PLISKA
STUDIA MATHEMATICA
BULGARICA
ПЛИСКА
БЪЛГАРСКИ
МАТЕМАТИЧЕСКИ
СТУДИИ

Pliska Stud. Math. Bulgar. **18** (2007), 195–212

PLISKA
STUDIA MATHEMATICA
BULGARICA

ENTROPY BASED APPROACH TO FINDING INTERACTING GENES RESPONSIBLE FOR COMPLEX HUMAN DISEASE

Valentin Milanov   Radoslav Nickolov

A challenging problem in human genetics is the identification and characterization of susceptibility genes for complex human diseases such as cardiovascular disease, cancer, hypertension and obesity. These conditions are likely due to the effects of high-order interactions among multiple genes and environmental factors. Genome-wide association studies, where hundreds of thousands of single-nucleotide polymorphisms (SNPs) are genotyped in samples of cases and controls, offer a powerful approach for mapping of complex disease genes. The classical statistical methods, parametric and non-parametric, are usually limited to small number of SNPs. Here we propose a new method based on a classical search algorithm - "sequential forward floating search", utilizing entropy based criterion function. Using simulated case-control data we demonstrate that the method has a high discovery rate under different models of gene-gene interaction, including pure interaction without main effects of the genes. The performance of the proposed method is also compared to a method recently advocated in the literature: multifactor dimensionality reduction (MDR).

1. Introduction

There are about $30,000 - 40,000$ genes in the human genome. With the recent report of the sequence that constitutes our genome and improving biological

2000 *Mathematics Subject Classification:* 62P10, 92D10, 92D30, 94A17, 62L10
Key words: entropy, SNP, genotype, genomewide, association, adaptive search

technology, we are now in a position to begin to detect the genes that predispose humans to complex diseases: cancer, diabetes, hypertension, obesity, etc. The search for genetic components of complex human diseases is based on studying the observed correlations between genetic markers and disease. One approach is the candidate-gene association study approach. Known genetic markers in the candidate gene are genotyped, and their association to the disease is tested with statistical association analysis methods. However there is a growing awareness that complex interactions among multiple genes and multiple environmental factors play an important role in determining an individual's risk of various common diseases. The idea is not new, [1] emphasized that the relationship between genes and biological end points is dependent on dynamic interactive networks of genes and environmental factors.

Genetic studies aim to determine which genetic polymorphisms play a functional role in the etiology of a disease. However, the general strategy for identifying Mendelian disease genes has largely been unsuccessful when applied to identifying susceptibility genes for common complex multifactorial diseases [2]. This is primarily due to the fact that the Mendelian approach requires each susceptibility factor to have a large independent main effect on disease risk because in general only one genetic locus is investigated at a time. Complex diseases are common, with unknown modes of inheritance and arise as a result of multiple mechanisms: common alleles with small to moderate effects, rare alleles with moderate to large effects, complex gene-gene and gene-environmental interactions [3, 4], etc. The gene-gene interaction is known as epistasis.

Epistasis-interaction between different genes

How can an interaction or epistasis be defined? There are two main interpretation of the notion of epistasis. *Biological interaction* usually corresponds to a situation in which the qualitative nature of the mechanism of action of a factor is affected by the presence or absence of an other factor [5]. Epistasis is the control of a phenotype by two or more genes. This genetic interaction often involves the masking of the phenotypic effects of one gene by the effects of a second gene. A gene is epistatic when its presence suppresses the effect of the other gene.

The above definition of epistasis can not be directly applied to binary traits. In human genetics, the phenotype of interest is often qualitative and usually dichotomous, indicating presence or absence of disease. Mathematical models for the joint action of two or more loci focus on the penetrance, or the probability of developing disease given genotype. *Statistical interaction* between loci requires a dependent effect, where the risk associated with a genotype at a locus is dependent on a genotype at another locus. As example consider two risk loci (A

and B). Locus A with two possible alleles, A or a, and locus B with possible alleles, B or b. Suppose that a predisposing allele is required at both loci in order to exhibit the trait, i.e. one or more copies of both allele A and allele B are required. Then, when the effects of both loci are considered, one can obtain the penetrance table shown in Table 1.

	BB	Bb	bb
AA	1	1	0
Aa	1	1	0
aa	0	0	0

Table 1: Penetrance of multilocus genotype: probability of disease given the genotypes at the considered marker loci

In the above table, the effect of allele A can only be observed when allele B is also present: without the presence of B, the effect of A is not observable. The effect at locus A would appear to be 'masked' by that at locus B. By analogy with the example in Table 1, we might say that locus B is epistatic to locus A, since when the genotype b/b is present at locus B, the effect of the alleles at locus A is not observable. However, one can equally say that locus A is epistatic to locus B, since when the genotype a/a is present at locus A, the effect of the alleles at locus B is not observable. Although not many epistatic interactions have been described to date [6, 7, 8, 9], there is little doubt that many such interactions exist.

