10-3-2012

Reconstructability of Epistatic Functions

Martin Zwick
Portland State University, zwick@pdx.edu

Joe Fusion
Portland State University

Beth Wilmot
Oregon Health & Science University

Citation Details

Zwick M, Fusion J and Wilmot B: Reconstructability of Epistatic Functions. journal of Molecular Engineering and Systems Biology 2012, 1:4 http://dx.doi.org/10.7243/2050-1412-1-4


Journal of
Molecular Engineering & Systems Biology
ISSN 2050-1412

HOAJ
Herbert Open Access Journals

Short report Open Access

Reconstructability of Epistatic Functions

Martin Zwick[1*], Joe Fusion[1] and Beth Wilmot[2]
Correspondence: zwick@pdx.edu
[1]Systems Science Graduate Program, Portland State University, Portland OR 97207-0751
[2]Division of Bioinformatics and Computational Biology, Department of Medical Informatics and Clinical Epidemiology and Oregon Clinical and Translational Research Institute, Oregon Health & Sciences University, 3181 S.W. Sam Jackson Park Rd., Portland, 97239-3098.

Abstract

Background: Reconstructability Analysis (RA) has been used to detect epistasis in genomic data; in that work, even the simplest RA models (variable-based models without loops) gave performance superior to two other methods. A follow-on theoretical study showed that RA also offers higher-resolution models, namely variable-based models with loops and state-based models, likely to be even more effective in modeling epistasis, and also described several mathematical approaches to classifying types of epistasis.

Methods: The present paper extends this second study by discussing a non-standard use of RA: the analysis of epistasis in quantitative as opposed to nominal variables; such quantitative variables are, for example, encountered in genetic characterizations of gene expression, *e.g.,* eQTL data. Three methods are investigated for applying variable- and state-based RA to quantitative dependent variables: (i) k-systems analysis, which treats continuous function values as pseudo-frequencies, (ii) b-systems analysis, which derives continuous values from binned DVs using expected value calculations, and (iii) u-systems analysis, which treats continuous function values as pseudo-utilities subject to a lottery. These methods are demonstrated and compared on synthetic data.

Results: The three methods of k-, b-, and u-systems analyses, both variable-based and state-based, are then applied to a published SNP dataset. A preliminary search is done with b-systems analysis, followed by more refined k- and u-systems searches. The analyses suggest candidates for epistatic interactions that affect the level of gene expression. As in the synthetic data studies, state-based RA is more powerful than variable-based RA.

Conclusions: While the previous RA studies looked at epistasis in nominal (or discretized) data, this paper shows that RA can also analyze epistasis in quantitative expression data without discretizing this data. Since RA can also model epistasis in frequency distributions and detect linkage disequilibrium, its successful application here also to continuous functions shows that it offers a flexible methodology for the analysis of genomic interaction effects.

Keywords: epistasis, gene-gene interactions, gene expression, eQTL, Reconstructability Analysis, information theory, graphical models, OCCAM, bioinformatics, k-systems analysis, u-systems analysis, function decomposition, state-based modeling.

Background

Reconstructability Analysis (RA) is a modeling methodology developed in the systems science community [1-6] based on the work of Ashby [7]. It has recently been used to detect epistasis in genomic data [1], where its performance was superior to multifactor dimension reduction and neural nets. Only the simplest RA models were used, namely variable-based models without loops, so the potential value of RA for studying epistasis was not fully explored. A follow-on paper [2] showed (a) that epistasis can be modeled by RA even more effectively with more refined variable-based models with loops and with state-based models (which nearly always have loops), (b) that in classifying epistasis one can use multiple rather than single models, and (c) that the RA classification of types of epistasis can be augmented with structures from other graphical modeling techniques.

This second [2] study still did not examine all the ways that RA can be used to study epistasis. In particular, it did not consider the use of RA to analyze epistasis affecting *quantitative* as opposed to *nominal* measures, as is encountered in the study of gene expression, *e.g.,* in eQTL data. This paper sets out the theory for such non-standard use of RA, showing with examples how analysis is done. Systematic testing on simulated data or analysis of real data for the purpose of articulating specific hypotheses about epistatic interactions are tasks left for the future. The aim of this paper is to explain and further develop RA methodology for analyzing functions, and thus lay the groundwork for such future investigations.

Methods

Methodological distinctions

RA is a fusion of information theory and graph theory developed

doi: 10.7243/2050-1412-1-4

Table 1. Methodological distinctions for RA

1. DATA TYPE	frequency or probability distribution
	function
	explicit function
	sample data
2. MODEL TYPE	variable-based (VB)
	without loops
	with loops
	state-based (SB) (usually has loops)
3. SYSTEM TYPE	directed system (input/output distinction)
	deterministic
	non-deterministic
	neutral system (no i/o distinction)
4. REPRESENTATION	b-systems (distribution or function)
	k-systems (function)
	u-systems (function)
5. COMPUTATIONAL TASK	search (exploratory) over models
	fit, test (confirmatory), or use one model
6. COMPOSITION ALGORITHM	maximum entropy (standard) composition
	Fourier composition

within the systems science community that partially overlaps other 'graphical models' methods that are better known, such as log-linear modeling and Bayesian networks. **Table 1** summarizes various distinctions within RA methodology, and reviews of this methodology are available [**8**,**9**]. Because these distinctions are not completely orthogonal, their discussion below does not simply follow the table sequentially, but the table is still a useful organizing framework. The first distinction points to the purpose of this paper: to explore the use of RA to analyze functions; but distributions are also discussed as part of necessary preliminaries. All the alternatives within the other methodological distinctions are relevant to the analysis of functions.

Model types

In standard uses of RA, both input and output variables (independent and dependent variables) are nominal with an arbitrary number of states. The input variables might be SNPs, for example, and an output variable might be the presence or absence of disease, or its surrogate, case vs. control. To illustrate how RA is applied to epistasis, let the input variables be A, B, C, etc. and the output variable Z, and let the epistatic interaction be between A, B, and Z, with C having no relation to Z. This is captured in model m_1 = ABC:ABZ, which is a variable-based model without loops, the only type of RA model used in the [**1**] study. This model has two relations, ABC and ABZ, the colon ":" meaning "independent of." The first relation allows for but does not necessarily assert association between the input variables. The second is the epistatic relation, which asserts that A and B jointly predict Z. In all models, the order of relations and the order of variables within a relation are arbitrary, *e.g.,* ABC:ABZ is the same as ZAB:CBA.

