Chapter 2
Exploring *Interestingness* in a Computational Evolution System for the Genome-Wide Genetic Analysis of Alzheimer's Disease

Jason H. Moore, Douglas P. Hill, Andrew Saykin, and Li Shen

Abstract Susceptibility to Alzheimer's disease is likely due to complex interaction among many genetic and environmental factors. Identifying complex genetic effects in large data sets will require computational methods that extend beyond what parametric statistical methods such as logistic regression can provide. We have previously introduced a computational evolution system (CES) that uses genetic programming (GP) to represent genetic models of disease and to search for optimal models in a rugged fitness landscape that is effectively infinite in size. The CES approach differs from other GP approaches in that it is able to learn how to solve the problem by generating its own operators. A key feature is the ability for the operators to use expert knowledge to guide the stochastic search. We have previously shown that CES is able to discover nonlinear genetic models of disease susceptibility in both simulated and real data. The goal of the present study was to introduce a measure of *interestingness* into the modeling process. Here, we define interestingness as a measure of non-additive gene-gene interactions. That is, we are more interested in those CES models that include attributes that exhibit synergistic effects on disease risk. To implement this new feature we first pre-processed the data to measure all pairwise gene-gene interaction effects using entropy-based methods. We then provided these pre-computed measures to CES as expert knowledge and as one of three fitness criteria in three-dimensional Pareto optimization. We applied this new CES algorithm to an Alzheimer's disease data set with approximately 520,000 genetic attributes. We show that this approach discovers more interesting models with the added benefit of improving classification accuracy. This study demonstrates the applicability of CES to genome-wide genetic analysis using expert knowledge derived from measures of interestingness.

J.H. Moore (✉) • D.P. Hill • A. Saykin • L. Shen
The Geisel School of Medicine at Dartmouth, One Medical Center Drive, HB7937, Lebanon, NH 03756, USA
e-mail: Jason.H.Moore@Dartmouth.edu; douglas.hill@Dartmouth.edu

R. Riolo et al. (eds.), *Genetic Programming Theory and Practice XI*, Genetic and Evolutionary Computation, DOI 10.1007/978-1-4939-0375-7_2,

Keywords Computational evolution • Genetic epidemiology • Epistasis • Gene-gene interactions

1 Introduction

The genetic analysis of Alzheimer's disease has had mixed results despite the availability of genome-wide measures of genetic variation, reviewed by Bertram and Tanzi (2012). The single best genetic risk factor is variation in the Apolipoprotein E (ApoE) gene with odds ratios (OR) between 3 and 20 depending on the particular combination of alleles. Aside from this strong genetic risk factor, approximately 9 others have been found with much smaller OR between 0.8 and 1.3. Some of the unexplained heritability of this common disease is likely due to complex gene-gene interactions or epistasis. This type of genetic effect cannot be predicted by the effects of single genetic variants and has been largely ignored by genetic and epidemiological studies (Moore and Williams 2009). Several recent studies have highlighted the importance of gene-gene interactions in Alzheimer's disease and thus provide a foundation for further investigation using data mining and machine learning methods that are ideally suited to detecting nonlinear effects of attribute combinations (Combarros et al. 2009; Lehmann et al. 2012; Bullock et al. 2013). Although promising, these studies only explored pairwise gene-gene interactions. The search for higher-order gene-gene interactions in an effectively infinite search space is a significant statistical and computational problem (Moore et al. 2010).

The overarching goal of this study is to explore data mining and machine learning alternatives to parametric statistical methods for the genetic analysis of complex human diseases. In particular, we are interested in computational intelligence methods that are able to learn how to solve genetic analysis problems as a human would. This general approach involves pre-processing the data to identify useful information, implementing a machine learning algorithm that is able to model nonlinear effects, implementing a stochastic search algorithm that is able to exploit expert knowledge and implementing post-processing methods that are able to enhance statistical and biological interpretation of the results. We have previously developed computational intelligence methods for genetic analysis that are based on genetic programming (GP). Genetic programming is an automated computational discovery tool that is inspired by Darwinian evolution by natural selection (Koza 1992; Banzhaf et al. 1998). The goal of GP is to 'evolve' computer programs to solve complex problems. This is accomplished by first generating or initializing a population of random computer programs that are composed of the basic building blocks needed to solve or approximate a solution to the problem. Genetic programming and its many variations have been applied successfully in a wide range of different problem domains including bioinformatics (Fogel and Corne 2003) and genetic analysis (Moore et al. 2010). GP is an attractive approach to the genetic analysis problem because it is inherently flexible, stochastic, parallel and easily adapted to exploit expert knowledge. The goal of the present study was to

build on a GP-based computational evolution strategy (CES) for genetic analysis. We introduce here a measure of interestingness into the modeling process. Here, we define interestingness as a measure of nonlinear gene-gene interactions. That is, we are more interested in those CES models that include attributes that exhibit synergistic effects on disease. To implement this new feature we first pre-processed the data to measure all pairwise gene-gene interaction effects using entropy-based methods. We then provided a small subset of these pre-computed measures to CES as expert knowledge and as one of the fitness criteria in three-dimensional Pareto optimization. We applied this new CES algorithm to an Alzheimer's disease data set with approximately 520,000 genetic attributes.

