Mask Functions for the Symbolic Modeling of Epistasis Using Genetic Programming

Ryan J. Urbanowicz
Dartmouth College
1 Medical Center Dr.
Hanover, NH 03755,USA

Bill C. White
Dartmouth College
1 Medical Center Dr.
Hanover, NH 03755,USA

Jason H. Moore
Dartmouth College
1 Medical Center Dr.
Hanover, NH 03755,USA
Jason.H.Moore@dartmouth.edu

ABSTRACT

The study of common, complex multifactorial diseases in genetic epidemiology is complicated by nonlinearity in the genotype-to-phenotype mapping relationship that is due, in part, to epistasis or gene-gene interactions. Symobolic discriminant analysis (SDA) is a flexible modeling approach which uses genetic programming (GP) to evolve an optimal predictive model using a predefined collection of mathematical functions, constants, and attributes. This has been shown to be an effective strategy for modeling epistasis. In the present study, we introduce the genetic "mask" as a novel building block which exploits expert knowledge in the form of a pre-constructed relationship between two attributes. The goal of this study was to determine whether the availability of "mask" building blocks improves SDA performance. The results of this study support the idea that pre-processing data improves GP performance.

Categories and Subject Descriptors

J.3 [**Computer Applications**]: Life and Medical Sciences—*biology and genetics*

General Terms

Algorithms, Design, Human Factors

Keywords

Genetic Analysis, Genetic Epidemiology, Genetic Programming, Symbolic Discriminant Analysis, Symbolic Regression, Function Set, Two-Locus Model, Genetic Mask

1. INTRODUCTION

Advancing laboratory techniques such as DNA microarrays [30] and Gene Chips [13] are driving the massive growth of biomedical data without a paralleled advancement in the analytical and computational methods utilized to interpret this information. The challenge for genetic epidemiologists will be to develop statistical and computational methods that are able to identify subsets of genetic attributes that classify and predict clinical endpoints. In the 1930's, Sir Ronald Fisher et al. [2] developed linear discriminant analysis (LDA) as a tool for classifying discrete endpoints using information about multiple attributes or variables. LDA linearly combines measurements of multiple explanatory variables into a single value or discriminant score that can be used to classify observations. The major disadvantage of LDA is the assumption of linearity, which means that the model needs to be pre-specified, and only the coefficients for each linear predictor are estimated from the data. In the early 90's, Koza et al. [5, 6] developed symbolic regression as a means of identifying regression equations that would not need to be pre-specified. Symbolic regression uses genetic programming (GP) machine learning methodology to identify optimal symbolic regression models. Most recently, Moore et al. [24, 21] extended symbolic regression into a method called symbolic discriminant analysis (SDA) where the symbolic model is used to generate symbolic discriminant scores for each observation in each group from which the classification error can be estimated for the model. The disadvantages of SDA include a large computational requirement, a potentially complex function output, and no guarantee that the GP will find the optimal solution. To begin addressing these shortcomings, Moore et al. [17] outlined a 5-step SDA method for the automated detection, characterization, and interpretation of epistasis in population-based data. The first step in this method employs a full factorial experimental design to optimize search parameters for running SDA. One of the key parameters is the selection of the function set building blocks to make available to SDA, including arithmetic operators $(+, -, *, /)$, relational operators $(=, !=, <, >, <=, >=, max, min)$ and Boolean operators (AND, OR, NOT, IF, XOR).

In the present study we evaluate "masks" as a novel function set of building blocks. We use the 5-step SDA method as a framework to test the ability of masks to facilitate modeling of epistatic interactions. A mask function set is made up of two-locus interaction models intended to provide SDA with a pre-constructed relationship between pairs of explanatory variables or attributes. There are 512 possible two-locus, two allele, two-phenotype, fully penetrant disease models [12]. Beyond providing the GP with the entire set of 512 unique masks, we explore the utility of smaller, simpler mask function sets assembled by (1) selecting only the models which reduce the genetic redundancy present in entire set, (2) selecting models believed to be of general bi-

GECCO'08, July 12–16, 2008, Atlanta, Georgia, USA.

$\{f_{ij}\} =$

Locus 2 \ Locus 1	AA	Aa	aa
BB	f_{11}	f_{12}	f_{13}
Bb	f_{21}	f_{22}	f_{23}
bb	f_{31}	f_{32}	f_{33}

Figure 1: A generalized 3-by-3 penetrance table. The row label gives the three possible genotypes of the first disease locus (i.e. AA, Aa, aa) and the column label gives the genotypes for the second locus (i.e. BB, Bb, bb). The table element (f_{ij}) (penetrance) is the probability of being affected with the disease when the genotype at the first locus is i, and that of the second locus is j.

