Exploring Biological Networks with Cytoscape Software

Natalie Yeung,[1] Melissa S. Cline,[2] Allan Kuchinsky,[3] Michael E. Smoot,[4] and Gary D. Bader[1]

[1]University of Toronto, Donnelly Centre for Cellular and Biomolecular Research, Toronto, Ontario, Canada
[2]Department of Molecular, Cell, and Developmental Biology, University of California, Santa Cruz, California
[3]Agilent Technologies, Santa Clara, California
[4]Department of Bioengineering, University of California, San Diego, La Jolla, California

ABSTRACT

Cytoscape is a free software package for visualizing, modeling, and analyzing molecular and genetic interaction networks. As a key feature, Cytoscape enables biologists to determine and analyze the interconnectivity of a list of genes or proteins. This unit explains how to use Cytoscape to load and navigate biological network information and view mRNA expression profiles and other functional genomics and proteomics data in the context of the network obtained for genes of interest. Additional analyses that can be performed with Cytoscape are also discussed. *Curr. Protoc. Bioinform.* 23:8.13.1-8.13.20. © 2008 by John Wiley & Sons, Inc.

Keywords: network visualization • network analysis • systems biology • protein interactions • biological network

INTRODUCTION

Cytoscape is a free, open-source software platform used to graphically visualize biological networks (Shannon et al., 2003). Networks contain nodes, representing objects (such as proteins), and connecting edges representing relationships between them (such as physical interactions). Importantly, Cytoscape allows the integration of experimental and other relevant data—such as Gene Ontology annotation (*UNIT 7.2*) and gene expression profiles—stored as node and edge attributes, in a network context. These can be mapped to visual attributes (such as node shape or edge color) allowing data to be visualized in a network context in many useful ways.

The most basic Cytoscape task is to visualize a network created from interaction data (Basic Protocol). Expression data can then be loaded as node attributes and visualized on the network by mapping node attributes to node colors (Alternate Protocol). These tasks are graphically summarized in the protocol flowchart (Fig. 8.13.1). Many analyses can be performed using Cytoscape plug-ins, which are downloadable extensions of the main Cytoscape software. Plug-ins add functionality, e.g., fetching network data from public sources and analyzing network topology to find biologically interesting patterns. Cytoscape can be downloaded for desktop use on Windows, Mac OS X, and Linux machines (Support Protocol 1). Other UNIX platforms that support recent versions of Java are also supported.

All protocols and figures shown use Cytoscape 2.5.2 (the most recent version, as of July 2007).

Current Protocols in Bioinformatics 8.13.1-8.13.20, September 2008
Published online September 2008 in Wiley Interscience (www.interscience.wiley.com).
DOI: 10.1002/0471250953.bi0813s23

