Floating Feature Selection for Multiloci Association of Quantitative Traits in Sib-pairs Analysis

H. Brunel, A. Perera, A. Buil, M. Sabater-Lleal, J. C. Souto, J. Fontcuberta, M. Vallverdú, J. M. Soria, P. Caminal

Abstract—**Finding association between genotypic differences and disease traits has become one of the main objectives in current genetic research. It has been published that some of the underlying factors in the dynamics of the coagulation process have a genetic compound, showing significant hereditability. This is the case of the Factor VII. In this work, we propose a method for selecting sets of Single Nucleotide Polymorphisms (SNPs) of the *F7* gene that are significantly related with the phenotype (Factor VII levels). The methodology is applied to the sib pairs from the GAIT project sample. The method consists of an adapted Sequential Floating Feature Selection (SFFS) algorithm. This algorithm is applied with two relevance criteria, one linear and one non linear. The SNPs sets found with linear models are included in the sets found with non linear techniques. The results fit in with previous results in clinical area.**

I. Introduction

One of the main goals of association studies is to discover genetic loci that are responsible for the hereditability of complex diseases [1]. Complex diseases are caused by the interaction of multiple genes and environmental factors. In the global burden of complex diseases, cardiovascular diseases represent the majority. In particular, ischemia and venous or arterial thromboses are cardiovascular complex diseases. They are produced by the suppression of the blood circulation in a vein or artery due to a clot or an obstruction of these blood vessels. During the coagulation process, a set of proteins in the blood plasma respond in cascade to form fibrin strands which strengthen the platelet plug. These proteins are called coagulation factors. It has been demonstrated that alterations in some of these coagulation factors have a genetic compound, showing significant hereditability [2]. For example, Factor V Leiden is a variant of Factor V produced by a mutation on the gene encoding this protein. It has been published that coagulation Factor VII (FVII) has also a genetic effect on disorders of hemostasis. The Factor VII is a vitamine K- dependent protein that plays an important role in the initiation of the coagulation process. The GAIT project (Genetic Analysis of Idiopathic Thrombophilia) is a family-based study of the genetics of thrombosis in the Spanish population [3]. In this project it has been demonstrated that the observed phenotypic variations in the factor VII levels in plasma are mostly due to the genetic variability in the *F7* gene.

The genotype-phenotype association studies rely on genetic markers. Single Nucleotide Polymorphisms (SNPs) are the most commonly used [4]. SNPs are positions in the genome where there is a mutation (normally a substitution of one base by another) which has been conserved generation by generation achieving a frequency of more than 1% of the population. There also exist other genetic variability sources that are starting to attract much attention such as Copy Number Variants (CNVs). CNVs are segments of the DNA that are 1kb or larger in size, present a variable number of times as copies in a genome [5].

The association between polymorphisms and phenotypes can depend on the population structure or homogeneity, and on the relationships between individuals [6]. This is the reason why researchers introduced family-based studies. Family-based studies can be divided into two categories which are sib-pairs analysis and family studies. Sib pairs tend to present more homogeneity of age and environment than other pairs of relatives [7].

In bioinformatics, the motivation of using feature selection (FS) has grown in the last decades, both in linkage and association studies. There exist two approaches for the feature selection problem [8]. Filter techniques are generally applied as a pre-processing step of the prediction model. In this case, the feature selection is independent of the response variable. On the other hand, wrapper approaches take into account the variable to be predicted during the feature search strategy. Feature selection consists on searching sets of relevant features from an original set, without losing significant information. The optimal solution to this problem involves an exhaustive search, which is computationally unfeasible [9]. This is why FS methods are known as suboptimal algorithms. Sequential algorithms are the most

This work has been partially supported by Spanish Ministerio de Ciencia y Tecnología through the CICYT grant TEC2007-63637/TCM, by Instituto de Bioingeniería de Cataluña through the grant 01-8/07/IBEC, by Ramon y Cajal program from the Spanish Ministerio de Educación y Ciencia and also by the grant Redes Temáticas de Investigación Cooperativa (RETIC) Cardiovascular (RECAVA) Exp-06/0014/0016RD from Fondo Investigación Sanitaria.

