MIP models for parsimonious haplotype estimations

Daniele Catanzaro (ULB - GOM)

March 26, 2010

Haplotyping estimation from aligned Single Nucleotide Polymorphism (SNP) fragments has attracted more and more attention in the recent years due to its importance in analysis of many fine-scale genetic data. Its application fields range from mapping of complex disease genes to inferring population histories, passing through designing drugs, functional genomics and pharmacogenetics. The literature proposes a number of estimation criteria to select a set of haplotypes among possibles alternatives. Usually, such criteria can be expressed under the form of objective functions, and the sets of haplotypes that optimize them are referred to as optimal. One of the most important estimation criteria is the pure parsimony which states that the optimal set of haplotypes for a given set of genotypes is the one having minimal cardinality. Finding the minimal number of haplotypes necessary to explain a given set of genotypes involves solving an optimization problem, called the Pure Parsimony Haplotyping (PPH) estimation problem, which is notoriously NP-Hard. In this talk we present an integer programming model for HLA association studies based on the parsimony criterion. The model is simple, compact, easy to implement and able to handle datasets containing up to 200 phenotypes. Computational experiments carried out on patients affected by psoriasis and severe alopecia areata show that the model is consistent with the experimental haplotype frequencies, showing, for the considered diseases at least, a high reliability of the predictions. These promising results give perspective on computer-aided association studies and encourage the development of efficient exact computational approaches for haplotype estimation.