MISS: a non-linear methodology based on mutual information for genetic association studies in both population and sib-pairs analysis

Helena Brunel1,2,3,*, Juan-Josep Gallardo-Chacon2,3, Alfonso Buil4, Montserrat Valverdú2,3, José Manuel Soria4, Pere Camina2,3 and Alexandre Perera2,3

1 Institut de Bioenginyeria de Catalunya, 2Departament d’Enginyeria de Sistemes, Automàtica i Informàtica Industrial, Universitat Politècnica de Catalunya, Pau Gargallo, 5, 08028 Barcelona, 3CIBER de Bioenginyeria, Biomateriales y Biomedicina and 4Unitat de Genòmica de Malalties Complexes, Institut de Recerca de l’Hospital de la Santa Creu i Sant Pau, Sant Antoni Maria Claret, 167, 08025 Barcelona, Spain

ABSTRACT

Motivation: Finding association between genetic variants and phenotypes related to disease has become an important vehicle for the study of complex disorders. In this context, multi-loci genetic association might unravel additional information when compared with single loci search. The main goal of this work is to propose a non-linear methodology based on information theory for finding combinatorial association between multi-SNPs and a given phenotype.

Results: The proposed methodology, called MISS (mutual information statistical significance), has been integrated jointly with a feature selection algorithm and has been tested on a synthetic dataset with a controlled phenotype and in the particular case of the FVII gene. The MISS methodology has been contrasted with a multiple linear regression (MLR) method used for genetic association in both, a population-based study and a sib-pairs analysis and with the maximum entropy conditional probability modelling (MECPM) method, which searches for predictive multi-locus interactions. Several sets of SNPs within the FVII gene region have been found to show a significant correlation with the FVII levels in blood. The proposed multi-site approach unveils combinations of SNPs that explain more significant information of the phenotype than their single effect on the protein FVII levels in blood.

Availability: The code is available at http://sisbio.recerca.upc.edu/MISS_0.2.tar.gz.

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on February 1, 2010; revised on April 1, 2010; accepted on April 25, 2010

1 INTRODUCTION

One of the main challenges of current genetic research is to study the association between genotypes and phenotypes and, in particular, to identify genetic factors responsible for the heritability of complex traits. The two principal strategies for finding genetic variants related to diseases are linkage analysis and association studies (Emahazion et al., 2001). Linkage occurs when multiple loci are physically related and are inherited together jointly with the disease. Linkage analyses take profit of this effect through measuring the cosegregation of the polymorphisms and a phenotype through a family or set of families. Genetic association looks for correlations between two investigated factors, typically a DNA sequence variability (a genotype) and a trait (a phenotype). In the last decades, technological advances in human genetics have allowed association studies to work with large numbers of genetic markers. Association studies may be affected by the structure or homogeneity of the population under study. In particular, common genetic information between relatives may introduce a bias in the data. The main goal of family studies is to explore this dependence in order to extract additional information. One of the first approaches for family association studies was the analysis of sib-pairs that present more homogeneity of age and environment than other pairs of relatives and are relatively easy to ascertain (Kruglyak and Lander, 1995). However, these methods have been extended to small pedigree (nuclear families) analysis and later on to extended families (Bishop and Williamson, 1990).

The genetic variability represents around 1% of the DNA sequence (Feuk et al., 2006). The genetic variants are responsible for differences between individuals such as physical appearance, susceptibility to disease or response to medical treatments. Among the different types of genetic variants, single nucleotide polymorphisms (SNPs) are the most common type of genetic variation used for the study of genetic diseases (Mooney, 2005; Su, 2007). SNPs are defined as single nucleotide positions in the genome where there is a mutation (i.e. the substitution of one base by another), which is observed in >1% of the population (Brookes, 1999). It is assumed that at most one mutation could have occurred by another), which is observed in >1% of the population (Brookes, 1999). It is assumed that at most one mutation could have occurred at a given locus in the short human evolution (Sarkis et al., 2007). Consequently, only two of the four common nucleotides may be found in a given SNP position, the ancestral one (A1) and the mutated one (A2).

