Multifactor dimensionality reduction software for detecting gene–gene and gene–environment interactions

*Lance W. Hahn, Marylyn D. Ritchie and Jason H. Moore**

Program in Human Genetics and Department of Molecular Physiology and Biophysics, Vanderbilt University Medical School, Nashville, TN 37232-0700, USA

ABSTRACT

Motivation: Polymorphisms in human genes are being described in remarkable numbers. Determining which polymorphisms and which environmental factors are associated with common, complex diseases has become a daunting task. This is partly because the effect of any single genetic variation will likely be dependent on other genetic variations (gene–gene interaction or epistasis) and environmental factors (gene–environment interaction). Detecting and characterizing interactions among multiple factors is both a statistical and a computational challenge. To address this problem, we have developed a multifactor dimensionality reduction (MDR) method for collapsing high-dimensional genetic data into a single dimension thus permitting interactions to be detected in relatively small sample sizes. In this paper, we describe the MDR approach and an MDR software package.

Results: We developed a program that integrates MDR with a cross-validation strategy for estimating the classification and prediction error of multifactor models. The software can be used to analyze interactions among 2–15 genetic and/or environmental factors. The dataset may contain up to 500 total variables and a maximum of 4000 study subjects.

Availability: Information on obtaining the executable code, example data, example analysis, and documentation is available upon request.

Contact: moore@phg.mc.vanderbilt.edu

Supplementary information: All supplementary information can be found at http://phg.mc.vanderbilt.edu/Software/MDR.

INTRODUCTION

An important goal of human genetics is to identify DNA sequence variations or polymorphisms in human genes that confer an increased risk to particular diseases. This is a difficult challenge for common, complex multifactorial diseases such as essential hypertension that are likely the result of interactions between multiple genetic and environmental factors (Kardia, 2000; Moore and Williams, 2002). Such gene–gene and gene–environment interactions are difficult to detect and characterize using traditional parametric statistical methods such as logistic regression because of the sparseness of the data in high dimensions. That is, when interactions among multiple variables are considered, there are many contingency table cells that have very few or no data points. This is referred to as the curse of dimensionality (Bellman, 1961) and can lead to parameter estimates that have very large standard errors resulting in an increase in type I errors (Concato *et al.*, 1996; Peduzzi *et al.*, 1996; Hosmer and Lemeshow, 2000). In addition, detecting gene–gene and gene–environment interactions using traditional procedures for fitting logistic regression models can be problematic leading to an increase in type II errors and a decrease in power. For example, forward selection is limited because interactions are only tested for those variables that have a statistically significant independent main effect. Those DNA sequence variations that have an interaction effect but not a main effect will be missed. With backward elimination, a complete model that includes all main effects and all interaction terms may require too many degrees of freedom. Stepwise procedures are more flexible than either forward selection or backward elimination but can also suffer from requiring too many degrees of freedom. Detecting interactions among variables is a well-known problem in data mining (Freitas, 2001) and an increasingly recognized problem in human genetics (Templeton, 2000; Moore and Williams, 2002).

To address concerns about inaccurate parameter estimates and low power for identifying interactions in relatively small sample sizes, we developed a nonparametric and genetic model-free approach called multifactor dimensionality reduction or MDR that uses a data reduction strategy (Ritchie *et al.*, 2001). With MDR, multilocus

*To whom correspondence should be addressed.

genotypes are pooled into high risk and low risk groups, effectively reducing the dimensionality of the genotype predictors from N dimensions to one dimension. The new one-dimensional multilocus genotype variable is evaluated for its ability to classify and predict disease status using cross-validation and permutation testing. The MDR approach is model-free in that it does not assume any particular genetic model and is nonparametric in that it does not estimate any parameters. We have demonstrated that MDR is able to identify evidence for high-order gene–gene interactions in the absence of any statistically significant independent main effects in simulated data (Ritchie *et al.*, 2001, 2003), in sporadic breast cancer (Ritchie *et al.*, 2001), and in essential hypertension (Moore and Williams, 2002). Further, we have outlined a mathematical proof that no method will discriminate between multilocus clinical endpoints better than MDR using multilocus genotypes (Hahn and Moore, 2003). In the present study, we describe recent extensions to the MDR method and a software package for implementing MDR in case-control and discordant sib-pair study designs.

