Original Paper

Human Heredity

Hum Hered 2003;56:73–82
DOI: 10.1159/000073735

Received: April 22, 2003
Accepted after revision: June 17, 2003

The Ubiquitous Nature of Epistasis in Determining Susceptibility to Common Human Diseases

Jason H. Moore

Program in Human Genetics, Department of Molecular Physiology and Biophysics, Vanderbilt University Medical School, Nashville, Tenn., USA

Key Words
Gene-gene interactions · Multifactor dimensionality reduction · Biological systems moduling

Abstract

There is increasing awareness that epistasis or gene-gene interaction plays a role in susceptibility to common human diseases. In this paper, we formulate a working hypothesis that epistasis is a ubiquitous component of the genetic architecture of common human diseases and that complex interactions are more important than the independent main effects of any one susceptibility gene. This working hypothesis is based on several bodies of evidence. First, the idea that epistasis is important is not new. In fact, the recognition that deviations from Mendelian ratios are due to interactions between genes has been around for nearly 100 years. Second, the ubiquity of biomolecular interactions in gene regulation and biochemical and metabolic systems suggest that relationship between DNA sequence variations and clinical endpoints is likely to involve gene-gene interactions. Third, positive results from studies of single polymorphisms typically do not replicate across independent samples. This is true for both linkage and association studies. Fourth, gene-gene interactions are commonly found when properly investigated. We review each of these points and then review an analytical strategy called multifactor dimensionality reduction for detecting epistasis. We end with ideas of how hypotheses about biological epistasis can be generated from statistical evidence using biochemical systems models. If this working hypothesis is true, it suggests that we need a research strategy for identifying common disease susceptibility genes that embraces, rather than ignores, the complexity of the genotype to phenotype relationship.

Introduction

A central goal of genetic epidemiology is the identification of loci with alleles or genotypes that confer an increased susceptibility to human disease. As the focus shifts away from rare Mendelian diseases towards common complex diseases, it is increasingly clear that epistasis or gene-gene interactions will likely be an important component of genetic architecture. While few disagree that complex interactions between both genetic and environmental factors will play a role in some disease etiologies, there is disagreement about how common such interactions are likely to be and their importance relative to independent main effects.

KARGER

Fax + 41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

Accessible online at:
www.karger.com/hhe

Jason H. Moore, PhD
Program in Human Genetics, Department of Molecular Physiology and Biophysics
519 Light Hall, Vanderbilt University Medical School
Nashville, TN 37232-0700 (USA)
Tel. +1 615 343 5852, Fax +1 615 343 8619, E-Mail moore@phg.mc.vanderbilt.edu

In this paper, we propose as a working hypothesis that epistasis is a ubiquitous component of the genetic architecture of common human diseases and that complex interactions are more important than the independent main effects of any one susceptibility gene. This working hypothesis is based on both historical and emerging research results. First, the idea that epistasis is important is not new. In fact, the recognition that deviations from Mendelian ratios are due to interactions between genes has been around for nearly 100 years. Second, the ubiquity of biomolecular interactions in gene regulation and biochemical and metabolic systems suggest that relationship between DNA sequence variations and clinical endpoints is likely to involve gene-gene interactions. Third, positive results from studies of single polymorphisms typically do not replicate across independent samples. This is true for both linkage and association studies. Fourth, gene-gene interactions are commonly found when properly investigated.

If this working hypothesis is true, it suggests that we need a research strategy for identifying common disease susceptibility genes that embraces, rather than ignores, the complexity of the genotype to phenotype relationship. We first present in detail the research results that preceded the formulation of this working hypothesis. We then review a new nonparametric and genetic-model free strategy called multifactor dimensionality reduction (MDR) for identifying high-order gene-gene interactions in association studies. Strategies such as MDR will be necessary if we are to make the intellectual leap from single-locus to multilocus association studies. Finally, we end with some ideas about how we might begin to understand the mechanisms by which epistasis influences disease susceptibility through a hierarchy of interacting genetic and biochemical networks.

Epistasis Is Not a New Idea

The concept of epistasis or gene-gene interaction has been around for at least 100 years and was recognized as an explanation for deviations from simple Mendelian ratios. William Bateson [1909] has been credited by Hollander [1955] and more recently by Phillips [1998] as the first to use the term epistasis that literally translated means 'resting upon'. A commonly used textbook definition of epistasis is one gene masking the effects of another gene [e.g. Neel and Schull 1954; Griffiths et al., 2000]. A classic example of epistasis comes from studies of the shape of seed capsules from crosses of a weedy plant called the shepard's purse *(Bursa bursa-pastoris)* by Shull [1914]. In this study, crosses from doubly heterozygous plants yielded Mendelian ratios of 15 triangular capsules to one oval capsule. It is generally believed that there are two pathways with dominant loci that lead to the triangular shape. It is only when both pathways are blocked by recessive alleles is the oval-shaped seed capsule produced. This is an example of a recessive-by-recessive interaction.

The shepard's purse example from Shull [1914] is an example of biological epistasis. That is, the gene-gene interaction has a biological basis. This is exactly what Bateson [1909] had in mind when he coined the term. This is in contrast to the concept of statistical epistasis or epistacy that was first used by Fisher [1918] to describe deviations from additivity in a statistical model. Making biological inferences about epistasis from statistical models can be difficult, although there are some approaches that attempt to do so [e.g. Cheverud and Routman 1995]. Concepts of biological and statistical epistasis have also been presented from an evolutionary biology perspective [e.g. Wade et al., 2001].

