12-6-2010

Reconstructability Analysis of Epistasis

Martin Zwick

Portland State University, zwick@pdx.edu

Citation Details

Reconstructability Analysis of Epistasis[1]

MARTIN ZWICK

Systems Science Graduate Program

Portland State University

Portland OR 97207-0751

(503) 725-4987

Fax: (503) 725-8489

email: zwick@pdx.edu

Summary

The literature on epistasis describes various methods to detect epistatic interactions and to classify different types of epistasis. Reconstructability Analysis (RA) has recently been used to detect epistasis in genomic data (Shervais et al., 2010). This paper shows that RA offers a classification of types of epistasis at three levels of resolution (variable-based models without loops, variable-based models with loops, state-based models). These types can be defined by the simplest RA structures that model the data without information loss; a more detailed classification can be defined by the information content of multiple candidate structures. The RA classification can be augmented with structures from related graphical modeling approaches. RA can analyze epistatic interactions involving an arbitrary number of genes or SNPs and constitutes a flexible and effective methodology for genomic analysis.

INTRODUCTION

RA is a modeling methodology developed in the systems community (Klir, 1976 1985; Conant, 1981; Krippendorff, 1981; Krippendorff, 1986; Broekstra, 1979; Cavallo, 1979; and others) based on the work of Ashby (Klir, 1986). It uses set theory or information theory to assess models whose structures are defined by graph theory. These hypergraph structures specify which relationships between variables satisfactorily model the data, and allow one to posit relationships that are not merely dyadic (two-way), but of arbitrary ordinality (triadic, tetradic, etc.). In the set-theoretic version of RA (SRA), the input data is a set theoretic relation, that is, a subset of the Cartesian product of the sets of values of the variables. In the information-theoretic version of RA (IRA), the input data

is a frequency or probability distribution. In both, a model is the relation or distribution –

henceforth the word "relation" will be used for either – that maximizes entropy subject to

the constraints of the model's structure. IRA partially overlaps log-linear methods and

logistic regression (Bishop et al., 1978; Knoke & Burke, 1980), and Bayesian networks

and other graphical models (Lauritzen, 1996). In the areas of overlap, RA and these other

methods are typically equivalent. For example, in many contexts the maximum entropy

solutions of RA give identical results to the maximum likelihood solutions of other

methods.

IRA, given a frequency distribution, is an information-theoretic approach to

statistical multivariate analysis, but it can alternatively be given a probability distribution,

with no sample size, for which the analysis is non-statistical. Set-theoretic mappings can

also be treated probabilistically (and non-statistically) with IRA. In its "k-systems"

version (Jones, 1985), IRA can analyze continuous functions of nominal variables by

rescaling the functions and treating them as probability distributions. IRA also has a

Fourier version (Zwick, 2004b) which, in conjunction with the k-systems approach,

resembles regression. The possibility of additional versions of RA is also implicit in

generalized information theory (Klir, 2005) which includes fuzzy distributions.

From the input data one can generate projections; for example, a relation ABZ

generates projections AB, AZ, BZ, A, B, and Z. Projection drops variables by doing a

logical "or" (in SRA) or a summation (in IRA) over the values of the projected variables.

If data is decomposable without information loss into a set of projections (lower-

dimensional distributions or set-theoretic relations) represented by a hypergraph, then the

hypergraph – more precisely, the projections it indicates – determine a calculated

distribution or relation that satisfactorily models the observed data. The possible

hypergraphs define a lattice of structures, and a major concern of RA is how best to

search this lattice for good models. Here, a "lattice" is a partially ordered graph that has

a single upper node (structure) where all variables are included in one relation and a

single lower node (structure) where variables or sets of variables are distributed into

separate relations, i.e., are independent of one another; this is explained more fully below

in *Lattice of Structures*. The RA lattice can also be augmented by graphical models

related to but not encompassed by the RA formalism. By contrast, other methods (e.g.,

logistic regression) that are similar or possibly equivalent to RA in the estimation of

individual models often do not explicitly articulate this lattice of possible models or

provide heuristics for searching it. When applied to genomic data, the RA lattice offers a

taxonomy of epistasis types, where a type is defined coarsely by the simplest structure

that fits the data without loss or defined in more detail by the vector of information values

for different decompositions.

       This paper amplifies the previous use of RA (Shervais et al., 2010) to detect

epistasis in genomic data. This amplification consists of (i) a more complete description

of RA methodology than offered in that earlier study, (ii) the use of RA to define a

taxonomy of epistasis types, and (iii) new extensions of RA methodology. Since the

focus of this paper is on RA methodology itself, no attempt is made to provide definitive

expositions of the similarities and differences between RA and other methods (this is the

subject of ongoing work); nonetheless, the relationships between RA and other methods are briefly considered in the final DISCUSSION section.

METHODS

The following is a compressed explanation of RA; for additional details, see the overview articles of Zwick (2001, 2004a).

Lattice of Structures

Consider two genes or SNPs, A and B, and a disease state, Z. We can think about this simple three variable ABZ relation in two ways: (1) we can regard A and B as inputs (independent variables) and Z as an output (a dependent variable), thus defining a "directed" system; or (2) we can abstain from making an input/output distinction among the variables, thus defining a "neutral" system. ABZ is the data, and it has projections (sub-relations) AB, AZ, BZ, A, B, and Z. A structure is a non-redundant set of projections used to model (compress) data. For a three variable system, the RA lattice of structures is shown in *Figure 1* for a neutral as well as for a directed system.

The top structure is the data itself, the "saturated" model. AB:Z and A:B:Z are alternative bottom (independence) structures for the directed and neutral lattices, respectively. The colon, ":" means "independent of." AB:Z, the bottom structure for the directed system lattice, means that Z is independent of AB; the AB relation allows for the possibility that A and B are associated. A:B:Z, the bottom structure for the neutral system lattice, means that all three variables are independent of one another. (In log-linear notation, one might write {AB}{Z} and {A}{B}{Z} instead of using colons.) The

choice of bottom structure reflects the essential difference between the directed and
neutral system lattices. The directed system lattice only depicts possible relations
between the input variables and the output. It ignores associations between the inputs by
allowing for all possible associations, and does so by including in every structure in the
lattice one relation involving all the inputs. This allows the statistical testing of model
differences that involve only input-output associations, and it insures that all models in
the lattice are hierarchically related to its bottom model, AB:Z.   By contrast, the neutral
system lattice includes all structures hierarchically related to the bottom model, A:B:Z,
where all variables are independent of one another, so this lattice considers not only
input-output relations but also the presence or absence of relations among inputs. From
another point of view, in the neutral system lattice every variable takes its turn as an
output.

Structure AB:AZ allows for associations between A and B and between A and Z;
using this structure to model data asserts that A is a predictor of Z; similarly for AB:BZ.
Since in epistasis A and B are inputs and Z is an output, this suggests using the directed
system lattice, but one structure in the neutral lattice, AZ:BZ, normally not considered for
directed systems, is also of interest; this is commonly known as a "Naïve Bayes" model.
In general, structures of the form PQ:QR mean that the PQ relation is independent of the
QR relation. Since these relations overlap in Q, they are not completely independent;
what the structure really says is that P and R are "conditionally independent" of one
another, conditioned on Q.   PQ:QR does not imply anything about whether P and R are
non-conditionally independent of one another. For example, A and B might be

independent of one another in data ABZ, but associated with one another in structure

AZ:BZ. However, if A and B are independent of one another in data ABZ, they remain

so in structure AB:AZ:BZ. Relations present in a structure indicate which relations in the

data are imposed on and thus satisfied by the model.

Of the nine structures in *Figure 1*, one, namely AB:AZ:BZ, has a loop (is cyclic);

to illustrate, this structure and the ones above and below it are shown in *Figure 2*. A

structure has a loop if something remains after (i) removing variables that occur in only

one relation, (ii) removing redundant relations, and iterating (i) and (ii) (Krippendorff,

1986). Structures with loops have greater computational requirements than those without

loops; they also cannot be interpreted in terms of conditional independence. Structures

are characterized by their complexity. For IRA this means degrees of freedom, which is

discussed below; for SRA, complexity is a more complicated notion that is beyond the

scope of this paper.

The three structures at level 2 and level 3 just permute the variables, so there are 5

general structures (where permutations are not distinguished) and 9 specific structures

(where permutations are distinguished), of which 5 (shown in bold in *Figure 1*) constitute

the directed system lattice. *Table 1* indicates how rapidly the numbers of structures scale

up as the number of variables increases. The fraction of structures that have loops also

markedly increases with the number of variables. For more than a handful of variables,

exhaustive search of all structures becomes impossible, and one of the distinguishing

features of RA is its explicit consideration of the problem of searching large lattices.

