

# *Statistical Applications in Genetics and Molecular Biology*

---

*Volume 9, Issue 1*

2010

*Article 18*

---

## Reconstructability Analysis as a Tool for Identifying Gene-Gene Interactions in Studies of Human Diseases

Stephen Shervais\*

Patricia L. Kramer<sup>†</sup>

Shawn K. Westaway<sup>‡</sup>

Nancy J. Cox\*\*

Martin Zwick<sup>††</sup>

\*Eastern Washington University, sshervais@ewu.edu

<sup>†</sup>Oregon Health & Science University, kramer@ohsu.edu

<sup>‡</sup>Oregon Health & Science University, westaway@ohsu.edu

\*\*University of Chicago, ncox@medicine.bsd.uchicago.edu

<sup>††</sup>Portland State University, zwick@pdx.edu

# Reconstructability Analysis as a Tool for Identifying Gene-Gene Interactions in Studies of Human Diseases\*

Stephen Shervais, Patricia L. Kramer, Shawn K. Westaway, Nancy J. Cox, and Martin Zwick

## Abstract

There are a number of common human diseases for which the genetic component may include an 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 (RA) uses Shannon's information theory to detect relationships between variables in categorical datasets. We applied RA to simulated data for five different models of gene-gene interaction, and 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  $\geq 80\%$ . We applied RA to a real dataset of type 2 non-insulin-dependent diabetes (NIDDM) cases and controls, and closely approximated the results of more conventional single SNP disease association studies. In addition, we replicated prior evidence for epistatic interactions between SNPs on chromosomes 2 and 15.

**KEYWORDS:** epistasis, reconstructability analysis, information theory, gene interaction modeling, OCCAM, genetics, bioinformatics

---

\*This work was supported by the National Institutes of Health (AG026916 to PK, U01 HL08471 and P60 DK20595 to NC).

## INTRODUCTION

Significant advances have been made over the last two decades in developing analytic methods and bioinformatics tools for detecting single genes that are necessary for, or contribute to, human diseases. For the most part, however, diseases with a “simple” genetic etiology are relatively rare. Common diseases (e.g., hypertension, cancer, dementia) are the result of DNA sequence variations in multiple genes, at least some of which may interact in a non-additive, or epistatic, fashion, and thus have a substantially more complex genetic etiology. Early genome-wide association results have identified associations with only modest main effects. To account for highly familial traits, one must assume that there are many more loci with very modest main effects, more rare variants with larger effects, or gene-gene and/or gene-environment interactions that are more prominent elements of the genetic component of these diseases. Detecting gene-gene interactions can be problematic in the absence of main effects. As the number of possible candidate genes to consider increases, the number of combinations likewise increases, and one is soon faced with “the curse of dimensionality” (Bellman, 1961). In this paper we show that for a modest number of genes (or SNPs) and for only pairwise interactions among them, this problem is manageable. It is likely that the methods used in this paper, based on Reconstructability Analysis (RA), can be effective for more variables and for more than pair-wise interactions; however, the scaling limitations of RA applied to epistasis have not yet been explored.

Elsewhere in the bioinformatics literature, epistatic interaction of two genes in a simulated dataset has been detected by first creating a computationally complex 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. This approach has been used with multifactor dimensionality reduction (MDR) (Hahn et al, 2003) and artificial neural networks (ANN) (Ritchie et al, 2004). With reconstructability analysis (RA), we use a computationally simpler approach to identify the genes involved in epistasis, and our results are superior to the performance of MDR and ANN on this simulated dataset. While the variables identified in the structure could be used to select features for a neural network or some other prediction technique, the RA structure can also be used directly for disease prediction.

We begin with an overview of the theory on which RA is based, and then discuss recent approaches to the simulation of gene interactions. Next, we describe the genetic models we used and the datasets they generated. We then present the comparative performance of RA on simulated data, followed by an application of RA to an existing set of real data. We also detail the feasibility of application of our approach to larger scale studies, including genome-wide

association studies (GWAS). RA, based in information theory, partially overlaps log-linear (LL) methods and thus logistic regression (LR); since LR is popular for modeling gene-gene interactions, we also briefly discuss the relationship between RA and LR.

## **METHODS**

### **RECONSTRUCTABILITY ANALYSIS**

Reconstructability analysis (RA) is an information- and graph-theoretic methodology which originates with Ashby (Ashby, 1964) and was subsequently developed by several others (Broekstra, 1979; Cavallo, 1979; Cellier and Yandell, 1987; Conant, 1981; Jones, 1985; Klir, 1976; Krippendorff, 1981; Krippendorff, 1986). An account of its origin (Klir, 1986) and compact summaries (Zwick, 2001; Zwick, 2004) are also available. RA resembles log-linear methods (Bishop, et al, 1978) used widely in the social sciences, and where RA and log-linear methodologies overlap they are equivalent (Knoke and Burke, 1980; Krippendorff, 1986). RA also overlaps with Bayesian networks. In RA, a probability or frequency distribution or a set-theoretic relation is decomposed into component distributions or relations (Klir, 1985). When applied to the decomposition of frequency distributions, RA does statistical analysis. RA can model problems both where inputs (independent variables) and outputs (dependent variables) are distinguished (called directed systems) and where this distinction is not made (neutral systems). Being based on information theory, which ignores metric information in the variables being analyzed, RA is a natural methodology for nominal, e.g., genomic, data. By contrast, certain other machine learning methods which have been used to study epistasis, such as neural nets (Ritchie et al, 2004; Ritchie et al, 2003) or support vector machines (Chen et al, 2008), inherently presuppose metric information and are thus less naturally suited for genomic analyses.

For example, assume a directed system with inputs (genes or SNPs) 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 as ABCDZ. RA decomposes such a distribution into projections such as ABCD and ABZ, which taken together define an RA structure ABCD:ABZ that is less complex (fewer degrees of freedom) than the data. This structure defines a calculated frequency (or probability) distribution  $ABCDZ_{ABCD:ABZ}$ , obtained by maximum entropy composition of the ABCD and ABZ projections, and compared with the observed ABCDZ. While the data itself, ABCDZ, also called the “saturated model,” asserts that all four inputs jointly predict Z, the ABCD:ABZ structure asserts that only A and B jointly predict Z, while C and D have no predictive relationship with Z. A and B predict Z simply

via the conditional probability distribution  $p(Z|AB)$ , obtained from the calculated  $ABCDZ_{ABCD:ABZ}$  distribution. (The ABCD component of the ABCD:ABZ structure allows relationships between the inputs themselves, but we are not interested in these relationships.)

