

## PIPE and Gaggle Hector Ramos

<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](mailto:hramos@systemsbiology.org).

We will be using all the proteins from the last SBEAMS tutorial that had a ProteinProphet probability  $> 0.9$ . We will use the Protein Information and Property Explorer (PIPE) to visually explore the functions and relationships of the proteins in the set.

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 0.8.90. (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 the PIPE**

1. **Optional:** Log in. Click the “Log in” tab and create an account using the same username and password from the previous tutorial (or any of your choice). This will save the different datasets we create along the way.

#### **Broadcasting Data In**

2. Open the bookmark you created in the last tutorial that points to the result set display in SBEAMS. On that page, make sure that the “mmarelli – pxproteome -> YeastORF\_20040422” search batch is selected, and the two probability constraints are set to “ $\geq 0.9$ ”, then click the “QUERY” button. Notice that the Firegoose recognizes that the page contains a list of 207 proteins (technical detail: the page has embedded “micro formats”).



4. In the Firegoose, select “PIPE” as the target. Press Broadcast. This will open a new instance of the PIPE in a separate Firefox tab. The PIPE will detect the incoming broadcast and present you with the following dialog, which you should fill out as follows (please pay particular attention to the given name of the dataset):

**Data broadcast detected**

Please provide the following information for the incoming dataset.

I would like to:

Create a new dataset from the incoming dataset

Use the incoming dataset to select a subset of rows of a currently existing dataset

---

Name for Dataset:

Organism:

Content:  Proteins  Genes

Data Type:

5. Press OK to see the data set be imported into the PIPE.

## II. Looking up Gene IDs

1. From the Data -> Summary tab, click the [px.proteome.sbeams](#) link to open the dataset.
2. Use the toolbar that appears to select “Operations” -> “ID Mapping”. In the dialog that appears, select “Yeast ORFs” (if not already selected) for the starting ID type and check “Entrez Gene ID”, “Gene Name”, and “Description” for Return Values. Press OK.

## III. Broadcasting to Web Resources/Databases

1. Now find the radio buttons in the column titles for your table. There are four: Proteins, Entrez Gene ID, Gene Symbol, and Description. Your selection here determines the names (that is, which column) will be broadcasted from this page to other pages (or applications). You should select [Entrez Gene ID](#).
2. Find the Gaggle toolbar near the top of your browser window and select [Entrez Gene](#) as the target (and [px.proteome.sbeams](#) as the Data source), like this:



3. Press “[Broadcast](#)”. You will see the NCBI Entrez Gene index page for the genes. Click into these descriptions and explore any information you find relevant.
4. Return to the PIPE tab. Select the “Proteins” column. Change the Firegoose broadcast target to [KEGG Pathway](#) and press [Broadcast](#).

Many of our genes are not found in the KEGG database. If you scroll down, however, you will begin to see listings of metabolic pathways to which some of our genes did map. Find "Oxidative phosphorylation" and click on it. Notice that 12 of our proteins mapped to this pathway (red text on green background). Other yeast genes are black text on green background; missing genes are black text on white background.

**This concludes** the "strictly in the browser" part of the tutorial. We left several other target online databases unexplored. Feel free to query them on your own time (String is a cool one).

Our goal in the next part of the tutorial will be to explore the documented interactions between the subset of our proteins that are localized in the peroxisome. Here you will see how the Firefox extension (FireGoose) and the PIPE will communicate with java programs that run outside the browser on your computer.

### **IV. Perform GO Enrichment and Explore the Results in Cytoscape**

3. Connect the FireGoose to the Boss. Click the down arrow next to the "Gaggle" label and select "Connect to Gaggle". The bottom right corner of your Firefox browser should indicate whether or not you connected successfully. Also verify by looking at the boss and seeing "Firegoose" on its list of connected geese.

