PHD THESIS PROPOSAL

Heterogeneous biological data integration through inference of mixed graphical model

Intégration de données hétérogènes par inférence de modèle graphique mixte

Christophe Ambroise[1*] and Svetlana Gribkova[2]

*Correspondence: christophe.ambroise@univ-evry.fr
[1]http://www.math-evry.cnrs.fr/members/cambroise/welcome, LaMME, université d'Évry val d'Essonne, , Évry, France
Full list of author information is available at the end of the proposal

Keywords: mixed graphical model; large dimension; exponential family; data integration

Context

Trees play a crucial role in human life and ecosystems. Forests are a pivotal element in the adaptation and mitigation of global warming. In this context, the study of mechanisms for adapting trees to environmental conditions is an important scientific and societal challenge. Modern "omics" technologies make it possible to obtain genomic, transcriptomic, epigenetic and other data on trees placed under different environmental conditions. An integrative statistical analysis of "omic" data involves jointly studying the available datasets in order to highlight different molecular regulatory mechanisms that interact and influence the phenotype.

Les arbres jouent un rôle crucial dans la vie de l'homme et des écosystèmes. Les forêts constituent l'élément charnière dans l'adaptation et l'atténuation du réchauffement climatique. Dans ce contexte l'étude de mécanismes d'adaptation des arbres aux conditions environnementales constitue un enjeux scientifique et sociétal important. Les technologies "omiques" modernes permettent d'obtenir des données génomiques, transcriptomiques, épigénétiques, etc. sur des arbres placés dans différentes conditions environnementales. Une analyse statistique intégrative de données "omiques" consiste à étudier conjointement les jeux de données disponibles afin de mettre en lumière différents mécanismes de régulation moléculaire qui interagissent et déterminent le phénotype.

Project description

Networks are a natural way of describing how biological variables interact. The numerous statistical methods of inferences of existing networks generally assume that the network is sparse (few effective interactions exist among all the possibilities) and sometimes make the hypothesis of the existence of an underlying structure (

group or hierarchy of variables). These a priori make it possible both to reduce the size of the problem and to offer a synthetic summary of the important interactions between the variables.

The aim of this thesis is to propose network inference methods from qualitative and quantitative variables (phenotype, genotype, methylome, transcriptome) taking into account a possible group structure or hierarchy between variables. Mixed graphical models represent a well suited model to infer relationships between variables of different nature. The research will consider a general sub-class of graphical models where the node-wise conditional distributions arise from exponential families. The inference of conditional independence between the variables of interest should allow a better understanding of the phenotypic plasticity of trees. The high dimension statistical context (few samples compared to the number of variables) and the difference in nature of the variables considered constitute the major difficulties of this research work at the interface.

Les réseaux constituent un façon naturelle de décrire comment des variables biologiques, phénotype, méthylome, transcriptome ... interagissent. Les nombreuses méthodes statistiques d'inférences de réseaux existantes supposent généralement sur que le réseau est creux (peu d'interactions effectives existent parmi l'ensemble des possibles) et font parfois l'hypothèse de l'existence d'une structure sous-jacente (groupe ou hiérarchie de variables). Ces a priori permettent à la fois de réduire la taille du problème et d'offrir un résumé synthétique des interactions importantes entre les variables.

L'objectif de cette thèse est de proposer des méthodes d'inférence de réseau à partir de variables qualitatives et quantitatives (phénotype, génotype, méthylome, transcriptome) prenant en compte une éventuelle structure de groupe ou hiérarchie entre variables. Les modèles graphiques mixtes représentent un modèle bien adapté pour inférer des relations entre des variables de nature différente. La recherche considérera une sous-classe générale de modèles graphiques où les distributions conditionnelles nodales proviennent de familles exponentielles. L'inférence de l'indépendance conditionnelle entre les variables d'intérêt devrait permettre une meilleure compréhension de la plasticité phénotypique des arbres. Le contexte statistique de grande dimension (peu d'échantillons par rapport au nombre de variables) et la différence de nature des variables considérées constituent les difficultés majeures de ce travail de recherche à l'interface.

Collaboration

This research is mainly motivated by the project ANR EPITREE (https://www6.inra.fr/epitree-project/Le-projet-EPITREE) which aims to study the impact of epigenetics (DNA methylation), gene expression and allelic variation in the mechanisms of adaptation of trees to the local environment and their phenotypic plasticity. Given the structure of the ANR project, the research in applied mathematics is at the interface of ecology and bioinformatics and thus involves collaborations with ecologists and bioinformaticians from Bordeaux, Orléans and Clermont-Ferrand.

Cette recherche est principalement motivée par le projet ANR EPITREE (https://www6.inra.fr/epitree-project/Le-projet-EPITREE) qui vise à étudier l'impact de l'épigénétique (méthylation ADN), de l'expression des gènes et de la variation allélique dans les mécanismes d'adaptation des arbres à l'environnement local et de leur plasticité phénotypique. De part la structure du projet ANR, la recherche se situe à l'interface des mathématiques, de l'écologie et la bio-informatique et implique des collaborations avec des chercheurs écologues et bio-informaticiens de Bordeaux, Orléans et Clermont-Ferrand.

Profile and skills required

The applicant will need to have a strong background in mathematics and statistics (typically Master's level or école d'ingénieur) as well a real motivation for applications in life sciences. An experience in computer programming is also requested.

Le candidat recherché devra avoir un solide bagage mathématique et statistique, issu typiquement M2 ou d'une école d'ingénieur ainsi qu'un intérêt certain pour les applications en sciences du vivant. Une expérience en programmation sera également nécessaire.

Author details

[1] http://www.math-evry.cnrs.fr/members/cambroise/welcome, LaMME, université d'Évry val d'Essonne, , Évry, France. [2] https://www.lpsm.paris/pageperso/gribkova/,LPMA, université Paris-Diderot, Paris, France.

References

1. Laby, R., Roueff, F., Gramfort, A.: Anomaly detection and localisation using mixed graphical models. arXiv preprint arXiv:1607.05974 (2016)
2. Meinshausen, N., Bühlmann, P.: High-dimensional graphs and variable selection with the lasso. The annals of statistics, 1436–1462 (2006)
3. Plomion, C., Bastien, C., Bogeat-Triboulot, M.-B., Bouffier, L., Déjardin, A., Duplessis, S., Fady, B., Heuertz, M., Le Gac, A.-L., Le Provost, G., *et al.*: Forest tree genomics: 10 achievements from the past 10 years and future prospects. Annals of forest science **73**(1), 77–103 (2016)
4. Ravikumar, P., Wainwright, M.J., Lafferty, J.D., *et al.*: High-dimensional ising model selection using l1-regularized logistic regression. The Annals of Statistics **38**(3), 1287–1319 (2010)
5. Yang, E., Ravikumar, P., Allen, G.I., Liu, Z.: Graphical models via univariate exponential family distributions. Journal of Machine Learning Research **16**(1), 3813–3847 (2015)