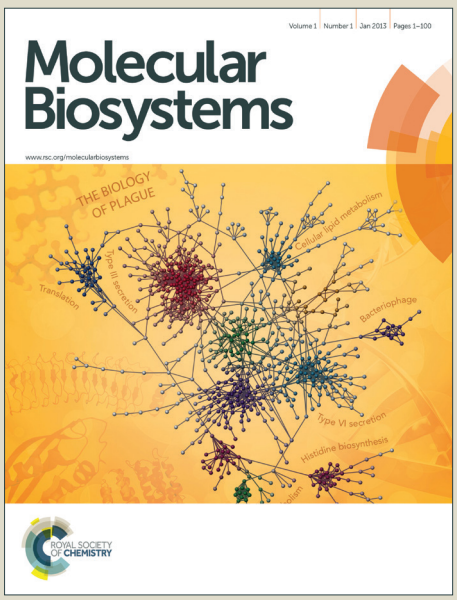
Molecular BioSystems

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems

Informative Bayesian Model Selection: A Method for Identifying Interactions in Genome-Wide Data

Mehran Aflakparast[1, 2], Ali Masoudi-Nejad[1*], Joseph H. Bozorgmehr[1], Shyam Visweswaran[3]

1. Laboratory of Systems Biology and Bioinformatics (LBB), Institute of Biochemistry and Biophysics, University of Tehran, Iran.
2. Department of Mathematics, Faculty of Sciences, VU University, Amsterdam, Netherlands
3. Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA 15260, USA

Running head: Informative Bayesian model selection for GWA data

***Corresponding Author**

Ali Masoudi-Nejad, Ph.D.
Laboratory of Systems Biology and Bioinformatics (LBB)
Institute of Biochemistry and Biophysics
University of Tehran, Tehran, Iran
E-mail: amasoudin@ibb.ut.ac.ir
www: http://LBB.ut.ac.ir
Tel: +98-21-6695-9256
Fax: +98-21-6640-4680

Abstract

In high-dimensional genome-wide (GWA) data, a key challenge is to detect genomic variants that interact in a nonlinear fashion in their association with disease. Identifying such genomic interactions is important for elucidating the inheritance of complex phenotypes and diseases. In this paper, we introduce a new computational method, called Informative Bayesian Model Selection (IBMS) that leverages correlation among variants in GWA data due linkage disequilibrium to identify interactions accurately in a computationally efficient manner. IBMS combines several statistical methods including canonical correlation analysis, logistic regression analysis, and a Bayesians statistical measure of evaluating interactions. Compared to BOOST and BEAM that are two widely used methods for detecting genomic interactions, IBMS had significantly higher power when evaluated on synthetic data. Furthermore, when applied to Alzheimer's disease GWA data, IBMS identified previously reported interactions. IBMS is a useful method for identifying variants in GWA data, and software that implements IBMS is freely available online from http://lbb.ut.ac.ir/Download/LBBsoft/IBMS.

Keywords: single nucleotide polymorphisms; genetic interactions; epistasis; genome-wide association studies; canonical correlation analysis; logistic regression; Bayesian model selection

Background

The elucidation of genetic variants that underlie complex phenotypes and diseases such as Alzheimer's disease remains a challenging problem. The most common type of genetic variation is the single nucleotide polymorphism (SNP) that results when a single nucleotide is replaced by another in the DNA sequence. The development of high-throughput genotyping technologies that simultaneously measure many thousands of SNPs have resulted in more than 600 genome-wide association (GWA) studies. However, many of the identified SNPs in GWA studies have only a small to moderate effect on the susceptibility of the disease (1, 2). One possible explanation for this observation is that interactions among SNPs including non-linear interactions may account for stronger effects. Non-linear interactions among genetic variants including SNPs are also known as epistatic interactions, and some progress has been made in recent years in developing computational and statistical methods for identifying such interactions in GWA data (2, 3). Methods that identify epistatic interactions in high-dimensional data such as GWA data have to address several challenges such as multiple testing, low power, and false positive rates. In typical GWA studies that measure more than a million SNPs, the number of potential epistatic interactions grows exponentially in the number of SNPs (4) and any interaction detection method has to address the problem of examining such a large number of potential interactions in an efficient fashion.

A characteristic of GWA data is the presence of extensive correlation among SNPs due to linkage disequilibrium (LD). Exploiting this correlation can help in reducing the number of SNPs to be examined for potential interactions. Two general categories of methods for reducing the number of variables (such as SNPs) are often used, namely, variable selection and variable extraction (5). Variable selection methods such as filter and wrapper methods select an optimal subset of

variables from the original set of variables. In contrast, variable extraction methods such as principal component analysis transform the original variables into a smaller set of more informative variables that retain the greatest amount of variation (6, 7).

Variables selection methods can be categorized into univariate and multivariate methods. Univariate variable selection methods have been used in analyzing GWA data (8) because they are computationally efficient. These methods primarily identify main effects of SNPs and ignore correlations or interactions among them. Multivariate variable selection methods such as Relief have been applied to GWA data too because of their ability to consider additional effects beyond main effects (9-11). While being effective in discarding irrelevant SNPs, Relief is unable to eliminate redundant SNPs. Thus, an important drawback of currently used variable selection methods for GWA data is that they may select a subset of correlated SNPs (1, 12, 13).

Computational methods including combinatorial methods have recently been developed to identify and characterize epistatic interactions (14). Combinatorial methods search over all possible combinations of SNPs to identify combinations that are predictive of the phenotype of interest. Multifactor dimensionality reduction (15-17) and the Bayesian combinatorial method (BCM) (18) are examples of combinatorial methods that identify associations between multiple SNPs and a phenotype by examining higher-order interactions among SNPs in case-control data. However, such methods that examine all possible subsets of SNPs can be applied only to data that consist of a few SNPs and are impractical for high-dimensional GWA data.

In this paper, we develop and evaluate a computationally efficient method called Informative Bayesian Model Selection (IBMS) that detects both SNP-SNP interactions and interactions between two groups of SNPs (e.g., a group may consist of SNPs that map to a gene). Given

grouped SNPs, this method consists of two main stages: 1) calculate group interactions that lead to weighting the two groups and their corresponding SNPs, and 2) identify interacting SNPs using the weights and a stochastic search strategy. IBMS combines canonical correlation analysis, logistic regression analysis, and BCM to efficiently identify epistatic SNPs in GWA data. Using synthetic data, we compare IBMS to two powerful and widely used methods for detecting genetic interactions, namely BOOST and BEAM. Furthermore, we apply IBMS to a late-onset Alzheimer's disease GWA dataset that contains over 300,000 SNPs.