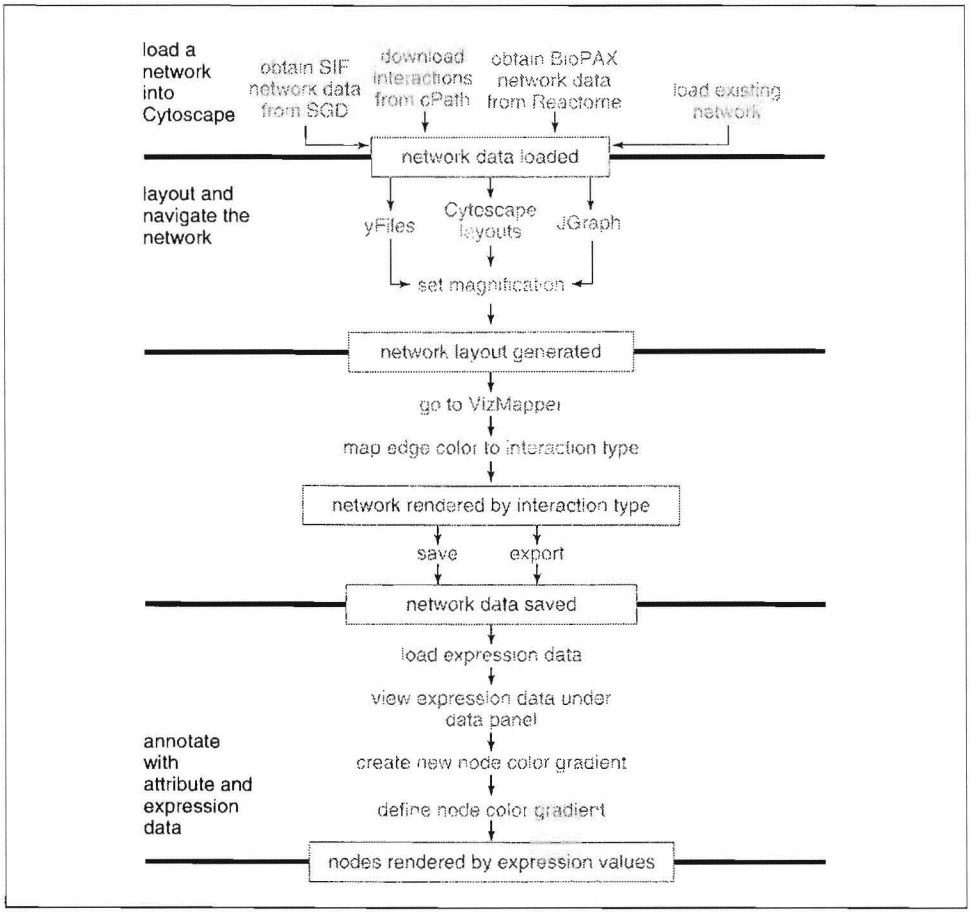

Figure 8.13.1 Flowchart summarizing the protocols defined in this unit.

BASIC PROTOCOL

VISUALIZE A NETWORK

This protocol outlines the steps necessary to create, lay out, and view networks in Cytoscape, along with tips for navigating the network and setting custom visual properties. Four network data loading methods are described; the first three involve downloading network data from online databases, while the fourth describes loading an existing local file.

Necessary Resources

Hardware

Computer with 1 GHz CPU or higher, a high-end graphics card, 60 MB of available hard disk space, at least 512 MB of free physical RAM (for networks up to 5000 edges; at least 1 GB of RAM (for larger networks) and a minimum screen resolution of 1024 × 768 (recommended; requirements depend on the size of the networks to be imported and analyzed)

Internet connection to obtain network data from online databases (not necessary to visualize interaction data from a local file)

Software

Operating System: Windows, Mac OS X, Linux, or another platform that supports Java

Java 2 Platform, Standard Edition, version 5.0 or higher (*http://java.sun.com/javase/downloads/index.jsp*).

Internet browser: e.g., Microsoft Internet Explorer (*http://www.microsoft.com*), Mozilla Firefox (*http://www.mozilla.org/firefox*), or Apple Safari (*www.apple.com/safari*), if downloading network files

Table 8.13.1 File Formats Supported by Cytoscape

File format	Description	Related URL
SIF (Simple Interaction Format)	Text format invented for Cytoscape (see Fig. 8.13.2)	*http://www.cytospace.org*
CYS (Cytoscape session file)	Default Cytoscape file format, containing both interaction data and visual properties	*http://www.cytospace.org*
GML (Graph Markup Language)	Standard network file format supported by multiple generic network software packages	*http://www.infosun.fml.uni-passau.de/Graphlet/GML*
XGMML (eXtensible Graph Markup and Modeling Language)	Standard XML format similar to but preferred over GML, since it can contain more information	*http://www.cs.rpi.edu/~puninj/XGMML*
SBML (Systems Biology Markup Language)	Standard XML format for representing mathematical pathway models	*http://sbml.org/documents*
PSI-MI (Proteomics Standards Initiative-Molecular Interaction format)	XML standard format for molecular interactions supported by molecular interaction databases	*http://www.psidev.info/index.php?q = node/60*
BioPAX (Biological PAthway eXchange)	Standard format for pathway information supported by multiple pathway databases	*http://www.biopax.org*

