Multifactor dimensionality reduction: An analysis strategy for modelling and detecting gene–gene interactions in human genetics and pharmacogenomics studies

*Alison A. Motsinger and Marylyn D. Ritchie**

Center for Human Genetics Research, Vanderbilt University Medical Center, 519 Light Hall, Nashville, TN 37232 0700, USA
* *Correspondence to*: Tel: +1 615 343 5851; Fax: +1 615 343 8619; E-mail: ritchie@chgr.mc.vanderbilt.edu

Date received (in revised form): 9th January 2005

Abstract

The detection of gene–gene and gene–environment interactions associated with complex human disease or pharmacogenomic endpoints is a difficult challenge for human geneticists. Unlike rare, Mendelian diseases that are associated with a single gene, most common diseases are caused by the non-linear interaction of numerous genetic and environmental variables. The dimensionality involved in the evaluation of combinations of many such variables quickly diminishes the usefulness of traditional, parametric statistical methods. Multifactor dimensionality reduction (MDR) is a novel and powerful statistical tool for detecting and modelling epistasis. MDR is a non-parametric and model-free approach that has been shown to have reasonable power to detect epistasis in both theoretical and empirical studies. MDR has detected interactions in diseases such as sporadic breast cancer, multiple sclerosis and essential hypertension.

As this method is more frequently applied, and was gained acceptance in the study of human disease and pharmacogenomics, it is becoming increasingly important that the implementation of the MDR approach is properly understood. As with all statistical methods, MDR is only powerful and useful when implemented correctly. Concerns regarding dataset structure, configuration parameters and the proper execution of permutation testing in reference to a particular dataset and configuration are essential to the method's effectiveness.

The detection, characterisation and interpretation of gene–gene and gene–environment interactions are expected to improve the diagnosis, prevention and treatment of common human diseases. MDR can be a powerful tool in reaching these goals when used appropriately.

Keywords: *epistasis, multifactor dimensionality reduction, gene–gene interactions, gene–environment interactions, pharmacogenomics*

Introduction

One of the biggest challenges in human genetics is identifying polymorphisms, or sequence variations, that present an increased risk of disease. In the case of rare, Mendelian single-gene disorders, such as sickle-cell anaemia or cystic fibrosis, the genotype to phenotype relationship is easily apparent, because the mutant genotype is explicitly responsible for disease. In the case of common, complex diseases, such as hypertension, diabetes or multiple sclerosis, this relationship is extremely difficult to characterise because disease is likely to be the result of many genetic and environmental factors. In fact, epistasis, or gene–gene interaction, is increasingly assumed to play a crucial role in the genetic architecture of common diseases.[1–3] This challenge is equally present in studies of pharmacogenomics.[4]

The dimensionality involved in the evaluation of combinations of many such variables quickly diminishes the usefulness of traditional, parametric statistical methods. Referred to as the curse of dimensionality,[5] as the number of genetic or environmental factors increases and the number of possible interactions increases exponentially, many contingency table cells will be left with very few, if any, data points. In logistic regression analysis, this can result in increased type I errors and parameter estimates with very large standard errors.[6] Traditional approaches using logistic regression modelling are limited in their ability to deal with many factors and simultaneously fail to characterise epistasis

models in the absence of main effects, due to the hierarchical model-building process.[7] This leads to an increase in type II errors and decreased power.[8] This is a particular problem with relatively small sample sizes. Because sample collection is time-consuming and expensive, the decreased power can make the cost of effective studies prohibitive with traditional analytical methods.

In order to address these concerns, a novel statistical method, multifactor dimensionality reduction (MDR), was developed. MDR reduces the dimensionality of multilocus data to improve the ability to detect genetic combinations that confer disease risk. MDR pools genotypes into 'high-risk' and 'low-risk' or 'response' and 'non-response' groups in order to reduce multidimensional data into only one dimension. Because it is a non-parametric method, no hypothesis concerning the value of any statistical parameter is made. It is also a model-free method, so no genetic inheritance model is assumed.[9]

MDR was designed to detect gene–gene or gene–environment interactions in datasets with categorical independent variables, such as single nucleotide polymorphisms (SNPs) and other sequence variations (insertions, deletions etc), as well as environmental data that can be represented as categorical variables. The endpoint, or dependent variable, must be dichotomous such as case/control for studies of human disease. Pharmacogenomics data can also be analysed with MDR, in terms of 'response/non-response' or 'toxicity/no toxicity'. MDR is appropriate for any data type with two distinct clinical endpoints.

MDR has been used to identify interactions in the absence of any significant main effects in simulated data. In addition, MDR has identified interactions in a variety of different real datasets, including sporadic breast cancer,[9] essential hypertension,[7] type 2 diabetes,[10] atrial fibrillation,[11] amyloid polyneuropathy[12] and coronary artery calcification.[13] Each of these studies was the first of its kind to explore complex interactions and thus needs to be replicated in additional datasets. Studies with simulated data (of multiple models of different allele frequencies and heritability) have also shown that MDR has high power to identify interactions in the presence of many types of noise commonly found in real datasets (including missing data and genotyping error), whereas errors such as heterogeneity (genetic or locus) and phenocopy diminish the power of MDR.[14] Additionally, a mathematical proof has shown that, due to the relationship between MDR and a naïve Bayes classifier, MDR is optimally efficient in discriminating between clinical endpoints using multilocus genotype data.[15]

As with any type of statistical method, the effectiveness of MDR is dependent on its proper implementation. Because this method is used more frequently, and was gained acceptance in the study of human disease, it is becoming increasingly important that the implementation of the MDR approach is properly understood. Although the details of the software package have been published,[9,16] there are few resources available to guide a user through the details of the method itself. Concerns regarding dataset structure (including sample size, balance of cases and controls and structure of family data) must be considered before using MDR. Subsequently, issues involving configuration parameters can affect the results of analysis and must be carefully considered (such as threshold values and cross-validation parameters). Performing hypothesis testing on an MDR model requires permutation testing. The proper execution of permutation testing in reference to a particular dataset and configuration is essential to the method's effectiveness.