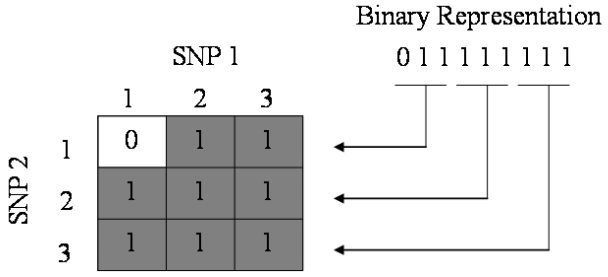

Figure 2: In this mask representation of the codominant model, single nucleotide polymorphisms (SNPs) are the loci of interest and the specific SNP polymorphism (1-3) represents the genotypes (i.e. AA, Aa, aa). The values (1-3) are arbitrarily assigned to code for a specific combination of the two SNP alleles an individual possesses at a given loci. For example 1 = G/G, 2 = G/T, 3 = T/T. The resulting 3-by-3 matrix of discrete outcomes can be represented as a 9-digit binary number where the digits are ordered by position as indicated by the arrows.

ological interest, and (3) selecting models which are customized to interactions detected in the given dataset using Multifactor Dimensionality Reduction (MDR). The utility of mask building blocks was assessed for both simulated and real case/control datasets. Each dataset contained known 2 way non-additive attribute interactions combined with a single attribute main effect. The real case/control dataset has been previously studied [17, 31, 19] and most recently utilized to test and validate the 5-step SDA method itself [17]. Masks are designed to save SDA the time and effort of evolving complex interaction models by intelligently providing such relationships in a pre-constructed fashion such that the most difficult task left to the GP is the selection of the appropriate attributes. We aim to evaluate masks as a novel function set and to determine whether expert knowledge at the level of building block selection can improve the modeling of epistatic interactions.

2. METHODS

2.1 Genetic Masks

The idea behind a "mask" is to provide SDA with a pre-constructed interaction relationship between pairs of explanatory variables or attributes. A geneticist might view a "mask" as being a disease model involving two genetic loci. As discussed by Li et al. [12], these two-locus models have been widely used in the study of complex diseases since they are a natural choice if the underlying disease mechanism involves two or more genes. A two-locus model is typically represented by a 3-by-3 penetrance table where each cell represents a combination of two inherited genotypes from different loci (see Figure 1). Penetrance is defined as the probability of disease given a particular genetic state.

In the most general case, the penetrance (f_{ij}) of a given table cell ranges from 0-1. If the value of f_{ij} is limited to the discrete values $'0'$ (not at all penetrant) and $'1'$ (fully penetrant), we can categorize the nine-parameter space to $2^9 = 512$ distinct models. The implementation of masks as building blocks for the GP requires the functional translation of a two-locus model into a binary representation of the penetrance table. We use the following notation to label each of these 512 fully penetrant two-locus models:

$$'modelnumber'_{10} = (f_{11}f_{12}f_{13}f_{21}f_{22}f_{23}f_{31}f_{32}f_{33})_2$$

Where the subscript of 2 or 10 indicates whether the number is represented as binary or decimal. Now, each position in the penetrance table may be represented as a single digit within a nine digit binary number. Consider the following example: In this study, the discrete outcomes 0 and 1 represent healthy and disease state, respectively. For the two locus model which represents codominance, position f_{11} is the only combination where both genes have the recessive genotype so $f_{11} = 0$ and all other f_{ij}s equal one. Figure 2 indicates how the codominant model would be expressed as a mask in this study.

Considering this representation, if an individual has genotype = 1 for both SNP 1 and SNP 2, the mask function would output the value $'0'$, while for any other combination of codes it would output a $'1'$. In model discovery, a mask "building block" is somewhat analogous to a simple black box. For a given subject, two SNP genotypes and the binary representation of a mask make up the input values which yield a discrete output of zero or one. We can view this as a form of constructive induction, where the genotypes of two SNP attributes are being combined into a single discrete attribute. If the disease model was comprised of a mask function alone, this output would correspond to the prediction of disease status (zero = no disease, one = disease).

Determining what set of masks to provide SDA is an important concern. The ability of masks to improve SDA modeling would most likely depend on the subset of masks selected to make up the function set. The first mask set tested in this study consists of all 512 possible models (M_512). Li et al. [12] classified progressively smaller groups of two-locus models, characterized by a step-wise reduction in the redundancy of interaction information which is present when all 512 models are considered. Here, we discuss how three sub-

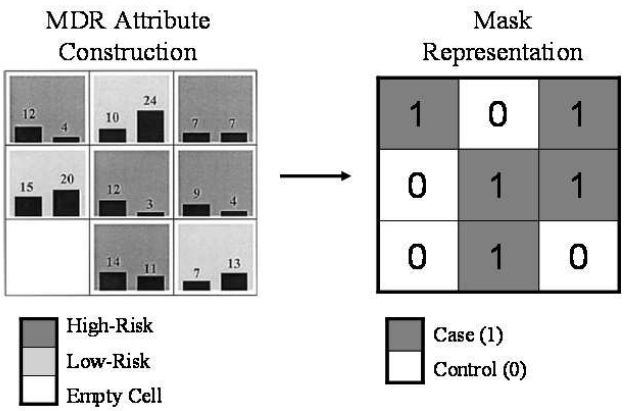

Figure 3: The bars in the MDR attribute construction represent the number of cases(left bar) and controls (right bar) in the dataset that possess a given combination of genotypes. Empty cells represent genotype combinations not found in the dataset. For simplicity, the are treated as low-risk/control cells.

