Proceedings of the 2007 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2007)

Gene-Gene Interaction Tests Using SVM and Neural Network Modeling

N. Matchenko-Shimko, M.P. Dubé
Université de Montréal and the Montreal Heart Institute
5000 Belanger- Research Centre
Montreal, QC, H1T1C8 Canada

Abstract - **Artificial neural networks (ANN) and support vector machine (SVM) modeling offer promise in the analysis of genotype-phenotype correlation in genetic association studies. In particular, we are interested in studying single nucleotide polymorphisms (SNPs) as genetic markers as predictors of a dichotomous disease outcome. The problem we are investigating is that of gene-gene and gene-environment interactions as determinants of the expression of complex diseases. This study builds on our previous work for a single gene testing procedure developed and presented earlier [1].**

As for single SNPs pre-selection [1], we rely on ANN sensitivity analysis algorithms to detect potential pairs of interacting SNPs associated with the disease outcome. The statistical test for SNP interaction is computed using a bootstrap technique and is based on the measure of the predictive significance of two SNPs from the change in the ANN error function (SVM regression error) when these two SNPs are removed from the ANN or SVM genotype-phenotype models. To investigate the power to detect and test gene-gene interactions we simulated genotypes including two interacting loci with low marginal effects, incomplete penetrance and phenocopies according to three different models of interaction.

I. INTRODUCTION

The main focus of genetic-epidemiological studies has shifted towards the analsysis of complex diseases such as cancers, various nutritional and metabolic disorders (obesity, diabetes), cardiovascular diseases (coronary diseases, hypertension), cerebrovascular diseases (atherosclerosis, stroke) and psychiatric illnesses (schizophrenia). Unlike rare genetic disorders which are characterized by a single gene, common diseases have a multifactorial nature [2-3]. They arise from the combined action of many genes, environmental factors, and risk-conferring behaviors, all of which may be associated with disease risk primarily through nonlinear interactions [4-6].

As these multifactorial diseases are influenced by multiple incompletely penetrant variants, each contributing gene by itself is expected to have a small effect size. The absolute increase in prevalence associated with each gene's risk factor is expected to be correspondingly small [7]. Moreover, the effect of any one factor (genetic or environmental) may be obscured or confounded by the effects of others [2, 8 - 10].

Gene-gene interaction models can exhibit minimal independent main effects, but produce an association with disease primarily through interactions [8, 11-13]. In this case the risk associated with a genotype at one susceptibility locus is dependent on a genotype at another susceptibility locus and thus an interaction between genetics factors is characterized by a dependant effect. There are several possible epistatic models which can occur in complex diseases. A synergistic epistatic effect occurs when the combined risk is greater than would be expected if susceptibility loci were independent. A permissive epistatic interaction occurs when only the allele carriers at susceptibility loci show an increased risk for disease. In an antagonistic epistatic effect, the combined risk is lower than expected. A balanced epistatic effect is a special case of antagonistic effects where the risk for risk-allele carriers at both loci is actually lower than for risk allele carriers at only one locus [14].

Association studies using case-control methodology and high-density maps of SNPs are now recognized as being essential for the identification of genetic variants that influence susceptibility to common disease [3,15-17], as they are intrinsically more powerful than linkage analysis in detecting weak genetic effects. Biallelic SNPs have many advantages over other types of polymorphism in the genetic dissection of complex traits and diseases and for population-based gene identification studies due to their abundance in the genome, positions throughout the genome in coding/non-coding DNA areas, relative stability (low mutation rate) and easy and efficient genotyping [2].

The primary focus of current statistical approaches in the dissection of complex diseases is to detect individual factors with non-negligible main effects in traditional single SNP analysis. However, the ability to detect any single locus is a function of the relative risk of that locus alone [14], even if the overall disease risk can be modeled as the product of risks at several independent risk loci. If single-locus effects are small relative to the interaction effects, then the joint analysis of multiple loci explicitly allowing for interaction between loci effects is a clear improvement over traditional approaches.

The determination of a single best strategy for the detection of susceptibility loci in a multilocus model is complicated due to the unknown number of interacting loci, the form of interaction causing the disease [18] and the high marker density or the number of potential interactions. For example, an increase in the number of genetic markers (such as hundreds of thousands of SNPs in a genome-wide scan) exponentially increases the number of possible interaction terms for SNP pairs [11, 13] and an exhaustive search is a computational burden.

There are several approaches to reducing the number of interactions to analyze. One of them is to select candidate SNPs to study and to limit the analysis to biologically plausible interactions between genes in related pathways

(involved in a biological interaction, found in the same pathway or involved in the same regulatory network) [14], or to the markers in regions containing modest linkage signals. Another approach is to restrict interaction analysis to SNPs with detectable main effects from single locus studies.

The complicated nature of gene-gene interaction requires the development of higher-level, flexible, model-free, noise-robust tools to account for the possibly non-linear interactions between complex disease factors and outcomes, as well as for the genetic heterogeneity, high phenocopy rate and incomplete penetrance of the genetic factors. In a previous article [1] we used artificial neural network and support vector machine techniques to test the association between single nucleotide polymorphisms (SNPs) and a dichotomous disease outcome in a population-based case-control association study. We have selected these "black-box" models for their high flexibility in modeling non-linear functions between input and output variables, their outstanding pattern recognition capability, the significant discrimination power, signal filtering and high classification performance. In this study we present an extension to both the ANN and SVM regression models with applications to the pre-selection of SNP-SNP combinations and to test the significance of potential interactions. Our artificial neural network (ANN) model of genotype-phenotype correlation is represented by a fully connected 3-layered feedforward neural network with input nodes corresponding to the number of genotyped SNPs, a hidden layer of nodes and a single output unit, corresponding to the affection status. In the SVM model SNPs are represented by the SVM input patterns and the affection status with SVM targets or labels. The selection of ANN characteristics (number of hidden layer, hidden layer nodes, learning parameters) is detailed in [1].