Methods and Materials

Algorithmic Methods

This section provides background information on the Bayesian combinatorial method (BCM) which uses a Bayesian statistic for measuring genetic interactions, and canonical correlation analysis (CCA) which measures the linear relationship between two multidimensional variables. It then describes the informative Bayesian model selection (IBMS) method which is based on BCM, CCA and logistic regression analysis (LRA).

Bayesian Combinatorial Method

BCM searches over combinations of SNPs to identify combinations that have a strong statistical association with the phenotype. Specifically, it exhaustively searches over all possible combinations of SNPs and identifies combinations with a high posterior probability using a Bayesian statistical method (18). BCM has several advantages including the ability to handle sparse and unbalanced data, ability to deal with nonlinear interactions, and is computationally efficient.

In BCM, an interaction model M is defined as a set of probabilities $\boldsymbol{\theta}_c$ that is represented as $P(Z|\boldsymbol{g} = (g_1, g_2, .., g_c))$ for phenotype Z, given a combination of SNP genotypes $\boldsymbol{g}$. For a given $\boldsymbol{g}$ value, a multinomial distribution is assumed for Z (binomial, if Z has only two states). Assuming that the parameters of all multinomial distributions i.e. $\boldsymbol{\theta}_c$ a priori follow a Dirichlet distribution, a posterior estimate for $\boldsymbol{\theta}_c$ is obtained. The Bayes theorem is used to compute the score of an interaction model as follows:

$$P(M|Data) \propto P(Data|M)P(M) \qquad (1)$$

where $P(M)$ is the prior probability of model M, which is assumed to be uniform over all models and $P(Data|M)$ is the marginal likelihood, which is evaluated with the following equation:

$$P(Data|M) = \int P(Data|M, \boldsymbol{\theta}_c)\, P(\boldsymbol{\theta}_c|M) d\boldsymbol{\theta}_c \qquad (2)$$

where $P(Data|M, \boldsymbol{\theta}_c)$ is the distribution of the data for a given genotype-phenotype table. Figure 1 presents an example of a genotype-phenotype table that gives counts obtained from data for an interaction model with two SNPs (denoted SNP1 and SNP2) and a binary phenotype (e.g., case and control).

[Figure 1]

A binomial distribution for each column (i.e., the combination of genotypes for SNP1 and SNP2) is assumed. Thus, $P(Data|M, \boldsymbol{\theta}_c)$is obtained by multiplying nine independent binomial distributions in Figure 1.

The closed form for $P(Data|M)$ is given by the following equation and was originally derived by Cooper and Herskovits (19):

$$P(Data|M) = \prod_{i=1}^{I}\left(\frac{(\alpha_i-1)!}{(n_i+\alpha_i-1)!}\prod_{j=1}^{J}\frac{(n_{ij}+\alpha_{ij}-1)!}{(\alpha_{ij}-1)!}\right) \quad (3)$$

where α_{ij} are the hyperparameters of a Dirichlet distribution with $\sum \alpha_{ij} = \alpha_i$. I is the number of genotype combinations (e.g., nine for a model with two SNPs), J is the number of phenotype states (e.g., two for case-control data), n_i is the number of samples for a given genotype combination of an epistatic model, and n_{ij} is the number of samples for a given phenotype state j and genotype combination i of a model. Assuming that the prior distribution $P(M)$ is uniform over all possible models and the hyperparameters of the Dirichlet distribution are all set to 1, the following expression gives the score that is used by BCM for an interaction model:

$$Score_{BCM}(M) = \prod_{i=1}^{I}\left(\frac{(J-1)!}{(n_i+J-1)!}\prod_{j=1}^{J} n_{ij}!\right) \quad (4)$$

BCM produces a posterior probability of association of a combination of SNPs of interest with the phenotype. The higher this probability the stronger is the interaction model's association with the phenotype. If it is desired to obtain a small list of high probability combination of SNPs pairs a threshold of posterior probability ≥ 0.95 may be used. A major limitation of BCM is that it searches exhaustively over all possible combinations of SNPs and hence it does not scale up to high-dimensional data. The IBMS method overcomes this limitation by computing an informative prior over models (instead of the uniform prior used in BCM) and by performing stochastic search over combinations of SNPs (instead of the exhaustive search used in BCM).

Canonical Correlation Analysis

Canonical correlation analysis (CCA) was developed by Hotelling (20, 21) for characterizing relationships among multiple dependent and independent variables. Given two sets of variables,

$\boldsymbol{X} = (X_1, X_2, .., X_p)$ and $\boldsymbol{Y} = (Y_1, Y_2, .., Y_q)$ that are measured on the same set of n objects, CCA constructs pairs of new variables as linear combinations of $\boldsymbol{X}$ and $\boldsymbol{Y}$ such that the correlation in the new pair is maximized. The new variables are defined in such a way that they explain the largest amount of variance in the data. CCA outputs a number of estimated equations called canonical functions; each canonical function provides two canonical variants representing the optimal linear combinations of $\boldsymbol{X}$ and $\boldsymbol{Y}$, and the canonical correlation coefficient $\boldsymbol{R}$ which represents the linear relationship between them. The first canonical function identifies linear combinations of original variables that yield the largest canonical correlation coefficient; the second identifies linear combinations of original variables that are not correlated with the first pair of canonical variants and yield the second largest canonical correlation coefficient; and so on. The output of CCA include (1) canonical weights that are defined as coefficients that are assigned to the original variables (also called standardized coefficients), (2) canonical loadings that are defined as correlations between the original variables and their corresponding canonical variants, and (3) canonical cross-loadings that are defined as the correlation between original variables of one set $\boldsymbol{X}$ and canonical variants of the other set $\boldsymbol{Y}$. In multivariate analysis, when the original variables increase in their correlation with each other canonical loadings and canonical cross-loadings are more often employed in interpreting the results (22).

CCA has several characteristics that make it appropriate to be used as part of IBMS. The problem of multiple testing is mitigated by CCA by limiting the inflation of Type I error. IBMS creates groups of SNPs (a *group* of SNPs is defined as a set of SNPs that map to a particular *gene* or a set of SNPs that are in LD in a genomic *region*) and performs CCA on every pair of groups. In IBMS, we use canonical cross-loadings and canonical loadings to measure the significance of each SNP in its group. CCA helps to identify significant gene-gene or region-

region interactions, which in turn avoids testing the large number of all possible SNP-SNP interactions. This is described in detail in the next section.

