Efficient identification of epistatic effects in multifactorial disorders

Orlando Anunciação[1], **Susana Vinga**[1,2], **Arlindo L. Oliveira**[1]*
1 INESC-ID / Instituto Superior Técnico, Universidade de Lisboa, Portugal
2 IDMEC/LAETA, Instituto Superior Técnico, Universidade de Lisboa, Portugal
* E-mail: aml@inesc-id.pt

Abstract

Complex diseases are typically caused by the combined effects of multiple genetic variations. When epistatic effects are present, these genetic variations show stronger effects when considered together than when considered individually. To identify groups of single nucleotide polymorphisms (SNP) that can be of use in the explanation of multifactorial conditions, we look for pairs of SNPs with a high value of information interaction. We show that standard classification methods or greedy feature selection methods do not perform well on this problem. We propose a computationally efficient method that uses the information interaction value as a figure of merit, and compare it with state of the art methods (BEAM and SNPHarvester) in artificial datasets simulating epistatic interactions. The results show that the method is powerful and efficient, and more effective at detecting pairwise epistatic interactions than existing alternatives. We also present results of the application of the method to the WTCCC breast cancer dataset. We found 89 statistically significant pairwise interactions with a p-value lower than 10^{-3}. Somewhat unexpectedly, almost all the SNPs involved in pairs with high value of information interaction also have moderate or high marginals, a result that may imply that the search for more complex interactions may be more effectively conducted by looking only at SNPs which, by themselves, have correlations with the condition under study.

1 Introduction and Related Work

Genome Wide Association Studies (GWAS) aim at discovering associations between genetic factors and specific conditions. In GWAS, hundreds of thousands of Single Nucleotide Polymorphisms (SNPs) are analyzed to determine whether they are associated with the disease or conditions of interest. These analyses are usually performed using single SNP statistical tests. This approach has severe limitations since epistatic interactions of SNPs are very important in determining susceptibility to complex diseases. Existing methods for SNP interaction discovery perform poorly when marginal effects of disease loci are weak or absent. The problem is that the individual effects of the interacting SNPs may be too small to be detected with the most commonly used statistical methods. Therefore, there is a need for more powerful methods that are able to identify interactions between SNPs with low marginal effects.

A number of different methods have been used to find epistatic interactions, including statistical methods (e.g. ATOM [1]), search methods (e.g. BEAM [2] or SNPHarvester [3]), regression methods (e.g. Lasso Penalized Logistic Regression [4]) and machine learning methods (e.g. MegaSNPHunter [5]) or decision tree based methods. Classic methods such as Logistic Regression have been pointed as appropriate methods to consistently estimate the strength of association between a predictor and disease [6] [7]. However it has also been noticed that logistic regression has limited power for modelling high-order non-linear interactions that are likely important in the etiology of complex diseases [8]. Multivariate regression is not suited for problems with hundreds of thousands of variables. Intermediate strategies that lead to fast computation while preserving the spirit of multivariate regression have been explored [4] [9]. The lasso penalty is an effective device for continuous model selection, especially in problems where the number of variables far exceeds the number of observations. The problem is that even these intermediate regression strategies are not prepared for dealing with the interactions between a large number of variables. That is why these methods are usually combined with filtering strategies such as selecting SNPs with high marginals and building the regression model considering only the interactions between these high-marginal SNPs [4].

Classification methods are a natural choice to the problem of identifying subsets of variables that influence a specific phenotype. We used WEKA to test several classification methods such as Alternating Decision Trees [10], Voted Perceptrons [11] and Support Vector Machines [12]. We also tested the SMLR method (Sparse Multinomial Logistic Regression) [13].

