

Genotype Susceptibility And Integrated Risk Factors for Complex Diseases

Weidong Mao, Dumitru Brinza, Nisar Hundewale, Stefan Gremalschi and Alexander Zelikovsky

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 discrete optimization methods to predict the genotype susceptibility for complex disease. The proposed combinatorial methods have been applied to publicly available genotype data on Crohn’s disease and autoimmune disorders for predicting susceptibility to these diseases. The result of predicted status can be also viewed as an integrated risk factor. The quality of susceptibility prediction algorithm has been assessed using leave-one-out and leave-many-out tests and shown to be statistically significant based on randomization tests. The best prediction rate achieved by the prediction algorithms is 69.5% for Crohn’s disease and 63.9% for autoimmune disorder. The risk rate of the corresponding integrated risk factor is 2.23 for Crohn’s disease and 1.73 for autoimmune disorder.

Index Terms—complex diseases, prediction methods, susceptibility, risk factors, genotypes.

I. INTRODUCTION

RECENT improvement in accessibility of high-throughput genotyping brought a great deal of attention to disease association and susceptibility studies[20]. High density maps of single nucleotide polymorphism (SNPs)[11] as well as massive genotype data with large number of individuals and number of SNPs become publicly available[8], [9], [12]. A catalogue of all human SNPs is hoped to allow genome-wide search of SNPs associated with genetic diseases.

Success stories when dealing with diseases caused by a single SNP or gene were reported. But 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[16], [2]. In general, a single SNP or gene may be impossible to associate because a disease may be caused by completely different modifications of alternative pathways. Furthermore, there are no reliable tools applicable to large genome ranges that could rule out or confirm association with a disease. It is even difficult to decide if a particular disease is genetic, e.g., the nature of Crohn’s disease has been disputed [1]. Although answers to above questions may not explicitly help to find specific disease-associated SNPs, they may be critical for

disease prevention. Indeed, knowing that an individual is (or is not) susceptible to (or belong to a risk group for) a certain disease will allow greatly reduce the cost of screening and preventive measures or even help to completely avoid disease development, e.g., by changing a diet.

Disease association analysis is usually searching for a SNP which can be a statistically significant risk factor. In epidemiology, the importance of the integrated risk factor is commonly measured by risk rates. The risk rate is the ratio of disease incidence proportion in the population exposed to the risk factor to that in the non-exposed population. Unfortunately, the traditional direct statistical association so far is unsatisfactory and arguably is not applicable to complex diseases [6]. Even if an individual SNP has a significant relative risk rate it may give only negligible absolute increase of probability, e.g., from 10 in a million to 20 in a million. Note that the cumulative power of several SNPs is difficult to assess because of SNP linkage. So it would be desirable to have a tool that would integrate different genetic risk factors resulted in high disease prediction rate and high risk rate.

This study is devoted to the problem of assessing accumulated information targeting to predict genotype susceptibility to complex diseases with significantly high accuracy and statistical power. In this paper, we first give several discrete optimization based algorithms for prediction disease susceptibility. We then compare leave-one- and leave-many-out tests demonstrating that prediction accuracy of suggested methods is sufficiently resilient to discarding case/ control data implying that leave-one-out test is a trustworthy accuracy measure. The randomization techniques have been used for computing the statistical significance level of proposed methods and resulted prediction weights. We show that prediction rate and statistical significance are well correlated.

The proposed methods are applied to two publicly available data: Crohn’s disease [8] and autoimmune disorder [18]. In the leave-one-out cross-validation tests the proposed linear programming (LP) based method achieves prediction rate of 69.5%(p-value below 2%) and 61.3%(p-value below 62%) and the risk rates of 2.23 and 0.98, respectively. We also show that SVM methods used in [14], [19] are not much worse than our proposed LP-based method.

The next section formally defines the problem and describes several universal and adhoc methods for predicting genotype susceptibility to complex diseases. Section III describes real case/control data sets, discusses prediction and risk rate measures and compares results for several susceptibility prediction methods. We draw the conclusion in the last section.

