Published on **MacDevCenter** (http://www.macdevcenter.com/)
http://www.macdevcenter.com/pub/a/mac/2004/08/20/bioinformatics.html
See this if you're having trouble printing code examples

Systems Biology

by Robert Jones
08/20/2004

Think of the DNA in a cell, the genome, as a set of blueprints. The proteins are the molecular machines encoded in that genome. Their interactions with other proteins, with DNA, with water, and all the other molecules constitute the production lines in the factory of the living cell. Energy production, protein synthesis, signaling, and a hundred other processes, all involve an exquisite choreography of this molecular machinery. Exploring this big picture of cellular function at the fine resolution of the molecules themselves is what "Systems Biology" is all about.

The grand vision is to integrate information from all the remarkable sources that we have available today to explore ever more complex aspects of biology. It is only by grasping the entire molecular complexity of a process that we can hope to understand the function of the brain, the development of an embryo, and the changes that take place in a disease like breast cancer.

That's the *grand* vision. The reality is a little tamer. A lot of effort today is spent characterizing the proteins in the cell and figuring out which ones interact with each other. Other groups are using microarray-based gene expression experiments to show how sets of genes are turned on and off in response to stimulus. And some groups try to integrate all the data to produce the "big picture" that everyone wants to see.

I'll introduce what I think of as the five components of Systems Biology and then describe a hands-on example that lets you explore a protein network. I will finish up with a set of resources than you can use to delve further into this emerging field of study.

1. Dissection

You can't just look at the sequence of a protein and tell what it interacts with. You need to do some work in the lab. Typically, this means identifying individual types of proteins in the cytoplasm, tagging them with some chemical "label," and using that to track where in the cell they are located and what other proteins they bind to.

Scientists have been doing this for years with specific proteins, but the current efforts combine a variety of new biochemical techniques with automation to improve throughput. This new burst of progress has earned this field a new name, Proteomics, to distinguish it from good old protein biochemistry. This echoes the emergence of Genomics from molecular biology.

Mass spectrometry (MS) is a major tool in Proteomics because of its ability to identify the components of complex mixtures of proteins. The technology behind mass spectrometry has made some amazing advances in the past few years, but the basic idea remains the same. You separate

molecules on the basis of how much they weigh, with a resolution of a single atomic mass unit.

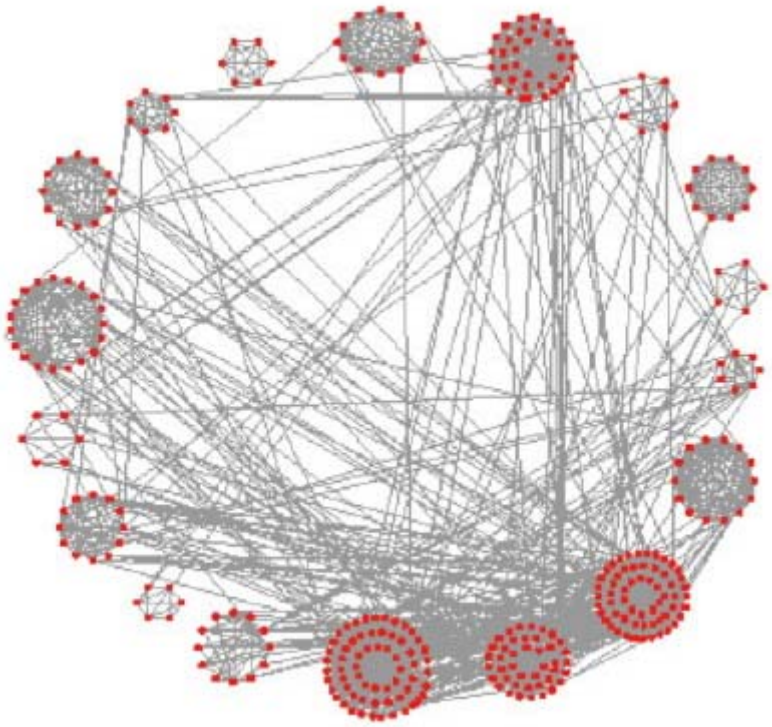
Some of the buzzwords to look out for in proteomics include:

MALDI-QTOF Mass Spectrometry (and many variants thereof)
Isotope-Coded Affinity Tag (ICAT) Analysis
Yeast Two-Hybrid Screens
Green Fluorescent Protein (GFP) tagging
2-D Polyacrylamide Gel Electrophoresis (2-D PAGE)

2. Visualization

The next step is to take all these one-to-one protein interactions and build a graph, or network, that represents the entire set. Each node represents a protein and each edge represents a known interaction. In principle, we can locate the proteins that we know are involved in a specific process and, from their interactions, perhaps discover additional proteins. We can define “pathways” or “cascades” of proteins that, for example, transmit a signal from the cell surface to the nucleus or that cooperate to construct a complex molecule.

The problem is that most networks involve hundreds of proteins. Displaying all of these can result in a tangled mess that is un-interpretable. Here is a relatively simple network that shows protein interactions in yeast.



Yeast network displayed in Osprey

The display of complex graphs is not just an issue for Systems Biology and algorithms from other fields are being brought to bear on our networks. It's an interesting mix of graph theory, visualization, and user interface design. We need a way to view the entire network that is comprehensible. Also, we hope to find a way to limit our view to specific subgraphs, hiding or collapsing the rest of the network

when it is not relevant. And finally, we need to interact with individual nodes and edges to view any annotation associated with them. While the tools that I describe below are making significant advances, there is still a lot of work to be done before they become really useful.

3. Integration

Interaction networks are part of the puzzle. They show us the "circuit diagram." We want to understand how the network operates and how it responds to changes in the inputs. For many processes we already have a lot of relevant information from "conventional" cell biology, microarray experiments, etc. What we're working on is a way to integrate all these data together, with the interaction network as one possible framework on which to display everything.

Most of the processes that we are interested in include some component of gene regulation as well as protein interactions. For example, detection of a protein on the surface of a cell may trigger a cascade of protein interactions that results in one or more genes being expressed in the genome. This interplay of the "worlds" of proteins and DNA is perhaps the biggest challenge for data integration. Whenever a number of proteins interact to accomplish a specific process, chances are that some of the genes that encode them will be expressed in some coordinated manner. So it is reasonable to overlay microarray gene expression data on the protein interaction network and look for correlations.

In the cell itself, most of the processes that we care about have been studied for many years in individual labs and the results have been written up in countless scientific papers. As a result, a major source of knowledge on interactions and regulation is the scientific literature.

It's an interesting phenomenon that biology today dines at two tables. It draws heavily on the specific data from the databases but still finds its interpretations of those data in the scientific literature. Automated extraction of the knowledge embedded in the literature remains a distant goal. But some success has been had in extracting specific terms, like gene names, from papers and inferring interactions where pairs of gene names are frequently found together.

The problem lies in the diversity of the (mostly) English language text of these papers and the false positive rate for these inferred interactions is high. With the abstracts of more than 14 million papers available on the PubMed site at NIH, textual analysis is getting a lot of attention and is still an area where a bright idea can make a big impact.

4. Simulation

An ultimate goal of Systems Biology is understand a complex biological process in sufficient detail that we can build a computational model of it. That would let us run simulations of its behavior and gain a quantitative understanding of its function.

This goal has been pursued for quite some time now. Enzyme kinetics is an area of biochemistry that quantifies the mechanisms and rates at which proteins catalyze the chemical reactions of their substrates. Some of the seminal work in that field was done almost a hundred years ago by the pioneers of biochemistry. Today we have models that describe how certain proteins operate in exquisite detail. But making the next step in complexity, from two or three proteins to even a small network,

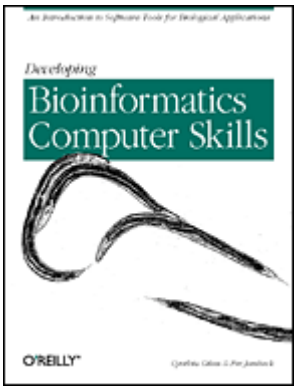
Related Reading

is proving incredibly difficult.

