Projet de thèse

1. Titre et encadrement de la thèse

Titre: Caractérisation de cibles thérapeutiques dans un programme génique tumoral

Co-supervision:
- Nicolas Champagnat (nicolas.champagnat@inria.fr), Pierre Vallois (pierre.vallois@univ-lorraine.fr), IECL, Université de Lorraine and Inria Nancy – Grand Est
- Laurent Vallat (laurent.vallat@chru-strasbourg.fr), Inserm UMR S1113, Université de Strasbourg

Financement: Région Grand Est.

2. Contexte

Dans la plupart des cancers, l'accumulation d'aberrations génétiques altère progressivement les programmes géniques, conduisant à des comportements aberrants et à la prolifération cellulaire incontrôlée. Diverses méthodes statistiques ont été proposées afin de découvrir les réseaux de régulation de gènes qui sous-tendent ces programmes géniques. L'identification de gènes cibles dont une modification d'expression permettrait de moduler le programme génique pourrait conduire à de nouvelles approches thérapeutiques dans le cancer. Dans le cas des leucémies lymphoïdes chroniques, il est possible de collecter des données *temporelles* d'expression de gènes, qui peuvent être utilisées pour inférer un réseau de régulation de gènes *dynamique* (Vallat et al. 2013).

Pour l'inférence des réseaux de gènes, diverses méthodes statistiques existent, qui reposent sur une matrice d'interactions entre gènes (réseaux de corrélation, réseaux bayésiens, modèles de régression, cf. Marbach et al., 2010, Allouche et al., 2013). Les méthodes de régression linéaire (ou non-linéaire) pénalisée (par exemple par LASSO ou Dantzig selector) sont plus adaptées à notre contexte, car elles fournissent des réseaux parcimonieux permettant de prédire l'effet d'une action sur un gène particulier.

3. Objectifs

La thèse portera sur les différents problèmes mentionnés ci-dessus, à savoir : l'amélioration de la méthodologie développée par Vallat et al. (2013) ; la conception et la mise en œuvre de modèles et techniques d'estimation pour les réseaux de groupes de gènes ; la conception d'un plan d'expérience pour la validation des réseaux inférés. Par ailleurs, les méthodes devront être étendues pour traiter simultanément des données d'expression d'ARN et de protéines. L'implémentation numérique des méthodes développées, leur test sur deux jeux de données déjà collectés par l'équipe de Laurent Vallat et leur comparaison avec les logiciels disponibles sur Bioconductor sont des caractéristiques essentielles du projet.

Une première période sera consacrée à l'apprentissage par le doctorant des outils et connaissances nécessaire au projet de thèse. Puisqu'il s'agit d'un projet pluridisciplinaire, cette étape est cruciale et représente un travail important. Il s'agit d'acquérir les notions relatives au contexte biologique, aux différents modèles de réseaux de gènes, aux méthodes d'inférence associées, et aux logiciels d'inférence de réseaux de gènes, particulièrement la librairie R CASCADE (Jung et al., 2013) qui met en œuvre la méthode d'inférence de Vallat et al. (2013). D'autres librairies reliées seront également étudiées, comme par exemple TDARACNE et GRENITS (disponibles sur la base de données de logiciels en bioinformatique Bioconductor). Le doctorant s'appuiera sur une première étude bibliographique relative à l'inférence de réseaux de gène actuellement en cours de réalisation dans l'équipe.

Dans Vallat et al. (2013), les auteurs ont déterminé des clusters de gènes et ont ensuite estimé par LASSO un modèle linéaire dynamique en exploitant les différents clusters de gènes. Cette étude

laisse plusieurs questions en suspens, notamment en ce qui concerne la classification non-supervisée, qui est réalisée sans tenir compte du rôle des clusters dans le modèle de réseau de gènes, et pour ce qui concerne la validation du réseau inféré. Ce dernier point est crucial dans l'optique de la prédiction de l'effet de la modification d'expression de certains gènes sur le comportement global du réseau. Bien que les données soient en grande dimension, puisqu'elles mesurent l'expression de plus de 20 000 gènes, elles portent habituellement sur un faible nombre de patients et un faible nombre de temps de mesure. De ce fait, il est irréaliste d'espérer pouvoir inférer le réseau de gènes complet. Nous serons donc conduits à faire des hypothèses restrictives pour proposer un modèle statistique adapté aux données qui soit prédictif, par exemple en étudiant la dynamique de groupes de gènes plutôt que de chaque gène individuellement. Du point de vue expérimental, la validation du réseau de gènes inféré et de ses prédictions pourra être réalisée par invalidation biologique expérimentale de quelques gènes ou groupes de gènes bien choisis dans le réseau.