The coinheritance of alleles on haplotypes of tightly linked loci leads to associations between these alleles in the population, known as *linkage disequilibrium* (LD). Linkage disequilibrium can be detected in population samples. A pair of loci is said to be in linkage disequilibrium when, in a sample of individuals, their joint haplotype frequencies deviate from those expected under independence. For example, consider two closely spaced loci 1 and 2 on a chromosome with alleles A, a, and B, b. Suppose p_A, p_a and p_B, p_b are the frequencies of these alleles in the population. Let the frequency of the AB haplotype is p_{AB}. Two loci are in independent or in *linkage equilibrium* if $p(AB) = p\,(A)\,p(B)$.

Many indirect association studies employ a dense map of polymorphisms to identify associated trait loci, and some even allow scanning all the genes throughout the genome, [10]. These usually utilize biallelic single nucleotide polymorphisms (SNPs) as markers. SNPs are variations in DNA sequence where one of the four nucleotides is substituted for another (for example, C for A). SNPs are the most frequent type of polymorphism in the genome, and they make up the majority of markers in a whole-genome association map.

Multi-locus methods are specifically designed to find multiple disease loci that may influence the disease by intricate genetic patterns, gene-gene interaction and gene-environment interactions. One of these methods is the Multifactor Dimensionality Reduction (MDR) [11]. Multifactor-Dimensionality Reduction (MDR) is a special case of patterning and recursive partitioning (PRP)

PRP is extension of the classification and regression tree (CART) method. In PRP method individuals are assigned to genotype groups based on their multilocus genotypes and then using the resulting classification as a predictor variable in a *recursive partitioning* [12]. PRP starts by creating a categorical variable for genotype group (which is referred to as pattern) such that individuals with identical multi-locus genotypes are assigned to one and the same group. The groups are formed using subsets of SNPs. For example, considering two SNPs, A and B, there are nine levels of the pattern variable corresponding to each cell (genotype) in Table 2.

SNP 1\ SNP 2	AA	Aa	aa
BB	$AABB$	$AaBB$	$aaBB$
Bb	$AABb$	$AaBb$	$aaBb$
bb	$AAbb$	$Aabb$	$aabb$

Table 2: Each level of the pattern variable corresponds to two-locus genotype - 9 possible

In general, if the total number of loci is N, then there are N choose n ways of choosing n SNPs at a time. Therefore N choose n genotype groups (pattern variables) are created correspondingly. Each one of these pattern variables can be included as potential predictors in RP. MDR is a special case of PRP in which (1) tree growth is restricted to a single split, and (2) misclassification error (i.e. the proportion of people incorrectly classified) is used as the measure of impurity. This result follows directly from the fact that MDR yields a partition of the data into two groups in a manner that minimizes the misclassification rate. The six steps in MDR are described with Figure 1 in the appendix.

The MDR algorithm has reasonable power to detect epistasis, [13]. However, the best multilocus predictor is discovered using an exhaustive search, which makes it not applicable to large number of predictors.

To address this and other limitations, a flexible computational framework for detecting and interpreting gene-gene interactions has recently been proposed [14]. In the first step entropy-based measures of information gain are used to select interesting predictors from the pool of possible candidates, a set of thousands of single-nucleotide polymorphisms (SNPs).

Recently, other methods incorporating entropy measures have been proposed, [15].

In this article, first we define entropy based measure of association between candidate set of polymorphisms and the disease status. Second we propose novel search procedure using the defined measure of association in float search algorithm for identifying susceptibility genes for complex human diseases among thousands of candidate genes, in particular gene-gene interaction using balance case-control samples. Third we compare the proposed method to MDR under three main scenarios.

2. Methods

Entropy as measure of association of group of m polymorphisms with disease status

The statistical entropy of X is defined as

$$H(X) = -\sum_{i=1}^{N} p_i \log p_i$$

where X is random vector with probability distribution $P(X = x_i) = p_i$, for $i = 1...N$, N is the number of possible values of X.

We use the concept of entropy to define measure of association of $m \geq 2$ independent candidate marker loci and the disease status. We consider SNP markers.

Denote $g_1, \ldots, g_G$, the set of all multilocus genotypes formed using m SNP loci, where the allele at the the *ith* locus is either allele 1 or allele 2. Since at each locus three genotypes can occur, (1/1, 1/2, and 2/2) the total number of genotypes is $G = 3^m$.

