# Genetic Programming Neural Networks as a Bioinformatics Tool for Human Genetics

Marylyn D. Ritchie[1], Christopher S. Coffey[2], and Jason H. Moore[1]

[1]Center for Human Genetics Research, Department of Molecular Physiology and Biophysics, Vanderbilt University, 519 Light Hall, Nashville, TN 37232
{ritchie, moore @chgr.mc.vanderbilt.edu}
[2]Department of Biostatistics, University of Alabama at Birmingham, Ryals Public Health Bldg., Rm. 327M, Birmingham, AL, 35294
{CCoffey@ms.soph.uab.edu}

**Abstract.** The identification of genes that influence the risk of common, complex diseases primarily through interactions with other genes and environmental factors remains a statistical and computational challenge in genetic epidemiology. This challenge is partly due to the limitations of parametric statistical methods for detecting genetic effects that are dependent solely or partially on interactions. We have previously introduced a genetic programming neural network (GPNN) as a method for optimizing the architecture of a neural network to improve the identification of gene combinations associated with disease risk. Previous empirical studies suggest GPNN has excellent power for identifying gene-gene interactions. The goal of this study was to compare the power of GPNN and stepwise logistic regression (SLR) for identifying gene-gene interactions. Using simulated data, we show that GPNN has higher power to identify gene-gene interactions than SLR. These results indicate that GPNN may be a useful pattern recognition approach for detecting gene-gene interactions.

## 1 Introduction

One goal of genetic epidemiology is to identify genes associated with common, complex multifactorial diseases. Success in achieving this goal will depend on a research strategy that recognizes and addresses the importance of interactions among multiple genetic and environmental factors in the etiology of diseases such as essential hypertension [1, 2]. One traditional approach to modeling the relationship between discrete predictors such as genotypes and discrete clinical outcomes is logistic regression [3]. Logistic regression is a parametric statistical approach for relating one or more independent or explanatory variables (e.g. genotypes) to a dependent or outcome variable (e.g. disease status) that follows a binomial distribution. However, as reviewed by Moore and Williams [2], the number of possible interaction terms grows exponentially as each additional main effect is included in the logistic regression model. Thus, logistic regression is limited in its ability to deal with interactions involving many factors. Having too many

K. Deb et al. (Eds.): GECCO 2004, LNCS 3102, pp. 438–448, 2004.

independent variables in relation to the number of observed outcome events is a well-recognized problem [4, 5] and is an example of the curse of dimensionality [6].

In response to this limitation, Ritchie et al. [7] developed a genetic programming optimized neural network (GPNN). Neural networks (NN) have been utilized in genetic epidemiology, however, with little success. A potential weakness in the previous NN applications is the improper selection of NN architecture. GPNN was developed in an attempt to improve upon the trial-and-error process of choosing an optimal architecture for a pure feed-forward back propagation neural network. The GPNN optimizes the inputs from a larger pool of variables, the weights, and the connectivity of the network including the number of hidden layers and the number of nodes in the hidden layer. Thus, the algorithm attempts to generate optimal neural network architecture for a given data set. This is an advantage over the traditional back propagation NN in which the inputs and architecture are pre-specified and only the weights are optimized.

Although previous empirical studies suggest GPNN has excellent power for identifying gene-gene interactions, a comparison of GPNN with a traditional statistical method has not yet been performed. The goal of the present study was to compare the power of GPNN and stepwise logistic regression (SLR) for identifying gene-gene interactions using data simulated from a variety of gene-gene interaction models. This study is motivated by the number of studies in human genetics where SLR has been applied. We wanted to determine if GPNN is more powerful than the status quo in the field. We find that GPNN has higher power to detect gene-gene interactions than stepwise logistic regression. These results demonstrate that GPNN may be an important pattern recognition tool for studies in genetic epidemiology.

