GENETIC PROGRAMMING
THEORY AND PRACTICE VIII

Edited by
RICK RIOLO
Center for the Study of Complex Systems
University of Michigan

TRENT MCCONGAHY
Solido Design Automation

EKATERINA VLADISLAVLEVA
Department of Mathematics and Computer Science, University of Antwerp

Kluwer Academic Publishers
Boston/Dordrecht/London

Chapter 1

EXPLOITING EXPERT KNOWLEDGE OF PROTEIN-PROTEIN INTERACTIONS IN A COMPUTATIONAL EVOLUTION SYSTEM FOR DETECTING EPISTASIS

Kristine A. Pattin[†], Joshua L. Payne[†], Douglas P. Hill, Thomas Caldwell, Jonathan M. Fisher, and Jason H. Moore

Dartmouth Medical Scool, One Medical Center Drive, HB7937, Lebanon, NH 03756 USA
[†]*These authors contributed equally to this work.*

Abstract The etiology of common human disease often involves a complex genetic architecture, where numerous points of genetic variation interact to influence disease susceptibility. Automating the detection of such epistatic genetic risk factors poses a major computational challenge, as the number of possible gene-gene interactions increases combinatorially with the number of sequence variations. Previously, we addressed this challenge with the development of a computational evolution system (CES) that incorporates greater biological realism than traditional artificial evolution methods. Our results demonstrated that CES is capable of efficiently navigating these large and rugged epistatic landscapes toward the discovery of biologically meaningful genetic models of disease predisposition. Further, we have shown that the efficacy of CES is improved dramatically when the system is provided with statistical expert knowledge. We anticipate that biological expert knowledge, such as genetic regulatory or protein-protein interaction maps, will provide complementary information, and further improve the ability of CES to model the genetic architectures of common human disease. The goal of this study is to test this hypothesis, utilizing publicly available protein-protein interaction information. We show that by incorporating this source of expert knowledge, the system is able to identify functional interactions that represent more concise models of disease susceptibility with improved accuracy. Our ability to incorporate biological knowledge into learning algorithms is an essential step toward the routine use of methods such as CES for identifying genetic risk factors for common human diseases.

Keywords: Computational Evolution, Genetic Epidemiology, Epistasis, Protein-Protein Interactions

1. Introduction

Recent developments in high-throughput genotyping technologies have allowed for inexpensive and dense mappings of the human genome. These mappings often comprise single nucleotide polymorphisms (SNPs), which are single nucleotide pairs that vary among people. SNPs are of interest because they constitute the most abundant form of genetic variation in human populations, and thus offer the potential to act as reliable markers of disease-causing genetic variants. Genome-wide association studies (GWAS), measuring 10^6 or more SNPs per individual, are becoming a standard methodology for the detection of genetic risk factors of human disease. Though these studies have generated a wealth of data, few have successfully identified single sequence variants that are highly predictive of clinical endpoints. Moreover, recent analyses of the robustness and evolvability of regulatory and proteomic interaction networks (Albert et al., 2000; Aldana et al., 2007; Jeong et al., 2001) suggest that phenotypic aberrations, such as disease state, rarely result from single points of failure, but more often from the confluence of a number of interacting components. Taken together, these observations suggest that common human diseases possess complex genotype-phenotype maps, with multiple interacting genetic factors influencing disease susceptibility (Moore, 2003; Moore and Williams, 2009).

The development of computational methods that aid in the discovery and characterization of epistatic interactions in GWAS datasets is therefore of the utmost importance (Cordell, 2009; Moore and Williams, 2009; Moore et al., 2010). Specifically, there are two important computational challenges that need to be addressed (Moore et al., 2010). First, we need data mining, machine learning and computational intelligence algorithms that are capable of modeling nonlinear relationships between multiple SNPs and clinical endpoints such as disease susceptibility. Second, we need powerful search algorithms that are able to identify optimal nonlinear models in large, rugged fitness landscapes. Unfortunately, exhaustive methods that enumerate all possible SNP combinations are infeasible for this problem, because the number of possible combinations grows exponentially with the number of SNPs. For example, in an analysis of one million candidate SNPs, there are 5×10^{11} pair-wise combinations and 1.7×10^{17} three-way combinations to consider.

For the purpose of this study, we focus on machine learning and computational approaches to this problem. Notable examples of prior work include multifactor dimensionality reduction (Ritchie et al., 2001) and random chemistry (Eppstein et al., 2007). In addition, artificial evolution approaches, such as genetic programming (Moore and White, 2007), have been investigated. In their canonical form, these methods have demonstrated limited success, due to their reliance on the presence of building blocks. As the primary characteristic of an epistatic genetic architecture is the absence of individual genetic effects, these artificial evolution methods lack the critical components needed to piece together an effective solution. In an effort to provide these critical components, several studies have investigated the effect of including statistical expert knowledge (Greene et al., 2009c; Greene et al., 2009d; Moore and White, 2006; Moore and White, 2007), as derived from a family of machine learning techniques known as Relief (Kononenko, 1994). This expert knowledge consists of a population-level estimate of the probability with which a SNP is associated with disease status, via individual or interaction effects. These probabilities are then be used to bias variation operators (Greene et al., 2009c) or population initialization (Greene et al., 2009d) toward SNPs that are thought to be associated with disease, effectively seeding the population with the necessary building blocks.

