

A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility

Jason H. Moore^{a,b,c,d,e,*}, Joshua C. Gilbert^a, Chia-Ti Tsai^f, Fu-Tien Chiang^f, Todd Holden^a, Nate Barney^a, Bill C. White^a

^aComputational Genetics Laboratory, Department of Genetics, Dartmouth–Hitchcock Medical Center, One Medical Center Dr., 706 Rubin Bldg, HB7937, Lebanon, NH 03756, USA

^bDepartment of Community and Family Medicine, Dartmouth Medical School, Lebanon, NH, USA

^cDepartment of Biological Sciences, Dartmouth College, Hanover, NH, USA

^dDepartment of Computer Science, University of New Hampshire, Durham, NH, USA

^eDepartment of Computer Science, University of Vermont, Burlington, VT, USA

^fDivision of Cardiology, Department of Internal Medicine, National Taiwan University Hospital, Taipei, Taiwan

Received 2 August 2005; received in revised form 15 November 2005; accepted 23 November 2005

Available online 2 February 2006

Abstract

Detecting, characterizing, and interpreting gene–gene interactions or epistasis in studies of human disease susceptibility is both a mathematical and a computational challenge. To address this problem, we have previously developed a multifactor dimensionality reduction (MDR) method for collapsing high-dimensional genetic data into a single dimension (i.e. constructive induction) thus permitting interactions to be detected in relatively small sample sizes. In this paper, we describe a comprehensive and flexible framework for detecting and interpreting gene–gene interactions that utilizes advances in information theory for selecting interesting single-nucleotide polymorphisms (SNPs), MDR for constructive induction, machine learning methods for classification, and finally graphical models for interpretation. We illustrate the usefulness of this strategy using artificial datasets simulated from several different two-locus and three-locus epistasis models. We show that the accuracy, sensitivity, specificity, and precision of a naïve Bayes classifier are significantly improved when SNPs are selected based on their information gain (i.e. class entropy removed) and reduced to a single attribute using MDR. We then apply this strategy to detecting, characterizing, and interpreting epistatic models in a genetic study ($n = 500$) of atrial fibrillation and show that both classification and model interpretation are significantly improved.

© 2005 Elsevier Ltd. All rights reserved.

Keywords: Gene–gene interactions; Constructive induction; Multifactor dimensionality reduction; Entropy; Machine learning; Data mining

1. Introduction

A central goal of human genetics is to understand the mapping relationship between DNA sequence variations and susceptibility to disease in order to improve diagnosis, prevention, and treatment. Success in this endeavor will

*Corresponding author. Computational Genetics Laboratory, Department of Genetics, Dartmouth–Hitchcock Medical Center, One Medical Center Dr., 706 Rubin Bldg, HB7937, Lebanon, NH 03756, USA. Tel.: +1 603 653 9939; fax: +1 603 653 9900.

E-mail address: jason.h.moore@dartmouth.edu (J.H. Moore).

depend critically on the degree of nonlinearity in the mapping between genotype to phenotype. Nonlinearities can arise from phenomena such as locus heterogeneity, phenocopy, and the dependence of genotypic effects on environmental factors (i.e. gene–environment interactions or plastic reaction norms) and genotypes at other loci (i.e. gene–gene interactions or *epistasis*). Epistasis has been recognized for many years as deviations from Mendelian segregation ratios (Bateson, 1909) or deviations from additivity in a linear statistical model (Fisher, 1918) and is likely due, in part, to canalization or mechanisms of stabilizing selection that evolve robust gene networks

(Waddington, 1942, 1957; Gibson and Wagner, 2000; Proulx and Phillips, 2005).

Epistasis has been defined in multiple different ways (e.g. Hollander, 1955; Phillips, 1998; Brodie, 2000). We have reviewed two types of epistasis, biological and statistical (Moore and Williams, 2005). Biological epistasis results from physical interactions between biomolecules (e.g. DNA, RNA, proteins, enzymes, etc.) and occurs at the cellular level in an individual. This type of epistasis is what Bateson (1909) had in mind when he coined the term. Statistical epistasis on the other hand occurs at the population level and is realized when there is interindividual variation in DNA sequences. The statistical phenomenon of epistasis is what Fisher (1918) had in mind. The relationship between biological and statistical epistasis is often confusing but will be important to understand if we are to make biological inferences from statistical results (Moore and Williams, 2005). The focus of the present study is the detection and characterization of statistical epistasis in human populations. It is the promise of systems biology to deliver an etiological understanding of epistasis (Moore and Williams, 2005; Moore et al., 2005; Moore, 2005).

Statistical epistasis is difficult to detect and characterize in human studies due to its inherent nonlinearity. In its extreme form, epistasis can occur in the absence of detectable independent effects of any one polymorphism. This presents several very difficult computational and statistical challenges, especially in the context of genome-wide association studies (Moore and Ritchie, 2004). First, modeling nonlinear interactions requires special analytical methods because parametric statistical approaches such as logistic regression can have less power for detecting interactions than independent main effects (Moore and Williams, 2002). Second, a lack of independent main effects has significant implications for genome-wide association studies with thousands of single-nucleotide polymorphisms (SNPs) because computational search strategies utilizing greedy hill-climbing algorithms rely on main effects. How will a specific nonlinear interaction be detected when an effectively infinite number of combinations will need to be evaluated? Exhaustive searches will not be possible and greedy algorithms will not be effective. Finally, models relating combinations of SNPs to disease susceptibility will be inherently difficult to interpret due to the high dimensionality. It has been argued that epistasis is a ubiquitous component of the genetic architecture of common human diseases (Templeton, 2000; Moore, 2003). It is time for ‘retooling’ our analytical approaches (Thornton-Wells et al., 2004) and research and training strategies (Sing et al., 2003).

In response to the need for powerful analytical strategies for detecting epistasis we have previously developed a non-parametric and genetic model-free data mining method called multifactor dimensionality reduction (MDR) (Ritchie et al., 2001; Hahn et al., 2003; Ritchie et al., 2003a,b; Hahn and Moore, 2004; Moore, 2004). With MDR, multilocus genotypes are pooled into high-risk and low-

risk groups, effectively reducing the dimensionality of the genotype predictors (i.e. attributes) from N dimensions to one dimension. The process of defining a new attribute as a function of two or more other attributes is referred to as constructive induction and was first developed by Michalski (1983). The new multilocus genotype attribute can be evaluated for its ability to classify and predict disease status. The MDR algorithm (Hahn et al., 2003) has reasonable power to detect epistasis (Ritchie et al., 2003a,b) but carries out knowledge discovery in a very specific manner using an exhaustive search and a single classifier to identify the optimal combination of polymorphisms for predicting a discrete disease endpoint. We present here a more flexible framework for implementation of MDR, and other constructive induction algorithms, that first selects interesting combinations of polymorphisms using entropy measures (Jakulin and Bratko, 2003; Jakulin et al., 2003) for evaluating and visualizing the information gain (IG) associated with considering attribute interactions. Once interesting SNPs are selected they can then be used to construct new attributes using MDR. Once a new attribute or set of attributes is constructed using MDR it can then be evaluated using any number of machine learning algorithms. Finally, we introduce entropy-based interaction graphs and interaction dendrograms (Jakulin and Bratko, 2003; Jakulin et al., 2003) as visual tools for interpreting epistasis models. We present here this multi-step approach to detecting epistasis and its application to both simulated and real epidemiological data.