# 2 Methods

## 2.1 A Genetic Programming Neural Network Approach

GPNN was developed to improve upon the trial-and-error process of choosing an optimal architecture for a pure feed-forward back propagation neural network (NN) [7]. Optimization of NN architecture using genetic programming (GP) was first proposed by Koza and Rice [8]. The goal of this approach is to use the evolutionary features of genetic programming to evolve the architecture of a NN. The use of binary expression trees allow for the flexibility of the GP to evolve a tree-like structure that adheres to the components of a NN. Figure 1 shows an example of a binary expression tree representation of a NN generated by GPNN. The GP is constrained such that it uses standard GP operators but retains the typical structure of a feed-forward NN. While GP could be implemented without constraints, the goal was to evolve NN since they were being explored as a tool

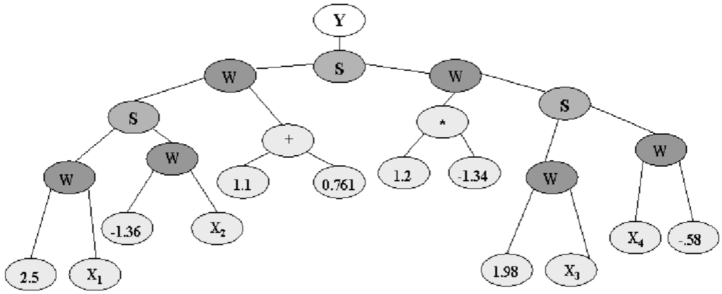


**Fig. 1.** An example of a NN evolved by GPNN. The Y is the output node, S indicates the activation function, W indicates a weight, and $X_1$-$X_4$ are the NN inputs.

for genetic epidemiology. Thus, we wanted to make an improvement to a method already
being used. A set of rules is defined prior to network evolution to ensure that the GP tree maintains a structure that represents a NN. The rules used for this GPNN implementation are consistent with those described by Koza and Rice [8]. The flexibility of the GPNN allows optimal network architectures to be generated that consist of the appropriate inputs, connections, and weights for a given data set.

The GPNN method has been described in detail [7]. The steps of the GPNN method are shown in Figure 2 and described in brief as follows. First, GPNN has a set of parameters that must be initialized before beginning the evolution of NN models. These include an independent variable input set, a list of mathematical functions, a fitness function, and finally the operating parameters of the GP. These operating parameters include number of demes (or populations), population size, number of generations, reproduction rate, crossover rate, mutation rate, and migration [7]. Second, the data are divided into 10 equal parts for 10-fold cross-validation. Here, we will train the GPNN on 9/10 of the data to develop a NN model. Later, we will test this model on the 1/10 of the data left out to evaluate the predictive ability of the model.

Third, training of the GPNN begins by generating an initial population of random solutions. Each solution is a binary expression tree representation of a NN, similar to that shown in Figure 1. Fourth, each GPNN is evaluated on the training set and its fitness recorded. Fifth, the best solutions are selected for crossover and reproduction using a fitness-proportionate selection technique, called roulette wheel selection, based on the classification error of the training data [9]. Classification error is defined as the proportion of individuals where the disease status was incorrectly specified. A predefined proportion of the best solutions will be directly copied (reproduced) into the new generation. Another proportion of the solutions will be used for crossover with other best solutions. The new generation, which is equal in size to the original population, begins the cycle again. This continues until some criterion is met at which point the GPNN stops. This