"Epistasis" normally connotes a joint association of two or more input variables with an output variable (conceivably more than one output variable, but only one is considered here) that cannot be represented as the result of independent assciations of the individual inputs with the output. For three variables, three RA structures of *Figure 1*, namely ABZ, AB:AZ:BZ, and AZ:BZ, can model associations of A and B with Z that cannot be represented as resulting from independent AZ and BZ associations. ABZ and AB:AZ:BZ always meet this criterion, while AZ:BZ meets it if and only if A and B are not mutually independent in the calculated distribution for this model; more about this below. For four variables, there are 17 structures in which the output depends on all three inputs (*Table 2*). 8 of the 9 of the four-variable directed lattice structures have loops (the data doesn't); one of the 8 additional neutral lattice structures has a loop.

Table 2 illustrates the fact that RA can analyze epistasis involving an arbitrary number of inputs, but the discussion below is restricted to three variable structures, which suffices to explain the method. The next section discusses how structures are used to model data, after which subsequent sections give examples of the three epistasis types shown in *Figure 1*. The RA lattice is then augmented with structures from related graphical modeling formalisms, and a refined version of RA that is state-based is introduced.

Analysis of Data

One of the ways that RA differs from other methods that are similar to it is the variety of types of data that it can analyze. In SRA, the data is a set-theoretic relation or

mapping. In IRA, it is either a probability distribution, a frequency distribution, or a

function. A set-theoretic relation is a set of *observed* nominal (A_i, B_j, Z_k) states which is

a subset of all *possible* (A_i, B_j, Z_k); in a mapping, $A \otimes B \rightarrow Z$, for every (A_i, B_j) there is

only one Z_k. An IRA distribution allows all (A_i, B_j, Z_k) states to occur but assigns a

probability or frequency (possibly 0) to each. Having no sample size, a probability

distribution, $p(A_i, B_j, Z_k)$ is non-statistical, while a frequency distribution $f(A_i, B_j, Z_k) =$

$N\,p(A_i, B_j, Z_k)$ is statistical. Instead of joint probabilities, $p(A_i, B_j, Z_k)$, the data may be

given in terms of conditional probabilities, $p(Z_k \mid A_i, B_j)$. When Z_k is a state of disease,

this conditional probability is "penetrance," and given such data, IRA analyzes a joint

distribution that assumes uniform $p(A_i, B_j)$, i.e., a distribution where all joint probabilities

are equal. In k-systems IRA (Jones, 1985), the data is a function $A \otimes B \rightarrow Z$, where A

and B are discrete, but Z is continuous; this variant of RA does a linear transformation of

Z so it can be treated as a probability.

Applied to particular data, a structure yields a *model*, m, which has a *calculated*

relation ABZ_m, whose entropy, H, is maximum subject to constraints of the projections

included in the structure. The subscript, m, indicates a structure in the lattice applied to

some data; unsubscripted terms refer to the data itself, the "saturated" model. The

entropy is the Hartley entropy for SRA or the Shannon entropy for IRA:

SRA: $H(ABZ_m) = \log_2 |ABZ_m|$

IRA: $H(ABZ_m) = -\Sigma\Sigma\Sigma\ p(ABZ_m) \log_2 p(ABZ_m)$

where $|ABZ_m|$ is the cardinality of the calculated set-theoretic relation, and $p(ABZ_m)$ is the calculated probability distribution for model m (which could be written equivalently as $p_m(ABZ)$).

For example, $ABZ_{AB:AZ:BZ}$ is the relation that has maximum entropy – in IRA, the distribution that is maximally uniform; in SRA, the relation with the maximum number of (A_i, B_j, Z_k) states – while having its AB, AZ, and BZ projections agree with the AB, AZ, and BZ projections of the data. Given data, when one speaks of the AB:AZ:BZ model, one means the calculated relation for the AB:AZ:BZ structure. Another way of thinking about the AB:AZ:BZ model is that it indicates which projections of the data we know. (This is the same in log-linear modeling.) The number of independent parameters needed to specify these projections is the degrees of freedom (df) of the model. The parameters are obtained from the data directly and not fitted. In IRA, for the AB:AZ:BZ model, they are the smallest subset of $p(A_i, B_j)$, $p(A_i, Z_k)$, and $p(B_j, Z_k)$ values sufficient to specify these three projections. The computation of ABZ_m is algebraic if the model does not have loops, but requires iteration if it does; time and space requirements vary with sample size in the former situation, but pose a greater burden in the latter.

Generating and assessing an RA model involves three steps: (i) *projection*, in which projections of the data specified by a structure are obtained, e.g., AB, AZ, BZ are obtained from ABZ; (ii) *composition*, in which the calculated model relation is generated by maximizing entropy subject to the projection constraints, e.g., $ABZ_{AB:AZ:BZ}$ is obtained

from AB, AZ, and BZ; and (iii) *evaluation*, in which the calculated relation is assessed by

being compared to the data, e.g., $ABZ_{AB:AZ:BZ}$ is compared to ABZ.

In the evaluation step, model m can be characterized by I_m, the normalized

information that it captures, which is 0 for the bottom (independence) structure and 1 for

the top (data):

$$I_m = [H(ABZ_{ind}) - H(ABZ_m)] / [H(ABZ_{ind}) - H(ABZ)]$$

(This is a variation on the Kullback-Leibler information distance.) ABZ_{ind} is the

calculated relation for the independence model, which is either A:B:Z (for neutral

systems) or AB:Z (for directed systems). For directed systems, it is useful to quantify

how predictable Z is if one knows A and B; this is expressed as the reduction of the

entropy of Z, knowing A and B, in the calculated distribution for model m:

$$\%\Delta H_m(Z|AB) = 100 [H(Z) - H_m(Z|AB)] / H(Z) = I_m \ \%\Delta H(Z|AB)$$

where $\%\Delta H(Z|AB)$ is the %entropy reduction in Z for the data. Reduction of entropy is

the nominal data analog of variance explained. For IRA, entropy reduction is related to

the model's conditional probability distribution, i.e., to penetrance. For models without

loops, this relation is algebraic:

$$H_m(Z|AB) \ = - \Sigma\Sigma \ p(AB) \ \Sigma \ p_m(Z|AB) \ \log_2 p_m(Z|AB).$$

Both the information and entropy reduction measures do not involve a sample size, so they are non-statistical. For IRA, given a sample size, the Likelihood-ratio Chi-square,

$$L^2_m = 1.3863 \text{ N } [H(Z) - H_m(Z|AB)] = 1.3863 \text{ N } [H(ABZ_{ind}) - H(ABZ_m)]$$

allows one to assess the p-value for the entropy reduction, given the difference in degrees of freedom between the model and independence. Degrees of freedom of a structure is the sum of the degrees of freedom of its relations, corrected for overlaps (Krippendorff ,1986); equivalently (the third equality in the equation that follows), it is the sum, over all relations and sub-relations in the structure, of the product of the cardinalities of the variables minus one (Knoke & Burke, 1980):

$$df_{AB:AZ:BZ} \quad = df_{AB} + df_{AZ} + df_{BZ} - df_A - df_B - df_Z$$

$$= (|AB|-1) + (|AZ|-1) + (|BZ|-1) - (|A|-1) - (|B|-1) - (|Z|-1)$$

$$= (|A||B|-1) + (|A||Z|-1) + (|B||Z| - 1) + (|A|-1) + (|B|-1) + (|Z|-1)$$

Calculating a p-value is not the only way to trade off information and complexity. One can alternatively use the Akaike (AIC) or Bayesian Information Criteria (BIC), which linearly combine these two factors. This is quite different from the way these factors are traded off in the Chi-square calculation of p-values. AIC and BIC also do not require that models being compared are hierarchically nested. The above L^2_m takes the

bottom (independence) model as the reference, but one could also choose the *top* (data) as the reference and assess the error in the model relative to the data with

$$L^2_m = 1.3863 \ N \ T_m = 1.3863 \ N \ [\ H \ (ABZ_m) - H(ABZ) \ ],$$

where T_m is the information theoretic transmission for the model. T_m is the difference between the entropy of the model and the entropy of the data, i.e., the error in the model. When m is the independence model, T_m is also known as "mutual information."

It is sometimes convenient to write equations in terms of transmissions rather than entropies. For example the equation for L^2 given above for testing the difference between a model and independence can be written equivalently as

$$L^2_m = 1.3863 \ N \ (T_{ind} - T_m)$$

T can be thought of not only as an the error in a model but also as a measure of asssociation between variables in the data. For example, $T_{AB:Z}$ is the error in the AB:Z model; equivalently, it is the association between AB and Z (which also means the entropy reduction in Z given the A and B inputs). $T_{AB:Z} - T_{AB:AZ}$ is the difference between the errors of the independence model, AB:Z, and the AB:AZ model, and this transmission difference equals $T_{A:Z}$, the association between A and Z, ignoring B. Transmission, like entropy, can be conditioned on variables. For example, $T_{A:Z|B}$ is the association between A and Z, conditioned on B; it equals $T_{AB:BZ}$. Assume that A is only

indirectly associated with Z via B, which in turn is directly associated with Z. $T_{A:Z|B}$ will

then be zero, but $T_{B:Z|A}$ will not. RA can analyze data where inputs are associated and

can distinguish between direct associations with a disease variable and indirect

associations due to linkage disequilibrium.

