

Reconstructability Analysis of Genetic Loci Associated with Alzheimer disease

Patricia Kramer PhD
Shawn K. Westaway PhD
Department of Neurology
Oregon Health and Sciences University,
Portland, OR, USA
kramer@ohsu.edu
westaway@ohsu.edu

Martin Zwick PhD
Systems Science Department,
Portland State University,
Portland, OR, USA
zwick@pdx.edu

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

Abstract—*Reconstructability Analysis (RA) is an information- and graph-theory-based method which has been successfully used in previous genomic studies. Here we apply it to genetic (14 SNPs) and non-genetic (Education, Age, Gender) data on Alzheimer disease in a well-characterized Case/Control sample of 424 individuals. We confirm the importance of APOE as a predictor of the disease, and identify one non-genetic factor, Education, and two SNPs, one in BINI and the other in SORCS1, as likely disease predictors. SORCS1 appears to be a common risk factor for people with or without APOE. We also identify a possible interaction effect between Education and BINI. Methodologically, we introduce and use to advantage some more powerful features of RA not used in prior genomic studies.*

Keywords: *Reconstructability Analysis, Alzheimer Disease, genetics, bioinformatics, OCCAM*

I. INTRODUCTION

The genetic component to complex human diseases can include direct effects of single genes or multiple genes acting independently, the epistatic interaction of multiple genes, and the interaction of genes with the environment. Detecting these interactions with standard statistical tools is difficult, because effects may be small or very complex or because there may be interaction effects where there are minimal or no main effects.

Significant advances have been made over the last two decades in developing analytic methods and bioinformatics tools for detecting single genes that are necessary for, or contribute to, human diseases. For the most part, however, diseases with a “simple” genetic etiology are relatively rare. Common diseases (e.g., hypertension, cancer, dementia) are the result of DNA sequence variations in multiple genes, at least some of which may interact in a non-additive, or epistatic, fashion, and thus have a substantially more complex genetic etiology. Early genome-wide association results have identified

associations with only modest main effects. To account for highly familial traits, there are many more loci with very modest main effects, more rare variants with larger effects, or gene-gene and/or gene-environment interactions that are a more prominent element of the genetic component of these diseases. For example, despite recent advances in the use of genome-wide association studies (GWAS) to identify genetic variants, or SNPs, associated with Alzheimer disease (AD), variants identified to date have modest main effects and account for only a fraction of the genetic risk.

Reconstructability Analysis (RA) is an information- and graph-theory-based method that is superior to current methods in several respects. Prior work has shown that RA is capable of detecting low levels of genetic interactions, despite high noise levels, in simulated data, reliably detecting interactions in heritabilities as low as 0.008, with as many as 50 noise genes [1]. RA outperformed earlier work which used neural nets [2] and multifactor dimension reduction [3]. Further, in single SNP tests on real data on diabetes, RA closely approximated results obtained by traditional linkage analysis in predicting both case/control status [4] and non-parametric linkage (NPL) categories [5]. In cross-chromosome tests [6], RA confirmed the association between the chr2 *NIDDM1* region and the chr15 *CYP19* region, and detected a multi-SNP association between *NIDDM1* and *CAPN3* on chr15 [7], supporting the suggestion that *CAPN3* contributes to susceptibility to diabetes [8]. Based on our past success with the diabetes dataset, in this paper we use RA to investigate genetic and non-genetic factors and gene-gene epistasis in AD. Our goal is to extend our understanding of Alzheimer disease and to confirm earlier studies that demonstrated the usefulness of RA.

II. RECONSTRUCTABILITY ANALYSIS

Reconstructability analysis (RA) is an information- and graph-theoretic methodology originated by Ashby [9] and further developed by others [10]–[17]. An account of its origin [18] and compact summaries [19], [20] are also available. In RA, a probability or frequency distribution or a set-theoretic relation is decomposed into component distributions or relations [21]. When applied to the decomposition of frequency distributions, RA does statistical multivariate analysis and resembles log-linear (LL) methods [22], used widely in the social sciences, and closely related logistic regression (LR) techniques. RA also overlaps with Bayesian networks (BN). Where these methodologies overlap, they are mathematically equivalent. However, RA, LL, and BN each have unique capacities not commonly available in the other two [23]. For example, RA, but not LL or BN, can be applied to set-theoretic relations and to arbitrary functions of nominal variables; RA also has a state-based version [24], [25], a finer-grained modeling approach than the standard variable-based approach, and also a Fourier version [26]. RA and LL, but not BN, can utilize models with loops and can address not only problems where IVs (independent variables, inputs) and DVs (dependent variables, outputs) are distinguished (called directed systems), but also problems where this distinction is not made (neutral systems). LR, as implemented in PLINK [27], which is widely used for genomic analysis, is less general than RA. Both RA and BN explicitly conceptualize the lattice of graphical models and have been computationally adapted for exploratory modeling, which is not as easily done in LL and LR. However, all these methodologies are inherently designed for nominal variables, and are thus natural for analyzing genomic data. (They can also be applied to quantitative or ordinal variables by binning.) By contrast, certain other machine learning methods such as neural nets [2], [28] or support vector machines [29], presuppose metric information and are thus less inherently suited for genomic analyses.

