Alternative Cross-Over Strategies and Selection Techniques for Grammatical Evolution Optimized Neural Networks

Alison A. Motsinger
Center for Human Genetics Research, Department of Molecular Physiology & Biophysics, Vanderbilt University, Nashville, TN, USA 37232
1-615-322-0834
motsinger@chgr.mc.vanderbilt.edu

Lance W. Hahn
Center for Human Genetics Research, Department of Molecular Physiology & Biophysics, Vanderbilt University, Nashville, TN, USA 37232
1-615-343-8616
hahn@chgr.mc.vanderbilt.edu

Scott M. Dudek
Center for Human Genetics Research, Department of Molecular Physiology & Biophysics, Vanderbilt University, Nashville, TN, USA 37232
1-615-343-8616
dudek@chgr.mc.vanderbilt.edu

Kelli K. Ryckman
Center for Human Genetics Research, Department of Molecular Physiology & Biophysics, Vanderbilt University, Nashville, TN, USA 37232
1-615-343-6549
ryckman@chgr.mc.vanderbilt.edu

Marylyn D. Ritchie
Center for Human Genetics Research, Department of Molecular Physiology & Biophysics, Vanderbilt University, Nashville, TN, USA 37232
1-615-343-6549
ritchie@chgr.mc.vanderbilt.edu

Categories and Subject Descriptors
I.5.1 [**Pattern Recognition**]: Models – *neural nets.*

General Terms: Algorithms

Keywords: Grammatical evolution, neural networks, crossover, selection

1. INTRODUCTION

One of the most difficult challenges in human genetics is the identification and characterization of susceptibility genes for common complex human diseases. The presence of gene-gene and gene-environment interactions comprising the genetic architecture of these diseases presents a substantial statistical challenge. As the field pushes toward genome-wide association studies with hundreds of thousands, or even millions, of variables, the development of novel statistical and computational methods is a necessity. Previously, we introduced a grammatical evolution optimized NN (GENN) to improve upon the trial-and-error process of choosing an optimal architecture for a pure feed-forward back propagation neural network. GENN optimizes the inputs from a large 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 automatically generates optimal neural network architecture for a given data set.

Like all evolutionary computing algorithms, grammatical evolution relies on evolutionary operators like crossover and selection to learn the best solution for a given dataset. We wanted to understand the effect of fitness proportionate versus ordinal selection schemes, and the effect of standard and novel crossover strategies on the performance of GENN.

2. METHODS

2.1 Grammatical Evolution Neural Networks (GENN)

Details of the GENN method has previously been described in detail in Motsinger et al 2006 [1].

2.2 Selection Techniques

There are two main classes of selection techniques: fitness proportionate and ordinal selection. For this study, we wanted to test the impact these two types of selection have on the performance of GENN. To compare fitness based selection to ordinal based, tournament selection was tested and compared to roulette wheel selection for its effect on the performance of GENN.

2.3 Crossover Strategies

One criticism of GE is the use of a seemingly destructive single-point crossover operator. To address this concern, our group has developed two alternative crossover strategies that more strictly maintain building blocks than standard one-point GA crossover.

Typically in a GA, a simple one-point crossover is used, where a crossover point is chosen on two binary strings (between codons), and corresponding segments of the string are swapped between the two parent strings. GENN was initially implemented using a standard two-point crossover during the GA. This method will be referred to as a "standard" crossover (Std.). The first new crossover strategy, "linear homology

GECCO'06, July 8–12, 2006, Seattle, Washington, USA.
ACM 1-59593-186-4/06/0007.

crossover" (L.H.) looks for matching (as defined functionally by the grammar) codons in the grammar. In the first step of this crossover, a site along the linear chromosome of Parent #1 is randomly selected and the codon at that site is translated by the grammar. A random point along the chromosome of Parent #2 is selected, and then the chromosome is scanned (randomly either left to right or right to left) and the codon transcribed by the grammar until a match is found for the codon on Parent #1. After a match is found, crossover occurs between these two matching codons. The second new method, in theory, preserves the building blocks more than either standard or linear homology crossover. This second new method, called "tree-based" crossover (T.B.), swaps functionally analogous trees. The linear genome is transcribed by the grammar, and the grammar is then translated into functional trees. Then functionally analogous branches (subtrees with identical root nodes) are identified, and crossover occurs between whole branches.

