Implementing FastSTRUCUTRE Algorithm with Linked Loci Extension

Michael Li

# **Introduction**

The STRUCTURE algorithm, created by Pritchard et al. in 2000 is a model-based clustering algorithm for inferring population structure from multilocus genotype. This method, which was later improved upon with extensions such as the linked loci model published by Falush et al. in 2003 has made a significant impact on the biological research community. It can be used for inferring and assigning individuals to distinct populations, studying hybrid zones, identifying admixed individuals, and estimating population allele frequencies from data with a significant amount migrant or admixed samples.

FINISH THIS PART

# **Model and Algorithm**

## **Basic Algorithm**

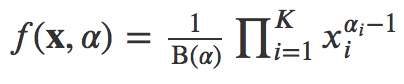
The STRUCUTRE algorithm uses the idea of MCMC (Markov Chain Monte Carlo) convergence in the form of Gibbs Sampling to sample and converge on local maximum likelihood points. STRUCTURE takes as input a collection of N samples which are each genotyped at L loci. It is assumed that the samples represent a mix of K populations, and one of the goals of the algorithm is to correctly assign the individuals to these populations.

In the most basic no-admixture model (not implemented in this project), each individual originates from one of the populations, which each has its own set of allele frequencies at each locus which is predicted by the algorithm. However, the obvious drawback to this most basic form is that in reality individuals are likely to have recent ancestors from more than one population. As an improvement, the admixture model was introduced, which assumes an individual receives some proportion of its ancestry from each population. In this model *qk(i)* represents the proportion of sample *i*’s genome that can be attributed to population *k*. The admixture model also makes it possible for an individual’s different allele copies to come from different populations, and to account for this *zl(i,a)* is introduced. *zl(i,a)* represents the ancestral population for the *a*th allele copy (in the case of diploid individuals *a* would be either 1 or 2) at locus *l* from individual *i*. An additional variable *pklj* representing the frequency of allele *j* at locus *l* in population *k* is needed in order for the sampling of *z*.

The algorithm starts off by first initializing all the *z*’s through random sampling from a uniform prior. The next step is just to iterate through the Gibbs Sampling process until completion, which I have defined as a set number of iterations. Each iteration follows this series of steps:

1. Run through all of the *z*’s and calculate *nklj*, which represents the counts of allele *j* at locus *l* in population *k*.
2. Run through all of the *z*’s and calculate *mk(i)*, which represents the counts of alleles in sample *i* from population *k*.
3. At each locus, *pklj* is sampled from the Dirichlet distribution, with the parameters (for alleles) set to λ + *nklj*.

The Dirichlet distribution is a multivariate distribution where values of the variants sum to 1, and concentration parameters αi for each variable i, where



1. For each sample *qk(i)* is sampled from the Dirichlet distribution, with the parameters (for populations) set to α + *mk(i)*.

## **Linked Loci Extension**

# **Results**

# **Discussion**

# **References**

1. Pritchard J. K., Stephens M., Donnelly P., 2000 Inference of population structure using multilocus genotype data. Genetics 155: 945–959.
2. Falush D, Stephens M, Pritchard JK. 2003 Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567–87.
3. Porras-Hurtado, Liliana et al. 2013 An overview of STRUCTURE: applications, parameter settings, and supporting software. Frontiers in genetics vol. 4 98. 29
4. Guo, Yuxin et al. 2018 Autosomal DIPs for population genetic structure and differentiation analyses of Chinese Xinjiang Kyrgyz ethnic group. Scientific reports vol. 8, 11054. 23
5. Raj, A., Stephens, M., Pritchard, J.K. 2014 fastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics 197: 573–589