Manuscript received December 30, 2005. This work was supported in part by the NIH Award 1 P20 GM065762-01A1. Dumitru Brinza was supported by GSU Molecular Basis of Disease Fellowship.

Weidong Mao, Dumitru Brinza, Nisar Hundewale and Stefan Gremalschi are with the Department of Computer Science, Georgia State University, Atlanta, GA 30303, USA {wmao, dima, nhundewale, stefan}@cs.gsu.edu.

Alexander Zelikovsky is with the Department of Computer Science, Georgia State University, Atlanta, GA 30303, USA (phone: (404) 651-0676; fax: (815) 642-0052; e-mail: alexz@cs.gsu.edu).

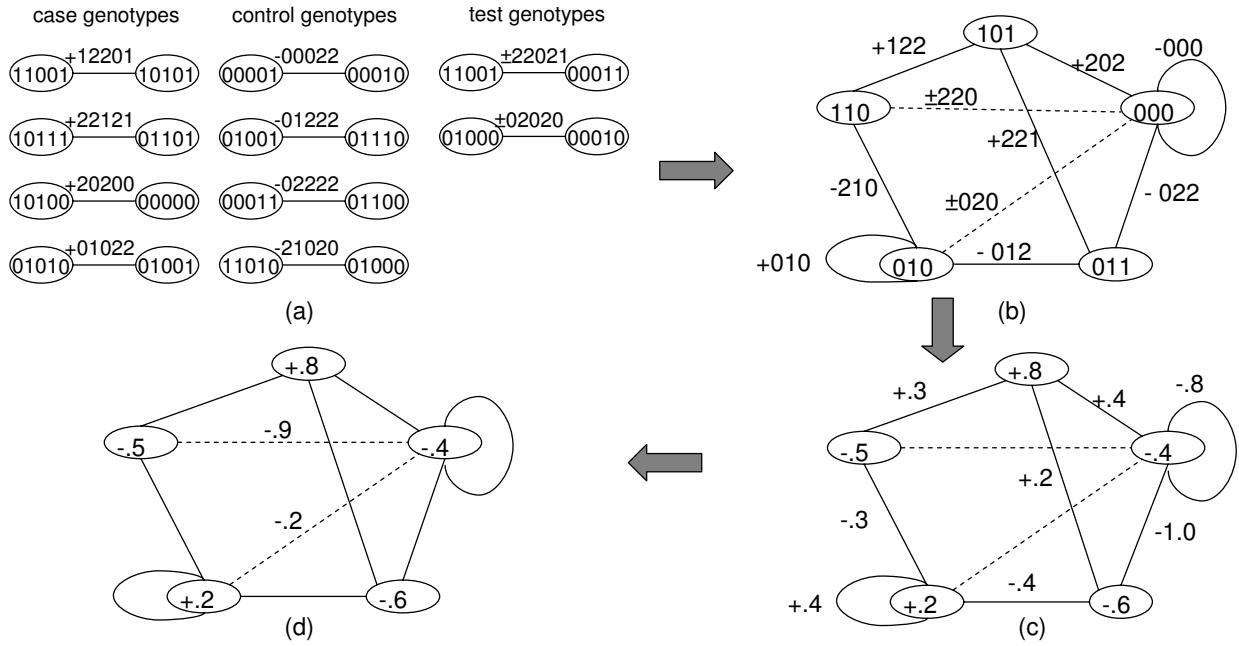


Fig. 1. LP-based Prediction Method. (a) The set of case, control and test genotypes are phased resulting in the sparse graph with vertices-haplotypes and edges-genotypes. (b) The last two SNPs are dropped without collapsing case and control edges resulting in a denser graph. (c) The LP finds optimal weights for vertices-haplotypes. (d) The status of test genotypes is predicted from the sign of the sum of weights of their endpoints.

II. PREDICTION METHODS FOR GENOTYPE SUSCEPTIBILITY

In this section we first describe the input and the output of prediction algorithms and how to predict genotype susceptibility. We then describe several universal and adhoc prediction methods.

Specifications of prediction 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., $s(g_n)$.

Below we describe several universal prediction methods. These methods are adaptations of general computer-intelligence classifying techniques.