The following discussion summarizes the basic ideas of RA. Consider a directed system with IVs (genes or SNPs or covariates) A, B, C, and D, and DV the disease status Z. Consider an observed frequency distribution $f(A, B, C, D, Z)$ which we write as ABCDZ. RA decomposes such a distribution into projections such as ABCD and ABZ, which when taken together define an RA model $m = ABCD:ABZ$ that is less complex (fewer degrees of freedom) than the data. This model defines a calculated frequency (or probability) distribution $ABCDZ_m$, obtained by maximum entropy composition of ABCD and ABZ, which is compared with the observed ABCDZ. While the data itself, ABCDZ, also called the “saturated model,” allows all four IVs to jointly predict Z, the ABCD:ABZ structure allows only A and B to jointly predict Z, with C and D having no predictive relationship with Z. A and B predict Z via the conditional probabilities $p_m(Z|AB)$ derived from the $ABCDZ_m$ distribution. (In all models, the order of the components and the order of the variables within each component is arbitrary, so, for example, ABCD:ABZ = BZA:CDBA.) The ABCD component of ABCD:ABZ assures that models that will be compared to one another all involve the same set of variables and will be hierarchically related; it also allows – but does not identify – associations between the IVs

themselves. For visual simplicity, this component that includes all the IVs is omitted in models described below.

If ABCD:ABZ is a good model, one can equivalently say that the “transmission” or “mutual information” $T_m(AB:Z) = H(Z) - H_m(Z|AB)$ is high, while the transmission $T_m(CD:Z)$ is low. H is uncertainty (Shannon entropy), so transmission here is reduction of uncertainty about Z. Dividing $T_m(AB:Z)$ by $H(Z)$ (and multiplying by 100) gives $\%H_m(Z|AB)$, the %uncertainty reduction of Z, knowing A and B. Uncertainty reduction for nominal variables is analogous to %variance explained for continuous variables, but one difference between the two is that because of the logarithm term in the expression for Shannon entropy even a small uncertainty reduction can correspond to a large effect size. For the reference “independence” model ABCD:Z, in which no IV predicts Z, $\%H_m(Z) = 0$. Uncertainty reduction can be assessed for statistical significance with the Chi-square distribution.

Because uncertainty reduction is an information theoretic measure of predictive efficacy that is not calculated by most other methods, it is useful to supplement it with %correct, a generic measure of predictive accuracy that is commonly produced by most modeling methods. %correct roughly follows uncertainty reduction, but the two are not precisely co-linear.

For the purposes of this study, there are three different classes of RA models: variable-based (VB) models without loops, variable-based models with loops, and state-based (SB) models. These allow coarse, refined, and ultra-refined modeling, respectively, graphically depicted in Figure 1 [23]. Conversely, these three classes are applicable to many variables, a modest number of variables, and few variables, respectively. The bold lines in the figure indicate how complex a model might be acceptable using each class of model

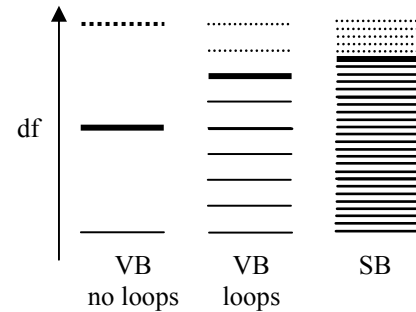


Figure 1. Degrees of refinement of RA models.

Models such as ABCDZ, ABCD:ABZ, and AB:Z each have a “single predicting component,” i.e., a component that includes Z and only one subset of the IVs and whose probability distribution is used to predict Z. Such models are loopless, and do “feature selection” or “dimensionality reduction.” By contrast, consider model ABCD:ABZ:CDZ; its the second and third components allow Z to be predicted by A and B jointly and also, separately, by C and D jointly. The two predicting components are integrated by a maximum entropy algorithm, and from the integrated (calculated) distribution, one obtains the conditional distribution for the model, $p_m(Z|ABCD)$. This conditional distribution is different from

the conditional distribution obtainable directly from the data, and it is the calculated distribution that is used for prediction.

