Published in final edited form as:

Proceedings (IEEE Int Conf Bioinformatics Biomed). 2011 ; : 696–702. doi:10.1109/BIBMW.

A Hybrid Bayesian Network/Structural Equation (BN/SEM)

Modeling Approach for Detecting Physiological Networks for Obesity-related Genetic Variants

Christine W. Duarte¹, Yann C. Klimentidis¹, Jacqueline J. Harris¹, Michelle Cardel², and José R. Fernández²

Christine W. Duarte: cduarte@uab.edu

¹Section on Statistical Genetics, Department of Biostatistics, University of Alabama at Birmingham

²Department of Nutrition Sciences, University of Alabama at Birmingham

Abstract

GWAS studies have been successful in finding genetic determinants of obesity. To translate discovered genetic variants into new therapies or prevention strategies, molecular or physiological mechanisms need to be discovered. One strategy is to perform data mining of data sets with detailed phenotypic data, such as those present in dbGAP (database of Genotypes and Phenotypes) for hypothesis generation. We propose a novel technique that combines the power and computational efficiency of existing Bayesian Network (BN) learning algorithms with the statistical rigor of Structural Equation Modeling (SEM) to produce an overall system that searches the space of potential networks and evaluates promising candidates using standard SEM model selection criteria. We demonstrate our method using the analysis of a candidate SNP data set from the AMERICO sample, a multi-ethnic cross-sectional cohort of roughly three hundred children with detailed obesity-related phenotypes. We demonstrate our approach by showing genetic mechanisms for three obesity-related SNPs.

1. Introduction

There is a growing epidemic of obesity in the United States that is a major concern for public health because obesity is a risk factor for cardiovascular disease, diabetes, cancer, and many other diseases that are leading causes of death, disability, and health care expenditure in the United States. Tremendous public investment has been made into finding genetic determinants of obesity, most recently in the form of GWAS (genome-wide association) studies. GWAS studies have resulted in the discovery of many associated genetic loci; for instance, the GIANT consortium has discovered or validated a total of 32 SNPs associated with body mass index (BMI) [1, 2].

While many successes in the genetics of obesity have occurred, the overall assessment of the contribution of GWAS studies to obesity research remains controversial, largely due to the lack of translation of basic scientific discoveries into improved health outcomes for patients. There are several possible reasons for the failure to translate basic discoveries in the genetics of obesity into improvement of methods for prevention and treatment for obesity-related traits. One is that the overall explained variance of the discovered loci is not high. For instance, only 1–2% of the variance of BMI is explained by the 32 loci found in the GIANT study [1], and thus most of the genetic variation in obesity remains undiscovered as of yet. However, there are examples in many GWAS studies in which discovered genetic variants, even of small effect, yield new insight into the pathophysiology of disease. For instance, in

Type II diabetes (T2D), while only 10% of the disease heritability is accounted for by the set of discovered genetic variants, novel pathways have been discovered with previously-unknown roles in Diabetes, and the role of β -cell dysfunction in T2D has been highlighted as a result of these studies [3].

For GWAS of BMI, some of the newly-discovered variants are in genes expressed in the central nervous system, and key hypothalamic pathways of energy balance are potentially implicated [1]. However, many of the discovered variants are in or near genes that are poorly-understood, and the genetic mechanisms of these variants is unknown. In order to translate discovered genetic variants into new therapies or prevention strategies, first molecular and/or physiological mechanisms of each variant needs to be discovered. However, it is difficult to know where to start when exploring potential genetic risk mechanisms, especially when little is known about a gene. One potential strategy is do use data mining of existing data sets in which detailed phenotypic data is available, and use machine learning and/or statistical modeling of this data for hypothesis generation. This strategy has much promise because of the many large, publicly-available data sets available in dbGAP.

