The path not taken

If known pathways represent preferred paths through global protein networks, large-scale experiments should focus on how to recognize these paths among all of the alternatives.

Edward M. Marcotte

> Two roads diverged in a yellow wood,
> And sorry I could not travel both
> And be one traveler, long I stood...
> *Robert Frost, 1915*

Molecular biologists face a curious quandary: after decades of deriving genetic and biochemical pathways, it is not clear if these traditional models are compatible with the emerging picture of extensively interconnected cellular networks. These model pathways are constructed to show the cellular machinery deemed relevant to a particular task, such as signaling or metabolic flux. In contrast, cellular networks are defined globally from large-scale protein interaction and gene expression measurements. So, the pathway is a model; the network, an experimental observation. The source of the quandary is the multitude of alternative pathways suggested by the cellular networks. Currently, the preference for a given path through the network is poorly understood, and the ability of our model pathways to accurately portray flux through complex networks is an open question.

This interplay between network and model is the subject of an article in a recent issue of *Science*. In the paper, Idekar *et al.*[1] refine a model of the galactose utilization pathway of yeast by knocking out each gene in the pathway, then measuring gene and protein expression with DNA microarrays and mass spectrometry, respectively. The resulting ream of data is integrated with all known protein–protein and protein–DNA interactions relevant to the galactose pathway; all of these data are then used to iteratively improve the model of galactose utilization.

This process allows the authors to implicate several new proteins in galactose utilization. As is common with these types of functional inferences, the exact roles the new proteins may play are unclear; instead the new proteins show a statistical association with known proteins of the galactose utilization pathway. As well as adding new proteins to the pathway, the authors suggest a potential feedback mechanism by which galactose-1-phosphate levels could regulate the pathway. Interestingly, they confirm a second observation[2] that levels of messenger RNA (mRNA) and protein expression are only poorly correlated—thus necessitating the collection of both types of expression data to fully characterize the system.

Idekar *et al.* also define a four-step process for integrating data from large-scale experiments into an intelligible biological model. The steps can be paraphrased as the following: first, define which genes are known to be in the pathway; second, perturb each gene in the pathway and measure the corresponding global cellular response; third, integrate the measured mRNA and protein expression data with all previously known protein–protein and protein–DNA interactions; and fourth, try to explain why the model deviates from the observations, then test specific hypotheses with additional perturbations.

However, all of this work to improve a simplified model begs the question: why not abandon the old representation of pathways and instead work directly with the networks? This requires a certain philosophical adjustment. We construct such simplified pathways because we find cellular processes too complex to comprehend in their entirety. Unfortunately, networks such as the protein interaction network pictured at the top of Figure 1 represent, to our best current understanding, the cellular reality. Can discrete pathways be consistent with these networks? For instance, when the proteins of the pathway at the bottom of Figure 1 are mapped into the cellular protein interaction network, many additional interactions are seen for each of the proteins in the pathway. At first glance, this abundance of alternative interactions suggests that a pathway representation is hopelessly inadequate.

An intriguing alternative exists, though: known pathways may represent the preferred paths for flux through cellular networks. Several lines of evidence support this notion. At the very least, the pathways are legitimate,