Considering the multilocus genotypes as states of random vector X we define the entropy of the controls and cases correspondingly as $E_{n,G}$ and $E_{d,G}$ as

$$E_{n,G}(X) = -\sum_{i=1}^{G} p_{g_i,n} \log p_{g_i,n} \quad , \quad E_{d,G}(X) = -\sum_{i=1}^{G} p_{g_i,d} \log p_{g_i,d} \tag{1}$$

where $p_{g_i,n}, p_{g_i,d}$ are the multilocus genotype frequencies in normal and diseased individuals respectively.

If none of the candidate SNPs are associated with the disease the distribution of multilocus genotypes in normal and diseased group would not exhibit significant difference in the frequency of the genotype g_i, or $p_{g_i,n} \approx p_{g_i,d}$.This will result in similar entropy signals, $E_{normal,G}(X) \approx E_{disease,G}(X)$ in both groups. In contrary if we assume that the SNPs are associated with the disease status then we expect to observe a pattern for the multilocus genotypes that are particular for the diseased individuals-high risk genotypes. The order in the system in genotype states in the group of diseased individuals will result in relatively small estimate of $E_{disease,G}(X)$ compared to $E_{normal,G}(X)$. In the extreme case when only one multilocus genotype is high-risk, then the frequency of this genotype in the diseased group of individuals is going to be high, close to 1, and the other genotypes will have really small frequencies.

We propose to use the difference between the entropy of genotype states of normal individuals and entropy of genotype states of diseased individuals as a measure of association between the set of m SNPs and disease status.

$$\begin{aligned} (2) \quad ED(X,\ G,\ m) &= E_{normal,G}(X) - E_{disease,G}(X) = \\ &= -\sum_{i=1}^{G} p_{g_i,normal} \log p_{g_i,normal} + \sum_{i=1}^{G} p_{g_i,disease} \log p_{g_i,disease} \end{aligned}$$

In general bigger margin of the two signals will indicate stronger association between the group of m selected polymorphisms and the disease status.

Standardized Entropy Criterion

In practise association studies aim to select set of polymorphisms among two or more candidate sets that describes the best way the association between the trait of interest and the genetic variation.

We defined criterion to compare two different candidate sets of polymorphisms (SNPs) using the defined entropy measure as in 2, in samples of $2n$ unrelated individuals: n cases and n controls**.** We assume that underlying population is homogeneous. The sets may have some polymorphisms in common. To answer the question which one has stronger association with the disease status we need a uniform *measure* to compare them. For these reasons we propose Standardized Entropy Difference (SED) measure in the form

$$(3) \qquad SED(X,\ G,\ m) = \frac{E_{normal,G}(X) - E_{disease,G}(X)}{E_{normal,G}(X)}$$

If $n_{g_i,n}, n_{g_i,d}$ are the numbers of individuals having genotype g_i in the given sample, for normal and diseased individuals respectively, for $1 \leq i \leq G$. Then using the maximum likelihood estimates of the genotype frequency, $p_{g_i,n}, p_{g_i,d}$ one can use as criterion the estimated SED measure to decide which of the candidate sets of polymorphisms has stronger association with the disease.

Search procedure utilizing the entropy as a measure of association, Modified Adaptive Float Search (MAFS)

One of the challenges that search algorithms need to deal identifying functional polymorphisms in complex disease is the unknown number of genes interacting. We describe a method based on modification of suboptimal search algorithm, Adaptive Sequential Forward Floating Search (ASFFS) as in [16]. The ASFFS is feature selection method for finding a subset of d features from a given set of D measurements, $d < D$. In contrast with the classical floating search methods which use only single feature adding or removing, respectively, in the number of features o added or removed at the time is determined adaptively according to r-actual generalization limit which is always smaller than user prespecified absolute generalization limit $r_{\max}$. As result the nearer, the current subset size, k, is to the final one, d,the higher is the generalization limit, this can be determined by setting up parameter b and checking if $|d - k| < b$. Terminating condition is $k \geq d + \Delta$, where Δ is determined heuristically. For details of ASFFS see Figure 2 in the appendix.

The search procedure ASFFS can not be directly utilized for identifying disease associated polymorphisms due to the fact that number of features d is predetermined, in practise the number of associated polymorphisms is unknown. We propose use of criterion function that allows to compare not only sets with the same cardinality but also sets with different number of features. We achieve this by using the uniform (not depending on the number of features in the set) criterion SED. The MAFS search procedure consists of steps that one encounter in the ASFFS with few modifications. First, when a candidate of set of k features is selected it is not evaluated versus the best set of k features so far but versus the overall best set of size K so far, in general $K \neq k$. In this way the search procedure goes through less steps since local improvements of the best set of size k are not replacing the current set. The nested effect is not an issue since the algorithm floats according to ASFFS. The algorithm uses the same parameters as the ones used in the ASFFS and one additional terminating parameter Λ, number of features that has been added to the best set of features without successfully improving the criterion function. We have $\Lambda < r_{\max}$, where $r_{\max}$ was the maximum generalization limit. The method can allow higher generalization level than the general ASFFS, since it has reduced run time.