2. Methods

2.1. A flexible framework for detecting, characterizing, and interpreting epistasis

There are four steps to our approach to detecting, characterizing, and interpreting epistasis in genetic studies of disease susceptibility. The first step is to select interesting combinations of SNPs using methods such as measures of entropy (Jakulin and Bratko, 2003; Jakulin et al., 2003) to assess IG. The second step is to construct new attributes from those selected in step one using constructive induction algorithms such as MDR. The third step is to develop and evaluate a classification model using the newly constructed attribute(s) with an appropriate machine learning strategy or statistical method such as a naïve Bayes classifier. The fourth and final step is to interpret the final epistasis model using visual methods such as interaction graphs and interaction dendrograms (Jakulin and Bratko, 2003; Jakulin et al., 2003). It is important to note that a wide variety of different methods and algorithms can be used at each step. This is important to remember as no one strategy is likely to be universally powerful. We describe each of these steps in turn below and the specific algorithms and methods employed in this study.

Step 1: Attribute selection using entropy-based measures of IG and interaction. The goal of the first step is to select

interesting attributes or SNPs from the pool of possible candidates. This can be accomplished using any number of filter methods such as a χ^2 -test of independence or ReliefF (Robnik-Siknjica and Kononenko, 2003). Here, we explore the use of entropy-based measures of IG to select SNPs that might interact.

Jakulin and Bratko (2003) have provided a metric for determining the gain in information about a class variable (e.g. case-control status) from merging two attributes together over that provided by the attributes independently. This measure of IG allows us to gauge the benefit of considering two (or more) attributes as one unit. While the concept of IG is not new (McGill, 1954), its application to the study of attribute interactions has been the focus of several recent studies (Jakulin and Bratko, 2003; Jakulin et al., 2003). Consider two attributes, A and B , and a class label C . Let $H(X)$ be the Shannon entropy (see Pierce, 1980) of X . The IG of A , B , and C can be written as Eq. (1) and defined in terms of Shannon entropy (Eqs. (2) and (3)).

$$IG(ABC) = I(A; B|C) - I(A; B), \quad (1)$$

$$I(A; B|C) = H(A|C) + H(B|C) - H(A, B|C), \quad (2)$$

$$I(A; B) = H(A) + H(B) - H(A, B). \quad (3)$$

The first term in (1), $I(A; B|C)$, measures the *interaction* of A and B . The second term, $I(A; B)$, measures the *dependency* or correlation between A and B . If this difference is positive, then there is evidence for an attribute interaction that cannot be linearly decomposed. If the difference is negative, then the information between A and B is redundant. If the difference is zero, then there is evidence of conditional independence or a mixture of synergy and redundancy.

To implement attribute selection, entropy-based IG is estimated for each individual attribute (i.e. main effects) and each pairwise combination of attributes (i.e. interaction effects). Pairs of attributes are sorted and those with the highest IG, or percentage of entropy in the class removed, are selected for further consideration. Other algorithms and metrics such as ReliefF (reviewed by Robnik-Siknjica and Kononenko, 2003) could be used to select interesting attributes.

Step 2: Constructive induction using MDR. Once interesting SNPs are selected, they can then be used in conjunction with constructive induction algorithms to generate new attributes that capture interaction information. Here, we use the MDR algorithm for constructive induction.

MDR was developed as a non-parametric and genetic model-free data mining strategy for identifying combination of SNPs that are predictive of a discrete clinical endpoint (Ritchie et al., 2001; Hahn et al., 2003; Ritchie et al., 2003a,b; Hahn and Moore, 2004; Moore, 2004). The MDR method has been successfully applied to detecting gene–gene interactions for a variety of common human diseases including, for example, sporadic breast cancer

(Ritchie et al., 2001), essential hypertension (Williams et al., 2004), atrial fibrillation (AF) (Tsai et al., 2004), myocardial infarction (Coffey et al., 2004), type II diabetes (Cho et al., 2004), prostate cancer (Xu et al., 2005), schizophrenia (Qin et al., 2005), and familial amyloid polyneuropathy (Soares et al., 2005). The MDR method has also been successfully applied in the context of pharmacogenetics and toxicogenetics (Wilke et al., 2005a,b). Implementation of MDR is computationally complex with reliance on exhaustive search algorithms, cross-validation (Hastie et al., 2001; Coffey et al., 2004), and permutation testing (Good, 2000). However, at the heart of the MDR approach is a constructive induction algorithm that creates a new attribute by pooling genotypes from multiple SNPs. Incorporating the kernel of the MDR approach in this multistep framework provides many more options for using MDR as part of a data mining strategy.

Constructive induction using the MDR kernel is accomplished in the following way. Given a threshold T , a multilocus genotype combination is considered high-risk if the ratio of cases to controls exceeds T , else it is considered low-risk. When the number of cases and controls are equal it is intuitive to set T equal to one. Bayes rule can also be used. Genotype combinations considered to be high-risk are labeled G_1 while those considered low-risk are labeled G_0 . This process constructs a new one-dimensional attribute with levels G_0 and G_1 . This new attribute can then be evaluated using any machine learning strategy method in Step 3. The MDR kernel has been integrated as a filter module into an open-source and freely available data mining software package called Weka (Witten and Frank, 2000; Frank et al., 2004). This special version of Weka (Weka-CG) can be downloaded for free from <http://www.epistasis.org>.

Step 3: Classification and machine learning. Once a new multilocus attribute is constructed in Step 2, it can then be evaluated using any machine learning method amenable to discrete data. We selected a naïve Bayes classifier here because we have shown previously that the classifier used in the MDR software package (Hahn et al., 2003) is analogous to a naïve Bayes classifier (Hahn and Moore, 2004). The naïve Bayes classifier uses probabilities to link hypotheses to events described by a list of attributes. More precisely, Mitchell (1997) defines the naïve Bayes classifier as (4)

$$v_{NB} = \arg \max_{v_j \in V} P(v_j) \prod_{i=1}^n P(a_i|v_j), \quad (4)$$

where v_j is one of a set of V classes and a_i is one of n attributes describing an event or data element. The class associated with a specific attribute list is the one, which maximizes the probability of the class and the probability of each attribute value given the specified class. The standard way to apply the naïve Bayes classifier to genotype data would be to use the genotype information

for each individual as a list of attributes to distinguish between the two hypotheses “The subject is high-risk.” and “The subject is low-risk”. An advantage of the naïve Bayes classifier is its simplicity and its basis in probability theory.

However, it is important to note that other methods such as decision trees or support vector machines can be used. At this point, an odds ratio for the single multilocus attribute can also be estimated using logistic regression to facilitate a traditional epidemiological analysis and interpretation. Evaluation of the predictor can be carried out using cross-validation (Hastie et al., 2001) and permutation testing (Good, 2000), for example. Cross-validation is a useful approach for limiting false-positives by assessing the generalizability of models (Coffey et al., 2004).