H. Brunel is at Institut de Bioenginyeria de Catalunya, Centre de Recerca en Enginyeria Biomèdica, Departament d'Enginyeria de Sistemes , Automàtica i Informàtica industrial, Universitat Politècnica de Catalunya, Pau Gargallo 5, 08028 Barcelona, Spain. Email: helena.brunel@upc.edu.

A. Perera, M. Vallverdú and P. Caminal are at Centre de Recerca en Enginyeria Biomèdica, Departament d'Enginyeria de Sistemes, Automàtica i Informàtica Industrial, Universitat Politècnica de Catalunya, Pau Gargallo 5, 08028 Barcelona, Spain, and at the "CIBER de Bioingeniería, Materiales y Nanomedicina (CIBER-BBN). Email: {alexandre.perera, montserrat.vallverdu, pere.caminal}@upc.edu.

A. Buil, M. Sabater-Lleal, J.C. Souto, J. Fontcuberta and J.M. Soria are at Unitat d'Hemostàsia i Trombosi, department d'Hematologia, Hospital Sant Pau, Sant Antoni Marià Claret 167, 08025 Barcelona, Spain. Email: {abuil, msabater, jsouto, jfontcuberta, jsoria}@santpau.es

commonly used. The first sequential algorithms were Sequential Forward Selection (SFS) and Sequential Backward Selection (SBS). Both of them produce the effect of finding redundant sets of features [10]. *Plus r- take away l* methods were developed for avoiding this problem. These algorithms combine *r* SFS and *l* SBS. Afterwards, the floating algorithms were introduced (SFFS: Sequential Forward Floating Selection and SBFS: Sequential Backward Floating Selection). These algorithms also combine forward and backward steps but dynamically and without depending on the *r* and *l* parameters. Wrapper algorithms can also be divided in two categories deterministic or randomized methods. While deterministic methods always produce the same result, randomized methods depend on a random element that could produce different sets on every run.

On the other hand, feature selection algorithms depend on the relevance criterion used to decide if a set of features is significantly representative. Depending on the correlation measure, the relevance criterion can be classified as linear or nonlinear. While Pearson coefficient measures linear correlations, mutual information can measure both linear and non linear correlations between variables [11]. Moreover, the information theory does not require a numerical representation of the genetic space. Information theory was born in telecommunications [12] but it has been applied to genetic investigations. In particular it has been applied to genotype-phenotype association studies [13].

In this work, we apply a version of the SFFS algorithm for finding sets of SNPs of the *F7* gene related with the phenotype (FVII levels in plasma). This has been done in a family–based study with a sib-pairs approach. We compare the SNPs *F7* sets found with two criteria. The nonlinear criterion is based on the mutual information statistical significance [14], and the linear criterion is based on linear regression models [15].

II. Materials and methods

A. GAIT Sample

The *F7* gene is located on the chromosome 13 of the human genome. It is about 13000 bases long among which 50 polymorphisms have been identified. The GAIT sample is composed by 399 individuals of 21 extended families. There are 12 families affected of thrombophilia. For each individual, we have the symbolic measures of the 50 SNPs, and also the quantitative value of the FVII levels in blood. We have selected 345 sib pairs and we have established genotypic distance between them by a methodology based on Identity-by-State methods (IBS). We have also established the phenotypic distance to the difference between the values of the FVII levels. The phenotypes differences represent a continuous variable that follows a normal distribution with *mean*=2.5 and *standard deviation* =34