Cytoscape 2.5.2, downloaded from *http://cytoscape.org* (see Support Protocol 1 to install a local copy)

Files

No external files required for downloading network data from online databases
Local files (if used): e.g., Microsoft Excel (`.xls`) or text files containing interaction data arranged in columns; example files available in the Cytoscape/sampleData folder created during Cytoscape installation (Support Protocol 1), some of which can be opened, viewed, and edited in a plain text editor such as Notepad or TextEdit; see Table 8.13.1 for standard supported file formats

Load a network into Cytoscape

1. Open Cytoscape by clicking the Cytoscape icon created during installation.

   This step is for users who used the automatic install program. See Support Protocol 1 for the installation procedure and a description of the user interface.

   For alternate methods of opening Cytoscape, see Support Protocol 1, step 5.

2. The Cytoscape desktop will appear (Fig. 8.13.3).

   The toolbar at the top of the desktop contains command buttons with tooltips (the name of the function will appear when the mouse hovers over the button for more than a few seconds). The center of the screen, which is blank when Cytoscape is started, will display networks as they are loaded.

   At the left of the screen is the Control Panel, which has four major tabs: the Network tree viewer, the VizMapper, a network Editor, and a basic Filters function.

   The Network tree viewer displays a list of all loaded networks and the number of nodes and edges that they contain. It also contains the Network overview panel at the bottom of the tab, which shows the current network with a blue box highlighting the portion currently being viewed.

   The VizMapper controls the node, edge, and global network visual properties of Cytoscape networks, and facilitates user-defined mapping of attribute data to visual properties.

galFiltered.sif

```
YMR043W pd YGR108W
YMR043W pd YDR461W
YMR043W pd YNL145W
YMR043W pd YJL157C
YMR043W pd YFL026W
YHR171W pp YDR412W
YHR171W pp YNR007C
YNR050C pp YMR138W
YJL203W pp YOL136C
YLL021W pp YLR362W
YGL013C pd YJL219W
YGL013C pd YOL156W
YKL101W pp YBR160W
YCL067C pd YIL015W
YCL067C pp YMR043W
YCL067C pd YDR461W
YCL067C pd YFL026W
YBL005W pd YJL219W
YDR323C pp YOR036W
YGL008C pd YMR043W
YER040W pd YPR035W
YER040W pd YGR019W
YNL098C pp YLR310C
YGL115W pp YGL208W
YPL089C pp YHR030C
YJL089W pd YER065C
```

Figure 8.13.2 A few lines from the `galFiltered.sif` protein interaction network file included in the `Cytoscape/sampleData` directory. The first and last columns contain node IDs, while the middle column defines an edge type.

3. Obtain new or load existing network data, using one of the following methods:

   a. Obtain yeast network data from the Saccharomyces Genome Database (SGD, Christie et al., 2004; Support Protocol 2).

   b. Obtain network data using the cPath database (Support Protocol 3).

   c. Obtain a biological pathway from the Reactome database (Support Protocol 4).

   d. Load an existing network data file (Support Protocol 5).

4. Click Import to load the network. Cytoscape will display a progress screen as it loads the data.

5. Check that loading status is successful, and then click Close.

   Steps 1 through 4 can be repeated multiple times, i.e., many networks can be loaded in separate Cytoscape windows.

6. To switch between networks, click on the filenames in the Network tab of the Control Panel (Fig. 8.13.3). Note that only one session can be loaded at a time.

Layout and navigate the network

7. If a network is not displayed after the data is successfully loaded, create a view of the network by selecting the Edit→Create View menu option.

   Small networks will have a view automatically created when they are loaded, while large networks (i.e., thousands of nodes and edges) will be loaded without a view. Larger networks are usually slower and harder to work with, due to their need for greater computational resources. However, they can be reduced to a selected subset of nodes and edges using the Filters function and then viewed as a smaller network. Filters are described in more detail in step 11e of this protocol.

