# Optimal Use of Biological Expert Knowledge from Literature Mining in Ant Colony Optimization for Analysis of Epistasis in Human Disease

Arvis Sulovari, Jeff Kiralis, and Jason H. Moore

Dartmouth-Hitchcock Medical Center, Lebanon
New Hampshire, 03756, United States
arvissulovari@gmail.com, {Jason.H.Moore,Jeff.Kiralis}@dartmouth.edu
http://www.epistasis.org

**Abstract.** The fast measurement of millions of sequence variations across the genome is possible with the current technology. As a result, a difficult challenge arise in bioinformatics: the identification of combinations of interacting DNA sequence variations predictive of common disease [1]. The Multifactor Dimensionality Reduction (MDR) method is capable of analysing such interactions but an exhaustive MDR search would require exponential time. Thus, we use the Ant Colony Optimization (ACO) as a stochastic wrapper. It has been shown by Greene et al. that this approach, if expert knowledge is incorporated, is effective for analysing large amounts of genetic variation[2]. In the ACO method integrated in the MDR package, a linear and an exponential probability distribution function can be used to weigh the expert knowledge. We generate our biological expert knowledge from a network of gene-gene interactions produced by a literature mining platform, Pathway Studio. We show that the linear distribution function of expert knowledge is the most appropriate to weigh our scores when expert knowledge from literature mining is used. We find that ACO parameters significantly affect the power of the method and we suggest values for these parameters that can be used to optimize MDR in Genome Wide Association Studies that use biological expert knowledge.

## 1 Introduction

Human geneticists are now able to measure millions of DNA sequence variations across large patient sample datasets. These large datasets present a challenge in the field of informatics: which variations can be used to predict susceptibility to common human disease such as cancer? What makes this challenge even more difficult is the fact that susceptibility to a given disease cannot always be determined by the action of a single gene, but rather the action of multiple interacting genes. Moore argues that non additive interactions, known as epistasis, are likely to be ubiquitous in common human disease [2]. Moore's argument relies on four

L. Vanneschi, W.S. Bush, and M. Giacobini (Eds.): EvoBIO 2013, LNCS 7833, pp. 129–140, 2013.

important concepts: the notion of epistasis is grounded in almost one century of scientific literature, molecular interactions between proteins are ubiquitous in biological systems, a single locus model is insufficient for explaining the etiology of common human diseases, and when scientists have tried to find epistasis using powerful computational and biostatistical methods, they have often been able to find examples of it. If we want to find predictors of common human disease, we need to employ methods which take into consideration the complexity of biological systems.

Data from biological systems is noisy due to the inherent complexity of these systems. The noise is primarily due to the fact that disease states of subjects with the same values for the relevant attributes could be different. Moreover, the fitness landscape is rugged because the models that contain less than all of the relevant attributes may perform worse than the surrounding noise [1].

The International HapMap Consortium suggests that approximately $3 \cdot 10^6$ carefully selected SNPs (i.e. single nucleotide polymorphisms) may be necessary and sufficient to capture all variation among the human population [3]. If this were true, we would expect $\binom{3 \cdot 10^6}{2} = 4.5 \cdot 10^{12}$ potential epistatic pairwise interactions. Biological systems provide inspiration for much more efficient machine learning algorithms.

Greene et al. have shown that the ACO method can be used effectively for human genetics problems when expert knowledge is used [1]. We used biological expert knowledge extracted from literature mining and rigorously examined the two different weighing functions of expert knowledge within the ant colony system in MDR to suggest good parameters for later use in Genome Wide Association Studies (GWAS). We believe that the approach of using knowledge from literature mining to facilitate MDR's quest in finding epistatic models underlying common human disease has not been explored before. Most importantly, this method has potential to provide more biologically relevant findings with regard to epistasis than previous MDR approaches.

## 2 Literature Mining Using Pathway Studio

Pathway Studio is a software application developed for navigation and analysis of biological pathways by Ariadne Genomics [4]. This software comes with a database of more than 100,000 interaction types, regulation and modification events between proteins, cell processes and small molecules. The database has been compiled by MedScan, a text-mining tool, to the whole PubMed. MedScan pre-processes text input from the user to extract the relevant sentences which are then subjected to Natural Language Processing (NLP). The pre-processing step uses a manually curated biological dictionary of synonyms. The NLP kernel deduces the syntactic structure of the sentences and establishes logical relationship between concepts. Finally, the results are matched against the functional ontology to produce biologically interpretable data [4].

