Selection of representative SNP sets for genome-wide association studies: A metaheuristic approach

Article *in* Optimization Letters · August 2011

DOI: 10.1007/s11590-011-0419-7

Selection of Representative SNP Sets for Genome-Wide Association Studies: A Metaheuristic Approach

Gürkan Üstünkar - Süreyya Özöğür-Akyüz - Gerhard W. Weber - Yeşim AydınSon - Christoph M. Friedrich

Abstract After the completion of Human Genome Project in 2003, it is now possible to convey the research studies to associate genetic variations in the human genome with common and complex diseases. The Single Nucleotide Polymorphism (SNP) biomarkers across the complete sets of DNA, or genomes, of 11 different populations are scanned for revealing genetic risk factors and quantitative traits associated with human diseases. The current challenge is to utilize the genome data efficiently and to develop tools that improve our understanding of etiology of complex diseases. Many of the algorithms needed to solve these problems are strongly supported by management science and operations research (OR) methods. One application is to select a subset of SNPs from the whole SNP set that is informative and small enough to convey subsequent association studies. In this paper, we present an OR application for representative SNP selection that makes use of our novel Simulated Annealing (SA) based feature selection algorithm. We hope that our work will facilitate reliable identification of SNPs that are involved in the etiology of complex diseases and ultimately supporting timely identification of genomic disease biomarkers, and development of personalized medicine approaches and targeted drug discoveries.

Keywords Simulated Annealing - Feature Selection - Representative SNP Selection - SNP-Complex Disease Association - Bioinformatics - OR in Computational Biology

Gürkan Üstünkar

Informatics Institute, Middle East Technical University, Ankara, Turkey

e-mail: e145307@metu.edu.tr

Süreyya Özöğür-Akyüz

Department of Mathematics and Computer Science, Bahçeşehir University, Istanbul, Turkey

e-mail: sureyya.akyuz@bahcesehir.edu.tr

Gerhard W. Weber

Institute of Applied Mathematics, Middle East Technical University, Ankara, Turkey

e-mail: gweber@metu.edu.tr

Yeşim Aydın Son

Informatics Institute, Middle East Technical University, Ankara, Turkey

e-mail: yesim@metu.edu.tr

Christoph M. Friedrich

Department of Computer Science, University of Applied Sciences and Arts, Dortmund, Germany

e-mail: christoph.friedrich@ fh-dortmund.de

1 Introduction

The human genome can be viewed as a sequence of 3.3 billion letters from the nucleobase set {A, C, G, T}. The nucleotide sequence in a population is same for more than 99% of the positions on the genome. However, humans possess unique genetic composition in about 1% of their genome [19]. These genetic variations include different nucleotide occurrences, called Single Nucleotide Polymorphisms (SNPs – 'snips') if occurred in more than 1% of a given population, deletion/insertion of one or more nucleotides, or variations in the number of multiple nucleotide repetitions (see Figure 1).

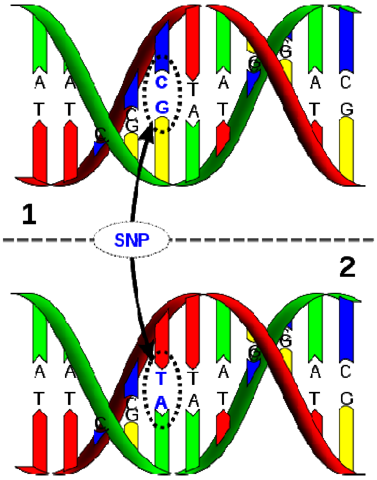

Fig.1 A C/T Polymorphism between two DNA Sequences[1]

Recently, there has been increasing research to find genetic markers, haplotypes[2], and potentially other variables that together contribute to a disease and serve as good predictors of the observable disease phenotypes. Complex diseases are typically associated with multiple genetic loci and several external (e.g., environmental) factors. Therefore, it is essential to investigate all polymorphisms located in the functional regions of candidate genes [6, 33], and integrate the information about the network of genes involved in biological systems of major physiological importance, such as lipid metabolism, cellular adhesion, inflammation, and others [32] for thorough analysis of these kind of diseases.

Association studies are among the promising ways of dealing with the problem of finding disease causing variants and such association studies typically make use of SNPs as they are the most common form of genetic variations and they can represent an individual's genetic variability in greatest detail [30]. However, the enormous number of SNPs (estimated more than 11 million) makes it infeasible to gather information and perform analysis on all the SNPs in the human genome. Thus, while performing a disease association study, the geneticist would want to experimentally test for association by only considering a subset of

[1] Figure excerpted from http://en.wikipedia.org/wiki/Single-nucleotide_polymorphism

[2] A *haplotype* is a combination of alleles (DNA sequences) at different places (loci) on the chromosome that are transmitted together. A haplotype may be one locus, several loci, or an entire chromosome.

the entire SNP set and not all SNPs, thereby considerably saving resources (alternatively, increasing the power of the statistical tests by increasing the number of individuals) as well as making the problem computationally feasible. Therefore, selecting a subset of SNPs that is informative enough to perform association studies but still small enough to reduce the analysis workload, to which we refer as ***representative SNP selection***, has become an important step for disease-gene association studies.