Part of the problem is that we cannot replicate most systems in vitro, in the test tube, in the way that we can with purified enzymes and their substrates. Another part of the problem is that many systems involve the regulation of gene expression and protein synthesis, each of which involve a huge number of different proteins and many, many unknown interactions. People have had some success in modeling specific regulatory networks in bacteria or yeast at a qualitative level but no general approach has emerged.

Eric Davidson's work on sea urchin development at CalTech is a dramatic example of where we might end up. The early stage development of an embryo involves exquisite cascades of regulation. The switching on and off of genes in those first few hours determines the fate of the early cells, whether they give rise to the nervous system, the gut or the muscles of the organism. Sea urchin happens to be an excellent experimental system in which to study development.

Through years of painstaking work, Eric and his group have identified many of the genes involved in the early stages of embryo development. They know which genes are turned on at which stage and can determine how each of them is regulated. Now they are getting to the really fun part and have assembled all their data into a network that resembles an electronic circuit with a series of gates.



Developing Bioinformatics Computer Skills
By Cynthia Gibas, Per Jambeck

Table of Contents
Index
Sample Chapter
Author's Article

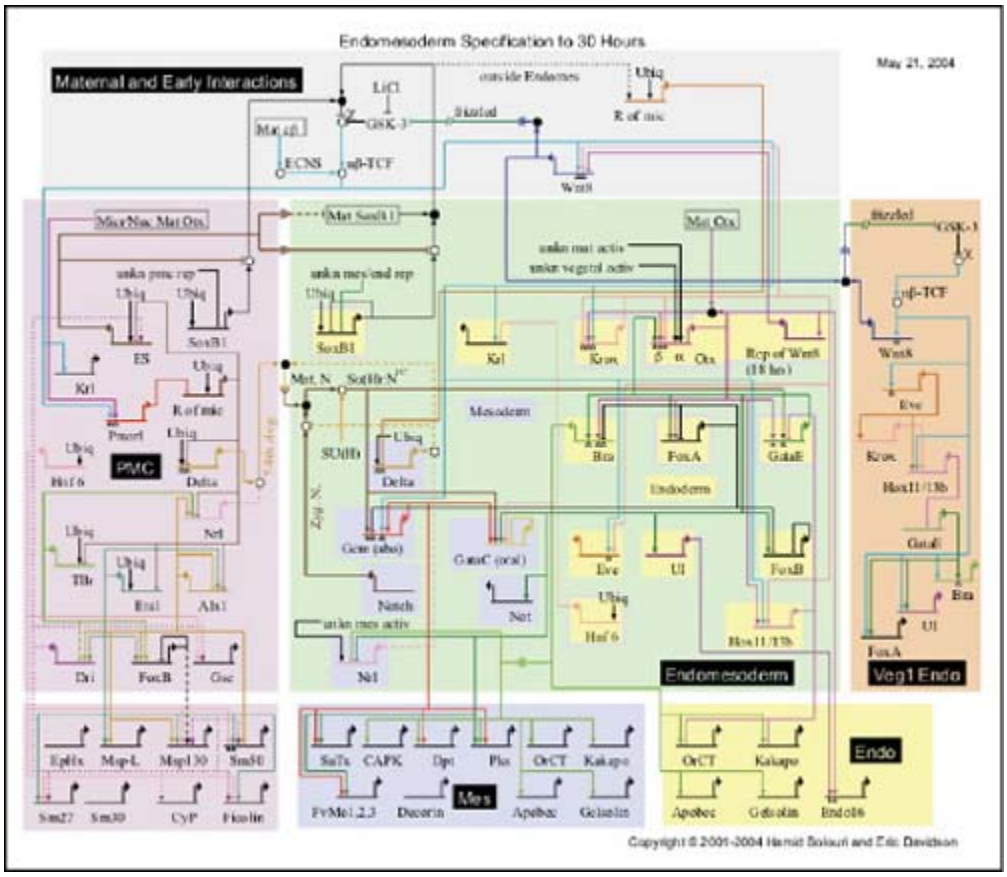
Read Online--Safari
Search this book on Safari:

Go

Only This Book

Code Fragments only

Eric's group has worked with the Institute for Systems Biology in Seattle to develop software to display their network. You can download ISB BioTapestry here: http://sugp.caltech.edu/endomes/ The biology behind the network is a bit too involved for it to make a good hands on example for this article (translation: it's too complicated for me), but it's worth taking a look. Fire up the application, click on one of the "document" icons on the left panel, such as "PMC hourly" and then use the "Hours" control at the bottom left to see how genes are turned on and off during the first few hours of development of the fertilized egg.