Here we queried all the genes corresponding to the SNPs in our dataset (Section 5). The output from Pathway Studio provided us with information on the number of interactions for each gene. The number of connections for each gene was averaged across all the present types of interactions to give an expert knowledge score. This method represents one way of processing the biological knowledge from Pathway Studio into expert knowledge recognized by the ACO method in MDR. Our processing method considered SNPs which belonged to genes with many interactions as more important than those with less interactions. The ant system integrated into MDR used Pathway Studio as its source of expert knowledge.

## 3 Multifactor Dimensionality Reduction Platform

Greene et al. developed an ACO framework to be available in version 2.0 and later of the Multifactor Dimensionality Reduction (MDR) software package [1]. This package provides a user friendly cross-platform Java GUI appropriate for genome-wide genetic analysis. In short, MDR groups multilocus genotypes in high-risk and low-risk groups, reducing the genotype predictors' dimensionality from $n$ to *1*. The new one dimensional multilocus-genotype variable is evaluated for its ability to classify and predict case-control status through cross-validation. The MDR method has been developed as a non-parametric and model-free genetic data mining strategy for identifying combinations of SNPs that are predictive of discrete clinical endpoint [7]. The MDR method has been successfully applied to detect gene-gene interactions in a variety of human diseases: breast cancer [7], type 2 diabetes [9], rheumatoid arthritis [8], and coronary artery disease [10].The MDR method is described in detail by Moore, et al. [5].

## 4 The MDR Ant Colony Optimization (ACO) Approach

The idea of using ants as an inspiration for machine learning algorithms is not new. Dorigo showed in 1991 that ants could be used as a search strategy by providing positive feedback [17]. In an ant system ants explore the landscape of possible solutions by leaving a trace of pheromones on each solution they find, depending on the quality of that discovery. Over time the pheromones evaporate and their signal weakens. The quantity of pheromone left on each discovery made by an ant determines the likelihood that the same region will be explored in the future by other ants. Dorigo and Stützle discuss how incorporation of a priori information can be used to derive heuristic information that biases the probabilistic decision taken by the ants [18]. ACO is one of the techniques of swarm intelligence, a relatively new domain within AI research, that has proven to be competitive with traditional techniques of data mining [19]. Moore et al. discovered that incorporation of a priori knowledge into machine learning algorithms is crucial if these algorithms are to succeed at genome-wide genetic analysis [20].

In the ant system integrated within the MDR package, the goal is to select the SNPs (i.e. attributes) which effectively determine an individual's risk of disease. We use Pathway Studio scores (Section 2) as biological expert knowledge.

The ACO method allows the user to select an exponential or linear function for weighing the scores. Below we discuss each of the resulting probability distributions. We assume that there are $n$ attributes $A_1, \ldots, A_n$ with $A_i$ having expert knowledge score $S_i$. We label the attributes so that $S_1 \leq S_2 \leq \ldots \leq S_n$.

### 4.1 Exponential Weighing

With exponential weighing, the probability that attribute $A_i$ is selected is given by the exponential function [1]

$$P(A_i) = \frac{1}{\sum\limits_{k=1}^{N} \theta^{-S_k}} \theta^{-S_i}, \tag{1}$$

where $\theta$ is the user-adjustable parameter, satisfying $0 < \theta \leq 1$, and here the expert knowledge scores $S_i$ are normalized so that they lie between 0 and 2.

As Greene et al. noted, if $\theta$ is near 1, attributes with a high expert knowledge score are only slightly more likely to be chosen than those with a lower score. Otherwise, for instance, if $\theta$ is 1/3, the attributes with a high score are much more likely to be chosen than those with lower scores.

