Grammatical Evolution Decision Trees for Trio Designs

Amanda English*
Department of Statistics
North Carolina State University
Raleigh, NC 27607
1 919 515-3574
amengli2@ncsu.edu

Holly Petruso*
Department of Statistics
North Carolina State University
Raleigh, NC 27607
1 919 515-3574
hepetrus@ncsu.edu

Chong Wang*
Department of Statistics
North Carolina State University
Raleigh, NC 27607
1 919 515-3574
cwang19@ncsu.edu

ABSTRACT
The detection of gene-gene and gene-interactions in genetic association studies is an important challenge in human genetics. The detection of such interactive models 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 the previous applications of GEDT have been limited to case-control studies with unrelated individuals. While this study design is popular in human genetics, other designs with related individuals offer distinct advantages. Specifically, a trio-based design (with genetic data for an affected individual and their parents collected) can be a powerful approach to mapping that is robust to population heterogeneity and other potential confounders. In the current study, we extend the GEDT approach to be able to handle trio data (trioGEDT), and demonstrate its potential in simulated data with gene-gene interactions that underlie disease risk.

Categories and Subject Descriptors
I.6 **[Simulation and Modeling]**: Miscellaneous

General Terms
Algorithms, design, experimentation

Keywords
Grammatical evolution, decision trees, gene-gene interactions, human genetics, trio designs

1. INTRODUCTION
One of the primary goals of human genetics is identifying genetic and environmental risk factors of disease [1]. Modern genotyping technologies have enabled the researcher to readily incorporate large numbers of genetic variables (often single-nucleotide polymorphisms (SNPs)) into epidemiological studies of disease.

The large number of SNPs available (often hundreds of thousands or millions) creates important statistical and computational challenges [2]. These challenges are magnified by the complexity of the types of genetic models that are thought to cause these diseases. It is intuitive that common, complex diseases are due to a myriad of genetic and environmental factors, and by interactions of these factors [3]. Finding such interactive models is a particular challenge for both traditional statistical models and modern data-mining techniques. An excellent review of the challenges presented by such models can be found in [2].

To address these challenges, a number of new data mining approaches have been developed (REF). Many of these methods use evolutionary computation to build classifiers that predict disease, as well as to perform feature selection in large data [4]. A number of EC algorithms (including genetic algorithms (GA), genetic programming (GP), and grammatical evolution (GE)) have been used to optimize a range of classifiers (neural networks, naïve Bayes classifiers, etc.) to detect complex genetic models that predict human diseases [5-7]. Recently, we described an application of grammatical evolution (GE) to optimize decision trees (DT) to detect interactions in association studies that has shown initial successes [8].

These initial studies have been promising, but have all been focused on applications in a limited type of study design – case/control studies where unrelated individuals with and without a disease of interest are collected. While this is a very popular study design, other types of designs are often used and the GEDT approach needs to be extended to analyze data from other designs.

In the current study, we expand the previously described GEDT approach to trio-based designs. In the trio design, individuals with a disease are recruited along with their parents (both mother and father) and genetic data is collected [9]. There are several advantages and disadvantages to this type of design as compared to case/control designs [9]. Trio designs are more robust to the influence of many confounding factors (such as population stratification and environmental factors), but can be difficult to collect especially for late onset diseases where parents may not be available [9].

In the current study, we extend the GEDT design to trio data by generating "pseudo-controls" from parental data in the trio design, and then comparing variant frequencies in between cases and pseudo-controls (trioGEDT). We demonstrate this approach in simulated datasets exhibiting gene-gene interactions that predict disease. We show that trioGEDT is a promising new extension of the GEDT approach for trio designs.

2. METHODS
2.1 Trio Designs and Pseudo-Controls
The basic idea of a trio design is to collect an affected individual and their two parents. This basic design is diagrammed in Figure

GECCO'12 Companion, July 7–11, 2012, Philadelphia, PA, USA.

