Lecture Notes for 20.320 Fall 2012

Network Modeling

Ernest Fraenkel

In this lecture we will explore ways in which network models can help us to understand better biological data. We will explore how networks can be used to uncover groups of proteins and/or genes that carry out a particular biological function. For example, we will use networks to find kinase substrates and to identify the function of protein sequences. In all these cases, we can take advantage of an area of computer science known as graph theory, which provides a number of sophisticated tools for studying networks.

Networks can be formally defined as a set of nodes (also called vertices) that are connected by edges. Typically a biological system is represented by creating a node for each protein or gene of interest, and using edges between the nodes to represent a relationship between them. For example, we could represent a signaling pathway of the type you have been studying in this course using the diagram below. Here edges represent kinase-substrate interactions.

Identifying Kinase Substrates

Sequence motifs can represent the preferred targets of an isolated kinase domain, but predictions based on these motifs suffer from obvious limitations. As we have discussed in previous lectures, the enzymatic specificity of the kinase represents only one component of the total specificity. Other protein-protein interactions between the substrate and the kinase also contribute to specificity, including SH2 and SH3 domains, D domains and DEF domains. In addition, closely related proteins can have related specificities. Thus, the motif results for members of the same kinase family may be extremely similar. Sequence features alone cannot

guarantee that we have identified the right family member. The ideal solution would combine knowledge of all the mechanisms of specificity that we have already outlined. Of course, the experimental data are still too limited to provide all the necessary information about docking motifs, scaffolds, interactions domains, subcellular localization,etc. for this approach to be practical.

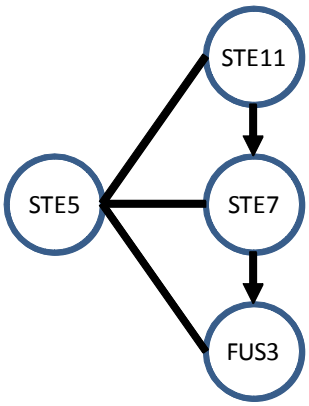
An interesting approach leveraging what we ***do*** know to identify kinase targets was reported by Linding et al. (2007). Cell 129:1415-26. Using a network model, they are able to integrate many different types of data to find the kinase that was most likely to have phosphorylated a particular peptide. In order to understand their approach, we will first cover a few basics of graph theory in the context of biological networks.

Where do the edges come from?

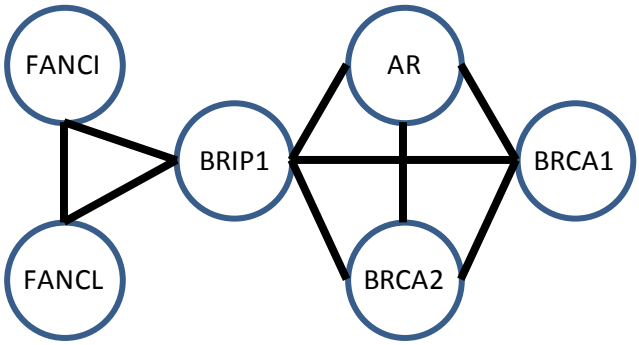
The tools that we will use from graph theory are extremely general, and make no assumptions regarding what the nodes and edges represent. We have already used an example in which the edges represent kinase-substrate interactions. 

Note that the edges have arrows to represent the direction from kinase to substrate. This is an example of a directed graph.

In other cases, we might wish to represent physical associations of proteins, even if they do not carry out any enzymatic reaction. In the example above, we could add the scaffold protein Ste5, as shown below. Here the edges without arrows (undirected edges) represent protein-protein interactions for which there is no particular meaning to a direction. Mass-spectrometry based methods and two-hybrid screens have identified hundreds of thousands of potential protein-protein interactions. The set of all known protein-protein interactions is often called the "interactome." Because these data are so vast, the only efficient way to use them is through graph theory.



We can also build networks in which the edges represent more abstract types of information. For example, we could make a network in which nodes are genes and edges connect genes that cause the same phenotype or disease. In the network below, FANCI, FANCL and BRIP1 are all associated with the blood disorder Fanconi anemia. BRIP1 is also associated with breast cancer, along with AR, BRCA1 and BRCA2. In this case, the entire network is undirected.



You may notice that the above graph is organized so that the genes associated with breast cancer are grouped together and those associated with Fanconi anemia are in a separate group. While this is a trivial example, clustering the nodes on a very complicated graph can often provide useful information about the meaning of nodes. Graph theory provides us with methods for analyzing networks to find such structure.