For the ANN technique, we used an evaluation procedure that measures the predictive significance of two interacting SNPs, based on the change in the error function when the two inputs corresponding to these tested SNPs are removed from the network. Two ANNs, one with all inputs and the other with 2 tested inputs removed are run in parallel, and the change in error is calculated as a function of the relative out-of-sample performance of these two networks. Inference is performed via bootstrapping (resampling with replacement). The measure of statistical significance for the 2-SNPs interaction test in the SVM technique is based on the change in the SVM regression error when these 2 SNPs (SVM variables) are removed from the model. A single SNP test is performed in parallel to the interaction test to control for the main effects of individual SNPs.

The proposed testing procedure is free of genetic modeling as it relies on the predictive importance of SNPs on the disease outcome and doesn't incorporate any specific type of gene-gene interaction. It is expected to perform equally well for a permissive and antagonistic epistatic interaction, as long as the interaction signal is sufficiently strong to be detected. The same holds true for the algorithms used in SNPs pre-selection, as they are based on the same ANN technique.

II. DATA

We used three different SNP-SNP interaction models with small marginal effects at each locus, as presented by Marchini et al. [18]: 1) an additive model, with effects within and between loci; 2) a complementary gene model, of explicit interaction with the odds of disease at baseline value for both loci, and multiplicative effect between and within loci when both loci have at least one disease-associated allele; 3) a complementary model of explicit interaction, similar to model 2, with the odds of disease at baseline value for both loci and a threshold of disease effect when both loci have at least one disease-associated allele. The models are depicted in Fig.1 in terms of the odds of disease for each combination of genotypes at two loci, parameterized baseline effect α, and genotypic effect θ. In model 1, the effects of loci A and B are reflected in θ_1 and θ_2 correspondingly, while in models 2 and 3 both loci have the same effect size (i.e., $\theta_1 = \theta_2 = \theta$).

We set a disease prevalence to $p = 0.01$ and a marginal parameter λ to the range 0.2 -1.0, corresponding to a single locus marginal effect – relative risk of 1.2 – 2.0 (suggested by empirical studies in humans) [15, 18]. The genotypic effect θ for each model was calculated based on the analytical formulas developed in [18] and expressing a marginal parameter of a locus as a function of the genotypic effects θ, the baseline value α and the allele frequency of the other locus. Fixing the last two parameters and varying the genotypic effect parameter allowed calculating the marginal parameters for the corresponding genotypic effects. Working backwards, the same formulas allowed determination of the magnitude of the interaction effects, corresponding to a desirable single locus marginal parameter.

We used the SNaP software [19] to simulate datasets of 100 marker genotypes with the corresponding affection status in equal number n of cases and controls with a single pair of unobserved causative loci, each of which is in linkage disequilibrium (LD) with one of the genotyped markers. n ranges from 1000 to 4000, disease allele frequency (DAF) and disease associated marker allele frequency (MAF) match and vary from 0.05 (rare) to 0.25 (common), the linkage disequilibrium measure r^2 was varied as 0.4, 0.6, 0.8, 1.0, and the marginal parameter λ as 0.2, 0.5 and 1.0 (~ odds ratio = 1.2, 1.5, 2.0). The SNaP output datasets were transformed to input formats of our ANN software [1] and the SVM^{light} software [20]. Coded genotypes were linearly scaled to the range [0, 1]. Phenotypes were labeled as "-1" for controls and "1" for cases for SNP-SNP interaction tests based on the SVM technique, and rescaled to the [0, 1] range for the SNP-SNP interaction tests using the ANN approach. Samples were shuffled and datasets were split into training (50%), testing (25%) and validation (25%) sets for ANN and into training (50%) and testing (50%) sets for the SVM approach. Sample shuffling was used to randomize the distribution of association patterns to all sets, which can be biased by uneven proportion of cases/controls. Same post-shuffled data set was used with both techniques; training was performed on the same subset of data and finally ANN error-function and SVM regression error were calculated on the entire data set.

Multiplicative within and between loci

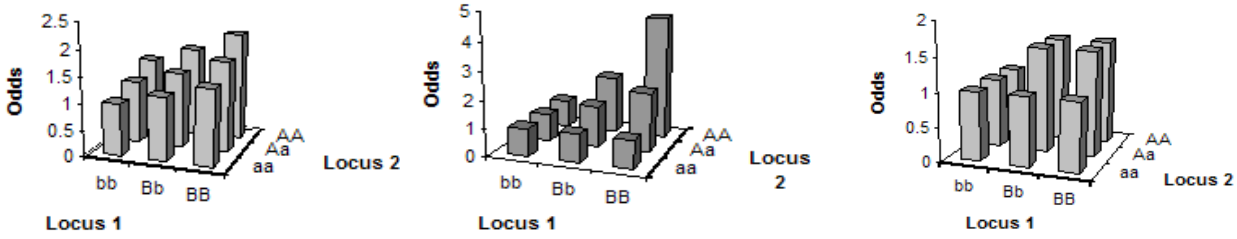
	BB	Bb	bb
AA	$\alpha(1+\theta_1)^2(1+\theta_2)^2$	$\alpha(1+\theta_1)^2(1+\theta_2)$	$\alpha(1+\theta_1)^2$
Aa	$\alpha(1+\theta_1)(1+\theta_2)^2$	$\alpha(1+\theta_1)(1+\theta_2)$	$\alpha(1+\theta_1)$
aa	$\alpha(1+\theta_2)^2$	$\alpha(1+\theta_2)$	α