1, where a traditional pedigree visualization structure is shown (with males represented as squares, females as circles, and offspring shown as lower in a family tree). The SNP genotypes of all individuals are collected and can be used for association. If there is not association between a SNP and the disease, Mendel's law of segregation holds. This can be used to understand the expected frequencies of alleles in the affected offspring in the case samples collected. Mendel's law of segregation states that every individual possesses a pair of alleles (assuming diploidy) for any particular trait and that each parent passes a randomly selected copy (allele) of only one of these to its offspring. The offspring then receives its own pair of alleles for that trait. In the case of Figure 1, Mendel's law specifies that the probability of an AA genotype is the same as the probability of an Aa genotype. However Mendel's law does not hold for allele transmissions if individuals are selected for a trait, which is associated with the marker genotype. Thus, to form a test of association between disease and the marker, we compare the genotypes actually observed in the affected offspring to those expected when

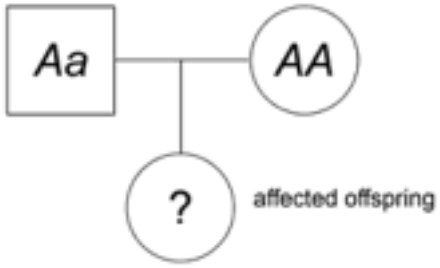

Figure 1. Allele Transmissions in Trios

Mendel's law holds. This idea can be formalized by considering all transmissions of a particular allele, say A, from an As parent to its affected offspring. Parents who are homozygous, AA or aa, contribute no information to the test, since the allele that will be transmitted is defined with absolute certainty by the parental genotype, e.g. an AA parent will always transmit an "A" allele to its offspring. These parents are non-informative. Each parent has a pair of alleles, one which is transmitted to the affected offspring, and one that is not. Table 1 summarizes the numbers of transmitted and non-transmitted alleles from each possible parental genotype. That is, there are w AA parents, all of whom transmitted A; likewise, aa parents transmitted an a. The two off-diagonals of the table are the heterozygous parents, Aa. Of those, x transmitted an A and y transmitted an a. Note that the table total is w+x+y+z=2n, where n is the number of offspring, and 2n is the number of parents. However, the homozygous parent does not contribute information to the test because they can only transmit one type of allele. If Mendel's Law of segregation holds, then for a heterozygote parent, the probability of transmitting A is the same as the probability of transmitting a, So the counts in the two off diagonal cells are equal in expectation. Any association method can then use this expectation as the null hypothesis and compare the observed allele counts to the expected. This approach is the basic concept behind the Transmission Disequilibrium Test (TDT) and its many successful derivatives [10].

To be able to use this concept in GEDT, we first generate cases and "pseudo-controls" as a preprocessing step to the data before running it through the GEDT process. In order to make the

Table 1. Multilocus penetrance function for XOR model.

	Parental Allele Not Transmitted		
		A	a
Transmitted	A	w	x
Parental Allele	a	y	z

comparison of the frequency of transmitted vs. untransmitted alleles, we scripted a preprocessing step to use the case data as collected, and to make a pseudo-control with the untransmitted alleles of each of the two parents in the trios. For example, in Figure 1, if the case had genotype "Aa", then the other alleles from the parents (one A allele from each) would form the untransmitted genotype. So one case with Aa and the pseudo-control with genotype AA would be used in further analysis. So for each trio, a total of two observations would be taken for analysis. By generating these pseudo-controls, the allele counts included in the analysis are equivalent to the contingency table in Table 1. After this pre-processing, the data can be run in GEDT as previously described.

2.2 Grammatical Evolution Decision Trees

Grammatical Evolution (GE) is a form of evolutionary computing that allows the generation of computer programs using grammars [11]. GE is inspired by the biological processes of transcription and translation, in which the genetic material (DNA) is transcribed into RNA and then the RNA is translated into a protein. By imitating the biological process of transcription and translation, GE separates "genotype" from "phenotype" and allows for greater diversity in populations than other methods. The heritable material in GE is the binary string, which is divided into codons (typically composed of 8 binary digits), and undergoes crossover and mutation. The binary chromosome string is transcribed into an integer string and these integer values are then translated by a mapping function into an appropriate production rule (a decision tree) from the grammar definition. GE is used to optimize the architecture and the recursive nature of a decision tree. The grammar used for GEDT has been previously described [8].

