

**Computational Methods for Single Point and Multipoint Analysis of Genetic Variants
Associated with a Simulated Complex Disorder in a General Population**

Christopher A. Shoemaker¹, Manish Pungliya^{2*}, Michael A. Sao Pedro^{1*}, Carolina Ruiz¹,
Sergio A. Alvarez³, Matthew Ward¹, Elizabeth F. Ryder², and Julia Krushkal^{2#}

¹Department of Computer Science and ²Department Biology and Biotechnology, Worcester
Polytechnic Institute, 100 Institute Road, Worcester, MA 01609; ³Department of Computer
Science, Wellesley College, Wellesley, MA 02481

* The second and the third author contributed equally to this work

Correspondence to: Julia Krushkal (Tel. 508-831-6140, Fax 508-831-5936, e-mail
krushkal@wpi.edu)

Abstract

Several techniques for association analysis have been applied to simulated genetic data for a general population. We describe and compare the performance of three single point methods and two multipoint approaches rooted in machine learning and data mining.

Running Title: Population-based association methods

Key Words: Allelic Linkage Disequilibrium, Genotypic Disequilibrium, Multivariate Data Visualization, Association Rule Mining, Artificial Neural Networks

Introduction

Given recent advances in collecting human genetic variants, there is a strong need for computational methods that can identify functional genetic variants in populations. In this report, we describe and compare the single point disequilibrium approach applied to individual alleles and genotypes, and two multipoint data mining techniques. We also present computer visualization techniques for viewing the clusters of loci that are correlated with disease status.

Most single point methods employed in this study were applied to single nucleotide polymorphism (SNP) data and considered qualitative traits. The multipoint methods, which employed additional data, were Artificial Neural Networks (ANN), used to infer a predictive model of the simulated disease, and Association Rule Mining (ARM), used to search for a descriptive model. Neural networks are useful in a variety of problems, and they are particularly good at identifying functional relationships in data. We applied ARM and ANN to identify functional mutations and predict the disease state.

We suggest possible strategies for reducing the dimensionality of the data and combining various techniques. We did not have the GAW12 answers at the time of this study.

Materials and Methods

Individuals Included in the Study. Three data sets were used: (1) 8250 pedigree founders from all 50 replicates; (2) 165 pedigree founders from the "best" replicate 42; and (3) 1000 living individuals from replicate 42. The general population was chosen to reduce the correlation of shared genetic material among pedigree founders. We report here mainly the results for data set 1. The results obtained for data sets 2 and 3 were consistent with those from data set 1.

Computational Techniques. Computational methods used in this study are listed in Table I. In order to reduce the dimensionality of the data, we included only those SNPs for which at least one founder in each replicate varied from the ancestral allele. Only 715 of the 9515 original SNPs met this criterion.

Allelic Linkage Disequilibrium. Each of the 715 SNPs and the 2855 short tandem repeats (STRs) were analyzed for Allelic Linkage Disequilibrium (ALD) for differences in the number of alleles in unaffected and affected individuals in all replicates combined using the chi-square test. The 100 STRs and the 100 SNPs with the lowest p -values were selected as input for the ARM (see below). ALD was also applied to data set 1 to compare the outcome with results of other techniques.

Genotype Disequilibrium. Each of the 715 SNPs was analyzed for the presence of Genotypic Disequilibrium (GD) using the chi-square test for the difference between the numbers of genotypes of affected and unaffected individuals. The Bonferroni correction was applied to account for multiple testing. When the significance level for the 715 SNPs was set at 0.05, the significance level for each individual SNP was determined to be 7.17×10^{-5} .

Multivariate Data Visualization. We have developed a Multivariate Data Visualization (MDV) program, Scansort (Ward, unpublished; available at matt@cs.wpi.edu), which sorts the SNP loci using differences between genotype frequencies of affected and unaffected subjects. The program generates one record for each SNP as follows: {diff, 11h, 12h, 22h, 11s, 12s, 22s, d11, d12, d22, ID}, where ID is the SNP index in the original order; ijh is the frequency of unaffected (healthy) individuals with genotype ij ; ijs is the frequency of affected (sick) individuals with genotype ij ; and $dij = |ijh - ijs|$. diff is the sorted SNP index computed as $\text{diff} =$

d11+d12+d22. The resulting array of SNPs was presented in graphical form using the visualization software XmdvTool [Ward, 1994] (available at <http://davis.wpi.edu/~xmdv>).