Step 4: Interpretation using interaction graphs and dendrograms. The final step is to interpret the multilocus model of disease susceptibility. This can be accomplished by building an interaction graph and an interaction dendrogram using the entropy estimates from Step 1 with the algorithms described by Jakulin and Bratko (2003). Interaction graphs are comprised of a node for each attribute with pairwise connections between them. The percentage of entropy removed (i.e. IG) by each attribute is visualized for each node. The percentage of entropy removed for each pairwise Cartesian product of attributes is visualized for each connection. Thus, the independent main effects of each polymorphism can be quickly compared to the interaction effect. Additive and non-additive interactions can be quickly assessed and used to interpret the MDR model, which consists of distributions of cases and controls for each genotype combination. Positive entropy values indicate synergistic interaction while negative entropy values indicate redundancy.

Interaction dendrograms are also a useful way to visualize interaction (Jakulin and Bratko, 2003). Here, hierarchical clustering is used to build a dendrogram that places strongly interacting attributes close together at the leaves of the tree. Jakulin and Bratko (2003) define the following dissimilarity measure, D (5), that is used by a hierarchical clustering algorithm to build a dendrogram. The value of 1000 is used as an upper bound to scale the dendrograms.

$$D(A, B) = |I(A; B; C)|^{-1} \text{ if } |I(A; B; C)|^{-1} < 1000 \\ 1000 \text{ otherwise.} \quad (5)$$

Using this measure, a dissimilarity matrix can be estimated and used with hierarchical cluster analysis to build an interaction dendrogram. This facilitates rapid identification and interpretation of pairs of interactions.

The algorithms for the entropy-based measures of IG are implemented in the Orange machine learning software package which is written in Python and provided for free as open-source (e.g. Curk et al., 2005). The entropy-based interaction graphs and interaction dendrograms were also implemented using Orange using Python scripts written by Jakulin and Bratko (2003).

2.2. Simulation study

The goal of the simulation study was to illustrate the usefulness of the strategy outlined above for detecting, characterizing, and interpreting gene–gene interactions. We developed one-, two-, and three-locus models of disease susceptibility in the form of penetrance functions that define the probability (p) of disease (D) given a particular genotype (G) or genotype combination (i.e. $p[D|G]$). The one- and two-locus models were selected from those described previously by Li and Reich (2000). We use here the naming convention of Li and Reich (2000) for the penetrance functions. All models use SNPs with two alleles of equal frequency. Genotype frequencies are consistent with Hardy–Weinberg proportions. Fig. 1A illustrates the one-locus penetrance function. This first model (M63) is a dominant model of disease susceptibility with a

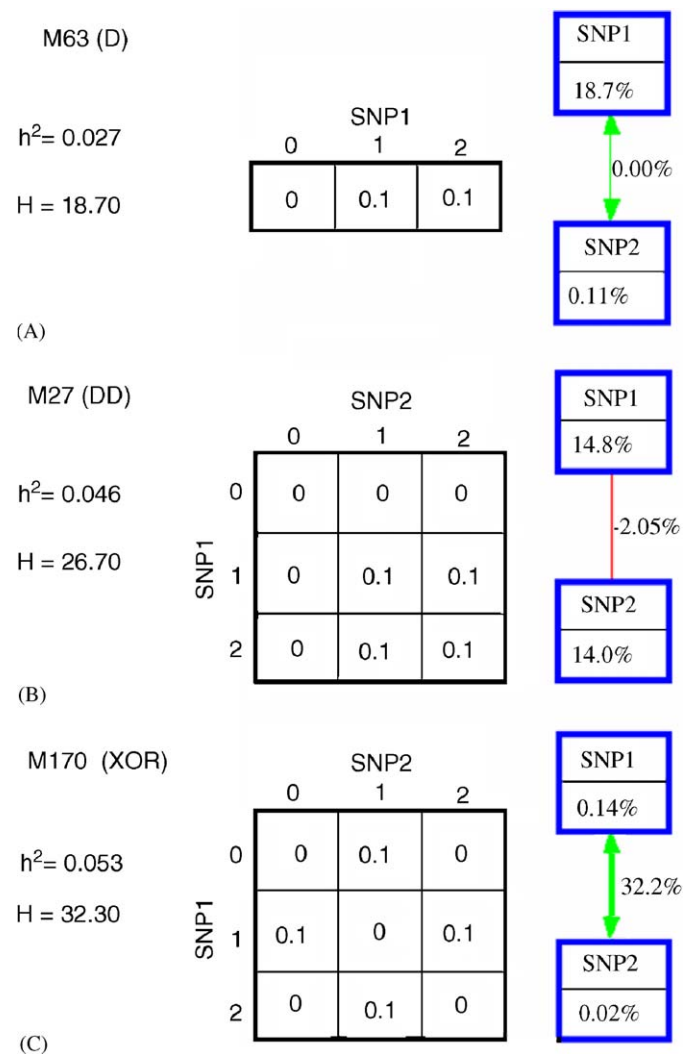


Fig. 1. Penetrance functions (tables), broad-sense heritability (h^2), total entropy (H), and interaction graphs for a dominant main effect model (A), a dominant-by-dominant interaction model (B), and a nonlinear interaction model based on the XOR function (C). Note that the entropy estimates in the interaction maps recapitulate the genetic models.

broad-sense heritability of 0.027. Figs. 1B and C illustrate two different two-locus interaction models. The first model (M27) specifies a dominant-by-dominant interaction with a heritability of 0.046. This model specifies an interaction effect and independent main effects. Model M170 specifies a nonlinear interaction in the absence of independent main effects using an XOR function that is not linearly separable. Here, a subject is at risk of disease if they inherit a heterozygous genotype at one locus, the other locus, but not both. The heritability for this model is 0.053.

We constructed a three-locus model using a composite penetrance function (P_c) of the form $P_c = aP_1 + bP_2$ where a and b sum to one and represent the relative weighting of two penetrance functions (P_1 and P_2) that combine additively. Here, we used weights of $a = 0.6$ and $b = 0.4$ to specify a larger genetic effect that is additive with a smaller effect. The three-locus model we selected combines the M170 (P_1) model with the M63 (P_2) model so that there is a two-locus interaction that is additive with a smaller single-locus main effect (XOR + D). This model has a heritability of 0.042. All models were selected to have heritability in the range of 0.02–0.10 that might be expected for a common, complex disease in which not all susceptibility factors are accounted for. These represent small genetic effect sizes. The detailed penetrance function for the three-locus model is not shown due to space concerns but is available upon request. We used each genetic model to simulate 200 cases and 200 controls, which represent a common, yet moderate, sample size for genetic studies. The simulation methods used are similar to those used by Ritchie et al. (2003a, b).