One natural way to select SNPs that are relevant is to compute the mutual information between every SNPs and the target phenotype in order to find single SNP associations to the phenotype. A SNP A is associated with the phenotype with a user-defined threshold T if and only if $I(A;Y) > T$. $I(A;Y)$ is the mutual information and expresses the amount of information that is shared between A and Y. Fleuret published a fast binary feature selection technique (that we will call the Fleuret method) based on Conditional Mutual Information Maximization (CMIM) [14] that we also applied to our problem of detecting interacting SNPs with low marginals. This method is based on picking features that maximize their mutual information with the class to predict, conditional to any feature already picked. This method aims at selecting features that are both individually informative and two-by-two weakly dependant. The goal of this method is to select a small subset of features that carries as much information as possible. To measure the amount of information carried by the features, the Fleuret method uses Conditional Mutual Information. This information theory measure is based on the Entropy $H(U)$ of a random variable, which quantifies the uncertainty of U. The conditional entropy $H(U|V)$ quantifies the remaining uncertainty of U, when V is known. If U is a deterministic function of V then $H(U|V) = 0$. On the other hand if U and V are independent, knowing V does not tell anything about U and $H(U|V) = H(U)$. The Conditional Mutual Information is given $I(U;V|W) = H(U|W) - H(U|W,V)$. This value can be seen as the difference between the average remaining uncertainty of U when W is known and the same uncertainty when both W and V are known. If V and W carry the same information about U, the conditional mutual information is zero. On the other hand if V brings information about U that is not already contained in W, $I(U;V|W)$ is different from zero.

2 The Information Interaction Pairwise Search Method

Our method is based on the application of a systematic search over all possible pairs of SNPs. Exhaustive search came as the only possibility after noticing that greedy classifier based methods and greedy feature selection methods performed poorly. If no information is available to guide the choice of the first relevant locus, that means all the pairs of SNPs need to be considered, to decide which ones may be interacting in a way that is relevant.

The choice of information interaction as the metric to evaluate if a pair of SNPs is associated with the phenotype arose as a natural option. We define that a pair of SNPs (A,B) is associated with the phenotype with a user-defined threshold T if and only if $I(A;B;Y) > T$. I is the information interaction (or synergy) and expresses the amount of information bound up in a set of variables, beyond that which is present in any subset of those variables, $I(A;B;Y) = I(A;B|Y) - I(A;B) = I(A;Y|B) - I(A;Y) = I(B;Y|A) - I(B;Y)$.

If the information that SNP A provides about class Y is higher if we know SNP B than it is if we do not know SNP B, then this additional information is the interaction information or synergy between the two variables A and B with respect to class Y.

To achieve higher efficiency, we used bit-level operations. It was necessary to convert genotypes into binary variables. Each SNP is therefore coded with 2 bits which is enough since we use 3 values for the genotypes and 1 value for missing data. This encoding of the SNPs is not the same that was used in other works such as [15] or [16]. The reason for this is that the tool we developed uses bitwise operations that make the execution much faster. We adapted the source code from the Fleuret method in order to calculate the value of the information interaction over all possible pairs of SNPs. With this approach we benefited from the efficient calculations of conditional mutual information that was already developed.

The statistical significance of the results obtained was computed by applying permutation testing. A test statistic, which is computed from the dataset, is compared with the distribution of permutation values. These permutation values are computed similarly to the test statistic, but under a random rearrangement of the labels of the dataset. The distribution of the test statistic computed in the permutation tests was also used to define thresholds that corresponded to a given level of significance.

3 Artificial and Real World Datasets

We performed experiments using both artificial and real world datasets. To perform controlled tests of methods that aim at finding interactions with low marginals, we need to have a dataset that simulates this type of interactions. For this, we used the artificial datasets that were proposed by the authors of the SNPHarvester algorithm [3]. In particular, we used the datasets in which there are multiple disease loci without marginal effects, i.e., no individual SNP was correlated with the condition. If the method can detect these interactions in the artificial datasets, then it will presumably be able to find similar types of interactions in real datasets such as the Wellcome Trust Breast Cancer Dataset, if they are present.