ALGORITHM

For ease of discussion, let us first consider the case in which MDR is used without cross-validation. In this instance, the MDR algorithm uses the complete dataset to identify the variables (i.e. genetic and/or environmental factors) that show the strongest association with disease status. In this algorithm, there are two parameters required from the user: (1) N, the number of variables to be selected at one time; and (2) T, the threshold ratio of affected individuals to unaffected individuals that is used to distinguish high-risk genotype combinations from low-risk genotype combinations. Typically, the MDR analysis is repeated using a range of values for N with part of the model discovery process being the selection of a best N. It should be noted that a single-locus analysis for main effects can be conducted with MDR by setting N to one. Selecting a best value for T depends on the goals of the analysis. The influence of T on the results and inferences is an active area of investigation (Hahn and Moore, 2003).

Now, let us consider MDR analysis with cross-validation. Before dividing the data into a training set and a testing set, the MDR program randomly shuffles the order of the observations in the dataset using a random seed supplied by the user. This reduces the risk of biasing the cross-validation due to nonrandom ordering of the data (e.g. the first half of the data consisting of affected subjects and the second half unaffected subjects). After the randomization, the individuals are arranged so that disease status alternates. If the number of affected and unaffected individuals is unbalanced, the alternation

ceases before reaching the end of the dataset. When using matched data, such as data from a matched case-control study, randomizing and alternating the data is undesirable because it disrupts the matching. In such a case, the randomization and alternation can be turned off with the RANDOMSHUFFLE parameter. The type of cross-validation used by MDR (NUMBERCVINTERVALS) is determined by the number of cross-validation intervals as defined by the user. When the number of affected and unaffected individuals are not equal, the excess individuals are distributed across the cross-validation groups as evenly as possible. The cross-validation analysis can be conducted several times using different random seeds and the results averaged to avoid spurious results due to chance divisions of the data (Ritchie *et al.*, 2001). Cross-validation allows estimation of the prediction error of a model by leaving out a portion of the data as an independent test set (Hastie *et al.*, 2001). With 10-fold cross-validation, the data are divided into 10 equal parts, the model is developed on 9/10 of the data (i.e. the training data) and then evaluated on the remaining 1/10 of the data (i.e. the independent testing data). This is repeated for each possible 9/10 and 1/10 of the data and the resulting ten prediction errors are averaged. Leave-one-out cross validation (LOOCV) is also a common strategy for estimating prediction error (Hastie *et al.*, 2001). With LOOCV, a single observation is left out as the independent test data.

Figure 1 illustrates the general steps involved in implementing the MDR method for case-control study designs using 10-fold cross-validation. The same procedure is equally applicable to discordant sib-pair study designs. The first step of MDR involves partitioning the data into some number of equal parts for cross-validation. In step two, a set of N genetic and/or discrete environmental factors is selected from the list of all factors. In step three, the N factors and their multifactor classes or cells are represented in N-dimensional space. For example, for two polymorphisms, each with three genotypes, there are nine two-locus genotype combinations. Then, the ratio of the number of cases (or affected sibs) to the number of controls (or unaffected sibs) is evaluated within each multifactor cell. In step four, each multifactor cell in N-dimensional space is labeled as high-risk if the ratio of cases to controls meets or exceeds some threshold T (e.g. $T = 1.0$) and low-risk if the threshold is not exceeded. In this way, a model for cases and controls (or affected and unaffected sibs) is formed by pooling those cells labeled high-risk into one group and those cells labeled low-risk into another group. This reduces the N-dimensional model to one dimension (i.e. one variable with two multifactor classes; low risk and high risk). In the case where there are cases but no controls, the cell is labeled high-risk. Likewise, when there are controls

but no cases, the cell is labeled low-risk. In the final model, the user can decide whether this is an accurate assignment. All possible combinations of N factors are evaluated sequentially for their ability to classify affected and unaffected individuals in the training data and the best N-factor model is selected in step five. In step six, the independent test data from the cross-validation is used to estimate the prediction error of the best model selected in step five. Steps one through six are repeated 10 times with the data split into 10 different training and testing sets. The data are only randomized and alternated once at the beginning of the analysis. Once MDR identifies the best combination of factors, the final step is to determine which multifactor levels (e.g. genotypes) are high risk and which are low risk using the entire dataset. This final evaluation is conducted with a ratio threshold that is determined by the ratio of the number of affected individuals in the dataset divided by the number of unaffected individuals in the dataset. Note that the user-defined threshold ratio (T) is not used in this final evaluation. This final threshold is used to adjust the results for an unbalanced number of cases and controls. The user can reassign high-risk and low-risk labels in a posthoc analysis of the output if needed. Also note that, in this implementation of MDR, only variables with a maximum of three levels (e.g. three genotypes) are allowed.

