

PIPE 2 and Gaggle

Hector Ramos

<http://pipe2.systemsbiology.net/>

(PIPE1: <http://pipe.systemsbiology.net/>)

Please follow along in this tutorial as we go through it in the class, or go your own pace if you choose. Feel free to add your own notes. If have suggestions or bug reports that might be useful for others, please email them to hramos@systemsbiology.org.

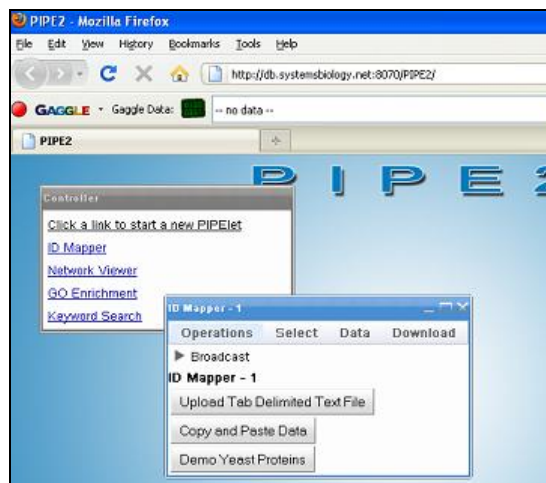
We will be using a set of Yeast proteins derived from a real ISB experiment. The researcher was interested in identifying any potential protein complex in the sample. These proteins had a protein prophet probability of > 0.9 . We will use PIPE2, the Firegoose, Entrez gene, Kegg and STRING to explore the functions of and interactions between these proteins and come to a conclusion about any potential protein complex.

Before we get started, be sure to make sure that the most recent Gaggle Firefox extension is installed on your Firefox browser. At time of writing, this was version 0.8.204. (This step is already taken care of for the course laptops. For other computers, see <http://gaggle.systemsbiology.org/docs/geese/firegoose>).

I. Loading Data into PIPE2

For the sake of this tutorial, we will simply press a button to load our example set of proteins. However, options for importing your own lists of proteins (at a later time) include: broadcasting directly from ProteinProphet (through Firegoose), uploading a tab delimited text file, or copy and pasting tab delimited text directly into a new instance of the IDMapper PIPElet.

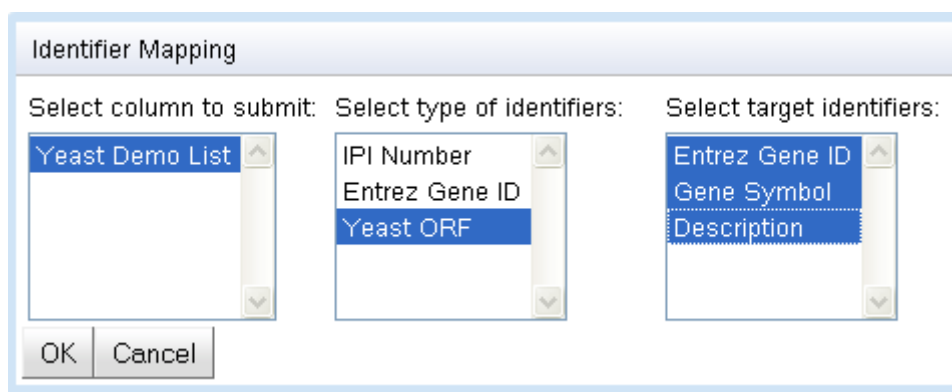
1. Within Firefox, go to: <http://pipe2.systemsbiology.net/>
2. Open a new instance of an IDMapper PIPElet.
3. Click the “Demo Yeast Proteins” button to load the set of proteins we will be working on in this tutorial.



II. Looking up Gene IDs

Entrez Gene IDs are used for a variety of functional annotations. In this step, we'll look up the Gene IDs for our yeast proteins, along with other bits of information.

