Combinatorial Search Methods for Multi-SNP Disease Association

Dumitru Brinza, Jingwu He, and Alexander Zelikovsky
Department of Computer Science, Georgia State University, Atlanta, GA 30303, USA
E-Mail: { dima , jingwu, alexz }@cs.gsu.edu

***Abstract*— Recent improvements in the accessibility of high-throughput genotyping have brought a deal of attention to genome-wide association studies for common complex diseases. Although, such diseases can be caused by multi-loci interactions, locus-by-locus studies are prevailing. Recently, two-loci analysis has been shown promising (Marchini et al, 2005), and multi-loci analysis is expected to find even deeper disease-associated interactions. Unfortunately, an exhaustive search among all possible corresponding multi-markers can be unfeasible even for small number of SNPs let alone the complete genome.**

In this paper we first propose to extract informative (indexing) SNPs that can be used for reconstructing of all SNPs almost without loss (He and Zelikovsky, 2006). In the reduced set of SNPs, we then propose to apply a novel combinatorial method for finding disease-associated multi-SNP combinations (MSCs). Our experimental study shows that the proposed methods are able to find MSCs whose disease association is statistically significant even after multiple testing adjustment. For (Daly et al, 2001) data we found a few unphased MSCs associated with Crohn's disease with multiple testing adjusted p-value below 0.05 while no single SNP or pair of SNPs show any significant association. For (Ueda et al, 2003) data we found a few new unphased and phased MSCs associated with autoimmune disorder.

I. INTRODUCTION

Recent improvement in accessibility of high-throughput DNA sequencing brought a great deal of attention to disease association and susceptibility studies. Successful genome-wide searches for disease-associated gene variations have been recently reported [10], [11]. However, complex diseases can be caused by combinations of several unlinked gene variations. In this paper we address computational challenge of searching such gene combinations.

Disease association studies analyze genetic variation across exposed to a disease and healthy individuals. The difference between individual DNA sequences occurs at a single-base sites, in which more than one allele is observed across population. Such variations are called single nucleotide polymorphisms (SNPs). The number of simultaneously typed SNPs for association and linkage studies is reaching 250,000 for SNP Mapping Arrays [1]. High density maps of SNPs as well as massive DNA data with large number of individuals and number of SNPs become publicly available [2]. Diploid organisms, like human, have two near-identical copies of each chromosome. Most genotyping techniques (e.g., SNP Mapping Arrays [1]) do not provide separate SNP sequences (*haplotypes*) for each of the two chromosomes. Instead, they provide SNP sequences (*genotypes*) representing mixtures of two haplotypes – each site is defined by an unordered pair of allele readings, one from each haplotype – while haplotypes are computationally inferred from genotypes [6], [4]. To genotype data we refer as unphased data and to haplotype data we refer as phased data. The disease association study analyze data given as genotypes or haplotypes with disease status.

Disease association analysis searches for a SNP with frequency among exposed individuals considerably higher than among unexposed individuals. Only statistically significant SNPs (whose frequency distribution has p-value less than 0.05) are reported. Successful as well as unsuccessful searches for SNPs with statistically significant association have been recently reported for different diseases and different suspected human genome regions (see e.g. [5]). Unfortunately, reported findings are frequently not reproducible on different populations. It is believed that this happens because the p-values are unadjusted to multiple testing – indeed, if the reported SNP is found among 100 SNPs then the probability that the SNP is associated with a disease by mere chance becomes roughly 100 times larger.

Since complex common diseases can be caused by multi-loci interactions two-loci analysis can be more powerful than traditional one-by-one SNP association analysis [9]. Multi-loci analysis is expected to find even deeper disease-associated interactions. In this paper we suggest to search for disease-associated MSCs in the genotype/haplotype data.

An exhaustive search among MSCs usually is infeasible even for small number of SNPs let alone the genome-wide studies.

In order to handle data with large number of SNPs we extract informative (indexing) SNPs that can be used for reconstructing of all other SNPs using multiple linear regression based method [8]. However, exhaustive searching for all possible SNP combinations is still very slow. The main contribution of this paper is a novel combinatorial method for finding disease-associated MSCs applied to index SNPs.