For epistasis involving two inputs and an output, *Figure 1* indicates three possible

models: (1) ABZ, the data itself, (2) AB:AZ:BZ, and (3) AZ:BZ, the third of which is

normally relevant only for neutral systems. These structures define a taxonomy of

epistasis, as follows. If the data cannot be decomposed without information loss to (2), it

is here called Type 1 epistasis. If it can be decomposed without loss to (2) but not to any

lower structure, it is Type 2; if it can be decomposed without loss to (3) but not to a lower

structure, it is Type 3. This is summarized in *Table 3*.

Examples of all three types of epistasis are found in the literature. The first type

is straightforward. When ABZ does not have any lossless decomposition, it inherently

has a triadic relation (interaction effect) involving inputs A and B and output Z. This is

epistasis in its strongest form. By contrast, AB:AZ:BZ and AZ:BZ do not have any

triadic relation, but only two dyadic input-output relations; AB:AZ:BZ specifies an

additional relation between the inputs.

The strength of Type 1 epistasis is quantified by $T_{AB:AZ:BZ}$. (The DISCUSSION

section below mentions a different information-theoretic measure that has been

incorrectly used to quantify a triadic interaction.) To test for the significance of such

epistasis, one computes a Chi-square p-value from $L^2_{AB:AZ:BZ}$ and $\Delta df = df_{ABZ} - df_{AB:AZ:BZ}$; if the difference is significant, AB:AZ:BZ does not fit the data. The other hypotheses are similarly tested. If the data can be decomposed still further (to AB:AZ, AB:BZ, or AB:Z) without loss, then both inputs are not associated with the output.

Type 2 epistasis means that the data can be decomposed to AB:AZ:BZ but not lower. Structure AB:AZ:BZ does *not* actually assert non-zero associations between A and B, A and Z, and B and Z; it merely allows for such associations. Similarly, structure AZ:BZ does *not* actually assert a zero association between A and B, despite the absence of the AB relation. Rather, these structures indicate which projections of the data are used to generate the calculated relation, i.e., which projections constrain the *composition* (entropy maximization) step of RA. There can exist data that must be modeled by AB:AZ:BZ because A and B are *not* in fact associated, where the $ABZ_{AZ:BZ}$ distribution would incorrectly show them to be associated, so the AB projection of the data is needed in the model to impose the non-existence of an association. This is discussed further below in Example #2 and also when Bayesian networks are introduced.

While the effects of A and B on Z cannot be separated in Type 1 epistasis because of the 3-way interaction effect, and in Type 2 epistasis because models with loops have no closed form algebraic solution, these effects can be separated in Type 3 epistasis, the weakest type, if A and B are mutually independent in the calculated distribution for this model. In this case, the entropy reduction in Z due to A and B together that is achieved

by AZ:BZ is simply the sum of the entropy reductions due to A and B separately. The

general expression is the following:

$$\Delta H_{AZ:BZ}(Z|AB) = [H(Z) - H(Z|A)] + [H(Z) - H(Z|B)] - [H(A) + H(B) - H_{AZ:BZ}(AB)].$$

If A and B are mutually independent in the model distribution, the rightmost term in

brackets drops out. Since $H(Z) - H(Z|A) = \Delta H_{AB:AZ}(Z|A)$ and $H(Z) - H(Z|B) =$

$\Delta H_{AB:BZ}(Z|B)$, this gives $\Delta H_{AZ:BZ}(Z|AB) = \Delta H_{AB:AZ}(Z|A) + \Delta H_{AB:BZ}(Z|B)$. Using the

proportionality of entropy reduction and normalized information noted earlier, when the

effects of A and B on Z can be separated, as follows:

$$I_{AZ:BZ} = I_{AB:AZ} + I_{AB:BZ}.$$

The above analysis suggests an information theoretic definition of epistasis as involving

entropy reduction in an output that is not the sum of the entropy reductions of the inputs.

(This would be an nominal data analog of defining epistasis in terms of non-additivity of

variance reductions.) This inequality is inherently true for ABZ; it can be expected to

hold for AB:AZ:BZ since useful algebraic equalities cannot be derived for models with

loops; it is true also for AZ:BZ if A and B are not independent in the model distribution.

Additive independence (or lack of it) for entropy reduction corresponds to

multiplicative independence (or lack of it) for penetrance. For structure ABZ, $p(Z|AB)$

cannot be written in terms of a product of $p(Z|A)$ and $p(Z|B)$ because this misses the

triadic interaction effect. For AB:AZ:BZ, this cannot be done because such an algebraic

relation is not derivable for a model with loops. For AZ:BZ, however, which does not

have a loop, one has

$$p_{AZ:BZ}(Z|A_i\,B_j\,) = p(A_i\,Z)\,p(B_j\,Z) \,/\, [\,p(Z)\,p(A_i\,B_j\,)\,]$$

This equation does not exhibit multiplicative independence of i- and j-terms (or additive

independence if one takes the logarithms of both sides), but if $p(A_iB_j) = p(A_i)\,p(B_j)$, i.e.,

A and B are mutually independent, it becomes

$$p_{AZ:BZ}(Z|A_i\,B_j\,) = p(Z|A_i)\,p(Z|B_j) \,/\, p(Z)\,],$$

which shows multiplicative independence (or, taking logarithms, additive independence).

So AZ:BZ might or might not exhibit epistasis, in the strict sense of the term, depending

on the presence or absence of association between the inputs. If IRA data is given as

penetrance values, i.e., as a conditional and not a joint distribution, the absence of

association between inputs is in effect assumed by default; this is true also for SRA data

in the form of a mapping. This means that in these cases, AZ:BZ does not manifest

epistasis, in the strict sense of the term, as the discussion of Example 1 below indicates.

RESULTS

Calculations for the examples below were done by the IRA program, OCCAM

(Willett & Zwick, 2004; Fusion et al., 2010; Zwick, 2010), and by separate SRA and

state-based IRA (Johnson, 2005) programs also developed at Portland State University.

Example #1: Type 3 epistasis

Table 4 presents data from Table 1 of Cordell (2002); B and G there are here

renamed A and B. The data is a genotype-to-phenotype mapping, i.e., not a frequency or

probability distribution. Mappings are naturally analyzed with SRA, but IRA can be used

instead by assigning equal probability to all (A_i, B_j, Z_k) states that occur and zero

probability to states that do not occur. For this data, IRA and SRA decompose to the

same structure; in other data, IRA sometimes decomposes data further than SRA.

IRA gives the results at the top of *Table 5*. Since RA is here analyzing a

probability distribution based on a mapping, there are no p-values. The simplest structure

that fits the data is AZ:BZ. Example #1 thus illustrates epistasis of Type 3. The SRA

analysis of Example #1 is shown at the bottom of *Table 5*. AZ:BZ is again identified as

the simplest model that fits the data, but other models don't have the same information

values as in the IRA analysis.

In the IRA results in *Table 5*, $I_{AZ:BZ} = I_{AB:AZ} + I_{AB:BZ}$, which means that entropy

reductions due to A and B are additive, so this data does not exhibit epistasis in the strict

sense of the term. It is, however, included in this paper as a "Type 3 epistasis" because

Cordell (2002) mentions it as an example of Bateson's definition of epistasis.

Additivity of entropy reductions is likely to be related to the distinction between

interactions that are absolutely absent, that are removable, and that are essential and not

removable (Wu et al., 2009). A removable interaction is one that is not significant on an

odds ratio (OR) or log(OR) or some other risk scale. The example given by Wu of a

removable interaction − revealed by the absence of an interaction in the F-Test of the

log(OR) scale − is classified by RA as Type 3 epistasis.

Example #2: Type 2 epistasis.

Cordell's (2002) second example of epistasis, shown below in *Table 6*(a), is a

penetrance table, but because the penetrance values are only 0 or 1, it could also be

considered a set-theoretic genotype-to-phenotype mapping. With either interpretation,

the table does not provide frequencies for the different genotypes. Assuming p(a) = p(A)

= p(b) = p(B) = .5 and Hardy-Weinberg equilibrium between independent loci A and B,

one obtains from the conditional probabilities of *Table 6(a)* the joint probabilities of

Table 6(b), where p(ABZ) = p(A) p(B) p(Z|AB).

IRA results on *Table 6* are shown at the top of *Table 7*. The simplest structure

that fits the data with no information loss is AB:AZ:BZ. This illustrates epistasis of Type

2. SRA applied to *Table 6* gives different and inferior results, shown at the bottom of

Table 7. IRA decomposes the data of Example #2 further than SRA. IRA analysis of the

conditional (penetrance) distribution of *Table 6* (as opposed to the joint distribution)

gives results very similar but not identical to *Table 7* (top).