Some variations in the DNA may be responsible for the development of diseases. Traditional genetic techniques have
facilitated the identification of several hundred human genes where there are mutations that lead to a Mendelian disease. However, they are not always useful for studying diseases that depend on the interaction of many gene sequences and the environment. These are known as complex diseases, such as Alzheimer disease or most cardiovascular diseases. For such diseases, multi-loci analysis is expected to be more powerful than the traditional locus-by-locus SNP association studies (Brinza et al., 2006), detecting even more interactions. Moreover, complex diseases generally involve greater difficulties in phenotype definition. The genetic heterogeneity is often closely associated with 'intermediate' phenotypes (Souto, 2003). An intermediate phenotype is a biological and measurable variable that index some aspects of disease risk and susceptibility, such as weight or some protein levels in blood.

Traditional statistical methods are not always useful for the detection of the multiple and combinatorial interactions that come into play in complex diseases. Therefore, the development of computational and statistical methods for detecting combinations of multiple SNPs is of clinical interest. These methods can be split in two basic aspects: the relevance criterion that determines how well a set of SNPs represents the observed variability in the phenotype and the search method used in the selection algorithm (Haldorsson et al., 2004). As many other fields in bioinformatics, genetic association and linkage studies require to use techniques from other engineering sciences. Genetic association can be approached from a pattern recognition point of view. Actually, finding association between genetic variants and a phenotype can be seen as a feature selection (FS) procedure, in the sense of selecting genetic variants with a relevance criterion of association with the phenotype.

The optimal solution to the FS problem requires searching all the possible combinations of features, which is computationally unfeasible (Jain and Zongker, 1997). Thus, FS methods are known to be suboptimal. FS algorithms can be classified depending on their search organization that can be sequential or random (Molina et al., 2002). Sequential algorithms are the most commonly used. Simple sequential algorithms are based on adding or removing features to the selection set by using forward and backward steps, respectively. The most common variants are called sequential forward selection (SFS) and sequential backward selection (SBS). This strategy does not take into account the correlations between features and may produce the effect of finding redundant sets of features. In order to avoid this problem, algorithms that combine forward and backward steps have been proposed. Plus r - take away l algorithm combines r SFS and l SBS. Floating variants were introduced [sequential forward floating selection (SFFS) and sequential backward floating selection (SBFFS)] in order to combine forward and backward steps dynamically (Somol et al., 1999). On the other hand, the search organization may also be driven by a random factor such as in genetic algorithms (GAs) or random search (Shah and Kusiak, 2004). In the last decades, the motivation of using FS in bioinformatics has grown, not only for association but also in other contexts (Saegs et al., 2007). In the domain of SNP analysis, FS has been used for tag SNP selection using linkage disequilibrium (LD) measures (Carlson et al., 2004), multiple linear regression (MLR; He and Zelikovsky, 2006) or sequential search (Dawy, 2006).

These multi-loci LD measures (\(D^2\) and \(r^2\)) can also be expressed as a linear correlation measure (Zaykin et al., 2008). Other methods for measuring the linear dependencies between variables are based on multi-linear regression MLR models (He and Zelikovsky, 2006) or the Student’s \(t\) statistical test and subsequently the Fisher’s \(F\) statistical test (Zhou and Wang, 2007). MLR-based models are the most commonly used method for measuring the linear dependencies between SNPs and phenotypes in the context of genetic association studies. However, these dependencies may not be only linear so that non-linear similarity measures may be appropriate as a relevance criterion of the FS algorithm (Cheng et al., 2007), such as support vector machines (SVMs), neural networks (NNs) and other classifiers from the machine learning field as found in Miller et al. (2009), who has proposed a search algorithm using maximum entropy conditional probability modelling (MECPM) or other proposals employing a boosting classifier based in CARTs (Wan et al., 2009).

Measures from information theory (IT) take into account both linear and non-linear correlations between variables. The basic measure of IT is the entropy, which is an uncertainty measure. The mutual information (MI) is a similarity measure from IT, which do not require a mathematical representation of the genetic space. It has been applied in several fields in bioinformatics (Schneider and Stephens, 1990). In the domain of SNP analysis, a sequential search algorithm based on MI has already been proposed in a genetic association context (Dawy, 2006). Zhang et al. (2009) propose a multi-loci LD measure based on MI for haplotype tagging SNPs selection.