Screenshot taken from ISB BioTapestry

5. Perturbation

An important use of interaction networks is to predict how a system will respond to specific changes in its environment or to a genetic defect in one of its components. For example you might explore how normal cells in a tissue become malignant by looking at the effect of perturbations on a network of relevant proteins. Armed with a hypothesis you can go back into the lab and see if it holds up in the real world.

A classic approach to experimental perturbation is to knock out a specific gene by a targeted mutation and see what happens. This has been used for decades but the technologies of today allow us to generate vast numbers of mutations and to monitor the expression of thousands of genes. Rather than looking at specific responses to single mutations we can now look at everything going on in the cell. This "wide-angle" view lets us see changes in things that we never thought to look at before.

Data Sources

High throughput proteomics technologies are yielding a huge amount of data. No doubt about it, proteomics represents a major advance. But this is not the same as DNA sequencing where we have, in effect, digital information. Comparing and integrating proteomics data is challenged by variation from cell to cell and by the ambiguity in the results that emerge when different techniques are compared. Put simply, the data are messy.

Probably the largest database of protein interactions is the Biomolecular Interaction Network Database (BIND) database, based at Mt. Sinai Hospital in Toronto (http://www.blueprint.org/bind/bind.php).

Currently this has around 96,000 interactions between 34,000 sequences from 871 organisms. A nice feature of this site is the tutorial page that shows you three small networks. These are presented in their real biological context. You can find those here: http://www.blueprint.org/bind/bind_tutorials.html. Click on the link at the bottom of each page to see the specific interactions described and then click through for more detailed information including the supporting evidence. Note that certain of the image maps do not work as advertised at the time of writing. Also based at Mt.Sinai is "The GRID," which we will access as part of our worked example. The relationship between these two groups is not clear to me.

Another major source is the Database of Interacting Proteins (DIP) from David Eisenberg's group at UCLA (http://dip.doe-mbi.ucla.edu/). DIP focuses on validated protein-protein interactions and is free for non-commercial access. As an aside, I hope to return to the issues of "free" access to data and software within bioinformatics in the future. It is a messy area that can cause all sorts of problems for non-academic users like myself.

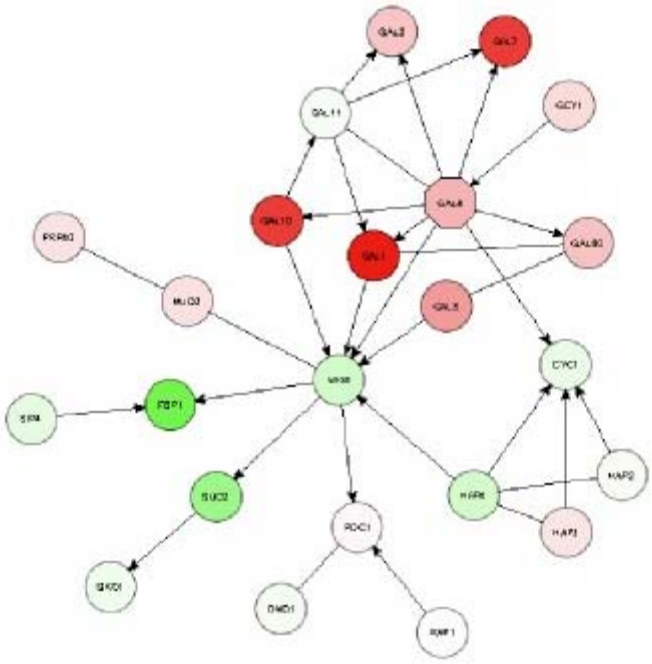
This site has links to some of the other interaction databases:
http://www.hgmp.mrc.ac.uk/GenomeWeb/prot-interaction.html

Software

Most of the software in Systems Biology is targeted to either the specifics of proteomics analysis, interpreting mass spectra, etc., or to the visualization of interaction networks. I am going to focus on the latter.

