# Optimization of Grammatical Evolution Decision Trees

Kristopher Hoover*, Rachel Marceau*, Tyndall Harris, Nicholas Hardison, David Reif, Alison Motsinger-Reif
North Carolina State University
Bioinformatics Research Center
Department of Statistics
919 515-3574
*Equal Contribution
{kmhoover, remarcea, tpharris, nhardis, dmreif, aamotsin}@ncsu.edu

## ABSTRACT
The detection of gene-gene and gene-environment interactions in genetic association studies presents a difficult computational and statistical challenge, especially as advances in genotyping technology have rapidly expanded the number of potential genetic predictors in such studies. The scale of these studies makes exhaustive search approaches infeasible, inspiring the application of evolutionary computation algorithms to perform variable selection and build classification models. Recently, an application of grammatical evolution to evolve decision trees (GEDT) has been introduced for detecting interaction models. Initial results were promising, but relied on arbitrary parameter choices for the evolutionary process. In the current study, we present the results of a parameter sweep evaluating the power of GEDT and show that improved parameter choices improves the performance of the method. The results of these experiments are important for the continued optimization, evaluation, and comparison of this and related methods, and for proper application in real data.

## Categories and Subject Descriptors
I.5.0 [**Computing Methodologies**]: Pattern Recognition

## General Terms
Performance

## Keywords
Grammatical evolution, decision trees, gene-gene interactions, human genetics, parameter sweep, genetic association studies

## 1. INTRODUCTION
Grammatical evolution decision trees (GEDT) was previously introduced as a new method to detect gene-gene and gene-environment interactions in genetic association studies (where the goal is to detect variants that predict common disease). GEDT uses grammatical evolution to evolve decision trees that classify disease status, and has been used on a limited range of simulations [1]. While initial results are promising, the initial parameter settings for the GE application have been arbitrary. We hypothesize that optimization of the GEDT approach through parameter sweeps of the evolutionary operators implemented will result in improved performance of the method. In the current study, we use simulated genetic data to evaluate the impact of different parameter settings for GEDT. The results of these sweeps will be crucial for the improved application of GEDT for continued methods development and application in real data.


*GECCO'11*, July 12–16, 2011, Dublin, Ireland.
ACM 978-1-4503-0690-4/11/07.

## 2. METHODS

### 2.1 Grammatical Evolution Decision Trees (GEDT)
The implementation of GE to evolve DTs has been previously described [2]. GE is used to perform variable selection in high-throughput data (Here, the input variables are single nucleotide polymorphisms, or SNPs) and optimize DTs that best classify the disease outcome. This is performed in a cross-validation framework, where classification error is the fitness function, and cross-validation consistency is used for final model selection.

### 2.2 Data Simulation
Penetrance functions are used to represent epistatic genetic models for simulation, where penetrance defines the probability of disease given a particular genotype combination by modeling the relationship between genetic variations (SNPs) and disease risk. For each individual subject within a dataset, a total of 100 SNPs were simulated, where two of the SNPs are associated with the outcome, and 98 are nuisance (noise) SNPs. Case-control data was simulated such that each dataset included 250 cases and 250 controls, and 100 datasets were generated for each model. We used a well-described epistatic model exhibiting interaction effects in the absence of main effects for the current study. This model, based on the nonlinear XOR function, was initially described by Li and Reich and genotypes were generated according to Hardy-Weinberg proportions [3]. In both models, p (the major allele frequency) = q (the minor allele frequency) = 0.5. GenomeSim software described by Dudek *et al* [4] was used to simulate the data. Additionally, to assess the false positive rate, 100 replicates of null data were simulated with no disease model. The same number of SNPs and individuals was generated, but case-control status was randomly assigned.

### 2.3 Data Analysis and Parameter Sweeps
GEDT was used to analyze each of the 200 simulated datasets described above. Each dataset was evaluated across a set of parameter sweeps. Table 1 lists the parameters considered in the current study and the values used in the sweep. There were 5 parameters considered, with the full combinatorics of all parameter values also implemented, resulting in a total of 72 parameter combinations used for analysis.

**Table 1. Parameter Values Used in the Parameter Sweep**

| Parameter | Possible Values |
|---|---|
| Population Size | 250, 500, 750 |
| Number of Generations | 250, 400, 550 |
| Crossover Rate | 0.8, 0.9 |
| Mutation Rate | 0.01, 0.05 |
| Selection method | Tournament; Roulette |

GEDT is implemented in C++ and Perl, and run on quad-core Core2 Xeon processors (8 processors, each at 3 GHz and with 4GB of memory). Software and user instructions are available from the following website: www4.stat.ncsu.edu/~motsinger.

