Titre du projet	**BINGO2 : Découverte de connaissances par et pour des requêtes inductives dans des applications en post-génomique**

Résumé

S'appuyant sur le cadre prometteur des bases de données inductives, ce projet a pour but la conception et la réalisation de nouvelles méthodes et outils pour la découverte de connaissances à partir de bases de données afin d'éviter le schéma "trop de motifs après trop de données" qui est malheureusement bien trop classique dans les processus d'extraction de connaissances. Convaincus qu'une recherche amont en fouille de données doit s'intéresser à des scénarios du monde réel, nous avons décidé de nous focaliser sur quelques processus d'extraction de connaissances à partir de données relevant de la biologie moléculaire, comme par exemple la découverte de groupes de synexpression ou encore celle de sites de fixation de facteurs de transcription. Le principe général des bases de données inductives repose sur l'idée que les processus d'extraction de connaissances à partir de bases de données peuvent être considérés comme des processus d'interrogation, i.e. des séquences de requêtes, qui exploitent à la fois les données ou les motifs et modèles sous jacents dans celles-ci. Les requêtes devant retourner des motifs ou modèles sont appelées *requêtes inductives.* Le développement du cadre des bases de données inductives nécessite d'identifier quelles primitives doivent être utilisées pour construire les requêtes et comment de telles requêtes peuvent être évaluées par l'intermédiaire de solveurs (i.e. d'outils basés sur des algorithmes de fouille de données sous contraintes). Une remarque importante concernant l'état de l'art est que la plupart des recherches en fouille de données sous contraintes ne se préoccupent pas de la question de *l'origine des contraintes.* Lorsqu'un analyste sait qu'il peut spécifier des contraintes sur des motifs ou des modèles, est-il possible de l'aider dans son processus de requêtes en le guidant lors de la définition des contraintes relevant du domaine de son problème ?

Les partenaires du projet BINGO2 ont déjà coopéré au développement du cadre des bases de données inductives au sein du projet BINGO (relevant de l'ACI Masse de Données) et qui s'est terminé fin 2007. Dans BINGO2, nous nous intéressons plus particulièrement à déterminer comment de la connaissance (partielle) du domaine (éventuellement elle-même découverte par l'intermédiaire de requêtes inductives) peut être utilisée pour aider à la découverte de connaissances dans les bases de données. Les résultats attendus portent sur la fouille de données séquentielles (notamment par l'apprentissage de distances et l'inférence grammaticale), la découverte de motifs dans les textes et les données 0/1, la fouille de données contraintes (pouvant porter sur des données en plusieurs dimensions et des sources de données hétérogènes) et un site web pour les biologistes pour l'analyse des données SAGE.

Partenaires

UCBN / GREYC
UJM / LaHC
INSA de Lyon / LIRIS
UCBL / CGMC
http://bingo2.greyc.fr/

Coordinateur

Bruno CRÉMILLEUX
Bruno.Cremilleux@info.unicaen.fr

Aide de l'ANR 320 000€

Début et durée Janvier 2008 – 36 mois

Référence ANR-07-MDCO-014