If ABCD:ABZ is a good structure, one can equivalently say that the “transmission” or “mutual information”  $T(AB:Z) = H(Z) - H(Z|AB)$  is high, while the conditional transmission  $T_{AB}(CD:Z)$  is zero.  $H$  is uncertainty (Shannon entropy), a measure of variability for non-metric variables that is analogous to variance for metric variables. So transmission is reduction of uncertainty about  $Z$ . Dividing  $T(AB:Z)$  by  $H(Z)$  gives  $\% \Delta H(Z)$ , the % uncertainty reduction of  $Z$ ; this is what is reported in Tables 2 and 4-7. By contrast, the structure ABCD:Z, called the “independence model” asserts that no input predicts  $Z$ ;  $T(AB:Z)$  is the predictive information in ABCD:ABZ relative to this independence model. The reduction of uncertainty for any structure can be assessed for statistical significance with the Chi-square distribution. For neutral systems, in which one does not distinguish between inputs and outputs, the independence model is A:B:C:..., i.e., inputs are no longer lumped in a single component.

There are different classes of RA structures. Structures such as ABCDZ, ABCD:ABZ, and AB:Z all have a “single predicting component,” i.e., one subset of the variables, including output  $Z$ , whose probability distribution can be used to predict  $Z$ . Such structures are “loopless,” and a loopless structure essentially picks out a single subset of the inputs that predicts the output. (It thus does “feature selection” or “dimensionality reduction.”) By contrast, the second and third components of structure ABCD:ABZ:CDZ assert that  $Z$  is predicted by  $A$  and  $B$  jointly and also, separately, by  $C$  and  $D$  jointly. The two predicting relations are integrated via the maximum entropy formalism, and from the integrated distribution that is calculated, one obtains a conditional distribution,  $p(Z|ABCD)$ , that is different from the conditional distribution obtainable directly from the data, and this calculated model conditional distribution is used for prediction.

Loopless structures can represent the fact that an input predicts the output even when controlling for all other inputs. Thus for example, for data ABCDZ, a high  $T(A:Z)$  says that  $A$  predicts  $Z$ , while a high  $T_{BCD}(A:Z) = T(ABCD:BCDZ)$  says that  $A$  predicts  $Z$  even when controlling for  $B$ ,  $C$ , and  $D$ . Thus RA could in principle be used to assess linkage disequilibrium from data.

If the non-predicting component that consists of all the inputs is included, structures with more than one predicting component necessarily have loops. RA calculations for structures without loops are simpler and faster, and the search for epistasis in this paper considered only loopless structures. For simplicity, the non-predicting components (in the present case, ABCD) are omitted below in Tables 2, 4, 5, and 6, which present results on simulated and real data. For loopless models with a small number of predicting inputs, one can easily examine all

possible structures – this is what was done in this study – and as the number of variables increases, this approach remains practical given sufficient computational resources. However, with more than a few variables the exhaustive evaluation of all models is prohibitive if models with loops are also assessed, since these models do not have algebraic solutions but require iterative computation (e.g., with Iterative Proportional Fitting); for details, see Zwick (2004). In the present study, loop structures are used to identify the character of the epistatic interaction. More specifically, after a loopless epistatic model was identified, the difference between the loopless structure and its nearest descendent loop structure was assessed for significance, as illustrated below in Table 7.

Different types of epistasis are associated with different RA structures. Consider a three variable system, ABZ. The data itself (the saturated model) might have a triadic interaction effect of A and B with Z that cannot be decomposed to a simpler structure without information loss. This would be the strongest type of epistasis. All of the simulated models in Table 1 show this type of epistasis, and so do the two-SNP structures found in the real data (Figure 3 and Table 6), i.e., in all these instances, the difference between ABZ and AB:AZ:BZ is significant at the .05 level. There are at least two other types of epistasis that can be modeled by simpler RA structures (AB:AZ:BZ and AZ:BZ) – these show up, for example, in the tables of Cordell (2002) – but a systematic examination of the relationship between RA classifications of epistasis and conventional distinctions (usually based on regression) is beyond the scope of this study.

A different information theoretic measure of a three-way interaction effect is given by the transmission difference,  $T_B(A:Z) - T(A:Z)$ , which indicates how much B increases or decreases the mutual information between A and Z. This quantity (which also equals  $T_A(B:Z) - T(B:Z)$ ) has been used as a measure of gene-environment interaction, in addition to  $T(A:B:Z)$  (for three variables) which is the constraint in the data relative to a neutral system independence model (Chanda et al, 2007). It should be noted, however, that this transmission difference can be either positive or negative, and that a triadic interaction may be present even if it is zero. The strength of the triadic interaction is thus not properly measured by this transmission difference; rather, it is measured by the information distance between ABZ and AB:AZ:BZ (Krippendorff, 2009).

The RA structures discussed above are all “variable based”, i.e., defined in terms of subsets of variables. RA encompasses also “state-based” structures (Zwick and Johnson, 2004), such as  $A_1B_2Z:B_1C_3Z$  defined in terms of information-rich input states ( $A_1B_2$  and  $B_1C_3$ ), that have either lower or higher predictive power than is typical. In this study, a SNP state is a coding of the diploid genotype, i.e., the genotype homozygous in one allele is recoded as state 1, heterozygous genotypes as state 2, and the genotype homozygous in the other allele as state 3. The  $A_1B_2Z$  component in the structure means that  $p(Z|A_1B_2)$  is

significantly different than  $p(Z)$  i.e., that genotype  $A_1B_2$  is at either higher or lower risk for disease than average. This type of analysis resembles multifactor dimensionality reduction (MDR), which has been used to study epistasis (Velez, et al., 2007). However, state-based RA was not used in this study.

Calculations were made using RA software programs developed at Portland State University (Portland, Oregon) now integrated into the OCCAM package (named for the principle of parsimony and “Organizational Complexity Computation and Modeling”). The earliest program was developed by Hosseini et al (1986); reviews of RA methodology (Zwick, 2001; Zwick, 2004), a list of recent RA papers (Zwick, 2009) and a description of the OCCAM architecture (Willett and Zwick, 2004) are available.

## **SIMULATED DATA**

Moore et al (2002) 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 et al (2004) to test their genetic programming approach to gene interaction modeling. These models, and the heritability values (the proportion of phenotypic variance attributable to genetic variance) calculated by Ritchie et al. based on the definition from Culverhouse et al (2002), are shown in Table 1. In each model, each cell represents the probability of disease given the particular combination of genotypes, for example,  $p(D | \text{gene}_1, \text{gene}_2)$  is the probability of disease, where  $\text{gene}_1$  has genotypes AA, Aa, and aa, and  $\text{gene}_2$  has genotypes BB, Bb, and bb (in these papers, these genes are labeled G and H).

All models assume 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 models were used to test the ability of RA to detect gene-gene interactions.

We created 30 datasets from each genetic model for each of two conditions. First, for comparability with existing work, we used two associated SNPs and eight non-associated, or “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, biomedical datasets are often divided into two equal groups: case subjects who are known to have the condition of interest, and control subjects selected from the general population. 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 20%, then the proportion of XX/YY in the cases should be 11% =  $(0.25*0.25*0.1)/[(0.5*0.5*0.2)+(0.25*0.25*0.1) + \text{terms for other genotypes with non-zero penetrances}]$ . For both groups, allele patterns for the non-contributing SNPs can be assigned at random.