A decision tree is a hierarchical decision-making model that consists of internal decision nodes and terminal leaf nodes [12]. Internal decision nodes represent attributes of an individual, whereas leaf nodes represent the class the individual belongs to. The root node either corresponds to an initial criterion or an attribute of an individual. The nodes within the hierarchical structure are connected via directed edges. Each outgoing edge from an internal node corresponds to the value of the attribute that the node represents. Decision trees can model data that has non-linear relationships between variables, and are a 'white-box' approach in the sense that it is easy to interpret decision trees. Once a tree is built, it can be translated to IF-THEN statements.

The implementation of GE for the model building and variable selection for decision tree modeling has been previously described in detail [8]. Grammatical Evolution Decision Tree is a six-step process, outlined in Figure 2. These steps are briefly described below:

Initialization: GEDT has a set of parameters for the genetic algorithm, the cross-validation, training dataset, and testing

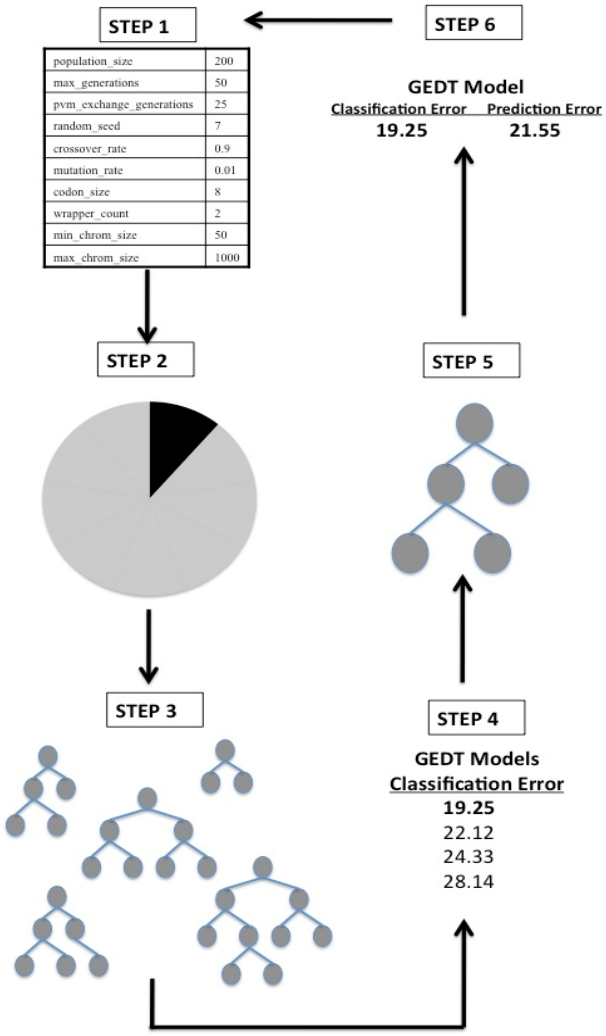

Figure. 2. An overview of the GEDT process that shows the six-step process of initialization, cross-validation, training, fitness evaluation using balanced error, natural selection (tournament) and testing (evaluating prediction error).

dataset, etc. These parameters must be initialized in the configuration file such as population size, maximum number of generations, crossover ratio, and mutation rate, etc.

Cross-Validation: the data are divided into 10 parts of equal size. 9/10 of the data is used to train the GEDT model, while the remaining 1/10 of the data is used to test the model and estimate the prediction error, which refers to the number of samples in the 1/10 of the data that are incorrectly classified using the GEDT model. Each time a different 1/10 of the data is left out for testing and the process is repeated 10 times. Cross-validation consistency is a measure of the number of times each variable appears in the GEDT model across 10 cross-validations. The consistency is used in the calculation of power results.