1. In the PIPE, click the Data tab and select the px.proteome.sbeams tab and select the "Proteins" column. In the toolbar, click "Operations" then "GO Enrichment". In the Dialog box that appears, select the Cellular Component option, and click the "OK" button. The Properties panel expands and shows you a progress bar that estimates the time it will take the operation to complete.

This action submits the job to a back-end cluster server. Calculation of the enrichment typically takes less than a minute.

**(BUG ALERT:** When the GO Enrichment operation completes, Firefox may display a "Open with.." or "save to" dialog. There is a bug in some versions of Firefox that keeps the "OK" button disabled. If this happens, simply click the "OK" button, and that will enable it. Then click it again to proceed.)

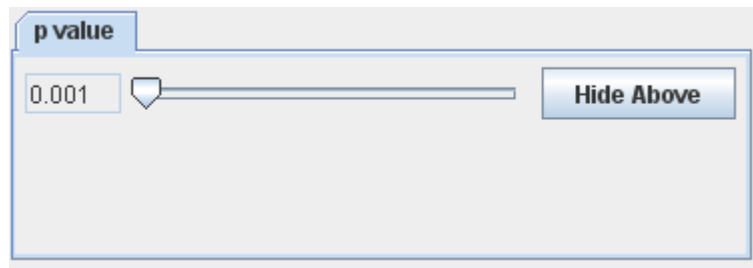
2. A Cytoscape window will soon appear containing a hierarchical network of successively more specific cellular component nodes: general towards the top, more specific below. The size of each node reflects the number of genes annotated to that Cellular Component and the color indicates the p-value of the enrichment. (Please see the discussion of the hypergeometric distribution in the lecture slides for background on this). You will often be interested in large, dark red nodes. These nodes represent the components of a cell whose constituents seem to be over-represented in your list of genes. At any time, you may hold your mouse over any node to see what the tiny little text written inside of it says.

3. Connect the Firegoose to the Boss. You may have noticed a Gaggle "Boss" window appear (and then become minimized) once the Cytoscape window completed loading. This is

automatic behavior in the Gaggle. When a Goose is launched, if there is no boss detected, one is automatically launched. Now, to connect the Firegoose to it, so that all these pieces of software can communicate, click the down arrow next to the “Gaggle” label and select “Connect to Gaggle”. The red light bulb should turn green if the Firegoose connected without any problems.

4. Back to the Cytoscape window. This is a big network and it may seem rather daunting. Lets focus on the most significant GO categories for our dataset

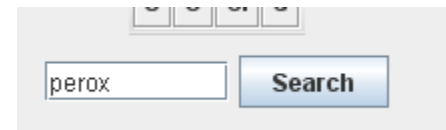
i) First, slide the p-value filter slider-bar all the way down to 0.001. Then press “Hide Above”. This removes all GO categories who’s p-value is greater than 0.001.



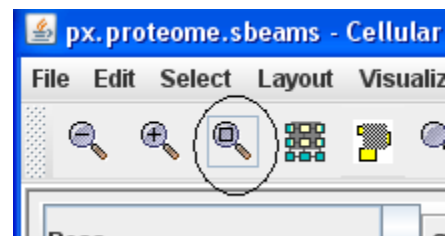
ii) To get a better view of what is depicted, increase the size of the window (maximize) and from the menu bar, do: Layout -> Layout whole graph

Now let’s focus on categories in and around “peroxisomal”.

iii) In the search box, type “perox” and press enter. This will highlight all nodes with “perox” in their name (don’t worry if you can’t find them right now).



iv) Now click the “Zoom Selected Region” button to zoom in to the selection.



v) Use the scroll bars to scroll up and get an idea of the context of these peroxisomal GO categories.

If you still don’t have a clear view of the text, you can drag the nodes around and reorganize them.