Here we first propose to extract informative (indexing) SNPs that can be used for reconstructing of all SNPs almost without loss or lossless for large enough number of SNPs [8]. In the reduced set of SNPs, we propose to apply a novel combinatorial method for finding disease-associated multi-SNP combinations. Our experimental study shows that the proposed methods are able to find MSCs whose disease association is statistically significant even after multiple testing

D.B. and J.H. were partially supported by Georgia State University Molecular Basis of Disease Fellowship. A.Z. was partially supported by NIH Award 1 P20 GM065762-01A1 and US CRDF Award MOM2-3049-CS-03.

adjustment. For Daly et al [3] data we found a few unphased MSCs associated with Crohn's disease with multiple testing adjusted p-value below 0.05 while no single SNP or pair of SNPs show any significant association. For Ueda et al [7] data we found a few new unphased and phased multi-SNP combinations associated with autoimmune disorder.

The rest of the paper is organized as follows. Section II introduces necessary notations and formally describes the searching problem for statistically significant disease-associated MSCs. Section III describes the exhaustive and combinatorial search algorithms for this problem and Section IV gives the results for the real data [3], [7].

II. Notations and Problem Formulation

In this section we introduce all necessary notations and formally describe the search of statistically significant disease-associated MSCs.

Our data is a set of n individuals described by values of m SNPs and its disease status. When individuals are represented by genotypes then SNP values belong to $\{0, 1, 2\}$, where 0's and 1's denote homozygous sites with major allele and minor allele, respectively, and 2's stand for heterozygous sites. When individuals are represented by haplotypes then SNP values belong to $\{0, 1\}$, where 0's and 1's denote major allele and minor allele, respectively.

A *multi-SNP combination (MSC)* C is given by a subset of SNPs $snp(C)$ ($snp(C)$ is a subset of the set of all m SNPs) and their values. The subset of individuals whose values match values of C on the SNPs from $snp(C)$ is denoted $set(C)$ and its size is denoted $|C|$. The set $set(C)$ is partitioned into two subsets: $exposed(C)$ consisting of exposed individuals and $unexposed(C)$ consisting of unexposed individuals.

We are interested in the probability that $set(C)$ is partitioned into $exposed(C)$ and $unexposed(C)$ by mere chance. According to binomial distribution, the p-value equals:

$$p(C) = \sum_{k=0}^{|exposed(C)|} \binom{n}{k} q^k (1-q)^{|C|-k} \qquad (1)$$

where $q = \frac{|exposed(all\ individuals)|}{n}$ is the probability of being exposed. The MSC C is statistically significant if the $p(C) < 0.05$.

The p-value computed by formula 1 is correct only if a single MSC is tested. Since statistically significant MSCs are searched among many such combinations, the computed p-value requires adjustment for multiple testing. The standard Bonferroni correction adjusts p-value by multiplying it by the number of the number of tests, i.e., number of tested MSCs. However, the Bonferroni correction is overly pessimistic, e.g., for finding one significant SNP among 100 we should multiply its p-value by 100; as a result, SNP should have $p < 0.0005$ in order to be statistically significant. Similarly this factor grows to 10^4 for 2-SNP combinations. Instead, we compute multiple testing adjustment using more accurate but computationally extensive randomization method. 10^4 times we repeat the following: (1) randomize the status of individuals (by random swapping) and (2) find the 500th smallest p-value of MSCs . This p-value corresponds to the multiple testing adjusted $p = 0.05$.

Formally the searching problem is as follows

Disease-Associated MSC Search. Given a population of n genotypes (or haplotypes) each containing values of m SNPs and disease status (exposed or unexposed), find all MSCs with multiple-testing adjusted p-value of the frequency distribution below 0.05.

III. Searching Methods for Disease-Associated MSCs

In this section we first discuss the exhaustive search for the MSCs. Next, we briefly describe multiple linear regression method [8] for extracting informative index SNPs. Then we propose the new combinatorial search algorithm and its faster implementation.

