# The Use of Multifactor Dimensionality Reduction to Detect Epistasis Among Potential Causal Genes of Alcoholism

by Laura Mustavich

Epistasis, the interaction among genes, is ubiquitous among common, complex, and multifactorial diseases. Therefore it has become necessary to develop methods to detect epistasis, the motivation for one such method, multifactor dimensionality reduction (MDR). We introduce the algorithm of MDR, its strengths and weaknesses, and finally illustrate the results of applying MDR to alcoholism. We compare these results to those from logistic regression, a commonly used alternative to MDR, and discuss methodological issues of MDR, and future work centered around these issues.

**Outline**

## Introduction

*Complex Diseases and Epistasis*

Most common, hereditary diseases are complex; caused by multiple genes, often interacting with one another. This interaction, termed *epistasis*, occurs when an allele at one locus masks the effect of an allele at another locus, and is illustrated with the example of hair color in mice, determined by the two-locus system depicted in the table below[1]:

| | | Genotype at locus G | | |
|---|---|---|---|---|
| | | g/g | g/G | G/G |
| Genotype at locus B | b/b | *White* | *Grey* | *Grey* |
| | b/B | *Black* | *Grey* | *Grey* |
| | B/B | *Black* | *Grey* | *Grey* |

As you can see from the table, considering locus B individually (column 3 of the table), allele B is dominant to allele b, since it confers *Black* hair color, even if only one copy of the B allele is present. Similarly, allele G is dominant to allele g at locus G, since it confers *Grey* hair color, even if only one copy of the G allele is present (row 3 of the table). If you consider both loci simultaneously, however, you can see that all mice who possess at least one copy of the G allele at locus G are *Grey*, despite their genotype at locus B. Allele B of locus B, is then said to be epistatic to allele G at locus G, and thus do not act independently of one another.

*Failure of Traditional Approaches*

Although traditional gene-hunting approaches, such as linkage and association analyses, have been very successful in discovering the genes responsible for rare Mendelian diseases, caused by a single gene, they have proven unsuccessful in determining the causal gene networks of complex, multifactorial diseases. This is largely due to the fact that, since so many genes interact to cause complex diseases, the effect of any single individually is so negligible that it is difficult to detect, for traditional methods were not designed to take interactions into account, but rather to detect strong, single-effects. As the emphasis in human genetics has shifted away from rare Mendelian disorders, to common complex diseases such as cancer, cardiovascular, metabolic, and psychiatric diseases, it has become apparent that we must develop methods specifically aimed at targeting the epistasis which is so ubiquitous among genes causing these diseases, if we are to uncover their genetic etiology, and ultimately, cure them.

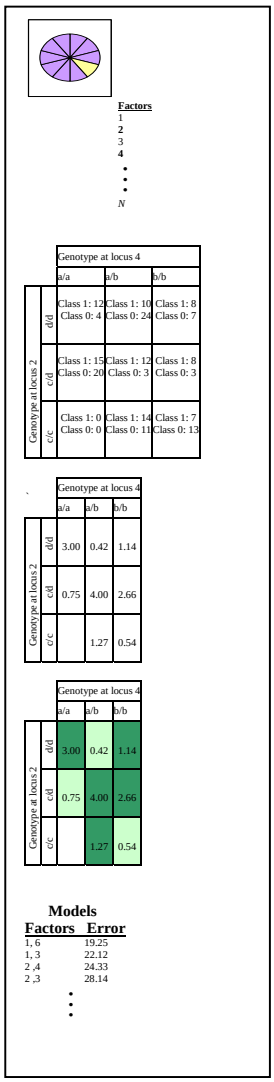
## Multifactor Dimensionality Reduction

*Algorithm*

Multifactor Dimensionality Reduction (MDR) is a data mining approach, developed by Marylyn D. Ritchie and colleagues from Vanderbilt University, to identify interactions among discrete variables that influence a binary outcome. It is a non-parametric alternative to traditional statistical methods such as logistic regression. While it was driven by the need to improve the power to detect gene-gene interactions, it can be applied to any set of discrete variables which may predict class. It is used to determine the optimal *k*th order model (the interaction of the set

of *k* variables that best predict the class) among the *N* possible variables. The algorithm is as follows[3]:

0) First, divide the records into 10 distinct subsets for later cross-validation.
1) Select a set of *k* variables from the total set of *N* variables.
2) Form a *k*-dimensional contingency table, where $l_i$ is the length of the *i*th dimension, and is the number of levels of the *i*th variable, where $i = 1, 2, \ldots, k$. (eg. for a second order model with 3 levels for one variable, and 4 levels for another, this process would form a 3 x 4 contingency table.) In each cell of the table (with the corresponding values for each of the *k* variables), record the number of records in each class (cases or controls) in the training set.
3) Calculate the ratio of the two classes in each cell
4) Label each cell as "high-risk" or "low-risk", according to whether the case-control ratio is above a pre-specified threshold. (This is the dimensionality reduction step, since it reduces *k*-dimensional space to 1 dimension with 2 levels.)
5) Use these labels to classify individuals as cases or controls in the testing set, and calculate the misclassification rate.
6) Repeat steps 1 – 5 across all training and testing sets.
7) Repeat steps 1 – 6 for all possible sets of *k* variables. There are $\binom{N}{k}$ of them.
8) Repeat steps 1-7 for any desired value of *k* for $k = 1, 2, \ldots, N$.




The best model is the one which minimizes the prediction error; the average misclassification rate across all 10 cross-validation subsets, and which maximizes the cross-validation consistency; the number of times a particular model had the lowest prediction error across cross-validation subsets. The significance of both these estimates can be assessed by permutation testing[3].

*Strengths*

In terms of multifactorial diseases, the original area of application for MDR, the main strength of the algorithm is that it facilitates the simultaneous detection and characterization of multiple genetic loci associated with a clinical trait by reducing the dimensionality of the multilocus data. MDR not only identifies genes which interact epistatically to cause a disease, but also indicates which variants of each gene interact to cause the clinical outcome. MDR is also non-parametric, since no parameters are estimated, which eliminates the uncertainty introduced by the parameter estimates of parametric methods, such as logistic regression. Furthermore, it assumes no particular genetic model, which is extremely useful when there is no *a priori* knowledge of the genetic system. Lastly, the false-positive rate, which is often a problem in traditional gene-hunting approaches, is minimized due to multiple testing[4].

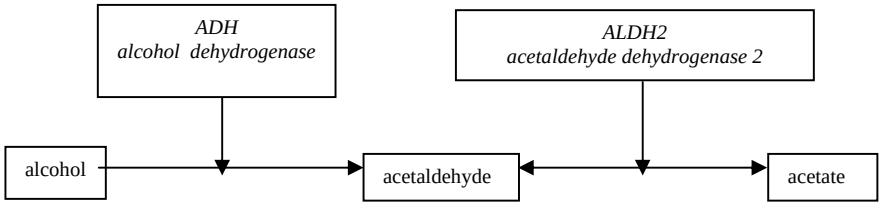
*Weaknesses*

One major weakness of MDR, however, is that, like many methods, MDR suffers from the curse of dimensionality. Predictive ability decreases as the dimensionality increases due to a decreasing number of samples for each combination of attributes. No search method is explicitly incorporated into the algorithm, but instead, an exhaustive search is implied and used in the implementation, rendering the computational intensity of MDR its most serious downfall. This computational intensity is not largely affected by sample size, but is largely a function of the number of attributes, *N*, and the order of the model, *k*. For biallelic loci, there are 3 possible genotypes for each locus, and $\binom{N}{k}$ possible *k*th order models, totalling $\binom{N}{k}3^k$ comparisons. The computational complexity is therefore tremendous when there are greater than 10 loci[4].

**Application to Alcoholism**

*Background*

Alcoholism is a complex disease, with which many genes have been found to be associated, including several classes of *ADH* (*alcohol dehydrogenase*), as well as *ALDH2* (*acetaldehyde dehodrogenase 2*). These genes encode enzymes of the same name which catalyze the steps from alcohol to acetaldehyde, and from acetaldehyde to acetate, respectively, in the metabolism of alcohol.




The gene *TAS2R38*, on the other hand, encodes a bitter taste receptor which confers the ability to taste the compound phenylthiocarbamide (PTC), which tastes bitter to those that can taste it. The ability to taste PTC is correlated with one's willingness to drink alcohol: non-tasters of PTC tend to perceive alcohol as sweet, and therefore tend to drink more alcohol, while tasters of PTC tend to perceive alcohol as bitter, and therefore tend to drink less alcohol. Since one must be willing to drink alcohol in order to become an alcoholic, the variant of *TAS2R38* is related to risk of alcoholism, however, no direct link has been found.

We hope to establish a genetic association between *TAS2R38* and alcoholism, by showing that the risk of alcoholism can be predicted by the form of *TAS2R38*, genes already found to be associated with alcoholism, and other taste receptor genes.

