Description détaillée du sujet intitulé

« Méthodes de clustering pour l'analyse et l'exploration des réseaux de régulation biologique impliquant les petits ARNs non codants »

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 Biologie Computationnelle

Directeurs de thèse :
Isabelle Dutour, Serge Dulucq
Contacts : isabelle.dutour@labri.fr, serge.dulucq@labri.fr

Le sujet de cette thèse porte sur la modélisation et l'analyse des réseaux de régulation biologique impliquant des ARNs non codants (sRNA pour Small RNA). Plus spécifiquement, le sujet vise à contribuer à la caractérisation des réseaux de régulations bactériens impliquant les sRNA en proposant de nouvelles méthodes de clustering permettant d'identifier automatiquement et visuellement des « potentiels » groupes de régulation d'intérêt fonctionnel pour la cellule. Ces candidats ainsi découverts pourront être alors proposés à nos partenaires biologistes pour une validation expérimentale de l'interaction sRNA-mRNA.

Le développement récent de nouvelles technologies de séquençage (NGS) a révolutionné les domaines « omiques » et notamment la génomique fonctionnelle. Il est aujourd'hui possible d'obtenir l'ensemble des sRNA (non codant pour des protéines) pour des bactéries non plus par prédictions bioinformatiques mais par séquençage expérimental. Certains de ces sRNA sont des acteurs essentiels dans la régulation de la traduction des mRNA. La preuve expérimentale d'une interaction entre un sRNA et un mRNA nécessite un travail expérimental de longue haleine, et ne peut pas être envisagée pour le nombre grandissant de sRNA découverts. La bioinformatique semble donc essentielle pour prédire des couples d'interaction sRNA-mRNA d'intérêt fonctionnel pour la cellule qui pourront, au final, être proposés comme candidats pour une validation expérimentale.

Nous avons proposé une démarche permettant l'annotation haut débit des sRNA d'un organisme par une approche semi-automatique en deux étapes principales (article en préparation) :

1. Prédiction des mRNA cibles de sRNA donnés, à l'aide d'un outil de prédiction d'interactions (IntaRNA [1]) choisi parmi les nombreux outils existants pour son bon compromis entre sensibilité et spécificité sur un jeu de données test.
2. Filtrage des prédictions précédentes, qui contiennent beaucoup de faux positifs, par deux approches successives complémentaires :
    a. Caractérisation fonctionnelle des sRNA grâce à une méthode d'enrichissement (DAVID [2]) sur les groupes de mRNA cibles. L'analyse conjointe des annotations extraites de bases de données telles que la GO (Gene Ontology) ou

KEGG (voies métaboliques) permet d'exhiber des sous-groupes de gènes partageant une même annotation et dont la présence conjointe dans le groupe est statistiquement significative et vraisemblablement pertinente biologiquement.

b. Visualisation interactive du graphe d'interactions sRNA-mRNA, enrichi des annotations de bases de données. L'article décrivant le premier prototype de ce nouveau logiciel, rNAV, a été présenté en conférence internationale par Romain Bourqui en octobre dernier [3]. rNAV permet notamment de naviguer efficacement par voisinage au sein du réseau qui peut contenir plusieurs milliers de nœuds et d'arêtes, et d'appliquer différents filtres sur les données (selon les noms de gènes, les zones d'appariement, les annotations fonctionnelles, etc).

L'objectif de cette thèse est de poursuivre ces travaux en mettant en œuvre de nouveaux développements méthodologiques permettant d'identifier automatiquement et visuellement des groupes de régulation d'intérêt fonctionnel pour la cellule, cad des clusters d'interactions sRNA-mRNA partageant significativement les mêmes annotations biologiques issues de l'enrichissement. La figure 1 ci-après montre une capture d'écran d'un (extrait de) réseau de régulation visualisé au sein de rNAV. On y distingue des regroupements intéressants autour de chaque sRNA. Cela permet de cibler les interactions avec les mRNA dont on veut explorer les annotations fonctionnelles. Cependant, cette exploration est aujourd'hui réalisée « manuellement » à l'aide de filtres, et non algorithmiquement.