For those interested, **Microbody** is defined in the Gene Ontology as: Cytoplasmic organelles, spherical or oval in shape, that are bounded by a single membrane and contain oxidative enzymes, especially those utilizing hydrogen peroxide (H<sub>2</sub>O<sub>2</sub>).

vi) Right click on the “peroxisome” node and click on the “genes” tab of the Node Browser window that pops up. This shows which of our genes were annotated to the

Peroxisome cellular component. Our next goal will be to save this list of genes back in the PIPE and use the PIPE and Gaggle to learn more about them.

5) Close the Node Browser.

### **V. Exploring Relationships Between the Peroxisomal Genes**

1) In the Cytoscape window, select “Firegoose” as your gaggle broadcast target. Make sure only the peroxisome node is selected, and then push the “G” button to send the list of gene names in this node to the Firegoose.



Go back to the Firefox browser with the PIPE open. Your Firegoose browser extension should now look like this:



2) Broadcast the list of gene names into the PIPE where it will be saved for future reference.

To do this:

- i) In the firegoose, select “PIPE” as the target and press the Broadcast button.
- ii) Fill the resulting dialog box as follows (attn: Name & Organism), and then press OK:

### Data broadcast detected

Please provide the following information for the incoming dataset.

I would like to:

Create a new dataset from the incoming dataset

Use the incoming dataset to select a subset of rows of a currently existing dataset

---

Name for Dataset:

Organism:

Content:  Proteins  Genes

Data Type:

### Inspect Gene Interactions

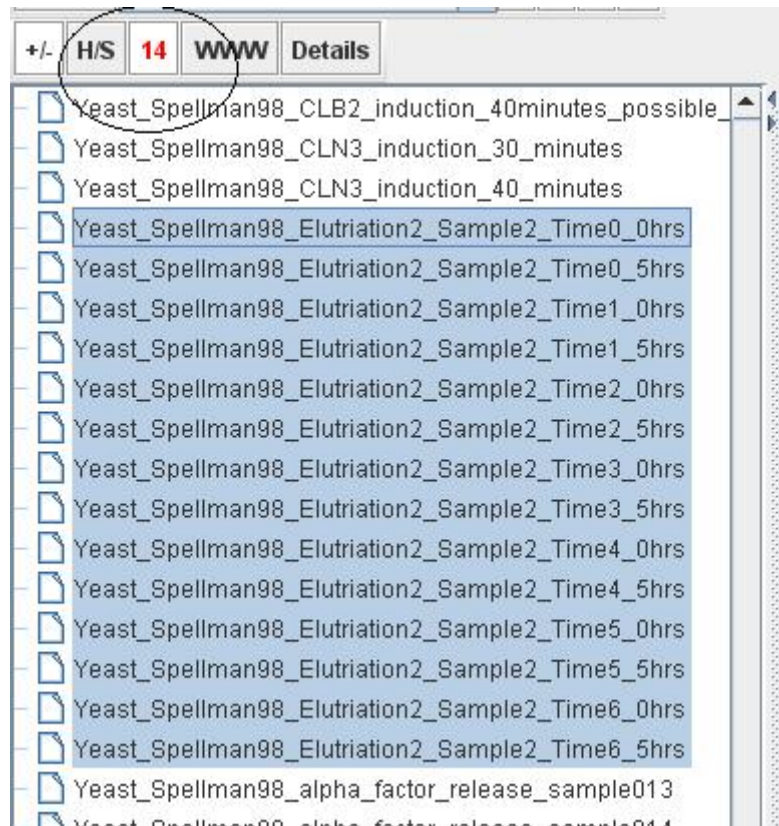
3) To see interactions between these genes, try broadcasting this list to **EMBL String** (a target in the Firegoose). Click on one of the edges of the resulting network to get details of the interaction. Do not close this tab; we'll need it again soon.

### Inspect Gene Expression Patterns

4) In the PIPE, click the "Gaggle Apps" tab and click the "Yeast MicroArray Data (from SMD)" link to open the Data Matrix Viewer (DMV) with data from the Stanford MicroArray Database.

5) When the DMV loads, expand the navigation tree on the left to the bottom level.