The recognition that epistasis plays an important role in the genotype to phenotype relationship by Bateson [1909] and others [e.g. Wright, 1932, 1934] early in the 20th century suggests that epistasis is indeed an important phenomenon in biology. This is especially true since it has withstood the test of time. It is reasonable to assume that epistasis also plays an important role in human biology and thus human disease susceptibility. While few disagree that epistasis plays a role in disease susceptibility, many disagree about just how common and how important epistasis is. In the following sections, we formulate the working hypothesis that epistasis is likely to be very common and likely to play a central role in determining an individual's risk of disease.

Biomolecular Interactions Are Ubiquitous

Perhaps the most important body of evidence leading to the working hypothesis that epistasis is ubiquitous is the fact that biomolecular interactions are a ubiquitous part of gene regulation, signal transduction, biochemical networks, and homeostatic developmental and physiological pathways. Consider the regulation of transcription. The proteins that are required to regulate transcription of protein-encoding genes can be classified into three groups. The first group is comprised of proteins that activate or repress transcription by binding to specific promoter and enhancer sequences. The second group is comprised of proteins such as RNA polymerase II that are common to all transcriptional complexes. The third group is comprised of proteins that interact with activators and repressors to enhance their effects. As discussed by Martinez [2002], it is cooperative protein-protein and protein-DNA interactions that are critical for the regulation of transcription by proteins from all three groups. Indeed, Gibson [1996] has shown using statistical thermodynamic models that nonadditive interactions are a natural property of transcriptional regulation. The importance of biomolecular interactions is also seen at the level of translation. Gallie [2002] stresses that extensive protein-protein and protein-RNA interactions are necessary for the ribosomes, mRNA, tRNAs, and other accessory factors to form a functional translation complex. Signal transduction is equally dependent on biomolecular interactions [Tyson et al., 2003] as are biochemical and metabolic networks [Michal, 1999]. As reviewed by Pogun [2001], the importance of considering interactions for understanding biological systems was recognized very early on by Bernard [1865] in the context of physiological homeostasis. Taken together, phenotypes can be viewed as the result of vast interconnected biological networks and systems that are driven by biomolecular interactions. Given these biological complexities, Templeton [2000] argues that epistasis must be ubiquitous for determining disease susceptibility.

Evolutionary theory and developmental biology suggests that epistasis is ubiquitous. For example, Gibson [1996] has shown that nonadditive interactions are a natural property of transcriptional regulation and thus may lead to networks of mutations that compensate for one another to produce homeostatic developmental pathways. This stabilization or buffering of developmental pathways by multiple interacting genes has been referred to as canalization [Waddington, 1942, 1957; Gibson and Wagner, 2000]. As Rice [1998] suggests, nonlinear interactions among polymorphisms from multiple different pathways make it possible for canalization to evolve. As discussed by Hansen [2003], increased trait variability is expected when genes enhance each other's effects while decreased variability is expected when genes compensate one another. The idea that gene-gene interactions increase robustness in the presence of mutations has been validated in yeast experiments [Wagner, 2000]. One result of canalization is that the trait variability we observe in populations is partly the result of patterns of epistasis. Canalization or stabilizing selection has important implications for human health because the result is reduced trait variability. Thus, canalization produces human populations that are healthy through epistasis, pleiotropy, and gene redundancy. Physiological homeostasis is a good example of this [Pigliucci, 2001].

The functional consequence of canalization is that extreme phenotypes due to single mutations are rare. Thus, mutations in genes from multiple pathways are necessary for extreme values of biological traits (e.g. disease endpoints). This suggests that multiple polymorphisms and their interactions will need to be considered if we are going to identify disease susceptibility genes. This is certainly true for understanding diseases such as essential hypertension [Williams et al., 2000; Moore and Williams, 2002] and for understanding evolution in general [e.g. Wolf et al., 2000]. It is interesting to note that stabilizing selection may give rise to linkage disequilibrium. This is because selection will favor compensatory alleles at particular loci when there are one or more loci with alleles that confer extreme values of the trait [Smith, 1998]. The relationship between alleles at different human genes will be important to characterize in the context of biological function.

Canalization and stabilizing selection also appear to be important for rare Mendelian diseases. For example, Roberts [1975] reviewed a Danish study by Morch [1941] that estimated the fertility of individuals with achondroplasia is one-fifth that of their unaffected siblings. Thus, the equilibrium allele frequency is very close to the mutation rate for that locus. Is it possible that compensatory mutations stabilize the phenotypes associated with rare Mendelian diseases? It was noted early on that the cystic fibrosis (CF) phenotype is not completely determined by mutations in the *CFTR* gene [Kerem et al., 1990]. In fact, it is now believed that variation in CF phenotype is due to complex interactions between multiple genes from different pathways [Salvatore et al., 2002]. The same is true for sickle-cell anemia [Templeton, 2000] and a number of other Mendelian disorders such as Hirschsprung disease, neurofibromatosis type 2, craniosynostosis, and adrenal hypoplasia congenita [Dipple and McCabe, 2000]. Thus, epistasis is important and even common for determining phenotypes for a number of rare Mendelian diseases.

Single-Locus Results Do Not Replicate

There is a growing awareness in genetic epidemiology that the results of genetic linkage and association studies of common multifactorial diseases do not replicate across multiple samples. In fact, there are typically more negative than positive results for most candidate genes examined. Hirschhorn et al. [2002] recently reviewed more than 600 association studies in which the same polymorphism had been studied in three or more independent samples. Of all these studies, there were only six results that were consistently replicated. This trend has been true for association studies of late-onset Alzheimer disease. As Finckh [2003] reviews, more than 100 different candidate genes for Alzheimer disease have been evaluated without any convincing evidence aside from that observed for the *Apolipoprotein E* gene. A similar review has been conducted by Altmuller et al. [2002] for 101 genetic linkage studies of 31 different diseases. Altmuller et al. [2002] found that fewer than half of the studies showed evidence for linkage and even fewer replicated across studies. Based on this ensemble of association and linkage studies, it appears as though there are few genes that have consistent large effects on multifactorial disease susceptibility in different populations. This can partly be explained by locus heterogeneity and partly by epistasis. Under an epistasis model, any given polymorphism will need to be considered in the context of other polymorphisms. If considered individually, an independent main effect may only be observed when the allele frequencies in a particular population yield a statistically detectable main effect. Under this model, a functional polymorphism considered individually may appear to be either a false positive or perhaps a true positive under a locus heterogeneity model. Indeed, Moore and Williams [2002] have hypothesized that the lack of replication of single-locus results in studies of essential hypertension is because epistasis effects are more important than independent main effects.

