A Model Free Method to Generate Human Genetics Datasets with Complex Gene-Disease Relationships

Casey S. Greene, Daniel S. Himmelstein, and Jason H. Moore

Dartmouth Medical School, Lebanon, NH 03756, USA
{Casey.S.Greene,Daniel.S.Himmelstein,Jason.H.Moore}@dartmouth.edu
http://www.epistasis.org

Abstract. A goal of human genetics is to discover genetic factors that influence individuals' susceptibility to common diseases. Most common diseases are thought to result from the joint failure of two or more interacting components instead of single component failures. This greatly complicates both the task of selecting informative genetic variations and the task of modeling interactions between them. We and others have previously developed algorithms to detect and model the relationships between these genetic factors and disease. Previously these methods have been evaluated with datasets simulated according to pre-defined genetic models. Here we develop and evaluate a model free evolution strategy to generate datasets which display a complex relationship between individual genotype and disease susceptibility. We show that this model free approach is capable of generating a diverse array of datasets with distinct gene-disease relationships for an arbitrary interaction order and sample size. We specifically generate six-hundred pareto fronts; one for each independent run of our algorithm. In each run the predictiveness of single genetic variation and pairs of genetic variations have been minimized, while the predictiveness of third, fourth, or fifth order combinations is maximized. This method and the resulting datasets will allow the capabilities of novel methods to be tested without pre-specified genetic models. This could improve our ability to evaluate which methods will succeed on human genetics problems where the model is not known in advance. We further make freely available to the community the entire pareto-optimal front of datasets from each run so that novel methods may be rigorously evaluated. These 56,600 datasets are available from http://discovery.dartmouth.edu/model_free_data/.

1 Introduction

Advances in genotyping technologies are changing the way geneticists measure genetic variation. It is now technologically feasible to measure more than one million variations from across the human genome. Here we focus on SNPs, or single nucleotide polymorphisms. A SNP is a single point in a DNA sequence that differs between individuals. A major goal in human genetics is to link the

C. Pizzuti, M.D. Ritchie, and M. Giacobini (Eds.): EvoBIO 2010, LNCS 6023, pp. 74–85, 2010.

state of these SNPs to disease risk. The standard approach to this problem is to measure the genotypes of people with and without a disease of interest across hundreds of thousands to millions of SNPs. Each of these SNPs is then tested individually for an association with the disease of interest. The goal is to discover SNPs that reliably predict disease susceptibility across many samples [1,2]. This approach has had limited success and the discovery of robust single SNP associations has been difficult to attain [3,4,5]. Even in the cases where single SNP associations have validated in independent samples, the SNPs often cannot be combined into effective classifiers of disease risk [6]. These studies, by only examining the association of single SNPs, ignore complex interactions that may be critical to understanding disease susceptibility.

The term for complex gene-gene interactions that influence disease susceptibility is epistasis. It is now recognized that studies which ignore epistasis may neglect informative markers [7,8,9,10]. Furthermore, epistasis is thought to play a critical role in the understanding of disease because of the complexity present in cellular and biological systems [11] and because it has been well characterized for other complex traits [7,12]. Detecting and characterizing epistasis in all but small datasets is difficult. Examining gene-disease relationships in the context of epistasis requires the consideration of the joint effect of SNPs and an exhaustive analysis of all possible interactions requires the enumeration of every potential set of SNPs [13]. When datasets contain many SNPs, combinatorial methods which evaluate each such combination are not feasible [14].

In human genetics we have, therefore, been confronted by a chicken and egg problem. We believe that it is likely that complex interactions occur, but without methods to detect these interactions in large datasets, we lack the ability to find them. Without found and validated interactions that lead to disease we lack the ability to test new methods on actual genetic datasets. Thus far the problem has been approached with datasets simulated according to hypothetical genetic models as in Velez et al. [15] and Greene et al. [10] among many others. Methods are tested for their ability to find a disease model placed in the datasets. This approach is useful but limited by the diversity and representativeness of the genetic models. The work we present here uses an evolution strategy to generate datasets containing complex genetic interactions that lead to disease without imposing a specific genetic model on the datasets.