Epistasis might be analyzed with more complex models, namely those with loops. Consider the variable-based model *with* a loop: m_2 = ABC:AZ:BZ. The loop in this model is the cycle of interactions between A and B (within ABC), B and Z, and Z and A. The first relation again allows association between input variables; there are now two predicting relations: one in which A predicts Z, and the other in which B predicts Z. These two interactions are integrated with the maximum entropy principle, but there is no three-way interaction effect, so epistasis in m_2 is of a weaker type than what exists in m_1 [**2**]. A more complex way to analyze epistasis is to use state-based models. Consider, for example, a state-based model m_3 = ABC:Z:A_1B_1Z:$A_2B_3Z_1$ (which has a loop), where subscripts indicate particular states of the indicated nominal variables. The first relation again allows for association between input variables. The Z component imposes the marginal distribution of the output variable. The third and fourth relations represent the power of the specific state A_1B_1 to predict Z and an anomalously high or low probability of the $A_2B_3Z_1$ state. Variable-based models with loops and state-based models can more effectively characterize epistasis than variable-based models without loops because they make finer discriminations. Though their structures are more complex, they make more economical use of their degrees of freedom [**2**].

Three RA steps: projection, composition, evaluation

In standard variable-based or state-based RA, the data is usually an observed frequency distribution f(ABCZ) = N p(ABCZ), where N = sample size; or the data could be given as a probability distribution p(ABCZ) without any sample size. p(ABCZ) means the distribution $p(A_i,B_j,C_k,Z_l)$, for all i,j,k,l; the short notation will generally be used except where the long notation is needed for clarity. This is for information-theoretic (or probabilistic) RA. There also exists a set-theoretic ("possibilistic") RA where the data is a set-theoretic relation or mapping; this type of RA is of limited interest for studying epistasis, so this methodological distinction is not included in **Table 1**.

Assume as before that Z depends on A and B but not C. An RA model (*e.g.,* m_1, m_2, or m_3) stipulates a *calculated* ABCZ distribution, p_m(ABCZ), that is generated from the projections of the data specified by the relations present in the model. The first step in constructing a model is projection, in which these relations are calculated from the data. In the composition step that follows, the relations are integrated using maximum entropy to obtain the calculated distribution for the model. For example, for data ABCZ and model ABC:ABZ, the projection step calculates ABC and ABZ, and the composition step joins these relations into the calculated $p_{ABC:ABZ}$(ABCZ) = p(ABC) p(ABZ) / p(AB). For models with loops, however, composition is not algebraic, but must be iterative. What composition does is generate a calculated distribution that is as uniform as possible (has maximum entropy) as long as its calculated projections for

doi: 10.7243/2050-1412-1-4

the relations in the model agree with the corresponding projections of the data. As indicated in **Table 1**, there is a composition algorithm that is an alternative to maximum entropy, and which has been tested in a preliminary way [**10**], but this algorithm hasn't been implemented in the RA software used in this project, and isn't considered in this paper.

The evaluation step assesses model goodness, which depends on (i) the difference between the model distribution and some reference distribution and (ii) the degrees of freedom of the model relative to that reference. The reference model is usually either the data, ABCZ, or the independence model, ABC:Z, in which no input is associated with the output. For "neutral systems," discussed below, the reference is instead sometimes the uniform distribution. If the reference is the data, one wants the calculated p_m to agree with the observed p. The difference between these two distributions (the error) is usually given in terms of the difference between calculated and observed Shannon entropies of these distributions, $H_m(ABCZ) - H(ABCZ) = T_m$, where T_m is "transmission," also known as the Kullback-Leibler distance. If the reference is the independence model, one wants the calculated p_m to differ as much as possible from a calculated p_{ind}, as long as the difference is significant statistically, so that the model captures a maximal fraction of the information (I) in the data. Model goodness with respect to difference from the reference distribution is often tabulated as $\%I = 100 (H_{ind} - H_m) / (H_{ind} - H_{data})$ where ind=independence; when the reference is the uniform distribution, $\%I = 100 (H_{uniform} - H_m) / (H_{uniform} - H_{data})$.

Model use; neutral systems

Model m is *used* by obtaining the conditional distribution $p_m(Z|AB)$, from $p_m(ABCZ)$, which allows one to predict Z from A and B. Because Z is nominal in standard RA, using this conditional distribution for prediction is doing a *classification*. Here, RA is closely related (but not necessarily completely equivalent) to log-linear methods, logistic regression, and Bayesian networks. The presence of a sample size allows models and their predictions to be tested for statistical significance.

So far, discussion has concerned "directed" systems, which have inputs and at least one output. RA also includes the analysis of "neutral systems," where inputs and outputs are not distinguished; in effect, all variables are inputs (or outputs). For the analysis of quantitative functions, one RA representation – "k-systems" analysis – uses neutral systems; the two other RA representations – "b-systems" and "u-systems" analysis – use directed systems. Neutral systems can also be used to detect association between inputs, *i.e.,* linkage disequilibrium among SNPs.

Software

RA used in this study is implemented in the program OCCAM [**11**,**12**], named after "Organizational Complexity Computation and Modeling" and the famous razor (principle of economy) sometimes spelled this way. OCCAM is web-accessible, and available for use by contacting the author (Zwick). State-based searches reported in this paper were done with an auxiliary program [**13**], but state-based searches have recently been incorporated into OCCAM.

Data

Consider data where Z is not nominal but quantitative (continuous), as occurs for example in measures of gene expression, and distinguish two types of data: (a) the data is given as an *explicit function* of nominal inputs A, B, and C, *i.e.,* as $Z_{ijk} = f(A_i, B_j, C_k)$, and (b) it is *sample data* consisting of input and output values for N cases, {A(c), B(c), C(c), Z(c)}, c = 1 to N, where A(c) is some nominal A_i value, and similarly for B and C.

An explicit function is a simple mapping of nominal arguments onto real values, not a sample of input and output values where the same input might be observed to have multiple output values. If $f(A_i, B_j, C_k)$ is penetrance, *i.e.,* the probability of disease given the inputs, one can regard it as the conditional probability $p(Z_1 | A_i, B_j, C_k)$, where Z is a binary output with Z_0 being the non-diseased state and Z_1 being the diseased state. By thus adding a nominal output variable, the data is amenable to standard probabilistic analysis. If, however, f is given as an arbitrary function, one does not have a probability distribution. The problem is no longer one of classification, but is the modeling of a continuous function. Log-linear methods and Bayesian networks are not applicable, but RA, in its k-systems or u-systems forms, can still be used.

In sample data, input states can repeat with different function values. One can bin (discretize) the output Z values, and thus convert the data into a probability distribution to which standard RA (and log-linear methods and Bayesian networks) apply. Unless there is a binning scheme standard for the type of data being examined, a number of bins is chosen, and bin boundaries are assigned to equalize, as much as possible, the number of data points falling in each bin. ("Equal sampling" binning makes more efficient use of the data than "equal interval" binning.) A reasonable minimum default for the number of bins is three, since this allows the detection of nonlinear effects (which two bins would not). If sample size permits, more bins for Z will give greater precision in calculated Z values for the b-systems approach. In the results reported below on real data, four bins were used. Input variables that are continuous must also be binned. An input that in preliminary calculations is highly predictive can be rebinned with more bins; and one could select a number of bins that optimizes the reduction per input bin of the entropy of Z [**14**]. In general, the number of bins assigned to all the variables is a "scarce resource", since the complexity of models that can be considered is limited, for any sample size, by the cardinalities of the variables.