However, the set of statistical and computational methods for data analysis are limited. While many powerful and robust methods for GWAS analysis have been developed, the vast majority of methods are unigenic and univariate, as required for the high dimensionality of GWAS data. Notable exceptions include methods for multi-trait analysis [4, 5], epistatic methods [6], and pathway-based methods for collapsing association signals for genes or pathways for improved power [7]. However, none of these methods are optimized for finding detailed mechanisms of action for genetic variants that likely involve the combined influence of multiple, interacting phenotypes. Here we propose a powerful, robust, statistically-interpretable multivariate statistical modeling system that can be used to dissect the genetic effects of risk variants on obesity-related diseases and conditions.

The method demonstrated here will allow investigators to more productively mine the currently-available data sets of genotypes and phenotypes for obesity-related traits in order to discover testable hypotheses for physiological mechanisms by which genetic variants lead to an obesity phenotype. Once these new hypotheses for genetic mechanisms are proposed, they can be tested and confirmed in the laboratory or in epidemiological studies, and the resulting knowledge can be used to develop novel strategies for prevention and treatment of obesity-related diseases.

2. Materials and Methods

2.1 Description of Americo Study

A total of 298 children, age 7 to 12 years (53% male), were recruited as part of a cross-sectional cohort study examining population differences in metabolic phenotypes among healthy children (no major illnesses or medical diagnoses). Race/ethnicity was determined by the parents of the subjects who could classify their children into either of these categories: African American (AA; n=96), Hispanic American (HA; n=78), European American (EA; n=114), or Bi-racial (n=6). All children were pubertal stage ≤ 3 as assessed by a pediatrician according to the criteria of Marshall and Tanner [8]. Informed assent and consent were obtained from children and parents respectively, as approved and regulated by the University of Alabama at Birmingham Institutional Review Board. All measurements were taken between 2004 and 2008 at the University of Alabama at Birmingham General Clinical Research Center (GCRC) and Department of Nutrition Sciences. Phenotypes consisted of anthropometric measurements, insulin measurements, glucose and lipids, blood pressure, body composition assessed by dual-energy x-ray absorptiometry (DXA), Physical

Activity by Accelerometer. Details of the study protocol and phenotype collection can be found elsewhere [9, 10].

2.2 Genotyping and Univariate Association Analysis

96 SNPs with validated associations from prior GWAS studies for BMI, hypertension, Type II Diabetes, and other obesity-related traits were selected for genotyping in the Americo sample [2, 11–17] (See Supplementary Table 2). DNA was obtained from all study participants. Information on 96 variants identified and replicated by GWAS for BMI, Type II Diabetes, Hypertension, and other obesity-related traits are listed in Supplementary Table 2. DNA was genotyped using Illumina Golden Gate at the UAB Heflin Genotyping Core. Each SNP was individually tested using an additive model for association with each trait, adjusting for age, Tanner stage, sex, racial/ethnic group, and height² (for total body fat and total lean mass).

2.3 Specific Algorithm

The hybrid Bayesian Network Structural Equation Modeling (BN/SEM) approach that we have implemented consists of the following steps:

1. For each SNP, find the set of associated traits at a pre-determined p -value threshold (after correcting for covariates).
2. For the SNP, covariates, and associated traits from the previous step, discover the highest-scoring network using DEAL (see details below).
3. Export DEAL network to MPLUS and fit network and obtain model diagnostics.
4. If model fit is appropriate, stop. Otherwise, make modifications to network until model fit is adequate.

The entire procedure including data import, univariate association analysis, Bayesian Network discovery using DEAL, export of discovered network to MPLUS, and MPLUS model evaluation, has been implemented in a custom R program which is available upon request. Final models from MPLUS were visualized using Cytoscape (<http://www.cytoscape.org>).

2.4 DEAL Algorithm Details

DEAL is a package for learning Bayesian Networks in the R programming language (<http://cran.r-project.org/web/packages/deal/index.html>). DEAL uses a Bayesian framework for learning networks of discrete and continuous variables using a conditional Gaussian likelihood with conjugate updating implemented used a heuristic search strategy. Banlists can be used to disallow edges directed into exogenous variables (such as the SNP and the covariates). The “learn” and “autosearch” functions are used for initializing and searching for the best network, respectively, and default prior distributions are used for all variables.

