An associative classification based approach for detecting SNP-SNP interactions in high dimensional genome

Suneetha Uppu
Department of Computing
Curtin University
Perth, Australia
Suneetha.uppu@postgrad.curtin.edu.au

Aneesh Krishna
Department of Computing
Curtin University
Perth, Australia
A.Krishna@curtin.edu.au

Raj P.Gopalan
Department of Computing
Curtin University
Perth, Australia
R.Gopalan@curtin.edu.au

***Abstract*— There have been many studies that depict genotype-phenotype relationships by identifying genetic variants associated with a specific disease. Researchers focus more attention on interactions between SNPs that are strongly associated with disease in the absence of main effect. In this context, a number of machine learning and data mining tools are applied to identify the combinations of multi-locus SNPs in higher order data. However, none of the current models can identify useful SNP-SNP interactions for high dimensional genome data. Detecting these interactions is challenging due to bio-molecular complexities and computational limitations. The goal of this research was to implement associative classification and study its effectiveness for detecting the epistasis in balanced and imbalanced datasets. The proposed approach was evaluated for two locus epistasis interactions using simulated data. The datasets were generated for 5 different penetrance functions by varying heritability, minor allele frequency and sample size. In total, 23,400 datasets were generated and several experiments are conducted to identify the disease causal SNP interactions. The accuracy of classification by the proposed approach was compared with the previous approaches. Though associative classification showed only relatively small improvement in accuracy for balanced datasets, it outperformed existing approaches in higher order multi-locus interactions in imbalanced datasets.**

Keywords*— *Epistasis, multi-locus, associative classification, SNP-SNP interactions

I. INTRODUCTION

There are many complex diseases where no single factor on its own has been identified as a predictor for the cause or presence of the disease. Rather, a multitude of interacting factors seem to be at play, making these diseases complex. Many complex diseases are believed to be caused by the combination of genetic, environmental and lifestyle factors. Some examples include breast cancer, diabetes, cystic fibrosis, multiple sclerosis, obesity, asthma, sickle cell anaemia and Alzheimer's disease [1]. About 99% of the human genome is identical, with only 1% variations in the DNA sequences of the chromosome [2]. Single Nucleotide Polymorphisms (SNPs) are the variations in the sequence of DNA when a single nucleotide sequence is altered. It has been estimated that about 12 million SNPs occur along the 3-billion-base human genome [3]. The consequences of SNPs may influence disease susceptibility or response to drugs or differences in physical appearance. A number of studies have focused on the role of SNPs and their associations in revealing the genetic epidemiology of disease susceptibility.

The conventional single locus SNP methods [4, 5] such as Chi-squared test or Fisher's exact test, are used to identify the SNPs which have high association with a particular disease. Highly ranked SNPs are considered to be highly associated with the disease. These parametric strategies ignored genetic and environmental factors involved in complex multi-collinearity with other loci. Hence, alternative methods were proposed which include logistic regression (LR)[6], penalized logistic regression [7], Monte Carlo logic regression [8], full Bayesian logic regression[9] and genetic programming for association studies [10]. The major challenges of these parametric methods are model specific and they are computationally complex due to the curse of dimensionality (that is, n way interactions require 2^n regression equations). Consequently, these limitations increased an intensive research towards data mining and machine learning approaches for identifying multi-locus SNP interactions in high dimensional data.

Multi-Dimensionality Reduction (MDR) [11-13] is an effective data reduction model focused on identifying interactions in the high dimensional search space using constructive induction. This non-parametric model missed a few disease causal interacting SNPs due to pooling too many multilocus genotype cells together. The model is restricted to a small number of SNPs. The model classifies each cell either as high risk or low risk without assessing the proportion of the risk level. The efficiency of the method is reduced in the presence of high genetic heterogeneity and phenocopy. Hence, several extended strategies are proposed by addressing these limitations. Some of the extended strategies of MDR are Odd Ratio based MDR (OR-MDR) [14], Generalized MDR (GMDR) [15], Model based MDR (MB-MDR) [16-18] and Robust MDR (RMDR) [19]. These approaches are successfully applied over genetic data [11, 14-16, 19].