Method overview

The details of the MDR method have been published previously.[9,14,16] Briefly, MDR is described here and is shown in Figure 1. In step one, the dataset is divided into multiple partitions for cross-validation. MDR can be performed without cross-validation; however, this is rarely done due to the potential for over-fitting.[17] Cross-validation[18] is an important part of the MDR method, because it tries to find a model that not only fits the given data but can also predict on future, unseen data. Since attainment of a second dataset for testing is time-consuming and often cost-prohibitive, cross-validation produces a testing set from the given data to evaluate the predictive ability of the model produced. In the case of ten-fold cross-validation, the training set comprises 90 per cent of the data, whereas the testing set comprises the remaining 10 per cent of the data.

In step two, a set of n genetic and/or environmental factors are selected. The n factors and their possible multifactor classes are represented in n-dimensional space; for example, for two loci with three genotypes each, there are nine possible two-locus-genotype combinations. Then, the ratio of the number of cases to the number of controls is calculated within each multifactor class. Each multifactor class in n-dimensional space is then labelled as 'high risk' if the cases to controls ratio meets or exceeds some threshold (eg ≥ 1), or as 'low risk' if that threshold is not exceeded, thus reducing the n-dimensional space to one dimension with two levels ('low risk' and 'high risk'). Among all of the two-factor combinations, a single model that has the fewest misclassified individuals is selected. This two-locus model will have the minimum classification error among the two-locus models. In order to evaluate the predictive ability of the model, prediction error is estimated using the testing set. Mathematically, the calculation of prediction error and classification error is the same, but the portion of the dataset used to calculate the metric is different. Classification error is calculated on the training set, whereas prediction error is calculated on the testing set. Both metrics measure the number of individuals whose clinical endpoint has been incorrectly specified by the MDR model.

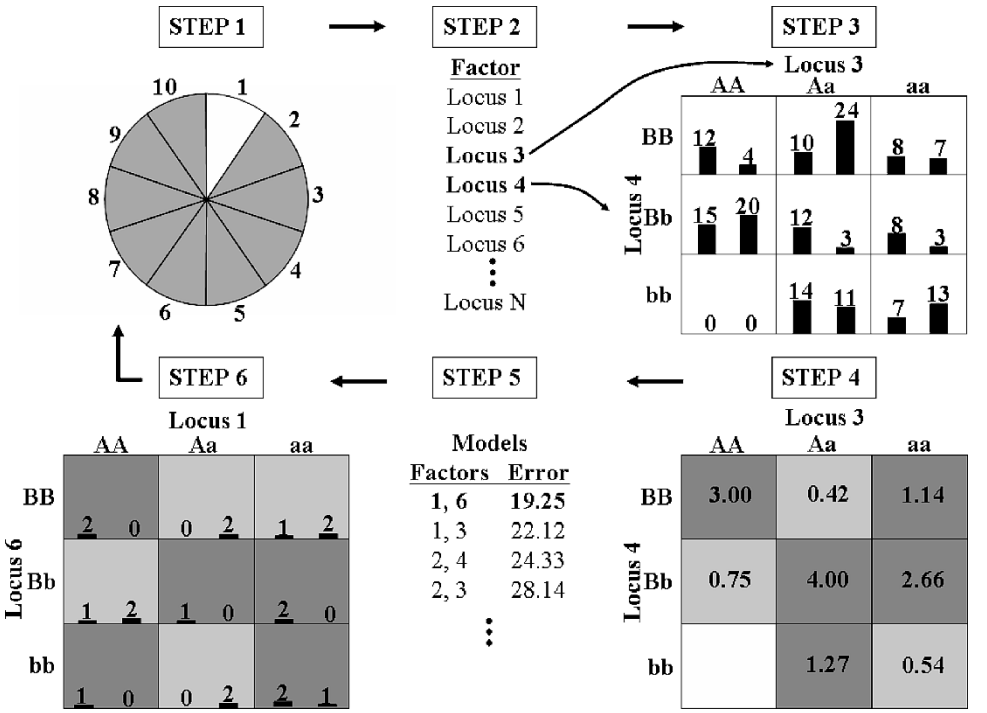

Figure 1. Summary of the general steps to implement the MDR method (adapted from Ritchie *et al.*[9]) In step one, the data are divided into a training set and an independent testing set for cross-validation. In step two, a set of *n* factors is then selected from the pool of all factors. In step three, the *n* factors and their possible multifactor cells are represented in *n*-dimensional space. In step four, each multifactor cell in the *n*-dimensional space is labelled as high risk if the ratio of affected individuals to unaffected individuals exceeds a threshold of one, 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 one through to six 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, whereas light-shaded cells represent low-risk genotype combinations. White cells represent genotype combinations for which no data were observed.

