

Computational population-based techniques in identifying genetic variants associated with simulated complex disorder in a general population

M. Sao Pedro¹, C. Shoemaker¹, M. Pungliya², C. Ruiz¹, S.A. Alvarez^{1,3}, M. Ward¹, M. Stevens¹, E.F. Ryder², and J. Krushkal²

¹Department of Computer Science and ²Department Biology and Biotechnology, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609. Tel. (508) 831-6140, Fax (508) 831-5936, e-mail krushkal@wpi.edu

³Current affiliation: Department of Computer Science, Wellesley College, Wellesley, MA 02481

Abstract

Several techniques for association analysis have been applied to the simulated data set for the general population (Problem 2 of the Genetic Analysis Workshop 12). We have focused our efforts on the pedigree founders who did not have any living parents. Entire pedigrees were also used in several methods to compare whether the inclusion of the offspring is beneficial. Association methods have been applied to test whether single nucleotide polymorphisms (SNPs) and microsatellite markers were correlated with a qualitative disease state. We compare here the performance of genotype-specific association analysis, allele-specific association analysis (linkage disequilibrium mapping), multivariate data visualization, association rule mining, and artificial neural networks. We did not have the answers to the problem. The performance of our various statistical analyses is evaluated by estimating the marker correlations between different methods, and the performance of our data mining approaches is measured by their degree of descriptiveness in the case of association rules, and by the accuracy of the classification predictions in the case of artificial neural networks.

Introduction

Despite the recent advances in collecting human genetic variants and the abundance of genetic databases of biallelic and microsatellite polymorphisms, there is a strong need for computational methods that would identify functional genetic variants in populations. A large number of recent studies have focused on family-based linkage and association analyses such as the transmission/disequilibrium test and its modifications (e.g., Spielman et al. 1993; Allison 1997; Lazzeroni and Lange 1998). Population-based studies, on the other hand, often focus on linkage disequilibrium analysis of individual alleles, using one genetic marker at a time (e.g. Collins and Morton 1998; Elston 1998). We investigated the performance of several existing and new association techniques in population-based samples using pedigree founders and, in some cases, all pedigree members. In this report, we compare the performance of the disequilibrium approach applied to individual alleles and genotypes with multipoint data mining techniques such as association rules and neural nets. We suggest possible strategies for reducing the data dimensionality and combining various techniques to identify functional genetic mutations.

Materials and Methods

Individuals Included in the Study

The data sets used in this analysis are described in Table 1. Data sets 1 and 2 contain the founders of the pedigrees from the general population. There were 165 such living individuals in each replicate that had no living parents and were unrelated to each other. Data sets 3 and 4 contain all living individuals in each pedigree. Data sets 1 and 4 included all replicates 1 through 50, while data sets 2 and 3 focused on the "best" replicate 42.

Table 1. Subjects included in the four different data sets used in the study.

Data set	1	2	3	4
Number of people	8250	165	1000	50000
Replicate number	1 through 50	42	42	1 through 50
Founders	Included	Included	Included	Included
Other living individuals	Not included	Not included	Included	Included
Dead relatives	Not included	Not included	Not included	Not Included

Genetic variants analyzed

We have analyzed both biallelic sequence variants and microsatellite markers. To reduce the dimensionality of the data, we selected only those sequence polymorphisms that were present in the pedigree founders in all 50 replicates. There were 715 such SNPs after the data reduction, while the number of SNPs before the data reduction was 9515. The distribution of these polymorphisms between the seven candidate genes is shown in Table 2, where A signifies the first gene, B the second, and so on.

Table 2. Number of occurrences of sequence polymorphisms from each gene in the 9515 SNPs present prior to the data reduction and the 715 SNPs shared among replicates.

Gene	Number of SNPs prior to the data reduction	Number of SNPs after the data reduction
A	3403	157
B	1642	90
C	1101	99
D	2305	121
E	70	37
F	629	34
G	365	177
<i>Total</i>	9515	715

All 2855 microsatellite markers were included at the start of the analysis. After the initial screening by the allelic linkage disequilibrium (see below), the amount of data was reduced to the 100 SNPs and 10 microsatellite markers with the top scores showing the closest association to the disease state.