Informative Bayesian Model Selection

The IBMS method combines CCA and LRA with the interaction model score used in BCM. IBMS uses a two-stage approach to first identify pairs of groups of SNPs that interact and then identifies individual SNPs across a pair of groups that interact. This approach achieves computational efficiency by partitioning SNPs into groups and then selecting SNP combinations from highly weighted groups to be evaluated with the BCM score.

After grouping SNPs (two approaches to grouping are described in the next section), in the first stage we use CCA to measure the informativeness of each SNP using the loading and cross-loading values. Then we use LRA to identify group-group interactions that are significantly associated with the phenotype. A weight is assigned to each group by calculating the frequency of having significant interactions with the remaining groups. Therefore, assuming *Group* as a random variable, a discrete probability distribution is obtained. In the second stage, IBMS stochastically searches the space of interaction models using an Independent Metropolis-Hastings algorithm to identify significant SNP-SNP interactions. In the following sections we describe the IBMS method in more detail (see Figures 2 and 3).

Grouping of SNPs

We use two ways to partition SNPs:

1. Partition SNPs according to their associated genes such that SNPs on a gene are in a single group. This grouping leads to identifying gene-gene interactions prior to identifying SNP-SNP interactions.
2. Partition SNPs based on LD so that SNPs in a group have high LD. Typically, such a group contains SNPs that lie adjacent to each other in some genomic region. This may be approximated by grouping together a constant number of SNPs successively along the genome.

Determining which partitioning scheme to use depends on the goals of the analyses and the computational costs. If the goal is to identify gene-gene and SNP-SNP interactions, then partitioning SNPs according to genes is used. If the goal is to identify only SNP-SNP interactions, then partitioning SNPs based on LD is preferred.

In addition, a partitioning scheme should balance the sample size (i.e., total number of cases and controls) and group size (i.e., number of SNPs or variables in each group). In CCA, the ratio of sample size to the number of variables plays an important role in the significance of statistical findings. Very small ratios will not represent the correlations well, thus obscuring any meaningful relationships. Very large ratios, on the other hand, may lead to inflated statistical significance. Since, our main focus was to develop a procedure of detecting interaction effects rather than proposing a specific procedure to partition SNPs, we selected the partitioning method and its parameters based on limited analyses of synthetic data. However, in order to comprehensively consider the effect of these parameters, sensitivity analysis should be performed.

Other ways of partitioning SNPs can also be considered. For instance, similarity of genes based on function and biological process (as defined, e.g., by the gene ontology), genes grouped by biological pathways (as defined, e.g., in the KEGG knowledge base), can be used to group SNPs.

Stage 1: Weighting

We use a novel approach for weighting SNPs and groups of SNPS using a combination of CCA and LRA (see Figure 2). The steps used in deriving the weights are as follows:

1. Let $\boldsymbol{S} = (s_1, s_2, .., s_L)$ be the set of all SNPS in the data. Partition $\boldsymbol{S}$ into m groups such that:

   $$\boldsymbol{S} = \bigcup_{i=1}^{m} \boldsymbol{S}_i \ , \ \boldsymbol{S}_i \cap \boldsymbol{S}_j = \emptyset \qquad \boldsymbol{S}_i = (s_{1i}, s_{2i}, ..., s_{n_i i}),\ i \neq j = 1, .., m$$

2. Apply CCA to every pair of groups. Each application of CCA results in a set of canonical variables for each group. Select the first optimal canonical variable for each group, namely, U and V which account for the largest amount of variation. Then apply LRA and perform the Wald test for a single coefficient in order to test interaction of groups (i.e.., $H_0: \beta_3 = 0$) in association with the phenotype as follows:

   $$logit\big(P(Z_i = 1)\big) = \beta_1 U_i + \beta_2 V_i + \beta_3 U_i V_i \qquad (5)$$

3. From the LRA tests, for a group $\boldsymbol{S}_i$ use the frequency of statistically significant interaction effects to determine a weight of informativeness using the following expression:

   $$P(\boldsymbol{S}_i) = \frac{\sum_{j=1(j\neq i)}^{m} I_{ij}}{\sum_{i=1}^{m}\sum_{j=1(j\neq i)}^{m} I_{ij}} \qquad (6)$$

   where I_{ij} is an indicator variable that signifies whether the null hypothesis $H_0: \beta_3 = 0$ is rejected (i.e., $I_{ij} = 0$) or not.

4. For $i = 1, \ldots, m$ calculate:

$$g_j(s_{il}) = \left|cl_j(s_{il})\right| + \left|ccl_j(s_{il})\right| \qquad j = 1, \ldots, m \quad l = 1, \ldots, n_i \qquad (7)$$

$$\omega_j(s_{il}) = \frac{g_j(s_{il})}{\sum_{p=1}^{n_i} g_j(s_{ip})} \qquad l = 1, 2, \ldots, n_i \qquad (8)$$

where $\omega_j(s_{il})$ is a measure of informativeness for the lth SNP in the ith group, when CCA is applied to $\boldsymbol{S}_i$ and $\boldsymbol{S}_j$. The canonical loading and canonical cross-loading for the lth SNP in the ith group, when CCA is applied to $\boldsymbol{S}_i$ and $\boldsymbol{S}_j$, are denoted as $cl_j(s_{il})$ and $ccl_j(s_{il})$ respectively.

Stage 2: Stochastic Search

In the second stage, using the outputs from the first stage, we use stochastic search to score a set of interaction models (see Figure 3). The steps in the search are as follows:

1. For a predefined c to construct c-way interaction models, start the Metropolis algorithm by sampling c different groups $(\boldsymbol{S}_1^{(0)}, \ldots, \boldsymbol{S}_c^{(0)})$ from all the groups using the probability distribution given by Equation 6.
2. Choose c SNPs, $(s_1^{(0)}, \ldots, s_c^{(0)})$, one SNP from each group, as elements of the first interaction association model $M^{(0)}$. Then calculate the model probability using the probability distribution from the previous stage as follows:

$$P\left(M = M^{(0)}\right) = \frac{\sum_{k=1}^{c} h(s_k^{(0)})}{\Phi_c} \qquad (9)$$

$$h(s_k^{(0)}) = \frac{\sum_{j=1(j\neq i)}^{c} \omega_j(s_k^{(0)})}{\sum_{j=1(j\neq i)}^{c} \sum_{p=1}^{n_k} \omega_j(s_{kp})} \quad k = 1, 2, \ldots, cl_k = 1, 2, \ldots, n_k \qquad (10)$$

where Φ_c is a constant that represents the sum of weights for any possible c-way model in the data. Determining this value can be ignored, since in the following steps it appears both in the numerator and the denominator of the fraction.