Basic Graph Theory

We will begin by defining some of the important terms used in graph theory:

Graph Terminology

- G=(V,E)
- Undirected vs. directed
- Weights – numbers assigned to each edge
- Degree(v) – number of edges incident on v
  - In-degree and out-degree
- Path from a to b is a series of vertices <a, v0, …, b> where edges exist between sequential vertices
- Path length = sum of edges weights (or number of edges) on path.

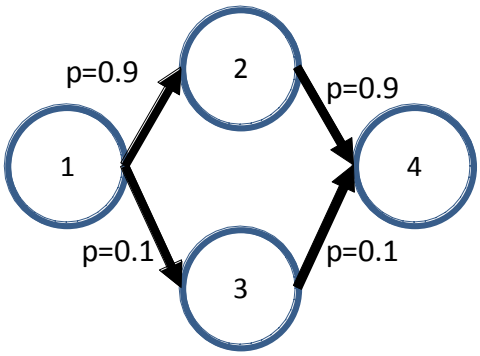
Edge weights

It is sometimes helpful to be able to associate a number with each edge in a graph. These numbers are called weights. For example, we might want to use these numbers to represent the strength of an interaction in terms of free energy. In the case of Networkin and several other examples of interest, the edges are derived from many different data sources, and are likely to include both high-confidence and dubious links between proteins. In these types of situations, it is helpful to associate each edge with a number that represents our confidence in the interaction.

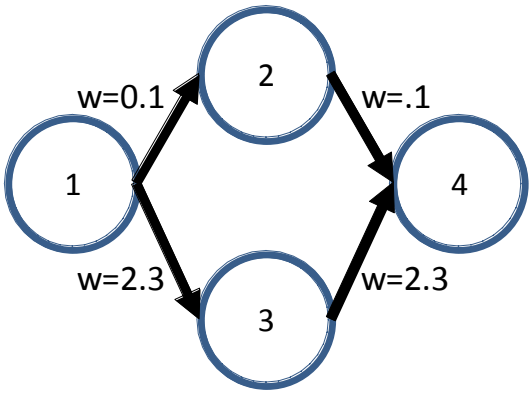
Distance and Shortest paths

We will soon see that it is useful to be able to find the shortest path between two nodes in a graph. The length of a path in a network is defined as the sum of the weights of the edges in a path. Consider the graph below, representing two possible paths by which node 1 causes the phosphorylation of node 4. The path either goes through node 2 or node 4. Let's assume that edges from node 1 to node 2 and node 2 to node 4 are based on extremely good data and have been estimated to have a probability of 0.9 of being correct. The remaining edges are low confidence, each with a probability of 0.1. We obviously have much greater confidence that the pathway uses node2 rather than node3.

Let's assign to each edge a weight that corresponds to the negative logarithm of the probability, as shown below.



Let's assign to each edge a weight that corresponds to the negative logarithm of the probability, as shown below. (I have used the natural log, but any base would be fine.) Now, the edges that have high confidence have low weights and vice versa.



The length of the path 1->2->4 is 0.2, while the length of 1->3->4 is 4.6. So the shortest path corresponds to the high confidence one. In fact, by setting the weight of an edge to $W_{edge} = -\log(P_{edge})$, the sum of the weights corresponds to the product of the probabilities associated with each edge in the path. Therefore, the shortest path will equal the most probable.

There are a number of rapid methods for computing the shortest path on graphs. For more information, see "Introduction to Algorithms" by Cormen, Leiserson, Rivest and Stein.