B. IBS methodology

Given a SNP position, the sib-pair analysis is based on determining if two sibs share 0, 1 or 2 alleles Identical-by-Descent (IBD) in this position [7]. Two alleles are IBD if one is a copy of the other or if they are both copies of the same ancestral [16]. In practice, the number of shared alleles IBD in a given position is difficult to calculate because the allelic measurement of the parents is not always available. Moreover, the IBD methodology is more appropriate for linkage analysis. The genotypic difference between sib pairs can also be inferred from Identity-by-State (IBS) methods. Two alleles are Identical-by-State if there are the same alleles, regardless of their ancestral origin. For quantitative traits, the phenotypic difference between sib pairs should present a correlation with the number of alleles IBS they share. In order to compare it with the phenotypic distance, a genotypic distance is established between each sib pair and at each SNP position. The IBS methodology estimate a probability distribution of sharing 0, 1 or 2 alleles IBS [17]. As it is described in the next section, the mutual information measure takes into account the relative probability of each symbol in the variables. Thus, given a SNP position, the genotypic distance between two sibs is set to the number of shared alleles IBS. The distance between two identical homozygous genotypes (e. g. AA and AA) is *d=0.* The distance between an homozygous and an heterozygous genotype (e. g. AA and AC) is *d=1*. The distance between two opposite homozygous genotypes (e. g. AA and CC) is *d=2*. When there are missing values in a SNP, the genotypic distance between sib pairs in this position is estimated from allelic frequencies. The missing values are set to the most frequent genotype. These symbolic registers (0, 1, 2) of the genotypic distances between sib pairs are used to represent the phenotypic distances. The feature selection algorithm is used with mutual information or linear regression models.

C. Mutual Information

The mutual information measures the generalized correlation between genotypic differences in SNP positions and phenotypic distances. It is computed directly from the symbol frequencies. *I(S, f)* denotes the mutual information between the differences on the SNPs of a set S against the phenotypic differences (1):

$$\begin{aligned} I(S,f) &= \sum_S \sum_f p(S,f)\log_2\left(\frac{p(S,f)}{p(S)p(f)}\right) \\ &= H(S) + H(f) - H(S,f) \end{aligned} \tag{1}$$

where *p(S)* and *p(f)* are the probability distribution functions of *S* and *f* respectively. *p(S,f)* is the joint probability distribution function for *S* and *f*. *H(S)* y *H(f)* represents the entropies of *S* and *f*. *H(S, f)* is the joint entropy of *S* and *f*. The mutual information is a symmetric and nonnegative measure. *I(S,f)*=0 if and only if the two variables are statistically independent and there are no finite simple effects. Even when a random feature is added to the set *S*, an increase of mutual information is produced, due to the finite sample size. Thus, a statistical hypothesis test is necessary to discriminate SNPs. This test compares the increase of information produced by a SNP with a null distribution of mutual information. The resulting p-value helps to decide if the SNP contributes with significant information to a set of SNPs *S*.

SNPs have different allelic frequencies that depend on the evolution of the mutation over a given population. The informative contribution of a SNP about the phenotype depends strongly on these allelic frequencies. Thus, the null distribution must take into account the allelic frequencies of the SNP to be compared with. A surrogate data technique generates copies of a SNP destroying its individual order so that the allelic frequencies are respected. Hence the null distribution is generated by a certain number of surrogate copies of the SNP under study.

D. Linear Regression

The multiple linear regression model intends to learn about linear relations that exist between a set of independent variables $S = \{S_i\}_{i=1..n}$ and an observed variable, the phenotype *f (2)*.

$$f = \beta_0 + \beta_1 S_1 + \Lambda + \beta_n S_n + \varepsilon \qquad (2)$$

where β_i are the coefficient of the model. They represent the independent contribution of the independent variables to the prediction of *f*. ε *is* the error of the model. The method estimates the β_i coefficients that minimize ε. A *t-Student* statistical test is applied to determine the significance of the linear correlation between the SNPs and the phenotype. A p-value helps to decide if the correlation between S_i and *f* is statistically significant.

Afterwards, a Fisher hypothesis test (*F-Test*) is applied. The null hypothesis supposes the nullity of the slope of the regression line (β_i=0). The resulting p-value determines if the independent variables (S_i), jointly, have a significant predictive capacity over the dependent variable *f*.

E. Sequential Floating Feature Selection

The algorithm of feature selection is an adapted version of the SFFS algorithm which finds not only the optimal solution but also several relevant sets of features. This algorithm returns sets of SNPs that are jointly able to represent the information of the phenotype (FVII levels). The SFFS algorithm has been applied with two relevance criteria. The linear criterion is based on the statistical significance of a multiple linear regression model, as defined in section 2.*D*. The nonlinear criterion is based on the statistical significance of the mutual information as described in section 2.*C*. The algorithm starts with the empty set S = {}. Features are added in S when the p-value associated to the information contribution of the feature is greater than a given threshold. Inversely, features are removed from S when the p-value associated to the information loss is lower than the threshold.

