing Reticulate Evolution*, Annals of Combinatorics 8(4) p.391-408, 2005.

[12] David Bryant et Daniel H. Huson : *Application of Phylogenetic Networks in Evolutionary Studies*, Molecular Biology and Evolution 23(2) p.256-264, 2006, `www.splitstree.org`.

[13] Neil Robertson et Paul D. Seymour : *Graph minors II : algorithmic aspects of tree-width*, Journal of Algorithms 7 p.309-322, 1986.

[14] Mike Steel : *The complexity of reconstruction trees from qualitative characters and subtrees*, Journal of Classifications 9 p.91-116, 1992.

[15] Yinglei Song, Chunmei Liu, Russell L. Malmberg et Liming Cai : *Phylogenetic network inferences through efficient haplotyping*, 6th Workshop on Algorithms in BioInformatics (WABI), 2006.

[16] Alexei Drummond, *Modeling virus evolution and population dynamics*, DIMACS Working Group on Phylogenetic Trees and Rapidly Evolving Pathogens, 2004.

[17] Vincent Berry, Sylvain Guillemot, François Nicolas, Christophe Paul : *On the Approximation of Computing Evolutionary Trees*, 11th Annual International Conference on Computing and Combinatorics (COCOON 2005), LNCS 3595 p.115-125, 2005.

[18] Philippe Gambette et Daniel H. Huson : *Improved layout of phylogenetic networks*, à paraître dans TCBB.

[19] Philippe Gambette : *Les graphes 2-intervallaires*, Mémoire de master MPRI, Université Paris 7, 2006.