Training GEDT: an initial population of random solutions is generated, each one corresponding to a decision tree model. For every rule, the minimum depth required to complete the mapping process is calculated. Only those production rules whose minimum depth is less than or equal to the remaining allowed depth are chosen.

Fitness evaluation: each decision tree model is evaluated on the training set and its fitness is recorded. To avoid bias caused by using unbalanced data, we have used balanced accuracy in the analysis of the GEDT process. The fitness function is calculated as the average of sensitivity and specificity which makes the fitness measure standardized. Sensitivity is calculated as the ratio of the correctly classified case samples to the total number of case samples, whereas specificity is calculated as the ratio of corrected classified control samples to the total number of control samples.

Natural selection: in this step, the best solutions are selected for cross-over and reproduction using the tournament technique. Few individuals from the population are selected at random and tournament selection is run. The one with the best fitness is chosen to be the winner of the tournament and is then selected for the cross-over process. Duplication, mutation and cross-over are applied to winners to generate the next generation. The new generation is equal in size to the original population and this new generation begins the cycle again. The process continues until DT model with zero classification error or until a predefined limit on the number of generations is reached. At the end of this evolution cycle, the overall best solution is identified as the optimal DT model.

Testing: the 1/10 of the data remaining is used to test the overall best solution obtained in the previous step. The prediction error is calculated to measure the overall model fitness and the classification error is calculated to evaluate and adjust the architecture of the DT model. The overall goal of this learning process is to find the DT model that can not only accurately classify the data at hand, but can also predict on future data.

Steps ii) to vi) are performed ten times with the same configuration settings and the prediction error is estimated as an average across the 10 cross-validations.

2.3 Data Simulation

To evaluate the potential of GEDT when extended to trio data, a single purely epistatic/interactive genetic model was generated for the simulation experiment. In a purely epistatic model, no single genetic variant can solely predict a phenotype. Rather, a specific combination of two or more genes serves as the predictor of the phenotype. These epistatic models, also known as interaction models, are quickly becoming crucial to understanding the true etiology behind genetic disorders [13]. Penetrance functions are numerical, matrix-like representations of epistatic genetic models. Within the function, the penetrance represents the probability of disease given a specific genotype combination. In the current study, we use a fully penetrant model where all disease risk is explained by genetic variation. While this may be overly optimistic for a real disease scenario, this extreme model is important in testing the potential of the trioGEDT approach. Table 2 shows the specific penetrance function used in the study, which is a nonlinear XOR function, initially described by Li and Reich [14]. Each genetic variant that is modeled represents a single nucleotide polymorphism (SNP). In the current study, a total of 100 SNPs per individual were simulated. Two of these 100 SNPs are signal SNPs which predict the disease loci, while the other 98 SNPs serve as "noise" loci. This was done to test not only the model building potential of trioGEDT, but to also test its performance in regards to feature selection. All simulations were performed using GenomeSim software, originally described by Dudek *et* al [15]. A total of 100 datasets were generated under

this simulated model, so that we could assess the power of trioGEDT to detect the interactive disease causing SNPs. For each dataset, a total of 200 trios were simulated (so 200 cases and 400 parents that were then processed to datasets of 300 cases and 300 pseudo-controls).

Table 2. Penetrance function for XOR model.

	BB	Bb	Bb
AA	0	1.0	0
Aa	1.0	0	1.0
aa	0	1.0	0

2.4 Implementation

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 authors upon request, or linked from the following website: www4.stat.ncsu.edu/~motsinger. For these experiments, the parameter setting of GEDT included: population of 750, crossover of 0.8, mutation of 0.5, tournament selection, generation size of 550. These parameters were identified as optimal from a previous parameter sweep experiments.

3. RESULTS

The results of the simulations indicated that the trioGEDT approach could detect the correct loci in 99/100 of the simulated datasets. This corresponds to an empirical power of 99%. The correct model was missed in only one of the 100 datasets.