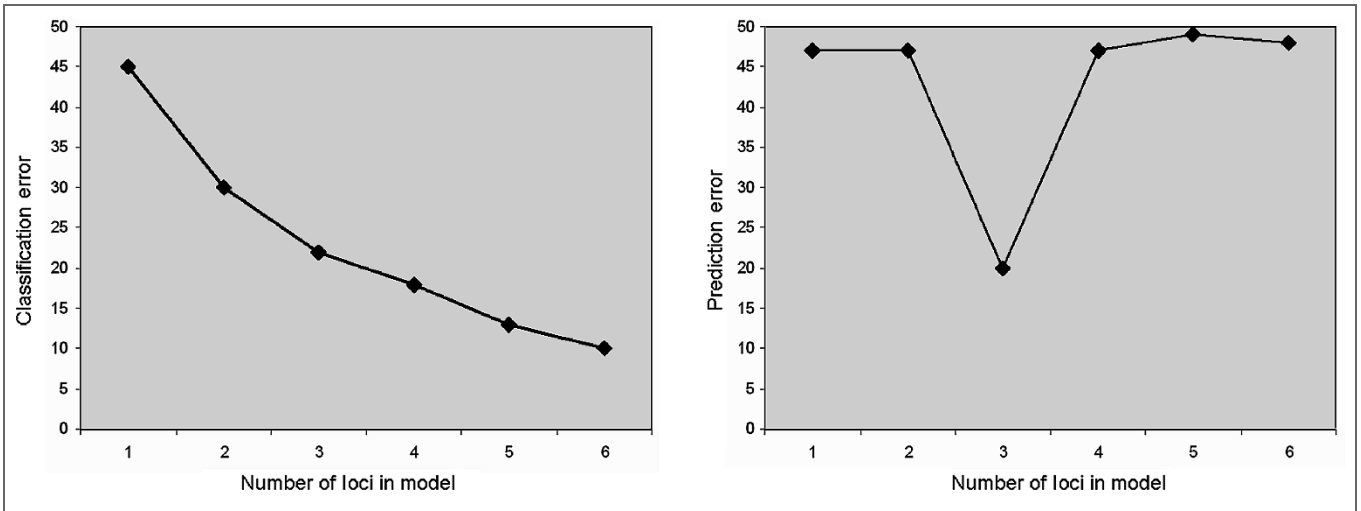

Figure 2. Example of trend of classification error(2A) and prediction error(2B) when the number of loci in a model increases. The classification error continues to get smaller and smaller, which indicates over-fitting. The prediction error will average around 50 per cent and will drop for the best model.

For studies with more than two factors, the steps of the MDR method are repeated for each possible model size (two-factor, three-factor etc), if computationally feasible. The result is a set of models, one for each model size considered. From this set, the model with the combination of loci and/or discrete environmental factors that maximises the cross-validation consistency and minimises the prediction error is selected. Cross-validation consistency is a measure of the number of times an MDR model is identified in each possible 90 per cent of the subjects.[9] When cross-validation consistency is maximal for one model and prediction error is minimal for another model, statistical parsimony is used to choose the best model. In model selection, it is crucial that prediction error, and not classification error, be used. This is due to over-fitting observed with classification error. As the number of loci evaluated increases, the classification error will always decrease. This phenomenon is shown in Figures 2A and 2B.

Hypothesis testing of this final best model can then be performed by evaluating the magnitude of the prediction error through permutation testing. Permutation testing is described in more detail below (Figures 3 and 4).

More recently, less emphasis has been put on choosing a single final model. Significance levels are assigned to each model in the final set using the permutation-testing procedure, then all significant models are reported. This new approach attempts to use all information within the final set of models.

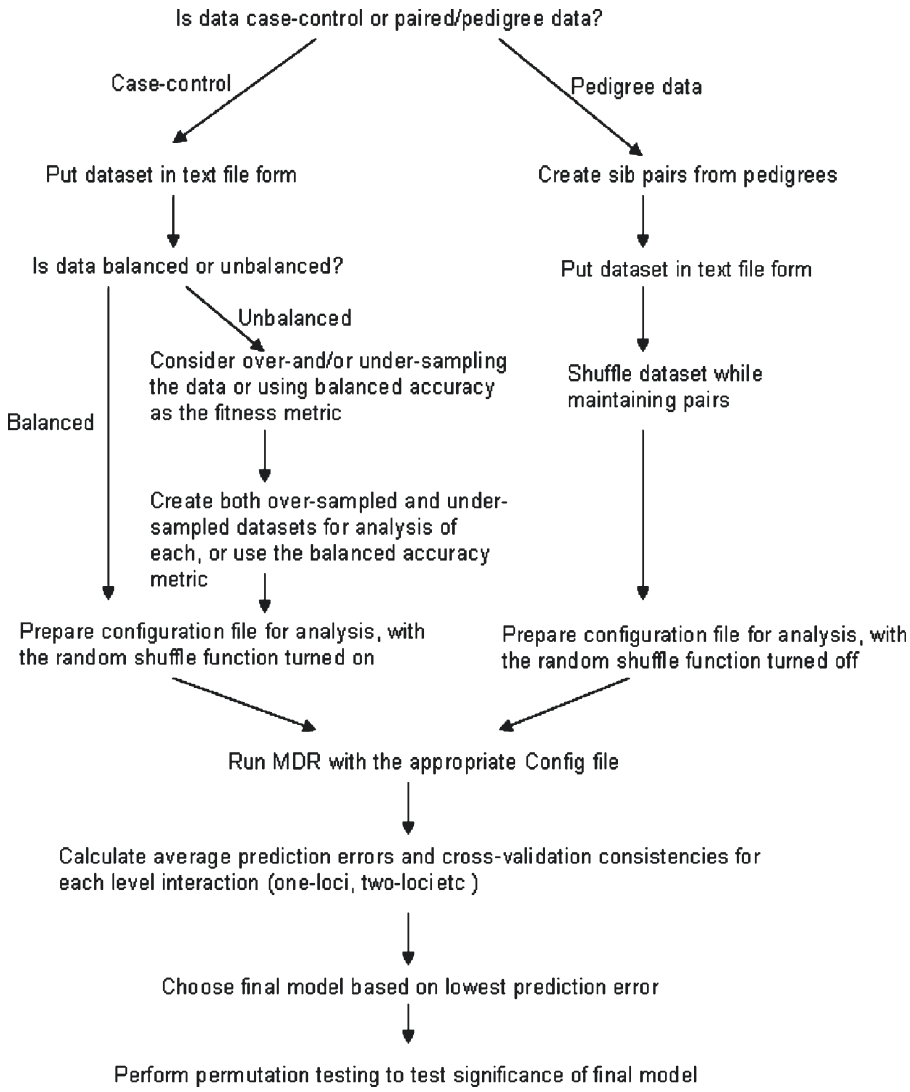

Figure 3. Flow chart of multifactor dimensionality reduction (MDR) procedure. The flow chart outlines the thought process that must be completed for each data analysis with MDR. The steps in analysis vary with different dataset structures and characteristics, and the flowchart guides the user through the decision-making process associated with any particular analysis.