2 Evolution Strategies

Evolution Strategies are algorithms modeled after natural evolution. The combination of several key evolutionary concepts, such as natural selection, population sizes, and mutation rates, produces algorithms capable of finding sought after members of a solution space [16]. Multiple generations allow evolution strategies to direct their results towards an ideal solution by preserving beneficial mutations. One key difference between mutation driven strategies, like the one implemented in our method, and genetic algorithms is the absence of recombination [17]. Recombination, the computational equivalent of genetic crossover, consists

of creating new individuals in a population by combining the characteristics of multiple members of the previous generation. Since recombination relies on the exchange of discrete blocks of information, crossover is only appropriate when clear building blocks can be defined [18]. Here it is unclear whether a building block would be a set of individuals or a set of SNPs. Therefore, because it is unclear what the proper building blocks would be, we do not use recombination. Evolutionary algorithms that lack recombination have proven themselves equally as powerful in certain instances and remain able to solve complex problems [19]. For the purpose of our study, we are faced with the challenge of evolving a difficult problem to solve, namely, datasets that have a high-order interaction with no main or two-way effects. Previously, others have used evolutionary algorithms to create problems that are hard for a specific heuristic to solve [20,21,22]. One novelty of the present study is our use of evolutionary algorithms to find problems without assuming a specific search algorithm or model.

3 Multi-objective Optimization and Pareto Optimality

Multiobjective problems are those for which practitioners wish to maximize or minimize two or more, often competing, characteristics of solutions. Evolutionary algorithms have been used to solve multi-objective optimization problems for more than twenty years [23,24,25]. These strategies are thought to be well suited to multi-objective problems because the population can carry out a search with solutions that succeed for different objectives [26]. The drawback of this approach is that assigning a single fitness score that encompasses every objective is difficult. Effectively using linear combinations of the objective scores for each objective requires knowledge about the problem and the fitness landscape which is unlikely to be available before a thorough analysis is performed. It is possible, however, to optimize many objectives without *a priori* knowledge by considering non-dominated (i.e. Pareto optimal) solutions as highly fit individuals. A non-dominated solution is one for which there is no solution that is better in all objectives. Here we use an approach focused on Pareto optimal solutions similar to one described by Goldberg [25]. In our approach, we use all Pareto optimal solutions as parents for the next generation, which would be equivalent to using only rank 1 individuals from Goldberg's approach. With this strategy we can explore the Pareto front of solutions which optimize each of our many objectives. We can then provide a number of model free datasets which are optimal with respect to our distinct objectives from a single run of the algorithm.

4 Multifactor Dimensionality Reduction (MDR)

Multifactor Dimensionality Reduction (MDR) is a widely used and a powerful model free method to detect and model gene-gene interactions associated with disease [27,28]. At the core of the MDR approach is an attribute construction algorithm that creates a new attribute by pooling genotypes from multiple SNPs. Constructive induction using the MDR kernel is accomplished in the following

way. Given a threshold T, a multilocus genotype combination is considered high-risk if the ratio of cases (subjects with disease) to controls (healthy subjects) exceeds or equals T, otherwise it is considered low-risk. Genotype combinations considered to be high-risk are labeled G1 while those considered low-risk are labeled G0. This process constructs a new one-dimensional attribute with levels G0 and G1. Here we use MDR to evaluate both high order and low order interactions. Our goal is to minimize the accuracy of this new variable for low-order combinations and maximize the accuracy of this variable for high-order combinations. Because MDR assumes no specific model, our dataset generation method is also model free.

5 A Model Free Dataset Generation Method

The first step in our process of generating datasets is to create random datasets. For each, we initialize a specified number of people (our sample) and provide them with random genotypes at three, four, or five SNPs. This number of SNPs can be arbitrarily assigned and is the order of the predictive interaction we wish to generate. We randomly assign these a case-control status. Current genotyping platforms are targeted towards measuring bi-allelic SNPs, i.e. those with two alleles. These SNPs can exist in one of three states. Here we indicate the states as 0, 1, and 2.