Previous versions of MDR required each individual in the dataset to have observed data for each variable. We have now modified MDR to accept missing data by defining a new level for each variable to be used when missing data is encountered. Thus, instead of each factor having three levels, there is now a fourth level encoding the missing data point. Thus, if three factors are being modeled, and one of them has missing data, information from the other two can now be incorporated into the analysis. All missing data should be encoded by a specific number (e.g. 9) that is different from the encoding used for the factor levels (e.g. 0, 1, and 2). The software described here includes the new missing data feature.

As described by Ritchie *et al.* (2001) and Moore *et al.* (2002b), it may be of interest to use cross-validation consistency as a measure of evidence for a particular model in addition to the prediction error. That is, how often were the same genes or same set of genes selected across different cross-validation datasets? The reasoning is that the functional factors should consistently be found regardless of how the data are divided for cross-validation. Finally, an empirical P-value for the result can be determined using one of a number of permutation testing strategies (Good, 2000). Since there are multiple ways to conduct the permutation testing, we leave this part of the analysis to the user. One strategy is to randomize the case and control labels in the original dataset multiple times to create a set of permuted datasets. MDR can then be run on
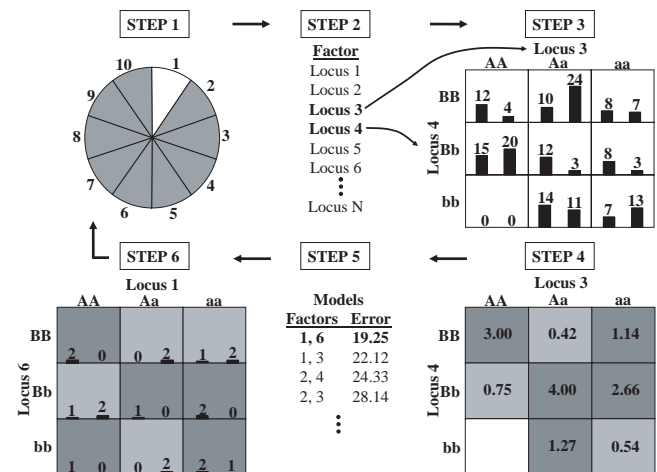
Fig. 1. Summary of the general steps involved in implementing the MDR method (adapted from Ritchie *et al.*, 2001). In step one, the data are divided into a training set (e.g. 9/10 of the data) and an independent testing set (e.g. 1/10 of the data) as part of cross-validation. In step 2, a set of N genetic and/or discrete environmental factors is then selected from the pool of all factors. In step three, the N factors and their possible multifactor classes or cells are represented in N-dimensional space. In step four, each multifactor cell in the N-dimensional space is labeled as high-risk if the ratio of affected individuals to unaffected individuals (the number in the cell) exceeds some threshold T (e.g. $T = 1.0$) and low-risk if the threshold is not exceeded. In steps five and six, the model with the best misclassification error is selected and the prediction error of the model is estimated using the independent test data. Steps 1 through 6 are repeated for each possible cross-validation interval. Bars represent hypothetical distributions of cases (left) and controls (right) with each multifactor combination. Dark-shaded cells represent high-risk genotype combinations while light-shaded cells represent low-risk genotype combinations. No shading or white cells represent genotype combinations for which no data was observed.

each permuted dataset and the maximum cross-validation consistency and minimum prediction error identified for each saved and used to create an empirical distribution for estimation of a P-value. This is the approach we use in the example run described below.

