Bayesian Graphical Models for Gene-Environment Interaction

Paola Sebastiani[1], Harold Bae[1], Avery McIntosh[1], Stefano Monti[2]
[1] Department of Biostatistics, Boston University School of Public Health, 801 Massachusetts Avenue, Boston MA 02118
[2]Department of Medicine, Section of Computational Biomedicine, Boston University School of Medicine, 715 Albany Street, Boston MA 02118

Abstract

Bayesian directed graphical models have been described as multivariate models that accommodate many interacting variables and therefore can be useful to describe many interacting genetic and non-genetic variants and their association with a complex genetic trait. We review different concepts of interactions and show that directed graphical models can conveniently represent biological interactions. We show how reading off these relations from a directed graph uses conditional independence between variables, and we review simple algorithms to check if two variables in a directed acyclic graph are conditionally independent given a third variable.

Key Words: Bayes theorem, directed acyclic graph, graphical models, Markov property, marginal and conditional independence.

1. Introduction

The time of genome-wide association studies and simple genetic association analyses is winding down, and there is strong interest in translating the vast amount of genetic data that have been generated in the past few years into practical knowledge. Some of the goals of this new phase of research are to gain a better understanding of the biological mechanisms that link genotype to phenotype, to translate genetic findings into therapeutics, and to use genetic data for personalized medicine through the development of accurate genetic risk prediction models (Bush and Moore 2012). However, while genetics may play an important role in shaping predisposition to disease, this role may be small compared to the effect of other non-genetic risk factors. Therefore, crucial to the success of these new investigations is the ability to factor in the effect of the environment and of non-genetic risk factors in in-depth analysis of the relations between genotypes and phenotypes.

In previous work, we proposed using Bayesian directed graphical models, also known as Bayesian networks, to dissect the genetic basis of a complex phenotype (Sebastiani, Ramoni, Nolan, Baldwin and Steinberg 2005), and we used a simple Bayesian directed graphical model in the analysis of a genome-wide association study of exceptional longevity (Sebastiani, et al. 2012a). In this manuscript we show how Bayesian graphical models can be used to integrate the association of many genetic and non-genetic risk factors with one or more phenotypes and clarify the types of interactions that can be captured by this model formalism. We will show some simple examples of networks that describe additive and interacting variables.

2. Bayesian Graphical Models

A Bayesian graphical model is a vector of random variables with a joint probability distribution that factorizes according to the Markov properties represented by a directed acyclic graph. Figure 1 shows an example Bayesian graphical model with 6 variables (represented by the nodes in the graph). The nodes in the network are often called children and parents based on the edge directions: For example X_1 is a parent of X_2, and X_2 is a child of X_1. A node can have multiple parents and multiple children. The figure summarizes the local and global Markov properties represented by the directed acyclic graph (Lauritzen 1996). The local Markov property states that the probability of a variable conditioned on its parents is independent of its non-descendants. The global Markov property states that the probability of a variable conditioned on its Markov blanket is independent of all the remaining variables in the graph. Note that the directions of the edges represent the factorization of the joint probability distribution of the variables that is consistent with the local Markov property, and should not be interpreted as causal relations (Cowell, Dawid, Lauritzen and Spiegelhalter 1999). Causality in directed graphical models is discussed at length in (Pearl 2010).