2.5 SEM Modeling Details

Structural models were fit in MPLUS (www.statmodel.com) using MLM estimation (i.e. the Satorra-Bentler chi-squared test) which is designed to be robust in the presence of normality deviations in the modeling of continuous traits. All of the endogenous variables were continuous. Covariates that were categorical were coded as dummy variables. Requirements for an acceptable model fit included: non-significant chi-squared goodness of fit statistic, RMSEA (root mean square error of approximation) < 0.06, CFI (comparative fit index) > 0.95, SRMR (standardized root mean residual) < 0.08, and model parameters significant at a $p < 0.1$ for most or all of the edges.

3. Results

The univariate association results are shown in Supplementary Table 3. Of the 96 SNPs that were genotyped in the Americo panel, 2 were discarded due to missingness of one of the three genotypic classes (rs1111875 and rs328). For the other SNPs, 50 had two or more associated traits and thus were submitted for network discovery in DEAL. Here we present the network results for three representative genetic variants: one associated variant for three different GWAS traits: Type II Diabetes (T2D), body mass index (BMI), and blood pressure/hypertension (BP/HT). Table 1 shows the univariate association results for all traits with $p < 0.05$ for the three specific SNPs (rs4402960 in the gene IGF2BP2, rs2681492 in the gene ATP2B1, and rs7561317 in the gene TMEM18). Supplementary Table 1 lists the phenotypes from the Americo study used in this analysis, and Supplementary Table 2 lists the SNPs genotyped in the sample.

Figure 1 shows the Bayesian Network proposed using DEAL for rs4402960 in the gene IGF2BP2 (left) and the final network evaluated in MPLUS (right). The model evaluation details are shown in Table 2. Edges from AA (dummy variable for African American ethnicity) and the SNP were added to dtotlean (total lean mass) in the final network. The final model shows that the SNP has independent mechanisms for effect on each of the associated traits, total lean mass and insulin sensitivity. Figure 2 shows the Bayesian Network proposed using DEAL for rs2681492 in the gene ATP2B1 (left) and the final network in MPLUS (right). The edge from AA (African American ethnicity) to acwdmod (moderate activity level) was removed in the final network. See Table 3 for model evaluation details. This discovered model shows support for a mechanism in which genetic risk for diastolic blood pressure is mediated by activity level.

Finally, Figure 3 shows the discovered Bayesian Network proposed for the SNP rs7561317 in TMEM18 (left) and the final model in MPLUS (right), with details of the final model fit given in Table 4. taat was removed from the final model because it is highly correlated with saat ($r=0.99$) and caused difficulties in model fitting in MPLUS. Edges from iaat, saat, and taat into BMI were replaced by a single edge from saat which was found to be sufficient for a good fit, and the covariates EA, AA, and sex were added to BMI. The covariate AA was added to dtotlean. For iaat, edges from snp, EA, and AA were removed, and the covariates sex and tanner were added. For saat, the covariates tanner and sex were added, and the taat edge was obviously removed. Overall, this model shows support for a mechanism in which the genetic effect of the TMEM18 SNP on BMI is mediated primarily through abdominal adiposity variables with additional modulation by total lean mass.

4. Discussion

We demonstrate a novel computational system that proposes physiological mechanisms of genetic risk for associated variants discovered in GWAS studies using analysis of phenotypically-rich data sets. Our system combines the power of Bayesian Network discovery algorithms for searching over the large set of potential networks with the statistical rigor of Structural Equation Modeling for assessing adequate model fit. We demonstrate our approach using preliminary results from a candidate SNP study in the Americo data set, a multi-ethnic cross-sectional study of children with detailed obesity-related phenotypes, giving three discovered genetic mechanisms as an illustration of our approach. Our future work on this computational system will include a detailed simulation study to assess the power and sample size requirements for networks of varying complexity, as well as analysis using larger data sets such as MESA, the Multi-Ethnic Study of Atherosclerosis (MESA) SHARE, which is publicly-available through dbGAP (<http://www.ncbi.nlm.nih.gov/gap>).