While such statistical expert knowledge has improved the applicability of artificial evolution methods for epistasis analysis, it has been suggested (Banzhaf et al., 2006) that the incorporation of greater biological realism may offer further performance improvements. Specifically, Banzhaf et al. (2006) have called for the development of open-ended computational evolution systems (CES) that attempt to emulate, rather than ignore, the complexities of biological systems. Paying heed to this call, we have recently developed a hierarchical, spatially-explicit CES that allows for the evolution of arbitrarily complex solutions and solution operators, and includes population memory via archives, feedback loops between archives and solutions, and environmental sensing (Moore et al., 2008; Moore et al., 2009; Greene et al., 2009b; Payne et al., 2010). Analyses of this system have demonstrated its ability to identify complex disease-causing genetic architectures in simulated data, and to recognize and exploit useful sources of expert knowledge. Specifically, we have shown that statistical expert knowledge, in the form of ReliefF scores, can be incorporated via environmental sensing (Greene et al., 2009b) and population initialization (Payne et al., 2010) to improve system performance. In addition to the statistical expert knowledge provided by machine learning techniques, there are numerous sources of biological expert knowledge that could be used to improve the performance of CES. For example, information regarding biochemical pathways, regulatory networks, and protein-protein interactions (PPIs) could be used to bias CES toward pathways or gene combinations that are known to interact experimentally. Indeed, such an approach has already proven successful in other GWAS (Askland et al., 2009; Emily et al., 2009).

Here, we investigate the use of biological expert knowledge of PPIs, extracted from the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database. To accommodate these data in CES, we extend two previously developed variation operators, so that solution construction is biased toward the inclusion of experimentally verified molecular interactions. We compare the performance of CES with and without this form of biological expert knowledge in the task of identifying an artificial two-locus epistatic genetic relationship in the background of real human genetic data.

2. Conceptual Overview of the Problem

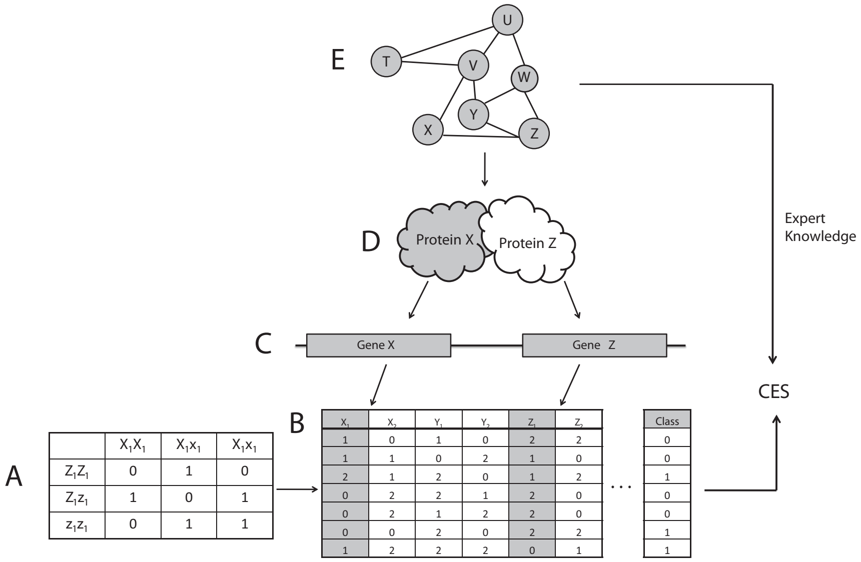
Since the structure of a protein interaction network is directly influenced by genetic variation, protein-protein interaction information may prove a valuable source of expert knowledge for CES. In Figure 1-1, we provide an overview of how we test this hypothesis in the current study. We evaluate the ability of CES to detect an artificial two-locus epistatic signal, generated from a genetic penetrance table (Fig 1-1A), that we placed in a real genetic bladder cancer background (Fig 1-1B). We selected two SNPs located in two separate genes (Fig 1-1C) to represent the epistatic signal such that this gene pair exhibited a validated protein-protein interaction (Fig 1-1D) in the STRING database (Fig 1-1E). Since we wish to understand the sensitivity of CES to PPI information, we generated data for interaction scenarios of varying strength, as described below.

3. Methods

In this section, we first present our computational evolution system, highlighting the algorithmic adjustments made to accommodate protein-protein interaction information. We then discuss the database from which this biological expert knowledge was drawn, and the genetic data used for performance testing. Lastly, we present our experimental design.