Since we assumed that A and B are independent in constructing *Table 6(b)*, *Table*

7 shows that $I_{AZ:BZ} = I_{AB:AZ} + I_{AB:BZ}$, as in Example #1. However, in this case, AZ:BZ

does not fit the data. We need the AB relation in AB:AZ:BZ to guarantee that AB

exhibits no association. If we had modeled the data in *Table 6* with AZ:BZ, we would

have obtained the $ABZ_{AZ:BZ}$ distribution shown in *Table 8*. Its AB projection is shown

there on the right, and it differs from the AB projection of the data shown on the right of

Table 6(b). This illustrates the point made earlier that if data is accurately fit by

AB:AZ:BZ but not by AZ:BZ, this does not mean that A and B are associated; in the

present situation, it means the opposite: that A and B are *not* associated, and the AB

relation in the model is needed to assure this.

Cordell (2002) gives another table – her Table 3 – to illustrate epistasis (a

heterogeneity model), but this table is equivalent to her Table 2 (epistasis "in a general

sense") if states are suitably relabeled, and need not be discussed.

Example #3: Type 1 epistasis (synthetic probability data)

Example #3 comes from an RA study of epistasis (Shervais et al., 2010). The

simulated penetrance data of Model 5 from Table 1 of that study are shown here in *Table*

9 (left). Since penetrance values are continuous and not only 0 or 1, this data can be

analyzed only by IRA and not also by SRA. This penetrance table was constructed so

that there is no main effect of either A or B on Z, and the construction assumed p(a) =

p(A) = p(b) = p(B) = 0.5 and Hardy-Weinberg equilibrium. With these assumptions, the

joint distribution is given in *Table 9* (right). As in Example #2, A and B are mutually

independent.

IRA results on the joint distribution of *Table 9* (right) are shown in *Table 10*.

Because the data were constructed with no main effects, every decomposition has no

information at all. This is the *strongest possible* example of Type 1 epistasis. If IRA is

instead done on the conditional distribution of *Table 9* (left), in effect assuming the AB

frequencies are uniform, similar but not identical results are obtained.

Example #3 was one of five synthetic datasets evaluated in the Shervais et al.

(2010) study. In all of these datasets there was only one epistatic pair, and this fact was

assumed to be known in earlier work on these datasets and in the Shervais study. All five

datasets showed Type 1 epistasis and the same results as *Table 10*. RA performance on

these datasets is summarized in *Table 11*. When 8 or 50 noise SNPs were added, RA

detected the correct epistatic pair 100% of the time for models 1-4, and close to that for

model 5. By contrast, previous work using only 8 noise genes found one of the two

active genes 47% of the time (Ritchie et al., 2004) and both genes only 19% of the time

(Hahn et al., 2003); Hahn used multifactor dimension reduction and Ritchie used neural

nets. The Shervais et al. study also evaluated results in terms of p-values: a gene pair was

counted a false positive if it was not the correct pair yet had a p-value less than 0.000, and

a false negative if it was correct yet did not meet this criterion; *Table 11* gives the error

rates in 1350 tests. The Ritchie et al. and Hahn et al. studies did not report information

on false positives.

Example #4: Type 1 epistasis (real frequency data)

       Examples #1-3 were non-statistical since sample size is meaningless for set-

theoretic relations and undefined for probability distributions. For real data in the form

of frequencies, not probabilities, statistical considerations enter. *Table 12* is from

Shervais et al., (2010), which replicated prior evidence (Cox et al., 1999) for epistatic

interactions in type 2 non-insulin-dependent diabetes between SNPs on chromosomes 2

and 15.

       The IRA analysis of this joint frequency distribution is shown in *Table 13*. p-

values of the models relative to the data (not to independence) are calculated from their

L^2 and Δdf (degrees of freedom) values. The analysis shows that *Table 12* exemplifies

Type 1 epistasis, since the identity of AB:AZ:BZ with the ABZ data can be rejected with

confidence (p=.013). The data cannot be decomposed without significant loss of

constraint. The strength of the constraint in the data is $\%\Delta H(Z|AB) = 8.52\%$, which is

sizeable since entropy involves a logarithm term. Note that the entropy reduction of

AB:AZ:BZ is much smaller, namely 4.27%; this difference shows the strength of the

purely triadic interaction. In the Shervais et al. (2010) study, all of the 36 candidate

epistatic SNP pairs showed Type 1 epistasis.

Unlike Example #3, information here does not immediately drop to zero upon

decomposition to AB:AZ:BZ; this model retains 50% of the information relative to

AB:Z. Possible constraint in AB is suggested by $I_{AB:Z}$ =.135 relative to A:B:Z; this is

conceivable since the data was not filtered to avoid linkage disequilibrium. However, the

p-value for A:B relative to AB (these two distributions are shown in *Table 12* on the

right) is .415, so the non-identity of A:B and AB cannot be asserted.

A Finer Information-Vector Taxonomy

Examples #3 and #4 above are both Type 1 epistasis, but the difference between

their RA results shows that within this type of epistasis, one can distinguish between

instances of epistasis by noting how rapidly information declines as one goes down the

lattice of structures. Reading the IRA results of *Table 10* and *Table 13* from top to

bottom and left to right, gives two *vectors* of information values, as follows:

	ABZ	AB:AZ:BZ	{AB:AZ, AB:BZ}, AZ:BZ	AB:Z, {AZ:B, BZ:A}	A:B:Z
Example #3:	[1.0,	.00,	{.00, .00}, .00,	.00, {.00, .00},	.00]
Example #4:	[1.0,	.57,	{.42, .28}, .43,	.14, {.28, .15},	.00]

where information values are here relative to A:B:Z and values within the curly brackets,

{}, are for models that merely permute the input variables. The information vector

characterizes the data by how rapidly information is lost as ABZ is decomposed. From

this perspective, Example #3 is clearly the most extreme case of epistasis: all the

interaction between A, B, and Z is triadic. There is no main effect due to A or to B and

no AB association, i.e., there are no dyadic constraints at all.   The ABZ relation is

maximally non-decomposable.

Any particular vector of values defines an equivalence class which one can

consider a type of epistasis. A taxonomy based on these classes makes finer

discriminations than those that just note how far down the lattice one can go and still

have 100% information. The idea of such a finer taxonomy based on equivalence classes

of the information vector can be illustrated by applying it to the classification of Li and

Reich (2000), who showed that for two locus penetrance tables having only 0 and 1

values, there are 50 types of epistasis, i.e., 50 different ways (considering symmetries)

that the nine values in the penetrance table can be assigned to either 0 or 1. An IRA

analysis of the penetrance tables of these 50 types yields the taxonomy shown in *Table*

14. There are 11 equivalence classes (a-k) of the information vector. Ignoring k, the last

of these, which is not actually epistatic because Z depends on only one input, there are 5

equivalence classes including 26 models within Type 1 epistasis (ABZ), and 5

equivalence classes including 22 models within Type 2 epistasis. This information vector

approach does not depend on penetrance values being only 0 or 1; it could classify tables

of continuous penetrance values, as done by Hallgrímsdottir and Yuster (2008).

The Li-Reich classification, with its 50 types, is a refined classification which

allows for biological interpretation, but this number of types is somewhat large, and will

scale up greatly with additional input variables. For two inputs, the three type

classification of ABZ, AB:AZ:BZ, and AZ:BZ is perhaps too coarse, but the first two

types expand into 10 equivalence classes, so RA provides both coarse and medium

classifications. (State-based RA, discussed below, provides a fine classification that is

comparable to that of Li-Reich.) For three inputs, the coarse RA classification gives 17

types (*Table 2*), and the approach of Li and Reich would give too many.

Extending the RA Lattice

Extending the RA lattice with related formalisms*.* In Example #4, the null hypothesis that

A and B are independent could not be rejected, and in Example #3, the inputs were

independent by construction. Despite this fact, both examples represent Type 1 epistasis:

the triadic interaction between A, B, and Z makes it impossible to decompose ABZ

without loss. Still, the ABZ model cannot express the fact that A and B might be

independent. Another class of models, Bayesian networks (BN), also known as recursive

models, includes a model that can assert this, namely that $p_m(ABZ) = p(A)\, p(B)\, p(Z|AB)$

is close enough to the data, for which $p(ABZ) = p(AB)\, p(Z|AB)$. This BN model is here

written as $AB_{A:B} :\ _{AB}Z$, where the second term means Z conditioned on AB. This model

asserts independence between A and B, but in the $p(Z|AB)$ term it also asserts a triadic

interaction. This model is not encompassed in the RA lattice, although one might regard

it as a multi-lattice RA model: it is ABZ on the 3-variable lattice, but A:B on the 2-input

lattice.

        While the RA lattice does not encompass this BN structure, the BN lattice is also

incomplete: it does not include the RA structure AB:AZ:BZ, since standard Bayesian

networks do not allow loops. RA and BN thus augment one another. There is yet

another structure, applicable to epistasis, which RA does not encompass. Recall the point

in the discussion of Example #2 that model AB:AZ:BZ might fit the data better than

AZ:BZ because the AB relation in AB:AZ:BZ imposes the *non-association* present in the

data that AZ:BZ does not impose. One can thus also consider an $AB_{A:B}$:AZ:BZ model, a

hybrid between RA and BN. This model's calculated relation has maximum entropy

subject to the constraints that its AZ and BZ projections agree with those of the data, but

its AB projection must agree with $AB_{A:B}$, and not the AB of the data. This class of

models is known as recursive hierarchical and block recursive models (Lauritzen, 1996).