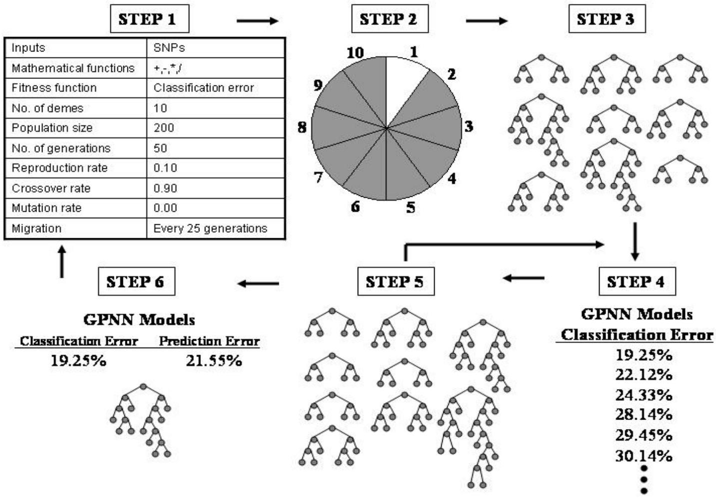


**Fig. 2.** The Steps of the GPNN algorithm

criterion is either a classification error of zero or the maximum number of generations having been reached. A “best-so-far” solution is chosen after each generation. At the end of the GPNN evolution, the one “best-so-far” solution is selected as the optimal NN. Sixth, this best GPNN model is tested on the 1/10 of the data left out to estimate the prediction error of the model. Prediction error is a measure of the ability to predict disease status in the 1/10 of the data. Steps two through six are performed ten times with the same parameters settings, each time using a different 9/10 of the data for training and 1/10 of the data for testing.

The results of a GPNN analysis include 10 GPNN models, one for each split of the data. In addition, a classification error and prediction error is recorded for each of the models. A cross-validation consistency can be measured to determine those variables which have a strong signal in the gene-gene interaction model [7, 10, 11, 12]. Cross-validation consistency is the number of times a particular combination of variables are present in the GPNN model out of the ten cross-validation data splits. Thus a high cross-validation consistency, ~10, would indicate a strong signal, whereas a low cross-validation consistency, ~1, would indicate a weak signal and a potentially false positive result.

## 2.2 Data Simulation

The goal of the simulation was to generate data sets that exhibit gene-gene interactions for the purpose of evaluating the power of GPNN in comparison to the power of SLR. We simulated a collection of models varying several conditions including number of interacting genes, allele frequency, and heritability. Heritability is defined in the broad sense as the proportion of phenotypic variation that is attributed to genetic factors. Loosely, this means the strength of the genetic effect. Thus a higher heritability will be a larger effect and easier to detect. Heritability is calculated using equations described in [13]. Additionally, we used a constant sample size for all simulations. We selected the sample size of 200 cases (individuals with disease) and 200 controls (individuals without disease) because this is a typical sample that is used in many genetic epidemiology studies.

As discussed by Templeton [14], epistasis, or gene-gene interaction, occurs when the combined effect of two or more genes on a phenotype could not have been predicted from their independent effects. It is anticipated that epistasis is likely to be a ubiquitous component of the genetic architecture of common human diseases [15]. Current statistical approaches in human genetics focus primarily on detecting the main effects and rarely consider the possibility of interactions [14]. In contrast, we are interested in simulating data using different epistasis models that exhibit minimal independent main effects, but produce an association with disease primarily through interactions. In this study, we use penetrance functions as genetic models. Penetrance functions model the relationship between genetic variations and disease risk. Penetrance is defined as the probability of disease given a particular combination of genotypes.

To evaluate the power of GPNN and SLR for detecting gene-gene interactions, we simulated case-control data using a variety of epistasis models in which the functional genes are single-nucleotide polymorphisms (SNPs). We selected models that exhibit interaction effects in the absence of any main effects. Interactions without main effects are desirable because they provide a high degree of complexity to challenge the ability of a method to identify gene-gene interactions. If main effects were present, it could be difficult to evaluate whether particular genes were detected due to the main effects or the interactions or both. In addition, it is likely that a method that can detect interacting genes in the absence of main effects will be able to detect main effect genes as well.

