Networks in Biology

Rogan Carr

Abstract

This essay describes the recent trend in studying networks and its application to biology. It begins with an overview of network theory and its applicable results and concludes with a specific look at the network of protein-protein interactions.

Introduction

In this paper, I intend to discuss the field of network theory that has become popular in the past ten years and look into how it is being applied to biology. Biology is a unique field because there exists enormous amounts of data but no ready way to process it all. The network approach to systems looks promising for biology because network theory is designed to take large amounts of data concerning relationships between objects, which biology is rife with, and return statistical information which can be compared with existing hypotheses.

After a brief discussion about the different areas of biology that network theory has touched, I will focus on recent research that examines the protein-protein interaction network. This is the network that consists of proteins in a given life form and maps out the relationships that they have with one another, be it chemically or merely physically. The network of protein-protein interactions promises to be interesting because of the failure of the reductionist method in biology. The reductionist method says that all the complexity of life is hidden in the genome, so if we can understand and map out the genome of an organism, we will be able to understand that organism in its entirety.

As the study of genetics became popular, so did the idea that the answer to understanding biology lay in mapping out the genome. It was thought that since there is one protein per gene, then the totality of an organism could be described through understanding its genome, its appearance and its behavior completely determined. Unfortunately, it became clear soon after the human genome was mapped out that this simple picture couldn't be the case: the simple yeast has about six thousand genes; the fruit fly fourteen thousand; the human thirty thousand [1].

The reductionist idea that the solution to understanding an organism was to simply study its genome falls flat when faced with these numbers. A gene contains the information to encode one protein. If proteins acted independently, then we would expect the number of genes to scale with the complexity of the organism. That a single-celled organism like yeast should have about half the genes of a fruit fly and about one-fifth the genes of a human being does not explain this complexity.

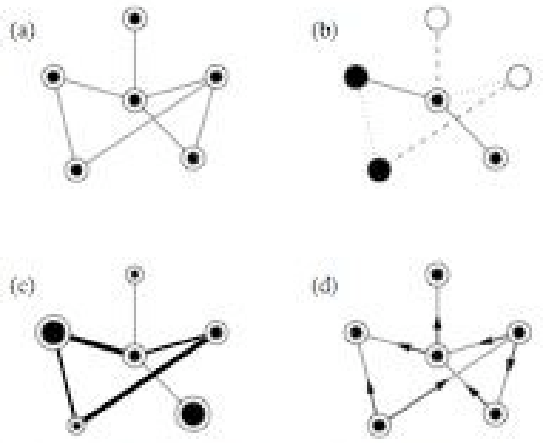
This non-linear rise in complexity with increasing number of genes must be due to interactions overlooked by the reductionist viewpoint in biology. One of the best candidates for the explanation of increasing complexity is that protein-protein interactions build up the interesting nature that we see; it is not the number of genes that controls the complexity, but the interactions between the elements which the genes encode for.

By studying the protein-protein interaction network, we can gain rough, statistical data that will hopefully reveal some of the character that builds up this complexity.

Overview of Networks

Network theory, also called graph theory in mathematics, is an old theory, stretching back to Euler's solution of the Königsberg bridge problem. Until about ten years ago, it used to be primarily the domain of mathematicians working on graph theory and social scientists studying social groups [2]. Developments in computing and statistics have allowed researchers to examine large amounts of data quickly and have resulted in a renaissance in network theory.

Network theory deals with collections of items called *nodes* (or vertices in the mathematical literature), which are connected by lines called edges [Fig 1]. The nodes represent the objects we are interested in, while the edges represent relationships between these objects. These relationships may be reciprocal or may point in only one direction, resulting in a *directed* network. Networks may also contain more complicated edges, having attributes such as weight, which represent the relative strength of the connection, or even connect more than two nodes. Two nodes are said to be *connected* if there is a path of edges drawn between them. [2, 3, 4]



Examples of various types of networks: (a) an undirected network with only a single type of vertex and a single type of edge; (b) a network with a number of discrete vertex and edge types; (c) a network with varying vertex and edge weights; (d) a directed network in which each edge has a direction.

Figure 1 (From [2])

Commonly, networks will be simple. An example of a typical network is that of friendships. The nodes represent individuals, while the links represent who are friends with whom. These edges may be directed if 'friendship' is interpreted in the strict sense (whom one person names a friend may not feel the same way), or be undirected if interpreted loosely. A social scientist may also want to weight the edges based on how long the two have been friends, how well they know each other, or how much they actually like one another. By visually inspecting the graph, a researcher can learn quite a bit about the social hierarchy under question.