This article is focused on a gene involved in the coagulation cascade. Cardiovascular diseases represent the major cause of death in the world population (Mackay and Mensah, 2004). In particular, ischaemia and venous or arterial thromboses are produced by a blood clot or an obstruction that blocks the blood circulation in a vein or an artery. During the coagulation process, a set of proteins in the blood plasma respond in cascade to form fibrin clots. These proteins are referred to as coagulation factors. Levels of coagulation factors in blood represent a set of intermediate phenotypes that may be a good starting point for identifying genes involved in disease risk for thromboses and ischaemia. It has been demonstrated that some of the coagulation factors have a genetic compound and show significant hereditability (Souto et al., 2000). For example, Factor V Leiden is a variant of factor V produced by a mutation on the gene that codes this protein (F5). Factor V Leiden is the most common hereditary disorder of the coagulation process (Stefano et al., 1998). It has also been published that coagulation Factor VII (FVII) has a genetic effect on disorders of haemostasis (Soria et al., 2005). The genetic variability in F7 gene is the most responsible for observed phenotypic variations in FVII levels.

The main goal of this work is to propose a non-linear method based on IT for finding association between genetic variants (SNPs) and intermediate phenotypes involved in complex diseases and more particularly in thrombosis. This methodology is applied for finding association between multiple sites in the F7 gene and the corresponding intermediate phenotype, factor FVII levels. The proposed method is compared with linear methods for genetic association based on MLR and with a non-linear approach MECPM.
This study has been developed with three datasets. Two of them correspond to real datasets of SNPs of the F7 gene whereas the third one is synthetic.

Coagulation factor VII (FVII) is a vitamin K-dependent protein that plays an important role at the top of the coagulation cascade. The GAIT (Genetic Analysis of Idiopathic Thrombophilia) project is a family-based study of the genetics of thrombosis in the Spanish population (Soria et al., 2005; Souto et al., 2000). In this project, it has been proven that the genetic variability in the F7 gene is the major responsible for factor FVII-level variations. The F7 gene is located at the segment 13q34, on the chromosome 13 of the human genome. It is about 13000 bases long among which around 50 polymorphisms have been identified. The GAIT study is composed of 398 individuals of 21 extended families, 12 of them affected by thrombophilia. For each individual, the symbolic measures of the SNPs (A1/A1, A1/A2, or A2/A2) of the F7 gene are available, as well as the quantitative measures of the phenotype (FVII levels in blood).

On one hand, a population-based study has been performed with 93 independent individuals, selected from the GAIT database. These individuals are called founders and they do not share genetic information from common ancestors. The phenotypes have been discretized using a direct plug-in approach, where kernel functions are used to estimate the optimal bandwidth and bandwidths (Wand, 1996). In this case, the factor FVII levels in blood have been discretized into eight categories.

On the other hand, a sib-pairs analysis has been developed with 345 pairs of sibs, selected from the GAIT database. Generally, family studies are based on comparing the genotypic information of two individuals within the same family. The first approach was to establish a genotypic distance by determining if two individuals share 0, 1 or 2 alleles identical-by-descent (IBD) at a given position (Kruglyak and Lander, 1995). Two alleles are IBD, if one is a copy of the other or if both of them are copies of the same ancestral (Weeks and Lange, 1992). In practice, it is not always possible to estimate the number of alleles shared IBD at a given position because the allelic measurements of the ancestors are not always available. Identity-by-state (IBS) methods also estimate the genotypic differences between sib-pairs. Two alleles are IBS if they are the same allele, regardless of their ancestral origin. The IBS methodology estimates a probability distribution of sharing 0, 1 or 2 alleles IBS by looking at the allelic frequencies (Bishop and Williamson, 1990). For avoiding the computing of these probabilities, the genotypic distance is established directly from the number of alleles shared IBS, as follows. The distance between two identical homozygous genotypes (e.g. A1A1 and A1A1) is d=0. The distance between an homozygous and an heterozygous genotype (e.g. A1A1 and A2A1) is d=1. The distance between two opposite homozygous genotypes (e.g. A1A1 and A2A2) is d=2. For quantitative traits, the number of alleles IBS that two sibs share should present a correlation with the difference of their phenotypes. Thus, the genotypic distance, computed for each sib-pair and each SNP of the F7 gene, is compared with the phenotypic distance which is the difference between the phenotype of each sib-pair and each SNP, as the distance between the two selected SNPs and the phenotype, without assuming the intrinsic properties of the F7 gene.

2 METHODS

2.1 Datasets

This study has been developed with three datasets. Two of them correspond to real datasets of SNPs of the F7 gene whereas the third one is synthetic.

