Data Mining in Genome Wide Association Studies

Tom Burr
Los Alamos National Laboratory, USA

INTRODUCTION

The genetic basis for some human diseases, in which one or a few genome regions increase the probability of acquiring the disease, is fairly well understood. For example, the risk for cystic fibrosis is linked to particular genomic regions. Identifying the genetic basis of more common diseases such as diabetes has proven to be more difficult, because many genome regions apparently are involved, and genetic effects are thought to depend in unknown ways on other factors, called covariates, such as diet and other environmental factors (Goldstein and Cavalleri, 2005).

Genome-wide association studies (GWAS) aim to discover the genetic basis for a given disease. The main goal in a GWAS is to identify genetic variants, single nucleotide polymorphisms (SNPs) in particular, that show association with the phenotype, such as "disease present" or "disease absent" either because they are causal, or more likely, because they are statistically correlated with an unobserved causal variant (Goldstein and Cavalleri, 2005). A GWAS can analyze "by DNA site" or "by multiple DNA sites. " In either case, data mining tools (Tachmazidou, Verzilli, and De Lorio, 2007) are proving to be quite useful for understanding the genetic causes for common diseases.

BACKGROUND

A GWAS involves genotyping many cases (typically 1000 or more) and controls (also 1000 or more) at a large number (10^4 to 10^6) of markers throughout the genome. These markers are usually SNPs. A SNP occurs at a DNA site if more than one nucleotide (A, C, T, or G) is found within the population of interest, which includes the cases (which have the disease being studied) and controls (which do not have the disease). For example, suppose the sequenced DNA fragment from subject 1 is AAGC**C**TA and from subject 2 is AAGC**T**TA. These contain a difference in a single nucleotide. In this case there are two alleles ("alleles" are variations of the DNA in this case), C and T. Almost all common SNPs have only two alleles, often with one allele being rare and the other allele being common.

Assume that measuring the DNA at millions of sites for thousands of individuals is feasible. The resulting measurements for n_1 cases and n_2 controls are partially listed below, using arbitrary labels of the sites such as shown below. Note that DNA site 3 is a candidate for an association, with T being the most prevalent state for cases and G being the most prevalent state for controls.

	123	456	789 ...
Case 1:	AA**T**	CT**A**	TAT ...
Case 2:	A* **T**	CT**C**	TAT …
...			
Case n_1:	AA**T**	CT**G**	TAT ...
Control 1:	AA**G**	CT**A**	TTA ...
Control 2:	AA**G**	CT**A**	TTA ...
...			
Control n_2:	AA**G**	CT**A**	TTA ...

Site 6 is also a candidate for an association, with state A among the controls and considerable variation among the cases. The * character (case 2) can denote missing data, an alignment character due to a deletion mutation, or an insertion mutation, etc. (Toivonen et al., 2000).

In this example, the eye can detect such association candidates "by DNA site." However, suppose the collection of sites were larger and all n_1 cases and n_2 controls were listed, or that the analysis were "by haplotype." In principle, the haplotype (one "half" of the genome of a paired-chromosome species such as humans) is the entire set of all DNA sites in the entire genome. In practice, haplotype refers to the sequenced sites, such as those in a haplotype mapping (HapMap,

2005) involving SNPs as we focus on here. Both a large "by DNA site" analysis and a haplotype analysis, which considers the joint behavior of multiple DNA sites, are tasks that are beyond the eye's capability.

Using modern sequencing methods, time and budget constraints prohibit sequencing all DNA sites for many subjects (Goldstein and Cavalleri, 2005). Instead, a promising shortcut involves identifying haplotype blocks (Zhang and Jin, 2003; Zhang et al., 2002). A haplotype block is a homogeneous region of DNA sites that exhibit high linkage disequilibrium. Linkage disequilibrium between two DNA sites means there is negligible recombination during reproduction, thus "linking" the allelic states far more frequently than if the sites evolved independently. The human genome contains regions of very high recombination rates and regions of very low recombination rates (within a haplotype block). If a haplotype block consists of approximately 10 sites, then a single SNP marker can indicate the DNA state (A, C, T, or G) for each site in the entire block for each subject, thus reducing the number of sequenced sites by a factor of 10.

The HapMap project (HapMap, 2005) has led to an increase from approximately 2 million known SNPs to more than 8 million. Many studies have reported low haplotype diversity with a few common haplotypes capturing most of the genetic variation. These haplotypes can be represented by a small number of haplotype-tagging SNPs (htSNPs). The presence of haplotype blocks makes a GWAS appealing, and summarizes the distribution of genetic variation throughout the genome. SNPs are effective genetic markers because of their abundance, relatively low mutation rate, functional relevance, ease of automating sequencing, and role as htSNPs. The HapMap project is exploiting the concept that if an htSNP correlates with phenotype, then some of the SNPs in its "association block" are likely to be causally linked to phenotype.

MAIN THRUST

Data Mining

Data mining involves the extraction of potentially useful information from data. Identifying genomic regions related to phenotype falls within the scope of data mining; we will limit discussion to a few specific data mining activities, which can all be illustrated using the following example. Consider the 10 haplotypes (rows) below (Tachmazidou et al., 2007) at each of 12 SNPs (columns). The rare allele is denoted "1," and the common allele is denoted "0." By inspection of the pattern of 0s and 1s, haplotypes 1 to 4 are somewhat distinguishable from haplotypes 5 to 10. Multidimensional scaling (Figure 1) is a method to display the distances between the 45 pairs of haplotypes (Venables and Ripley, 1999). Although there are several evolutionary-model-based distance definitions, the Manhattan distance (the number of differences between a given pair of haplotype) is defensible, and was used to create all three plots in Figure 1. The top plot in Figure 1 suggests that there are two or three genetic groups. Ideally, if there are only two phenotypes (disease present or absent), then there would be two genetic groups that correspond to the two phenotypes. In practice, because common diseases are proving to have a complex genetic component, it is common to have more than two genetic groups, arising, for example, due to racial or geographic subdivision structures in the sampled population (Liu et al., 2004).

haplotype1	1	0	0	0	1	0	1	0	0	0	1	0
haplotype2	1	0	0	0	1	0	1	0	0	0	0	0
haplotype3	1	0	0	0	0	0	1	0	0	0	0	0
haplotype4	1	0	0	0	0	0	0	0	0	1	0	0
haplotype5	0	1	0	0	0	1	0	0	0	0	0	1
haplotype6	0	1	0	0	0	1	0	0	0	0	0	0
haplotype7	0	0	0	1	0	1	0	0	0	0	0	0
haplotype8	0	0	0	1	0	1	0	0	1	0	0	0
haplotype9	0	0	0	1	0	1	0	1	1	0	0	0
haplotype10	0	0	1	0	0	1	0	0	0	0	0	0

The data mining activities described below include: defining genetics-model-based distance measures; selecting features; control of the false alarm rate; clustering in the context of phylogenetic tree building, and genomic control using genetic model fitting to protect against spurious association between halplotype and disease status.

Defining Genetics-Model-Based Distance Measures

Effective haplotype blocks in a GWAS requires small "within-block" variation relative to "between-block" variation. Variation can be defined and measured in several ways. One way is the Manhattan distance be-