The authors of the SNPHarvester algorithm used 60 different epistatic models as the base for the generation of datasets in which disease loci do not have marginal effects. These epistatic models, firstly used in [17] use different parameter values for parameters such as heritability (h^2) and minor allele frequency (MAF). Heritability ranges from 0.025 to 0.4 and MAF ranges from 0.2 to 0.4. 100 datasets were generated for each disease model, each one with 200 cases, 200 controls and 1000 SNPs. In these datasets there is only a pair of interacting SNPs in positions 1 and 10 from left to right. The datasets in which there are multiple disease loci without marginal effects try to simulate the expectation that there might exist multiple SNP-SNP interactions in the association studies. Eight hybrid models were used for the generation of the datasets. Each hybrid model is a mixture of five pure epistatic models with the same heritability and MAF. For example if a hybrid model HM1 consists of models $m_1 ... m_5$, the first interaction is based on model m_1, the second interaction based on model m_2 and so on. Thus there are five interactions in the HM that are simulated independently. 100 datasets were simulated for each hybrid model and each dataset contains 200 cases, 200 controls and 1000 SNPs. The 10 SNPs that belong to the 5 pairwise interactions are in positions (1,100), (201,300), (401,500), (601,700) and (801,900). Since these datasets with multiple loci are closer to what we may expect in reality, we will report the results of our methods in these datasets generated with hybrid models.

The Wellcome Trust Case Control Consortium (WTCCC) breast cancer dataset is a large real world dataset with 1045 cases, 1476 controls and 15436 SNPs. We used only the SNPs that were also used in the WTCCC original publication [18] after data cleaning.

4 Experimental Results and Discussion

The application of the classification methods to our selected artificial dataset did not obtain good results. None of the methods achieved an accuracy above 56.5%, which means that they were unable to idenfify with any accuracy the relevant SNPs. We also performed experiments using the feature selection methods in WEKA, but none of the variables selected by the feature selection methods corresponded to the disease loci. To understand why classifiers performed so poorly, we performed a study in which we started by training a classifier with the 10 disease loci, added one unrelated variable at each step and retrained the algorithm using the same parameters. The results was that the accuracy rapidly degraded with the addition of SNPs that are unrelated with the condition (see Figure 1, left.). Existing classification methods are not capable of identifying the relevant SNPs when many irrelevant ones are present.

We also used an implementation of the Fleuret Method available on the author's web page. This tool is able to perform feature selection with the Fleuret method and train/test a bayesian or a perceptron classifier. We were interested only in checking if the feature selection method was able to find the interacting SNPs of two artificial datasets, one with two interacting loci and the other with a total of 10 disease loci. However, the Fleuret method was not able to detect any of the interacting SNPs on these two datasets.

We finally applied the Information Interaction method proposed in this article to all possible pairs of SNPs. Using the values obtained in the permutation tests, it was possible to define a threshold that could distinguish, with high confidence, between true disease loci pairs and not associated SNPs. Figure 1, right, shows the results obtained with the Information Interaction method on one artificial dataset. The distinction between true disease loci and other SNP pairs is very explicit making it possible to detect the 5 pairwise interactions. We applied the method to all the datasets that were generated with different simulation parameters. We could then compare the results obtained with the results of SNPHarvester and BEAM methods. The results of the comparison are shown on Figure 2, left. It is clear that the Information Interaction method (IIM) outperforms both SNP Harvester and BEAM in terms of power. In Hybrid Model 8 our method has more problems in discovering the interactions because the signal is more hidden in the noise. However, even in those harder conditions, our method performed better than SNP Harvester and BEAM.This improved recall value takes place without a significant increase on the number of false positives. In a total of 800 datasets, SNP Harvester