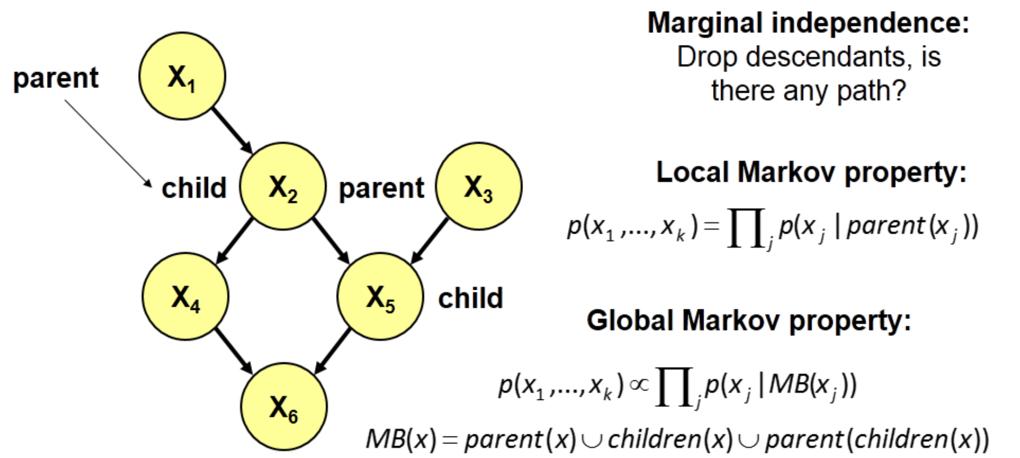
parent
X_1
child
X_2
parent
X_3
X_4
X_5
child
X_6
Marginal independence:
Drop descendants, is there any path?
Local Markov property:
$p(x_1,...,x_k) = \prod_j p(x_j \mid parent(x_j))$
Global Markov property:
$p(x_1,...,x_k) \propto \prod_j p(x_j \mid MB(x_j))$
$MB(x) = parent(x) \cup children(x) \cup parent(children(x))$

Figure 1: Example of a Bayesian directed graphical model in which nodes represent random variables, directed edges represent conditional dependencies, and the graph represents local and global Markov properties used to factorize the joint probability distribution of the variables. The sets of parents and children of a node X are denoted by parent(X) and children(X). MB(X) denotes the Markov blanket of a node that comprises all parents, children, and parents of the children nodes.

Bayesian graphical models have been described as models that show the effect of many interacting variables (Friedman, Linial, Nachman and Pe'er 2000, Needham, Bradford, Bulpitt and Westhead 2007), but the word interaction is often used to describe mutual associations and does not necessarily match the notion of statistical interaction, or the various definitions of interactions used in epidemiology and public health (Clayton 2009, Thomas 2010) (See Table 1 for a summary). Our goal is to use graphical models to describe the joint effect of many genetic and non-genetic factors on a trait. Therefore it is important to clarify the type of interactions that can be represented in a graphical model, and to map prototypical graphs to standard epidemiology concepts of association, confounding, mediation and interaction. Here, we will focus on interaction, and the detection of interacting variables in a graphical model.

3. Notions of Interaction

Statistical interaction is model dependent: in a parametric statistical model

$$Y = f(X, \theta, \varepsilon)$$

describing the effects of covariates X on an outcome Y through parameters θ and errors ε, interaction terms are those parameters that describe departure from a model with only main effects. This definition makes interaction dependent on the measurement scale. For example, when the parametric model is linear regression:

$$Y = \theta_0 + \sum_j x_j \theta_j + \sum_{h,k} x_h x_k \theta_{hk} + \varepsilon$$

the interaction terms θ_{hk} represent the lack of fit of the model with only main effects θ_h in the linear scale (*additive interaction*), and interaction can disappear after transformation of the data. In a logistic regression model for a binary outcome Y:

$$\text{logit}(P(Y=1 \mid X)) = \log \frac{P(Y=1 \mid X)}{1 - P(Y=1 \mid X)} = \theta_0 + \sum_j x_j \theta_j + \sum_{h,k} x_h x_k \theta_{hk}$$

the interaction terms θ_{hk} represent the lack of fit of the model with only main effects θ_h in the logistic scale (*multiplicative interaction*). Interaction in the linear scale does not imply interaction in the multiplicative scale and vice versa. This issue has received substantial attention in the past (Jewell 2003).

Table 1: Definition of Interaction in Epidemiology	
Statistical interaction	Departure from a model with only main effects. This definition is model based, and is linked to the parametric model used for analysis
Biological interaction	A biologic response produced by the simultaneous exposure to two or more agents that differs from the combined response to the agents when applied independently.
Effect modification	The effect of one factor changes as the other factor changes
Quantitative interaction	The effect of one factor changes as the other factor changes (effect modification) but the direction of effects does not.
Qualitative interaction	The effect of one factor changes as the other factor changes (effect modification) and the direction of effects also changes.

