Reconstructability Analysis As A Tool For Identifying Gene-Gene Interactions In Studies Of Human Diseases

Stephen Shervais
College of Business and
Public Administration
Eastern Washington University
Cheney, WA 99004
sshervais@ewu.edu

Martin Zwick
Systems Science Ph.D. Program
Portland State University
Portland, OR 97207
zwick@pdx.edu

Patricia Kramer
Department of Neurology
Oregon Health and
Sciences University
Portland, OR 97239
kramer@ohsu.edu

Abstract – *There are a number of human diseases that are caused by the epistatic interaction of multiple genes. Detecting these interactions with standard statistical tools is difficult, because there may be an interaction effect, but minimal or no main effect. Reconstructability analysis uses Shannon's information theory to detect relationships between variables in categorical datasets. We apply reconstructability analysis to data generated by five different models of gene-gene interaction, with heritability levels from 0.053 to 0.008, using 200 controls and 200 cases. We find that even with heritability levels as low as 0.008, and with the inclusion of 50 non-associated genes in the dataset, we can identify the interacting gene pairs with an accuracy of 80% or better.*

Keywords : Epistasis, reconstructability analysis, information theory, gene-gene interaction, gene interaction modeling, Occam, genetics.

1 Introduction

Many human diseases are caused by the epistatic interaction of multiple genes. Detecting these interactions with standard statistical tools is difficult, because there may be an interaction effect, but minimal or no main effect.

Great advances have been made over the last two decades in developing analytic methods for detecting single genes that are necessary, and often sufficient, to cause human diseases. For the most part, however, diseases with such a "simple" genetic etiology are relatively rare. Common diseases (e.g., hypertension, cancer, dementia) are the result of DNA sequence variations in multiple genes, many of which interact in a non-additive, or "epistatic", fashion, and thus have a substantially more complex genetic etiology. Detecting gene-gene interactions with traditional analytic methods is problematic in the absence of main effects and the presence of sizeable interaction effects. As the number of possible candidate genes to consider increases, the number of combinations likewise increases, and one is soon faced with what has been referred to by Bellman [1] as "the curse of dimensionality." In this paper we show that for interactions involving only a pair of genes, this problem may well be manageable.

Elsewhere in the literature, epistatic interaction of two genes in a synthetic dataset has been detected by first creating a tool that could solve the problem of predicting the disease based on gene data, and then examining the structure of the solution to determine which genes were contributing to it. For example, Hahn [2], used this approach with multifactor dimensionality reduction, while Ritchie [3] applied it with artificial neural networks. With reconstructability analysis (RA), we examine the structure of the data directly to detect relationships in subsets of the variables, and use these to define the structural model. If the purpose of the exercise is disease prediction, the variables identified in the structural model can then be used in, for example, a neural network; or RA might itself be used for this purpose.

We use the term *model* in two different ways in this paper. When we speak of a *genetic* model, we refer to the penetrance tables (1-5) given below. Penetrance is defined as the probability that an individual gets a disease, given that he/she has a particular constellation of alleles at the two interacting gene loci. This probability is a function of the respective allele frequencies. When we speak of a *structural* model, we refer to models generated by the RA software, some of which are better descriptors of the data than others.

We start by describing recent approaches to the simulation of gene interactions. Then we briefly discuss the theory behind reconstructability analysis. Next, we describe the genetic models we used, and the datasets they generated. Finally, we show that RA successfully identifies the interacting gene pairs in a range of situations.

2 Gene Interaction Modeling

In 2002, Moore [4], introduced a genetic algorithm tool for producing genetic models characterized by equal marginal penetrance values for all gene pairs and maximum variance among penetrance values. Five of the genetic models represented in that paper were later used by Ritchie

Presented at IEEE Systems, Man, and Cybernetics meeting, Hawaii, Oct 10-12, 2005

[3], to test their genetic programming approach to gene interaction modeling. These models, and the heritability values calculated by Ritchie are shown in tables 1-5. In each model, each cell represents the probability of disease given the particular combination of genotypes, for example, p(D | gene$_1$, gene$_2$), where gene$_1$ has alleles GG, Gg, and gg, and gene$_2$ has alleles HH, Hh, and hh.

All models are in Hardy-Weinberg equilibrium, in which the frequency of any particular genotype is determined by the product of the frequencies of alleles involved and not by evolutionary forces such as natural selection or sampling error. The frequencies of the two alleles at each gene are equal., which maximizes genotype variation. Given these two conditions, all five models exhibit significant interaction effects, but no marginal gene effects. Models were not designed with reference to any predetermined biological considerations. These five genetic models were used to test the ability of reconstructability analysis to detect gene-gene interactions.