Since a model includes a non-predicting component containing all the IVs (e.g., ABCD in model ABCD:ABZ), models with more than one predicting component necessarily have loops. (The simplest example of a two predicting component model that has a loop is BA:AZ:ZB.) For loopless models with a small number of predicting IVs, one can easily examine all possible structures. However, with more than a few variables, exhaustive evaluation of all models becomes prohibitive if models with loops are considered, since these models do not have algebraic solutions but require iterative computation [26].

The models discussed above are “variable based”, i.e., defined in terms of subsets of variables. RA also includes “state-based” models [24], such as ABCD:A₁B₂Z:B₁C₃Z, which are instead defined at least partially in terms of information-rich specific states of variables, the states being indicated by subscripts. In this study, a SNP state codes a diploid genotype, i.e., the genotype homozygous in one allele is coded as state 1, heterozygous genotypes as state 2, and the genotype homozygous in the other allele as state 3. Z encodes case (Z=1) vs. control (Z=0). The A₁B₂Z component in the above model means that $p_m(Z_1|A_1B_2)$ is significantly different than $p(Z_1)$ i.e., that genotype A₁B₂ is at higher or lower risk for disease than average. (This could be expressed in terms of $p(Z_0)$; since Z has two states, $p_m(Z_0|A_1B_2) = 1 - p_m(Z_1|A_1B_2)$.) State-based RA resembles multifactor dimensionality reduction (MDR), which has been used to study epistasis [30], and some implementations of logistic regression. Most state-based models have loops, but a definitive algorithm for loop detection in this class of models has not yet been established.

Variable-based models with or without loops and state-based models can be evaluated in terms of uncertainty reduction, which is tested for significance relative to independence with a Chi-square p-value. Consider choosing the model with the greatest uncertainty reduction that is significant in this way. This “Cumulative-p” criterion, however, always overfits, and needs to be augmented by the more stringent condition that every step from independence is statistically significant. This is one way to select a best model, namely the most uncertainty-reducing model which is cumulatively significant and whose path from the independence model is also significant at every step. We call this the “Incremental-p best model.” We also use two other criteria to define best models: BIC, the Bayesian Information Criterion [31], and AIC, the Akaike Information Criterion [32]. All three criteria penalize the model for complexity (Δdf relative to independence), i.e., trade off uncertainty reduction and model simplicity, in different ways. AIC and BIC integrate these two considerations linearly, quite different from the way they are integrated in a Chi-square p-value calculation. Of these three criteria, BIC is the most conservative, penalizing complexity the most severely, so the interactions in the BIC best model are the most reliable. Incremental-p and AIC are less conservative criteria that select more complex models; sometimes Incremental-p selects a more complex model than AIC; sometimes the reverse is true. BIC never overfits; AIC or Incremental-p sometimes overfit. In this study, what is actually

calculated is $\Delta AIC = AIC(\text{reference}) - AIC(\text{model})$, similarly for ΔBIC ; good models have high ΔAIC or ΔBIC .

Like other methods, RA allows one to control for particular IVs. For example, a high $T(A:Z)$ says that A predicts Z, while a high $T_C(A:Z) = T(AC:CZ)$ says that A predicts Z even when controlling for C. $T_C(A:Z) = H(Z|C) - H(Z|AC)$, so A predicts Z even when controlling for C when the uncertainty of Z knowing C is reduced by knowing also A. The significance of this reduction can be assessed by a Chi-square p-value. Controlling for some variables while calculating associations between others is easiest to grasp for loopless models, but it can also be applied to models with loops.

Calculations were done using the RA software program developed at Portland State University (Portland, Oregon) called OCCAM (named for the principle of parsimony and also “Organizational Complexity Computation and Modeling”). The earliest program was developed by Zwick and Hosseini [33]; reviews of RA methodology [19], [20], a list of recent RA papers, an OCCAM manual [34] and a description of OCCAM architecture [35] are available.

III. METHODOLOGY

A. The Data

Subjects were recruited from aging research cohorts collected over twenty years at the Layton Aging and Alzheimer’s Center at Oregon Health & Science University (OHSU) (Portland, Oregon). Stringent criteria were used to ascertain well-characterized cases and controls. All subjects were deceased and had been evaluated for cognitive decline and dementia within 12 months prior to death. In addition, all were at least 65 years of age at the time of death, had an autopsy, were of Caucasian ancestry and had DNA available for SNP genotyping. Controls were defined as clinically non-demented individuals with autopsy confirmation of no AD neuropathology, and cases were defined as clinically demented individuals with autopsy-confirmed high levels of AD neuropathology. A total of 437 individuals met these criteria. The study was approved by the IRB at OHSU.