2 Computational Evolution

It has been suggested that the incorporation of greater biological realism into GP may improve its ability to solve complex, real-world problems. 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 biotic systems. With this in mind, 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 and Williams 2009; Greene et al. 2009a,b; 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 (Moore and White 2007), can be incorporated via environmental sensing (Greene et al. 2009b) and population initialization (Payne et al. 2010) to improve system performance. In addition, we recently showed that biological expert knowledge in the form of protein-protein interactions could be used to guide CES toward valid gene-gene interaction models (Pattin et al. 2010). We also showed how visualization of CES results could improve the modeling process (Moore et al. 2011). More recently, we have demonstrated how two-dimensional Pareto optimization can be used to help guide the search (Moore et al. 2013). Here, we introduce a measure of *interestingness* and show how it can be incorporated as expert knowledge and into a three-way Pareto optimization. We briefly introduce both of these in turn below.

Pre-processing Measure of Interestingness

A central goal of this study is to identify new genetic models of Alzheimer's disease that have not been revealed in prior studies. We are particularly interested in those

models that are comprised of nonlinear gene-gene interactions that are not predicted from the independent effects of single genetic variants. We used here a measure of gene-gene interaction introduced to the genetics community by Moore et al. (2006) that is based on information theory or entropy. Specifically, we measure *interaction information* as the information gain due to the synergistic effects of two genetic variants after the individual effects have been subtracted out. This provides a measure of interestingness that can be used to help guide the proposed CES methodology toward models that are new and different from what has been previously discovered.

Pareto Optimization

A common approach for addressing overfitting in data mining and machine learning is to use cross-validation as an estimate of the generalizability of a model. Unfortunately, implementation of cross-validation methods in conjunction with stochastic methods such as GP can be complex given these algorithms are likely to find different models in each division of the data. Pareto optimization (reviewed by Lamont and VanVeldhuizen (2002)) offers a viable alternative and has been shown to be quite effective in the context of GP (Smits and Kotanchek 2004). Pareto optimization balances several different model objectives that are each treated equally. We have previously used classification accuracy and model size as our two objectives (Moore et al. 2013). Here, we add a third dimension defined by the interaction information measure of interestingness. For a given GP population, models for which there are no better as measured by accuracy, model size and interestingness are selected. This subset of Pareto-optimal models is referred to as the Pareto front. The goal of the present study was to introduce the use of interestingness as a third dimension in Pareto optimization of CES. This allows CES to explore models that might have good interestingness but poor accuracy or error.

Post-processing of CES Models

We have previously demonstrated that post-processing CES results can improve model discovery (Moore et al. 2011, 2013). In other words, there is value in analyzing the results of a CES run and extracting knowledge from that analysis that can be used to improve interpretation of CES models. Here, we use network analysis and visualization to help interpret the genetic effects in CES models (Hu et al. 2013).

3 Methods

In this section, we first present a summary of our computational evolution system (CES) for open-ended genetic analysis of complex human diseases. We then discuss our implementation of the pre-processing methods, Pareto optimization and post-processing extensions and their application to Alzheimer's disease.

Computational Evolution System

In Fig. 2.1, 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. 2.1D), which compete with one another within