Table 1. Model 1 Penetrance Values (heritability = 0.053)

	Table Penetrance			Margin penetrance
	GG (.25)	*Gg* (.50)	*gg* (.25)	
HH (.25)	0.00	0.10	0.00	0.05
Hh (.50)	0.10	0.00	0.10	0.05
hh (.25)	0.00	0.10	0.00	0.05
Margin Penetrance	0.05	0.05	0.05	

Table 2. Model 2 Penetrance Values (heritability = 0.051)

	Table penetrance			Margin penetrance
	GG (.25)	*Gg* (.50)	*gg* (.25)	
HH (.25)	0.00	0.00	0.10	0.025
Hh (.50)	0.00	0.05	0.00	0.025
hh (.25)	0.10	0.00	0.00	0.025
Margin penetrance	0.025	0.025	0.025	

Table 3. Model 3 Penetrance Values (heritability = 0.026)

	Table penetrance			Margin penetrance
	GG (.25)	*Gg* (.50)	*gg* (.25)	
HH (.25)	0.00	0.04	0.00	0.02
Hh (.50)	0.04	0.02	0.00	0.02
hh (.25)	0.00	0.00	0.08	0.02
Margin penetrance	0.02	0.02	0.02	

Table 4. Model 4 Penetrance Values (heritability = 0.012)

	Table penetrance			Margin penetrance
	GG (.25)	*Gg* (.50)	*gg* (.25)	
HH (.25)	0.00	0.02	0.08	0.03
Hh (.50)	0.05	0.03	0.01	0.03
hh (.25)	0.02	0.04	0.02	0.03
Margin penetrance	0.03	0.03	0.03	

Table 5. Model 5 Penetrance Values (heritability = 0.008)

	Table penetrance			Margin penetrance
	GG (.25)	*Gg* (.50)	*gg* (.25)	
HH (.25)	0.00	0.04	0.08	0.04
Hh (.50)	0.06	0.04	0.02	0.04
hh (.25)	0.04	0.04	0.04	0.04
Margin penetrance	0.04	0.04	0.04	

These five genetic models were used to test the ability of reconstructability analysis to detect gene-gene interactions.

3 Reconstructability Analysis

Reconstructability analysis (RA) derives from Ashby [5], and was developed by Broekstra, Cavallo, Cellier, Conant, Jones, Klir, Krippendorff, and others. An account of its origin is given by Klir [6]; an extensive bibliography is available in [7], and a compact summary of RA is available in publications by Zwick [8], [9]. RA resembles log-linear (LL) methods [10], used widely in the social sciences, and where RA and LL methodologies overlap they are equivalent [11], [12]. In RA [13], a probability or frequency distribution or a set-theoretic relation is decomposed (compressed, simplified) into component distributions or relations. When applied to the decomposition of frequency or probability distributions, RA does statistical analysis. RA can model problems both where "independent variables" (inputs) and "dependent variables" (outputs) are distinguished (called *directed systems*) and where this distinction is not made (*neutral systems*).

For example, assume we have a directed system, with inputs (genes) A, B, C and D, and output the disease status, Z. Consider an observed frequency distribution f(A, B, C, D, Z) which we write more simply as ABCDZ. RA decomposes such distributions into sets of projections such as ABCD and ABZ, which taken together define a structural model ABCD:ABZ, which is less complex (has fewer degrees of freedom) than the data. This model defines a *calcu-*

lated distribution $ABCDZ_{ABCD:ABZ}$, obtained by maximum entropy composition of the ABCD and ABZ projections, that can be compared with the observed ABCDZ. The ABCD:ABZ model asserts that A and B jointly predict Z, while C and D have no predictive relationship with Z. (The ABCD component of the model allows relationships between the inputs themselves, but we are not interested in such relationships.) If ABCD:ABZ is a good model, one could equivalently say that the "transmission" or "mutual information" T(AB:Z) is high, while the transmission T(CD:Z) is zero. By contrast, the model ABCD:Z, called the "independence model" asserts that no input predicts Z. The data itself, ABCDZ, also called the "saturated model," asserts that all four inputs jointly predict Z. Thus ABCDZ has 100% of the predictive information in the data; ABCD:Z has 0% of the predictive information in the data; and all models between these two have some intermediate measure of information content. The information content of any model can be assessed for statistical significance with the Chi-square distribution.

There are different classes of RA models. Models such as ABCDZ, ABCD:ABZ, and AB:Z all have a "single predicting component," i.e., only one component that includes the output, Z. Such models are "loopless," and a loopless model essentially picks out a single subset of the inputs that predicts the output. By contrast, the model ABCD:ABZ:CDZ asserts that Z is predicted by A and B jointly and also, separately, by C and D jointly. Models with more than one predicting components have loops. RA calculations for models without loops are simpler and faster, and all structural models considered in this paper are loopless. For simplicity, the non-predicting component (in the present case, ABCD) will from now on be omitted from the model specification.