*Results*

In order to do this, I used Multifactor Dimensionality Reduction software[2], an implementation of the MDR algorithm, found on www.sourceforge.net, to analyze a sample of cases (alcoholics) and controls (non-alcoholics) from three East Asian populations: the Ami,

Atayal, and Taiwanese. The 120 individuals were genotyped for 98 markers (single nucleotide polymorphisms) within several genes: *ALDH2*, all *ADH* genes, and 2 taste receptor genes, *TAS2R16* and *TAS2R38* (PTC). Due to extensive computation time, I was forced to restrict the number of markers, and was advised to use markers solely within the *ADH1C* gene, and the 2 taste receptor genes, leaving me with 36 attributes, and considered models only up to order 4. After dropping incomplete records, I was left with 79 individuals.

I initially ran MDR on all three populations combined. The table below shows the results of the best model for each order model, from 1 to 4, along with its classification accuracy, cross-validation consistency, and significance, as evaluated by the sign test.

| Order | Model | Training Bal. Acc. | Testing Bal. Acc. | Sign Test (p) | CV Consistency |
|---|---|---|---|---|---|
| 1 | X.04..ADH1C.dwstrm.Te | 0.6049 | 0.4278 | 0 (1.0000) | 5/10 |
| 2 | X.07..TAS2R16.C_11431<br>X.04..ADH1C.dwstrm.Te | 0.7076 | 0.4438 | 3 (0.9453) | 6/10 |
| 3 | X.07..TAS2R16.C_11431<br>X.04..ADH1C.dwstrm.Te<br>X.04..ADH1C.rs3762896 | 0.785 | 0.3186 | 1 (0.9990) | 4/10 |
| 4 | X.07..TAS2R16.C_11431<br>X.07..PTC.C_8876291_1<br>X.07..PTC.C_8876482_1<br>X.04..ADH1C.dwstrm.Te | 0.8453 | 0.3564 | 2 (0.9893) | 6/10 |

As you can see, none of the models were significant, as they performed worse in terms of testing accuracy, than if the individuals were classified randomly. Thus, I did not explore these models further.

Instead, I considered the populations separately, in the event that my poor results above were due to population admixture. I obtained similar results for the Atayal and Taiwanese, but more interesting results for the 30 Ami individuals, depicted below:

| Order | Model | Training Bal. Acc. | Testing Bal. Acc. | Sign Test (p) | CV Consistency |
|---|---|---|---|---|---|
| 1 | X.07..TAS2R16.C_11431 | 0.7331 | 0.4598 | 5 (0.6230) | 5/10 |
| 2 | X.07..TAS2R16.C_11431<br>X.04..ADH1C.C_2688508 | 0.8284 | 0.3476 | 2 (0.9893) | 3/10 |
| 3 | X.07..TAS2R16.C_11431<br>X.07..PTC.C_8876467_1<br>X.04..ADH1C.C_2688508 | 0.9688 | 0.9545 | 10 (0.0010) | 10/10 |
| 4 | X.07..TAS2R16.C_11431<br>X.07..TAS2R16.C_11431.1<br>X.07..PTC.C_8876467_1<br>X.04..ADH1C.C_2688508 | 0.9722 | 0.8712 | 8 (0.0547) | 9/10 |

While the first and second order models performed quite poorly, the fourth order model was much better, with the third order model indisputably the best, with its 95% accuracy, 10/10 cross-validation consistency, and significance by the sign test. This indicates that *TAS2R16*, *TAS2R38*(PTC), and *ADH1C*, may all interact to effect alcoholism susceptibility, at least in the

Ami population. The positions of the three markers involved in the model, may give us clues to how these genes might be interacting.

*Logistic Regression*

Because MDR has been posited as a much stronger alternative to other methods to detect epistasis, especially logistic regression, I decided to analyze the same data set with logistic regression in order to see if this method would confer similar results. My initial attempt at an exhaustive search algorithm for logistic regression, took way too much computation time in R (much longer than MDR), even for models only up to order 4. Therefore, I incorporated a greedy algorithm to find the best *k*th order model for each $k = 1,2,\ldots,N$, where each model only considered the interaction term of the selected attribute from the previous order model, with one additional attribute. I chose this search method over a genetic programming method since each subsequent order model in the MDR results included the attributes of the previous model. Because this algorithm proved much faster than the exhaustive search used in the MDR software, I was able to find the best model for each possible order. The results are shown below:

| Model.Order | Additional.Attribute | AIC |
|---|---|---|
| 1 | X.07..TAS2R16.C_11431 | 38.68188593 |
| 2 | X.07..TAS2R16.C_11431.1 | 37.59538615 |
| 3 | X.07..PTC.C_8876482_1 | 36.67571859 |
| 4 | X.07..TAS2R16.C_75402 | 36.67571859 |
| 5 | X.07..PTC.C_9506256_1 | 36.67571859 |
| 6 | X.04..ADH1C..EcoRI.In | 36.67571859 |
| 7 | X.04..ADH1C..TATAAA | 36.67571859 |
| 8 | X.04..ADH1C.66bp.InDe | 36.67571859 |
| 9 | X.04..ADH1C.C_2645744 | 36.67571859 |
| 10 | X.04..ADH1C.C_2688487 | 36.67571859 |
| 11 | X.04..ADH1C.C_2688509 | 36.67571859 |
| 12 | X.04..ADH1C.C_2688511 | 36.67571859 |
| 13 | X.04..ADH1C.C_2688547 | 36.67571859 |
| 14 | X.04..ADH1C.dws.InDel | 36.67571859 |
| 15 | X.04..ADH1C.ex6.fn.RF | 36.67571859 |
| 16 | X.04..ADH1C.ex8.fn.RF | 36.67571859 |
| 17 | X.04..ADH1C.new.Ex8.f | 36.67571859 |
| 18 | X.04..ADH1C.rs1789920 | 36.67571859 |
| 19 | X.04..ADH1C.rs1789924 | 36.67571859 |
| 20 | X.04..ADH1C.rs2165671 | 36.67571859 |
| 21 | X.07..TAS2R16.C_32911 | 37.44732682 |
| 22 | X.07..TAS2R16.C_29144 | 37.50110886 |
| 23 | X.04..ADH1C.rs1583977 | 37.17992136 |
| 24 | X.04..ADH1C.rs1042026 | 38.2315879 |
| 25 | X.07..PTC.C_8876467_1 | 38.75680141 |
| 26 | X.04..ADH1C.rs1051643 | 39.58595557 |
| 27 | X.04..ADH1C.rs1001713 | 37.00813606 |
| 28 | X.04..ADH1C.C_2688508 | 37.6020125 |
| 29 | X.04..ADH1C.rs3762896 | 35.88084916 |
| 30 | X.04..ADH1C.rs2646012 | 36.3481602 |
| 31 | X.07..PTC.C_9506826_1 | 35.91206284 |
| 32 | X.04..ADH1C.rs4513578 | 35.15543197 |

| | | |
|---|---|---|
| 33 | X.07..PTC.C_8876291_1 | 35.51478143 |
| 34 | X.04..ADH1C.dwstrm.Te | 36.36629911 |
| 35 | X.04..ADH1C.Ex5.HaeII | 36.96935466 |
| 36 | X.07..PTC.C_9506827_1 | 38.26345394 |

**Discussion**

According to the Akaike Information Criterion (AIC), the third order model is a local optimim, relative to other order models, similar to the MDR results. Many subsequent order models have the same AIC, however, we select the most parsimonious model. Additionally, like the MDR results, the X.07..TAS2R16.C_11431 marker is included in the model, and is also the first selected attribute. Logistic regression and MDR, however, converged upon different attributes for the other two markers in the model. Future simulations must be done to determine which method is most powerful, and gives the most accurate results, in addition to comparisons with other methods.

As you can see from the results, however, the 32$^{nd}$ order model is the best, according to AIC, although I doubt how realistic this model is. With 3 different genes involved, it is improbable that these genes physically interact at 32 different locations among the genes. However, this is something that would have to be verified experimentally. Not surprisingly, all three markers implicated by the MDR results are included in this 32$^{nd}$ order model.

While future work should initially focus on simulation comparisons between MDR and other methods, to determine whether this new method is worth any further attention, subsequent work should focus on appropriate search methods, such as greedy algorithms and genetic programming, which I could not test here due to time constraints, since this is the main weakness of the MDR implementation. While genetic programming search methods have been very successful, and should definitely be compared with greedy algorithms, as applied to MDR, I hypothesize that greedy algorithms will outperform genetic programming methods due to the fact that lower order models are subsumed by higher order models in the MDR results. As MDR still shows promise as a method useful in detecting epistasis, I will surely explore the incorporation of greedy algorithms into MDR in the future, to improve a potentially invaluable method in the world of human disease.

References

1. Cordell, HJ. (2002) Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, 11: 2463-2468.
2. Hahn, LW, MD Ritchie, and JH Moore. (2003) Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics*, 19: 376-382.
3. Moore, JH. (2003) The Ubiquitous Nature of Epistasis in Determining Susceptibility to Common Human Diseases. *Hum Hered*, 56: 73-82.
4. Ritchie, MD, LW Hahn, N Roodi, LR Bailey, WD Dupont, FF Parl, and JH Moore. (2001) Multifactor-Dimensionality Reduction Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer. *Am. J. Hum. Genet.*, 69: 138-147.