Titre : Tri d'informations biomédicales publiques : prise en compte de leur qualité
Encadrant(s) : Sarah Cohen-Boulakia et Christine Froidevaux
Lieu/Equipe : LRI, équipe Bioinformatique
En collaboration avec : Médecins et biologistes de l'Institut Curie
Contact scientifique : cohen@lri.fr
Pré requis : Une curiosité pour la biologie est attendue. La maîtrise du JAVA est nécessaire.

Description

Ce travail se place dans le contexte de l'étude de maladies génétiques et en particulier des cancers pour lequel l'analyse de nouveaux résultats expérimentaux passe nécessairement par la recherche d'informations issues de nombreuses sources de données. La croissance extrêmement rapide du nombre de sources publiques disponibles (plus de 1 000 recensées en 2009) associée à la croissance exponentielle du nombre de données disponibles dans ces sources rend la phase de recherche d'informations de plus en plus complexe et fastidieuse.

Les données biologiques ont deux caractéristiques. D'abord, elles sont sous la forme de fiches d'annotations qui représentent des expertises, elles sont donc hautement complémentaires mais peuvent aussi être divergentes, et selon les utilisateurs on peut leur associer différents niveaux de fiabilité. Ensuite, elles sont reliées par des références croisées. Une même information (fiche) peut alors être obtenue en suivant différents *chemins* de références croisées, chacun composé de plusieurs sources.

L'objectif des systèmes d'intégration est de proposer une vue unifiée de grands ensembles d'informations. Plusieurs systèmes d'intégration de données exploitent aussi ces chemins alternatifs comme par exemple, BioGuide (Cohen-Boulakia *et al.*), Orchestra (Ives *et al.*), ou Biozon (Yona *et al.*). Néanmoins, dans ce type de systèmes, une requête qui recherche *les articles scientifiques traitant du rôle de l'herceptine dans le cancer du sein* renvoie en Novembre 2009 en moyenne plus de 1 600 réponses.

Identifier rapidement les informations les plus pertinentes dans cette masse d'informations est un réel défi à relever pour tendre vers une meilleure compréhension des mécanismes du cancer et dans la recherche de cibles thérapeutiques.

La pertinence des informations à retourner dépend du profil de l'utilisateur : On peut souhaiter connaître en priorité les informations les plus *stables* (qui décrivent des résultats connus et obtenus par plusieurs chemins), ou encore les informations les plus *rares* ou *nouvelles* tout en étant issues de sources en lesquelles l'utilisateur a une forte *confiance*.

Le travail de stage comportera 3 tâches majeures :

(1) En collaboration avec les partenaires biologistes
   - recenser différents critères de qualité associés aux données (par exemple, niveaux de confiance dans les sources)
   - compléter les *gold standards* qui ont été collectés : constitution d'une base de requêtes d'intérêt et résultats attendus sous la forme de listes de groupes d'informations attendues[1].

(2) Concevoir différentes fonctions de tris et d'extraction des informations retournées en réponse à une requête. Ces fonctions devront pouvoir être paramétrées pour prendre en compte différents critères de qualité des informations retournées.

(3) Implémenter dans le cadre du système BioGuide (codé en JAVA) les fonctions de tri.

[1] Par exemple, *pour la requête Q1 qui recherche les gènes associés à une maladie M, un utilisateur cherchant des informations stables devrait retrouver les gènes g1, g2 et g3 avant les gènes g4 et g5.*