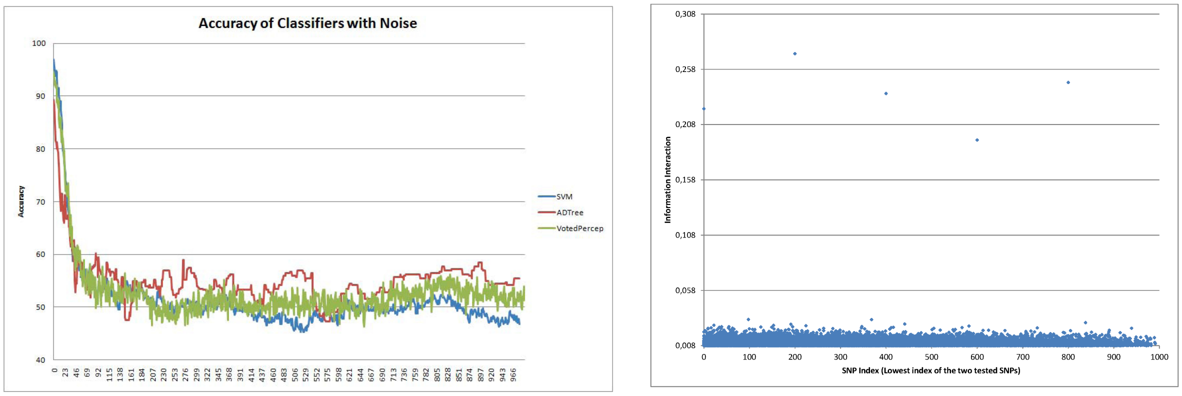

Figure 1. Left: Evolution of the accuracy of classification methods with the addition of irrelevant SNPs. Right: Information Interaction values for all pairs of SNPs (X_i, X_j) in artificial dataset. The 5 interacting pairs are very explicit, with information interaction values above 0.19.

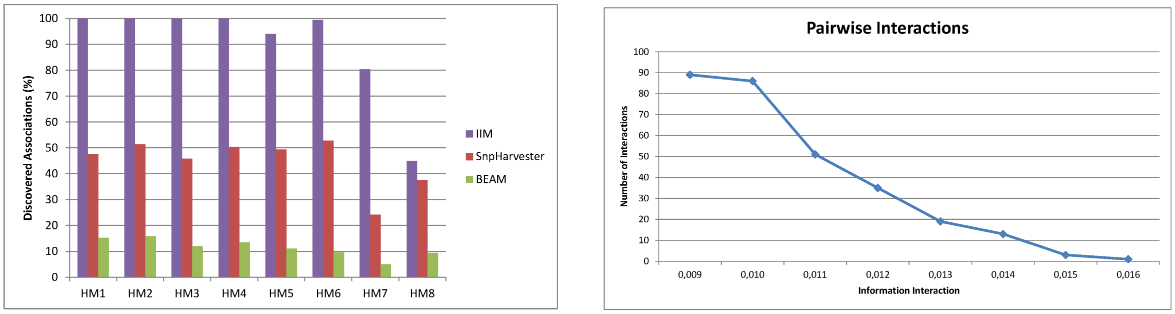

Figure 2. Left: Fraction of associations found by Information Interaction Search, SNPHarvester and BEAM. Right: Accumulated number of pairs of SNPs with Information Interaction higher than a given value, in the WTCCC dataset. Values of 0.009 correspond to a p-value of 10^{-3}.

discovers a total of 5 false positives while our method discovers 8.

After the pre-processing of the WTCCC Breast Cancer dataset, we applied the information interaction method. To determine the threshold T to use, we ran 1000 permutations tests. In each permutation test we applied the information interaction measure to all possible pairs of variables. At the end we selected the highest value h of information interaction between all pairs of SNPs discovered in the 1000 permutation tests. We then used a value of T greater than h so that each interaction discovered is statistically significant at the 0.001 level of confidence. In our permutation procedure h was found to be $h = 0.008875$. We then fixed $T = 0.009$. With this threshold, there are 89 pairs of SNPs that have an Information Interaction value higher than 0.009 and, therefore, a p-value of 10^{-3}. We also applied the mutual information method to the WTCCC Breast Cancer dataset. A total of 90 different SNPs were found to share information with the phenotype with a mutual information value above the threshold obtained for this method with a p-value of 10^{-3}. This p-value is automatically corrected for multiple hypothesis testing, given that we use a permutation based method to compute the statistical significance.