When dealing with many variables, one may want to consider a number of N-factor models where N may vary across a range of factor counts. The MDR software can be run several times with the configuration reflecting a different number of factors each time. Each run of MDR will produce a single model that maximizes the number individuals with the proper risk assignment. Single best models are selected from among each of the one-factor, two-factor, three-factor, four-factor, up to N-factor combinations. Among this set of best multifactor models, the combination of genetic and/or discrete

environmental factors that minimizes the prediction error and/or maximizes the cross-validation consistency is selected and evaluated using permutation testing. When two or more models have the same prediction error and cross-validation consistency, statistical parsimony can be used to select the smaller model as the more likely candidate. It is up to the user to select the final model using these criteria or others.

For large datasets and/or high-order models, the exhaustive consideration of all possible factor combinations can become computationally infeasible. When the number of combinations to be evaluated exceeds computational feasibility, machine-learning methods such as parallel genetic algorithms (Cantu-Paz, 2000) must be employed as has been done for the cellular automata method (Moore and Hahn, 2002a,b).

IMPLEMENTATION

The MDR software is available in a Linux version (compiled and benchmarked on a PC with a 600 MHz Pentium-III running Red Hat 2.2.5-15) and a Sun version (compiled and benchmarked on a SPARC Ultra-80 running SunOS 5.8). It was written in C and compiled with the GNU C compiler. The executable files and some example datasets are freely available to not-for-profit organizations and studies upon request. Benchmark measurements are available in the supplemental documentation at http://phg.mc.vanderbilt.edu/Software/MDR. In addition to the platform used, the following analysis parameters significantly altered computation time: (1) the number of factors considered for a model; (2) the number of cross-validation intervals; (3) the number of individuals in a dataset; and (4) the number of variables in the dataset.

Input format

The input file can be in one of two formats: (1) text format; and (2) restricted pre-makeped format. In text format, a single individual's data are represented on a single line. The first value in a line is the individual's disease status and all following values are assumed to be genotypic or discrete environmental data. The values must be separated by non-numeric delimiters such as spaces, tabs or commas. Although disease status is usually 1 for affected and 0 for unaffected, the user can specify the value associated with affected individuals in the configuration file. Comments may be included at the beginning of the data file by simply initiating the line with an alphabetic or non-numeric symbol.

Pre-makeped format refers to the input files used by the utility makeped (Terwilliger and Ott, 1994). For a given line, characters 1 through 16 are used to describe the pedigree ID and other attributes of the individual; character 17 indicates disease status (1 is unaffected, 2 is affected). Note that this is more restrictive than makeped

in that the individual attributes in the data must be exactly 16 characters. The genetic variables are described in terms of alleles with each allele separated by a space.

Datasets may contain up to 4000 individuals with up to 500 factors or variables. Genetic and environmental variables can have two or three levels (e.g. 0 and 1 or 0, 1 and 2) plus an additional level (e.g. 9) for missing data. The order of the variables is not important unless MDR discovers multiple models with the same misclassification error. In the current implementation of MDR, only the first model is reported when a tie between models occurs. Reporting all tied models will be implemented in a future version of the software.

Example run

To provide an example of data analysis using the MDR program, we simulated a case-control sample dataset with 200 affected individuals (cases) and 200 unaffected individuals (controls) with 10 single nucleotide polymorphisms or SNPs for each individual. The goal of our simulation was to create a dataset that represented a disease etiology that was due to two interacting SNPs. Frankel and Schork (1996) and Moore *et al.* (2002a) have described a complex two-variable gene–gene interaction model in which *aaBB*, *AaBb*, and *AAbb* are the high-risk genotype combinations. When the allele frequencies are equal, there is a strong interaction effect on disease risk in the absence of any main or independent effect for either of the genetic variations. This means that the genetic variations do not independently affect disease risk. Thus, each genetic variation only has an effect on disease risk in the context of the other genetic variation. Such gene–gene interactions are believed to play an important role in determining an individual's risk for developing common diseases such as essential hypertension (Moore and Williams, 2002) and sporadic breast cancer (Ritchie *et al.*, 2001). The following penetrance values were used to define the probability (P) of disease (D) given each specific combination of genotypes from the two functional SNPs: $P(\mathrm{D}|AAbb) = 0.1$, $P(\mathrm{D}|AaBb) = 0.05$, $P(\mathrm{D}|aaBB) = 0.1$, $P(\mathrm{D}|\text{ others}) = 0$. The other eight SNPs in the simulated data are not functional and therefore represent potential false-positives.

