

# A Combinatorial Method for Predicting Genetic Susceptibility to Complex Diseases

Weidong Mao    Jingwu He    Dumitru Brinza    Alex Zelikovsky  
Department of Computer Science  
Georgia State University  
Atlanta, GA 30303, USA

**Abstract**—Recent improvements in the accessibility of high-throughput genotyping have brought a great deal of attention to disease association and susceptibility studies. This paper explores possibility of applying combinatorial methods to disease susceptibility prediction. The proposed combinatorial methods as well as standard statistical methods are applied to publicly available genotype data on Crohn’s disease and autoimmune disorders for predicting susceptibility to these diseases. The quality of susceptibility prediction algorithm is assessed using leave-one-out and leave-many-out tests - the disease status of one or several individuals is predicted and compared to their actual disease status which is initially made unknown to the algorithm. The best prediction rate achieved by the proposed algorithms is 77.78% for Crohn’s disease and 64.99% for autoimmune disorders, respectively.

## I. INTRODUCTION

Recent improvement in accessibility of high-throughput genotyping brought a great deal of attention to disease association and susceptibility studies [1]. High density maps of single nucleotide polymorphism (SNPs) [2] as well as massive genotype data with large number of individuals and number of SNPs become publicly available [3], [4], [5]. By now most of analysis of the new data is undertaken in statistics community [6], [7]. Different line of attack on disease susceptibility adhered to computational community with an emphasis on designing rather than analytical methodology.

The main goal of disease association and susceptibility analysis is to identify gene variations or, in general, haplotypes and genotypes which are susceptible to a particular disease. There are basically two main steps in disease susceptibility analysis: (i) the population haplotyping and (ii) identifying of the haplotypes/genotypes susceptible to the diseases. Unfortunately, existing methods for the step (i) introduce substantial noise drastically decreasing association strength of methods applied for step (ii) [7].

The main obstacle in population haplotyping is that the methods of inferring two haplotypes from individual data are too expensive [8]. The vast majority of the data are in the *genotype form*. For each bi-allelic SNP, genotypes specify whether the corresponding individual is homogeneous, i.e., the both haplotypes have the same allele, indicating present allele (major, referred as 0, or minor, referred as 1), or the corresponding individual is heterogeneous, i.e.,

the two haplotypes have different alleles (referred as 2). Several statistical, combinatorial and hybrid methods have been successfully applied to the haplotype inference problem also referred as phasing problem [9]. We use GERBIL [9] and PHASE[10] to infer haplotypes for population.

The traditional direct statistical association so far is unsatisfactory and arguably is not applicable to complex disease since it mostly relies on an assumption that the disease is caused by a single Mendelian gene [6]. Indeed, some complex diseases, such as psychiatric disorders, are characterized by a non Mendelian, multifactorial genetic contribution with a number of susceptible genes interacting with each other [11], [12]. Disease association analysis usually results in claims that a presence of a given SNP considerably increases the risk of a certain disease which are of limited use for disease susceptibility because susceptible SNPs are usually linked and do not, therefore, have an increased cumulative impact as it would be expected from the independent SNPs.

The observed weakness of statistical methods may lead to a quite plausible assertion that each case of complex diseases may have a unique chain of genetic as well as environmental elements [6]. On the contrary, this study tries to assess accumulated information using combinatorial tools hoping that there exist certain combinatorial dependencies in haplotypes which are scattered all over lengthy sequences and are difficult to recover and, therefore, have not yet been (statistically) analyzed. This paper explores possibility of applying combinatorial methods to known case/control studies with the hope to reliably (to certain extent) predict disease susceptibility.

We first describe several prediction algorithms which are mostly based on combinatorial optimization. Then we describe the leave-one-out test for comparative estimation of prediction quality of a discrimination algorithm as well as bootstrapping strategy for estimation of significance of obtained results.

We apply proposed methods to two data sets. The first data set consists of case/control study of Crohn’s disease [3] of 129 family trios. The other set for autoimmune disorder [13] consists of 1036 unrelated case/control individuals. We achieved correct prediction rate of 77.78% and 64.99%, respectively. After applying bootstrapping we obtain with 95% confidence the correct prediction rate of 75.38% for

Crohn's disease. We have also performed a *Monte-Carlo test* by running our methods on Crohn's disease's data with randomly swapped case/control markers. The average prediction rate falls to 50% for all proposed methods. This confirms predominating genetic susceptibility of Crohn's disease [14], high association of the chosen haplotype region with Crohn's disease as well as capabilities of the proposed methods to detect such susceptibility.