The sample size allows one to calculate statistical significance, and quantitative Z predictions can be recovered by a simple expected value calculation as discussed below.

doi: 10.7243/2050-1412-1-4

Since one is analyzing a distribution where input states can have different binned output values, the problem is stochastic. This approach is referred to as "b-systems" analysis ("b" for binning). Or, one can convert sample data to an explicit function by averaging the f(c) values for each (A_i, B_j, C_k) input state; the data is then the function $Z_{ijk} = f_{ave}(A_i, B_j, C_k)$, which can be modeled with either k-systems or u-systems analysis, as explained below. In doing this, sample size information is discarded, and the analysis is non-statistical and deterministic.

For both explicit functions and sample data, an important issue is whether or not function values are available for the *complete* input space. It is assumed in this paper that the data provides a fully defined function. How well the b-, k-, and u-systems approaches reconstruct partially defined functions requires a separate study.

b-systems analysis

The b-systems approach is just standard RA with one additional calculation that obtains a continuous predicted value for the output variable. The continuous prediction of Z for any input state is obtained from the conditional distribution $p(Z_l \mid A_i,B_j,C_k)$ by calculating Z's expected value for that input state by summing over bins. Each bin, b, has assigned to it some continuous value, Z_b, which could be (i) the midpoint of the range of Z values that define this bin or – preferably – (ii) the mean or (iii) the mode of the Z values that were discretized to this bin. (Bin values can be assigned in OCCAM using its rebinning option.) Z's expected value is $\langle Z_{ij} \rangle = \Sigma_b Z_b\, p(Z_b \mid A_i,B_j)$. The entropy of this conditional distribution for each input state is a measure of the confidence that should be associated with this expected Z value; one can also calculate a variance for this expected value.

Binned gene expression data were used in a preliminary RA study [15] but expected value predictions were not examined there. That study employed only simple variable-based models without loops, but it successfully replicated and extended SNP-gene expression associations in a public dataset [16].

Table 2. Hallgrímsdottir and Yuster (HY) data
f is the original function (data). f_{k6}–f_{k8} are functions reconstructed from state- and variable-based k-systems models in **Table 3**. f_{u6}–f_{u9} are functions reconstructed from state- and variable-based u-systems models in **Table 5**, further below.

			k-systems			u-systems		
A	B	f	f_{k6}	f_{k7}	f_{k8}	f_{u6}	f_{u8}	f_{u9}
0	0	145.6	151.4	143.1	151.4	145.6	145.6	145.6
0	1	157.1	151.4	101.2	151.4	157.1	157.1	157.1
0	2	0	0.0	58.3	0.0	30.3	19.6	0.0
1	0	123.1	123.1	87.9	123.1	123.1	123.1	123.1
1	1	29.6	36.4	62.2	24.5	30.3	19.6	24.5
1	2	33.2	36.4	35.8	24.5	30.3	19.6	24.5
2	0	18.6	36.4	56.3	24.5	30.3	19.6	24.5
2	1	16.5	36.4	39.8	24.5	30.3	19.6	24.5
2	2	83.9	36.4	22.9	83.9	30.3	83.9	83.9
		rms-dev	18.53	40.17	5.44	21.41	8.69	4.72
		R^2	0.90	0.52	0.99	0.86	0.98	0.99
		Δdf	3	4	4	3	4	5
reference for Δdf			uniform			AB:Z		
		models	k_6=A_0B_2:A_0:A_1B_0			u_6= AB:Z:$A_0B_1Z_0$:$A_0B_0Z_1$:$A_1B_0Z_1$		
			k_7=A:B			u_8= u_6:$A_2B_2Z_1$		
			k_8= k_6:A_2B_2			u_9= u_6:$A_2B_2Z_1$:$A_0B_2Z_1$		

k-systems analysis

The use of "k-systems" analysis to compress (decompose) functions of nominal variables was first proposed by Jones [17], who joined it to state-based analysis, but the k-systems approach can also be used with variable-based analysis [18]. This approach applies to data that is an explicit function of nominal variables. If sample data is given, it is converted to an explicit function by averaging function values for every input state. The within-input-state variation in the function value and the frequencies of the input states are ignored.

The central idea in k-system analysis is to linearly scale the function (its average value for every input state) so it can be treated as a pseudo-probability, do standard reconstructability analysis on this distribution, and then do an inverse scaling of reconstructed pseudo-probabilities to get an approximation to the original function. Jones called the scaled function "k" in honor of George Klir, one of the pioneers of RA, hence the name "k-systems analysis." Here the scaled function is called p, because it is treated as if it were a probability. Scaling is done as follows:

p(ABC) = a f(ABC) + b, where a and b are constants, such that $0 \le p(ABC) \le 1$ for all states of ABC, and $\Sigma\Sigma\Sigma\, p(ABC) = 1$

b is chosen to bring negative values of f to zero or to a slightly positive value (negative values are not relevant to gene

Table 3. State- and variable-based k-systems analysis of HY data
H = Shannon entropy, %I = % information, df = degrees of freedom. Variable-based models are in italics. Models are neutral systems.

	Model	H	%I	Δdf
10	*AB*	*2.596*	*100.0%*	*8*
9	A_0B_2:A_0:A_1B_0:A_2B_2:A_2	2.597	99.8%	5
8	A_0B_2:A_0:A_1B_0:A_2B_2	2.607	98.2%	4
7	*A:B*	*2.981*	*32.9%*	*4*
6	A_0B_2:A_0:A_1B_0	2.681	85.2%	3
5	A_0B_2:A_0	2.795	65.3%	2
4	*A*	*3.069*	*17.6%*	*2*
3	*B*	*3.082*	*15.3%*	*2*
2	A_0B_2	3.000	29.6%	1
1	*Uniform*	*3.170*	*0.0%*	*0*

doi: 10.7243/2050-1412-1-4

expression, but this discussion applies to functions in general); a is then determined by the normalization condition that the probabilities sum to 1. The scaled p(ABC) is then subjected to standard neutral system RA. Suppose the variable-based model AB:BC with its calculated distribution, $p_{AB:BC}(ABC)$, is recommended by the analysis. The reconstructed function is obtained by the inverse scaling, *i.e.,*

$$f_{AB:BC}(ABC) = (p_{AB:BC}(ABC) - b) / a$$

One can as well use state-based modeling to decompose an observed p(ABC) and obtain a calculated $p_m(ABC)$. Jones joined scaling-decomposition-rescaling with state-based modeling, but these two procedures are separable.