978-1-4799-7502-0/14 $31.00 © 2014 IEEE
DOI 10.1109/BIBE.2014.29

Tree based classification algorithms such as decision trees [20] and random forests (RF) [21] are high dimensional predictive models which uncover the interactions between genes that do not have strong marginal effects. However, RF contradictorily requires a marginal effect in at least one of the SNP interacting pair. This reduces the power of identifying the interactions between SNPs using RF models. The extended RF strategies are EpiForest [22], Random Jungle (RJ) [23], RFCouple [24], SNPInterForest [25], Stratified sampling RF (SRF) and TRM [26]. These approaches are successfully applied and analysed over real genetic data [23, 25-28].

Support Vector Machine (SVM) [29] is a potentially powerful pattern matching supervised learning approach. The approach not only provides interpretable results but also differentiate cases and controls. However, it may not cope well with missing data and the efficiency is reduced in the presence of genetic heritability. Neural Networks (NN) [29] resemble a directed graph which performs powerful pattern recognition and classification. They are capable of handling large volumes of data by addressing genetic heterogeneity, polygenic inheritance, high phenocopy rates and incomplete penetrance. However, it gives spurious results in three dimensional models and their output binary trees are difficult to interpret.

Despite their limitations, the existing approaches identify the existence of major proportions of interacting genes at multi-locus. However, none of these models could expose SNPs at a locus which can have a stronger association with a disease and a weaker association for another disease. In a few cases, a SNP may not be directly associated with the disease, but may influence the nearest genes to be associated with the disease. Further, the accuracy of the current models is degraded in imbalanced datasets by increasing the classification errors. Hence, there is no single model which can uncover the complexity of genetic architecture by identifying disease causal SNPs and their interactions between genes.

In this paper, a new approach based on associative classification is implemented to identify the interactions more effectively than the existing methods. The proposed approach will classify the subjects by determining the complexity of interactions and their associations with the disease. The goal of this study is to evaluate the proposed approach on the simulated data by varying heritability, minor allele frequencies and case control ratio. The study identified SNP-SNP interactions of order 2 for both balanced and imbalanced datasets. Finally, the approach is validated in terms of accuracy and compared with previous methods under the same simulated scenarios.

Associative classification is briefly reviewed and then applied to the present problem in Section II - A. Data Simulation scenarios and Data Analysis are explained in Section II - B and Section II - C. Results and discussion are included in Section III. Section IV contains the conclusion and some directions for further work.

II. METHODS

A. Associative classifier

Associative Classifiers (AC) have been successfully implemented in data mining to build more efficient and accurate classifiers than traditional techniques [30]. The approach integrates association rule discovery and classification. In general, association rules generated from frequent itemsets are used to classify data based on the class labels. The steps involved in the Associative classification are: a) Identifying frequently occurring conjunctions of attribute – value pairs (frequent itemsets) in training dataset b) Generating class based association rules (CARs) from frequent itemsets, which satisfy confidence and support criteria c) Pruning and ranking these CARs to organise for the classification, and d) Classifying the test dataset into predefined class labels. Some of the proposed ACs are Classification Based on Association (CBA) [31], Classification based on Multiple Association rules (CMAR) [32], Classification based on Predictive Association Rules (CPAR) [33], and Live and Let Live (L^3) [34].

