SNP-set Detection and Association Test with Hamming Distance Information

Charlotte Wang

Institute of Epidemiology and Preventive Medicine, National Taiwan University

Two major challenges have arisen in recent genetic association studies. One is the need to develop methods to reduce the intractably large numbers of genetic variants in genomic data to a more computationally manageable number, and the other is to find ways to increase the power of statistical tests used in association studies. Tackling these problems with a SNP-set analysis can be an efficient solution. However, most current association studies constructed possible marker-sets based on tests of pre-specified SNP-sets or on tests through sliding windows for whole genome. No combined procedures for defining SNP-sets, followed by association analysis between the SNP-sets and the disease of interest, have been proposed.

To construct SNP-sets, we proposed a Hamming distance-based clustering algorithm (HD-Cluster), which employs Hamming distance to measure the dissimilarity between strings of SNP genotypes and evaluates whether the given SNPs should be clustered. With the SNP sets obtained, a Hamming distance-based association test (HDAT) was developed to examine susceptibility to the disease of interest. This test assesses whether the similarity in genotypes between a diseased and a normal individual differs from the similarity between two individuals with the same disease status.

The statistical properties of the proposed methods are discussed, and illustrated with simulations and applications. In simulation studies, the results showed that the HD-Cluster can identify correct clustering patterns and is also an efficient algorithm. This method can be applied not only to genetic data, but also to categorical data in general. For the proposed test, the HDAT works well regardless of the sample size, effects of SNPs within the given set, and the signal-to-noise ratio (proportion of the number of disease-associated SNPs to the number of neutral SNPs). Moreover, for genotyping data of coronary artery disease (CAD) from the WTCCC, our proposed methods successfully identified one SNP-set containing four SNPs that have been reported in literatures to associate with the disease.

To conclude, the proposed clustering algorithm and association test have demonstrated reliable and satisfactory performance. In our proposed methodology, no inference of haplotypes is needed, and SNPs under consideration do not need to be linked. In addition, this test works well for a SNP-set containing both SNPs with a deleterious effect and those with a protective effect, and for a set containing many neutral SNPs.

Keywords: association test, clustering analysis, dendrogram, Hamming distance, similarity, SNP set.