Cytoscape (http://www.cytoscape.org/) is a joint effort between groups at the Institute for Systems Biology (ISB) in Seattle, Univ. California in San Diego and Memorial Sloan Kettering Cancer Center in New York. They have built a general purpose network visualization tool with plug-in capabilities.

The idea is that people will contribute plug-ins that allow integration and overlay of different types of data. The stable version is 1.1.1 but they have an alpha v2.0 available which includes a new open-source graph library. They have a couple of tutorials that you can work through on their web site and I encourage you to look at those. Here is a screenshot of a simple interaction network from yeast, which has gene expression data overlaid on the nodes, with green indicating relatively high expression and red indicating low expression.



Screenshot taken from Cytoscape

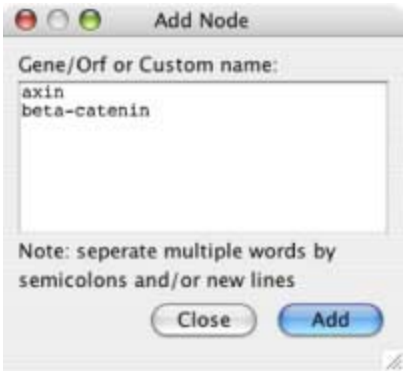
While this sort of software can produce some impressive networks, it can be difficult to demonstrate a simple, practical use of the data. So although Cytoscape is an impressive tool, I'm going to use a different application for our example. Osprey is a Java application from Mike Tyers' group at Mt.Sinai in Toronto. It is free for academics and commercial users can get a free trial license for 30 days. Their home page is here: http://biodata.mshri.on.ca:80/osprey/servlet/Index

In our example, we're going to look at a signaling pathway that plays a critical role in development of the fly embryo (it's the "Wnt" pathway for those in the know). Names of genes and proteins are used interchangeably, which can be a bit confusing. We care about protein interactions but often times we use genetic techniques to discover them.

We'll start by defining two proteins of interest and then use the database to find other proteins that both of them interact with. This is an example of exploring the large dataset to discover interactions that might not otherwise be apparent.

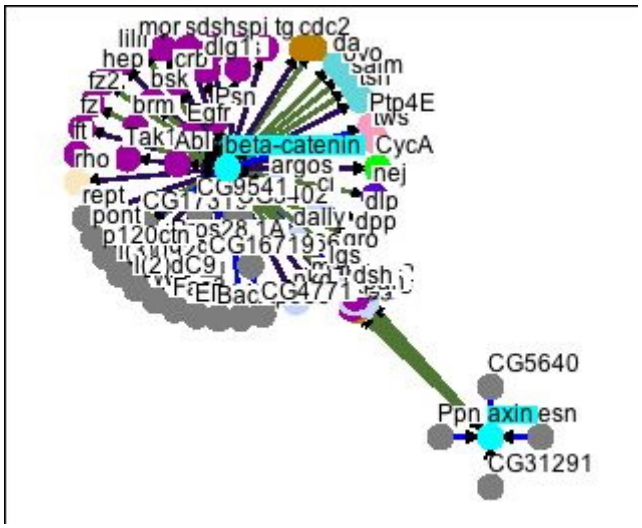
Fire up the application and when "Database Connection Settings" pops up, select "Fly Grid" and click "Continue." This configures Osprey to use the GRID database of interaction in the Fruit Fly, Drosophila.

You will then see blank "canvas" with a toolbar on the top and a panel to the left. We will add two nodes to the canvas where each node represents a gene from Drosophila. Click on the red circle with a black cross in the toolbar to bring up the "Add Node" window. Enter the gene names "axin" and "beta-catenin" as shown and click "Add."



