VIEWPOINT

Molecular Networks: The Top-Down View

Dennis Bray

Network theory can give a useful overview of how a biological system works. But to make testable predictions, we need the details.

The exhilarating progress of the past decade has brought an unprecedented wealth of quantitative information on living systems, from genomic sequences to protein structures and beyond. But although technical advances make data collection ever easier, investigators are increasingly concerned by their inability to gain a bigger picture. How can this growing mountain of facts be assimilated, and where will the new ideas come from that will help us gain a broader perspective?

One possibility, recently popular, is to treat living cells as a network. Everything in a living cell is, of course, connected to everything else, and interactions between macromolecules through multiple noncovalent bonds are the very fabric of life. It is therefore an attractive notion that, by taking a top-down view of protein-protein interactions, enzymatic pathways, signaling pathways, and gene regulatory pathways, we will gain a better perspective of how they work. Disciplines such as engineering and the social sciences have used networks to analyze their data for many years. So why shouldn't molecular biologists use insights gained from these other fields to learn more about cells? Is this approach useful, and if so, what can it teach us?

The most basic feature of any network is its architecture, which places boundaries on how it acts and how it might have been formed. If you arrange a large collection of nodes (representing molecules in our case) in two dimensions, for example, you can connect them up in a variety of ways—by linking nearest neighbors in a regular fashion, or by selecting them at random and joining them together. A third strategy—which is of great contemporary interest because it seems to correspond to many naturally occurring networks—is to give a few of the nodes a very large number of connections and allow the rest to have relatively few (Fig. 1). These three networks will exhibit different global features, even if it is assumed that they contain the same number of nodes and the same number of connections (*1*). The number of connections per node for both the regular and random networks, for example, will have a roughly normal distribution with an average value that gives a characteristic "scale" to the network.

By contrast, the nodes in the third type of network range from a very few highly connected species to a large number of weakly connected species. Characteristically, the number of molecules (N) with a given number of connections (k) falls off as a power law: $N(k) \sim k^{-g}$, where g is between 2 and 3. Because $N(k)$ does not show a characteristic peak value, this type of network is often referred to as "scale free." The average distance between any two nodes of a scale-free network (the number of intervening connections) is almost as small as the random network. On the other hand, the extent to which neighbors of a node are themselves connected (referred to as its clustering coefficient, or "cliquishness") is almost as large as in a regular network. Scale-free networks are best known in sociology, where they have been shown to represent networks of friends in a population and are sometimes referred to as "small-world networks." Estimates of the distance between any two of the several billion sites on the World Wide Web are said to be close to 20 intermediate links (*2*).

A flurry of recently published results indicates that protein-protein interactions also have the features of a scale-free network. The interior of a living cell is an aqueous slurry based in large part on multiprotein complexes. Some complexes have been isolated and studied in detail but many more remain uncharted, either because they are insoluble or because they depend on fragile, transient liaisons that fall apart as soon as one tries to isolate them. Understanding the nature of these complexes, where they are located, and how they work is crucial for an understanding of the cell. Consequently, investigators have been encouraged to develop fast, high-throughput techniques such as yeast two-hybrid screens and affinity chromatography followed by mass spectrometry to detect which proteins bind to which. Other methods have been devised by which protein associations can also be deduced from genomic data. Unfortunately, each method has its drawbacks and none gives complete or unambiguous data. Side-by-side comparisons of data obtained by different methods show limited reproducibility, and there are serious concerns that what is examined might be only a subset of the complexes (*3*). But accuracy can be improved by combining data from different sources, and the results all indicate that protein interaction networks are small-world networks (*4, 5*). That is, some proteins serve as hubs for very large numbers of interactions, whereas the others, the majority, act more like simple links and participate in one or a few complexes.