To formulate SNP interactions as an AC problem, let D be a relation of tuples, whose schema is represented by n distinct attributes SNP_1, SNP_2,….., SNP_n and a class attribute C. Let C be a finite set of class labels with case c_1 and control c_2 respectively, where, $c_i \in C$. The attributes are treated as categorical where the class labels are known in training data instances in D and the class labels are unknown in testing data instances. Each instance tuple in D is represented as $t_i = (v_{i1}, v_{i2}, \ldots, v_{in}, c_i)$ where v_{i1} is an item value for SNP_1, v_{i2} for SNP_2, etc and c_i is a class label. Association rule R is generated in the form of $X \rightarrow Y$ which matches a tuple $t \in D$ when $X \subseteq t$. X is the antecedent which represents interacting SNPs associated with the class label and Y the consequent which represents either case or control. Support and Confidence are the two parameters used to measure the quality of association rules. Support is the number of tuples in D containing $X \cup Y$ and confidence is the number of tuples matching $X \cup Y$ divided by the number of tuples containing X. The CARs are organised and ranked by computing support and confidence along with the rule cardinality (measure of number of elements of the rule). Redundant and noisy rules are discarded in the rule pruning phase that passes minimum support and confidence thresholds. Several pruning techniques (such as pessimistic error, database coverage, chi-square, redundant rule and lazy pruning) are adopted to reduce the size of AC. Hence, most significant and high quality rules are selected to form a more accurate and efficient classification model. These rules are used to classify the test data. Finally, the accuracy of classification for the dataset is calculated as the ratio of number of objects correctly classified to the total number of objects in the test data.

B. Data Simulation

The goal of this simulated study is to detect interactions between multi-locus SNPs using the AC

approach. Two simulated scenarios are considered to evaluate the accuracy of AC with the previous approaches in the absence of main effect.

Scenario I

In the first scenario, six two locus epistasis (gene-gene interactions) models with different penetrance values are simulated for 20 SNPs with two functional SNPs (P1 and P2) and 18 independent non-functional SNPs. Case-control datasets are simulated with 200 cases and 200 controls in accordance to Hardy-Weinberg proportions [13]. Figure 1 represents the overview of model dependent allele frequencies along with their penetrance tables. A simple model of two alleles (p and q) necessarily sums to unity. That is, p+q = 1 where p is minor allele frequency and q is the alternative allele frequency. Model 1 is based on nonlinear XOR function described by [35, 36] in which all high risk genotype combinations (AaBB, Aabb, AABb and aaBb) have a penetrance value of 0.1. Model 2 is described by [36, 37] in which high risk genotype combinations (AAbb, AaBb and aaBB) have penetrance values 0.1, 0.05 and 0.1 respectively. Other four models are described by [36] with Minor Allele Frequencies (MAFs) of 0.25, 0.25, 0.1 and 0.1 respectively. Cases and controls of 1:1, 1:2, 1:4, and 1:6 ratios are generated for 400 samples. 100 datasets are simulated for each model in order to evaluate the power of AC by estimating the number of times the approach successfully identified two functional SNPs. In total, 2,400 datasets are generated and analysed in the absence of main effect.

Fig.1: Penetrance functions and minor allele frequencies used to simulate case-control data exhibiting multilocus SNP-SNP interactions in absence of main effects [13].

Model 1, p=0.5, q=0.5

	BB	Bb	bb
AA	0	0.1	0
Aa	0.1	0	0.1
aa	0	0.1	0

Model 2, p=0.5, q=0.5

	BB	Bb	bb
AA	0	0	0.1
Aa	0	0.05	0
aa	0.1	0	0

Model 3, p=0.25, q=0.75

	BB	Bb	bb
AA	0.08	0.07	0.05
Aa	0.1	0	0.1
aa	0.3	0.1	0.04

Model 4, p=0.25, q=0.75

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

Model 5, p=0.1, q=0.9

	BB	Bb	bb
AA	0.7	0.05	0.02
Aa	0.05	0.09	0.01
aa	0.02	0.01	0.03

Model 6, p=0.1, q=0.9

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

Scenario II

In the second scenario, datasets are replicated as in the simulated study performed by Velez, D.R., [12] with 20 SNPs. Among these 2 SNPs are functional and 18 SNPs are non- functional. The two locus interaction models are generated from publicly available tool GAMETES [38]. The tool generates randomly pure and strict n-locus disease models with specified heritability, minor allele frequency and quantile population size. In this simulated scenario, two locus epistasis models are distributed across seven heritability (0.01, 0.025, 0.05, 0.1, 0.2, 0.3 and 0.4) and two different minor allele frequencies (0.2 and 0.4). Five models for each 14 heritability-allele frequency combinations are generated to develop 70 models in accordance to Hardy-Weinberg proportions. The penetrance tables are generated for these 70 models in the absence of main effect. One hundred datasets are generated for each model with sample size of 400. The case-control ratios of the samples are 1:1, 1:2, and 1:4. In total, 21,000 datasets are generated and analysed to identify the two way interactions in the absence of main effect.

