

# Ant Colony Optimization for Genome-Wide Genetic Analysis

Casey S. Greene, Bill C. White, and Jason H. Moore

Dartmouth College, Lebanon, NH, USA

{Casey.S.Greene,Bill.C.White,Jason.H.Moore}@dartmouth.edu

**Abstract.** In human genetics it is now feasible to measure large numbers of DNA sequence variations across the human genome. Given current knowledge about biological networks and disease processes it seems likely that disease risk can best be modeled by interactions between biological components, which can be examined as interacting DNA sequence variations. The machine learning challenge is to effectively explore interactions in these datasets to identify combinations of variations which are predictive of common human diseases. Ant colony optimization (ACO) is a promising approach to this problem. The goal of this study is to examine the usefulness of ACO for problems in this domain and to develop a prototype of an expert knowledge guided probabilistic search wrapper. We show that an ACO approach is not successful in the absence of expert knowledge but is successful when expert knowledge is supplied through the pheromone updating rule.

## 1 Introduction

Researchers in the biological and biomedical sciences are now capable of generating enormous amounts of data. In human genetics it is now technically and economically feasible to measure more than one million DNA sequence variations from across the human genome. Here we focus on the single nucleotide polymorphism or SNP which is a single point in a DNA sequence that differs among people. It is anticipated that at least one SNP occurs approximately every 100 nucleotides across the  $3 \times 10^9$  nucleotide human genome. An important goal in human genetics is the determination of which of the millions of SNPs are useful for predicting who is at risk for common diseases. This “genome-wide” approach is expected to revolutionize the genetic analysis of common human disease. The charge for computer science and bioinformatics is the development of algorithms for the detection and characterization of SNPs which are predictive of human health and disease. Success in this endeavor will be difficult due to nonlinearity in the genotype-to-phenotype mapping relationship that is due, in part, to epistasis or nonadditive gene-gene interactions. The implication of epistasis from a data mining point of view is that SNPs need to be considered jointly in learning algorithms rather than individually. The challenge of modeling attribute interactions has been previously described [1]. Due to the combinatorial magnitude of this problem, intelligent analysis strategies are needed.

## 1.1 Concept Difficulty

Combining the difficulty of modeling nonlinear attribute interactions with the challenge of attribute selection yields for this domain what Goldberg [2] calls a needle-in-a-haystack problem. That is, there may be a particular combination of SNPs that together with the right nonlinear function are a significant predictor of disease susceptibility. Considered individually they may not look any different than thousands of other SNPs not involved in the disease process. Under these models, the learning algorithm is truly looking for a genetic needle in a genomic haystack. These epistatic interactions are thought to be widespread, perhaps ubiquitous, among risk factors for these common human diseases [3]. A recent report from the International HapMap Consortium [4] suggests that approximately 300,000 carefully selected SNPs may be necessary to capture all of the relevant variation across the Caucasian human genome. Assuming this is true (it is probably a lower bound), we would need to scan  $4.5 \times 10^{10}$  pairwise combinations of SNPs to find a genetic needle. The number of higher order combinations is astronomical.

## 1.2 Ant Colony Optimization

Ant colony optimization (ACO) is a positive feedback approach to search modeled on the behavior of ants [5]. Ant colony optimization is attractive for the area of human genetics because it is a straightforward population based approach to search which is easily parallelizable. Ant colony systems have previously been applied to the mining of biological data. Parpinelli et al. [6] demonstrate their AntMiner system as a rule discovery method on biological data. Here we begin to develop a probabilistic search wrapper which can be integrated into the publicly available Multifactor Dimensionality Reduction (MDR) software. But is ant colony optimization suitable for a problem like this? Without expert knowledge the answer would seem to be no. There is no reason to expect that an ACO or any other wrapper method will perform better than a random attribute selector because there are no building blocks for this problem when accuracy is used as a metric of quality. The accuracy of any given classifier looks no better than any other with just one of the two correct SNPs in the model. Indeed, we have observed this in the field of genetic programming [7,8]. Subsequent work has shown that by integrating expert knowledge into a genetic programming scheme, it is possible to develop a wrapper that is able to perform better than a random attribute selector [9,10]. Fortunately the ACO metaheuristic is amenable to the inclusion of heuristic information. Work here examines whether or not it is possible to integrate expert knowledge, our heuristic information, into an ACO framework to develop a wrapper which performs well in this domain.

