Gene-gene and gene-environment interactions underlying complex traits and their detection

**Abbreviations:** MDR, multifactor dimensionality reduction method; CART, classification and regression trees; MARS, multivariate adaptive regression spline; GPNN, genetic programming optimized neural network

**Introduction**

No genes or environmental factors are isolated from the interactive genomic and epigenomic networks in shaping a biological phenotype. Non intuitivity and nonlinearity are a natural property of the network’s architecture. Consequently, the existence of interactions among genes, called gene-gene (also known as epistatic) interactions, and between genes and environmental factors (broadly defined as all non-genetic exposures), called gene-environment (GE) interactions, is the normal rather than an exception. Several converging lines of evidence have pointed to the dominant role of interactions in the inherited traits; in particular, epistatic and GE interactions are considered as one of the primary culprits for missing heritability, referred to the majority of the genetic variation that is not yet identified by the more than a decade’s practice of genome-wide association studies. Identification of background-specific factors among genes in combination with lifestyles and environmental exposures is an important scientific topic in genetics, breeding, and genetic epidemiology.

A high degree of context dependence of genetic architecture likely results in a relatively weak marginal genotype-phenotype correlations for complex traits, making traditional univariate approaches that test for association one factor at a time futile. The multi factorial strategies are thus critical in hunting highly mutually dependent factors underlying a trait. However, such a search has to face a significant obstacle called “the curse of dimensionality”, a problem caused by the exponential increase in volume of possible interactions with the number of factors to consider. The conventional regression methods, established by the extension under the concept of single factor-based approaches, are hardly appropriate for tackling ubiquitous yet elusive interactions because of several problems, heavy computational burden (usually computationally intractable), increased Type I and II errors, and reduced robustness and potential bias as a result of highly sparse data in a multi factorial model. Diverse novel approaches such as data mining and machine learning have been explored recently for various kinds of phenotypes, namely, Bayesian belief network, tree-based algorithms including multivariate adaptive regression spline (MARS), classification and regression trees (CART) or recursive partitioning methods and random forests approach, pattern recognition approaches including neural network strategies such as the parameter decreasing method (PDM) and genetic programming optimized neural network (GPNN), genetic algorithm strategies, and cellular automata (CA) approach, support vector machine (SVM), penalized regression, and Bayesian methods.

Among these methods emerged recently, data reduction approaches (a constructive induction strategy) such as the multifactor dimensionality reduction method (MDR), the combinatorial partitioning method, and the restricted partition method are promising to address the multidimensionality problems. Rather than modeling the interaction term per se as with regression methods, a data reduction strategy seeks for a pattern in a combination of factors/attributes of interest that maximizes the phenotypic variation it explains. It treats the joint action as a whole, coinciding to the very original epistasis coined by Bateson, offering a solution that avoids decomposition as in regression methods where the number of interaction parameters grows exponentially as each new variable is added. It also has a straightforward correspondence to the concept of the phenotypic landscape that unifies biological, statistical genetics and evolutionary theories. Notably the pioneering MDR method has sustained its popularity in detection of interactions since its launch.

Several extensions of the MDR have been made for analyzing different traits, e.g., binary, count, continuous, polytomous, ordinal, time-to-onset, multivariate and others, as well as combinations of those, and also entertaining various study designs including homogeneous and admixed unrelated-subject and family as well as mixtures of them. Such extensions include to inclusion of covariates, to survival data, to multivariate phenotypes, to multi-categorical or ordinal phenotypes, to case-control study in structured populations, to family study, and to unified analysis of both unrelated and related samples. With these extensions, the MDR-type methods offer a powerful tool for handling the breadth of data types and addressing statistical issues associated with study design and sampling scheme.

Despite the methodological progresses in detection of multifactor interactions, there are still difficult computational challenges and multiple hypothesis testing problems in practice, especially for detecting high-order interactions for the large-scale such as whole genome data. Further theoretical and computational work is required for effective identification of interacting factors underlying the complex traits.

**Acknowledgement**

The author thanks Guo-Bo Chen, Hai-Ming Xu, Xi-Wei Sun, and Lei Yan for their contributions to the development of GMDR. This project was supported in part by NIH Grant DA025095 to X.-Y.L.

**Conflict of interest**

The author declares no conflict of interest.
Funding

None.

References


40. Bateson W. Mendel’s principles of heredity. UK: Cambridge University Press; 1909.