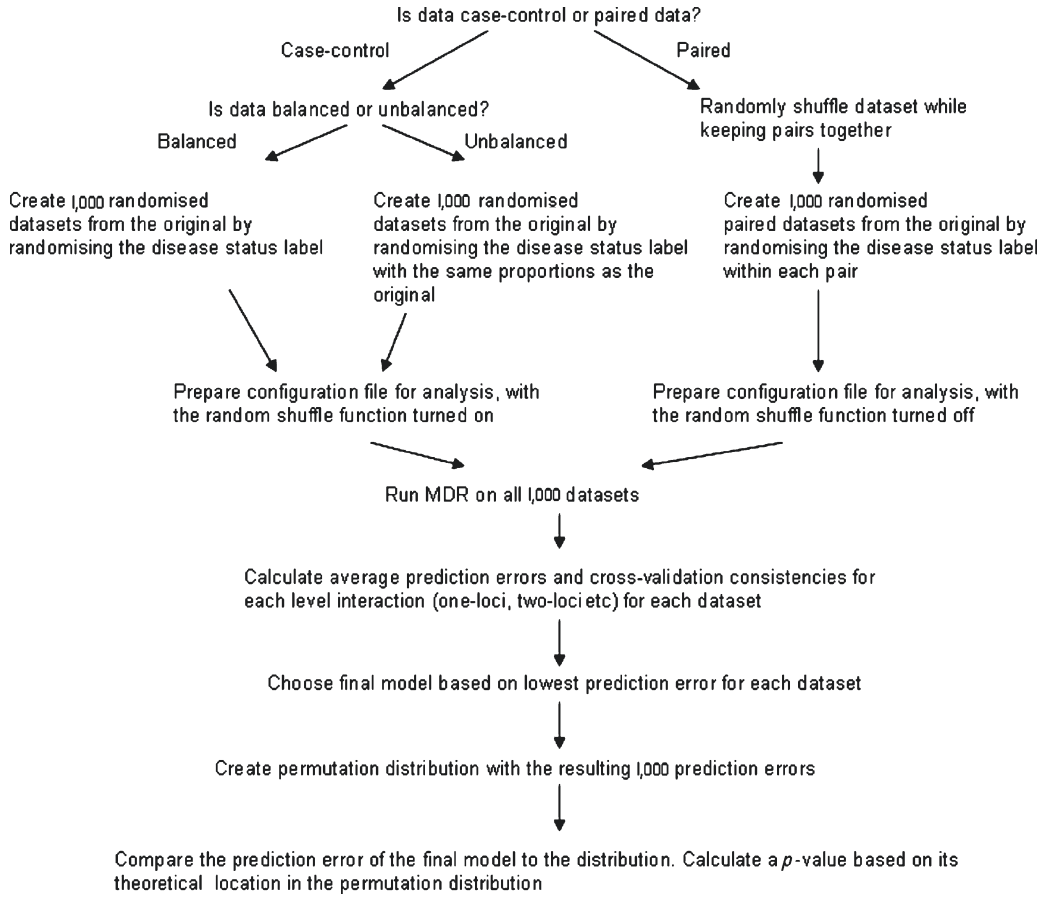

Figure 4. Flow chart of permutation testing procedure. The flow chart carries the user through the process of permutation testing step by step, for any type of dataset analysis.

Because the end goal of the MDR method is hypothesis generation, this approach may be preferred to reduce the risk of false negatives.

Implementation

The original distributed version of MDR was available as Linux, Solaris, or MAC OS command line software. Presently, MDR software is being distributed in a Java software package with a graphical user interface. The most current open-source version is available at www.epistasis.org/mdr.html. MDR has also been incorporated into the Weka-CG software, which is available from the same website. In addition, a 'C' library is under development for users to plug MDR into their own software packages.

Dataset issues

Figure 3 displays the steps considered in the MDR procedure which will be covered in detail in the next four sections. When designing complex genetic and pharmacogenomic studies, the structure and size of a dataset is very important. MDR can be easily applied to case-control and discordant sibling pair study designs with little or no dataset modification. Appropriate datasets for MDR will include any number of genetic and environmental independent variables, along with two distinct clinical endpoints (dependent variables). MDR was originally designed to find interactions in studies of disease risk, but it is applicable to any type of dataset with two outcome levels. Efforts are underway to expand MDR to include more than two endpoints, because this can be done in other contingency table methods, but this is currently a restriction in the MDR software.

For case-control data with unrelated individuals, the order of individuals within the dataset is irrelevant because the data will be randomly shuffled during cross-validation. If the dataset consists of family/sibling data, or population-based matched case-control data, the order of individuals is very important. In such cases, the pairs must be kept together within the dataset during cross-validation splitting. These data should not be randomly shuffled during MDR analysis.

Pedigree data can be more complicated. Currently, pedigrees must be converted to sibling pair data for analysis, and there are several options to handle such datasets.

The first option is to use all possible affected–unaffected pairs from each family. This would allow individuals to be represented multiple times in a dataset. The other option is to consider only one randomly chosen affected–unaffected pair from each pedigree. Currently, such datasets are handled on a case by case basis, and further work is being done to determine the appropriateness of each approach.