groups were obtained each of which were tested in this study. The first step towards constructing intelligent mask subsets involves the removal of all zero and single locus mask models which are of no value in representing two-locus interaction. Mask models were excluded if all f_{ij}s in the matrix were either zero or one, or if f_{ij}s did not change with row or with column. The second mask set in this study consisted of 96 models (M_96) obtained by additionally removing models where the first and second locus could be exchanged, or where two alleles within the first (or second) locus could be exchanged. This was done to remove positional redundancy. The third mask set, consisting of 48 models (M_48), additionally removed masks from the 96 mask subset that represented an inverse of the affection status ($0 = 1, 1 = 0$). The fourth mask set consisted of 6 models (M_6) defined and studied by Neuman and Rice [26], singled out as being of being of potential biological interest. They include a recessive-recessive, dominant-dominant, recessive-dominant, modifying-effect, threshold, and exclusive OR model. We will evaluate whether the availability of these different mask functions improves SDA modeling of complex genetic relationships.

2.2 Mask Selection by Multifactoral Dimensionality Reduction (MDR)

A potentially superior approach to the implementation of masks would require a quick and simple data pre-processing step. Multifactoral dimensionality reduction (MDR) will conveniently construct and evaluate all possible two-locus interaction models for each pair of attributes in the dataset. From this, any number of best two-locus interaction models may be selected to make up the mask set. In essence, MDR allows us to select a set of masks customized to a given dataset. MDR was developed as a nonparametric, model-free data mining strategy for identifying combinations of SNPs that are predictive of a discrete clinical endpoint [19, 28, 15, 16]. The MDR method has been successfully applied to detecting gene-gene interactions for a variety of common human diseases including adverse drug reactions [32]. While MDR was designed to perform attribute construction for any reasonable number of loci, the following explanation of MDR will be given from the perspective of constructing two-locus models. At the heart of the MDR approach is an attribute construction algorithm that creates a new attribute by pooling genotypes from any given pair SNPs. Constructive induction using the MDR kernel is accomplished in the following way. Given a threshold T, a two-locus genotype combination (eg. AABb) is considered high-risk if the ratio of cases (subjects with disease) to controls (healthy subjects) exceeds or equals T, otherwise it is considered low-risk. Genotype combinations considered to be high-risk are labeled $G1$ while those considered low-risk are labeled $G0$. This process constructs a new one-dimensional attribute with levels $G0$ and $G1$ which represent an output of 0 or 1 respectively when formatted as a mask (Figure 3). The MDR method is described in more detail by Hahn et al. [4].

Using MDR, three mask sets were assembled and evaluated in addition to the ones previously described in section 2.1. Using the landscape feature of MDR the prediction accuracies for all possible two-locus combinations can be assessed to rapidly rank all two-locus models. Prediction accuracy is simply the percentage of subjects whose status is correctly predicted by a given model. From this ranked list, the first MDR mask set included only the single best two way interaction model (MDR_1), the second included the 5 best interaction models (MDR_5), and the third included the 10 best interaction models (MDR_10). In all of the above cases, only the interaction framework information was captured by the mask representation. In other words, the mask representation did not retain the knowledge of which attributes were used to construct that framework. This was done so that the GP could assign any pair of attributes to be the input loci for any given mask.

2.3 Symbolic Discriminant Analysis

Symbolic discriminant analysis uses symbolic regression to generate models from which symbolic discriminant scores are generated so that classification error can be estimated for the model. SDA is able to automatically identify an optimal functional form and coefficients of discriminant functions that may be linear or nonlinear [24, 21, 17, 27, 20, 14]. This is accomplished by providing a list of mathematical functions and a list of explanatory variables that can be used to build discriminant scores. Here, GP is used to perform a parallel search for a combination of functions and variables that optimally discriminates between two endpoint groups. GP permits the automatic discovery of symbolic discriminant functions that can take any form defined by the functions provided.

There are two key advantages of SDA over traditional multivariate methods. First, SDA does not pre-specify the functional form of the model. The basic mathematical building blocks (and mask functions examined in this paper) are defined and then flexibly combined with explanatory variables to derive the best discriminant function. The second advantage of SDA is the automatic and unbiased selection of variables from a potential list of thousands. This differs from traditional model fitting which involves stepwise procedures that enter a variable into the model and then keep it in the model if it has a statistically significant marginal or independent main effect [25].