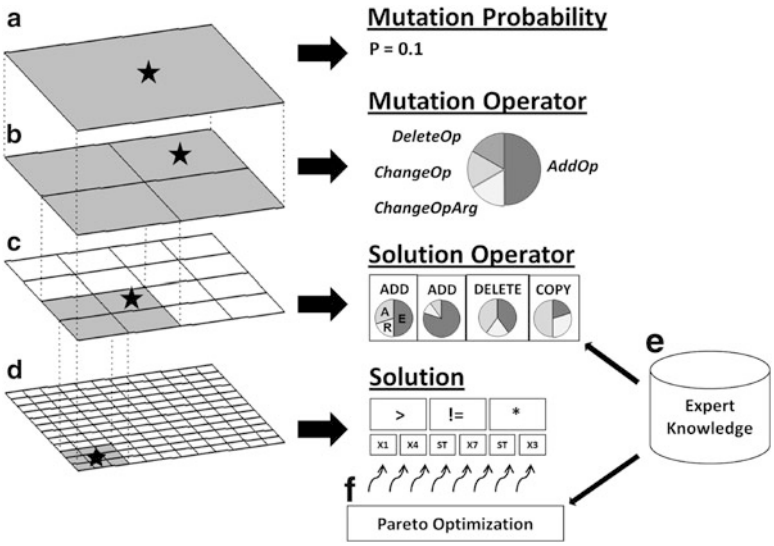

Fig. 2.1 Visual overview of our computational evolution system for discovering symbolic discriminant functions that differentiate disease subjects from healthy subjects using information about single nucleotide polymorphisms (SNPs). The hierarchical structure is shown on the *left* while some specific 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 or SNP) 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 consists of some combination of the ADD, DELETE and COPY functions each with their respective set of probabilities that define whether attributes are added, deleted or copied randomly, using an attribute archive (memory) or just randomly. In this implementation of CES, we use pre-processed expert knowledge (*E*) with Pareto optimization (*F*) to help reduce overfitting. The top two levels of the hierarchy (*A* and *B*) exist to generate variability in the operators that modify the solutions. This system allows operators of arbitrary complexity to modify solutions. A 12 × 12 grid is shown here as an example. A 36 × 36 grid was used in the present study

spatially-localized, overlapping neighborhoods. The second layer of the hierarchy contains a lattice of arbitrarily complex solution operators (Fig. 2.1C), which operate on the solutions in the lower layer. The third layer of the hierarchy contains a lattice of mutation operators (Fig. 2.1B), 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. 2.1A). CES includes a source of expert knowledge (Fig. 2.1E) that can be used to with the solution operators and as part of Pareto optimization (Fig. 2.1F). CES also possesses an attribute archive, which stores the frequencies with which attributes are used. The solution operators can then exploit these data to bias the construction of solutions toward frequently utilized attributes. We did not use the attribute archive in the present study.

Solution Representation, Fitness Evaluation, Selection, and Pareto Optimization

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. 2.1). The function set contains $+, -, *, /, \%, <, <=, >, >=, ==, !=$, where % denotes protected modulus. Operands are either SNPs, constants, or the output of another element in the stack.

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 the medians of these two distributions. Each of the possible relationships between S_0 and S_i $(<, <=, >=, >)$ is tested across all individuals, and the one with best overall accuracy is chosen to classify whether individuals are healthy or diseased.

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:

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

Solution length can be assessed in several ways. The number of elements in the classifier is the most straightforward. Since many solutions leave results on the stack that do not contribute to the classification, we can define "number of relevant elements" as only those contributing to the result. Finally we can count the number of unique SNPs in the relevant elements. We have chosen to use this as the measure of length in the present study as it makes the resulting solutions easier to analyze.

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. In all previous CES implementations election has been both synchronous and elitist, such that the solution of highest fitness within a given neighborhood was always selected to repopulate the focal site of that neighborhood. In the present study, selection proceeds in two stages modeled after Pareto domination tournaments and fitness sharing described by Horn et al. (1994). We used classification accuracy, number of attributes in the model and interaction information as the axes in the Pareto optimization. Here, the sum of the interaction information for all pairs of attributes in a model is the measure of interestingness. First all dominated solutions and solutions evaluating to a constant are removed from competition. A solution is dominated if there exists any solution of lesser or equal length with lesser classification error, or lesser length and equal error. If no solution survives this stage, one of the nine is chosen with equal probability. If more than one solution survives, each is assigned a probability and a roulette wheel selection is made. Higher selection probability is assigned to a solution if there are relatively fewer solutions of that length in the lattice, in order to prevent convergence on solutions of a single length. For the present results we made the probability inversely proportional to the square of the number of existing solutions of the same length. 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 1–15 elements subject to the constraint that they produce a valid output that is not constant for all input. The functions are selected at random with uniform probability from the function set.

Solution Operators

CES allows for the evolution of arbitrarily complex variation operators used to modify solutions. This is achieved by initializing the solution operator lattice (Fig. 2.1C) with a set of basic building blocks which can be recombined in any way to form composite operators throughout the execution of the program. The action of some of these operators is influenced by any of several types of expert knowledge (EK) that CES recognizes. In this study we have used one type of EK, Association EK. Association EK is used to help CES to more quickly find solutions using specific combinations of attributes or SNPs. Here, we used a measure of interaction information as the expert knowledge. Adding and altering attributes is based on a lookup table that is constructed from the strength of interactions between pairs of attributes. Because 521,028 attributes have over 2.7×2^{11} pairs, it was impossible to pre-compute and store all pairs in the memory available. We pre-computed all pairs but stored only the 93,606 most strongly interacting pairs, a tiny fraction of the total. The following are the building blocks and the way they are influenced by Association EK.

1. ADD: Inserts a randomly generated element into the solution at a randomly selected position. If the element immediately before this position is one of the pairs of strongly interacting elements, preferentially chooses one of these interacting elements.
2. ALTER: Modifies either the function or an argument of a randomly selected element. If it chooses an attribute as the new argument, that attribute is selected as above in ADD.
3. COPY: Within a randomly selected neighboring solution, randomly selects an element and inserts it into a randomly selected position in the focal solution.
4. DELETE: Removes an element from a randomly selected position or a position that is probabilistically selected by the lookup table.
5. REPLACE: Within a randomly selected neighboring solution, randomly selects a source position. In the focal solution, randomly selects a destination position. Replaces everything between the destination position and the end (root) of the focal solution with everything between the source position and the end of the source solution.