Because we wish to generate datasets with high concept difficulty, one of our goals is to minimize simple genetic effects. To do this, we use MDR to evaluate the best possible low-order predictors. We then, under a Pareto strategy, select those datasets which minimize the low-order predictiveness. In the case of our work here we focus on minimizing the predictiveness of all single SNPs and all two SNP combinations. Our next goal is to maximize the predictiveness of higher order combinations. To do this we evaluate the predictiveness of all the SNPs for every individual using MDR. Because we are maximizing the predictiveness of all the SNPs in the data, we only need to test a single attribute combination (i.e. that of all SNPs) and this task is computationally simple. Using MDR in this way gives us an accuracy, which we then use a Pareto strategy to maximize. Thus in this specific case our evolution strategy exploits Pareto optimization to minimize the single locus and two-locus predictiveness while maximizing the three, four, or five locus predictiveness. Specifically we use all Pareto optimal solutions as parents for the next generation. To generate offspring these parents are duplicated and, at each SNP for each individual, the value can be changed according to the mutation rate.

For SNPs not under selective pressure in humans, these states exhibit what is called Hardy-Weinberg equilibrium (HWE). Hartl and Clark provide an excellent overview of the Hardy Weinberg principle [29]. Because most SNPs are not under selection, deviations from HWE have historically been used as a marker of genotyping error [30]. The implicit assumption is that a SNP which is not in HWE is more likely to be a genotyping error than a SNP under selection. Examinations of early genetic association studies suggested that these concerns

may be well founded [31]. As genotyping methods improve and genotyping error is reduced, it becomes more likely that these SNPs are under selective pressure and less likely that deviations from HWE are due to genotyping error, and thus it becomes less likely that geneticists will filter SNPs which deviate from HWE. Indeed new methods have been developed which use the principles of Hardy-Weinberg equilibrium to detect an association between a genotype and disease [32]. Because the field is currently in transition we provide two sets of datasets, one set where we optimize for non-significant HWE genotype frequencies and one where we do not. In both cases we have initialized the frequencies of the genotype states as under HWE but selection can alter these frequencies.

By using a test of HWE as an additional Pareto criterion we can generate datasets containing SNPs that would not be filtered by currently used quality control measures. In this way we develop datasets where there is a model free but complex relationship between genotype and disease. With a wide array of datasets we can then test the ability of novel methods to detect and characterize complex epistatic relationships without making assumptions about the underlying genetic model. Because the result of each run is a set of Pareto optimal solutions, users can pick solutions with a wide array of difficulties to use while evaluating novel methods. For the set of results where we attempted to preserve Hardy-Weinberg equilibrium we actually minimize disequilibrium. Specifically we minimize the chi-square statistic which measures deviation from HWE. Because this expands our pareto front from three dimensions to four it dramatically increases the size of the pareto front. To insure that each parent has an opportunity to generate a reasonable number of offspring, we limit the number of parents taken to the next generation to one hundred when we are also optimizing for Hardy-Weinberg equilibrium. When there are more than one hundred individuals on the front, we choose the individuals in the "elbow" of the pareto front (i.e. non-extreme individuals). This tie-breaker keeps individuals which are good in regards to more than one dimension at the cost of those which excel in a single dimension.

At the conclusion of each run we have a front of pareto optimal datasets. Because comparing entire pareto fronts is difficult we wish to provide, in addition to the front, a single member of the pareto-optimal group which can represent the run. We have done this by picking the individual dataset with the smallest euclidean distance from the best values obtained for each measure.