3. For a defined value of c, the stochastic search will consider c groups. Accordingly, the search method will need the results of the weighting stage of $\binom{c}{2} = \frac{c!}{(c-2)!2!}$ analyses (i.e., one CCA for each pair of groups) to assign weights to the SNPs of the selected groups. Each CCA calculation outputs two sets of weights such that each set contains the weights for the SNPs of a group. Hence, $2 \times \binom{c}{2}$ columns of weights are reported. This means that c-1 weights are assigned for every SNP in its group. The total weight of a SNP is computed as the sum of the c-1 weights, which forms the numerator in Equation 10.
4. Sample c SNPs $(s_1^{(1)},\ldots,s_c^{(1)})$ from the selected groups in step 1, one SNP from each group, as the elements of the next interaction association model $M^{(1)}$. Then calculate the following ratio:

5. $$\lambda = ln\left(\frac{P(\text{Data}|M^{(1)})P(M^{(1)})}{P(\text{Data}|M^{(0)})P(M^{(0)})}\right) \quad (11)$$

where $P(\text{Data}|M^{(1)})$ is the BCM score given by Equation 4, and $P(M^{(1)})$ is the prior probability of an interaction model given by Equation 10.

6. If $\lambda \geq 0$: update model: $M^{(1)} \rightarrow M^{(0)}$, else, update model with the probability λ.
7. Repeat steps 3-4 until the number of predefined within-group iterations is reached. Then report the last k resulting interaction models.

8. Repeat steps 1-5 until the number of predefined among-groups iterations is reached.
9. Compare all the resulting interaction models obtained from step 5 using steps 3-4, and report the final *k* interaction models.

Experimental Methods

This section provides details of the synthetic and the GWA datasets and the comparison algorithmic methods used in our experiments.

Synthetic SNP data

The synthetic datasets that we used were generated from a set of 70 epistatic models which were previously developed (17) and used in evaluating interaction detection methods (17, 23-25). The models have two minor allele frequency (MAF) values of 0.2 and 0.4 and seven heritability (H) values of 0.01, 0.025, 0.05, 0.10, 0.20, 0.30, and 0.40 (Velez et al., 2007 provides a detailed description of these genetic models).

For each model, 100 datasets were generated for each of four sample sizes (200, 400, 800 and 1600) where each dataset contains equal number of case and control samples. The epistatic models were used to generate a pair of epistatic SNP values, and a set of 18 SNPs that were assigned random values was appended to simulate SNPs that are non-informative with respect to the case/control status. Thus, each sample contained values for 20 SNPs of which only two SNPs were functionally related to the case/control status. These synthetic datasets are available online at http://discovery.dartmouth.edu/epistatic_data/#VelezDataModels.

Alzheimer's Disease GWA data

Alzheimer's disease (AD) is the commonest neurodegenerative disease associated with aging and the commonest cause of dementia (26). AD affects about 3% of all people between ages 65 and 74, about 19% of those between 75 and 84, and about 47% of those over 85. AD is characterized by adult onset of progressive dementia that typically begins with subtle memory failure and progresses to a slew of cognitive deficits like confusion, language disturbance and poor judgment (27).

AD is typically divided into early-onset Alzheimer's disease (EOAD) in which the onset of disease is before 65 years of age and late-onset Alzheimer's disease (LOAD) in which the onset is at 65 years of age or later. EOAD is rare and exhibits an autosomal dominant mode of inheritance. The genetic basis of EOAD is well established, and mutations in one of three genes (amyloid precursor protein gene, presenelin 1, or presenelin 2) account for most cases of EOAD (28).

LOAD is widespread and is estimated to affect almost half of all people over the age of 85. LOAD is believed to be a disease with both genetic and environmental influences, and elucidating the role of genetic factors in the pathogenesis and development of LOAD has been a major focus of research for more than a decade. One genetic risk factor for LOAD that has been consistently replicated is the apolipoprotein E (APOE) locus (29) determined by the combined genotypes at the loci rs429358 and rs7412. In the past few years, GWASs have identified several additional genetic loci associated with LOAD.

The LOAD GWA data we used were collected and analyzed originally by Reiman et al. The genotype data collected about 1411 samples that contained 861 cases diagnosed with late-onset Alzheimer's disease (LOAD) and 550 controls; 644 were APOE ε4 carriers (one or more copies

of the ε4 allele) and 767 were non-carriers. Of the 1411 samples, the status of case/control was neuropathologically determined from brain tissue in 1047 samples and was determined clinically in 364 samples. In this dataset, 61% (861 of 1411) had LOAD. For each individual, the genotype data consist of 502,627 SNPs that were measured on an Affymetrix chip; the original investigators analyzed 312,316 SNPs after applying quality controls. We used those 312,316 SNPs, plus two additional APOE SNPs from the same study namely, rs429358 and rs7412.

Evaluation of IBMS

We evaluated three aspects of IBMS on synthetic data. First, we assessed its ability to correctly rank SNPs in data generated from different genetic models and of varying sample sizes. Second, we compared the performance of IBMS with BCM on the hardest to detect interactions that are generated by genetic models that have low H and low MAF values. Third, we compared the performance of IBMS to two commonly used genetic interaction methods, namely, BEAM and BOOST. Although several other methods for detecting genetic interactions have been described in the literature, we restricted our evaluation to BEAM and BOOST, since these methods have been shown to be superior and scalable over other methods.

BEAM is a Bayesian-based epistasis detection method that partitions genetic markers into three categories (30). The first category contains markers assumed to have no impact on the phenotype, the second category contains associated markers that are assumed to have main effects, and the third category contains markers that are assumed to have main and interaction effects. BEAM uses a novel Bayesian statistic to exhaustively score interactions among markers. The BEAM software is available from http://www.fas.harvard.edu/~junliu/BEAM.

BOOST is an exhaustive search method that identifies two-locus interactions in GWA data using log-linear models. It proposes an upper bound for the likelihood ratio test statistic to prune insignificant epistatic interactions. This procedure approximates the test statistic which reduces the computational cost to a considerable degree. Moreover, it uses a Boolean representation of the genotype data which allows efficient collection of counts for genotype-phenotype tables using logic operations (31). The BOOST software is available from http://bioinformatics.ust.hk/BOOST.html.

Results

This section describes the results obtained from applying IBMS and the comparison methods to synthetic data and the results obtained from applying IBMS to the LOAD GWA data.

