Michael Li
October 30, 2018

CS 4775 Project Proposal

Problem:

My plan for the final project is to reimplement the STRUCTURE algorithm for dividing a set of genotypes into separate populations. STRUCTURE, whose basic algorithm was first described by Pritchard, Stevens, and Donnelly[1] is one of the most widely used tools by researchers for population analysis. It uses Bayesian iterative methods and Markov Chain Monte Carlo (MCMC) estimation to place samples into groups based on similarity. Then, if I still have time, I will try to alter the method to account for correlations between linked loci described by Falush et al[2], which would lead to advantages such as allowing for more accurate estimation of statistical uncertainty and detection of older admixture events.

Algorithms/Models:

- MCMC
- Gibbs Sampling
- Linked Loci

Implementation:

I plan to iterate through the algorithm by forming a Gibbs sampler (sampling from Dirichlet Distribution) for allele frequencies and admixture proportions, and then sampling the ancestral population of each allele from this data. This will be repeated many times (~100,000 iterations), with MCMC thinning included to account for possible dependence between adjacent iterations. The user will be able to identify k, the number of populations that they want the algorithm to separate the samples into. I plan on testing my algorithm by using the original data set used in the example analyses from the Pritchard, Stephens, and Donnelly article. When completed I would then play around with parameters and implement linked loci in order to achieve the best results.

---

[1] Pritchard JK, Stephens M, Donnelly P; Genetics. 2000 Jun; 155(2):945-59.

[2] Daniel Falush, Matthew Stephens and Jonathan K. Pritchard; GENETICS *August 1, 2003 vol. 164 no. 4 1567-1587*

Additionally:
Porrashurtado L, et al. An overview ofSTRUCTURE: applications, parameter settings, and supporting software. Frontiers in Genetics. 2013; 4-98.