Chapter 7.19
A Haplotype Analysis System for Genes Discovery of Common Diseases

Takashi Kido
HuBit Genomix, Inc., Japan

ABSTRACT

This chapter introduces computational methods for detecting complex disease loci with haplotype analysis. It argues that the haplotype analysis, which plays a major role in the study of population genetics, can be computationally modeled and systematically implemented as a means for detecting causative genes of complex diseases. In this chapter, the author provides a review of issues on haplotype analysis and proposes the analysis system which integrates a comprehensive spectrum of functions on haplotype analysis for supporting disease association studies. The explanation of the system and some real examples of the haplotype analysis will not only provide researchers with better understanding of current theory and practice of genetic association studies, but also present a computational perspective on the gene discovery research for the common diseases.

INTRODUCTION

In recent years, much attention has been focusing on finding causative genes for common diseases in human genetics. (Badano & Katsanis, 2002; Daly, 2001; Fan & Knapp, 2003; Gabriel et al., 2002) These findings of causative genes of common diseases including diabetes, hypertension, heart disease, cancer, and mental illness are expected to be opening doors for realizing new diagnoses and drug discoveries. A promising approach for the gene discovery on common diseases is to statistically examine genetic association between the risk of common diseases and DNA variations in human populations. While single nucleotide polymorphisms (SNPs), the most common genetic variation, are widely used for this genetic association study, haplotypes, the combination of closely linked SNPs on a chromosome, has been shown to have pivotal roles in the study of the genetic basis of disease (Clark, 2004; Niu, 2004; Schaid, 2004). The main purpose of this report is to provide a comprehensive review of

haplotype analysis in genetic association studies on complex, common diseases and provide the computational framework which enables us to carry out successful high-throughput genome-wide association study.

In addition to the review of the recent developments of haplotype analysis, the author presents the design, implementation, and application of a haplotype analysis system for supporting genome-wide association study. While there are some useful tools or programs available for haplotype analysis (Kitamura et al., 2002; Niu, Zin, Zu, & Liu, 2002; Sham & Curtis, 1995; Stephens, Smith, & Donnelly, 2001), little work has been reported for a comprehensive analysis pipleline for large-scale and high-throughput SNPs screening which fully integrate these functions. HAPSCORE (Zhang, Rowe, Struewing, & Buetow, 2002) is one of the few examples of those pipeline systems; however, it does not include some analysis functions such as automatic linkage disequilibrium (LD) block partitioning and disease association analysis tools. In this report, the author presents a system, LDMiner (Higashi et al., 2003), which represents the pioneer pipeline system that integrates a comprehensive spectrum of functions related to haplotype analysis. This report introduces the details of LDMiner and shows some examples of haplotype analysis with LDMiner, which helps to explain the theory and practice on population-based association study for common diseases.

BACKGROUND

Genetic Variations and Common Diseases

The progress on human genome science is opening doors for the discovery of new diagnostics, preventive strategies, and drug therapies for common complex diseases including diabetes, hypertension, heart disease, cancer, and mental illness. Analysis of human genome primarily focuses on variations in the human DNA sequence, since these differences can affect the potential risk of disease outbreaks or the effectiveness of a drug treatment of the diseases.

A common method for determining the genetic differences between individuals is to find single nucleotide polymorphisms (SNPs). A SNP is defined as a DNA sequence variation referring to an alteration of a single nucleotide (A, T, G, C). SNPs represent the most common genetic variations. In fact, there are millions of SNPs in the human genome (Kruglyak & Nickerson, 2001), and it is estimated that there will be on average one SNP every 1,000 base pairs. SNPs are caused when nucleotides replicate imperfectly or mutate. Although most of these SNPs have no ostentatious impacts on the survival of the species, certain SNPs may confer beneficial effects allowing species to evolve and to adapt to new environments more successfully, while certain others may be detrimental. These SNPs are passed on from generation to generation. After hundreds of years, some SNPs become established in the population.

A number of instances are known for which a particular nucleotide at a SNP locus (i.e., a particular SNP allele) is associated with an individual's propensity to develop a disease. For example, it has been reported that functional SNPs in the lymphotoxin-alpha gene were associated with susceptibility to myocardial infection by means of a large-scale case-control association study using 92,788 SNP markers (Ozaki et al., 2002). There have also been a number of reports that show some SNPs in certain genes can determine whether a drug can treat a disease more effectively in individual with certain genotypes compared to those who do not carry such SNPs. For example, Cummins et al. (2004) reported that there is a strong association in Han Chinese between a genetic marker, the human leukocyte antigen HLA-B*1502, and Stevens-Johnson syndrome induced by carbamazepine, a drug commonly prescribed