We developed 30 datasets for each of the five genetic models and each of the two noise conditions. The choice of 30 was arbitrary – since OCCAM does not yet provide a platform for the automatic testing of multiple randomly generated datasets, and had to be run manually one dataset at a time. The interacting SNPs 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 OCCAM. By specifying suitable parameters, OCCAM searches through the lattice of structures for structural models of a particular class which have high information content (high reduction in output uncertainty). In this instance, OCCAM was asked simply to output all loopless structures with two predicting inputs, ordered by their information content. While in general using more datasets would be preferable, the extremely successful performance of OCCAM on these 30, as indicated below, suggests that more datasets were not necessary.

## **NIDDM DATA**

Cox et al. (1999) identified an interaction between genetic loci in the *NIDDM1* region on chromosome 2 (chr2) and a region in the vicinity of *CYP19* on chromosome 15 (chr15), based on correlation of nonparametric linkage (NPL) scores. Subsequently, they identified a significant association between susceptibility to NIDDM and a set of 16 SNPs within, or close to, the calpain-10 gene (*CAPN10*) in the *NIDDM1* region (Horikawa et al, 2000). Later, using the same subject and SNP dataset, but weighting subjects in terms of the chr2 NPL scores derived from linkage studies based on their family data, Tsalenko and associates (Tsalenko et al, 2003) applied an information-theoretic approach, also based on mutual information and equivalent to the use of loopless structures in RA, and identified a set of 15 informative SNPs in the *CAPN10/GPR35* region.

**Table 1.** Penetrance and Heritability of Simulated Genetic Models

	Table Penetrance			Margin Penetrance
<b>Model 1</b> (heritability = 0.053)	<i>AA</i> (.25)	<i>Aa</i> (.50)	<i>aa</i> (.25)	
<i>BB</i> (.25)	0.00	0.10	0.00	0.05
<i>Bb</i> (.50)	0.10	0.00	0.10	0.05
<i>bb</i> (.25)	0.00	0.10	0.00	0.05
Margin Penetrance	0.05	0.05	0.05	
<b>Model 2</b> (heritability = 0.051)	<i>AA</i> (.25)	<i>Aa</i> (.50)	<i>aa</i> (.25)	
<i>BB</i> (.25)	0.00	0.00	0.10	0.025
<i>Bb</i> (.50)	0.00	0.05	0.00	0.025
<i>bb</i> (.25)	0.10	0.00	0.00	0.025
Margin Penetrance	0.025	0.025	0.025	
<b>Model 3</b> (heritability = 0.026)	<i>AA</i> (.25)	<i>Aa</i> (.50)	<i>aa</i> (.25)	
<i>BB</i> (.25)	0.00	0.04	0.00	0.02
<i>Bb</i> (.50)	0.04	0.02	0.00	0.02
<i>bb</i> (.25)	0.00	0.00	0.08	0.02
Margin Penetrance	0.02	0.02	0.02	
<b>Model 4</b> (heritability = 0.012)	<i>AA</i> (.25)	<i>Aa</i> (.50)	<i>aa</i> (.25)	
<i>BB</i> (.25)	0.00	0.02	0.08	0.03
<i>Bb</i> (.50)	0.05	0.03	0.01	0.03
<i>bb</i> (.25)	0.02	0.04	0.02	0.03
Margin Penetrance	0.03	0.03	0.03	
<b>Model 5</b> (heritability = 0.008)	<i>AA</i> (.25)	<i>Aa</i> (.50)	<i>aa</i> (.25)	
<i>BB</i> (.25)	0.00	0.04	0.08	0.04
<i>Bb</i> (.50)	0.06	0.04	0.02	0.04
<i>bb</i> (.25)	0.04	0.04	0.04	0.04
Margin Penetrance	0.04	0.04	0.04	

The data and subject sample here was based on the sample used in these studies. It comprised 220 Mexican-American individuals from Starr County, Texas, including one case from each of 108 multiplex NIDDM families, and 112 population-based controls, and contained SNP genotype data for 76 SNPs on chr2 and 126 SNPs on chr15, a more dense set of SNPs than had been used in the previous studies. Because the dataset is balanced in terms of numbers of cases and controls, and because structures are identified on the basis of information theoretic measures, as opposed to predictive accuracy, there are no sensitivity vs. specificity issues, such as those considered by Velez et al (2007).

## **DATA PREPROCESSING**

For ease of indexing, markers were numbered sequentially; specifically, chr2 markers were numbered A1-A76, and chr15 markers B1-B126. Once a marker was given a number, the number was retained throughout the study. The original dataset contained 88,880 data points (220 subjects\*202 SNPs\*2 alleles/SNP). Of these, approximately 7000 points were missing (~8%). The OCCAM software itself does not yet fill in missing data (though it allows these either to be skipped or to be coded for as an additional state), so preprocessing of missing data must be done as a preliminary step. This was done as follows. All missing data occurred in pairs (i.e., genotypes). Four SNPs on chr2 and one on chr15 were dropped because they had no data at all; five SNPs with a minor allele frequency (MAF)  $\leq .01$  (four on chr2 and one on chr15) were also dropped. Furthermore, we selected a set of 39 SNPs on chr15 that included and extended beyond the *CYP19* locus, since our purpose was to replicate previous results regarding interaction between that region and chr2 loci. If a marker was dropped in the case dataset, it was also dropped from the controls, and vice versa. Subjects were dropped if they were missing data on more than 50% of the markers on chr2 or chr15. This included 8 subjects, one from cases and seven from controls. If a record was dropped from the chr2 dataset, it was also dropped from chr15, and vice versa.

At this point, imputing values for missing genotype data required information on the linkage disequilibrium (LD) structure of the regions spanned by the SNPs. We used the computer program PLINK (Purcell et al, 2007) to estimate all pairwise  $r^2$  values for the 68 SNPs on chr2 and for the 39 SNPs on chr15. We resolved missing genotype data in two steps. First, there was a total of 9 SNPs (5 on chr2, 4 on chr15) with > 5% missing data in the combined case and control dataset. Of these, seven (3 on chr2, 4 on chr15) were in strong LD ( $r^2 > .90$ ) with at least one SNP with no or minimal missing data. We elected to drop these seven SNPs, rather than replace their missing genotypes with those from SNPs in strong LD, because RA does not consider physical location or weighting. The remaining two SNPs (A14, A70) were not in strong LD ( $r^2 < .65$ ) with any

other SNPs; since they had nearly 10% missing data, we elected to drop these from the analysis. The final set of markers consisted of 63 chr2 SNPs and 35 chr15 SNPs.

Second, the majority of SNPs had < 5% missing genotypes. None of these SNPs was in strong LD ( $r^2 < .7$ ) with proximal SNPs in the dataset. Since standard imputation methods perform less reliably as LD declines, we replaced missing data in these cases by a random number draw against the allele frequency distribution for each SNP, calculated separately for cases and controls. Because of the relatively small sample size, we found that replacing missing data with different random draws caused differences in results. To correct for this, we created 10 datasets with independently-generated fillers for the missing data. Each of these datasets was then processed by OCCAM, and results were compared. In the discussion that follows only structures that appeared in all 10 runs were considered. Once missing data was replaced, each genotype was recoded into a single, three-valued data SNP state, as explained above. Finally, the dependent variable ( $Z$ ) was added, coding cases as 1 and controls as 0.