2.4 Genetic Programming

Genetic programming is an automated computational discovery tool that is inspired by Darwinian evolution and natural selection [6, 7, 9, 8, 1, 11, 10]. The goal of GP is evolve computer programs to solve problems. This is accomplished by first generating random computer programs that are composed of the building blocks needed to solve or approximate a solution to a problem. Each randomly generated program is evaluated and the good programs are selected and recombined to form new computer programs. This process of selection based on fitness and recombination to generate variability is repeated until a best program or set of programs is identified. The advantage of GP and other evolutionary computing algorithms is that they carry out a parallel or beam search of the fitness landscape by considering hundreds or thousands of solutions simultaneously. Recombination makes it possible to sample multiple peaks in a rugged fitness landscape. In the present study, symbolic discriminant functions are represented in the computer as expression trees. Each node in the tree is a function while each leaf in the tree is either an attribute or a constant. Constants made available for GP included $(-2, -1, 0, 1, 2)$[17]. The fitness of a tree is measured by the accuracy of the symbolic discriminant function applied to a dataset. Variability is introduced at each generation by randomly recombining or swapping pieces of trees and by introducing random mutations. Here, we used a fixed recombination frequency of 0.9 and a fixed mutation frequency of 0.01. Selection of trees during evolution was carried out using a three-way tournament. With this approach, three trees or models are randomly selected with replacement from the population. The tree with the best fitness then becomes a candidate for recombination and/or reproduction. Parameters such as the population size, the number of generations, the function set, and the depth of the trees were all optimized using a full factorial experimental design.

2.5 Cross Validation Strategy

SDA is a powerful and flexible modeling strategy. However, like any supervised machine learning method, SDA is susceptible to overfitting [14]. Here, we employed a three-way cross validation (CV) strategy that is similar to the approach described by Rowland [29]. With this CV strategy the data are randomly divided into three equal parts labeled training, testing, and validation. Here, the best n SDA models from a single GP run are selected based on their accuracy in the training set. These n models are then used to make predictions in the testing set. The n models are sorted based on their testing accuracy and the best model selected. The single best SDA model is then evaluated using the validation set. The validation accuracy is a measure of the generalizability of the best model. In this study we set $n = 20$. Once the final best model for any particular run was selected, we reported the average of the training, testing, and validation error for that model. This prevents spurious results due to unusual chance partitions of the data. It is this average accuracy that is used to compare best models across different SDA runs.

2.6 Symbolic Modeling with SDA

The initial parameter sweep step of the 5-step SDA method was utilized in this study as a platform to test the ability of masks to contribute to the success of building models directed at detecting, characterizing, and interpreting epistasis [17]. Since no one stochastic search algorithm is optimal for every fitness landscape, the 5 step method aims to conduct an intelligent search of the fitness landscape under an optimal set of parameters for the GP. The five steps include (1) employment of a full factorial experimental design to optimize search parameters, (2) carrying out a coarse-grained search using genetic programming (GP), (3) generating expert knowledge by statistically modeling the best solutions, (4) carrying out a fine-grained stochastic search using an estimation of distribution algorithm based on what is learned in step three, and (5) using function mapping and interaction dendrograms to interpret symbolic models. This study will utilize the first step as a rapid evaluation of masks for solving the complex genetic modeling problem with and without the availability of a mask function set. The 5-step SDA method is described in more detail by Moore et al. [17].

Table 1: Summary of the function sets available to the GP.

Set	Consists of...
1	Arithmetic $(+, -, *, /)$
2	Relational $(=, !=, <, >, <=, >=, max, min)$
3	Arithmetic and Relational
4	Boolean $(AND, OR, NOT, NOR, IF, XOR)$
5	Arithmetic and Boolean
6	Relational and Boolean
7	Arithmetic, Relational, and Boolean
8	The set of "Masks" being evaluated.
9	Arithmetic and Masks
10	Relational and Masks
11	Arithmetic, Relational, and Masks
12	Boolean and Masks
13	Arithmetic, Boolean, and Masks
14	Relational, Boolean, and Masks
15	Arithmetic, Relational, Boolean, and Masks

The goal of Step 1 is to determine the optimal parameter settings for the GP using a full factorial experimental design. We considered population sizes of 100, 500, and 1000 trees, generation lengths of 100, 500, and 1000 iterations, and tree depths of one, two, and three. In addition, we considered fifteen different combinations of mathematical functions and masks. Table 1 identifies seven different sets of mathematical functions used previously in SDA modeling (function sets 1-7) along with eight new function sets which incorporate "masks" (function sets 8-15). These four parameters yield a total of 405 level combinations. For each level combination we ran the GP using 10 different random seeds and for each run recorded the best model along with its average accuracy. A total 4050 runs were performed. The average model prediction accuracies were obtained for all 270 runs representative of each function. The function set with the highest average accuracy score within sets 1-7 and 8-15 were selected as the best "no mask" and "mask" function sets, respectively and a t-test for independent samples was used to detect significant differences in the score distributions. To determine which of the parameters were a significant predictor of GP performance we employed a four-way ANOVA for fixed effects. Tukey's HSD was used for posthoc analysis to

Table 2: Key findings and parameters.