Genome-wide SNP data for all subjects was obtained from the NIH-sponsored Alzheimer Disease Genetics Consortium (ADGC). Imputed genotypes, provided by the ADGC, were used to replace any missing data. For this study, we selected 15 SNPs, most of which represent genes that have been reported to be associated with AD in published genome-wide association studies (GWAS) (as summarized on the Alzforum website, www.alzgene.org). One SNP and 13 subjects were dropped due to excessive missing data, yielding an initial data set with a sample size of 424, including 221 cases and 203 controls. Missing data for which no imputed data was available were then handled in two different ways: (1) they were excluded from the data, i.e., these subjects were also dropped, slightly reducing the sample size further, or (2) they were treated as a fourth genotype. When (1) was used, this is noted in results below; otherwise (2) was done.

B. Analysis

Our strategy involved the following four steps.

Step 1: We looked at variable-based loopless models to see what these models suggest about the most predictive single IVs.

Step 2: We then searched among variable-based models with loops, and proposed three best models for the AD data using the three criteria of BIC, AIC, and Incremental-p. These models are the principal results of this study.

Step 3: In the models of Step 2, both direct and indirect effects of the IVs on Z (CaseControl) can contribute to the interactions that were found, so we did a series of calculations where we controlled for one or more of the IVs while looking for associations of the remaining IVs with Z.

TABLE I. VARIABLES

Variable Name	ID	Gene	Chromosome	Comment
APOE	Ap	APOE	19	1 = Allele 4 present; 0 = absent
Gender	Sx	n/a		1 = M, 0 = F
Education	Ed	n/a		Grade: 0 = <9, 1 = 9-12, 2 = >12 years
Age at last examination	Ag	n/a		0 = 60<75, 1 = 75<90, 2 = 90+ years
rs1801133	A	MTHFR	1	
rs3818361	B	CR1	1	Missing 3
rs7561528	C	BIN1	2	
rs744373	D	BIN1	2	
rs6943822	E	RELN	7	
rs4298437	F	RELN	7	
rs7012010	G	CLU	8	
rs11136000	H	CLU	8	
rs10786998	J	SORCS1	10	Missing 9
rs11193130	K	SORCS1	10	Missing 11
rs610932	L	MS4A6A	11	
rs3851179	M	PICALM	11	
rs3764650	N	ABCA7	19	Missing 2
rs3865444	P	CD33	19	Missing 9
CaseControl	Z	n/a		Case = 1, Control = 0

Step 4: Finally, narrowing our IV set to the four salient predictors from the previous three steps, we looked at state-based models to see if these models suggested interaction effects. Since state-based models are more refined (specific) and thus more powerful than variable-based models even with loops, it is possible that these models can pick out interactions that are too subtle to be detected with variable-based models.

All statistical tests in this paper used a 0.05 cut-off for significance. The p values reported in the tables below are “cumulative” p-values, i.e., tested for the model compared to the independence model. The tables also indicate whenever

models listed are not incrementally significant, i.e., significant for each step ascending from the reference of independence.

This study differs methodologically from our previous use of RA to analyze diabetes data [7] in two ways: (1) There we utilized only variable-based models without loops, while here we exploit the more powerful variable-based models with loops and state-based models. (2) Here we also use RA methods for controlling for some variables by looking at conditional associations and by partitioning the data into separate values of important variables, specifically Ap.

IV. RESULTS

Step 1. An examination of the simplest loopless models, namely those with a single predicting IV, yielded the following table (Table II) of the reduction of the uncertainty of Z, given the IV. The table reports all IVs whose p-value ≤ 0.05 . As expected, APOE (Ap) is the top uncertainty reducer. The next most predictive IV, Education, reduces uncertainty much less.

TABLE II. BEST SINGLE IV PREDICTORS OF CASECONTROL

IV	% $\Delta H(Z IV)$	p	%correct
Ap	9.1	0.000	67.3
Ed	3.5	0.000	56.7
C	2.6	0.001	57.9
K	2.5	0.001	56.4
J	1.5	0.015	54.7
Ag	1.2	0.036	55.7
L	1.1	0.047	54.5
none	--	--	51.8

Subjects with missing J and K values were excluded here; sample size = 413.