Epistasis Is Commonly Found when Properly Investigated

Templeton [2000] has suggested that epistasis is commonly found when properly investigated. Why is epistasis so difficult to detect? What is the proper way to detect epistasis? Epistasis is difficult to detect and characterize using traditional parametric statistical methods such as linear and logistic regression because of the sparseness of the data in high dimensions. That is, when interactions among multiple polymorphisms are considered, there are many multilocus genotype combinations that have very few or no data points. This phenomenon has been referred to as the curse of dimensionality [Bellman, 1961] and, for methods such as logistic regression, can lead to parameter estimates that have very large standard errors resulting in an increase in type I errors [Concato et al., 1993; Peduzzi et al., 1996; Hosmer and Lemeshow, 2000]. In addition, detecting gene-gene interactions using traditional procedures for fitting regression models can be problematic leading to an increase in type II errors and a decrease in power. For example, forward selection is limited because interactions are only tested for those variables that have a statistically significant independent main effect. Those DNA sequence variations that have an interaction effect, but no or minimal main effect, will be missed. With backward elimination, a complete model that includes all main effects and all interaction terms may require too many degrees of freedom. Stepwise procedures are more flexible than either forward selection or backward elimination but can also suffer from requiring too many degrees of freedom. Detecting interactions among variables is a well-known challenge in statistics and data mining [Freitas, 2001].

What is a proper method for detecting epistasis? The answer to this question is currently unknown, however, there are several alternatives to linear and logistic regression that have been developed. Cheverud and Routman [1995] developed a method for detecting epistasis that is based on ideas of biological or physiological epistasis instead of Fisherian statistical epistasis that is dependent on population allele frequencies. Application of the physiological epistasis approach to studies of quantitative traits in mice has routinely identified evidence for epistasis. For example, Leamy et al. [2002] carried out a genetic study of fluctuating asymmetry of mandible size in mice. The authors detected overwhelming evidence of epistasis with more than 30 separate statistically significant nonadditive gene-gene interactions. It is interesting to note that no statistically significant independent main effects were detected for any of the loci in this study [Leamy et al., 2002]. Thus, most of the genetic variation for this trait can be attributed to epistatic effects. Many of these effects would not have been detected using the Fisherian approach due to lack of sufficient power [Leamy et al., 2002].

In addition to the physiological epistasis methods of Cheverud and Routman [1995], exploratory data analysis methods such as the combinatorial partitioning method (CPM) have been developed for generating hypotheses about epistatic effects on quantitative trait variation [Nelson et al., 2001]. The CPM of Nelson et al. [2001] simultaneously considers multiple polymorphic loci to identify combinations of genotypes that are most strongly associated with variation in the quantitative trait. Genotypes from multiple loci are pooled into a smaller number of classes, thereby addressing the increased dimensionality associated with modeling of interactions. Nelson et al. [2001] applied CPM to modeling the relationship between 18 diallelic loci from six cardiovascular disease susceptibility genes and interindividual variability in plasma triglycerides. This study identified nonadditive epistatic interactions between multiple loci in the absence of independent main effects. Moore et al. [2002a, c] have also applied CPM as an exploratory data analysis tool for generating hypotheses about the role of epistasis in the genetic architecture of plasma levels of tissue plasminogen activator (t-PA) and plasminogen activator inhibitor (PAI-1), two enzymes involved in the formation and degradation of blood clots. In these studies, CPM identified evidence of epistasis in the absence of main effects. No evidence of epistasis was detected using the Fisherian analysis of variance approach.

Several alternatives to logistic regression for discrete clinical endpoints have also been developed. For example, Hoh et al. [2000] and Hoh and Ott [2001] have developed a combination of sequential and resampling methods for summing association statistics to detect combined effects of multiple SNPs. This approach uses standard statistics in a novel way to detect multilocus effects. Application of this method to a coronary artery restenosis case-control dataset yielded a highly significant interaction among seven SNPs from seven different genes [Zee et al., 2002]. These associations would not have been identified using standard logistic regression analysis due to a lack of degrees of freedom for estimating all the interactions terms. The use of logic functions for defining new variables that can be included in a logistic regression analysis may also be useful [Kooperberg et al., 2001].

Pattern recognition and machine learning approaches have also been explored for the detection of epistasis. For example, Moore and Hahn [2002a, b] have developed an approach that utilizes the spatial and temporal information processing abilities of cellular automata to identify combinations of SNPs that interact to influence disease risk. Using simulated data, these studies have demonstrated that cellular automata, combined with parallel genetic algorithms for optimization, have very good power for identifying nonlinear interactions among multiple SNPs in the absence of any detectable independent main effects. Ritchie et al. [2003] have developed a neural network strategy for identifying gene-gene interactions. Previous attempts to use neural networks in genetic studies have considered fixed network architectures. Ritchie et al. [2003] have shown using simulated data that different neural network architectures are optimal for different epistasis models. In response to this, Ritchie et al. [2003] developed a machine learning approach that uses genetic programming for optimizing the architecture of the neural network in addition to the weights and the SNP inputs. Simulation studies suggest that this genetic programming neural network strategy has more power for identifying gene-gene interactions than the traditional approach of selecting one or a set of fixed neural network architectures prior to analysis. While promising, these pattern recognition approaches need to be validated with more simulation studies and then applied to real data.

