

A Dynamic Programming Algorithm for Haplotype Block Partitioning and Its Application in Association Studies

Kui Zhang, Ph.D.
University of South California

Abstract:

Recent studies showed a haplotype block structure for the human genome such that it can be divided into discrete blocks of limited haplotype diversity. A small fraction of SNPs (tag SNPs) can be used to distinguish a large fraction of the haplotypes in each block. These tag SNPs can be extremely useful for association studies in that it may not be necessary to genotype all the SNPs. We develop a dynamic programming algorithm to partition the haplotypes into blocks. The algorithm is guaranteed to find the haplotype blocks with the minimum number of tag SNPs required to account for most of the common haplotypes in each block. We apply this algorithm to the chromosome 21 data of Patil et al. [Science 294, 1719-1723]. Using the same criteria as in Patil et al., we identify a total of 3,582 tag SNPs and 2,575 blocks that are 21.5% and 37.7% smaller, respectively, than those identified using a greedy algorithm of Patil et al. When the tag SNPs instead of all the SNPs are used to reduce the genotyping effort in association studies, an important question is how much power is lost. We develop the following simulation strategy to quantitatively assess the power loss. First, case-parental or case-control samples are generated based on a disease model. Second, a small fraction of samples are selected to determine the haplotype blocks and the tag SNPs by our dynamic programming algorithm. Third, the statistical power of tests is evaluated based on three kinds of data: (1) all of the SNPs and the corresponding haplotypes; (2) the tag SNPs and the corresponding haplotypes; (3) the same number of randomly chosen SNPs as the number of tag SNPs and the corresponding haplotypes. We study the power of different association tests with a variety of disease models and block partitioning criteria. Our study indicates that the genotyping efforts can be significantly reduced by the tag SNPs without much loss of power. Depending on the specific block partitioning algorithm and the disease model, on average, when the identified tag SNPs are only 25% of all the SNPs, the power is reduced by only 9%. Linear Models for Controlling Background Signal in Microarray Gene Expression Studies.