## 2 The Proposed Ant Colony Optimization Algorithm

Ant colony optimization is a particularly appropriate framework for this problem because of its simplicity and the ease with which expert knowledge can be included. For our work with genetic programming we developed specialized fitness

functions [7], recombination [8] and mutation operators [10]. With ACO it is conceptually much simpler to include expert knowledge. For this ACO metaheuristic we have elected to include expert knowledge as an additional component of the pheromone update rule. When the accuracy of the classifier is identical, ants that choose SNPs with better expert knowledge will contribute more pheromone to those paths than SNPs with lower expert knowledge scores.

## 2.1 Implementation

This ACO is implemented in C++. For the purposes of this power analysis solutions consist of pairs of attributes. The solution kept as the result is the pair of attributes with the highest balanced accuracy according to MDR (detailed in section 3). MDR analysis is performed through version 0.2.5 of the libmdr open source C library available from [www.epistasis.org](http://www.epistasis.org)

## 2.2 Pheromone Updating with Expert Knowledge

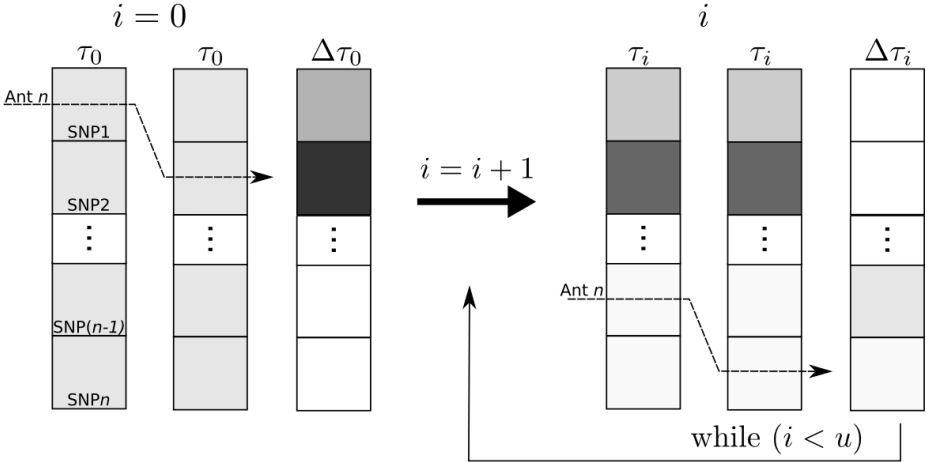
We have discovered in our work with genetic programming that expert knowledge is critical for machine learning algorithms to be successful with this problem [7]. Here we apply principles discovered in our genetic programming work to the ACO arena and structured our pheromone updating rule such that expert knowledge is provided within the pheromone update rule. The pheromone is updated for each SNP,  $a$ , according to the following function after the  $i$ th update:

$$\tau_{a,i+1} = \tau_{a,i}\rho + \Delta\tau_{a,i} \quad (1)$$

$\Delta\tau_{a,i}$  is the additional pheromone contributed by ants during this update cycle and  $\rho$  is the evaporation factor.  $\Delta\tau_{a,i}$  is obtained as a combination of the MDR accuracy and the expert knowledge information for each ant,  $k$ , of  $m$  total ants that contain attribute  $a$ :

$$\Delta\tau_{a,i} = \sum_{k=1}^m Q_{a,b} \alpha E_a^\beta \quad (2)$$

where  $Q$  is the MDR accuracy of a model containing that attribute  $a$  and the other attribute,  $b$ , chosen by ant  $k$ . In the case that  $a$  and  $b$  are the same SNP the MDR accuracy is set to zero to push the metaheuristic away from SNPs that have a strong main effect without an epistatic effect.  $E$  is the expert knowledge information for attribute  $a$ , and  $\alpha$  and  $\beta$  are coefficients that determine the relative weighting of  $Q$  and  $E$ . In this case Tuned ReliefF (TuRF) weights (see Section 4) are used as the expert knowledge [11]. This update rule is used through  $u$  total updates. This serves as an easy to understand pheromone updating rule which incorporates expert knowledge for ACO in the field of genetic analysis. This approach is similar to the use of heuristic information in other ACO approaches [5], except that here the heuristic alters the amount of pheromone deposited instead of modifying the likelihood of an ant selecting a path given pheromone information. This means that the initial search is very exploratory and that good SNPs by our heuristic information should be more heavily selected towards the end as pheromone information accumulates.



