Journal of Biomedical Informatics xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Generating a robust statistical causal structure over 13 cardiovascular disease risk factors using genomics data

Azam Yazdani [a,*], Akram Yazdani [a], Ahmad Samiei [b], Eric Boerwinkle [a]

[a] Human Genetics Center, UTHealth School of Public Health, 1200 Pressler Street, Suite E-447, Houston, TX 77030, United States
[b] Department of Software Systematic, D-14482 Potsdam, Germany

ARTICLE INFO

Article history:
Received 27 July 2015
Revised 19 January 2016
Accepted 22 January 2016
Available online xxxx

Keywords:
Causal network
Data integration
Cardiovascular disease risk factors
Partial correlation
Conditional independency
Granularity DAG

ABSTRACT

Understanding causal relationships among large numbers of variables is a fundamental goal of biomedical sciences and can be facilitated by Directed Acyclic Graphs (DAGs) where directed edges between nodes represent the influence of components of the system on each other. In an observational setting, some of the directions are often unidentifiable because of Markov equivalency. Additional exogenous information, such as expert knowledge or genotype data can help establish directionality among the endogenous variables. In this study, we use the method of principle component analysis to extract information across the genome in order to generate a robust statistical causal network among phenotypes, the variables of primary interest. The method is applied to 590,020 SNP genotypes measured on 1596 individuals to generate the statistical causal network of 13 cardiovascular disease risk factor phenotypes. First, principal component analysis was used to capture information across the genome. The principal components were then used to identify a robust causal network structure, GDAG, among the phenotypes. Analyzing a robust causal network over risk factors reveals the flow of information in direct and alternative paths, as well as determining predictors and good targets for intervention. For example, the analysis identified BMI as influencing multiple other risk factor phenotypes and a good target for intervention to lower disease risk.

0. Introduction

Interindividual variation in disease susceptibility is influenced by genetic variants, which can be organized into a defined biologic pathways or data-driven associative networks [1–3]. By identifying variables correlated with the primary endpoint of interest, we are able to classify individuals and predict future disease. Going beyond partial correlations and evaluating causal relationships among variables plays an essential first step in risk prediction, thereby promoting more efficacious treatment of current disease and prevention of future disease. By changing the level of a causal variable (e.g. LDL-cholesterol levels), we are able to change the risk of future disease (e.g. coronary heart disease), which may not be the case for mere associated variables (e.g. HDL-cholesterol levels) [4]. In the case of a randomized intervention, such as a clinical trial, identification of causation is conceptually straight forward. However, in observational studies, which represent the majority of most large-scale epidemiologic studies, causal inference is more complex. In most applications, especially "big data" applications, causal inference is embodied in Directed Acyclic Graphs (DAGs), where any inference is based on an estimated graph (i.e. nodes and edges). DAGs are illustrations of causal relationships among the variables. Mendelian randomization is an established approach to identify causal relationships [5–8] and it is natural in a biomedical setting to integrate genomics and phenotypic information to help establish directionality within a network of phenotypes. We apply this technique in large data sets from different granularities to achieve robust causal graphs (i.e. DAGs). In the present context, granularities are defined as hierarchical levels with different quiddity that the causal relationship between them is known, e.g. they are reflecting different levels of biologic organization and measurement (genomic and phenotypic, [4]). In the application shown here, we use data from a deeper granularity, the genome, to generate a robust statistical causal network among 13 risk factor phenotypes. Inclusion of genotypes in the analysis of phenotypes (e.g. plasma glucose levels) provides two advantages: first, genotypes are assumed to be measured without error, and second, there is a natural order between these granularities (genome variation → phenotype variation; G → P) and this knowledge helps identify robust directionality in the upper granularity.

* Corresponding author at: UTHealth School of Public Health, 1200 Herman Pressler, Houston, TX 77030, United States. Tel.: +1 713 500 9808.
E-mail address: azam.yazdani@uth.tmc.edu (A. Yazdani).

http://dx.doi.org/10.1016/j.jbi.2016.01.012
1532-0464/

Using genome information is a promising approach to identify directionality that is less susceptible to confounding. Previous applications in data integration using gene expression data and genotypes have followed a similar logic [9–12]. For example, Mehrabian et al. [9] integrated genotypic and phenotypic data in a segregating mouse population to generate causal relationships. Aten et al. [11] introduced an algorithm to estimate directionality among nodes in a DAG by applying information from selected single nucleotide polymorphisms (SNPs). In this study, we apply the concept of granularity in a comprehensive manner and extract information from a deeper granularity, here the genome, to achieve a robust causal network among variables of interest in the upper level of granularity, here cardiovascular risk factor phenotypes. To go beyond using a sample of SNPs, which are incomplete and may introduce instability in the study results [13], the method of principal components is used to extract information across the genome. Integration of genome information embedded in the deeper granularity and captured using principal component analysis with phenotype information in the upper granularity results in a robust causal network among the phenotypes, and we call this algorithm Granularity Directed Acyclic Graph (GDAG).

We first briefly review the theory of graphical causal inference and introduce the granularity framework and the GDAG algorithm. The utility of this approach is introduced by application to a data set including 13 cardiovascular disease risk factors and 590,020 SNP genotypes measured on 1596 individuals and then the estimated structure is further interpreted. Use of information from the genome level of granularity allowed us to robustly generate the statistical causal network among the phenotypes. A discussion of the GDAG algorithm and the results is provided.