3. Data Simulation

To evaluate the power of MAFS for detecting gene-gene interactions, we simulated case-control data using several main scenarios. Under each scenario models involving independent loci without single main effects were considered. If main effects were present, it could be difficult to evaluate whether particular loci were detected because of the main effects, or because of the interactions, or both that is why interactions without main effects only were considered in this study. Under these settings the degree of complexity is higher and it is good way to test the ability of the method to identify gene-gene interactions. All genotypes were generated according to Hardy-Weinberg equilibrium. Statistically HWE means that the alleles for the next generation for any given individual are chosen independently. Consider the simplest case of a single locus with two alleles A and a with allele frequencies of p and q, respectively. HWE predicts that the genotypic frequencies for the AA homozygote to be p^2, the Aa heterozygote to be $2pq$, and the other aa homozygote to be q^2.

Scenario 1. Four different two-locus epistasis models in which the functional loci are single-nucleotide polymorphisms (SNPs) as in [11]. The four models were generated using the epistasis model discovery method in [17], using allele frequencies of $p = 0.25$ and $q = 0.75$ for models 1 and 2, Tables 3 and 4 and allele frequencies of $p = 0.1$ and $q = 0.9$ for models 3 and 4, for details see Tables 5 and 6. For the four models the combination of genotypes formed using the 2 SNPs exhibiting interaction in the absence of independent main effects.

	BB	Bb	bb
AA	0.08	0.07	0.05
Aa	0.10	0	0.10
aa	0.03	0.10	0.04

Table 3: Model 1, Penetrance values, p=0.25, q=0.75 minor and major allele frequencies.

For example consider model 1 with penetrance functions given in Table 3, to show that there is no single main effect of each locus we need to show that

	BB	Bb	bb
AA	0	0.01	0.09
Aa	0.04	0.01	0.08
aa	0.07	0.09	0.03

Table 4: Model 2, Penetrance values, p=0.25, q=0.75 minor and major allele frequencies.

	BB	Bb	bb
AA	0.07	0.05	0.02
Aa	0.05	0.09	0.01
aa	0.02	0.01	0.03

Table 5: Model 3, Penetrance values, p=0.1, q=0.9 minor and major allele frequencies.

	BB	Bb	bb
AA	0.09	0.001	0.02
Aa	0.08	0.07	0.005
aa	0.003	0.007	0.02

Table 6: Model 4, Penetrance values, p=0.1, q=0.9 minor and major allele frequencies.

$$f_{AA} = f_{Aa} = f_{aa} \quad \text{or} \quad f_{AA} \approx f_{Aa} \approx f_{aa} \tag{4}$$
$$\text{and}$$
$$f_{BB} = f_{Bb} = f_{bb} \quad \text{or} \quad f_{BB} \approx f_{Bb} \approx f_{bb}$$

To obtain $f_{AA} = P(D/AA)$, the probability of being diseased given the genotype AA at the locus with possible alleles A, a. Using the conditional and total probability formulas and the fact that the loci are independent, probability of

$$f_{AA} = f_{AA,BB}\ p_B^2 + 2f_{AA,Bb}\ p_B\ p_b + f_{AA,bb}\ p_b^2 = 0.0593$$

Similarly can be obtained $f_{Aa} = 0.0624$and $f_{aa} = 0.0618$. The close marginal penetrance functions indicate that there is no effect of the locus A alone. In a similar way one can check that the locus B has no main effect alone.

For each of the four epistasis model we simulated 100 datasets. Each dataset consisted of 200 cases and 200 controls, for each individual genotype at 10 independent SNP loci were simulated. The first two SNPs $1, 2$ were the functional ones. Each SNP had two alleles with the common allele having a frequency of 0.75,or 0.9, as described above. To evaluate the method performance under different sample sizes we generate 100 data sets with 100 cases and 100 controls and 100 data sets with 50 cases and 50 controls.