**Fig. 1.** In our ACO metaheuristic ants explore a pair of SNPs. From that pair of SNPs a  $\Delta\tau_i$  is calculated as shown in equation 2 for each SNP.  $\Delta\tau_i$  is a combination of the quality of the pair,  $Q$ , and the expert knowledge score for that SNP,  $E$ . Shading here is representative of the strength of pheromone,  $\tau$ , which begins evenly distributed and changes with each update, ( $i$ ).

### 2.3 Parameter Settings

We restrict the amount of the search space which can be explored to examine how well the algorithm performs given the ability to search a small number of the possible interactions. In each run 5000 total ants explore the search space. This means that at most approximately 1% of the total possible interactions (i.e. the search space) are examined. The power analysis (i.e. how often the correct answer is found) is performed with 250 ants per update for 20 updates, 500 ants per update for 10 updates and 1000 ants per update for 5 updates. The parameter  $\alpha$  is fixed at 1 and  $\beta$  is tested at 0, 1, and 2. When  $\beta$  equals 0, the expert knowledge weighting factor becomes 1 and does not affect the updating of the pheromone, thus it is possible to examine the impact of expert knowledge on the ant colony optimization approach in this domain. The evaporation parameter  $\rho$  is held constant at 0.5.

## 3 Multifactor Dimensionality Reduction (MDR) for Attribute Construction

Multifactor dimensionality reduction (MDR) was developed as a nonparametric and genetic model-free data mining strategy for identifying combination of SNPs that are predictive of a discrete clinical endpoint [12,13,14,15]. The MDR method has been successfully applied to detecting gene-gene interactions for a variety of common human diseases including adverse drug reactions [16]. At the heart 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  $G1$ . It is this new single variable that is returned by the MDR function as the quality,  $Q$ , for the ACO metaheuristic. Moore et al. [14] describe the MDR method in more detail. Open-source MDR software is freely available from [www.epistasis.org](http://www.epistasis.org)

## 4 Expert Knowledge from Tuned ReliefF (TuRF)

Our goal is to provide an external measure of attribute quality that can be used as expert knowledge for pheromone updating by the ACO metaheuristic. Here the external measure used is statistical, but it can just as easily be biological. There are many statistical and computational methods for determining the quality of attributes. Our goal is to use a method that is capable of identifying attributes that predict class primarily through dependencies or interactions with other attributes. Kira and Rendell [17] developed an algorithm called Relief that is capable of detecting attribute dependencies.

Relief estimates the quality of attributes through a nearest neighbor algorithm that selects neighbors (instances) from the same class and from the different class based on the vector of values across attributes. Weights ( $W$ ) or quality estimates for each attribute ( $a$ ) are estimated based on whether the nearest neighbor (nearest hit,  $H$ ) of a randomly selected instance ( $R$ ) from the same class and the nearest neighbor from the other class (nearest miss,  $M$ ) have the same or different values. This process of adjusting weights is repeated for  $m$  instances. The algorithm produces weights for each attribute ranging from -1 (worst) to +1 (best). Kononenko [18] improved upon Relief by choosing  $n$  nearest neighbors instead of just one. This new ReliefF algorithm has been shown to be more robust to noisy attributes and missing data [19] and is widely used in data mining applications [19].

We developed a modified ReliefF algorithm for the domain of human genetics called Tuned ReliefF (TuRF). We have previously shown that TuRF is significantly better than ReliefF in this domain [11]. The TuRF algorithm systematically removes attributes that have low quality estimates so that the ReliefF values of the remaining attributes can be re-estimated. We apply TuRF as described by Moore and White [11] to each dataset. Here TuRF scores compose the expert knowledge component of the ACO metaheuristic,  $E$ .

