Building Bayesian Networks from GWAS Statistics Based on Independence of Causal Influence

Lu Zhang*, Qiuping Pan*, Xintao Wu* and Xinghua Shi†
*University of Arkansas
Email: lz006@uark.edu, qpan@email.uark.edu, xintaowu@uark.edu
†University of North Carolina at Charlotte
Email: x.shi@uncc.edu

***Abstract*—Genome-wide association studies (GWASs) have received an increasing attention to understand genotype-phenotype relationships. In this paper, we study how to build Bayesian networks from publicly released GWAS statistics to explicitly reveal the conditional dependency between single-nucleotide polymorphisms (SNPs) and traits. The key challenge in building a Bayesian network is the specification of the conditional probability table (CPT) of an variable with multiple parent variables. We employ the Independence of Causal Influences (ICI) which assumes that the causal mechanism of each parent variable is mutually independent. Specifically, we derive a formulation from the Noisy-or model, one of the ICI models, to specify the CPT using the released GWAS statistics. We prove that the specified CPT is accurate as long as the underlying individual-level genotype and phenotype profile data follows the Noisy-or model. We empirically evaluate the Noisy-or model and its derived formulation using data from openSNP. Experimental results demonstrate the effectiveness of our approach.**

I. Introduction

Genome-wide association studies (GWASs) have received an increasing attention due to its promising potential in genetic diagnostics, drug development and personalized medicine. GWAS examine genetic variation from different individuals to find out whether certain single-nucleotide polymorphisms (SNPs) occur more frequently in people with a particular trait/disease. Many SNPs have been reported to be associated with various complex diseases. Genotypes of SNPs at individual level are sensitive and thus are usually under controlled access guided by data access agreements. However, most of the GWAS statistics and SNP-trait associations are publicly available. For example, the GWAS catalog [1] collects and publicly releases literature-derived GWAS statistics, including pair-wise SNP-trait associations and related statistics.

Modeling the associations of genetic variants with traits has been reported to aid the understanding the relationship between genotypes and phenotypes. Particularly, Bayesian networks have been demonstrated to be powerful for such modeling to dissect complex (e.g. gene interactions) or causal relationships between SNPs and associated traits [2]–[4]. However, these methods require raw genotypes of SNPs and such information may not always be available. Instead, a recent study [5] proposed a method to build a Bayesian network using only the released GWAS statistics for characterizing SNP-trait associations. In their method, each trait or SNP is represented as a node in the network, and a two-layered Bayesian network is built with no arc either among either the trait nodes or among the SNP nodes. Nonetheless, this work suffers from a significant limitation in the sense that, in its constructed two-layered Bayesian network, the arcs go from trait nodes to SNP nodes. The direction of these arcs contradicts to the fact that a trait in GWAS is usually treated as a dependent variable and a SNP is considered as an independent variable, which means the arcs should point from SNP nodes to trait nodes. In this paper, we correct the orientation of the arcs and study how to build an accurate Bayesian network from GWAS statistics.

The key challenge of building a Bayesian network using GWAS statistics is that, when a dependent variable (i.e., trait) has associations with multiple independent variables (i.e., SNPs), the Bayesian network needs to specify the conditional probability table (CPT) of the trait conditional on every value combination of its associated SNPs. However, GWAS statistics only provide the independent associations of a SNP with its associated trait. The information about epistatic interactions among multiple SNPs that bring about joint effect on a trait is rather limited. Additionally, complex traits are commonly associated with many SNPs. Therefore, it is a combinatorial problem for specifying CPTs because the number of the conditional probability distribution values in the CPT is exponential to the number of SNPs associated with a trait.