At their most basic, networks are created by attaching new nodes randomly to each other based on a probability that each given pair will be connected. This is known as a *random graph*. After a network is assembled, there are many properties that can give insight into the structure of the network. [2, 3]

The first property of a network researchers commonly look at is its *degree distribution*. The degree of a node is just the number of edges attached to it. The degree distribution reveals information about the network's topology: Is the network mainly composed of objects with the same number of edges connecting them, or are there some nodes (called *hubs*) that have a disproportionately large number of edges? [2, 3, 4]

Other properties that can give broad statistical information about the network are the mean geodesic length, the amount of clustering, mixing, degree correlations and community structure. The mean geodesic length refers to the famous "six degrees of separation." What is the average shortest-distance between two points on the network? Clustering is a way of measuring the formation of triangles on a network. In terms of a social network, this would measure the probability for a given person to be friends with his friend's friends. Mixing shows if nodes of different types have a preference of which kind they link to. Similarly, degree correlations reveal how nodes of different degrees connect. Do nodes of high degree tend to connect to others of high degree, or mostly to low? Finally, a measure of community structure tells if the network tends to break up into large groups of high connectivity, or if the network is more uniform. [2]

In a random graph, most of these properties are determined by the method for assigning randomness to the network. They are therefore random, and can't give much information about the graph. The one property that is not random in a random graph is the average number of links between a two nodes. Random graphs are *small world* networks, meaning that they can be crossed from one side to another in a finite and small (roughly the logarithm of the number of nodes) amount of edges. [2]

Networks have become popular research tools in the last ten or so years because of advances in computing and statistical analysis. This has allowed researchers to look at the properties of large networks in the real world with accuracy and ease. The surprising result that experiments have shown is that most real world networks are not random graphs, but rather have many interesting properties [Fig 2].

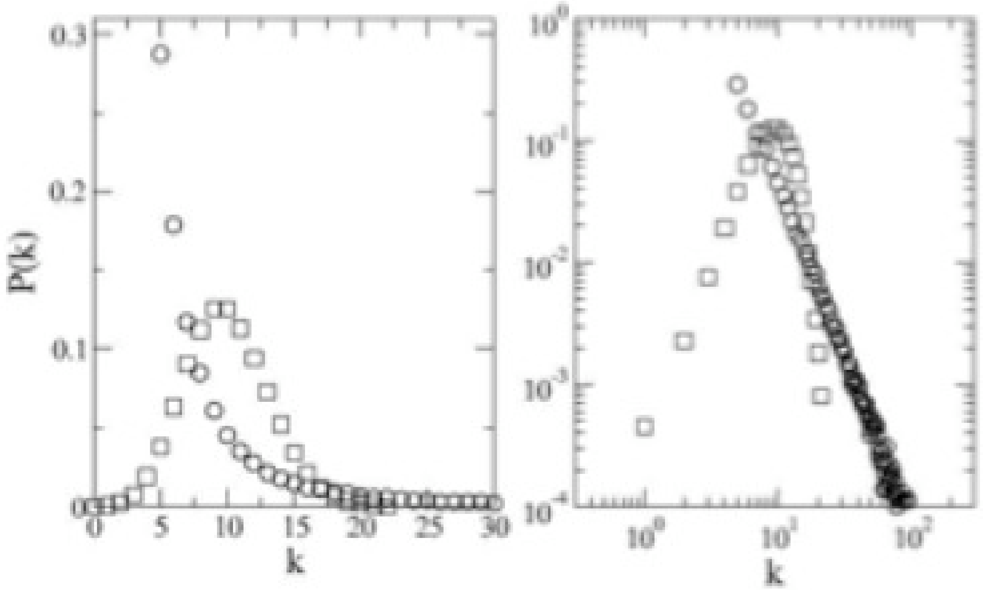


Figure 2. A comparison between degree distributions a random graph (squares) and a scale-free graph (circles). Note the linear scale on left, log-log on right. (From [4])

This is due to two main effects: Networks in the real world are not static; they constantly have nodes added and removed. Real world networks are also not formed randomly. Due to relationships between the nodes, when nodes are added, they will attach preferentially to certain other nodes. These two factors result in evolving networks that have all sorts of interesting properties. [5, 2]

For example, many networks studied recently cannot be described by giving the average number of edges a vertex has. The degree distributions follow a power law, meaning that any given node could be connected to one, ten or even one thousand other nodes, and as a result, stating the average number of connections per node would be meaningless. These networks are referred to as *scale-free*, and are extremely common and popular to study [6].

This scale-free-ness is due to the fact that when nodes are added to a network, they do so with what researchers refer to as preferential attachment. In networks that follow a power law degree distribution, new nodes added to the network will be much more likely to attach to existing nodes that have high degrees than to those that have low degrees. This can be seen in the example of social networks that when people meet other people, they are more likely to meet outgoing people with many friends than loners.