Calculations were made using the RA software programs developed at Portland State University, now integrated into the package Occam (for the principle of parsimony and as an acronym for "Organizational Complexity Computation And Modeling"). The earliest of these programs was developed by Zwick and Hosseini [14]; a review of RA methodology is offered in [8], [9]; a list of recent RA papers in the PSU group is given in [15]. A description of the OCCAM architecture is given in [16].

4 Methodology

The methodology used in this paper differs significantly from that described elsewhere in the literature. The standard approach is to develop a tool to predict the presence or absence of the disease as accurately as possible, and then to examine the structure of the resulting tool. This requires a training/testing process, with many-fold validation. For example, [17] used genetic programming techniques to create artificial neural networks that could separate cases from controls, and then reported out the genes used as inputs by the NNs. Our approach is more direct. We measure the information content of the dataset directly, and then select a model based on that. This means we have no training/testing dataset as such, and do not require cross-validation.

For this study, we selected five models of gene-gene interactions from the literature, [4], [17], and created thirty datasets from each genetic model for each of two conditions. First, for comparability with existing work, we used two associated genes (actually, single nucleotide poly morphisms – SNPs) and eight noise SNPs. Second, to demonstrate the scalability of the approach, we used two associated SNPs and 50 noise SNPs. Unlike most machine learning problems, which seek to identify patterns in a random selection of the general population, medical datasets are often divided into two equal groups: cases and controls. Cases are subjects who are known to have the condition at issue. Controls are selected from the general population, and are known to not have the condition. The distinction is important when using penetrance tables to develop datasets, because the two groups must be handled differently. Control allele patterns appear with the frequencies associated with the general population, and show zero penetrance. For example, pattern XX should appear for approximately 25%, and XX/YY for approximately 6.25% of the controls. The allele patterns for the cases appear with frequencies associated with both the relative penetrance and the overall population frequencies. If, for example, allele pattern XX/YY had a penetrance of 10%, and allele pattern Xx/Yy had a 20% penetrance, then the proportion of XX/YY in the cases should be 12.5% = (.25 * .25 * .10)/(.50 * .50 * .20). For both groups, the allele patterns of the non-contributing genes can be assigned at random.

We developed 30 datasets for each of the five genetic models and each of the two noise inputs. The interacting genes were assigned on the basis of the penetrance tables, while the values of noise SNPs were assigned at random.

The datasets were then run through the PSU Occam software. By specifying suitable parameters, Occam searches through the Lattice of Structures for structural models of a particular class which have high information content. In the present instance, Occam was asked simply to output all loopless models with two predicting inputs, ordered by their information content.

Table 6 (below) shows an example of Occam output for Dataset 11 of Model 5. Column 1 identifies the data model. Column 2 identifies the information content of the subject model for the 8 noise SNP dataset. Column 3 identifies the information content of the subject model for the 50 noise SNP dataset. The two active genes are A and B. The table is arranged in decreasing information content order, based on the 8 noise SNP results. The first five rows are the top five models identified in the 8 noise SNP dataset. Model HZ was the single-gene model with the highest information, and the information levels of models AZ and BZ are included to show that using even a correct single gene provides no usable information.

Table 6. Sample Occam Output, Model 5

	Information Content	
MODEL	8 Noise SNP	50 Noise SNP
ABZ	0.0722	0.0662
ACZ	0.0411	0.0038
CGZ	0.0236	0.0143
DJZ	0.0206	0.0073
BGZ	0.0080	0.0042
HZ	0.0061	0.0036
AZ	0.0016	0.0007
BZ	0.0006	0.0002
Z	0.0000	0.0000

5 Results and Discussion

A summary of the results is shown in Table 7. Column 1 indicates the genetic model used, and Column 2 lists the heritability of the disease in the genetic model (from [1] and [2]). Columns 3 and 4 indicate the percentage of the thirty datasets in which RA was able to identify the correct gene model. In all but the lowest heritability model, RA was able to consistently identify the two active SNPs, in the first experiment out of a total of ten, and in the second, out of a total of 52, based on datasets containing 200 cases and 200 controls. Even for the lowest heritability model and the highest number of noise SNPs, RA was successful in identifying the two active genes 80% of the time. This compares favorably with previous work, which (with only eight noise genes) found one of the two active genes 47% of the time [17], and both genes 19% of the time [2].