6 Experimental Design and Analysis

Our first task was to determine a useful mutation rate. Because this system is driven by mutation, we need to employ an effective mutation rate to evolve good solutions. We examined mutation rate by evaluating rates of 0.05, 0.04, 0.03, 0.02, 0.01, 0.008, 0.006, 0.004 and 0.002 for sample sizes of 500 and 1000 using a fixed number of generations (750) and a fixed population size (1000). We then used the four-way testing accuracy to evaluate how far the evolution had progressed. We used these results to pick an appropriate mutation rate for the number of generations and population size we use here.

Our second task was to evaluate whether our Pareto evolution strategy outperformed a random search. We generated two million random datasets and compared the resulting Pareto front to the pareto fronts generated at the end of evolution in our system. We tested the significance of the differences observed for the accuracies from the fronts from evolved runs and those from the randomly generated runs. We statistically tested these differences with Hotelling's T-test and considered the differences significant when the p-value was less than or equal to 0.05. This means that we would consider results significant only one time out of twenty when there was no significant effect.

Our final task was to generate sets of Pareto optimal datasets each exhibiting three-way interactions, four-way interactions and five-way interactions. In each case we maximized three, four, and five-way MDR accuracies respectively while minimizing one and two way accuracies. We further generated datasets both under pressure to maintain Hardy-Weinberg equilibrium and irrespective of HWE. For each parameter setting (three, four, and five-way interactions with and without HWE) we generated 100 sets of datasets for a total of 600 sets of pareto optimal datasets. In total we have generated more than 50,000 datasets with a complex gene-disease relationship and made these datasets available to researchers as described in section 8.

7 Results

Our parameter sweep of mutation rates showed that, for this problem, a mutation rate of 0.004 led to the greatest success for datasets of five hundred people and 0.002 led to the greatest success for datasets of one thousand people. Large scale parameter sweeping with the sample sizes that we wished to generate was infeasible, but because the optimal mutation rate was related to the sample size we estimated that for the situation where we wish to evolve datasets containing 3000 individuals with a complex gene-disease relationship, a mutation rate of 0.001 would work, although because of the indirectness required to perform the parameter sweep it is not necessarily optimal.

We compared the results from our evolution strategy to a random search. The results are presented in Figure 1 and Table 1. Figure 1 shows the pareto front generated during a single evolved run and the pareto front generated by a random search over the same number of datasets. It is clear that solutions from the pareto front from the evolved run are generally much better than the randomly generated datasets. As Table 1 shows, the evolution strategy consistently outperforms the random search. Furthermore, as Table 1 shows, we were able to consistently generate datasets with a complex gene-disease relationship that lack low order predictors in a model free manner. In each case the differences between the pareto front from two million random datasets and that obtained at the end of our evolution strategy was highly significantly different ($p < 0.001$) indicating that these differences are not likely to be due to chance.

Figure 2 provides some insight into the difficulty of the problem. Minimizing the one-way accuracies and two way accuracies happens relatively quickly and

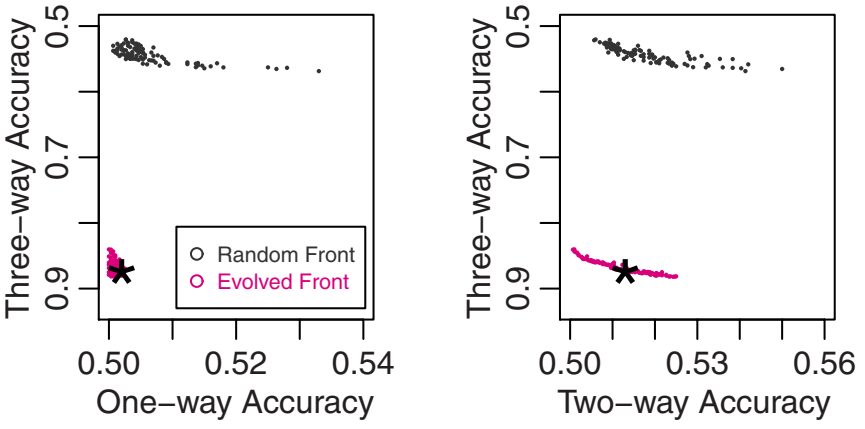

Fig. 1. This figure shows two dimensional projections of the three dimensional pareto front obtained at the end of one run of the evolution strategy and the pareto front obtained by randomly generating two million datasets. The three-way accuracy, which is maximized, is plotted against the one and two-way accuracies, which are minimized. The pareto front from the evolved run is clearly better than the pareto front from random initialization of two million datasets. The star shows the single dataset from the front chosen as the result of the run, which is used to compare across all runs in Table 1.