The k-systems RA approach can be compared to the analysis of variance without replication, also used to study the association between nominal inputs and a quantitative output that is unique for each input state. Jones [**19**] and Gouw and Jones [**20**] discuss these alternative methodologies and argue for the merits of their approach, which differs from multivariate experimental design methods in making no assumptions about the functional form of the relation between inputs and output or of the distribution of errors. However, k-systems analysis is not statistical, and does not offer confidence estimates, so its results need to be subjected to cross-validation studies in which test data assess the generalizability of models obtained from training data.

The following tables show the k-systems analysis of a function of two variables, and the efficacy of state- as compared to variable-based modeling. **Table 2** is data from Hallgrímsdottir and Yuster [**21**], who classify types of two-locus epistasis. Since the lowest function value is 0, the function needs only ratio scaling by a to convert it to a pseudo-probability. A and B have cardinality 3, so the data's degrees of freedom (df) is 8, assuming knowledge of the average f.

Table 3 shows the results of both state- and variable-based k-systems analysis. The results show that the function can be represented by many fewer degrees of freedom if state-based models are used. State-based model #6, $A_0B_2{:}A_0{:}A_1B_0$, with Δdf=3, has one fewer degree of freedom than the variable-based model #7, A:B; yet model #6 captures 85.2% of the information in the data, while model #7 captures only 32.9%. State-based model #8, with the same Δdf=4 as model #7, captures 98.2%, almost three times the information of model #7. Similarly, for Δdf=2, state-based model #5, $A_0B_2{:}A_0$ has about four times the information of variable-based models #3 and #4. The reconstructed functions for models #6, #7, and #8 are shown in **Table 2** where closeness of these functions to f is indicated by rms-dev and R^2. In **Table 3**, closeness of the reconstructed functions to f is given by entropy-based %I.

Since k-systems analysis is non-statistical – there is no sample size that can be meaningfully used – the choice of a particular model and its reconstructed function must be made on the basis of user-provided criteria, based only on %I and Δdf. In the present case, models #6 and #8 are plausible candidates, but #7 is not. State-based analysis of continuous output functions is more powerful than variable-based analysis.

u-systems analysis

The u-systems approach, like k-systems analysis, applies to functions defined uniquely by their inputs. The essential idea underlying this approach is borrowed from expected value calculations in decision theory [**13**,**18**]: any utility value can be regarded as the expected value of an appropriate lottery. Similarly, any function value can be generated as an expected value of an appropriate probability distribution. Thus, for example, if we wish our expected value to equal the actual function value, *i.e.,*

$$\langle Z \rangle_{ij} = p(Z_0|A_iB_j)\, Z_0 + p(Z_1|A_iB_j)\, Z_1 = f(A_i,B_j),$$

this can be accomplished by setting the representative values for the bins to the minimum and maximum values of the function, *i.e.,* $Z_0=f_{min}$ and $Z_1=f_{max}$, and by defining

$$p(Z_1|A_iB_j) = (f(A_i,B_j) - f_{min})/(f_{max} - f_{min}); \; p(Z_0|A_iB_j) = 1 - p(Z_1|A_iB_j).$$

Table 4. u-systems form of Hallgrímsdottir and Yuster (HY) data

A	B	f	$p(Z_0\|AB)$	$p(Z_1\|AB)$
0	0	145.6	0.072	0.927
0	1	157.1	0.000	1.000
0	2	0	1.000	0.000
1	0	123.1	0.216	0.783
1	1	29.6	0.810	0.189
1	2	33.2	0.792	0.207
2	0	18.6	0.882	0.117
2	1	16.5	0.891	0.108
2	2	83.9	0.072	0.927

Table 5. State- and variable-based u-systems analysis of HY data
Variable-based models are in italics. Models are directed systems.

	Model	H	%I	Δdf
10	*ABZ*	*3.679*	*100.0%*	*8*
9	$AB{:}Z{:}A_0B_1Z_0{:}A_0B_0Z_1{:}A_1B_0Z_1{:}A_2B_2Z_1{:}A_0B_2Z_1$	3.684	99.0%	5
8	$AB{:}Z{:}A_0B_1Z_0{:}A_0B_0Z_1{:}A_1B_0Z_1{:}A_2B_2Z_1$	3.708	93.9%	4
7	*AB:AZ:BZ*	*4.004*	*31.9%*	*4*
6	$AB{:}Z{:}A_0B_1Z_0{:}A_0B_0Z_1{:}A_1B_0Z_1$	3.767	81.4%	3
5	$AB{:}Z{:}A_0B_1Z_0{:}A_0B_0Z_1$	3.874	59.0%	2
4	*AB:AZ*	*4.078*	*16.3%*	*2*
3	*AB:BZ*	*4.090*	*13.8%*	*2*
2	$AB{:}Z{:}A_0B_1Z_0$	4.007	31.2%	1
1	*AB:Z*	*4.156*	*0.0%*	*0*

doi: 10.7243/2050-1412-1-4

For the (A_i, B_j) value for which the function is maximum, $p(Z_1) = 1$ and $p(Z_0) = 0$; for the (A_i, B_j) value for which the function is minimum, $p(Z_1) = 0$ and $p(Z_0) = 1$.

Applying this approach to the Hallgrímsdottir and Yuster [21] data of **Table 2** gives the $p(Z_0|A_iB_j)$ and $p(Z_1|A_iB_j)$ values shown in **Table 4**. To get ABZ data, these conditional probabilities, p(Z|AB), must be multiplied by some p(AB). Since the Hallgrímsdottir-Yuster data is an explicit function, p(AB) is taken to be a constant.

In applications where sample data is available, one has frequencies for the different input states. In k-systems analysis, these are not usable, but in u-systems analysis, one could multiply p(Z|AB) by the input state probability, p(AB). This does a weighting of function values; alternatively, weights based on variances of function values for different input states might be used. Here no weighting has been done. The results of variable- and state-based u-systems analysis of the HY data are provided by **Table 5**; the reconstructed functions for state-based models #6, #8, and #9 are given in **Table 2**.

Once again, results show that state-based analysis is more powerful than variable-based analysis. This is indicated by the higher information content of state-based model #5 compared to the information contents of variable-based models #4 and #3, all of which have $\Delta df=2$; also by the higher information content of state-based model #8 compared to that of variable-based model #7, both of which have $\Delta df=4$.

The u-systems analysis of **Table 5** resembles but is still different from the k-systems analysis of **Table 3**. **Table 6** shows the sequence of states added in these two searches. The analytical relationship between these two representations and the arguments for using one versus the other are under investigation. Still, the results of these two different mathematical approaches are in rough agreement.