The rest of the paper is organized as follows. Section II describes several statistical and proposed combinatorial susceptibility prediction methods. Section III describes and discusses prediction susceptibility results for real data.

## II. PREDICTION ALGORITHMS FOR DISEASE SUSCEPTIBILITY

We will first describe the input and the output of a prediction algorithm. Then we describe implemented algorithms including the closest neighbor and statistical heuristics and three proposed graph-based algorithms.

Data sets have  $n$  genotypes and each has  $m$  SNPs. The input for a prediction algorithm includes:

- (G1) Training genotype set  $g_i = (g_{i,j}), i = 0, \dots, n-1, j = 1, \dots, m, g_{i,j} \in \{0, 1, 2\}$
- (G2) Disease status  $s(g_i) \in \{-1, 1\}$ , indicating if  $g_i, i = 0, \dots, n-1$ , is in case (1) or in control (-1), and
- (G3) Testing genotype  $g_n$  without any disease status.

The input data can also be phased, then each genotype is represented by a pair of haplotypes.

We will refer to the parts (G1-G2) of the input as *training set* and to the part (G3) as the test case. The output of prediction algorithms is the disease status of the genotype  $g_n$ , i.e.,  $s(g_n)$ .

Haplotype-based algorithms are supposed to more accurately take in account combinatorial structure of data but they suffer from the noise contributed by uncertainty in phasing of original genotype data. Our experiments show that adaptation of genotype-based algorithms to haplotypes only slightly affects accuracy of prediction indicating high quality of our phasing method.

Below we describe the prediction algorithms which have been implemented and verified on both data sets. Although there is a huge literature on classification and prediction algorithms and general approaches such as neural networks [15], SVM [16] and classification trees [17], we decided to confine ourselves mostly to statistical and combinatorial algorithms. We first describe the following well-known universal prediction method.

(1) *Closest Genotype Neighbor*. For the genotype  $g_n$ , find the closest genotype  $g_i$  using Hamming distance and then set  $s(g_n) \leftarrow s(g_i)$ .

(2) *The LOD-based Prediction Algorithm*. This is a standard statistics method based on allele frequency. For each SNP  $j = 1, \dots, m$ , we find the allele frequency  $fd_j(0)$  of 0's in the case population. Similarly, for the same SNP  $j$ , we find the allele frequency score  $fd_j(1)$  of 1's and  $fd_j(2)$  of 2's in the case population and the corresponding frequency in the control population ( $fh_j(0), fh_j(1), fh_j(2)$ ). Then we

compute LOD score (likelihood of odds ratio) of each SNP as follows.

$$LOD_j(i) = \log \frac{fd_j(i)}{fh_j(i)}, \quad i = 0, 1, 2$$

For genotype  $g_n$ , if the cumulative LOD score of all SNPs  $\sum_{j=1}^m L_j(g_{n,j})$  is greater than 0, then the output disease status  $s(g_n)$  is set to 1 ( $g_n$  is predicted to be in case population) and -1, otherwise.

(3) *Graph-based Prediction Algorithms*. The two methods are based on the following *genotype graph*  $X = \{H, G\}$ , where the vertices  $H$  are distinct haplotypes and the edges  $G$  are genotypes each connecting its two haplotypes (vertices).

When applying graph heuristics to  $X$ , we found that it is necessary to increase the density of  $X$ . This can be achieved by dropping certain SNPs (or, equivalently, keeping only certain tag SNPs). Indeed, dropping a SNP may result in collapsing of certain vertices in  $X$ , i.e., different vertices become identical. Collapsing vertices may also result in collapsing certain edges (genotypes). A SNP dropping is not allowed if that results in collapsing edges from case and control populations, but collapsing of edges from the same population is allowed.

A simple greedy strategy consists of (1) traversing all the SNPs and (2) dropping a SNP if it is allowed that will result in keeping a minimal subset of SNPs which do not collapse genotypes from opposite populations. Unfortunately, in the original graph  $X$  we may already have collapsed edges from opposite populations - in fact, Daly *et al* data contain such pair of genotypes. In this case we just remove the both genotypes in training set. Our experiments show that on average, we are left with 22 tag SNP's out of 103 for Daly *et al* [3] data and 34 tag SNP's out of 108 for Ueda *et al* [13] data.