Definitions of interaction used in genetic epidemiology are often more qualitative in nature, and often invoke the concept of "biological interaction". See Table 1 and references (Rothman, Greenland and Walker 1980, Thomas 2010).

4. Probabilistic Interactions in Graphical Models

Consider a simple example with 3 variables: G is a discrete variable that represents the alleles of a gene (e.g., G=AA, AB, or BB); E is a binary variable that represents exposure/no exposure to a non-genetic risk factor; and P is a binary variable that represents presence/absence of a phenotype. Suppose that the probabilistic relations between the 3 variables are represented by a directed graph. We provide a definition of interaction between *G* and *E* in their association with *P*, and show examples of graphical models for interaction between *G* and *E*.

As in graphical log-linear models (Whittaker 1990, Lauritzen 1996) we can say that *G* and *E* do not interact in their association with *P* if the joint probability distribution of the three variables factorizes as:

$$p(G,E,P) = f(G,P)h(E,P)$$

where *f(·)* and *h(·)* are two functions that take positive values. The factorization of the joint probability distribution into a product of two factors in which *G* and *E* never appear together is equivalent to conditional independence of *G* and *E* given *P* (written $G \wedge E|P$ (Whittaker 1990). The factorization $p(G,E,P) = f(G,P)h(E,P)$ can be interpreted as the property that the joint effect of *G* and *E* on *P* (measured by the joint probability of *G*, *E*, and *P*) is equivalent to the combined actions of the individual factors (measured by the product of *f(G, P)* and *h(E,P)*). Therefore, we can use conditional independence of two variables given a third variable to denote a lack of biological interactions of the two variables in their joint association with the third variable.

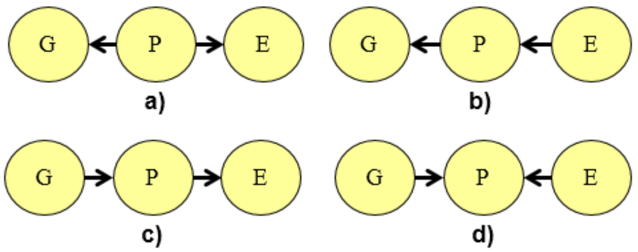

Figure 2: Four directed graphical models that display the mutual relation between gene alleles (node G), environmental exposure (node E), and expression of a phenotype (node P). The Markov property represented by the graph a) is conditional independence of G and E given P. Graphs b) and c) show conditional independence of G and E given P. Graph d) shows conditional dependence of G and E given P. The first three graphs show pairwise associations, and lack of interaction between G and E in their association with P. The graph in d) shows joint dependence of P on both G and E, and interaction between G and E in their association with P.

Figure 2 shows four directed graphical models with examples of no interactions between G and E in their association with P (graphs a, b and c), and an example of interaction between G and E (graph d). The different graphs represent the probability distributions that could be estimated from data, conditionally on the study design, rather than causal relations. As an example, the graphical structure in a) is appropriate to represents the associations between gene, environment and a phenotype that can be estimated from a case-control study (Sebastiani, et al. 2012a). The graphical structure in d) is appropriate for modeling the dependence of P on G and E using data from a prospective study.

The model represented by graph a) has been used to represent the joint effect of many genes on a complex trait, although it has been erroneously described as a model with many interacting genes (Okser, et al. 2010). The graphical structure does not represent interactions between the children nodes of P but only pairwise associations. We showed that prediction of a complex genetic trait based on such a graphical model is equivalent to prediction based on a logistic regression model with the genetic effects of multiple genes summarized into a genetic risk score. We also showed how this graphical model can be extended to include “pairwise” interactions. See the discussion in (Sebastiani, Solovieff and Sun 2012b) for additional details. The simple structure can be extended to represent pleiotropic genes (i.e. genes that are associated with different, apparently unrelated phenotypes). Hartley et al (Hartley, Monti, Liu, Steinberg and Sebastiani 2012) developed a method to infer these extended structure from genome-wide genotype data and evaluated the approach in simulated and real data. The method is implemented in the computer program PleioGrip (Hartley and Sebastiani 2013) that is freely available from the Boston University digital library (http://hdl.handle.net/2144/4367).