We next wanted to evaluate whether SNPs selected using entropy measures and then used to construct new attributes using MDR improves classification of discrete clinical endpoints. For each dataset, we created two new datasets. The first new dataset consisted of the functional SNPs plus the class variable. The second dataset consisted of just a single attribute constructed using MDR in addition to the class variable. Using 10-fold cross validation, we compared the accuracy, sensitivity, specificity, and precision of a naïve Bayes classifier between the datasets for each two-locus and three-locus genetic model. Each of these measures is a function of the percentage of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Accuracy is defined as $(TP + TN) / (TP + TN + FP + FN)$. Sensitivity or recall is defined as $TP / (TP + FN)$ while specificity is defined as $TN / (TN + FP)$. Precision or the positive predictive value is defined as $TP / (TP + FP)$. Means were compared using a corrected resampled t -test (Nadeau and Bengio, 2003) and were considered statistically different when the p -value was less than or equal to the acceptable type I error rate of 0.05. These analyses were carried out using the Weka software package (Witten and Frank, 2000; Frank et al., 2004) with our MDR kernel (Weka-CG, <http://www.epistasis.org>).

2.3. Application to an atrial fibrillation dataset

We have previously carried out an MDR analysis of 250 patients with documented non-familial structural AF and 250 controls that were matched to cases on a 1-to-1 basis with regard to age, gender, presence of left ventricular dysfunction, and presence of significant valvular heart disease (Tsai et al. 2004). The *ACE* gene insertion/deletion (I/D) polymorphism, the *T174M*, *M235T*, *G-6A*, *A-20C*, *G-152A*, and *G-217A* polymorphisms of the *angiotensinogen* gene, and the *A1166C* polymorphism of the *angiotensin II type I receptor* gene were studied. Here, we applied the four-step approach described above to this dataset. We first estimated the IG for each SNP and each pair of SNPs. The pair of SNPs and the single SNP with the highest IG were selected and used to construct a new attribute with MDR. As described above, we used 10-fold cross validation and a corrected resampled t -test (Nadeau and Bengio, 2003) to compare the accuracy, sensitivity, specificity, and precision of a naïve Bayes classifier between a dataset with the three best SNPs and a dataset with the single MDR-constructed attribute. Means were considered statistically different when the p -value was less than or equal to the acceptable type I error rate of 0.05.

3. Results

3.1. Analysis of simulated data

Figs. 1 and 2 illustrate the overall class entropy explained by the functional loci from each model along with its associated interaction graph. In each case, the patterns of entropy recapitulate the main and/or interaction effects for each genetic model. For example, the interaction map for the three-locus model in Fig. 2 shows a clear interaction effect of SNP1 and SNP2 while SNP5 has a clear main effect. Note that SNP1 and SNP2 do not by themselves remove any significant class entropy. This is expected because the XOR model (M170) has, by design, no independent main effects. These results suggest that entropy-based measures of IG will be useful for selecting SNP subsets because the pattern of IG recapitulate a wide variety of different genetic effects including the extreme example of interaction in the absence of main effects.

Table 1 summarizes the mean and standard deviation of each performance measure for each genetic model. For the M27 (DD) model we found no statistically significant differences among any of the performance measures. In fact, the performance measures were all identical. Constructing a new variable using MDR made no difference in classification using a naïve Bayes classifier. This result is expected since the SNPs in these models have significant main effects on disease susceptibility that can be modeled effectively using a classifier such as naïve Bayes. However, constructive induction using MDR did significantly improve classification accuracy, specificity, and precision for the M170 (XOR) model. Here, the naïve Bayes classifier

was only able to effectively model the nonlinear interaction between the two SNPs when a new variable was constructed using the MDR algorithm. Similar results were found for the three-locus model that included the M170 model additively with a dominant locus. Thus,

MDR significantly improved classification in the presence of epistasis or nonlinear gene–gene interaction.

3.2. Analysis of an atrial fibrillation dataset

We have demonstrated using several multilocus genetic models that attribute selection using entropy measures followed by constructive induction using MDR improves classification with a naïve Bayes classifier when nonlinear interactions are present. How will this approach perform when applied to a noisy dataset from a real epidemiological study? We first review our previous results from an exhaustive MDR analysis of an AF dataset and then present results from the comprehensive strategy outlined here.

A single-locus association analysis by Tsai et al. (2004) revealed significant results for three polymorphisms in the *angiotensinogen* gene. An exhaustive MDR analysis was used to evaluate interactions among all possible subsets of the eight polymorphisms. The best model consisted of two polymorphisms from the *angiotensinogen* gene (*T174M*, *M235T*) and a single polymorphism from the *ACE* gene (*I/D*). This three-locus model had a perfect cross-validation consistency of 10 and a prediction error of 37.26. Both were significant at the 0.001 level based on a 1000-fold permutation test. Interestingly, only one of the polymorphisms in the MDR model had a significant main effect.

Fig. 3 illustrates the interaction graph for these polymorphisms based on entropy-based measures of IG. It is clear from the interaction graph that the *M235T* polymorphism has a main effect that is independent of the other loci. It is also clear that the *T174M* and the *ACE I/D* polymorphisms have an interaction effect in the absence of a main effect. Fig. 4 illustrates the interaction dendrogram

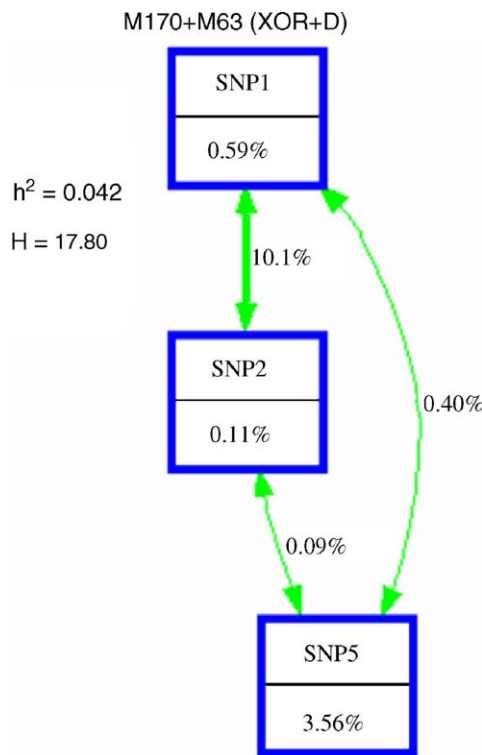


Fig. 2. Interaction map, broad-sense heritability (h^2), and total entropy (H) for a three-locus model consisting of a nonlinear interaction (XOR) that is additive with a dominant main effect (D).