Closest Genotype Neighbor (CN). For the test genotype g_t , find the closest (with respect to Hamming distance) genotype g_i in the training set, and set the status $s(g_t)$ equal to $s(g_i)$.

Support Vector Machine Algorithm (SVM). Support Vector Machine (SVM) is a general learning system based on recent advances in statistical learning theory. SVMs deliver state-of-the-art performance in real-world applications and have been used in case/control studies [14], [19]. We use SVM-light [5] with the radial basis function with $\gamma = 0.5$.

Random Forest (RF). A random forest is a collection of CART-like trees following specific rules for tree growing, tree combination, self-testing, and post-processing. We use Leo Breiman and Adele Cutler's original implementation of RF version 5.1 [3]. This version of RF handles unbalanced data to predict accurately. RF tries to perform regression on the specified variables to produce the suitable model. RF uses bootstrapping to produce random trees and it has its own cross-validation technique to validate the model for prediction/classification.

CDPG: Tomita [17] introduced the Criterion of Detecting Personal Group (CDPG) for extracting risk factor candidates (RFCs). RFCs are extracted using binomial test and random permutation tests. CDPG performs exhaustive combination analysis using case/control data and assumes the appearance of case and control subjects belonging to a certain rule as a series of Bernoulli trials, where two possible outcomes are case and control subjects with some probabilities.

We now describe two ad hoc prediction methods (i.e., classifying techniques taking in account the nature of the classification problem). The first method is 2-SNP method [13] and the second method is a variation of the LP-based method [15].

Most Reliable 2 SNP Prediction [13] (MR). This method chooses a pair of adjacent SNPs (site of s_i and s_{i+1}) to predict the disease status of the test genotype g_t by voting among genotypes from training set which have the same SNP values as g_t at the chosen sites s_i and s_{i+1} . They chose the 2 adjacent SNPs with the highest prediction rate in the training set.

LP-based Prediction Algorithm (LP). This method 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) (see Figure 1(a)).

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). Discarding a SNP is not allowed if it results in collapsing edges from case and control populations, but collapsing of edges from the same population is allowed (see Figure 1(b)).

A simple greedy strategy consists of traversing all the SNPs and dropping a SNP if it is allowed. The resulted set of SNPs is a minimal subset of SNPs which do not collapse genotypes from opposite disease status. 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. Only such original collapsing is allowed - the status of such edges is assumed to be the one of majority of genotypes. Our experiments show that on average, we are left with 21 tag SNP's out of 103 for Daly *et al* [8] data and 29 tag SNP's out of 108 for Ueda *et al* [18] data (see description of the next section). The selected set tag SNPs are better candidates for being disease associated, in fact only such tag SNPs were used in the prediction methods with the highest accuracy.

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 $s(g_n)$. The LP-based 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 (see Figure 1(c)). 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 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 , $-1 \leq p_i \leq 1$ such that for any genotype-edge $e_{ij} = (h_i, h_j)$, $s(e_{ij})(p_i + p_j) \geq 0$ 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

$$\sum_{e_{ij}=(h_i, h_j)} s(e_{ij})(p_i + p_j) \quad (1)$$

The above formulation with the objective (1) is the linear program which can be efficiently solved by a standard linear

TABLE I
CONFUSION TABLE.

	True Data (Golden Standard)		
	Case	Control	
pCS	True Positive TP	False Positive FP	Positive Prediction Rate PPR= TP/(TP+FP)
pCO	False Negative FN	True Negative TN	Negative Prediction Rate NPR= TN/(FN+TN)
	Sensitivity TP/(TP+FN)	Specificity TN/(FP+ TN)	Accuracy (TP+TN)/(TP+FP+FN+TN)

program solver such as GNU Linear Programming Kit (GLPK) [10].

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, if the sum is strictly negative, it is attributed to the control (see Figure 1(d)), otherwise $s(g_n)$ is assigned according to 2-SNP prediction algorithm [13].

III. QUALITY OF SUSCEPTIBILITY PREDICTION METHODS

In this section we describe the two real case/control population samples and the results of leave-one-out and leave-many-out cross-validation tests estimating susceptibility prediction methods on these sets.