Synthetic data results

Using the synthetic datasets we examined the highest scoring two-locus interactions using IBMS and two comparison methods, namely, BOOST and BEAM. Figure S1 gives the detailed results for 70 pure epistasis models without main effects. For each genetic model, we defined *power* of the method as the proportion of the 100 replicate datasets for which the method ranked the two interacting SNPs as the top two SNPs. For a small number of models, such as the models with H and MAF both set to 0.2, the statistical power of BOOST is slightly higher than IBMS. This almost always happened with stochastic search methods compared to exhaustive search methods.

For all other genetic models IBMS outperforms both BOOST and BEAM. The power of all methods was affected by H and MAF in that lower values resulted in lower power. However,

when the values for H and MAF are low (e.g., H = 0.01, MAF = 0.2) IBMS has nearly double the power as BOOST or BEAM.

The average power of the three methods over all 70 genetic models shows that IBMS performs better than the two comparison methods (see Table 1). In particular, at low sample sizes BOOST and BEAM had lower power compared to IBMS. This is due to the nature of IBMS which provides a complementary combination of a SNP-weighting linear function and a Bayesian scoring non-linear function.

All experiments were conducted with a desktop computer with a 2.26 GHz CPU and 4 GB of RAM. Table 2 gives the average running time for each method on 100 datasets with 20 SNPs and different sample sizes. The average running times of IBMS are the lowest among all methods. BEAM has higher running times compared to other two methods. It is possible to reduce the running time of BEAM by adjusting its MCMC parameters, but reducing the running time typically leads to lower power. On the other hand, BOOST has lower running times due to its pruning and Boolean operation techniques.

We also compared the performance of IBMS to that of BCM on the five most challenging genetic models, namely, models 55-59 that have low H and low MAF values. The powers of the two methods on these models are shown in Table 3, and the running times for the two methods are shown in Figure 2. The results show that IBMS achieves higher power with lesser lower running times compared to BCM. The reduced running time is due to IBMS examining fewer interaction models compared to BCM. The higher power is because of the non-constant value of the interaction model probability which makes IBMS scoring function distinct from that of BCM.

Since, the weighting stage is the critical component of IBMS, we evaluated IBMS in its ability to rank SNPs. For this analysis, we used the synthetic datasets with different combinations of H and MAF values. First, we partitioned a dataset into two groups such that each group contained one of the interacting SNPs. Then, we applied CCA and obtained weights for all SNPs in either of the two groups and ranked SNPs according to their weights such that a SNP with the highest weight was assigned a rank of one and so on. After examining 100 datasets of each genetic model, we determined the average rank for each of the two interacting SNPs over 100 runs (see Figure 4). The results show that the interacting SNPs are highly weighted on the whole. This becomes more pronounced as the sample size grows. For instance, IBMS achieves excellent performance in ranking SNPs for the sample size of 1600 with an average rank of close to one for the interacting SNPs.

[Figure 4]

LOAD GWA data results

We used a LOAD GWA dataset to demonstrate the application of IBMS on a genome scale dataset. This dataset contained 234,665 SNPs from 861 cases and 550 controls. We used IMPUTE (http://mathgen.stats.ox.ac.uk/impute/impute.html) for imputing the missing genotypes. We applied the genotypic test using the chi-square statistic with 2 degrees of freedom using the PLINK software, and retained 76,755 SNPs with p-values less than 0.2 for further analysis.

To identify potential epistatic interactions, we applied IBMS to identify pairs of interacting SNPs. In the first stage, we partitioned the SNPs into 295 groups with each group containing 260 adjacent SNPs (with the exception that the last group contained 315 SNPs). CCA was applied to every pair of groups. Then, the first canonical variables, one for each one group, was extracted

and analyzed with LRA to detect group-interactions. The LRA model contained two SNP variable terms and one SNP interaction term, and a t-test was performed on all pairs of group-variables (i.e., canonical variables) to statistically test the interaction term. After Bonferroni correction for multiple testing, we used a p-value threshold of $\frac{0.05}{\binom{295}{2}}$ to reject or accept the interaction term. The application of the first stage showed that there was a significant interaction between the groups containing SNPs mapped to the APOE and GAB2 genes. This interaction was previously reported by the original authors (34). The APOE group of SNPs obtained a higher weight than the GAB2 group.

In the second stage, the stochastic search identified several SNPs that interact with rs7412 which is a well characterized SNP on the APOE gene that is associated with LOAD (32-34). Table S1 gives the top 50 high scoring SNPs that interact with SNP rs7412. In particular, five previously reported SNPs namely rs901104, rs4291702, rs71158590, rs4945261, and rs2510038 in the GAB2 gene obtained low ranks and high interaction scores (see Table S1).

The running time for the weighting stage was approximately 4 hours on a desktop computer with a CPU of 2.66 MHz and RAM of 4 GB running the 32-bit Windows 7 operating system. The running time for the stochastic search stage was approximately 17 hours.

Discussion

A range of methods have been described in the literature to identify true disease associated genetic variants in high-dimensional GWA data that contain a large and highly redundant set of SNPs. Such data present several analytic challenges especially in the detection of interacting SNPs. In this paper, we described and evaluated a computationally efficient method called

IBMS that identifies SNP-SNP interactions and interactions between two groups of SNPs in high-dimensional GWA data. The IBMS method is a two-stage method that combines CCA, LRA, and BCM to detect epistatic SNPs.

The results demonstrate the utility of IBMS in identifying epistatic interactions. Compared to existing methods such as BOOST and BEAM, IBMS performed better in identifying interacting SNPs. Moreover, it is computationally efficient for application to high-dimensional GWA data.

The IBMS method can be considered as a variable selection method. Most existing variable selection methods that are applied to SNP data focus on identifying a subset of SNPs with good classification accuracy, and are typically guided by the main effects of the individual SNPs. IBMS, however, weights SNPs based on their main effects as well as on their interaction effects. Thus, SNPs that are selected by IBMS may provide even better classification accuracy by the inclusion of interaction effects.

A key advantage of IBMS is in addressing the multiple-testing problem. In GWA data, as the number of SNPs increases, the number of tests for association of single SNPs or their combinations with the disease becomes astronomical. Therefore, the results become unreliable because of the large number of false positives. In IBMS we tackled this problem by partitioning the SNPs into groups, and reduced the number of tests to a great degree by performing analyses for only two-group interactions. By grouping SNPs based on adjacency on the genome we take advantage of the correlation due to LD between nearby SNPs. By grouping SNPs into groups according to their associated genes, and weighing the informativeness of each group, we can identify more informative groups of SNPs that can facilitate the selection of candidate genes for future biological experiments.