Table 1
Statistical comparison of performance measures for data from different genetic models

Model	Measure	Data ^a		<i>p</i> -value
		No MDR	MDR	
M27 (DD)	Accuracy	0.72 (0.05)	0.72 (0.05)	>0.05
	Sensitivity	0.44 (0.10)	0.44 (0.10)	>0.05
	Specificity	1 (0.00)	1 (0.00)	>0.05
	Precision	1 (0.00)	1 (0.00)	>0.05
M170 (XOR)	Accuracy	0.49 (0.10)	0.76 (0.05)	<0.001
	Sensitivity	0.45 (0.17)	0.51 (0.10)	>0.05
	Specificity	0.53 (0.20)	1 (0.00)	<0.001
	Precision	0.49 (0.13)	1 (0.00)	<0.001
M170 + M63 (XOR + D)	Accuracy	0.55 (0.07)	0.71 (0.08)	0.02
	Sensitivity	0.41 (0.11)	0.65 (0.12)	<0.001
	Specificity	0.69 (0.12)	0.77 (0.09)	>0.05
	Precision	0.57 (0.11)	0.74 (0.09)	0.05
Atrial fibrillation	Accuracy	0.57 (0.07)	0.63 (0.06)	0.02
	Sensitivity	0.81 (0.11)	0.74 (0.09)	>0.05
	Specificity	0.33 (0.12)	0.53 (0.09)	<0.001
	Precision	0.55 (0.05)	0.61 (0.05)	0.007

^aMean (standard deviation) of each performance measure estimated from 10 cross-validation datasets.

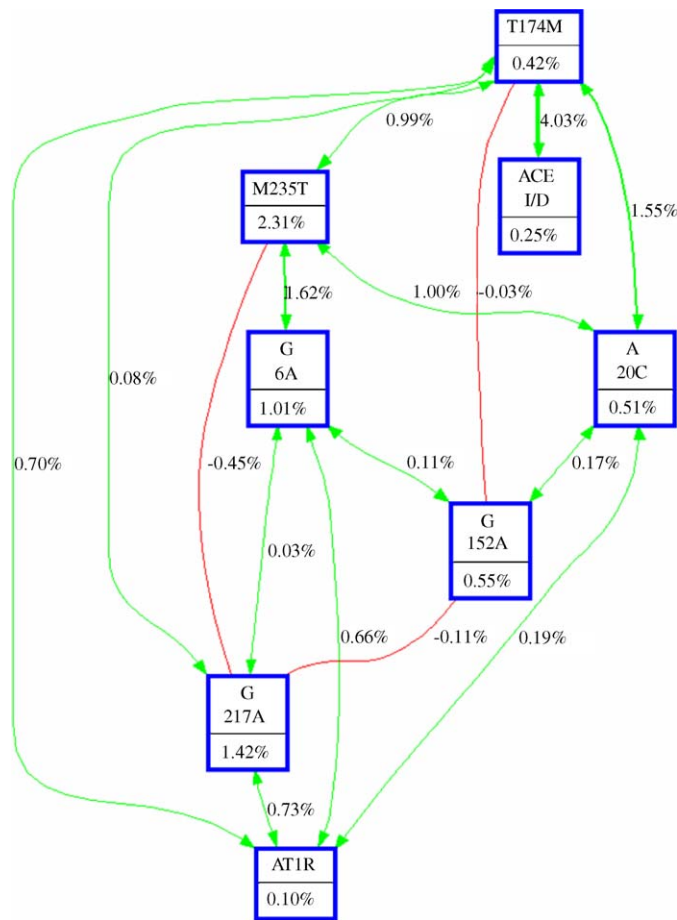


Fig. 3. Interaction graph for the atrial fibrillation dataset. Note that the *T174M* and *ACE I/D* polymorphisms jointly explain (i.e. remove) the most entropy. The *M235T* polymorphism has the largest univariate effect.

for these polymorphisms. The hierarchical cluster analysis clearly shows that the *T174M* and the *ACE I/D* polymorphisms have the strongest synergistic interaction effect. Thus, the entropy analysis identifies these three polymorphisms as the top candidates to be carried forward to the constructive induction step using MDR.

In the next step, we employed MDR to create a new attribute and then compare the classification using the three original polymorphisms with the new constructed attribute. Using 10-fold cross validation and a paired *t*-test, we compared the accuracy, sensitivity, specificity, and precision of a naïve Bayes classifier. Table 1 summarizes the results of this comparison. We found that the MDR constructed attribute significantly improved the accuracy, specificity, and precision of the naïve Bayes classifier. This is similar to the results obtained from the simulations with the three-locus model that consisted of a two-locus nonlinear interaction and an additive main effect. In addition to improving classification, we now have more information for the interpretation of this model. The interaction graph and interaction dendrogram illustrated in Figs. 3 and 4 clearly show the independent main effect of the *M235T* polymorphism and the nonlinear interaction of

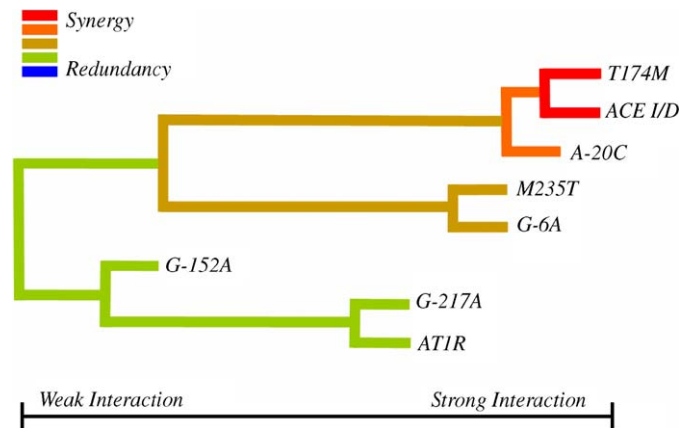


Fig. 4. Interaction dendrogram for the atrial fibrillation dataset. It is clear that the *T174M* and *ACE I/D* polymorphisms have the strongest synergistic interaction.

the *T174M* and *ACE I/D* polymorphisms. This interpretation was not obvious from the distribution of cases and controls across the 27 three-locus genotype combinations in the MDR model reported by Tsai et al. (2004) and illustrated in Fig. 5A.

4. Discussion and conclusions

We have presented a comprehensive and flexible framework for the detection, characterization, and interpretation of epistasis in genetic studies of disease susceptibility. This strategy first selects interesting combinations of polymorphisms based on their interaction information estimated using the entropy-based measures proposed by Jakulin and Bratko (2003) and Jakulin et al. (2003). Once interesting subsets of attributes are selected they can be reduced to single attribute using constructive induction methods such as MDR and then modeled using machine learning and classification. In addition to showing that entropy-based measures are useful for selecting subsets of polymorphisms we have also shown that the interaction graphs and interaction dendrograms (Jakulin and Bratko, 2003; Jakulin et al., 2003) constructed using these measures are useful for model interpretation. Additive and non-additive effects can be quickly identified and interpreted.