2.4 Data Simulation

The intention of the data simulations for this power study was to mimic gene-gene interaction, or epistasis, in case-control genetic data to evaluate GENN using penetrance functions. Penetrance defines the probability of disease given a particular genotype combination by modeling the relationship between genetic variations and disease risk. We simulated case-control data using models exhibiting interaction effects in the absence of main effects. Two different allele frequencies were chosen for our simulations (0.8/0.2 and 0.6/0.4). For each dataset, 100 SNPs were generated per individual, with 500 cases and 500 controls per dataset. A range of heritability (proportion of the total phenotype that is due to genetic effects) values was selected including 5%, 10%, 15%, 20%, and 25%. Datasets were simulated using software described by Moore et al 2002 [2]. All possible combinations of allele frequencies and heritability values were simulated, resulting in ten models. The penetrance functions used in this study are available from the authors upon request. One hundred datasets were generated per model. Dummy variable encoding was used for each dataset, where *n-1* dummy variables were used for *n* levels.

2.5 Data Analysis

The selection techniques and crossover strategy options were incorporated into GENN as options in the configuration file. GENN was then used to analyze all 10 epistasis models with all combinations of the two selection techniques and three crossover options. The other configuration parameter settings remained identical between the analyses and included: 10 demes, migration every 25 generations, population size of 200 per deme, 50 generations, crossover rate of 0.9, and a reproduction rate of 0.1.

3. RESULTS

Table 1 lists the power results for all ten epistasis models under the six different configuration combinations. Power was estimated as the percentage of times GENN correctly identified the correct model (with no false positive loci) over the hundred datasets per model. An ANOVA analysis comparing the results of the six different configuration indicated there is not a significant difference between the analyses (p=0.9853).

Table 1. Power (%) of GENN Analyses

Model		Configuration Parameters					
Minor Allele Freq.	Heritability	Roulette Wheel Selection			Tournament Selection		
		Std	L.H.	T.B.	Std	L.H.	T.B.
0.2	5%	99	99	99	98	99	99
0.2	10%	77	93	82	84	87	79
0.2	15%	55	53	62	66	66	64
0.2	20%	91	93	93	89	95	94
0.2	25%	84	85	74	77	88	76
0.4	5%	90	97	93	94	98	95
0.4	10%	98	100	99	97	99	100
0.4	15%	99	100	99	96	100	99
0.4	20%	94	92	95	98	94	99
0.4	25%	98	99	98	100	100	99

4. DISCUSSION

These results show that the performance of GENN is not significantly affected by the implementation of different crossover strategies or selection techniques. The relative equivalence of these results implies that even though the single-point standard crossover is frequently criticized for not maintaining building blocks during the evolutionary process, the characteristic is not a detriment to its performance. By forcing the maintenance of building blocks through different types of crossover strategies, no significant gain in performance is seen in this study.

5. ACKNOWLEDGMENTS

This work was supported by National Institutes of Health grants HL65962, GM62758, and AG20135.

6. REFERENCES

[1] Motsinger AA, Dudek SM, Hahn LW, and Ritchie MD. *Comparison of Neural Network Optimization Approaches for Studies of Human Genetics.* Lecture Notes in Computer Science, 3907: 103-114. 2006.

[2] Moore J, Hahn L, Ritchie M, Thornton T, White B. *Application of genetic algorithms to the discovery of complex models for simulation studies in human genetics.* Langdon, WB, Cantu-Paz, E, Mathias, K, Roy, R, Davis, D, Poli, R, Balakrishnan, K, Honavar, V, Rudolph, G, Wegener, J, Bull, L, Potter, MA, Schultz, AC, Miller, JF, Burke, E, and Jonoska, N. Proceedings of the Genetic and Evolutionary Algorithm Conference. 1150-1155. 2002. San Francisco, Morgan Kaufman Publishers.