Use of Phenotypic Data

715 SNPs and 2855 microsatellite markers were studied for the association with the disease affection status, a qualitative trait. The neural net approach and the association rule mining approach also considered environmental factors and quantitative traits Q1-Q5 in addition to SNPs and microsatellite markers.

Computational Techniques Used to Find Significant Markers

Each of the 715 SNPs shared among the pedigree founders in 50 replicates and every microsatellite marker was analyzed for allelic linkage disequilibrium in data set 4 using the chi-square test. The 100 markers and the 100 SNPs with the lowest p-values from this analysis were selected as input for the association rule mining tools. Each of the 715 SNPs was also analyzed for the presence of genotypic disequilibrium in data set 1 using the chi-square test for the difference between the genotype frequencies in healthy and diseased people. The significance level for each individual SNP was determined by the Bonferroni correction. In addition, these 715 SNPs were also analyzed by multivariate data visualization in data set 1 to sort the SNPs by the sum of the absolute values of the differences among their genotype frequencies in healthy and diseased people. The resulting array of SNPs was presented in graphical form using the visualization software package XmdvTool (Ward 1994). More detailed description of the genotype-

specific association analysis and the multivariate data visualization is provided in a separate report by Pungliya et al.

Additionally, two multipoint analysis techniques rooted in the machine learning and data mining literature were used to analyze the data for the founders of the general population. Artificial Neural Networks (ANN) were used to infer a predictive model of the simulated disease, while Association Rule Mining (ARM) was used to search for a descriptive model. We comment on some of the experimental results obtained with these techniques below. Further details, including a discussion of the relative merits of ANN and ARM, are provided in a separate GAW12 abstract by Shoemaker et al. For the ANN experiments, a fully connected feedforward network with one hidden layer was used. Inputs were extracted from data for the general population; different combinations of input features were considered, including SNP data only, SNP and microsatellite marker data, and the latter together with data for the environmental factors and quantitative attributes. A single output unit representing the disease affection status was used. Weights of the connections between network units were initialized to small random values. The network was then trained using the method of error backpropagation (Rumelhart and Hinton 1987). In order to obtain a model that is more readily interpretable by humans, association rule mining was attempted using the standard Apriori algorithm (Agrawal et al. 2000). We used two available tools for mining association rules from the general population data, IBM's Intelligent Miner and the University of Singapore's CBA system (Liu et al. 1998). The large number of attributes for the GAW12 data set led to very high running times for the rule mining process, making it very difficult to obtain useful rules. This is not surprising, as the Apriori algorithm was designed for applications with fewer attributes and seeks to exhaustively find all rules that satisfy specified minimum support and confidence constraints; thus, it fails to take advantage of the special structure of the problem at hand. We then concentrated on mining association rules over replicate 42 using the 100 markers and the 100 SNPs with the lowest p-values together with affection status and all environmental factors and qualitative traits. We report the results of those experiments below.

Results

Statistical Analyses

We compared the most significant SNPs generated by allele-specific linkage disequilibrium analysis from data set 4 with those generated by genotype-specific disequilibrium analysis from data set 1. 39% of the 100 most significant SNPs from data set 4 were also found in the 100 most significant SNPs from data set 1. This compares to a 14% match expected by chance between the top 100 items of two randomly ordered lists of 715 items. Both approaches indicated that candidate genes A, B and F were associated with the disease. When allelic linkage disequilibrium was used, the top 100 SNPs sorted by their chi-square values included 69 SNPs from gene A, 20 from gene B, and 11 from gene F. When genotype-specific disequilibrium was investigated, the list of SNPs with significant p-values included 107 SNPs from gene A, 52 from gene B, and 13 from gene F. Additional analysis of genotype disequilibrium in genes A and B using data sets 2 and 3 confirmed the likely role of these genes in affecting the disease state.

For data set 1, the order of the 715 SNPs sorted by their chi-square values according to their genotype-specific disequilibrium was also compared to the SNP order obtained by multivariate data visualization using the sum of absolute differences in genotype frequencies in healthy and affected individuals. The correlation coefficient between these two sorted orders of polymorphisms was 0.427.