To deal with this issue, in this paper, we propose to employ the assumption of Independence of Causal Influences (ICI) which is widely used in building Bayesian networks from data [6]. ICI assumes that, when there are multiple parent variables, the causal mechanism of each parent variable is mutually independent, so that the combined influence of the multiple parents is decomposable into a series of independent influence of each parent variable. The ICI model is usually learned through the process of finding the parameters that make the model fit the data best [7], and hence cannot be learned from the statistics directly. Among different ICI models, we pay special attention to the Noisy-or model [8]. We derive an equivalent formulation from the Noisy-or model which can be used to specify the CPT from the released GWAS statistics where the underlying genotypes can be unknown. We prove that, the specified CPT is accurate as long as the individual-level profile data follows the Noisy-or model. To evaluate our proposed method, we use the profile data from openSNP [9], a platform that allows customers to share and publish

978-1-5090-1610-5/16/$31.00 ©2016 IEEE

their genetic test results. We extract the statistics from the profile data following the GWAS data release procedure, and construct the Bayesian network from the statistics. We then empirically evaluate the fitness of the Noisy-or model and its derived CPT specification formulation.

II. The GWAS Catalog and Statistics

In GWAS, participants are divided into two groups: individuals with the trait/disease (case group) and matched individuals without the trait/disease (control group). Each individual is genotyped using microarray technology or more recently sequencing technologies. Each SNP (only biallelic SNPs are considered in most GWAS) can be viewed to have two possible alleles. The allele that is more frequent in the case group comparing with the control group is called the risk allele (e.g., A), and the other one is called the non-risk allele (e.g., G). Each individual carries a pair of alleles inherited from both parents and the genotype refers to the two alleles an individual has for a particular SNP. The genotype that contains two risk alleles is called be a homozygote for risk alleles (e.g., AA), the genotype that contains two non-risk alleles is called a homozygote for non-risk alleles (e.g., GG), and the genotype that contains one risk allele and one non-risk allele is called a heterozygote (e.g., AG).

TABLE I: The Contingency Table on Genotypes.

	AA	AG	GG	Total
Cases	r_0	r_1	r_2	R
Controls	s_0	s_1	s_2	S
Total	n_0	n_1	n_2	N

TABLE II: The Contingency Table on Alleles.

	A	G	Total
Cases	$2r_0+r_1$	r_1+2r_2	$2R$
Controls	$2s_0+s_1$	s_1+2s_2	$2S$
Total	$2n_0+n_1$	n_1+2n_2	$2N$

In a typical process of a GWAS, firstly a genotype profile dataset is generated by genotyping the individuals in the case group and the control group. For each SNP, the genotype frequency is counted over the two groups to obtain a 3×2 contingency table as shown in Table I (r_0 denotes the number of individuals in the case group with genotype AA and so forth). Then, the genotype frequency is transformed into the 2×2 allele frequency table as shown in Table II. After that, statistical tests such as chi-square test, are performed on the allele contingency table to investigate whether there is an association between the SNP and the trait. In addition to a p-value indicating the significance of the association, GWAS also reports odds ratios that measure the difference of frequency of an allele in the case versus control group. If the odds ratio is larger than 1, it indicates that the risk allele is more frequent in the case group than it is in the control group. Finally, the trait and its significantly associated SNPs are reported, along with the risk allele type and related statistics. The GWAS catalog extracts these information from literature, and releases the curated collection of SNP-trait associations.

III. Constructing A Two-Layered Bayesian Network From GWAS Statistics

We follow a procedure similar to that in [5] to construct a Bayesian network from the GWAS catalog, with the orientation of the arcs corrected to reflect the real direction of SNP-trait associations. The first step is to identify useful information from the GWAS catalog. Formally, we can extract a set of traits $\mathcal{T}$ and a set of SNPs $\mathcal{S}$ from the GWAS catalog. For each trait T, we denote the control group by $T = 0$, and the case group by $T = 1$. Then, we can extract a subset of m associated SNPs $\mathbf{S} = \{S_1, \cdots, S_m\}$. For each SNP $S_j \in \mathbf{S}$, we denote the SNP allele as $S_j^a = \{0, 1\}$, where 0 represents the non-risk allele and 1 represents the risk allele. The risk allele frequency in the control group denoted by $f_j^c(1)$, and the odds ratio denoted by O_j can be extracted from the GWAS catalog. Its risk allele frequency in the case group (denoted by $f_j^t(1)$) can be derived using $f_j^c(1)$ and O_j as follows.