## 5 Fisher's Exact Test

Fisher's exact test is a significance test appropriate for categorical count data [20]. The resulting  $p$ -value denotes the likelihood that an association of the observed magnitude is likely by chance alone. For our use we arrange the results in a 2x2 contingency table:

**Table 1.** 2x2 contingency table for power analysis

	Success	Failure
Parameter Set 1 ( $PS1$ )	# Successful with $PS1$	# Unsuccessful with $PS1$
Parameter Set 2 ( $PS2$ )	# Successful with $PS2$	# Unsuccessful with $PS2$

With this contingency table we can detect whether the association between success (power) at different parameter settings ( $PS1$  and  $PS2$ ) is likely due to chance alone. The resulting  $p$ -value for this test can be interpreted as the likelihood of seeing a difference among powers of the size observed without an association.

## 6 Data Simulation

The goal of the simulation study is to generate artificial datasets with high concept difficulty to evaluate the power of ACO in the domain of human genetics. We first develop 30 different penetrance functions (i.e. genetic models) that define a probabilistic relationship between genotype and phenotype where susceptibility to disease is dependent on genotypes from two SNPs in the absence of any independent effects. The 30 penetrance functions include groups of five with heritabilities of 0.025, 0.05, 0.1, 0.2, 0.3, or 0.4. These heritabilities range from a very small to a large genetic effect size. Each functional SNP has two alleles with frequencies of 0.4 and 0.6.

**Table 2.** Penetrance values for an example epistasis model

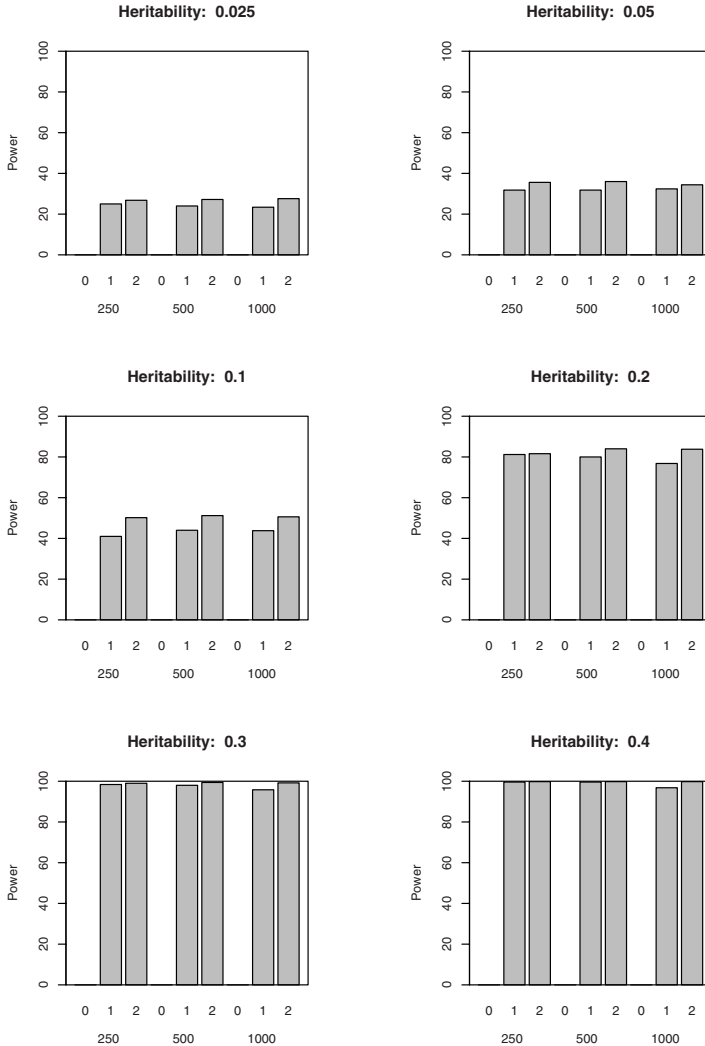
	AA (0.36)	Aa (0.48)	aa (0.16)
BB (0.36)	0.077	0.656	0.880
Bb (0.48)	0.892	0.235	0.312
bb (0.16)	0.174	0.842	0.106

