

Data Mining Tools for Genotype-Phenotype Correlation

Jianhua Liu, PhD¹; Jason Buskirk, BS¹; Jennifer Santangelo, BS¹; Ryan J. Deiter, BS¹; Audrey Papp²; Glen E. Cooke, MD²; Wolfgang Sadee, Dr.rer.nat²; Jyoti Kamal, PhD¹

¹The Information Warehouse; ²Department of Cardiovascular Medicine
The Ohio State University Medical Center, Columbus, OH 43201, USA

Abstract

Single Nucleotide Polymorphisms (SNPs) may be the key to diagnosing and treating certain diseases. A preliminary study was conducted at The Ohio State University Medical Center Information Warehouse to correlate such SNPs with a selected group of lab values for cardiology patients. Early results show that data mining tools can be valuable for understanding such correlations, but further refinement of the methodology and data preparation is needed to fully realize such value.

Introduction

Single Nucleotide Polymorphisms (SNPs) are DNA sequence variations in the genome that differ by a single nucleotide (A, T, C or G). Scientists have long believed that SNPs may hold the key to formulating the diagnosis and treatment of certain diseases, as well as predicting individual responses to treatments. It is evident that SNPs can be an effective tool for identifying multiple genes associated with diseases such as cancer and vascular disease. One such example is the association of Alzheimer's disease with two SNPs in the apolipoprotein E (*ApoE*) gene (1). At the Ohio State University Medical Center (OSUMC), we have utilized data mining tools to correlate SNPs from cardiology patients with their lab data stored in the Information Warehouse (IW).

Methods

Genomic (SNP) and phenotypic data was collected for subjects with cardiovascular disease. The captured phenotypic data includes the maximum lab reading for Total Cholesterol (CHOL), Triglyceride (TRIG), Low Density Lipid (LDL), and C-Reactive Protein (CRP). These results were marked as high if they were above 200, 250, 140, and 3.0, respectively. In addition, the Master Death Index published by the Social Security Administration was available. For this study, we defined early death (E_DEATH) as that occurring five or more years prior to the life expectancy at birth for the corresponding gender group as published by the CDC in 2003. Analysis was performed using the Oracle 10gR2 platform with data mining option. The Attribute of Importance (AI) and Frequent Itemset (FI) data mining techniques were employed.

Results

Genomic data from 47 SNPs associated with 1469 subjects with cardiovascular disease was collected. In total, 881 subjects had at least one lab observation and 102 subjects had expired. AI was performed over all SNPs targeting the previously described phenotypes. FI was employed to find any correlations among genotypes (Ta-

ble 1) and phenotypes, as well as the combination of these two (Table 2). FI analyses of SNPs among subjects with additional high lab values alone were also conducted.

Table 1: Frequent Itemset of Genotypes

Itemset	Support	Length	Rank
S08(C,C), S41(T,T)	622	2	1
S08(C,C), S14(T,T)	576	2	2
S08(C,C), S35(C,C), S41(T,T)	503	3	1
S08(C,C), S14(T,T), S41(T,T)	494	3	2

Table 2: Frequent Itemset of Genotype-Phenotype

Itemset	Support	Length	Rank
S08(C,C), CHOL	129	2	1581
S41(T,T), CHOL	119	2	1800
S08(C,C), S41(T,T), CHOL	111	3	8660

Discussion

The AI results show that genes with high population fraction that is homozygous for one allele tend to contribute more toward any given phenotype. While this might be significant for the cardiovascular population as a whole, it is noteworthy that the importance of SNPs using different phenotypes as target attributes varies. While genes with low population allele pairs tend to be disregarded by pure statistical calculations, some evenly distributed ones are among the top of the list. In the FI study, 42% (622/1469) of the subjects were found to have S08(C,C)-S41(T,T) association, potentially leading researchers to examine these genes in more detail. Most interesting, however, are those itemsets listed in Table 2. Of the 233 subjects found to have high CHOL, more than 50% were shown to have S08(C,C). Also, although 171 subjects were marked as having high TRIG, none were found to be included in the itemsets of this study.

Future Work

The described work is illustrative of new methods currently being developed for determining the contribution of less populated alleles in a gene. In our future work, we plan to recruit a contrasting population of randomly selected subjects in order to gain further insight into the results of this preliminary study.

Reference

1. Saunders AM, Strittmatter WJ, Schmechel D, St George-Hyslop PH, Pericak-Vance MA, Joo SH, Rosi BL, Gusella JF, Crapper-McLachlan DR, Alberts MJ, Hulette C, Crain B, Goldgaber D and Roses AD (1993) Association of apolipoprotein E allele epsilon 4 with late-onset familial and sporadic Alzheimer's disease. *Neurology* 43:1467-1472.