Application of Genetic Algorithms to the Discovery of Complex Models for Simulation Studies in Human Genetics

Jason H. Moore, Lance W. Hahn, Marylyn D. Ritchie, Tricia A. Thornton, Bill C. White

Program in Human Genetics
Department of Molecular Physiology and Biophysics
519 Light Hall
Vanderbilt University
Nashville, TN 37232-0700
{moore, hahn, ritchie, thornton, bwhite}@phg.mc.vanderbilt.edu

Abstract

Simulation studies are useful in various disciplines for a number of reasons including the development and evaluation of new computational and statistical methods. This is particularly true in human genetics and genetic epidemiology where new analytical methods are needed for the detection and characterization of disease susceptibility genes whose effects are complex, nonlinear, and partially or solely dependent on the effects of other genes. Despite this need, the development of complex genetic models that can be used to simulate data is not always intuitive. In fact, only a few such models have been published. In this paper, we present a strategy for identifying complex genetic models for simulation studies that utilizes genetic algorithms. The genetic models used in this study are penetrance functions that define the probability of disease given a specific DNA sequence variation has been inherited. We demonstrate that the genetic algorithm approach routinely identifies interesting and useful penetrance functions in a human-competitve manner.

1 INTRODUCTION

One goal of human genetics is to identify genes that confer an increased risk of disease in certain individuals. The identification of disease susceptibility genes has the potential to improve human health through the development of new prevention, diagnosis, and treatment strategies. Although achieving this goal is an important public health endeavor, it is not easily accomplished for common diseases, such as essential hypertension, due to the complex multifactorial nature of the disease (Kardia, 2000; Moore and Williams, 2002). That is, in such cases, risk of disease is due to a complex interplay between multiple genes and multiple environmental factors. The identification of genes that influence risk of disease only through complex interactions with other genes (i.e. gene-gene interactions) and/or environmental factors (i.e. gene-environment interactions) remains a statistical and computational challenge (Templeton, 2000; Moore and Williams, 2002). The statistical challenge is to consider high-dimensional interactions without loss of degrees of freedom while the computational challenge lies in the size and complexity of the search space. Gene-gene interactions are examples of attribute interactions, a major challenge for data mining (Freitas, 2001).

Several new methods have been developed in an attempt to address the statistical and computational challenges of detecting and characterizing complex disease susceptibility genes. These methods can be classified as either data reduction approaches or pattern recognition approaches. Data reduction methods such as the multifactor dimensionality reduction or MDR approach (Ritchie et al., 2001) seek to reduce the dimensionality of the problem in order to facilitate exploratory data analysis and hypothesis testing. MDR reduces multiple predictor variables to a single variable, thereby reducing the dimensionality of the problem. In contrast, pattern recognition and machine learning strategies such as neural networks (Lucek et al., 1998; Saccone et al., 1999) and cellular automata (Moore and Hahn, 2002) consider the full dimensionality of the data by considering patterns of DNA sequence variations. Although these methods are promising, the power of these approaches for identifying gene-gene and gene-environment interactions has not been fully evaluated. The evaluation of power is best accomplished using simulated data.

The goal of this study was to develop a genetic algorithm (GA) strategy for discovering complex genetic models in the form of penetrance functions that can be used to simulate data for the evaluation of new statistical and computational methods. Penetrance functions define the probability of disease given a particular combination of DNA sequence variations has been inherited. Penetrance functions of interest in this study exhibit gene-gene or attribute interactions in the absence of independent main effects. We begin in Section 2 with an overview of genetic models in terms of penetrance functions. In Section 3, we describe our GA approach to discovering

complex genetic models. A summary and discussion of the results are presented in Sections 4 and 5 respectively. The conclusions are presented in Section 6. The results presented in this paper demonstrate a GA strategy is capable of routinely identifying interesting and useful genetic models in a human-competitve manner.