Dataset	Mask?	Best Mask Set	Average Accuracy	FS	Depth
XOR-R	No		0.641786	6	3
XOR-R	Yes	MDR_5	0.671864	12	3
XOR-D	No		0.670635	3	2
XOR-D	Yes	MDR_5	0.691973	9	2
AF	No		0.600094	1	3
AF	Yes	MDR_512	0.606329	13	3

determine the optimal settings for the population size, generation length, and tree depth parameters. All results were considered statistically significant at a type I error rate of 0.05.

2.7 Datasets

Two simulated datasets named XOR-R and XOR-D were tested in this study along with a real case/control dataset. Both simulated datasets contain 10 SNP attributes across 200 cases and 200 controls. XOR-R is a three locus model combining a nonlinear interaction that is not linearly separable (XOR), with a recessive main effect model (R). XOR-D is a three locus model combining a nonlinear interaction that is not linearly separable (XOR), with a dominant main effect model (D). These two locus models are drawn directly from the enumerated set of models described by Li et al. [12]. These models share a heritability of approximately 0.05 which is within the range that might be expected for a common, complex disease in which not all susceptibility factors are accounted for. The simulation and dataset assembly methods are further discussed by Moore et al. [19].

The real case/control data set comes from a study by Tsai et al. [31] which analyzed 250 patients with documented nonfamilial structural atrial fibrillation (AF) and 250 controls that were matched to cases on a 1-to-1 basis with regard to age, gender, presence of left ventricular dysfunction, and presence of significant valvular heart disease. The ACE gene insertion/deletion (I/D) polymorphism, the T174M, M235T, G6A, A-20C, G152A, and G217A polymorphisms of the angiotensinogen gene, and the A1166C polymorphism of the angiotensin II type I receptor gene were included as attributes. Moore et al. [17] utilized this dataset to evaluate the 5-step framework for modeling with SDA.

2.8 Software and Hardware

The SDA, and GP algorithms were both programmed in Java as part of an open-source Symbolic Modeler software package available by request[18]. Perl was used to call the Java programs multiple times and parse the output. All statistical analyses were performed using STATISTICA. Open-source MDR software is freely available from www.epistasis.org. All GP runs were conducted in parallel using 100 processors from the DISCOVERY supercomputer at Dartmouth College (http://discovery.dartmouth.edu).

3. RESULTS

A parameter sweep was completed for each of the 7 mask sets discussed (M_512, M_96, M_48, M_6, MDR_1, MDR_5, MDR_10) and was repeated for all three datasets described above. In all, 21 parameter sweeps were performed. Figure 5 gives an example of the average prediction accuracies for each function set over a single parameter sweep. In this example, Tukey's HSD post-hoc analysis indicated that all "mask" function sets yielded significantly higher average prediction accuracies than function sets without masks. Additionally, a t-test for independent samples, comparing the best "mask" and "no mask" function set accuracies from Figure 4 indicates that the "mask" function set yields models with significantly higher average prediction accuracies ($P << 0.001$). For each of the 21 parameter sweeps the best "mask" and "no mask" function sets were identified as previously described (data not shown). As would be expected, the same "no mask" function set was consistently found to be the best within each dataset analysis (data not shown). Figure 5 compares the different mask sets tested for each dataset examined. From this figure it is clear that the availability of masks in modeling the given datasets tended to improve the overall average prediction accuracies. Analysis of the XOR-R dataset indicates that the presence of mask sets M_512, M_96, M_48, MDR_1, MDR_5, and MDR_10 each yield significantly higher average model prediction accuracies when compared to the best function set with no masks ($P << 0.001$). Analysis of the XOR-D dataset indicates that the presence of mask sets M_96, M_48, MDR_1, MDR_5, and MDR_10 each yield significantly higher average model prediction accuracies than without masks ($P << 0.001$). Analysis of the real atrial fibrilation dataset indicates that the presence of mask sets M_512, M_96, M_48, MDR_5, and MDR_10 each yield significantly higher average model prediction accuracies than without masks ($P < 0.05$). The only case where a mask set performed significantly worse ($P < 0.05$) was mask set M_6 for the XOR-D dataset. The key results taken from the 21 parameter sweeps performed are summarized in Table 2, where the best mask set indicates which mask set performed best for the given dataset and the average accuracy indicates the average model accuracy across all 270 runs of the best function set indicated. The optimized GP parameters of

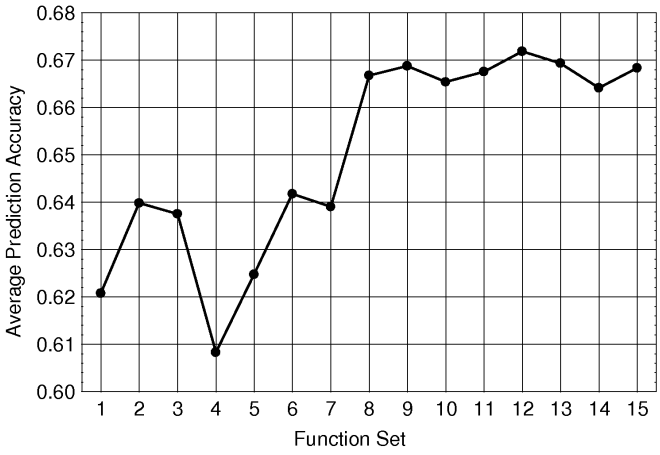

Figure 4: Average prediction accuracies for each function set over a single parameter sweep. Function sets 1-7 represent those without masks available, while 8-15 represent those with masks available. The dataset and mask set represented in this figure include XOR-R and MDR_5 respectively.