Computational Evolution System

In Figure 1-2, we provide a graphical overview of CES, which is both hierarchically organized and spatially explicit. The bottom level of the hierarchy consists of a lattice of solutions (Fig. 1-2d), which compete with one another within spatially-localized, overlapping neighborhoods. The second layer of the hierarchy contains a lattice of arbitrarily complex solution operators (Fig. 1-2c), which operate on the solutions in the lower layer. The third layer of the hierarchy contains a lattice of mutation operators (Fig. 1-2b), which modify the solution operators in the second layer, and the highest layer of the hierarchy governs the rate at which the mutation operators are modified (Fig. 1-2a). CES includes environmental noise (Fig. 1-2h), which perturbs the attribute values of the solutions with probability p_{noise}, as they are read from the input data. Intermediate values of p_{noise} allow for the escape of local optima, and improve classification power (Greene et al., 2009a). In this study, $p_{noise} = 0.07$. CES also possesses an attribute archive (Fig. 1-2g), which stores the frequencies with which attributes are used. The solution operators can then exploit these



	X_1X_1	X_1x_1	X_1x_1
Z_1Z_1	0	1	0
Z_1z_1	1	0	1
z_1z_1	0	1	1

X_1	X_2	Y_1	Y_2	Z_1	Z_2
1	0	1	0	2	2
1	1	0	2	1	0
2	1	2	0	1	2
0	2	2	1	2	0
0	2	1	2	2	0
0	0	2	0	2	2
1	2	2	2	0	1

Class
0
0
1
0
0
1
1

Figure 1-1. Overview of data generation and the integration of biological expert knowledge in CES. To assess the performance of CES on human data when provided with expert knowledge from protein-protein interactions, we merged an artificial two-locus epistatic signal with a real genetic dataset. This artificial signal was generated from a penetrance table (A) and represented by two SNPs selected from the bladder cancer dataset (B). Here, these SNPs are denoted X_1 and Z_1, and are located on different genes (C) whose protein products are a validated protein-protein interaction (D) in the STRING database (E).

data to bias the construction of solutions toward frequently utilized attributes. To conduct a more transparent analysis of the influence of biological expert knowledge on system performance, we prohibit the use of the archive in this study.

Solution Representation, Fitness Evaluation, and Selection

Each solution represents a classifier, which takes a set of SNPs as input and produces an output that can be used to assign diseased or healthy status. These solutions are represented as stacks, where each element in the stack consists of a function and two operands (Fig. 1-2). The function set contains $+, -, *, \%, <, \leq, >, \geq, ==, \neq$, where $\%$ denotes protected modulus (i.e., $x \ \% \ 0 = x$ (Langdon, 1998)). Operands are either SNPs, constants, or the output of another element in the stack (Fig. 1-2).

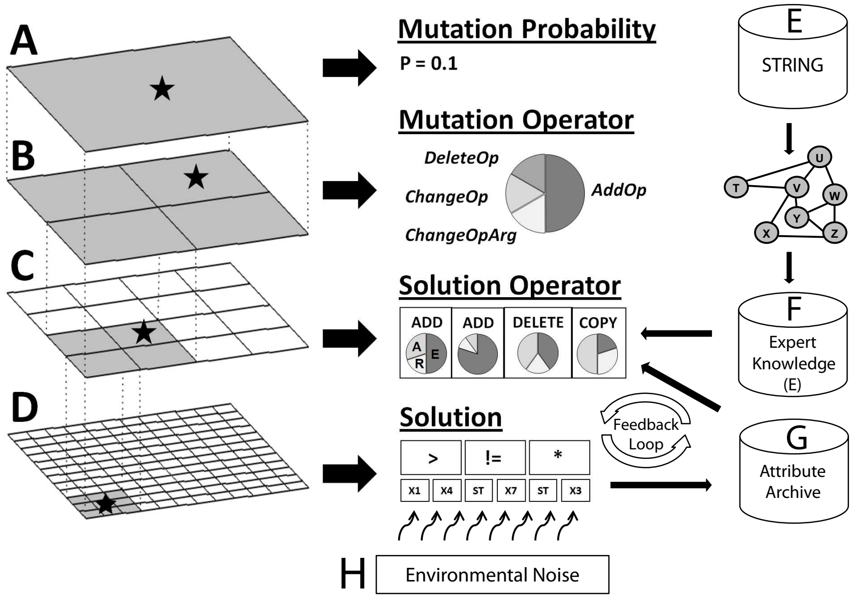


Figure 1-2. Visual overview of our computational evolution system for discovering symbolic discriminant functions that differentiate disease subjects from healthy subjects using information about single nucleotide polymorphisms (SNPs). The hierarchical structure is shown on the left while some specific examples at each level are shown in the middle. At the lowest level (D) is a grid of solutions. Each solution consists of a list of functions and their arguments (e.g. X1 is an attribute) that are evaluated using a stack (denoted by ST in the solution). The next level up (C) is a grid of solution operators that each consist of some combination of the ADD, ALTER, COPY, DELETE, and REPLACE functions and their respective set of probabilities that define whether expert knowledge (F) based on protein-protein interaction information from the STRING database (E) is used instead of a random generator (denoted by R in the probability pie). The attribute archive (G) is derived from the frequency with which each attribute occurs among solutions in the population. For this study, use of the attribute archive was prohibited. Finally, environmental noise (H) perturbs the data to prevent over fitting. The top two levels of the hierarchy (A and B) exist to generate variability in the operators that modify the solutions. A 12 × 12 grid is shown here as an example.