Thus, as more powerful alternatives to traditional parametric statistical approaches are developed, the detection of epistasis becomes easier. As these studies and others have illustrated, epistasis is commonly found when powerful analytical approaches are employed. In the next section, we review our experience with the multifactor dimensionality reduction (MDR) method for identifying gene-gene interactions in epidemiological study designs.

Multifactor Dimensionality Reduction for Detecting Epistasis

The multifactor dimensionality reduction (MDR) method was developed specifically to improve the power to detect gene-gene interactions in epidemiological study designs over that provided by logistic regression [Ritchie et al., 2001]. The MDR approach is nonparametric in that no parameters are estimated and is free of any assumed genetic model. Figure 1 illustrates the general procedure to implement the MDR method. In step one, the data are divided into a training set (e.g. 9/10 of the data) and an independent testing set (e.g. 1/10 of the data) as part of cross-validation. Second, 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 for the number of cases to the number of controls is calculated within each multifactor class. Each multifactor class in *n*-dimensional space is then labeled as 'high risk' if the cases to controls ratio meets or exceeds some threshold (e.g. ≥ 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'). The collection of these multifactor classes composes the MDR model for the particular combination of factors. Among all of the two factor combinations, a single MDR 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

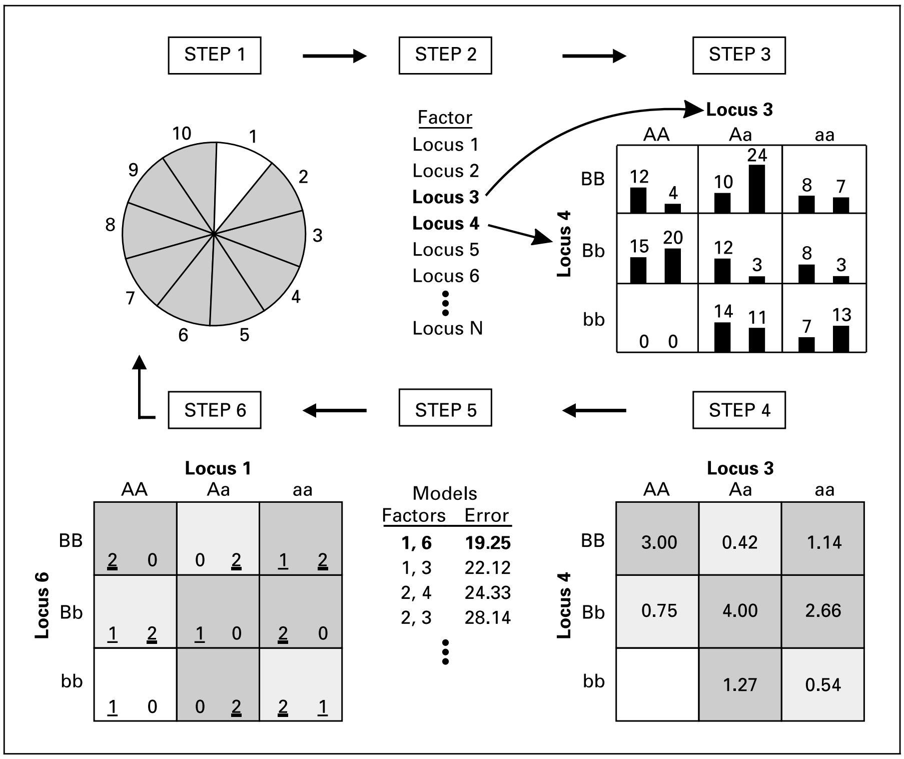


Fig. 1. Summary of the general steps involved in implementing the MDR method [adapted from Ritchie et al., 2001]. In step one, the data are divided into a training set (e.g. 9/10 of the data) and an independent testing set (e.g. 1/10 of the data) as part of cross-validation. In step two, a set of *n* genetic and/or discrete environmental factors is then selected from the pool of all factors. In step three, the *n* factors and their possible multifactor classes or cells are represented in *n*-dimensional space. In step four, each multifactor cell in the *n*-dimensional space is labeled as high-risk if the ratio of affected individuals to unaffected individuals (the number in the cell) exceeds some threshold T (e.g. $T = 1.0$) and low-risk if the threshold is not exceeded. In steps five and six, the model with the best misclassification error is selected and the prediction error of the model is estimated using the independent test data. Steps one through 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 while light-shaded cells represent low-risk genotype combinations. No shading or white cells represent genotype combinations for which no data was observed.

model, prediction error is estimated using 10-fold cross-validation. This entire procedure is performed ten times, using different random number seeds, to reduce the chance of observing spurious results due to chance divisions of the data.

For studies with more than two factors, the steps of the MDR method are repeated for each possible model size (i.e. each number of loci and/or environmental factors), 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 maximizes the cross-validation consistency and minimizes the prediction error is selected. Cross-validation consistency is a measure of the number of times a particular set of loci and/or factors are identified across the cross validation subsets [Ritchie et al., 2001; Moore et al., 2002b; Moore, 2003] and is measured in the following way. For each 10-fold cross validation, the number of times the same set of loci/factors were identified across the 10 data subsets is recorded. The minimum value is one if the combination of factors occurs in only one subset and the maximum value is 10 if the same combination of loci/factors was identified across all 10 subsets. Thus, there are 10 possible values of the measure of consistency ranging from 1 to 10, where 10 is considered strong evidence in favor of a multifactor association. Prediction error is a measure of how well the MDR model predicts risk status in the independent test sets. The prediction error is calculated as the average of the prediction errors across each of the 10 cross validation subsets. Hypothesis testing of the best model(s) can then be performed by evaluating the magnitude of the cross-validation consistency and prediction error estimates using permutation testing. Here, the disease labels are randomized and the entire MDR analysis repeated.