5. Ascertaining Probabilistic Interactions in Graphical Models

The concept of probabilistic interaction requires three actors: the two agents and a response of interest. Interaction (or lack of it) between the two agents has to be interpreted relative to the association of the two agents with the response. In a Bayesian graphical model with many variables, say $G_1,...,G_g, E_1,...,E_e, P_1,...,P_k$, two genes may interact in their association with one of the possible phenotypes of interest, and may not interact in their association with a different phenotype. Although the ascertainment of interacting variables essentially requires determining whether the variables are conditionally independent given the response, establishing conditional independence that are not directly represented by the local and global Markov properties may be daunting. Several algorithms have been proposed for this purpose. One approach uses the concept of d-separation, where “d” stands for “directional,” see (Cowell, Dawid, Lauritzen and and Spiegelhalter 1999). There are applets over the WWW that can verify whether two variables in a graph are conditionally independent given a third variable by checking for d-separability (see for example: http://www.phil.cmu.edu/~wimberly/dsep/dSep.html). An alternative algorithm requires some simple graph manipulations described in the table below. The proof that the graph manipulation can actually detect conditional independence can be found in (Barber 2011). Figure 3 shows an example.

Table 2: Are X and Y conditionally independent given Z?

Step 1: select ancestral graph	Remove all nodes in the graph that are not X, Y, Z, and their ancestors. The ancestors of X, Y and Z are the nodes in the graph from which there is a directed path to X, Y and Z.
Step 2: moralization of the ancestral graph	Add undirected edges between any pair of nodes with a common child, and drop the direction of the remaining directed edges.
Step 3: pruning the ancestral graph	Remove all edges from the node Z
Step 4: separation	In the remaining graph, look for a path linking X and Y. If there is no such a path, then X and Y are conditionally independent of Z

The algorithm can also be applied to check whether two variables are marginally independent by letting variable Z = ∅. For example, one can easily show that G_1 and G_2 are marginally independent.

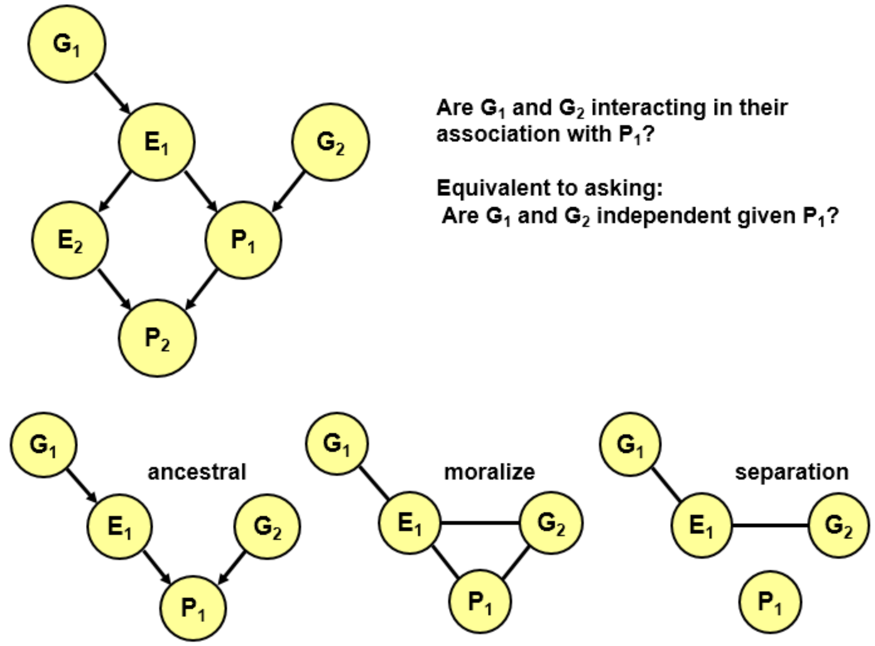
G₁
E₁
G₂
E₂
P₁
P₂
Are G₁ and G₂ interacting in their association with P₁?
Equivalent to asking:
Are G₁ and G₂ independent given P₁?
ancestral
moralize
separation

Figure 3: Application of the algorithm in Table 2 to ascertain whether G_1 and G_2 interact in their association with P_1. Since there is a path between G_1 and G_2 in the pruned moralized ancestral graph, then the two variables interact. However, note that the two nodes are marginally independent.