There were two parts to this experiment. First, we looked only at structures containing single SNPs from chr2, which allowed comparison of results with those of Horikawa et al (2000) and Tsalenko et al (2003). Second, we considered only those structures that demonstrated epistasis by having one SNP on chr2 and one on chr15, which allowed comparison of results with those of Cox et al (1999).

## RESULTS

### SIMULATED DATA

Table 2 shows an example of OCCAM output for Dataset 11 of Model 5. Column 1 identifies the structure, and column 2 indicates  $\% \Delta H(\text{output})$ , the percentage of the uncertainty in  $Z$  reduced by knowing the inputs in the structure for the 8 noise SNP dataset. For the 50 noise SNP dataset, column 3 identifies the structure and column 4 tabulates the uncertainty reduced. The two active genes are A and B. The table is arranged in decreasing order of uncertainty reduction. Structure HZ is the single-gene structure with the highest  $\% \Delta H(\text{output})$ , and the  $\% \Delta H(\text{output})$  levels of structures AZ and BZ are included to show that using even a correct single gene provides almost no usable information.

**Table 2.** Sample OCCAM Output, Model 5

8 Noise SNPs		50 Noise SNPs	
Structure	%Uncertainty reduced	Structure	%Uncertainty reduced
ABZ	7.22	ABZ	6.62
ACZ	4.11	CGZ	1.43
CGZ	2.36	DJZ	0.73
DJZ	2.06	BGZ	0.42
BGZ	0.80	ACZ	0.38
HZ	0.61	HZ	0.36
AZ	0.16	AZ	0.07
BZ	0.06	BZ	0.02
Z	0.00	Z	0.00

A summary of all five genetic models is shown in Table 3. Column 1 indicates the genetic model used, and column 2 lists the heritability of the disease in this genetic model (Moore et al, 2002; Ritchie et al, 2004). Columns 3 and 4 indicate the percentage of the 30 randomly constructed datasets for each genetic model in which RA was able to identify the correct gene structure. 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 10, 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.

The error rate of this approach appears to be extremely small. Since each test of a SNP pair in our simulated data is independent, we can merge all the tests for a given model. For the two active SNP + eight noise SNP test, for example, that gives a total of 1350 tests ( $\text{comb}(10,2) \times 30$ ) within each model. Since we know *a priori* that there is only one active SNP pair, we call a SNP pair a false positive if it is not the AB pair yet still meets an  $\alpha \leq 0.000$  criterion (where  $\alpha$  is the probability of error if one rejects the null hypothesis that the output is independent of the inputs), and our AB pair is counted as a false negative if it doesn't meet that criterion. Model 5 had 5 false positives (where SNP pairs other than AB were identified as significant), and 2 false negatives (where the AB pair was rejected), for an error rate of 0.004. Model 4 had 9 false positives and no false negatives, for an error rate of 0.005. Models 1, 2, and 3 had no errors.

**Table 3.** Effectiveness of RA in identifying gene-gene interactions

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%

By comparison, previous work with only eight noise genes found one of the two active genes only 47% of the time (Ritchie et al, 2004) and both genes only 19% of the time (Hahn et al, 2003). Neither study reported on false positives. Hahn used multifactor dimension reduction and Ritchie used neural nets. RA significantly outperforms these earlier studies.

## NIDDM DATA

### SINGLE SNP SEARCH

The results of the first part of this experiment are shown in Table 4 and Figure 1. To compare our results with those of Horikawa et al (2000), we took the top 16 markers that appeared in all 10 experiments. This gave us a  $\% \Delta H(\text{output})$  cutoff of 1.25. All but one had  $\alpha < 0.10$  (where, as above,  $\alpha$  is the probability of error if one rejects the null hypothesis that the output is independent of the inputs; it is for a confirmatory test on one structure, and does not correct for the number of structures being examined in the exploratory search). Thirteen of the 16 markers we identified were among the 16 markers identified by Horikawa et al (2000) as significantly associated with disease (p-value < .05).

Although a number of the SNPs in Figures 1 and 2 are not in LD with other SNPs ( $r^2 < .3$ ) in the tables, two LD blocks exist, with pairwise  $r^2$  ranging from .51-.98. One spans *GPR35* (A37, A38, A52) and another spans the distal half of *CAPN10* (A19, A30, A43, A48, A69, A56). Two additional SNPs near *GPR35* (A35, A40) are in strong LD. However, only one SNP from each of the *GPR35* and *CAPN10* LD blocks is represented in the list of epistatic pairs with chromosome 15 (Figure 3).

**Table 4.** Single SNP Structures

Structure	%Uncertainty reduced*	$\alpha^*$
A37	2.897	0.01
A35	2.817	0.02
A40	2.726	0.02
A26	2.691	0.02
A38	2.493	0.03
A46	2.420	0.01
A25	2.283	0.04
A23	2.233	0.04
A18	2.221	0.04
A52	2.102	0.05
A22	2.087	0.05
A55	1.621	0.09
A51	1.610	0.10
A16	1.589	0.10
A54	1.374	0.13
A44	1.249	0.08

\*average of 10 data runs

**Table 5.** SNP Structures for predicting NPL

Structure	%Uncertainty reduced*	$\alpha^*$
A19	6.55	0.01
A30	6.51	0.01
A56	5.72	0.02
A48	5.59	0.03
A33	5.21	0.02
A31	4.77	0.04
A69	4.12	0.07
A36	4.09	0.06
A43	4.07	0.09
A28	4.07	0.07
A10	3.84	0.10
A49	3.66	0.09
A38	3.52	0.11
A9	3.46	0.11
A68	3.40	0.12