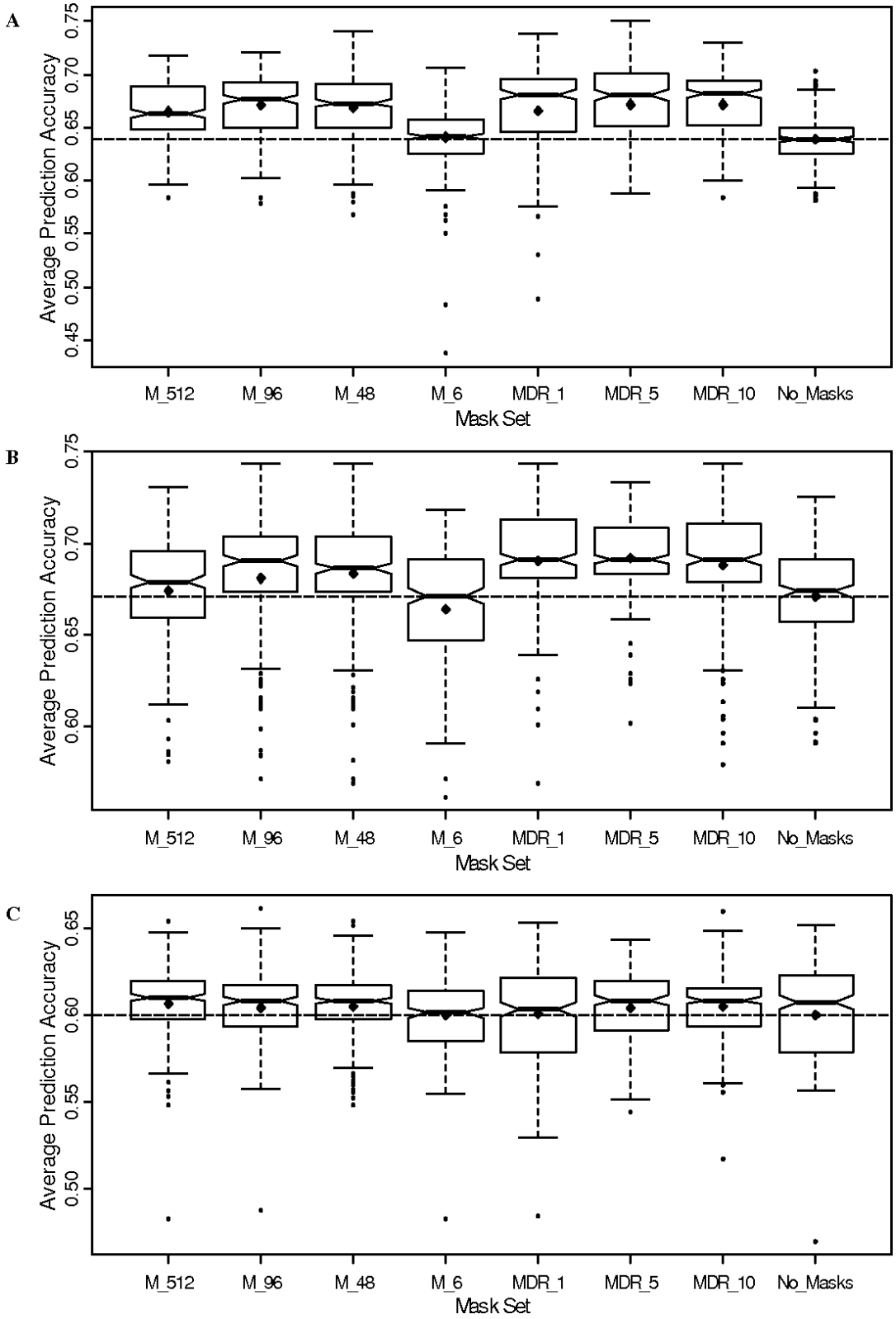

Figure 5: A comparison of the average prediction accuracy distributions for the A.) XOR-R dataset, B.) XOR-D dataset, and C.) atrial fibrilation dataset. Each box-plot represents a distribution of 270 prediction accuracy values for the function set selected as the "best" for a given combination of dataset and mask set. The diamonds represent the average prediction accuracy for the respective distribution. The horizontal dotted line gives the average prediction accuracy "no mask" distribution to which all others are compared.

function set (FS) and depth are also indicated in Table 2. The optimized number of generations and population size were both found to be 500 for each of the three datasets. For both of the simulated dataset analyses, the availability of mask set MDR_5 yielded the highest average model prediction accuracy. For the atrial fibrillation dataset analysis, the availability of mask set MDR_5 was again successful in significantly improving average prediction accuracy, but the availability of mask set M_512 yielded the highest average prediction accuracy. The parameters identified in Table 2 can be used in a course grain search as part of the 5-step approach described by Moore et al. [17].