The MDR method is currently limited to loci and environmental factors with two or three categories, but no more. Thus, it is ideally suited for SNPs with two alleles and discrete environmental factors or other covariates. Simulation studies have indicated that MDR has greater than 80% power to detect interactions in the absence of main effects for a variety of epistasis models, even in the presence of significant genotyping error and/or missing data [Ritchie et al., 2003]. It should be noted that power was significantly reduced when the data had significant locus heterogeneity. This is a common problem for all data analysis methods. Improving the power to detect epistasis in the presence of heterogeneity is an active area of investigation. An MDR software package is available for the UNIX and Linux operating systems and is described in detail by Hahn et al. [2003].

Application of MDR to case-control datasets has routinely yielded evidence of epistasis in the absence of main effects. For example, Ritchie et al. [2001] identified a statistically significant interaction among four SNPs from three estrogen metabolism genes for sporadic breast cancer. Again, this interaction was detected in the absence of independent main effects for any of the four SNPs. Evidence for epistasis has also been detected for other common disease such as essential hypertension [Moore and Williams, 2002].

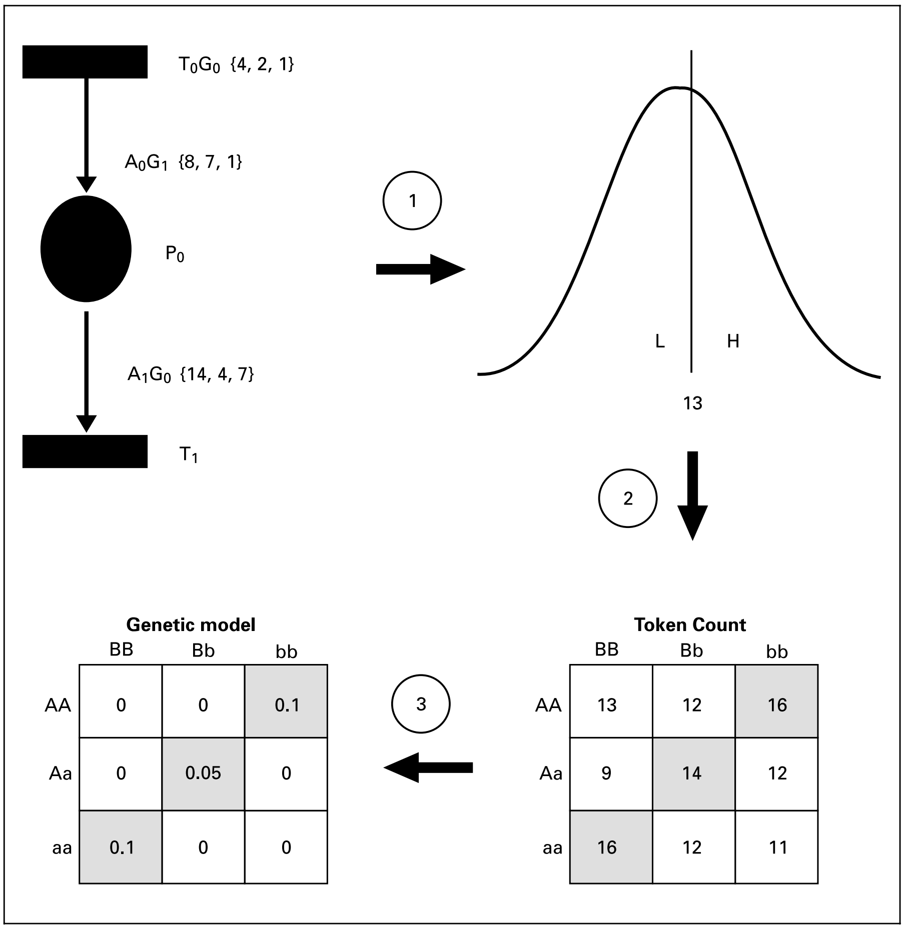
A concern with any data driven method is the increased potential for false-positives due to multiple looks at the data. It should be noted that MDR deals with multiple testing issues through combination of cross-validation and permutation testing. Cross-validation is useful because it is able to estimate the prediction error of a model. Models with false-positive loci usually do not generalize very well and thus will have a prediction error of approximately 0.50. Permutation testing is used to verify that the estimated prediction error would not be expected under the null hypothesis. In our experience with both real and simulated data, this combined computational approach greatly reduces type I errors. For example, we carried out both an MDR and a conditional logistic regression analysis of three SNPs from three candidate genes in a study from the Physician's Health Study consisting of 373 males with cardiovascular disease and 373 healthy male controls [Coffey et al., 2003a]. The logistic regression analysis identified a statistically significant interaction between two of the three SNPs. Although MDR identified evidence that the same two SNPs had an interactive effect, the estimated prediction error was not statistically different than the expected error of 0.50 under the null hypothesis. These results suggested that this result would not generalize to other datasets. Indeed, the same analysis was conducted in a second independent dataset from the same population and the positive logistic regression did not replicate. Thus, cross-validation and permutation testing are important for reducing type I errors. Unfortunately, they are rarely employed for evaluating gene-gene interactions results [Coffey et al., 2003b].

Finally, we would like to note that there is a concept in computer science called 'no free lunch'. The idea is that no one computational or statistical method will be optimal for every dataset. A successful data analysis will likely involve the combined use of multiple data analysis methods that have different strengths and weaknesses. While MDR and the other methods mentioned show promise for detecting epistasis, it is important to realize that more method development and evaluation is needed before a suite of complementary tools can be made available to the research community.