Scenario 2. One three-locus epistasis model generated using the epistasis model discovery method in [17]. Genotypes at 10 independent SNP loci were simulated, the first three SNPs $1, 2, 3$ functional. Each SNP had two alleles with the common allele having a frequency of 0.5.The penetrance functions are given in Table 7. Similar to derivation of marginal penetrance function for the two-locus model one can obtain the marginal penetrance for each locus, e.g. f_{AA}, f_{Aa}, f_{aa} for locus A. The independent main effect of the AA genotype is $f_{AA} = P(D/AA) = 0.475$. However, the probability of disease given the genotype combination $AAbbCC$ is 0.1 while the probability of disease given genotype $AABBcc$ is 1.0. Hence it is clear that the effect of the genotype AA is dependent on the genotypes at the other two loci. The independent main effects for all three loci are given below.

$$f_{AA} = 0.475, \quad f_{Aa} = 0.487, \quad f_{aa} = 0.475$$

$$f_{BB} = 0.475, \quad f_{Bb} = 0.481, \quad f_{bb} = 0.487$$

$$f_{CC} = 0.475, \quad f_{Cc} = 0.487, \quad f_{cc} = 0.475$$

To evaluate the methods performance under different sample sizes we generate 100 data sets with 200 cases and 200 controls, 100 data sets with 100 cases and 100 controls and 100 data sets with 50 cases and 50 controls.

	CC			Cc			cc		
	AA	Aa	aa	AA	Aa	aa	AA	Aa	aa
BB	0.4	0.9	0.7	0.2	0.2	0.6	1.0	0.4	0.5
Bb	0.9	0.0	0.9	0.9	0.9	0.0	0.3	0.1	0.6
bb	0.1	0.2	0.6	0.6	0.6	0.3	0.3	0.9	1.0

Table 7: Penetrance values for 3-locus genotypes formed using 3 SNPs exhibiting interaction in the absence of independent main effects, p=0.5 q=0.5 minor and major allele frequencies.

Scenario 3. One four-locus epistasis model generated according to [17]. Again 10 independent SNPs were simulated to form each individual multilocus genotype., the first four SNPs $1, 2, 3, 4$ were simulated to be functional. Each SNP had two alleles with the common allele with frequency of 0.5.The penetrance functions are given in Table 8.

100 data sets with 200 cases and 200 controls, 100 data sets with 100 cases and 100 controls and 100 data sets with 50 cases and 50 controls were generated.

4. Results

We ran MAFS with different parameter setting to perform search for each one of the scenarios described in the previous section. Parameters are set to values that allow search for high order interaction. Most important factor in determination of generalization level $r_{\max}$ for each scenario has been the available number of sampled individuals. Due to the small number of candidate SNPs in scenario $1 - 3$, 10, we set the terminating parameter to $\Lambda = 8$, which practically leads to search for the best set among set consisting of $2 - 9$ SNPs. The entire set

		CC				Cc				cc		
		AA	Aa	aa		AA	Aa	aa		AA	Aa	aa
DD												
	BB	0.9	1.0	0.0		0.9	0.1	0.1		1.0	0.9	1.0
	Bb	0.9	0.0	0.8		1.0	0.0	1.0		0.1	0.0	0.0
	bb	0.1	1.0	0.9		1.0	0.1	0.8		0.9	0.9	1.0
Dd												
	BB	0.0	0.0	0.8		0.1	1.0	0.1		0.0	1.0	0.1
	Bb	1.0	0.1	1.0		0.0	0.8	0.7		0.0	1.0	0.0
	bb	0.1	1.0	0.0		1.0	0.3	0.1		0.0	0.0	0.2
dd												
	BB	0.9	0.1	0.0		1.0	0.0	0.0		0.9	1.0	0.7
	Bb	0.9	0.0	0.0		0.0	0.5	1.0		0.0	1.0	1.0
	bb	1.0	0.9	1.0		1.0	0.2	0.3		0.9	0.0	0.0

Table 8: Penetrance values for 4-locus genotypes formed using 4 SNPs exhibiting interaction in the absence of independent main effects, p=0.5 q=0.5 minor and major allele frequencies.

consisting of 10 SNPs is of no interest. Parameter settings for the first three scenarios are given in Table 9.

We ran MAFS with these parameters for all scenarios. The best model using MAFS has been selected among the best models of size $2, 3, \ldots, 8$. The model with maximum SED value among is chosen as the best model overall.

The power of MAFS under each model is estimated with the discovery rate, the number of times when MAFS identified best model containing the functional SNPs out of each set of 100 datasets.

Under scenarios $1-3$ datasets were analyzed also with MDR method. Using MDR we conducted an exhaustive search of all possible $2-5$-locus interactions for the two-locus models (scenario 1), $2-7$ locus interaction for the three-locus model (scenario 2), and $2-8$ for the four-locus model (scenario 3). We chose to evaluate interactions up to that order since with a 9-locus model for example every individual will have a unique genotype and thus the prediction error estimates will not be that accurate. We applied the MDR algorithm using a cases-to-controls threshold ratio of $1:1$ as described in [11]. Each dataset was analyzed using 10-fold cross-validation. The power of MDR under each model was estimated with the discovery rate, the number of times MDR identified the functional SNPs out of each set of 100 datasets.

