TITRE: ADN, Algorithms for Distributed biological Networks

EQUIPE/THEME: MabioVis / Modélisation et Comparaison de séquences et structures biologiques ou Génomique comparative, Modélisation, Analyse de données biologiques

DIRECTEURS: Macha Nikolski ou Pascal Durrens

COURRIELS: macha@labri.fr <mailto:macha@labri.fr>,pascal.durrens@labri.fr <mailto:pascal.durrens@labri.fr>

MOTS-CLES: Calcul distribué sur cloud, Fouille de données et classification, Inférence de réseaux biomoléculaires, Génomique comparée, Comparaison de réseaux métaboliques

DIRECTEURS HABILITES: Macha Nikolski ou Pascal Durrens

DESCRIPTION du SUJET:
Le projet doctoral proposé s'intègre dans le contexte du projet ADN fédérant l'ensemble des chercheurs de l'équipe MabioVis. Dans ce cadre, deux axes ont été identifiés pouvant chacun conduire à un travail de thèse.
A) Analyse comparée de réseaux métaboliques : L'analyse bioinformatique du métabolisme d'espèces nouvellement séquencées requiert d'appréhender leurs différences sur un plan fonctionnel, en s'appuyant sur l'analyse comparée de leurs réseaux métaboliques codés sous forme de graphes. La comparaison de réseaux présente plusieurs applications allant de l'identification de fonctions qui leur sont communes/spécifiques à la compréhension de différences essentielles dans le phénotype. Cette comparaison peut se faire autant pour les données génomiques que métagénomiques. Nous envisageons plusieurs approches multi-niveaux. (1)Une comparaison locale où l'on est intéressé par répondre à des questions sur des gênes spécifiques. En effet, ici on s'intéressera au développement d'un modèle statistique permettant d'évaluer la présence ou absence d'une famille de protéines à partir de lectures obtenues par NGS. Ce modèle devra palier à des défauts des travaux précédents qui ignorent les biais de séquençage et de représentativité des lectures (Sun et al., 2007; Sharon et al., 2009). (2) Une comparaison plus globale au niveau fonctionnel s'intéressera à comparer les réseaux métaboliques. Ici on essayera de s'affranchir de la limitation ensembliste présente dans les travaux actuels.(Overbeek, Begley et al. 2005; Dinsdale, Edwards et al. 2008; Markowitz, Szeto et al. 2008). On développera d'une part de nouveaux algorithmes basés sur le principe l'édition, de détection de motifs ainsi que de nouvelles signatures pour les réseaux métaboliques. Ce dernier volet requiert à gérer des masses conséquentes de données et les approches modernes tournent vers les environnements distribués (grid, cloud). Par exemple, MetaShark (Pinney et al., 2005) identifie des enzymes dans les données métagénomiques en multipliant les exécutions sur une grille de calcul. On s'intéressera donc à adapter des méthodes existantes ou en développer de nouvelles dans ce cadre de calcul distribué.
B) Formalisation de l'analyse multidimensionnelle appliquée à des donnéesgénomiques: Les masses de données produites par les travaux de génomique comparative représentent une très grande variété d'entités biologiques et de relations entre eux mises en évidence ou utilisées au cours des différentes analyses bioinformatiques. Ces relations n-aires sont multiples: voisinage, recouvrement, inclusion, synténie, différence, provenance, composition, similarité ? et font l'objet de nombreuses ontologies qui

règlement leur association et en donne une sémantique. Les données manipulées représentent un volume très important (un génome de 15 Megabases peut générer jusqu'à 100 Gigaoctets de données) et en croissance superlinéaire grâce aux avancées technologiques. Cette masse de relations n-aires contient implicitement des relations recherchées par les utilisateurs biologistes, qui peuvent être obtenues par projection selon une dimension d'intérêt : par exemple, l'orthologie est la conjonction d'une relation d'homologie et d'une relation de synténie, une famille de protéines phylogénétique peut être faite par agglomération consensuelle de regroupements de relations de similarité et de structure en motifs, etc ? Dans la fouille de données, plusieurs méthodes sont destinées à analyser des données décrites dans les espaces multidimensionnels. Ces dernières s'adaptent naturellement aux problèmes que nous avons abordés en développant des algorithmes ad hoc (clustering de protéines, inférence de fusions/fissions de gènes, extrapolation de réseaux métaboliques). Un premier travail consiste à revoir les solutions algorithmiques que nous avons proposées à la lumière de l'état de l'art en analyse multidimensionnelle. Parmi les approches qui ont été proposées ces dernières années, notons les concepts de cube de données (Gray et al 1997) et de skyline (Börzsönyi et al 2001). Récemment, de nouvelles propositions ont plaidé pour leur combinaison (Yiu et al 2011, Raïssi et al 2010); cette combinaison ne se fait pas de manière automatique et nécessite de nouvelles études. Il s'agira donc de réaliser :

1.Une partie analyse théorique, qui dépend de la classe de requêtes que l'équipe se pose et qui caractérisera ces classes en utilisant un formalisme adapté.
2.Une partie expérimentale, qui propose un modèle physique de stockage qui permet d'optimiser les requêtes identifiées ci-dessus, notamment par le biais de modèles NoSQL.