### 4.2 Linear Weighing

For linear weighing the probability that attribute $S_i$ is selected is given by

$$P(A_i) = mS_i + b \tag{2}$$

for some constants $m$ and $b$. We require that $m \geq 0$ so that $P(A_1) \leq P(A_2) \leq \ldots \leq P(A_n)$. This assures that attributes with larger expert knowledge scores are more apt to be selected.

The constraints $\sum_{i=1}^{n} P(A_i) = 1$ and $P(A_1) \geq 0$, and the requirement $m \geq 0$ are satisfied only when:

$$m \in \left[0, \frac{1}{\sum_{i=1}^{n}(S_i - S_1)}\right] \quad \text{and} \quad P(A_N) \in \left[\frac{1}{n}, \frac{S_N - S_1}{\sum_{i=1}^{n}(S_i - S_1)}\right].$$

Here $P(A_n)$ is the probability of selecting the attribute with the highest expert knowledge score.

The parameter $M_p \in [0, 1]$ adjusts $m$ and $P(A_n)$ so that:

$$m = \frac{M_p}{\sum_{i=1}^{N}(S_i - S_1)} \quad \text{and} \quad P(A_n) = \frac{1}{N} + M_p \left( \frac{S_i - S_1}{\sum_{i=1}^{N}(S_i - S_1)} - \frac{1}{N} \right).$$

Both of these functions have the following pheromone update procedure:

$$\delta\tau_{a,i} = \sum_{k=1}^{m} Q_{a,b} \cdot S_a^{\beta}$$

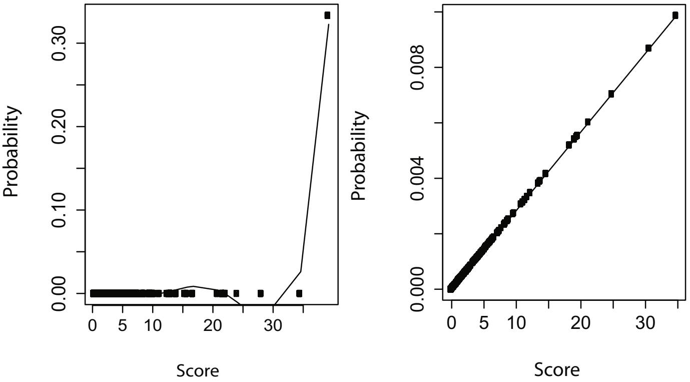

**Fig. 1.** Assignment of probabilities for each expert knowledge score being chosen according to the exponential distribution (left panel) and linear distribution (right panel). In the left panel, the probability of being chosen of any score below the maximum score is zero, which makes this function inappropriate for our expert knowledge. On the other hand, the linear probability distribution function assigns non-zero probabilities to many more attributes. The solid line on the left represents a polynomial best fit.

where $\delta\tau_{a,i}$ is the change in pheromone strength between updates. $Q_{a,b}$ is the MDR accuracy for a model containing both attributes $a$ and $b$, while $S_a$ is the biological expert knowledge from Pathway Studio for attribute $a$ and $\beta$ is a weighing exponent for the expert knowledge, $E_a$.

Figure 1 shows the distribution of probabilities for each expert knowledge score according to both distribution functions.

## 5 Data Simulation and Analysis

A genotype study conducted by Andrews et al. produced a SNP dataset of 1421 SNPs in approximately 400 hypothesized cancer-related genes from the SNP500 database [11]. This dataset contains 893 controls and 617 subjects with bladder cancer. Here we replaced a random set of 100 SNP-SNP pairs from the original dataset along with the class values (i.e. case/control status) with two synthetic epistatic SNPs and their respective new class values.

The generation of the synthetic SNPs and their class values was done using the GAMETES algorithm [12] [13]. This algorithm generates SNP datasets of various population sizes, heritabilites and allele frequencies. For every run of the GAMETES, there were three different models, in each one of which we had one epistatic SNP pair.

In our case, we generated nine models of disease risk, each containing two relevant epistatic SNPs. These models spanned three heritabilities (0.05, 0.1, and 0.2). For each heritability GAMETES generated three models. All these models exhibit no main effects when the SNPs have a minor allele frequency of 0.4, which were the conditions we used for generating our data. This means that the effects in each dataset will be due to epistatic interactions and not main effects caused by a single SNP. We generated 100 datasets for each model.