Association Rule Mining. Association rules [Hájek et al., 1966; Agrawal et al., 1993] identify collections of data attributes that are statistically related in the underlying data. An association rule is of the form $X \Rightarrow Y$ where X and Y are disjoint conjunctions of propositions. The *confidence* of the rule is the conditional probability of Y given X , $\Pr(Y|X)$, and the *support* of the rule is the prior probability of X and Y , $\Pr(X \text{ and } Y)$. Here probability is taken to be the observed frequency in the underlying data. The ratio $\Pr(Y|X)/\Pr(Y)$, called the "lift" of the rule, is often used as a measure of how interesting the rule is. The further this ratio is from 1, the more likely/unlikely X makes Y .

We mined association rules in data sets 1 and 3 to find correlations among SNPs, STRs, environmental factors, quantitative traits Q1-Q5, and affection status. To reduce the dimensionality of the data, we used the top 100 SNPs and 100 STRs identified by ALD. Since we were specifically interested in classifying for presence or absence of the disease, we focused on mining those rules that had only affection status values on the right-hand side of the rule.

We used IBM's Intelligent Miner for Data and CBA (Classification Based Associations) [Liu et al., 1998]. In this report, we focus on CBA. Since these tools were developed mainly for market basket analysis, we translated the GAW12 data into a transactional format. Continuous attributes were discretized into intervals. These intervals were determined by finding the splits of the attribute value range that maximized the information gain with respect to the affection status.

Artificial Neural Networks. ANN are a class of models of distributed computation inspired by biological systems of neurons. An ANN is defined by a weighted directed acyclic

graph (G,E,w) , where G is the set of nodes (*processing units*), corresponding roughly to neurons or functional groups of neurons; E is the set of edges of the graph, representing neuronal interconnections; and w is a real-valued function defining connection weights. Neural networks are capable of a rudimentary form of "learning", associated with changes in the weights. Training may be performed by the method of error backpropagation [e.g., Rumelhart et al. 1986].

We employed a feedforward network architecture with two layers of processing units. The inputs feed directly into the units of the first, or "hidden" layer. The hidden units feed into the units of the second, or "output" layer. The network topology is "fully connected", meaning that all pairs of (input, hidden) units are connected, as are all (hidden, output) pairs. Inputs were extracted from data for the pedigree founders. Different combinations of input features were considered, including SNP data only, SNP and STR data, and the latter together with data for the environmental factors and quantitative traits Q1-Q5. A single output unit representing the 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. Convergence to an equilibrium state generally took place after a few hundred training iterations.

Results

Distribution of top rank SNPs among candidate genes. There was a very good agreement between the SNPs identified as important by the three single point methods, which are sensitive to allelic or genotypic differences between the affected and unaffected individuals (Table II). We calculated Spearman rank correlation coefficients for the SNPs ranked according to ALD, GD, and MDV, using data set 1. For the 100 most significant SNPs identified by each method, r_s

= 0.920 for GD and MDV, 0.933 between ALD and GD, and 0.975 between ALD and MDV, corresponding to $p < 0.001$.

All single point approaches indicated that candidate genes 1, 2, and 6 were associated with the disease. None of the SNPs from other genes ("false positives") were detected. When ALD was used in data set 1, the top 100 SNPs included 69 SNPs from gene 1 (44% of the total number of SNPs from that gene), 20 from gene 2 (22%), and 11 from gene 6 (32%). When using GD, the list of SNPs with significant p -values ($p < 7.17 \times 10^{-5}$) included 107 SNPs from gene 1 (68%), 52 from gene 2 (58%), and 13 from gene 6 (32%).