4. DISCUSSION AND CONCLUSION

The introduction of masks as a novel function set represents the application of expert knowledge at the basic level of building block selection. The value of incorporating expert knowledge in GP has been examined from a number of perspectives [22, 23, 3]. Success in these studies, along with simple intuition suggest that by utilizing all available knowledge about a problem, a better solution might be more quickly identified in the seemingly infinite search space. While Reif et al. [27] examined the ability of complex function sets such as square, square root, sine, and cosine to improve SDA modeling, masks make up the most complex and customizable building blocks assessed to date.

This introductory evaluation of masks has utilized a parameter sweep as a test platform and average prediction accuracy as a metric of comparison. There are a number of observations and conclusions to draw from this study. First, the availability of mask building blocks successfully improved the average prediction accuracies of complex disease models generated for both simulated and real datasets. It should be noted that while this improvement is statistically significant, it is a relatively small improvement in terms of number of samples correctly classified. Table 2 reflects a classification improvement of approximately 3%, 2%, and 0.6% for datasets XOR-R, XOR-D, and AF respectively following the introduction of masks. The argument could be made that any improvement in the ability to model and correctly classify patients is of value, offering greater evidence that attributes in the model are markers of or contributors to disease. For the purposes of this study it important to note simply that function sets including masks performed as well if not better than other function sets in the framework used to test them. As such, masks represent a customizable alternative to standard function sets which may prove to have a considerably larger impact on success in other problem spaces yet to be explored.

Almost all mask sets were successful in improving average prediction accuracy in a given dataset with the exception of M_6 which consistently performed poorly with respect to all other mask sets. This would intuitively suggest that the ability of mask building blocks to improve model prediction accuracy is dependent on the inclusion of masks correctly representing trends found in the dataset. While the models selected for M_6 were not valuable in modeling the datasets examined, the utility of these models can not be discounted when considering other complex problems.

The mask sets which performed consistently better than no mask sets across all datasets examined included M_96, M_48, MDR_5, and MDR_10. In the case of M_96 and M_48 it seems that valuable genetic interaction relationships were preserved after reducing the redundancy present from all 512 possible models as described by Li et a. [12]. A function set with fewer masks results in a smaller search space of potential combinations of building blocks. In the interest of saving GP time in the construction of models, it is advantageous to keep function sets as small as possible. In the case of MDR_5 and MDR_10 which represent mask sets customized to interactions detected in the dataset, it seems we have identified an efficient method for the exploitation of expert knowledge in the construction of building blocks.

MDR masks perform better or as well as other mask sets, indicating that the simplest and most effective implementation of masks involves the customized construction of a mask set using MDR. While the number of "best" MDR models to include in mask sets was chosen arbitrarily, the results suggest that inclusion of only the single best model may limit the success of MDR-masks. Alternatively, for the datasets analyzed, MDR_10 showed no significant improvement over MDR_5 suggesting that while greater than one best two-locus model was valuable there is some threshold at which the addition of further two-locus MDR models fails to be of any benefit. This threshold is likely at or around 5 two-locus models as indicated by the success of MDR_5.

While we have addressed the question of whether masks improve average model prediction accuracy in an SDA parameter sweep, the potential advantage of masks extends much further. We hypothesis that utilizing genetic masks as model building blocks will additionally hasten and simplify the interpretability of complex interaction models for the prediction of disease state. Since masks represent a preconstructed, two-locus interaction framework, we intuitively expect that the interpretability of models including them will be improved. To pursue the potential value of masks further, our future work will include continued assessment of the above datasets utilizing the optimized parameters indicated in Table 2. The 5-step analysis outlined by Moore provides the framework for this continued assessment.

5. ACKNOWLEDGMENTS

This work was supported by NIH grants LM009012 and AI59694.

6. REFERENCES

[1] W. Banzhaf, P. Nordin, R. E. Keller, and F. D. Francone. *Genetic programming: an introduction: on the automatic evolution of computer programs and its applications*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1998.

[2] R. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936.

[3] C. Greene, B. White, and J. Moore. An Expert Knowledge-Guided Mutation Operator for Genome-Wide Genetic Analysis Using Genetic Programming.

[4] L. Hahn, M. Ritchie, and J. Moore. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions, 2003.

[5] J. Koza. The Genetic Programming Paradigm: Genetically Breeding Populations of Computer Programs to Solve Problems. *Dynamic, Genetic, and Chaotic Programming: The Sixth-Generation*, 1992.

[6] J. R. Koza. *Genetic programming: on the programming of computers by means of natural selection*. MIT Press, Cambridge, MA, USA, 1992.

[7] J. R. Koza. *Genetic programming II: automatic discovery of reusable programs*. MIT Press, Cambridge, MA, USA, 1994.