Reducing biological and statistical redundancy from hundreds of thousands of SNPs is the key for representative SNP selection. Dealing with many dependent association tests is one of the emerging issues on the statistical and computational side. SNP vs. disease data, in addition to being large, redundant, diverse and distributed, has three important characteristics posing challenges for the data analysis and modeling: (1) heterogeneity, (2) a constantly evolving biological nature and (3) complexity. Therefore intelligent methods are needed to find SNPs associated with the disease and extract biologically relevant subsets. The problem of SNP selection has been proven to be NP-hard in general [2] and current selection methods possess certain restrictions and require use of heuristics for reducing the complexity of the problem.

OR methods have been used recently to the problem of representative SNP selection [10, 34, 35, 36, 37, 38, 39]. In this paper, we present a method for selecting representative SNP subset for stronger association with complex disease after following an integrative biological scoring and filtering approach. An OR class novel feature selection method based on Simulated Annealing (SA) has been developed for representative SNP selection, in which we try to maximize tagged SNP prediction while minimizing cardinality of the selected SNP subset. The methods introduced in this paper intend to contribute to a better understanding of the etiology of complex diseases and support reliable and timely identification of disease causing variants. It is also our aim to encourage future efforts to make modern optimization methods, supported by OR and its scientific community, become useful in the research subject of this paper and in computational biology, bioinformatics, medicine and heathcare in general.

The reminder of this paper is organized as follows. In Section 2, we give a formal definition of the problem. In Section 3, we review some of the existing methods for representative SNP selection. In Section 4, we present our methodology and algorithm. In Section 5, we describe our data sets and give comparative analysis of our work and existing feature selection schemes. Finally we conclude the paper in Section 6 and discuss future research directions.

2 Problem Definition

The aim of the representative SNP selection approach is to find a minimal subset of SNPs, whose allele[3] information can explain the whole set of SNPs in the candidate region under study (a whole chromosome, or a target region) to the greatest detail. A formal definition of the problem can be stated as follows: Let $S = \{ SNP_1 , ..., SNP_n \}$ be a set of n SNPs in a candidate region and $G = \{g_1, ..., g_m\}$ be a data set of m genotypes, where each genotype g_i consists of the consecutive allele information of the n SNPs: SNP_1 , ... , SNP_n. For simplicity we represent $g_i \in G$ be a vector of size n whose vector element is 0 when the allele of

[3] An *allele* is defined as one of two or more forms of the DNA sequence of a particular gene.

a SNP is homozygous dominant[4], 1 when it is heterozygote[5] and 2 when it is homozygous recessive[6]. Alternatively, data can be gathered from a case-control study. In this case, researcher would also have a hand on the phenotype data for the particular patient and this information can be matched with genotype information. Phenotype variable will be represented by P and it takes the value -1 if it belongs to case group and 1 if it belongs to control group. Matrix A in Figure 2 represents such data.

Suppose that the maximum number of SNPs that can be selected is k (which can be alternatively be a variable for the problem), and a function $f(R|G,P)$ evaluates how well the allele information of SNPs in subset $R \subset S$ retains the allele information of all SNPs in S based on the genotype data G and classification performance of selected set R on disease phenotype are.

	SNP$_1$	SNP$_2$	SNP$_3$	....	SNP$_n$	P
g_1	0	0	0	....	1	1
g_2	2	1	1	....	1	1
g_3	2	0	0	....	2	-1
.	.	.	.	....	.	.
.	.	.	.	....	.	.
.	.	.	.	....	.	.
.	.	.	.	....	.	.
.	.	.	.	....	.	.
g_m	1	0	2	....	1	1

Fig. 2 SNP-Genotype Matrix *A*

Given S, G, P and k the Representative SNP Selection problem is the following optimization problem:

$$\max F(R|G,P)$$

$$\begin{aligned} \text{subject to: } & R \subset S, \\ & |R| \leq k, \\ & k > 0, \\ & k \text{ integer.} \end{aligned}$$

To solve *Representative SNP Selection problem*, one needs to find an optimal subset of SNPs, R, of size less than or equal to k, based on the given evaluation function $F(R|G,P)$. From a set theoretic point of view, it is computationally intractable to examine all possible subsets of the given set of SNPs to select a set of representative markers, except for very small data sets. The problem is proven to be NP-hard [2]. To cope with this difficulty, it is possible to

[4] An individual that is *homozygous dominant* for a particular trait carries two copies of the allele that codes for the dominant trait.

[5] A person possessing two different forms of a particular gene, one inherited from each parent.

[6] An individual that is *homozygous recessive* for a particular trait carries two copies of the allele that codes for the recessive trait.

divide Representative SNP selection into three largely independent steps: (1) identifying genomic segments where the selection will be performed; (2) defining a measure to quantify how well a set of SNPs can predict all observed and/or unobserved SNPs; and (3) searching a minimum set of Representative SNPs that meets a desired threshold.

3 Related Work

Application of statistical hypothesis-testing procedures is the basic approach for finding genotype-phenotype associations. The null hypothesis to be tested is that there is no difference between two study groups with respect to the genotype frequencies (i.e. genotype proportions) observed in each group. The chi-square and Fisher's exact tests may be applied in this task [23]. Odds-ratios are also commonly used to indicate differences between groups on the basis of their genotype frequencies. Methods for multiple testing (such as Bonferroni or False Discovery Rate) in high-dimensional settings can be applied when many SNPs are considered simultaneously.

