Network cleanup

Babak Alipanahi & Brendan J Frey

The erroneous links in networks inferred from data can be efficiently eliminated under certain conditions.

Networks offer an alluring simplicity for representing complex systems of interacting parts[1]. But when networks are constructed from biological data through statistical inference, it is often unclear how faithfully they represent the real systems. In many cases, true links between nodes are obscured by a sea of noise in the form of erroneous links. In this issue, two studies by Feizi *et al.*[2] and Barzel *et al.*[3] describe efficient, easily implemented methods for identifying and removing erroneous links, thereby producing more accurate networks. Both papers demonstrate the application of their techniques to large-scale practical problems, such as the DREAM5 gene regulatory network inference challenge[4]. In addition, Feizi *et al.*[2] explore other applications by analyzing networks of interacting residues in protein structures and social networks of scientists.

The problem of erroneous links in inferred networks was described in 1921 by the geneticist—and founder of the field of network inference—Sewall Wright, who said, "The degree of correlation between two variables can be calculated by well-known methods, but when it is found it gives merely the resultant of all connecting paths of influence"[5]. As an example, suppose one gene directly controls a second gene, which in turn directly controls a third gene. Correlation analysis will erroneously indicate that the first gene directly influences the third gene. Other methods for linking variables, such as mutual information and distance correlation, are limited by the same problem. The goal of network inference is to identify the direct links and their strengths while suppressing the indirect, or transitive, associations. This problem is difficult because experimental techniques often cannot distinguish between direct and indirect effects.

Feizi *et al.*[2] and Barzel *et al.*[3] tackle the problem of network inference from conceptually quite different starting points. Feizi *et al.*[2] view the measured correlations as a consequence of flows along the edges in the true network. In contrast, Barzel *et al.*[3] treat the measured correlations as small perturbations that result from adding up the small perturbations induced along edges in the true network. In both cases, the authors turn the seemingly intractable problem of network inference into easily implemented algorithms that invert these processes to obtain the true network from the measured correlations.

To illustrate the methods, we have implemented them, applied them to a 'toy' gene regulatory network problem and compared the results with the known true network of interactions (**Fig. 1a**).

Both methods account for how the total, measured effect of a source node on a target node is mediated by the direct neighbors of the target (**Fig. 1b**). In addition, if the source is directly connected to the target node, that direct effect is accounted for too. This accounting is not quite correct because of loops, but it is reasonably accurate if the strengths of indirect effects decay substantially as they propagate around the loops.

As a concrete example, consider the network shown in **Figure 1a**, in which circles represent genes and links between genes

