Titre

Titre : Classification hiérarchique sous contrainte de contiguïté pour l'analyse de données Hi-C.

Enjeux scientifiques

La classification ascendante hiérarchique (CAH) est une méthode d'analyse exploratoire des données très largement utilisée. Son cadre naturel est celui dans lequel les données d'entrée sont fournies sous la forme d'une matrice de distances. Un critère de lien (*linkage*) définit alors la manière dont les distances entre classes sont calculées à partir de cette distance. Si ce cadre est bien connu, particulièrement lorsque la matrice d'entrée est une distance euclidienne, il reste à approfondir et à élargir pour des cas moins standards qui ont des applications pratiques importantes.
En particulier, ce stage s'intéressera à l'étude et l'extension de la CAH pour l'analyse de données issues de la génomique. Dans ce cadre, les objets à classer sont des positions du génome (loci, gènes, groupes de gènes), organisés par leur position sur un chromosome. Dans plusieurs applications typiques en génomique, ces positions sont également liées entre elles par des données définissant une similarité ou une dissimilarité entre paires de positions : c'est le cas pour les données Hi-C, issues du séquençage haut-débit, qui correspondent à des mesures de la proximité spatiale (dans la cellule) entre paire de groupes de loci (appelés bins). Dans ces cas, trouver des classes de positions *contiguës*, qui respectent l'organisation du chromosome, a des applications importantes : les classes de loci contigus du déséquilibre de liaison permettent d'améliorer la sélection de variables dans les études d'association[1] et les classes de bins contigus obtenues à partir de données Hi-C peuvent s'interpréter en terme de domaines fonctionnels (TAD)[2]. Dans ce dernier cas, l'aspect hiérarchique de la classification est, en outre, particulièrement bien adapté au contexte biologique puisqu'on attend un enroulement hiérarchique des chromosomes[3].
Le stage abordera donc les problématiques suivantes :

1. il analysera les extensions possibles de la CAH standard à des données décrites par des noyaux (matrices de similarités définies positives), des similarités quelconques et des dissimilarités euclidiennes ou non euclidiennes. Des travaux existent[4] qui font le lien entre certaines de ces versions dans le cas particulier des noyaux et des distances induites par celles-ci mais le cas de similarités quelconques est encore largement non étudié, de même que l'utilisation de critères de lien différent du critère de Ward ;
2. il analysera des données réelles et mettra en relation classification telle que produite par la CAH contrainte et TADs tels que fournis par les outils de bioinformatiques existants pour aborder ce problème. Cette partie s'appuiera sur un package **R** déjà développé par les encadrants et sera appliqué à un problème lié à la mortalité péri-natale chez le cochon pour lequel on dispose de données Hi-C à deux stades du développement embryonnaire. Cette application doit permettre d'améliorer la compréhension des mécanismes physiologiques

---

1 Dehman A., Ambroise C., Neuvial P. (2015) Performance of a blockwise approach in variable selection using linkage disequilibrium information. *BMC Bioinformatics*, **16**, 148.

2 Fotuhi Siahpirani A., Ay F., Roy S. (2016) A multi-task graph-clustering approach for chromosome conformation capture data sets identifies conserved modules of chromosomal interactions. *Genome Biology*, **17**, 114.

3 Fraser J., Ferrai C., Chiariello A.M., Schueler M., Rito T., Laudanno G., Barbieri M., Moore B.L., Kraemer D.C.A., Aitken S., Xie S.Q., Morris K.J., Itoh M., Kawaji H., Jaeger I., Hayashizaki Y., Carninci P. Forrest A.R.R., The FANTOM Consortium, Semple C.A., Dostie J., Pombo A., Nicodemi M. (2015) Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Molecular Systems Biology*, **11, 852.**

4 Chavent M., Kuentz-Simonet V., Labenne A., Saracco J. (2017) Hierarchical clustering with two dissimilarity matrices: the ClustGeo2 R package. *Preprint.*
Miyamoto S., Abe R., Endo Y., Takeshita J. (2015) Ward method of hierarchical clustering for non-Euclidean similarity measures. *In Proceedings of the VIIth International Conference of Soft Computing and Pattern Recognition (SoCPaR 2015)*.
Dehman A. (2015) Spatial Clustering of Linkage Disequilibrium Blocks for Genome-Wide Association Studies. Thèse de doctorat de l'université Paris Saclay.

permettant la survie à la naissance en identifiant les loci concernés par la réorganisation spatiale de la chromatine en fin de gestation.

Le stage a donc des aspects appliqués et des aspects plus théoriques ou méthodologiques et pourra, à son issue, déboucher sur une poursuite en thèse.

Localisation

Unité MIAT, INRA de Toulouse et Institut de Mathématiques de Toulouse (Université Toulouse III)

Encadrants

Nathalie Villa-Vialaneix (MIAT, INRA, Toulouse)
Pierre Neuvial (IMT, Université Toulouse 3)