Two dots will appear on the canvas. Left-click on either of them and the left panel will display information about the gene. An important thing to notice is the list of alternate names for each of these genes. What we are calling "Axin" is also known as "axn," "din," "CT6340," and "0442/30." Gene nomenclature is a nightmare throughout biology and perhaps nowhere more so than in the Drosphila community where researchers went through a phase of giving genes names like "disheveled," "Van Gogh," and "Mothers Against Decapentaplegic." The products of a bunch of no-good graduate students with too much time on their hands, if you ask me! This site lists more examples of their tomfoolery: http://www.arches.uga.edu/~jpetrie/genes.html.

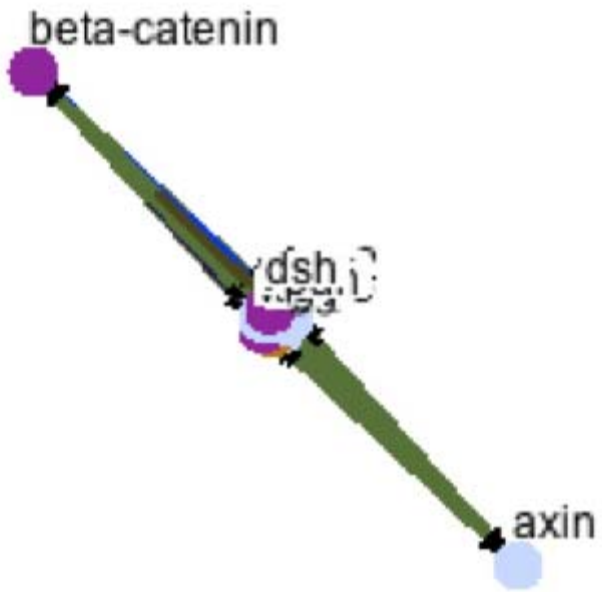
But back to our example! Holding down the left button on a selected node lets you move it around the canvas. Place the two nodes so that they are separate but still have plenty of space around them. Select both nodes with Ctrl-Z or by sweeping them. Then go to the Insert menu and select "All interactions for selected nodes." Osprey will then go out to the Fly GRID database and fetch all nodes that are connected to these two. When it is finished you will have something like this, depending on your node placement.



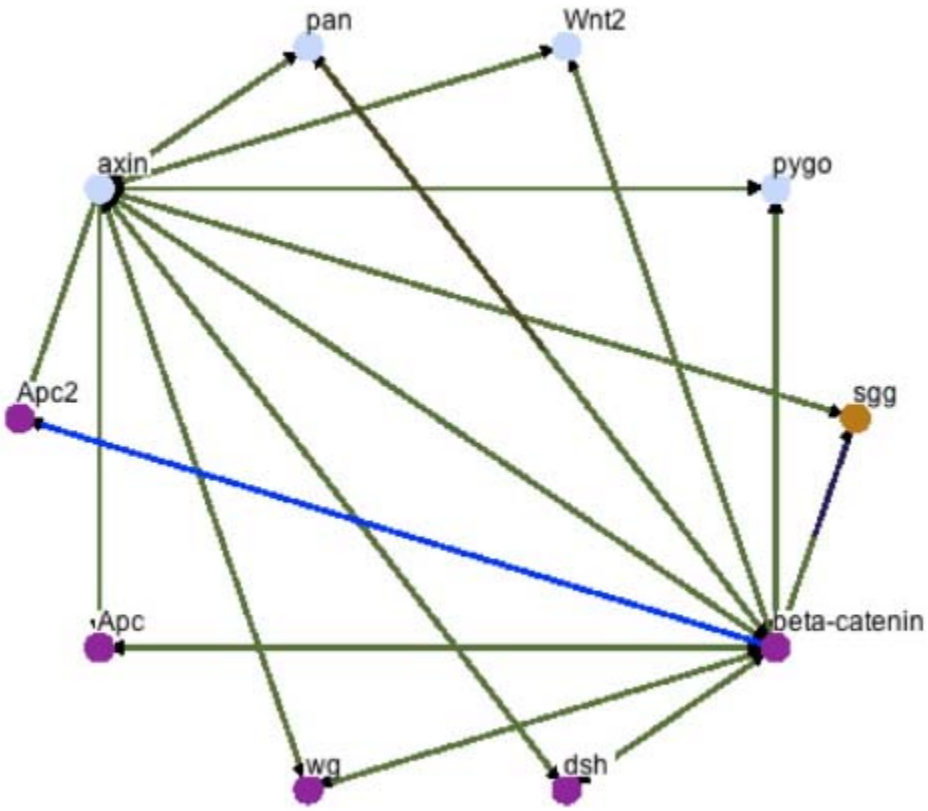
Screenshot taken from Osprey

This looks like the seed head of a dandelion! It is telling us that Beta-catenin has a lot of interactions but Axin has only a few.