One particular type of pedigree data is triad data, where the genotypes of the parents and an affected child are known. In this case, 'pseudo-controls' must be created because this approach will enable evaluation of the genotypes that were transmitted to the affected child in comparison with genotypes that were non-transmitted. This is done using allele data from the two parents to create a new 'child' with the alleles that were not transferred to the real affected child. For example, if the mother had genotype 'Aa' for a particular gene, and the affected child had received the 'A' allele, the pseudo-control would receive the 'a' allele from the mother. This would be done for every gene or SNP from both parents. Sibling pairs would be created from the pseudo-control and affected child for analysis.

Sample size requirements for MDR are not yet known. A total sample size of 400 individuals has been shown to have excellent power to detect two-locus interactions for a specific set of epistasis models simulated in datasets of ten total SNPs.[9] Larger sample sizes are needed for higher-order interactions. There is no theoretical formula for power calculations for MDR, so more thorough empirical estimates for sample size and power are needed. Preliminary simulation studies have demonstrated that datasets smaller than 50 cases and 50 controls show a decrease in power and, in addition, begin to show an upward bias and inflated variance on the prediction error estimates (unpublished data). Currently, more simulation studies are underway to understand the influence of different effect sizes and sample sizes on the power of the MDR method.

If the dataset is not balanced in the number of cases and controls, variations on the MDR configuration parameters may be utilised. When analysing such a dataset, there are several options. First, over-sampling or under-sampling might be considered.[9,19] Over-sampling involves randomly re-sampling the under-represented class of individuals within the dataset until the number of cases and controls are equal. Secondly, under-sampling involves randomly removing members of the over-represented class of individuals from the dataset until it is balanced. There is no particular rule for whether over-sampling or under-sampling is generally preferable. Currently, research is being done with simulated data to understand the implications of over- or under-sampling. Initial observations indicate that either over- or under-sampling is preferred over analysing an unbalanced dataset with a greater than 2:1 ratio of cases to controls or vice versa (manuscript in preparation). In many datasets, convergence of results following over- and under-sampling demonstrates a strong signal. If results vary widely among the sampling datasets, it may indicate a weak signal within the dataset (manuscript in preparation). There are risks associated with using over- or under-sampling techniques. Over-sampling can introduce false associations due to the particular samples that were over-sampled. In addition, this can provide a false sense of higher power. Under-sampling is a mechanism by which data are thrown away. Again, this can lead to the introduction of a false association, as a result of which samples are thrown out, or this can reduce power due to a smaller sample size. Thus, although these techniques are used in the literature,[19,20] they can be dangerous.

A potentially more conservative alternative that has been proposed for analysing unbalanced data is adjusting the MDR threshold value. The threshold value defines the ratio that determines the disease risk status assignment to a particular multi-locus genotype combination. Typically, this value is set to 'one'. The idea behind modifying this parameter was to correct for the chance that a multifactor combination could be assigned a classification of 'high risk' or 'low risk' simply because of the numerical dominance of one disease class in the dataset. When the threshold is adjusted, the calculation of classification error will also need to be modified to accommodate the unbalanced data, such as using a balanced accuracy metric. Further research is being conducted to understand fully the implications of adjusting the threshold, as well as addressing other potential solutions for unbalanced data, such as new fitness functions.

MDR configuration parameters

After dataset formatting, the next step is to establish configuration parameters for data analysis.[16] There are several parameters that must be individually established for each new dataset. A random seed (which can be any random number) must be selected for the random shuffle function used for cross-validation. Random shuffling reduces the risk of biasing cross-validation due to non-random ordering of data. This same random seed should be used in permutation testing, which will be discussed below. The next parameter is the number of loci considered. This describes the number of factors considered in each interaction model. For example, if 'loci considered' is set from '2–5', MDR will test for all two-factor, three-factor etc, up to five-factor interactions.

Currently, when dealing with missing data, MDR includes this missing data as an additional genotype level. This is not problematic when there is a small amount of missing data. If there is a large percentage of data missing, however, it can overwhelm the solutions, and MDR can model the missing data more so than the genotype data. Thus, caution should be used when a dataset has a large amount of missing data. Instead, one can use data imputation techniques in the data

manipulation module of the Java MDR software. Alternative missing data solutions are currently being investigated.

The final configuration issue to consider is the cross-validation parameter. There are multiple types of cross-validation, each with its own advantages and disadvantages, from 'leave one out cross-validation' (LOOCV) to 'N'-fold cross-validation to no cross-validation.[18] LOOCV is where only one individual is left out of the training group for model validation. 'N'-fold cross-validation involves partitioning the data into 'N' groups, where one group is used for testing and the remaining groups are used in training. For MDR, ten-fold cross-validation has traditionally been used. Even though this technique is computationally intensive and the estimate of the prediction error may be biased, its smaller variance makes it well suited to the end goal of MDR, which is hypothesis generation. In the original version of MDR (and original MDR paper),[9] the dataset had to be perfectly divisible by the cross-validation interval, typically ten. This often meant that a few individuals had to be thrown out of a dataset. Current versions of MDR do not have this restriction. Now, the dataset is divided into partitions as evenly as possible, without losing any data. Current simulations are underway to explore different types of cross-validation for evaluating power, type I error, bias and variance. Regardless of the type of cross-validation selected, it is recommended that cross-validation be used because it has been shown to be so important to prevent over-fitting.[17]

Performing MDR analysis

Using the MDR software is very straightforward after all decisions regarding configuration parameters have been made. There are a few issues that influence computation time: the number of factors considered for a model (the dimension of interaction), the number of individuals in a dataset, the number of factors/variables considered for each individual and the number of cross-validation intervals. These variables increase computation time exponentially due to the combinatorial aspect of the algorithm.