To generate a variety of epistasis models for this study, we selected three criteria for variation. First, we selected epistasis models with a varying number of interacting genes: either two or three. Previous studies had only investigated the power of GPNN using two-gene models [7]. We speculate that common diseases will be comprised of complex interactions among many genes. The number of interacting genes simulated here may still be too few to be biologically relevant. However, few, if any complex gene-gene interaction models are known at this time. Next, we selected two different allele frequencies. An allele frequency of 0.2/0.8 was selected so that we could evaluate the ability of GPNN in situations where there is a relatively rare allele. In addition, the frequency of 0.4/0.6 was selected to allow for the situation where both alleles are relatively common. Finally, we selected a range of heritability values including 3%, 2%, 1.5%, 1%, and 0.5%. These heritability values fall into the realm of very small genetic effects. In comparison, the heritability of many common diseases is much higher. For example, Alzheimer's disease is estimated to have heritability exceeding 60% [16] while breast, colorectal, and prostate cancers are 27%, 35%, and 42% respectively [17]. We chose to simulate data using epistasis models with such small heritability values to test the lower limits of GPNN. Based on previous studies, GPNN has over 80% power when the heritability is between 2%-5% [7]. For this particular study, we wanted to explore even smaller genetic effects to identify the point at which GPNN loses power.

We generated models using software described by Moore et al. [18]. We selected models from all possible combinations of number of interacting genes, allele frequency, and heritability, resulting in 20 total models. The penetrance tables for combinations of two SNPs are shown in Tables 1-10. The penetrance tables for the three SNP models are available from the authors by request. All 20 models were selected because they exhibit interaction effects in the absence of any main effects when genotypes are generated using the Hardy-Weinberg equation. Although the biological plausibility of these models is unknown, they represent the worst-case scenario for a disease-detection method because they have minimal main effects. If a method works well with minimal main effects, presumably the method will continue to work well in the presence of main effects.

**Table 1.** Model 1 – Two SNPs, allele frequency 0.2/0.8, $h^2 = 0.030$

| | *AA* | *Aa* | *aa* |
|---|---|---|---|
| *BB* | 0.0998 | 0.0984 | 0.0022 |
| *Bb* | 0.0933 | 0.0996 | 0.0002 |
| *bb* | 0.0028 | 0.0000 | 0.0574 |

Table 1 is an example of a penetrance function for a two-gene epistasis model with no main effects. Each gene is a single SNP with two alleles and three genotypes. In this example, the alleles each have a biological population frequency of $p = 0.2$ $q = 0.8$ with genotype frequencies of $p^2$ for *AA* and *BB*, $2pq$ for *Aa* and *Bb*, and $q^2$ for *aa* and *bb*, consistent with Hardy-Weinberg equilibrium. Thus, assuming the frequency of the *AA* genotype is 0.16, the frequency of *Aa* is 0.32, and the frequency of *aa* is 0.64, then the marginal penetrance of *BB* (i.e. the effect of just the *BB* genotype on disease risk) can be calculated as (0.04 * 0.0998) + (0.32 * 0.0984) + (0.64 * 0.0022) = 0.03. This means that the probability of disease given the *BB* genotype is 0.03, regardless of the genotype at the other genetic variation. Similarly, the marginal penetrance of *Bb* can be calculated as (0.04 * 0.0933) + (0.32 *0.0996) + (0.64 * 0.0002) = 0.03. Note that for this model, all of the marginal penetrance values (i.e. the probability of disease given a single genotype, independent of the others) are equal, which indicates the absence of main effects (i.e. the genetic variations do not independently affect disease risk). This is true despite the table penetrance values not being equal. Here, risk of disease is greatly increased by inheriting one of the following high-risk genotype combinations: *AABB, AABb, AaBB, AaBb*, and slightly increased by inheriting genotype combination *aaBb*.

Each data set consisted of 200 cases and 200 controls. We simulated 100 data sets of each model consisting of the functional SNPs and either seven or eight non-functional SNPs for a total of ten SNPs. This resulted in 2000 total datasets. We used a dummy variable encoding for the genotypes where *n-1* dummy variables are used for *n* levels (or genotypes) [19]. Based on the dummy coding, these data would have 20 input variables.