Each solution produces a discrete output S_i when applied to an individual i. Symbolic discriminant analysis (Moore et al., 2002) is then used to map this output to a classification rule, as follows. The solution is independently applied to the set of diseased and healthy individuals to obtain two separate distributions of outputs, $S^{diseased}$ and $S^{healthy}$, respectively. A classification threshold S_0 is then calculated as the arithmetic mean of these two distributions. The corresponding solution classifies individual i as diseased if $S_i > S_0$ and healthy otherwise. Solution accuracy is assessed through a comparison of predicted and actual clinical endpoints. Specifically, the number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) are used to calculate accuracy as

$$A = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right). \tag{1.1}$$

Solution fitness is then given as a function of accuracy and solution length L

$$f = A + \frac{\alpha}{L}, \tag{1.2}$$

where α is a tunable parameter used to encourage parsimony (for all experiments reported here, $\alpha = 0.001$).

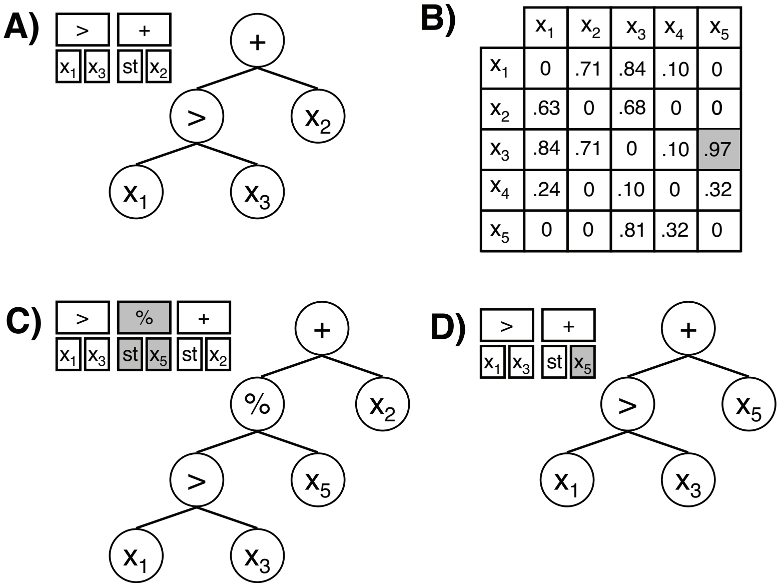
The population is organized on a two-dimensional lattice with periodic boundary conditions. Each solution occupies a single lattice site, and competes with the solutions occupying the eight spatially adjacent sites. Selection is both synchronous and elitist, such that the solution of highest fitness within a given neighborhood is always selected to repopulate the focal site of that neighborhood. Reproduction is either sexual or asexual, as dictated by the evolvable solution operators that reside in the next layer of the hierarchy.

The population is initialized by randomly generating solutions with one, three, and seven elements, in equal proportions. Functions are selected at random with uniform probability from the function set and SNP attributes are selected using an enumerative scheme. Specifically, SNPs are drawn with uniform probability and without replacement until all SNPs are represented. The attribute pool is then regenerated as a new random permutation of the SNPs, and the process is repeated, until the attribute requirements of all solutions are satisfied.

Solution Operators

CES allows for the evolution of arbitrarily complex variation operators. This is achieved by initializing the solution operator lattice (Fig. 1-2c) with a set of five basic building blocks (COPY, REPLACE, DELETE, ADD, ALTER), which can be recombined in any way to form new operators. The COPY operator inserts a random element of the focal solution stack into the stack of a randomly chosen neighboring solution. The REPLACE operator extracts a sequence of random length from a neighboring solution stack and overwrites a randomly chosen sequence of the focal solution stack with that information, and the DELETE operator removes an element from the focal solution stack.