If the RA lattice of *Figure 1* is reduced to the three epistatic structures defined

above as well as these two additional models structures, this gives *Figure 3*, which

defines five epistatic types. With this altered taxonomy, Example #2 should be

reclassified as Type 5 instead of Type 2, and Example #3 should be reclassified as Type 4

instead of Type 1. Types 1 & 4 and Types 2 & 5 each constitute a pair, where the second

of each pair asserts the absence of association between the inputs, and the first permits

linkage disequilibrium. Example #4 might also be regarded as being Type 4, instead of

Type 1, since the independence of A and B could not be rejected.

Extending the lattice with state-based RA*.* The RA lattice of *Figure 1* can be extended

within the RA formalism itself using "state-based" RA. This extension was first

introduced by Jones (1985), as part of his k-systems analysis, and was later integrated

into the mainstream IRA formalism (Johnson & Zwick, 2000; Zwick & Johnson, 2004).

RA, as discussed so far, which can now be labeled "variable-based" RA, uses structures

that are subsets of variables, e.g., AB:AZ:BZ. State-based RA uses structures that can

also specify *particular states* of one or more variables, e.g., $AB:A_1Z_2:B_2Z$. It identifies

which variable *states* are salient, i.e., informationally rich, not merely which variables are

salient. State-based RA resembles the Li-Reich (2000) and Hallgrímsdottir & Yuster

(2008) epistasis classifications, and, like the latter, can analyze continuous penetrance

values.

Within variable-based RA, models with loops can make finer discriminations than

models without loops; state-based RA makes still finer discriminations, as depicted in

Figure 4. Adding loops to a variable-based model or using state-based RA may allow

one to choose a more complex – and thus more predictive – model. (In the figure dotted

lines represent models too complex to be statistically significant; a thick solid line

represents the most complex model that is significant in each of the three model types.)

There is another way a state-based model can be superior to a variable-based model: it

may have more information than a variable-based model (without or with loops) having

the same or even greater complexity (df), as is illustrated below.

The down side of the additional refinement of the state-based approach is that, as

the number of variables increases, its lattice of structures grows even more explosively

than the variable-based lattice (*Table 1*). A variable-based structure is defined

independently of variable cardinalities, so the number of structures in the variable-based

lattice is also independent of these cardinalities. But state-based structures are defined in

terms of individual states, so the number of structures in the state-based lattice expands

greatly with higher variable cardinalities. Computationally, this is handled by doing a

greedy search that successively adds single parameters ($\Delta df=1$) to a starting model, which

is typically variable-based and often the bottom reference model.

State-based RA is illustrated in *Table 15* by its application to Example #3 (the

synthetic data of *Table 9*), which exhibited Type 1 (or Type 4) epistasis. In *Table 15*

variable-based models are interspersed among the state-based models and shown in

italics. Since the data are simulated probabilities, there is no sample size, and one must

select a model by some other means of trading off information and complexity. The table

shows that while variable-based RA indicates that any decomposition loses *all* the

information in the data, state-based RA indicates that a simplification (shaded) of 5

degrees of freedom (from Δdf of 8 to 3) still retains about 91% of the information.

Applying state-based RA to Example #4 (diabetes data of *Table 12*), also an

instance of Type 1 epistasis, gives the results shown in *Table 16*. The left-most column

provides an arbitrary index for each model. The column to its right gives I_m. After Δdf

relative to AB:Z, two p-values are given: p_{cum} and p_{incr}. p_{cum} is the cumulative p-value

relative to the constant reference of the independence model (m=0). p_{incr} is the

incremental p-value relative to the model indexed by m_{incr}; for state-based models this is

the next lower state-based model. Variable based models, in italics, are added to the table

for purposes of comparison. Considering only variable-based loopless models (in italics

with extra spaces), there are no models between $\Delta df = 2$ and $\Delta df = 8$. Considering also

the variable-based model with loops, AB:AZ:BZ, there are no models between $\Delta df = 2$

and $\Delta df = 4$ and between $\Delta df = 4$ and $\Delta df = 8$. (For two inputs and one output there is

only one model with loops, but as the number of inputs increases, most models do have

loops, as was shown dramatically in *Table 1*.) As *Table 16* indicates, state-based models

offer intermediate, more refined, options. For three variables, this effect is modest, but

for more variables it is more substantial, as schematically suggested by *Figure 4*.

If one requires for model acceptance that both cumulative and incremental p-

values be less than 0.05, then the shaded structure in *Table 16*, namely $AB:Z:A_2B_1Z:A_3Z$,

marked by a #, is the best state-based model, and the model, AB:AZ, marked by a *, is

the best variable-based model . This state-based model identifies a triadic effect resulting

from the interaction of specific states A_2, B_1, and Z (since Z is dichotomous, a specific Z

state does not need to be specified), and also identifies a main effect due to genotype A_3.

This model has the same Δdf as the best variable-based model but captures far more

information (.765 vs. .327). One might also note that the state-based model with $\Delta df = 3$

(m = 5) almost meets the $p \leq .05$ standard and captures as much as .907 of the

information. In summary, *Table 16* shows that state-based RA considerably augments

and refines the variable-based analysis of epistasis.

These facts illustrate the capacity of state-based RA to capture more information

with simpler models than variable-based RA. Because of their refinement, state-based

models are likely to have greater power and give fewer false positives than variable-

based models; similarly, within variable-based RA, using models with loops is likely to

have greater power and give fewer false positives than using loopless models.   However,

models with loops take longer to compute than those without loops (because calculated

relations must be generated iteratively if there are loops), and state-based models take

longer to search than variable-based models (because there are many more state-based

models).

DISCUSSION

This paper describes the use of Reconstructability Analysis to study epistasis. It

does not discuss all variations of RA, e.g., the k-systems method (Jones, 1985) that

decomposes continuous functions (not necessarily distributions) of discrete arguments or

the Fourier version of RA (Zwick, 2004b) that minimizes square error rather than

maximizing entropy in the composition step. When both of these variations are

combined, if an output is a sum of quantitative functions of subsets of the inputs, RA will

identify the subsets, i.e., the variables participating in interaction effects without

requiring any assumptions about the mathematical form of the interactions. For example,

if $Z = f_1(A, B) + f_2(B,C)$, this variant of RA will give model AB:BC. Here, independence

is additive not multiplicative, as it normally is in probabilistic systems, and RA comes to

resemble regression and thus standard models of epistasis (Cordell, 2002).   Or, the k-

systems method can be used with standard maximum entropy composition to analyze

continuous phenotypes; when this is done, the lattice of structures is defined only by the

input variables. Nor does this paper describe other possible augmentations of RA offered

by the graphical models literature. For example, there is another class of graphical

models, known as simset models (Studeny, 2004) which could further augment the RA

lattice of structures; simset models allow multiple simultaneous structural hypotheses.

As noted in the Introduction, IRA is graph theory plus information theory, where graph theory defines structures and information theory (for SRA, set theory) evaluates them with data. While information theoretic methods have been used in genomic analysis, (Tsalenko, 2003; Dawy et al., 2006; Chanda et al., 2007; Dong et al., 2007; Kang et al., 2008), these methods have not yet fully exploited the capacities of information theory, nor have they involved searches through large lattices of models. For example, Tsalenko et al. (2003) used an information theoretic approach that involves only single variable loopless models. Information theory is also not always properly applied. For example, Chanda et al., (2007) use two information theoretic measures to quantify interaction:

(a) $-H(A) - H(B) - H(Z) + H(AB) + H(AZ) + H(BZ) - H(ABZ)$

(b) $H(A) + H(B) + H(Z) - H(ABZ)$

The first of these measures equals $T_{A:Z|B} - T_{A:Z}$ or equivalently $T_{B:Z|A} - T_{B:Z}$ and can be either positive or negative. Both positive and negative values can reflect the presence of an ABZ triadic interaction, since one input alters the association of the other input with the output. One might perhaps consider taking the absolute value of this quantity, but an interaction may be present even if this this quantity is zero. For example, assume that B has two states. $T_{A:Z|B} - T_{A:Z} = [\, p(B_1)\, T_{A:Z|B1} + p(B_2)\, T_{A:Z|B2}\,] - T_{A:Z}$. $T_{A:Z|B}$ is an average; $T_{A:Z|B1}$ might be bigger than $T_{A:Z}$ while $T_{A:Z|B2}$ might be smaller, or vice versa, so their sum might be zero even though the association of A and Z is affected by both

states of B. There is thus no value of this measure that is a definitive indication that a

triadic interaction is absent, hence this measure does not properly quantify such an

interaction. The strength of the triadic interaction needs to be measured instead by

$H(ABZ_{AB:AZ:BZ})$ - $H(ABZ)$, the entropy of the model minus the entropy of the data, which

is always positive. ($H(ABZ_{AB:AZ:BZ})$) cannot be written algebraically in terms of the

entropy of the data and its projections, since the model has a loop.) To test for the

significance of this entropy difference, a p-value is calculated from $L^2_{AB:AZ:BZ}$ and $\Delta df =$

$df_{ABZ} - df_{AB:AZ:BZ}$. This issue – the correct way to quantify interactions – has been

elucidated by (Krippendorff, 2009). The second of the Chanda et al. (2007) measures is

$T_{A:B:Z}$, which measures the total constraint in ABZ, not only the constraint involving Z.

