

Research Statement

Kui Zhang

My current research interests focus on developing mathematical, statistical and computational methods to solve scientific problems in genetics and molecular biology. In the following, I will first summarize my research according to different research areas and then outline my research plan in the next few years.

Summary

Mathematical, statistical and computational problems in human genetics

Identifying the genetic variants that are predisposing to complex diseases or traits, such as hypertension, heart disease and diabetes, etc., is one of the most important problems in Human Genetics. Traditional linkage analysis methods based on family data are successful in mapping genetics mutations for rare diseases. When the disease gene and marker loci are very close, for example, within 1cM, the linkage usually could not be detected. So it is not applicable to fine mapping disease genes. For this reason, using linkage disequilibrium (also referred as allelic association) methods to map complex disease genes have great of interest recently. Among of them, family-based association methods have great promises in mapping genes of complex traits. My colleagues and I have developed statistical methods to test linkage and association simultaneously for quantitative traits in general pedigrees and assess the power of transmission disequilibrium tests for quantitative traits. Simulation studies have shown that our methods have better power than existing methods for many cases considered.

Linkage disequilibrium is the key to fine mapping the complex disease genes. It is also known that the haplotype approaches for fine mapping complex disease genes are generally powerful than single-marker based methods. With the completion of human genome and availability of a dense genome-wide map of SNPs markers, it is possible to study the linkage disequilibrium and haplotype patterns in the whole genome. As an initial step, we have developed a method to identify all compatible haplotypes in a general pedigree using genotype data at a set of tightly linked SNPs. This method had been implemented in the software to distribute to the scientific community. The simulation studies have shown that our method outperforms other available methods.

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 have developed 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. With limited resources, the investigators would like to restrict the number of tag SNPs used in their study. We first formulated this problem as finding a block partition with a fixed number of tag SNPs that can cover the maximal percentage of a genome. Then we have developed two dynamic programming algorithms

to solve this problem, which are fairly flexible to take into account the knowledge of functional polymorphisms.

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 have proposed a simulation strategy to quantitatively assess the power loss. 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. We also used the similar simulation method to assess the power loss using tag SNPs in quantitative locus (QTL) mapping. It surprised us greatly to see that the results are substantially different.

In linkage studies, individual genome scans generally have low power to detect QTL and provide imprecise estimates of their location and effect, especially when the effect is small. As a consequence, follow-up for fine mapping and positional cloning is problematic. Although multiple studies of the same QTL have been conducted, investigators often evaluate scans other than their own when deciding which regions merit further investigation, but they have limited options for formally integrating the data. We have developed an empirical Bayes analytic method to integrate information from multiple genome scans. The simulation results indicate that the empirical Bayes method can account for between-study heterogeneity, estimate the QTL location and effect more precisely, and provide narrower confidence intervals than results from any single individual study.

Functional genomics

With the completion of the human genome project and the rapid development of high through technologies, such as microarrays, yeast-two-hybrid systems, large-scale deletion experiments, etc., a large amount of biological data has been generated and certainly more data will be generated in the near future. Extracted information from these data can provide some insight into the function of genes and better understanding of the genetic networks underlying complex biological processes. We have developed a simulation-based approach to assessing the reliability of gene clusters identified from different clustering algorithms. The method has been implemented in the software to distribute to the scientific community. My colleagues and I have developed a method to predict protein functions based on the protein interaction network. We are in the process of developing model-based approaches to identify biologically meaningful gene clusters using gene expression data incorporating existing biological information for genes. My colleagues and I are working on a project to assess the relationship between the sequence similarity and the expression similarity of genes using the microarray data downloaded from the public database.