Multivariate Data Visualization. We identified the SNPs with high values of the sorted index, diff, and classified them according to their genotype values. Figure 1 presents the relationships among the sorted order of SNPs, the genotype frequencies, and their differences between the affected and unaffected individuals. The data are displayed using parallel coordinates, where each dimension or variable is a vertical axis, and each multivariate data point forms a polyline across the axes. Clusterings are apparent on the graph. SNPs exhibiting little variation between affected and unaffected subjects generate lines near the bottom of axes corresponding to diff, d11, d12, and d22. The most significant variations can be seen in the upper lines in axes showing values of d11 and d22.

Association Rule Mining. The association rules obtained from our experiments provided a descriptive model correlating some of the attributes. Some examples are shown in Table III. CBA discovered 17860 rules with support above 1% and confidence above 80% in data set 1, and 18996 such rules in data set 3. A subset of 120 rules, containing the best classifying rules, was selected by CBA (see [Liu et al., 1998] for details on the classifier construction). The

resulting classifier achieved a classification error of 16.8% when tested over the training data set 3, which constitutes a reduction of the 25.9% error rate expected from a random classifier.

Artificial Neural Networks. Our ANN experiments showed good predictive accuracy, which increased with the amount of available information. For example, SNP-only accuracy was lower than that obtained when STR data were incorporated. During training, prediction error decreased from an initial value of 25%-30% to values less than 1% for some of the experiments. This error was obtained when testing the networks on the same data as that used for training.

When the data were split into disjoint training and testing sets using subsets of replicates 1-5 and 42, the error obtained after training (10%-15%) was still significantly lower than that corresponding to random predictions (25.9%). Five different training pairs were used, all of which resulted in similar error rate. Examples included using replicates 1-4 as the training set and replicate 5 as the test set; using replicates 1-4 as the training set and 42 as the test set; and splitting replicate 42 into training and test sets.

Figure 2 shows the internal representations of the unaffected and affected groups as learned by an ANN. The ANN learned a two-dimensional "feature space" in which the two groups corresponded to different (but overlapping) regions; unaffected individuals were clustered toward the upper left corner, while affected ones tended to lie closer to the lower right corner. Further processing of the feature space within the ANN predicted affection status by determining a suitable decision boundary in this space. Because the regions occupied by the unaffected and affected groups overlap, some prediction error remained.

Discussion

We correctly identified the importance of candidate genes 1, 2, and 6 to the disease state. Because these genes contained multiple SNPs, it is difficult to identify causative mutations for the disease as opposed to those changes that may be in linkage disequilibrium with a deleterious mutation. Many of the SNPs located near the mutations with the lowest p -values also have low p -values, most likely because of the linkage disequilibrium between SNP variants. In gene 1 position 557 has the lowest p -value. At least seven other SNPs at the 5' end of that gene also have $p \leq 6.29 \times 10^{-8}$, suggesting a possible functional role of the 5' end of gene 1 in the disease.

We pursued the hypothesis that the polymorphisms showing the lowest p -values (e.g., SNP position 557 in gene 1, 6805 and 7332 in gene 6, and 4894 in gene 2) were the most likely candidates causing the disease. However, GAW12 answers reveal that position 557 in gene 1 is important to the disease state, but neither position 6805 nor 7332 in gene 6 is.

Our multipoint analyses illustrate the importance of two different machine learning/ data mining approaches. These methods have a great advantage that they are able to utilize all the available information including SNPs, STRs, quantitative traits and environmental factors. ANN trained using the method of error backpropagation provide accurate predictive models. Great reductions in the error rate of ANN were achieved for predicting the disease status compared to simply guessing the majority class. However, it is not easy to extract descriptive information from an ANN trained to predict the disease status. ARM is intended to produce descriptive models that can be readily interpreted. In contrast to other multipoint methods of analysis such as regression, ARM has the advantage that it discovers associations among the data without assuming any particular model. However, standard ARM algorithms spend an inordinate amount of time attempting to find all association rules with support and confidence above minimum

levels specified *a priori*. In addition, because there were fewer affected than unaffected individuals in the data, there was less support for rules that implied that a person would be affected. We are currently developing ARM algorithms better suited to genetic analysis, based on our work in a different application domain [Lin et al., 2000].