Sample Size	Parameter	Settings
	d	9
100(200)	$r_{\max}$	3
	b	8
	Λ	8
	d	9
400	$r_{\max}$	4
	b	8
	Λ	8

Table 9: Parameter settings for scenario 1-3 using MAFS. d - number of features in the targeted set, r-max- maximal generalization level, b neighborhood parameter, lambda - number of features added to the best set of features with no improvement of the criterion function.

Power Comparison

We compared the power of the proposed search procedure, MAFS, with the MDR. The results are summarized in Tables 10 and 11.

MDR outperforms MAFS for models for 3 and 4 in the case of sample size 400 individuals, and for model 3 when the sample size is 200.

However the proposed search procedure, overall, is more powerful than MDR especially for relatively small samples.

MAFS considerably outperformed MDR under scenario 2 and 3.

Model		Method					
		MAFS	MDR	MAFS	MDR	MAFS	MDR
	Sample size	100		200		400	
1		0.29	0.18	0.58	0.26	0.73	0.33
2		0.35	0.47	0.53	0.69	0.91	0.73
3		0.18	0.14	0.29	0.36	0.35	0.51
4		0.21	0.20	0.30	0.27	0.37	0.47

Table 10: Power of MAFS and MDR under scenario 1.

Scenario		Method					
		MAFS	MDR	MAFS	MDR	MAFS	MDR
	Sample size	100		200		400	
2		0.29	0.00	0.59	0.23	0.91	0.63
3		0.13	0.00	0.68	0.02	0.94	0.56

Table 11: Power of MAFS and MDR under scenario 2 and 3.

5. Discussion

Common variants with small individual effects might contribute more substantially to disease risk through nonadditive interactions among loci. In situation like this one should be aware that examining only a single locus at a time, these effects might be missed. Because only very common variants will be found in combination at a measurable frequency, the study of gene-gene interactions in common disease is implicitly most relevant to the second approach of studying common variants.

Our proposed search procedure target the detection of population-level association between interacting genes predisposing to a complex diseases from a case-control sample. The procedure utilizes uniform criterion function that is designed to exploit for association, simultaneously, set of independent biallelic markers (SNPs), typed on the entire genome. The uniformity of the criterion function allows to compare between candidate set of polymorphisms of different cardinality. The goal is to optimally use the information of combination of polymorphisms.

We evaluated the discovery rate the proposed search procedure through simulations. In our simulation scenarios we used models of interacting susceptibility loci without significant main effects. The MAFS discovery rate was compared the one of MDR.

We found that the MAFS with the proposed criterion, especially in the case of small sample size, is more powerful than MDR. The power decreases with the number of interacting loci, which is result of poor estimates of the multilocus genotype frequencies.

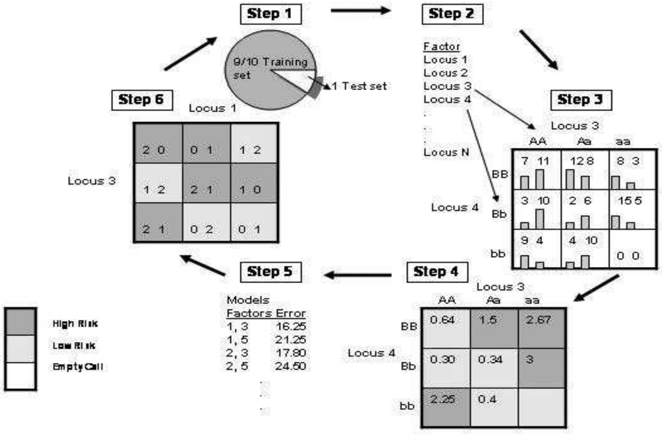
Method can be used as a screening tool for genomewide screening by extending it to finding few best set of polymorphisms.

Similar schemes as the one in [14] can utilize the search procedure as first step of the analysis of large number of candidate polymorphisms. In many cases under scenario 2, the three-locus epistasis model and four locus interaction models the right model has not been selected (second best) but model of two loci, the method may pick few best models for example the first 3 and then these be analyzed with other methods.

However, in our present study, in order to avoid the spurious associations due to stratification or admixture we assumed that the samples of cases and controls either come from a homogeneous population or are properly matched for ethnicity. MAFS does not assume particular genetic model, that is, no mode of inheritance needs to be specified. This is important for diseases in which the mode of inheritance is unknown and likely very complex.

There are certain disadvantages and limitations of MAFS. First MAFS models can be difficult to interpret. In contrast with MDR, which classifies the multilocus genotypes as high/low risk, MAFS does not provide clear interpretation of the diseased genotypes or how the selected markers explain the pattern of disease status. Second, in its current form, MAFS can be applied only to case-control studies that are balanced (i.e., that have the same number of cases and of controls).