Two-locus interaction, multiplicative effect

	BB	Bb	bb
AA	$\alpha(1+\theta)^4$	$\alpha(1+\theta)^2$	α
Aa	$\alpha(1+\theta)^2$	$\alpha(1+\theta)$	α
aa	α	A	α

Two-locus interaction, threshold effect

	BB	Bb	Bb
AA	$\alpha(1+\theta)$	$\alpha(1+\theta)$	A
Aa	$\alpha(1+\theta)$	$\alpha(1+\theta)$	A
aa	α	α	A



model 1 model 2 model 3

Fig.1. Two locus disease models as described in [18]. The odds of disease for each genotype pair at two loci are presented on the top and the illustrations of the genotypic risks for the corresponding models are presented below. Both loci have equal effects, disease allele frequency DAF = 0.25 for both loci, marginal parameter $\lambda = 0.2$, the derived genotypic effects θ, corresponding to a marginal parameter $\lambda = 0.2$, for each model are set to 0.2, 0.4 and 0.53 respectively and the odds baseline value $\alpha = 1.0$.

III. The Search for Significant Gene-Gene Interactions

The pre-selection of SNP-SNP (or SNP-environment) interacting pairs associated with a high risk of disease was conducted according to two strategies: i) an interaction-based strategy, or search over all possible pairs of factors and ii) a 2-stage approach comprising the identification of the set of most significant single-factor effects at the first stage, and evaluating all possible two-way interactions among the selected factors at the second stage.

The interaction-based strategy involves an exhaustive search which is time consuming but was computationally feasible for 100 SNPs. It is based on the ANN input sensitivity analysis using the input clamping technique [21, 22]. The clamping technique applied to 2 inputs consists of comparing the error made by the network with the original pattern to the error made when selected inputs are clamped to fixed values (in general the average value specific to each of tested inputs) for all patterns. The greater increase in the error corresponds to the greater importance of a tested input pair.

The first step of the 2-stage approach is a single factor search including the analytical input sensitivity analysis [1, 22-24] based on the calculation of partial derivatives of the network output with respect to the network inputs. The computational procedure is very fast and allows selecting a subset of the most significant single factors. The second stage is the same as the interaction-based strategy applied only to the factors selected during the first stage. This combined strategy decreases the computational burden via "dimension reduction" of interaction-based search, as fewer loci are included into datasets.

The power of a search strategy depends, besides other common factors, on the ratio of individual factor effects and the 2-factor interaction effects. If one of the interacting factors doesn't reach a significant threshold (i.e. the marginal effect is low) but the combination of both loci is statistically significant, then the interaction-based approach outperforms the combined strategy. As all three interaction models assume some level of marginal effect, we started with the 2-stage strategy for efficiency. In cases where the 2-stage strategy failed to detect at least one locus, we applied the interaction-based approach and compared the performance of both strategies.

IV. Testing Procedure to Detect Interacting Loci

To test for the significance of SNP-SNP interaction as determinants of disease outcome, we followed the testing procedure applied to a single SNP factor as previously detailed in [1], by extending it to the case of two loci. For the SVM testing procedure, we built two SVM models. Input vectors $\mathbf{X^1}$ of the first SVM have a dimension equal to the number of SNPs in the sample, while input vectors $\mathbf{X^2}$ of the second SVM have the pair of tested SNPs excluded from all samples. The null hypothesis is formulated in the following way: if a pair of tested SNPs is not important in the context of the disease model, then removing it from the model will not increase the error of the SVM regression. Therefore, the regression error function of the second SVM with the irrelevant SNPs excluded will be the same or smaller than the error function of the first SVM model with all inputs included:

$$H_0 : \varphi_1(\mathbf{X^1}) \geq \varphi_2(\mathbf{X^2}), \quad (1)$$

where $\varphi_1(\mathbf{X^1})$ and $\varphi_2(\mathbf{X^2})$ are the regression error functions of the first and the second SVM models correspondingly. The quantity

$$m_0 = \varphi_1(\mathbf{X^1}) - \varphi_2(\mathbf{X^2}) \quad (2)$$

is used as a statistic for the purpose of testing the relevance of a particular pair of inputs. With the help of the bootstrap technique, we tested the validity of the null hypothesis.

To avoid false positive results of gene-gene interaction detection we used 2 additional SVM models of the same configuration as the models described above, each one with only one input excluded: the first SVM model doesn't include the first input from the tested pair and the second one excludes the second input belonging to the same tested pair. Correspondingly we have 2 other null hypotheses for testing each input separately [1]:

$$H_{01} : \varphi_1(\mathbf{X}^1) \geq \varphi_3(\mathbf{X^3}) \quad (3)$$

$$H_{02} : \varphi_1(\mathbf{X^1}) \geq \varphi_4(\mathbf{X^4}), \quad (4)$$

where $\varphi_3(\mathbf{X^3})$ is the regression error function of the SVM model with all inputs except the first input from the tested pair of inputs, and $\varphi_4(\mathbf{X^4})$ is the regression error function of the SVM model with all inputs except the second input belonging to the pair of inputs in question.

The corresponding statistics for testing the relevance of these single inputs are:

$$m_1 = \varphi_1(\mathbf{X^1}) - \varphi_3(\mathbf{X^3}) \quad (5)$$

$$m_2 = \varphi_1(\mathbf{X^1}) - \varphi_4(\mathbf{X^4}) \quad (6)$$

Using the bootstrap technique, we tested the validity of the null hypothesis (1) along with supporting null hypotheses (3) and (4).

The procedure is realized in the following steps:

1) Using the original sample, train all four SVM models on the training dataset (50% of the samples size).

2) Calculate the original statistics (2), (5) and (6) on the entire dataset, including the training and testing datasets.

3) Draw a sample Z_T^* from $\{Z_1^*,...,Z_n^*\}$ with replacement from the original dataset and repeat steps 1 and 2 for Z_T^*.

