Title

Hierarchical clustering with contiguity constraints for Hi-C data analysis

Description of the PhD proposal

Hierarchical ascendant clustering (HAC) is a widely used exploratory statistical analysis. In its basic presentation, input data are given as distance matrices between objects and a linkage criterium defines distances between clusters based on the distances between input objects. This framework is well-known when the distance is Euclidean but less standard cases, including non Euclidean dissimilarities or arbitrary similarities, are still to be better understood and studied. In particular, this internship will focus on the specific case of an extension of HAC that is well suited to genomic data. In this type of applications, objects to be clustered are positions of the genome (loci, genes, groups of genes…), organized by their place along the genome sequence. In several typical applications in genomics, these positions are also related to each others by an external information that provides insight on a similarity between them. For instance, this is the case for Hi-C data, obtained from Next Generation Sequencing (NGS), that are measure of the spatial proximity (in the cell) between pairs of loci (usually called bins). In this case, it makes sense to find clusters that are composed of *contiguous* bins because these clusters are in adequation with the chromosomic organization and can be interpreted in terms of functional domains (TAD)[1]. They are also in adequation with what we know about the physical hierarchical organization of the chromosomes within the cell[2].

The internship aims at addressing the following issues:

1. it will study possible extensions of HAC to data described by kernels, arbitrary similarity measures or arbitrary dissimilarities. Previous articles[3] make a link between some of these variants in the specific case of kernels and of distances they induce but the case of arbitrary similarities is still an open issue, as well as the use of linkage criteria other than Ward's linkage;
2. it will analyze real data and will study the relations between results obtained by HAC and TADs as obtained by bioinformatics tools that address this issue. This part will be based on an already developed **R** packages that is maintained by the supervisors of this internship. The work will focus on an application linked to perinatal mortality in pigs for which Hi-C data have been obtained at two embryonic development stages. This application aims at a better understanding of the physiological mechanisms linked to birth survival by identifying loci related to the spatial reorganization of the chromatin at the end of the gestation.

The internship thus has applicative and theoretical or methodological aspects and could be the starting point of a PhD subject.

---

1 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.

2 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.**

3 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.

Location

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

Supervisors

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