

A Critical Assessment of the Bayesian Approach to Causal Network Inference

Jeremy Purvis

March 20, 2006

A recent paper appearing in *Science* (1) describes a noninvasive, computational approach to elucidating biological network pathways. The method, known as Bayesian network inference, constructs probabilistic graphical models of the dependency relationships among multivariate data (for an overview of Bayesian network analysis, see (2)). In this study, Sachs *et al.* measured the abundance of 11 proteins and lipids by flow cytometry, imposing nine different perturbation conditions, to generate a hypothetical network model of downstream signaling in activated human T cells. Statistically speaking, their Bayesian network is simply a collection of multivariate joint probability distributions assembled into graphical form. While these graphs offer an attractive way to represent complex dependencies among a set of variables, any implications about the underlying biological processes should be regarded with caution.

Data Collection and Experimental Methods An impressive aspect of this study was the volume and quality of measurements. Multicolor flow cytometry (3) quantifies the abundance of up to 11 signaling molecules in individual cells simultaneously and *in vivo*. By fluorescently labeling the compounds of interest, the instantaneous levels of each compound were recorded independently, on a cell-by-cell basis. The latter feature proved to be invaluable for extracting the most correlative information from the data. Sachs demonstrates this by noting that when abundance levels were averaged across cell populations, much of the predicted network detail was obscured (loss of five edges compared to graph results from individual cell analysis).

Less impressive, yet still informative, was the method of intervention. Although the perturbations imposed on the system were modeled as *ideal* (affecting a single known molecule), only seven out of the nine reagents were

reported as ‘specific’ (see Table 1 in (1)). Of these seven, only one has a single molecular target. At best, this technical compromise might necessitate a larger data set in order to achieve the same level of statistical confidence. At worst, nonideal interventions could perturb the network in unknown ways and produce misleading results. We must assume the interventions used in this study are well characterized and do not exhibit unknown behaviors.

Hypothesis Discovery Bearing this wealth of multivariate data, the authors sought to construct a graphical depiction of intermolecular dependencies that best explained the correlations observed from flow cytometry. The general idea is that undirected graph structure (representing strictly correlative relationships) can be deduced from the unperturbed measurements, while controlled interventions to the system will cause a disruption in these correlations in such a way that directed edges (representing causal relationships) on the graph may be inferred. Their exact method of inference, as described in (4, 5), operates under a few key assumptions. Briefly, a Bayesian network over a set \mathbf{X} (in this work, the signaling proteins) consists of a directed acyclic graph (DAG) G whose vertices correspond to the random variables in \mathbf{X} , and parameters $\Theta_{\mathbf{m}}$ that describe the conditional distribution for each variable given its immediate parents. By discretizing all possible graph structures, the network space was searched to generate hypothesized dependencies for bayesian analysis. Finally, the posterior probability of each dependency was computed as the sum of the posterior probabilities of all DAGs in which this dependency existed.

There is one glaring difficulty with this approach: the number of graphs increases superexponentially with the number of variables, which makes exhaustive enumeration of all possible graph structures impractical. To deal with this problem, the authors sampled “representative” networks from the complete graph space using a non-parametric bootstrap method (6). These samples were then used to compile a list of frequently occurring *features*. A feature is a property such as “ $X \rightarrow Y$ is in the network” or “given Y , Z is conditionally independent of X ” that can be compared among hypothetical networks.

To summarize, high-scoring graphs were found by sampling a fraction of the hypothetical network space, discretizing each network into small subnetwork building blocks, looking for commonly occurring features, and compiling them into a tractable set of chimeric network structures. At this point, one may reasonably ask whether there are plausible network structures that were either undetected by the bootstrap method or composed of too many

‘unrepresentative’ features. In either case, they would fail to appear as hypotheses for subsequent Bayesian analysis. In a separate study, Friedman *et al.* showed that this bootstrap method has a high rate of false negatives (Friedman, 2000). Thus, one cannot be confident that these networks uniquely depict the underlying physical dependencies, even if high-scoring models are found which explain the data reasonably well.

Even if data collection and hypothesis generation were acceptably carried out, the final goal of constructing a meaningful causal graph presents a serious challenge. In determining the most probable network structures, there are two priors to consider: parameters and graph structure. Both are central to Bayesian network inference and must be carefully selected. In particular, the choice of distributions for these priors should be judged according to their physical reasonability. In the next two sections, I examine the derivation of these priors and note any discrepancies between the authors’ choices and the biological entities they represent.

Parameter Priors As described in (7), a continuous vector-valued parameter $\Theta_{\mathbf{m}}$ represents the conditional probability that maximizes the likelihood that the data D was generated by the Bayesian network $\mathbf{B}^* = (\mathbf{G}, \Theta_{\mathbf{m}})$. Under the bayesian model, each parameter θ_i has a Dirichlet distribution and depends only on the structure of the model that is local to the parameter (i.e., X_i and its parents). We can interpret this, in physical terms, to mean that the dependency relationship between one signaling molecule and another is contingent on their ability to interact. If we accept that the physical basis for dependency is determined by structural properties of molecules (size, binding affinities, etc.), then this is a reasonable choice for parameter priors. The relationships may be directed, implying causation, or undirected, implying correlation.