2 PENETRANCE FUNCTIONS AS GENETIC MODELS

Penetrance functions represent one approach to modeling the relationship between genetic variations (i.e. variation in the DNA sequence of a gene) and risk of disease. Penetrance is simply the probability of disease given a particular combination of genotypes. A single genotype is determined by one allele (i.e. a specific DNA sequence state) inherited from the mother and one allele inherited from the father. For most genetic variations, only two alleles (*A* or *a*) exist in the biological population. Therefore, because the order of the alleles is unimportant, a genotype can have one of three values: *AA*, *Aa* or *aa*. Penetrance functions define the probability of disease for all genotypes for one or more genetic variations. Once the penetrance functions are specified, genetic data can easily be simulated for people with the disease and for people without the disease. For example, the penetrance function for an autosomal recessive disease (i.e. a disease that requires two copies of the same allele) such as cystic fibrosis in which only one of the three genotypes leads to disease might look like Table 1. Here, individuals who inherit the *AA* or *Aa* genotypes have zero probability of disease while individuals who inherit the *aa* genotype are certain to have the disease. From this simple recessive Mendelian model, data can simply be simulated by giving affected individuals *aa* genotypes and unaffected individuals *AA* or *Aa* genotypes, in proportion to their defined population frequencies.

Table 1. Penetrance values for three genotypes from a gene acting under an autosomal recessive disease model.

AA	*Aa*	*aa*
0	0	1

More complex genetic models can be developed by assigning disease risk to more than one genotype from one or more genetic variations. Table 2 illustrates a penetrance function that relates two genetic variations, each with two alleles and three genotypes, to risk of disease. In this example, the alleles each have a biological population frequency of $p = q = 0.5$ 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 (Hartl and Clark 1997). Thus, assuming the frequency of the *AA* genotype is 0.25, the frequency of *Aa* is 0.5, and the frequency of *aa* is 0.25, then the marginal penetrance of *BB* (i.e. the effect of just the *BB* genotype on disease risk) can be calculated as (0.25 * 0) + (0.5 * 0) + (0.25 * 1) = 0.25. This means that the probability of disease given the *BB* genotype is 0.25, regardless of the genotype at the other genetic variation. Similarly, the marginal penetrance of *Bb* can be calculated as (0.25 * 0) + (0.5 * 0.5) + (0.25 * 0) = 0.25. 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 exactly two high-risk alleles (e.g. *a* and *b* are defined as high risk). Thus, *aa*/*BB*, *Aa*/*Bb*, and *AA*/*bb* are the high-risk genotype combinations. This model was first described by Frankel and Schork (1996). What makes this model complex is the absence of a main effect for either of the genetic variations. Thus, each genetic variation only has an effect on disease risk in the context of the other genetic variation. Such gene-gene interactions are believed to play an important role in determining an individual's risk for developing common diseases (Moore and Williams, 2002; Templeton, 2000).

Table 2. Penetrance values for combinations of genotypes from two genes exhibiting interactions but not main effects.

	Table penetrance values			Margin penetrance values
	AA (.25)	*Aa* (.50)	*aa* (.25)	
BB (.25)	0	0	1	.25
Bb (.50)	0	.50	0	.25
bb (.25)	1	0	0	.25
Margin penetrance values	.25	.25	.25	

Table 3. Penetrance values for combinations of genotypes from two genes exhibiting interactions but not main effects.

	Table penetrance values			Margin penetrance values
	AA (.25)	*Aa* (.50)	*aa* (.25)	
BB (.25)	0	1	0	.50
Bb (.50)	1	0	1	.50
bb (.25)	0	1	0	.50
Margin penetrance values	.50	.50	.50	

The gene-gene interaction model described in Table 2 was developed by trial and error. That is, a human derived this model by substituting various allele frequencies and

penetrance functions until a model was found that had attribute or gene-gene interaction effects without independent main effects. This is one of only a few complex genetic models that have been described in the literature. The scarcity of complex genetic models in the literature is primarily due to the extraordinary combinatorial complexity of the problem, as has been discussed by Culverhouse et al. (2002). Effectively, there are an infinite number of possible penetrance functions that could be developed for just two genetic variations. Only some of these models would exhibit a complex relationship with disease risk. The size of the search space precludes the human-based trial and error approach as well as exhaustive computational searches without specific restrictions and assumptions about the allele frequency and penetrance function values. For example, Li and Reich (2000) enumerated every possible penetrance function using probability values restricted to zero and one. This yielded a manageable 2^9 total models. Only one of these models exhibits interaction effects in the absence of main effects (see Table 3). Culverhouse et al. (2002) have also enumerated a restricted set of models. The goal of the present study was to develop a machine intelligence approach to discovering complex genetic models in the form of penetrance functions. The next section describes the GA approach we used.