The solution operators reside on a periodic, toroidal lattice of coarser granularity than the solution lattice (Fig. 2.1C). Each site is occupied by a single solution operator, which is assigned to operate on 3×3 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. For this purpose we assigned the fitness of a solution as we have done in previous studies: balanced accuracy with a small penalty for number of elements. We did not adapt a Pareto tournament to the selection of solution operators.

Mutation Operators

The third level of the hierarchy contains the mutation operators, which are used to modify the solution operators (Fig. 2.1B). 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 (Fig. 2.1A).

Alzheimer's Disease Data

The data used in this study came from the Alzheimer's Disease Neuroimaging Initiative (ADNI) that began on October 1, 2004. The study takes functional MRIs every 6–12 month of patients in three categories: those who are neuro-typical, those with mild cognitive impairment, and those with Alzheimer's disease. A total of 521,028 single-nucleotide polymorphisms (SNPs) were measured across the human genome in a total of 740 subjects. Here, we used neuro-typical patients as the control subjects and those with mild cognitive impairment or Alzheimer's disease as the case subjects creating a binary class or outcome. The goal of the modeling exercise is to identify the optimal subset of SNPs along with the optimal mathematical model that is predictive of the binary class.

Pre-processing, Experimental Design and Post-processing

The goal of this study was to apply CES to the genetic analysis of Alzheimer's disease. We first pre-processed the data by estimating the interaction information for all pairs of SNPs as described by Moore et al. (2006). We considered pairs of SNPs that have higher interaction information more interesting. This pre-processed interestingness measure was used as expert knowledge (Attribute EK) in the CES solution modifiers and as an additional axis in a three-way Pareto optimization.

Each CES run was conducted with a 36×36 grid of solutions for 2,000 generations. CES was implemented in a hierarchical framework inspired by the age-layered population structure algorithm or ALPS (Hornby 2006). Here, we implemented a depth six binary tree where each node represents a CES run with the leaves of the tree representing the initial runs. Each higher node run is initialized with Pareto optimal solutions from the lower nodes. This was performed 10 times with different sets of random seeds. This analysis was repeated with and without expert knowledge and with two or three-way Pareto optimization where the two-way Pareto had only classification accuracy and model size as axes. The three-way Pareto optimization included interaction information as the measure of interestingness. Thus, we had 4 treatment groups each with 10 runs. We used two-way analysis of variance (ANOVA) to compare the mean classification accuracy and interaction information among the two expert knowledge groups and the two Pareto groups. We also considered the interaction effect of both expert knowledge and Pareto. All results were considered statistically significant at the 0.05 significance level.

We reported the best models discovered from each set of runs under each of the four experimental conditions. We used the visualization of statistical epistasis networks (ViSEN) method developed by Hu et al. (2013) to visualize the two-way and three-way gene-gene interactions among SNPs in the best CES models. This allows us to interpret the nature of the genetic effects in a visual manner.

4 Results

Figure 2.2 summarizes the average classification error (1-accuracy, y-axis) of the most accurate models, model size measured by number of attributes (x-axis) and the interaction information or interestingness summed across all pairs of SNPs in a model (size of circle). The results are shown according to whether two-way or three-way Pareto optimization was used with or without expert knowledge. As expected, three-way Pareto with expert knowledge showed the highest interaction information and thus the highest interestingness. An unexpected result was that these models also had the lowest average classification error. Thus, the most interesting models were also the ones that classified disease best. The ANOVA results confirm this. We found that two-way vs. three-way Pareto had a significant effect on both classification error ($p = 0.006$) and interaction information ($p < 0.001$). We also found that whether we included expert knowledge had a significant effect on classification error ($p < 0.001$) and interaction information ($p < 0.001$). In addition, Pareto and expert knowledge had a strong interaction effect on classification error ($p < 0.001$) and interaction information ($p < 0.001$).