$$f_j^t(1) = \frac{O_j \cdot f_j^c(1)}{O_j \cdot f_j^c(1) + 1 - f_j^c(1)}. \quad (1)$$

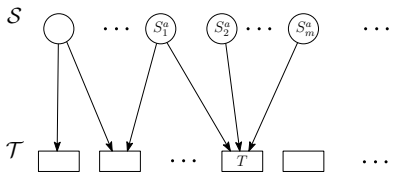

Fig. 1: A Two-layered Bayesian network of SNPs and associated traits.

The second step is to build the structure of the Bayesian network based on the extracted SNP-trait associations. If an SNP is associated with a trait, then we add an arc connecting the two nodes pointing from the SNP node to the trait node. Since the trait-trait or SNP-SNP associations cannot be acquired from the GWAS catalog, the arcs between traits and those between SNPs are prohibited, as shown in Figure 1.

The last step is to specify the CPT stored at each node. First, we need to specify the prior probability of each SNP which can be thought of as the frequency of each SNP in a population. In our paper we acquire the prior probability of each trait $P(T)$ which can be obtained from literature or the Internet, and then use it to compute the prior probability of a SNP $P(S_j^a)$. Thus, the prior probability of each SNP S_j is computed as

$$P(S_j^a = s_j) = P(S_j^a = s_j | T = 0)P(T = 0) + P(S_j^a = s_j | T = 1)P(T = 1). \quad (2)$$

Second, we need to specify the conditional probability of each trait given its associated SNPs. We compute $P(T = 0|\mathbf{S}^a = \mathbf{s})$ as follows.

$$P(T = 0|\mathbf{S}^a = \mathbf{s}) = \frac{P(T = 0) \prod_{j=1}^m P(S_j^a = s_j | T = 0)}{\prod_{j=1}^m P(S_j^a = s_j)}. \quad (3)$$

We prove in the next section that, the above computation of $P(T = 0|\mathbf{S}^a = \mathbf{s})$ is accurate as long as the profile data follows the Noisy-or model. Note that the maximum likelihood estimates of $P(S_j^a = s_j|T = 0)$ and $P(S_j^a = s_j|T = 1)$ are precisely the allele frequencies $f_j^c(s_j)$ and $f_j^t(s_j)$ that can be extracted from the GWAS catalog. Therefore, the

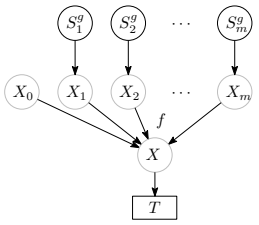

Fig. 2: The ICI model

Bayesian network can be constructed with the information solely extracted from the GWAS catalog, without knowledge of the raw genotypes.

IV. Modeling SNP-Trait Associations

A. Data Representation

Assume that we are given the profile data $\mathcal{D} = \{\cdots, \mathbf{d}^l, \cdots\}$ where each record $\mathbf{d}^l$ contains the traits and SNP genotypes of an individual. Consider a trait T that is associated with m SNPs $\mathbf{S} = \{S_1, \cdots, S_m\}$. As stated in Section II, each individual carries a combination of alleles for an SNP in his genotype profile. For distinguishing the allele and the genotype for each SNP S_j, we denote the SNP genotype as $S_j^g = \{0, 1, e\}$, where 0 represents the homozygote for non-risk allele, 1 represents the homozygote for risk allele, and e represents the heterozygote. As shown in Table II, the connection between the allele frequency and the genotype frequency is given as follows, where $s = \{0, 1\}$ and $t = \{0, 1\}$. Note that we use the same symbol s to denote the allele type (e.g., A) and its corresponding homozygote (e.g., AA).