Acknowledgments

The authors would like to thank the Americo study participants and families, and acknowledge the following NIH grants: R01-DK067426, NIH T32-HI007457, 5P30DK056336-09, CA47888, and P60-DK079626. The opinions expressed are those of the authors and not necessarily those of the NIH or any other organization with which the authors are affiliated.

References

1. Hebebr J, et al. Chipping away the 'missing heritability': GIANT steps forward in the molecular elucidation of obesity - but still lots to go. *Obes Facts*. 2010 Oct.vol. 3:294–303.
2. Speliotes EK, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet*. 2010 Nov.vol. 42:937–948. [PubMed: 20935630]
3. Billings LK, Florez JC. The genetics of type 2 diabetes: what have we learned from GWAS? *Ann N Y Acad Sci*. 2010 Nov.vol. 1212:59–77. [PubMed: 21091714]
4. Ferreira MAR, Purcell SM. A multivariate test of association. *Bioinformatics*. 2009 Jan 1.vol. 25:132–133. [PubMed: 19019849]
5. Yang QO, et al. Analyze Multivariate Phenotypes in Genetic Association Studies by Combining Univariate Association Tests. *Genetic Epidemiology*. 2010 Jul.vol. 34:444–454. [PubMed: 20583287]
6. Wan X, et al. BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *Am J Hum Genet*. 2010 Sep 10.vol. 87:325–340. [PubMed: 20817139]
7. Wang K, et al. Analysing biological pathways in genome-wide association studies. *Nat Rev Genet*. 2010 Dec.vol. 11:843–854. [PubMed: 21085203]
8. Marshall WA, Tanner JM. Growth and physiological development during adolescence. *Annu Rev Med*. 1968; vol. 19:283–300. [PubMed: 4297619]
9. Casazza K, et al. Differential influence of diet and physical activity on components of metabolic syndrome in a multiethnic sample of children. *J Am Diet Assoc*. 2009 Feb.vol. 109:236–244. [PubMed: 19167950]
10. Casazza K, et al. Physical fitness, activity, and insulin dynamics in early pubertal children. *Pediatr Exerc Sci*. 2009 Feb.vol. 21:63–76. [PubMed: 19411712]
11. Willer CJ, et al. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat Genet*. 2009 Jan.vol. 41:25–34. [PubMed: 19079261]
12. Meyre D, et al. Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in European populations. *Nat Genet*. 2009 Feb.vol. 41:157–159. [PubMed: 19151714]
13. Thorleifsson G, et al. Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. *Nat Genet*. 2009 Jan.vol. 41:18–24. [PubMed: 19079260]
14. Voight BF, et al. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat Genet*. 2010 Jul.vol. 42:579–589. [PubMed: 20581827]
15. Franceschini N, et al. Recent findings in the genetics of blood pressure and hypertension traits. *Am J Hypertens*. 2011 Apr.vol. 24:392–400. [PubMed: 20948529]
16. Rivadeneira F, et al. Twenty bone-mineral-density loci identified by large-scale meta-analysis of genome-wide association studies. *Nat Genet*. 2009 Nov.vol. 41:1199–1206. [PubMed: 19801982]
17. Teslovich TM, et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*. 2010 Aug 5.vol. 466:707–713. [PubMed: 20686565]

Figure 1.

DEAL-proposed network (top) and final network from MPLUS (bottom) for rs4402960 in IGF2BP2 which was previously shown to be associated with Type II Diabetes (T2D). Edges from AA and snp to dtotlean were added to final network. See Table 2 for model statistics. dtotlean is total lean mass normalized by height squared, and si is insulin sensitivity (see Supplementary Table 2 for more details on phenotypes).

Figure 2.

DEAL-proposed network (top) and final network from MPLUS (bottom) for rs2681492 in ATP2B1 which was previously shown to be associated with blood pressure and hypertension (BP/HT). The edge from AA to acwdmod was removed from the final network. See Table 3 for model statistics. acwdmod is a measure of physical activity, and AVE_DIA is average diastolic blood pressure (see Supplementary Table 2 for more details on phenotypes).