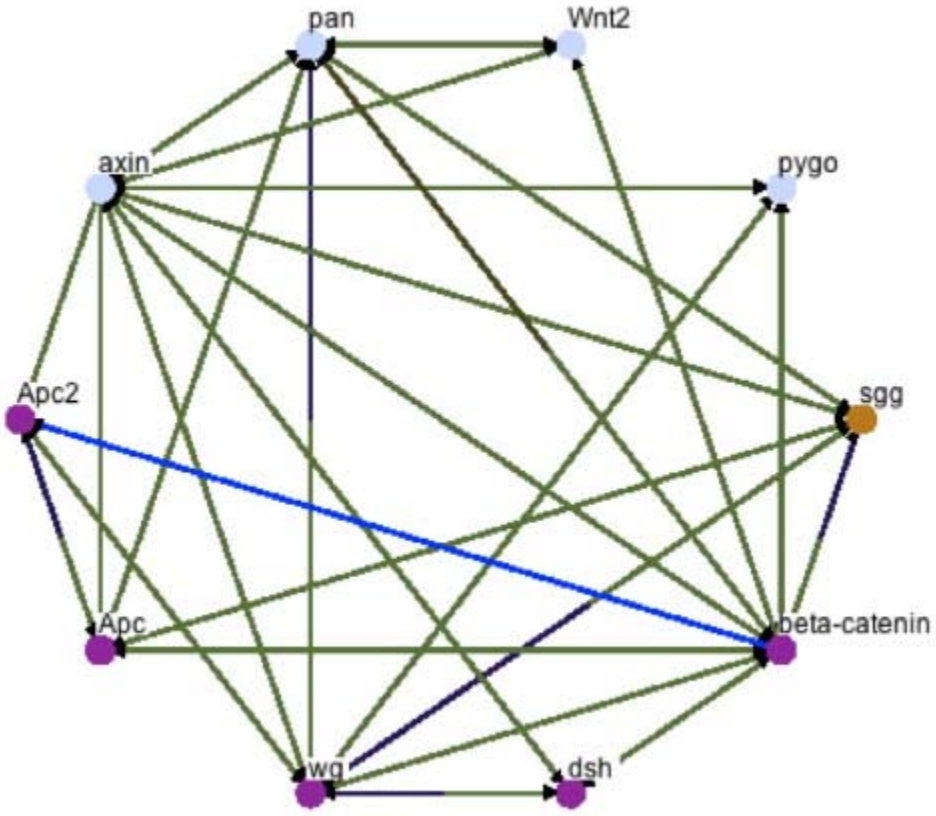
What we care about are proteins that interact with both Axin and Beta-catenin. We can remove everything but these using the filters in the lower left panel of the application. Under "Connection Filters" click on "Minimum". Enter "2" in the popup window and click "filter".



That has simplified things, but all our nodes are collapsed on each other. We need to update the layout of the nodes in order to see what's going on. Select all the nodes, go to "Layout" -> "Circular" -> "One Circle" to spread things out.



Finally we'll add a bit more complexity by selecting all the nodes and going to "Insert" -> "Only interactions within selected nodes."



The color of each node represents the primary function of that gene or protein. You will see reference in the information panel to “GO component,” “GO process.” etc. “GO” refers to the Gene Ontology project, set of three ontologies that try and categorize biological processes, structures, etc. to provide a framework and controlled vocabulary for molecular biology. In our example the purple nodes are involved in cell organization and the light blue ones are involved in signal transduction, communication between and within cells.

