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

Christophe Ambroise¹, Alia Dehman², Pierre Neuvial³, Guillem Rigaill⁴ and Nathalie Vialaneix⁵

¹LaMME, Evry • ²Hyphen-stat, Toulouse ³Institut de Mathématiques de Toulouse/CNRS • ⁴IP2, CNRS/INRA • ⁵INRA MIAT

Motivation: Regionally-structured genomic data

Genome-Wide Association Studies (GWAS)

- loci: SNP
- similarity: linkage disequilibrium
- regions: LD/haplotype blocks

Chromosome contact maps (Hi-C)

- loci: binned genome positions
- similarity: contact intensity
- regions: TAD, A/B compartments

Key 1: Ward’s linkage in constant time

Distance between clusters: Ward’s linkage

\[ \delta(C, C') = \frac{S(C)}{|C|} + \frac{S(C')}{|C'|} - \frac{S(C \cup C')}{|C \cup C'|}, \quad S(C) = \sum_{(i,j) \in C} s_{ij} \]

Key 2: Storing candidate fusions in a min-heap

Min heap

- partially ordered binary tree
- nodes = candidate merges
- ordering given by the linkage \( \delta \)
- next candidate fusion is the root of the heap

Complexity

- \( O(|p|) \) in space
- \( O(p \log(p)) \) in time

Implementation

R package adjclust³

- plots of similarity, dendrogram and clustering
- wrappers for SNP or Hi-C data analyses
- model selection by broken stick⁴ or slope heuristic⁵

GWAS: inferring linkage disequilibrium blocks

Band approximation

Quality index: proportion of of approximation vs \( h \)

Scalability

Data from [⁷]

Hi-C: inferring Topologically Associated Domains

Influence of bandwidth

DI around clusters

Directionality Index (DI, [⁷]) values are expected to show a sharp variation at TADs boundaries

Data from [⁷] and [⁹]

References

³ C. Ambroise and others, Adjclust: Adjacency-Constrained Clustering of a Block-Diagonal Similarity (2018).