$$P(S_j^a = s|T = t) \;=\; P(S_j^g = s|T = t) + \frac{1}{2}P(S_j^g = e|T = t). \quad (4)$$

This equation can be extended to a combination of multiple SNPs, given by

$$P(\mathbf{S}^a = \mathbf{s}|T = t) \;=\; \sum_{j=0}^{m} \frac{1}{2^j} \sum_{\mathbf{s}' \in \pi_j(\mathbf{s})} P(\mathbf{S}^g = \mathbf{s}'|T = t), \quad (5)$$

where $\pi_j(\mathbf{s})$ denotes the all the possible genotype combinations that replace j homozygotes in $\mathbf{s}$ with heterozygotes.

B. Independence of Causal Influence

We describe the Independence of Causal Influence in our context. In the Bayesian network that exhibits ICI, each SNP S_j is connected with a hidden variable X_j, which represents the "effective value" of S_j on T. The connection between S_j and X_j is defined via various stochastic or deterministic functions. Then, X_js are combined using some deterministic function $f(\cdot)$. Usually, in order to be a decomposable function, $f(\cdot)$ is required be associative and commutative. Besides, an additional hidden variable X_0 is added to represent some background knowledge, resulting a combination function $X = f(X_0, X_1, \cdots, X_m)$. Finally, another stochastic or deterministic function is applied to X to get the value of T. The structure of the general formulation of the ICI models is shown in Figure 2. In the following, we introduce the Noisy-or model and derive Equation (3) from its formulation.

C. The Noisy-or Model

The Noisy-or model can be considered as a generalization of the deterministic OR relation. In this model, each hidden variable X_j is a binary variable taking values of 0 and 1. The connection between each pair of S_j and X_j is defined as the following probabilistic distribution:

$$\text{for each } j,\; P(X_j = 0|S_j^g = s_j) = \begin{cases} 1 & \text{if } s_j = 0, \\ \theta_j(s_j) & \text{otherwise,} \end{cases}$$

where $s_j = \{0, 1, e\}$, and $\theta_j(s_j)$ is called the noise parameter representing the probability that the presence of the risk allele of S_j would be effective in the occurrence of the trait. It is also defined that

$$P(X_0 = 0) \;=\; \theta_0,$$

which is called a leak probability that allows the trait to occur when the risk alleles of all SNPs are absent. Then, $f(\cdot)$ is defined as the deterministic OR function that takes all X_js as the input, i.e.,

$$f(X_0 = x_0, X_1 = x_1, \cdots, X_m = x_m) \;=\; x_0 \vee x_1 \vee \cdots \vee x_m.$$

Finally, T directly takes the value of the output of $f(\cdot)$. Straightforwardly, T equals 0 if and only if all X_js take the value of 0. Thus, the probability of $T = 0$ given $\mathbf{S}^g = \mathbf{s}$ is given by

$$\begin{aligned} P(T = 0|\mathbf{S}^g = \mathbf{s}) \;&=\; P(X_0 = 0) \prod_{j:s_j \neq 0} P(X_j = 0|S_j^g = s_j) \\ &=\; \theta_0 \prod_{j:s_j \neq 0} \theta_j(s_j). \end{aligned}$$

To learn the Noisy-or model, let each record $\mathbf{d}^l = \{t^l, \mathbf{s}^l\}$ represent the trait and genotype profile of an individual. The objective function is typically formalized as the maximization of log-likelihood of the model given the observed data: $\sum_{l=1}^{|\mathcal{D}|} \log P(\{T, \mathbf{S}\} = \mathbf{d}^l)$. According to [7], maximizing the log-likelihood is equivalent to maximizing the conditional log-likelihood $\sum_{l=1}^{|\mathcal{D}|} \log P(T = t^l|\mathbf{S} = \mathbf{s}^l)$. Following a similar procedure in [7], the Noisy-or model can be learned using the EM algorithm [10]. We include the EM algorithm with the derived formulas in our technical report [11].

The following lemma and proposition derive Equation (3) from the Noisy-or model.