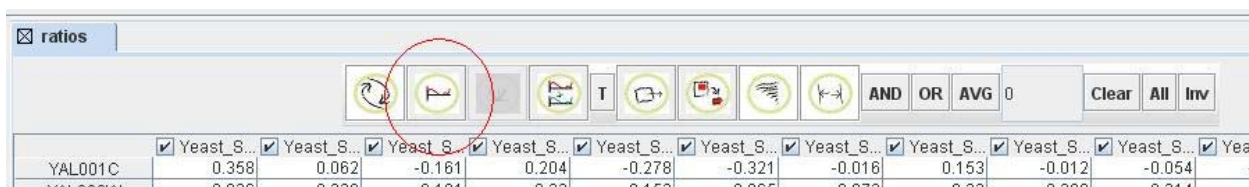
6) Highlight the following experiments, and then press the red "14" button to load the data pertaining to these experiments (it takes a moment to load):



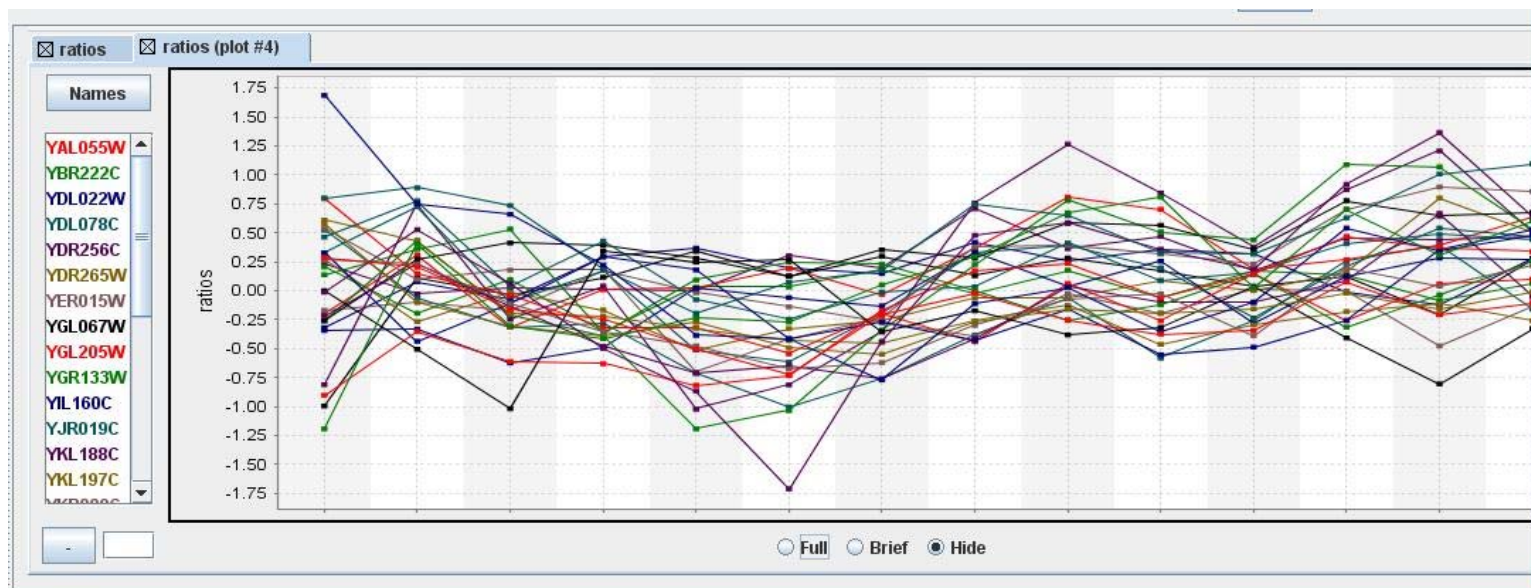
Next we'll want to see the plots of gene expression of the peroxisomal proteins identified in the previous section over these 6.5 hours.