6. Conclusions

Directed graphical models provide a flexible model formalism to describe complex biological interactions between genetic and non-genetic factors and their effects on multiple phenotypes. Biological interactions are described using probability to define the joint effects of many variables and the combined individual effects. Interaction is described by conditional dependency and lack of interaction is described by conditional independence. The same approach based on conditional independence can be used to describe mediation of genetic and environmental effects. Tools to easily ascertain conditional independence of variables in a network that are not captured by the local and global Markov properties are needed to easily characterize the set of interacting variables. In addition, simple summaries are also needed to quantify the effect of interactions.

Acknowledgements

This work was funded by the National Institute on Aging (NIA U19-AG023122), the National Heart Lung Blood Institute (R21HL114237), and the National Institute of General Medicine (T32 GM74905).

References

Barber, D. (2011), *Bayesian Reasoning and Machine Learning*, Cambridge ; New York: Cambridge University Press.

Bush, W. S., and Moore, J. H. (2012), "Chapter 11: Genome-Wide Association Studies," *PLoS Comput Biol*, 8, e1002822.

Clayton, D. G. (2009), "Prediction and Interaction in Complex Disease Genetics: Experience in Type 1 Diabetes," *PLoS Genet*, 5, e1000540.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and and Spiegelhalter, D. J. (1999), *Probabilistic Networks and Expert Systems.*, New York: Springer Verlag.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. (1999), *Probabilistic Networks and Expert Systems.*, New York:: Springer Verlag.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000), "Using Bayesian Networks to Analyze Expression Data," *J Comput Biol*, 7, 601-620.

Hartley, S. W., Monti, S., Liu, C. T., Steinberg, M. H., and Sebastiani, P. (2012), "Bayesian Methods for Multivariate Modeling of Pleiotropic Snp Associations and Genetic Risk Prediction," *Front Genet*, 3, 176.

Hartley, S. W., and Sebastiani, P. (2013), "Pleiogrip: Genetic Risk Prediction with Pleiotropy," *Bioinformatics*, 29, 1086-1088.

Jewell, R. (2003), *Statistics for Epidemiology*, Boca Raton: CRC/Chapman and Hall.

Lauritzen, S. L. (1996), *Graphical Models*, Oxford: Clarendon Press.

Needham, C. J., Bradford, J. R., Bulpitt, A. J., and Westhead, D. R. (2007), "A Primer on Learning in Bayesian Networks for Computational Biology," *PLoS Comput Biol*, 3, e129.

Okser, S., et al. (2010), "Genetic Variants and Their Interactions in the Prediction of Increased Pre-Clinical Carotid Atherosclerosis: The Cardiovascular Risk in Young Finns Study," *PLoS Genet*, 6.

Pearl, J. (2010), "An Introduction to Causal Inference," *Int J Biostat*, 6, Article 7.

Rothman, K. J., Greenland, S., and Walker, A. M. (1980), "Concepts of Interaction," *Am J Epidemiol*, 112, 467-470.

Sebastiani, P., Ramoni, M. F., Nolan, V., Baldwin, C. T., and Steinberg, M. H. (2005), "Genetic Dissection and Prognostic Modeling of Overt Stroke in Sickle Cell Anemia," *Nat Genet*, 37, 435-440.

Sebastiani, P., et al. (2012a), "Genetic Signatures of Exceptional Longevity in Humans," *PLoS ONE*, 7, e29848.

Sebastiani, P., Solovieff, N., and Sun, J. X. (2012b), "Naive Bayesian Classifier and Genetic Risk Score for Genetic Risk Prediction of a Categorical Trait: Not So Different after All!," *Front Genet*, 3, 26.

Thomas, D. (2010), "Gene-Environment-Wide Association Studies: Emerging Approaches," *Nat Rev Genet*, 11, 259-272.

Whittaker, J. (1990), *Graphical Models in Applied Multivariate Statistics.*, New York: John Wiley & Sons.