In this study, the ADD and ALTER operators were modified to incorporate the biological expert knowledge of PPIs (Fig. 1-3). Specifically, the ADD operator (Fig. 1-3c) places a randomly chosen function and its arguments into



	x_1	x_2	x_3	x_4	x_5
x_1	0	.71	.84	.10	0
x_2	.63	0	.68	0	0
x_3	.84	.71	0	.10	.97
x_4	.24	0	.10	0	.32
x_5	0	0	.81	.32	0

Figure 1-3. (A) In CES, solutions are represented using stacks, which naturally translate into the parse tree representations commonly used in genetic programming (Poli et al., 2008). To provide the variation operators with the biological expert knowledge of protein-protein interactions, we explicitly maintain an interaction matrix (B), where each element corresponds to the confidence score of that interaction. These scores are used to bias the selection of attributes in the solution operators, as illustrated for the (C) ADD and (D) ALTER operators. (C) The ADD operator chooses a random location in the stack and inserts a new element (shown in gray), which consists of a randomly generated function and two attributes. SNP attributes are selected by finding the nearest existing upstream SNP in the stack (x_3) and then choosing an interacting SNP from the PPI data (B) with probability proportional to the confidence score. In this case, x_5 is selected. (D) The ALTER operator chooses a random element in the stack and mutates its function, or one of its attributes. If a SNP is selected for mutation, then a new SNP is chosen in the same fashion as the ADD operator.

an arbitrary location of the focal solution stack (Fig. 1-3a). The first argument is selected with uniform probability from three choices: stack output, constant, or attribute. If the argument is chosen to be an attribute, then the nearest upstream SNP in the list representation of the stack is identified. The PPI data for that SNP is then queried, and one of its interacting SNPs is selected (Fig. 1-3b). If the second argument is also chosen to be an attribute, then the PPI data corresponding to the first attribute is queried, and another interacting SNP is selected.

The ALTER operator (Fig. 1-3d) randomly chooses an element from the focal solution stack (Fig. 1-3a) and mutates either its function, or one of its input arguments. If the function is chosen for mutation, it is replaced by a randomly chosen function. If an input argument is selected for mutation, it can be replaced by the stack output, a constant, or an attribute. If the argument is chosen to be an attribute, then the corresponding SNP is selected in the same fashion as in the ADD operator.

In both the ADD and ALTER operators, SNPs were selected from the PPI data with probability proportional to their confidence scores. We used an exponential scaling function to increase the probability with which SNP pairings of high confidence were selected. Let $\vec{I}$ denote the vector of k SNPs with which SNP i interacts, and $\vec{S}$ denote the corresponding vector of confidence scores. The probability with which an interacting SNP $\vec{I_j}$ is selected is given by

$$p_{j|i} = \phi^z \tag{1.3}$$

where

$$z = \frac{(k-1)(\max(\vec{S}) - \vec{S_j})}{(\max(\vec{S}) - \min(\vec{S}))}. \tag{1.4}$$

This mapping of raw confidence scores to selection probabilities has two main features: (i) selection probability decrease exponentially with decreasing confidence, and (ii) all probabilities fall in the range $(0, 1)$. The severity of the exponential decrease is controlled by ϕ. For all experiments considered herein, $\phi = 0.95$.

Solution operators possess evolvable probability vectors that determine the frequency with which functions and attributes are selected at random, from expert knowledge sources, or archives (Fig. 1-2). To highlight the influence of biological expert knowledge on system performance, functions were always chosen at random, and attributes were always selected based on the biological expert knowledge of PPIs, except for the control experiments, where attributes were selected at random.

The solution operators reside on a periodic, toroidal lattice of coarser granularity than the solution lattice (Fig. 1-2). Each site is occupied by a single solution operator, which is assigned to operate on 3x3 sub-grid of solutions. These operators compete with one another in a manner similar to the competition among solutions, and their selection probability is determined by the fitness changes they evoke in the solutions they control.

Mutation Operators

The third level of the hierarchy (Fig. 1-2b) contains the mutation operators, which are used to modify the solution operators. These reside on a toroidal lattice of even coarser granularity, and are assigned to modify a subset of the solution operators below. The mutation operators are represented as three-element vectors, where each element corresponds to the probability with which a specific mutation operator is used. These three mutation operators work as follows. The first (DeleteOp) deletes an element of a solution operator; the second (AddOp) adds an element to a solution operator, and the third (ChangeOp) mutates an existing element in a solution operator. The probabilities with which these mutation operators are used undergo mutation at a rate specified in the highest level of the hierarchy (Figure 1-2a).

STRING

There are a number of publicly available protein-protein interaction databases. For this study, we used the Search Tool for the Retrieval of Interacting Genes / Proteins (STRING) (Jenson et al., 2009), which incorporates PPI information from a number of widely used interaction databases, and currently contains over 2.5 million proteins from 630 different organisms. STRING is freely available, and provides a transparent application programming interface. Queries to the database require the specification of a protein identifier and return a list of all interaction partners, along with a numeric confidence score for each interaction. These confidence scores range from 0 to 1 and are based on a variety of factors, including experimental data and co-occurrence relationships found using text-mining applications. For a detailed description of the scoring methods, see (von Mering et al., 2005).

