Case(Control)-Free Multi-SNP Combinations in Case-Control Studies

Dumitru Brinza[1] **and Alexander Zelikovsky**[1]

1 Introduction

Several genome-wide searches for disease-associated gene variations have been recently reported (Spinola et al, 2006; Herbert et al, 2006). However, complex diseases can be caused by combinations of several unlinked gene variations. Unfortunately, an exhaustive search among all possible corresponding multi-SNPs can be unfeasible even for small number of SNPs let alone the complete genome.

Disease association analysis searches for a SNP with frequency among diseased individuals (cases) considerably higher than among non-diseased individuals (controls). In this work we first explore the problem of searching for *the most disease-associated and the most disease-resistant multi-gene interactions* for a given case-control study. In the previously published work [3] we proposed fast complimentary greedy search (CGS) which finds multi-SNP combinations with non-trivially high association on real data. Here we present two new approaches which are slightly slower than CGS but aimed to find more MSCs associated with a disease. One is k-level CGS (k-CGS) which is a modification of CGS with fixed k SNPs. Second one is k-level Alternating Closure Search (k-ACS) which starts with a set defined by k-length MSC found by the exhaustive search and minimizes its p-value by reducing the subset of controls and keeping the subset of cases as large as possible. We compare proposed methods with CGS on case/control datasets of Crohn's disease (Daly et al, 2001), autoimmune disorder (Ueda et al, 2003), tick-borne encephalitis (Barkash et al, 2006), and rheumatoid arthritis (GAW15 NARAC 18q, 2006).

When dealing with common diseases, it is necessary to search and analyze multiple independent causes each resulted from interaction of multiple SNPs scattered over the entire genome. Exploiting k-CGS and k-ACS methods for searching associated risk and resistance factors, we address the *disease susceptibility prediction problem*. We use greedy version of proposed in [3] optimum clustering formulation to extract independent causes. Then we use model-fitting algorithm that transforms clustering algorithm into susceptibility prediction algorithm. We compare the accuracy of the proposed prediction methods (k-CGSP and k-ACSP) with previously known methods using leave-one-out (LOO) test.

2 Disease Association Search

Risk and resistance factors representing gene variation interaction can be defined in terms of SNPs as follows. A *multi-SNP combination* (MSC) C is a subset of SNP-columns (denoted $snp(C)$) and the values of these SNPs, 0, 1, or 2, where 0 or 1 stands for major or minor allele in homozygous SNPs, and 2 stands for heterozygous SNPs. The subset of individuals-rows whose restriction on columns of $snp(C)$ coincide with values of C is denoted $cluster(C)$.

Maximum Case(Control)-Free Cluster Problem [3]. Find a maximum size cluster C containing only cases or controls.

To solve control- (case-) free cluster problem we propose k-CGS method which starts with clusters defined by each k-SNP combination and greedily add to this combination one by one SNPs with allele value removing a set of genotypes with highest ratio of controls over cases (cases over controls) until all controls (cases) are removed.

An MSC C is closed if it includes all alleles common to all individuals having C. When searching for disease risk factor, the k-ACG method starts with clusters defined by k-SNP combinations and minimizes their p-value by recursively removing from the closure of case-subset the closure of the control-subset and produce a set for the next iteration. The procedure stops when all controls are removed or removing does not minimize the p-value. For resistance factor we follow the same procedure considering that cases are controls and vice versa. The p-values of found MSCs are multiple testing adjusted via randomization.

3 Susceptibility Prediction based on case/control-free MSCs

Disease Susceptibility Prediction Problem. Given a sample population S (a training set) and one more individual $t \notin S$ with the known SNPs but unknown disease status (testing individual), find (predict) the unknown disease status.

Disease Clustering Problem. Given a population sample S, find a partition $\mathcal{P}$ of S into clusters $S = S_1 \cup \ldots \cup S_k$, with disease status 0 or 1 assigned to each cluster S_i, minimizing $entropy(\mathcal{P}) = -\sum_{i=1}^{k} \frac{|S_i|}{|S|} \ln \frac{|S_i|}{|S|}$ for a given bound on the number of individuals who are assigned incorrect status in clusters of the partition $\mathcal{P}$, $error(\mathcal{P}) < \alpha \cdot |\mathcal{P}|$ [3].

Clustering-based Model-Fitting Prediction Algorithm unknown disease status as follows. It sets the status to case or control and then run the corresponding clustering algorithm that covers either the entire or almost all (e.g., 70-95%) individuals. The predicted status corresponds to the clustering with less entropy.