4) Compute the bootstrap statistics (7-9) as in step 2:

$$m_0^* = \varphi_1(\mathbf{X^{*1}}) - \varphi_2(\mathbf{X^{*2}}) \quad (7)$$

$$m_1^* = \varphi_1(\mathbf{X^{*1}}) - \varphi_3(\mathbf{X^{*3}}) \quad (8)$$

$$m_2^* = \varphi_1(\mathbf{X^{*1}}) - \varphi_4(\mathbf{X^{*4}}) \quad (9)$$

5) Replicate steps 3 and 4 *N* times (usually 100 or 1000).

6) Calculate the proportion of the positive bootstrap statistics m_0^*, m_1^* and m_2^* created in step 4.

7) Reject the null hypothesis (1) if the original test statistics m_0 is < 0 and the proportion of positive bootstrap statistics m_0^* is < 0.05, and this proportion is smaller than the proportion of each positive bootstrap statistics m_1^* and m_2^*; otherwise, fail to reject.

We used the equivalent approach for the ANN-based tests with four optimized ANN networks corresponding to the four SVM models, three similar statistics expressing the difference in the error function of these networks and the bootstrap technique to validate the null hypotheses.

Gene-environment interaction testing is handled in a similar way by treating an environment variable as a locus. Overview of the SVM algorithm and the selection of the SVM model parameters used are described in Appendixes A and B correspondingly.

V. RESULTS

We investigated the power to detect interacting SNPs as predictors of disease outcome by using ANN and SVM modeling according to the following parameters: the type of SNP-SNP interaction model, the marginal effect size, the samples size, the frequency of the disease allele/environmental factor, and the extent of linkage disequilibrium (LD) between the unobserved causative locus and one of the genotyped markers.

First, we defined the sample sizes providing minimal power requirements for gene-gene interaction detection for a fixed genotype effect size and fixed marginal effects of two disease loci or marginal heterozygote odds ratios at both loci. We calculated that 4000 cases and 4000 controls are required for low marginal effect sizes of 1.2; 2500 cases and 2500 controls provide enough power to detect an interaction of genes with marginal effects of 1.5; and the number of samples can be reduced to 1000 cases and 1000 controls if the marginal effect increases to 2.0. The same sample sizes have been used for the pre-selection of SNP pairs. In general, the minimal dataset size is defined by the single genotype size and can be reduced by this single genotype size decrease [1].

Pre-selection

According to our results, both pre-selection strategies performed equally well across a variety of parameters and models, except for some datasets with i) low marginal effects of 1.2, or ii) the combination of low disease allele frequencies at the surrogate marker and low LD, with disease allele of marginal effect sizes of 1.5 and 2.0, corresponding to a reduced marker marginal effect below 1.5 and 2.0 (Table 1). Most simulated gene-gene interactions that were undetected with the combined 2-stage pre-selection strategy are not statistically significant and some of them did not pass pre-selection with both strategies. Some statistically significant SNP-SNP interactions with marginal effects of 1.2, that are not detected with the combined 2-stage pre-selection strategy were detected when using the 2-gene interaction-based strategy.

Gene-gene interaction testing

The results of the gene-gene interaction tests for the simulated datasets are presented in Fig. 2 for all three models: the additive model, and the two complementary models of explicit gene-gene interactions: a multiplicative model and a multiplicative with threshold model. The statistical power to detect gene-gene interaction using the ANN technique is plotted on the left, and the corresponding results for the SVM technique are presented on the right. DAF and MAF are set to be equal in our simulated data and we used them as interchangeable parameters.

The power to detect SNP-SNP interactions is strongly correlated with the marginal effect size of disease loci, the sample size, allele frequency of the disease loci and linkage

TABLE I

LIST OF DATASETS FOR WHICH THE TWO-STAGE PRE-SELECTION STRATEGY FAILED TO DETECT AT LEAST ONE SIMULATED CAUSATIVE LOCUS. BOTH DISEASE SNPS (OR SURROGATE MARKERS) WERE SIMULATED TO HAVE EQUAL MARGINAL EFFECTS, DISEASE ALLELE FREQUENCIES AND LD WITH THE DISEASE-ASSOCIATED SNP.

Disease model	n (cases)	Disease allele marginal effect (OR)	Disease allele frequency	r^2 (linkage disequilibrium)	Interaction based strategy	2-stage strategy	Statistical significance
2	4000	1.2	0.05	0.8	Y	N	No
	4000	1.2	0.05	1	Y	N	No
	4000	1.2	0.1	0.8	Y	N	Yes
	4000	1.2	0.15	0.8	Y	N	Yes
3	4000	1.2	0.05	0.8	N	N	No
	4000	1.2	0.05	1	N	N	No
	4000	1.2	0.1	0.8	Y	N	Yes
	4000	1.2	0.15	0.8	Y	N	Yes
1	2500	1.5	0.05	0.4	Y	N	No
	2500	1.5	0.05	0.6	Y	N	No
	2500	1.5	0.2	0.4	Y	N	No
2	2500	1.5	0.2	0.4	Y	N	No
2	1000	2	0.05	0.4	Y	N	No
3	1000	2	0.05	0.4	N	N	No
3	1000	2	0.05	0.6	Y	N	No

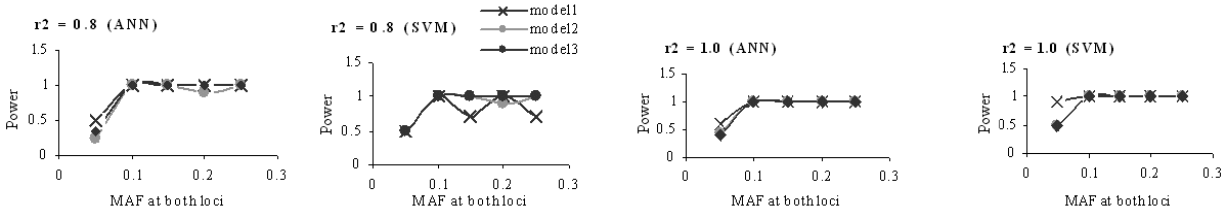


Fig. 2a. Power to detect gene-gene interaction using the ANN and SVM modeling for marginal effects of disease loci equally set to 1.2 and for a dataset of 4000 cases and 4000 controls, at a significance threshold of 0.05. Model1 is the additive model; model2 and model3 correspond to complementary gene models of explicit interaction, with multiplicative effects between and within loci when both loci have at least one disease-associated allele (model2), and a threshold of disease effect when both loci have at least one disease-associated allele (model3).

