Adjacency-constrained hierarchical clustering of a band similarity matrix with application to genomics

Christophe Ambroise[1], Alia Dehman[2], Michel Koskas[3], Pierre Neuvial[4], Guillem Rigaill,[1,5] and Nathalie Vialaneix[6]

[1] Laboratoire de Mathématiques et Modélisation d'Évry, UMR CNRS 8071, Université d'Évry Val d'Essonne, 23 boulevard de France, 91037 Évry, France
christophe.ambroise@univ-evry.fr

[2] Hyphen-stat, 195 Route d'Espagne, 31036 Toulouse, France
alia.dehman@hyphen-stat.com

[3] UMR518 AgroParisTech/INRA, , 16 rue Claude Bernard, 75231 Paris Cedex 05, France
michel.koskas@agroparistech.fr

[4] Institut de Mathématiques de Toulouse, UMR5219 CNRS, Université de Toulouse, UPS IMT, F-31062 Toulouse Cedex 9, France
pierre.neuvial@math.univ-toulouse.fr

[5] Institute of Plant Sciences Paris Saclay IPS2, CNRS, INRA, Gif sur Yvette, France.
guillem.rigaill@inra.fr

[6] MIAT, Université de Toulouse, INRA, Castanet-Tolosan, France
nathalie.vialaneix@inra.fr

Genetic information is coded in long strings of DNA organized in chromosomes. High-throughput sequencing now allow to study biological phenomena along the whole genome at a very high resolution (for example using RNAseq, DNAseq, ChipSeq, HiC…). In most cases, we expect neighboring positions to behave similarly and using this *a priori* information is a way to tackle the complexity of genome-wide analyses. It is common practice to partition each chromosome into relevant regions, because the obtained regions hopefully correspond to biological relevant or interpretable units (genes, binding site, TADs, ...) and also because the obtained partition can be used as a dimensionality reduction method and allows statistical methods to be used more easily at the region level than at the position level. In most cases, regions of interest are however unknown and should be discovered using the data. A popular way to do so is to aggregate neighboring and similar positions based on a measure of similarity between pairs of positions, in a hierarchical way that is close to the biological organization of the genome.

This article focuses on a modification of the classical hierarchical agglomerative clustering (HAC), where only adjacent clusters (according to the ordering of positions within a chromosome) can be merged. Adjacency-constrained HAC is implemented in the R package **rioja**. Our main contribution with respect to existing works is an efficient implementation of adjacency-constrained HAC in the case where the similarity between genomically distant objects can be considered as negligible. We propose an algorithm that is almost linear in time and space with respect to the number of objects to be clustered. It uses a sparse band strategy based on pre-computations of certain cumulative sums of similarities, combined with a min-heap approach to efficiently store and maintain a list of candidate merges. This algorithm is implemented in the R package **adjclust**, which is available at https://CRAN.R-project.org/package=adjclust. We provide applications to SNP and Hi-C datasets.