Intégration de données transcriptomiques volumineuses et hétérogènes via des calculs de similarité de graphe

Modalités pratiques

La thèse décrite dans ce document devrait débuter en octobre ou novembre 2017. Les candidats intéressés doivent se faire connaître au plus tard fin mai 2017 (voir partie « Contacts » de ce document) en envoyant un CV et une lettre de motivation, en français ou en anglais.

Résumé

Une question d'intérêt majeur en biologie est de comprendre les relations de régulation entre les gènes : elles permettent d'améliorer la compréhension du fonctionnement global de la cellule et d'identifier les gènes impliqués dans un phénomène d'intérêt comme une maladie. Le développement des technologies d'acquisition haut-débit a permis la collecte importante sur le fonctionnement des organismes vivants à divers niveaux d'échelle (génome, transcriptome, cellule, ...). Cependant, la combinaison de ces informations pose de nouveaux défis : les informations sont non seulement nombreuses et volumineuses mais aussi hétérogènes car collectées dans des conditions expérimentales différentes par des techniques expérimentales différentes et à des niveaux différents du vivant. Aussi, cette masse d'information reste très sous-exploitée. Notre objectif est de proposer une méthode à base de graphes pour pouvoir reconstruire un réseau de régulation génique à partir d'informations hétérogènes.

Présentation scientifique du projet de recherche

En biologie, le développement récent des technologies d'acquisition haut-débit a permis la collecte d'informations sur le fonctionnement des organismes vivants. Cependant, les données collectées sont non seulement volumineuses mais aussi très hétérogènes car obtenues dans des conditions expérimentales différentes, par des techniques expérimentales différentes et à des niveaux différents du vivant (différents 'omiques). Aussi, cette masse d'information reste **très largement sous-exploitée**, faute d'outils d'analyse adéquats pour intégrer ces informations.

Afin de comprendre les relations de régulationexistantes entre les gènes ([Zhang & Mallick, 2013; Montastier et al., 2015]), nous souhaitons proposer une solution pour **l'intégration de données transcriptomiques** pour la reconstruction de réseaux de régulation. En effet, les données transcriptomiques (ou données d'expression de gènes) donnent un instantané exhaustif de l'état d'une cellule dans un tissu donné, un organisme donné et des conditions environnementales données. Le site GEO (Gene Expression Omnibus https://www.ncbi.nlm.nih.gov/geo) est largement utilisé pour la mise à disposition publique de ce type de données.

Toutefois, si les méthodes permettant de re-construire un réseau (ou graphe) de régulation à partir de données transcriptomiques sont maintenant bien établies ([Friedman et al., 2008; Allen & Liu, 2012], par exemple), celles-ci permettent d'obtenir un réseau pour chaque expérience mais pas de combiner les résultats de plusieurs expériences. Des travaux ont abordé la question de l'intégration d'expériences diverses : une première approche consiste à utiliser des modèles

statistiques incorporant diverses expériences par des contraintes adéquates [Chiquet et al., 2011; Mohan et al., 2012; Danaher et al., 2013; Villa-Vialaneix et al., 2013] mais ces approches permettent surtout de construire des réseaux joints correspondant à des conditions expérimentales différentes (cas et contrôle par exemple) dans une même expérience. Une approche alternative consiste à travailler au niveau des graphes inférés à partir des diverses expériences et à utiliser la fréquence d'apparition d'une arête [Ballouz et al., 2015]. Si l'approche montre une amélioration de la qualité de l'inférence, elle est très simpliste et ne tient compte ni de la structure globale de chacun des graphes, ni de leurs pertinences par rapport à l'ensemble (similarité/atypicité)

Notre objectif, est de proposer une approche alternative permettant **d'intégrer des réseaux de régulation** inférés à partir d'expériences transcriptomiques diverses. Notre approche se base sur la combinaison de méthodes permettant d'obtenir des similarités entre graphes à partir de méthodes de détection de sous-graphes similaires. Le projet comporte donc des aspects de modélisation biostatistique (inférence de réseaux, fouille de données sur les graphes ou les données décrites par des similarités) et des aspects de recherche informatique (algorithme de calcul de similarités entre graphes , et incorporant éventuellement des informations de contexte, comme par exemple les fonctions connues des gènes). Il utilisera des données transcriptomiques publiques en se concentrant sur un ou plusieurs cas d'étude dans lesquels les données de plusieurs expériences, obtenues avec des technologies de mesures différentes sur un même organisme et un même tissu, seront utilisées. Pour chacune de ces expériences, un réseau de co-expression sera inféré et l'objectif sera donc de synthétiser ces informations hétérogènes. L'objectif du projet est d'étendre ces approches pour déterminer de manière automatique les similarités entre sous graphes. L'approche consisterait donc à comparer différents graphes en les considérants comme des sous-ensembles fortement couplés. L'enjeu d'une telle approche est d'améliorer la qualité des calculs de similarité pour mieux détecter les parties de graphes similaires et ensuite utiliser ces informations de proximité dans un but d'intégration des différents graphes ou sous-graphes. L'identification de sous parties ouvre la possibilité aux algorithmes de s'ajuster différemment selon les sous-ensembles des graphes.