3 THE GENETIC ALGORITHM

3.1 OVERVIEW OF GENETIC ALGORITHMS

Genetic algorithms have been shown to be a very effective strategy for implementing beam searches of rugged fitness landscapes (Goldberg, 1989). Briefly, this is accomplished by generating a random population of models or solutions, evaluating their ability to solve a particular problem, selecting the best models or solutions, and generating variability in these models by exchanging model components among different models. The process of selecting models and introducing variability is iterated until an optimal model is identified or some termination criteria are satisfied. This general procedure was inspired by the problem solving abilities of evolution by natural selection in biological populations. Using similar language, GAs operate using populations of chromosomes (models) that undergo selection according to fitness, reproduction, recombination, and mutation.

3.2 DESCRIPTION OF OUR GENETIC ALGORITHM

3.2.1 SOLUTION REPRESENTATION

A solution or model consists of a set of nine penetrance values or probabilities on the interval from zero to one in increments of 0.001. Thus, the entire search space consisted of 10^{27} possible models. Each penetrance value represents the probability of disease given a particular combination of two genotypes. Each of the nine real-valued probabilities was encoded as 32 bits for a total GA chromosome length of 288 bits.

3.2.2 FITNESS FUNCTION

Fitness was determined by maximizing the variance of the table penetrance values (V_t) and minimizing the variance of the marginal penetrance values (V_m). Maximizing V_t ensures that we identify interesting patterns of genotypes while minimizing V_m ensures the size of the main effect of each genotype is small. We stopped the GA when a model satisfied both $V_t \geq 0.1$ and $V_m \leq 0.0001$. These values were selected to ensure interaction effects without main effects of each genetic variation.

3.2.3 GA PARAMETERS

Table 4 summarizes the GA parameters used in this study. We ran the GA a total of 100,000 times with each run consisting of a maximum of 10,000 generations.

Table 4. GA parameters.

Objective	Discover complex models
Fitness function	$V_t - V_m$
Number of runs	100,000
Stopping criteria	$V_t \geq 0.1$ and $V_m \leq 0.0001$
Population size	200
Generations	10,000
Selection	Stochastic uniform sampling
Crossover	Uniform, by variable
Crossover probability	0.60
Mutation	Gaussian
Mutation probability	0.01

3.2.4 SOFTWARE AND HARDWARE

Our GA implementation used GAlib, a C++ class library for UNIX, Windows and Mac operating systems (http://lancet.mit.edu/ga/). Coarse-grained parallelism, utilizing 10 processors to perform 10 sets of 10,000 runs, for a total of 100,000 runs, used the MPICH parallel programming library on a 110-node Beowulf-style parallel computing cluster running Linux.

4 RESULTS

The GA was run for a total of 100,000 times, and the best model was saved from each. Of the 100,000 best models discovered by the GA, there were no duplicates. Thus, each model was unique. We first wanted to know whether penetrance function models that have been previously described in the literature were discovered by

the GA. The GA-generated model illustrated in Table 5 (V_t = .154764 and V_m = .000044) is very similar to the model shown in Table 2 while the model illustrated in Table 6 (V_t = .21157 and V_m = .000082) is very similar to the model shown in Table 3. Subtle variations of the models shown in Tables 2 and 5 were discovered in 13 out of the 100,000 GA runs. Similarly, subtle variations of the models shown in Tables 3 and 6 were discovered in three out of the 100,000 GA runs. Thus, the GA routinely discovered models that have been described previously.

Table 5. GA-generated model similar to the previously described model in Table 2.

	Table penetrance values			Margin penetrance values
	AA (.25)	*Aa* (.50)	*aa* (.25)	
BB (.25)	.083	.076	.964	.29
Bb (.50)	.056	.508	.085	.30
bb (.25)	.977	.098	.062	.30
Margin penetrance values	.30	.29	.31	

Table 6. GA-generated model similar to the previously described model in Table 3.