The best approach for identifying important genetic loci may be a combination of methods. ALD or GD can be used for initial screening of the genome data, to eliminate those polymorphisms that are clearly unrelated to the disease status. As a next step, MDV can be used, which is a powerful interactive tool that can identify clusters of SNPs of interest by investigating the relationships between the sorting order and differences in frequencies in individuals with a particular genotype. As a final step, multipoint techniques (ARM and ANN) should be applied, which will take into account information from many SNPs, STRs, quantitative traits, and environmental factors. ARM may be crucial in identifying functional mutations, while ANN addresses the important issue of predicting disease status. We are developing strategies for combining these techniques into a multistep approach to association analysis.

Acknowledgements

We thank the anonymous reviewers for their comments. This work was supported by a grant from the Research Development Council at Worcester Polytechnic Institute.

References

- Agrawal R, Imielinski T, Swami A. 1993. Mining association rules between sets of items in large databases, Proc. of the ACM SIGMOD Conference on Management of Data, ACM, Washington, D.C. p 207-216
- Hájek P, Havel I, Chytil M. 1966. The GUHA method of automatic hypotheses determination. Computing 1:293-308.
- Lin W, Alvarez SA, Ruiz C. 2000. Collaborative recommendation via adaptive association rule mining. WebKDD-2000 Workshop on Web Mining for E-Commerce, 6th International Conference on Knowledge Discovery and Data Mining, Boston, MA.
- Liu B, Hsu W, Ma Y. 1998. Integrating classification and association rule mining. Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining, August 1998, p 80-86.
- Rumelhart D, Hinton G, Williams R. 1986. Learning internal representations by backpropagation. In: Parallel Distributed Processing, vol. I. MIT Press.
- Ward MO. 1994. XmdvTool: Integrating multiple methods for visualizing multivariate data. In: Proc. of Visualization '94. p 326-333.

Table I. Population-based methods used

Method	Abbreviation	Data set	Markers
Allelic Linkage Disequilibrium	ALD	1	715 SNPs, 2855 STRs
Genotype Disequilibrium	GD	1-3	715 SNPs
Multivariate Data Visualization	MDV	1-3	715 SNPs
Association Rule Mining	ARM	1,3	Top 100 SNPs, Top 100 STRs
Artificial Neural Networks	ANN	1*,2	715 SNPs, 2855 STRs

*Replicates 1-5 only for data set 1

Table II. Top 15 SNP loci determined by different techniques using 8250 founders

Rank	ALD	GD	MDV	ARM(support)	ARM(confidence)
1	557¹	557¹	557¹	3456¹	3835¹ , 3534 ^{1*}
2	76¹	76¹	76¹	4315 ¹ , 3456¹	3835¹ , 5757^{1*}
3	1553¹	1553¹	2619¹	3502, 3456¹	7577 ¹ , 6903 ^{1*}
4	2619¹	2619¹	1553¹	4681 ¹ , 3456¹	7890 ¹ , 6903 ^{1*}
5	3853 ¹	3853 ¹	3573¹	1553¹	7281¹ , 6903 ^{1*}
6	3742 ¹	3742 ¹	3835¹	3502 ¹ , 1553¹	76¹ , 15935 ^{1*}
7	3573¹	3573¹	3853 ¹	4681 ¹ , 1553¹	557¹ , 15935 ^{1*}
8	3835¹	3835¹	3742 ¹	4315 ¹ , 1553¹	17478 ^{1*}
9	3456¹	3456¹	3456¹	14839 ¹ , 3456¹	2942¹ , 11180^{1*}
10	5757¹	5757¹	5757¹	2619¹	2942¹ , 15775 ^{1*}
11	7281¹	7281¹	7281¹	14839 ¹ , 1553¹	76¹ , 18909 ^{1*}
12	2942¹	2942¹	2942¹	2619¹ , 1553¹	557¹ , 18909 ^{1*}
13	2923 ¹	2923 ¹	2923 ¹	3573¹	596¹ , 5961 ²
14	11180¹	11180¹	11180¹	3456¹ , 1553¹	189¹ , 5961 ²
15	189¹	189¹	1478 ¹	14839 ¹ , 2619¹	596¹ , 4977 ²

Note. ARM results are provided in two separate columns, with top 15 rules sorted by support, and top 15 rules sorted by confidence. Only ARM results restricted to SNPs are shown. Gene numbers are provided in superscript. An asterisk (*) shows classification rules that all have

100% confidence and should all be considered to have rank equal 1. SNPs identified as having rank of 15 or less by all the four methods are shown in bold.