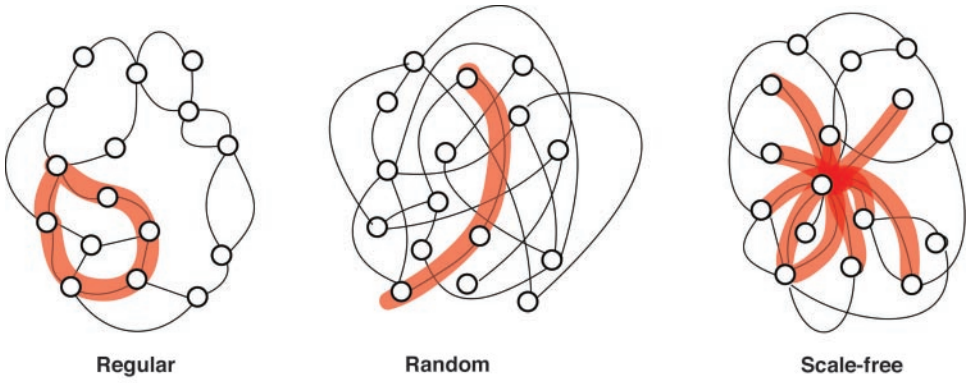

Fig. 1. The same set of nodes can be linked together in many different ways, three of which are illustrated here. A regular network, with nearest neighbors connected, tends to be "cliquish," having local groups of highly interconnected nodes. A random network lacks cliquishness but is easily traversed because the number of steps between any two nodes is relatively small. Scale-free networks, distinguished by the presence of a few highly connected nodes, are both cliquish and easily traversed. Most networks of interest to biologists are, of course, much larger than those shown here and require graph-theoretic methods for their analysis.

Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK.

Probably the best characterized molecular network that exhibits scale-free properties is that of the interlinked pathways of metabolism. Pathways of enzymatically catalyzed reactions that interconvert the hundreds of small molecules of a cell are very well known and extensively documented. Indeed, the familiar biochemical wall chart of intermediary metabolism is the oldest and best established example of a molecular network. Thirty years ago, Kacser and his colleagues pioneered mathematical methods for the analysis of metabolic networks, representing individual small molecules such as pyruvate or citrate as nodes and the enzyme reactions that interconvert them as connections (*6*). These methods allowed them to deduce global features, such as the contributions made by different steps to the overall flux of the pathway, or the way that changing one step would affect the flux through another step at a remote part of the network. This body of work, now known as metabolic control analysis, stands as a pioneering example of how global features can be distilled from a large body of network data (*7*). Recent graph-theoretic approaches to this same body of information have shown that metabolism also has the properties of a scale-free network. Some molecules, such as pyruvate or coenzyme A, are large hubs, whereas the average molecule undergoes just one or two reactions. The number of catalytic steps required to go from any one compound to any other is surprisingly small, and metabolic networks have a high clustering coefficient, which suggests the presence of local cliques or clusters of connected molecules (*8*).

By itself, the fact that a network has scale-free properties is of limited use to biologists. Power laws occur very widely in nature and can have many different mechanistic origins. If we wish to obtain testable biological insights, we must probe further into the substructure of the network. One way forward is to focus on local clusters or cliques in a network and ask how these are themselves arranged. Can we resolve metabolic networks into hierarchical subsystems of highly interconnected reactions sharing similar functions? Can we relate protein interactions to RNA expression data or to cellular location? Eventually, such a top-down analysis leads us to the same modules and motifs identified in other, more reductionist approaches, as described by Alon (*9*).

The networks mentioned so far are "isotropic" in the sense that they do not have a well-defined input and output. But there are other kinds of networks whose primary function is to transform a set of inputs into a second set as output. This is the case with neural networks, originally developed as simple models of how parts of the brain function. Neural networks have a remarkable ability to learn different patterns of inputs by changing the strengths of their connections (*10*). They are widely used in a variety of tasks of machine recognition. From the standpoint of a living cell, the closest approximation to a neural network is probably found in the pathways of intracellular signals (*11*). Multiple receptors on the outside of a cell receive sets of stimuli from the environment and relay these through cascades of coupled molecular events to one or a number of target molecules (associated with DNA, for example, or the cytoskeleton). Because of the directed and highly interconnected nature of these reactions, the ensemble as a whole should perform many of the functions commonly seen in neural networks. Thus, in aggregate, the signaling pathways of a cell are capable of recognizing sets of inputs and responding appropriately, with their connection "strengths" having been selected during evolution. One combination of external conditions might trigger cell division, for example, whereas another might cause differentiation. From what we know of neural nets, it seems that some signaling molecules in the pathway should perform the function of "hidden units" that embody, in their state of activity, an abstraction of the outside world. It also seems reasonable that the networks of cell signaling reactions should, like highly connected neural networks, be resistant to damage and continue to function even if some of their connections are severed.