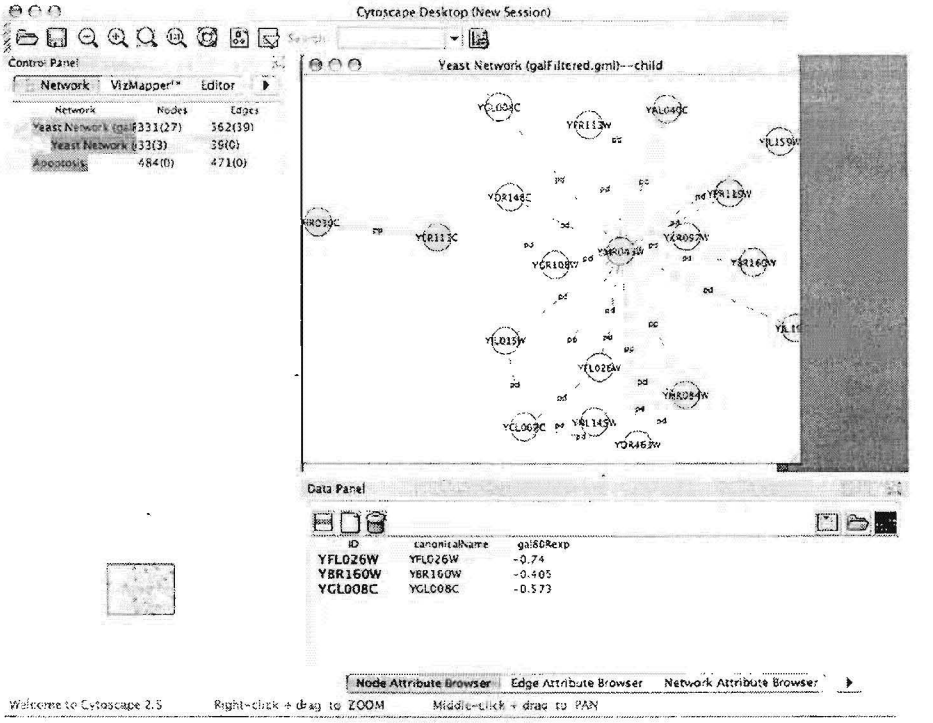

Figure 8.13.3 The basic Cytoscape user interface.

8. Apply a layout using the Layout menu. Applying a layout to a network moves the positions of nodes and edges to reduce overlap, provide a clearer visual representation of the data, and make the structure of the network more interpretable.

   Cytoscape offers a set of tools for automated layout, using a variety of algorithms, e.g., hierarchical, circular, and attribute-based layouts. In addition to automatic layout for the entire network, some of the tools optionally operate on selected parts of a network.

   Different layouts are tailored for different types of networks. Hierarchical layouts work better on tree-like networks, circular layouts work better if the network is circular, and force-directed type layouts—including the Cytoscape "Force-Directed" layout—are better for well connected networks. Force-directed layout algorithms model edges as springs and nodes as like-charged particles, so nodes repel each other and edges spring, but connect nodes at a preferred length. After a short simulation of this physical system, the layout produces a network layout where nodes do not overlap but are not too far away from each other.

   While most layouts do not consider information about the network other than the connectivity, some attribute-based layouts are available that place nodes and edges based on their attributes. Examples include using edge weights to calculate edge length or clustering nodes with common annotations together (Garcia et al., 2007).

   Networks can also be manually laid out by selecting single or multiple nodes and dragging them across the screen.

   The Rotate and Scale functions can be applied to the whole network or a subset of it, while Align, Distribute, and Stack (which allow aligning, evenly distributing, or stacking selected nodes in space on the canvas) require some or all nodes to be selected. To select nodes, click on each one while holding down the Shift key or click and drag to select an area containing the node set.