Networkin

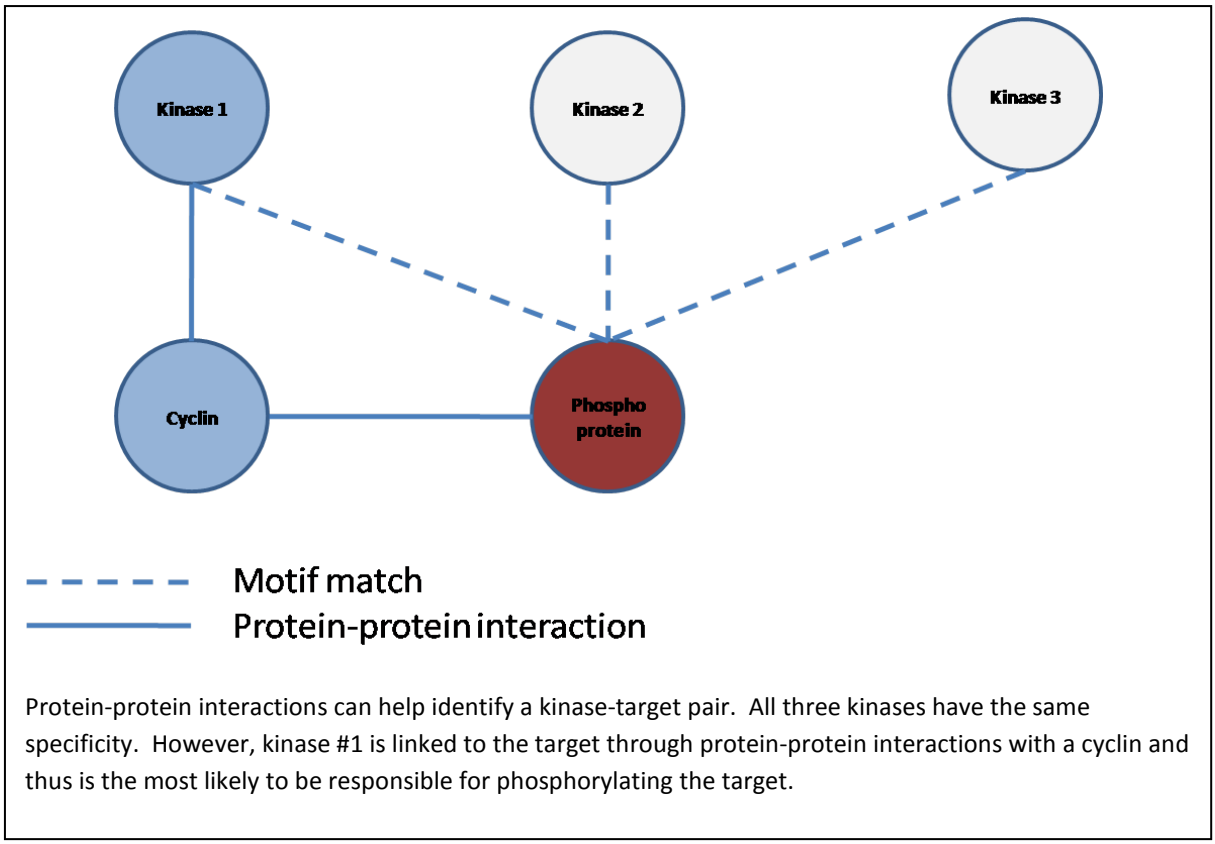
The process begins with an experimentally determined phosphorylation site, such as those determined by mass-spectrometry and seeks to identify the kinase that modified it. The sequence surrounding the site is scored using a library of dozens of motifs representing the substrate specificities of kinases. Each top scoring motif is treated as representing a family of closely related potential kinases. This family is identified by comparing the sequence of the kinase associated with the motif and all other kinases in the same species with the BLASTP program. The result is a family of potential modifiers of the phosphorylation site.

How can we figure out which of these family members is the right one without detailed knowledge of the docking motifs, etc.? The algorithm, called NetworKIN, relies on a database of probabilistically weighted functional associations between proteins organized in a graph. The graph is defined as a collection of vertices, or nodes, representing proteins and edges between vertices that represent physical interactions between the two connected vertices.

The network uses several types of data to determine whether an edge exists between two vertices. For example, the algorithm depends in part on curated databases that report well-characterized enzymatic or signaling pathways. However, these databases contain a very small fraction of all the interactions in the cell, and thus need to be supplemented with other types of data. Additional data come from high-throughput experiments that report physical interactions (mass-spec and two-hybrid experiments), computational methods that attempt to automatically read the literature, and methods that examine gene expression data to find genes that are expressed under similar conditions. Obviously, these data are not all equally reliable, and the algorithm associates each edge with the probability that the interaction is real. A graph with numbers associated with each edge is called a weighted graph.

To find the most likely kinase for the target, NetworKIN computes the shortest path between the target and each potential kinase, where the length of the path is the sum of the weights associated with each edge in the path. Since the network is weighted by the probabilities of interaction, this corresponds to the most probable set of interactions between the kinase and the substrate. All the candidate kinases can then be ranked and the top few considered as the most likely.

What is the value of this network based approach? Consider the case of cyclin dependent kinase (CDK). The sequence specificities of the CDKs are not diverse enough to uniquely identify the kinase responsible for targeting a protein. However, there may be protein-protein interactions between both a CDK and a cyclin and between the cyclin and a target that strongly implicate one kinase over the others. This approach can be extremely valuable in generating hypotheses for the kinases responsible for phosphorylation changes observed in cells as they respond to various stimuli, including disease, toxins, or metabolic stress.