The algorithm combines the forward and backward steps as described below:

1. Initialization of the set *S*={},
2. **Forward step***:* for each available SNP S_i, the p-value associate to its information contribution to *S* is computed
3. For each significant SNP, a new forward (2) search is started from the new set S:= S+{S_i}. The forward step (2) is repeated whereas there are significant SNPs.
4. **Backward step***:* For each SNP S_i in S, the p-value associated to the loss of information when removing S_i from S is computed.
5. For each nonsignificant SNP, a new backward search (4) is started from the new set S:= S-{S_i}. The backward step (4) is repeated whereas there are nonsignificant SNPs and the set S has more than one SNP.
6. Go to step 2.
7. If there are neither significant SNPs in step 3 nor nonsignificant SNPs in step 5, the search is stopped.

III. RESULTS

The results presented in this section were obtained with certain simulation parameters. The significance threshold has been fixed to 0.005 for both linear and nonlinear methodologies. The phenotypic differences have been discretized to 16 classes, according to the methodology proposed in [18]. For the nonlinear criterion, the null distribution of mutual information has been generated with n=3000 random features by surrogate data technique. A p-value is computed from this null distribution using density estimation as proposed in [19]. When finding a set, we compute its p-value as a set by surrogating the phenotypic differences.

In figure 1, the mutual information (MI) between genotypic and phenotypic differences is shown. This mutual information value is compared to the MI value corresponding to the significance threshold (pval = 0.005).

Sets of SNPs obtained with the linear regression models are presented in Table I. SNPs sets obtained using the statistical significance of the mutual information are presented in Table II.

TABLE I
SNPS SETS AND ITS P-VALUES OBTAINED WITH LINEAR MODELS

SNP sets	p-value
73, -401	2.51e-5
698, -401	4.51e-6

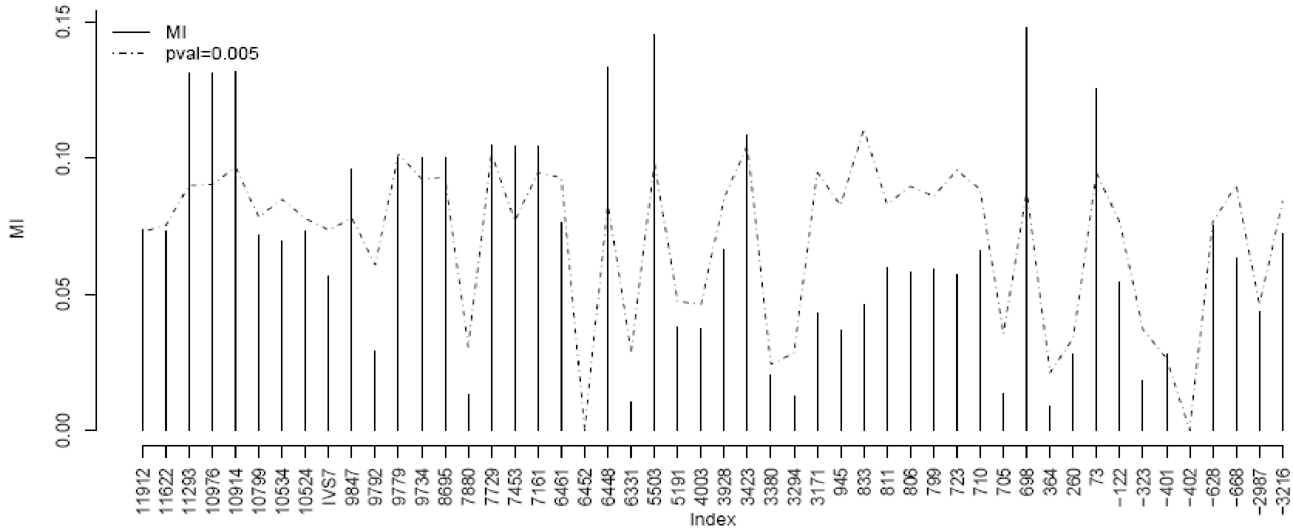

Fig. 1. Mutual Information (MI) of genotypic differences at SNP positions of the *F7* gene against phenotypic differences on the FVII levels. The dotted line represents the corresponding mutual information value associated to a significance level of 0.005.