Biological Systems Modeling of Epistasis

As described above, the working hypothesis that epistasis is ubiquitous is partly based on the ubiquity of biomolecular interactions in biological systems. If we are to understand how statistical epistasis observed at the population level arises from biological epistasis, we need to develop strategies for documenting the transfer of genetic information through the hierarchy of genes, regulatory networks, biochemical systems, physiological systems, and clinical endpoints. Simulation of complex biological systems can play an important role in characterizing the relationship between statistical and biological epistasis by developing computational models of reality that can be used to perform thought experiments with the goal of generating testable hypotheses [Di Paolo et al., 2000]. Hypotheses derived from such simulations or from actual biological data can then be tested through perturbation of model organism systems [Jansen, 2003].

As a first step towards understanding the hierarchical relationship between genes, regulatory networks, biochemical systems, physiological systems, and clinical endpoints, we have developed an automated approach to generating Petri net models of both real and hypothetical biochemical systems that are consistent with patterns of epistasis as defined at the population level by penetrance functions [Moore and Hahn, 2003a, b, c]. As reviewed by



Genetic model

	BB	Bb	bb
AA	0	0	0.1
Aa	0	0.05	0
aa	0.1	0	0

Token Count

	BB	Bb	bb
AA	13	12	16
Aa	9	14	12
aa	16	12	11

Fig. 2. Summary of the steps involved in generating Petri net models of biochemical systems that are consistent with patterns of disease susceptibility among multilocus genotypes. In step one, a Petri net model is generated with genotype-dependent elements. In this example, the first transition (T_0) is dependent on genotypes (G_0) at locus 0. The values in the brackets represent the rate at which the transition fires for each of three genotypes. The arcs (A_0, A_1) are also genotype-dependent. Once a Petri net model is constructed, it is then executed for some number of iterations and the final count of the number of molecules (i.e. tokens) at place 0 (P_0) compared to a threshold value (e.g. 13). In step two, each Petri net is assigned a risk status. If the number of molecules is equal to or greater than the threshold value, that genotype combination is labeled 'high-risk'. If the number of molecules is less than the threshold value, that genotype combination is labeled 'low-risk'. In step three, the risk status of each genotype combination is compared to that from genetic model (e.g. a penetrance function). The number of correct assignments is used as a measure of the fitness of that Petri net. This process is iterated to yield a Petri net model that is consistent with the genetic model.

Desel and Juhas [2001], Petri nets are a type of directed graph that can be used to model discrete dynamical systems. Goss and Peccoud [1998] demonstrated that Petri nets could be used to model molecular interactions in biochemical systems. The goal of identifying Petri net models of biochemical systems that are consistent with observed patterns of epistasis is accomplished by developing Petri nets that are dependent on specific genotypes from two or more genetic variations. Here, specific components of the Petri net are genotype-dependent yielding different Petri net behavior. Each Petri net model is related to the genetic model using a discrete version of a threshold or liability model. For each model, the quantity of a given molecular species is recorded and, if a certain threshold is exceeded, a risk assignment is made. If the quantity does not exceed the threshold, the alternative risk assignment is made. In the studies by Moore and Hahn [2003a, b, c], the high-risk and low-risk assignments made by the discrete threshold from the output of the Petri net can then be compared to the high-risk and low-risk genotypes from the genetic model. A perfect match indicates the Petri net model is consistent with the pattern of epistasis observed in the genetic model. The Petri net then becomes a model that relates the genetic variations to risk of disease through an intermediate biochemical network. A detailed example is given in figure 2.

Moore and Hahn [2003a, b, c] have shown that this approach can routinely identify Petri net models of biochemical systems that are consistent with patterns of epistasis for a wide range of two- and three-locus epistasis models. What can be learned from these computational modeling exercises? Most importantly, we can generate hypotheses about the complexity of the biochemical systems underlying observed patterns of epistasis. For example, Moore and Hahn [2003a, b] found that extremely simple Petri net models could explain a variety of two-locus epistasis models. Most of the biochemical systems models discovered consisted of only one molecular specie with dynamic transitions to and from that specie. With the three-locus models [Moore and Hahn, 2003c], we discovered that at least one locus had pleiotropic effects in each of the biochemical systems models discovered. This was a ubiquitous property and might indicate that pleiotropy is a necessary feature of higher-order epistatic effects. Indeed, this has been observed for the simulation of gene regulatory networks [Gibson, 1996]. These results might provide some insight into the functional properties of actual biological systems. In the very least, interesting trends can become formal hypotheses that need be validated using observational data and experimental systems.

Conclusions

We have provided a review of several lines of evidence that support the formation of a working hypothesis that epistasis is a ubiquitous component of the genetic architecture of common human diseases and that complex interactions are more important than the independent main effects of any one susceptibility gene. While we have focused primarily of epistasis in this review, we fully recognize that many other phenomena such as genetic heterogeneity and phenocopy significantly contribute to the complexity of the genotype to phenotype mapping. The successful testing of the working hypothesis in genetic studies of human disease will be dependent on the availability of powerful statistical methods that are less susceptible to the limitations of traditional parametric approaches such as linear and logistic regression. Many of these statistical and computational tools are just now becoming available. Initial application of many of these new methods suggests that epistasis is commonly found. However, it will be important to make these tools available to the research community so they can be applied to the detection of epistasis effects for a number of different diseases. We anticipate future results will support this working hypothesis.

