

## 1. Introduction

- Recent improvement in accessibility of high-throughput genotyping brought a great deal of attention to disease association and susceptibility studies.
- High density maps of single nucleotide polymorphism (SNPs) as well as massive genotype data with large number of individuals and number of SNPs become publicly available.
- The main goal of disease association analysis is to identify gene variations or, in general, haplotypes which contribute to the risk of a particular disease.
- The linkage disequilibrium (LD) approach has been widely used in disease association studies based on individual SNP.
- A single SNP or gene may be impossible to associate because a disease may be caused by completely different modifications of alternative pathways.
- More attempts treat multi-marker haplotypes as the variable for correlation.
- There are some universal methods such as statistics and classification algorithms can be used for diseases prediction.
- We proposed some ad hoc methods for disease susceptibility prediction which have a higher prediction rate.

## 2. Disease Susceptibility Prediction

### Training Set

- Training genotype set  $g_i = (g_{ij}), i = 0, \dots, n-1; j = 1, \dots, m, g_{ij} \in \{0, 1, 2\}$ .
- 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).

### Test

- Test genotype  $g$ , with unknown disease status.

### Output

- A Disease status of the test genotype  $s(g)$ .
- The input data can also be phased, then each genotype is represented by pair of haplotypes
- Genotypes are phased to haplotypes through PHASE or GERBIL.

### DISEASE TAGGING

- Genotype and haplotype may contain redundant information or "noise".
- Disease tagging = find important SNPs or subset of genotype/haplotype which are responsible for diseases.

#### Genotype Tagging

- Find minimum number of SNPs (tags) such that for any pair of case and control genotypes there exists at least one tag where these two genotypes are different.

#### Tagging Pairs of Haplotype

- Phase each genotype  $g = \{h1, h2\}$ .
- Find minimum number of SNPs (tags) such that for any pair of case and control genotypes  $g = \{h1, h2\}$  and  $g' = \{h1', h2'\}$ , there exists at least one tag for which either  $h1\{t\} \neq h1'\{t\}$  and  $h1\{t\} \neq h2'\{t\}$  or  $h2\{t\} \neq h1'\{t\}$  and  $h2\{t\} \neq h2'\{t\}$
- The tagging problem can be resolved by greedy algorithm or linear program

### UNIVERSAL METHODS

#### Closest Neighbors

For the test genotype  $g$ , using Hamming distance to find the closest genotype  $g_i$  with in training set, and then set  $s(g) = s(g_i)$ .

### Support Vector Machine (SVM)

SVM is a new generation learning system based on recent advances in statistical learning theory. SVMs deliver state-of-the-art performance in real-world applications such as text categorization, hand-written character recognition, image classification, bioinformatics.

### Random Forest

Random Forest is an algorithm for classifying. It uses large number of individual decision trees and decides the class by choosing the mode (most frequently occurring) of the classes as determined by the individual trees.

### AD HOC METHODS

#### LP – based Prediction Algorithm

- Assumption: certain haplotypes are susceptible to the disease while others are resistant to the disease.
- The genotype susceptibility is assumed to be a sum of susceptibilities of its two haplotypes.
- 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.
- Linear program: 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  $e_{ij}$

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

The total sum of absolute values of genotype weights is maximized.

- Linear program (1)-(3) can be solved by CPLEX.
- If LP-based algorithm finds non-zero sum of haplotype weights for  $g_r$  then  $s(g)$  is assigned accordingly. Otherwise,  $s(g_r)$  is assigned according to the following prediction algorithm.

#### 2 SNPs

- This method chooses a pair of adjacent SNPs (site of  $S_j$  and  $S_{j+1}$ ) to predict the disease status of the test genotype  $g_r$  by voting among genotypes from training set which have the same SNP values as  $g_r$  at the chosen sites  $S_j$  and  $S_{j+1}$ .
- There are two ways to choose the pair of SNPs.

#### Most reliable 2 SNPs

- 2 SNPs which have a highest prediction rate in training set with the prediction algorithm of majority voting.

#### Most significant 2 SNPs

- 2 SNPs which have the smallest  $p$ -value in training set. The  $p$ -value is computed based on the binomial distribution among case and control population.

$$P_p(n|N) = \binom{N}{n} p^n (1-p)^{N-n}$$

$N$ : the total number of case and control  
 $n$ : number of case or control  
 $p$ : probability of case or control

## 3. Results

### LEAVE-ONE OUT TEST

- 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 to get the prediction rate.

Data Set	Population	Prediction Methods					
		Closest Neighbor	SVM	Random Forest	LP	MR	MS
I	Sensitivity	54.17	72.22	69.44	86.11	68.75	70.14
	Specificity	58.85	62.55	62.14	72.84	63.38	64.61
	<b>Total</b>	<b>57.11</b>	<b>66.15</b>	<b>64.85</b>	<b>77.78</b>	<b>65.38</b>	<b>66.67</b>
II	Sensitivity	43.75	75	66.15	65.97	67.45	78.39
	Specificity	65.95	53.99	64.57	64.42	61.5	62.73
	<b>Total</b>	<b>57.72</b>	<b>61.78</b>	<b>64.96</b>	<b>64.99</b>	<b>63.71</b>	<b>68.53</b>

### STATISTICAL SIGNIFICANCE OF SUSCEPTIBILITY PREDICTION ALGORITHMS

- The significance level for the prediction algorithms is computed based on the randomization test.
- Randomization testing is a computationally extensive way of determining whether the null hypothesis is reasonable.
- Null hypothesis: the observed prediction rate can be obtained just by chance.
- If the prediction rate is significant (evaluated by  $p$ -value), then the null hypothesis is rejected.
- We generate totally 5000 random instance by keeping the same genotypes as in the real instance but with swapped case and control status between 10000 pairs of genotypes randomly chosen from the entire sample.
- $K$  is the number of instances with higher prediction rate than the observed prediction rate on the real instance.
- $P = K / 5000$
- If  $p < 5\%$ , then this provides strong evidence that the null hypothesis is not true;
- If  $p < 1\%$ , then the evidence is much stronger.
- The one-side  $p$ -value for LP method on data set I and II is 0.02 and 0.04, respectively.

## 4. Conclusions

- Improvement in accessibility of high-throughput genotyping brought a great deal of attention to disease association and susceptibility studies.
- Complex diseases are associated with multi-markers (haplotypes).
- Statistics methods for single-marker (SNP) are not applicable to complex disease.
- We introduce some universal methods for classification and disease discrimination.
- We propose several *ad hoc* genetic susceptibility prediction algorithms and Linear Programming algorithms provide a higher prediction rate.
- We improve the statistics methods on susceptibility prediction.
- We apply our methods to Crohn's disease and autoimmune disorders and achieve a correct prediction rate of 77.78% and 68.53%, respectively.
- Leave-one-out and randomization tests are used to validate proposed algorithms.