In the modified datasets containing the synthetic SNPs, the noise was provided by the 1419 biological SNPs. Here we mapped Pathway Studio's expert knowledge scores for each gene, from a total of 397 genes, into expert knowledge scores for each respective SNP. As for the synthetic SNPs, we assigned them three different scores: upper 10%, upper 1%, and upper 0.1% cut-off values according to the overall ranking of the scores. We then provided these scores to the ant system which converted them into selection probabilities using the linear distribution function. (See Section 4.2)

We explored five major parameters of the ant system: maximum probability, $\beta$, retention factor, and number of ants and updates. Maximum probability (i.e. the slope of the linear probability function) was assigned values of 10%, 50% and 90%. Beta was assigned four different values: 0, 1, 2, and 4. Ants and updates were each assigned values of 100, 200, 400, and 800. The retention factor determines how much weight is given to information from the previous iterations relative to the most recent iteration. We considered retention factors of 0.1, 0.5 and 0.9. A total of 640,000 parameter combinations was explored. We considered a high number of total interactions between parameters in order to assure the discovery of the two epistatic SNPs. The sweep of all MDR parameters was done on a 1300-processor cluster at Dartmouth College. To determine statistical significance, we used logistic regression. Logistic regression allows for a rigorous examination of the effect that one or more continuous factors (i.e. parameters) have on the success rate, the number of runs that selected the two synthetic SNPs over the total runs. We used the R statistical programming language to run logistic regression [14]. We assessed all single and pairwise effects of

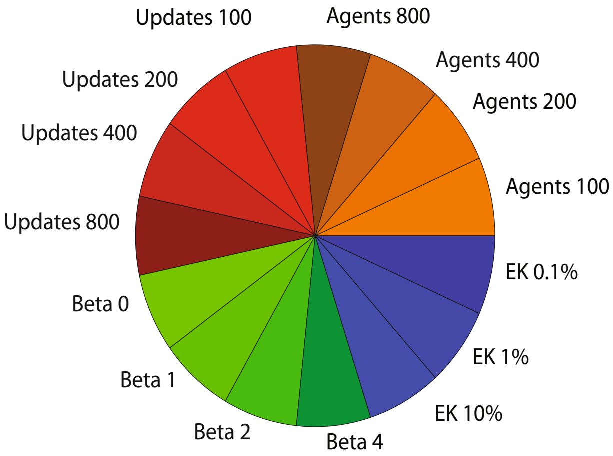

**Fig. 2.** Legend for the segment plots in Figure 3

all parameters on the success rate. Results of logistic regression were considered significant when $p \leq 0.05$.

## 6 Experimental Results

We found that a retention factor of 0.9 and a maximum probability of 90% are the best parameters for fine tuning the ACO method. These values support the findings by Greene et al. [1]. Hence, in the data presented below we fixed these two parameters and looked at combinations of the other ACO parameters: $\beta$, number of ants, number of updates, and the expert knowledge scores for the two