Figure 2.3 illustrates the overall best model discovered by CES. This model had a classification accuracy of 0.738 (error = 0.262) and consisted of 7 attributes or SNPs. We selected this model as a compromise between model size, accuracy and interestingness. It was discovered by CES run with the three-way Pareto

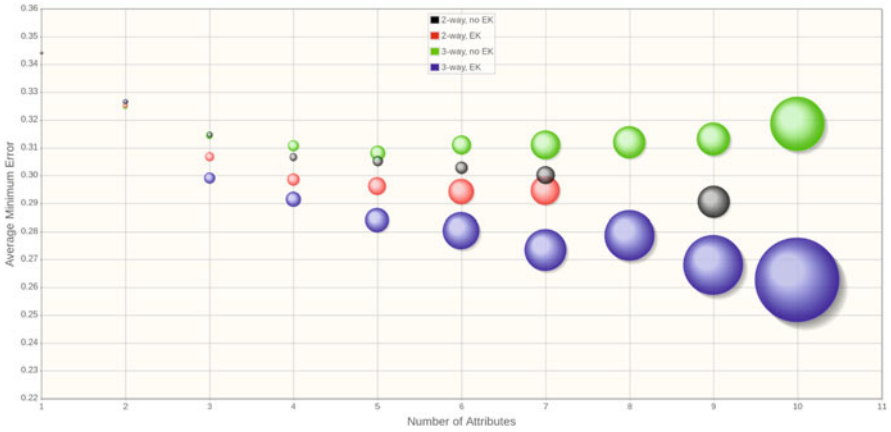

Fig. 2.2 Visual summary of the CES results. Shown are the average interaction information scores (*size of circle*) among the best models on the Pareto front for the 10 CES runs. Results are shown by the average minimum error (1-accuracy – y-axis) and the number of attributes in the model (x-axis). Results are also shown according to the CES method used. *Black* indicates two-way Pareto optimization with no expert knowledge from pre-processing the data for interaction information scores. *Red* indicates two-way Pareto optimization with expert knowledge. *Green* indicates three-way Pareto optimization with no expert knowledge. *Blue* indicates three-way Pareto optimization with expert knowledge. Note that three-way Pareto optimization with expert knowledge achieves the lowest error across model sizes and has the highest interestingness

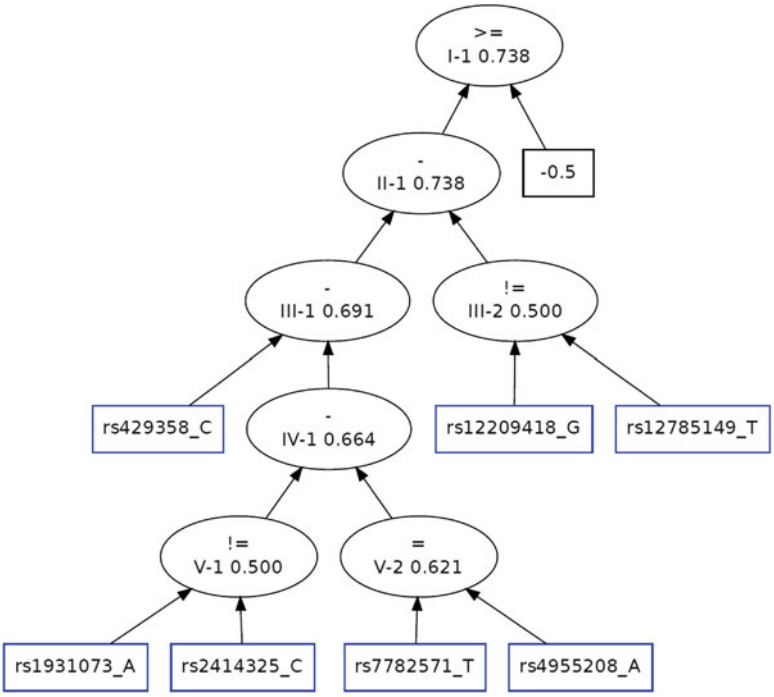

Fig. 2.3 The overall best Pareto-optimal model discovered by CES using three-way Pareto optimization and expert knowledge. The model includes attributes (*rectangles*), constants (*squares*) and mathematical functions or nodes (*ovals*). Each model outputs a discriminant score for each subject in the data set. These scores are then used for classification in a discriminant analysis. The numbers shown within each oval are the classification accuracies at each level in the tree

optimization and with using expert knowledge to guide the search. Figure 2.4 shows a statistical interpretation of the model. Note the large independent effect of SNP rs429358 which by itself has an accuracy of about 0.656. Each of the other six SNPs in the model has weak independent effects and strong pairwise interactions with at least one other SNP. Together they improve the accuracy of the classifier from 0.656 for a model with just the large independent effect to 0.738 for the collection of all seven attributes or SNPs.

5 Summary and Discussion

Alzheimer's disease is likely the result of complex interactions among many genetic and environmental factors. We have demonstrated here how a computational evolution system (CES) can be used to identify new models of disease susceptibility in genome-wide genetic studies with hundreds of thousands of attributes. Despite the size and complexity of the search space, CES was able to find models with high accuracy.