After that, SNPs C and K are less predictive than Ed. This is followed by Age and SNP L. The next best predictors, SNPs A and G (these are variable names, not alleles), with $p \leq 0.05$, are weaker still; these are not included in the table, but do show up in a model discussed below. In addition to reporting uncertainty reductions, we report the %correct that predictions achieve. The bottom entry of the table indicates that the independence model (which doesn't reduce uncertainty of Z at all) has %correct = 51.8; this is the result of always predicting the majority state of Z, which in our data is Control.

Of the seven IVs listed in the table, J and K, which are in the same gene, are extremely tightly associated, with % ΔH of 89.2% or 88.4%, depending on which of the two is used to predict the other, and with %correct in predictions (both ways) of 97.3%. Despite this close association K is a better predictor of Z than J. (That this shows up much more strongly in % ΔH than in %correct is due to the log term in the former measure.) Given this tight association, one of the two can be omitted from further consideration; we chose J to be dropped. Each of the SNP pairs, C and D, E and F, and G and H are similarly in a single gene and are tightly associated; again, one of each pair is more predictive of Z, namely C, E, and G.

Step 2. Searching among models with loops yielded the best variable-based models listed in Table III. (For visual simplicity, in this table and in the rest of this paper, the model component that includes all the IVs is omitted.) The BIC model is the most solid result of this study. If we had considered only models without loops, the BIC model would have been simply ApZ, i.e., we would have missed the importance of Ed and K. This illustrates the point made by Figure 1.

TABLE III. ^a
BEST VARIABLE-BASED MODELS USING THE THREE CRITERIA

Criterion	Model	% ΔH	% Correct	Δdf	^b p
BIC	ApZ:EdZ:KZ	15.6	70.5	5	0.00
AIC	ApAZ:EdZ:KZ:CZ	19.8	73.4	11	0.00
Incremental-p	ApZ:EdZ:KZ:CZ:LZ	18.3	71.2	9	0.00

a. Subjects missing K values were excluded; sample size = 413. b. The p-value is the cumulative p relative to independence. All three models are incrementally significant.

The BIC model includes the 1st, 2nd, and 4th best single IV predictors of Z shown in Table II. The two less conservative criteria, namely AIC and Incremental-p, add A, C, and L as possible predictors; of these, considering Table II and also results presented below, C is the most likely to be reliable.

Step 3. Since we are concerned about possible associations among the IVs, we did a series of calculations that illuminate the effects of such associations. Some of these calculations explicitly or implicitly controlled for some IVs while looking at associations of other IVs with Z. We considered four types of control calculations, summarized as follows.

3.1 Control for *all other* (17) IVs, while looking at the association of one IV with Z.

3.2 Control for the covariates, Ag, Ed, Sx, and for Ap, or subsets of these four IVs, using loopless models.

3.3 Control for the covariates, Ag, Ed, Sx, and for Ap, or subsets of these four IVs, using models with loops.

3.4 RA analysis for the two specific values of Ap.

3.1. Calculations that control for some IVs while looking at associations of other IVs with Z do so by comparing two different models. Because of our small sample size, such comparisons were not statistically significant when association of any individual IV with Z was controlled for *all other* 17 IVs.

3.2. Our small sample size also did not indicate any significant predictor of Z when controlling for the three covariates and Ap. This calculation compares model ApAgEdSxZ with model ApAgEdSxYZ, where Y is one other IV. None of these comparisons was significant. However, controlling for fewer IVs yielded results. Controlling for Ag, Ed, and Sx, we found that Ap, not surprisingly, is a significant predictor, and the only one. Controlling for Ap and Ed, we found C to be the only significant predictor. Controlling for Ap alone, Ed, C, K, and A were all significant individual predictors, in that order. These results are consistent with those of Tables II and III.

3.3. It is possible to control for IVs also using models with loops by selecting a reference model that has the IVs one wants

to control for, and testing whether adding a new IV is significant. To obtain this reference model, an analysis was first done on the four IVs, Ap, Ag, Ed, and Sx, that we want to control for. The results are shown in Table IV.

TABLE IV. SELECTING A MODEL WITH LOOPS FOR CONTROL CALCULATIONS

Criterion	Model	% ΔH	% Correct	Δdf	P
BIC	ApZ:EdZ	11.3	70.0	3	0.00
AIC	ApSxZ:EdZ:AgZ ^a	13.2	70.0	7	0.00
Incremental-p	ApZ:EdZ:AgZ	12.3	70.0	5	0.00

a. Not incrementally significant (sample size 424)

We selected the Incremental-p model, ApZ:EdZ:AgZ, which is intermediate in complexity, as the reference for control calculations, and searched for additional IV predictors that are significant relative to this reference. We found that adding one predictor, K or C or J, was incrementally significant; adding two sequentially, either K and C, or C and J, was also significant, not surprisingly since J and K are tightly associated. If we instead select the BIC model, ApZ:EdZ, as the reference model for these control calculations, the best predictors to add to this model are K or C or J, in that order. If we select an even simpler model, namely ApZ, as the reference, the best predictor to add are Ed or K or C or J, in that order. These results are consistent and support the proposition that of the IVs in the best models reported in Table III, the predictive effects of K and C are independent of and not due to associations with Ap and Ed.