In addition to statistical hypothesis testing in which causative SNPs are identified, one may chose use classification models for genotype-phenotype association modeling. This can be done by representing different genotypes for a particular SNP as inputs and phenotype as label. Different statistical and machine learning techniques, such as logistic regression and support vector machines, can be applied for this purpose. Not only the genotype information extracted from multiple SNPs but also information related to environmental exposure factors and other biomarkers can be incorporated by introducing multivariable statistical and machine learning models in this context. Tagging and different feature selection procedures are useful to improve the prediction performance of multiple-SNPs models. The former can be applied to problems with a large number of SNPs in which haplotype data is present. Feature selection is recommended to reduce the number of highly-correlated SNPs, in which high Linkage Disequibrium[7] (LD) makes it difficult to select true disease causing variant. These methods are presented in the subsequent sections.

3.1 Statistical Methods

In order to select a subset of SNPs in genome-wide complex disease association studies, various statistical measures and testing based approaches have been introduced specific to problem domain. The paper [25] proposes a sliding window approach, which made use of combination of p-values from multiple independent tests by making use of

$$X^2 = -2 \sum_{i=1}^{\mathrm{m}} log(p_i) \sim X_{2m}^2 .$$

Here, p_i denotes p-value of association between SNP_i and disease presence and m is the number of SNPs in the sliding window. It is shown that test statistic X^2 follows a Chi-square distribution with *2m* degrees of freedom. The basic advantage of this approach is that it takes into account the ordering of SNPs on the chromosome and allows detection of chromosome regions with significant associations by merging adjacent windows [9, 29]. However an implicit assumption is made that the distance between any two adjacent SNPs is constant.

[7] *Linkage disequilibrium* is the non-random association of alleles at two or more loci.

Other scan statistics have been developed that also considers the ordering and spacing of SNPs on the chromosome [8, 12, 21, 31]. For example in [31] a two step procedure was presented for calculating chromosomal scan statistic:

(1) identify SNP clusters,

(2) extract clusters with significant disease association.

It is assumed here that position of each SNP follows a Poisson distribution. Therefore length between two adjacent SNPs is assumed to have exponential distribution and distance between two particular SNP is assumed to follow a Gamma distribution. Using these assumptions one can identify the clusters of SNPs by testing the hypothesis that whether the observed length between a set of SNPs is equal or less than the expected length. If the hypothesis is rejected then this group of SNPs is identified as a cluster. Then to test the significance of disease association for a particular cluster Pearson Chi-square p-values are calculated. However this type of scan approaches has the disadvantage that they do no incorporate gene-gene interactions.

3.2 Tagging and Machine Learning Methods

One obvious observation from the formal definition of representative SNP selection problem is the selected subset's dependence on the function *F*. In the literature, various objective functions have been defined to represent the allele information of genotypes in *G* using SNPs in *S* and solve the problem accordingly. One can classify the proposed approaches into three categories according to how they try to measure the allele information of genotypes: (1) Haplotype Diversity based approaches, (b) Pairwise Association based approaches, (3) Predicting Tagged SNPs.

Haplotype Diversity based approaches are inspired by the fact that DNA can be partitioned into discrete blocks such that within each block high LD is observed and between blocks low LD is observed [7, 26]. As a result of this feature, number of distinct haplotypes consisting of the SNPs is very small across a population. Hence, one would try to find the subset of SNPs, which are responsible for the "limited haplotype diversity" in order to find the representative SNP set. Different studies have been conveyed to see how well diverse haplotypes can be distinguished depending on a selected "diversity measure" and chose the best one. A detailed explanation on the different type of measures used in the literature can be seen in [11, 14, 15, 17]. The usual approach in these methods is to exhaustively list and search through every subsets of H. Therefore, only a small number of SNPs can be analyzed. To cope with this problem, efficient heuristics have been proposed using Dynamic Programming [34, 35, 36, 37, 38], Principal Component Analysis [13, 22, 24] and Greedy Algorithm [39]. Although haplotype diversity based methods are simple to implement they depend on the block partitioning method used for a target locus. In addition, the union of the candidate SNP sets for each block may not be an optimal set for the overall locus.

Pairwise Association based approaches are based on the principle that all the SNPs in the target locus are highly associated with at least one of the SNPs in the selected SNP subset. This way, although a SNP that can be used to predict a disease causing variant may not be selected as a representative SNP, the association may be indirectly assumed from the selected SNP that is highly associated with it. The association between SNPs can be estimated using LD. The common solution approach for these methods is to cluster the SNPs into different subsets and choose a representative SNP (or SNPs) from each cluster [1, 4, 5]. Although with their $O(cgs^2)$ complexity (c being number of clusters, g being number of genotypes and s being number of SNPs) pairwise association methods are so much faster than haplotype diversity based methods, they have a major shortcoming as they cannot explain multi-SNP dependencies [2] and they tend to select more tag SNPs [16].

Predicting Tagged SNPs is motivated by the idea of reconstructing the genotype data from an initial set of selected SNPs in order to minimize the error of prediction for unselected SNPs. Those prediction methods have a certain advantage over Pairwise Association methods as they would take multi-SNP dependencies into consideration. The paper [2] proposes a measure called "Informativeness" and used dynamic programming to solve the problem of finding the optimal subset of SNPs that can best predict the remaining (tagged) SNPs. Let $E^{s}_{i,j}$ be the event that genotypes g_i and g_j have a different allele at SNP s, and $E^{S}_{i,j}$ be the event that genotypes g_i and g_j have a different allele at some SNP in S. To measure how well a set of SNPs, $S = \{SNP_1, ..., SNP_k\}$, can predict the SNP, s, the used measure is as follows:

$$I(S,s) = P_{i \neq j}(E^{S}_{i,j} \mid E^{s}_{i,j}).$$

A more recent approach for use of dynamic programming is proposed by [10] by fixing the number of representative SNPs for each tagged SNP to 2. The paper [20] improves over the current predicting based method by allowing multi-allelic prediction (instead of bi-allelic) and not restricting the number of representative SNPs. They proposed a heuristic algorithm that uses the probabilistic framework of Bayesian networks to effectively identify a set of predictive SNPs.