Table 1. A summary of the accuracies obtained for the evolution strategies and random search. HWE indicates whether or not the datasets were selected based on their conformance to Hardy-Weinberg equilibrium. The one-way and two-way accuracies are always minimized and the n-way accuracy is maximized. The accuracies are presented as mean and (standard deviation).

	Parameters			Results		
n-way	Gen.	Pop.	HWE	One-way (sd)	Two-way (sd)	n-way (sd)
Three-way	2000	1000	No	0.502 (0.001)	0.511 (0.007)	0.886 (0.023)
	2000	1000	Yes	0.504 (0.002)	0.509 (0.003)	0.680 (0.024)
	1	2000000	No	0.506 (0.006)	0.518 (0.009)	0.543 (0.012)
Four-way	2000	1000	No	0.502 (0.001)	0.510 (0.003)	0.897 (0.018)
	2000	1000	Yes	0.507 (0.003)	0.513 (0.003)	0.673 (0.009)
	1	2000000	No	0.507 (0.004)	0.519 (0.006)	0.571 (0.011)
Five-way	2000	1000	No	0.502 (0.001)	0.510 (0.002)	0.895 (0.009)
	2000	1000	Yes	0.511 (0.003)	0.518 (0.003)	0.693 (0.008)
	1	2000000	No	0.507 (0.004)	0.520 (0.005)	0.612 (0.011)

