Evaluation an association stud and spatial clustering methods applied to DNA data

Makoto Tomita*
Tokyo Medical and Dental University, Tokyo, Japan - mtomita@ism.ac.jp

Abstract

In statistical genetics, SNPs, haplotype and diplotype are the major polymorphic markers in DNA data. Therefore, in recent years, the relative frequencies of different haplotypes have been estimated, their structures have been examined, and haplotypes have been used for association studies. We evaluated haplotype analyses, e.g., identification of haplotype blocks and haplotype association studies.

Ke words: SNPs data, DNA marker, maximum likelihood estimation, association study, haplotype.

1. Introduction

Single nucleotide polymorphisms (SNPs) are the most abundant form of genetic variation. Almost all SNPs are bi-allelic, and typically one allele is present in the majority of the chromosomes of a population, and the alternative variant (i.e., the minor allele) is present with less frequency. SNPs are promising tools for mapping susceptibility mutations that contribute to complex diseases. Most SNPs can be used as surrogate markers for positional cloning of genetic loci, because of the allelic association which is well-known as linkage disequilibrium (LD).

In early research, linkage disequilibrium analysis for SNP data is particularly important. Methods of trait mapping based on theories of linkage disequilibrium analysis have been developing quickly in recent years. In DNA sequences, domain “hotspots” exist at which recombinations have occurred briskly. Conversely, large domains with infrequent recombinations in which linkage disequilibrium is maintained also exist. Such domain called a “haplotype block” or “LD block”. Although the value of D' represents one of the disequilibrium parameters important for identifying haplotype blocks.

On the other hand, recently, the association has been actively studied between genotype and phenotype in post genomic research. Here ‘genotype’ means not only genotype itself but also haplotypes and diplotype configurations that are estimated from genotypes of the sample, and ‘phenotype’ indicates qualitative or quantitative variables which may be related to some specific diseases. Quantitative phenotype variable, called QTL (quantitative trait locus), includes covariate such as BMI and glucose level.

Some algorithms have been proposed so far to analyze the association between the genotype information and the quantitative phenotypic QTL. The algorithm $QTLhaplo$ (Shibata, et al., 2004) deals with the association between the genotype and the univariate phenotype, assuming the normality of the conditional distribution of the phenotype given the genotype information. The likelihood is calculated on the basis of frequencies of diplotype configurations (joint probability of the frequencies of the haplotypes that compose the dipolotype) and the density function of a normal distribution. The algorithm $QTLmarc$ (Kamitsuji & Kamatani, 2006) has been proposed for multivariate analysis of multiple quantitative responses, however, it can deal with only the case where each subject’s haplotype is determined uniquely from its genotype. It is doubtful whether it can evaluate the association properly in general cases. Therefore, it is valuable to develop a general method of association analysis for multivariate quantitative responses. We extend the algorithm $QTLhaplo$ so that it can deal with the association between the genotype and multiple quantitative variables assuming three types of models, i.e., the dominant, recessive and additive models.

2. Concluding remarks

We consider the case where there are three loci with two kinds of alleles as genotypes and two quantitative phenotype variables. As a genotype input data set, an actual data set for 44 subjects was downloaded from the Hapmap project, and the information of the region (107,189 loci) of the X chromosome was used. Among large number of loci, 40 loci with linkage disequilibrium were selected by Tomita, et al. (2008), where they studied this area of the X chromosome from Hapmap project. Note that we do not have any information on haplotype as the data set does not contain the phase information and that there is no missing observation.

So far we discussed the comparison of the powers between our method and QTLmarc. In actual data analysis, however, the objective is to find out if there exists any haplotype which is closely related to the phenotypes assuming an appropriate one among the dominant, recessive and additive models. For this purpose we may operationally choose the model with the highest significance. Note that, if we wish to use the AIC statistics we can compute them up to an additive constant based on the values of likelihood ratio statistics, because the log-likelihood statistics for the null models are common among the dominant, recessive and additive models.

There are two advantages of our method compared to the $QTLmarc$ algorithm. One is that our method can treat genotype data with stochastically determined diplotypes and the other is that we can assume any model among dominant, recessive and additive models. It is expected that our method will be useful in association studies of complex diseases such as schizophrenia and autism, where the causes of the diseases are not yet resolved and there exist multiple candidate responses.

Acknowledgment
This work was partly supported by KAKENHI (Grant-in-Aid for Scientific Research).

References

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B. et al. (2002) The structure of haplotype blocks in the human genome, Science, 296:2225-2229.

Kamatani, N., Sekine, A., Kitamoto, T., Iida, A., Saito, S., Kogame, A. et al. (2004) Large-scale single-nucleotide polymorphism (SNP) and haplotype analyses, using dense SNP Maps, of 199 drug-related genes in 752 subjects: the analysis of the association between uncommon SNPs within haplotype blocks and the haplotypes constructed with haplotype-tagging SNPs, American Journal of Human Genetics, 75:190-203.

Kamitsuji, S. & Kamatani, N. (2006) Estimation of haplotype associated with several quantitative phenotypes based on maximization of are under a receiver operating characteristic (ROC) curve, Journal of Human Genetics, 51:314-325.

Shibata, K., Ito, T., Kitamura, Y., Iwaaki, N., Tanaka, H. & Kamatani, N. (2004) Simultaneous estimation of haplotype frequencies and quantitative trait parameters: applications to the test of association between phenotype and diplotype configuration, Genetics, 168:525-539.

Tomita, M., Hatsumichi, M. & Kurihara, K. (2008) Identify LD Blocks Based on Hierarchical Spatial Data, Computational Statistics & Data Analysis, 52:1806-1820.

Tomita, M., Kurihara, K. & Moon, S. H. (2012) An Application to Select Tag Loci by Using Hierarchical Structures of DNA Markers, Journal of the Korean Data Analysis Society, 13:2749-2762.

Tomita, M., Hashimoto, N. & Tanaka, Y. (2011) Association Study for the Relationship Between a Haplotype or Haplotype Set and Multiple Quantitative Responses, Computational Statistics & Data Analysis, 55:2104-2113.