1. Inside this new PIPElet, click the menu item "Operations" then "ID Mapping" to bring up the Identifier Mapping dialog box.
2. In the Identifier Mapping dialog box, make the following selections, then press OK:
 - Column to submit: Yeast Demo List
 - Type of identifiers: Yeast ORF
 - Target Identifiers: Entrez Gene ID, Gene Symbol, and Description (use the "Control" button to multi-select). The Identifier Mapping Dialog box should look like this:



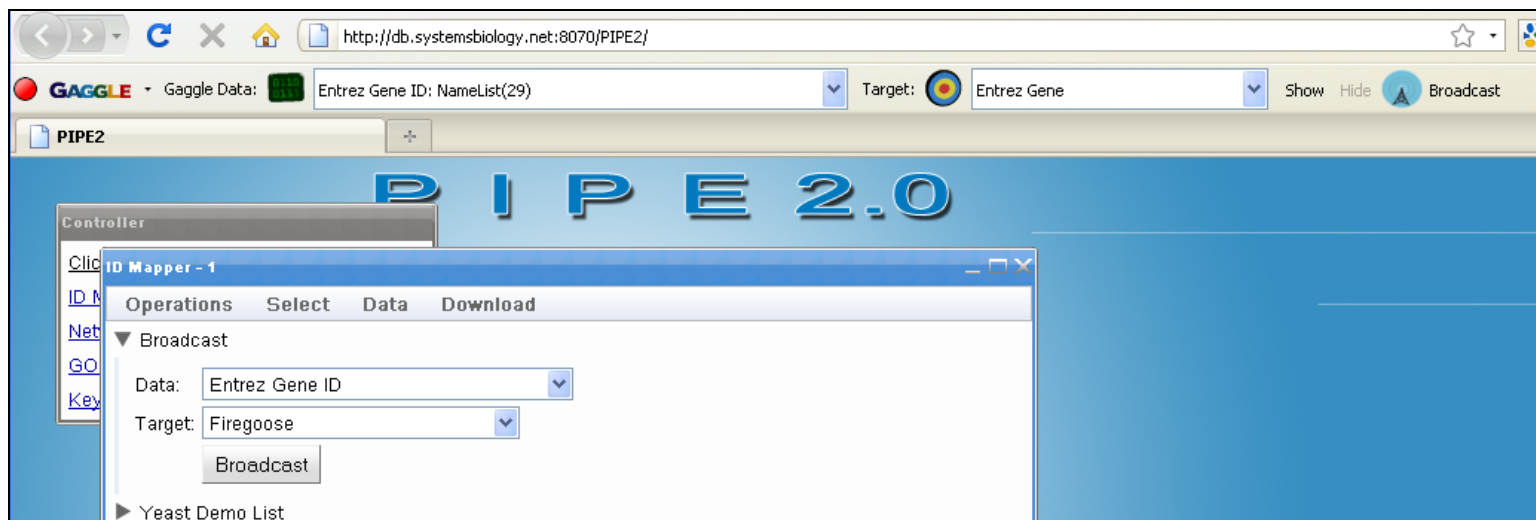
III. Entrez Gene Database (www)

Here we will query the online Entrez Gene database to gather a bit more information about some of our proteins.

To do this, we will use the "broadcasting" mechanisms of both PIPE2 and the Firegoose/Gaggle.

1. Send the Entrez Gene IDs to the Firegoose. Click the "Broadcast" arrow in the ID Mapper PIPElet to reveal the source and target fields. For data source, select "Entrez Gene ID". For target, select "Firegoose". Click "Broadcast" to send the list to the Firegoose.
2. Find the Gaggle toolbar (Firegoose) near the top of your browser window and select Entrez Gene as the target (and "Entrez Gene ID: NameList(29)" as the Data source), like this:

PIPE2 and Gaggle -- Tutorial



3. Press “Broadcast” button on the Firegoose. You will see the NCBI Entrez Gene index page for the genes. Click into these descriptions to find the answers to the following questions:

How many interactions are noted for each of the following genes (estimations are perfectly OK)?

1. FBAI - _____
2. HXK2 - _____
3. MVDI - _____

Bonus question: How many of those interactions are with other genes in our list?

(You really don’t have to answer that, but we’ll visually answer this question in a minute)

IV. KEGG Database (www)

Let’s see what KEGG has to say about our list of proteins.