Partitioning of SNPs into groups plays an important role in IBMS, and one limitation of our method is that we have considered only two simple approaches for grouping SNPs. In future work, we plan to explore new approaches for grouping SNPs based on combining different sources of biological data. Such new grouping methods have the potential to further improve the performance of IBMS.

Another limitation of IBMS is that although the interaction scoring function is non-linear, the CCA function is linear. Thus, IBMS considers only linear relations for group-group interactions. In future work, we plan to extend IBMS to use non-linear canonical correlation methods.

In conclusion, we have developed a computationally efficient and accurate method for detecting interactions among genomic variants in high-dimensional data. We hope that researchers will find IBMS to be a useful tool in the analyses of GWA data and that future extensions will lead to additional improvements.

Author contributions

MA and AMN designed the study; MA developed the algorithmic methods and performed the experiments and the analyses; SV helped with the analysis; MA, SV, JHB, and AMN drafted the manuscript; AMN and SV edited the manuscript.

Competing interest statement

The authors declare no competing financial interests.

Data archiving

This article does not report new empirical data. The software for IBMS, the documentation and the tutorial, are available from the following URL: http://lbb.ut.ac.ir/Download/LBBsoft/IBMS

References

1. Moore JH, Asselbergs FW, Williams SM. Bioinformatics challenges for genome-wide association studies. *Bioinformatics*. 2010; 26(4):445-455.

2. Wang Y, Liu G, Feng M, Wong L . An empirical comparison of several recent epistatic interactions detection methods. *Bioinformatics*. 2011; 27(21):2936-2943.

3. Moore JH, Wiliams SM. Epistasis and its implications for personal genetics. *Hum Genet*. 2009; 85:309-3290.

4. Bellman R, Kalaba R. A mathematical theory of adaptive control processes. *Proc Natl Acad Sci USA*. 1959; 45:1288-1290.

5. Steen KV . Travelling the world of gene-gene interactions. *Briefings in Bioinformatics*. 2011;10:1-19.

6. Jolliffe IT. *Principal Component Analysis*. 2nd ed., Springer series in statistics. 2002.

7. Guyon I, Gunn S, Nikravesh M, et al. Feature extraction, foundations and applications. In: Guyon I, Gunn S, Nikravesh M, Zadeh L, (eds). Series Studies in Fuzziness and Soft Computing. Physica-Verlag: Springer. 2006.

8. Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics*. 2007; 23:2507-2517.

9. Kira, K. and L. A. Rendell. The feature selection problem: traditional methods and new algorithm. 1992; In: Proceedings of AAAI'92.

10. Robnik-Sikonja, M. and I. Kononenko. An adaptation of Relief for attribute estimation in regression. In: D. H. Fisher (ed.): *Machine Learning: Proceedings of the Fourteenth International Conference* (ICML'97). 1997; 296–304.

11. Robnik-Sikonja, M. and I. Kononenko. Theoretical and Empirical Analysis of ReliefF and RReliefF. M*achine Learning Journal*. 2003; 53:23-69.

12. Cordell HJ. Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet. 2009; 10:392-404.

13. Liang Y, Kelemen A. Statistical advances and challenges for analyzing correlated high dimensional SNP in genomic study for complex diseases. *Statistics Surveys*. 2008; 2:43-60.

14. Heidema AG, Boer JMA, Nagelkerke N, Mariman ECM, Van der AD, Feskens EJM. The challenge for genetic epidemiologists: how to analyze large numbers of SNPs in relation to complex diseases. *BMC Genet*. 2006; 7:23.

15. Hahn LW, Ritchie MD, Moore JH. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics*. 2003; **19:**376-382.

16. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet*. 2001; **69:**138-147.

17. Velez DR, White BC, Motsinger AA, Bush WS, Ritchie MD, Williams SM, et al. A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. *Genetics Epidemiology*. 2007; 31(4):306-315.

18. Visweswaran S, Wong AL, Barmada MM. A Bayesian Method for Identifying Genetic Interactions. *AMIA Annu Symp Proc. 2009;* 673-677.

19. Cooper GF, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*. 1992; 9(4) 309-347.

20. Hotelling H. The most predictable criterion. *The Journal of Educational Psychology. 1935;* 26(2):139-143.

21. Hotelling H. Relations between two sets of variates. *Biometrika*. 1936; 28:321-377.

22. Nimon K, Henson RK, Gates MS. Revisiting Interpretation of Canonical Correlation Analysis: A Tutorial and Demonstration of Canonical Commonality Analysis. *Multivariate Behavioral Research*. 2010; 45:702-724.

23. Wan X, Yang C, Yang Q, Xue H, Tang NL, Yu W. Detecting two-locus associations allowing for interactions in genome-wide association studies. *Bioinformatics*. 2010 Oct 15; 26(20):2517-25.

24. Jiang X, Barmada MM, Cooper GF, Becich MJ. A Bayesian method for evaluating and discovering disease loci associations. *PLoS ONE*. 2011; 6(8):e22075.

25. Aflakparast M, Salimi H, Gerami A, Dubé M-P, Visweswaran S, Masoudi-Nejad A. Cuckoo search epistasis: a new method for exploring significant genetic interactions. *Heredity*. 2014; 112, 666–674.

26. Goedert M, Spillantini MG. A century of Alzheimer's disease. *Science*. 2006; 314(5800):3777-781.

27. Bertram L, Lill CM, Tanzi RE. The genetics of Alzheimer disease: back to the future. *Neuron*. 2010; 68:270-281.

28. Avramopoulos D. Genetics of Alzheimers disease: Recent advances. *Genome Med*. 2009; 1-3

29. Coon KD, Myers AJ, Craig DW, Webster JA, Pearson JV, Lince DH, et al. A high-density whole-genome association study reveals that APOE is the major susceptibility gene for sporadic late-onset Alzheimer's disease. . *J Clin Psychiatry* 2007; 68(4):613.

30. Zhang X, Liu JS. Bayesian inference of epistatic interactions in case-control studies. *Nat. Genetics*. 2007; 39:1167-1173.

31. Wan X, et al. BOOST: a fast approach to detecting gene-gene interactions in genome-wide case- control studies. *Am. J. Hum. Genetics*. 2010; 87:325-340.

32. Combarros O, et al. Epistasis in sporadic Alzheimers disease. *Elsevier*, Neurobiology of Aging. 2009; 30:1333-1349.

33. Shi H, et al. Analysis of Genome-Wide Association Study (GWAS) data looking for replicating signals in Alzheimers disease (AD). *Int J MolEpidemiol Genet*. 2010; 1(1):53-66.