A major advantage of this approach is its flexibility. For example, the constructive induction step can be carried using methods other than MDR. In fact, MDR is one of many novel methods that have been developed to combine attributes. All constructive induction algorithms fall into one of four general categories based on their function. The MDR approach is an example of hypothesis-driven constructive induction (HCI) in which the representation space (i.e. attributes) is first transformed using some set of functions or operators. The AQ17 method is another example that uses HCI (Wnek and Michalski, 1994). Data-driven constructive induction (DCI) is implemented by applying a set of operators, seeing how well the new representation aids your classifier, and the sequentially

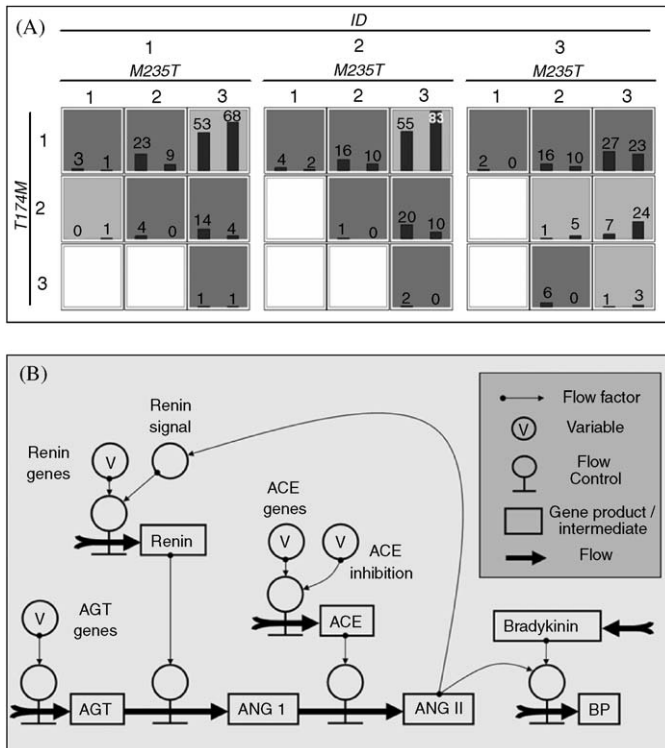


Fig. 5. This figure illustrates the statistical model for a previously reported multilocus association between three polymorphisms in the *AGT* and *ACE* genes and susceptibility to atrial fibrillation (Tsai et al., 2004). The distribution of cases (left bars) and controls (right bars) are illustrated for each multilocus genotype combination. Dark-shaded cells are considered “high-risk” for disease while light-shaded cells are considered “low-risk”. White cells indicate no subjects with those genotype combinations were observed in the dataset. Also illustrated (B) is a summary of a computational model of the renin-angiotensin system as adapted from Takahashi et al. (2003) and Takahashi and Smithies (2004). A more detailed form of this model has been used to carry out simulations of the changes in this pathway that are involved in blood pressure regulation (Takahashi et al., 2003). Expansion of this model may facilitate computational thought experiments (e.g. Moore et al., 2005) that can be used to generate hypotheses about the relationship between *AGT*, *ACE* and susceptibility to atrial fibrillation.

applying more operators until the representation has full coverage (i.e. all instances are correctly classified). The hierarchy induction tool or HINT (Zupan et al., 1998) is an example of a DCI algorithm (Bloedorn and Michalski, 1998). Knowledge-based constructive induction (KCI) is different from both HCI and DCI in that it uses domain specific knowledge to construct attributes. As a simple example, genotypes for a SNP with a recessive effect could be coded as 0 (*AA*, *Aa*) and 1 (*aa*) to form a new attribute. The Automated Mathematician program (Lenat, 1983, 1997) is an example of a KCI algorithm. Finally, two or more of these approaches can be combined to form a multistrategy constructive induction (MCI) algorithm. For example, applying MDR in a DCI framework might be good for dealing genetic heterogeneity because different attributes are constructed for different portions of the data.

It is important to note that constructive induction algorithms such as MDR can be applied using preprocess-

sing or interleaving (Hu, 1998). The approach we have presented here uses preprocessing. That is, information theory is used to select interesting attributes for further consideration by constructive induction and machine learning. However, it is also possible to combine attribute selection and construction into an iterative process called interleaving. Here, an interesting set of attributes is selected and a new attribute constructed. This new constructed attribute is then placed back into the dataset and the process repeated. Interleaving will be a valid approach to explore within the framework presented here because it allows hierarchical interaction models to be constructed. This will be very important for genome-wide association studies (Hirschhorn and Daly, 2005; Wang et al., 2005) with 300 000 or more SNPs because it will not be possible to identify a five-locus epistatic model due to the computational limitations of exhaustive searches (Moore and Ritchie, 2004). If the five-locus interaction is decomposable into lower order models it may be possible to computationally detect it through an interleaving approach to constructive induction. If the five-locus interaction is not decomposable then our best bet is to use expert knowledge to guide the process via a KCI algorithm.

There are many additional variations that are possible within this general framework. For example, once a new attribute is constructed with MDR it can be added back to the attribute list and the full dataset analysed using any classifier combined with a deterministic search algorithm such as best-first (Michalewicz and Fogel, 2000) or a stochastic search algorithm such as simulated annealing (Kirkpatrick et al., 1983) or a genetic algorithm (Goldberg, 1998). This will be important as we move to genome-wide association studies (Moore and Ritchie, 2004; Marchini et al., 2005). All of these options and more can be implemented as part of the user-friendly Weka (Witten and Frank, 2000; Frank et al., 2004) and Orange (Curk et al., 2005) software packages.

The availability of flexible algorithms and software for detecting, characterizing, and interpreting epistasis will play an important role in understanding susceptibility to common human diseases (Hoh and Ott, 2004; Thornton-Wells et al., 2004). Other approaches to detecting epistasis have been proposed and can all be combined with constructive induction algorithms such as MDR. For example, MDR could be integrated with neural network approaches such as those proposed by Ritchie et al. (2003a, b, 2004).

Finally, no discussion of statistical epistasis is complete without addressing its relationship with biological epistasis. Ultimately, it will be important to make etiological inferences from our multilocus models of disease susceptibility. Fig. 5 illustrates the challenge of making inferences about biological epistasis in a pathway from statistical epistasis in a population-based model. Panel A illustrates the distribution of cases and controls across the 27 genotype combinations from the three *AGT* and *ACE* genes polymorphisms that were previously identified by

MDR as significant predictors of AF (Tsai et al., 2004). How does the distribution of ‘high-risk’ and low-risk’ genotype combinations relate to the renin–angiotensin system summarized in Panel B? Other studies of common disease such as type I (Cordell et al., 1995, 2001) and type II (Cox et al., 1999, 2004) diabetes have struggled with similar issues in the search for functional genes. Regardless, making inferences about etiology from any genetic model is a significant challenge (Page et al., 2003). This is especially true for studies of humans. Indeed, Moore and Williams (2005) have suggested that developing an understanding of the relationship between statistical and biological epistasis will be most successful in less complex unicellular organisms where systems biology will have the biggest impact in the short term. This is evident in recent systems-level studies of epistasis in yeast (e.g. Segre et al., 2005).