1. Return to the PIPE2 tab in the Firefox browser. In the IDMapper PIPElet’s broadcast panel, select “Yeast Demo List” as the data and “Firegoose” as the target, and press “Broadcast”. In the Firegoose, change the Firegoose broadcast target to KEGG Pathway and press Broadcast.

2. In the resulting Pathway Search Results page, notice that 15 of our proteins are mapped onto the “Metabolic pathways” item. Click on it to open the pathways image.

In this image, the lines represent transitions/reactions (catalyzed by enzymes) of one compound transitioning into another (the circles). The red lines are those enzymes (proteins) contained in our list. If you hold the mouse over any of these objects on the screen, you will get more information about it. There are also labels scattered throughout describing the pathway in its proximity.

Which pathway has the highest concentration of our genes around it (most red lines in its proximity)? _____

What are the names of 2 of those proteins?


1. _____
2. _____

3. Press “Back” on the browser to go back to the Pathway Search Results.

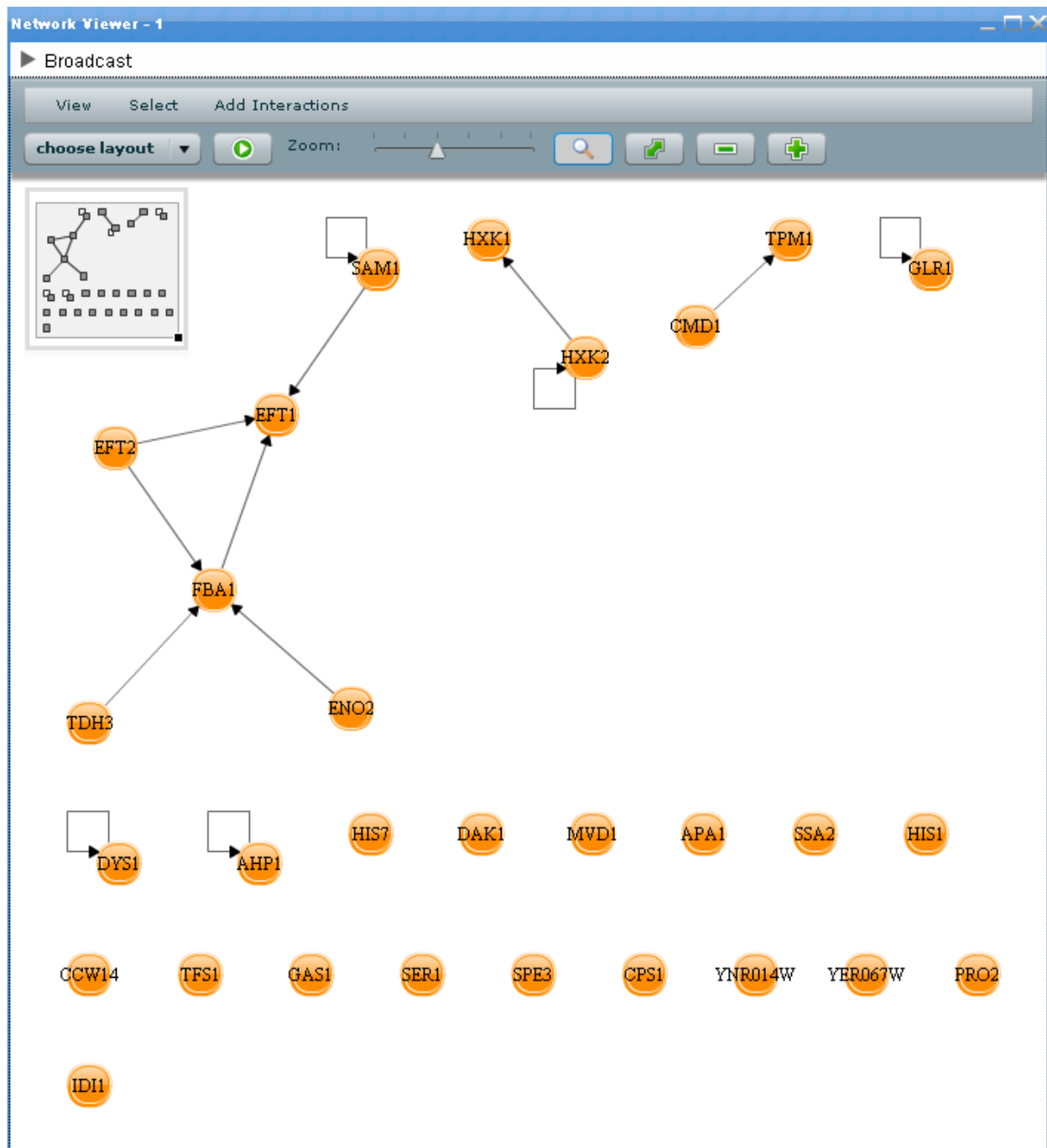
What are items #2 and #3 on that list, and what 3 proteins do they have in common? (hint: press “show all objects”)