Table III. Example association rules discovered using the CBA data mining tool

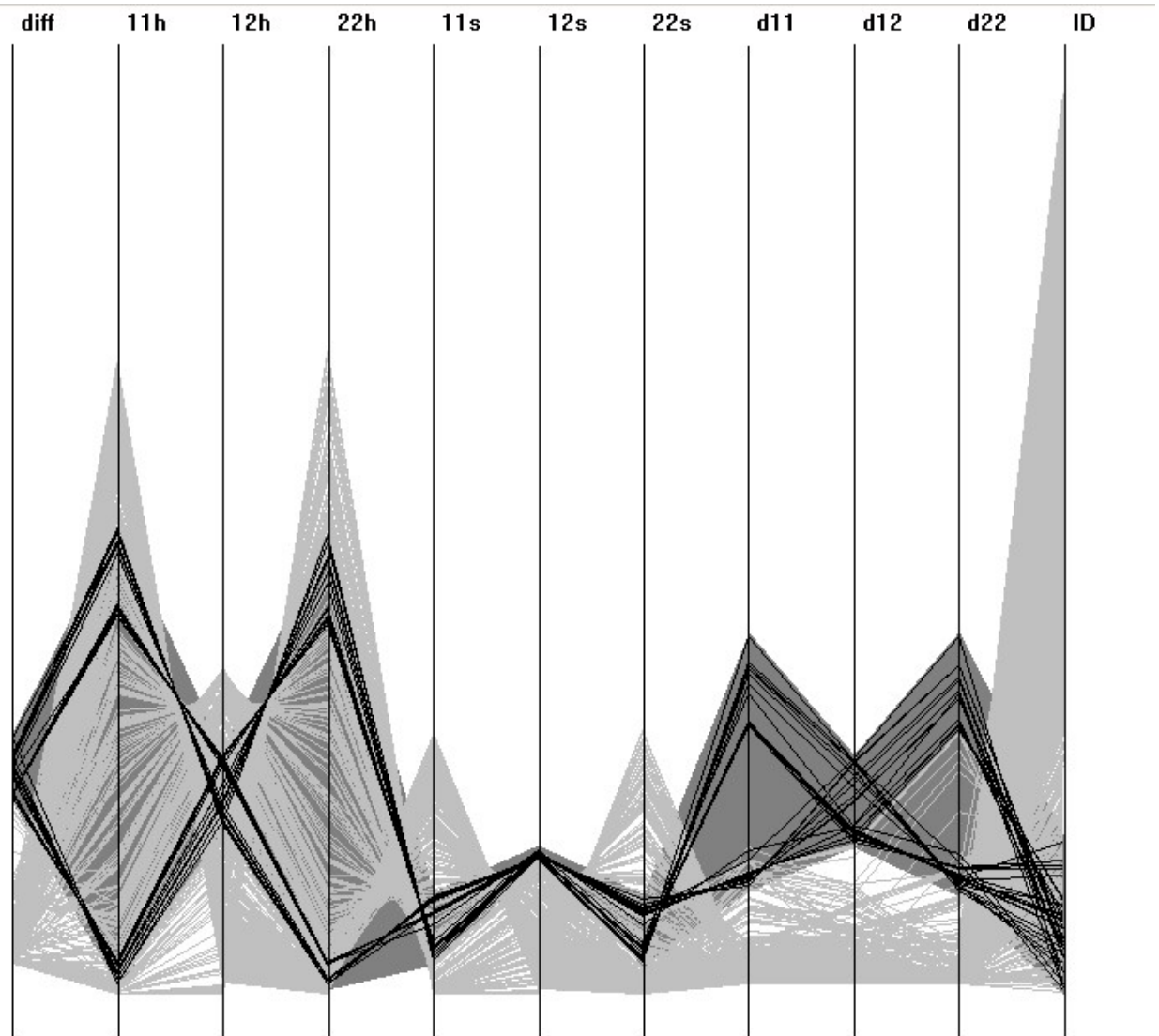
Data set	Association Rule	Support	Confidence	Lift
1	(STR D3G165=05) and (SNP 6903 gene 1=12) \Rightarrow unaffected	9.08%	85.24%	1.15
1	(SNP 557 gene 1=11) and (SNP 15935 gene 1=22) \Rightarrow affected	0.04%	100.00%	3.29
1	(Sex = male) and (SNP 557 gene 1=22) \Rightarrow unaffected	28.66%	93.11%	1.26
3	(STR D16G038=02) and (SNP 14061 gene 1=22) \Rightarrow unaffected	3.70%	100.00%	1.35
3	(Q1 in [0,16.45]) and (SNP 1553 gene 1=11) \Rightarrow unaffected	9.5%	100.00%	1.35
3	(Q2 > 25.19) and (Environmental Factor 2 = 0) \Rightarrow affected	2.8%	75.68%	3.86