Data Sets. The data set Daly *et al* [8] 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 missing genotype data and haplotypes have been inferred using 2SNP phasing method [4]. The highest risk rate for single SNP is 2.7.

The data set of Ueda *et al* [18] 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. Similarly, the missing genotype data and haplotypes have been inferred. The highest risk rate for single SNP is 1.9.

Cross-validation Tests. In the leave-one-out cross-validation, the disease status of each genotype in the data set is predicted while the rest of the data is regarded as the training set. In the leave-many-out cross-validation, n individuals are uniformly at random picked up from the data set, marked and put back, where n is the size of the data set. This way, approximately 2/3 of the individuals are picked at least once and marked while the rest will not be marked. The training set consists of marked data and the testing set consists of unmarked data.

Quality Measures. In cross-validation tests, the predicted and the actual disease statuses are compared and the standard confusion matrix is filled (see Table I). Predicted cases and predicted controls are notated by pCS and pCO respectively. We report sensitivity, specificity, and accuracy of the prediction methods. We also report the the risk rate of the corresponding integrated risk factor associated with each prediction method. It is computed as the the ratio of the probability of developing

disease among those predicted susceptible to the probability of developing disease among those predicted non-susceptible [7]:

$$\text{Risk Rate} = \frac{TP}{TP + FP} / \frac{FN}{TN + FN}$$

We report the 95% confidence intervals (CI) for accuracy and risk rate, for leave-one-out test 95% CI is computed using bootstrapping. We also compute significance level, p -value, for the accuracy of prediction algorithms computed using 5000 randomized instances. On the randomized instances, the average prediction rate for SVM and RF has been 60% and for all other methods except has been 50%.

Results and Discussion. Table II compares 6 different prediction methods for both data sets. Column C denotes performed cross-validation tests, LOO stays for leave-one-out test and LMO stays for leave-many-out test. For leave-one-out test, the best accuracy is achieved by LP – 69.5% on Daly *et al.* [8] data and by MR – 63.9% on Ueda *et al.* [12] data. For leave-many-out test, the accuracy only slightly degrades showing resiliency to the size of the data. The risk rates for the integrated risk factor associated with prediction methods are comparable with risk rates for individual SNPs – for the first data set, 2.23 (LP method) vs 2.7 and for the second data set, 1.73 (RF method) and 1.64 (MR method) vs 3.2. The good performance of SVM and certain other universal methods indicate that they can possibly be adjusted to improve specific ad hoc methods for prediction of susceptibility to complex diseases.

IV. CONCLUSION

In this paper, we discuss motivation behind the genotype susceptibility studies. The developed ad hoc susceptibility prediction method based on linear programming is shown to have high prediction rates and high relevant risk rate for associated integrated risk factors for two completely different case/control studies for Crohn's disease [8] and autoimmune disorders [12]. The extensive computational results show great potential of the proposed prediction methods. In our future work we are going to continue validation of the proposed method.

REFERENCES

- [1] Anderson, M. (2001) 'Crohn's: An Autoimmune or Bacteria-Related Disease?', *The Scientist*, 22:15-16.
- [2] 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.
- [3] Breiman, L. and Cutler, A. <http://www.stat.berkeley.edu/users/breiman/RandomForests/>
- [4] Brinza, D. and Zelikovsky, A. (2006) 2SNP: Scalable Phasing Based on 2-SNP Haplotypes, *Bioinformatics*, 22(3):371-3.
- [5] Joachims, T. <http://svmlight.joachims.org/>
- [6] Clark AG. (2003) 'Finding Genes Underlying Risk of Complex Disease by Linkage Disequilibrium Mapping', *Curr Opin Genet Dev.*, 13(3):296-302.
- [7] Clinical Epidemiology Glossary, <http://www.med.ualberta.ca/ebm/define.htm>.
- [8] 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.
- [9] 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.

TABLE II

THE COMPARISON OF SENSITIVITY, SPECIFICITY, ACCURACY AND RISK RATE WITH 95% CONFIDENCE INTERVALS (CI) AND p -VALUE FOR 6 PREDICTION METHODS FOR TWO REAL DATA SETS.