While the widespread use of new methods may validate the idea that epistasis is common, the future lies in understanding the etiology of epistasis at the biological level. Once we understand the biological mechanisms of epistasis, only then can we develop better diagnosis, prevention, and treatment strategies. Understanding the etiology of epistasis will require a combined computational and experimental approach. Computer simulations and modeling can generate important etiological hypotheses that can then be validated using model organisms. It is clear that a combined effort among statistical geneticists, epidemiologists, bioinformaticists, and molecular geneticists is needed to test this working hypothesis and benefit from its validation.

Acknowledgments

This work was supported by National Institutes of Health grants HL65234, HL65962, GM31304, AG19085, and AG20135. We would like to thank Lance Hahn, Marylyn Ritchie, Tricia Thornton, Bill White, and the attendees of the 2003 Oberwolfach Conference on Medical Statistics (Current Developments in Statistics Methodology for Genetic Architecture of Complex Diseases) for many stimulating discussions about the detection and characterization of epistasis. We would also like to thank two anonymous reviewers for their very thoughtful comments.

References

Altmuller J, Palmer LJ, Fischer G, Scherb H, Wjst M: Genomewide scans of complex human diseases: true linkage is hard to find. Am J Hum Genet 2001;69:936–950.

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

Bellman R: Adaptive Control Processes. Princeton, Princeton University Press, 1961.

Bernard C: Introduction à l'étude de la médecine expérimentale, par M. Claude Bernard. Paris, J.B. Baillière et fils, 1865.

Cheverud JM, Routman EJ: Epistasis and its contribution to genetic variance components. Genetics 1995;139:1455–1461.

Coffey CS, Hebert PR, Ritchie MD, Krumholz HM, Morgan TM, Gaziano JM, Ridker PM, Moore JH: 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. Submitted, 2003a.

Coffey CS, Hebert PR, Krumholz HM, Williams SM, Moore JH: Reporting of model validation procedures in studies of genetic interactions. Submitted, 2003b.

Concato J, Feinstein AR, Holford TR: The risk of determining risk with multivariable models. Ann Intern Med 1993;118:201–210.

Desel J, Juhas G: What is a Petri net? Informal answers for the informed reader; in Ehrig H, Juhas G (eds): Lecture Notes in Computer Science 2128. Berlin, Springer, 2001, pp 1–25.

Di Paolo EA, Noble J, Bullock S: Simulation models as opaque thought experiments; in Dedau MA, McCaskill JS, Packard NH, Rasmussen S (eds): Artificial Life VII: Proceedings of the Seventh International Conference on Artificial Life. Cambridge, The MIT Press, 2000.

Dipple KM, McCabe ER: Modifier genes convert 'simple' Mendelian disorders to complex traits. Mol Genet Metab 2000;71:43–50.

Finckh U: The future of genetic association studies in Alzheimer disease. J Neural Transm 2003; 110:253–266.

Fisher RA. The correlation between relatives on the supposition of Mendelian inheritance. Trans R Soc Edinburgh 1918;52:399–433.

Freitas AA: Understanding the crucial role of attribute interaction in data mining. Artif Intel Rev 2001;16:177–199.

Gallie DR: Protein-protein interactions required during translation. Plant Mol Biol 2002;50: 949–970.

Gibson G: Epistasis and pleiotropy as natural properties of transcriptional regulation. Theor Popul Biol 1996;49:58–89.

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

Goss PJ, Peccoud J: Quantitative modeling of stochastic systems in molecular biology by using stochastic Petri nets. Proc Natl Acad Sci USA 1998;95:6750–6755.

Griffiths AJF, Miller JH, Suzuki DT, Lewontin RC, Gelbart WM: An Introduction to Genetic Analysis. New York, WH Freeman, 2000.

Hahn LW, Ritchie MD, Moore JH: Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics 2003;19:376–382.

Hansen TF: Is modularity necessary for evolvability? Remarks on the relationship between pleiotropy and evolvability. BioSystems 2003;69: 83–94.

Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K: A comprehensive review of genetic association studies. Genet Med 2002;4:45–61.

Hoh J, Ott J: A train of thoughts on gene mapping. Theor Popul Biol 2001;60:149–153.

Hoh J, Wille A, Zee R, Cheng S, Reynolds R, Lindpaintner K, Ott J: Selecting SNPs in two-stage analysis of disease association data: A model-free approach. Ann Hum Genet 2000;64:413–417.

Hollander WF: Epistasis and hypostasis. J Hered 1955;46:222–225.

Hosmer DW, Lemeshow S: Applied Logistic Regression. New York, John Wiley & Sons Inc., 2000.

Jansen RC: Studying complex biological systems using multifactorial perturbation. Nat Rev Genet 2003;4:145–151.

Kerem E, Corey M, Kerem BS, Rommens J, Markiewicz D, Levison H, Tsui LC, Durie P: The relation between genotype and phenotype in cystic fibrosis-analysis of the most common mutation (delta F508). N Engl J Med 1990;323: 1517–1522.

Kooperberg C, Ruczinski I, LeBlanc ML, Hsu L: Sequence analysis using logic regression. Genet Epidemiol 2001;21:S626–S631.

Leamy LJ, Routman EJ, Cheverud JM: An epistatic genetic basis for fluctuating asymmetry of mandible size in mice. Evolution 2002;56:642–653.

Martinez E: Multi-protein complexes in eukaryotic gene transcription. Plant Mol Biol 2002;50: 925–947.

Michal G: Biochemical Pathways: An Atlas of Biochemistry and Molecular Biology. New York, Wiley, 1999.

Moore JH: Cross validation consistency for the assessment of genetic programming results in microarray studies; in Raidl G, Meyer J-A, Middendorf M, Cagnoni S, Cardalda JJR, Corne DW, Gottlieb J, Guillot A, Hart E, Johnson CG, Marchiori E (eds): Lecture Notes in Computer Science 2611. Berlin, Springer-Verlag, 2003, pp 99–106.