9. Adjust the viewing area and magnification of the network. There are six methods for navigating across a network:

   a. *Zoom out:* View a larger region of the network by clicking on the button depicting a magnifying glass with a minus (–) sign.

   b. *Zoom in:* View a smaller region of the network in greater detail by clicking on the button depicting a magnifying glass with a plus (+) sign.

   c. *Zoom to a selected region:* View a selected subset of the network by clicking on the button depicting a magnifying glass with a dotted rectangle.

   d. *View the entire network:* See the entire network at once by clicking on the button depicting a magnifying glass labeled 1:1.

   e. *Pan across the network:* View different portions of the network by clicking and dragging the blue box shown in the Network Overview in the lower left-hand corner of Cytoscape.

   f. *Continuous zoom:* Zoom in and out of a network by right clicking the mouse and dragging the mouse up and down over the network view.

10. Create a child network (new network containing a subset of the original parent) by selecting nodes and/or edges and then going to File→New→Network→From selected nodes, all edges. A second window will appear containing the new network.

11. Select nodes and edges of interest using one of the following methods:

    a. Hold down the Shift key while clicking on nodes and edges.

    b. Click and drag to select a region of the network.

    c. Use the options provided under Nodes and Edges in the Select menu.

    d. Use the Quick Find search box provided in the Cytoscape toolbar (see Fig. 8.13.4).

       Quick Find provides a fast way to select nodes or edges that share an attribute value or range of values. The default search is by node name, so typing the first few letters of a node name in the search box will bring up a list of all matching node names. Click on the configuration icon directly to the right of the search box to change the search to another node or edge attribute. For numerical attributes, the search box will change into a slider that allows the selection of a numerical range (see Fig. 8.13.5).

    e. Create and apply a filter using the Filters tab in the Control Panel.

       The Filters tab features a user-friendly interface that is also accessible from the funnel icon on the Cytoscape toolbar (see Fig. 8.13.4).

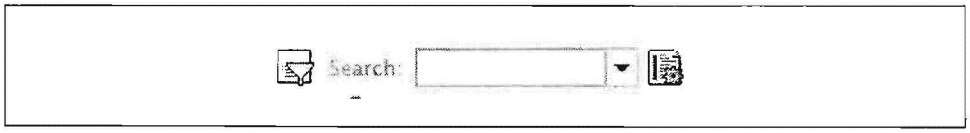

Figure 8.13.4 The funnel icon at the left of the figure opens the Filters tab in the Control Panel. Next to it is the Quick Find search box, and the Quick Find configuration icon is at the far right.

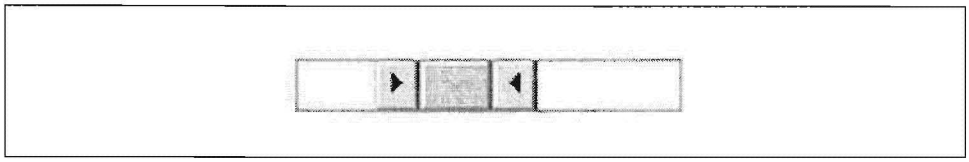

Figure 8.13.5 The Quick Find search box can filter numerical attributes by dragging the two triangles to define minimum and maximum values. All nodes or edges falling within this range will be selected.

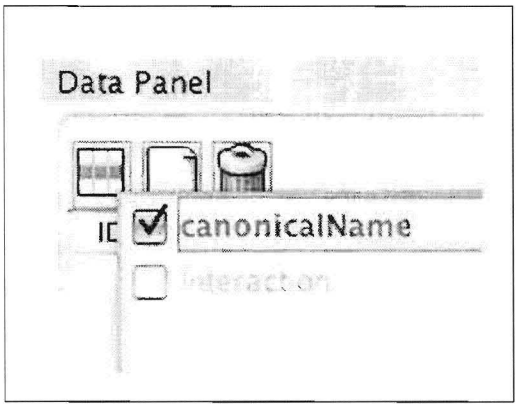

Figure 8.13.6 The Select Attributes icon is found at the far left of the Data Panel (the rectangle with a grey horizontal stripe). When clicked, a list of attributes appears. These will be displayed in the Data Panel if they are checked.