34. Reiman EM, et al. GAB2 alleles modify Alzheimers risk in APOE E 4 Carriers. *Neuron*. 2007; 54:713-720

Table 1. Average power of the three interaction methods obtained by averaging over 70 genetic models.

Sample size	IBMS	BOOST	BEAM
1600	0.96	0.94	0.67
800	0.90	0.85	0.64
400	0.84	0.69	0.41

Table 2. Average running times in seconds for 100 datasets for the three interaction methods.

Sample size	IBMS	BCM	BOOST	BEAM
1600	11 s	17 s	11 s	150 s
800	7 s	11 s	8 s	85 s
400	5 s	8 s	6 s	55 s

Table 3. Comparison of the power of IBMS and BCM on the five most challenging synthetic genetic models.

Sample size	Method	Model 55	Model 56	Model 57	Model 58	Model 59
1600	IBMS	0.67	0.75	0.72	0.97	0.51
	BCM	0.66	0.71	0.68	0.96	0.51
800	IBMS	0.31	0.34	0.27	0.64	0.21
	BCM	0.27	0.29	0.28	0.56	0.19
400	IBMS	0.13	0.12	0.12	0.27	0.10
	BCM	0.07	0.11	0.10	0.27	0.05

Figure Legends

Figure 1. An example of genotype-phenotype table with 2 SNPs and a phenotype with two states (e.g., case and control). The counts in the table are obtained from a dataset of genotypes that have been measured on a group of cases and controls.

Figure 2. Flowchart showing the weighting stage of IBMS**.**

Figure 3. Flowchart showing the stochastic search stage of IBMS.

Figure 4. Performance of IBMS in ranking SNPs on synthetic data using the first stage of the method. It gives the average rank of the interacting SNPs, namely SNP1 and SNP2 over 100 datasets under different genetic models. The lower the average rank value, the higher the informativeness the SNP.

Figure 1

SNP1	AA	AA	AA	Aa	Aa	Aa	aa	aa	aa
SNP2	BB	Bb	bb	BB	Bb	bb	BB	Bb	bb
Case	n_{11}	n_{21}	n_{31}	n_{41}	n_{51}	n_{61}	n_{71}	n_{81}	n_{91}
Control	n_{12}	n_{22}	n_{32}	n_{42}	n_{52}	n_{62}	n_{72}	n_{82}	n_{92}
Total	n_1	n_2	n_3	n_4	n_5	n_6	n_7	n_8	n_9