Table 2 summarizes the penetrance values to three significant digits for one of the 30 models. The values in parentheses are the genotype frequencies. Each of the models is used to generate 100 replicate datasets with a sample size of 1600. Each dataset consists of an equal number of case (disease) and control (no disease) subjects. Each pair of functional SNPs is combined within a genome-wide set of 998 randomly generated SNPs for a total of 1000 attributes. A total of 3,000 datasets are generated and analyzed.

## 7 Experimental Design and Statistical Analysis

For each set of 100 datasets and for each set of parameters we count the number of times the correct two functional attributes are selected as the best model by our ACO implementation. This count, expressed as a percentage, is an estimate

of the power of the method. This percentage represents how often ACO meta-heuristic finds the answer that we know is present. We compare the significance of power estimates between the methods (e.g.  $\beta = 0$ ,  $\beta = 1$ , and  $\beta = 2$ ,) by performing fisher's exact test [20]. Results are considered statistically significant when  $p \leq 0.05$ .

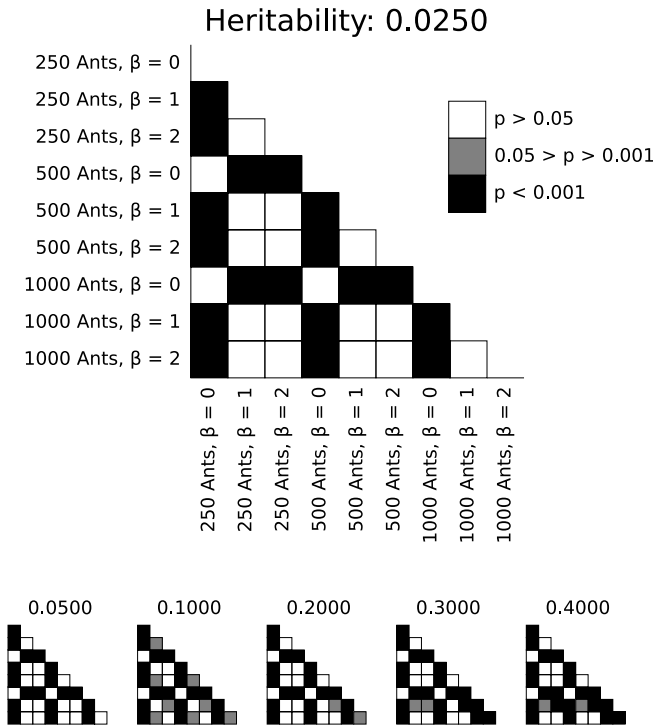


**Fig. 2.** The average power across heritabilities. Each group of three bars shows one combination of ants and updates (250 ants/20 updates, 500 ants/10 updates, 1000 ants/5 updates respectively). Within the ant and update groups the beta value is 0, 1, and 2 from left to right.

## 8 Experimental Results

Figure 2 summarizes the average power (% success) for each method. Each bar represents the power averaged over 500 datasets (5 models with 100 datasets each). Power represents the number of times out of 100 that the ACO finds the right two attributes. These results clearly show that the ACO approach is unable to successfully find the correct pair of attributes when expert knowledge is not used as the power is very low for all cases where the expert knowledge weighting parameter,  $\beta$ , is set to zero. These results also show that the ACO metaheuristic is frequently successful when  $\beta$  is one or two, showing the critical need for expert knowledge.

To assess the reliability and robustness of these results quantitatively we use fisher’s exact test (Section 5). As figure 3 shows, the difference between values of  $\beta$  of 0 and values of  $\beta$  of 1 and 2 is highly significant in all circumstances.