With the exception of the AIC model in Table III, which is not incrementally significant, in all of the models we have considered so far, the effect of each IV is independent of the effects of the other IVs. That is, so far, we do not see any interaction (epistatic) effects; each component of all of these models involves only one IV and Z. In Step 3.4 and in Step 4, however, we do find such interaction effects.

3.4. Finally, we repeated the RA analysis setting Ap=0 or Ap=1, and the results are shown in Table V. We note that since Ed, K, or C showed up as predictors in this analysis for Ap=0, this cannot be due to association with Ap, since here Ap is fixed. Second, K occurs as a shared risk factor for both people who have APOE and those who do not, but these two groups seem also to have some different specific risk factors, namely Ed, C, and Ag for those who do not have APOE, and A and G (these are variable names, not alleles) for those who do. Third, the predictive power of Ap=1 models is stronger than the predictive power of Ap=0 models, as expected since the Ap=1 models are dominated by the risk allele. (The predictive power of Ap=1 models is also stronger than the models of Table III, all of which include Ap as a variable). The Ap=0 models are a little more heterogeneous with respect to what's causing AD, since these models doesn't include the APOE risk factor. Fourth, both the Ap=0 and Ap=1 results suggest interaction effects; EdC and AgK in the former, and AK in the latter; and these effects are in models that are incrementally significant.

TABLE V. BEST VARIABLE-BASED MODELS FOR SUBJECTS WITHOUT OR WITH APOE

Criterion	Model	% ΔH	% Correct	Δdf	P
<i>Ap = 0</i>					
BIC	EdZ:CZ	9.4	69.3	4	0.00
AIC	EdCZ:AgKZ	19.1	75.7	16	0.00
Incremental-p	EdCZ:AgKZ	19.1	75.7	16	0.00
<i>Ap = 1</i>					
BIC	AKZ	23.5	79.5	8	0.00
AIC	AKZ:AgZ ^a	32.8	80.8	14	0.00
Incremental-p	AKZ	23.5	79.5	8	0.00

a. Not incrementally significant

Step 4. Finally, we analyzed the data also with state-based (SB) RA models, restricting the analysis to only the four IVs, Ap, Ed, K, and C, where missing K values were excluded (sample size 413). In the results obtained, all three best models (BIC, AIC, and Incremental-p) were the same, namely the $\Delta df=5$ model (omitting for clarity the ApEdCK:Z part of the model),

$$Ap_0Z : Ap_0Ed_0Z : K_2Z : Ap_0Ed_2C_2Z : Ap_0Ed_1C_2K_1Z$$

This model has 5 specific states added in the order listed to the independence model ApEdCK:Z. This model is better than the $\Delta df=5$ BIC model of Table III, namely ApZ:EdZ:KZ. It has a higher uncertainty reduction ($\% \Delta H(Z|ApEdCK) = 19.4\%$ compared to $\% \Delta H(Z|ApEdK) = 15.6\%$ for the earlier BIC model), a higher ΔBIC value (80.7 here, compared to 59.1 for the earlier model), and a very slightly higher %correct (70.7 compared to 70.5). The SB model has a p-value relative to the reference of 0.00, and the incremental p-values in the five steps are all 0.000, except for the 5th step, which is 0.003.

One should not be confused by the SB analysis picking out Ap_0 rather than Ap_1 ; a state is information-rich if it either increases or decreases penetrance over 'the average'. Because the model actually includes the IV component, ApEdCK, and the marginals, Z, which together constitute the independence model, and because the cardinalities of both Ap and Z are 2, the first interaction here, Ap_0Z , is equivalent to a simple ApZ variable-based component, which adds only 1 degree of freedom to the independence model. The third component, K_2Z , is also simple, but it is a real state-based component since the cardinality of K is 3; knowing K_2Z and also ApEdKC and Z does not tell us $p(K_0Z)$ or $p(K_1Z)$. It was noted above that K is predictive of Z; this SB model tells us more specifically that it is K_2 that is especially predictive.