We analyzed the data using 10-fold cross-validation and configured MDR to consider one through six variables at a time. Table 1 describes examples of parameter settings used by MDR. To ensure that the analysis was not influenced by a chance division of the data (i.e. an order effect) or by initial conditions, we ran the analysis 10 times using 10 different random number seeds. A portion of an output file generated by MDR is shown in Figure 2. The text in the figure was produced when two variables were considered for a solution and the random seed three was used. Only the last (i.e. 10th) cross-validation result and the final result are shown. The tenth cross-validation

Table 1. Descriptive list of MDR parameters

Parameter name	Example value	Parameter description
INPUTFILE	SampleData.dat	The data file to be analyzed
INDIVIDLIMIT	400	The number of individuals to be analyzed
INPUTTYPE	1	Data format (0 – text, 1 – pre-makeped)
AFFECTEDINPUTVALUE	1	Code for 'Affected' disease status
LEGALMAXGENOTYPE	2	Highest legal genotype value
LEGALMINGENOTYPE	0	Lowest legal genotype value
MDRRANDOMSEED	3	Random seed used for shuffling the data
RANDOMSHUFFLE	OFF	Randomly shuffle the data
LOCICONSIDERED	2	The number of loci considered when forming and evaluating a model.
FORCELOCI	OFF	Allows a user to force MDR to consider only user-specified loci.
THRESHOLDRATIO	1.0	Threshold for associating a ratio with high-risk status
NOTRECOGNIZEDRESPONSE	-1	How should novel patterns be treated during testing? (-1 - unknown, 0 - as unaffected, 1 as affected).
TIECELLVALUE	1	How should ratios be labeled when they equal the threshold ratio? (-1 – unknown, 0 – unaffected, 1 – affected)
SHOWBESTPARTITION	OFF	Display the components of the best model
SHOWCOMBOMISCLASS	OFF	Show all model misclassification rates.
SHOWMAINEFFECT	OFF	Display a table of the number of individuals grouped by locus and genotype
VERBOSE	OFF	Display statements about MDR progress
NUMBERCVINTERVALS	10	The number of cross-validation intervals

analysis found the two functional SNPs (one and six) as the best two-factor model for the 360 individuals in the training data with a classification error of 19.44%. This model predicted a disease status for all 40 individuals in the test dataset with a prediction error of 17.5%. In some cases, individuals in the testing data may include a multilocus genotype combination that has not been encountered in the training data. To deal with this possibility, a parameter (NOTRECOGNIZEDRESPONSE) can be set in the configuration file so that genotype combinations not encountered during training are assigned 'unaffected' disease status during testing. As a result, all individuals would be classified as either affected or unaffected. If the parameter is set to assign an 'unknown' disease status to patterns encountered during testing that were not encountered during training, then the number of individuals not classified during testing can be greater than zero.

The final result includes a description of the model from the ten cross-validation intervals with the lowest prediction error and how well the model performs on the whole dataset. Additional statistics reported are mean classification and prediction errors across the ten best models.

The results of the MDR analysis for each number of factors considered are presented in Table 2. The model with the lowest prediction error and highest cross-validation consistency was selected for each number of factors considered. The reported cross-validation consistency is the number of cross-validation intervals (maximum of 10) that a particular SNP combination was chosen

```
========================Subgroup # 10 of 10==========

#10:Loci: [ 1   6 ]
#10:                              Training        Evaluating
#10:                          (1 group of 360) (1 group of 40)
#10:Statistics describing the Best model found:
#10: Misclassification Errors:         19.44%       17.50%
#10: # of Individuals that couldn't be classified:        0
========================Final Result ================
During model generation a threshold ratio of 1.000000
was used (as specified in the configuration file.)
During this final evaluation of the model and
data set, 1.000000 was used as the threshold ratio.
Final Loci:  1  6
# 2:         Complete Dataset
# 2:             (400)
# 2: Misclass:       19.25%
                                   Training   Evaluating
Mean misclass across all best models:  19.25%     19.25%
```

Fig. 2. This is the last of 10 cross-validation intervals and the final result of an MDR analysis log file. The two-factor model with the best evaluation performance included variables 1 and 6. The lowest evaluation misclassification occurred in the second cross-validation interval and the final misclassification for the two-factor model is 19.25% for the whole dataset.

by MDR averaged across the 10 runs. The average classification and prediction errors are the averages across all cross-validation intervals and all runs. The model that minimized prediction error and maximized the cross-validation consistency was the two-factor model that included the correct functional SNPs, one and six.