**Fig. 3.** Fisher’s exact test  $p$ -values for assessing whether the differences among groups seen in the bar graph is significant. Values of  $p$  between 0.05 and 0.001 mean that between one time out of twenty and one time out of one-thousand, a difference of that magnitude is expected by chance alone. Values of  $p$  below 0.001 mean that less than one time out of one-thousand, a difference of that magnitude is expected by chance alone. Order of parameter settings is retained between the example heritability (0.0250) and the other heritabilities shown.

This quantitatively confirms that the expert knowledge factor,  $E$  is a crucial component in the success of the ACO metaheuristic in this domain. At a heritability of 0.100 where the largest performance difference occurs among different weightings of  $\beta$ , it is apparent that the difference between the powers for values of  $\beta = 1$  and  $\beta = 2$  is significant. This suggests that a higher weighting of  $\beta$  seems to be advantageous for this problem. In addition at higher heritabilities the difference between the 1000 ants/5 updates power with a  $\beta$  of 1 becomes significantly different from the power with other parameter settings, which also suggests that a high weighting of expert knowledge is more appropriate in these cases, especially when the number of ants is high and the number of updates low.

## 9 Discussion and Conclusion

Our results show that ACO is a viable approach to this problem when an expert knowledge is added in to the pheromone updating rule. This suggests that ACO may be an appropriate search strategy when exhaustive analysis is impossible. These results are also encouraging given the relative simplicity of this approach and the relatively high power given that only about 1% of the dataset can be explored by the metaheuristic. These results indicate a power somewhat greater than the previously used genetic programming approaches [7,9,10].

In this case the pheromone updating rule is a combination of the classifier accuracy and the expert knowledge from TuRF. Modifications such as a rank based ant system [21] or a *MIN - MAX* ant system [22,23] warrant investigation as these approaches may be better able to deal with this type of data. Also warranting more investigation is wide sweep of the  $\beta$ , expert knowledge weighting, parameter which leads to increased power at low (0.100) heritabilities. Merkle et al. show that dynamically altering the heuristic weighting factor,  $\beta$ , during the search can lead to greater success for a resource-constrained project scheduling problem [24]. Perhaps a similar approach is appropriate here to better balance exploration and exploitation.

Work now can focus on a number of areas within the ACO metaheuristic. What is the best way to initialize the pheromone? Are there more appropriate ant systems or updating rules for this problem? We have seen in the field of GP that by developing highly tuned operators it is possible to keep the power of the approach high while exploring a much smaller search space. Is it possible and advantageous to develop similar tuned approaches in the field of ant colony optimization while keeping parameters for the approach conceptually simple enough for users of the MDR software to understand?

In this work the building blocks of outside knowledge are obtained by pre-processing data with TuRF. For the realm of genetic studies, outside knowledge can also be obtained from the numerous public databases available to geneticists. Tools are being developed which integrate knowledge across these public databases and generate information about relationships between genes and disease in the context of protein interactions [25]. Future work will also focus on integrating multiple distinct expert knowledge types and sources. For the ACO

metaheuristic the question arises, is it better to include all types of outside knowledge in the same run in the same large pheromone updating rule, or is it better to use a reinitialization strategy once convergence occurs that takes advantage of different sources of expert knowledge in phases? Here we insert the heuristic information into the pheromone updating rule. We have found that given domain specific knowledge and an approach which takes advantage of this knowledge, it is possible for an ACO strategy to succeed, even for a needle-in-a-haystack problem. This indicates that ACO may be a useful wrapper for genome wide analysis of common human diseases with a complex genetic architecture.

**Acknowledgements.** This work was supported by NIH grants LM009012 and AI59694. The authors would like to thank Chantel Sloan, Anna Tyler and Ryan Urbanowicz for their careful reading of the manuscript.