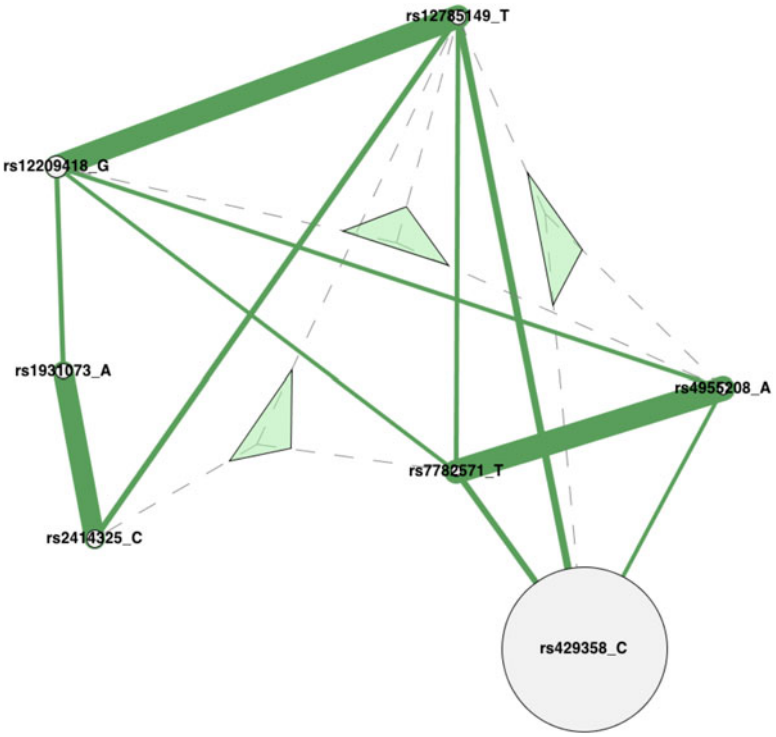

Fig. 2.4 Visualization of the statistical interaction network among the attributes in the overall best CES model. The size of the circle is proportional to the independent effects of each genetic variant on disease risk. The lines connecting each node are proportional to the interaction information for that pair of variants. The *triangles* represent the three-way information gain. Note that the genetic variants with the strongest pairwise interactions are children of the same functions in the CES classification model

The overall best model discovered consisted of seven attributes or SNPs achieving an accuracy of 0.738. The SNP with the large independent effect (rs429358) is located in the ApoE gene that is a known strong risk factor for Alzheimer disease. CES was able to improve on this local minimum by adding six additional SNPs to the model that each is part of a relatively strong gene-gene interaction pair. Interestingly, each pair of strong gene-gene interactions was co-located in the model as children of the same function. SNP rs1931073 is in an intergenic region near the PPAP2B gene that is involved with cell adhesion and cell-cell interactions. SNP rs2414325 is in a gene called UNC13C for which not much is known about its function. SNP rs7782571 is near the ISPD gene that is known to be mutated in rare diseases such as Walker-Warburg syndrome that is known to have brain anomalies. SNP rs4955208 is in the OSBPL10 gene that codes for an intracellular lipid receptor. SNP rs12209418 is in the PKIB gene, a protein kinase inhibitor, and is associated with neuronitis or inflammation of the neurons.

SNP rs12785149 is in the FAM107B gene whose function is not well known. Of these seven genes, only ApoE is present in the Alzgene database (Bertram et al. 2012) that provides an unbiased catalog of known genetic risk factors for Alzheimer's disease. The other six could represent new and novel discoveries. Based on their biology, it is plausible that all of them could be related to the disease process. For example, OSBPL10 is involved with lipid metabolism which is a known component of Alzheimer's disease pathobiology. At this point they remain novel hypotheses that will need to be tested in other data.

We introduced here the concept of interestingness defined as the degree of gene-gene interaction in a given CES model. Interestingness was introduced into CES in two different ways. First, pre-computed pairwise interaction information scores were used as expert knowledge during the model building process. Second, we used interaction information as an additional axis in a three-way Pareto optimization that also included model error and size. This allowed CES to explore models with high interestingness but low accuracy thus promoting diversity. Interestingness has been explored previously for use with data mining (see survey by Geng and Hamilton (2006)). Geng and Hamilton (2006) review nine specific criteria for determining whether a model or result is interesting. The first is conciseness or parsimony. The second is coverage (i.e. applies to a large subset of the data). The third is reliability that is measured by the accuracy or error of a classifier. The fourth is peculiarity that measures how far away a finding is from others. The fifth is diversity that measures how different the elements of a model are. The sixth is novelty (i.e. the result is new). The seventh is *surprisingness* that measures how unexpected the result is based on prior knowledge. The eighth is utility that measures how useful the result is. The final criterion is actionability that measures how applicable a result is to a particular domain. Each of these criteria can be grouped into objective and subjective categories. For example, conciseness, coverage, reliability peculiarity and diversity are all objective measures because they can be computed using an algorithm or mathematical function. On the other hand, novelty, *surprisingness*, utility and actionability are all subjective and dependent on the experience and knowledge of the particular domain expert. In this study we used interaction information as a pre-processed measure of interestingness that could be used to guide the CES. This is an objective measure because we are computing a specific measure from the data. However, it can also be seen as a subjective measure because some in the field do not think gene-gene interactions are important.