C. Data Analysis

The datasets for both scenarios are analysed using the latest MDR software tool available from www.epistasis.org. The data is exhaustively evaluated for all the possible two locus interactions between SNPs using the naïve Bayes classifier. Balanced accuracy is estimated using 10 fold cross validation for both training and testing data. Finally, a best model with high testing accuracy and high cross validation consistency is selected. The power of MDR has been estimated by the number of times the functional SNPs are identified in 100 datasets of each model. The final results are statistically evaluated with a 1000 fold permutation test and whose p-values are compared with 0.05 in determining the significance of the findings. The datasets for both scenarios are analysed using the associative classifier. The accuracy of AC algorithms is analysed using 10 fold cross validation and the disease causal SNPs are identified. Among the many methods and software implementations that have been used to investigate the interactions between SNPs, the most prominent approaches for identifying genetic effects in the presence of interactions are MDR, RF, SVM and NN. Further, Naïve Bayes algorithm is also considered in this paper as it is a well-established machine learning method and has been successfully applied in analysing GWAS data. Both scenarios are analysed using RF, SVM, NN and Naïve Bayes algorithms. Ten Fold cross validation is performed to reduce the possibility of biased estimation due to the division of data.

III. Results and Discussion

Several experiments were performed over 23,400 datasets to evaluate the accuracy of AC over other approaches. The goal of this study is to determine whether AC is a better approach for identifying the higher order SNP interactions in the absence of main effect. The approach considers the ratio of cases and controls for each SNP combination at different loci. It generates statistically significant genotype combinatorial associations in terms of rules based on cases and controls. Predicting class labels of test objects from these rules retains higher accuracy in genetic combinations that contribute to a disease. Despite the increase in accuracy, the approach will still reduce the false positive error by permutation testing under the null hypothesis. The results have been obtained on two simulated scenarios to identify complex associations between genotype and phenotype.