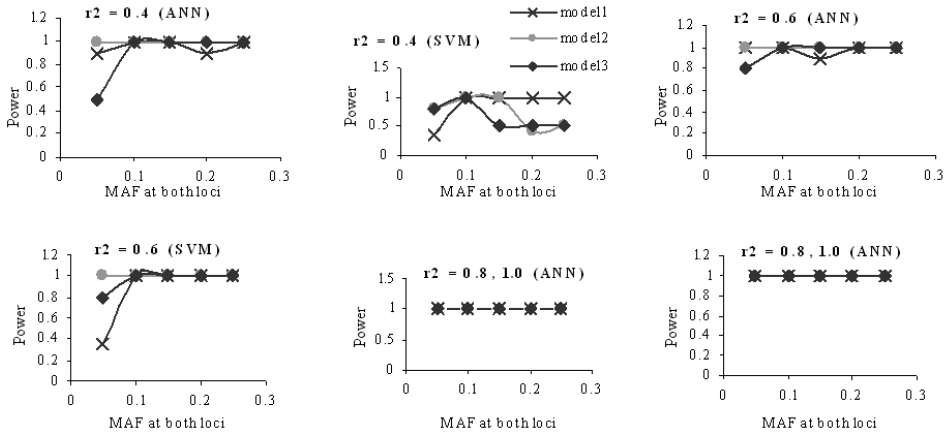
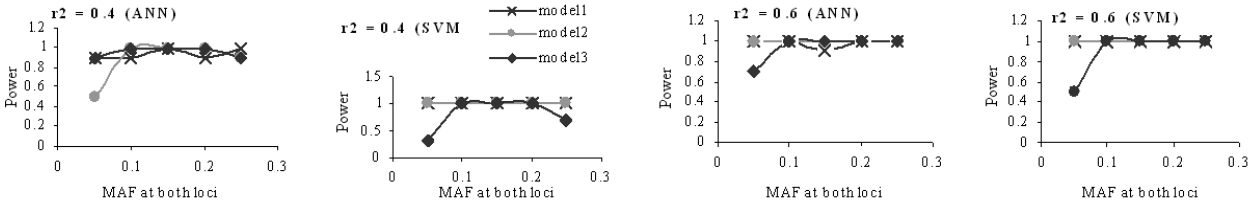


Fig. 2b. Power to detect gene-gene interactions by using the ANN and SVM techniques. The marginal effect is set to OR=1.5. The dataset has 2500 cases and 2500 controls. The significance threshold is set to 0.05.



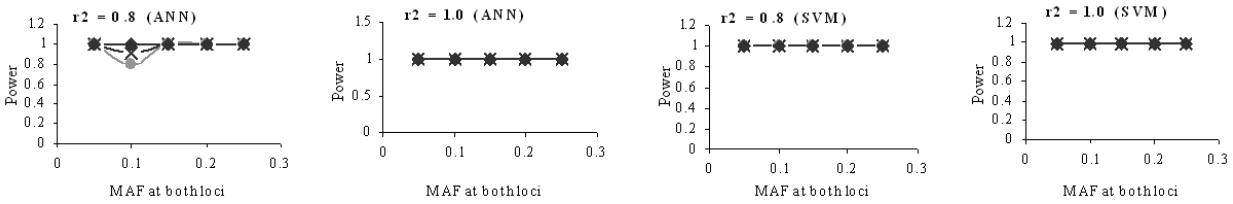


Fig. 2c. Power of gene-gene interaction detection using the ANN and SVM techniques for marginal effects of disease loci equally set to OR=2.0, for 1000 cases and 1000 controls, and with a significance threshold set to 0.05.

disequilibrium between disease and marker loci. This statement holds generally across all gene-gene interaction models and for both testing techniques. Thus a gene-gene interaction involving SNPs with low marginal effect sizes at each disease locus requires a much larger sample size for detection than an interaction of two genes with moderate marginal gene effect sizes.

The power to detect interaction increases with increasing disease allele frequency (or MAF), regardless of the interaction model or the machine learning technique used. The effect of the disease allele frequency on the power is more pronounced in the combination with low LD and/or low marginal gene effects. This impact of the low disease allele frequency can be compensated by an increase in sample size (1000 cases) for a moderate marginal gene effect but can be insufficient for a low marginal effect of OR=1.2. Another possible solution is to minimize the genotype size by selecting a smaller number of candidate SNPs for the interaction study.

The effect of LD on the power to detect interaction was investigated by comparing the power of the causal SNP to that of the surrogate marker in LD with the causal SNP when the causal SNP is removed from the study. As expected from our previous study [1], the detection of interaction decreased in the presence of incomplete LD between the disease SNP (absent from the testing set) and the SNP in LD (tested), as well as when marker alleles decrease the marginal heterozygote ratio at a disease locus. The marginal effect size of a surrogate marker allele that is in LD with the causative allele and has the same allele frequency is equal to the marginal effect size of the causative allele when the LD measure $r^2 = 1.0$ and is smaller than the marginal effect size of the causative allele when $r^2 < 1.0$. It was shown in [1] that the DAF/MAF mismatch lowers the marginal effect of a surrogate marker allele. We found the effect of LD on the power of interaction detection to be model-independent and similar for both the ANN and SVM techniques.