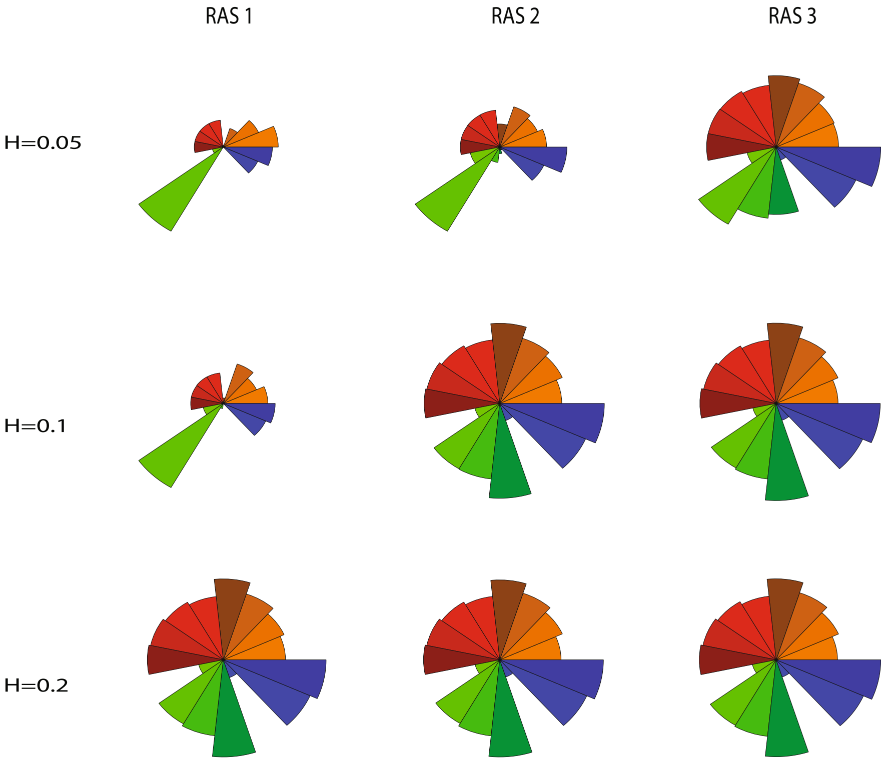

**Fig. 3.** Results of the simulations on all 9 models. Each plot summarizes the results from 19,200 sweep runs and the size of each sector in a plot represents the success rate of each respective parameter across all those runs. $Beta = 1$ is the most important parameter in the first four datasets along with the highest expert knowledge score. In the other 5 models, $Beta = 4$ yields the highest success along with the highest expert knowledge score and the highest number of ants. Each model was simulated by GAMETES under a Minor Allele Frequency of 0.4. Each row corresponds to a different heritability (H) and each column corresponds to a different quantile of the Relative Allele Signals (RAS).

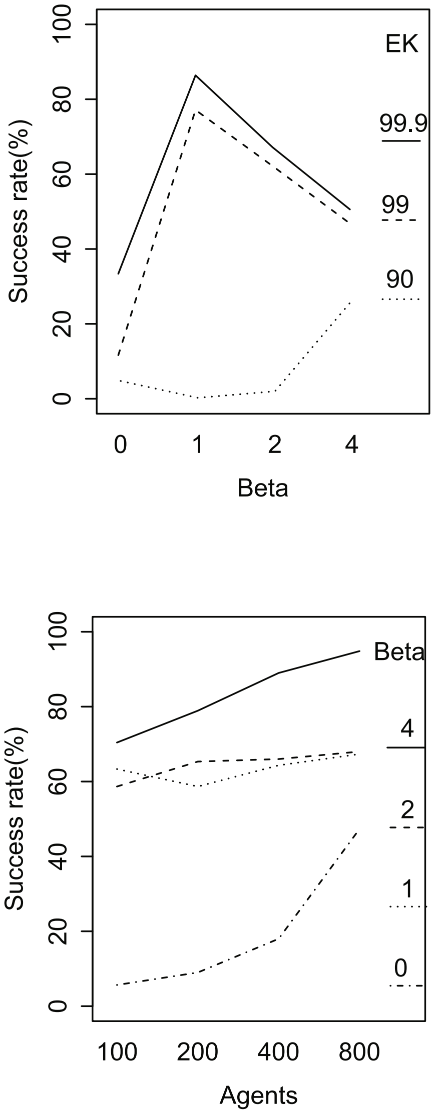

**Fig. 4.** Two of the pairwise parameter interaction plots that were found to be significant by logistic regression in models of heritability 0.05 (top panel) and 0.2 (bottom panel). The intersections or the non-parallel segments infer a pairwise interaction between the parameters which is also confirmed by logistic regression at a significance level of $p \leq 0.05$.