Machine Learning / Data Mining Results

The association rules obtained from our experiments provided a descriptive model correlating some of the attributes in replicate 42. These rules described interesting patterns. For example, we found that people in replicate 42 who have a value of 08 in marker D10G108 and a value of 22 for SNP 11911 in gene A are predicted to be healthy with a confidence of 98.246%, and this pattern is supported by 5.7% of the population in this replicate. Also, people with a value of 02 in marker D16G038 and a value of 22 in SNP 14061 in gene A are predicted to be healthy with a support of 3.7% and 100% confidence.

We describe the ANN results separately, as this technique is intended to provide a predictive model, rather than a descriptive model of the simulated disease. The inferred model used multiple SNP and marker attributes in order to provide predictions of the affection status of individuals; information regarding the relative importance of single attributes was not directly provided. Our ANN experiments showed good

predictive accuracy, with increasing accuracy as the amount of information available increased; thus, SNP-only accuracy was lower than that obtained when marker data was incorporated, for example. During training, prediction error decreased from an initial value of 25%-30%, corresponding roughly to the fraction of diseased individuals among the founders of the general population, to values less than 1% for some of the experiments. These results should be cautiously interpreted, as this error was obtained when testing the networks on the same data as that used for training. When the available data were split into disjoint training and testing sets, the error obtained after training was still significantly lower than that corresponding to random predictions; the absolute levels were in the 10%-15% range.

More details on the results of the association rule mining and neural networks approaches and a comparison of the two are provided in a separate abstract by Shoemaker et al.

Discussion

Various techniques described here show the importance of genes A, B, and F, and especially gene A in affecting the disease status. However, there was some lack of agreement between the lists of top performing SNPs obtained in various statistical analyses. This discrepancy could be explained by, first, sampling differences in various data sets that may play a role when alleles located in linkage disequilibrium with the disease are studied. Second, technique-specific differences may also affect the results of the analysis.

We found that standard association rule mining tools are incapable of handling genomic data directly given the high dimensional nature of these data. In work in progress, some of the authors of the current study are developing data mining algorithms that will be capable of rapidly discovering significant association rules in this context; this work builds on previous work by Lin et al. (2000).

The multipoint techniques described here were able to incorporate both microsatellite markers and biallelic sequence polymorphisms. We believe that such a multipoint approach may be crucial in solving the problem of identifying important functional mutations. Single-point analysis, on the other hand, has proved useful for initial screening and dimensionality reduction of the genome data.

Acknowledgements

We thank K. Ferzoco for assistance with this project. This work was supported by the grant "Computational Algorithms for Analysis of Genomic Data" from the Research Development Council at the Worcester Polytechnic Institute.

Bibliography

- Agrawal, R., T. Imielinski, A. Swami (2000) Mining Association Rules between Sets of Items in Large Databases, Proc. of the ACM SIGMOD Conference on Management of Data, ACM, Washington, D.C., pages 207-216
- Allison, D.B. (1997) Transmission-disequilibrium tests for quantitative traits. *Am. J. Hum. Genet.* 60:676-90.
- Collins, A., and N. E. Morton (1998) Mapping a disease locus by allelic association. *Proc. Natl. Acad. Sci. USA* 95: 1741-1747
- Elston, R. C. (1998). Linkage and association. *Genetic Epidemiology* 15: 565-576
- Lazzeroni, L.C., and K. Lange. 1998. A conditional inference framework for extending the transmission/disequilibrium test. *Hum Hered.* 48:67-81.
- Lin, W., S.A. Alvarez, and C. Ruiz (2000) Collaborative Recommendation via Adaptive Association Rule Mining, WebKDD-2000 Workshop on Web Mining for E-Commerce, Sixth International Conference on Knowledge Discovery and Data Mining, Boston, MA Aug. 2000
- Liu, B., W. Hsu, and Y. Ma (1998) Integrating Classification and Association Rule Mining, Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, August 1998, pp 80-86.
- Rumelhart, Hinton, Williams (1986) "Learning Internal Representations by Backpropagation", in *Parallel Distributed Processing*, vol. I. MIT Press

Spielman, R.S., R. E. McGinnis, and W. J. Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet.* 1993 52:506-16.

Ward, M.O. (1994). XmdvTool: Integrating Multiple Methods for Visualizing Multivariate Data", *Proc. of Visualization '94*, pp. 326-333.