V. Exploring interactions (Yeast-2-Hybrid + STRING)

Here we explore interactions through network views.

1. In PIPE2, open a new instance of a Network Viewer PIPElet.
2. Back in IDMapper -I, in the broadcast panel, select “Whole Spreadsheet” as the data source and “Network Viewer – I” as the target, then hit “Broadcast”.
3. In “Network Viewer – I”, in the menu bar, select “Add Interactions” -> “Yeast” -> “Add Yeast Two-Hybrid Interactions”.
4. Press the layout button: 
5. From the menu bar, click “View” -> “Set Node Labels ->” -> “Gene Symbol”.

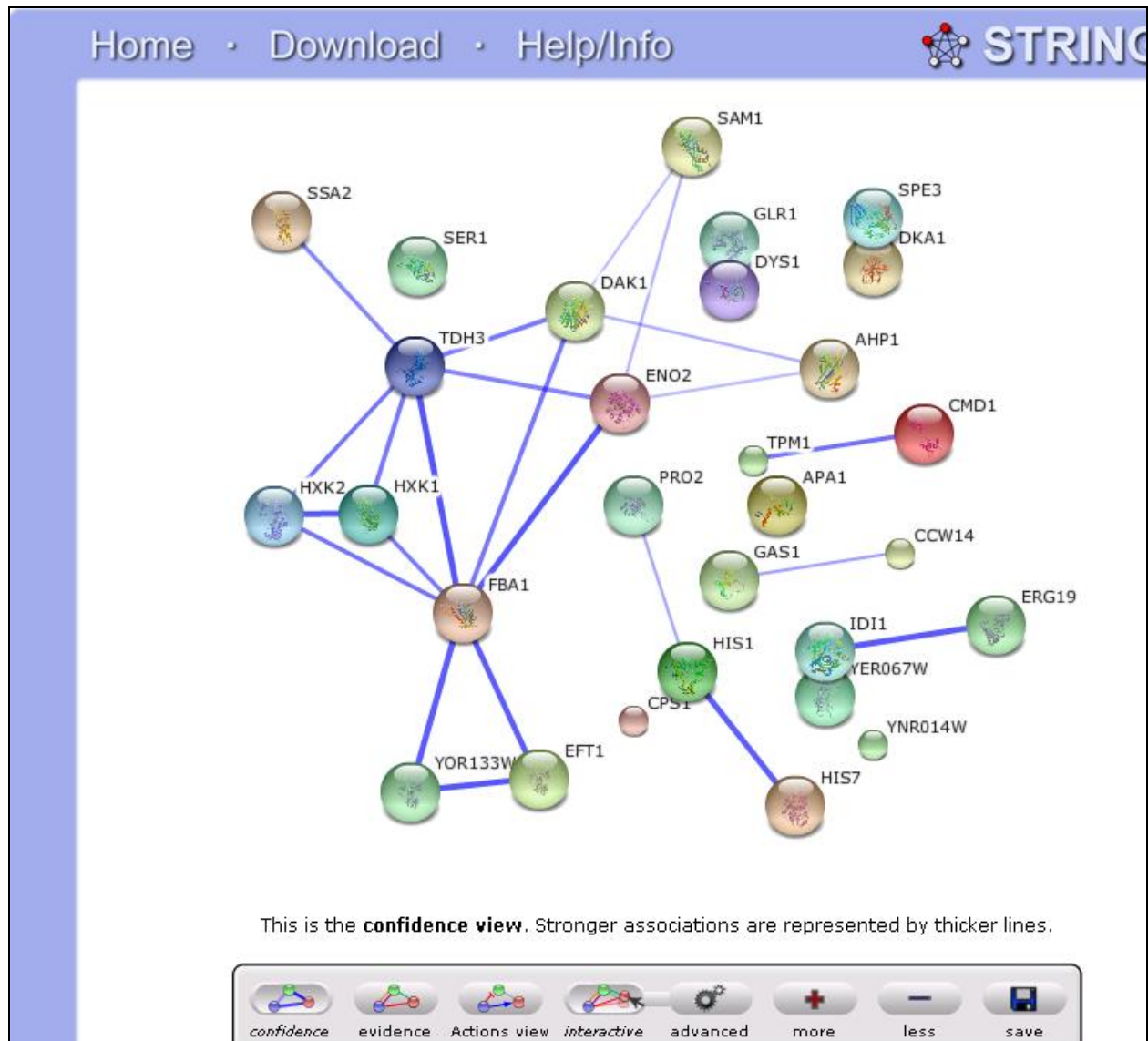
Your Network Viewer PIPElet should look something like this:



