Abstract for:

Approaches For Inferring Past Population Size Changes From Genome-wide Genetic Data

From Christoph Theunert.

The history of populations or species is of fundamental importance in a variety of areas. Gaining details about demographic, cultural, climatic, or political aspects of the past may provide insights that improve the understanding of how populations have evolved over time and how they may evolve in future. Different types of resources can be informative about different periods of time.

One especially important resource is genetic data, either from a single individual or a group of organisms. Since genetic material gets passed on from generation to generation, traces of past events can still be detected in today's genetic data. Using computational and population genetics methods, sequences from whole genomes can be scanned for traces of such events and, therefore, assist in new interpretations of historical details of populations or groups of interest.

This thesis focuses on the detection and interpretation of past population size changes. Two approaches to infer particular parameters from underlying demographic models are described. The first part of this thesis introduces two summary statistics which were designed to detect fluctuations in size from genome-wide Single Nucleotide Polymorphism (SNP) data. Demographic inferences from such data are inherently complicated due to recombination and ascertainment bias. Hence, two new statistics are introduced: allele frequency-identity by descent (AF-IBD) and allele frequency-identity by state (AF-IBS). Both make use of linkage disequilibrium information and exhibit defined relationships to the time of the underlying mathematical process. A fast and efficient Approximate Bayesian Computation framework based on AF-IBD and AF-IBS is constructed that can accurately estimate demographic parameters. These two statistics were tested for the biasing effects of hidden recombination events, ascertainment bias, and phasing errors. The statistics were found to be robust to a variety of these tested biases. The inference approach was then applied to genome-wide SNP data to infer the demographic histories of two human populations.

The second part of this thesis introduces a new way of summarizing information from the site frequency spectrum. Commonly applied site frequency spectrum based inference methods make use of allele frequency information from individual segregating sites. Our newly developed method, the 2 point spectrum, summarizes allele frequency information from all possible pairs of segregating sites, thereby increasing the number of potentially informative values from the same underlying data set. These additional information are then incorporated into a Markov Chain Monte Carlo framework. This allows for a high degree of flexibility and implements an efficient method to infer population size trajectories over time. We tested the method on a variety of different simulated data sets from underlying demographic models. Results indicate that this non-parametric 2 point spectrum method can accurately infer the extent and times of past population size changes and, therefore, correctly estimates the history of temporal size fluctuations. Furthermore, the initial results suggest that the amount of required data and the accuracy of the final results are comparable with other publicly available non-parametric methods. An easy to use command line program was implemented and will be made publicly available.