Human genetics research has been focused almost exclusively on detecting single attribute effects on disease risk that completely ignore the complexity of the genotype-phenotype relationship. Our primary objective here was to further develop a computational evolution system or CES for the discovery of new and novel genetic associations that embrace the complexity of the problem. We were encouraged by the results of this study and think measures of interestingness have a very important role to play as we attempt to bring expert knowledge back into the modeling process. Our future studies will further explore the role of interestingness for improving computational intelligence modeling strategies in this domain.

Acknowledgements This work was supported by NIH grants LM011360, LM009012, LM010098 and AI59694. We would like to thank the participants of present and past Genetic Programming Theory and Practice Workshops (GPTP) for their stimulating feedback and discussion that helped formulate some of the ideas in this paper.

References

Banzhaf W, Francone FD, Keller RE, Nordin P (1998) Genetic programming: an introduction on the automatic evolution of computer programs and its applications. Morgan Kaufmann, San Francisco

Banzhaf W, Beslon G, Christensen S, Foster J, Képès F, Lefort V, Miller J, Radman M, Ramsden J (2006) From artificial evolution to computational evolution: a research agenda. Nat Rev Genet 7:729–735

Bertram L, Tanzi RE (2012) The genetics of Alzheimer's disease. Prog Mol Biol Transl Sci 107:79–100. doi:10.1016/B978-0-12-385883-2.00008-4

Bullock JM, Medway C, Cortina-Borja M, Turton JC, Prince JA, Ibrahim-Verbaas CA, Schuur M, Breteler MM, van Duijn CM, Kehoe PG, Barber R, Coto E, Alvarez V, Deloukas P, Hammond N, Combarros O, Mateo I, Warden DR, Lehmann MG, Belbin O, Brown K, Wilcock GK, Heun R, Kolsch H, Smith AD, Lehmann DJ, Morgan K (2013) Discovery by the epistasis project of an epistatic interaction between the GSTM3 gene and the HHEX/IDE/KIF11 locus in the risk of Alzheimer's disease. Neurobiol Aging 34(4):1309.e1–1309.e7. doi:10.1016/j.neurobiolaging.2012.08.010

Combarros O, van Duijn CM, Hammond N, Belbin O, Arias-Vasquez A, Cortina-Borja M, Lehmann MG, Aulchenko YS, Schuur M, Kolsch H, Heun R, Wilcock GK, Brown K, Kehoe PG, Harrison R, Coto E, Alvarez V, Deloukas P, Mateo I, Gwilliam R, Morgan K, Warden DR, Smith AD, Lehmann DJ (2009) Replication by the epistasis project of the interaction between the genes for IL-6 and IL-10 in the risk of Alzheimer's disease. J Neuroinflammation 6:22. doi:10.1186/1742-2094-6-22

Fogel GB, Corne DW (eds) (2003) Evolutionary computation in bioinformatics. Morgan Kaufmann, San Francisco

Geng L, Hamilton HJ (2006) Interestingness measures for data mining: a survey. ACM Comput Surv 38(3). doi:10.1145/1132960.1132963, http://doi.acm.org/10.1145/1132960.1132963

Greene CS, Hill DP, Moore JH (2009a) Environmental noise improves epistasis models of genetic data discovered using a computational evolution system. In: Proceedings of the 11th annual conference on genetic and evolutionary computation, GECCO'09, Montreal. ACM, New York, pp 1785–1786. doi:10.1145/1569901.1570160, http://doi.acm.org/10.1145/1569901.1570160

Greene CS, Hill DP, Moore JH (2009b) Environmental sensing of expert knowledge in a computational evolution system for complex problem solving in human genetics. In: Riolo RL, O'Reilly UM, McConaghy T (eds) Genetic programming theory and practice VII. Genetic and evolutionary computation. Springer, Ann Arbor, chap 2, pp 19–36

Horn J, Nafpliotis N, Goldberg DE (1994) A niched pareto genetic algorithm for multiobjective optimization. In: Proceedings of the first IEEE conference on evolutionary computation, IEEE world congress on computational intelligence, Orlando, vol 1, pp 82–87. doi:10.1109/ICEC.1994.350037, http://dx.doi.org/10.1109/ICEC.1994.350037

Hornby GS (2006) ALPS: the age-layered population structure for reducing the problem of premature convergence. In: Proceedings of the 8th annual conference on genetic and evolutionary computation, GECCO'06, Seattle. ACM, New York, pp 815–822. doi:10.1145/1143997.1144142, http://doi.acm.org/10.1145/1143997.1144142

Hu T, Chen Y, Kiralis JW, Moore JH (2013) ViSEN: methodology and software for visualization of statistical epistasis networks. Genet Epidemiol 37(3):283–285. doi:10.1002/gepi.21718