Moore JH, Hahn LW: A cellular automata approach to detecting interactions among single-nucleotide polymorphisms in complex multifactorial diseases. Pac Symp Biocomput 2002; 7:53–64.

Moore JH, Hahn LW: Cellular automata and genetic algorithms for parallel problem solving in human genetics; in Merelo JJ, Panagiotis A, Beyer H-G (eds): Lecture Notes in Computer Science 2439. Berlin, Springer-Verlag, 2002, pp 821–830.

Moore JH, Hahn LW: Grammatical evolution for the discovery of Petri net models of complex genetic systems; in Cantu-Paz E, et al (eds): Lecture Notes in Computer Science, Berlin, Springer-Verlag, 2003, pp 2412–2413.

Moore JH, Hahn LW: Evaluation of a discrete dynamic systems approach for modeling the hierarchical relationship between genes, biochemistry, and disease susceptibility. Discrete Contin Dyn Sys, 2003b, in press.

Moore JH, Hahn LW: Petri net modeling of high-order genetic systems using grammatical evolution. BioSystems, 2003c, in press.

Moore JH, Lamb JM, Brown NJ, Vaughan DE: A comparison of combinatorial partitioning and linear regression for the detection of epistatic effects of the ACE I/D and PAI-1 4G/5G polymorphisms on plasma PAI-1 levels. Clin Genet 2002a;62:74–79.

Moore JH, Parker JS, Olsen NJ, Aune T: Symbolic discriminant analysis of microarray data in autoimmune disease. Genet Epidemiol 2002b;23: 57–69.

Moore JH, Smolkin ME, Lamb JM, Brown NJ, Vaughan DE: The relationship between plasma t-PA and PAI-1 levels is dependent on epistatic effects of the ACE I/D and PAI-1 4G/5G polymorphisms. Clin Genet 2002;62:53–59.

Moore JH, Williams SM: New strategies for identifying gene-gene interactions in hypertension. Ann Med 2002;34:88–95.

Morch ET: Chondrodystrophic Dwarfs in Denmark. Copenhagen, Munksgaard, 1941.

Neel JV, Schull WJ: Human Heredity. Chicago, University of Chicago Press, 1954.

Nelson MR, Kardia SL, Ferrell RE, Sing CF: A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. Genome Res 2001; 11:458–470.

Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR: A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996;49:1373–1379.

Phillips PC: The language of gene interaction. Genetics 1998;149:1167–1171.

Pigliucci M: Phenotypic Plasticity. Baltimore, The Johns Hopkins Press, 2001.

Pogun S: Are attractors 'strange', or is life more complicated than the simple laws of physics? BioSystems 2001;63:101–114.

Rice SH: The evolution of canalization and the breaking of von Baer's laws: Modeling the evolution of development with epistasis. Evolution 1998;52:647–656.

Ritchie MD, Hahn LW, Moore JH: Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. Genet Epidemiol 2003;24: 150–157.

Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH: Multifactor dimensionality reduction reveals high-order interactions among estrogen metabolism genes in sporadic breast cancer. Am J Hum Genet 2001; 69:138–147.

Ritchie MD, White BC, Parker JS, Hahn LW, Moore JH: Optimization of neural network architecture using genetic programming improves detection and modeling of gene-gene interactions in studies of human diseases. BMC Bioinformatics 2003, in press.

Roberts DF: Fertility, mortality and culture: the changing pattern of natural selection. In: The Role of Natural Selection in Human Evolution. New York, American Elsevier Publishing Company, 1975.

Salvatore F, Scudiero O, Castaldo G: Genotype-phenotype correlation in cystic fibrosis: The role of modifier genes. Am J Med Genet 2002; 111:88–95.

Shull GH: Duplicate genes for capsule form in BURSA bursa Bastoris. J Ind Abst Vererb 1914;12:97–149.

Smith JM: Evolutionary Genetics. New York, Oxford University Press, 1998.

Templeton AR: Epistasis and complex traits; in Wolf J, Brodie III B, Wade M (eds): Epistasis and the Evolutionary Process. New York, Oxford University Press, 2000.

Tyson JJ, Chen KC, Novak B: Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. Curr Opin Cell Biol 2003;15:221–223.

Waddington CH: Canalization of development and the inheritance of acquired characters. Nature 1942;150:563–565.

Waddington CH: The Strategy of the Genes. New York, MacMillan, 1957.

Wade MJ, Winther RG, Agrawal AF, Goodnight CJ: Alternative definitions of epistasis: dependence and interaction. Trends Ecol Evol 2001; 16:498–504.

Wagner A: Robustness against mutations in genetic networks of yeast. Nat Genet 2000;24:355–361.

Williams SM, Addy JH, Phillips JA 3rd, Dai M, Kpodonu J, Afful J, Jackson H, Joseph K, Eason F, Murray MM, Epperson P, Aduonum A, Wong LJ, Jose PA, Felder RA: Combinations of variations in multiple genes are associated with hypertension. Hypertension 2000;36: 2–6.

Wolf JB, Brofie III ED, Wade MJ: Epistasis and the Evolutionary Process. New York, Oxford University Press, 2000.

Wright S: The role of mutation, inbreeding, crossbreeding and selection in evolution. Proc 6th Intl Congr Genet 1932;1:356–366.

Wright S: Physiological and evolutionary theories of codominance. Am Nat 1934;68:25–53.

Zee RY, Hoh J, Cheng S, Reynolds R, Grow MA, Silbergleit A, Walker K, Steiner L, Zangenberg G, Fernandez-Ortiz A, Macaya C, Pintor E, Fernandez-Cruz A, Ott J, Lindpainter K: Multi-locus interactions predict risk for post-PTCA restenosis: An approach to the genetic analysis of common complex disease. Pharmacogenomics J 2002;2:197–201.