TABLE II

SNPs sets and its p-values obtained with mutual information statistical significance

SNPs sets	p-value	SNPs sets	p-value
11293, 4003 ,-401	<0.0005	6448, 698	<0.0005
11293,-401, -628	<0.0005	6448, -628, -401	<0.0005
10976, 4003 ,-401	<0.0005	3423, -401, -628	<0.0005
10976, -401, -628	<0.0005	698	<0.0005
10914,4003,-401	<0.0005	698, 2987	<0.0005
10914, -401, -628	<0.0005	**-401, 698**	**<0.0005**
9779, 4003, -401	<0.0005	-401, 5503, -628	<0.0005
73, -401, 6448	**<0.0005**	-628, 7453, -2987	<0.0005
-628, 73	<0.0005	-628, 7161, -2987	<0.0005
9734, 4003, -401	<0.0005	-628, -2987, 7729	<0.0005
8695, 4003, -401	<0.0005	-2987, 73	<0.0005

We observe that linear models find 2 sets of 2 SNPs while 22 sets of 2 or more SNPs are found with nonlinear methods with strong statistical significance levels. Sib pair's analysis is useful for reducing the variability in the data. This is the reason why the method is more powerful than using nonrelated individuals regarding the significance levels [13].The two sets obtained with the statistical significance of linear regression models are reproduced by using mutual information statistical significance, as shown in bold in Table II. The set (73, -401) is included in a set found using nonlinear measures with the addition of the 6448 SNP.

In previous biological researches about the *F7* gene, we found that SNPs 73 and -401 are included in a set of SNPs with molecular evidence of effect on the phenotype [2]. This agrees with one of the sets found in both linear and nonlinear methods. In the other hand, previous studies on the Spanish population have demonstrated that several SNPs (-2987, -628, -401, 73, 698 and 10976) found by our methodologies are functional polymorphisms over the *F7* gene [20]. We can emphasize on SNPS -401, 73, 698 and 6448 as the most relevant. In particular, we can observe in Table II that the SNP -401 appears recurrently in most of the sets. The information of this SNP is statistically significant, as we can see in figure 1, but it is not the most significant one. It means that the information contribution of this SNP is statistically significant and independent of the information of the other SNPs. It has been demonstrated that SNP -401 is a functional polymorphism in the promoter region of the *F7* gene, related to the phenotype [21], [22]. The mutated allele (T instead of G) in this SNP position is associated with lower plasma levels of FVII protein. The mutation of this SNP is also strongly related to the binding properties of the protein complexes.

IV. Conclusion

In this work we have proposed an adapted Sequential Forward Floating Selection algorithm for obtaining sets of SNPs of the F7 gene significantly related with the phenotype (FVII levels). The algorithm finds several relevant sets of SNPs and it has been applied with two different methodologies. The first one is based on a multiple linear regression model and the second one is based on a nonlinear measure (mutual information). The nonlinear criterion returns a greater number of statistically significant sets than the linear models. The SNPs sets found with linear model are included in some of the SNPs sets obtained with the mutual information criterion. In particular, the SNPs set (73, -401), reproduced by both criteria is one of appears in previous biological publications in a set of SNPs that present molecular evidence of effect on the phenotype. Most of correlated SNPs have been reported as functional polymorphisms related to the Factor VII levels. IN particular, the -401 SNP appears in most of the sets found by the algorithm. This SNP brings statistically significant information about the phenotype. This information is also independent of the information brought by other SNPs.

ACKNOWLEDGMENT

CIBER-BBN is an initiative of the Spanish ISCIII.

REFERENCES

[1] Z. Dawy et al. “Gene mapping of complex diseases”. *IEEE Signal Processing Magazines,* vol. 24, no. 1, 2007, pp. 83-90.

