Requêtes alternatives dans le contexte d'un entrepôt de données génomiques

Christine Froidevaux*, Frédéric Lemoine*

*LRI, CNRS UMR 8623, Université Paris Sud 11, France
{chris,lemoine}@lri.fr,

Résumé. Afin d'aider les biologistes à annoter des génomes, ce qui nécessite l'analyse, le croisement, et la comparaison de données provenant de sources diverses, nous avons conçu un entrepôt de données de génomique microbienne. Nous présentons la structure globale flexible de l'entrepôt et son architecture multi-niveaux et définissons des correspondances entre ces niveaux. Nous introduisons ensuite la notion de requête alternative et montrons comment le système peut construire l'ensemble des requêtes alternatives à une requête initiale. Pour cela, nous indroduisons un mécanisme d'interrogation qui repose sur l'architecture multi-niveaux, et donnons un algorithme de calcul des requêtes alternatives.

1 Introduction

Avec l'entrée dans l'ère post-génomique, l'avancée du séquençage de génomes et l'utilisation de plus en plus massive d'expériences à haut débit produisent une quantité gigantesque de données biologiques. La conception de systèmes de gestion de données pour stocker et interroger cette information devient cruciale, en particulier dans le domaine de l'annotation fonctionnelle des génomes, qui consiste en l'attribution d'une fonction biologique aux produits de chaque gène. Cette tâche est indispensable pour savoir quels gènes sont impliqués dans certains processus (e.g la pathogénicité pour les génomes microbiens).

C'est dans ce contexte que nous avons conçu l'entrepôt de données génomiques Microbiogenomics[1], dont l'objectif est de rassembler des données de génomique microbienne, pour l'annotation fonctionnelle (ou la ré-annotation) de génomes microbiens (Lemoine et al., 2007). Pour réaliser cette tâche d'annotation, les biologistes ont besoin d'une grande variété de données (telles que des données fonctionnelles, d'homologie, de voies métaboliques, etc.) qui se trouvent dans diverses sources de données dispersées sur le web. Leur travail consiste à naviguer dans les sources de données, trouver des gènes / protéines homologues à leurs gènes / protéines d'étude, comparer les données qui proviennent de ces différentes sources et finalement prendre une décision quant à la fonction de leurs protéines d'intérêt.

Notre objectif est de pouvoir effectuer des calculs sur les données, ainsi que d'appliquer des techniques de fouille de données telles que l'extraction de règles d'associations. C'est pourquoi nous avons choisi une architecture d'entrepôt de données, bien adaptée à ces tâches. Notre entrepôt est spécifique et ne suit pas la définition classique d'un entrepôt de données de

[1] http ://microbiogenomics.u-psud.fr