This effect of preferential attachment in a network can be likened to the effect of interactions on a gas of atoms. If you just have a collection of non-interacting atoms distributed randomly, they will float around as a gas. Once interactions between different atoms are turned on, all kinds of interesting properties will appear. What was formerly a classical gas can now be a solid or a liquid, or a Bose-Einstein condensate. By tuning the parameters that control how new vertices are attached to a network, the network will diverge from a random graph to a more interesting state with properties far from the equilibrium of a random graph.

Networks in Biology

The network approach to studying systems has been applied to many different biological systems in recent years. While the network approach can’t give any information about the actual physical processes taking place, it plays a necessary role in studying the overall behavior of the system. This approach is useful for biologists, because networks properties can reveal interesting features of the system at hand.

By looking at the network of objects they wish to study, researchers can learn about the organizational structure of the system, the evolution of the system and how the organization of the system affects the function [4]. For example, the statistical behavior of a certain type of organism as a node on the network may lead to better understanding of how the organism evolved by revealing pressures on the organisms or features about the organism previously unknown.

Biological networks that have been studied span many orders of magnitude in size. The examples range from ecological food webs such as predator-prey networks, to physical networks such as networks of blood vessels (and vascular networks), neural networks and bio-chemical networks such as the network of metabolic pathways and the genetic regulatory network. [2, 4]

Another biological network of interest is the network of physical interactions between proteins, the protein-protein interaction network [Fig. 3]. Researchers have begun to study it in recent years in the hopes to get a better understanding of the inner workings of the cell [7]. The field of proteomics, as it is called, is interesting, because it contains researchers from laboratory biology, bioinformatics and physics all working on the same problem.

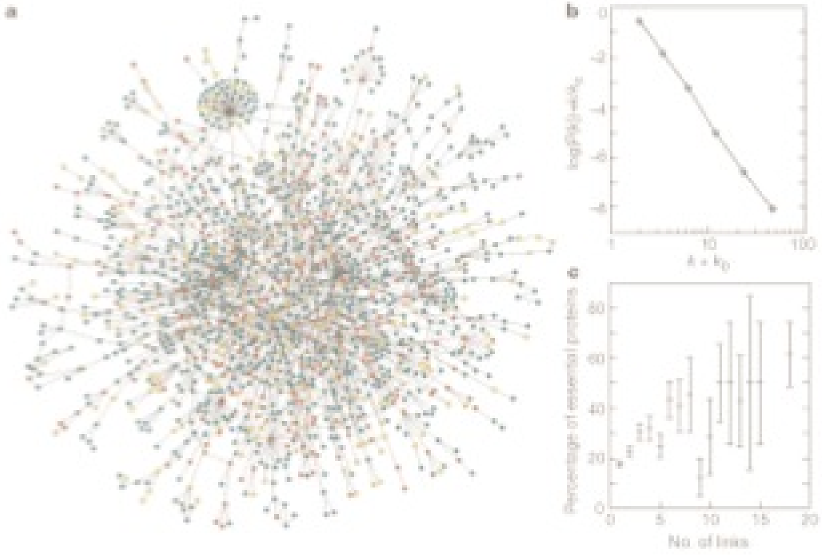


Figure 3. a) Map of protein-protein interactions for a yeast cell. b) Degree distribution. c) The fraction of proteins necessary for life with degree k, versus degree k. (From [8])

The process to create the network begins in the laboratory. Biologists have developed a method for screening protein-protein interactions called the two-hybrid method. The two-hybrid method uses live cells to do all the actual work by taking advantage of the cell's own reporting system to inform about interactions. [9]

After the protein-protein interaction screening has taken place, the researchers enter the information into a database, where it can be accessed by others. From here, the protein-protein interaction network is recreated. While there are still problems with the data due to the experimental process that determines it, it is still a good representation of the underlying data [10, 7]

While the protein-protein interaction network has been studied for many different types of organisms, including yeast, the fruit fly, the roundworm and viruses, the networks do show an overall similarity. The protein-protein interaction network seems to be characterized by a scale-free network (that is, power law degree distribution) with a modular structure [4, 7, 8] [Fig. 4]. While it is unclear whether the modularity is due to incomplete data [7], it is clear that the network has at least one large component, and that the degree distribution must be a power law.