\*average of 10 data runs

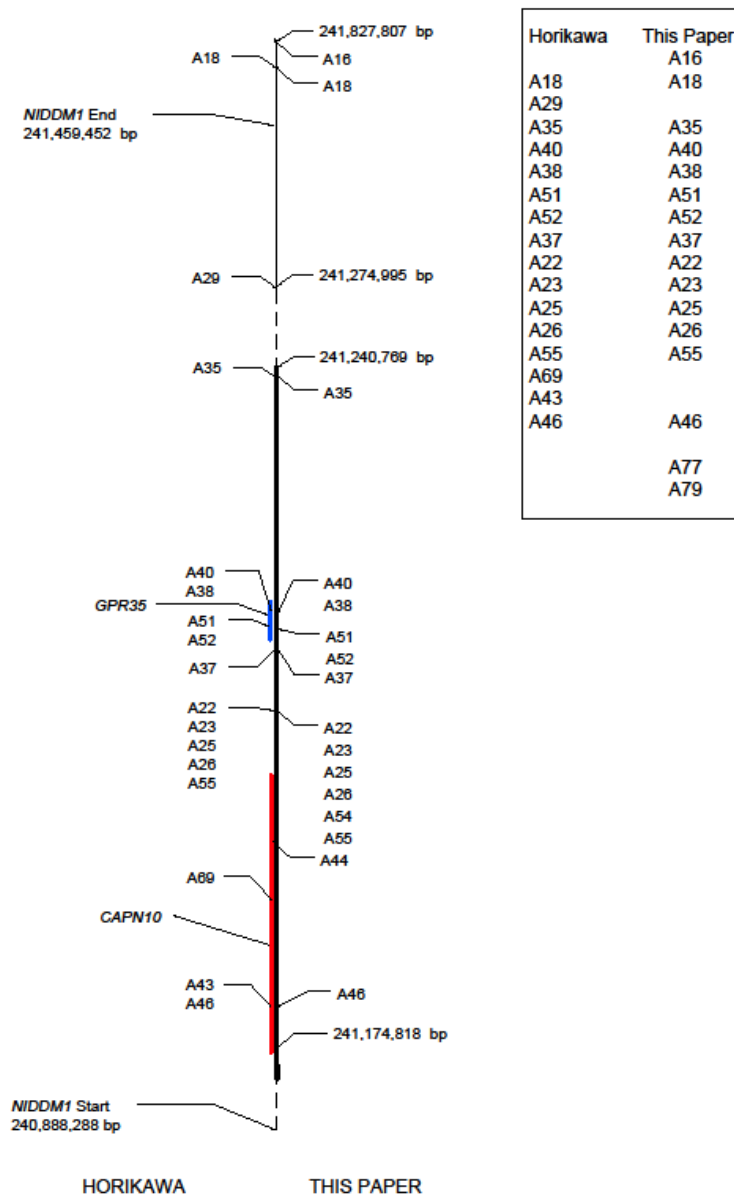
Our results are also consistent with the chr2 results in Tsalenko et al. (2003) (Table 5, Figure 2). Of our top 15 markers, 12 had  $\alpha \leq 0.10$ , and 10 of those were among the 15 top scoring SNPs in their study. We dropped two SNPs (A59, A65), included in the top 15 SNPs in their study because of missing data issues. These SNPs are in strong LD ( $r^2 > .90$ ) with A56 which is included in both lists; thus we effectively found 12 of the 15 SNPs in Tsalenko et al (2003).

## EPISTASIS

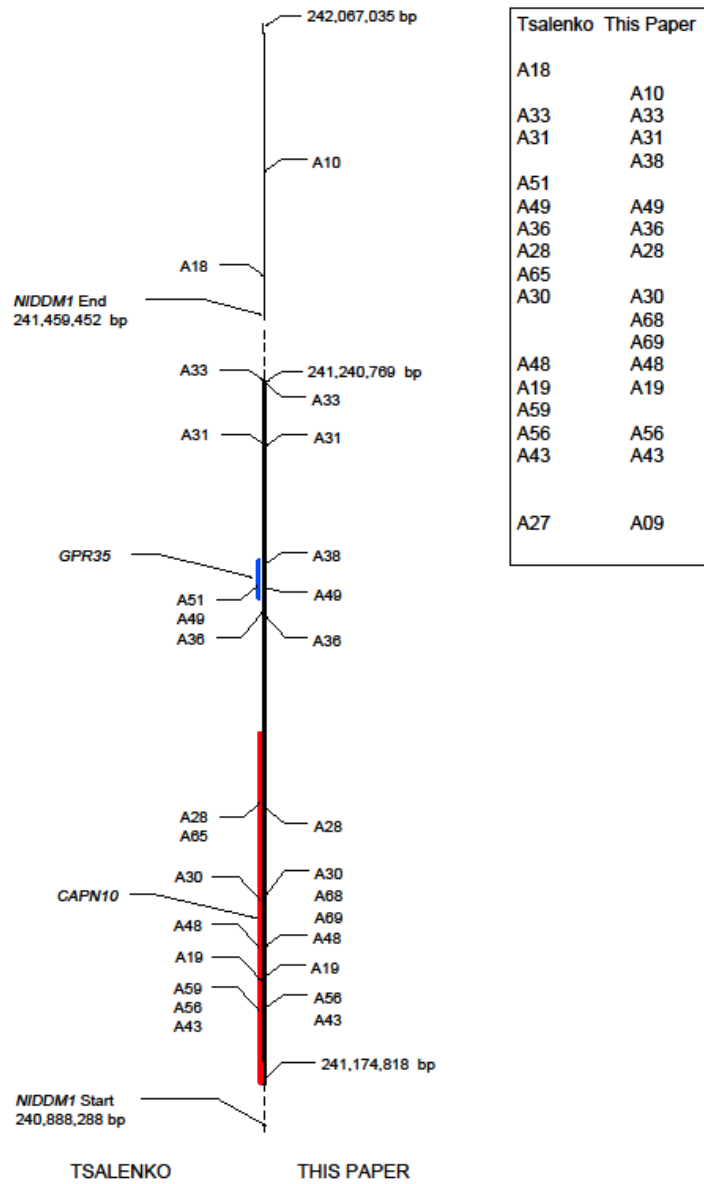
Cox et al (Cox et al, 1999) identified an epistatic interaction between the chr2 *NIDDM1* region and a region in the vicinity of *CYP19* on chr15, based on correlation of NPL scores. The *CYP19* region is relatively small (130Kb), encompassing two SNPs (B3 and B4) in our original dataset, while the chr15 region tested is large (~10,000 Kb). The original marker, D15S119, which showed linkage to the *CYP19* region, is ~2 Mb distal to *CYP19* (NCBI Map Viewer, Build 37.1); it is encompassed by our markers B39, B43 and B49.

Epistatic structures are shown in Table 6 and Figure 3. Our criteria required that the structure appear in all 10 data runs, and that  $\alpha \leq 0.10$ . All 36 resulting structures had  $\alpha < 0.02$ . In these 36 structures, 30 involved chr2 SNPs with strong main effects, as reported in Table 4. Also, in these 36 structures, 14 chr2 SNPs interacted with 21 chr15 SNPs. Of these 14 chr2 SNPs, 8 had strong main effects, and of these 8, A35 was involved in many (10) epistatic structures. A total of 9 structures bracket the *CYP19* region (those including B39, B43, B44, B124 or B125); the chr15 markers in six of these structures (those including B39, B43 and B49) flank *D15S119*. One structure, A22B3, involved a direct interaction with *CYP19*.

In addition, a cluster of SNPs (A22, A25, and A26 in the region between *GPR35* and *CAPN10*) were involved in 10 structures showing interactions between that region of chr2 and the chr15 *CAPN3/ZFP106* region. In other findings, structure A26B32 demonstrates an interaction between the chr2 *GPR35/CAPN10* region and the chr15 gene HNF6, a transcription factor in which mutations lead to a form of diabetes called “maturity onset diabetes of the young” (MODY).



**Figure 1.** Horikawa Comparison. This figure maps part of the *NIDDM1* region of chromosome 2. Markers found by Horikawa on the left side of the map, and by this study on the right. Highlighted genes are *CAPN10* (red) and *GPR35* (blue). The insert shows all markers found by each study, in chromosomal order. This study identified 13 of the 16 SNPs identified by Horikawa et al (2000).