	Table penetrance values			Margin penetrance values
	AA (.25)	*Aa* (.50)	*aa* (.25)	
BB (.25)	.094	.905	.097	.51
Bb (.50)	.967	.097	.937	.52
bb (.25)	.027	.990	.080	.51
Margin penetrance values	.50	.52	.52	

Table 7. A GA-generated model.

	Table penetrance values			Margin penetrance values
	AA (.25)	*Aa* (.50)	*aa* (.25)	
BB (.25)	.967	.314	.137	.43
Bb (.50)	.313	.312	.742	.43
bb (.25)	.129	.779	.075	.42
Margin penetrance values	.43	.42	.44	

Our second question was whether the GA routinely generated new and interesting models. All of the models identified by the GA exhibited gene-gene interactions with minimal or no main effects. In fact, other than the class of models illustrated in Tables 5 and 6, none have been described previously in the literature. Thus, approximately 99,987 models are unique. Tables 7-10 illustrate four of the new models discovered by the GA.

Table 8. A GA-generated model.

	Table penetrance values			Margin penetrance values
	AA (.25)	*Aa* (.50)	*aa* (.25)	
BB (.25)	.967	.139	.799	.51
Bb (.50)	.057	.655	.627	.50
bb (.25)	.974	.544	.019	.52
Margin penetrance values	.51	.50	.52	

Table 9. A GA-generated model.

	Table penetrance values			Margin penetrance values
	AA (.25)	*Aa* (.50)	*aa* (.25)	
BB (.25)	.017	.451	.711	.42
Bb (.50)	.520	.571	.039	.41
bb (.25)	.640	.053	.949	.43
Margin penetrance values	.41	.43	.42	

Table 10. A GA-generated model.

	Table penetrance values			Margin penetrance values
	AA (.25)	*Aa* (.50)	*aa* (.25)	
BB (.25)	.954	.256	.360	.44
Bb (.50)	.010	.731	.300	.45
bb (.25)	.801	.093	.808	.44
Margin penetrance values	.46	.44	.45	

The model summarized in Table 7 (V_t = .106238 and V_m = .000052) indicates that individuals with genotype

combinations of *AA/BB*, *Aa/bb*, and *aa/Bb* are at highest risk of disease while those with *AA/Bb*, *Aa/BB*, and *Aa/Bb* are at intermediate risk. The remaining individuals are at relatively low risk. This nonlinear pattern of high-risk and low-risk genotype combinations is indicative of gene-gene interactions. Risk of disease is not significantly different between single genotypes (represented by margin penetrance values), confirming an absence of main effects.

The models summarized in Table 8 (V_t = .140427 and V_m = .000091), Table 9 (V_t = .110712 and V_m = .000098), and Table 10 (V_t = .120743 and V_m = .000035) have different nonlinear combinations of genotypes associated with varying risk of disease. Again, none of the genotypes in these models is associated with disease risk independent of the other genotypes. This indicates gene-gene or attribute interaction in the absence of main effects.

5 DISCUSSION

In the present work, we focused on genetic models with just two genetic variations. However, we anticipate that genetic models incorporating more than just two genetic variations will be useful in simulation studies since most common diseases are likely to be influenced by many genes. This is evident in the study by Ritchie et al. (2001) that identified a combination of four genetic variations that is associated with risk of sporadic breast cancer in a complex nonlinear manner. Our future studies will focus on expanding the GA to search for combinations of three or more genetic variations that exhibit attribute interaction in the absence of main effects. Further, it will be important to explore a range of allele frequencies as well as methods for categorizing models into similar classes.

Human genetics is undergoing an information explosion and a comprehension implosion. In fact, our ability to measure genetic information, and biological information in general, is far outpacing our ability to interpret it. As demonstrated in this study, machine intelligence strategies such as GAs hold promise for dealing with genetic data that is high-dimensional and complex. However, the present study is not the first to apply evolutionary algorithms to a genetics problem. In fact, evolutonary algorithms have been used to optimize data analysis approaches in genetic epidemiology studies (Congdon et al., 1993; Carlborg et al., 2000; Tapadar et al., 2000; Moore and Hahn, 2002), gene expression studies (Moore and Parker, 2001; Moore et al., 2001; Parker and Moore, 2001), and studies of gene networks (Koza et al., 2001). We anticipate an increase in applications of GAs in the field of human genetics as more investigations begin to focus on the challenge of simulating and analyzing complex, high-dimensional genetic data.