Koza JR (1992) Genetic programming: on the programming of computers by means of natural selection (complex adaptive systems), 1st edn. A Bradford Book. MIT Press, London. http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0262111705

Lamont GB, VanVeldhuizen DA (2002) Evolutionary algorithms for solving multi-objective problems. Kluwer Academic, Norwell

Lehmann DJ, Schuur M, Warden DR, Hammond N, Belbin O, Kolsch H, Lehmann MG, Wilcock GK, Brown K, Kehoe PG, Morris CM, Barker R, Coto E, Alvarez V, Deloukas P, Mateo I, Gwilliam R, Combarros O, Arias-Vasquez A, Aulchenko YS, Ikram MA, Breteler MM, van Duijn CM, Oulhaj A, Heun R, Cortina-Borja M, Morgan K, Robson K, Smith AD (2012) Transferrin and HFE genes interact in Alzheimer's disease risk: the epistasis project. Neurobiol Aging 33(1):202.e1–202.e13. doi:10.1016/j.neurobiolaging.2010.07.018

Moore JH, White BC (2007) Tuning ReliefF for genome-wide genetic analysis. In: Proceedings of the 5th European conference on evolutionary computation, machine learning and data mining in bioinformatics, EvoBIO'07, Valencia. Springer, Berlin/Heidelberg, pp 166–175. http://dl.acm.org/citation.cfm?id=1761486.1761502

Moore JH, Williams SM (2009) Epistasis and its implications for personal genetics. Am J Hum Genet 85(3):309–320. doi:10.1016/j.ajhg.2009.08.006, http://dx.doi.org/10.1016/j.ajhg.2009.08.006

Moore JH, Parker JS, Olsen NJ, Aune TM (2002) Symbolic discriminant analysis of microarray data in autoimmune disease. Genet Epidemiol 23(1):57–69

Moore JH, Gilbert JC, Tsai CT, Chiang FT, Holden T, Barney N, White BC (2006) A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. J Theor Biol 241(2):252–261. doi:10.1016/j.jtbi.2005.11.036, http://dx.doi.org/10.1016/j.jtbi.2005.11.036

Moore JH, Andrews PC, Barney N, White BC (2008) Development and evaluation of an open-ended computational evolution system for the genetic analysis of susceptibility to common human diseases. In: Marchiori E, Moore JH (eds) EvoBIO'08, Naples. Lecture notes in computer science, vol 4973. Springer, pp 129–140

Moore JH, Asselbergs FW, Williams SM (2010) Bioinformatics challenges for genome-wide association studies. Bioinformatics 26(4):445–455. doi:10.1093/bioinformatics/btp713

Moore JH, Hill DP, Fisher JM, Lavender N, Kidd LC (2011) Human-computer interaction in a computational evolution system for the genetic analysis of cancer. In: Riolo R, Vladislavleva E, Moore JH (eds) Genetic programming theory and practice IX. Genetic and evolutionary computation. Springer, Ann Arbor, chap 9, pp 153–171. doi:10.1007/978-1-4614-1770-5-9

Moore JH, Hill DP, Sulovary A, Kidd L (2013) Genetic analysis of prostate cancer using computational evolution, pareto-optimization and post-processing. In: Riolo RL, Moore JH, Ritchie MD, Vladislavleva K (eds) Genetic programming theory and practice X. Genetic and evolutionary computation. Springer, Ann Arbor, pp 87–101

Pattin KA, Payne JL, Hill DP, Caldwell T, Fisher JM, Moore JH (2010) Exploiting expert knowledge of protein-protein interactions in a computational evolution system for detecting epistasis. In: Riolo R, McConaghy T, Vladislavleva E (eds) Genetic programming theory and practice VIII. Genetic and evolutionary computation, vol 8. Springer, Ann Arbor, chap 12, pp 195–210. http://www.springer.com/computer/ai/book/978-1-4419-7746-5

Payne J, Greene C, Hill D, Moore J (2010) Sensible initialization of a computational evolution system using expert knowledge for epistasis analysis in human genetics. In: Exploitation of linkage learning in evolutionary algorithms. Springer, Ann Arbor, chap 10, pp 215–226

Smits G, Kotanchek M (2004) Pareto-front exploitation in symbolic regression. In: O'Reilly UM, Yu T, Riolo RL, Worzel B (eds) Genetic programming theory and practice II. Springer, Ann Arbor, chap 17, pp 283–299. doi:10.1007/0-387-23254-0-17

http://www.springer.com/978-1-4939-0374-0

Genetic Programming Theory and Practice XI
Riolo, R.; Moore, J.H.; Kotanchek, M. (Eds.)
2014, XIV, 227 p. 68 illus., 32 illus. in color., Hardcover
ISBN: 978-1-4939-0374-0