2.2 Linear method

The reference linear method chosen for selecting SNPs correlated with the phenotype is a MLR model (He and Zelikovsky, 2006). MLR tries to fit a model that represents the linear relations existing between a set of independent variables \( S = \{S_i\} \), where \( S_i \) are SNPs, and an observed variable, the phenotype \( f \) as in (1).

\[
f = \beta_0 + \beta_1 S_1 + \cdots + \beta_n S_n + \epsilon
\]

where \( \beta_i \) are the regression coefficients and \( \epsilon \) is the error of the model. The method estimates the values of \( \beta_i \) that minimize \( \epsilon \). Each coefficient \( \beta_i \) represents the individual contribution of a SNP \( S_i \) for the prediction of \( f \).

2.2.1 Correlation of one SNP against the phenotype

The correlation of each SNP \( S_i \) against the phenotype \( f \) is represented by the regression coefficient \( \beta_i \). The statistical significance of this correlation is determined by a Student’s t-test. The null hypothesis supposes the nullity of the corresponding regression coefficient \( (\beta_i = 0) \). Given a set of SNPs \( S \) and a SNP \( S_i \), the Student’s t-test over the regression coefficient \( \beta_i \) is used to determine if \( S_i \) individually adds information about the phenotype \( f \) with respect to \( S \). The significance of the correlation of the total set \( S + S_i \) is described in the next section.

2.2.2 Correlation of a set of SNPs against the phenotype

A Fisher hypothesis test (F-test) is applied to determine if the total set of SNPs is significantly correlated with the phenotype. In this case, the null hypothesis supposes the nullity of the slope of the regression line, i.e. all the regression coefficients at the same time, \( (\beta_i = 0) \). The resulting P-value is used to determine if the SNPs \( S_i \), jointly, have a significant predictive linear capacity over the phenotype \( f \). Once a SNP \( S_i \) is added to the selection set \( S \), the F-test is used to determine if the total set \( S + S_i \) is significantly linearly correlated with the phenotype \( f \).

2.3 Non-linear method

One of the greatest advantages of the MI is that it measures both linear and non-linear correlations between variables. Moreover, MI can be computed directly from the symbol frequencies without a previous numerical representation of genetic data, which involves a loss of information. The MI between a SNP or a set of SNPs \( S \) and a phenotype \( f \) is defined in (2).

\[
I(S|f) = \sum p(S)p(f)\log\left(\frac{p(S|f)}{p(S)p(f)}\right) = H(S) + H(f) - H(S,f)
\]

where \( p(S) \) and \( p(f) \) are the probability distribution functions of \( S \) and \( f \), respectively, and \( p(S|f) \) is the joint probability distribution function for \( S \) and \( f \). \( H(S) \) and \( H(f) \) represent the entropies of \( S \) and \( f \), respectively, and \( H(S,f) \) is the joint entropy of \( S \) and \( f \) (Shannon, 1948). The MI is symmetric and non-negative, \( I(S|f) = 0 \) yields if and only if the two variables are statistically independent and there are no finite sample effects.

2.3.1 Finite sample size effects

Information theoretic measures are biased by finite sample size effects. When the sample size is finite, the MI value is larger than the real information value (3).

\[
I(S|f) > 0
\]

The main effect of the finite sample size in this methodology is that given a set of SNPs \( S \), any other SNP \( S_i \), even a random SNP, will add information.
about the phenotype to the set (4).

\[ IS + S_i(f) \geq IS(f) \]  

(4)

Thus, a statistical hypothesis test is necessary to discriminate SNPs and to know if the gain of information about the phenotype is not only due to the finite sample size problem. This test compares the MI of a SNP against the phenotype with a null distribution of MI. The resulting \( P \)-value helps to decide if the SNP shares a significant amount of information with the phenotype. The null distribution of MI has to show the behaviour of the MI in random cases.

