Combinatorial Optimization in an Alignment Problem for Mixed-Membership Clustering in Population Genetics

Xiran Liu[1], Naama M. Kopelman[2], Noah A. Rosenberg[1,3,*]
[1]Institute for Computational and Mathematical Engineering, Stanford University,
[2]Holon Institute of Technology, [3]Department of Biology, Stanford University

Motivation

- In **population genetics**, unsupervised clustering is widely used to infer **ancestry proportions** from genetic data.
- Aligning and interpreting clustering results produced by different algorithms, or even multiple runs of the same algorithm, is hard.

Challenges:

- label-switching
- arbitrariness in number of clusters
- genuine multimodality

Goals:

- a unified measure to quantify the cost of alignment
- a formalized pipeline to perform efficient alignment.

Background: Population Structure

- Membership coefficient: the proportion of the individual's genome originating in each of the K clusters (inferred populations).

Data 1: Membership coefficients of 978 human individuals inferred based on 791 loci (Fortier *et al.* 2020)

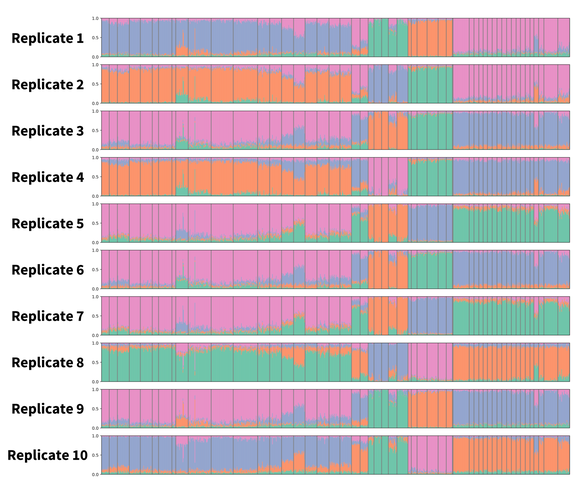

Fig 1. STRUCTURE-based ancestry inference with 4 clusters. 10 replicates are from different runs on the same individuals. Each color represents a different inferred cluster (generally associated with a prior population).

Data 2: Membership coefficients of 600 chicken individuals inferred based on 27 dinucleotide microsatellite loci (Rosenberg *et al.* 2001)

Cost of Misalignment

- Individual membership under Dirichlet model
  - In replicate 1: $p \sim \mathrm{Dir}(\boldsymbol{a})$, $\boldsymbol{a} = (a_1, \ldots, a_K)$
  - In replicate 2: $q \sim \mathrm{Dir}(\boldsymbol{\Phi}(\boldsymbol{a}))$, with permutation $\Phi(\boldsymbol{a})$.
- Contribution of an individual to the distance between two replicates, 1 and 2:

$$X = \|p - q\|_2^2$$

- Mean of X when two replicates are aligned, i.e., $\boldsymbol{a} = \Phi(\boldsymbol{a})$:

$$A_{\boldsymbol{a},\boldsymbol{a}} = \frac{4\sum_{i=1}^{K-1}\sum_{j=i+1}^{K} a_i a_j}{(a_0+1)a_0^2}$$

- Mean of X when two replicates are possibly permuted: $A_{\boldsymbol{a},\Phi(\boldsymbol{a})}$
- Mean contribution of an individual to the cost of misalignment:

$$C_{\boldsymbol{a},\Phi(\boldsymbol{a})} = \frac{A_{\boldsymbol{a},\Phi(\boldsymbol{a})} - A_{\boldsymbol{a},\boldsymbol{a}}}{2} = \frac{1}{a_0^2}\sum_{i=1}^{K} a_i(a_i - a_{\Phi(i)})$$

Demonstration on Real Data

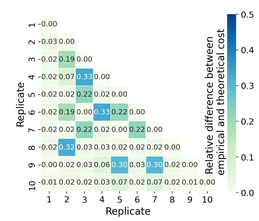

Fig 2. Relative difference between empirical and theoretical costs demonstrated on human population structure replicates.

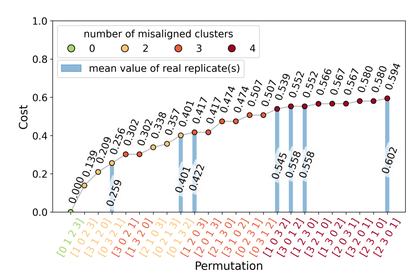

Fig 3. Theoretical misalignment cost as a function of permutations compared to the empirical cost of real replicates following the optimal permutation identified by an existing method, *CLUMPP*.

Efficient Alignment via Optimization and Network Algorithms

- Existing alignment methods, including *CLUMPP* (Jakobsson & Rosenberg 2007) and *CLUMPAK* (Kopelman *et al.* 2016), are less efficient when K is large.

Pairwise alignment:

- Membership matrix: P of size $N \times K_1$ and Q of size $N \times K_2$
- Distances between clusters from two replicates: $C(p_i, q_j)$
- Alignment indicator matrix: W, $W_{ij} \in \{0,1\}$
- Objective:

$$\mathrm{argmin}_W \sum_{i=1}^{K_1}\sum_{j=1}^{K_2} W_{ij} C_{ij}$$

→ integer linear programming (ILP) problem

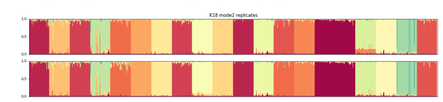

Fig 4. Inferred chicken ancestry from two clustering replicates under optimal alignment.

Mode detection:

- Node: clustering replicate
- Edge weight: membership similarity between aligned replicates
- Finding groups of replicates with high similarity amongst them → community detection problem
- Mode consensus: mean membership of replicates in the mode

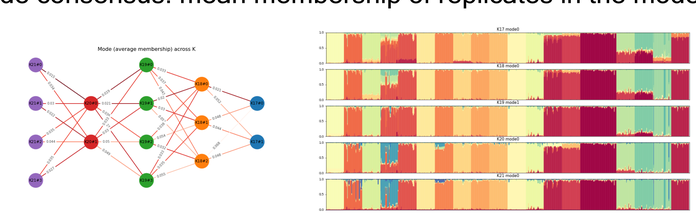

Fig 5. Alignment across mode consensus with different K: optimal alignment outcome as a multipartite graph (left) and example of consensus ancestry for aligned modes (right).

Results and Conclusions

- For fixed Dirichlet parameters, the misalignment cost increases with the Hamming distance between permutations (Fig 3).
- The misalignment cost accurately reflects the empirical cost on low variance data (Fig 2), thus can contribute to improving cluster alignment algorithms seeking to find optimal permutations of replicates.
- Clustering alignment can be performed with high efficiency and accuracy through numerical optimization and community algorithms.
- Visualization of clustering modes in multipartite graph helps interpret the genuine multimodality of the inferred population structure.

Next steps:

- Incorporate leader clustering as an optional heuristic method to achieve higher efficiency when number of replicates is large.
- Compile the alignment pipeline into a distributable package easily accessible to researchers in the population genetics community.