synthetic SNPs. There were a total of 19,200 runs for each model amounting to a 172,800 runs overall. The results of these runs per model are shown in Figures 2 and 3 [15]. Segment plots are used to visually summarize the results for each model. Although the exact contribution of each parameter to the overall success rate cannot be assessed, the relative contribution of each parameter within and between models can be easily determined. The highest expert knowledge score for the two epistatic SNPs (i.e. upper 0.1%) yielded the highest success rates across all datasets, which was to be expected. The number of agents had a slightly different behaviour. In the first two models of heritability of 0.05, the highest number of agents yielded the lowest success rate which seemed counter-intuitive at first. However, this behaviour was also to be expected since the signal of the synthetic SNPs was weaker than the biological noise in the first two models. Indeed, in the other seven models where the synthetic SNPs had a stronger signal, the highest number of ants yielded the highest success rate among the ant population sizes within models. Each of the four different ant update values had a near-uniform effect on the success rate. As for the expert knowledge weighing factor, $\beta$, we noticed an interesting behaviour. In all datasets with heritability of 0.05 and in the first two models with heritability of 0.1 for the synthetic SNPs, $\beta = 1$ had the highest effect on the success rate among the betas. This value of $\beta$ was the same as the one suggested by Greene et al. [1]. However, in the remaining 5 models, $\beta = 4$ had the highest effect on the success rate among the betas. To understand better this behaviour of the $\beta$ parameter, we ran MDR exhaustively on the first four models (data not shown). We found that our two synthetic SNPs were not the best two-way model chosen by MDR in the first four models, which were the same models where $\beta = 1$ had the highest success rate among the betas. The more exponential weight we put on our expert knowledge by increasing $\beta$, the lower the success rate became in each of the first four models where the synthetic SNPs seemed to be undetectable even by exhaustive MDR. This could be due to amplification of noise in those models when $\beta > 1$.

While segment plots can be very useful in visualizing vast amounts of data, they do not give us statistical details on the effect of single-parameter or pairwise interactions which is why we used logistic regression. Logistic regression showed that all single-parameter effects are significant at $p \leq 0.001$ across all three heritabilities with the exception of the number of ants and updates. These results agreed with the summary from the segment plots. Moreover, logistic regression showed that pairwise parameter interactions significant at $p \leq 0.05$ were $\beta$:expert knowledge in all models with heritability of 0.05, ants:expert knowledge in all models with heritability 0.1 and all possible pairwise parameter combinations in all models of heritability of 0.2 with the exception of ants:updates. To visualize the pairwise parameter interactions, we used interaction plots (Figure 4). Based on these results, we suggest parameter settings of $\beta = 1$, in the case of a weak epistatic signal and $\beta = 4$ otherwise, retention factor of 0.9 and maximum probability of 90%. The number of ants and updates would be best if set at the highest value given the computational constraints. Our recommended settings for the $\beta$ parameter are different from those of Greene et al. which is most likely due to the different source of expert knowledge and the fact that the noisy SNPs in our data are biological and not simulated.

## 7 Discussion

Our study highlights the importance of utilizing biological expert knowledge in guiding GWAS. Here we presented one method of integrating biological expert knowledge from Pathway Studio into the ACO algorithm within MDR. The interactions found by Pathway Studio in the literature have more biological relevance than those generated by statistical methods alone. Recent studies have emphasized the importance of using biologically relevant expert knowledge in computational methods attempting to detect epistasis in genome wide genetic analysis [6] [21]. Our approach can yield biologically relevant results as defined by the current literature.

We chose the linear function to weigh the expert knowledge scores extracted from Pathway Studio as it presented one important advantage over the exponential function: it assigned non-zero probabilities of being chosen to more attributes (i.e. SNPs) compared to the exponential function. The linear function guarantees us that MDR will explore a bigger space of the solutions' landscape and yet spend less time compared to an exhaustive run. The solutions considered in this landscape also have a high biological relevance due to the source of expert knowledge.

We observed several interactions between the ant system parameters. Both logistic regression and segment plots helped us understand and visualize the effect that each parameter as well as pairwise combinations of parameters had on the overall success rate of the ACO.

Alternatives to processing our expert knowledge from Pathway Studio have been considered. We could make the scores even more biologically relevant by calculating the expert knowledge scores for every pairwise interaction in our dataset instead of calculating them for every single SNP, in order to estimate the relevance of the interactions using mutual information scores. The latter approach would also require a modification of the current ACO method in MDR as it currently only accepts scores for individual SNPs. Another improvement can be done on the function used to weigh the expert knowledge scores. We chose the linear function because of its superior representation of scores over the exponential function. However, these two functions do not present the only two heuristics' probability functions that can be used. In fact, as Dorigo and Stützle discuss in their book, the ACO algorithm could have other additional features, such as the Model Based Search [18] which is yet to be explored.