Here we propose two clustering algorithms based on the k-CGS and k-ACS association searches. For given case/control study genotype data and a cluster defined by k-SNP MSC we first find the largest case-free and control-free subclusters. Individuals

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

covered by the largest of this two subclusters are clustered with corresponding disease status and removed from the initial set. We apply the above procedure to the resulted set until 70% to 99% of individuals are covered. The percentage is chosen according to the best prediction in the internal leave-one-out test on training data.

4 Experimental Results

Since we search MSCs among all SNPs, we adjust the computed p-value to multiple testing by generating 10^4 data samples and randomizing the status of individuals. The 500th smallest p-value over all computed p-values corresponds to the multiple testing adjusted p = 0.05.

Table 1: Disease-associated multi-SNPs combinations found by combinatorial search methods.

Search method	SNP combination with minimum p-value			MT-unadjusted p corresponding to adjusted p=0.05	Number of SNP combinations with MT-adjusted p<0.05	runtime sec.
	case frequency	control frequency	unadjusted p-value			
Crohn's disease (Daly et al, 2001) (cases=144, controls=243, SNPs=103)						
CGS	0.06	0.00	1.4×10^{-4}	9.75×10^{-9}	0	0.05
1-CGS	0.06	0.00	1.4×10^{-4}	1.8×10^{-8}	0	2.00
1-ACS	0.19	0.04	4.5×10^{-6}	3.7×10^{-6}	2	1.00
Autoimmune disorder (Ueda et al, 2003) (cases=378, controls=646, SNPs=108)						
CGS	0.02	0.00	3.4×10^{-4}	8.6×10^{-7}	0	0.10
1-CGS	0.05	0.00	4.4×10^{-8}	5.9×10^{-8}	2	3.00
1-ACS	0.43	0.28	9.2×10^{-5}	1.8×10^{-8}	0	2.00
Tick-borne encephalitis (Barkash et al, 2006) (cases=26, controls=65, SNPs=58)						
CGS	0.19	0.00	1.9×10^{-3}	4.4×10^{-5}	0	0.01
1-CGS	0.31	0.00	4.4×10^{-5}	1.2×10^{-5}	0	0.20
1-ACS	0.31	0.00	4.4×10^{-5}	5.6×10^{-5}	1	0.10
Rheumatoid arthritis, (GAW15 NARAC 18q, 2006) (cases=460, controls=460, SNPs=2300)						
CGS	0.13	0.00	9.3×10^{-19}	1.2×0^{-18}	1	1.00
1-CGS	0.14	0.00	2.9×10^{-20}	6.7×10^{-19}	9	2000.00
1-ACS	0.08	0.01	1.6×10^{-8}	2.3×10^{-7}	2	1230.5

The leave-one-out cross-validation test on three real datasets shows that CGSP performs better than k-CGSP and k-ACSP and other previously known prediction algorithms. However, k-CGSP and k-ACSP maintain their accuracies in leave-many-out test when CGSP accuracy drastically falls down by 15-20%.

Table 2: Leave-one-out cross validation results of three methods for three real data sets.

Dataset	Quality	Prediction Methods			
		SVM	CGSP	1-CGSP	1-ACSP
Crohn's disease Daly et al, 2001	sensitivity	20.8	70.8	53.4	72.5
	specificity	88.8	76.6	85.2	81.9
	accuracy	**63.6**	**84.4**	**76.1**	**76.7**
	runtime (h)	3.0	0.08	8.30	5.10
autoimmune disorder Ueda et al, 2003	sensitivity	14.3	86.7	46.3	63.4
	specificity	88.2	92.7	85.7	80.7
	accuracy	**60.9**	**90.6**	**73.5**	**72.2**
	runtime (h)	7.0	0.20	22.70	18.50
tick-borne encephalitis Barkhash et al, 2006	sensitivity	11.4	65.4	56.9	61.9
	specificity	93.2	99.9	88.2	89.8
	accuracy	**72.2**	**90.1**	**86.1**	**84.3**
	runtime (h)	0.2	0.01	0.08	0.04

References

[1] 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.

[2] Barkhash, A., Perelygin, A., Brinza, D., Pilipenko, P., Bogdanova, YU., Romaschenko, A., Voevoda, M. and Brinton, M. (2006) Genetic Resistance to Flaviviruses, *The Fifth Intl. Conf. on Bioinformatics of Genome Regulation and Structure (BGRS'06)*.

[3] Brinza, D. and Zelikovsky, A. (2006) Combinatorial Methods for Disease Association Search and Susceptibility Prediction, *6th Workshop on Algorithms in BioInformatics (WABI 2006)*, **LNB 4175**, 286–297.