Comparing the rms-deviations and R^2s given in **Table 2** for the u-systems reconstructed functions (for models #6, #8, and #9) to the same measures given in **Table 2** for the k-systems reconstructed functions (for models #6 and #8) suggests that, on the HY data, k-systems reconstruction performs slightly better than u-systems reconstruction. However, a general conclusion about the relative merits of these two approaches to reconstructing continuous functions must await comparative studies on more datasets.

Comparing b-, k-, and u-systems analyses

Three different approaches are thus available for the analysis of functions such as might be obtained from gene expression data: b-systems analysis, which is a slight extension of standard probabilistic analysis, which requires that function values be binned; k-systems analysis, which rescales the function so it can be treated as a pseudo-probability; and u-systems analysis, which treats the function value as the expected value of a lottery. The salient properties of these approaches are summarized in **Table 7**.

Comments:

1. The input-output relation in u-systems analysis is deterministic in that only the average function value for any input state is used; however, it is stochastic in that the analysis itself is probabilistic (the output has two states).

2. The use of input state repeats is possible in u-systems analysis in that the probabilities of input states or the variances of their function values could be used to weight the conditional probabilities of the output states.

3. Analysis with b-systems is statistical in that likelihood ratio values, p-values, and information measures (*e.g.,* the Akaike and Bayesian Information Criteria) are relevant to the choice of model. These measures are not relevant and thus cannot be used for model selection for k- or u-systems, since a sample size is either unavailable or irrelevant.

4. The use of exact function values in k-systems and u-systems is an advantage since artifacts introduced by binning are avoided; b-systems analysis may give results that are sensitive to the exact procedures and parameters used for discretization.

5. The range limits of a u-systems reconstructed function can be extended by choosing bin values beyond the minimum and maximum function values in the data. By contrast, the range of a b-systems reconstructed function is limited by those bins populated by the data, while the range of a k-systems reconstructed function is not inherently limited.

Table 6. Sequence of added states in k- and u-systems analysis of HY data.

	k	u
5	A_2	$A_0B_2Z_1$
4	A_2B_2	$A_2B_2Z_1$
3	A_1B_0	$A_1B_0Z_1$
2	A_0	$A_0B_0Z_1$
1	A_0B_2	$A_0B_1Z_0$

Table 7. Comparing the b-, k-, and u-systems approaches
Preferred values of attributes are in bold.

attribute	b-systems	k-systems	u-systems
system type	directed	neutral	directed
presence of loops	more common	**less common**	more common
input-output relation	**stochastic**	deterministic	deterministic[1]
use of input state repeats	**yes**	no	**possible[2]**
statistical tests[3]	**yes**	no	no
use of exact function values	no (binned)[4]	**yes (of average)**	**yes (of average)**
reconstruction range limited	yes (by bins)	**no**	yes but arbitrary[5]

doi: 10.7243/2050-1412-1-4

Results: A real data example

In a preliminary study [22], RA was used to analyze data on states for 75 SNPs in eight genes and expression values for these genes. The expression values were binned, but the study aimed only at detecting SNPs that predicted expression, so continuous expression functions of the SNP states were not also generated. This study used only variable-based models without loops, and its primary results were models with only one predicting SNP. This simple analysis detected associations between SNP alleles and binned gene expression values, replicating and extending SNP–gene expression associations in a public data set [16]. Only the data from SNPs and expression values from the published positive *cis*-acting (local) SNP–gene expression associations were included. RA identified all but two *cis*-acting SNPs (false negative rate = 2.7%) and detected additional *cis*-acting SNPs.

To further explore the use of RA for quantitative functions, continuous gene expression values were generated from this data in a two-step process: the first step did a coarse search among models using b-systems analysis, the second step did a finer analysis of the results of the first step using the k- and u-systems approaches.

Step 1: b-systems analysis. A b-systems search through the space of models, including models with loops, was done with OCCAM. The expression values were discretized into four bins. Before doing this search, cases with missing data for salient variables were eliminated from the data, since OCCAM currently treats missing data as an additional state of the affected variable, and this extra state (a) can introduce artifacts into the analysis and (b) adds complexity to the model which raises the threshold for statistical significance. The original data had N=193; after case elimination, N=83, yet despite this drop in sample size, more models were significant because models were simpler. The search results are shown in **Table 8**. The search was done for three levels of complexity; at each level, the best 25 models are selected from the larger set of models which were generated by adding one relation to the best 25 models of the preceding lower level.

A promising model from the results of this search was selected for more detailed k- and u-systems analyses. The selection was based on the OCCAM results summary, which tabulates "best models" using different model selection criteria. The most conservative of these criteria is (i), the Bayesian Information Criterion (BIC), which favors simple (low Δdf) models. Just as in the Wilmot [22] study, where BIC favored loopless models having only one SNP predictor, the search results identified a set of five equivalent models, of form IV:XZ, where Z is the binned expression value for gene KIF1B, IV is a component that contains all the independent variables (the 75 SNPs), and X is any one of the {N, O, P, Q, R} SNPs, all of which are in KIF1B and in strong linkage disequilibrium (LD) with one another. IV:OZ was arbitrarily chosen to represent this set of models.

A less conservative criterion (ii) was also used that selects the model with the highest %I that is statistically significant ($p<.05$) relative to a reference of independence, and for

Table 8. b-systems variable-based model search for gene KIF1B

Models are directed systems. "(n models)" means n equivalent models, not listed. LR and p are likelihood ratio and significance relative to the IV:Z reference; p_{incr} is incremental significance relative to the immediate model 'ancestor'. Good models have low H, Δdf, p, p_{incr}, and high ΔLR, %I, ΔAIC, ΔBIC.

Model	Level	H	Δdf	ΔLR	p	%I	ΔAIC	ΔBIC	p_{incr}
IV:XZ:YZ:MZ (25 models)	3	7.46	12	60.3	0.000	32.7%	36.3	7.1	0.077
IV:XZ:YZ (25 models)	2	7.53	8	51.9	0.000	28.1%	35.9	16.4	0.000
IV:XZ (5 models)	1	7.66	4	36.9	0.000	20.0%	28.9	19.2	0.000
IV:WZ (3 models)	1	7.67	4	35.2	0.000	19.1%	27.2	17.5	0.000
13 other models									
IV:YZ (4 models)	1	7.86	4	14.0	0.007	7.6%	6.0	-3.7	0.007
IV:Z	0	7.98	0	0.0	1.000	0.0%	0.0	0.0	0.000
Model selection criteria									
i. Best Model(s) by ΔBIC:									
IV:XZ	1	7.66	4	36.9	0.000	0.1999	28.9	19.2	0.000
ii. Best Model(s) by %I with $p < 0.05$ and also all $p_{incr} < 0.05$:									
IV:XZ:YZ	2	7.53	8	51.9	0.000	0.2811	35.9	16.4	0.000
iii. Best Model(s) by ΔAIC:									
IV:XZ:YZ:MZ	3	7.46	12	60.3	0.000	32.7%	36.3	7.1	0.077
iv. Best Model(s) by %I with $p < 0.05$:									
IV:XZ:YZ:MZ	3	7.46	12	60.3	0.000	32.7%	36.3	7.1	0.077

doi: 10.7243/2050-1412-1-4

Table 9. SNPs in models selected

The SNP that represents the LD group is in bold

	X SNPs		Y SNPs
N	rs946501	**G**	**rs6433464**
O	**rs9332414**	H	rs11674895
P	rs17034643	I	rs10195413
Q	rs121242	J	rs4972643
R	rs12120191	K	rs10930654
		L	rs4144329
M	rs1995969		

Table 11. State-based b-systems analysis of KIF1B

Variable-based models are in italics. Models are directed systems.