$T_{A:B:Z}$ will be non-zero even when Z is independent of both A and B if A and B are

mutually associated (in linkage disequilibrium). It is $T_{AB:Z}$ that measures the constraint

that involves Z, and that is why AB:Z, not A:B:Z, is taken as the bottom reference for

directed systems. So measure (b) also does not properly quantify the triadic interaction.

Readers will note similarities between RA and other methods for studying

epistasis, e.g., logistic regression (LR). LR applied to nominal input variables, where

dummy variables code variable states, is the same as log-linear (LL) modeling; where

these formalisms overlap, RA, LL, and LR are equivalent. Still, RA employs entropy and

transmission measures not normally reported in LL or LR, and these measures are useful

and intuitively easy to understand. LR does not normally evaluate the structure AZ:BZ,

which can model epistasis, because it is not hierarchically related to AB:Z. RA is also

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). When

genotypes are coded nominally, inputs with three or more states are sometimes recoded

with two or more binary variables, and with this type of dummy variable coding, LR

resembles state-based RA. Whether the two are equivalent is under investigation, but

even if they are, there remain differences, at least computational ones: LR maximizes

likelihood, typically with the Newton-Raphson algorithm, while RA maximizes entropy

with Iterative Proportional Fitting. LR software sometimes uses the Wald test instead of

the more robust likelihood test. More critically, LR software is usually not designed for

exploratory purposes and is sometimes unable to handle interactions between many

variables. As already noted, LR as a methodology does not explicitly articulate the lattice

of possible models or provide heuristics for searching it. More generally, RA's fusion of

information theory and graph theory connects it strongly with the graphical models

literature. In its graph theory aspects RA explicitly considers the lattice of possible

structures and offers heuristics for searching this lattice. Also RA has a set-theoretic

version, can analyze continuous outputs, and has a Fourier-based variation. In summary,

while RA and LR (and LL) may be identical where they overlap, RA has distinctive

features, both theoretical and computational, which make it useful for the study of

epistasis.

RA and other nominal data methods are inherently more appropriate to studying

genomic data than other approaches such as neural nets (Ritchie et al., 2003) or support

vector machines (Chen et al., 2008) that presuppose metric information. The predictive

relation in an RA (or LL, LR, or BN) model is precisely the conditional probability of the

discrete output, given the discrete inputs. Since conditional probability of the diseased

output state is penetrance, information theory is a natural and transparent way to

represent relations between genotype and phenotype. Also, the entropy of the nominal

output variable is analogous to variance for continuous variables, and %entropy reduced

is analogous to %variance explained, so the core concepts of RA are intuitively grasped.

By contrast, a neural network fits data via hard-to-interpret weights, and usually does not

include 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.

An earlier study utilizing both simulated and real data (Shervais, 2010) showed

that RA can be used as a tool in genomics research. In that study, RA performed better

than two other multivariate methods (multifactor dimension reduction and neural nets) in

detecting epistasis in simulated data and was also applied successfully to detecting

epistasis in type 2 non-insulin-dependent diabetes data. This paper follows up that

previous study (i) by putting the methods used there in a more encompassing framework,

(ii) by showing that RA has additional capacities not used in that study, and (iii) by

introducing innovations in RA methodology that enhance its potential value for genomic

research. The models that RA offers at different levels of refinement, as shown in *Figure

4*, and the variations in RA methodology discussed above make RA a very flexible

methodology, suitable for studies of genome-wide association, gene-expression, disease

risk factors, and other biomedical applications. For example, one could, in a GWAS shift from fast searches of coarse variable-based loopless models to slow searches of fine variable-based models with loops to slower searches of ultra-fine state-based models while progressively reducing the number of SNPs under consideration. Results of a GWAS of two interacting loci using logistic regression (Marchini et al., 2005) suggest that variable-based loopless RA models would also have the power to analyze an initially very large number of SNPs, since ABZ models in RA are equivalent to fully saturated LR models. By first reducing the number of SNPs with loopless model analysis, it would then be possible to examine epistatic interactions involving many more than two SNPs, using variable-based models with loops and state-based models. Having this spectrum of models and having multiple methodological variants is a distinctive asset of RA.

References

Bishop, Y., Feinberg, S., & Holland, P. (1978) Discrete Multivariate Analysis.

Cambridge: MIT Press.

Chanda, P., Zhang, A., Brazeau, D., Sucheston, L., Freudenheim, J.L., Ambrosone, C.,

and Ramanathan, M. (2007) Information-theoretic metrics for visualizing gene-

environment interactions. Am. J. Hum. Genet. 81, 939-963

Chen, S.H., Sun, J., Dimitrov, L., Turner, A.R., Adams, T.S., Meyers, D.A., Chang, B.L.,

Zheng, S.L., Grönbert, H., Xu, J., & Hsu, F.C. (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., Frigge M, Nicolae D.L., Concannon P., Hanis C.L., Bell G.I., Kong, A. (1999)

Loci on chromosomes 2 (NIDDM1) and 15 interact to increase susceptibility to diabetes

in Mexican Americans, Nat. Genet. 21, 213-215.

Dawy, Z., Goebel, B., Hagenauer, J., Adnreoli, C., Meitinger, T., & Mueller, J.C. (2006) Gene Mapping and Marker Clustering Using Shannon's Mutual Information. IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 3, no. 1, 47-56.

Dong, C., Chu, X., Wang, Y., Wang, Y. Jin, L., Shi, T, Huang, W., & Li, Y. (2007) Exploration of gene-gene interaction effects using entropy-based methods. European Journal of Human Genetics 1-7.

Fusion, J., Willett, K. & Zwick, M. (2010) OCCAM: A Reconstructability Analysis Program. http://www.sysc.pdx.edu/download/papers/woccaman.pdf

Hallgrímsdottir, I.B. & Yuster, D.S. (2008) A complete classification of epistatic two-locus models. BMC Genetics 9, 17-32.

Hahn, L.W., Ritchie M.D., & Moore J.H. (2003) Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics 19, 376-382.

Johnson, M. (2005) State-Based Systems Modeling: Theory, Implementation, and Applications. Ph.D. Dissertation, Portland State University.

Johnson, M. & Zwick, M. (2000) State-Based Reconstructability Modeling For Decision Analysis. In Proceedings of The World Congress of the Systems Sciences and ISSS 2000,

(eds. J.K. Allen & J.M. Wilby), Toronto, Canada: International Society for the Systems Sciences. http://www.sysc.pdx.edu/download/papers/isss_jo_zw.pdf

Jones, B. (1985) Reconstructability Analysis for General Functions. Internat. J. Gen. Sys. 11, 133-142.

Kang, G., Yue, W., Zhang, J., Chi, Y., Zuo, Y. & Zhang, D. (2008) An entropy-based approach for testing genetic epistasis underlying complex diseases. Journal of Theoretical Biology 250, 362-374.

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. New York: Plenum Press.

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. Beverly Hills: Sage.

Krippendorff, K. (1981) An Algorithm for Identifying Structural Models of Multivariate Data. Internat. J. Gen. Sys. 7, 63-79.

Krippendorff, K. (1986) Information Theory. Beverly Hills: Sage.

Krippendorff, K. (2009) Information of interactions in complex systems. Internat. J. Gen. Sys. 38, 669-680.

Lauritzen, S.L. (1996) Graphical Models (Oxford Statistical Science Series), Oxford University Press.

Li, W. & Reich, J. (2000) A complete enumeration and classification of two-locus disease models. Hum. Hered. 50(6), 334-49.

Marchini, J., Donnelly, P., & Cardon, L.R. (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. Nature Genetics 37, 413-417.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J. Sklar, P. de Bakker, P.I., Daly, M.J., Sham, P.C. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559-575.

Ritchie, M., Coffey, C.S., & Moore, J.H. (2004) Genetic Programming Neural Networks as a Bioinformatics Tool for Human Genetics. Genetic and Evolutionary Computation Conference. Seattle, 438-448.

Ritchie, M.D., White, B.C., Parker, J.S., Hahn, L.W., & Moore, J.H. (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, 28-41.

