Analyse symbolique de données multi-génomiques dans un cadre RDF réparti
Directeurs de thèse : D. Sherman, P. Durrens
Couriel: david.sherman@labri.fr, pascal.durrens@labri.fr
Equipe : Modèles et Algorithmes pour la bioinformatique et la visualisation

Nature du sujet:

Renforcement d'une thématique scientifiquement reconnue :
Analyse informatique de génomes ; interrogation et exploration de masses de données

Développement d'une thématique émergente :
Calcul sur grille en génomique

Projet transverse
Informatique et biologie

Description synthétique du sujet (maximum 10 lignes):

Grâce au développement de méthodes informatiques et d'algorithmes de calcul efficaces, la fouille de données multi-génomiques est devenu un outil standard très puissant. On le formalise ici par une analyse symbolique des relations n-aires, où les opérations sont réalisées sur divers serveurs de calcul reliés par Internet qui se décrit dans le W3C Resource Description Framework. Le système proposé prend une requête formulé par l'expert, identifie les ressources nécessaires et compile la requête sous forme de stratégie (workflow) à exécute sur une grille de calcul. L'approche sera validée sur un ensemble de 20 génomes complets et des requêtes-types élaborés en collaboration avec un Consortium de biologistes européens.

Description détaillée du sujet:

Grâce au développement de méthodes informatiques et d'algorithmes de calcul efficaces, la fouille de données génomiques est devenue un outil standard très puissant pour la science, utilisé dans divers domaines d'application tels que le développement de biotechnologies, la médecine et bien sûr la recherche fondamentale.

Le cadre multi-génomique présente un défi particulier, car ses méthodes sont basées sur les relations n-aires entre les gènes et non pas sur des propriétés de gènes en tant qu'individus ou en relation avec un seul autre gène de référence. Trois éléments interviennent. En premier, les algorithmes de fouille ou de reconnaissance de formes, nécessaires pour identifier les relations :ils sont très coûteux en temps de calcul, et sont souvent des approximations aux problèmes NP-complets. En deuxième, le besoin de maîtriser de très grandes masses de données et d'importantes ressources en calcul : tout les deux sont nécessaires pour traiter le problème à l'échelle de plusieurs, voire des centaines, séquences de génomes complets. Troisième, le besoin de logiciels d'interrogation et d'exploration : ils ramènent les données à une échelle compréhensible et permet de suivre un raisonnement dans l'inférence de résultats.

Le sujet proposé touche aux trois éléments mais surtout aux deux derniers. Dans le cadre d'un grand projet national et sur la base d'un fonds important de méthodes informatiques réalisées à façon et de données originales, nous allons formaliser le problème de fouille de données sous forme d'une analyse symbolique des relations n-aire multi-génomiques, définir des méthodes informatiques correspondants aux opérations logiques sous-jacentes, et réaliser de modules logiciels pour évaluer et valider l'approche sur de vraies données. Conceptuellement, le système proposé prend une requête formulée par l'expert, identifie les ressources nécessaires et compile la requête sous forme de stratégie (workflow) à exécuter sur une grille de calcul.

Cette approche fait appel à un grand nombre de logiciels existants, disponibles soit sur notre grappe de calcul, soit grâce à leur mise à disposition par les principaux centres de ressources en Europe, Japon et les États-Unis. Chacun déclare ses propriétés et paramètres en Web Services Description Language (WSDL) ou plus généralement dans le Resource Description Framework (RDF), et la compilation de la requête doit tenir compte de ses propriétés pour bien déployer la stratégie sur la grille.

L'approche sera validée sur un ensemble de 20 génomes complets et des requêtes- types élaborés en collaboration avec un Consortium de biologistes européens.

Références :

1: Sherman D, Durrens P, Iragne F, Beyne E, Nikolski M, Souciet JL. Genolevures complete genomes provide data and tools for comparative genomics of hemiascomycetous yeasts. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D432-5.

2: Dujon B, Sherman D, Fischer G, Durrens P, Casaregola S, et al. Genome evolution in yeasts. Nature. 2004 Jul 1;430(6995):35-44.

3: Stevens RD, Robinson AJ, Goble CA. myGrid: personalised bioinformatics on the information grid. Bioinformatics. 2003;19 Suppl 1:i302-4.

4: Hull D, Wolstencroft K, Stevens R, Goble C, Pocock MR, Li P, Oinn T. Taverna: a tool for building and running workflows of services. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W729-32.

5: Spudich G, Fernández-Suárez XM, Birney E.
Genome browsing with Ensembl: a practical overview. Brief Funct Genomic Proteomic. 2007 Sep; 6(3):202-19. Epub 2007 Oct 29.

6: Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics. 2005 Aug 15;21(16):3439-40.