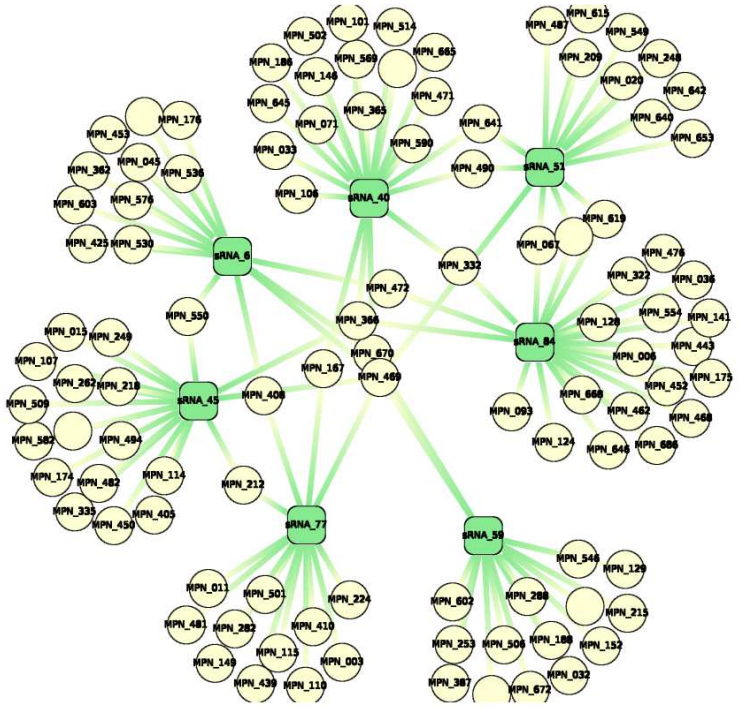

Figure 1 : Extrait d'un réseau de régulation, capture d'écran du logiciel rNAV. Les sRNA sont représentés par des carrés verts et les mRNA par des ronds jaunes.

Proposer un outil générique comme rNAV de construction puis de visualisation et d'exploration dynamique de réseaux d'interactions « clustérisés », combinant un ensemble d'enrichissements (GO, KEGG, etc), serait une réelle avancée pour la communauté. Parmi les défis que nous aurons à relever pour atteindre cet objectif, il y aura deux volets essentiels :

- Un premier volet concernera l'intégration pertinente des données biologiques hétérogènes issues des différentes sources d'information (interactions, enrichissement) avec l'objectif de favoriser la création de clusters d'intérêt. Un verrou concerne en particulier la redondance dans les annotations provoquée par les aspects hiérarchiques de certaines classifications biologiques (par exemple la GO, Wang et al l'ont décrit dans [4]). Pour résoudre ce problème, plusieurs approches peuvent être envisagées en exploitant la topologie, à la fois des classifications et des graphes de régulation : soit pendant la phase d'enrichissement elle-même, soit à son issue en proposant une métrique de similarité sémantique entre les annotations obtenues.

- Un second volet se focalisera sur le développement d'une approche de clustering à la volée exploitant le graphe contenant la totalité des annotations enrichies. Pour répondre à cette question, nous nous intéresserons à l'extension d'algorithmes de *link communities* afin de permettre le regroupement des faisceaux d'arêtes basé sur la métrique du premier volet. Par exemple, l'extension et l'adaptation d'algorithmes de *winding roads* [5] pourrait permettre d'inférer de manière dynamique les sous-graphes correspondant à des clusters de gènes partageant des annotations, et ainsi simplifier le dessin du graphe pour explorer dynamiquement les hubs en intégrant d'autres caractéristiques biologiques (critère de voisinage ou données sur la position des interactions). La représentation visuelle et l'exploration interactive des réseaux clustérisés seront intégrées au logiciel rNAV et réalisées en étroite collaboration avec des collègues du thème EVADOM de notre équipe experts en visualisation d'informations.

Le contexte général de ce projet s'inscrit dans la continuité d'une collaboration avec des biologistes experts des mollicutes (A. Blanchard et P. Sirand-Pugnet, UMR1090, INRA, Bordeaux) [6], pour développer des approches expérimentales pour la validation des interactions sRNA-mRNA. En effet, il est connu que le génome des mollicutes, qui sont des bactéries dites minimales, contient très peu de facteurs transcription. Les sRNA sont donc de très bons candidats pour jouer un rôle crucial dans la régulation de ces bactéries.

[1] A. Busch, A. S. Richter, and R. Backofen. Intarna: efficient prediction of bacterial srna targets incorporating target site accessibility and seed regions. Bioinformatics, 24(24):2849–2856, Dec 2008.
[2] D.W. Huang, B. T. Sherman, Q. Tan, J. Kir, D. Liu, D. Bryant, Y. Guo, R. Stephens, M. W. Baseler, H. C. Lane, and R. A. Lempicki. David bioinformatics resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic Acids Res, 35(Web Server issue):W169–W175, Jul 2007.
[3] Dubois J., Ghozlane A., Thébault P., Dutour I., Bourqui R., Genome-wide detection of sRNA targets with rNAV, Proc. in 3rd IEEE Symposium on Biological Data Visualization, 13-14 October 2013, Atlanta, USA.
[4] Wang J, Zhou X, Zhu J, Gu Y, Zhao W, Zou J, Guo Z. GO-function: deriving biologically relevant functions from statistically significant functions. Briefings in Bioinformatics 2012;13:216-227.
[5] Lambert A., Dubois J., Bourqui R., Pathway Preserving Representation of Metabolic Networks, Computer Graphics Forum special issue on 13th Eurographics/IEEE-VGTC Symposium on Visualization, 1021-1030, 2011.
[6] Sirand-Pugnet P, Lartigue C, Marenda M, Jacob D, Barré A, Barbe V, Schenowitz C, Mangenot S, Couloux A, Segurens B, de Daruvar A, Blanchard A, Citti C, Being pathogenic, plastic, and sexual while living with a nearly minimal bacterial genome. PLoS Genet. 2007 May 18;3(5):e75.