These confidence scores were maintained within CES as an interaction matrix (Fig. 1-3b). In the dataset considered herein, these scores were always symmetric, though this need not be the case in general. These scores were scaled and normalized by row to produce an attribute selection probability. The average number of interaction partners per SNP in our dataset (i.e., the average number of non-zero entries per row in the interaction matrix) was approximately 208, with a maximum of 811 and a minimum of 2. The average confidence score was 0.44, with a maximum of 0.99 and a minimum of 0.15.

Data Generation

In order to conduct controlled experiments that (i) highlight the sensitivity of our method to the strength of a PPI score, and (ii) allow for an assessment of system performance on human data, we merged an artificial two-locus epistatic signal with a real genetic dataset. These genetic data were originally collected in an effort to ascertain the genetic risk factors of bladder cancer (Andrew et al., 2008), and consist of 1,423 SNPs found in 394 genes across 491 cases and 791 controls. All of the 394 genes in this dataset were represented by multiple SNPs. However, we made the conservative assumption that the disease status associated with the two interacting genes was only associated with a single SNP on each gene (Fig. 1-1). In real data, it is probable that many, if not all, SNPs on these genes would be associated with disease status. This artificial, two-locus epistatic signal was generated from a penetrance table as described in (Culverhouse et al., 2002), with a minor allele frequency of 0.2 and a heritability of 0.4. The SNPs selected to exhibit this signal represented PPI scores of varying strength. Specifically, we considered four separate cases, with confidence scores ranging from 0.916 to 0.998. While these confidence scores are very close in absolute value, the latter score corresponds to an attribute selection probability that is approximately ten times that of the former. This results from the previously described exponential scaling function used for attribute selection.

Experimental Design

We investigated the effect of including biological expert knowledge of PPIs in CES through a direct comparison with two controls. In the first control, CES was not provided with expert knowledge of any form. In the second control, the two epistatic SNPs associated with disease status were chosen such that their corresponding gene products exhibited no functional interaction, i.e., the PPI associated with the two interacting SNPs did not exist in the STRING database.

For each interaction scenario, we generated 100 independent datasets, where the embedded epistatic signal was generated anew from the same penetrance function. The functionally interacting SNPs were therefore the same within each group of 100 datasets, but the individuals possessing the disease-state allele combination differed between these 100 datasets. This allowed for an assessment of the extent to which CES could identify the disease-causing genetic variants when associated with a variety of human genetic backgrounds. For each of the 100 datasets, we performed 100 independent replications. At the end of each of the 100 replications, we calculated the frequency with which all pairs of SNPs co-occurred in the best solution found by CES. We considered the system successful if the most common pairing was the embedded epistatic interaction.

4. Results

In all experiments, the results of the two controls were statistically indistinguishable. We therefore only report the data corresponding to the control where CES was not provided with the biological expert knowledge of PPIs. In this case, CES was unable to correctly identify the two functional SNPs for any confidence score. In contrast, when provided with biological expert knowledge, CES was able to correctly identify the two functional SNPs as the most frequent pair in the vast majority of the datasets considered (Table 1-1). For confidence scores greater than 0.96, CES found the correct SNPs in 100% of the datasets.

Table 1-1. Percentage of datasets in which CES successfully identified the correct SNP pairings as the most frequent, for the four confidence score scenarios considered.

Confidence Score of PPI	With Expert Knowledge	Control
0.998	100%	0%
0.963	100%	0%
0.933	96%	0%
0.916	80%	0%

For lower confidence scores, the percentage of successful trials decreased, but only to a minimum of 80%.

In Figure 1-4, we show the distributions of solution accuracy (A) and length (L) for those solutions that identified the two most frequent SNPs. These distributions were qualitatively similar for all confidence score scenarios, so we only depict the highest confidence case. For our controls, the most frequently identified SNP pairing was never the correct pair (Table 1-1), so these data correspond to incorrect models. However, the data corresponding to CES with biological expert knowledge pertain to models that successfully identified the correct SNP pair in the vast majority of cases. Within these cases, CES identified models of disease predisposition that were both more accurate and more concise (Figure 1-4).

Specifically, the average model accuracy of CES observed using the biological expert knowledge of PPIs was 0.79 while the average model accuracy without these data was only 0.64. The distribution of model accuracy for CES with expert knowledge was bimodal, with a smaller mode of less accurate solutions centered around an accuracy of 0.66. This mode corresponds to solutions that identified the correct two SNPs, but in a non-predictive model. For example, the model $x_1\%z_1$ ($A = 0.64$) contained the correct SNPs but was unable to accurately predict disease status. The average number of replicates containing the most frequent SNP pair was much larger when CES used PPI data (27.48 models) than when it did not (13.18 models). This stems from the fact that the models identified by CES without expert knowledge were always incorrect and inaccurate. In these cases, the most frequent SNP pairing was determined by chance, and the number of replicates containing this pairing was therefore significantly reduced, relative to the case of CES with expert knowledge.

For both CES with and without expert knowledge, the distributions of solution length (L) were bimodal (Figure 1-4). However, the lower mode was much larger, and the higher mode much smaller, when CES used biological expert knowledge than when it did not. This results in an average solution length of 13.08 for CES with expert knowledge and an average solution length of 21.14 for CES without expert knowledge.