**Table 2.** Model 2 - Two SNPs, allele frequency 0.2/0.8, $h^2 = 0.020$

| | *AA* | *Aa* | *aa* |
|---|---|---|---|
| *BB* | 0.0786 | 0.0003 | 0.0967 |
| *Bb* | 0.0010 | 0.0013 | 0.1001 |
| *bb* | 0.0948 | 0.0998 | 0.0428 |

**Table 3.** Model 3 - Two SNPs, allele frequency 0.2/0.8, $h^2 = 0.015$

| | *AA* | *Aa* | *aa* |
|---|---|---|---|
| *BB* | 0.0276 | 0.0942 | 0.0287 |
| *Bb* | 0.0941 | 0.0996 | 0.0226 |
| *bb* | 0.0277 | 0.0198 | 0.0657 |

**Table 4.** Model 4 - Two SNPs, allele frequency 0.2/0.8, $h^2 = 0.010$

| | *AA* | *Aa* | *aa* |
|---|---|---|---|
| *BB* | 0.0884 | 0.0894 | 0.0307 |
| *Bb* | 0.0710 | 0.0036 | 0.0737 |
| *bb* | 0.0368 | 0.0711 | 0.0404 |

**Table 5.** Model 5 - Two SNPs, allele frequency 0.2/0.8, $h^2 = 0.005$

| | *AA* | *Aa* | *aa* |
|---|---|---|---|
| *BB* | 0.0539 | 0.0732 | 0.0416 |
| *Bb* | 0.007 | 0.0207 | 0.0685 |
| *bb* | 0.0732 | 0.066 | 0.044 |

**Table 6.** Model 6 - Two SNPs, allele frequency 0.4/0.6, $h^2 = 0.030$

| | *AA* | *Aa* | *aa* |
|---|---|---|---|
| *BB* | 0.0848 | 0.0754 | 0.0053 |
| *Bb* | 0.0705 | 0.0135 | 0.0967 |
| *bb* | 0.0118 | 0.0937 | 0.0131 |

**Table 7.** Model 7 - Two SNPs, allele frequency 0.4/0.6, $h^2 = 0.020$

| | *AA* | *Aa* | *aa* |
|---|---|---|---|
| *BB* | 0.0093 | 0.0281 | 0.0902 |
| *Bb* | 0.0491 | 0.0763 | 0.0063 |
| *bb* | 0.0625 | 0.0161 | 0.0824 |

**Table 8.** Model 8 - Two SNPs, allele frequency 0.4/0.6, $h^2 = 0.015$

| | *AA* | *Aa* | *aa* |
|---|---|---|---|
| *BB* | 0.0381 | 0.0151 | 0.073 |
| *Bb* | 0.0485 | 0.0618 | 0.0067 |
| *bb* | 0.0288 | 0.0209 | 0.0693 |

**Table 9.** Model 9 - Two SNPs, allele frequency 0.4/0.6, $h^2 = 0.010$

| | *AA* | *Aa* | *aa* |
|---|---|---|---|
| *BB* | 0.0465 | 0.0368 | 0.0706 |
| *Bb* | 0.0666 | 0.0691 | 0.02 |
| *bb* | 0.0314 | 0.0329 | 0.0818 |

**Table 10.** Model 10 - Two SNPs, allele frequency 0.4/0.6, $h^2 = 0.005$

| | *AA* | *Aa* | *aa* |
|---|---|---|---|
| *BB* | 0.0161 | 0.0514 | 0.0573 |
| *Bb* | 0.0287 | 0.0442 | 0.0614 |
| *bb* | 0.0867 | 0.0511 | 0.0253 |