In summary, we have described a multistep analytical approach for the detection, characterization, and interpretation of epistasis in genetic and epidemiologic studies of common human diseases. At the heart of this approach is the use of constructive induction algorithms such as MDR that are capable of capturing information about nonlinear interactions among multiple attributes (i.e. SNPs) in a single constructed attribute. The power of this strategy is the ability to plug and play different attribute selection methods, different constructive induction algorithms, and different machine learning strategies. We believe that this is a realistic framework for detecting epistasis because no one method or approach will be ideal for all human diseases. The ability to take multiple analytical paths through this framework will facilitate knowledge discovery from patterns of results across different strategies with different strengths and weaknesses. Perhaps the greatest challenge will be the interpretation of epistasis models. The interaction graphs and interaction dendrograms we have introduced to this domain provide powerful visual tools for the statistical interpretation of epistasis effects. Templeton (2000) and Wade (2001) have argued that epistasis is commonly found when properly investigated. The analytical framework described here takes us a step closer to realizing this.

Acknowledgments

This work was supported by National Institutes of Health (USA) Grants AI59694, HD047447, RR018787, and HL65234, and two grants from the National Taiwan University Hospital, Taiwan, ROC (92N016 and 93N004). We thank Dr. Aleks Jakulin for his assistance with the interaction dendrograms.

References

Bateson, W., 1909. *Mendel's Principles of Heredity*. Cambridge University Press, Cambridge.

Bloedorn, E., Michalski, R.S., 1998. Data-driven constructive induction. *IEEE Intell. Syst.* 13, 30–37.

Brodie III, E.D., 2000. Why evolutionary genetics does not always add up. In: Wolf, J., Brodie, III, B., Wade, M. (Eds.), *Epistasis and the Evolutionary Process*. Oxford University Press, New York, pp. 3–19.

Cho, Y.M., Ritchie, M.D., Moore, J.H., Park, J.Y., Lee, K.U., Shin, H.D., Lee, H.K., Park, K.S., 2004. Multifactor-dimensionality reduction shows a two-locus interaction associated with Type 2 diabetes mellitus. *Diabetologia* 47, 549–554.

Coffey, C.S., Hebert, P.R., Ritchie, M.D., Krumholz, H.M., Morgan, T.M., Gaziano, J.M., Ridker, P.M., Moore, J.H., 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 Bioinform.* 4, 49.

Cordell, H.J., Todd, J.A., Bennett, S.T., Kawaguchi, Y., Farrall, M., 1995. Two-locus maximum lod score analysis of a multifactorial trait: joint consideration of IDDM2 and IDDM4 with IDDM1 in type 1 diabetes. *Am. J. Hum. Genet.* 57, 920–934.

Cordell, H.J., Todd, J.A., Hill, N.J., Lord, C.J., Lyons, P.A., Peterson, L.B., Wicker, L.S., Clayton, D.G., 2001. Statistical modeling of interlocus interactions in a complex disease: rejection of the multiplicative model of epistasis in type 1 diabetes. *Genetics* 158, 357–367.

Cox, N.J., Frigge, M., Nicolae, D.L., Concannon, P., Hanis, C.L., Bell, G.I., Kong, A., 1999. Loci on chromosomes 2 (NIDDM1) and 15 interact to increase susceptibility to diabetes in Mexican Americans. *Nat. Genet.* 21, 213–215.

Cox, N.J., Hayes, M.G., Roe, C.A., Tsuchiya, T., Bell, G.I., 2004. Linkage of calpain 10 to type 2 diabetes: the biological rationale. *Diabetes* 53 (Suppl 1), S19–S25.

Curk, T., Demsar, J., Xu, Q., Leban, G., Petrovic, U., Bratko, I., Shaulsky, G., Zupan, B., 2005. Microarray data mining with visual programming. *Bioinformatics* 21, 396–398.

Fisher, R.A., 1918. The correlations between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinburgh* 52, 399–433.

Frank, E., Hall, M., Trigg, L., Holmes, G., Witten, I.H., 2004. Data mining in bioinformatics using Weka. *Bioinformatics* 20, 2479–2481.

Gibson, G., Wagner, G., 2000. Canalization in evolutionary genetics: a stabilizing theory? *BioEssays* 22, 372–380.

Goldberg, D.E., 1998. *Genetic Algorithms in Search, Optimisation and Machine Learning*. Addison-Wesley Publishing Company, Inc., Reading, Massachusetts.

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

Hahn, L.W., Moore, J.H., 2004. Ideal discrimination of discrete clinical endpoints using multilocus genotypes. *In Silico Biol.* 4, 183–194.

Hahn, L.W., Ritchie, M.D., Moore, J.H., 2003. Multifactor dimensionality reduction software for detecting gene–gene and gene–environment interactions. *Bioinformatics* 19, 376–382.

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

Hirschhorn, J.N., Daly, M.J., 2005. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* 6, 95–108.

Hoh, J., Ott, J., 2004. Genetic dissection of diseases: design and methods. *Curr. Opin. Genet. Dev.* 14, 229–232.

Hollander, W.F., 1955. Epistasis and hypostasis. *J. Hered.* 46, 222–225.

Hu, Y.-J., 1998. Constructive induction: covering attribute spectrum. In: Liu, H., Motoda, H. (Eds.), *Feature Extraction, Construction and Selection: A Data Mining Perspective*. Kluwer, Boston, pp. 257–272.

Jakulin, A., Bratko, I., 2003. Analyzing attribute interactions. *Lect. Notes Artif. Intell.* 2838, 229–240.

Jakulin, A., Bratko, I., Smrke, D., Demsar, J., Zupan, B., 2003. Attribute interactions in medical data analysis. *Lect. Notes Artif. Intell.* 2780, 229–238.

Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P., 1983. Optimization by simulated annealing. *Science* 220, 671–680.

Lenat, D.B., 1983. Learning from observation and discovery. In: Michalski, R.S., Carbonell, J.G., Mitchell, T.M. (Eds.), *Machine Learning: An Artificial Intelligence Approach*. Morgan Kaufmann, Los Altos, CA.