This SB model is also interesting because it includes in components #2, 4, and 5, interaction effect between states of Ap, Ed, and Z, between states of Ap, Ed, C, and Z, and between states of Ap, Ed, C, K, and Z. The last of these interaction effects is especially complex. Since this is our first application of state-based RA modeling to genomic data, and since our sample size is small, we hesitate to make assertions based on these findings. We note, however, that this SB model

supports the interaction effect between Ed and C that was found in the variable-based models for Ap=0 in Table V

V. DISCUSSION

In summary, our results suggest that:

1. APOE genotype, education level and *SORCSI* (K) are solid predictors of case/control status, because they appear in the most conservative (BIC) best model. *SORCSI* also gains support because it appears in both Ap=0 and 1 models. Although *SORCSI* has not been implicated as a susceptibility gene for AD in any GWAS studies, its potential importance in the pathophysiology of AD has been reported in two studies [36], [37].

2. *BINI* (C) is the next most likely predictor, since it appears in the Incremental-p and AIC best models of Table III, as well as the Ap=0 results of Table V. *BINI* has been reported to be a susceptibility gene for AD, and has been replicated in at least two independent GWAS studies (see www.alzforum.org)

3. Calculations that control for APOE genotype, education level and age suggest that *SORCSI* and *BINI* do not derive their predictive power indirectly via their associations with these controlling variables, but have independent predictive power as originally suggested by the models of Table III. *BINI* has been associated with lower episodic memory [38] and the effects of age on episodic memory are reported to be smaller in subjects with high educational levels compared to those with lower levels [39]. These reports lend support to our findings of an interaction between *BINI* and education level.

4. *MS4A6A*, *MTHFR* and *CLU* appear as possible predictors in models of Table III and Table V. Since these predictors were not supported by multiple different calculations, they must be regarded as only tentative.

5. The main models proposed here (Table III) do not show evidence of interaction effects between the genes we investigated. However, there are suggestions of such effects between Education and *BINI* (C) in both the analyses for APOE=0 (Table V) and in the state-based analysis. These suggestions must also be regarded as tentative.

A number of the SNPs we included in this study, for which significant associations with AD have been reported and replicated in previous GWAS studies, did not appear as predictors of AD. This may be due, at least in part, to the fact that in the GWAS studies, based on thousands of cases and controls, the majority of controls were still living and, thus, no neuropathological confirmation of control status was available.

Methodologically, we demonstrate the analytical capabilities of Reconstructability Analysis, and extend its uses beyond our earlier genomic studies.

REFERENCES

- [1] S. Shervais, Zwick, M., Kramer, P. "Reconstructability Analysis as a tool for identifying gene-gene interactions in studies of human diseases." Proceedings, IEEE International Conference on Systems, Man, and Cybernetics, Waikoloa, Hawaii, October 2005
- [2] M. Ritchie, et al., "Optimization of neural network architecture using genetic programming improves detection and modeling of gene-gene