This network has almost all nodes linked to all other nodes. In reality not all proteins will physically interact with each other. Some of the interactions shown here are inferred from genetic experiments and with some of those a single perturbation can have multiple indirect effects. So understanding the evidence behind each interaction is important.

The edges of the graph, the lines between the nodes, are colored according to the evidence that supports that interaction. You can click on these to get the detailed information. Click on the link between “pan” and beta-catenin and the upper left panel will show you the experimental technique(s) used to define the interaction.

Below that is a button called “PubMed”. This will open up your browser and point it to the “PubMed” database of biomedical literature at the National Library of Medicine. It will display abstracts for the papers that demonstrated the interaction and in many cases these will include links to the full text of the papers, some of which provide free access.

Click on other nodes and build out the network by adding new interactions. Keep the complexity manageable with the various filters and explore the literature that supports the interactions. Notice how some proteins are involved in huge numbers of interactions whereas others are quite limited. Get

a feel for the complexity of the data and the amount of work that has gone into its discovery.

Who are the Players in Systems Biology?

Systems Biology initiatives are popping up all over the place at the moment. These range from new standalone institutes to loose collaborations between existing labs. Here are a few of the leading lights in the field.

Institute for Systems Biology (ISB) in Seattle
http://www.systemsbiology.org
ISB is a non-profit institute set up by Lee Hood that works on bioinformatics, genomics and proteomics with an emphasis on new technologies. Lee is an eloquent evangelist for Systems Biology and his talks are well worth hearing if you get the chance.

MIT Computational and Systems Biology Initiative (CSBi)
http://csbi.mit.edu/
MIT has taken the approach of coordinating work in existing labs across campus.

Bio-X at Stanford University
http://biox.stanford.edu/
Bio-X is a new program that is bringing together biologists, physicians, engineers, chemists and computer scientists to work on big problems in biology. The program is a combination of campus labs and a central hub in the form of a dramatic new building.

Final Thoughts

By its very nature, System Biology demands input from a wide range of scientific disciplines. Every aspect of the work involves complex data management and analysis and that means there are plenty of opportunities for creative developers. The big centers are an obvious focus for the work but there are many smaller labs around the world that are broadening their horizons to make use of, and contribute to, these growing resources. Do some background reading, see who is working near you and see where your skills might be useful. This turbulent interface where different areas of science flow into each other is an exciting place for developers like us to work in.

Resources

Here are some important papers on Systems Biology with links to the free full text or PDF of each paper.

The digital code of DNA
Hood, L. and Galas, D. (2003) Nature 421, 444 – 448
http://www.nature.com/cgi-taf/DynaPage.taf?file=/nature/journal/v421/n6921/full/nature01410_fs.html

Regulatory gene networks and the properties of the developmental process.
Davidson, E.H., McClay, D.R. and Hood, L. (2003) Proc. Natl. Acad. Sci. USA 100:1475-1480.
http://www.pnas.org/cgi/content/full/100/4/1475

Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry.
Yuen Ho et al. (2002) Nature 415:180-3

http://www.nature.com/cgi-taf/DynaPage.taf?
file=/nature/journal/v415/n6868/full/415180a_fs.html&content_filetype=pdf

BIND: the Biomolecular Interaction Network Database.
Bader GD, Betel D, Hogue CW. (2003) Nucleic Acids Res. 31(1):248-50
http://www.blueprint.org/publications/docs/PMID12519993.pdf

Cytoscape: a software environment for integrated models of biomolecular interaction networks.
Shannon P, et al. (2003) Genome Res. 11:2498-504
http://www.genome.org/cgi/reprint/13/11/2498

Osprey: A Network Visualization System.
Breitkreutz, BJ., Stark, C., Tyers M. (2003) Genome Biology 2003 4(3):R22
http://genomebiology.com/content/pdf/gb-2003-4-3-r22.pdf

Robert Jones runs Craic Computing, a small bioinformatics company in Seattle that provides advanced software and data analysis services to the biotechnology industry. He was a bench molecular biologist for many years before programming got the better of him.

Return to MacDevCenter.com.

Copyright © 2004 O'Reilly Media, Inc.