

Prediction Methods for Inherited Disease

Weidong Mao¹, Jingwu He^{1,2}, Kelly Westbrook¹ and Alexander Zelikovsky^{1,3}

Keywords: Graph, Disease Predicting, Phasing

1 Introduction.

Recent improvements in the accessibility of high-throughput genotyping have brought a great deal of attention to disease association studies[6]. It is believed that more accurate disease association is achieved with inferred haplotypes rather than with directly available genotypes. 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. There are basically two main steps in disease association: (i) the population haplotyping and (ii) the haplotypes associating with diseases. We propose scalable methods for haplotyping family trio data and combinatorial methods for predicting susceptibility for complex diseases. We validated these methods by using a leave-one-out framework to estimate the predictive power of different predicting algorithms.

2 Methods.

We applied an integer linear program (ILP) to Pure-Parsimony phasing genotypes of trios. Compared to the two known phasing ILP formulations for phasing from [3] and [2], our ILP has much smaller variables, constraints and runtime.

Genotype/Haplotype Statistics. This method decide a given genotype is case or control based on the frequency of each allele in case/control.

Neighbor-Joining. This method use hamming distance to find the closest genotype(case or control) as a reference.

Graph Neighbor. The phased data is represented as the graph $X = \{H, G\}$, where vertices H are distinct haplotypes and the edges G are genotypes, each connecting its two haplotypes. Each edge has a case/control marker $m(g_i) \in \{-1, 1\}$, indicating it is healthy (1) or sick (-1). To predict the case/coontrol marker of a given edge, we want to find if the edge collapsed with any edge of X. if so, we can assign marker to this edge. Otherwise, we apply the following method for computing $m(g_i)$.

$m(g_i)$ attains -1 if

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

and 1, otherwise.

Disease Tagging. When applying graph heuristics to X , we found that it is necessary to increase density of X , i.e., average degree of a vertex. This is achieved by dropping certain SNP's (or, equivalently, keeping only certain tag SNP's). Dropping SNP's may result in

¹Department of Computer Science, Georgia State University, Atlanta, GA 30303. E-mail: {weidong,jingwu,kelly,alexz}@cs.gsu.edu.

² Supported by GSU Molecular Basis and Disease Fellowship

³ Partially supported by NIH Award 1 P20

collapsing of edges or vertices. A SNP can be dropped only if no collapsing edges from case and control exist after dropping. Our experiments show that on average, we are left with 20 tag SNP's for data of Daly et al[1].

3 Results.

The proposed methods for disease susceptibility prediction have been applied to Crohn's diseases data of Daly et al[1]. We applied the leave-one-out test for all methods. In this test, we remove the case/control marker for each sample and try to predict the marker by using the rest of population as training data. We also found that phasing results of PHASE[5],(P) and GERBIL[4](G) are not feasible. We fixed solutions to get feasible result of PHASE $P(f)$ and GERBIL $G(f)$. After deleted recombination, we can get result (P^*) and (G^*) for PHASE and GERBIL, respectively.

The best prediction rate for different phasing methods is 75-81% when recombinations are forbidden and 62-70% when recombinations are allowed by PHASE or GERBIL. We confirm the prediction rates by bootstrapping. We randomly sample 20 sick and 200 health genotypes as training data to predict all the others. The average prediction rate of 20/200 bootstrapping is 59.75% for sick population, 74.17% for health population, and 72.96% for total.

Recombination	Forbidden			Allowed			
Prediction Methods	G*	P*	ILP	G	G(f)	P	P(f)
Genotype Statistics	56.12	57.36	57.63	57.88	57.11	58.14	59.17
Haplotype Statistics	56.12	56.19	56.85	56.85	55.81	57.36	56.59
Neighbor-Joining	52.19	57.88	57.11	57.11	50.91	57.11	57.63
Graph Neighbor	77.48	74.65	81.62	62.44	64.51	67.21	70.24

Table 1: The comparison of the prediction rate for all methods

References

- [1] Daly,M., Rioux,J., Schaffner,S., Hudson,T. and Lander,E. (2001). High-resolution haplotype structure in the human genome. Nat Genet 29:229-232.
- [2] Brown,D. and Harrower,I. (2004). A New Integer Programming Formulation for the Pure Parsimony Problem in Haplotype Analysis. Algorithms in Bioinformatics, 4th International Workshop (WABI), 3240 volume of Lecture Notes in Bioinformatics, pp. 254-265.
- [3] Gusfield,D. (2003). Haplotype inference by pure parsimony. In R. Baeza-Yates, E. Chavez, and M. Crochemore, editors, 14'th Annual Symposium on Combinatorial Pattern Matching (CPM 2003), volume 2676 of Springer LNCS, pages 144–155.
- [4] Kimmel,G. and Shamir,R. (2005) GERBIL: Genotype resolution and block identification using likelihood. PNAS.2005; 102: 158-162.
- [5] Stephens,M., Smith, N.J., and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. Am. J. Hum. Genet.,(2001) 68:978-8.
- [6] 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.