Shervais, S., Kramer, P.L., Westaway, S.K., Cox, N.J., & Zwick, M. (2010) Reconstructability Analysis as a Tool for Identifying Gene-Gene Interactions in Studies of Human Diseases. Statistical Applications in Genetics and Molecular Biology 9(1), article 18. http://www.pdx.edu/sites/www.pdx.edu.sysc/files/media_assets/SAGMB.pdf

Studeny, M. (2004) Probabilistic Conditional Independence Structures (Information Science and Statistics), Springer.

Tsalenko, A., Ben-Dor, A., Cox, N. & Yakhini, Z. (2003) Methods for analysis and visualization of SNP genotype data for complex diseases. Pacific Symposium on Biocomputing, Lihue, Hawaii, 548-561.

Willett, K. and Zwick, M. (2004) A Software Architecture for Reconstructability

Analysis. Kybernetes 33, 997-1008.

http://www.sysc.pdx.edu/download/papers/kenpitf.pdf

Wu, C., Zhang, H, Liu, X. DeWan, A., Dubrow, R, Ying, Z, Yang, Y., & Hoh, J. (2009)

Detecting essential and removable interactions in genome-wide association studies.

Statistics and Its Interface 2, 161-170.

Zwick, M. (2001) Wholes and Parts in General Systems Methodology. In: The Character

Concept in Evolutionary Biology (ed. G. Wagner), pp. 237-256. New York: Academic

Press. http://www.sysc.pdx.edu/download/papers/wholesg.pdf

Zwick, M. (2004a) An Overview of Reconstructability Analysis. Kybernetes 33, 877-

905. http://www.sysc.pdx.edu/download/papers/ldlpitf.pdf

Zwick, M. (2004b) Reconstructability Analysis With Fourier Transforms. Kybernetes 33,

877-905. http://www.sysc.pdx.edu/download/papers/raftpitf.pdf

Zwick, M. (2010) Discrete Multivariate Modeling.

http://www.pdx.edu/sysc/faculty_zwick_publications.html

Zwick, M. & Johnson, M.S. (2004) State-Based Reconstructability Analysis. Kybernetes,

33, 1041-1052. http://www.sysc.pdx.edu/download/papers/mjpitf.pdf

Table 1. Numbers of structures

	Total # of variables	3	4	5	6
neutral	# general structures	5	20	180	16,143
neutral	# specific structures	9	114	6,894	7,785,062
directed, 1 output	# specific structures	5	19	167	7,580
directed, 1 output, no loops	# specific structures	4	8	16	32

Table 2. Epistatic structures for a 3 input, 1 output system

The 17 structures are listed according to their level of decomposition (the actual lattice, i.e., the parent-child relationships, is not shown.) Variable permutations are illustrated by the permutations of ABC:ABZ:BCZ which are ABC:ABZ:ACZ and ABC:ACZ:BCZ. The additional structures from the neutral lattice are naïve-Bayes-like.

Directed lattice structures (9 of 19)	Additional structures from neutral lattice (8)
ABCZ	–
ABC:ABZ:BCZ:ACZ	–
ABC:ABZ:BCZ +two permutations	ABZ:BCZ:ACZ
ABC:ABZ:CZ + two permutations	–
ABC:AZ:BZ:CZ	ABZ:ACZ+ two permutations
–	ABZ:CZ+ two permutations
–	AZ:BZ:CZ

Table 3. Observed & calculated relations in three types of epistasis

Type 1: $ABZ \neq ABZ_{AB:AZ:BZ}$

Type 2: $ABZ = ABZ_{AB:AZ:BZ} \neq ABZ_{AZ:BZ}$

Type 3: $ABZ = ABZ_{AB:AZ:BZ} = ABZ_{AZ:BZ} \neq$ any calculated relation for a lower structure

Table 4. Example #1: data

Cordell (2002) Table 1: "Example of phenotypes (e.g., hair colour) obtained from different genotypes at two loci interacting epistatically, under Bateson's (1909) definition of epistasis." Coding a/a, a/A, and A/A as states 1, 2, and 3, and similarly for B, and coding phenotypes White, Grey, and Black as 1, 2, 3 gives the $A \otimes B \rightarrow Z$ mapping on the right, where genotype AB maps onto phenotype Z.

	Genotype at locus B					B	1	2	3
Genotype at locus A	b/b	b/B	B/B						
a/a	White	Grey	Grey			1	1	2	2
a/A	Black	Grey	Grey		A	2	3	2	2
A/A	Black	Grey	Grey			3	3	2	2

Table 5. Example #1: IRA & SRA results

(The model is followed by its normalized information content. For IRA, the information about Z in A and B in model AZ:BZ is the sum of the information about Z in A in model AB:AZ and the information about Z in B in model AB:BZ.)

	IRA	
	ABZ 1.00	
	AB:AZ:BZ 1.00	
AB:AZ 0.25	AB:BZ 0.75	AZ:BZ 1.00
AB:Z 0.00	AZ:B 0.25	BZ:A 0.75
	A:B:Z 0.00	

	SRA	
	ABZ 1.00	
	AB:AZ:BZ 1.00	
AB:AZ 0.37	AB:BZ 0.74	AZ:BZ 1.00
AB:Z 0.00	AZ:B 0.37	BZ:A 0.74
	A:B:Z 0.00	

Table 6. Example #2: penetrance data

(a) (Cordell (2002) Table 2: "Example of a penetrance table for two loci interacting

epistatically in a general sense." a/a, a/A, A/A and b/b, b/B, B/B are recoded as 1, 2, 3.)

(b) Table converted to a joint ABZ distribution (Left: ABZ. Right: its AB projection.)

(a)

B	1	2	3
A 1	0	0	0
A 2	0	1	1
A 3	0	1	1

(b)

Z	1			2				B	1	2	3
B	1	2	3	1	2	3					
A 1	.00	.00	.00	.0625	.125	.0625		1	.0625	.125	.0625
A 2	.00	.25	.125	.125	.00	.00	A	2	.125	.25	.125
A 3	.00	.125	.0625	.0625	.00	.00		3	.0625	.125	.0625

Table 7. Example #2: IRA results

(The structure is followed by its normalized information content.)

	IRA	
	ABZ 1.00	
	AB:AZ:BZ 1.00	
AB:AZ 0.382	AB:BZ 0.382	AZ:BZ 0.764
AB:Z 0.00	AZ:B 0.382	BZ:A 0.382
	A:B:Z 0.00	
	SRA	
	ABZ 1.00	
	AB:AZ:BZ 0.47	
AB:AZ 0.26	AB:BZ 0.26	AZ:BZ 0.47
AB:Z 0.00	AZ:B 0.26	BZ:A 0.26
	A:B:Z 0.00	

Table 8. Example #2: ABZ$_{AZ:BZ}$ probability distribution

(The AB projection of ABZ$_{AZ:BZ}$ on the right differs from the AB projection of the data.)

Z		1			2						
B	1	2	3	1	2	3		B	1	2	3
A 1	0	0	0	.1429	.0714	.0357		1	.1429	.0714	.0357
2	0	.25	.125	.0714	.0357	.0179	A	2	.0714	.2857	.1429
3	0	.125	.0625	.0357	.0179	.0089		3	.0357	.1429	.0714

Table 9. Example #3: penetrance data & its joint distribution

(Synthetic data (Shervais et al 2010); a/a, a/A, A/A and b/b, b/B, BB are again coded as 1, 2, and 3. Left: penetrance table (heritability = 0.008). Right: ABZ joint distribution with above assumptions.)

B	1	2	3
A 1	.00	.04	.08
2	.06	.04	.02
3	.04	.04	.04

Z		1			2	
B	1	2	3	1	2	3
A 1	.00	.005	.005	.0625	.12	.0575
2	.0075	.01	.0025	.1175	.24	.1225
3	.0025	.005	.0025	.06	.12	.06

Table 10. Example #3: IRA results on penetrance table (Table 9)

(The model is followed by its normalized information content.)

	ABZ 1.00	
	AB:AZ:BZ 0.00	
AB:AZ 0.00	AB:BZ 0.00	AZ:BZ 0.00
AB:Z 0.00	AZ:B 0.00	BZ:A 0.00
	A:B:Z 0.00	

Table 11. Effectiveness of RA in identifying gene-gene interactions (synthetic data)

(The RA results below are from Shervais et al. (2010); Example #3 is model 5 in that

paper. Compare these to the results of dimension reduction (MDR) and neural nets (NN):

with 8 noise SNPs, NN detected at least one correct SNP 47% of the time (Ritchie et al.,

2004), and MDR detected the two correct SNPs 19% of the time (Hahn et al., 2003). FP

= false positives; FN = false negatives; Error rate = (FP+FN)/#tests; the Shervais paper

has typos in the error rates for models 4 and 5, as the FP and FN numbers given there

indicate.)