After collapsing the graph  $X$  we add the edge corresponding to the test-case genotype  $g_n$ . If the edge  $g_n$  collapses with another edge  $g_i$ , then we set the predicted disease status  $s(g_n) = s(g_i)$ . Otherwise, we apply one of the following two methods for computing the disease status  $s(g_n)$ .

*First Neighbor*:  $s(g_n)$  attains 1 if

$$\sum_{e \text{ adjacent to } g_n} s(e) > 0$$

and -1, otherwise. In other words, the predicted disease status is decided by voting among all adjacent edges.

*Second Neighbor*:  $s(g_n)$  attains 1 if

$$\sum_{e \text{ adjacent to } g_n} \left( s(e) - \frac{\sum_{e' \text{ adjacent to } e} s(e')}{\delta(e)} \right) > 0$$

and -1, otherwise, where  $\delta(e)$  is the number of edges adjacent to  $e$ . This method enhances the First Neighbor algorithm by taking in account the second neighbors.

(4) *LP-based Prediction Algorithm*. This method assumes that certain haplotypes are susceptible to the disease while others are resistant to the disease. The genotype susceptibility is then assumed to be a sum of susceptibilities of its two haplotypes.

We want to assign a positive weight to susceptible haplotypes and a negative weight to resistant haplotypes such that for any control genotype the sum of weights of its haplotypes is negative and for any case genotype it is positive. We would also like to maximize the confidence of our weight assignment which can be measured by the absolute values of the genotype weights. In other words, we would like to maximize the sum<sup>1</sup> of absolute values of weights over all genotypes.

Formally, for each vertex  $h_i$  (corresponding to haplotype) of the graph  $X$  we wish to assign the weight  $p_i$ ,

$$-1 \leq p_i \leq 1 \quad (1)$$

such that for any genotype-edge  $e_{ij} = (h_i, h_j)$ ,

$$s(e_{ij})(p_i + p_j) \geq 0 \quad (2)$$

where  $s(e_{ij}) \in \{-1, 1\}$  is the disease status of genotype represented by edge  $e_{ij}$ .

The total sum of absolute values of genotype weights is maximized

$$\sum_{e_{ij}=(h_i, h_j)} s(e_{ij})(p_i + p_j) \quad (3)$$

The formulation (1-3) is a linear program which can be efficiently solved by a standard linear program solver such as CPLEX [18] or LPSolve [19].

For the left-out testing genotype  $g_n$ , we compute the sum of weights of its haplotypes. If the sum is strictly positive, the genotype is attributed to the case, otherwise it is attributed to the control. Figure I

(5) *Combined Prediction Algorithm.* The disease status of the left-out testing genotype  $g_n$  is assigned as follows: if LP-based finds non-zero sum of haplotype weights for  $g_n$ , then  $s(g_n)$  is assigned accordingly. Otherwise,  $s(g_n)$  is assigned according to LOD-based prediction algorithm.

### III. RESULTS

In this section we first describe the testing framework for evaluation quality of prediction algorithms, then we provide experimental results over all prediction algorithms described in the previous section.

#### (1) *Estimating the Quality of Prediction Algorithms*

We have applied leave-one-out cross-validation to evaluate the quality of susceptibility-predicting algorithms as follows. We predict the disease status of each genotype in the given data set by applying the susceptibility-predicting algorithm to the rest of the data which is regarded as the training set. Then we compare the predicted susceptibility with the actual disease status. We report the prediction rate separately for cases and controls as well as for the entire population.

For verification purposes we also perform the following Monte-Carlo random test (MCT). The original association of  $n$  genotypes with markers is randomly scrambled, i.e.,

<sup>1</sup>Instead, we may maximize minimum absolute value over all genotype weights. Our experiments show that the results are quite similar for the both objectives.

we repeatedly randomly swap case and control markers and then run the prediction algorithm. The expected prediction rate should be 50% corresponding to random data prediction. Such test confirms that the programming implementation does not rely on illegal information, and that the genotype data contain disease susceptibility.

The confidence level of obtained data is confirmed by bootstrapping. We have performed a specified number of random samplings with replacement from the original data and then run our algorithms on each of these samples. The reported prediction rate is the worst observed in 95% of samples of the resulted distribution.

Another way of bootstrapping is to randomly choose 20 cases and 200 controls as training set and predict others, and repeat these samplings 100 times reporting the average prediction rate.