Once again, the global view gives us only a general impression of the performance of a network. Further progress demands that we descend in scale. Signaling pathways, like metabolism, are subdivided into smaller and more specialized subsystems, with the added complication that these are frequently located in distinct regions of the cell. Analysis of cell signaling pathways is subject to all the caveats noted above for the measurements of protein-protein interactions—no surprise here, because the formation of protein complexes is an essential signaling mechanism. Despite this, considerable progress has been made in identifying circuit design and the way in which small groups of receptors and enzymes can perform distinct computational tasks, such as amplification and coincidence detection (*12*).

Perhaps the most challenging molecular network in a cell is that governing gene expression. Thousands of genes, or in some species tens of thousands of genes, direct the formation of proteins, many of which then collaborate to control, in reciprocal fashion, the expression of other genes. Thirty years ago, Kauffman examined the properties of a theoretical network of genes and showed that they could generate highly complicated temporal patterns of gene expression (*13*). The assumption Kauffman made at the time, of a randomly connected network, is clearly incorrect. But the cascades of gene control uncovered in recent years have a baroque complexity capable of elaborating patterns at least as complicated. The regulatory gene network for the development of endoderm in a sea urchin embryo, for example, contains more than 40 genes linked into a complicated regulatory control system that changes state with cell location and developmental time (*14*). In light of this, the prospect of obtaining a truly global picture of the regulatory control system of a complete eukaryotic organism with many thousands of genes seems daunting. And yet a genome-wide analysis of the binding sites of transcription factors in the yeast *Saccharomyces cerevisiae* was recently achieved (*15*). This study not only documents potential pathways used by yeast cells to regulate gene expression but also identifies network motifs, the simplest units of network architecture.

There are clearly huge obstacles to overcome before we have a complete understanding of molecular networks. It is technically difficult to identify connections with a high degree of certainty, and harder still to make quantitative measurements of their strength. Even when the data have been obtained, novel and sophisticated methods are required to understand what they mean. But the results obtained so far are encouraging and demonstrate the need for analysis at multiple levels—from the global graph-theoretic view, through hierarchical levels of subsystems, down to individual network motifs. We have a new continent to explore and will need maps at every scale to find our way.

References

1. D. J. Watts, S. H. Strogatz, *Nature* **393**, 440 (1998).
2. A.-L. Barabási, *Linked: The New Science of Networks* (Perseus, Cambridge, MA, 2002).
3. C. von Mering *et al*., *Nature* **417**, 399 (2002).
4. H. Jeong, S. P. Mason, A.-L. Barabási, Z. N. Oltvai, *Nature* **411**, 41 (2001).
5. A. Wagner, *Mol. Biol. Evol.* **18**, 1283 (2001).
6. H. Kacser, J. A. Burns, *Symp. Soc. Exp. Biol.* **32**, 65 (1973).
7. D. A. Fell, *Understanding the Control of Metabolism* (Portland, London, 1997).
8. A. Wagner, D. A. Fell, *Proc. R. Soc. London Ser. B* **268**, 1803 (2001).
9. U. Alon, *Science* **301**, 1866 (2003).
10. G. E. Hinton, *Sci. Am.* **267**, 144 (September 1992).
11. D. Bray, *Nature* **376**, 307 (1995).
12. U. S. Bhalla, R. Iyengar, *Science* **283**, 381 (1998).
13. S. A. Kauffman, *J. Theor. Biol.* **22**, 437 (1969).
14. E. H. Davidson *et al*., *Science* **295**, 1669 (2002).
15. T. I. Lee *et al*., *Science* **298**, 799 (2002).