In our previous study [1], we compared the performance of the ANN and SVM techniques with a real dataset comprising approximately 40 genetic and non-genetic factors for 700 cases and 300 controls. We demonstrated that the SVM technique is superior to the ANN technique in detection of single significant SNPs as well as for two interacting factors due to the SVM learning algorithm advantage in finding global minima over local minima as in ANN. However, the SVM technique requires the removal of all but one marker from those markers that are in high LD with each other.

VI. Conclusion

We presented the ANN and SVM techniques applied to the detection and significance testing of gene-gene interactions with a complex disease outcome in a population based case-control study with different disease models involving two interacting causative loci. Both techniques offer the necessary power to detect rare to common single disease alleles of low to high effect sizes in samples of realistic size. The power of detection correlates with allele frequency of the disease-associated loci, with LD between causative and marker alleles, with DAF/MAF mismatch and the sample size. The minimal requirements for a successful study design are defined on the basis of results obtained with simulated datasets. The proposed algorithms are model-free.

Unfortunately, we anticipate that the application of both techniques to large genome-scan association studies would become time-intensive. This is mostly due to the pre-selection step for two interacting factors associated with a high risk of disease when applying the interaction based strategy (see III). This strategy involves an exhaustive iterative search over numerous subsets of SNPs (SNPs genotype "windows"). The 2-stage strategy of two interacting factors pre-selection is definitely much less time consuming than the interaction based strategy. It includes splitting a genome-wide set of SNPs into manageable sliding windows of genotypes (plus non-genetic factors) and conducting relatively fast single factors pre-selection at first, and then interaction-based two-factor pre-selection on the limited amount of SNPs. The latter approach has the risk of missing significant interactions with a weak single SNP effect (marginal effect size < OR=1.2). The testing part of a single SNP-SNP (SNP-environmental factor) pair takes about 1 hour using 100 SNPs for a dataset of 4000 samples and with 100 bootstraps on a single 3GHz cpu. However, the pre-selection and testing algorithms are amenable to parallelization and can run on a cluster of parallel computers.

We did not attempt to apply any other type of ANN architecture beyond a traditional 3-layered feedforward network for improving the classification performance of ANN technique. Our choice of learning algorithms and learning parameters was made to maximize the difference in performance between two networks while avoiding over fitting. Applying such advanced addition to the ANN optimization like evolutionary computing [29] can potentially improve the performance of the proposed ANN-based approach for the detection gene-gene interactions.

Of future interest, we plan to migrate to parallel computing and we wish to expand the two-locus interaction tests to an analogous class of three-locus models.

ACKNOWLEDGMENT

We are grateful to Yoshua Bengio for helpful discussions, and to Dana Aeschliman for comments and advice. This project was supported by Thomas J. Hudson, Genome Québec and Genome Canada, as well as the Montreal Heart Institute Foundation. MPD is supported by the Fonds de la Recherche en Santé du Québec (FRSQ).

REFERENCES

[1] N.Matchenko-Shimko, M.-P.Dubé, "Bootstrap Inference with Neural-Network Modeling for Gene-Disease Association Testing", *Proceedings of the 2006 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, 2006, *in press.*

[2] NJ.Schork, D.Fallin, S.Lancbury, "Single nucleotide polymorphism and the future of genetic epidemiology. Mini Review", *Clinical genetics*", Vol. 58, pp.250-264, 2000.

[3] ES. Lander, NJ. Schork, "Genetic dissection of complex traits". *Science*, Vol.265, pp.2037–2048, 1994.[4] AR.Templeton, "Epistasis and complex traits"**,** *Epistasis and Evolutionary Process"* Edited by: Wade M, Brodie III B, Wolf J.Oxford,Oxford University Press; 2000.

[5] JH. Moore and SM. Williams, "New strategies for identifying gene-gene interactions in hypertension", *Ann Med*, Vol.34, pp.88-95. 2002.

[6] P. Kiberstis, and L. Roberts, "Introduction to special issue. It's Not Just the Genes", *Science,* Vol 296, Issue 5568, p. 685, 2002

[7] KT. Zondervan, LR. Cardon, "The complex interplay among factors that influence allelic association", *Nat Rev Genet.*,Vol. 5(2),pp. 89-100,2004.

[8] WN.Frankel, NJ.Schork, "Who's afraid of epistasis?", *Nat Genet*. Vol. 14(4), pp.371-373,1996.

[9] Nicholas J. Schork, "Genetically Complex Cardiovascular Traits Origins, Problems, and potential solutions", *Hypertension,*Vol. 29, p.145, 1997.

[10] NJ. Schork, "Genetics of complex disease. Approaches, problems and solutions", *Am J Resp Crit Care Med,* Vol.156: S103-S109, 1997.

[11] SM.Williams, JL. Haines, JH. Moore, "The use of animal models in the study of complex disease: all else is never equal or why do so many human studies fail to replicate animal findings?" *Bioessays*, Vol. 26(2), pp.170-9, 2004

[12] R. Culverhouse, B.K. Suarez, J.Lin, & T. Reich, "A perspective on epistasis: limits of models displaying no main effect", *Am. J. Hum. Genet,* 70, pp.461–471, 2002

[13] J. Hoh, J. Ott, "Mathematical multi-locus approaches to localizing complex human trait genes", *Nat Rev Genet.*,Vol. 4(9), pp.701-9.2003.

[14] Christopher S. Carlson, Michael A. Eberle, Leonid Kruglyak & Deborah A. Nickerson, "Mapping complex disease loci in whole-genome association studies", *Nature,* Vol 429, pp.446-453, 2004.