Les mesures de ressemblance ainsi obtenues permettront d'avoir une vue d'ensemble des graphes inférés. Ces informations pourront être utilisées, dans un premier temps, pour des approches de fouilles de données décrites par des similarités (comme celles de [Olteanu & Villa-Vialaneix, 2015]), ce qui permettra de déterminer les réseaux typiques ou atypiques et éventuellement d'affiner l'intégration en retirant les cas trop marginaux. Enfin, une proposition d'un réseau consensus sera obtenue par calcul d'un réseau « moyen » au sens de la similarité introduite, par exemple (i.e., réseau dont la similarité moyenne à tous les réseaux est la plus grande).

Compétences requises

Master en informatique et mathématiques appliquées. Les compétences suivantes sont attendues :

- Conception, implantation et interrogation de bases de données relationnelles
- Intégration de données hétérogènes (ETL, …)
- Management et interrogation de données graphes (BD NO SQL)
- Algorithmes de matching de graphes

- Statistique, analyse de données
- Connaissances en traitement distribué (Map Reduce, Hadoop) serait un plus

Description du co-encadrement

Le projet est de nature pluridisciplinaire (biostatistique, théorie de graphes et management des données hétérogènes et volumineuses). Il sera co-encadré par Franck Ravat (IRIT, UT1), Nathalie Villa-Vialaneix (MIAT, INRA), Cassia Trojahn (IRIT, UT2) et Olivier Teste (IRIT, UT2).

MIAT (INRA) : l'inférence de réseaux de régulation et l'intégration de données biologiques hétérogènes sont des thématiques étudiées depuis plusieurs années par l'équipe SaAB. Ce projet est dans la continuité des travaux existant et aborde la question avec des outils situés en dehors des approches statistiques habituelles.

IRIT : les équipes SIG et Melodi ont déjà travaillé ensemble sur l'intégration de données hétérogènes et volumineuses afin de faciliter la prise de décision. Ce projet de thèse serait un complément intéressant au projet IBLID (sept 2016, Labex CIMI). IBLID (Integration of Big and Linked Data for On-Line Analytics) repose sur la représentation unifiée de données hétérogènes afin de faciliter la tâche des décideurs. De même, l'équipe SIG a proposé des solutions pour l'intégration de données hétérogènes à l'aide de graphes de données.

Contacts

Franck Ravat franck.ravat@irit.fr
Olivier Teste olivier.teste@irit.fr
Cassia Trojahn cassia.trojahn@irit.fr
Nathalie Villa-Vialaneix nathalie.villa-vialaneix@inra.fr

Références

[Allen & Liu, 2012] Allen, G. and Liu, Z. (2012). A log-linear graphical model for inferring genetic networks from high-throughput sequencing data. IEEE International Conference on Bioinformatics and Biomedicine (BIBM).
[Ballouz et al., 2015] Ballouz, S., Verleyen, W. & Gillis, J. (2015). Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. Bioinformatics, 31(13), 2123-2130.
[Chiquet et al., 2011] Chiquet, J., Grandvalet, Y. & Ambroise, C. (2011). Inferring multiple graphical structures. Statistics and Computing, 21(4), 537-553.
[Danaher et al., 2013] Danaher, P., Wang, P. & Witten, D. (2013). The joint graphical lasso for inverse covariance estimation across multiple classes. Journal of the Royal Statistical Society B, 76(2), 373-397.
[Friedman et al., 2008] Friedman, J., Hastie, T. & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. Biostatistics, 9(3), 432-441.
[Marx, 2013] Marx, V. (2013). Biology: the big challenge of big data. Nature, 498, 255-260.
[Megdiche et al., 2016] Megdiche, I. Teste, O. & Trojahn, C. (2016). An extensible linear approach for holistic ontology matching. International Semantic Web Conference, 1, 393-410.
[Mohan et al., 2012] Mohan, K., Chung, J.Y., Han, S., Witten, D., Lee, S.I. & Fazel,

M. (2012). Structured learning of Gaussian graphical models. In Proceedings of NIPS , Lake Tahoe, Nevada, USA.
[Montastier et al., 2015] Montastier, E., Villa-Vialaneix, N., Caspar-Bauguil, S., Hlavaty, P., Tvrzicka, E., Gonzalez, I., Saris, W., Langin, D., Kunesova, M. & Viguerie, N. (2015). System model network for adipose tissue signatures related to weight changes in response to calorie restriction and subsequent weight maintenance. PLoS Computational Biology, 11(1), e1004047.
[Olteanu & Villa-Vialaneix, 2015] Olteanu, M., & Villa-Vialaneix, N. (2015). On-line relational and multiple relational SOM. Neurocomputing, 147, 15–30.
[Villa-Vialaneix et al., 2013] Villa-Vialaneix, N., Vignes, M., Viguerie, N. & San Cristobal, M. (2013) Inferring networks from multiple samples with consensus LASSO. Quantitative Technology and Qualitative Management, 11(1), 39-60.
[Zhang & Mallick, 2013] Zhang, L. & Mallick, B. (2013). Inferring gene networks from discrete expression data. Biostatistics, 14(4), 708-722.