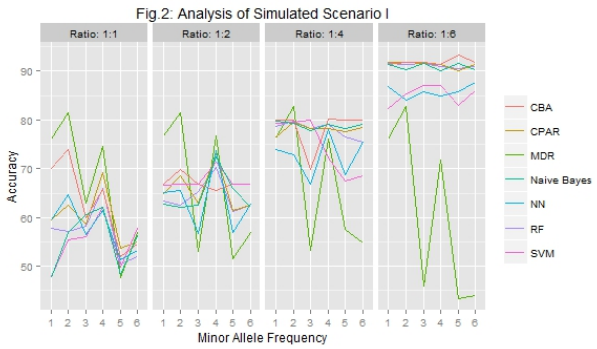

Fig.2: Analysis of Simulated Scenario I

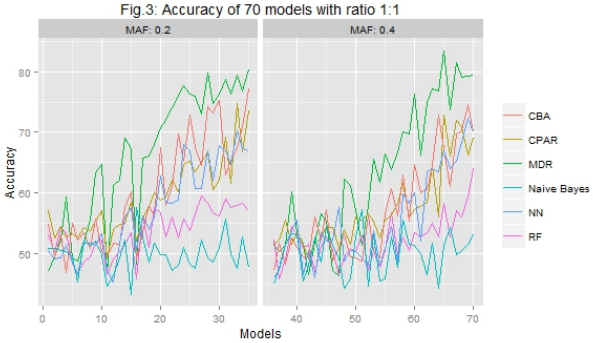

Fig.3: Accuracy of 70 models with ratio 1:1

Fig.4: Accuracy of 70 models with ratio 1:2

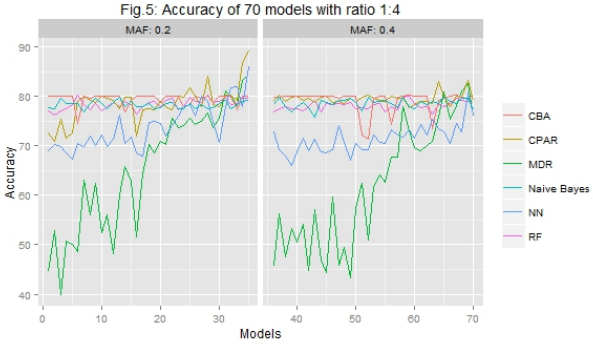

Fig.5: Accuracy of 70 models with ratio 1:4

In the first scenario, the approach is validated for both balanced and imbalanced datasets. Figure 2 shows the accuracy of AC over MDR, RF, SVM, NN and Naïve Bayes classifiers for the simulated scenario I with 400 samples of case-control ratios 1:1,1:2,1:4 and 1:6. On an average of 100 datasets for each model, MDR performed well for all 6 models in 1:1 ratio. However, AC performed better when allele frequencies are 0.1 and 0.9 exhibiting interactions in the absence of main effect. On average, in 1:2 ratio, there is a rise of 13% of accuracy over other algorithms for model 3, 5 and 6. It is observed that both for balanced and imbalanced data, AC is more accurate when the allele frequencies are 0.1 and 0.9. The accuracy of AC for 1:4 ratio is almost 16% higher than other existing approaches. It is observed that the accuracy is slightly reduced by about 2% in model 2 where allele frequencies are equal. The accuracy of AC is much higher than MDR in 1:6 ratios of all 6 models. Accuracy of AC is about 50% higher than MDR when MAF values are 0.1 and 0.25.

The results of the second scenario of simulation demonstrated that the AC performed well across a wide range of SNP-SNP interaction models. Figure 3 illustrates accuracy of AC over other approaches for 1: 1 ratio with MAF 0.2 and 0.4. MDR predominantly outperformed AC and other approaches. However, AC is more accurate than other approaches by up to 10% for allele frequencies 0.2 and 0.8 with heritability of 0.01. It also performed well when there is no genetic influence over the phenotype. MDR performed significantly better in balanced data compared to other methods when MAF is equal to 0.4. However, accuracy of AC improved up to 4% when heritability is 0.025. It also significantly performed better than other approaches when there is no genetic influence over the phenotype. Figure 4 graphically represents accuracy of AC for 1:2 ratio of sample size 400 with MAF 0.2 and 0.4 respectively. For average MAF values of 0.2 and 0.4, the accuracy of AC is higher by up to 14% and 16% respectively compared to MDR for heritability values of 0.01, 0.025, 0.05 and 0.1. It is also observed that, Naive Bayes' algorithm performed significantly better than MDR. However, on average AC was more accurate than Naive Bayes' for MAF values of 0.2 and 0.4 by up to 3% and 5% respectively. AC had the same accuracy as MDR for heritability 0.2, 0.3 and 0.4 for both MAF values (0.2 and 0.4). Figure 5 illustrates accuracy of AC for 1:4 ratios with MAF 0.2 and 0.4 respectively. AC predominantly outperformed in all 70 models compared with other existing approaches. These results demonstrate that the power of AC increases in imbalanced data with higher proportions of controls than cases.

IV. CONCLUSION

In this paper, association based classification approach was implemented for detecting interactions in balanced and unbalanced data. The approach was evaluated for two locus interactions using simulated data. The approach performed significantly better than the existing approaches in imbalanced data. However, the experimental results showed only small improvement in accuracy for balanced data. Further studies will investigate the performance of AC over three-way to ten-way genotype interactions and how these contribute to associated phenotype. The approach will be further applied to real data to confirm the success rate of identifying the interactions between SNPs in high dimensional genome.

Further, the empirical power of the approach will be determined in the presence of genotyping error, missing data, phenocopy and genetic heterogeneity.

REFERENCES

[1] D. J. Hunter, "Gene–environment interactions in human diseases," *Nature Reviews Genetics,* vol. 6, pp. 287-298, 2005.