(From Wikipedia: Elutriation, also known as air classification, is a process for separating lighter particles from heavier ones using a vertically-directed stream of gas or liquid (usually upwards). This method is profoundly used for particles with size ( $>1 \mu\text{m}$ ) The smaller or lighter particles rise to the top (overflow) because their terminal velocities are lower than the velocity of the rising fluid.)

- 7) Return to the PIPE and navigate to the px.yeast.prots dataset.
- 8) Make sure your Firegoose has “px.yeast.prots: NameList(30)” in the Gaggle Data field and “DMV” in the target field. Then hit the Broadcast Button.
- 9) In the DMV window, those genes will now be selected. Press the “Plot Selected Rows” button:



Click on “Hide”, and you should see the following:



We see that the expression of these genes appear to be correlated across the 6.5 hours of elutriation.

## VI. Looking for Plausible SRM Targets

In this exercise, we will start with the list of 30 peroxisomal genes from previous exercises and try to find 1 likely candidate protein that we would have expected to see in our results, and that later might be targeted in future (SRM) experiments.

- 1) In the PIPE, click on the “Gaggle Apps” tab and click the “Yeast Protein-Protein interactions (GenelDs)” link to open up a cytoscape window containing Yeast two-hybrid protein interaction data. Don’t worry if it looks like a big hairball right now, we will make some sense out of it.

2) In the PIPE, make sure the px.yeast.prots dataset is displayed, and in the firegoose, you see “px.yeast.prots: NameList(30)” as the data source. Then select “Yeast PP” as the target and hit “Broadcast”

3) Now maximize the cytoscape window containing the Yeast PP data and notice at the bottom it tells you that 17 nodes are selected. We are going to focus in on these nodes and their nearest neighbors.

4) From the “Select” menu bar, go Nodes -> First neighbors of selected nodes.

5) Now let’s hide all other nodes: Select -> Nodes -> Invert Selection.  
Then: Select -> Nodes -> Hide selection

6) re-layout the network with Ctrl-L.

Now we see our proteins and their first neighbors with experimentally verified interactions linking them together. Let’s first see which of these proteins are our proteins, and which were added when we expanded the network.

7) To do this, all we need to do is re-broadcast our list of proteins, and they become highlighted in the Cytoscape network. Return to PIPE, select “px.yeast.prots: NameList(30)” as the data source in the Firegoose, “Yeast PP” as the target, and hit broadcast. Now return to the Yeast PP cytoscape window.

Question: After exploring this network a little while, where our proteins are selected (gray) and other proteins that were not observed in our experiment are white, which white protein seems like a likely candidate for further investigation and possible targeting with SRM in the future? \_\_\_\_\_

### **See how EMBL STRING reports the interactions between this protein and our set of proteins.**

Open the Firefox tab containing the STRING network. Click the browser back button until you get to the page with the list of proteins that were originally submitted. Add your new protein to the end of the list and hit “GO !”. Follow the links to the network.

How many connections do you see between this new protein and our original list of 30?  
\_\_\_\_\_

### **See how this genes expression during elutriation matches up with the rest of the genes.**

## PIPE and Gaggle -- Tutorial

Back on the “Yeast PP” Cytoscape window, add the new protein to your selection by holding “Shift” and clicking it. Here you should have 18 selected proteins.

Broadcast it to “DMV”.

Go to the DMV window. The DMV application has selected the rows corresponding to the proteins you broadcasted. This now includes the new protein and the total number of selected proteins should be 31.

Click the “ratios” tab to bring up the spreadsheet containing all the expression levels.

Click the “plot selected rows” button once again to create a new plot, this time containing the new protein.

Find the protein in the list on the left and select it.

Do the expression levels of this new protein provide further evidence that this protein should be targeted in follow up SRM experiments? \_\_\_\_\_

That’s it. Thanks!