Our proposed algorithm falls into last group of methods and explained in detail in Section 4.

4 Proposed Methodology

Figure 3 depicts the overall methodology used to achieve representative SNP set. We first apply initial filtering based on quality control measures Minor Allele Frequency, missingness and Hardy-Weinberg equilibrium. After that we calculate multiple testing adjusted p-values of association. We then filter biologically relevant SNPs. Lastly, we apply our feature selection algorithm on different chromosomes and merge the selected SNP subsets to find representative SNP set.

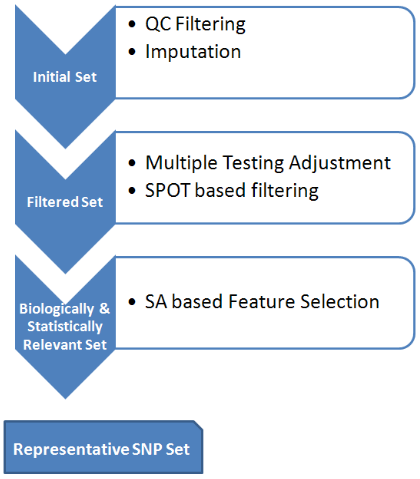
Initial Set
QC Filtering
Imputation
Filtered Set
Multiple Testing Adjustment
SPOT based filtering
Biologically & Statistically Relevant Set
SA based Feature Selection
Representative SNP Set

Fig. 3 Process Steps for Finding Representative SNP Set

4.1 Filtering Biologically Relevant SNPs

In order to gain insight from a Genome Wide Association Study (GWAS) in which hundreds of thousands SNPs are genotyped from subjects that are diagnosed with a disease (case) and healthy people (control) one needs to decrease the dimension to a manageable value. Unfortunately even with classification schemes that are optimized for large scale data it takes too much time to perform analysis. Therefore there is a need to extract biologically relevant data as an initial pass for analysis. This would be achieved by incorporating information from biological databases so that biologically relevant SNPs, such as those in genes related to the phenotype or with potentially non-neutral effects on gene expression such as a splice sites, are given higher priority.

SPOT[8] recently introduced the genomic information network (GIN) method [28] for systematically implementing this kind of strategy. GIN is a directed graph whose nodes are features from a biological database. The GIN represents a process: it begins with a SNP and ends in the terminal node with the determination of its overall prioritization score S. The overall score is a cumulative measure of biological relevance obtained by combining information across multiple domains. For example, if a SNP is in a gene then it will link to the gene node, which will increase the overall score. Once the overall score S determined, they rank SNPs from a GWAS for further study by $\frac{p}{10^S}$. We used SPOT for extracting biologically relevant SNPs from the whole SNP set after performing Case Control Association study to find p-values.

4.2 Simulated Annealing Based Feature Selection Scheme

In Section 2, we state that the optimal SNP subset depends on the selected evaluation function. It is also pointed out that maximizing the prediction accuracy of selected SNPs over unselected SNPs is an approach used in the literature for representative SNP selection. We set our goal as to find a minimum size set of representative SNPs and a prediction algorithm, such that the prediction error is minimized. Then our objective function becomes:

$$\sum_{i=1}^{n-k} NaiveBayes(G_R, G_{T_i}) + NaiveBayes(G_R, P),$$

where G_R denotes genotype data related with representative SNP set R, G_{T_i} denotes genotype data related with $SNP_i \in S\backslash R$ and

$$NaiveBayes(F,L) = argmax_L\ P(L = l) \prod_{i=1}^{n} p(F_i = f_i | L = l)$$

denotes a Naive Bayes classifier where F is the feature set (SNP set in our context) and L is the label. We calculate 5-fold Cross Validation (CV) based classification and find classification accuracy.

We used Simulated Annealing (SA) [18], which is a local search algorithm, to solve our problem. SA strives for the best solution starting of a randomly created solution. Each step of the SA algorithm replaces the current solution by a "nearby" solution. The new solutions are chosen depending on an evaluation function and a global parameter T (temperature). T value is gradually decreased during the process. Fundamental to the SA structure is the binary coding scheme. Let C_i represents i^{th} coding where each code containing n SNPs

[8] https://spot.cgsmd.isi.edu/submit.php

(dimension). Each code of the length n is a sequence over $\{0, 1\}^n$ (0 represents a non-selected SNP and 1 represents a selected SNP). For example, assume there is a code represented by C_i = {1, 0, 1, 0, 0, 1, 0}. In this encoding scheme SNP_1, SNP_3 and SNP_6 are selected SNPs. A neighbor for a coding scheme C_i is another coding scheme, which is one bit different than $C_{i.}$

We create a random binary coding of size n as an initial solution and test the accuracy of the solution using Naive Bayes by calculating the mean classification error for (n-k) supervised learning iterations, where k is the number of selected SNPs in a particular iteration. We run the algorithm for a certain amount of steps (user defined). We use a tradeoff between accuracy and the number of SNPs in the representative SNP set. Therefore we also try to minimize the number of chosen SNPs (k). The pseudocode of the algorithm is given in Algorithm 1.