1. Background

Assume a DAG $D = (v, \varepsilon)$ where v is a set of nodes with p elements which corresponds to a set of p random variables and ε is a set of edges which connect the nodes and shows the partial correlation between two corresponding variables. The existence of a directed edge between two nodes shows the causal relationship between the corresponding variables. Assume P is a joint probability distribution over the variables corresponding to the nodes in DAG $D = (v, \varepsilon)$. The underlying assumption for a DAG is the Markov condition over D and P [14]. D and P must satisfy the Markov condition: every variable Y_i, $i \in v$ is independent of any subset of its predecessors conditioned on a set of variables, corresponds to parents/immediate causes of node i,

$$Y_i \perp \{Y_k; i\&k \in v \setminus pa(i)\} | Y_{pa(i)},$$

where Y_k occurs before Y_i and parental set $pa(i) = pa_D(.)$ denotes the set of parents of node i relatives to the underlying structure of DAG D. For $j \in pa(i)$, we denote $j \to i$ or $Y_j \longrightarrow Y_i$.

A topology or skeleton of a DAG is a graph without direction and is obtained by identification of conditional (in)dependencies, see section "Identification the Topology of Nodes" below. Identification of directions is however a challenging problem due to the Markov equivalent property of observational data. Analysis of data in the upper granularity can identify only v-structures, two nonadjacent nodes pointing inward toward a third node. A complete assessment of directionality (i.e. statistical causal relationships) usually cannot be determined from such data alone, resulting in Markov equivalent DAGs [15,16]. Different DAGs on the same set of nodes are Markov equivalent (ME DAGs) if and only if they have the same topology and the same v-structures [17]. When the number of nodes grows, the number of ME DAGs can grow super-exponentially [18]. Complete determination of directionality over the corresponding set v is not, however, possible in most of cases.

2. The GDAG method

Identifying robust and complete directionality and showing flow of information is a difficult task, but can be facilitated by integration of different data types (i.e. granularities) where we know the direction of effect is from one granularity to the other. Assume we are seeking a DAG between two phenotypes Y_1 and Y_2. For this example, assume genome-wide information, related to the set (Y_1, Y_2) is captured in the variable X_1. Based on the results of an analysis assessing conditional independencies, we find that X_1 is correlated to Y_1 and is independent of Y_2 given Y_1, by notation $Y_2 \perp X_1 | Y_1$. Since genome sequence variation is a causal factor in phenotypic differences (and not the other way around), the direction of the effect is from X_1 to Y_1, as shown in DAG A in Fig. 1. Knowing the relationship between X_1 and Y_1 helps generate the directionality between Y_1 and Y_2 based on the property $Y_2 \perp X_1 | Y_1$, and the direction shows the flow of information is from Y_1 to Y_2, as shown in DAG B in Fig. 1. If we obtain $X_1 \perp Y_2 \;\&\; X_1 \not\perp Y_2 \,|\, Y_1$ by analysis of the data, then the direction of effect would be from Y_2 to Y_1, as shown in DAG C in Fig. 1, which represents a v-structure at Y_1.

To identify the direction among three variables in ME DAGs $Y_1 - Y_3 - Y_2$, we need to have at least two variables from the genome (i.e. a lower level of granularity, where G → P) influencing Y_1 and Y_2 or one variable from the genome influencing Y_3. By integrating multi-omics data from different granularities, we are able to derive causal inference that is less susceptible to confounding and, as a result, estimate causal networks robustly and uniquely. Partial information from a deeper granularity creates weak instrumental variables and may result in unstable structures in the upper granularity [13], and we may not be able to find a genome variable strongly associated with every phenotype under study [19]. Therefore, we go beyond inclusion of a sample of SNP marker genotypes and extract comprehensive information across the genome by application of principal component analysis (PCA) to reduce the dimensionality of the data while retaining most of the variation in the data set. Since PCA is an unsupervised approach, it avoids increasing false discovery using the same data twice. The steps of the GDAG algorithm are summarized as follows:

The GDAG Algorithm: Steps to identify a Granularity Directed Acyclic Graph (GDAG) over a set of variables of interest, Y, using data from a deeper granularity, X

1. Extract genome information by principal component analysis. Select the principal components responsible for a majority of genome variation, set X.
2. Estimate a topology over sets Y and X.[a]
3. If a variable in set X is linked to a variable in set Y, draw an arrow from the former to the latter.
4. Use the established directions from step 3, generate other directions using partial correlations recorded in step 2.[b]
5. If there is an undirected link between Ys, use rules in [20] to identify directionality.[c]

[a] Topology estimation is detailed in the following section.
[b] Presented at the beginning of this section.
[c] The supplementary information provides further details.

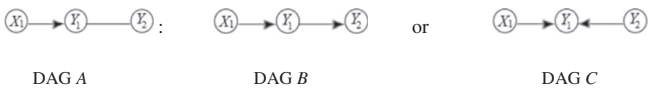

Fig. 1. DAG A is a representation of three connected variables as well as the knowledge about direction of the effect between two granularities where variable X_1 is from a deeper granularity. DAG B and DAG C represent direction identification based on analysis of data.