	Model	H	%I	Δdf	p	p_{incr}
7	*OGZ*	*4.128*	*100.0%*	*16*	*0.000*	*0.879*
6	$OG:Z:O_0Z_0:O_2Z_0:O_1Z_2:O_0G_0Z_1:O_2G_0Z_1$	4.151	93.4%	5	0.000	0.164
5	$OG:Z:O_0Z_0:O_2Z_0:O_1Z_2:\mathbf{O_0G_0Z_1}$	4.158	91.3%	4	0.000	**0.046**
4	$OG:Z:O_0Z_0:O_2Z_0:O_1Z_2$	4.174	86.8%	3	0.000	0.026
3	$OG:Z:O_0Z_0:O_2Z_0$	4.193	81.3%	2	0.000	0.000
2	$OG:Z:O_0Z_0$	4.273	58.3%	1	0.000	0.000
1	*OG:Z*	*4.475*	*0.0%*	*0*	*1.000*	*1.000*

Table 10. b- and k-systems reconstructions for gene KIF1B

f_{b1}-f_{b3} are reconstructed functions for variable-based b-systems analyses, whose details are not included. f_{b4} and f_{b5} are reconstructed functions from b-systems models #4 and #5 in **Table 11**. f_{k3} and f_{k4} are for k-systems models #3 and #4 in **Table 12**

			b-systems					k-systems	
			variable-based			state-based		state-based	
O	G	f	f_{b1}	f_{b2}	f_{b3}	f_{b4}	f_{b5}	f_{k3}	f_{k4}
0	0	2.34	2.40	2.36	2.37	2.39	2.36	2.39	2.39
0	1	2.41	2.40	2.41	2.40	2.39	2.40	2.39	2.39
0	2	2.41	2.40	2.40	2.40	2.39	2.40	2.39	2.39
1	0	2.21	2.22	2.20	2.19	2.22	2.22	2.25	2.25
1	1	2.28	2.22	2.24	2.25	2.22	2.22	2.25	2.25
1	2	2.25	2.22	2.19	2.19	2.22	2.22	2.25	2.25
2	0	1.84	2.04	2.03	2.04	2.04	2.04	1.84	1.87
2	1	1.90	2.04	2.04	2.04	2.04	2.04	1.84	1.87
2	2	1.76	2.04	2.03	2.00	2.04	2.04	1.84	1.76
		rms-dev	0.124	0.121	0.117	0.125	0.124	0.041	0.029
		R^2	0.904	0.913	0.917	0.905	0.915	0.970	0.985
		Δdf	4	8	16	3	4	2	3
		reference for Δdf	OG:Z					uniform	
		models	b_1=OG:OZ	b_2=OG:OZ:GZ	b_3=OGZ			k_3=O	
			b_4 = $OG:Z:O_0Z_0:O_2Z_0:O_1Z_2$					k_4=$O:O_2:G_2$	
			b_5 = $OG:Z:O_0Z_0:O_2Z_0:O_1Z_2:O_0G_0Z_1$						

which each incremental df increase from independence is also statistically significant (p_{incr}<.05). This "path-significance" criterion selected best models of the form, IV:XZ:YZ, which has a loop, where X is as above and Y is any one of the {G, H, I, J, K, L} SNPs, all in gene OLA1 and also in strong LD with one another. IV:OZ:GZ was arbitrarily chosen to represent this set of b-systems models. The two less conservative criteria, (iii) AIC, and (iv) cumulative but not also incremental significance, being more vulnerable to false positives, were not considered. The b-systems search thus yields two models IV:OZ and IV:OZ:GZ (and their equivalents). The SNPs represented by X and Y (and M in IV:XZ:YZ:MZ) are listed in **Table 9**. That the X variables are in strong LD with one another, as are the Y variables, is shown by an OCCAM neutral system run that looks only for associations between the SNPs (it drops the DV); this calculation yields the X and Y clusters. Also, in a directed system calculation where the DV is O or G, the other SNPs in each group maximally predict (%I=100%) their representative SNP.

Table 10 shows f(OG), the average expression for the OG states in the data, and f_{b1}(OG), f_{b2}(OG), and f_{b3}(OG), the expected-value reconstructions of three variable-based b-systems models (detailed results not shown), with their R^2 and rms-deviations relative to f(OG). This calculation only deleted cases where data was missing for O or G, so here N=187. f_{b1} and f_{b2} are the reconstructed functions for IV:OZ and IV:OZ:GZ, the best models in the search shown in **Table 8** according to BIC and path-significance criteria, respectively. f_{b3} is the reconstructed function for a triadic variable-based interaction between O, G, and Z. In the terminology of the earlier RA study of epistasis [2], IV:OGZ is Type 1 epistasis, the strongest type; IV:OZ:GZ is Type 2 epistasis, a weaker type. IV:OZ, having only one SNP predictor, is not epistatic at all. All three of these b-systems models agree well with the data, and the only slight improvement in R^2 (from 0.904 to 0.913 or 0.917) by adding G as a predictor (either via Type 2 or Type 1 epistasis) suggests that this addition may not be warranted, even though this addition was incrementally significant in the initial b-systems search (model IV:XZ:YZ in **Table 8**).

Table 10 also indicates the results of expected value reconstructions of two state-based b-systems models. The state-based analysis is shown in **Table 11**. State-based b-systems models are slightly better (higher R^2 and/or lower Δdf) than variable-based b-systems models. Note that adding the fourth state (going from model #4 to #5), in bold, which indicates an epistatic interaction, is incrementally (and cumulatively) significant.