Table 7. Effectiveness of reconstructability analysis in identifying gene-gene interactions with both active and inactive genes

Genetic Model	Heritability	% With Both Active Genes in the top RA Model (8 noise SNPs, n = 30)	% With Both Active Genes in the top RA Model (50 noise SNPs n = 30)
1	0.053	100%	100%
2	0.051	100%	100%
3	0.026	100%	100%
4	0.012	100%	100%
5	0.008	93%	80%

Ours is essentially a brute force approach. Since the genetic models were designed with no main effect, no single SNP is linked to the disease more than any other single SNP, so no single SNP measure gives any clue about the interaction effect involving the two active SNPs. One cannot, therefore, reduce the set of SNPs to consider by looking at any single SNP. However, *pairs* of SNPs that don't include *both* active SNPs will also show no effect, so one can simply look at all pairs of SNPs to find the active pair. Using the Occam software, this is not a particularly burdensome calculation; processing time for 400 records and 10 SNPs takes less than a second, and 400 records and 52 SNPs takes approximately 4 seconds on a Pentium-class PC.

In summary, reconstructability analysis can readily detect gene-gene interactions that predict disease in the absence of any main (single gene) effect and in the presence of noise genes.

BIBLIOGRAPHY

[1] R. Bellman, *Adaptive Control Processes*. Princeton University Press, Princeton, NJ, 1961.

[2] L. Hahn, M. Ritchie and J. Moore, “Multifactor dimensionality reduction software for detecting gene–gene and gene–environment interactions,” *Bioinformatics*, Vol. 19, No. 3, pp. 376-382, 2003.

[3] M. Ritchie, C. Coffey, and J. Moore, “Genetic Programming Neural Networks as a Bioinformatics Tool for Human Genetics,” Proceedings of the Genetic and Evolutionary Computation Conference, Seattle, pp. 438-448, 2004.

[4] J. Moore, L. Hahn, M. Ritchie, T. Thornton, and B. White, “Application of genetic algorithms to the discovery of complex genetic models for simulations studies in human genetics,” Proceedings of the Genetic and Evolutionary Algorithm Conference, New York, pp. 1150-1155, 2002.

[5] W. R. Ashby, “Constraint Analysis of Many-Dimensional Relations.” *General Systems Yearbook*, vol. 9, pp. 99-105, 1964.

[6] G. Klir, “Reconstructability Analysis: An Offspring of Ashby’s Constraint Theory”, *Systems Research*, 3 (4), pp. 267-271, 1986.

[7] G. Klir, ed. *Int. J. General Systems* Special Issue on GSPS, Vol. 24, Number 1 & 2, 1996. (includes an RA bibliography).

[8] M. Zwick, M. (2001) "Wholes and Parts in General Systems Methodology", in Wagner, G., ed., *The Character Concept in Evolutionary Biology*, Academic Press, New York pp. 237-256. 2001. http://www.sysc.pdx.edu/faculty/Zwick/research.html#wholes

[9] M. Zwick, "An Overview of Reconstructability Analysis", *Kybernetes*, Vol. 33, Number 5/6, pp. 877-905, 2004.; originally in *Proceedings of 12th International World Organization of Systems and Cybernetics and 4th International Institute for General Systems Studies Workshop*, Pittsburgh, March 24-26, 2002. http://www.sysc.pdx.edu/download/papers/ldlpitf.pdf

[10] Y. Bishop, S. Feinberg, and P. Holland, *Discrete Multivariate Analysis*. (MIT Press, Cambridge, MA, 1978.

[11] D. Knoke, P.J. Burke *Log-Linear Models*: *Quantitative Applications in the Social Sciences Monograph # 20*. Sage, Beverly Hills, CA, 1980.

[12] K. Krippendorff, *Information Theory: Structural Models for Qualitative Data: Quantitative Applications in the Social Sciences #62*, Sage, Beverly Hills, CA, 1986.

[13] G. Klir, *The Architecture of Systems Problem Solving*, Plenum Press, New York, NY, 1985.

[14] J. Hosseini, R. R. Harmon, and M. Zwick, "Segment Congruence Analysis Via Information Theory," in *Proceedings, International Society for General Systems Research*, Philadelphia, pp. G62 - G77, 1986.

[15] M. Zwick, "*Discrete Multivariate Modeling*" 2005. http://www.sysc.pdx.edu/res_struct.html

[16] K. Willett, and M. Zwick, "A Software Architecture for Reconstructability Analysis", *Kybernetes*, Vol. 33, Number 5/6, pp. 997-1008, 2004; originally in *Proceedings of 12th International World Organization of Systems and Cybernetics and 4th International Institute for General Systems Studies Workshop*, Pittsburgh, 2002. http://www.sysc.pdx.edu/download/papers/kenpitf.pdf

[17] M. Ritchie, B. White, J. Parker, L. W. Hahn and J. H. Moore, "Optimization of neural network architecture using genetic programming improves detection and modeling of gene-gene interactions in studies of human diseases," *BMC Bioinformatics*, 4:28, 2003. http://www.biomedcentral.com/1471-2104/4/28