Editorial

Taesung Park

Department of Statistics, Seoul National University,
Gwanak_1 Gwanak-ro, Gwanak-gu, Seoul, Korea 151-747
E-mail: taesungp@gmail.com

Jurg Ott

Key Laboratory of Mental Health,
Institute of Psychology, Chinese Academy of Sciences,
4A Datun Road, Beijing 100101, China
E-mail: ottjurg@psych.ac.cn

Biographical notes: Taesung Park received his BS and MS in Statistics from Seoul National University, Korea, in 1984 and 1986, respectively, and PhD in Biostatistics from the University of Michigan in 1990. He served as chair of the Bioinformatics Program, Seoul National University from April 2005 to March 2008. He is a Professor at the Department of Statistics, Seoul National University, Korea. He is the Director for the National Research Laboratory of Bioinformatics and Biostatistics. His research areas include microarray data analysis, genome-wide association studies, statistical genetics, missing data analysis and longitudinal data analysis.

Jurg Ott received his PhD Degree in Zoology from the University of Zurich, Switzerland, in 1967. He is a visiting professor at the Institute of Psychology, Chinese Academy of Sciences, China, and Professor Emeritus at the Rockefeller University, USA.

Data mining for high throughput data from genome-wide association studies.

This special issue of the International Journal of Data Mining and Bioinformatics is comprised of papers that are extended versions of papers selected at the workshop on 'Data mining for high throughput data from genome-wide association studies', jointly held at the IEEE International Conference on Bioinformatics & Biomedicine (BIBM) in 2010. The focus of the workshop was to introduce new machine learning approaches to high throughput data from genome-wide association studies (GWAS).

GWAS have become a popular strategy to discover genetic factors such as single nucleotide polymorphism (SNP) affecting common complex diseases. Many GWAS have successfully identified genetic risk factors associated with common diseases, and have achieved substantial success in unveiling genomic regions responsible for the various aspects of phenotypes.

However, identifying the underlying mechanism of disease-susceptible loci has proven to be difficult due to the polygenic and multiple-pathway nature of complex diseases. The newly identified genes from GWAS only explain a small portion of the

genetic factors in complex diseases. This rather limited finding is partly ascribed to the lack of multiple SNP-based analyses.

In this context, the analysis strategy for GWAS has been stepped up from the single SNP approach towards a multiple SNPs approach for understanding the complexity of genotype–phenotype association considering gene–gene and gene–environment interaction.

However, multiple SNPs-based analyses are complicated owing to computational burden, the large number of comparisons implied by even bivariate analysis, and a limited set of appropriate analysis tools.

Considering the current investment in GWAS, with large sample sizes and high costs, a more complete examination of the resulting data is warranted by using multiple SNPs-based analyses. Although multiple SNPs-based analyses are generally recognised as one solution to discover additional genetic factors and understand complex genetic components affecting disease susceptibility, several issues remain to be further investigated.

This special issue discusses the most challenging issues in multiple SNPs approaches, including gene–gene interaction, and introduces statistical and computational methods for data mining and machine learning for revealing hidden association networks of genotype–phenotype relationships.

The first paper, by Shen et al., proposed a data mining approach using support vector machines with L1 penalty for detecting gene–gene interactions. The second paper, by Kwon et al., proposed an efficient program called cuGWAM for performing multifactor dimensionality reduction (MDR) analysis for GWAS. cuGWAM uses a CUDA-enabled high-performance graphics processing unit for detecting gene–gene interactions.

The third paper, by Park et al., proposed a data mining approach to the discovery of multivariate phenotypes using association rule mining for GWAS. The fourth paper, by Sun and Ott, introduced multilocus association analysis under polygenic models. The fifth paper, by Lee et al., proposed a two-step approach for detecting multiple loci associated with the traits. The sixth paper, by Ahn et al., introduced an approach to selecting multiple loci associated with the drug responses. Finally, Han et al. demonstrated the effect of sample size on genome-wide population differentiation studies based on the SNPs.

The seven papers in this volume provide scientists with an overview on the recent advancements in multiple SNPs analyses in GWAS in the field of statistical genetics and bioinformatics. We hope the papers can encourage researchers towards a more extensive use of statistical genetics and bioinformatics techniques for research in biology and medical sciences.