**Figure 2.** Tsalenko Comparison. This figure maps part of the *NIDDM1* region of chromosome 2, and shows markers found to be important in predicting NPL scores in the study by (Tsalenko et al. 2003) on the left side of the map, and in this study on the right. Highlighted genes are *CAPN10* (red) and *GPR35* (blue). The insert shows all markers found by each study, in chromosomal order. This study identified 10 of the 15 SNPs identified by Tsalenko.

**Table 6.** Epistatic SNP Structures

Structure	% Uncertainty reduced*	$\alpha^*$
A26B32	8.78	0.00
A35B47	8.32	0.00
A41B44	7.92	0.00
A35B49	7.82	0.00
A25B15	7.61	0.01
A40B49	7.60	0.00
A37B39	7.59	0.01
A35B125	7.45	0.00
A26B45	7.35	0.01
A40B44	7.32	0.01
A40B125	7.26	0.01
A35B44	7.20	0.01
A23B39	7.18	0.01
A25B16	7.11	0.01
A66B44	7.05	0.01
A40B47	7.00	0.01
A60B44	6.97	0.01
A25B39	6.94	0.01
A47B44	6.93	0.01
A37B43	6.88	0.01
A35B124	6.87	0.01
A25B7	6.82	0.01
A55B46	6.81	0.01
A23B47	6.66	0.01
A35B14	6.61	0.01
A25B24	6.59	0.01
A18B56	6.57	0.01
A22B3	6.49	0.02
A35B31	6.48	0.01
A26B25	6.39	0.02
A37B33	6.25	0.02
A35B56	6.23	0.02
A35B28	6.15	0.02
A35B65	6.13	0.02
A62B44	6.12	0.02
A26B44	6.01	0.02

\*average of 10 data runs

Table 7 gives more detail, in the form of the %uncertainty reductions, for the A26B32 structure. Due to the effect of the loop, the uncertainty reduction in

Z for AB:AZ:BZ is slightly subadditive with respect to uncertainty reductions for AB:AZ and AB:BZ, but the uncertainty reduction for ABZ is much larger than the uncertainty reduction for AB:AZ:BZ. This indicates a large triadic interaction effect. In this table,  $\Delta df$  and  $\Delta LR$  are differences in degrees of freedom and likelihood ratio relative to A26B32:Z.  $\alpha^*$  is significance relative to A26B32:Z, showing that A26 is a significant a predictor of Z while B32 is not, and that A26B32Z is different from independence;  $\alpha^\#$  is significance relative to the loop structure A26B32:A26Z:B32Z, showing that the triadic interaction effect in A26B32Z is statistically significant. All epistatic pairs (Table 6 and Figure 3) exhibit such a triadic interaction effect. Table 8 gives the penetrance table for A26B32. The SNP state is shown in parentheses. Each cell (each genotype) has a penetrance value equal to #cases / (#cases + #controls), with the denominator (the frequency of that genotype) given in square brackets. Under these entries in italics and smaller font is a modified estimate of the penetrance followed by the 95% confidence interval shown in curly brackets; these are calculated by the adjusted Wald method (Sauro, 2006). There is a suggestion of possible heterosis (non-monotonicity with respect to allele dose) in this table; this is not definitive, given the small number of occurrences of the genotypes.

**Table 7.** Uncertainty reductions for A26B32Z

Structure	$\Delta df$	$\Delta LR$	$\alpha^*$	$\alpha^\#$	$\% \Delta H(Z)$
A26B32Z	8	27.00	0.0007	0.0047	9.10
A26B32:A26Z:B32Z	4	11.98	0.0174	1.0000	4.04
A26B32:A26Z	2	8.22	0.0164	–	2.77
A26B32:B32Z	2	4.20	0.1227	–	1.41
A26B32:Z	0	0.00	1.0000	–	0.00

The chr2 *GPR35* region itself interacts with chr15 genes *SEMA6D* (A41B44) and *GALK2* (A37B39). In addition to contributing two structures in the vicinity of *CYP19*, SNP A35 (at the end of the *NIDDMI* region) also interacts with chr15 genes *SEMA6D* (A35B47), and *USP8* (A35B49). From the perspective of chr15, the SNP associated with *SEMA6D*, SNP B44, is involved in eight epistatic structures on chr2, most of them within *CAPN10*.

**Table 8.** Penetrance table for A26B32Z

	B32(1)	B32(2)	B32(3)
A26(1)	.40 [117] .40 {.32,.49}	.51 [43] .51 {.37,.65}	1.00 [5] .86 {.60,1.00}
A26(2)	.62 [29] .61 {.44,.77}	.85 [13] .80 {.57,.97}	0.00 [2] .25 {0,.63}
A26(3)	1.00 [3] .80 {.47,1.00}	.50 [2] .50 {.09,.91}	----- [0]