6 CONCLUSIONS

The results of this study document the utility of GAs for the discovery of complex genetic models that can be used for simulation studies in human genetics. In fact, our GA discovered approximately 99,987 models that have not been previously described in the literature. Thus, the results are human-competitive and routine. To our knowledge, this is the first application of a machine intelligence approach to the discovery of complex genetic models such as penetrance functions.

Acknowledgments

This work was funded by National Institutes of Health grants HL65234, HL65962, GM31304, AG19085, and AG20135.

References

Carlborg O, Andersson L, Kinghorn B (2000): The use of a genetic algorithm for simultaneous mapping of multiple interacting quantitative trait loci. Genetics 155:2003-10.

Congdon CB, Sing CF, Reilly SL (1993): Genetic algorithms for identifying combinations of genes and other risk factors associated with coronary artery disease. In: Proceedings of the Workshop on Artificial Intelligence and the Genome. Chambery.

Culverhouse R, Suarez BK, Lin J, Reich T (2002): A perspective on epistasis: limits of models displaying no main effect. American Journal of Human Genetics 70:461-71.

Frankel WN, Schork NJ (1996): Who's afraid of epistasis? Nature Genetics 14:371-373.

Freitas AA (2001): Understanding the crucial role of attribute interaction in data mining. Artificial Intelligence Reviews 16:177-199.

Goldberg DE (1989): "Genetic Algorithms in Search, Optimization, and Machine Learning." Reading: Addison-Wesley.

Hartl DL, Clark AG (1997): "Principles of Population Genetics" Mass: Sinauer Assoc.

Kardia SLR (2000): Context-dependent genetic effects in hypertension. Current Hypertension Reports 2:32-38.

Koza JR, Mydlowec W, Lanza G, Yu J, Keane MA (2001): Reverse engineering of metabolic pathways from observed data using genetic programming. Pacific Symposium on Biocomputing 2001:434-445.

Li W, Reich J (2000): A complete enumeration and classification of two-locus disease models. Human Heredity 50:334-349.

Lucek P, Hanke J, Reich J, Solla SA, Ott J (1998): Multi-locus nonparametric linkage analysis of complex trait loci with neural networks. Human Heredity 48:275-284.

Moore JH, Hahn LW (2002): A cellular automata approach to detecting interactions among single-nucleotide polymorphisms in complex multifactorial diseases. Pacific Symposium on Biocomputing 2002:53-64.

Moore JH, Parker JS, Hahn LW (2001): Symbolic discriminant analysis for mining gene expression patterns. In De Raedt, L. and Flach, P. (eds), "Machine Learning: ECML 2001" Lecture Notes in Artificial Intelligence 2167:372-381, Berlin: Springer-Verlag.

Moore JH, Parker JS (2001): Evolutionary computation in microarray data analysis. In Lin, S. and Johnson, K. (eds), "Methods of Microarray Data Analysis" Boston: Kluwer Academic Publishers.

Moore JH, Williams SW (2002): New strategies for identifying gene-gene interactions in hypertension. Annals of Medicine 34:1-8.

Parker JS, Moore JH (2001): Dynamics based pattern recognition and parallel genetic algorithms for the analysis of multivariate gene expression data. Proceedings of the 2001 Genetic and Evolutionary Computation Conference Workshop Program, San Francisco, pp 433-436.

Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH (2001): Multifactor dimensionality reduction reveals high-order interactions among estrogen metabolism genes in sporadic breast cancer. American Journal of Human Genetics 69:138-147.

Saccone NL, Downey Jr. TJ, Meyer DJ, Neuman RJ, Rice JP (1999): Mapping genotype to phenotype for linkage analysis. Genetic Epidemiology. S17: S703-8.

Tapadar P, Ghosh S, Majumder PP (2000): Haplotyping in pedigrees via a genetic algorithm. Human Heredity 50:43-56

Templeton AR (2000): Epistasis and complex traits. In: Wade M, Brodie III B, Wolf J (eds) "Epistasis and Evolutionary Process" Oxford: Oxford University Press.