Current versions of MDR are constrained by the parameters discussed in the previous section, but work is in progress to expand MDR to more diverse datasets. One current development is to expand MDR to analyse data with more than two clinical endpoints, such as 'unaffected', 'mildly affected' and 'strongly affected'. The immediate relevance of such an extension could easily be seen in studies of many common medical conditions with multiple phenotypes, such as diabetes, blood pressure, etc. As mentioned earlier, this modification should not be too difficult because MDR is a contingency table method, which is a type of method often used for ordinal data.

Additionally, work is being done to expand the capability of MDR to capitalise further on pedigree data. MDR–PDT has been developed to merge the MDR algorithm with the pedigree disequilibrium test (PDT).[21] PDT was developed as a test for linkage disequilibrium. This merger will allow the application of MDR to complex pedigree data in the presence of family structure.

For large datasets with many individuals and/or loci, the combinatorial explosion involved in an exhaustive search of all multifactorial combinations exponentially extends computation time. Typically, datasets are analysed out to four- or five-way interactions. Power studies with moderately sized datasets indicate that MDR has excellent power to detect interactions up to this level, but power to detect higher-order interactions decreases. Also, the computation time required for analysis beyond this point becomes prohibitive. To try to resolve these issues and enable analysis of much higher-order interactions and much larger datasets, a parallel programming implementation of MDR is in development. Utilising parallel programming and parallel supercomputing technologies will allow analysis of larger datasets and higher-order interactions in reasonable time frames.

Permutation testing

Once a final MDR model or set of models has been chosen, permutation testing can be used to perform a hypothesis test and evaluate its statistical significance. The theory behind permutation testing is to create an empirical distribution of prediction errors that could be expected simply by chance. This distribution must be created for each individual dataset, mimicking the configuration parameters and dataset characteristics of the original MDR analysis.[22]

Permutation testing has similar considerations as a typical MDR analysis. If the dataset has a balanced ratio of cases and controls, the ratio of cases and controls in the randomised datasets should also be balanced. When analysing unbalanced data, the randomised datasets must reflect the same proportions of cases and controls. In addition, all configuration parameters used in the original analysis should be identical in permutation testing. This is to ensure that the permutation test mimics the original analysis, except for the random disease status label.

Once the randomised data sets are created and configuration issues are considered, an MDR analysis is performed on all randomised datasets. After the analysis of each dataset, the best model is extracted using the same criterion that was used for the original analysis. The prediction errors of the single best model from each analysis comprise the empirical distribution. The prediction errors within the distribution are sorted in ascending order because the lower the error, the better the model. Once the distribution is created, the final model from the original run can be evaluated. The location in the empirical distribution where the original error would fall directly translates into the *p*-value of the analysis. This omnibus permutation test may be a conservative method, but it is more likely to control for type I error, while not limiting power. As mentioned earlier, the primary goal of MDR is

hypothesis generation for future studies; however, one often wants some measure of how likely it is that the model or set of models detected by MDR would arise by chance. Permutation testing allows for the evaluation of statistical significance of one or a few MDR models.

Case series

The importance of the correct MDR implementation can be illustrated using a simulated dataset. SNP data were simulated, containing a three-locus gene–gene interaction model with no main effects, as described by Moore *et al.*[23] The epistasis model can be shown as a multilocus penetrance function, where the table values indicate the probability of disease, given a specific multilocus genotype combination $p(\mathrm{D}|\mathrm{AABBCC} = 0.07)$. This particular dataset included allele frequencies of 0.2 and 0.8 and a heritability of 1.5 per cent. The effect in the dataset was simulated using a three-locus interaction model between loci 3, 5 and 10. The model is shown in Table 1.

The dataset is unbalanced, with 200 controls and 50 cases. First, the dataset was run without any considerations for its unbalanced nature: without any manipulation of the data (ie no over- or under-sampling), without changing the threshold (leaving it at 1.0), and following all previously mentioned configuration recommendations. Single-locus through to five-locus interactions were considered in the analysis. The resulting best models for each level of interaction are listed in Table 2A. Based on the lowest prediction error and highest cross-validation consistency, the single-locus model would be chosen as the final model. The correct three-locus model was identified, but not chosen as the final best model due to the over-representation of controls within the dataset, skewing the assignment of disease risk status for each multi-locus combination.

To perform permutation testing properly, the randomised datasets must reflect the proportion of cases and controls in the real dataset. Permutation testing was done correctly, reflecting the unbalanced nature of the dataset as well as all configuration parameters used in the original analysis. The permutation distribution showed that the final model revealed by MDR analysis was not statistically significant.

To demonstrate the importance of proper permutation testing, randomised datasets were created for permutation testing without consideration for the unbalanced nature of the original dataset. When permutation testing was done in this manner, the final single-locus model was found to be significant. In fact, all five candidate models (single-locus through to five-locus models) were significant. This demonstrates the challenge presented by unbalanced datasets — disease risk status in each cell can be influenced by the numerical dominance of one affection class, making detection of a true signal difficult.

As mentioned previously, altering the threshold value has been suggested to deal with this challenge. To demonstrate the effect of altering the threshold value, the data were run again, but this time adjusting the threshold to reflect the proportions seen within the data. Because there were 50 cases and 200 controls, the threshold was set to 0.25, instead of 1.0. As discussed earlier, this produces unpredictable results, (also shown in Table 2B). The final model chosen from this run of MDR would be the two-locus model as it has the minimum prediction error and parsimony. However, it does not include even one of the three actual disease loci. Adjusting the threshold gave rise to an even worse performance than was seen in the original MDR run — the correct model was not identified even as the best three-locus model. Proper permutation testing, using the adjusted threshold value, revealed that the final model was not statistically significant. Using an alternative fitness metric to accommodate the unbalanced nature of the data, however, can improve this procedure. Balanced accuracy (or 1-balanced classification error) takes into account the ratio of cases to controls in the dataset. This metric is calculated by the equation [1 − ((sensitivity + specificity)/2)]. The results of the MDR analysis using a threshold of 0.25 and balanced accuracy as the fitness metric are shown in Table 2C. Here, the best model is the three-locus

Table 1. Three-locus penetrance table where values in bold indicate genotype frequencies and table values indicate penetrance. Penetrance is probability of disease given a particular genotype combination.