The null distribution of the MI depends on the size of the sample, the degrees of freedom of each particular SNP and the degrees of freedom of the phenotype (Dawy, 2006). A first approach to obtain the null distribution is to generate \( n \) random copies of the SNP under study. The resulting MI null distribution is a vector of the corresponding \( n \) MI values. The \( P \)-value associated to a SNP is calculated by using kernel density estimation for the null distribution and to compare the area under the curve until the MI value of the SNP with the total area under the curve. Another possible approach for obtaining the null distribution is to find an analytical expression. The null distribution of finite sample size MI has been already approximated to a gamma distribution (Goebel, 2005). However, the required regularity conditions are not always met (Szymczak et al., 2007). This approximation is imprecise when the sample size is small, if the number per class differs substantially or if the number of classes is small. It happens mostly when the number of SNPs in a set increases. In this case, the null distribution will also depend on the minor allelic frequency (MAF) of a particular SNP. This is the reason why the surrogate data approach is used for generating the null distribution seems to be appropriate. This technique generates copies of a SNP by destroying its individual order so that the allelic frequencies are respected.

2.3.2 Single SNP significance

In order to avoid the comparison between sets with different numbers of SNPs, this method proposes to evaluate the increase of information about the phenotype produced by a SNP against a given set of SNPs. It is especially interesting in the context of a FS algorithm. The null distribution associated to a SNP \( S_i \) is generated by \( N_c \) surrogate copies \( S_j \) of the SNP respecting its allelic frequencies. Hence, given a set of previously selected SNPs \( S \) and an SNP \( S_i \), the MI of the total set against the phenotype \( I(x; S_i; f) \) is compared with the null distribution of MI \( I(x; S_j; f); j = 1, \ldots, N_c \), generated by the MI of each random copy \( S_j \) of the SNP. The resulting \( P \)-value is the statistical significance of the gain of MI against the phenotype \( f \) produced by the SNP \( S_i \) on the set \( S \). This method has been called MESS (MI statistical significance). The main difference between this work and the one presented in Dawy (2006) is the surrogate technique and the sequential search algorithm used for selecting SNPs.

2.3.3 SNPs set significance

Once a set of SNPs is selected, another statistical test is applied in order to determine if the selection set of SNPs, jointly, have a significant MI against the phenotype. This test consists in comparing the MI of the set against the phenotype \( I(S; f) \) with a null distribution of MI generated with \( N_c \) surrogate copies of the phenotype. The resulting \( P \)-value determines if the selection set, as a set, is significantly related with the phenotype.

For both significance tests, the null distribution has been generated using \( N_c = 3000 \) surrogate copies. The \( p \)-values have been computed using kernel density estimation of the null distribution (Wand and Jones, 1994) and integrating the resulting density function.

2.4 Search method

Both MESS and MLR are relevance criteria to determine if the inclusion of a SNP \( S_i \) in a set \( S \) produces an increase of information about the phenotype. This relevance criterion will be used within a FS algorithm in order to select the sets of SNPs significantly related to the phenotype.

---

**Fig. 1.** SFFS algorithm for genetic association.

Here, an adapted version of the SFFS algorithm will be applied. The original sequential algorithms are one-solution algorithms and find only one suboptimal solution. The proposed version of SFFS finds all the possible solutions.

The multi-solution strategy consists in starting a new search for each significant SNP. This variant of the SFFS algorithm returns multiple sets of SNPs that are jointly able to represent the information of the phenotype. A multi-solution sequential FS algorithm based on relevance chains has already been proposed for SNP selection (Dawy, 2006). The main difference with this algorithm is the floating strategy that avoids finding redundant SNPs. When two SNPs with a similar variability among individuals are selected together, the selection set contains twice the same information about the phenotype. This occurs generally when SNPs are inherited together in the same haplotype. The proposed floating algorithms avoid this problem.

Floating algorithms combine forward and backward steps as in Figure 1. Forward steps add relevant features whereas backward steps allow to deflate the selection set, removing the redundant SNPs. The algorithm starts with a forward test due to the finite sample size effects. Only few SNPs are enough for recovering the whole information about the phenotype so that backward initial steps would remove all the SNPs, which involves starting as much new searches as the number of SNPs in the original set, which is computationally expensive. The SFFS algorithm do not depend on a required number of forward and backward steps, but would find SNPs as long as there is significance.

The algorithm obtains sets of SNPs that, together, show a significant correlation with the phenotype, but that do not share redundant information one with each other. Thus, SNPs appearing in the same set may have a different variability among individuals, and so may belong to different haplotype regions. The proposed algorithm has been applied in both population-based and sib-pairs analysis. For each of them, the traditional linear method have been compared with the proposed non-linear criterion. The significance threshold used in both relevance criteria has been set to 0.05. The main structure of the developed algorithm is described in Figure 1.