Le travail du doctorant visera premièrement à améliorer l'étape de classification de Vallat et al. (2013) en la réalisant conjointement à l'inférence du réseau de manière itérative. Cette approche nécessite de calculer des scores de confiance dans le modèle inféré, par exemple par Boostrap (Allouche et al., 2013). La deuxième partie de la thèse visera à exploiter la grande dimension des données recueillies en inférant seulement une matrice d'interaction entre l'expression agrégée de *groupes de gènes*. On peut espérer réduire ainsi l'erreur statistique d'estimation. La troisième partie de la thèse portera sur la validation des méthodes développées et sur la construction d'un cadre théorique pour proposer des expériences biologiques pour cette validation.

Bibliographie

- D. Allouche, C. Cierco-Ayrolles, S. de Givry, G. Guillermin, B. Mangin, T. Schiex, J. Vandel, M. Vignes. A panel of learning methods for the reconstruction of gene regulatory networks in a systems genetics context. In: A. de la Fuente (ed.), Gene Network Inference: Verification of Methods for Systems Genetics Data, Springer-Verlag Berlin, 2013.
- N. Jung, F. Bertrand, S. Bahram, L.Vallat, M. Maumy-Bertrand. Cascade: a R-package to study, predict and simulate the diffusion of a signal through a temporal gene network. Bioinformatics, 30(4), 571-3, 2014.
- D. Marbach, R.J. Prill, T. Schaffter, C. Mattiussi, D. Floreano, G. Stolovitzky. Revealing strengths and weaknesses of methods for gene network inference. PNAS 107(14), 6286-6291, 2010.
- L. Vallat, C.A. Kemper, N. Jung, M. Maumy-Bertrand, F. Bertrand, N. Meyer, A. Porcheville, J.W. Fisher III, J.G. Gribben, S. Barham. Reverse-engineering the genetic circuity of a cancer cell with predicted intervention in chronic lymphocytic leukemia. PNAS 110(2), 459–464, 2013.

4. Partenariat scientifique

La thèse se déroulera à l'IECL (Institut Élie Cartan de Lorraine), le laboratoire de Mathématiques de l'Université de Lorraine. Le groupe de probabilités et statistiques, composé de plus de 30 membres permanents, est le plus grand de la Région Grand Est. Deux équipes-projet Inria appartiennent à cette équipe : la première, BIGS (Biology, Genetics et Statistics) travaille sur la modélisation stochastique et statistique pour la biologie et la médecine ; la seconde, TOSCA (TO Simulate and CAlibrate stochastic models) travaille sur la modélisation stochastique, le contrôle stochastique et les méthodes numériques probabilistes, avec une grande partie des applications orientées vers la biologie (écologie, évolution, médecine, neurosciences).

Pierre Vallois (Professeur Université de Lorraine) et Nicolas Champagnat (Chargé de Recherche Inria) sont habitués à travailler avec des biologistes et des cliniciens. N. Champagnat a obtenu récemment un financement de l'ITMO Cancer pour un projet sur la modélisation dynamique de l'ADN circulant, avec des médecins de l'Institut de Cancérologie de Lorraine et du CHRU de Strasbourg.

Laurent Vallat est docteur en médecine (spécialisé en hématologie biologique et immunologie) et titulaire d'un doctorat en sciences (spécialité des bases moléculaires de l'oncogénèse). Il est MCU-PH en hématologie à la Faculté de médecine, médecin biologiste dans le service d'hématologie de l'hôpital de Strasbourg et il dirige un groupe de recherche dans l'unité Inserm UMR S_1113. Il a coordonné un projet ITMO Cancer qui a permis de mettre en place un modèle biologique ex vivo

de stimulation des cellules leucémiques et de générer un nouveau jeu de données d'expressions temporelles de gènes et d'abondance de protéines.

5. Conditions de réalisation de la thèse

Le doctorant sera membre de l'IECL et la thèse se déroulera dans l'Ecole Doctorale IAEM (Informatique, Automatique, Electronique-Electrotechnique, Mathématiques, ED 77) de l'Université de Lorraine, à Nancy. Laurent Vallat travaillant à Strasbourg, il sera important d'organiser des rencontres régulières, alternativement à Nancy ou Strasbourg. Nous planifions d'organiser ces rencontres au moins une fois par mois. L'étudiant sera également amené à se rendre régulièrement au CHRU de Strasbourg, au moins une fois toutes les deux semaines, où il disposera d'un bureau et de matériel informatique.

L'étudiant bénéficiera à l'IECL de matériel informatique, d'une bibliothèque de Mathématiques interne au laboratoire et d'un accès aux publications scientifique via les abonnements de l'IECL et de l'Université de Lorraine.