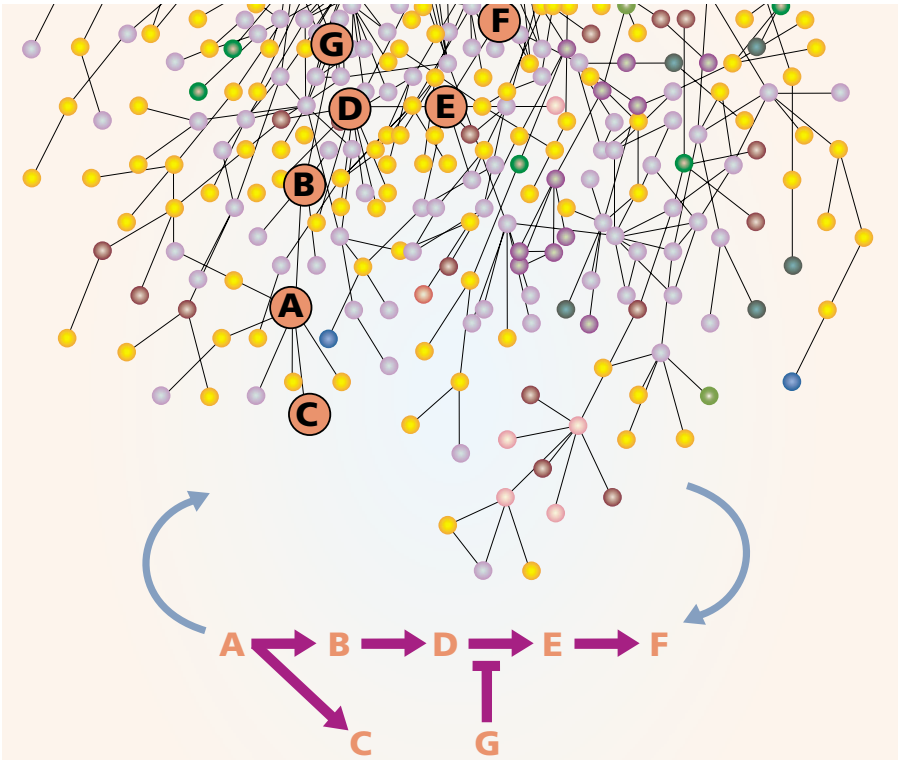

Figure 1. How do our traditional genetic and biochemical pathways measure up to the observed global co-expression and interaction networks? Mapping the genes from a defined genetic network (A–G, bottom) into a global network[7] (red circles labeled A–G, top) reveals many additional interactions and genetic relationships for each component of the genetic network. Two possible interpretations present themselves: either the defined genetic network is a dramatic oversimplification or it represents a preferred path through the global network.

Edward M. Marcotte is assistant professor in the Department of Chemistry and Biochemistry & Institute of Cell and Molecular Biology, 2500 Speedway, University of Texas, Austin, TX 78712 (marcotte@icmb.utexas.edu).

experimentally observed paths through the network. This is an important point, as it is not unreasonable that many paths through a cellular network may be energetically unfavorable and therefore not traversable. Also, a number of pathways have been verified by *in vivo* measurements of metabolic turnover using such methods as whole-cell nuclear magnetic resonance (NMR; e.g., as in ref. 3). Probabilistically speaking, we would expect preferred paths to be discovered by scientists more often than rare paths, suggesting that well-studied pathways are more commonly traversed paths in the network. Finally, metabolic pathways operate along free energy gradients, which we would not expect from random paths through the network.

If known pathways do represent preferred paths through global protein networks, the goal then becomes how to recognize these paths among all of the alternatives. To make this idea concrete, consider the large data sets collected by Idekar *et al.* Given these data, could one have predicted the galactose utilization pathway? Currently, the answer is no. This difficulty in recognizing such preferred paths *de novo* probably stems from systematic absence of important data, such as small-molecule concentrations or protein–metabolite interactions.

Although Idekar *et al.* gather quite a lot of data about concentrations of intracellular species, one class of molecules is conspicuously absent: the metabolites. Traditional biochemical pathways are defined as series of successive modifications to small molecules. To extract these sorts of pathways from global networks will require knowledge of the metabolite concentrations and protein–metabolite interactions. In a step in that direction, techniques have recently been developed to measure cellular concentrations of several hundred small molecules using gas chromatography/mass spectrometry of cell lysates[4]. We can anticipate that measuring metabolite levels along with gene and protein levels would greatly expand our ability to infer metabolic pathways.

So, these considerations of model pathways and metabolite profiling suggest a modification to the four-step strategy outlined by Idekar *et al*. It seems not unreasonable to perform such an analysis for all genes in yeast, thereby providing a global set of perturbation and knockout expression phenotypes. We might expand Idekar *et al*.'s scheme as follows: first, knock out every gene in the genome and measure global data on gene, protein, and metabolite expression levels; second, combine data with all previously known interactions, collected from genome-wide interaction screens[5,6] as well as from previous reports in the literature[7]. Include other interpretive frameworks, such as predicted transcriptional and functional gene networks[8]; third, embed known pathways into these networks; and fourth, try to model flux of metabolites along known pathways given the observed gene, protein, and metabolite expression data. Any rules learned in the last step can potentially be used to predict flux through new, currently unknown pathways. Should rules be constructed for finding pathways *de novo*, this data set would be the gold standard for pathway modelers, presumably providing virtually all of the raw data for modeling hundreds of pathways, including many not yet discovered. And much like Frost's traveler, we could begin to explore the paths less traveled.

1. Idekar, T. *et al. Science* **292**, 929–934 (2001).
2. Gygi, S.P., Rochon, Y., Franza, B.R. & Aebersold, R. *Mol. Cell. Biol*. **19**, 1720–1730 (1999).
3. Pasternack, L.B., Laude, D.A. Jr. & Appling, D.R. *Biochemistry* **33**, 7166–7173 (1994).
4. Fiehn, O. *et al. Nat. Biotechnol.* **18**, 1157–1161 (2000).
5. Uetz, P. *et al. Nature* **403**, 623–627 (2000).
6. Ito, T. *et al. Proc. Natl. Acad. Sci. USA* **98**, 4569–4574 (2001).
7. Xenarios, I., *et al. Nucleic Acids Res*. **29,** 239–241 (2001).
8. Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O. & Eisenberg. D. *Nature* **402**, 83–86 (1999).