Alg. 1 Simulated Annealing Representative SNP Selection Algorithm

```
**Input:**
$s_0$  initial randomly selected SNP set
$t$    simulated annealing parameter temperature.
$d$    simulated annealing parameter decreasing factor.
$c_{max}$  number of iterations.

**Output:**
$s_{best}$  representative SNP set

$s \leftarrow s_0; e \leftarrow E(s)$
$s_{best} \leftarrow s; e_{\mathrm{best}} \leftarrow e$
**For** $c = 1$ to $c_{\max}$
    $s_{new} \leftarrow \mathrm{neighbor}(s)$
    $e_{new} \leftarrow E(s_{new})$
**if** $e_{new} > e_{best}$ **then**
$s_{best} \leftarrow s_{new}; e_{best} \leftarrow e_{new}$
**if** $P(e_{new}, t) <$ **random()** **then**
$s \leftarrow s_{new}; e \leftarrow e_{new}; t = t * d$
**Next** $c$
**return** $s_{best}$
```

Here, s_0 is the initial randomly selected SNP set (R) and $E(s)$ is an evaluation function denoted by:

$$E(s) = w(\sum_{i=1}^{n-k} NaiveBayes(G_R, G_{T_i}) + NaiveBayes(G_R, P)) + (1 - w)k$$

presenting our objective function. Cardinality of the representative SNP set is denoted by k. We use two user specified arguments for the algorithm: c_{max} denotes the number of iterations and w $(0 \leq w \leq 1)$ denotes weight that specifies tradeoff. The smaller the w, the less SNPs will be chosen to represent overall SNP set S. Lastly, t, d and $P(E(S),t)$ (energy) denotes the simulated annealing parameters. Temperature is defined by t and decreasing factor is denoted by d. $P(E(S),t)$ is calculated by:

$$P(E(s),t) = \exp\left(-\frac{E(s)}{t}\right).$$

5 Experimental Study

5.1 Data Sets

We used two data sets for evaluation purposes. The first data set is whole genome association Rheumatoid Arthritis (RA) data from the North American Rheumatoid Arthritis Consortium (NARAC) including 868 cases and 1194 controls. The data was used in Genetic Analysis Workshop 16 (GAW 16). It consists of 545.080 SNP-genotype fields from the Illumina 550K chip. The second data set is whole genome association data for Alzheimer's disease (AD). The data was obtained from the Alzheimer's disease Neuroimaging Initiative (ADNI) database (www.loni.ucla.edu/ADNI). The used subset of the ADNI data includes 149 AD cases and 182 controls. It consists of 555.963 SNP-genotype fields from the Illumina 610Quad chip. We applied an initial quality control based filtering with PLINK [27] tool for both data sets and we have excluded those SNPs whose minor allele frequency is less than 0.01, missingness rate is greater than 0.1 and those individuals whose genotype missingness rate is greater than 0.1. Additionally, we excluded those SNPs whose Hardy-Weinberg p-value is less than 0.001. By doing that we managed to reduce the SNP size for first data set to 501.463 and second data set to 517.180. We also applied imputation to cope with missing alleles by using BEAGLE [3].

5.2 Biological Prioritization

We used SPOT for extracting biologically relevant SNPs from the whole SNP sets following a genome-wide association run. For our analysis we used multiple test adjusted p-values (False Discovery Rate) to calculated weighted p-value through SPOT. We used SPOT default values for ECR node scoring parameters, Gene node scoring parameters - link indexes for SNP/gene functional properties. We used 0.25 as p-value threshold for the first data set. We used first 10,000 SNPs from prioritization results of second data set (as using a multiple adjusted p-value threshold of 0.25 filtered out almost all of the SNPs). After removing heterozygous haploid genotypes for both data sets finally there remain 9,083 SNPs in the first data set and 9,998 SNPs in the second data set.

5.3 Results

In order to test the classification performance of the representative SNP set on phenotype variable, we first applied an initial split (80%-20%) on the filtered biologically and statistically relevant data and seperate training and testing set for our model. Therefore, we have randomly selected 1,650 patients for the training set and 588 patients for the test set for RA data and 265 patients for the training set and 66 patients for the test set for AD data. Following that, we ran our simulated annealing based representative SNP selection algorithm on the training set. As the algorithm is based on the idea of selecting the subset of SNPs, which best predicts the remaining SNPs for a *particular genomic region*; we ran the algorithm for each chromosome and merged the selected SNPs as the overall representative SNP set. The prediction performance (mean accuracy) of the selected SNPs (on unselected SNPs) for each chromosome is presented in Tables 1-3 below for each data set:

Table 1 Prediction Performance of Representative SNP Selection Algorithm: w = 0.3, t = 10, d = 0.1, c_{max} = 1,000

	Rheumatoid Arthritis			Alzheimer's Disease		
Chromosome	Initial	Selected	Prediction Accuracy	Initial	Selected	Prediction Accuracy
1	687	20	0.628	781	35	0.728
2	790	31	0.620	743	35	0.702
3	438	11	0.592	813	39	0.709
4	486	17	0.620	542	17	0.672
5	542	14	0.590	567	14	0.679
6	1,410	90	0.725	832	33	0.683
7	393	11	0.580	532	18	0.717
8	595	24	0.615	569	19	0.698
9	461	22	0.614	430	9	0.619
10	402	10	0.606	478	18	0.675
11	343	12	0.606	530	19	0.675
12	345	9	0.601	442	13	0.611
13	297	5	0.578	311	10	0.622
14	280	5	0.580	336	7	0.611
15	258	7	0.583	282	3	0.566
16	280	7	0.599	297	14	0.653
17	184	5	0.578	291	12	0.675
18	179	1	0.582	351	11	0.634
19	137	4	0.577	167	2	0.585
20	206	7	0.584	317	11	0.615
21	132	5	0.588	119	2	0.581
22	103	1	0.590	127	3	0.581
23	135	4	0.587	141	3	0.566
TOTAL	9083	322		9998	347	