[2] J. M. Soria et al. “The F7 gene and clotting factor VII levels: Dissection of human quantitative trait locus”. *Human Biology,* vol. 77, no. 5, 2005, pp. 561-75.

[3] J.C. Souto et al., “Genetic susceptibility to thrombosis and its relationship to physiological risk factors: the GAIT study”, Am. J. Hum Genet. Vol. 67, 2000, pp. 1452-59.

[4] B. V. Halldórsson, S. Istrail, F. M. De La Vega. “Optimal selection of SNP markers for disease association studies”. *Human Heredity,* vol. 58, 2004, pp. 190-202.

[5] B. Stranger et al., “Relative impact of nucleotide and copy number variation on gene expression phenotypes”. *Science,* vol. 315, 2007, pp. 848-53.

[6] M. George, X. Hongyan, G. Varghese. “Comparison of Family Based Association Tests using the North American Rheumatoid Arthritis Consortium Data”. *Genetic Analysis Workshop 15*, 2006.

[7] L. Kruglyak, E. S. Lander. “Complete multipoint sib-pair analysis of qualitative and quantitative traits”. *Am. J. Hum. Genet*, vol. 47, 1995, pp. 439-54.

[8] Y. Saeys, I. Inza, P. Larrañaga. “A review of feature selection techniques in bioinformatics”. *Bioinformatics,* vol. 23, no. 19, 2007, pp. 2507-17.

[9] A. Jain, D. Zongker, “Feature selection: evaluation, application, and small sample performance”, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 19, no. 2, 1997, pp. 153-8.

[10] P. Somol, P. Pudil, J. Novovicova , P. Paclik. “Adaptive floating search methods in feature selection” *Pattern Recognition Letters,* vol. 10, 1999,pp. 1157-63.

[11] M. Ng, L. Chan, “Informative gene discovery for cancer classification from microarray expression data”, *IEEE Workshop on Machine Learning for Signal Processing,* 2005, pp. 393-8.

[12] C. Shannon. “A mathematical theory of communication”. *The Bell Systems Technical Journal*, vol. 27, 1948, pp. 379-423.

[13] Z. Dawy et al., “Gene mapping and marker clustering using Shannon’s mutual information”, *IEEE transactions on computacional biology and bioinformatics,* vol. 3, no. 1, 2006, pp. 47-56.

[14] H. Brunel et al. “SNP Sets Selection under Mutual Information Criterion, Application to *F7*/FVII dataset”. *30th Annual international conference on the IEEE Engineering in Medicine and Biology Society,* 2008.

[15] J. He, A. Zelikovski. “MLR-tagging: informative SNP selection for unphased genotypes based on multiple linear regression”. B*ioinformatics,* vol. 22, no. 20, 2006, pp. 2558-61.

[16] D.E. Weeks, K. Lange. “A multilocus extension of the Affected-Pedigree-Member Method of linkage analysis”, *Am. J. Hum. Genet.,* vol. 50, 1992, pp. 859-68.

[17] T. Bishop, J. A. Williamson. “The power of Identy-by-State methods for linkage analysis". *Am. J. Hum. Genet,* vol. 46, 1990, pp. 254-65.

[18] M.P. Wand, “Data-based choice of histogram binwidth”, University of New South Wales, Australian Graduate School of Management Working Paper Series, no. 95–011. 1995

[19] M.P. Wand, M.C. Jones, *Kernel smoothing.* Chapman and Hall, London, 1995.

[20] M. Sabater-Lleal et al. “Complexity of the genetic contribution to factor VII deficiency in two Spanish families: clinical and biological implications”. *Journal of Hematology*, vol. 88, 2003, pp.906-13.

[21] M. Ferdinand et al. “A. Two common functional polymorphisms in the promoter region of the coagulation factor VII gene determining plasma factor VII activity and mass concentration”. *Blood*, vol. 93, no. 10, 1999, pp. 3432-41.

[22] D. Girelli et al. “Polymorphisms in the factor VII gene and the risk of myocardial infarction in patients with coronary artery diseases”. *The New England Journal of Medicine,* 2000, pp. 774-80.