Exhaustive Search. The search for disease-associated MSCs among all possible combinations can be done by the following *exhaustive search*. In order to find a MSC with the p-value of the frequency distribution below 0.05, we should check all one-SNP, two-SNP, ..., m-SNP combinations. The checking procedure takes $O(n \sum_{k=1}^{m} \binom{m}{k} 3^k)$ runtime for unphased combinations since there are 3 possible SNP values $\{0, 1, 2\}$. Similarly, for phased combination, the runtime is $O(n \sum_{k=1}^{m} \binom{m}{k} 2^k)$ since there only 2 possible SNP values. The exhaustive search is infeasible even for small number of SNPs and we limit ourselves with the small number of SNPs, i.e., instead of searching all MSCs, we search only containing at most $k = 1, 2, 3$ SNPs. We refer to k as *search level* of exhaustive search.

Indexing with MLR Tagging Method. In order to reduce the runtime of exhaustive search, we propose to decrease the size the input data set by extracting informative SNPs (further referred as *indexing SNPs*) from which one can reconstruct all other SNPs. In our experiments we use multiple linear regression tagging method of [8]. However, there is a tradeoff between the number of chosen SNPs and quality of reconstruction. Our strategy is to chose maximum number of index SNPs that still result in the reasonable runtime of the exhaustive search.

Combinatorial Search. Here we propose the following search method for disease-associated MSCs. It can find a combinations with large number of SNPs even for small search levels (here number of SNPs is not equivalent to searching level, see below).

Given a MSC C, sometimes one can decrease the frequency of C among unexposed population by adding SNPs to the $snp(C)$ which have the same value in all individuals from $exposed(C)$. Such addition will not affect $exposed(C)$ but may reduce $unexposed(C)$.

Formally, a MSC C' is an *exposed-closure* of MSC C, if $exposed(C') = exposed(C)$ and $|unexposed(C')|$ is minimized. Two MSCs C and C' are equivalent if $exposed(C) = exposed(C')$ and $unexposed(C) = unexposed(C')$. Equivalent MSCs can not be distinguished, and we will represent

the equivalence class by the MSC with the maximum number of specified SNPs. This representative C can be efficiently found by incorporating into $snp(C)$ *all* SNPs that have the same value across all individuals in $exposed(C)$.

The proposed combinatorial search finds the best p-value of the exposed-closure of each single-SNP, after that it searches for the best p-value among exposed-closure of all 2-SNP combinations and so on. The procedure stops after all exposed-closure of all k-SNP combinations $(k < m)$ are checked. The corresponding *search level* is the number of SNPs selected for exposed-closuring, e.g., on the level 2 of searching combinatorial search will test exposed-closure of all 2-SNP combinations for association with a disease. Because of the exposed-closure, for the same level of searching combinatorial search finds better association than exhaustive search. However, proposed combinatorial search is as slow as exhaustive search.

Speed-up of Combinatorial Search. A faster implementation of this method avoids checking MSCs which are not (and cannot lead to) statistically significant ones. Formally, a MSC C is called an $intersection$ of MSC C_1 and C_2 if $exposed(C) = exposed(C_1) \cap exposed(C_2)$ and $|unexposed(C)|$ is minimized. A MSC C is called *trivial* if its unadjusted p-value is larger than 0.05 even if the set $exposed(C)$ would be empty. Note that intersection of a trivial MSC with another is trivial.

A faster implementation of the combinatorial search is as follows:

1. Compute a set G_1 of all 1-SNP exposed-closed MSCs, exclude trivial combinations.
2. Compute sets G_k of all pairwise intersections of the MSCs from G_{k-1}, exclude trivial combinations and already existing in $G_1 \bigcup G_2 \bigcup .. \bigcup G_{k-1}$, $k = 2..N$.
3. For each G_k output MSCs whose unadjusted $p < 0.05$.

Still, in order to find all MSCs associated with a disease we have to check all possible SNP combinations with all possible SNP values. Our searching approach is also computationally intensive and step 2 from the algorithm can generate an exponential number of MSCs. However, exposed-closure avoids generation and checking of non-significant MSCs. Additionally, removing of trivial MSCs at each iteration of step 2 considerably reduces the number of newly generated MSCs. For example, for search level 2 our method is faster than level-2 exhaustive search, and returns all possible disease-associated 2-SNP combinations as well as the set of MSCs obtained by exposed-closure of 1- or 2-SNP combinations. In conclusion, proposed method is more efficient than the exhaustive search and can find MSCs associated with a disease on small search levels.

