Proteomics Informatics Course
Gaggle Tutorial

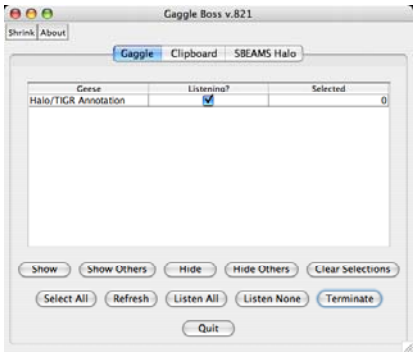
This tutorial, and the related exercises, explore many kinds of data for Halobacterium, including: proteomics and microarray measurements made in a time course and dosage experiment here at the ISB; a computationally-derived gene association network; KEGG and EMBL annotation, and NCBI Blastp results. The tutorial and exercises are concerned with exploring all of these data together.

The data and all of the bioinformatics tools used here are available from your web browser, with one exception: the statistical programming environment, ***R***, which is already installed on your laptop, and launches by double-clicking.

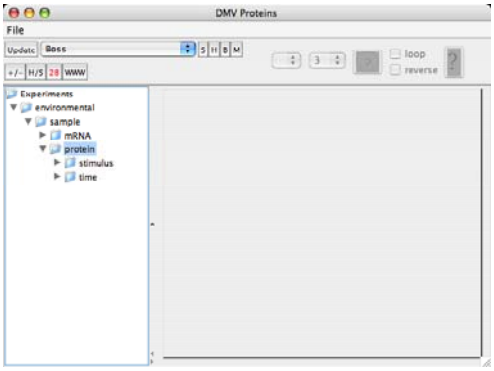
In this tutorial we will explore five Halobacterium succinate dehydrogenase proteins. All the steps described below are also presented in the companion web page, along with the Java Web Start links to launch the applications with their data. Please see:

http://gaggle.systemsbiology.net/projects/demos/proteomics-2006-oct

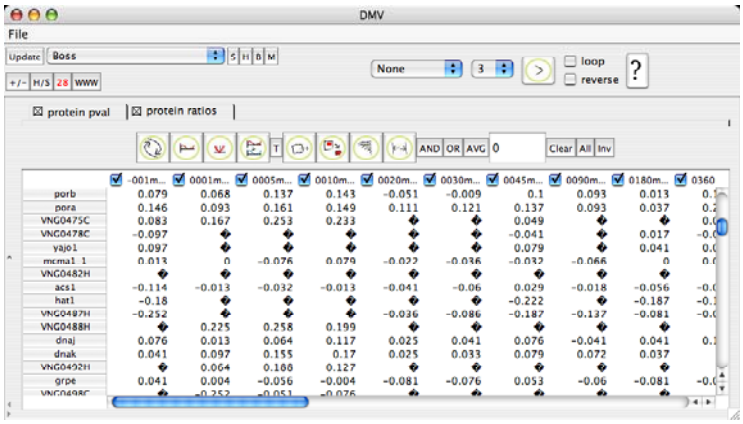
To begin, start the gaggle boss and wait for it to appear on your screen before you proceed. Now launch the ***Annotation Goose*** which can search Halo genes and proteins by orf name, gene name, and function. In its search box, type **sdh** and click the **Search** button. Five genes will be displayed:



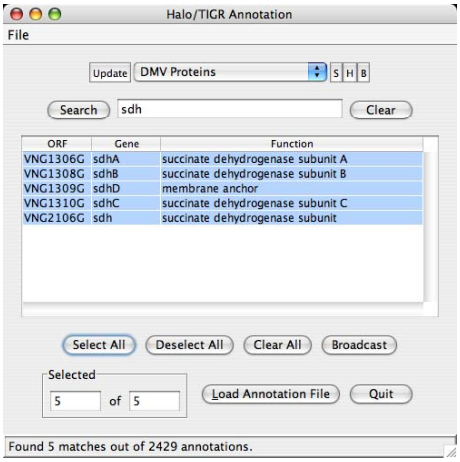
Now launch the ***DMV*** – the ***D***ata ***M***atrix ***V***iewer. Once it appears, expand the nested directory on the left until the ***environmental-sample-protein*** hierarchy is visible, then select ***protein*** and observe that **28** conditions (28 columns) of data are available:



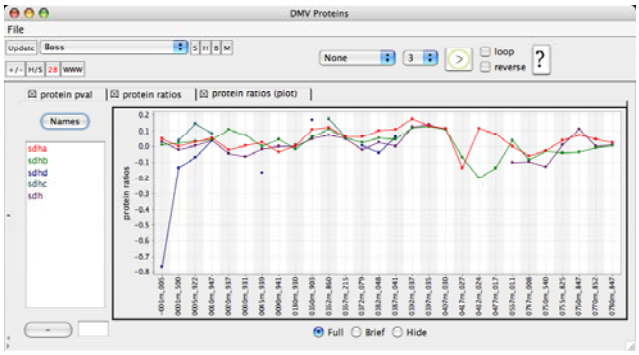
Now load these conditions by pressing the red button; hide the condition choosing panel; make sure you see a table like this:



Return now to the Annotation goose, select **DMV Proteins** as the target goose, and broadcast the five succinate dehydrogenase gene names to the **DMV.**



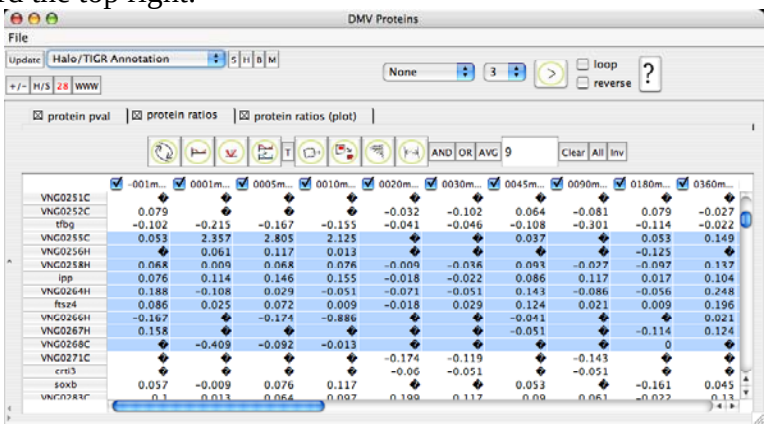
Returning to the **DMV,** and making judicious use of the tooltips on the buttons (which explain what the buttons do) plot protein ratios against experimental condition for the five rows selected by your broadcast, above, from the annotation goose.



Interlude: Broadcasting and other Gaggle Operations

It is worthwhile to spend some time practicing with the gaggle before exploring the data further. Try the following things:

1. In any goose, find the ***Update*** button, and click it. The goose contacts the gaggle boss to obtain an up-to-date list of the current geese.
2. This action updates the 'goose list' -- a pulldown menu to the immediate right of the ***Update*** button in every goose. Once you have refreshed that list via the update, you can then choose any goose as your current *target goose;* all subsequent broadcasts will be sent there. You can – in addition to broadcasting -- **S**how and **H**ide that target goose. This is very handy when, as is often the case, you have many geese running at the same time.
3. Bring the annotation goose (labeled **'Halo/TIGR Annotation'** unless you renamed it) to the foreground. Press its ***Update*** button. **S**how the boss, and in the boss's list of geese, edit the name of either goose, press the **Return** key to make sure the name change is registered, and then observe that your goose has the new name in its title bar.
4. Now raise the **DMV.** Click on the '**protein ratios**' tab. Select any number of rows in the ratios table; you must click or drag in the portion of the table where numbers are displayed, not on the protein names in the left colum. Note that the total number of rows currently selected is displayed toward the top right:

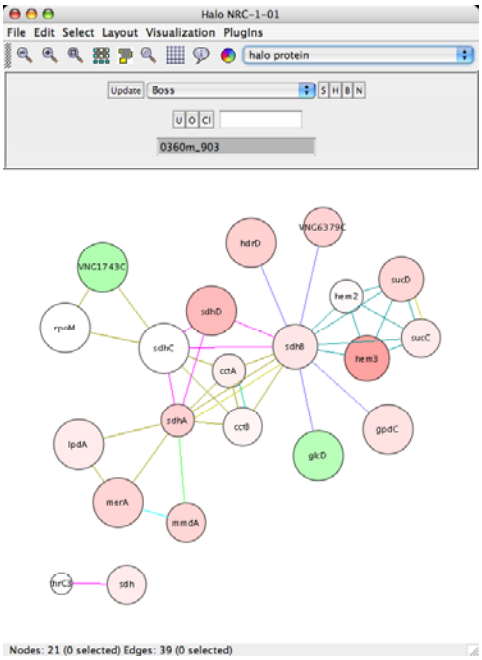


5. Broadcast these selected genes to the annotation goose. (You may wish to first raise the annotation goose, and **Clear** any names which may currently be in its search box. This ensures that the new search you will perform will only be for the names you just selected).
6. The **DMV** has many features; the tooltips on the buttons will help you to understand them. Experiment with the correlation finder, the plotter, the sub-matrix selector, and the various buttons which manipulate the current selection.
7. We strongly recommend that all broadcasts go to a specific goose. It is possible, however, to broadcast explicitly to the boss, and to rely upon the current settings of the boss to determine who hears that broadcast. Experiment with this by exploring the buttons in the boss.

Cytoscape: Add Network Views to the Gaggle

The Baliga Lab has a collection of computationally-derived gene associations. These may be viewed and explored in Cytoscape. Click on the **NRC-1 network** link on the web page to launch. Once it appears on your screen, try the following steps:

1. Find the succinate dehydrogenase proteins in the annotation goose. Update the goose selector, choose **Halo NRC-1** as the target, and then broadcast these five names to Cytoscape.
2. **S**how the Cytoscape goose.
3. Extend the Cytoscape selection (currently those same 5 genes, whose count is displayed at the lower left of the Cytoscape window) by selecting their '*first neighbors*'. You can do this by pressing **Ctrl-F** or from the **Selection->Nodes** menu; 21 nodes should now be selected.
4. Use **Ctrl-N** or the **Selection->To New Window** menu to place this selection in a new Cytoscape window; it will have a name **NRC-1-01** (or something similar). In this new window, make sure that the current visual style is '*halo protein*' – this ensures that the protein ratios and p values, which we are about to broadcast from the DMV to this small network, will be displayed in Cytoscape with appropriate colors and node sizes (indicating ratios and p-values, respectively of the proteins).



5. Return to the **DMV**, set the proper target goose, set the movie '*loop*' control, and a 1-second frame rate. Press the '**Run Movie'** button, raise the Cytoscape window, and watch the succinate dehydrogenase proteins change in color and size (ratio and p-value) as the **DMV** cycles through the 28 experimental conditions.

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

CONCLUSION

These exercises introduced the basics of the Gaggle. You have used three geese, each of which is good at presenting and manipulating (essentially) one kind of data. Be sure that you are very comfortable with these geese, and the various gaggle operations which connect them. You do not need to use the

boss often, but please experiment with it also. This practice prepares you for exploring other data using other geese -- which are available at the tutorial's web page, and which we will now briefly describe.

1. **The Firefox goose** (or **'FireGoose'**). We have written an extension to the popular Firefox web brower, which at present is primarily useful for broadcasting gene (or protein) names to and from KEGG. This extension is already installed in your course laptop; you will recognize the Gaggle toolbar near the top of the Firefox window.
2. **TIGR MeV.** We often broadcast experimental data – that is, numerical matrices representing mRNA or protein ratios – to MeV, to take advantage of its many capabilities, especially those for clustering data. The names of genes or proteins in selected clusters, if deemed interesting, may be broadcast back to the gaggle.
3. ***R.*** Though not for the faint of heart, the statistical and bioinformatics capabilities of the ***R*** programming environment are very powerful. We broadcast gene lists, numerical matrices, networks and associative arrays between ***R*** and other geese. We will demonstrate this briefly during the exercises portion of the afternoon.
4. **EMBL STRING WBI.** This goose will soon be retired in favor of the STRING capabilities we are adding to the FireGoose, but for right now, this is still useful. We typically broadcast gene names to STRING, find out what associations may exists, expanding the network as needed. Network 'edges' are annotated by STRING to include the experimental, computational, or Pubmed evidence they adduce for the association of any two proteins.
5. **Blastp.** The Halobacterium 'sequence goose' looks up amino acid or nucleotide sequences for names it receives via a gaggle broadcast. If the **auto** button is checked (which we recommend), and if **Blastp** is the sequence goose's target goose, then the sequence is sent automatically to NCBI's blastp website, and your browser displays a filled-in blastp form, poised to do the search when you press a button in the browser.