Bayesian Learning of Haplotype Block Variation

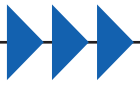
Greenspan, G. and Geiger, D.
Department of Computer Science, Technion

Observable haplotype blocks arise from the interaction between recombination hot-spots, bottleneck effects and genetic drift. The presence of recombination hot-spots in human chromosomes has been demonstrated by several recent high-resolution studies of SNP covariation. They separate between stretches of up to 100,000 base pairs in which almost no recombination takes place, so the SNPs lying between hot-spots act a single multi-site allele or `haplotype block'. A bottleneck occurs when a locally-reproducing population is descended from a small group of individuals, for example due to migration. As the new population grows, it will exhibit far less genetic variation within each block than expected for its size. These small populations also undergo significant genetic drift, in which the variation is decreased further by many generations of random mating.

An accurate statistical model of the haplotype blocks present in a chromosomal region can be used to strengthen the power of genetic association analysis, improve the accuracy of general haplotype resolution and further our understanding of the recombination process itself. Empirical studies of populations descended from a bottleneck, confirmed by our simulations, show that the haplotype blocks in a chromosomal region can be modelled by a dynamic Bayesian network. Each hidden variable corresponds to the ancestral source of a haplotype block, with first-order Markovian transition probabilities reflecting the recombination which has occurred at the hot-spots in the intervening generations. Each SNP observed in an individual chromosome depends upon the ancestral block from which it is descended, under a suitable mutation model.

We have developed a general tool which learns this dynamic Bayesian network model from raw SNP data. The problem differs from classical Markov model training in several ways. The location of hot-spots is not given, requiring a selection between 2^(loci-1) possible network topologies. The values for each hidden state must also be inferred, a difficulty compounded by the presence of failed measurements and the fact that only joint SNP measurements from pairs of chromosomes are often available. Utilizing an ML (maximum likelihood) approach leads to over-fitting, producing a model in which there are no recombination hot-spots and too many ancestral haplotypes. So we adopt the MDL (minimum description length) criterion, which seeks to minimize the number of bits required to represent data D with a model M, given by DL(M)-log2(Pr(D|M)).

Starting with no hot-spots, our search strategy iterates over possible hot-spot insertions (or de-



letions and nudges in later rounds), trying only those operations which improve our score more than their neighbors, repeating until no further improvement can be found. For a particular assignment of hot-spots, the haplotype blocks for each subject are obtained via a hierarchical EM procedure, which handles both joint unphased and failed measurements. The transition probabilities between the discovered blocks are inferred by EM, then block values are iteratively eliminated to further improve the DL score. Tests on both simulated and real-world data demonstrate our method's ability to recover the haplotype block distribution of a chromosomal region from phased or unphased samples. Our algorithm is guaranteed to converge and takes O(loci ^ 2*samples) time.

Future work will focus on improving the decisions made in the block value elimination stage, to deal with data from older populations in which many mutations have taken place. Avenues being explored include modifying the EM iterative procedure for MDL instead of ML, model-based cluster analysis, and phylogenetic tree pruning. An accurate choice of ancestor blocks will also allow our method to estimate site-specific mutation rates from the data observed.