#### (2) *Data Sets*

The data set Daly *et al* [3] is derived from the 616 kilobase region of human Chromosome 5q31 that may contain a genetic variant responsible for Crohn's disease by genotyping 103 SNPs for 129 trios. All offspring belong to the case population, while almost all parents belong to the control population. In entire data, there are 144 case and 243 control individuals.

The data set of Ueda *et al* [13] are sequenced from 330kb of human DNA containing gene CD28, CTLA4 and ICONS which are proved related to autoimmune disorder. A total of 108 SNPs were genotyped in 384 cases of autoimmune disorder and 652 controls.

#### (3) *Experimental Results*

For those haplotype-based methods, we use two different phasing algorithms PHASE[10] and GERBIL[9] to infer population haplotypes, but report result of GERBIL only, because PHASE gives similar result. The best prediction rate achieved by the combinatorial method for Crohn's disease and autoimmune disorder are 77.78% and 64.99%, respectively. From Table I we can find that our graph-based prediction algorithm give a better result than traditional statistics methods, and even better if we combine both of them.

The data set of Ueda *et al* [13] is unrelated case/control population, so the haplotype structure is much more complex than that of Daly *et al* [3], which are described in family trios. For haplotype-based prediction methods, the complexity will affect the prediction rate. As a result, the prediction rate for Daly *et al* is higher, as of 77.78% , while for Ueda *et al*, is 64.99% only.

The Figure 1 shows the distribution of the genotype weights for the LP-based prediction algorithm. The height of columns is proportional to the number of cases (positive height) and the number of controls (negative height). It is easy to see that cases prefer the right side while the controls prefer to be on the left side of the distribution.

In Table II we report bootstrapping rates, i. e., the 5th worst rate out of 100 runs (95% confidence) and different bootstrapping rates – averaged over 100 random samplings of 20 case and 200 control genotypes.

TABLE I

THE COMPARISON OF THE PREDICTION RATES OF 6 PREDICTION METHODS FOR CROHN'S DISEASE (DALY *et al*)[3] AND AUTOIMMUNE DISORDER (UEDA *et al*) [13]. GENOTYPE DATA ARE PHASED BY GERBIL [9].

Data Set	Population	Prediction Methods					
		Closest Neighbor	LOD	First Neighbor	Second Neighbor	Linear Programming	Combined Method
(Daly <i>et al</i> )	Case	54.17	47.22	67.39	67.13	62.19	86.11
	Control	58.85	64.20	57.38	56.96	84.72	72.84
	<b>Total</b>	<b>57.11</b>	<b>57.88</b>	<b>61.06</b>	<b>62.44</b>	<b>76.24</b>	<b>77.78</b>
(Ueda <i>et al</i> )	Case	43.75	56.51	22.66	41.15	30.47	65.97
	Control	65.95	54.75	83.74	64.42	84.97	64.42
	<b>Total</b>	<b>57.72</b>	<b>55.41</b>	<b>61.10</b>	<b>55.79</b>	<b>64.77</b>	<b>64.99</b>

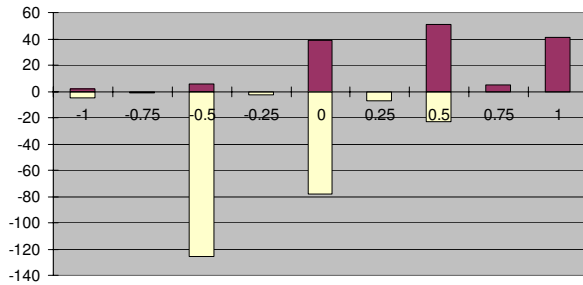


Fig. 1. Distribution of the genotype weights for the Linear Programming prediction algorithm. The dark columns over the median horizontal line correspond to the numbers of cases with the genotype weight in the range specified by the  $x$ -axis. The light columns below the median horizontal line correspond to the numbers of controls within respective genotype weight range.

#### IV. CONCLUSIONS

Recent improvements in the accessibility of high-throughput genotyping has brought a great deal of attention to disease association and susceptibility studies. Unfortunately, the current statistical methods do not reliably predict susceptibility for complex diseases. In this paper, we suggest a simple framework for estimation of the prediction quality of the algorithms predicting susceptibility to genetic diseases and a bootstrapping strategy for estimating the significance of the results obtained from prediction methods. We apply our methods to Crohn's case/control data of Daly *et al* and Autoimmune disorders and achieve a correct prediction rate of 77.78% and 64.99%, respectively. After performing a Monte-Carlo test by running our methods on the original data with randomly permuted case/control markers, the average prediction rate falls to 50% for all proposed methods. After applying bootstrapping, we obtain with 95% confidence the correct prediction rate of 75.71%.