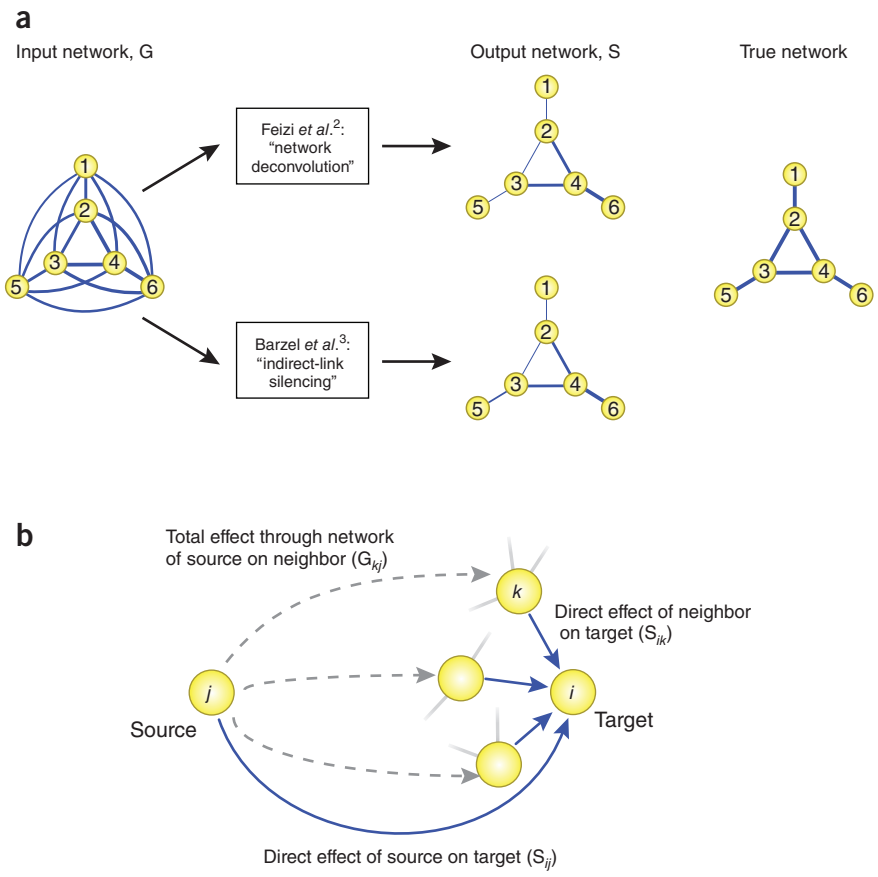

Figure 1 Cleaning up networks. (**a**) An example of how the erroneous links in an input network (G) derived using pairwise correlation can be suppressed using the methods described by Feizi *et al.*[2] (called network deconvolution) and Barzel *et al.*[3] (called silencing). The network of true relationships (S) is shown on the right. Edge thickness corresponds to the connection strength. (**b**) How the network of total, measured effects *G* is related to the unknown network of direct effects, *S*. The total effect of node *j* on node *i*, G_{ij}, can be obtained by adding the direct effect of *j* on *i* (S_{ij}) to the sum of the direct effects of each neighbor *k* on *i* (S_{ik}) multiplied by the total effect of *j* on *k* in the appropriate subnetwork (G_{kj}).

Babak Alipanahi and Brendan J. Frey are with the Departments of Electrical and Computer Engineering and the Donnelly Centre for Cellular and Biomolecular Research at the University of Toronto, Toronto, Ontario, Canada.
e-mail: babak@psi.toronto.edu or frey@psi.toronto.edu

Box 1 Mathematical basis of suppressing indirect effects in networks

A network can be represented as a matrix with each node corresponding to one row and column and edges represented as entries in the matrix. This allows mathematical analysis of the network using tools from linear algebra. Using notation from Barzel *et al.*[3], S is the true matrix of direct associations, where each entry S_{ij} is the rate of change of variable i with respect to variable j, assuming all other variables are held constant. The correlation matrix G is constructed from measurements of the total effect of each variable on every other variable, and it can be related to S as follows (also see **Fig. 1b**). The total effect of j on i, G_{ij}, can be obtained by summing up the effects mediated through the direct neighbors of i in the true network, S. Each direct neighbor, k, is connected to j through a subnetwork whose total effect is approximately equal to G_{kj}, giving a relationship used in both studies

$$G_{ij} = S_{ij} + \Sigma_{k:k\neq j} G_{kj} S_{ik}$$

Network inference entails finding a network S that satisfies the above equation for all $i \neq j$. Both Barzel *et al.*[3] and Feizi *et al.*[2] describe unique, but approximate, closed form solutions for S in terms of G.

It turns out that both approaches are related to the method of partial correlation, which is the correlation between the residual errors for two variables when they are linearly predicted from all other variables[5,6]. If the input G is a correlation matrix and $P = G^{-1}$, then for $i \neq j$, Feizi *et al.*'s solution has the form $S_{ij} = -P_{ij}$, Barzel *et al.*'s solution has the form $S_{ij} = -(1-\Sigma_{k\neq i} G^2_{ik}) P_{ij}$ and the partial correlation is $S_{ij} = -P_{ij}/\sqrt{(P_{ii}\, P_{jj})}$. So, the different methods scale the inverse correlation matrix differently. The proposed methods can be applied to other types of input matrix, such as one derived using mutual information.

represent regulatory effects. These effects may be mediated by different molecular mechanisms, such as direct binding of a transcription factor to the promoter of a target gene or phosphorylation of a target protein by a kinase. In the example, the effect of gene 1 on gene 4 can be broken down into the sum of three terms: (i) the direct effect of gene 2 on gene 4 multiplied by the total effect of gene 1 on gene 2, (ii) the direct effect of gene 3 on gene 4 multiplied by the total effect of gene 1 on gene 3 and (iii) the direct effect of gene 6 on gene 4 multiplied by the total effect of gene 1 on gene 6. If there were a direct link between genes 1 and 4, it would be included in the sum.