Model Structure Priors As stated above, the prior distribution of model parameters depends entirely on local graph structure, which itself is constrained to a well-defined probability distribution and characteristic structural properties. It is here that we encounter the most severe weakness in the model. Bayesian network inference allows for multiple simultaneous interactions to be modeled, but under the assumption of conditional independency. This assumption states that in the directed structure $X \rightarrow Y \rightarrow Z$ (Z de-

depends on Y and Y depends on Z), we cannot model a direct effect of X on Z . More precisely, X is conditionally independent of Z given \mathbf{Y} if the probability distribution of X conditioned on both \mathbf{Y} and Z is the same as the probability distribution of X conditioned only on \mathbf{Y} :

$$P(X|\mathbf{Y}, \mathbf{Z}) = \mathbf{P}(\mathbf{X}|\mathbf{Y}) \tag{1}$$

Is this a biologically tenable constraint? We need only to look at Figure 2 of (1) for a contradicting example of this type of dependency. Here, the causal relationship between three enzymes contains the substructure $\text{PKC} \rightarrow \text{RAS} \rightarrow \text{Raf}$ and also $\text{PKC} \rightarrow \text{Raf}$, which is strictly forbidden under the current Bayesian model. It should be immediately clear from this example that, within the broad context of complex biological networks, we cannot rule out the possibility that a molecule depends on both its parents *and* its grandparents (8). The inference method as proposed by Sachs may serve as a good first-order approximation, but improvements to the model structure must be devised in order to capture more realistic biological relationships.

Certainly, Sach’s model allows for manageable discretization of the network space. But we have not yet seen how each network is distributed, i.e., what are the prior probabilities for each network in the hypothesis space? Here Sachs used the standard Bayesian scoring metric (9) that rewards models for being simple (i.e., fewer arcs). From a modeling point of view, this is a logical decision since more complicated networks have an inherent advantage in achieving a better likelihood score, having more variables, parameters, etc. From a biologist’s perspective, however, this choice is slightly troubling. Living systems are notorious for functioning in ways that are neither simple nor optimal (10).

It is worth noting that the signaling network investigated in this paper represents decades of reductionist-approach experiments and hundreds of published studies. The prospect of high-throughput measurements and clever statistical methods replacing traditional approaches to network modeling as the dominant method of discovery is unlikely, since the success of the Bayesian method requires extensive knowledge of the system at hand in order to be effective. For example, to model the effect of a specific intervention, we must be able to identify the exact target of perturbation. Knowing

this target *a priori* means that it has already been identified in relation to the network, and thus it is trivial to propose it as a novel component.

Then there is the issue of graph structure. While the authors stress that their graphs represent “a statistical dependence of relative influence” and “do not necessarily imply a direct functional or physical connection,” one is prone to discern these graphs as if they were actual depictions of the physical interactions among protein signals. Directed acyclic graphs are an oversimplification of biological network structure and limit the selection of hypothetical graphs to a inflexible model with predefined structural limitations. That pathways cannot contain cyclic structures, for example, precludes a multitude of well-documented biological pathways (e.g., feedback loops) from being accurately modeled. Among the remaining networks that *can* be captured by DAGs, the range of hypothesized structures is limited to whatever sampling procedure is used to probe the network space, which can be substantial. We have seen that these sampling methods have their own set of problems. Consider this frustrating result: in order to calculate the probability that X causes Y , we must consider this causal relationship in isolation from the full network model. If the feature ‘ X causes Y ’ exists in the true physical network, but does not appear in enough sampled hypotheses, it will not become part of the final predicted network. Truly, it is a risky business probing biological signals and forcing them into structurally rigid graphical representations.

References

1. Sachs, K., O. Perez, D. Pe’er, D. Lauffenburger, and G. P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science* 308:523-529. (2005).
2. Pe’er, D. Bayesian Network Analysis of Signaling Networks: A Primer. *Science STKE* 2005:pl4 (2005).
3. Perez, O. D., G. P. Nolan. *Nature Biotechnology* 20:155 (2002).
4. Pe’er, D., A. Regev, G. Elidan, and N. Friedman. Inferring subnetworks from perturbed expression profiles. *Bioinformatics*. 17:S215-S224. (2001).
5. Yoo, C. a. C. G. F. *Uncertainty in Artificial Intelligence*. pp. 116-125. (1999).

6. Friedman, N., M. Linial, I. Nachman, and D. Pe'er. Using Bayesian networks to analyze expression data. *J. Comp. Biol.* 7:601-620. (2000).
7. Heckerman, D., C. Meek, and G. Cooper. A Bayesian approach to causal discovery. *Computation, Causation, and Discovery*. pp.141-166. C.
8. Sachs, K., D. Gifford, T. Jaakkola, P. Sorger, and D. Lauffenburger. Bayesian Network Approach to Cell Signaling Pathway Model. *Science STKE* 2002:pe38 (2002).
9. Heckerman, D. A Tutorial on Learning with Bayesian Networks. *Microsoft Research MSR-TR-95-06*. (1995).
10. Marinissen, M. J. and J. S. Gutkind. G-protein-coupled receptors and signaling networks: emerging paradigms. *Trends Pharmacol. Sci.*22:368-375. (2001).
11. Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Francisco. (1988) Glymour, G. F. Cooper, Eds. MIT Press, Cambridge, MA. (1999).