Sujet de thèse 2015-2016

Développement de méthodes d'intégration de sources de connaissances hétérogènes pour une annotation unifiée de groupes de gènes.

EQUIPE/THEME: MABioVis – Thème EvaDOME
DIRECTEURS: Patricia Thébault et Fleur Mougin

MOTS-CLES: *Bioinformatique, intégration, similarité sémantique, data mining, Web sémantique*

RESUME du SUJET:

Les récents progrès des nouvelles technologies de séquençage ouvrent de nouvelles voies pour l'étude du vivant grâce à une augmentation sans précédent de la quantité de données et de leur hétérogénéité. Cet amas d'information est essentiel pour permettre d'améliorer la compréhension des relations entre génotype et phénotype. Pour répondre à ces besoins et, en particulier, pour proposer des annotations pertinentes exprimant la fonction biologique de groupes de gènes reliés phénotypiquement, il est essentiel d'exploiter conjointement toutes ces informations.

Les approches classiques d'annotation de ces groupes de gènes reposent sur des méthodes statistiques d'enrichissement où les annotations résultantes ne retiennent que les termes surreprésentés sans prendre en considération les qualités et quantité variables des informations disponibles pour un organisme vivant. La quantité et la diversité des données à manipuler posent des défis d'intégration à plusieurs niveaux. En effet, ces informations issues de différentes sources sont à la fois de types très hétérogènes et reliées sémantiquement. Pour palier à ces difficultés, il est nécessaire de générer une annotation intégrée en prenant en compte les liens implicites qui existent entre ces informations, en éliminant les éventuelles redondances et en pondérant les termes proposés par une ou plusieurs sources. Dans ce contexte, nous envisageons deux axes de développements qui visent (1) à proposer de nouvelles mesures de similarité sémantique entre des sources d'information hétérogènes et (2) à développer de nouvelles approches de data mining avec pour but commun de déterminer l'ensemble restreint des annotations les plus pertinentes permettant d'expliquer la fonction biologique d'un groupe de gènes.

****

Description détaillée du sujet intitulé

« Développement de méthodes d'intégration de sources de connaissances hétérogènes pour une annotation unifiée de groupes de gènes. »

Unité de recherche :

LaBRI (Laboratoire Bordelais de Recherche en Informatique) UMR 5800
Equipe MABioVis - Modèles et Algorithmes pour la Bioinformatique et la Visualisation d'informations
Thème EvaDOME

Centre de recherche INSERM U897
ERIAS (Equipe de Recherche en Informatique Appliquée à la Santé)

Directeurs de thèse :

Patricia Thébault et Fleur Mougin
thebault@labri.fr et fmougin@labri.fr

Sujet de thèse

Les récents progrès des nouvelles technologies de séquençage ouvrent de nouvelles voies pour l'étude du vivant grâce à une augmentation sans précédent de la quantité de données et de leur hétérogénéité. Nous observons ainsi une croissance constante du nombre de données et de sources de connaissances dans le monde académique, la plupart du temps sous-exploitées. La conception de méthodes informatiques et bioinformatiques pour exploiter et comparer toutes ces informations est devenue un des challenges actuels et nécessite de nouveaux développements.

Par exemple, les données de type RNA-SEQ sont essentielles pour appréhender le transcriptome complet d'une cellule dans plusieurs conditions afin d'identifier des groupes conservés de gènes co-exprimés en lien avec le phénotype. Les approches classiques d'annotation de ces groupes de gènes reposent sur des méthodes statistiques d'enrichissement où les annotations résultantes ne retiennent que les termes surreprésentés sans prendre en considération les qualité et quantité variables des informations disponibles pour un organisme vivant. Par ailleurs, l'annotation fournie par des outils tels que DAVID [1] est brute, sans post-traitement des termes identifiés comme pertinents (on parlera plutôt de connaissances, en opposition aux données, correspondant aux produits de gènes annotés). Ainsi, les utilisateurs récupèrent généralement une liste à plat de termes d'annotation qui nécessiteraient d'être comparés pour disposer d'une annotation intégrée (sans redondance et avec une pondération en fonction du degré de confiance de la source de connaissances d'origine ou encore du nombre de sources spécifiant une annotation donnée).