These interactions come from UW's Yeast-2-Hybrid interaction dataset.

Lets see what STRING says about these proteins.

6. Back in IDMapper – I, broadcast the first column (By selecting “Yeast Demo List” from the broadcast panel data source field) of the data to the Firegoose, and from the Firegoose, broadcast to EMBL String. Press “Continue” until you get to this screen:



(Note: you may have to click the “confidence” icon to get this view.)

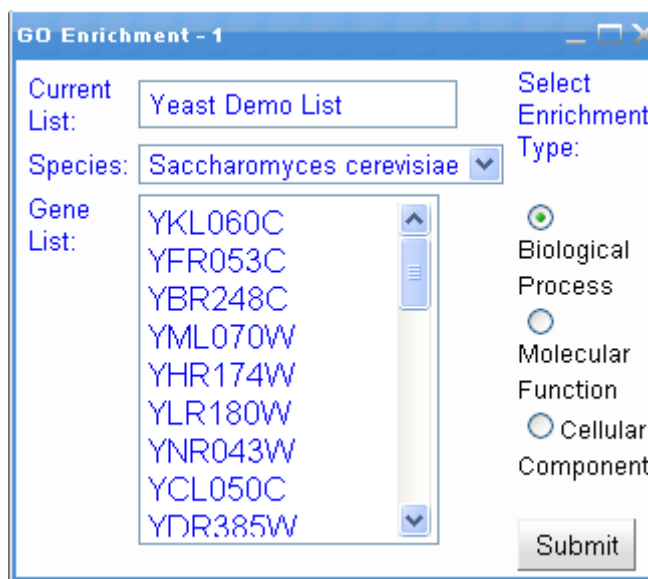
Locate the 3 proteins from the end of the last section (FBA1, HXK1, HXK2). Investigate the difference in connectivity between these 3 proteins. Where does STRING get the connections not found in PIPE2? (hint: click on the connecting edges) _____

We’ll come back to these networks in a minute.

VI. Functional Enrichment with Gene Ontology Categories

Gene Ontology enrichment tells you which GO categories are significantly enriched for a list of proteins/genes.

1. In PIPE2, open a GO Enrichment PIPElet. From IDMapper – I, broadcast the “Yeast Demo List” column to this new GO Enrichment PIPElet. The GO Enrichment PIPElet should look like this:



2. Hit “Submit”. (When you get good at PIPE2, you can multitask and do other things while this process is completing, but for now, just relax.)

We are enriching for biological process GO categories. The p-value for each GO category corresponds to the hypergeometric distribution value based on the 4 parameters: # of items in your list that mapped to that category, your list’s size, number of genes total (in the yeast genome) that map to the same category, and number of total genes possible in the organisms genome (for Yeast, ~6,000).

e.g., for “alcohol catabolic process”:

$$\text{hyperg}(6, 29, 67, 6000) = 4.33804668787027\text{e-}07$$

Notice that the results on the first page seem to also suggest a lot of sugar metabolic processes (like KEGG did).