[15] KE.Lohmueller, CL. Pearce, M. Pike, ES.Lander, JN. Hirschhorn, "Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease", *Nature Genet*, Vol. 33(2), pp.177-82,2003.

[16] H.J. Risch, "Searching for genetic determinants in the new millennium", *Nature*. Vol. 405, 847–856, 2000.

[17] L.J. Palmer, Cookson WOCM, "Using single nucleotide polymorphisms as a mean to understanding the pathophysiology of asthma", *Respiratory research,* Vol 2, No.2, pp.102-112, 2001.

[18] J. Marchini, L. Cardon, M. Phillips, and P. Donnelly, "Genome-wide strategies for detecting multiple loci influencing complex diseases", *Nature Genetics*, Vol. 37, pp.413–417, 2005.

[19] M. Nothnagel, "Simulation of LD block-structured SNP haplotype data and its use for the analysis of case-control data by supervised learning methods", *Am J Hum Genet*, vol. 71, (Suppl.)(4), pp. A2363,2002.

[20] T. Joachims, "Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning", B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.

[21] Timothy Masters, "Practical Neural Network Recipes in C++". Academic Press, Inc. 1993, pp.195-197.

[22] J.J. Montaño, A. Palmer, "Numeric sensitivity analysis applied to feedforward neural networks", Neural Computing and Applications, 12(2), pp. 119-125, 2003.

[23] JM. Zurada, A. Malinowski, I. Cloete, "Sensitivity analysis for minimization of input data dimension for feedforward neural network", *In: Proceedings IEEE International Symposium on Circuits and Systems 1994: IEEE, New York*, pp. 447–450.

[24] CM. Bishop, "Neural networks for pattern recognition", Oxford University Press, Oxford, 1995.

[25] VN Vapnik, "Statistical Learning Theory". John Wiley and Sons, 1998.

[26] A. Zien, G. Rätsch, S. Mika, B. Schölkopf, C. Lemmen, A. Smola, T. Lengauer, and K.-R. Müller, "Engineering support vector machine kernels that recognize translation initiation sites", *German Conference on Bioinformatics, 1999.*

[27] E. Osuna, R. Freund, and F. Girosi, "Support vector machines: training and applications", AI Memo 1602, MIT, May 1997.

[28] R. Burbidge, B.Buxton, "An introduction to support vector machines for data mining", SVM tutorials, 2001.

[29] M. D. Ritchie, B.C. White, J.S. Parker, L.W. Hahn, J.H. Moore, "Optimization of Neural Networks using Genetic Programming to Improve Detection and Modeling of Gene-Gene Interactions in Studies of Human Diseases". *Alwyn M. Barry* editor*, GECCO 2003 Proceedings of the Bird of a Feather Workshops, Genetic and Evolutionary Computation Conference, Chicago,* 2003, pp.72-74.

APPENDIX A. OVERVIEW OF THE SVM ALGORITHM

SVM is a supervised learning algorithm developed over the past decade by Vapnik and others [25] which employs a structural risk minimization (SRM) principle which minimizes an upper bound on the expected risk. Suppose we have a set of *l* examples presented by vectors:

$$(y_1, \boldsymbol{x}_1),...,(y_l, \boldsymbol{x}_l), \mathbf{x} \in \Re^n, y \in \{-1,+1\}$$

with only 2 classes. The task of classification is to find a rule which assigns an instance to one of these classes. One possible formalization of this class is to estimate a function $f : \Re^n \to \{-1,+1\}$. In linearly separable cases, the hyperplane which separates two different groups of input vectors with a maximum margin, is constructed by finding another vector **w** (weight vector) and a parameter *b* (bias) that minimizes $\|\mathbf{w}\|^2$ and satisfies the following conditions $\boldsymbol{w}^T \cdot \boldsymbol{x}_i + b \geq +1$ for the group 1 with positively labeled targets $y_i = +1$ and $\boldsymbol{w}^T \cdot \boldsymbol{x}_i + b \leq -1$ for the group 2 with negatively labeled targets $y_i = -1$, or

$$y_i[\boldsymbol{w}^T \cdot \boldsymbol{x}_i + b] \geq +1 \text{ ,i=1,...,}l. \quad (1)$$

Vectors **x** for which $y_i[\boldsymbol{w}^T \cdot \boldsymbol{x}_i + b] = 1$ are the support vectors which lie closest to the separating hyperplane. Here y_i is the group index; $\boldsymbol{w}$ is a vector normal to the constructed hyperplane, $|b|/\|\boldsymbol{w}\|$ is the perpendicular distance from the hyperplane to the origin and $\|\boldsymbol{w}\|$ is the Euclidean norm of $\boldsymbol{w}$. After the determination of $\boldsymbol{w}$ and *b*, a given vector **x** can be classified by:

$$f(\boldsymbol{x}) = sign\left[(\boldsymbol{w}^T \cdot \boldsymbol{x}) + b\right] \quad (2)$$

The optimization is a convex quadratic programming (QP) problem, which has a global optimum. Thus the problem of many local optima in the case of training like in the case in neural network is avoided. Parameters in QP solvers will affect only the training time but not the quality of the solution. The optimal solution is given by (4) has optimal weight vector

$\boldsymbol{w}^* = \sum_{i=1}^{i} \lambda_i^* \cdot y_i \cdot \boldsymbol{x_i}$ and bias $b^* = y_i \cdot \boldsymbol{w}^{*T} \cdot \boldsymbol{x}_i$ for any support vector $\mathbf{x}_i$. The decision function is transformed into:

$$f(\boldsymbol{x}) = sign\left[\sum_{i=1}^{i} \lambda_i^* \cdot y_i \cdot \boldsymbol{x}^T \cdot \boldsymbol{x_i} + b\right]. \quad (3)$$