[2] K. C. Koenen, "Genetics of posttraumatic stress disorder: review and recommendations for future studies," *Journal of traumatic stress,* vol. 20, pp. 737-750, 2007.

[3] S. F. Sheet, "Human genome project," US Department of Energy genome Program's biological and environmental research information system (BERIS).(Cited on 2010 July 28) available from: http://www.ornl.gov/sci/techresources/Human_Genome/.

[4] H. J. Cordell, "Detecting gene–gene interactions that underlie human diseases," *Nature Reviews Genetics,* vol. 10, pp. 392-404, 2009.

[5] J. H. Moore, F. W. Asselbergs, and S. M. Williams, "Bioinformatics challenges for genome-wide association studies," *Bioinformatics,* vol. 26, pp. 445-455, 2010.

[6] C. C. Chen, H. Schwender, J. Keith, R. Nunkesser, K. Mengersen, and P. Macrossan, "Methods for identifying SNP interactions: a review on variations of Logic Regression, Random Forest and Bayesian logistic regression," *Computational Biology and Bioinformatics, IEEE/ACM Transactions on,* vol. 8, pp. 1580-1591, 2011.

[7] M. Y. Park and T. Hastie, "Penalized logistic regression for detecting gene interactions," *Biostatistics,* vol. 9, pp. 30-50, 2008.

[8] C. Kooperberg and I. Ruczinski, "Identifying interacting SNPs using Monte Carlo logic regression," *Genetic epidemiology,* vol. 28, pp. 157-170, 2005.

[9] A. Fritsch and K. Ickstadt, "Comparing logic regression based methods for identifying SNP interactions," in *Bioinformatics research and development*, ed: Springer, 2007, pp. 90-103.

[10] R. Nunkesser, T. Bernholt, H. Schwender, K. Ickstadt, and I. Wegener, "Detecting high-order interactions of single nucleotide polymorphisms using genetic programming," *Bioinformatics,* vol. 23, pp. 3280-3288, 2007.

[11] M. D. Ritchie, L. W. Hahn, N. Roodi, L. R. Bailey, W. D. Dupont, F. F. Parl, and J. H. Moore, "Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer," *The American Journal of Human Genetics,* vol. 69, pp. 138-147, 2001.

[12] D. R. Velez, B. C. White, A. A. Motsinger, W. S. Bush, M. D. Ritchie, S. M. Williams, and J. H. Moore, "A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction," *Genetic epidemiology,* vol. 31, pp. 306-315, 2007.

[13] M. D. Ritchie, L. W. Hahn, and J. 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,* vol. 24, pp. 150-157, 2003.

[14] Y. Chung, S. Y. Lee, R. C. Elston, and T. Park, "Odds ratio based multifactor-dimensionality reduction method for detecting gene–gene interactions," *Bioinformatics,* vol. 23, pp. 71-76, 2007.

[15] X.-Y. Lou, G.-B. Chen, L. Yan, J. Z. Ma, J. Zhu, R. C. Elston, and M. D. Li, "A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence," *The American Journal of Human Genetics,* vol. 80, pp. 1125-1137, 2007.

[16] M. Calle, V. Urrea, G. Vellalta, N. Malats, and K. Steen, "Improving strategies for detecting genetic patterns of disease susceptibility in association studies," *Statistics in medicine,* vol. 27, pp. 6532-6546, 2008.

[17] M. L. Calle, V. Urrea, N. Malats i Riera, and K. Van Steen, "MB-MDR: model-based multifactor dimensionality reduction for detecting interactions in high-dimensional genomic data," 2008.

[18] T. Cattaert, M. L. Calle, S. M. Dudek, J. M. Mahachie John, F. Van Lishout, V. Urrea, M. D. Ritchie, and K. Van Steen, "Model-Based Multifactor Dimensionality Reduction for detecting epistasis in case–control data in the presence of noise," *Annals of human genetics,* vol. 75, pp. 78-89, 2011.

[19] J. Gui, A. S. Andrew, P. Andrews, H. M. Nelson, K. T. Kelsey, M. R. Karagas, and J. H. Moore, "A robust multifactor dimensionality reduction method for detecting gene–gene interactions with application to the genetic analysis of bladder cancer susceptibility," *Annals of human genetics,* vol. 75, pp. 20-28, 2011.