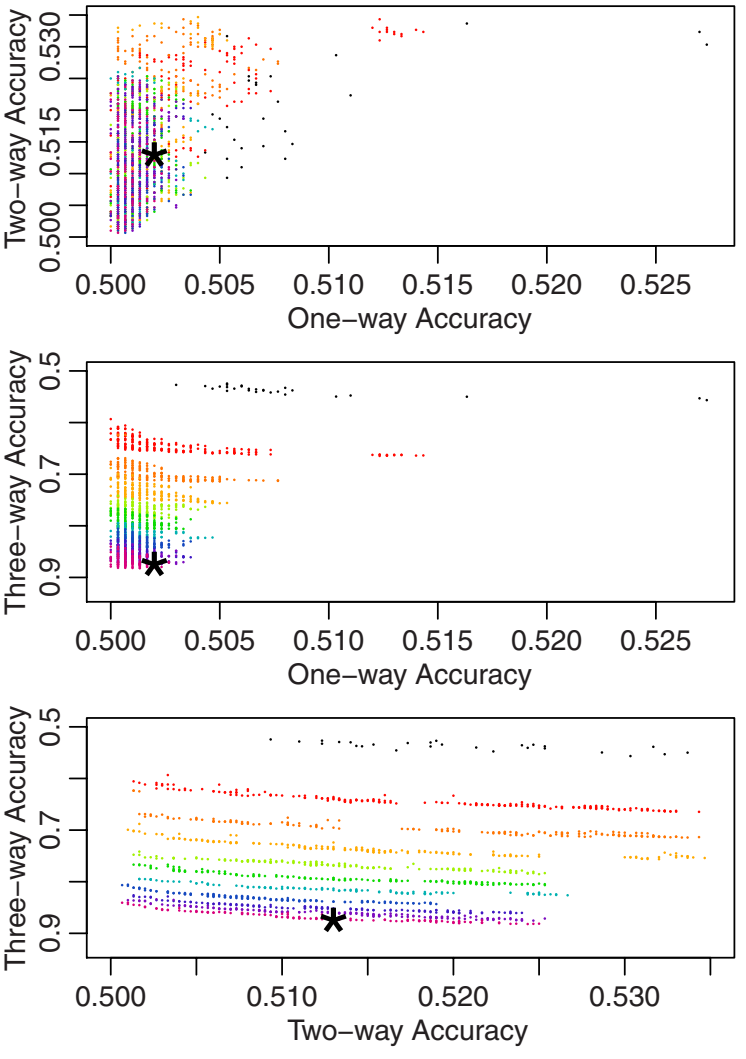

Fig. 2. The pareto front from a single run improves as the run proceeds. The pareto fronts are shown for each pairwise combination of objectives in each box at every 200 generations as shown by color. The star indicates the final solution used to evaluate the run.

within the first few hundred generations. Maximizing the higher order accuracies continues throughout the entire run and progress is still being made and the two-thousandth generation. The star indicates the dataset that was chosen from the front to represent the run. It is this one dataset that is taken from each run according to the euclidean distance strategy discussed in section 5 that is used to calculate the summary statistics in Table 1.

8 Dataset Availability

All pareto optimal datasets generated during these experiments are available from the website `http://discovery.dartmouth.edu/model_free_data/`. We provide a number of means of obtaining datasets. First, we provide archive files of "best of runs" obtained with and without Hardy-Weinberg equilibrium constraints. These representative datasets are obtained by choosing the dataset with the smallest euclidean distance between its own values and the optimum values obtained by all datasets on the pareto front as described in section 5. In addition to these representative datasets, for each run we provide all datasets that, at the end of the run, make up the complete Pareto front. To assist with the use of these datasets we further provide an information file for each run containing the characteristics of every dataset in the pareto-optimal front. From these files it is possible for investigators to develop suites of datasets that display certain characteristics (e.g. one and two way accuracies less than 52% and four-way accuracies of approximately 70%). Using datasets generated from our provided results, investigators can test novel methods across data exhibiting gene-disease relationships unconstrained by specific genetic models. The SNPs we provide can be combined with other noisy SNPs to represent a three, four, or five-way genetic interaction in a sea of noisy SNPs.

9 Discussion and Conclusions

Evolutionary computing has previously been used to generate epistatic (i.e. interaction based) models of a gene-disease relationship for both two-way [33] and higher order [34] interactions. Here we generate epistatic datasets in a model free manner. We also describe a novel evolution strategy for creating model free datasets and use this strategy to create 56,600 datasets with complex gene-disease relationships which we make publicly available. These datasets provide test beds for novel genetic analysis methods. By providing human genetics datasets with complex interactions that do not assume a model we hope to bypass the chicken and egg problem that has previously confronted the field. Methods tested on datasets generated in this manner may better generalize to true genetic association datasets.

Future work should focus on evaluating potential building blocks, so that crossover can improve the efficiency of the search. Potential building blocks include subsets of the individuals in each dataset, subsets of the SNPs in each dataset, or subsets of both individuals and SNPs in each dataset. It is not intuitive which approach is most useful, so these options should be fully explored. It may be that customized crossover operators are required to obtain useful genetic mixing for this problem.

Future work should also focus on making these datasets and others generated in this manner widely available. By dividing the resulting datasets into standardized testing and training datasets, it would be feasible to compare algorithms in a straightforward and objective manner. This comparison of algorithms would

provide a great deal of information about these methods to human geneticists attempting to understand the basis of common human disease. By providing open and publicly available datasets which do not assume a model but which contain a complex relationship between individual and disease, we hope to improve our understanding of commonly used methods in human genetics. We also hope to provide a framework for objectively testing future methods. Only when we can effectively judge methods across a comprehensive test suite of datasets can we develop methods likely to discover the underlying basis of common human diseases.

Acknowledgement

This work was supported by NIH grants LM009012, AI59694, HD047447, and ES007373. The authors would like to thank Peter Schmitt for his excellent technical assistance.