In non-linearly separable cases, SVM maps the input variable into a high dimensional feature space **z** by the function ϕ. By choosing a non-linear mapping *a priori* SVM finds an optimal linear separating hyperplane with the maximal margin in this higher dimensional space. The decision function (2) becomes (4):

$$f(x) = sign[\phi(\boldsymbol{x})^T \cdot \boldsymbol{w}^* + b^*] = sign\left[\sum_{i=1}^{i} \lambda_i^* \cdot y_i \cdot \phi(\boldsymbol{x})^T \cdot \phi(\boldsymbol{x_i}) + b^*\right].$$

The inner product $K(x_i, x_j) \equiv \phi(x_i)^T \cdot \phi(x_j)$ is called the kernel function. It allows constructing an optimal separating hyperplane in the feature space without explicitly performing calculations in this space. The decision function from (4)

$$f(x) = sign\left[\sum_{i=1}^{i} \lambda_i^* \cdot y_i \cdot K(\boldsymbol{x_i}, \boldsymbol{x_j}) + b^*\right]$$

has optimal bias $b^* = y_i \cdot \boldsymbol{w}^{*T} \cdot \phi(\boldsymbol{x}_i) = y_i - \sum_{j=1}^{l} y_j \lambda_i^* K(\boldsymbol{x}_j, \boldsymbol{x}_i)$ for any support vector $\mathbf{x}_i$. $\lambda_1^*, ..., \lambda_l^*$ are the Lagrange multipliers. SVM uses the following four basic kernels: i) linear, $K(x_i, x_j) \equiv x_i^T \cdot x_j$, ii) polynomial, $K(x_i, x_j) \equiv (\gamma x_i^T \cdot x_j + r)^d, \gamma > 0$, iii) radial basis function (RBF), $K(x_i, x_j) \equiv \exp(-\gamma \| x_i \cdot x_j \|^2), \gamma > 0$, and iv) sigmoid, $K(x_i, x_j) \equiv \tanh(\gamma x_i^T \cdot x_j + r)$, where γ, r, and d are kernel parameters.

In case of non-separable data (noisy data), training with zero-error leads to poor generalization as the learned classifiers are fitting idiosyncrasies of the noise in the training data. SVM allows misclassification of some data points. The separating hyperplane is subject to the following conditions: minimizing $\frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{l} \xi_i^k$ and satisfying $y_i[\boldsymbol{w}^T \cdot \phi(\boldsymbol{x_i}) + b] \geq +1 - \xi_i, \xi \geq 0, i = 1, ..., l$, where $\xi_1, ...\xi_l$ is a vector of slack variables that measure the amount of violation of the constrain (1); C is a regularization parameter that controls the trade-off between maximizing the margin and minimizing the training error term.

In a regression SVM, the regression task is to estimate the functional dependence of y on a set of independent variables **x**. As in other regression problems, the relationship between the independent and dependent variables is given by a deterministic function f plus the addition of some additive noise: $y = f(\mathbf{x}) + noise$. The task is then to find a functional form for f that can correctly predict new cases that the SVM has not been presented with before. This can be achieved by training the SVM model on a sample set, i.e., training set, a process that involves, like classification, the sequential optimization of an error function. The main difference is the type of a loss function employed:

$L^e(\mathbf{x}, y, f) = |y - f(\mathbf{x})|_e^p = \max(0, |y - f(\mathbf{x})| - \varepsilon)^p$ with $p \in \{1,2\}$.

The loss function only counts error predictions which are more than ε away from the training data. This loss function allows the concepts of margin to be carried over to the regression case keeping all of the nice statistical properties. Support vector regression also results in a QP.

APPENDIX B. SVM MODEL PARAMETERS

The process of determining the decision boundary is greatly influenced by the selection of the kernel and classifier parameters implicitly defining the structure of the high dimensional feature space where the maximal marginal hyperplane is found. The choice of kernel and kernel-related parameters is generally domain-specific [26] and involves choosing the similarity measure for the data, a representation of the data, and/or a hypothesis space for learning that reflects the prior knowledge about the problem in hand. Our genetic data are characterized by a week signal but strong noise, and our task is to detect significant SNPs by measuring the difference in performance between two SVMs. Therefore, to select the optimal kernel and SVM parameters, we recorded the MSE and the accuracy rate for two SVM models and used the difference in MSE and in accuracy rates between the two SVM models.

We applied the following kernels implemented in the SVM*light* software: local radial basis function (RBF), global linear and polynomial functions. All of them span a sufficiently rich hypothesis space [27, 28] and are positive and symmetrical. Local kernels attempt to measure the proximity of data samples and are based on a distance function, while global kernels are dot-product based. To find optimal parameters we first established the parameter ranges and did an exhaustive grid search over these ranges. Various combinations of parameters were tried on a coarse grid and on a gradually refined grid (refined resolution and boundary). The MSE error was large in the linear SVM, while the accuracy rate and MSE differences were small. This was expected due to the non-linear inputs-target relationship. The RBF SVM performed better in terms of accuracy rate and MSE error across the wide range of parameters spaces. However the MSE and accuracy rate differences were small and thus not suitable to extract the required effect. And finally the polynomial SVM of the second degree with the default regularization parameter $C = \frac{1}{avg(\boldsymbol{x} \cdot \boldsymbol{x})}$, $\gamma = 1.0$ and $r = 1.0$ produced optimal regression accuracy. The accuracy rate and MSE differences were larger than in the cases of linear and RBF kernels. γ and r variations exhibited no significant influence on performance when C was kept at the default value. The significant increase in the C parameter above the default value with γ and r kept constant at default values substantially increased the optimization time. Raising the polynomial kernel to the high power ($d=3$) significantly increased the optimization time and made no significant improvement in the performance.