The linux implementation of MAFS is available upon request to the authors.



Locus 4 \ Locus 3	AA	Aa	aa
BB	7 11	12 8	8 3
Bb	3 10	2 6	15 5
bb	9 4	4 10	0 0

Locus 4 \ Locus 3	AA	Aa	aa
BB	0.64	1.5	2.67
Bb	0.30	0.34	3
bb	2.25	0.4	

Models

Factors	Error
1, 3	16.25
1, 5	21.25
2, 3	17.80
2, 5	24.50
.	
.	
.	

Locus 3 \ Locus 1			
	2 0	0 1	1 2
	1 2	2 1	1 0
	2 1	0 2	0 1

Figure 1: Summary of 6 steps in the cross validation procedure involved in implementation of the MDR method: a set of n genetic factors is selected among N factors; then factors and their possible multifactor classes or cells are represented in n-dimensional space; each multifactor cell in n-dimensional space is labeled as either “high-risk”or “low-risk”, and the prediction error of each model is estimated. For each multifactor combination, hypothetical distributions of cases (left bars in boxes) and of controls (right bars in boxes) are shown.

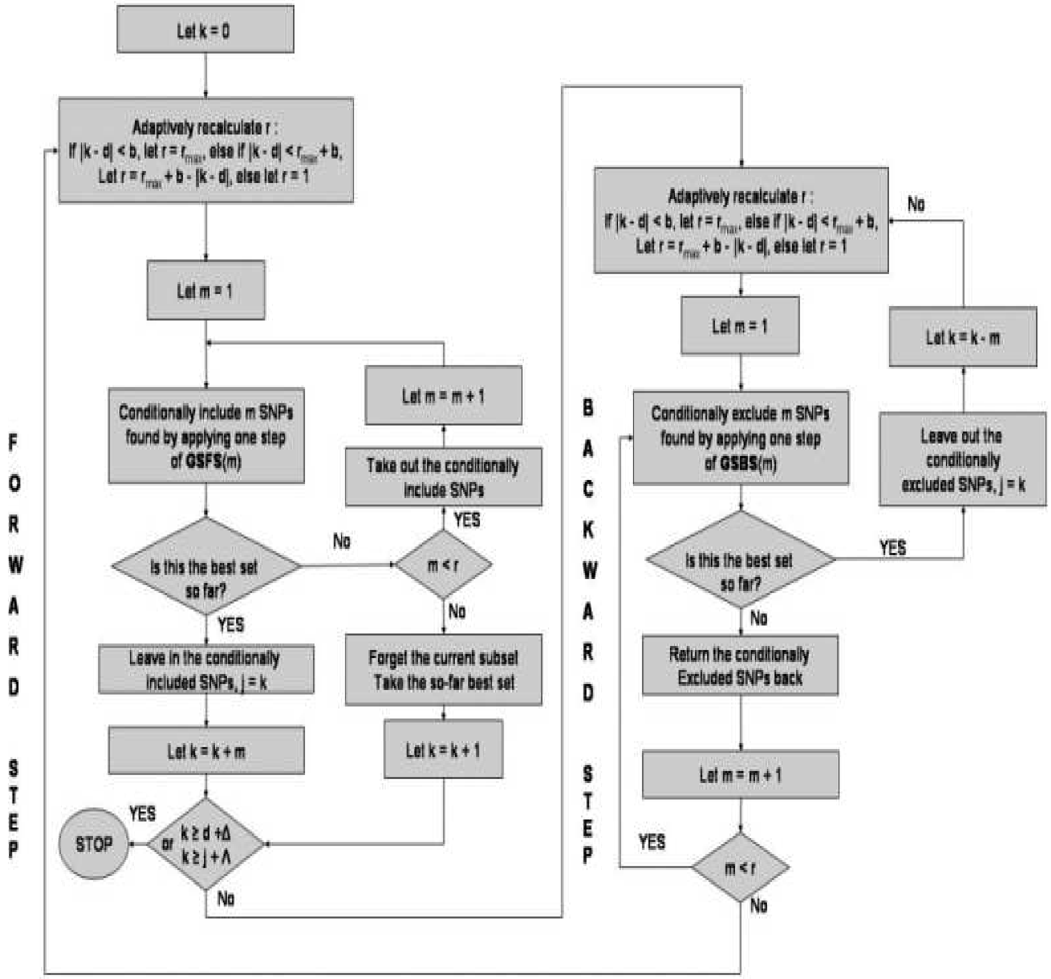

Let k = 0
Adaptively recalculate r :
If |k - d| < b, let r = r_max, else if |k - d| < r_max + b,
Let r = r_max + b - |k - d|, else let r = 1
Let m = 1
Conditionally include m SNPs found by applying one step of GSFS(m)
Let m = m + 1
Take out the conditionally include SNPs
YES
Is this the best set so far?
No
m < r
YES
No
Leave in the conditionally included SNPs, j = k
Forget the current subset Take the so-far best set
Let k = k + m
Let k = k + 1
YES
STOP
k ≥ d +Δ
or
k ≥ j + Λ
No
F O R W A R D S T E P
B A C K W A R D S T E P
Adaptively recalculate r :
If |k - d| < b, let r = r_max, else if |k - d| < r_max + b,
Let r = r_max + b - |k - d|, else let r = 1
No
Let m = 1
Let k = k - m
Conditionally exclude m SNPs found by applying one step of GSBS(m)
Leave out the conditionally excluded SNPs, j = k
Is this the best set so far?
YES
No
Return the conditionally Excluded SNPs back
Let m = m + 1
YES
m < r
No

Figure 2: Flow chart of ASFFS algorithm.

REFERENCES

[1] Wright S. The roles of mutation, inbreeding, cross breeding, and selection in evolution. *Proceedings of the 6th International Congress of Genetics* **1** (1932), 356–366.

[2] Altmuller J, Palmer LJ, Fischer G, Scherb H, Wjst M. Genomewide scans of complex human diseases: true linkage is hard to find. *Am J Hum Genet* **69** (2001), 936–950.

[3] Templeton AR Epistasis and complex traits. In: Wade M, Brodie III B, Wolf J, eds. *Epistasis and the Evolutionary Process.* NewYork; Oxford University Press, 2000, 41–57.

[4] Sing FC, Davignon J. Role of the apolipoprotein E in determining normal plasma lipid and lipoprotein variation. *Am. J. Hum. Jenet.* **37** (1985), 268-285.