## References

1. Freitas, A.A.: Understanding the crucial role of attribute interaction in data mining. *Artif. Intell. Rev.* 16(3), 177–199 (2001)
2. Goldberg, D.E.: *The Design of Innovation: Lessons from and for Competent Genetic Algorithms*. Kluwer Academic Publishers, Norwell (2002)
3. Moore, J.H.: The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Human Heredity* 56, 73–82 (2003)
4. The International HapMap Consortium: A haplotype map of the human genome. *Nature* 437(7063), 1299–1320 (2005); 10.1038/nature04226
5. Dorigo, M., Maniezzo, V., Coloni, A.: Positive feedback as a search strategy. Technical report 91-016, Dipartimento di Elettronica e Informatica, Politecnico di Milano (1991)
6. Parpinelli, R., Lopes, H., Freitas, A.: An Ant Colony Based System for Data Mining: Applications to Medical Data. In: *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2001)*, pp. 791–797 (2001)
7. Moore, J.H., White, B.C.: Genome-wide genetic analysis using genetic programming: The critical need for expert knowledge. *Genetic Programming Theory and Practice IV* (2007)
8. White, B.C., Gilbert, J.C., Reif, D.M., Moore, J.H.: A statistical comparison of grammatical evolution strategies in the domain of human genetics. In: *Proceedings of the IEEE Congress on Evolutionary Computing*, pp. 676–682 (2005)
9. Moore, J.H., White, B.C.: Exploiting expert knowledge in genetic programming for genome-wide genetic analysis. In: Runarsson, T.P., Beyer, H.-G., Burke, E.K., Merelo-Guervós, J.J., Whitley, L.D., Yao, X. (eds.) *PPSN 2006*. LNCS, vol. 4193, pp. 969–977. Springer, Heidelberg (2006)
10. Greene, C.S., White, B.C., Moore, J.H.: An expert knowledge-guided mutation operator for genome-wide genetic analysis using genetic programming. In: Rajapakse, J.C., Schmidt, B., Volkert, L.G. (eds.) *PRIB 2007*. LNCS (LNBI), vol. 4774, pp. 30–40. Springer, Heidelberg (2007)
11. Moore, J.H., White, B.C.: Tuning relief for genome-wide genetic analysis. In: Marchiori, E., Moore, J.H., Rajapakse, J.C. (eds.) *EvoBIO 2007*. LNCS, vol. 4447, pp. 166–175. Springer, Heidelberg (2007)

12. Moore, J.H.: Computational analysis of gene-gene interactions using multifactor dimensionality reduction. *Expert Review of Molecular Diagnostics* 4(6), 795–803 (2004)
13. Moore, J.H.: Genome-wide analysis of epistasis using multifactor dimensionality reduction: feature selection and construction in the domain of human genetics. In: *Knowledge Discovery and Data Mining: Challenges and Realities with Real World Data*. IGI (2007)
14. 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)
15. 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. *American Journal of Human Genetics* 69, 138–147 (2001)
16. Wilke, R.A., Reif, D.M., Moore, J.H.: Combinatorial pharmacogenetics. *Nature Reviews Drug Discovery* 4, 911–918 (2005)
17. Kira, K., Rendell, L.A.: A practical approach to feature selection. In: *Machine Learning: Proceedings of the AAA 1992* (1992)
18. Kononenko, I.: Estimating attributes: Analysis and extension of relief. In: *Machine Learning: ECML-1994*, vol. 94, pp. 171–182 (1994)
19. Robnik-Sikonja, M., Kononenko, I.: Theoretical and empirical analysis of relief and relief. *Mach. Learn.* 53, 23–69 (2003)
20. 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)
21. Bullnheimer, B., Hartl, R., Strauss, C.: A new rank-based version of the ant system: a computational study. *Central European Journal for Operations Research and Economics* 7(1), 25–38 (1999)
22. Stützle, T., Hoos, H.: MAX-MIN Ant System and local search for the traveling salesman problem. *IEEE International Conference on Evolutionary Computation 1997*, 309–314 (1997)
23. Stützle, T., Hoos, H.H.: MAX-MIN Ant System. *Future Generation Computer Systems* 16(8), 889–914 (2000)
24. Merkle, D., Middendorf, M., Schmeck, H.: Ant colony optimization for resource-constrained project scheduling. *IEEE Transactions on Evolutionary Computation* 6(4), 333–346 (2002)
25. Gonzalez, G., Uribe, J.C., Tari, L., Brophy, C., Baral, C.: Mining gene-disease relationships from biomedical literature: Weighting protein-protein interactions and connectivity measures. In: *Pacific Symposium on Biocomputing*, vol. 12, pp. 28–39 (2007)