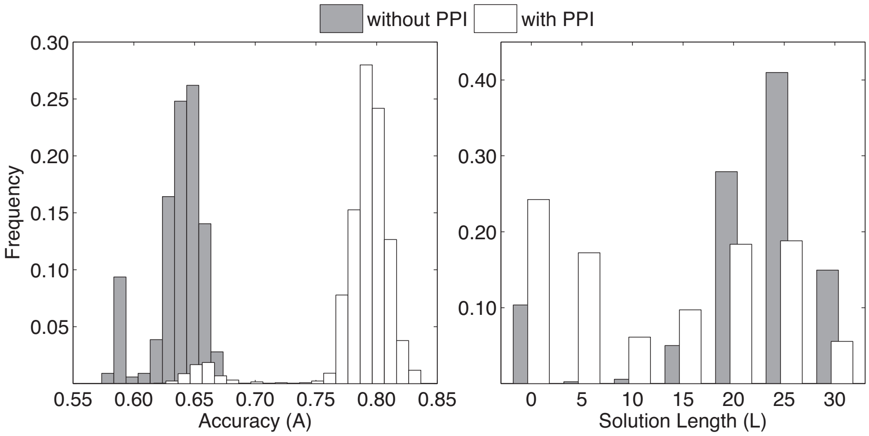


Figure 1-4. Frequency distributions of solution accuracy (left) and solution length (right), corresponding to the best final solutions found by CES with (white bars) and without (gray bars) the biological expert knowledge of protein-protein interactions. Data correspond only to those solutions that identified the most common SNP pairing of the 100 replications (see text). The bins are the same for both bar types, but are offset in the horizontal dimension for visual clarity.

5. Discussion

Our results demonstrate that CES exhibits the ability to identify simulated epistatic interactions in more concise models of disease susceptibility and with improved accuracy when using biological expert knowledge in the form of protein-protein interactions (PPI). In order to evaluate the influence of PPI confidence score strength on CES, we determined the frequency at which the system could identify our simulated two-locus interactions that represented a range of confidence scores. We found that CES achieved a success rate between 80% and 100% for low to high confidence scores. The genetic models discovered by CES with access to PPI information were significantly smaller and had significantly higher accuracies that those found by CES without expert knowledge. These results suggest that CES may be ready to tackle a wide range of real world data from studies designed to identify genetic predictors of susceptibility to common human diseases. Of course, real data brings its own unique challenges that will need to be addressed. These include low signal to noise ratio, noisy data due to laboratory errors, complex correlation structure and, of course, the challenge of providing a biological interpretation of CES models.

Although we focused exclusively on PPI as a source of biological knowledge in this study, the analysis of real world data will require giving CES access to other sources such as biochemical pathway information, gene ontology, and chromosomal location. In addition, it will be important to include other types of expert knowledge such as those derived from prior statistical or computational analyses. Our previous study showing that CES can learn to exploit a single source of knowledge from multiple different candidates suggests that a combination of biological and statistical sources of knowledge can be incorporated into the analysis of real world data (Greene et al., 2009b). Our future work with CES will focus on the application of this approach to the genetic analysis of human disease. Success in this domain will provide the ultimate validation of the hypothesis that the incorporation of complexity into genetic and evolutionary computation algorithms will facilitate solving complex problems such as those from the biomedical sciences (Banzhaf et al., 2006).

Acknowledgment

We would like to thank the participants of GPTP '09 for their insightful comments, many of which served to inspire the current study. This work was supported by NIH grants LM009012, LM010098 and AI59694.

References

Albert, R., Jeong, H., and Barabási, A.L. (2000). Error and attack tolerance of complex networks. *Nature*, 406:378–382.

Aldana, M., Balleza, E., Kauffman, S., and Resendiz, O. (2007). Robustness and evolvability in genetic regulatory networks. *Journal of Theoretical Biology*, 245:433–448.

Andrew, A.S., Karagas, M.R., Nelson, H.H., Guarrera, S., Polidoro, S., Gamberini, S., Sacerdote, C., Moore, J.H., Kelsey, K.T., Vineis, P., and Matullo, G. (2008). Assessment of multiple DNA repair gene polymorphisms and bladder cancer susceptibility in a joint italian and u.s. population: a comparison of alternative analytic approaches. *Human Heredity*, 65:105–118.

Askland, K., Read, C., and Moore, J.H. (2009). Pathway-based analyses of whole-genome association study data in bipolar disorder reveal genes mediating ion channel activity and synaptic neurotransmission. *Human Genetics*, 125:63–79.

Banzhaf, W., Beslon, G., Christensen, S., Foster, J.A., Képès, F., Lefort, V., Miller, J.F., Radman, M., and Ramsden, J.J. (2006). From artificial evolution to computational evolution: a research agenda. *Nature Reviews Genetics*, 7:729–735.