References

1. Chanock, S.J., Manolio, T., Boehnke, M., Boerwinkle, E., Hunter, D.J., Thomas, G., Hirschhorn, J.N., Abecasis, G., Altshuler, D., Bailey-Wilson, J.E., Brooks, L.D., Cardon, L.R., Daly, M., Donnelly, P., Fraumeni, J.F., Freimer, N.B., Gerhard, D.S., Gunter, C., Guttmacher, A.E., Guyer, M.S., Harris, E.L., Hoh, J., Hoover, R., Kong, C.A., Merikangas, K.R., Morton, C.C., Palmer, L.J., Phimister, E.G., Rice, J.P., Roberts, J., Rotimi, C., Tucker, M.A., Vogan, K.J., Wacholder, S., Wijsman, E.M., Winn, D.M., Collins, F.S.: Replicating genotype-phenotype associations. Nature 447(7145), 655–660 (2007)
2. McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P.A., Hirschhorn, J.N.: Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat. Rev. Genet. 9(5), 356–369 (2008)
3. Hirschhorn, J.N., Lohmueller, K., Byrne, E., Hirschhorn, K.: A comprehensive review of genetic association studies. Genet. Med. 4, 45–61 (2002)
4. Shriner, D., Vaughan, L.K., Padilla, M.A., Tiwari, H.K.: Problems with Genome-Wide association studies. Science 316(5833), 1840–1841 (2007)
5. Williams, S.M., Canter, J.A., Crawford, D.C., Moore, J.H., Ritchie, M.D., Haines, J.L.: Problems with Genome-Wide association studies. Science 316(5833), 1841–1842 (2007)
6. Jakobsdottir, J., Gorin, M.B., Conley, Y.P., Ferrell, R.E., Weeks, D.E.: Interpretation of genetic association studies: Markers with replicated highly significant odds ratios may be poor classifiers. PLoS Genetics 5(2), e1000337 (2009)
7. Templeton, A.: Epistasis and complex traits. In: Epistasis and the Evolutionary Process, pp. 41–57 (2000)
8. Moore, J.H.: The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Human Heredity 56, 73–82 (2003)
9. Moore, J.H., Williams, S.M.: Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. BioEssays 27(6), 637–646 (2005)