We found that the linear algebra used to derive the methods of Feizi *et al.*[2] and Barzel *et al.*[3] can be reformulated so as to obtain insights into the similarities and differences between the methods and into their relationship to previous work on partial correlation[5,6]. The methods differ in how the inferred, direct effect associated with each link is scaled (**Box 1**). The partial correlation scales each link according to its source and target. The method of Barzel *et al.*[3] scales the strength of each link according to its target, whereas the method of Feizi *et al.*[2] does not scale the links. The strength of each link is used as a proxy for the significance of the association, so links with small strengths are discarded. The scaling factors for partial correlation are determined from the inverse of the correlation matrix, and the scaling factors used by Barzel *et al.*[3] can be computed directly from the correlation matrix. As a consequence of these differences, the methods can output quite different solutions. Further work is required to determine the relative advantages of each approach.

How well can we expect these methods, and others based on partial correlation, to work, and what are their limitations? If the physical system being analyzed is linear and all variables are observed, a nonzero partial correlation between two variables indicates that they are dependent when all other variables are held constant[8]. Whether or not this implies a physically relevant link should be answered in the context of the problem of causal interpretations of networks, where an intervention-based approach may be needed to identify causal links with some guarantees[9].

Both proposed methods, and related ones, will be highly sensitive to missing variables because of Simpson's paradox, which states that a statistical relationship between two variables may be reversed or eliminated when additional variables are included[10]. For example, one may observe that the expression of gene 1 is negatively correlated with that of gene 2 but that after the expression value is adjusted to account for the cell cycle, gene 1 is positively correlated with gene 2. Yet, when additional adjustments are made for chromatin structure, the genes may no longer appear correlated. Another limitation is that the methods do not output confidence levels for links or, more generally, distributions over possible networks, which could be useful for downstream analyses.

The methods discussed here, and related ones, should be viewed as exploratory tools that can be applied to guide research rather than to draw scientific conclusions. As shown by the example in **Figure 1a**, the methods can be used to direct attention to links that are more likely to have direct, possibly causal effects, so that a more careful analysis, possibly including additional data, can be conducted. For example, by using these methods to infer networks of RNA-binding proteins[7], it is possible to identify potential regulators and incorporate them into an accurate regulatory model of splicing[11]. There are other frameworks that can be used to infer networks, including Bayesian reasoning, which would generate a distribution of possible networks, and information theory, which provides guarantees based on asymptotic analysis[12]. Ultimately, the processes of link hypothesis generation, causal testing and network refinement can be formalized using structural equation modeling[5,9] and structural causal modeling[9], two techniques from the statistics and artificial intelligence communities.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

1. Lopes, C.T. *et al. Bioinformatics* **26**, 2347–2348 (2010).
2. Feizi, S. *et al. Nat. Biotechnol.* **31**, 726–733 (2013).
3. Barzel, B. *et al. Nat. Biotechnol.* **31**, 720–725 (2013).
4. Marbach, D. *et al. Nat. Methods* **9**, 796–804 (2012).
5. Wright, S. *J. Agric. Res.* **20**, 557–585 (1921).
6. Raveh, A. *Am. Stat.* **39**, 39–42 (1985).
7. Ray, D. *et al. Nature* **499**, 172–177 (2013).
8. Pearl, J. *Sociol. Methods Res.* **27**, 226–284 (1998).
9. Pearl, J. *Stat. Surv.* **3**, 96–146 (2009).
10. Simpson, E.H. *J. Roy. Stat. Soc. B* **13**, 238–241 (1951).
11. Barash, Y. *et al. Nature* **465**, 53–59 (2010).
12. Santhanam, N.P. *et al. IEEE Trans. Inf. Theory* **58**, 4117–4134 (2012).