		Daly <i>et al.</i> [8]					
C	Quality measure	Prediction Methods					
		CN	SVM	RF	CDPG	MR	LP
L O O	sensitivity	45.5	20.8	34.0	68.8	30.6	37.5
	specificity	63.3	88.8	85.2	58.0	85.2	88.5
	accuracy	54.6	63.6	66.1	62.2	65.5	69.5
	95%-CI	±.9	±.5	±.6	±.8	±.9	±.5
	p -value	0.03	0.04	0.30	0.04	0.03	0.01
	risk rate	1.25	1.52	1.83	1.49	2.00	2.23
L M O	95%-CI	±.09	±.04	±.03	±.02	±.02	±.05
	sensitivity	45.9	18.0	30.0	59.7	28.0	36.0
	specificity	54.0	89.3	82.2	55.6	76.5	82.3
	accuracy	52.2	62.9	64.2	57.1	58.5	68.4
	95%-CI	±.9	±.5	±.5	±.9	±.9	±.05
	risk rate	0.99	1.45	1.67	1.47	1.15	2.01
	95%-CI	±.06	±.26	±.12	±.01	±.01	±.01

		Ueda <i>et al.</i> [12]					
C	Quality measure	Prediction Methods					
		CN	SVM	RF	CDPG	MR	LP
L O O	sensitivity	37.7	14.3	18.0	58.6	6.9	7.1
	specificity	64.5	88.2	92.8	61.7	97.2	91.2
	accuracy	54.8	60.9	65.1	60.5	63.9	61.3
	95%-CI	±.9	±.3	±.4	±.8	±.9	±.3
	p -value	0.04	0.70	0.73	0.05	0.04	0.62
	risk rate	1.05	1.15	1.73	1.67	1.64	0.86
L M O	95%-CI	±.01	±.03	±.03	±.01	±.01	±.03
	sensitivity	34.8	12.7	13.4	56.0	7.2	8.0
	specificity	64.8	90.5	83.5	56.9	89.4	82.7
	accuracy	53.4	61.8	62.4	56.6	58.4	59.3
	95%-CI	±.9	±.3	±.3	±.9	±.9	±.6
	risk rate	0.98	1.22	1.25	1.38	0.76	0.98
	95%-CI	±.06	±.03	±.03	±.01	±.01	±.01

- [10] GLPK (2000). GNU Linear Programming Kit. <http://www.gnu.org>.
- [11] The International HapMap Project, <http://www.hap.map.org>
- [12] 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.
- [13] Kimmel, G. and Shamir R. (2005) A Block-Free Hidden Markov Model for Genotypes and Its Application to Disease Association. *J. of Computational Biology*, Vol. 12, No. 10: 1243-1260.
- [14] Listgarten, J., Damaraju, S., Poulin B., Cook, L., Dufour, J., Driga, A., Mackey, J., Wishart, D., Greiner, R., and Zanke, B. (2004) Predictive Models for Breast Cancer Susceptibility from Multiple Single Nucleotide Polymorphisms. *mphClinical Cancer Research*, Vol. 10, 2725-2737, 2004.
- [15] Mao, W., He, J., Brinza D. and Zelikovsky, A. (2005) 'A Combinatorial Method for Predicting Genetic Susceptibility to Complex Diseases', *Proc. International Conf. of the IEEE Engineering in Medicine and Biology (EMBC'05)*, to appear.
- [16] Merikangas, K.R., Risch, N. (2003) 'Will the Genomics Revolution Revolutionize Psychiatry', *The American Journal of Psychiatry*, 160:625-635.
- [17] Tomita, Y., Yokota, M. and Honda, H. (2005) Classification method for prediction of multifactorial disease development using interaction between genetic and environmental factors, *IEEE computational systems bioinformatics conference*, abstract.
- [18] 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.
- [19] Waddell, M., Page, D., Zhan, F., Barlogie, B., and Shaughnessy, J. (2004) Predicting Cancer Susceptibility from Single Nucleotide Polymorphism Data: A Case Study in Multiple Myeloma. *Proceedings of BIOKDD 2005*, 05, 2005.
- [20] 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.