$h^2 = 0.03$		CC			Cc			cc		
		0.64			**0.32**			**0.04**		
		BB	**Bb**	**bb**	**BB**	**Bb**	**bb**	**BB**	**Bb**	**bb**
		0.64	**0.32**	**0.04**	**0.64**	**0.32**	**0.04**	**0.64**	**0.32**	**0.04**
AA	**0.64**	0.07	0.02	0.01	0.00	0.08	0.07	0.04	0.02	0.00
Aa	**0.32**	0.00	0.07	0.06	0.09	0.03	0.08	0.06	0.07	0.01
aa	**0.04**	0.05	0.01	0.08	0.06	0.01	0.10	0.10	0.02	0.05

Table 2 (A). Original MDR analysis of unbalanced data.

Number of loci in model	Best candidate model	Average cross-validation consistency	Average prediction error
1	1	10	20.00%
2	7, 8	7	21.28%
3	3, 5, 10	5	22.90%
4	2, 3, 5, 10	4	26.60%
5	1, 4, 7, 8, 9	4	29.66%
Final model: locus 1 — INCORRECT MODEL			

(B). Alnalysis using threshold adjustment only.

Number of loci in model	Best candidate model	Average cross-validation consistency	Average prediction error
1	5	5	40.80%
2	4, 9	3	30.53%
3	4, 5, 7	7	31.74%
4	3, 5, 8, 10	4	36.60%
5	1, 2, 4, 5, 9	5	40.00%
Final model: loci 4 and 9 — INCORRECT MODEL			

(C). Analysis using threshold adjustment and balanced accuracy.

Number of loci in model	Best candidate model	Average cross-validation consistency	Average prediction error
1	8	6	49.50%
2	10, 5	4	45.75%
3	10, 5, 3	10	29.97%
4	10, 5, 3, 2	7	33.69%
5	10, 8, 5, 4, 3	6	41.86%
Final model: loci 3, 5, 10 — CORRECT MODEL			

model because it has both the minimum prediction error and the maximum cross-validation consistency. Thus, this combination of fitness metric and adjusted threshold successfully identifies the correct model.

As recommended earlier, in situations with an unbalanced dataset both over- and under-sampling can be evaluated. This was done for the present example to demonstrate the effectiveness of this approach. To generate the over-sampled dataset, the 50 cases were randomly re-sampled with replacement until there were 200 cases balanced with 200 controls. MDR was run on this newly modified dataset with a threshold of 1.0, following all other configuration considerations recommended earlier. The results are shown in Table 2D. In this case, the correct three-locus model was successfully identified as both the best three-locus model and as the final model. Permutation testing (properly reflecting the over-sampled dataset and proper configuration parameters) revealed that the three-locus final model was statistically significant.

Simultaneously, under-sampling was also performed — resulting in a dataset of 50 cases and 50 controls (randomly

(D). Analysis of over-sampled data.

Number of loci in model	Best candidate model	Average cross-validation consistency	Average prediction error
1	8	9	46.25%
2	5, 10	3	46.75%
3	3, 5, 10	10	25.84%
4	3, 5, 8, 10	7	29.83%
5	3, 5, 6, 8, 10	5	25.88%
Final model: loci 3, 5, 10 — CORRECT MODEL			

selected from the original 200). This dataset was run and the results outlined in Table 2E. Again, the correct three-locus model was identified as the final model produced by MDR. Proper permutation testing revealed that the final model was again statistically significant.

By comparing the results from the two modified datasets, it becomes apparent that the three-locus model identified by each is the correct final model. The convergence of results from both analyses indicates a strong signal within the dataset. Neither the unbalanced dataset nor the adjusted threshold value without adjusting the fitness metric analysis was able to identify the correct model.

Another possible error in MDR implementation is the use of misclassification error for model selection, instead of prediction error. The results of this analysis are listed in Table 2F. The correct model was identified as the best three-locus model, but with no previous knowledge, the five-locus model would be chosen as the final model based on the lowest error. As mentioned previously, the use of a misclassification error for model selection results in

(E). Analysis of under-sampled data.

Number of loci in model	Best candidate model	Average cross-validation consistency	Average prediction error
1	6	5	60.00%
2	5, 10	10	32.56%
3	3, 5, 10	10	28.00%
4	3, 5, 7, 10	7	32.53%
5	3, 4, 5, 6, 10	4	39.95%
Final model: loci 3, 5, 10 — CORRECT MODEL			

(F). Analysis using misclassification error for model selection.

Number of loci in model	Best candidate model	Average classification error
1	1	20.00%
2	7, 8	18.79%
3	3, 5, 10	18.00%
4	2, 3, 5, 10	16.80%
5	1, 4, 7, 8, 9	15.60%
Final model: loci 1, 4, 7, 8, 9 — INCORRECT MODEL		

over-fitting of the data. As the number of loci in the candidate model increases, the misclassification error always decreases, as shown in this analysis.

This sample dataset demonstrates the importance of proper implementation of the MDR method. A simulated dataset was used so that the correct model was known and the deleterious effect of improper implementation could be readily apparent. These phenomena are also observed during the analysis of real data.

Conclusions

MDR is a novel and powerful statistical tool for detecting and modelling epistasis in the study of human disease and pharmacogenomics. In making this method more available and acceptable in the scientific community, it is important that the guidelines for use are well understood.