Filters serve as a more complex and flexible form of searching than Quick Find. Boolean and regular expression searches are supported, as well as all of the functionality available within Quick Find. More information on how to use filters can be found in the Filters chapter of the Cytoscape user manual (accessed from the Help menu online at http://www.cytoscape.org/cgi-bin/moin.cgi/Cytoscape_User_Manual; included in the Cytoscape installation directory).

In Cytoscape 2.5.2, these filters are restricted to AND *and* OR *Boolean expressions. Future versions of Cytoscape will extend this functionality.*

Set visual properties for network nodes and edges

12. Click on the Edge Attribute Browser tab in the Data Panel at the bottom of the screen to view the attributes associated with the edges in the network. By default, the edge ID (identifier) attribute is displayed.

    Selecting edges in the network will display their respective attribute values in the Edge Attributes tab. To view other attributes, click on the Select Attributes button at the left of the Data Panel to display a list of available attributes and highlight the desired ones by clicking on them (Fig. 8.13.6). Close the list by right-clicking or by clicking anywhere outside the box. To see the desired attributes for an edge, the edge must be selected in the view (see step 10, above, for selection details).

13. Open the VizMapper tab in the Control Panel and access VizMapper in one of three ways:

    Select the View→Open VizMapper menu option.
    Select the VizMapper icon in the main button bar.
    Click on the VizMapper tab in the Control Panel at the left of the screen.

    The VizMapper controls how visual properties, such as node or edge color, are assigned from attribute data.

14. Create a new visual style by clicking on the Options button at the top right of the VizMapper tab and selecting the Create new Visual Style . . . option. Enter a name for the new style.

    Once created, visual styles can be modified, saved, and applied to other networks. Alternatively, an existing similar visual style can be copied (using the Copy existing Visual Style . . . option) and then modified, which may take less time than defining a new one.

15. Set the colors of edges in the network to correspond to the type of interactions they represent:

    a. Select the visual attribute by double-clicking the Edge Color entry listed in the Unused Properties section of the Visual Mapping Browser. Edge Color will now appear at the top of the list, under the Edge Visual Mapping Category.

b. Select the network attribute by clicking on the cell to the right of Edge Color and choosing "interaction" from the drop-down list that appears.

c. Select an appropriate Mapping Type according to the data values of the network attribute; in this case, choose Discrete Mapper. All existing attribute values for "interaction" will then be displayed.

Two other attribute mapper types exist in addition to Discrete mappers. A Passthrough mapper directly passes through the attribute value to the visual attribute. This makes most sense for visual attributes such as labels. The other mapper type is the Continuous mapper, which maps a continuous data attribute to a continuous visual attribute, such as mRNA expression values mapped to a node color gradient.

d. Set the mapping relationship. Click the empty cell next to one of the interaction values. Buttons marked "..." (more detail) and "X" (delete) will appear on the right side of the cell. Click on the "..." button and select a color from the color palette. The change will immediately appear on the network.

A different color can be assigned for each value of the network attribute that exists.

This procedure can also be used to map any node data attribute to any node visual property.

Save and export the network

16. Save the network using the Save or Save As ... options in the File menu. This saves the entire Cytoscape session, including the network and all its Node and Edge attributes, as a Cytoscape-specific `.cys` file, which can then be opened for further viewing or editing at a later time.

Cytoscape session files can also be shared with collaborators or as supplementary material for a paper for viewing or editing.

17. Export the network as an image file using File→Export→Network View As Graphics.

A number of standard image types are supported. PDF format is recommended for publication-quality figures. Other options available include exporting the network to a standard interaction data file type for use in other software packages.

18. Exit Cytoscape by selecting File→Quit.

SUPPORT PROTOCOL 1

INSTALLING CYTOSCAPE LOCALLY

This support protocol provides instructions for downloading and installing Cytoscape, along with an introduction to the various components of its user interface.