Table 2. Results of an MDR analysis of the example dataset

No. of factors considered	Best candidate model	Average cross-validation consistency	Average classification error (%)	Average prediction error (%)
1	4	7.3	44.83	49.88
2	**1, 6**	**10.0**[a]	**19.25**[a]	**19.25**[a]
3	1, 2, 6	10.0	19.25	19.30
4	1, 3, 6, 7	3.4	17.74	24.08
5	1, 2, 3, 6, 7	4.6	14.24	25.37
6	1, 2, 3, 6, 7, 10	1.0	9.58	33.65

[a]$P < 0.001$.

Note that classification error decreases as the number of factors considered increases. This is due to higher-order models overfitting the data. Also note that as the model size increases beyond two, the prediction error increases. Thus, although the higher-order models are overfitting the data, they do a worse job predicting. The permutation testing indicated the cross-validation consistency and the prediction error are statistically significant at the 0.001 level. This indicates that among 1000 permuted datasets, no best models had a cross-validation consistency or a prediction error of the same magnitude as was observed for the original dataset. The configuration files, data files and log files for this example analysis are available upon request.

DISCUSSION AND CONCLUSIONS

We have developed a software package for implementing the multifactor dimensionality reduction (MDR) approach of Ritchie *et al.* (2001) for detecting and characterizing gene–gene and gene–environment interaction effects on risk of common complex multifactorial diseases. This implementation of MDR can be used to analyze interactions among up to 15 genetic and/or environmental factors in a maximum of 4000 study subjects and up to 500 total variables. By allowing up to 15 variables to play a role in the categorization, the model solution can contain all main effects and N-way interactions involving the 15 variables.

The greatest limitation of the MDR software is the combinatorial nature of the algorithm. Genetic datasets with hundreds or thousands of variables for each individual sample will quickly overwhelm MDR. Although, faster computers will allow the analysis of a slightly larger range of variables (greater than 100), novel approaches will be required to analyze datasets with several hundred or more variables. For example, analysis of a 50-variable dataset using five variables at a time requires assessing over two million variable combinations. Our Pentium III 600-MHz PC took just under 6 hours to perform these computations. If the dataset had contained 1000 variables instead, the number of combinations would increase to over eight

trillion. Datasets of this magnitude will be available in the near future, but the processing speed required to analyze them will not.

Given the combinatorial limitations of MDR, we are currently exploring machine learning strategies for selecting optimal combinations of genetic and environmental factors from among an effectively infinite search space. Evolutionary computation is a machine learning strategy that we have used successfully in genetic epidemiology (Moore *et al.*, 2002a; Moore and Hahn, 2002a,b) and genomics (Moore and Parker, 2001; Moore *et al.*, 2002b) and may be useful for optimizing MDR. Future enhancements of the MDR algorithm and software will include an evolutionary computing search algorithm for selecting genetic and environmental factors.

Multifactor dimensionality reduction is currently applicable only to case-control and discordant sib-pair study designs. However, many genetic studies are carried out using multigenerational family data. A future direction is to modify MDR to identify gene–gene and gene–environment interactions in large, complex pedigrees. For example, it may be possible to merge MDR with the pedigree disequilibrium statistic (PDT) of Martin *et al.* (2000). The PDT was developed specifically to provide a general test of linkage disequilibrium that can be applied to complex pedigrees even in the presence of population substructure.

Multifactor dimensionality reduction is a promising new approach for overcoming some of the limitations of logistic regression for the detection and characterization of gene–gene and gene–environment interactions. Previous empirical studies demonstrate that MDR has good power for identifying high-order interactions in simulated data (Ritchie *et al.*, 2001, 2003). Further, MDR has played an important role in the identification of gene–gene interactions in real data from case-control studies of sporadic breast cancer (Ritchie *et al.*, 2001) and essential hypertension (Moore and Williams, 2002). Additionally, a theoretical study has provided a proof that MDR is ideally suited for discriminating between binary clinical

endpoints using multilocus genotypes (Hahn and Moore, 2003). The availability of an MDR software package will enable this new method to be widely used for genetic epidemiology studies of common complex multifactorial diseases.