- interactions in studies of human diseases," *BMC Bioinformatics*, 4, 28 2003.
- [3] L.W. Hahn, et al., "Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions," *Bioinformatics*, 19, 376-382 2003.
 - [4] Y. Horikawa, et al., "Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus," *Nat Genet*, 26, 163-175 2000.
 - [5] A. Tsalenko, et al., "Methods for analysis and visualization of SNP genotype data for complex diseases," *Pacific Symposium on Biocomputing*, 548-561 2003.
 - [6] N.J.Cox, et al., "Loci on chromosomes 2 (NIDDM1) and 15 interact to increase susceptibility to diabetes in Mexican Americans," *Nat Genet*, 21, 213-215 1999.
 - [7] S. Shervais, Zwick, M., Kramer, P., Westaway, S., "Reconstructability Analysis as a tool for identifying gene-gene interactions in studies of human diseases." Presentation, International Genetic Epidemiology Society Annual Meeting, Kahuku, Hawaii, 2009
 - [8] K. Walder, et al., "Calpain 3 gene expression in skeletal muscle is associated with body fat content and measures of insulin resistance," *Int J Obes Relat Metab Disord*, 26, 442-449 2002.
 - [9] W.R. Ashby, "Constraint Analysis of Many-Dimensional Relations," *General Systems Yearbook*, 9, 99-105 1964.
 - [10] G. Broekstra "Nonprobabilistic constraint analysis and a two-stage approximation method of structure identification." *Proceedings of the Society for General Systems Studies*, Houston 1979.
 - [11] R.E. Cavallo The role of systems methodology in social science research. M. Nijhoff, Boston 1979.
 - [12] F. Cellier and D. Yandell "SAPS-II: A New Implementation of the Systems Approach Problem Solver," *Internat J Gen Sys*, 13, 307-322 1987.
 - [13] R.C. Conant, "Set-Theoretic Structure Modeling," *Internat J Gen Sys*, 7, 93-107 1981.
 - [14] B. Jones, "Reconstructability Analysis for General Functions," *Internat J Gen Sys*, 11, 133-142 1985.
 - [15] G. Klir, "Identification of Generative Structures in Empirical Data," *Internat J Gen Sys*, 3, 89-104 1976.
 - [16] K. Krippendorff, "An Algorithm for Identifying Structural Models of Multivariate Data," *Internat J Gen Sys*, 7, 63-79 1981.
 - [17] K. Krippendorff, *Information Theory* Sage, Beverly Hills 1986.
 - [18] G. Klir, "Reconstructability Analysis: An Offspring of Ashby's Constraint Theory Systems Research," 3, 267-271 1986.
 - [19] M. Zwick, "Wholes and Parts in General Systems Methodology." In Wagner, G. (ed), *The Character Concept in Evolutionary Biology*. Academic Press, New York, 237-256 2001.
 - [20] M. Zwick, "An Overview of Reconstructability Analysis," *Kybernetes*, 33, 877-905 2004.
 - [21] G. Klir, *The Architecture of Systems Problem Solving*. Plenum Press, New York 1985.
 - [22] Y. Bishop, et al., *Discrete Multivariate Analysis*. MIT Press, Cambridge 1978.
 - [23] M. Zwick, "Reconstructability Analysis of Epistasis." *Annals of Human Genetics*, vol. 75, issue 1, pp. 157-171. DOI: 10.1111/j.1469-1809.2010.00628.x 2011.
 - [24] M. Zwick, and Johnson, M.S. "State-Based Reconstructability Analysis," *Kybernetes*, 33, 1041-1052 2004.
 - [25] M. Johnson, "State-Based Systems Modeling: Theory, Implementation, and Applications." PhD Dissertation, Portland State University 2005.
 - [26] M. Zwick, "Reconstructability Analysis With Fourier Transforms." *Kybernetes*, 33, 877-905 2004. <http://www.sysc.pdx.edu/download/papers/raftpitf.pdf>
 - [27] S. Purcell, et al., "PLINK: a tool set for whole-genome association and population-based linkage analyses," *Am J Hum Genet*, 81, 559-575 2007.
 - [28] M. Ritchie, et al., "Genetic Programming Neural Networks as a Bioinformatics Tool for Human Genetics." *Genetic and Evolutionary Computation Conference*. Seattle, 438-448 2004.
 - [29] S. H. Chen et al., "A support vector machine approach for detecting gene-gene interaction," *Genet Epidemiol*, 32, 152-167 2008.
 - [30] D. Velez, et al., "A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction," *Genet Epidemiol*, 31, 306-315 2007.
 - [31] G. Schwarz, "Estimating the dimension of a model." *Ann. Stat.* 6: 461-464 1978.
 - [32] H. Akaike, "A new look at the statistical model identification". *IEEE Trans. Automat. Control* AC-19: 716-723 1974.
 - [33] J. Hosseini, et al., "Segment Congruence Analysis Via Information Theory." *International Society for General Systems Research*. Philadelphia, G62-G77 1986.
 - [34] J. Fusion, J., Willett, K. and Zwick, M., *OCCAM: A Reconstructability Analysis Program* 2012. <http://www.sysc.pdx.edu/download/papers/woccaman.pdf>
 - [35] K. Willett, and Zwick, M. "A Software Architecture for Reconstructability Analysis," *Kybernetes*, 33, 997-1008 2004.
 - [36] C. Reitz, et al., "SORCS1 Alters Amyloid Precursor Protein Processing and Variants May Increase Alzheimer's Disease Risk," *Ann Neurol*. 2011 January ; 69(1): 47-64. doi:10.1002/ana.22308
 - [37] C. Reitz, Lee JH, Rogers RS, Mayeux R, "Impact of Genetic Variation in SORCS1 on Memory Retention." *PLoS ONE* 6(10): e24588. doi:10.1371/journal.pone.0024588 2011.
 - [38] S. Barral, et al., "Genotype patterns at PICALM, CR1, BIN1, CLU, and APOE genes are associated with episodic memory." *Neurology* 78:1464-1471 2012.
 - [39] L. Angel, Séverine Fay, Badiâa Bouazzaoui, Alexia Baudouin, Michel Isingrini "Protective role of educational level on episodic memory aging: An event-related potential study". *Brain and Cognition* 74:312-323 2010.