[5] Siemiatycki J, Thomas DC Biological models and statistical interactions: an example from multistage carcinogenesis. *Int. J. Epidemiol.* **10** (1981), 383–387.

[6] Bolk S. et al. A human model for multigenic inheritance: phenotypic expression in Hirschsprungdisease requires both the RETgene and a new 9q31 locus. *Proc. Natl Acad. Sci. USA* **97** (2000), 268–273.

[7] Zetterberg H, Zafiropoulos A, Spandidos DA, Rymo L. & Blennow K. Gene–gene interaction between fetal MTHFR 677C>T and transcobalamin 776C>G polymorphisms in human spontaneous abortion. *Hum. Reprod.* **18** (2003), 1948–1950.

[8] Butt C. et al. Combined carrier status of prothrombin 20210A and factor XIII-A Leu34 alleles as astrong risk factor for myocardial infarction: evidence of a gene–gene interaction. *Blood* **101** (2003), 3037–3041.

[9] Tiret L. et al. Synergistic effects of angiotensin-converting enzyme and angiotensin-II type 1 receptorgene polymorphisms on risk of myocardial infarction. *Lancet* **344** (1994), 910–913.

[10] Hoh J, Ott J Scan statistics to scan markers for susceptibility genes. *PNAS* **97** (2000), 9615–9617.

[11] 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* **69** (2001), 138–147.

[12] Foulkes AS, DeGruttola V, Hertogs K Combining genotype groups and recursive partitioning: An application to HIV-1 genetics data. *JRSS C* **53** (2004), 311–323.

[13] Marylyn D. Ritchie, Lance W. Hahn, and Jason H. Moore Power of Multifactor Dimensionality Reduction for Detecting Gene-Gene Interactions in the Presence of genotyping Error, Missing Data, Phenocopy, and Genetic Heterogeneity. *Genetic Epidemiology* **24** (2003), 150–157.

[14] Jason H. Moore, Joshua C. Gilbert, Chia-Ti Tsai, Fu-Tien Chiang, Todd Holden, Nate Barney, Bill C. White A flexible computational frame work for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility *Journal of Theoretical Biology* **241** (2006), 252–261.

[15] Jinying Zhao, Eric Boerwinkle, Momiao Xiong An Entropy_based Statistics for Genomewide Association Studies. *Volume I. Berlin/Heidelberg/New York: Springer-Verlag*, 2005.

[16] P. Somol, P. Pudil, J. Novikova, P. Paclik Adaptive floating search methods in feature selection. *Pattern Recognition Letters* **15(11)** (1999), 1157–1163.

[17] Jason H. Moore, Lance W. Hahn, Marylyn D. Ritchie, Tricia A. Thornton, Bill C. White Routine discovery of complex genetic models using genetic algorithms. *Applied Soft Computing* **4** (2004), 79–86.

Valentin Milanov, Ph.D.
Department of Mathematics and Computer Science
Fayetteville State University
1200 Murchison Road
Fyetteville, NC 28301
e-mail: vmilanov@uncfsu.edu