- Lenat, D.B., 1997. On automated scientific theory formation: a case study using the AM program. In: Hayes, J.E., Michie, D., Mikulich, L.I. (Eds.), *Machine Intelligence*, vol. 9. Halstead Press, New York.
- Li, W., Reich, J., 2000. A complete enumeration and classification of two-locus disease models. *Hum. Hered.* 50, 334–349.
- Marchini, J., Donnelly, P., Cardon, L.R., 2005. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat. Genet.* 37, 413–417.
- McGill, W.J., 1954. Multivariate information transmission. *Psychometrika* 19, 97–116.
- Michalewicz, Z., Fogel, D.B., 2000. *How to Solve It: Modern Heuristics*. Springer, New York.
- Michalski, R.S., 1983. A theory and methodology of inductive learning. *Artif. Intell.* 20, 111–161.
- Mitchell, T., 1997. *Machine Learning*. McGraw-Hill, New York.
- Moore, J.H., 2003. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum. Hered.* 56, 73–82.
- Moore, J.H., 2004. Computational analysis of gene–gene interactions in common human diseases using multifactor dimensionality reduction. *Expert. Rev. Mol. Diagn.* 4, 795–803.
- Moore, J.H., 2005. A global view of epistasis. *Nat. Genet.* 37, 13–14.
- Moore, J.H., Ritchie, M.D., 2004. The challenges of whole-genome approaches to common diseases. *J. Am. Med. Assoc.* 291, 1642–1643.
- Moore, J.H., Williams, S.W., 2002. New strategies for identifying gene–gene interactions in hypertension. *Ann. Med.* 34, 88–95.
- Moore, J.H., Williams, S.W., 2005. Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *BioEssays* 27, 637–646.
- Moore, J.H., Boczek, E.M., Summar, M.L., 2005. Connecting the dots between genes, biochemistry, and disease susceptibility: systems biology modeling in human genetics. *Mol. Genet. Metab.* 84, 104–111.
- Nadeau, C., Bengio, Y., 2003. Inference for the generalization error. *Mach. Learn.* 52, 239–281.
- Page, G.P., George, V., Go, R.C., Page, P.Z., Allison, D.B., 2003. “Are we there yet?”: Deciding when one has demonstrated specific genetic causation in complex diseases and quantitative traits. *Am. J. Hum. Genet.* 73, 711–719.
- Phillips, P.C., 1998. The language of gene interaction. *Genetics* 149, 1167–1171.
- Pierce, J.R., 1980. *An Introduction to Information Theory: Symbols, Signals, and Noise*. Dover, New York.
- Proulx, S.R., Phillips, P.C., 2005. The opportunity for canalization and the evolution of genetic networks. *Am. Nat.* 165, 147–162.
- Qin, S., Zhao, X., Pan, Y., Liu, J., Feng, G., Fu, J., Bao, J., Zhang, Z., He, L., 2005. An association study of the *N*-methyl-D-aspartate receptor NR1 subunit gene (*GRIN1*) and NR2B subunit gene (*GRIN2B*) in schizophrenia with universal DNA microarray. *Eur. J. Hum. Genet.* 13, 807–814.
- Ritchie, M.D., Hahn, L.W., Roodi, N., Bailey, L.R., Dupont, W.D., Parl, F.F., 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., Moore, J.H., 2003a. Power of multifactor dimensionality reduction for detecting gene–gene interactions in the presence of genotyping error, phenocopy, and genetic heterogeneity. *Genet. Epidemiol.* 24, 150–157.
- Ritchie, M.D., White, B.C., Parker, J.S., Hahn, L.W., Moore, J.H., 2003b. Optimization of neural network architecture using genetic programming improves the detection and modeling of gene–gene interactions in studies of human diseases. *BMC Bioinform.* 4, 28.
- Ritchie, M.D., Coffey, C.S., Moore, J.H., 2004. Genetic programming neural networks as a bioinformatics tool in human genetics. *Lect. Notes Comput. Sci.* 3102, 438–448.
- Robnik-Siknja, M., Kononenko, I., 2003. Theoretical and empirical analysis of ReliefF and RReliefF. *Mach. Learn.* 53, 23–69.
- Segre, D., Deluna, A., Church, G.M., Kishony, R., 2005. Modular epistasis in yeast metabolism. *Nat. Genet.* 37, 77–83.
- Sing, C.F., Stengard, J.H., Kardia, S.L., 2003. Genes, environment, and cardiovascular disease. *Arterioscler. Thromb. Vasc. Biol.* 23, 1190–1196.
- Soares, M.L., Coelho, T., Sousa, A., Batalov, S., Conceicao, I., Sales-Luis, M.L., Ritchie, M.D., Williams, S.M., Nievergelt, C.M., Schork, N.J., Saraiva, M.J., Buxbaum, J.N., 2005. Susceptibility and modifier genes in Portuguese transthyretin V30M amyloid polyneuropathy: complexity in a single-gene disease. *Hum. Mol. Genet.* 14, 543–553.
- Takahashi, N., Smithies, O., 2004. Human genetics, animal models and computer simulations for studying hypertension. *Trends Genet.* 20, 136–145.
- Takahashi, N., Hagaman, J.R., Kim, H.S., Smithies, O., 2003. Minireview: computer simulations of blood pressure regulation by the renin–angiotensin system. *Endocrinology* 144, 2184–2190.
- Templeton, A.R., 2000. Epistasis and complex traits. In: Wolf, J., Brodie, III, B., Wade, M. (Eds.), *Epistasis and the Evolutionary Process*. Oxford University Press, New York, pp. 41–57.
- Thornton-Wells, T.A., Moore, J.H., Haines, J.L., 2004. Genetics, statistics and human disease: analytical retooling for complexity. *Trends Genet.* 20, 640–647.
- Tsai, C.T., Lai, L.P., Lin, J.L., Chiang, F.T., Hwang, J.J., Ritchie, M.D., Moore, J.H., Hsu, K.L., Tseng, C.D., Liau, C.S., Tseng, Y.Z., 2004. Renin–angiotensin system gene polymorphisms and atrial fibrillation. *Circulation* 109, 1640–1646.
- Wade, M.J., 2001. Epistasis, complex traits, and mapping genes. *Genetica* 112–113, 59–69.
- Waddington, C.H., 1942. Canalization of development and the inheritance of acquired characters. *Nature* 150, 563–565.
- Waddington, C.H., 1957. *The Strategy of the Genes*. MacMillan, New York.
- Wang, W.Y., Barratt, B.J., Clayton, D.G., Todd, J.A., 2005. Genome-wide association studies: theoretical and practical concerns. *Nat. Rev. Genet.* 6, 109–118.
- Wilke, R.A., Moore, J.H., Burmester, J.K., 2005a. Relative impact of CYP3A genotype and concomitant medication on the severity of atorvastatin-induced muscle damage. *Pharmacogenet. Genom.* 15, 415–421.
- Wilke, R.A., Reif, D.M., Moore, J.H., 2005b. Combinatorial pharmacogenetics. *Nat. Rev. Drug Discovery* 4, 911–918.
- Williams, S.M., Ritchie, M.D., Phillips III, J.A., Dawson, E., Prince, M., Dzhura, E., Willis, A., Semenza, A., Summar, M., White, B.C., Addy, J.H., Kpodonu, J., Wong, L.J., Felder, R.A., Jose, P.A., Moore, J.H., 2004. Multilocus analysis of hypertension: a hierarchical approach. *Hum. Hered.* 57, 28–38.
- Witten, I.H., Frank, E., 2000. *Data Mining*. Morgan Kaufman Publishers, San Francisco.
- Wnek, J., Michalski, R.S., 1994. Hypothesis-driven constructive induction in AQ17-HCI: a method and experiments. *Mach. Learn.* 14, 139–168.
- Xu, J., Lowery, J., Wiklund, F., Sun, J., Lindmark, F., Hsu, F.-C., Dimitrov, L., Chang, B., Turner, A.R., Adami, H.-O., Suh, E., Moore, J.H., Zheng, S.L., Isaacs, W.B., Trent, J.M., Gronberg, H., 2005. The interaction of four inflammatory genes significantly predicts prostate cancer risk. *Cancer Epidemiol. Biomarkers Prev.* 14, 2563–2568.
- Zupan, B., Bohanec, M., Demsar, J., Bratko, I., 1998. Feature transformation by function decomposition. *IEEE Int. Syst. Appl.* 13, 38–43.