Figure 3.

DEAL-proposed network (top) and final network from MPLUS (bottom) for rs7561317 in TMEM18 which was previously shown to be associated with BMI (body mass index). *taat* was removed because it is highly correlated with *saat* ($r=0.99$) and caused difficulties in model fitting in MPLUS. Edges from *iaat*, *saat*, and *taat* into BMI were replaced by a single edge from *saat* which was found to be sufficient for a good fit, and the covariates EA, AA, and sex were added to BMI. The covariate AA was added to *dtotlean*. For *iaat*, edges from *snp*, EA, and AA were removed, and the covariates sex and tanner were added. For *saat*, the covariates tanner and sex were added, and the *taat* edge was obviously removed. See Table 4 for model statistics. *iaat*, *saat*, and *taat* are intraabdomininal, subcutaneous, and total

abdominal adipose tissue, respectively, and dtotlean is total lean mass normalized by height squared (see Supplementary Table 2 for more details on phenotypes).

Table 1

Univariate association results for each phenotype studied (see Supplementary Table 1 for complete list of phenotypes used) with three SNPs associated with obesity-related traits that showed nominal significance ($p < 0.05$).

SNP	Gene	Trait	Est.	P
rs4402960	IGF2BP2	dtotlean	-0.03	0.0032
rs4402960	IGF2BP2	si	0.96	0.0033
rs2681492	ATP2B1	AVE_DIA	1.68	0.0442
rs2681492	ATP2B1	acwdmod	8.697	0.0472
rs7561317	TMEM18	bmi	-0.73	0.0183
rs7561317	TMEM18	dtotlean	-0.03	0.0173
rs7561317	TMEM18	iaat	-6.46	0.0265
rs7561317	TMEM18	saat	-19.27	0.0427

Table 2

Model fit statistics and estimated parameters (from MPLUS) for the network shown in Figure 1 (right) for rs4402960 in IGF2BP2.

Statistic	Value	df
Chi-Sq Test	0.694	2
RMSEA	0.000	
CFI	1.000	
SRMR	0.006	

Parameter		Est.	P-Val
DTOTLEAN	TANR	0.064	0.000
DTOTLEAN	SEX	−0.105	0.000
DTOTLEAN	AA	0.049	0.001
DTOTLEAN	SNP	−0.031	0.003
SI	SNP	0.950	0.028
SI	TANR	−1.601	0.000
SI	EA	1.750	0.000
SI	SEX	−1.034	0.015

Table 3

Model fit statistics and estimated parameters (from MPLUS) for the network shown in Figure 2 (right) for rs7561317 in TMEM18.

Statistic	Value	df	P-Val
Chi-Sq Test	1.098	3	0.778
RMSEA	0.000		
CFI	1.000		
SRMR	0.014		

Parameter		Est.	P-Val
AVE_DIA	AA	2.72	0.001
AVE_DIA	ACWDMOD	0.04	0.000
ACWDMOD	SNP	11.20	0.030
ACWDMOD	SEX	-18.07	0.000

Table 4

Model fit statistics and estimated parameters (from MPLUS) for the network shown in Figure 3 (right) for rs7561317 in TMEM18.

Statistic	Value	df
Chi-Sq Test	12.104	15
RMSEA	0.000	
CFI	1.000	
SRMR	0.021	

Parameter		Est.	P-Val
BMI	DTOTLEAN	9.222	0.000
BMI	SAAT	0.029	0.000
DTOTLEAN	TANNER	0.060	0.000
DTOTLEAN	SEX	−0.085	0.000
DTOTLEAN	IAAT	0.001	0.000
DTOTLEAN	AA	0.056	0.001
IAAT	SAAT	0.261	0.000
IAAT	SEX	−6.530	0.001
IAAT	TANNER	−3.300	0.012
SAAT	SNP	−19.929	0.003
SAAT	EA	−34.361	0.007
SAAT	AA	−48.125	0.000
SAAT	TANNER	8.564	0.363
SAAT	AGE	6.441	0.065
SAAT	SEX	26.291	0.012