IV. Results and Discussion

In this section we discuss the results of four methods for searching disease-associated MSCs on real phased and unphased datasets.

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* [7] 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.

The both datasets have been phased using 2SNP software [4]. The missing data (16% in [3] and 10% in [7]) have been imputed in genotypes from the resulted haplotypes. For each genotype dataset, we have also created corresponding haplotype dataset in which each individual is represented by a haplotype with the disease status inherited from the corresponding individual genotype.

Search Methods. We have compared the following 4 methods for search disease-associated MSCs.

- Exhaustive Search (**ES**);
- Indexed Exhaustive Search (**IES(30)**): exhaustive search on the indexed datasets obtained by extracting 30 indexed SNPs with MLR based tagging method [8];
- Combinatorial Search (**CS**);
- Indexed Combinatorial Search (**ICS(30)**): combinatorial search on the indexed datasets obtained by extracting 30 indexed SNPs with MLR based tagging method [8].

Each of these methods have been applied to the search levels 1,..,6; however, significant MSCs have been found only on levels 1 and 2 because adjusted p-value grows with level. The size of the datasets is enough large to make exhaustive search impossible even for a combination of 6 SNPs. All experiments were ran on Processor Pentium 4 3.2Ghz, RAM 2Gb, OS Linux – the runtime is given in the last column of Table I.

Performance Quality. The quality of searching methods is compared by the number of found statistically significant MSCs (see the 7th column of Table I). Since statistical significance should be adjusted to multiple testing, we report for each method and data set the 0.05 threshold adjusted for multiple testing (this threshold is computed by randomization and given in the third column of Table I). In the 4th, 5th and 6th columns, we give the frequencies of the best MSC among exposed and unexposed population and the unadjusted p-value, respectively.

Discussion. Comparing indexed counterparts with exhaustive and combinatorial searches shows that indexing is quite successful. Indeed, indexed search finds the same MSCs as non-indexed search but it is much faster and its multiple-testing adjusted 0.05-threshold is higher and easier to meet.

Comparing combinatorial searches with the exhaustive counterparts is advantageous to the former. Indeed, for unphased data [3] the exhaustive search on the first and second search levels is unsuccessful while the combinatorial search finds several statistically significant MSCs for the same searching level. Similarly, for unphased and phased data of

TABLE I

COMPARISON OF FOUR METHODS FOR SEARCHING DISEASE-ASSOCIATED MULTI-SNPS COMBINATIONS.