Table 2 Prediction Performance of Representative SNP Selection Algorithm: w = 0.5, t = 10, d = 0.1, c_{max} = 1,000

	Rheumatoid Arthritis			Alzheimer's Disease		
Chromosome	Initial	Selected	Prediction Accuracy	Initial	Selected	Prediction Accuracy
1	687	52	0.616	781	73	0.751
2	790	53	0.622	743	69	0.766
3	438	40	0.601	813	73	0.736
4	486	36	0.619	542	56	0.769
5	542	41	0.612	567	60	0.758
6	1,410	136	0.723	832	89	0.758
7	393	19	0.607	532	48	0.759
8	595	47	0.609	569	50	0.743

9	461	31	0.593	430	40	0.728
10	402	20	0.595	478	44	0.755
11	343	17	0.595	530	42	0.736
12	345	29	0.615	442	27	0.698
13	297	20	0.608	311	25	0.698
14	280	24	0.604	336	27	0.716
15	258	14	0.600	282	14	0.641
16	280	18	0.593	297	16	0.642
17	184	13	0.599	291	14	0.653
18	179	9	0.582	351	19	0.698
19	137	4	0.583	167	2	0.585
20	206	16	0.596	317	22	0.694
21	132	6	0.581	119	4	0.626
22	103	3	0.593	127	3	0.536
23	135	3	0.582	141	5	0.618
TOTAL	9083	651		9998	822	

Table 3 Prediction Performance of Representative SNP Selection Algorithm: w = 0.7, *t* = 10, *d* = 0.1, c_{max} = 1,000

	Rheumatoid Arthritis			Alzheimer's Disease		
Chromosome	Initial	Selected	Prediction Accuracy	Initial	Selected	Prediction Accuracy
1	687	114	0.629	781	116	0.823
2	790	120	0.620	743	100	0.789
3	438	83	0.592	813	124	0.804
4	486	91	0.620	542	81	0.770
5	542	93	0.590	567	70	0.779
6	1,410	217	0.709	832	126	0.770
7	393	54	0.612	532	94	0.791
8	595	97	0.606	569	76	0.752
9	461	91	0,610	430	69	0.773
10	402	65	0.606	478	80	0.770
11	343	62	0.606	530	83	0.809
12	345	48	0.601	442	61	0.767
13	297	52	0.578	311	53	0.734
14	280	39	0.580	336	67	0.728
15	258	32	0.583	282	45	0.729
16	280	47	0.599	297	55	0.740
17	184	20	0.578	291	36	0.716
18	179	21	0.582	351	41	0.731
19	137	15	0.584	167	28	0.692
20	206	29	0.584	317	50	0.737
21	132	21	0.587	119	10	0.686
22	103	16	0.590	127	15	0.659
23	135	6	0.587	141	12	0.658
TOTAL	9083	1433		9998	1492	

Using representative SNP selection algorithm we managed to decrease the dimensions considerably. For example for w = 0.5 the number of SNPs is decreased from 9,083 to 651 for RA data set and from 9,998 to 822 for AD data set. Average prediction accuracy (over not selected) of the selected SNP set for RA data is 0.605 and it is 0.698 for AD data. This means that although we decrease the dimension more than 90%, we do not introduce a considerably high information loss. To observe the classification performance of the selected set over the disease phenotype, we compared the performance against two filtering based attribute selection scheme from WEKA[9] tool set (Relief-F and Chi-Square) and randomly selected set of SNPs. In order to achieve that, we have selected the same set of SNPs for the test sets to that of training sets and applied a 10-fold Cross Validation (CV) run using Naive Bayes classifer as the supervised learning scheme. Results are presented in Tables 4-5 below:

Table 4 10-Fold CV Results for AD Data

	w = 0.3, 347 SNPs			w = 0.5, 822 SNPs			w = 0.7, 1492 SNPs		
Measure	SA-SNP	Chi-Square	Relief-F	SA-SNP	Chi-Square	Relief-F	SA-SNP	Chi-Square	Relief-F
Accuracy	0,5000	0,6061	0,7879	0,5606	0,5152	0,8788	**0,5455**	0,5455	0,5303
Recall	0,0000	0,5806	0,7419	0,1290	0,1613	0,8065	**0,0968**	0,0968	0,0000
NPV	0,5156	0,6286	0,7838	0,5500	0,5273	0,8462	**0,5410**	0,5410	0,5303
Precision	0,0000	0,5806	0,7931	0,6667	0,4545	0,9259	**0,6000**	0,6000	NA
Specificity	**0,9429**	0,6286	0,8286	**0,9429**	0,8286	0,9429	0,9429	0,9429	1,0000

Table 5 10-Fold CV Results for RA Data

	w = 0.3, 322 SNPs			w = 0.5, 650 SNPs			w = 0.7, 1433 SNPs		
Measure	SA-SNP	Chi-Square	Relief-F	SA-SNP	Chi-Square	Relief-F	SA-SNP	Chi-Square	Relief-F
Accuracy	0,5728	0,5607	0,6432	0,5558	0,5607	0,6189	**0,5728**	0,5607	0,4587
Recall	0,1713	0,0000	0,9227	0,0829	0,0000	0,9834	0,0497	0,0000	1,0000
NPV	0,5775	0,5607	0,8750	0,5632	0,5607	0,9625	0,5689	0,5607	1,0000
Precision	0,5439	NA	0,5567	0,4688	NA	0,5361	**0,6923**	NA	0,4480
Specificity	0,8874	1,0000	0,4242	0,9264	1,0000	0,3333	0,9827	1,0000	0,0346

Results reveal that our algorithm performs considerably better against well known filtering based attribute selection schemes especially when prediction accuracy is favored against minimizing cardinality of SNP set.