#### REFERENCES

[1] Zhang, K., Calabrese, P., Nordborg, M., Sun, F. (2002) 'Haplotype Block Structure and Its Applications in Association Studies: Power and Study Design', *The American Journal of Human Genetics*, 71:1836-1894.  
 [2] The International HapMap Project, <http://www.hapmap.org>  
 [3] Daly, M., Rioux, J., Schaffner, S., Hudson, T. and Lander, E. (2001) 'High Resolution Haplotype Structure in the Human Genome', *Nature Genetics*, 29:229-232.

TABLE II

THE COMPARISON OF THE PREDICTION RATES OF TWO PREDICTION METHODS (LINEAR PROGRAMMING AND COMBINED METHOD) ON DALY *et al*. WE REPORT BOOTSTRAPPING RATES, I. E., THE 5TH WORST RATE OUT OF 100 RUNS (95% CONFIDENCE) AND DIFFERENT BOOTSTRAPPING RATES – AVERAGED OVER 100 RANDOM CHOICES OF 20 CASE AND 200 CONTROL GENOTYPES.

Population	Prediction Methods			
	Linear Programming		Combined Method	
	95%	200/20	95%	200/20
Case %	58.04	56.65	83.33	77.78
Control %	82.71	80.66	71.19	72.84
Total %	73.58	71.76	75.71	74.67

[4] Gabriel, G., Schaffner, S., Nguyen, H., Moore, J., Roy, J., Blumenstiel, B., *et al.* (2002) 'The Structure of Haplotype Blocks in the Human Genome', *Science*, 296:2225-2229.  
 [5] Johnson, G.C.L., Esposito, L., Barratt, B.J., Smith, A.N., Heward, *et al.* (2001) 'Haplotype Tagging for the Identification of Common Disease Genes', *Nature Genetics*, 29:233-237.  
 [6] Clark AG. (2003) 'Finding Genes Underlying Risk of Complex Disease by Linkage Disequilibrium Mapping', *Curr Opin Genet Dev.*, 13(3):296-302.  
 [7] Zhao, H., Pfiffer, R. and Gail, MH. (2003) 'Haplotype Analysis in Population Genetics and Association Studies', *Pharmacogenomics*, 4:171-178.  
 [8] Goldstein, D. and Weale, M. (2001) 'Population Genomics: Linkage Disequilibrium Holds the Key', *Current Biology*, 11:576-579.  
 [9] Kimmel, G. and Shamir, R. (2005) GERBIL: Genotype Resolution and Block Identification Using Likelihood', *Proceedings of the National Academy of Sciences*, 102:158-162.  
 [10] Stephens, M., Smith, N.J., and Donnelly, P. (2001) 'A New Statistical Method for Haplotype Reconstruction from Population Data', *The American Journal of Human Genetics*, 68:978-988.  
 [11] Merikangas, KR., Risch, N. (2003) 'Will the Genomics Revolution Revolutionize Psychiatry', *The American Journal of Psychiatry*, 160:625-635.  
 [12] Botstein, D., Risch, N. (2003) 'Discovering Genotypes Underlying Human Phenotypes: Past Successes for Mendelian Disease, Future Approaches for Complex Disease', *Nature Genetics*, 33:228-237.  
 [13] Ueda, H., Howson, J.M.M., Esposito, L. *et al.* (2003) 'Association of the T Cell Regulatory Gene CTLA4 with Susceptibility to Autoimmune Disease', *Nature*, 423:506-511.  
 [14] Anderson, M. (2001) 'Crohn's: An Autoimmune or Bacteria-Related Disease?', *The Scientist*, 22:15-16.  
 [15] Serretti, A. and Smeraldi, E. (2004) 'Neural network analysis in pharmacogenetics of mood disorders', *BMC Medical Genetics*, 5:27.  
 [16] Vapnik, V.N. (1995) 'The Nature of Statistical Learning Theory', *Springer*.  
 [17] Basic Classification Trees, <http://www.ece.wisc.edu>  
 [18] ILOG CPLEX, <http://www.ilog.com>  
 [19] LPSolve, <http://www.netlib.org>