Search level	Search method	MT-unadjusted p corresponding to adjusted p=0.05	SNP combination with minimum p-value			Number of SNP combinations with MT-adjusted p<0.05	runtime sec.
			exposed frequency	unexposed frequency	unadjusted p-value		
Unphased Daly et al [3]							
1	ES	1.6×10^{-3}	0.31	0.16	1.8×10^{-3}	0	0.9
	IES(30)	3.9×10^{-3}	0.30	0.16	4.7×10^{-3}	0	0.5
	CS	5.1×10^{-5}	0.30	0.11	2.0×10^{-5}	2	1.0
	ICS(30)	2.2×10^{-3}	0.30	0.14	4.6×10^{-4}	1	0.6
2	ES	1.9×10^{-5}	0.30	0.13	3.1×10^{-4}	0	15.0
	IES(30)	1.0×10^{-4}	0.31	0.14	4.4×10^{-4}	0	1.0
	CS	1.5×10^{-6}	0.17	0.02	6.5×10^{-7}	2	7.0
	ICS(30)	5.0×10^{-5}	0.17	0.04	3.7×10^{-5}	1	0.4
Unphased Ueda et al [7]							
1	ES	1.3×10^{-3}	0.43	0.28	1.1×10^{-4}	2	1.0
	IES(30)	3.1×10^{-3}	0.43	0.28	1.1×10^{-4}	4	0.6
	CS	1.8×10^{-4}	0.43	0.28	9.2×10^{-5}	2	1.1
	ICS(30)	1.6×10^{-3}	0.43	0.28	1.1×10^{-4}	4	0.6
2	ES	2.7×10^{-6}	0.25	0.12	1.5×10^{-6}	2	30.0
	IES(30)	8.0×10^{-5}	0.25	0.12	1.5×10^{-6}	9	3.0
	CS	1.1×10^{-6}	0.16	0.06	8.5×10^{-7}	3	20.0
	ICS(30)	4.7×10^{-5}	0.25	0.12	1.1×10^{-6}	10	1.0
Phased Daly et al [3]							
1	ES	2.4×10^{-3}	0.52	0.40	9.7×10^{-3}	0	1.0
	IES(30)	7.2×10^{-3}	0.52	0.41	1.6×10^{-2}	0	0.6
	CS	1.3×10^{-4}	0.52	0.36	4.3×10^{-4}	0	1.1
	ICS(30)	1.6×10^{-2}	0.52	0.40	1.0×10^{-2}	1	0.7
2	ES	3.0×10^{-5}	0.05	0.01	1.4×10^{-3}	0	23.0
	IES(30)	1.7×10^{-4}	0.55	0.42	5.5×10^{-3}	0	3.0
	CS	7.0×10^{-7}	0.48	0.30	5.9×10^{-5}	0	17.0
	ICS(30)	5.8×10^{-5}	0.48	0.35	3.1×10^{-3}	0	1.0
Phased Ueda et al [7]							
1	ES	9.2×10^{-4}	0.65	0.53	3.2×10^{-4}	2	6.0
	IES(30)	5.3×10^{-3}	0.66	0.55	1.4×10^{-3}	2	2.0
	CS	8.3×10^{-4}	0.37	0.28	2.9×10^{-4}	5	6.2
	ICS(30)	7.4×10^{-2}	0.66	0.55	1.4×10^{-3}	10	2.1
2	ES	2.1×10^{-6}	0.17	0.09	6.8×10^{-7}	2	173.0
	IES(30)	1.7×10^{-4}	0.19	0.12	3.7×10^{-5}	2	16.0
	CS	5.0×10^{-7}	0.02	0.00	1.6×10^{-8}	8	75.0
	ICS(30)	9.5×10^{-5}	0.19	0.12	3.0×10^{-5}	2	5.7

[7] the combinatorial search found much more statistically significant MSCs than the exhaustive search for the same searching level.

We conclude that the proposed indexing approach and the combinatorial search method are very promising techniques for searching statistically significant diseases-associated MSCs which can lead to discovery disease causes. The next step in our research is biological validation of statistically significant MSCs discovered by proposed searching methods.

REFERENCES

[1] Affymetrix (2005) *http://www.affymetrix.com/products/arrays/*.

[2] International HapMap Consortium. (2003) The International HapMap Project. *Nature*, **426**, 789–796, *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.

[4] Brinza, D. and Zelikovsky, A. (2006) 2SNP: Scalable Phasing Based on 2-SNP Haplotypes, *Bioinformatics*, **22(3)**, 371–373.

[5] Clark AG. (2003) Finding Genes Underlying Risk of Complex Disease by Linkage Disequilibrium Mapping, *Curr Opin Genet Dev.*, **13(3)**, 296–302.

[6] 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**, 978-998.

[7] 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.

[8] He, J. and Zelikovsky, A. (2006) Tag SNP Selection Based on Multivariate Linear Regression, *Proc. of Intl Conf on Computational Science (ICCS 2006)*, to appear.

[9] Marchini, J., Donnelley , P. and Cardon, L.R, (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases, *Nature Genetics*, **37**, 413–417.

[10] Herbert, A., Gerry, N.P., McQueen, M.B. (2006) A Common Genetic Variant Is Associated with Adult and Childhood Obesity, *SCIENCE*, **312**, 279–284.

[11] Spinola, M., Meyer, P., Kammerer, S. et al. (2006) Association of the PDCD5 Locus With Lung Cancer Risk and Prognosis in Smokers, *American Journal of Clinical Oncology*, **24:11**.