ACKNOWLEDGEMENTS

REFERENCES

Bellman,R. (1961) *Adaptive Control Processes*. Princeton University Press, Princeton, NJ.

Cantu-Paz,E. (2000) *Efficient and Accurate Parallel Genetic Algorithms*. Kluwer, Boston.

Concato,J., Feinstein,A.R. and Holford,T.R. (1996) The risk of determining risk with multivariable models. *Ann. Int. Med.*, **118**, 201–210.

Frankel,W.N. and Schork,N.J. (1996) Who's afraid of epistasis? *Nature Genet.*, **14**, 371–373.

Freitas,A.A. (2001) Understanding the crucial role of attribute interaction in data mining. *Artif. Intel. Rev.*, **16**, 177–199.

Good,P. (2000) *Permutation Tests: a Practical Guide to Resampling Methods for Testing Hypotheses*. Springer, New York.

Hahn,L.W. and Moore,J.H. (2003) Multifactor dimensionality reduction is an ideal discriminator of discrete clinical endpoints using multilocus genotypes. Submitted.

Hastie,T., Tibshirani,R. and Friedman,J. (2001) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.

Hosmer,D.W. and Lemeshow,S. (2000) *Applied Logistic Regression*. Wiley, New York.

Kardia,S.L.R. (2000) Context-dependent genetic effects in hypertension. *Curr. Hypertens Reports*, **2**, 32–38.

Martin,E.R., Monks,S.A., Warren,L.L. and Kaplan,N.L. (2000) A test for linkage and association in general pedigrees: the pedigree disequilibrium test. *Am. J. Hum. Genet.*, **67**, 146–154.

Moore,J.H. and Parker,J.S. (2001) Evolutionary computation in microarray data analysis. In Lin,S. and Johnson,K. (eds), *Methods of Microarray Data Analysis*. Kluwer, Boston, pp. 21–35.

Moore,J.H. and Williams,S.W. (2002) New strategies for identifying gene–gene interactions in hypertension. *Ann. Med.*, **34**, 88–95.

Moore,J.H. and Hahn,L.W. (2002a) A cellular automata approach to detecting interactions among single-nucleotide polymorphisms in complex multifactorial diseases. *Pac. Symp. Biocomput.*, **7**, 53–64.

Moore,J.H. and Hahn,L.W. (2002b) Cellular automata and genetic algorithms for parallel problem solving in human genetics. In Merelo,J.J., Panagiotis,A. and Beyer,H.-G. (eds), *Lecture Notes in Computer Science*. Springer, Berlin, pp. 821–830.

Moore,J.H., Hahn,L.W., Ritchie,M.D., Thornton,T.A. and White,B.C. (2002a) Application of genetic algorithms to the discovery of complex genetic models for simulation studies in human genetics. In Langdon,W.B., Cantu-Paz,E., Mathias,K., Roy,R., Davis,D., Poli,R., Balakrishnan,K., Honavar,V., Rudolph,G., Wegener,J., Bull,L., Potter,M.A., Schultz,A.C., Miller,J.F., Burke,E. and Jonoska,N. (eds), *Proceedings of the Genetic and Evolutionary Computation Conference*. Morgan Kaufmann, San Francisco, pp. 1150–1155.

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

Peduzzi,P., Concato,J., Kemper,E., Holford,T.R. and Feinstein,A.R. (1996) A simulation study of the number of events per variable in logistic regression analysis. *J. Clin. Epidemiol.*, **49**, 1373–1379.

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

Ritchie,M.D., Hahn,L.W. and Moore,J.H. (2003) Power of multifactor dimensionality reduction for detecting gene–gene interactions in the presence of genotyping error, phenocopy, and genetic heterogeneity. *Genet. Epidemiol.*, in press.

Terwilliger,J.D. and Ott,J. (1994) *Handbook of Human Genetic Linkage*. Johns Hopkins University Press, Baltimore, pp. 15.

Templeton,A.R. (2000) Epistasis and complex traits. In Wade,M., Brodie III,B. and Wolf,J. (eds), *Epistasis and the Evolutionary Process*. Oxford University Press, New York, pp. 41–57.