2.5 The MISS package

The set of functions and algorithms developed for this study have been integrated in a package for the R statistical language (R Development Core Team, 2005). This package is called MISS and the version 0.2 is already built and available by request to the authors. The MISS library contains a documentation that includes examples of the use of the described algorithm and its underlying functions using a SNP dataset. MISS allows for parallel and distributed computing through MPI. Parallel implementation has been coded on top of snow R-package, authored by (Tierney et al., 2004).
A nonlinear methodology based on mutual information for genetic association studies

Two datasets have been taken into account for the gene study, one considering only unrelated individuals and the other one corresponding to sib-pairs. A total of 5 experiments have been carried out. Three of them correspond to founders data whereas the remaining two have been applied for sib-pairs. The proposed floating FS has been applied with the two described criteria, the linear one (MLR) and the non-linear one (MISS) in both datasets.

Moreover a third methodology, non-linear, corresponding to the MECPM which found single SNPs related with FVII levels in blood. Most of the SNPs obtained in the three cases are reported in the literature as functional variants related with FVII gene. It is important to remark that the sets obtained with the proposed floating search algorithm contain SNPs that do not give information about the phenotype but that

Table 1. Results obtained for founders data analysis

<table>
<thead>
<tr>
<th>Method</th>
<th>MLR</th>
<th>MISS</th>
<th>MECPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>151</td>
<td>48</td>
<td>61</td>
</tr>
<tr>
<td>( α )</td>
<td>6</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>( P )-value</td>
<td>( 10^{-3} )</td>
<td>( 10^{-3} )</td>
<td>( 10^{-1} )</td>
</tr>
</tbody>
</table>

\( N \) represents the number of sets obtained and \( α \) the average number of SNPs in a set. \( P \)-values are the order of magnitude of the obtained \( P \)-values and relevant sets of SNPs are those that present most statistical significance. Each entry on the left side of the columns represents a set of SNPs whereas SNPs in curly brackets are common SNPs appearing in all sets (left side of the columns).

3 RESULTS AND DISCUSSION

The methodology described above has been applied to both a real case corresponding to the F7 gene and to a simulated dataset.

3.1 Application to the F7 gene

Two datasets have been taken into account for the F7 gene study, one considering only unrelated individuals and the other one corresponding to sib-pairs. A total of 5 experiments have been carried out. Three of them correspond to founders data whereas the remaining two have been applied for sib-pairs. The proposed floating FS has been applied with the two described criteria, the linear one (MLR) and the non-linear one (MISS) in both datasets. Moreover a third methodology, non-linear, corresponding to the MECPM software has been applied for founders data. The results obtained for the population-based study are presented in Table 1, while Table 2 shows the results obtained in the sib-pairs analysis.

For each experiment, Tables 1 and 2 show the number of SNPs' sets obtained (\( N \)), the average number of SNPs in a set (\( n \)), the \( P \)-values order of magnitude and the most relevant sets of SNPs, corresponding to the most statistical significant ones. SNPs obtained in Tables 1 and 2 are represented in Figure 2, showing their location within the F7 gene.

Table 2. Results obtained for the sib-pairs analysis

<table>
<thead>
<tr>
<th>Method</th>
<th>MLR</th>
<th>MISS</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>( n )</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>( P )-value</td>
<td>( 10^{-6} )</td>
<td>( 10^{-18} )</td>
</tr>
</tbody>
</table>

\( N \) represents the number of sets obtained and \( n \) the average number of SNPs in a set. \( P \)-values are the order of magnitude of the obtained \( P \)-values and relevant sets of SNPs are those that present most statistical significance. Each entry on the left side of the columns.

\( Soria \) et al. (2007); \( H \) van't Hooft and Hermans (2009); MECPM is an available algorithm based on the maximum entropy principle. MECPM is designed for genotypic data. It has been applied only to the founders dataset given that sib-pairs IBS data are located in the promoter region or in splice sites as shown in Figure 2.

The SNPs' sets obtained using MLR are similar since they contain common SNPs. The SNPs that make the sets different (rs762636, rs36208415, rs36208416 and rs510317) contain similar information about the phenotype. These SNPs appear in Soria et al. (2005) in the same cluster of SNPs with a high probability of posterior effect on the phenotype and are located in the promoter region or in splice sites as shown in Figure 2.