VII. Integrating Annotation and Interaction Data

The Network Viewer is programmed to treat incoming GO terms uniquely. This might be useful in the following manner.

1. In the GO Enrichment – I PIPElet, select the “alcohol catabolic process” row of the table. This will add that category to the list of possible broadcast sources.

2. Open the Broadcast panel of the GO Enrichment – I PIPElet and select “alcohol catabolic process” as the data source and “Network Viewer – I” PIPElet as the target, like so:

GO Enrichment - 1

Gene Ontology Enrichment results for Biological Process


Data: alcohol catabolic process : (6)

Target: Network Viewer - 1

List name: Yeast Demo List (length: 29)

Broadcast

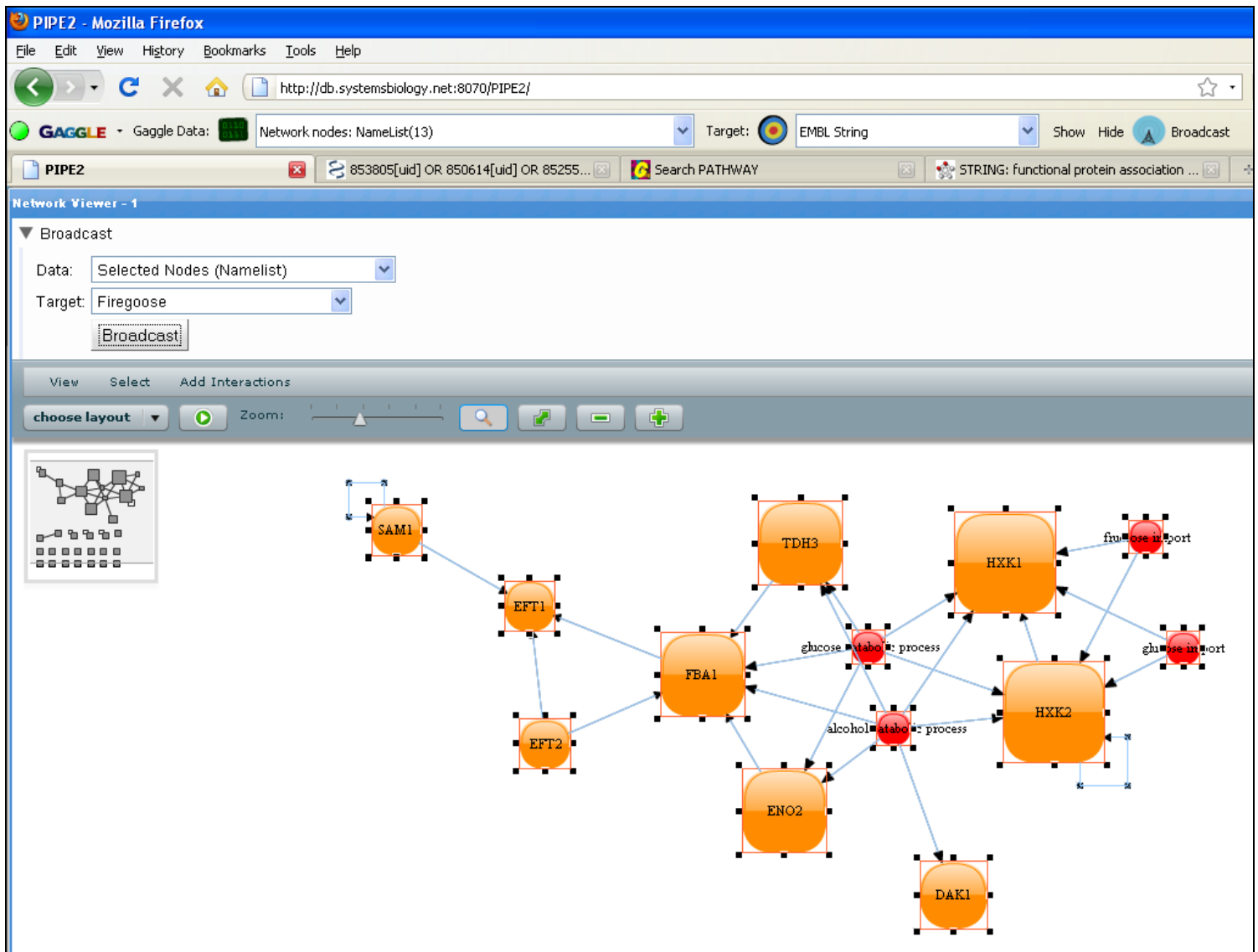
GO Term	GO ID	p-value	Gene Count	Category Size
alcohol catabolic process	GO:0046164	4.33804668787027e-07	6	67
glycolysis	GO:0006096	6.28108472362647e-07	5	38
regulation of cellular protein metabolic process	GO:0032268	2.47643026309800e-06	11	456
glucose catabolic process	GO:0006007	2.82100627244209e-06	5	51

