

TITRE DU PROJET : Filtres pour la recherche de similarité dans des structures arborescentes :

Application aux ARN

Laboratoire Demandeur : Laboratoire Bordelais de Recherche en Informatique

Numéro : UMR 5800

Directeur : Maylis Delest

Téléphone : 05 40 00 60 82 Mél. : delest@labri.fr

Responsable du Projet (coordonnées du chef de projet) : Pascal Ferraro

Téléphone : 05 40 00 35 07 Mél. : ferraro@labri.fr

Durée du Projet : 3 ans

Période d'exécution budgétaire prévisionnelle : 2011-2014

Résumé du Projet

Les ARN sont des molécules formées par une suite de nucléotides et présentes dans tout être vivant. La comparaison de structures secondaires d'ARN est un des outils indispensables à l'étude de ces molécules. Parmi les modèles déjà proposés pour modéliser ces structures, citons les séquences annotées, les 2-intervalles et les arbres. Les algorithmes permettant la comparaison de ces structures sont essentiellement basés sur leur édition et sont généralement polynomiaux. Hélas, le degré élevé du polynôme et la taille grandissante des bases de données ne permettent pas d'utiliser de manière réaliste ces méthodes pour la recherche d'une requête dans une base de données. Dans ce contexte, le travail proposé consiste à mettre en place des filtres rapides permettant de sélectionner les structures de la base potentiellement proche de la requête. En suivant les travaux réalisés en algorithmique des séquences, on pourra envisager des filtres sans perte ou bien avec perte. La mise en œuvre de tels filtres posera assurément des problèmes d'indexation de structures arborescentes ordonnées (voire celles de séquences annotées ou autres).

Contexte Biologique

Depuis quelques années, le rôle des ARN non-codants dans les différents processus cellulaires a été considérablement réévalué. On sait maintenant que de petits ARN non codants (comme les petits ARN nucléaires, les petits ARN nucléolaires, les micro-ARN) interviennent à tous différents niveaux de la régulation et de l'expression des gènes, par exemple. Il y a donc un besoin crucial d'outils informatiques pour aider à la manipulation et à la modélisation des ARN. Le but de ce projet est de fournir des outils à la fois efficaces et pertinents d'un point de vue biologique pour le problème essentiel de la comparaison d'ARN dans le cadre de grandes bases de données. Ces travaux devraient mener à la proposition d'outils standard qui ont vocation à devenir une référence dans la communauté.

Le projet repose sur une collaboration étroite entre trois équipes multidisciplinaires qui ont une longue expérience de l'algorithmique et de la biologie de l'ARN : l'équipe MaBioVis du LaBRI, le Centre de BioInformatique de Bordeaux – CBiB et le département de mathématique de l'Université Simon Fraser à Vancouver. Une co-tutelle de thèse pourrait être de plus envisagée avec ce dernier partenaire.

Objectifs du projet

La manipulation des grandes bases de données textuelles que constituent les dictionnaires, les génomes, ou les textes circulant sur le World Wide Web fait appel à des structures de données élaborées et à des algorithmes efficaces. Dans le cas des génomes, de nombreux outils appelés « filtres » ont été mis au point (BLAST, FASTA, QUASAR, SWIFT, NIMBUS; ...), leurs objectifs étant d'identifier rapidement les éléments de la base potentiellement proche de la requête.