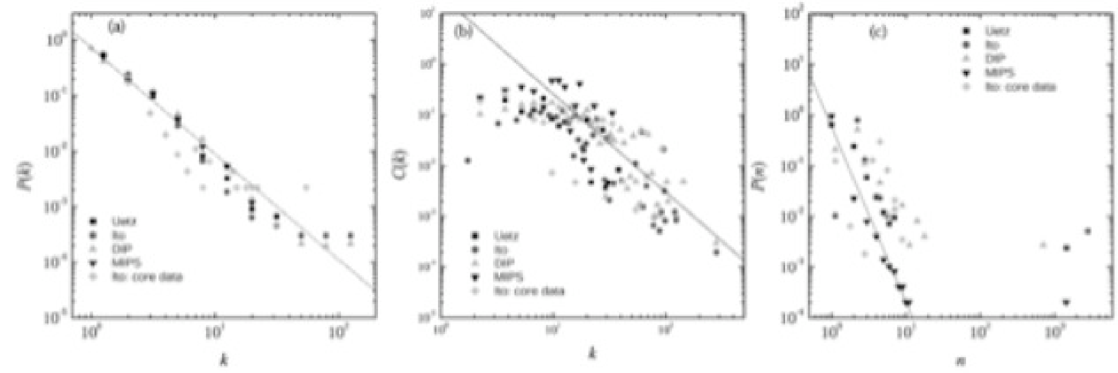


Figure 4. Characteristics of four different protein-protein interaction networks. a) Degree distribution. b) Clustering coefficient as a function of degree. c) Modular size distribution. Note the large component on the right-hand side. (From [7])

Specific studies of the network have also revealed novel information. It was determined that the network seems to form a somewhat community structure, with many cellular functions happening in modular components surrounding highly connected proteins [10]. Studies on the vulnerability of the cell to specific proteins confirmed this, showing that the necessity of a protein for life rose with the number of interactions that protein had with others [8].

That the necessity is correlated with the number of interactions is interesting because it implicitly shows us the structure of the complexity in the protein-protein network. If the cell complexity were built up just in the DNA, and not in the interactions, then the necessity of a protein for life would be correlated only with its function, not its interaction with others.

The study of the protein-protein interaction network is also interesting because it leads to its own questions: Does the network show physical localization? Why is there a scale-free network? The answer to the first question seems to be yes, while the second question is not so clear. One hypothesis is that the scale-free nature of the protein-protein interaction map arose due to gene-duplication. Gene-duplication occurs frequently in evolution, and would result in new proteins that interacted with the old subsets of proteins, causing the degree of old proteins to become large as their neighbors duplicate over time. [7] Thus the nature of the network can also be used to look at the evolution of the network.

Conclusion

Network theory has dramatically expanded in use in the last ten or so years mainly due to the widespread availability of computing resources. Combined with modern statistics methods, network theory has shown itself to be a great tool for weeding through large datasets and gaining information.

Biology is perhaps the field that will benefit most from this field, as it has so much data, but not any great ways to sort through it. The prime example of this is the functioning of cells. There exists an extreme amount of data about the objects within a cell, but there isn't a great understanding of why complexity builds up differently with different sets of genes. While there is a correlation between complexity and the number of genes, the relationship is more muddled.

It is in this setting that studies of protein-protein interaction networks and the like will do the most good. The network approach to systems, while not able to look at the actual function of its members, is a great way for highlighting interesting behaviors that would otherwise get lost in the data. Initial studies with the protein-protein interaction network in the last six years have shown that the proteins do tend to become more important for life as they interact more with one another. This hints that protein-protein interactions could play as much or more of a role that proteins do just by themselves.

As an interesting comment to end with, protein-protein interaction data currently only contains information regarding two-protein interactions. If two-protein interactions seem to be so important, then it is very likely that higher-order interactions will be playing a significant effect inside the cell. The current limit seems to be actually finding methods that can record these interactions.

References

[1] Drosophila Virtual Library. “Introduction to Drosophila” http://www.ceolas.org/fly/intro.html

[2] Newman, M. E. J. “The structure and function of complex networks.” Preprint: arXiv:cond-mat/0303516v1.

[3] Barabási, A.-L. and Bonabeau, E. (2003) “Scale-free networks.” Sci. Am. 288, 60–69

[4] Albert, Réka. “Scale-free networks in cell biology.” Preprint: arXiv:q-bio/0510054.

[5] L. A. N. Amaral, A. Scala, M. Barthélémy , and H. E. Stanley. (2000) “Classes of small-world networks.” Proceedings of the National Academy of Sciences Early Edition.

[6] Arita, M. (2005) “Scale-freeness and biological networks.” Journal of Biochemistry 138, 1-4.

[7] Yook, S.-H., Oltval, Z., Barabási, A-L. (2004) “Functional and topological characterization of protein interaction networks.” Proteomics 4, 928-942.

[8] Jeong, H., Mason, S. P., Barabási, A.-L., Oltvai, Z. N. (2001) “Lethality and centrality in protein networks.” Nature 411, 41-42.

[9] Pandey, A, Mann, M. (2000) “Proteomics to study genes and genomes.” Nature, 405, 837-846.

[10] Maslov, S. and Sneppen, K. (2002) “Specificity and Stability in Topology of Protein Networks.” Science 296, 910-913.