5 Conclusions and future work

We now summarize the main results of this study. In the artificial datasets no single SNP had any statistically significant association with the phenotype, the proposed method, based on information interaction, could find the relevant pairs of SNPs more frequently than any other state of the art methods such as SNPHarvester or BEAM. Since the method uses a very efficient computation mechanism, it can be applied to datasets with hundreds of thousands of SNPs using adequate computational resources. However, the permutation based assessment of statistical significance is computationally intensive, and needs to be replaced by a more efficient mechanism. We are currently working on methods that obtain the estimates of the p-values with a much smaller number of permutation tests. The application of the information interaction method to the WTCCC breast cancer dataset found 89 pairs of SNPs that, when considered together, share more information with the phenotype than when considered individually. A total of 49 different SNPs were involved in the 89 interactions. 48 of these 49 SNPs found with information interaction were also found with the mutual information method. The only exception was SNP rs660895 from MHC gene that was found with the information interaction method but not with the mutual information method. Our results could suggest that most epistatic interactions with relevance to breast cancer have moderate or high marginals.

References

1. Li M, Wang K, Grant S, Hakonarson H, Li C (2009) ATOM: a powerful gene-based association test by combining optimally weighted markers. Bioinformatics 25: 497–503.

2. Zhang Y, Liu J (2007) Bayesian inference of epistatic interactions in case-control studies. Nature genetics 39: 1167–1173.

3. Yang C, He Z, Wan X, Yang Q, Xue H, et al. (2009) SNPHarvester: a filtering-based approach for detecting epistatic interactions in genome-wide association studies. Bioinformatics 25: 504–511.

4. Wu T, Chen Y, Hastie T, Sobel E, Lange K (2009) Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics 25: 714–721.

5. Xiang W, Can Y, Qiang Y, Hong X, Nelson T, et al. (2009) MegaSNPHunter: a learning approach to detect disease predisposition SNPs and high level interactions in genome wide association study. BMC Bioinformatics 10: 13.

6. Heidema A, Boer J, Nagelkerke N, Mariman E, et al. (2006) The challenge for genetic epidemiologists: how to analyze large numbers of SNPs in relation to complex diseases. BMC Genetics 7: 23.

7. Nagelkerke N, Smits J, le Cessie S, van Houwelingen H (2005) Testing goodness-of-fit of the logistic regression model in case-control studies using sample reweighting. Statistics in Medicine 24: 121–130.

8. Moore J, Asselbergs F, Williams S (2010) Bioinformatics Challenges for Genome-Wide Association Studies. Bioinformatics 26: 445–455.

9. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67: 301–320.

10. Freund Y, Mason L (1999) The alternating decision tree learning algorithm. Proceedings of the 16th International Conference on Machine Learning : 124-133.

11. Freund Y, Schapire R (1999) Large margin classification using the perceptron algorithm. Machine learning 37: 277–296.

12. Vapnik V (1998) Statistical learning theory. Wiley-Interscience.

13. Krishnapuram B, Carin L, Figueiredo M, Hartemink A (2005) Learning sparse bayesian classifiers: multi-class formulation, fast algorithms, and generalization bounds. IEEE Trans Pattern Anal Machine Intell 27: 957–968.

14. Fleuret F (2004) Fast binary feature selection with conditional mutual information. The Journal of Machine Learning Research 5: 1531–1555.

15. McKinney B, Crowe Jr J, Guo J, Tian D (2009) Capturing the spectrum of interaction effects in genetic association studies by simulated evaporative cooling network analysis. PLoS Genetics 5: e1000432.

16. Moore J, Gilbert J, Tsai C, Chiang F, Holden T, et al. (2006) A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. Journal of theoretical biology 241: 252–261.

17. Velez D, White B, Motsinger A, Bush W, Ritchie M, et al. (2007) A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. Genetic Epidemiology 31: 306–315.

18. Burton P, Clayton D, Cardon L, Craddock N, Deloukas P, et al. (2007) Association scan of 14,500 nonsynonymous snps in four diseases identifies autoimmunity variants. Nature genetics 39: 1329–1337.