4. DISCUSSION

The results of this small-scale simulation show the potential of GEDT to be used for trio-based designs. The process of producing pseudo-controls and then analyzing this data with the GEDT approach was highly successful for this one simulated model. It is important to note that this simulation exhibited interactions without any main effects of individual SNPs. It is reasonable to assume that a method that does well with this extreme type of interaction would also perform well on models with main effects, though this assumption should be evaluated in future work. The work presented here is limited in its scope, but does provide an important proof of principle experiment that such an approach could be successful.

Future work should evaluate the performance of trioGEDT on a wider range of simulations. Future simulation experiments should evaluate the impact of different parameter choices on the performance of the method, as well as the performance of the method on a wider range of models. Genetic models to consider could include different modes of inheritance, different types of interaction models, etc. Additionally, it would be important to test the power of the method in a range of sample sizes. The computational aspects of the method should also be evaluated, and the method should be optimized to handle large-scale data.

Also, in any methods development work, it is important to compare the performance of the method to other similar approaches. The power should be compared to more traditional methods like the Transmission Disequilibrium Test (TDT) [9] as well as to more recent data-mining approaches such as the Multifactor Dimensionality Reduction Pedigree Disequilibrium Test (MDR-PDT) [16].

Finally, since the main point of methods development is for eventual application, the trioGEDT should be used on a real dataset in human genetics.

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 their faculty mentors on the project, Drs. Alison Motsinger-Reif and David Reif, as well as other participants in the program.

6. REFERENCES

[1] D. Altshuler, *et al.*, "Genetic mapping in human disease," *Science,* vol. 322, pp. 881-8, Nov 7 2008.

[2] J. H. Moore and M. D. Ritchie, "STUDENTJAMA. The challenges of whole-genome approaches to common diseases," *JAMA,* vol. 291, pp. 1642-3, Apr 7 2004.

[3] D. B. Goldstein, "Common genetic variation and human traits," *N Engl J Med,* vol. 360, pp. 1696-8, Apr 23 2009.

[4] A. A. Motsinger, *et al.*, "Novel methods for detecting epistasis in pharmacogenomics studies," *Pharmacogenomics,* vol. 8, pp. 1229-41, Sep 2007.

[5] X. Yao, "Evolutionary artificial neural networks," *Int J Neural Syst,* vol. 4, pp. 203-22, Sep 1993.

[6] A. A. Motsinger-Reif and M. D. Ritchie, "Neural networks for genetic epidemiology: past, present, and future," *BioData Min,* vol. 1, p. 3, 2008.

[7] J. Koza and J. P. Rice, "Genetic generation of both the weights and architecture for a neural network.," *IEEE Transactions,* vol. 2, 1991.

[8] A. A. Motsinger-Reif, *et al.*, "Grammatical evolution decision trees for detecting gene-gene interactions," *BioData Min,* vol. 3, p. 8, 2010.

[9] R. S. Spielman, *et al.*, "Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM)," *Am J Hum Genet,* vol. 52, pp. 506-16, Mar 1993.

[10] R. S. Spielman and W. J. Ewens, "The TDT and other family-based tests for linkage disequilibrium and association," *Am J Hum Genet,* vol. 59, pp. 983-9, Nov 1996.

[11] M. O'Neill and C. Ryan, *Grammatical Evolution*. Boston: Kluwer Academic Publishers, 2001.

[12] E. Alpaydin, *Introduction to Machine Learning*. Cambridge, MA: MIT Press, 2004.

[13] J. H. Moore, "The ubiquitous nature of epistasis in determining susceptibility to common human diseases," *Hum Hered,* vol. 56, pp. 73-82, 2003.

[14] N. Li, *et al.*, "[Identification of gene-gene interactions related to the etiology of complex disease: a multifactor dimensionality reduction-genotype pedigree disequilibrium test]," *Zhonghua Liu Xing Bing Xue Za Zhi,* vol. 28, pp. 1036-40, Oct 2007.

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

[16] E. R. Martin, *et al.*, "A novel method to identify gene-gene effects in nuclear families: the MDR-PDT," *Genet Epidemiol,* vol. 30, pp. 111-23, Feb 2006.