Figure 2

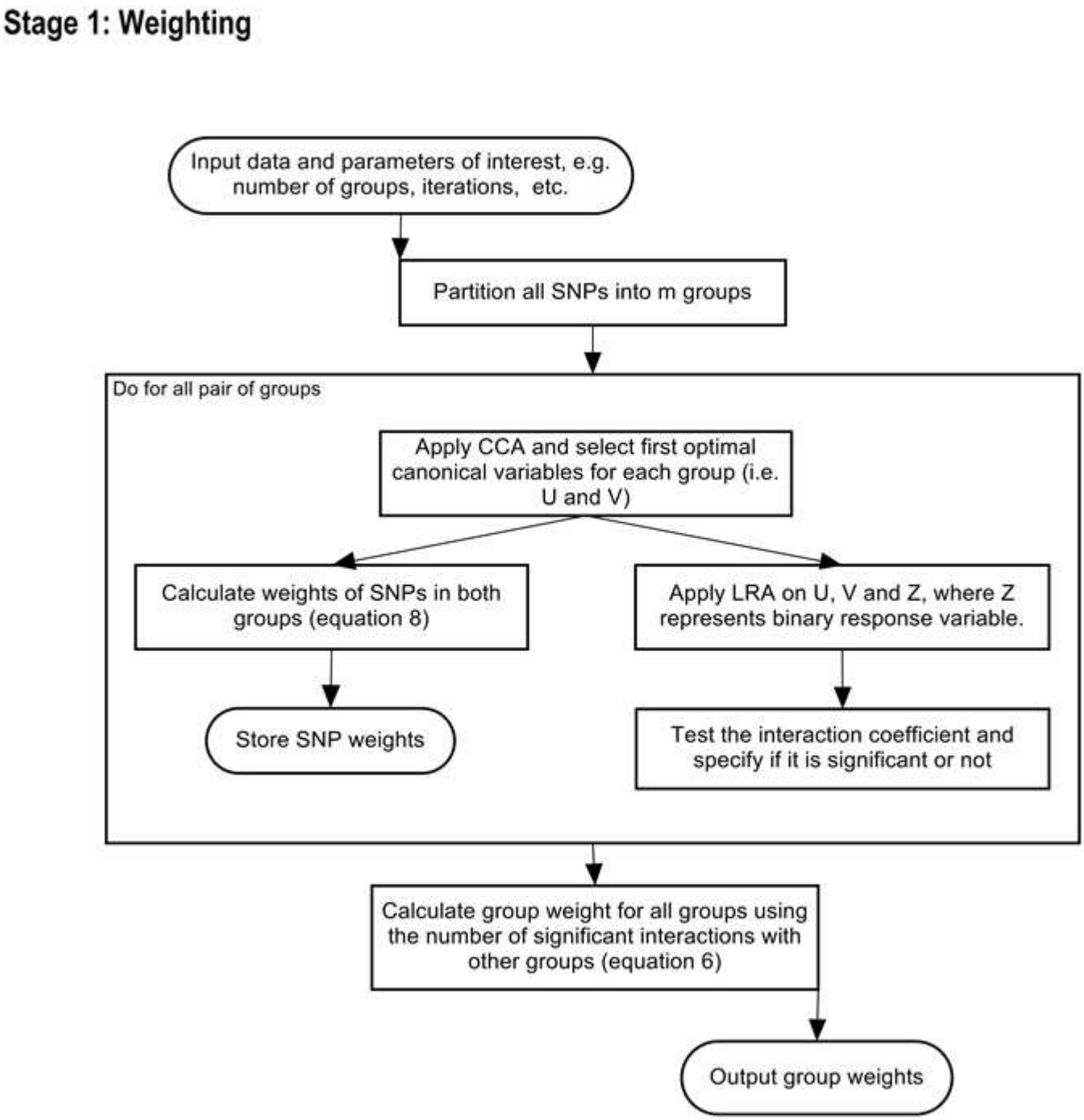


Figure 3

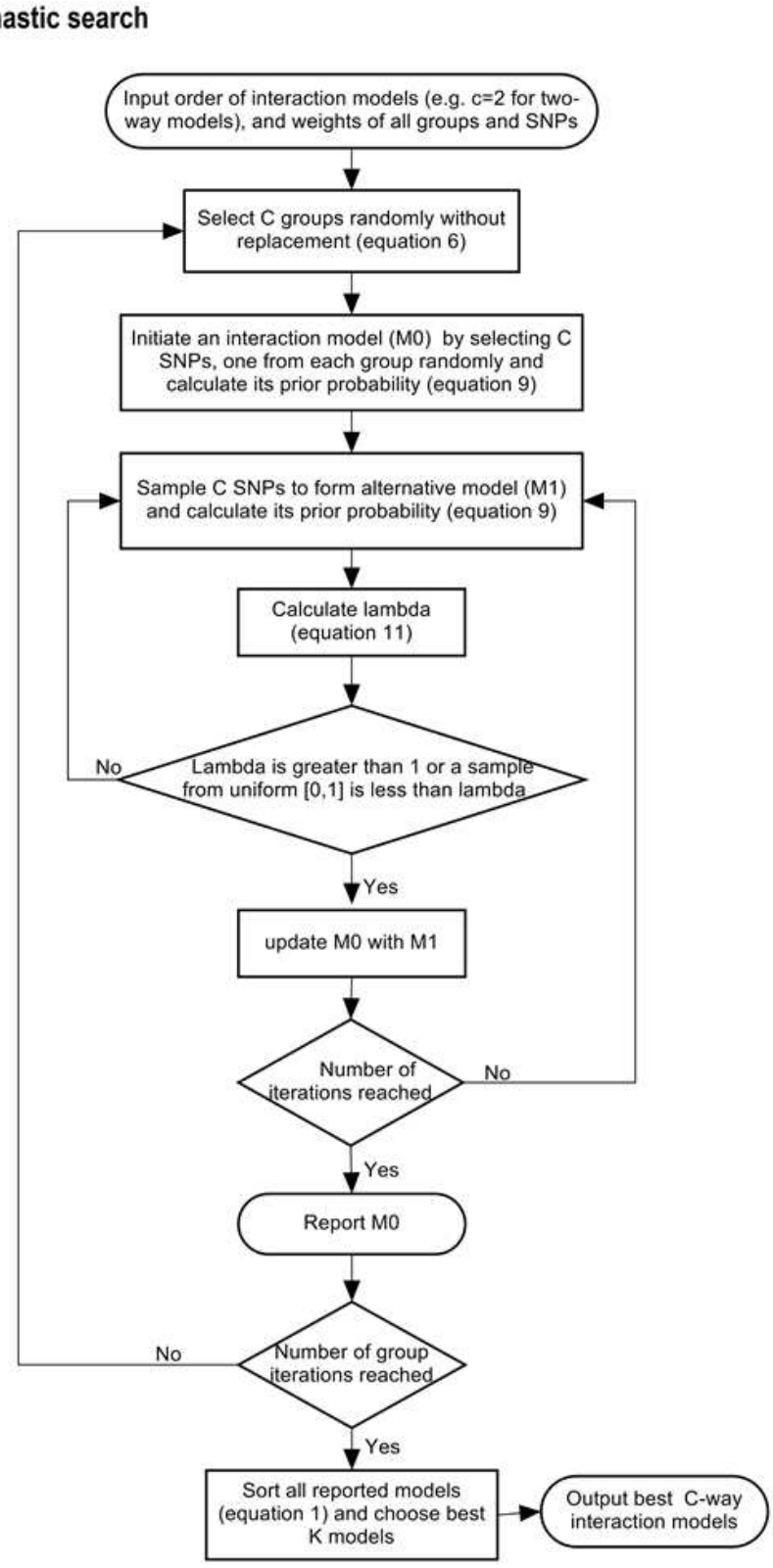
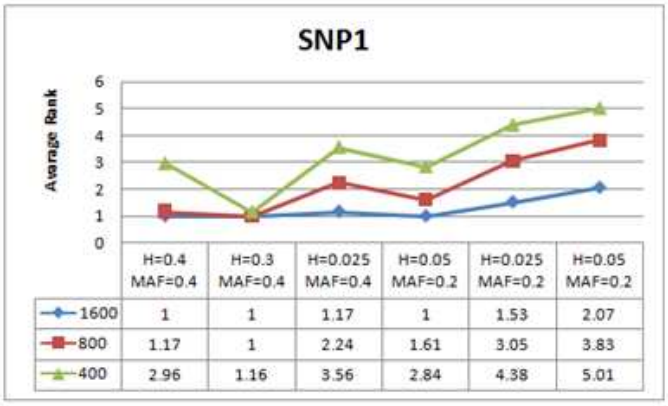
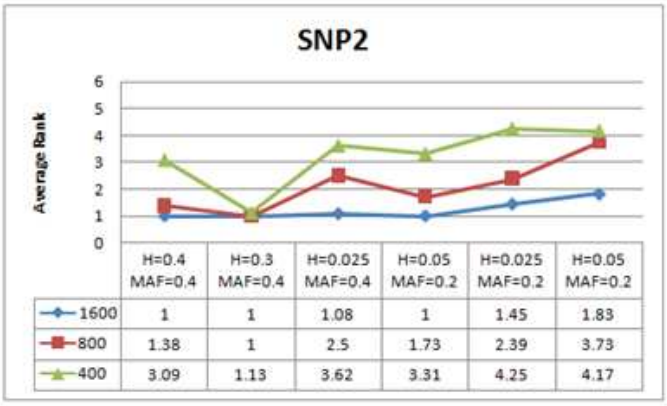


Figure 4



	H=0.4 MAF=0.4	H=0.3 MAF=0.4	H=0.025 MAF=0.4	H=0.05 MAF=0.2	H=0.025 MAF=0.2	H=0.05 MAF=0.2
1600	1	1	1.17	1	1.53	2.07
800	1.17	1	2.24	1.61	3.05	3.83
400	2.96	1.16	3.56	2.84	4.38	5.01



	H=0.4 MAF=0.4	H=0.3 MAF=0.4	H=0.025 MAF=0.4	H=0.05 MAF=0.2	H=0.025 MAF=0.2	H=0.05 MAF=0.2
1600	1	1	1.08	1	1.45	1.83
800	1.38	1	2.5	1.73	2.39	3.73
400	3.09	1.13	3.62	3.31	4.25	4.17