## 2.3 Data Analysis

Next, we used GPNN and SLR to analyze 100 data sets for each of the epistasis models. The GP parameter settings for GPNN included 10 demes, population size of 200 per deme, 50 generations, reproduction rate of 0.10, crossover rate of 0.90, mutation rate of 0.0, and migration every 25 generations. GPNN is not required to use all the variables as inputs. Here, GPNN performed random variable selection in the initial population of solutions. Through evolution, GPNN selects those variables that are most relevant. We calculated a cross-validation consistency for each SNP in each data set. This measure is defined as the number of times each SNP is in the GPNN model across the ten cross validation intervals. Thus, one would expect a strong signal to be consistent across all ten or most of the data splits, where a false positive signal may be present in only one or a few of the cross validation intervals. We estimated the power of GPNN as the number of times the correct functional SNPs had a cross-validation consistency that was higher than all other SNPs in the dataset, divided by the total number of datasets for each epistasis model. Either one or both of the dummy variables could be selected to consider a gene present in the model.

SLR is based on a statistical algorithm that determines the importance of variables and either includes them or excludes them from the model. The importance is determined by the statistical significance of the variable based on a chi-squared test [3]. Here, we used a p-Value of 0.20 to enter the model, and a p-Value of 0.10 to remain in the model. This type of model building procedure can also be referred to as hierarchical model building because to consider interactions among the variables, each variable must remain in the model due to its statistical significance on its own. Thus, using this approach, one can only detect interactions in the presence of main effects of each of the interacting variables. We performed this SLR procedure on each data set. We estimated power of SLR as the number of times the interaction term for the correct functional SNPs was selected in the final SLR model.

## 3 Results

The results of this study are shown in Table 11. Here, we list the 20 epistasis models sorted by number of genes, allele frequency, and heritability along the vertical axis. SLR has no power to detect the functional genes in any of the models studied. GPNN, on the other hand, has higher power than SLR for all of the epistasis models. The power of GPNN is higher for the models with two functional genes, and similarly for the models with higher heritability values.

## 4 Discussion

Identifying disease susceptibility genes associated with common complex, multifactorial diseases is a major challenge for genetic epidemiology. One of the dominating factors in this challenge is the difficulty in detecting gene-gene interactions with currently available statistical approaches. To deal with this issue, new statistical approaches have been developed such as the GPNN. GPNN has been shown to have higher power than a back propagation NN using simulated data generated under five two-gene epistasis models [7]. The goal of the current study was to compare the power of GPNN and SLR for detecting gene-gene interactions using data simulated from a variety of epistasis models. Computationally, GPNN is more burdensome than SLR. However, in human genetics the goal is to identify disease susceptibility genes. If one method is more powerful, even if it is more computationally expensive, it may be money well spent. Based on the results shown in Table 11, SLR had no power to detect a statistically significant interaction term. In comparison, GPNN had high power for most of the models examined. These results led to some skepticism that logistic regression (LR) may not be able to model the interactions that we had simulated. To be certain that LR was able to model these nonlinear interactions, we performed a forward selection LR analysis using only the two or three functional SNPs and their corresponding interaction term (Table 12). We estimated the power of LR using the number of data sets where the interaction term was statistically significant. In this study, LR had between 5-100% and 0-25% power for the two and three gene models respectively. Thus, LR was theoretically able to model these interactions. We conclude that LR may be a successful procedure when the selection of variables has been conducted prior to the modeling process. However, when variable selection and modeling is taking place simultaneously, GPNN may provide higher power to detect such gene-gene interaction effects.

While these results demonstrate the lower limits of GPNN's power to detect gene-gene interactions, there are still many more questions to be addressed. First, it will be important to extend the simulation studies to include more interacting genes, larger sample sizes and a larger range of higher heritability values. In addition, a larger set of epistasis models including those with a small degree of main effect would provide further evidence of the power of GPNN. Finally, it would be interesting to use a different model validation procedure, such as the three-way data split [20], instead of ten-fold cross validation.