Lemma 1: Let $P(T|\mathbf{S}^g)$ follow the Noisy-or model. Then for $\mathbf{S}^g$ we have

$$P(\mathbf{S}^g = \mathbf{s}|T = 0) \;=\; \prod_{j=1}^{m} P(S_j^g = s_j|T = 0), \quad (6)$$

and for $\mathbf{S}^a$ we also have

$$P(\mathbf{S}^a = \mathbf{s}|T = 0) \;=\; \prod_{j=1}^{m} P(S_j^a = s_j|T = 0). \quad (7)$$

Proposition 1: Let $P(T|\mathbf{S}^g)$ follow the Noisy-or model. Then we have

$$P(T = 0|\mathbf{S}^a = \mathbf{s}) = \frac{P(T = 0) \prod_{j=1}^{m} P(S_j^a = s_j|T = 0)}{\prod_{j=1}^{m} P(S_j^a = s_j)}.$$

Please refer to our technical report [11] for proof details.

V. Experiments

We conduct experiments to evaluate the model fitness in modeling SNP-trait associations. We use the profile data from openSNP where more than two thousand users over the world share their genotype and phenotype files. In summary, there are 395 traits and more than 2,000,000 SNPs in the openSNP database by Aug 15, 2016. We use openSNP of version 20151231 in the experiments. After filtering, we obtain a dataset containing 23 traits and 256,845 SNPs. Please refer to our technical report [11] for detailed data setup.

A. Experimental Results

1) Building a Bayesian Network from the Statistics: To build the Bayesian network, we extract for each trait the associated SNPs along with the risk allele types, risk allele frequencies and odds ratios. Then, we perform the fisher's exact test of independence to test whether the association between the trait and the SNP is significant. The threshold of the p-value is set as 4×10^{-5}. We discard the traits with zero associated SNP, as well as the traits with only one associated SNP as they have no effect in testing ICI model. As a result, we obtain 7 traits and 34 associated SNPs. Please refer to our technical report [11] for details.

TABLE III: The chi-square values, degree of freedom (df), p-value, RMSEA of the Noisy-or model

Trait	Chi-square	df	p-Value	RMSEA
Eye with Blue Halo	6.73	4	0.15	0.10
Hair on Fingers	14.46	14	0.41	0.02
Irritable Bowel Syndrome	5.24	4	0.26	0.05
ADHD	55.32	53	0.38	0.02
Astigmatism	132.55	123	0.26	0.02
Do You Grind Your Teeth	50.13	49	0.42	0.02
Enjoy Driving A Car	96.33	98	0.52	NA

2) Model Fitness: We first evaluate the fitness of the Noisy-or model. For each trait, we predict the observed number of individuals with a specific trait and specific SNP genotypes, i.e., $n(T, \mathbf{S}^g)$, by computing the predicted value as $\hat{n}(T, \mathbf{S}^g) = P(T|\mathbf{S}^g)n(\mathbf{S}^g)$, where $n(\mathbf{S}^g)$ is the observed total number of individuals with the SNP genotypes. We then compute the p-value to show the significance. The model is not rejected if p-value >0.05. We further compute the Root Mean Square Error of Approximation (RMSEA) values, which indicates good fitness if smaller than 0.1. As shown in Table III, the Noisy-or model is accepted with good fitness for all traits according to the p-values and RMSEA. This means that the Noisy-or model is good for modeling SNP-trait associations.

Then, we conduct the chi-square tests for $P(T = 0|\mathbf{S}^a = \mathbf{s})$ shown in Equation (3). Since the Noisy-or model fits the data well, we also expect a good fitness of its derived formulation. We predict the observed value of $n(T, \mathbf{S}^a)$ by computing the predicted value as $\hat{n}(T, \mathbf{S}^a) = P(T|\mathbf{S}^a)n(\mathbf{S}^a)$. Note that $n(T, \mathbf{S}^a)$ and $n(\mathbf{S}^a)$ are the counts of individuals with specific SNP alleles and can be obtained from $n(T, \mathbf{S}^g)$ and $n(\mathbf{S}^g)$ similarly to Equation (5). The chi-square values, degree of freedom, p-values and RMSEA are shown in Table IV. As expected, the results show a good fitness in general.

