Chapter II
Genome-Wide Analysis of Epistasis Using Multifactor Dimensionality Reduction: Feature Selection and Construction in the Domain of Human Genetics

Jason H. Moore
Dartmouth Medical School, USA

ABSTRACT

Human genetics is an evolving discipline that is being driven by rapid advances in technologies that make it possible to measure enormous quantities of genetic information. An important goal of human genetics is to understand the mapping relationship between interindividual variation in DNA sequences (i.e., the genome) and variability in disease susceptibility (i.e., the phenotype). The focus of the present study is the detection and characterization of nonlinear interactions among DNA sequence variations in human populations using data mining and machine learning methods. We first review the concept difficulty and then review a multifactor dimensionality reduction (MDR) approach that was developed specifically for this domain. We then present some ideas about how to scale the MDR approach to datasets with thousands of attributes (i.e., genome-wide analysis). Finally, we end with some ideas about how nonlinear genetic models might be statistically interpreted to facilitate making biological inferences.

THE PROBLEM DOMAIN: HUMAN GENETICS

Human genetics can be broadly defined as the study of genes and their role in human biology. An important goal of human genetics is to understand the mapping relationship between interindividual variation in DNA sequences (i.e., the genome) and variability in disease susceptibility (i.e., the phenotype). Stated another way, how does one or more changes in an individual's DNA sequence increase or decrease their risk of

developing a common disease such as cancer or cardiovascular disease through complex networks biomolecules that are hierarchically organized and highly interactive? Understanding the role of DNA sequences in disease susceptibility is likely to improve diagnosis, prevention and treatment. Success in this important public health endeavor will depend critically on the degree of nonlinearity in the mapping between genotype to phenotype. Nonlinearities can arise from phenomena such as locus heterogeneity (i.e., different DNA sequence variations leading to the same phenotype), phenocopy (i.e., environmentally determined phenotypes), 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*). It is this latter source of nonlinearity, *epistasis*, that is of interest here. Epistasis has been recognized for many years as deviations from the simple inheritance patterns observed by Mendel (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 (i.e., redundant) gene networks (Gibson & Wagner, 2000; Waddington, 1942, 1957; Proulx & Phillips, 2005).

Epistasis has been defined in multiple different ways (e.g., Brodie, 2000; Hollander, 1955; Philips, 1998). We have reviewed two types of epistasis, biological and statistical (Moore & Williams, 2005). Biological epistasis results from physical interactions between biomolecules (e.g., DNA, RNA, proteins, enzymes, etc.) and occur 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 & Williams, 2005).

The focus of the present study is the detection and characterization of statistical epistasis in human populations using data mining and machine learning methods. We first review the concept difficulty and then review a multifactor dimensionality reduction (MDR) approach that was developed specifically for this domain. We then present some ideas about how to scale the MDR approach to datasets with thousands of attributes (i.e., genome-wide analysis). Finally, we end with some ideas about how nonlinear genetic models might be statistically interpreted to facilitate making biological inferences.

CONCEPT DIFFICULTY

Epistasis can be defined as biological or statistical (Moore & Williams, 2005). Biological epistasis occurs at the cellular level when two or more biomolecules physically interact. In contrast, statistical epistasis occurs at the population level and is characterized by deviation from additivity in a linear mathematical model. Consider the following simple example of statistical epistasis in the form of a penetrance function. Penetrance is simply the probability (P) of disease (D) given a particular combination of genotypes (G) that was inherited (i.e., P[D|G]). A single genotype is determined by one allele (i.e., a specific DNA sequence state) inherited from the mother and one allele inherited from the father. For most single nucleotide polymorphisms or SNPs, only two alleles (e.g., encoded by *A* or *a*) exist in the biological population. Therefore, because the order of the alleles is unimportant, a genotype can have one of three values: *AA*, *Aa* or *aa*. The model illustrated in Table 1 is an extreme example of epistasis. Let's assume that genotypes *AA*, *aa*, *BB*, and *bb* have population frequencies of 0.25 while genotypes *Aa* and *Bb* have frequencies