Protein-protein interactions can help identify a kinase-target pair. All three kinases have the same specificity. However, kinase #1 is linked to the target through protein-protein interactions with a cyclin and thus is the most likely to be responsible for phosphorylating the target.

What are the limitations of such an approach? Obviously, if the matrices used to predict kinase targets are wrong, the results will be unreliable. In addition, there are likely to be some kinases with unknown specificities or specificities that differ from their homologs. Equally important, the data underlying the interaction network has many sources of error. These are usually classified by rates of "false positives", proteins reported to interact that don't, and "false negatives", ones without edges in the graph that do interact. Although it is usually possible to assess a false positive rate for high-throughput data, the rate of false negatives is usually unknown. (Why?) What are the effects of false positives and false negatives on the predictions?

Other applications of network approaches

We will not have time to discuss the many other applications of networks to biological problems. Network modeling has also been used successfully to help with protein annotation. A large fraction of genes have no known function. Earlier in the course we examined how sequence homology could be used to help solve this problem. However, many proteins remain unannotated. An alternative approach to determining protein function is sometimes referred to as "guilt by association." The idea is similar to that behind NetworKIN – proteins that are

functionally related will be closer to each other in the graph of known protein-protein interactions. With the development of high-throughput methods of detecting protein-protein interactions based on mass-spectrometry and two-hybrid technology, there are now hundreds of thousands of reported protein-protein interactions. Of course, not all of these are accurate.

If you are interested, you might also take a look at some of the publications from my lab that use these methods to reconstruct unknown signaling and response pathways:

Sources of Protein-Protein Interaction Data

There are many experimental methods for measuring protein-protein interactions. Professor White discussed some of these earlier in the semester. Each type of experiment is associated with different sources of noise, leading to false positive and false negative interactions. In lecture we discussed how non-specific binding in mass-spec experiments can be reduced by using TAP-tag purifications. In this approach, a "bait" protein is connected to calmodulin binding protein, which is linked by a TEV protease site to Protein A. The bait is bound to a solid support coupled to IgG. After washing, one elutes with the highly specific TEV protease. The eluate is then recaptured using calmodulin. After more washes, the sample is then eluted using EGTA, which prevents the binding of calmodulin and calmodulin binding protein. By using two highly specific and gentle elution steps (TEV protease and calmodulin), the protocol minimizes the number of proteins that are recovered simply because they are "sticky." Other sources of protein-protein interactions are yeast two-hybrid and protein complementation assays.

Recently, there has also been success in predicting protein-protein interactions. This is much harder than the predicting small molecule-protein interactions because of the potential structural rearrangements in the proteins. Relevant papers include the following: Tuncbag et al. (2011) Nature Protocols 6:1341 and Zhang et al (2012) Nature 590:556.

References:

Alon, U. (2007) Nat Review Genetics 8(6):450-61.

Berger et al. (2008). Cell 133: 1266-76.

Clackson and Wells (1995). Science 267:383-386.

Deremble and Lavery. (2005) Current Opinion in Structural Biology 15: 171-175.

Fong, et al. (2004). Genome Biology 5:R11.

Gnad, et al. Genome Biology 2007, 8:R250.

Grigoryan and Keating. Current Opinion in Structural Biology 18:1-7.

Grigoryan *et al.* (2009) Nature 458, 859-864

Kaplan, *et al.* (2005). PLoS Comput Biol. 1: e1.

Keating and Newman (2003). Science 300:2097-2101.

Kobe et al. (2005). Bioch. Biophys. Acta 1754:200-9.

Linding et al. (2007). Cell 129:1415-26.

Moreira, et al. (2007). Proteins 68: 803-812.

Noyes, et al. (2008). Cell 133:1277-89.

Rohs, et al. (2009). Nature. 461:1248-53.

Rohs, et al. (2010). Annu. Rev. Biochem. 79:233-69.

Seeman, et al. (1976). PNAS 73:804-8.

Tuncbag et al. (2011) Nature Protocols 6:1341.

Ubersax and Ferrell (2007). Nature Reviews Molecular Cell Biology 8:530-541.

Woods and Schier (2008). Nature Biotech. 26:650-1.

Yaffe et al. (2001). Nature Biotech. 19:348-353.

Zhang et al (2012) Nature 590:556.

20.320 Analysis of Biomolecular and Cellular Systems
Fall 2012