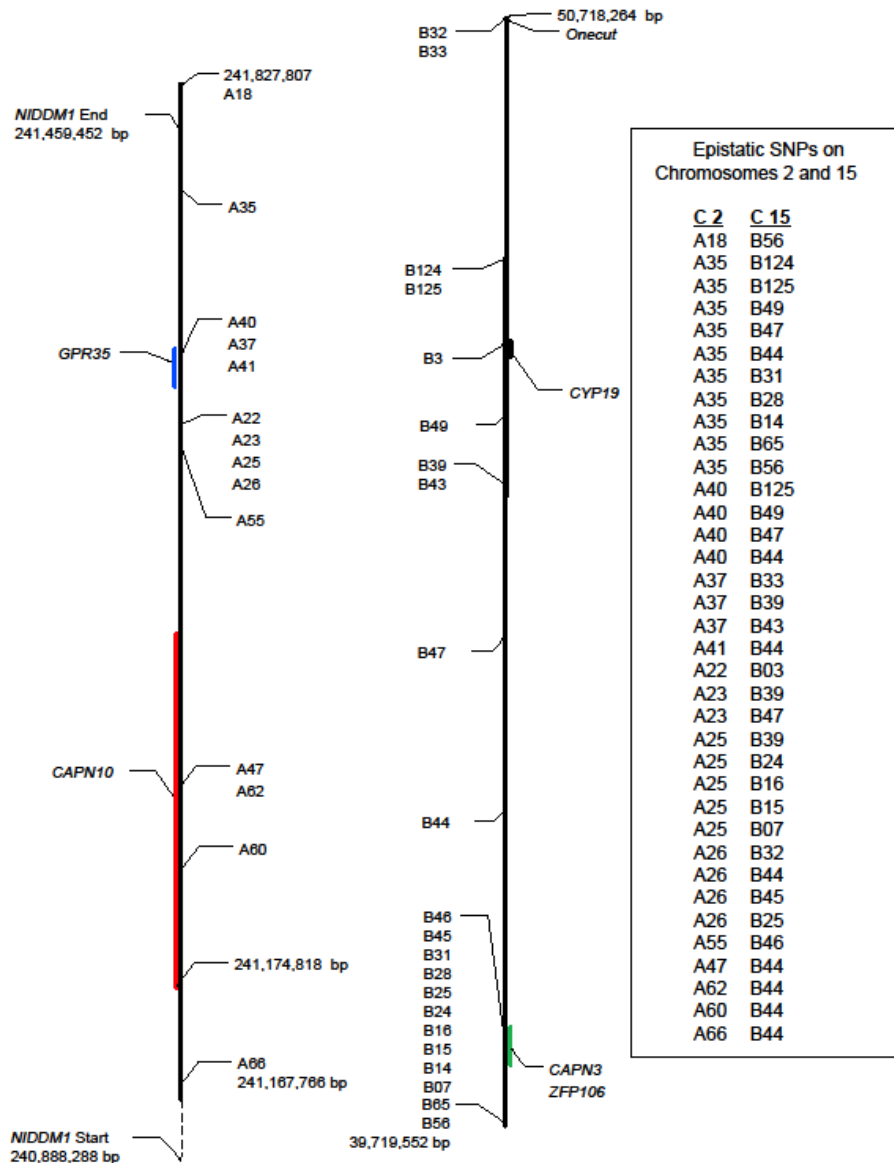
## DISCUSSION

RA is capable of detecting low levels of genetic interactions, despite high noise levels, in simulated data, reliably detecting interactions in heritabilities as low as 0.008, with as many as 50 intervening noise genes. As noted above, RA outperforms the earlier work of (Ritchie et al, 2003), who used neural nets, and (Hahn et al, 2003), who used multifactor dimension reduction.

To validate the use of RA for association and gene-gene interaction studies, we used a set of *NIDDM1* case and control subjects, for which there was evidence in the literature of an association between *NIDDM1* and a specific region on chromosome 2, as well as a possible interaction with a region on chromosome 15. For this specific purpose, the small sample size that we used (200 cases and 200 controls) is adequate. The SNP data available on these subjects had been generated on the basis of prior linkage and positional cloning studies. Although the rationale for selecting SNPs for association studies now encompasses a broader spectrum of considerations, and although larger sample sizes are needed for genome-wide association studies, these considerations have no consequence for the basic integrity of RA methodology.

In single SNP tests on real data, RA closely approximated results obtained in previous analyses of these data (Horikawa et al, 2000; Tsalenko et al, 2003). In cross-chromosome tests (Cox et al, 1999), we confirmed the association between the chr2 *NIDDM1* region and the chr15 *CYP19* region. We detected a multi-SNP association between *NIDDM1* and *CAPN3* on chr15; it has been suggested that *CAPN3* may contribute to susceptibility to diabetes (Walder et al, 2002).

In this study, SNPs are identified by an information theoretic measure calculated from the full dataset; that is, the data was not split into training and test sets to assess predictive accuracy. The study uses the simplest possible RA analysis, namely loopless models with only one or two inputs. This type of RA is equivalent to the information-theory approach used by Tsalenko et al (2003)



**Figure 3.** Epistasis Findings. This figure maps part of the *NIDDM1* region of chromosome 2, and the study region of chromosome 15, and shows epistatic markers found in this study. Highlighted genes are *CAPN10* (red), *GPR35* (blue), *CAPN3* (green), and *CYP19* (black). One of our markers (B4) falls within the *CYP19* region found by Cox et al (1999).

and is also related to the measures used by Chanda et al (2007) to study gene-environment interactions. More complicated RA methods, e.g., variable-based models with loops or state-based models, were not here employed to detect epistasis, although models with loops were used (Table 7) as reference models to show that epistasis is a triadic interaction effect rather than a cyclic linkage of variables, i.e., that it corresponds to ABZ and not AB:AZ:BZ.

There are other methods, such as log-linear (LL) models and Bayesian networks (BN), which resemble information theory approaches, which could be used to study epistasis. Where these probabilistic methods overlap, they are typically mathematically equivalent, although some differences exist. Standard Bayesian analysis does not consider multiple predicting component structures which have loops, and state-based modeling does not appear to exist in either the LL or BN toolkits. RA has additional features that distinguish it from BN and LL. It is applicable to set-theoretic relations and mappings whose analysis is non-statistical and has the possibility of other extensions via “generalized information theory” (Klir, 2005); set-theoretic RA might offer fast approximations for the analysis of contingency tables. RA also includes methods to analyze continuous outputs which could be applied to gene expression, and it has a Fourier variation that resembles non-linear regression and might be applied to epistasis.

Logistic regression (LR), which has been used to study epistasis, is closely related to log-linear (LL) methods. LR, applied to nominal (as opposed to continuous) input variables where dummy variables code the states of the input variables, is essentially the same as LL; in the area where these three formalisms overlap, RA, LL, and LR are all equivalent. Thus, for example, LR gives the same delta-likelihood-ratio and  $\alpha$  values as those shown in Table 7. Despite this equivalence, RA employs uncertainty and transmission measures not normally reported in LL or LR, and these measures are useful and intuitively easy to understand; recall, for example, the fact that conditional transmissions could allow one to encompass the effects of LD on SNP-disease associations. LR does not evaluate the structure AZ:BZ (as opposed to AB:AZ:BZ), which can sometimes model epistasis, because it is not hierarchically related to AB:Z, the independence reference structure for directed systems, but AZ:BZ is naturally encompassed in RA (and LL). RA is also definitely different from LR as implemented in the PLINK software (Purcell et al, 2007) which has been employed for the analysis of epistasis. PLINK regresses against allele dose, i.e., treats variables as quantitative rather than nominal, and is inappropriate when the dependence of disease on allele dose is not monotonic (or if monotonic, not linear); as noted earlier (Table 8), such non-monotonicity may occur in our data. But even when genotypes are coded as nominal states, LR does not fully overlap – and is thus not completely equivalent to – RA. The formal differences between RA and LR include the differences noted above between RA and LL: RA has a

set-theoretic version, can analyze continuous outputs, and has a Fourier-based variation. Further, RA is a fusion of information theory and graph theory, and connects strongly with the “graphical models” literature. In its graph theoretic aspects RA explicitly considers the lattice of all possible structures and offers heuristics for searching this lattice. For example, OCCAM is explicitly designed for exploratory modeling, though it can also be used more simply for confirmatory modeling. By contrast, LR software is usually not designed for exploratory purposes and is sometimes unable to handle interactions between many variables. Inputs with three or more states are sometimes recoded in terms of two or more binary variables, the meaning of whose states is inherently obscure. However, states and interactions between states can be explicitly coded in LR as separate effects whose interpretation is straightforward, and with such coding, LR resembles state-based RA. Whether the two are formally equivalent is under investigation, but even if they are, there remain computational differences: LR maximizes likelihood, typically with the Newton-Raphson algorithm, while RA maximizes entropy, equivalent here to maximizing likelihood, with Iterative Proportional Fitting. In summary, while RA and LR (and LL) are identical where they overlap, RA has distinctive features, both theoretical and computational, which make it useful for the study of epistasis.