Genetic model	Heritability	% both active genes in top RA model 8 noise SNPs n = 30	% both active genes in top RA model 50 noise SNPs n = 30	Error rate 8 noise SNPs n=30
1	0.053	100%	100%	0
2	0.051	100%	100%	0
3	0.026	100%	100%	0
4	0.012	100%	100%	.007
5	0.008	93%	80%	.005

Table 12. Example #4: joint frequency distribution, diabetes data (Cox et al., 1999)

(Left: a frequency distribution for SNPs A35 and B47 and disease, Z: Z=1 controls; Z=2

cases. Right: the AB projection of the data & its independence model distribution.)

Z		1			2	
B	1	2	3	1	2	3
A 1	18	32	5	27	30	13
2	27	13	2	9	21	5
3	2	6	2	1	1	0

AB			A:B		
45	62	18	49.1	60.2	15.8
36	34	7	30.2	37.1	9.7
3	7	2	4.7	5.8	1.5

Table 13. Example #4: IRA results

(Model, [I_m for neutral lattice; I_m and %$\Delta H_m(Z|AB)$ for directed lattice], **Δdf** and p-value

relative to the data, ABZ, *not* relative to independence. The big difference between

%$\Delta H_{ABZ}(Z|AB)$ = 8.52 and %$\Delta H_{AB:AZ:BZ}(Z|AB)$ = 4.27 indicates the strength of the

triadic interaction.)

	ABZ [1.00;1.00, 8.52] **0** 1.0	
	AB:AZ:BZ [0.569; 0.502, 4.27] **4** .013	
AB:AZ [0.418; 0.327, 2.79] **6** .009	AB:BZ [0.281; 0.169, 1.44] **6** .002	AZ:BZ [0.429,-] **8** .033
AB:Z [0.135,0.00] **8** .001	AZ:B [0.283,-] **10** .021	BZ:A [0.146,-] **10** .005
	A:B:Z [0.00,-] **12** .004	

Table 14. RA taxonomy of the Li-Reich penetrance tables

(EC = equivalence class based on information vector.)

EC	Type	Information vector								
		ABZ	AB:AZ:BZ	AB:AZ	AB:BZ	AZ:BZ	AB:Z	AZ:B	BZ:A	A:B:Z
a	ABZ	1	0	0	0	0	0	0	0	0
b	ABZ	1	0.16	0.07	0.07	0.15	0	0.07	0.07	0
c	ABZ	1	0.33	0.33	0.00	0.33	0	0.33	0	0
d	ABZ	1	0.42	0.20	0.20	0.40	0	0.20	0.20	0
e	ABZ	1	0.55	0.38	0.07	0.46	0	0.38	0.07	0
f	AB:AZ:BZ	1	1	0.33	0.33	0.67	0	0.33	0.33	0
g	AB:AZ:BZ	1	1	0.38	0.38	0.76	0	0.38	0.38	0
h	AB:AZ:BZ	1	1	0.69	0.07	0.76	0	0.69	0.07	0
i	AB:AZ:BZ	1	1	0.39	0.39	0.78	0	0.39	0.39	0
j	AB:AZ:BZ	1	1	0.60	0.20	0.79	0	0.60	0.20	0
k	AZ:B	1	1	1	0	1	0	1	0	0

EC	Li-Reich Models
a	84, 98
b	78, 85, 86, 94, 99, 106, 113, 114, 170
c	14, 21, 97, 28, 42, 70
d	10, 12, 17, 68
e	29, 30, 43, 101, 108
f	11, 13, 19, 26, 41, 69
g	27, 45, 186
h	15, 23, 57, 58, 59, 61
i	1, 2, 16
j	3, 5, 40, 18
k	7, 56

Table 15. Example #3: State-Based RA results (non-statistical)

(Unlike *Table 13*, Δdf here is given relative to the directed system independence model.

Results for variable-based models (italics) are from *Table 10*. Variable-based models

without loops are written with extra spaces.)

I_m	$\Delta df_{AB:Z}$	Structure
1.000	*8*	*A B Z*
1.000	4	$AB:Z:A_1B_1Z:A_2B_3Z:A_1B_3Z:A_2B_1Z$
0.0	*4*	*AB:AZ:BZ*
0.906	3	$AB:Z:A_1B_1Z:A_2B_3Z:A_1B_3Z$
0.756	2	$AB:Z:A_1B_1Z:A_2B_3Z$
0.0	*2*	*A B : A Z* or *A B : B Z*
0.535	1	$AB:Z:A_1B_1Z$
0.000	*0*	*A B : Z*

Table 16. Example #4: SB-RA results (statistical)

(Results for variable-based models are in italics; variable-based models without loops are written with extra spaces; # is the best state-based model; * is the best variable-based model; m = model index; m_{incr} = reference model for p_{incr}.)

m	I_m	$\Delta df_{AB:Z}$	p_{cum}	p_{incr}	m_{incr}	Structure
11	*1.000*	8	0.0014	0.9684	10	*A B Z*
10	1.000	7	0.0007	0.8736	9	$AB:Z:A_2B_1Z:A_3Z:A_1B_2Z:A_3B_3Z:B_3Z:A_3B_1Z:A_2B_2Z$
9	0.999	6	0.0003	0.5036	8	$AB:Z:A_2B_1Z:A_3Z:A_1B_2Z:A_3B_3Z:B_3Z:A_3B_1Z$
8	0.981	5	0.0002	0.3022	7	$AB:Z:A_2B_1Z:A_3Z:A_1B_2Z:A_3B_3Z:B_3Z:A_3B_1Z$
7	0.939	4	0.0001	0.3695	5	$AB:Z:A_2B_1Z:A_3Z:A_1B_2Z:A_3B_3Z$
6	*0.502*	4	*0.0129*	*0.1098*	3	*AB:AZ:BZ*
				0.0148	2	
5	0.907	3	0.0000	0.0575	4	$AB:Z:A_2B_1Z:A_3Z:A_1B_2Z$
4#	0.765	2	0.0001	0.0045	1	$AB:Z:A_2B_1Z:A_3Z$
3*	*0.327*	2	*0.0161*	*0.0161*	0	*A B : A Z*
2	*0.169*	2	*0.1188*	*0.1188*	0	*A B : B Z*
1	0.445	1	0.0008	0.0008	0	$AB:Z:A_2B_1Z$
0	*0.000*	0	*1.00*	*1.0000*		*A B : Z*

Figure 1. Lattice of Specific Structures for a 2 input, 1 output system

(The full set of structures is the lattice for neutral systems; the bold structures constitute

the lattice for directed systems with inputs A & B and output Z. The three structures that

model epistasis are boxed. The order of variables in a relation is arbitrary, e.g., AB = BA,

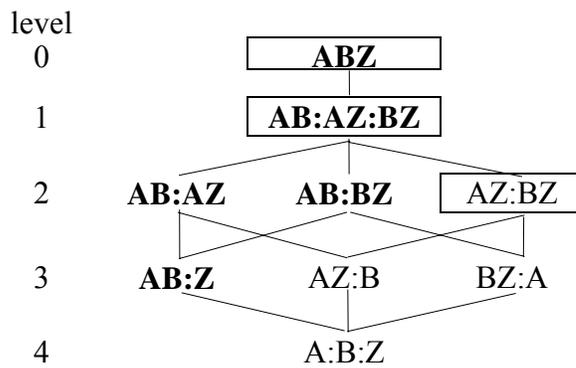and the order of relations in a structure is also arbitrary, e.g., AB:AZ:BZ = AZ:AB:BZ.)

level
0   ┌──────────────┐
                │   **ABZ**   │
                └──────────────┘
1   ┌──────────────┐
                │ **AB:AZ:BZ**   │
                └──────────────┘

2   **AB:AZ**   **AB:BZ**   ┌──────────┐
                                    │ AZ:BZ   │
                                    └──────────┘

3   **AB:Z**   AZ:B   BZ:A

4   A:B:Z

*Figure 2. **Examples of structures with and without loops***

(AB:AZ:BZ, which has a loop, is shown with ABZ above it and AB:BZ below it, which
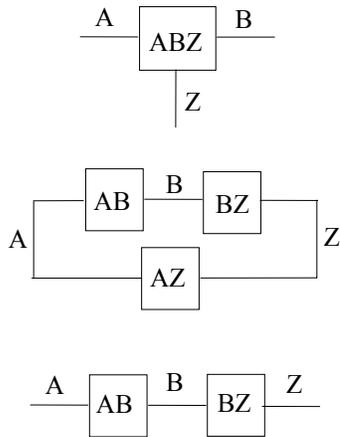
do not. Lines are variables and boxes are relations.)

Figure 3. RA, BN, and hybrid RA-BN 2 input, 1 output epistatic structures

(Five epistasis types are now indicated: 3 from RA plus 2 new types.)

level	RA models	BN & RA-BN hybrids	df
0	1. ABZ		17
1	2. AB:AZ:BZ	4. $AB_{A:B} : {}_{AB}Z$	13
2	3. AZ:BZ	5. $AB_{A:B} : AZ:BZ$	9

Figure 4 Degrees of refinement of RA models

(This is a general scheme for many variables; for 3 variables only one model has loops.)

Degrees of
freedom

Variable-based   State-based

Loopless   Allowing loops