Step 2.1: k-systems analysis. **Table 10** above includes models #3 and #4 from a k-systems analysis whose results are given below in **Table 12**. The first state selected (model #2) was O_2, whose average expression value is notably different from the other two O states. This state alone embodies 91.4% of the data. The second state selected (model #4) was O_0,

doi: 10.7243/2050-1412-1-4

Table 12. State-based k-systems analysis for KIF1B
Variable-based models are in italics. Models are neutral systems. Model #6 is the data. The reconstructed function for models #3 and #4 are shown above in **Table 10**

	Model	H	%I	Δdf
6	*OG*	*3.16090*	*100.0%*	*8*
5	O_2:O_0:O_2G_2:G_0	3.16091	99.9%	4
4	O_2:O_0:O_2G_2	3.16103	98.6%	3
3	O_2:O_0 = *O*	*3.16119*	*96.8%*	*2*
2	O_2	3.16168	91.4%	1
1	uniform	3.16993	0.0%	0

Table 13. State-based u-systems analysis for KIF1B
Variable-based models are in italics. Models are directed systems. Model #7 is the data. The reconstructed functions for models #2 and #3 are the same as f_{k2}. and f_{k3} in **Table 10**.

	Model	H	%I	Δdf
7	*OGZ*	*3.636*	*100.0%*	*8*
6	OG:Z:O_2Z_1:O_1Z_1:$O_2G_2Z_1$:$O_0G_0Z_1$:G_1Z_1	3.640	99.1%	5
5	OG:Z:O_2Z_1:O_1Z_1:$O_2G_2Z_1$:$O_0G_0Z_1$	3.645	98.2%	4
4	OG:Z:O_2Z_1:O_1Z_1:$O_2G_2Z_1$	3.660	95.2%	3
3	OG:Z:O_2Z_1:O_1Z_1 = *OG:OZ*	*3.683*	*90.6%*	*2*
2	OG:Z:O_2Z_1	3.735	80.1%	1
1	*OG:Z:*	*4.137*	*0.0%*	*0*

Table 14. Sequence of added states in b-, k-, and u-systems analyses of KIF1B data

	b	k	u
5	$O_2G_0Z_1$	O_1G_2	G_1Z_1
4	$O_0G_0Z_1$	G_0	$O_0G_0Z_1$
3	O_1Z_2	O_2G_2	$O_2G_2Z_1$
2	O_2Z_0	O_0	O_1Z_1
1	O_0Z_0	O_2	O_2Z_1

which increases %I by about 5%. The two state model O_2:O_0 is equivalent to the variable based model, O, since df(O)=2. The third state selected (model #4) was O_2G_2, indicating an epistatic interaction affecting KIF1B transcription involving a SNP (O) in that gene and a SNP (G) in a different gene, OLA1, but this increases %I only by about 2%, which undermines confidence in the reality of this epistatic effect. Since k-systems analysis is non-statistical, these findings can only be validated by tests in new data.

Step 2.2: **Table 13** shows the u-systems analysis of this data. State-based model #3, namely OG:Z:O_2Z_1:O_1Z_1, is equivalent to variable-based u-systems model OG:OZ, since O_0Z_1 is determined by the known states, O_2Z_1 and O_1Z_1 (because Z_1 is known from the Z component) and knowing all O_JZ_1 probabilities determines also the O_JZ_0 probabilities (because O is known from OG). Not surprisingly, therefore, the reconstructed function for this u-systems model is identical to that of the k-systems model #3, O. As in state-based k-systems analysis, the third added state, $O_2G_2Z_1$, posits an epistatic interaction, increasing %I by 4.6%, a slightly greater increase than observed in the k-systems analysis. This added state is in fact the same as what is added in the k-systems search, since $O_2G_2Z_1$ determines also $O_2G_2Z_0$ (because O_2G_2 is known from the OG component). Again, not surprisingly, the function reconstructed for this u-systems model is identical to that of k-systems model #4, O:O_2G_2. Beyond Δdf=3, however, u- and k-analyses diverge, but since R^2 for df=3 models is already 0.985, more complex models are not included in **Table 10**. The states added in b-, k-, and u-analyses are listed in **Table 14**.

In summary, b-systems analysis, both variable- and state-based, suggests an epistatic effect between O and G, supported by incremental significance of the models, but k-systems analysis – and perhaps also u-systems analysis – casts some doubt upon this. Since k- and u-systems analyses treat function values exactly, and in state-based models with great precision, their implications carry some weight. The point of these analyses, however, is not to ascertain whether there is an OG epistatic effect on KIF1B gene expression or not. The point is merely to illustrate the methodology of this continuous RA approach as applied to gene expression data; specifically to show that k- and u-systems approaches can reconstruct functions more accurately than the b-systems approach. A secondary purpose here is to demonstrate that state-based RA is more powerful than variable-based RA, which is strikingly shown by the HY analyses but also evidenced in the KIF1B analyses.

Discussion & Conclusions

This paper demonstrates the use of RA to analyze continuous functions, as opposed to frequency or probability distributions, and advances the methods for doing so. In tests on a simple function used in an earlier study on classifying epistasis [21] and on real data [22], k-systems and u-systems analyses yield similar but slightly different results, and have comparable modeling efficacy. The relative merits of these two non-standard RA approaches to decomposing functions, and their suitability for different data contexts, need to be explored further. In future software development, parallelization of the software will make possible the application of RA to GWAS. After this is done, the number of SNPs that RA can analyze will depend on the type of model. For loopless models, which have closed form solutions, 100,000s of SNPs could be considered; for variable-based models with loops, which require iterative methods, perhaps 1000s (certainly 100s); for state-based models, perhaps 100s (certainly 10s).

Since RA can model epistasis in frequency (and probability) distributions [1,2], and also detect LD, and since k- and u-systems variants of RA can be usefully applied also to functions obtained in studies of gene expression, RA

doi: 10.7243/2050-1412-1-4

constitutes a flexible and powerful modeling methodology for the study of epistasis. A variety of other approaches have been applied to modeling epistasis in gene expression [23-28]. Comparing RA to other approaches is beyond the scope of this paper, and will be the subject of future studies, but advantages of RA over some other methods can be stated briefly: (i) RA can be used for *both* nominal data and continuous function applications; this allows one to work within a single mathematical/computational framework, while some other methods are specific to only one these two applications; (ii) RA has three levels of refinement – coarse (variable-based models without loops), fine (variable-based models with loops), and ultra-fine (state-based models); this allows one to move smoothly within a single framework from broad search, *e.g.,* GWAS, involving very many variables to ultra-fine analysis that focuses on only a few, while some other methods (*e.g.,* [21]) are specific to one of these situations; (iii) RA explicitly considers the space of possible models in its use of hypergraphs, and is thus especially designed for exploratory searches, while some other methods are primarily confirmatory, and require the user to specify the models to be considered; (iv) Even where RA overlaps with – and to the extent of the overlap is obviously not superior to – other methods, it has distinctive features, so it complements these other methods (see, *e.g.,* the discussion of RA vs. Bayesian networks and other graphical models in [2]).