10. Greene, C.S., Penrod, N.M., Williams, S.M., Moore, J.H.: Failure to replicate a genetic association may provide important clues about genetic architecture. PLoS ONE 4(6), e5639 (2009)
11. Tyler, A.L., Asselbergs, F.W., Williams, S.M., Moore, J.H.: Shadows of complexity: what biological networks reveal about epistasis and pleiotropy. BioEssays 31(2), 220–227 (2009)
12. Shao, H., Burrage, L.C., Sinasac, D.S., Hill, A.E., Ernest, S.R., O'Brien, W., Courtland, H., Jepsen, K.J., Kirby, A., Kulbokas, E.J., Daly, M.J., Broman, K.W., Lander, E.S., Nadeau, J.H.: Genetic architecture of complex traits: Large phenotypic effects and pervasive epistasis. Proceedings of the National Academy of Sciences 105(50), 19910–19914 (2008)
13. Freitas, A.A.: Understanding the crucial role of attribute interaction in data mining. Artif. Intell. Rev. 16(3), 177–199 (2001)
14. Moore, J.H., Ritchie, M.D.: The challenges of Whole-Genome approaches to common diseases. JAMA 291(13), 1642–1643 (2004)
15. Velez, D.R., White, B.C., Motsinger, A.A., Bush, W.S., Ritchie, M.D., Williams, S.M., Moore, J.H.: A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. Genetic Epidemiology 31(4), 306–315 (2007)
16. Hoffmeister, F., Bäck, T.: Genetic algorithms and evolution strategies - similarities and differences. In: Schwefel, H.-P., Männer, R. (eds.) PPSN 1990. LNCS, vol. 496, pp. 455–469. Springer, Heidelberg (1991)
17. Bäck, T., Hoffmeister, F., Schwefel, H.: A survey of evolution strategies. In: Proceedings of the Fourth International Conference on Genetic Algorithms, pp. 2–9 (1991)
18. Goldberg, D.E.: The Design of Innovation: Lessons from and for Competent Genetic Algorithms. Kluwer Academic Publishers, Norwell (2002)
19. Greenwood, G., Shin, J.: On the evolutionary search for solutions to the protein folding problem. In: Fogel, G., Corne, D. (eds.) Evolutionary Computation in Bioinformatics, pp. 115–136. Elsevier Science, Amsterdam (2003)
20. van Hemert, J.I.: Property analysis of symmetric travelling salesman problem instances acquired through evolution. In: Raidl, G.R., Gottlieb, J. (eds.) EvoCOP 2005. LNCS, vol. 3448, pp. 122–131. Springer, Heidelberg (2005)
21. van Hemert, J.I.: Evolving combinatorial problem instances that are difficult to solve. Evolutionary Computation 14(4), 433–462 (2006)
22. Julstrom, B.A.: Evolving heuristically difficult instances of combinatorial problems. In: GECCO 2009: Proceedings of the 11th Annual conference on Genetic and evolutionary computation, pp. 279–286. ACM, New York (2009)
23. Schaffer, J.D.: Multiple objective optimization with vector evaluated genetic algorithms. In: Proceedings of the 1st International Conference on Genetic Algorithms, pp. 93–100. L. Erlbaum Associates Inc., Hillsdale (1985)
24. Richardson, J.T., Palmer, M.R., Liepins, G.E., Hilliard, M.: Some guidelines for genetic algorithms with penalty functions. In: Proceedings of the third international conference on Genetic algorithms, pp. 191–197. Morgan Kaufmann Publishers Inc., San Francisco (1989)
25. Goldberg, D.: Genetic algorithms in search, optimization and machine learning. Addison-Wesley Longman Publishing Co., Inc., Boston (1989)
26. Fonseca, C.M., Fleming, P.J.: An overview of evolutionary algorithms in multiobjective optimization. Evolutionary Computation 3, 1–16 (1995)

27. Ritchie, M.D., Hahn, L.W., Roodi, N., Bailey, L.R., Dupont, W.D., Parl, F.F., Moore, J.H.: Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am. J. Hum. Genet. 69(1), 138–147 (2001)
28. Moore, J.H., Gilbert, J.C., Tsai, C.T., Chiang, F.T., Holden, T., Barney, N., White, B.C.: A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. Journal of Theoretical Biology 241(2), 252–261 (2006)
29. Hartl, D.L., Clark, A.G.: Principles of Population Genetics, 3rd edn. Sinauer Associates, Sunderland (1997)
30. Hosking, L., Lumsden, S., Lewis, K., Yeo, A., McCarthy, L., Bansal, A., Riley, J., Purvis, I., Xu, C.: Detection of genotyping errors by Hardy-Weinberg equilibrium testing. Eur. J. Hum. Genet. 12(5), 395–399 (2004)
31. Xu, J., Turner, A., Little, J., Bleecker, E., Meyers, D.: Positive results in association studies are associated with departure from Hardy-Weinberg equilibrium: hint for genotyping error? Human Genetics 111(6), 573–574 (2002)
32. Ryckman, K.K., Jiang, L., Li, C., Bartlett, J., Haines, J.L., Williams, S.M.: A prevalence-based association test for case-control studies. Genetic Epidemiology 32(7), 600–605 (2008)
33. Moore, J.H., Hahn, L.W., Ritchie, M.D., Thornton, T.A., White, B.C.: Application of genetic algorithms to the discovery of complex models for simulation studies in human genetics. In: Proceedings of the Genetic and Evolutionary Computation Conference, pp. 1150–1155. Morgan Kaufmann Publishers Inc., San Francisco (2002)
34. Moore, J.H., Hahn, L.W., Ritchie, M.D., Thornton, T.A., White, B.C.: Routine discovery of complex genetic models using genetic algorithms. Applied Soft Computing 4(1), 79–86 (2004)