Applied research projects

Other than the methodology work in statistical genetics, I have been actively involved in several applied projects in the last few years. The aims of these projects are to identify the genetic components of phenotypes using SNP data and/or microarray data. As a statistical geneticist, I have collaborated with scientists in biomedical and biological fields through three ways. First, I have

been actively involved as a co-investigator in their grant preparation. The duties include setting appropriate study designs, calculating the sample size and power, adapting the available methods and developing the new methods for data analysis, and writing up the statistical analysis in the grant proposal. Second, we have provided the statistical support for their data analysis using the available methods. We have analyzed the gene expression data identified several genes that are associated with poor survival in Glioblastoma Multiforme, which will be published in the Journal of Neuro-Oncology. We have also analyzed the SNP data and identified several genes that are associated with some phenotypes (e.g., lipid responses to fenofibrate, congenital cytomegalovirus related hearing loss etc.). These results are presented in several papers and were submitted for publication. Currently, I am also working with two collaborators to identify the haplotype blocks and tag SNPs from complex disease studies. Third, we have developed new methods for their data analysis. Currently, we are working on the haplotype inference for the killer cell immunoglobulin-like receptor (KIR) genes. The majority of KIR genes are detected as either present or absent using locus-specific genotyping technology, which leaves missing data: whether a detected gene is present on one or both chromosomes remains unknown. Thus, the performance of methods for haplotype inference (e.g., EM, PHASE, etc) for KIR genes may be compromised. In the other hand, many known haplotypes and sub-haplotypes have been identified. To accommodate these, we have developed an EM-based method for haplotype inference by incorporating previously identified haplotypes and/or sub-haplotypes. Our simulation results showed that our method appears more useful than available methods in haplotype inference for KIR genes.

Future research plan

My research interests will still focus on developing mathematical, statistical and computational methods to help solve the scientific problems in genetics and molecular biology in the near future. With the completion of the sequences from the human genome and many other organisms, the advanced high throughput technologies for genes and proteins, such as microarray, yeast-two-hybrid systems, etc., new mathematical, statistical and computational methods are needed to extract useful information from the very large amount of data to be generated in the coming exciting years.

Fine mapping complex disease genes using linkage disequilibrium has been my research interest for the past few years, I will put a large portion of my effort on this field through a funded NIH/NIGMS grant R01-GM74913: haplotype analysis in linkage disequilibrium mapping. Specifically, (1) we will develop efficient algorithms to estimate haplotype frequencies and determine haplotype configurations for individuals in general pedigrees with a large number of tightly linked genetic markers, which is a natural extension of our previous work. (2) We will develop new statistics based on haplotype sharing that can combine information from multiple regions of interest and will design efficient algorithms to compute new statistics. We will investigate their properties using empirical and simulated data sets and will develop novel statistical methods based on them for mapping complex human genes. (3) Many methods for haplotype block partitioning and tag SNP selection have been proposed but their performances have not been extensively and carefully evaluated. We will evaluate the power using tag SNPs and compare it with the power of randomly selected SNPs and evenly distributed SNPs using simulated data sets as well as publicly available data sets. In this context we will compare different methods for tag SNP selection and investigate the effect of several critical issues in designing efficient and effective algorithms for tag SNP selection. (4) We will implement and validate the developed

methods in user-friendly software with web-based user interfaces and distribute the program via the Internet to the scientific community. At the same time, the analysis of genome-wide association data poses daunting challenges to scientists. We will explore the methodology developments and hope to make some progress in this area.

I will also pursue my research on functional genomics, especially on mining useful information from large amount of gene expression data and other data generated by advanced biology technologies. Many existing mathematical and statistical methods have been applied to extract information from such data. However, much useful information is ignored in many current implementations and most of them have not combined the different data sources. The achievements obtained from such data are far from the potential of these advanced technologies. We are developing more efficient methods to extract information from such data and hope to make some progress in this area.

Collaborating with people having very diverse background and constantly learning new things in other fields is one of the most attractive aspects of being a statistician. I have always enjoyed working with scientists in other fields, understanding real scientific questions, and developing mathematical and statistical tools to solve these problems. I will continue doing so in the future.