- Hit "Broadcast".
- Do the same thing (broadcast to Network Viewer) with the following GO categories:
 - glucose catabolic process
 - fructose import
 - glucose import (on second page – click the "right" arrow button at bottom)
- Go back to the "Network Viewer – I" PIPElet, maximize it (similar to windows on your desktop) and click the layout button: 

Now we have a cluster of proteins connected by direct interaction experiments (yeast-2-hybrid) and functional associations (GO terms). Let's see how STRING compares.

- Select the cluster in the Network Viewer by clicking and dragging across it.
- Expand the "Broadcast" panel of the Network Viewer PIPElet. Select "Selected Nodes (Namelist)" as the datasource and "Firegoose" as the target and hit "Broadcast". It should look something like this:

PIPE2 and Gaggle -- Tutorial



8. In the Firegoose, ensure “Network nodes: NameList(13)” is the data Source and “EMBL String” is the target, and hit “Broadcast”.

9. **Caution:** In String, select “*Saccharomyces cerevisiae*” as the organism and click “continue”. On the page following that, String tries to map all of your input to identifiers it recognizes. In particular, at the bottom of the page, you’ll notice that it also tried to map “alcohol catabolic process”, “fructose import”, “glucose catabolic process”, and “glucose import”. **Uncheck the mappings String attempted to make!** Then click “continue”.

10. Explore the String network. In particular, look at edges that are in String and not in PIPE2. Click on them and investigate the evidence they provide for those edges. That type of information is not in PIPE2 yet... perhaps one day.

VIII. Conclusion

No conclusive evidence for enrichment of any known protein complexes, however the co-occurrence of the 3 proteins FBA1, HXK1, and HXK2 in different annotation databases may warrant further experimental investigation into possible interactions.

IX. Suggestions for Further Exploration

Cellular component Gene Ontology category enrichment might also help in identifying any specific, known complexes within the cell that our list of proteins might seem to implicate. In conjunction with the Keyword Search PIPElet, one could potentially find proteins that were not in the original list, but that might be worth searching for either in their MS/MS search results (perhaps it had a lower protein prophet probability score than the .9 cutoff), or target that protein(s) in a future SRM experiment.

The Keyword Search PIPElet allows one to enter in a search term and then it returns categories (from the Gene Ontology and Uniprot/swissprot databases) which match the search term along with all the genes mapped to that category.

We didn't do this here because it didn't work well for this example list of proteins. However, it could be possible, for other lists, to enrich for cellular component, find that the list contains 22 of 26 proteins annotated to "transcription factor TFIID complex", open a Keyword Search PIPElet and enter "transcription factor TFIID complex", and find the remaining 4 proteins that belong to the transcription factor TFIID complex. Then you could take a second look at your MS search data to see if there is any supporting evidence at all for the existence of any one of those 4 proteins in your sample.

The Network Viewer contains many functions that we did not explore in this tutorial. Among them are: coloring or drawing (shape & size) of the nodes based on data attributes and filtering out nodes based on # of edges or on data attributes. So the Network Viewer can be used to view protein (or gene) expression values by setting a color gradient from low to high to be from red to green. These options are available in the "View" -> "Attribute Data -> Visual Cue mapping" menu item.