The SNPs' sets obtained using MISS also contain common information. SNPs that make the sets different (rs491098, rs510335 and rs561241) appear in the same cluster of SNPs in Soria et al. (2005). Moreover, both SNPs rs493833 and rs491098 belong to the fifth intron of the F7 gene. It is important to remark that the sets obtained with the proposed floating search algorithm contain SNPs that do not give information about the phenotype but that
Fig. 3. Two-loci interactions between SNPs. The histogram on the left represents the significance of the correlation of each individual SNP with the phenotype. The matrix at the centre shows the statistical significance of adding the SNPs in files to the SNPs in columns (black squares represent significant combinations of SNPs) and LD figure on the right represents the $r^2$ LD measure between each pair of SNPs.

Most of the SNPs found with MECPM have been reported in the literature as functional polymorphisms related with the phenotype. It can be observed in Table 2 that for sib-pairs data, MLR only found 3 sets of SNPs, whereas MISS found 50 sets, a similar number than using founders data. As for the population-based study, sets obtained using MLR contain more SNPs ($n = 4$) than sets obtained with MISS ($n = 3$). Conversely to the founders case, MISS obtains higher $P$-values than MLR. SNP sets obtained with MLR are also similar being differentiated by SNPs rs762636, rs762635 and jointly by rs2774033 and rs6039. SNPs rs762736 and rs762635 are in the same cluster in Soria et al. (2005), while SNP rs6039 appears in another cluster. Most of the sets are composed by SNPs that appear in different clusters in Soria et al. (2005), giving different information about the phenotype. In particular, SNP rs36209567, also known as A294V, appears in several sets and is located in the ninth exon of the F7 gene, producing an amino acid change in the resulting protein from an alanine to a valine. SNP rs36208758 is also located in the third exon but it is a missense mutation that does not produce any amino acid change in the resulting protein (Fig. 2).

### 3.2 Application to simulated data

The simulation study has been developed to validate the performance of our methodology in detecting true interactions between SNPs and a phenotype defined by an epistatic multiplicative model. For each method, we built six datasets of different size, corresponding to 5, 10, 15, 20, 25 and 50 SNPs and 85 samples. The datasets contain two SNPs of the F7 gene and random SNPs non-related with the phenotype, generated by a multiplicative model based on the information of the two SNPs. The correlation between the synthetic phenotype and the two selected SNPs, as a set, is non-linear as it has been built through a multiplicative model.

MLR only detects one of the SNPs, regardless of the size of the dataset. MECPM finds several SNPs as individual sets, including the two selected SNPs. In contrast, for each dataset size, MISS is able to detect sets containing both SNPs.

Table 3 shows the CPU time corresponding to each method. All computations were performed on a 12 Intel E7310 processors (4 MB Cache 1.60 GHz) with 32 GB random access memory. MISS was launched using snow on MPI mode over the 12 nodes. MECPM has been applied to a binary phenotype, giving faster and better results whereas with MLR and MISS it has been discretized into eight categories using Wand (1996). The parameters of MISS have been also adjusted in benefit of obtaining a right detection with the minimum computational cost. Thus, the null distribution has

<table>
<thead>
<tr>
<th>Size of the dataset</th>
<th>CPU time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MLR</td>
</tr>
<tr>
<td>5</td>
<td>0.2</td>
</tr>
<tr>
<td>10</td>
<td>0.3</td>
</tr>
<tr>
<td>15</td>
<td>0.5</td>
</tr>
<tr>
<td>20</td>
<td>0.7</td>
</tr>
<tr>
<td>25</td>
<td>0.8</td>
</tr>
<tr>
<td>50</td>
<td>2.8</td>
</tr>
<tr>
<td>47*</td>
<td>238</td>
</tr>
</tbody>
</table>
As this study is a local association analysis focused on the F7 gene, functional studies about F7 polymorphisms are used to validate our results. Most of the SNPs found have been reported in the literature as functional elements correlated with the phenotype. The results presented in Soria et al. (2005) were confirmed by the same group in Sabater-Lleal et al. (2007) by functional assays. Moreover, some of these results have been also replicated by association analysis and/or functional assays in van’t Hooft et al. (1999), Yang et al. (2007), Marchetti et al. (1993), Wulff and Hermann (2000), Feng et al. (2000) and Pollak et al. (1996).