**Table 11.** Power comparison of GPNN and Stepwise Logistic Regression (SLR)

| Model | | | Power (%) | |
|---|---|---|---|---|
| # Genes | Allele frequency | $h^2$ | GPNN | SLR |
| 2 | 0.2/0.8 | 0.030 | 100 | 0 |
| 2 | 0.2/0.8 | 0.020 | 94 | 0 |
| 2 | 0.2/0.8 | 0.015 | 97 | 0 |
| 2 | 0.2/0.8 | 0.010 | 81 | 0 |
| 2 | 0.2/0.8 | 0.005 | 24 | 0 |
| 2 | 0.4/0.6 | 0.030 | 100 | 0 |
| 2 | 0.4/0.6 | 0.020 | 99 | 0 |
| 2 | 0.4/0.6 | 0.015 | 99 | 0 |
| 2 | 0.4/0.6 | 0.010 | 77 | 0 |
| 2 | 0.4/0.6 | 0.005 | 16 | 0 |
| 3 | 0.2/0.8 | 0.030 | 99 | 0 |
| 3 | 0.2/0.8 | 0.020 | 94 | 0 |
| 3 | 0.2/0.8 | 0.015 | 22 | 0 |
| 3 | 0.2/0.8 | 0.010 | 4 | 0 |
| 3 | 0.2/0.8 | 0.005 | 3 | 0 |
| 3 | 0.4/0.6 | 0.030 | 75 | 0 |
| 3 | 0.4/0.6 | 0.020 | 35 | 0 |
| 3 | 0.4/0.6 | 0.015 | 20 | 0 |
| 3 | 0.4/0.6 | 0.010 | 3 | 0 |
| 3 | 0.4/0.6 | 0.005 | 1 | 0 |

The results of this study show that GPNN has higher power than SLR to detect gene-gene interactions in models with very small heritability values. Since most common diseases have overall heritability estimates greater than 20%, and GPNN was shown to have 100% power for heritability of 5% due to the genes examined [7], GPNN should have high power for detecting interactions in most common diseases. GPNN is likely to be a powerful pattern recognition approach for the detection of gene-gene interactions in future studies of common human disease.

**Table 12.** Power of Explicit Logistic Regression (LR)

| Model | | | Power (%) |
|---|---|---|---|
| # Genes | Allele frequency | $h^2$ | LR |
| 2 | 0.2/0.8 | 0.030 | 92 |
| 2 | 0.2/0.8 | 0.020 | 61 |
| 2 | 0.2/0.8 | 0.015 | 100 |
| 2 | 0.2/0.8 | 0.010 | 60 |
| 2 | 0.2/0.8 | 0.005 | 64 |
| 2 | 0.4/0.6 | 0.030 | 7 |
| 2 | 0.4/0.6 | 0.020 | 5 |
| 2 | 0.4/0.6 | 0.015 | 29 |
| 2 | 0.4/0.6 | 0.010 | 44 |
| 2 | 0.4/0.6 | 0.005 | 89 |
| 3 | 0.2/0.8 | 0.030 | 25 |
| 3 | 0.2/0.8 | 0.020 | 16 |
| 3 | 0.2/0.8 | 0.015 | 2 |
| 3 | 0.2/0.8 | 0.010 | 5 |
| 3 | 0.2/0.8 | 0.005 | 2 |
| 3 | 0.4/0.6 | 0.030 | 0 |
| 3 | 0.4/0.6 | 0.020 | 7 |
| 3 | 0.4/0.6 | 0.015 | 0 |
| 3 | 0.4/0.6 | 0.010 | 12 |
| 3 | 0.4/0.6 | 0.005 | 2 |

**Acknowledgements.** This work was supported by National Institutes of Health grants HL65234, HL65962, GM31304, AG19085, AG20135, and LM007450.