Power for all analyses was estimated under each epistatic model as the number of times the algorithm correctly identified the functional (disease causing) loci as the top 2 loci based on the cross-validation criteria discussed above out of each set of 100 datasets, without any false positive loci.

## 3. RESULTS

The power of GEDT to detect the functional loci is visualized for each individual parameter across the sweeps as box and whisker plots in Figure 1. In each boxplot, the whiskers are defined as points at 1.25 times the interquartile range. These results show general trends that increased parameter values lead to higher performance. To test for significant differences in the parameter changes in power, multiple linear regression analysis was performed, testing the impact of each parameter setting across all results. The results indicate significant differences in all but one of the parameter settings. Mutation rate ($p<0.0001$), selection method type ($p<0.0001$), population size ($p<0.0001$), and generation number ($p<0.0005$) all make a significant difference in the power. There was not a significant difference in results based on the crossover rate ($p=0.5699$).

The false positive rate of the method (found by determining the power to detect the model in the null data) was nearly always zero, which indicates a very low false positive rate (results not shown). These results show that differences in the power results are not due to inflation of the false positive rate, or bias towards the specified model.

## 4. DISCUSSION

The results of the current study show the relative performance of GEDT to detect a purely interactive genetic model under a range of different parameter settings. A small increase in mutation rate makes a very big difference in power. Crossover rate was the only parameter which did not have a significant p-value, but there may be a ceiling effect in the current study. Population size is shown to be very significant, with what appears to be a linear increase over populations. There is also a very significant difference in selection methods - tournament greatly outperformed roulette. The number of generations parameter also shows a strong increasing trend with power, even though it was the least significant of the significant parameters. From the parameters swept, future studies with data of similar characteristics should consider the following parameter configuration for optimal power and minimal computational cost: population of 750, crossover of 0.8, mutation of 0.5, tournament selection, generation size of 550.

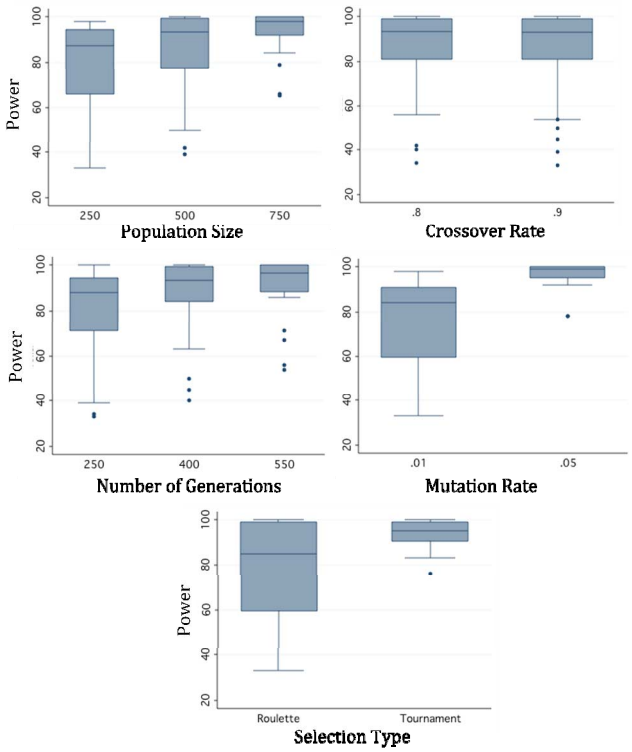


**Figure 1. Power Results for Each Parameter Swept. The box and whisker plots represent the distribution of power results across all other parameter values.**

These results show parameter settings for the current model and data structure, but should be extended for a wider range of data structures, particularly for larger datasets. Understanding how optimal parameter choices covary with different aspects of the data such as number of individuals, etc. will be important to understand as GEDT is applied to larger datasets.

## 5. ACKNOWLEDGMENTS

The research is based upon work supported by the National Science Foundation under CSUMS grant #DMS-0703392 (PI: Sujit Ghosh). The authors would like to thank the other participants of the Computation for Undergraduates in Statistics Program for their helpful input into the project.

## 6. REFERENCES

[1] O'Neill, M. and C. Ryan, Grammatical Evolution: Evolutionary automatic programming in an arbitrary language. 2003, Boston: Kluwer Academic Publishers.

[2] Motsinger-Reif, A.A., et al., Grammatical evolution decision trees for detecting gene-gene interactions. BioData Min, 2010. 3(1): p. 8.

[3] Li, W. and J. Reich, A complete enumeration and classification of two-locus disease models. Hum Hered, 2000. 50(6): p. 334-49.

[4] Dudek, S.M., et al., Data simulation software for whole-genome association and other studies in human genetics. Pac Symp Biocomput, 2006: p. 499-510.