Note. These rules were selected among those rules with either the highest support and/or the highest confidence to illustrate the interactions between simulated phenotypic and genotypic attributes. The last rule illustrates the interaction between Q2 and an environmental factor.

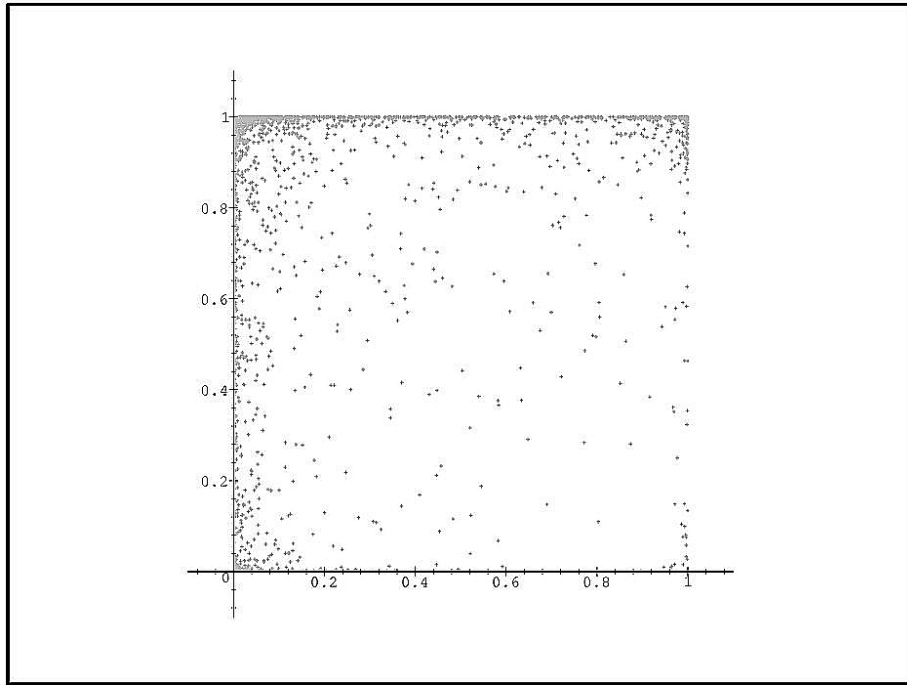
Figure legends

Figure 1. Relationships among sorted order of SNPs and differences in genotype frequencies in affected and unaffected individuals using 715 SNPs from 8250 pedigree founders. Black lines correspond to selected individual SNPs with the highest values of parameter diff. The dark gray area shows a range of values for selected SNPs. The remaining SNPs are presented by light grey lines.

Figure 2. Features learned by the Artificial Neural Network for unaffected (**A**) and affected (**B**) individuals. Each point represents an individual. The coordinates of a given point within the figure are the levels of activation induced in two processing units which constitute the hidden layer of the network, in response to the input data for the given individual.



A



B

