Title: Efficient Haplotype Inference on Pedigrees and Applications

Abstract: We study haplotype reconstruction under the Mendelian law of inheritance and the minimum recombination principle on pedigree data. We prove that the problem of finding a minimum-recombinant haplotype configuration (MRHC) is in general NP-hard. This is the first complexity result concerning the problem to our knowledge. Later on we show that MRHC is also NP-hard for tree pedigrees. An iterative algorithm based on blocks of consecutive resolved marker loci (called block-extension) is proposed. It is very efficient and accurate for data sets requiring few recombinants. A polynomial-time exact algorithm for haplotype reconstruction without recombinants is also presented. This algorithm first identifies all the necessary constraints based on the Mendelian law and the zero recombinant assumption, and represents them using a system of linear equations over the cyclic group Z_2. By using a simple method based on Gaussian elimination, we could obtain all possible feasible haplotype configurations. For genotypes with missing alleles, we develop an effective integer linear programming (ILP) formulation of the MRHC problem and a branch-and-bound strategy that utilizes a partial order relationship (and some other special relationships) among variables to decide the branching order. The partial order relationship is discovered in the preprocessing of constraints by considering unique properties in our ILP formulation. Non-trivial (lower and upper) bounds on the optimal number of recombinants are introduced at each branching node to effectively prune the search tree. When multiple solutions exist, a best haplotype configuration is selected based on a maximum likelihood approach. The ILP algorithm works for any pedigree structures, regardless of the number of recombinants, and effective for any practical size problems. We have implemented the above algorithms in a software package called PedPhase and tested them on simulated data sets as well as on a real data set. The results show that the algorithms perform very well. For example, our results of ILP algorithm on simulated data show that the algorithm could recover haplotypes with 50 loci from a pedigree of size 29 in seconds on a standard PC. Its accuracy is more than 99.8% for data with no missing alleles and 98.3% for data with 20% missing alleles in terms of correctly recovered phase information at each marker locus.

Haplotype information is much valuable for disease gene association mapping, which is a very important problem in biomedical research. We also develop a new algorithmic method for haplotype mapping of case-control data based on a density-based clustering algorithm, and propose a new haplotype (dis)similarity measure. The mapping regards haplotype segments as data points in a high dimensional space. Clusters are then identified using a density-based clustering algorithm. *Z*-score based on the numbers of cases and controls in a cluster can be used as an indicator of the degree of association between the cluster and the disease under study. Preliminary experimental results on an independent simulated data set, and on a real data set with the known disease gene location show that our method could predict the gene location with high accuracy, even when the rate of phenocopies is high.