These guidelines must also be understood when comparing MDR with other, more traditional methods such as logistic regression or classification and regression trees. To evaluate multiple methods accurately, the parameters defined for each method must be comparable. The range of loci interactions considered must be identical, along with cross-validation splits and permutation parameters.

Building on the success that MDR has already had, many of the performance features of the method are currently being studied. More extensive power studies are being performed to estimate the power of MDR in datasets with different sample sizes, effect sizes, number of factors and noise level attached to the true model. Additionally, other levels of N-fold cross-validation are being explored for their influence on power and computation time. Understanding the problems that can arise from over- and under-sampling, new fitness metrics are being explored to handle the problem of unbalanced data. The dissection of all performance features of MDR is a priority of future research.

The detection, characterisation and interpretation of gene–gene and gene–environment interactions are expected to improve the diagnosis, prevention and treatment of common human diseases. MDR can be a powerful tool in reaching these goals when used appropriately.

References

1. Moore, J.H. (2003), 'The ubiquitous nature of epistasis in determining susceptibility to common human diseases', *Hum. Hered.* Vol. 56, pp. 73–82.
2. Sing, C.F., Stengard, J.H. and Kardia, S.L. (2004), 'Dynamic relationships between the genome and exposures to environments as causes of common human diseases', *World Rev. Nutr. Diet.* Vol. 93, pp. 77–91.
3. Thornton-Wells, T.A., Moore, J.H. and Haines, J.L. (2004), 'Genetics, statistics and human disease: Analytical retooling for complexity', *Trends Genet.* Vol. 20, pp. 640–647.
4. Wilke, R.A., Reif, D.M. and Moore, J.H. (2005), 'Combinatorial pharmacogenetics', *Nat. Rev. Drug Discov.* Vol. 4, pp. 911–918.
5. Bellman, R. (1961), 'Adaptive Control Processes', Princeton University Press, Princeton, WJ.
6. Peduzzi, P., Concato, J., Kemper, E. *et al.* (1996), 'A simulation study of the number of events per variable in logistic regression analysis', *J. Clin. Epidemiol.* Vol. 49, pp. 1373–1379.
7. Moore, J.H. and Williams, S.M. (2002), 'New strategies for identifying gene–gene interactions in hypertension', *Ann. Med.* Vol. 34, pp. 88–95.
8. Moore, J.H. (2004), 'Computational analysis of gene–gene interactions using multifactor dimensionality reduction', *Expert Rev. Mol. Diagn.* Vol. 4, pp. 795–803.
9. Ritchie, M.D., Hahn, L.W., Roodi, N. *et al.* (2001), 'Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer', *Am. J. Hum. Genet.* Vol. 69, pp. 138–147.
10. Cho, Y.M., Ritchie, M.D., Moore, J.H. *et al.* (2004), 'Multifactor-dimensionality reduction shows a two-locus interaction associated with type 2 diabetes mellitus', *Diabetologia* Vol. 47, pp. 549–554.
11. Tsai, C.T., Lai, L.P., Lin, J.L. *et al.* (2004), 'Renin–angiotensin system gene polymorphisms and atrial fibrillation', *Circulation* Vol. 109, pp. 1640–1646.
12. Soares, M.L., Coelho, T., Sousa, A. *et al.* (2005), 'Susceptibility and modifier genes in Portuguese transthyretin V30M amyloid polyneuropathy: Complexity in a single-gene disease', *Hum. Mol. Genet.* Vol. 14, pp. 543–553.
13. Bastone, L., Reilly, M., Rader, D.J. and Foulkes, A.S. (2004), 'MDR and PRP: A comparison of methods for high-order genotype-phenotype associations', *Hum. Hered.* Vol. 58, pp. 82–92.
14. 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, missing data, phenocopy, and genetic heterogeneity', *Genet. Epidemiol.* Vol. 24, pp. 150–157.
15. Hahn, L.W. and Moore, J.H. (2004), 'Ideal discrimination of discrete clinical endpoints using multilocus genotypes', *In Silico Biol.* Vol. 4, pp. 183–194.
16. Hahn, L.W., Ritchie, M.D. and Moore, J.H. (2003), 'Multifactor dimensionality reduction software for detecting gene–gene and gene–environment interactions', *Bioinformatics* Vol. 19, pp. 376–382.
17. Coffey, C.S., Hebert, P.R., Ritchie, M.D. *et al.* (2004), 'An application of conditional logistic regression and multifactor dimensionality reduction for detecting gene–gene interactions on risk of myocardial infarction: The importance of model validation', *BMC Bioinformatics* Vol. 5, p. 49.
18. Hastie, T., Tibshirani, R. and Friedman, J.H. (2001), 'The Elements of Statistical Learning', Springer Verlag, Basel, Switzerland.
19. Weiss, G.M. and Provost, F. (2003), 'Learning when training data are costly: The effect of class distribution on tree induction', *J. Artif. Intell. Res.* Vol. 19, pp. 315–354.
20. Japkowicz, N. and Stephen, S. (2002), 'The class imbalance problem: A systematic study', *Intell. Data Anal. J.* Vol. 6, pp. 429–450.
21. Martin, E.R., Ritchie, M.D., Hahn, L.W. *et al.*, (2006), 'A novel method to identify gene–gene effects in nuclear families: The MDR-PDT', *Genet. Epidemiol.* Vol. 30, pp. 111–123.
22. Good, P. (2000), 'Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses', Springer-Verlag, New York, NY.
23. Moore, J., Hahn, L., Ritchie, M. *et al.* (2002), Application of Genetic Algorithms to the Discovery of Complex Models for Simulation Studies in Human Genetics', in: Langdon, W., Cantu-Paz, E., Mathias, K. *et al.* (eds.), Morgan Kaufman Publishers, San Francisco, CA, pp. 1150–1155.