[8] J. R. Koza. *Genetic Programming IV: Routine Human-Competitive Machine Intelligence*. Kluwer Academic Publishers, Norwell, MA, USA, 2003.

[9] J. R. Koza, D. Andre, F. H. Bennett, and M. A. Keane. *Genetic Programming III: Darwinian Invention & Problem Solving*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999.

[10] W. B. Langdon and R. Poli. *Foundations of Genetic Programming*. Springer-Verlag, 2002.

[11] W. B. Langdon and K. J. R. *Genetic Programming and Data Structures: Genetic Programming + Data Structures = Automatic Programming!* Kluwer Academic Publishers, Norwell, MA, USA, 1998.

[12] W. Li and J. Reich. A Complete Enumeration and Classification of Two-Locus Disease Models. *Human Heredity*, 50:334–349, 2000.

[13] R. Lipshutz, S. Fodor, T. Gingeras, D. Lockhart, et al. High density synthetic oligonucleotide arrays. *Nature Genetics*, 21(Suppl 1):20–24, 1999.

[14] J. Moore. Cross Validation Consistency for the Assessment of Genetic Programming Results in Microarray Studies. *Applications of Evolutionary Computing: EvoWorkshops 2003: EvoBIO, EvoCOP, EvoIASP, EvoMUSART, EvoROB, and EvoSTIM, Essex, UK, April 14-16, 2003: Proceedings*, 2003.

[15] J. Moore. Computational analysis of gene-gene interactions using multifactor dimensionality reduction. *Expert Rev Mol Diagn*, 4(6):795–803, 2004.

[16] J. Moore. Genome-wide analysis of epistasis using multifactor dimensionality reduction: feature selection and construction in the domain of human genetics. *Knowledge Discovery and Data Mining: Challenges and Realities with Real World Data*, 2006.

[17] J. Moore, N. Barney, C. Tsai, F. Chiang, J. Gui, and B. White. Symbolic Modeling of Epistasis. *Hum Hered*, 63(2):120–133, 2007.

[18] J. Moore, N. Barney, B. White, R. Riolo, T. Soule, and B. Worzel. Solving Complex Problems In Human Genetics Using. *Genetic Programming Theory and Practice {V}*, pages 69–86.

[19] J. Moore, J. Gilbert, C. Tsai, F. Chiang, T. Holden, N. Barney, and B. White. A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *Journal of Theoretical Biology*, 241(2):252–261, 2006.

[20] J. Moore and J. Parker. Evolutionary computation in microarray data analysis. *Methods of Microarray Data Analysis*, 2002.

[21] J. Moore, J. Parker, N. Olsen, and T. Aune. Symbolic discriminant analysis of microarray data in autoimmune disease. *Genetic Epidemiology*, 23(1):57–69, 2002.

[22] J. Moore and B. White. Exploiting expert knowledge in genetic programming for genome-wide genetic analysis. *Lecture Notes in Computer Science*, 4193:969–977, 2006.

[23] J. Moore and B. White. Genome-wide genetic analysis using genetic programming: The critical need for expert knowledge. *Genetic Programming Theory and Practice IV. New York, Springer*, 2006.

[24] J. H. Moore, J. S. Parker, and L. W. Hahn. Symbolic discriminant analysis for mining gene expression patterns. *Lecture Notes in Artificial Intelligence*, 2167:191–205, 2001.

[25] J. Neter, W. Wasserman, and M. Kutner. *Applied Linear Statistical Models: Regression, Analysis of Variance, and Experimental Designs*. Irwin Homewood, IL, 1990.

[26] R. Neuman and J. Rice. Two-locus models of disease. *Genet Epidemiol*, 9(5):347–65, 1992.

[27] D. Reif, B. White, N. Olsen, T. Aune, and J. Moore. Complex function sets improve symbolic discriminant analysis of microarray data. *Lecture Notes in Computer Science*, 2724:2277–2287.

[28] M. Ritchie, L. Hahn, N. Roodi, L. Bailey, W. Dupont, F. Parl, and J. Moore. Multifactor-Dimensionality Reduction Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer. *The American Journal of Human Genetics*, 69(1):138–147, 2001.

[29] J. Rowland. Model selection methodology in supervised learning with evolutionary computation. *Biosystems*, 72(1-2):187–196, 2003.

[30] M. Schena, D. Shalon, R. Davis, P. Brown, et al. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science(Washington)*, 270(5235):467–470, 1995.

[31] C. Tsai, L. Lai, J. Lin, F. Chiang, J. Hwang, M. Ritchie, J. Moore, K. Hsu, C. Tseng, C. Liau, et al. Renin-Angiotensin System Gene Polymorphisms and Atrial Fibrillation, 2004.

[32] R. Wilke, D. Reif, and J. Moore. Combinatorial pharmacogenetics. *Nat Rev Drug Discov*, 4(11):911–8, 2005.

APPENDIX

A. ADDITIONAL AUTHORS

Nate Barney (Dartmouth College, email:natebarney@gmail.com)