Pour cela, il est nécessaire d'intégrer en amont les sources de connaissances classiquement utilisées pour l'annotation de groupes de gènes en établissant des correspondances entre

leurs termes. Dans ce cadre, les technologies du Web sémantique et de l'initiative Linked Open data seront particulièrement utiles. Dans le domaine biomédical en particulier, des travaux se sont attachés à rendre disponibles certaines sources de connaissances dans des formats exploitables par les machines (RDF, OWL) avec la volonté de les interconnecter entre elles [2][3][4]. Cependant, les liens décrits entre les différentes sources de connaissances le sont au niveau des données; les cross-références entre identifiants permettent de récupérer des informations croisant différentes sources de connaissances, mais il n'y a pas d'intégration au niveau des connaissances elles-mêmes. En revanche, Callahan *et al.* ont exploité une ontologie comme pivot pour faire correspondre les connaissances de plusieurs sources dans Bio2RDF [5]. Ce travail, particulièrement intéressant mais qui en est encore à un stade préliminaire, n'intègre pas toutes les sources de connaissances permettant d'annoter les groupes de gènes et la description de certaines connaissances n'est pas assez précise.
Afin de fournir une annotation intégrée, ce projet vise à développer des méthodes de similarité sémantique et de data-mining pour améliorer l'interprétabilité biologique des groupes de gènes produits par des méthodes de classification non supervisée. Dans le cadre de l'analyse différentielle, deux niveaux d'étude seront envisagés :

1) Le premier consiste à développer des méthodes informatiques pour identifier les annotations les plus pertinentes pour un groupe de gènes donné, en s'appuyant sur l'ensemble des termes associés à chaque gène issus de sources de connaissances, telles que la GO et KEGG.
2) Le second niveau vise à évaluer le potentiel apport de cette représentation synthétique des annotations (premier niveau d'analyse) par rapport aux méthodes d'enrichissement classiquement utilisées en bioinformatique.

L'application de ces travaux nous permettra de tirer profit de la croissance constante du nombre de données issues des nouvelles technologies de séquençage. Par ailleurs, dans un contexte où les communautés bioinformatique et clinique nécessitent d'échanger des données, le besoin d'intégrer les données et connaissances omiques et phénotypiques est majeur. Dans ce contexte, nous proposerons de nouvelles stratégies d'annotation à haut débit à partir de groupes de gènes identifiés par des méthodes de *clustering* non supervisées [6].

Profil recherché :

Etudiant titulaire d'un master 2 de bioinformatique ou informatique, disposant des compétences suivantes :

- Connaissance des principaux logiciels et méthodes de bioinformatique
- Connaissance des données biologiques
- Maîtrise de l'environnement linux/unix
- Maîtrise des langages Python, Java

Références :

[1] D. W. Huang, B. T. Sherman, Q. Tan, J. Kir, D. Liu, D. Bryant, Y. Guo, R. Stephens, M. W. Baseler, H. C. Lane, et R. A. Lempicki, « DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists », *Nucleic Acids Res.*, vol. 35, n° Web Server issue, p. W169-W175, juill. 2007.

[2] F. Belleau, M.-A. Nolin, N. Tourigny, P. Rigault, et J. Morissette, « Bio2RDF: towards a mashup to build bioinformatics knowledge systems », *J. Biomed. Inform.*, vol. 41, n° 5, p. 706-716, oct. 2008.

[3] M. Samwald, A. Jentzsch, C. Bouton, C. S. Kallesøe, E. Willighagen, J. Hajagos, M. S. Marshall, E. Prud'hommeaux, O. Hassenzadeh, E. Pichler, et S. Stephens, « Linked open drug data for pharmaceutical research and development », *J. Cheminformatics*, vol. 3, n° 1, p. 19, 2011.

[4] M. J. García Godoy, E. López-Camacho, I. Navas-Delgado, et J. F. Aldana-Montes, « Sharing and executing linked data queries in a collaborative environment », *Bioinforma. Oxf. Engl.*, vol. 29, n° 13, p. 1663-1670, juill. 2013.

[5] A. Callahan, J. Cruz-Toledo, et M. Dumontier, « Ontology-Based Querying with Bio2RDF's Linked Open Data », *J. Biomed. Semant.*, vol. 4 Suppl 1, p. S1, avr. 2013.

[6] Thébault P, Bourqui R, Benchimol W, Gaspin C, Sirand-Pugnet P, Uricaru R, Dutour I. Advantages of mixing bioinformatics and visualization approaches for analyzing sRNA-mediated regulatory bacterial networks. Briefings in Bioinformatics. 2014