TABLE IV: The chi-square values, p-value, degree of freedom, RMSEA of Equation (3).

Trait	Chi-sqaure	df	P-Value	RMSEA
Eye with Blue Halo	1.14	NA	NA	NA
Hair on Fingers	2.75	1	0.10	0.12
Irritable Bowel Syndrome	0.84	NA	NA	NA
ADHD	12.59	29	0.99	NA
Astigmatism	43.58	47	0.62	NA
Do You Grind Your Teeth	5.35	15	0.99	NA
Enjoy Driving A Car	111.10	609	0.99	NA

VI. Conclusions

In this paper, we studied the construction of Bayesian networks from publicly released GWAS statistics. We employed the Independence of Causal Influences (ICI) which assumes the causal mechanism of each parent variable is mutually independent. We derived a formulation from the Noisy-or model, one of the ICI models, to specify the CPT using the released GWAS statistics, and developed a Bayesian Network construction algorithm based on the CPT specification formulation. We proved that, the specified CPT is accurate as long as the underlying individual-level genotype and phenotype profile data follows the Noisy-or model. Then, we empirically evaluated the fitness of the Noisy-or model as well as its derived formulation. The results showed good fitness for both methods, indicating that the constructed Bayesian network can accurately represent the conditional dependency between SNPs and traits.

Acknowledgements

The work is supported in part by US National Science Foundation (DGE-1523115 and IIS-1502273 to QP and XW, and DGE-1523154 and IIS-1502172 to XS).

References

[1] D. Welter *et al.*, "The nhgri gwas catalog, a curated resource of snp-trait associations," *Nucleic acids research*, vol. 42, no. D1, pp. D1001–D1006, 2014.

[2] X. Jiang, R. E. Neapolitan, M. M. Barmada, and S. Visweswaran, "Learning genetic epistasis using bayesian network scoring criteria," *BMC bioinformatics*, vol. 12, no. 1, p. 1, 2011.

[3] B. Han, X.-w. Chen, Z. Talebizadeh, and H. Xu, "Genetic studies of complex human diseases: Characterizing snp-disease associations using bayesian networks," *BMC systems biology*, vol. 6, no. Suppl 3, 2012.

[4] Z. Zeng, X. Jiang, and R. Neapolitan, "Discovering causal interactions using bayesian network scoring and information gain," *BMC bioinformatics*, vol. 17, no. 1, p. 1, 2016.

[5] Y. Wang, X. Wu, and X. Shi, "Using aggregate human genome data for individual identification," in *BIBM'13*. IEEE, 2013, pp. 410–415.

[6] D. Heckerman and J. S. Breese, "Causal independence for probability assessment and inference using bayesian networks," *IEEE Trans. Syst., Man, Cybern. A, Syst.,Humans*, vol. 26, no. 6, pp. 826–831, 1996.

[7] J. Vomlel, "Noisy-or classifier," *International Journal of Intelligent Systems*, vol. 21, no. 3, pp. 381–398, 2006.

[8] J. H. Kim and J. Pearl, "A computational model for causal and diagnostic reasoning in inference systems." in *IJCAI*, vol. 83, 1983, pp. 190–193.

[9] B. Greshake, P. E. Bayer, H. Rausch, and J. Reda, "Opensnp–a crowdsourced web resource for personal genomics," *PLoS One*, vol. 9, no. 3, p. e89204, 2014.

[10] G. McLachlan and T. Krishnan, *The EM algorithm and extensions*. John Wiley & Sons, 2007, vol. 382.

[11] L. Zhang, Q. Pan, X. Wu, and X. Shi, "Building bayesian networks from gwas statistics based on independence of causal influence," University of Arkansas, Tech. Rep. DPL-2016-003, 2016.