Our understanding of common human disease would be enhanced if more methods which take into consideration biologically relevant knowledge, similar to the approach we have presented, can be developed to detect epistasis in GWAS. If the epistasis quest of computational methods, such as MDR, is facilitated and directed by biologically relevant knowledge, then our preventative, diagnostic and treatment options will improve and could lead to better health and lower incidence of common disease.

**Acknowledgements.** The work was funded by grants R01 LM010098, LM009012 and AI59694 from the U.S.A National Institute of Health.

## References

1. Greene, C.S., Gilmore, J.M., Kiralis, J., Andrews, P.C., Moore, J.H.: Optimal Use of Expert Knowledge in Ant Colony Optimization for the Analysis of Epistasis in Human Disease. In: Pizzuti, C., Ritchie, M.D., Giacobini, M. (eds.) EvoBIO 2009. LNCS, vol. 5483, pp. 92–103. Springer, Heidelberg (2009)
2. Moore, J.H.: The ubiquitous nature of epistasis in determining susceptibility to common human disease. Human Heredity 56, 73–82 (2003)
3. The International HapMap Consortium: A second Generation human haplotype of over 3.1 million SNPs. Nature 449, 851–861 (2007)
4. Nikitin, A., Egorov, S., Mazo, I.: Pathway Studio-the analysis and navigation of molecular networks. Bioinformatics Oxford Journals 19(16), 2155–2157 (2003)
5. 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)
6. Cordell, H.J.: Detecting gene-gene interactions that underlie human diseases. Nature Review Genetics 10, 392–404 (2009)
7. Ritchie, M.D., et al.: Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. American Journal of Human Genetics 69, 138–147 (2001)
8. Julia, A., et al.: Identification of a two-loci epistatic interaction associated with susceptibility to rheumatoid arthritis through reverse engineering and multifactor dimensionality reduction. Genomics 90, 6–13 (2007)
9. Cho, Y.M., et al.: Multifactor-dimensionality reduction shows a two-locus interaction associated with type 2 diabetes mellitus. Diabetologia 47, 549–554 (2004)
10. Tsai, C.T., et al.: Reninangiotensin system gene polymorphisms and coronary artery disease in a large angiographic cohort: detection of high order genegene interaction. Atherosclerosis 195, 172–180 (2007)
11. Andrew, A.S., et al.: Bladder Cancer SNP panel predicts susceptibility and survival. Human Genetics 125(5-6), 527–539 (2009)
12. Urbanowicz, R.J., Kiralis, J., Sinnot-Armstrong, N.A., Heberling, T., Fisher, J.M., Moore, J.H.: GAMETES: a fast, direct algorithm for generating pure, strict, epistatic models with random architectures. BioData Mining 5(16) (2012)
13. `http://sourceforge.net/projects/gametes/`
14. `http://stat.ethz.ch/R-manual/R-patched/library/stats/html/glm.html`
15. `http://stat.ethz.ch/R-manual/R-patched/library/graphics/html/stars.html`
16. Sokal, R.R., Rohlf, F.J.: Biometry: the principles and practice of statistics in biological research, 3rd edn. W.H. Freeman and Co., New York (1995)
17. Dorigo, M., Maniezzo, V., Colorni, A.: Positive Feedback as a search strategy. Dipartimento di Elettronica e Informatica, Politecnico di Milano, Technical Reports, 91–116 (1991)

18. Dorigo, M., Stützle, T.: Ant Colony Optimization (2004)
19. Martens, D., et al.: Editorial Survey: Swarm Intelligence for Data Mining. Machine Learning 82(1), 1–42 (2011)
20. Moore, J.H., White, B.C.: Genome-wide genetic analysis using genetic programming: The critical need for expert knowledge. In: Genetic Programming Theory and Practice IV. Springer (2007)
21. Pattin, K., Moore, J.H.: Exploiting the proteome to improve the genome-wide genetic analysis of epistasis in common human diseases. Human Genetics 124(1), 19–29 (2008)