## References

1. Kardia S.L.R.: Context-dependent genetic effects in hypertension. Curr. Hypertens. Reports. 2 (2000) 32-38
2. Moore J.H. and Williams S.M.: New strategies for identifying gene-gene interactions in hypertension. Ann. Med. 34 (2002) 88-95
3. Hosmer D.W. and Lemeshow S.: Applied Logistic Regression. John Wiley & Sons Inc., New York (2000)
4. Concato J., Feinstein A.R., Holford T.R.: The risk of determining risk with multivariable models. Ann. Int. Med. 118 (1996) 201-210
5. Peduzzi P., Concato J., Kemper E., Holford T.R., Feinstein A.R.: A simulation study of the number of events per variable in logistic regression analysis. J. Clin. Epidemiol. 49 (1996) 1373-1379
6. Bellman R.: Adaptive Control Processes. Princeton University Press, Princeton (1961)
7. Ritchie M.D., White B.C., Parker J.S., Hahn L.W., Moore J.H.: Optimization of neural network architecture using genetic programming improves detection of gene-gene interactions in studies of human diseases. BMC Bioinformatics, 4 (2003) 28
8. Koza J.R. and Rice J.P.: Genetic generation of both the weights and architecture for a neural network. IEEE Press Vol II (1991) 397-404
9. Mitchell M.: An Introduction to Genetic Algorithms. MIT Press, Cambridge (1996)
10. Moore J.H.: Cross validation consistency for the assessment of genetic programming results in microarray studies. In: Lecture Notes in Computer Science Vol 2611 ed. by: Raidl, G, et al. Springer-Verlag, Berlin (2003) 99-106
11. Moore J.H., Parker J.S., Olsen N.J., Aune T.S.: Symbolic discriminant analysis of microarray data in autoimmune disease. Genet Epidemiol 23 (2002) 57-69
12. Ritchie M.D., Hahn, L.W., Roodi N., Bailey L.R., Dupont W.D., Parl F.F., Moore J.H.: Multifactor dimensionality reduction reveals high-order interactions among estrogen metabolism genes in sporadic breast cancer. Am. J. Hum. Genet. 69 (2001) 138-147
13. Culverhouse R., Suarez B.K., Lin J., Reich T.: A Perspective on Epistasis: Limits of Models Displaying No Main Effect. Am J Hum Genet 70 (2002) 461-471
14. Templeton A.R.: Epistasis and complex traits. In: Epistasis and Evolutionary Process. ed. by: Wolf J., Brodie III B., Wade M. Oxford University Press, Oxford (2000)
15. Moore J.H.: The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Hum Hered 56 (2003) 73-82
16. Ashford J.W. and Mortimer J.A.: Non-familial Alzheimer's disease is mainly due to genetic factors. J Alzheimers Dis. 4 (2002) 169-77
17. Hemminki K. and Mutanen P.: Genetic epidemiology of multistage carcinogenesis. Mutat. Res. 473 (2001) 11-21
18. Moore, J.H., Hahn L.W., Ritchie M.D., Thornton T.A., White B.C.: Application of genetic algorithms to the discovery of complex genetic models for simulations studies in human genetics. In: Proceedings of the Genetic and Evolutionary Algorithm Conference ed. by W.B. Langdon, E. Cantu-Paz, K. Mathias, R. Roy, D. Davis, R. Poli, K. Balakrishnan, V. Honavar, G. Rudolph, J. Wegener, L. Bull, M.A. Potter, A.C. Schultz, J.F. Miller, E. Burke, and N. Jonoska. Morgan Kaufman Publishers San Francisco (2002) 1150-1155
19. Ott J.: Neural networks and disease association. Am. J. Med. Genet. 105 (2001) 60-61
20. Rowland J.J. Generalisation and model selection in supervised learning with evolutionary computation. In: Lecture Notes in Computer Science Vol 2611 ed. by: Raidl, G, et al. Springer-Verlag, Berlin (2003) 119-130