6 Conclusion

In this paper, we have presented a novel representative SNP Selection algorithm based on the idea of maximizing prediction accuracy of selected SNP set over non-selected. We have developed a methodology based on simulated annealing to help prioritize SNPs according to biological relevance alongside with p-value of association. We have performed biological prioritization and SNP selection on real life data belonging to Rheumatoid Arthritis and Alzheimer's disease and got promising results by reducing the dimension without much information loss. We have performed a comparative study with two well known attribute selection schemes. Our algorithm performed reasonably well against filtering based approaches. We hope that our work will facilitate reliable identification of SNPs that are involved in the etiology of complex diseases ultimately supporting timely

[9] http://www.cs.waikato.ac.nz/~ml/weka

identification of genomic disease biomarkers, and development of personalized medicine approaches and targeted drug discoveries. We also hope that our work will encourage OR community to explore research subject and to discover application areas of much more advanced OR methods such as conic programming, multi-objective optimization, stochastic programming, and optimization in data mining and in computational statistics.

Acknowledgements

The Genetic Analysis Workshops (GAW) are supported by NIH grant R01 GM031575 from the National Institute of General Medical Sciences. This work is based on GAW 16 NARAC Rheumatoid Arthritis data that was gathered with the support of grants from the National Institutes of Health (NO1-AR-2-2263 and RO1-AR-44422), and the National Arthritis Foundation.

We gratefully acknowledge the sharing of the ADNI dataset by the ADNI investigators[10]. Data collection and sharing for the Alzheimer dataset of this project was funded by the Alzheimer's disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering and other organizations. For full acknowledgment of funding sources of the ADNI project see: http://www.adni-info.org.

References

[1] S. I. Ao, K. Yip, M. Ng, D. Cheung, P. Fong, I. Melhado, and P. C. Sham. Clustag: hierarchical clustering and graph methods for selecting tag SNPs. *Bioinformatics*, 21:1735 -1736, 2005.

[2] V. Bafna, B. V. Halldorsson, R. Schwartz, and A.G. Clark. Haplotypes and informative SNP selection algorithms: don't block out information. In *Proceedings of the Seventh International Conference on Research in Computational Molecular Biology*, 19-20, 2003.

[3] Sharon R. Browning and Brian L. Browning. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics*, 81:1084–1097, 2007.

[4] M. C. Byng, J. C. Whittaker, A. P. Cuthbert, C. G. Mathew, and C. M. Lewis. SNP subset selection for genetic association studies. *Annals of Human Genetics*, 67:543-556, 2003.

[5] C. S. Carlson, M. A. Eberle, M. J. Rieder, Q. Yi, L. Kruglyak, and D. A. Nickerson. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *American Journal of Human Genetics*, 74:106-120, 2004.

[6] M. Corbex, O. Poirier, F. Fumeron, D. Betoulle, A. Evans, J. B. Ruidavets, D. Arveiler, G. Luc, L. Tiret, and F. Cambien. Extensive association analysis between the CETP gene and coronary heart disease phenotypes reveals several putative functional polymorphisms and gene-environment interaction. *Genetic Epidemiology*, 19:64-80, 2000.

[10] List of investors: http://loni.ucla.edu//ADNI//Collaboration//ADNI_Authorship_list.pdf

[7] M. J. Daly, J. D. Rioux, S. F. Schaffner, T. J. Hudson, and E. S. Lander. High resolution haplotype structure in the human genome. *Nature Genetics*, 29:229-232, 2001.

[8] A. Dembo and S. Karlin. Poisson approximations for r-scan processes. *The Annals of Applied Probability*, 2:329-357, 1992.

[9] F. Dudbridge and B.P.C. Koeleman. Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies. *American journal of human genetics*, 75:424 – 435, 2004.

[10] E. Halperin, G. Kimmel, and R. Shamir. Tag SNP selection in genotype data for maximizing snp prediction accuracy. *Bioinformatics*, 21:195–203, 2005.

[11] J. Hampe, S. Schreiber, and M. Krawczak. Entropy-based SNP selection for genetic association studies. *Human Genetics*, 114:36–43, 2003.

[12] J. Hoh and J. Ott. Scan statistics to scan markers for susceptibility genes. *Proceedings of National Academy of Science*, 97:9615-9617, 2000.

[13] B. Horne and N. J. Camp. Principal component analysis for selection of optimal SNP-sets that capture intragenic genetic variation. *Genetic Epidemiology*, 26:11– 21, 2004.

[14] G. C. L. Johnson, L. Esposito, B. J. Barratt, A. N. Smith, J. Heward, G. Di Genova, H. Ueda, H. J. Cordell, I. A. Eaves, F. Dudbridge, R. C. J. Twells, F. Payne, W. Hughes, S. Nutland, H. Stevens, P. Carr, E. Tuomilehto-Wolf, J. Tuomilehto, S. C. L. Gough, D. G. Clayton, , and J. A. Todd. Haplotype tagging for the identification of common disease genes. *Nature Genetics*, 29:233-237, 2001.