Other methods, very different from RA/LL/LR/BN modeling, have also been used to analyze epistasis, namely neural networks and support vector machines. As noted earlier, these methods are actually designed for continuous variables, and so are intrinsically not suited to genomic data, which is nominal. Moreover, the predictive relation inherent in an RA (or LL, LR, or BN) structure is a conditional probability of the discrete output, given its discrete inputs. This conditional probability is precisely penetrance and thus is a natural and transparent way to represent relations between genotype and phenotype. By contrast, a neural network fits data via a set of hard-to-interpret weights, and is usually not accompanied by statistical assessment. Also, neural networks are designed for deterministic input-output relations, and often do not perform well when relations are stochastic, which is typically the case for genotype-phenotype relations, e.g., for epistatic pairs in the diabetes data analyzed here.

In this study, RA uses a simple 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 do not include both active SNPs will also show no effect, so one can simply look at all pairs of SNPs to find the active pair. For OCCAM, this is not a 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. We note that with these computational times, RA is feasible for approaches that have been previously described for potential analysis of interactions in GWAS. For example, consideration of all possible pairs of regions identified among the top signals (e.g. top 10,000 signals), or consideration of all pairs of regions across the genome for the top replicated signals, or considering all possible pairs of regions from a pre-identified set of loci for which there is prior evidence of biological interaction at the level of the protein and/or DNA would all be computationally feasible with RA. With some optimization of code, and use of parallel analysis, RA would be computationally feasible to apply in a genome x genome analysis of interaction, although there are clearly statistical issues in interpreting results of such a large number of tests.

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

## REFERENCES

- Ashby, W.R. (1964) *Constraint Analysis of Many-Dimensional Relations*, *General Systems Yearbook*, 9, 99-105.
- Bellman, R. (1961) *Adaptive Control Processes*. Princeton University Press, Princeton.
- Bishop, Y. et al (1978) *Discrete Multivariate Analysis*. MIT Press, Cambridge.
- Broekstra, G. (1979) *Nonprobabilistic constraint analysis and a two-stage approximation method of structure identification*. Houston.
- Cavallo, R.E. (1979) *The role of systems methodology in social science research*. M. Nijhoff, Boston.
- Cellier, F. and Yandell, D. (1987) SAPS-II: A New Implementation of the Systems Approach Problem Solver, *Internat J Gen Sys*, 13, 307-322.
- Chanda, P. et al (2007) Information-theoretic metrics for visualizing gene-environment interactions, *Am J Hum Genet*, 81, 939-963.
- Chen, S.H. et al (2008) A support vector machine approach for detecting gene-gene interaction, *Genet Epidemiol*, 32, 152-167.
- Conant, R.C. (1981) Set-Theoretic Structure Modeling, *Internat J Gen Sys*, 7, 93-107.

- Cordell, H.J. (2002) Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans, *Hum Mol Genet*, 11, 2463-2468.
- Cox, N.J. et al (1999) Loci on chromosomes 2 (NIDDM1) and 15 interact to increase susceptibility to diabetes in Mexican Americans, *Nat Genet*, 21, 213-215.
- Culverhouse, R. et al (2002) A perspective on epistasis: limits of models displaying no main effect, *Am J Hum Genet*, 70, 461-471.
- Hahn, L.W. et al (2003) Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions, *Bioinformatics*, 19, 376-382.
- Horikawa, Y. et al (2000) Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus, *Nat Genet*, 26, 163-175.
- Hosseini, J.C., Harmon, R.R., and Zwick, M. (1986) Segment Congruence Analysis Via Information Theory. *Proceedings, International Society for General Systems Research*. Philadelphia, G62-G77.
- Jones, B. (1985) Reconstructability Analysis for General Functions, *Internat J Gen Sys*, 11, 133-142.
- Klir, G. (1976) Identification of Generative Structures in Empirical Data, *Internat J Gen Sys*, 3, 89-104.
- Klir, G. (1985) *The Architecture of Systems Problem Solving*. Plenum Press, New York.
- Klir, G. (1986) Reconstructability Analysis: An Offspring of Ashby's Constraint Theory *Systems Research*, 3, 267-271.
- Klir, G. (2005) *Uncertainty and Information: Foundations of Generalized Information Theory*. Wiley-IEEE Press.
- Knoke, D. and Burke, P.J. (1980) *Log-Linear Models*. Sage, Beverly Hills.
- Krippendorff, K. (1981) An Algorithm for Identifying Structural Models of Multivariate Data, *Internat J Gen Sys*, 7, 63-79.
- Krippendorff, K. (1986) *Information Theory: Structural Models for Qualitative Data*. Sage, Beverly Hills.

- Krippendorff, K. (2009). Information of interactions in complex systems. *Internat J Gen Sys*, 38, 669-680.
- Moore, J. et al (2002) Application of genetic algorithms to the discovery of complex genetic models for simulation studies in human genetics. *Genetic and Evolutionary Algorithm Conference*. New York, 1150-1155.
- Purcell, S. et al (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses, *Am J Hum Genet*, 81, 559-575.
- Ritchie, M. et al (2004) Genetic Programming Neural Networks as a Bioinformatics Tool for Human Genetics. *Genetic and Evolutionary Computation Conference*. Seattle, 438-448.
- Ritchie, M.D. et al (2003) 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.
- Sauro, Jeff (2006). <http://www.measuringusability.com/wald.htm#exact>  
Based on Sauro J and Lewis, JR (2005) Estimating Completion Rates from Small Samples Using Binomial Confidence Intervals: Comparisons and Recommendations, Proceedings of the Human Factors and Ergonomics Society, 49th Annual Meeting, pp. 2100-2104, and Lewis JR and Sauro J (2006) When 100% Really Isn't 100%: Improving the Accuracy of Small-Sample Estimates of Completion Rates, *Journal of Usability Studies*, 3(1) pp. 136-150.
- Tsalenko, A. et al (2003) Methods for analysis and visualization of SNP genotype data for complex diseases. *Pacific Symposium on Biocomputing*. 548-561.
- Velez, D.R. et al (2007) A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction, *Genet Epidemiol*, 31, 306-315.
- Walder, K. et al (2002) Calpain 3 gene expression in skeletal muscle is associated with body fat content and measures of insulin resistance, *Int J Obes Relat Metab Disord*, 26, 442-449.
- Willett, K. and Zwick, M. (2004) A Software Architecture for Reconstructability Analysis, *Kybernetes*, 33, 997-1008.

- Zwick, M. (2001) Wholes and Parts in General Systems Methodology. In Wagner, G. (ed), *The Character Concept in Evolutionary Biology*. Academic Press, New York, 237-256.
- Zwick, M. (2004) An Overview of Reconstructability Analysis, *Kybernetes*, 33, 877-905.
- Zwick, M. (2009) Discrete Multivariate Modeling web page.  
<http://www.pdx.edu/sysc/research-discrete-multivariate-modeling>
- Zwick, M. and Johnson, M.S. (2004) State-Based Reconstructability Analysis, *Kybernetes*, 33, 1041-1052.