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
MZ provided the main contribution to this work: drawing on pioneering work of Bush Jones and in collaboration with Michael Johnson (his graduate student), he developed the theory for applying RA to continuous functions, did the reported calculations on the data, and wrote the paper. JF programmed RA in the software package (OCCAM) used as the primary methodological instrument in this research. BW provided the real SNP data, did the initial study that applied RA to gene expression data, and helped in the interpretation of results.

Authors' information
MZ is one of the principal Reconstructability Analysis theorists today; his many years of work on RA is documented in his web page, http://www.pdx.edu/sysc/research-discrete-multivariate-modeling . He has also directed the development of the RA software package, OCCAM, since its inception. JF, a graduate student in the Systems Science Graduate Program, has been the OCCAM programmer for a number of years. BW is a genomics researcher at the Oregon Health Sciences Researcher; she took MZ's RA course several years ago, and in preliminary studies showed that RA can detect epistasis in gene expression data.

Acknowledgements and funding
This project was supported in part by the Oregon Tax Check-Off Alzheimer's Research Fund administered by the Layton Aging & Alzheimer's Disease Center in collaboration with the Oregon Partnership for Alzheimer's Research.

Publication history
Received: 29-May-2012 Revised: 21-June-2012
Re-revised: 17-Aug-2012 Accepted: 28-Aug-2012
Published: 03-Oct-2012

References
1. Shervais, S. *et al.*: **Reconstructability analysis as a tool for identifying gene-gene interactions in studies of human diseases**. *Stat Appl Genet Mol Biol* 2010, **9**(1):Article18. | Article | PubMed
2. Zwick, M.: **Reconstructability analysis of epistasis**. *Ann Hum Genet* 2011, **75**(1):157-171. | Article | PubMed
3. Klir, G.: **Identification of Generative Structures in Empirical Data**. *Internat. J. Gen. Sys* 1976, **3:**89-104. | Article
4. Klir, G.: (1985). *The Architecture of Systems Problem Solving*. New York: Plenum Press.
5. Krippendorff, K.: **An Algorithm for Identifying Structural Models of Multivariate Data**. *Internat J Gen Sys* 1981, **7**(1):63-79. | Article
6. Krippendorff, K. (1986). *Information Theory.* Beverly Hills: Sage.
7. Klir, G.: **Reconstructability Analysis: An Offspring of Ashby's Constraint Theory**. *Systems Research* 1986, **3**(4)**:** 267-271. | Article
8. Zwick, M.: **Wholes and Parts in General Systems Methodology**. *The Character Concept in Evolutionary Biology* 2001, 237-256. New York: Academic Press. | Article
9. Zwick, M.: **An Overview of Reconstructability Analysis**. *Kybernetes* 2004, **33**(5/6):877-905. | Article
10. Zwick, M.: **Reconstructability Analysis With Fourier Transforms**. *Kybernetes* 2004, **33**(5/6):1026-1040. | Article
11. Willett, K. *et al.*: **A Software Architecture for Reconstructability Analysis**. *Kybernetes* 2044, **33**(5/6):997-1008. | Article
12. Fusion, J., Willett, K. and Zwick, M. (2011). **OCCAM: A Reconstructability Analysis Program**. | Article | PubMed Abstract | PubMed Full Text
13. Johnson, M.: **State-Based Systems Modeling: Theory, Implementation, and Applications**. 2005, 219:3183756. PhD Dissertation, Portland State University.
14. Zwick, M.: **Information-Theoretic Mask Analysis of Rainfall Time-Series Data**. *Advances in Systems Science and Applications*, 2005, **1**:154-159. | Article
15. Wilmot, B., Zwick, M., and McWeeney, S. (2008). **Reconstructability Analysis Detects Genetic Variation Associated with Gene Expression**. 12th QTL-MAS Workshop in Computational Genetics, Uppsala, Sweden, May 15-16. | Article | PubMed Abstract | PubMed Full Text
16. Myers, A. J. *et al.*: **A survey of genetic human cortical gene expression**. *Nat Genet* 2007, **39**(12):1494-1499. | Article | PubMed
17. Jones, B.: **Reconstructability Analysis for General Functions**. *Internat. J Gen Sys* 1985, **11**:133-142. | Article
18. Johnson, M. *et al.*: **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. | Article
19. Jones, B.: **K-systems Versus Classical Multivariate Systems**. *Internat. J Gen Sys* 1985, **12**(1):1-6. | Article
20. Gouw, D. *et al.*: **Using K-systems Theory to Compute True Main and Interaction Effects in a Design of Experiments**. *Cybernetics and Systems: An International Journal*, 1995, **26**(4):481-498. | Article
21. Hallgrimsdottir, I. B. & Yuster, D. S.: **A complete classification of epistatic two-locus models**. *BMC Genet* 2008, **9**:17. | Article | PubMed Abstract | PubMed Full Text
22. Wilmot, B. (2008). **Association of SNPs and Gene Expression Using Reconstructability Analysis: A Proof of Principle**. Oregon Clinical + Translational Research Institute unpublished report, April 8, 2008. | Article | PubMed Abstract | PubMed Full Text
23. Culverhouse, R., Klein, T. & Shannon, W.: **Detecting epistatic interactions contributing to quantitative traits**. *Genet Epidemiol* 2004, **27**(2):141-152. *Genet. Epidemiol*. **27**(2):141-52. | Article | PubMed
24. Wang, T. & Zeng, Z. B.: **Models and partition of variance for quantitative trait loci with epistasis and linkage disequilibrium**. *BMC Genet* 2006, **7**:9. | Article | PubMed Abstract | PubMed Full Text
25. Jannink, J. L.: **Identifying quantitative trait locus by genetic background interactions in association studies**. *Genetics* 2007, **176**(1):553-561. |

Article | PubMed Abstract | PubMed Full Text

26. Wang, T. & Zeng, Z. B.: **Contribution of genetic effects to genetic variance components with epistasis and linkage disequilibrium**. *BMC Genet* 2009, **10:**52. | Article | PubMed Abstract | PubMed Full Text
27. Michaelson, J. J., Alberts, R., Schughart, K. & Beyer, A.: **Data-driven assessment of eQTL mapping methods**. *BMC Genomics* 2010, **11:**502. | Article | PubMed Abstract | PubMed Full Text
28. Zhang, Y., Jiang, B., Zhu, J. & Liu, J. S.: **Bayesian models for detecting epistatic interactions from genetic data**. *Ann Hum Genet* 2011, **75**(1):183-193. | Article | PubMed

Citation:
Zwick M, Fusion J and Wilmot B: **Reconstructability of Epistatic Functions**. *journal of Molecular Engineering and Systems Biology* 2012, **1**:4.
http://dx.doi.org/10.7243/2050-1412-1-4