It has been observed that results obtained for founders data are different than results obtained in the sib-pairs analysis. Most of these differences are due to the differences in the datasets, containing different individuals and different genotypic measures. However, intrinsic differences in the variability of genetic data between individuals may also influence the results. The large variability present in founders genetic data may increase the false positive discovery (Lawrence et al., 2005). This has been observed with the high values of $N$ and $n$, especially with MLR. MISS is more conservative as it finds a similar number of SNP sets for both datasets ($N \sim 50$).

Contrarily to the genotypic variability, the variance of the phenotypic differences of the sib-pairs is higher ($V = 1159.7$) than the variance of the phenotypes of the founders ($V = 826.3$). This variability can only be expressed through the combinations of SNPs. This combinations are more easily found using MISS (50 combinations obtained) than using MLR (only 3 combinations).

Using the comparison with the synthetic dataset, it has been observed that neither MLR nor MECPM find the true positive corresponding to the combination of the first two SNPs whereas MISS is able to detect this interaction. Figure 3 illustrates this through a singular case (generated with founders data). It can be observed that SNPs rs9604025 and rs510335 are not individually significantly correlated with the phenotype, whereas the combination of these two SNPs is significantly related with the phenotype (red square). Moreover, this combination of SNPs is not detected as a significant one using MLR or MECPM. SNP rs9604025 appears in Soria et al. (2005) as a functional variant related to FVII levels. SNP rs510335 is a relevant SNP cited in many works related to the factor VII. The rare T allele is associated with lower plasma concentrations of FVII protein and fully activated FVII molecules (van’t Hooft et al., 1999). Moreover, the LD plot shows that these SNPs do not present a significant correlation. On one hand, Figure 3 demonstrates that the multi-site strategy detect genotype–phenotype associations that one-by-one SNP approaches do not detect. On the other hand, it has been shown that some of these combinatorial effects of SNPs on the phenotype are found using MISS whereas they are not detected with the other linear or non-linear approaches.

However, this accuracy is obtained by increasing the complexity and sacrificing the computational performance of the algorithm. This is not a critical point for local association studies like this, but it may become severer in a genome-wide association study (GWAS).

4 CONCLUSION

We have presented the MISS methodology, a non-linear method for genetic association based on the MI statistical significance of sets of SNPs against a phenotype under study. This method has been applied as a relevance criterion of a floating FS algorithm, proposed in the context of genetic association for complex diseases. MISS has been compared with MLR, a linear method commonly used for genetic association and with MECPM, an algorithm for searching predictive multi-loci interactions. The different methods have been tested with two samples of the GAIT project, corresponding to a population-based study and a sib-pairs analysis and also with a simulated dataset. In particular, the goal was to find association between SNPs of the F7 gene and factor VII protein plasma levels. It has been demonstrated that with GAIT project data and a synthetic dataset, MISS improves the results of traditional genetic association methods. On one hand, multi-site association improves the results obtained with one-by-one SNP association methods, showing that combinations of SNPs may contain information about the phenotype that single SNPs are not able to capture. On the other hand, the proposed non-linear method (MISS) is not only able to recover the results from other methods but it also improves them, finding correlations between genotype and phenotype not detected with the other methods. Several sets of SNPs of the F7 gene have been found as genetic features able to describe the information of the phenotype (FVII levels in blood). The results obtained in this study for the F7 gene have been validated by previous works using an experimental approach. Most of the obtained SNPs have been previously reported in the literature as functional variants that have a molecular effect on FVII levels. The application of MISS to the simulated dataset has proven its capacity of finding true associations against the other methods. However, this accuracy is obtained at the cost of an increased computational task of the algorithm that should be improved for its use in GWASs and for its application to other diseases or phenotypes.

Funding: Spanish Ministerio de Ciencia y Tecnología through the CICYT (grant TEC2007-63637/TCM, partially); Institut de Bioenginyeria de Catalunya (grant 01-8/07/IBEC, partially); Ramon y Cajal program from the Spanish Ministerio de Educación y Ciencia (partially); grants PI-08/0420 and PI-08/0756 (partially); “Programa d’Estabilització d’Investigadors de la Direcció d’Estratègia i Coordinació del Departament de Salut” from Generalitat de Catalunya (to J.M.S.); CIBER-BBN (an initiative of the Spanish ISCIII).

Conflict of Interest: none declared.

REFERENCES