[15] R. Judson, B. Salisbury, J. Schneider, A. Windemuth, and J.C. Stephens. How many SNPs does a genome-wide haplotype map require? *Pharmacogenomics*, 3:379–391, 2002.

[16] X. Ke, M. M. Miretti, J. Broxholme, S. Hunt, S. Beck, D. R. Bentley, P. Deloukas, and L. R. Cardon. A comparison of tagging methods and their tagging space. *Human Molecular Genetics*, 14:2757-2767, 2005.

[17] X. Ke and L. R. Cardon. Efficient selective screening of haplotype tag snps. *Bioinformatics*, 19(2):287–288, 2003.

[18] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.

[19] L. Kruglyak and D. A. Nickerson. Variation is the spice of life. *Nature Genetics*, 27(3):234–236, March 2001.

[20] P. H. Lee and H. Shatkay. Bntagger: Improved tagging SNP selection using bayesian networks. *ISMB (Supplement of Bioinformatics)*, 22:211–219, 2006.

[21] A.M. Levin, D. Ghosh, K. R. Cho, and S. L. R. Kardia. A model-based scan statistic for identifying extreme chromosomal regions of gene expression in human tumors. *Bioinformatics*, 21(12):2867–2874, 2005.

[22] Z. Lin and R. B. Altman. Finding haplotype tagging snps by use of principal components analysis. *American Journal of Human Genetics*, 75:850-861, 2004.

[23] E. Lindholm, O. Melander, P. Almgren, G. Berglund, C.D. Agardh, L. Groop, and M. Orho-Melander. Polymorphism in the MHC2TA gene is associated with features of the metabolic syndrome and cardiovascular mortality. *PLoS ONE*, 1:e64, 12 2006.

[24] Z. Meng, D. V. Zaykin, C. Xu, M. Wagner, and M. G. Ehm. Selection of genetic markers for association analyses using linkage disequilibrium and haplotypes. *American Journal of Human Genetics*, 73:115-130, 2003.

[25] B. M. Neale and P. C. Sham. The future of association studies: Gene-based analysis and replication. *American Journal of Human Genetics*, 75:353–362, 2004.

[26] N. Patil, A. J. Berno, D. A. Hinds, W. A. Barrett, J. M. Doshi, C. R. Hacker, C. R. Kautzer andD. H. Lee, C. Marjoribanks, and D. P. McDonough. Blocks of limited haplotype diversity revealed by high resolution scanning of human chromosome 21. *Science*, 294:1719-1722, 2001.

[27] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M.A.R. Ferreira, D. Bender, J. Maller, P. Sklar, de P.I.W. Bakker, M.J. Daly, and P.C. Sham. PLINK: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics*, 81:559–575, 2007.

[28] S.F. Saccone, R. Bolze, P. Thomas, J. Quan, G. Mehta, E. Deelman, J.A. Tischfield, and J.P. Rice. Spot: a web-based tool for using biological databases to prioritize SNPs after a genome-wide association study, 2010.

[29] S.R. Seaman and B. Muller-Myhsok. Rapid simulation of p values for product methods and multiple-testing adjustment in association studies. *American Journal of Human Genetics*, 76(3):399 – 408, 2005.

[30] B. S. Shastry. Snps in disease gene mapping, medicinal drug development and evolution. *Journal of Human Genetics*, 52(11):871–880, 2007.

[31] Y. Sun, A. Levin, E. Boerwinkle, H. Robertson, and S. Kardia. A scan statistic for identifying chromosomal patterns of snp association. *Genetic Epidemiology*, 30:627-635, 2006.

[32] N. Tahri-Daizadeh, D. A Tregouet., V. Nicaud, N. Manuel, F. Cambien, and L. Tiret. Automated detection of informative combined effects in genetic association studies of complex traits. *Genome Research*, 13:1952-1960, 2003.

[33] D. A. Tregouet, S. Barbaux, S. Escolano, N. Tahri, J. L. Goldmard, L. Tiret, and F. Cambien. Specific haplotypes of the p-selectin gene are associated with myocardial infarction. *Human Molecular Genetics*, 11:2015-2023, 2002.

[34] K. Zhang. Dynamic programming algorithm for haplotype block partitioning: application to human chromosome 21 haplotype data. *Proceedings of the 7th Annual International Conference on Research in Computational Molecular Biology (RECOMB)*, page 332-340, 2003.

[35] K. Zhang, P. Calabrese, M. Nordborg, and F. Sun. Haplotype block structure and its application to association studies: power and study designs. *American Journal of Human Genetics*, 71:1386-1394, 2002.

[36] K. Zhang, M. Deng, T. Chen, M. S.Waterman, and F. Sun. A dynamic programming algorithm for haplotype block partitioning. *In Proceedings of the National Academy of Sciences (PNAS)*, 99:7335-7339, 2002.

[37] K. Zhang and L. Jin. Haploblockfinder: haplotype block analyses. *Bioinformatics*, 19:1300-1301, 2003.

[38] K. Zhang, Z. Qin, T. Chen, J. S. Liu, M. S. Waterman, and F. Sun. Hapblock: haplotype block partitioning and tag SNP selection software using a set of dynamic progrmming algorithms. *Bioinformatics*, 21:131-134, 2005.

[39] P. Zhang, H. Sheng, and R. Uehara. A double classification tree search algorithm for index SNP selection. *BMC Bioinformatics*, 5:89, 2004.