[20] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques*: Morgan kaufmann, 2006.

[21] L. Breiman, "Random forests," *Machine learning,* vol. 45, pp. 5-32, 2001.

[22] R. Jiang, W. Tang, X. Wu, and W. Fu, "A random forest approach to the detection of epistatic interactions in case-control studies," *BMC bioinformatics,* vol. 10, p. S65, 2009.

[23] D. F. Schwarz, I. R. König, and A. Ziegler, "On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data," *Bioinformatics,* vol. 26, pp. 1752-1758, 2010.

[24] L. De Lobel, P. Geurts, G. Baele, F. Castro-Giner, M. Kogevinas, and K. Van Steen, "A screening methodology based on random forests to improve the detection of gene–gene interactions," *European journal of human genetics,* vol. 18, pp. 1127-1132, 2010.

[25] M. Yoshida and A. Koike, "SNPInterForest: a new method for detecting epistatic interactions," *BMC bioinformatics,* vol. 12, p. 469, 2011.

[26] H. Y. Lin, Y. Ann Chen, Y. Y. Tsai, X. Qu, T. S. Tseng, and J. Y. Park, "TRM: A Powerful Two-Stage Machine Learning Approach for Identifying SNP-SNP Interactions," *Annals of human genetics,* vol. 76, pp. 53-62, 2012.

[27] Q. Wu, Y. Ye, Y. Liu, and M. K. Ng, "SNP selection and classification of genome-wide SNP data using stratified sampling random forests," *NanoBioscience, IEEE Transactions on,* vol. 11, pp. 216-227, 2012.

[28] A. Bureau, J. Dupuis, K. Falls, K. L. Lunetta, B. Hayward, T. P. Keith, and P. Van Eerdewegh, "Identifying SNPs predictive of phenotype using random forests," *Genetic epidemiology,* vol. 28, pp. 171-182, 2005.

[29] R. Upstill-Goddard, D. Eccles, J. Fliege, and A. Collins, "Machine learning approaches for the discovery of gene–gene interactions in disease data," *Briefings in bioinformatics,* vol. 14, pp. 251-260, 2013.

[30] F. Thabtah, "A review of associative classification mining," *The Knowledge Engineering Review,* vol. 22, pp. 37-65, 2007.

[31] B. L. W. H. Y. Ma, "Integrating classification and association rule mining," in *Proceedings of the 4th*, 1998.

[32] W. Li, J. Han, and J. Pei, "CMAR: Accurate and efficient classification based on multiple class-association rules," in *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, 2001, pp. 369-376.

[33] J. Han, "CPAR: Classification based on predictive association rules," in *Proceedings of the third SIAM international conference on data mining*, 2003, pp. 331-335.

[34] E. Baralis, S. Chiusano, and P. Garza, "A lazy approach to associative classification," *Knowledge and Data Engineering, IEEE Transactions on,* vol. 20, pp. 156-171, 2008.

[35] W. Li and J. Reich, "A complete enumeration and classification of two-locus disease models," *Human heredity,* vol. 50, pp. 334-349, 2000.

[36] J. H. Moore, L. W. Hahn, M. D. Ritchie, T. A. Thornton, and B. C. White, "Application of genetic algorithms to the discovery of complex models for simulation studies in human genetics," in *Proceedings of the Genetic and Evolutionary Computation Conference/GECCO. Genetic and Evolutionary Computation Conference*, 2002, p. 1150.

[37] W. N. Frankel and N. J. Schork, "Who's afraid of epistasis?," *Nature genetics,* vol. 14, pp. 371-373, 1996.

[38] R. J. Urbanowicz, J. Kiralis, N. A. Sinnott-Armstrong, T. Heberling, J. M. Fisher, and J. H. Moore, "GAMETES: a fast, direct algorithm for generating pure, strict, epistatic models with random architectures," *BioData mining,* vol. 5, pp. 1-14, 2012.