Cordell, H.J. (2009). Detecting gene-gene interactions that underlie human diseases. *Nature Reviews Genetics*, 10:392–404.

Culverhouse, R., Suarez, B.K., Lin, J., and Reich, T. (2002). A perspective on epistasis: limits of models displaying no main effect. *American Journal of Human Genetics*, 70(2):461–471.

Emily, M., Mailund, T., Hein, J., Schauser, L., and Schierup, M.H. (2009). Using biological networks to search for interacting loci in genome-wide association studies. *European Journal of Human Genetics*, 17(10):1231–1240.

Eppstein, M.J., Payne, J.L., White, B.C., and Moore, J.H. (2007). Genomic mining for complex disease traits with random chemistry. *Genetic Programming and Evolvable Machines*, 8:395–411.

Greene, C.S., Hill, D.P., and Moore, J.H. (2009a). Environmental noise improves epistasis models of genetic data discovered using a computational evolution system. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 1785–1786.

Greene, C.S., Hill, D.P., and Moore, J.H. (2009b). Environmental sensing of expert knowledge in a computational evolution system for complex problem solving in human genetics. In Riolo, R., O-Reilly, U.M., and McConaghy, T., editors, *Genetic Programming Theory and Practice VII*, pages 19–36. Springer.

Greene, C.S., White, B.C., and Moore, J.H. (2009c). An expert knowledge-guided mutation operator for genome-wide genetic analysis using genetic programming. In *Lecture Notes in Bioinformatics*, volume 4774, pages 30–40.

Greene, C.S., White, B.C., and Moore, J.H. (2009d). Sensible initialization using expert knowledge for genome-wide analysis of epistasis using genetic programming. In *Proceedings of the IEEE Congress on Evolutionary Computation*, pages 1289–1296.

Jenson, L.J., M.Kuhn, Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., Bork, P., and von Mering, C. (2009). String 8 - a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Research*, 37:D412–D416.

Jeong, H., Mason, S.P., Barabási, A.L., and Oltvai, Z.N. (2001). Lethality and centrality in protein networks. *Nature*, 411:41–42.

Kononenko, I. (1994). Estimating attributes: analysis and extensions of RELIEF. In *European Conference on Machine Learning*, pages 171–182.

Langdon, W.B. (1998). *Genetic Programming and Data Structures: Genetic Programming + Data Structures = Automatic Programming!* Kluwer Academic Publishers Group.

Moore, J.H. (2003). The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Human Heredity*, 56:73–82.

Moore, J.H., Andrews, P.C., Barney, N., and White, B.C. (2008). Development and evaluation of an open-ended computational evolution system for the genetic analysis of susceptibility to common human diseases. In *Lecture Notes in Computer Science*, volume 4973, pages 129–140.

Moore, J.H., Asselbergs, F.W., and Williams, S.M. (2010). Bioinformatics challenges for genome-wide association studies. *Bioinformatics*, 26(4):445–455.

Moore, J.H., Greene, C.S., Andrews, P.C., and White, B.C. (2009). Does complexity matter? artificial evolution, computational evolution, and the genetic analysis of epistasis in common human diseases. In Riolo, R., Soule, T., and Worzel, B., editors, *Genetic Programming Theory and Practice VI*. Springer.

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

Moore, J.H. and White, B.C. (2006). Exploiting expert knowledge in genetic programming for genome-wide genetic analysis. In *Lecture Notes in Computer Science*, volume 4193, pages 969–977.

Moore, J.H. and White, B.C. (2007). Genome-wide genetic analysis using genetic programming: The critical need for expert knowledge. In Riolo, R., Soule, T., and Worzel, B., editors, *Genetic Programming Theory and Practice IV*, pages 11–28. Springer.

Moore, J.H. and Williams, S.M. (2009). Epistasis and its implications for personal genetics. *American Journal of Human Genetics*, 85:309–320.

Payne, J.L., Greene, C.S., Hill, D.P., and Moore, J.H. (2010). Sensible initialization of a computational evolution system using expert knowledge for epistasis analysis in human genetics. In Chen, Y.P., editor, *Exploitation of Linkage Learning in Evolutionary Algorithms*, pages 215–226. Springer.

Poli, R., Langdon, W.B., and McPhee, N.F. (2008). *A Field Guide to Genetic Programming*. Published via `http://lulu.com` and freely available at `http://www.gp-field-guide.org.uk`.

Ritchie, M.D., Hahn, L.W., Roodi, N., Bailey, L.R., Dupont, W.D., Parl, F.F., and Moore, J.H. (2001). Multifactor dimensionality reduction reveals high-order interactions among estrogen metabolism genes in sporadic breast cancer. *American Journal of Human Genetics*, 69:138–147.

von Mering, C., Jensen, L.J., Snel, B., Hooper, S.D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M.A., and Bork, P. (2005). String: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Research*, 33:D433–D437.