Necessary Resources

Hardware

Computer with 1 GHz CPU or higher, a high-end graphics card, 60 MB of available hard disk space, at least 512 MB of free physical RAM (for networks up to 5000 edges; at least 1 GB of RAM (for larger networks) and a minimum screen resolution of 1024 × 768 (recommended; requirements depend on the size of the networks to be imported and analyzed)

Internet connection (required to download Java and Cytoscape)

Software

Operating System: Windows, Mac OS X, Linux, or another platform that supports Java

Internet browser: e.g., Microsoft Internet Explorer (*http://www.microsoft.com*), Mozilla Firefox (*http://www.mozilla.org/firefox*), or Apple Safari (*www.apple.com/safari*)

Files

None required

1. If not already installed, download and install the Java 2 Platform, Standard Edition, version 5.0 or higher (*http://java.sun.com/javase/downloads/index.jsp*).
2. Go to *http://cytoscape.org* and click on the link marked All Releases, then Download Cytoscape 2.5.2 at the top right of the screen.
3. Accept the terms of the Lesser GNU Public License (LGPL), fill in the user registration form, and click the Proceed to Download button.
4. Click on the appropriate installation package to download it, and then double-click on the downloaded icon to start the installation process. Note that the directory in which Cytoscape is installed will be the directory in which Cytoscape initially starts.

   The installation package is roughly 40 MB in size and may take some time to download on slower Internet connections.

5a. *To launch Cytoscape:* Click on the icon created in the Cytoscape installation directory.

5b. *To launch Cytoscape by an alternative means:* Open in Windows by double-clicking `cytoscape.bat` or `cytoscape.jar`, or open in Linux and Mac OS X by running `cytoscape.sh` directly from the command line with various parameters described in the Cytoscape user manual (accessed from the Help menu online at *http://www.cytoscape.org/cgi-bin/moin.cgi/Cytoscape_User_Manual*; included in the Cytoscape installation directory).

6. The Cytoscape desktop will appear (Fig. 8.13.3).
7. To exit Cytoscape, use the File→Quit menu option.

OBTAIN YEAST NETWORK DATA FROM SACCHAROMYCES GENOME DATABASE (SGD)

SUPPORT PROTOCOL 2

The SGD provides physical and genetic interactions for yeast, which may be downloaded as a Cytoscape SIF file (see Table 8.13.1).

Necessary Resources

See Basic Protocol

1. Launch Cytoscape as in Basic Protocol, step 1, and go to *http://db.yeastgenome.org/cgi-bin/batchDownload* and scroll down to the section labeled Step 1: Your Input.
2. Under Enter Feature/Standard Gene names, enter a gene symbol such as `PPA2`.
3. Under Step 2, under the section labeled Other data, check the boxes for physical and genetic interactions and click Submit. The Web browser will be redirected to a page labeled Download Data.
4. Note the SIF filename near the right side of this page at a link labeled Name of Downloadable File, and click on the link to download the `.sif` file.
5. If the file is not automatically uncompressed during download, uncompress it.
6. Continue to Support Protocol 4, step 3 (Load an existing network data file) to import contents of the `.sif` file into Cytoscape.

SUPPORT PROTOCOL 3

OBTAIN NETWORK DATA USING THE cPath DATABASE

Another useful resource for Cytoscape data is the cPath database and Cytoscape plug-in (Cerami et al., 2006). Currently, the Cytoscape cPath plug-in draws data from the MINT (Zanzoni et al., 2002; *UNIT 8.5*) and IntAct (Hermjakob et al., 2004) databases.

Necessary Resources

See Basic Protocol

1. Launch Cytoscape as in Basic Protocol, step 1, and go to File→New→Network→Construct network using cPath . . . A window should appear, as shown in Figure 8.13.7.
2. Select the desired species in the species pull-down menu, which is set to All Organisms by default.
3. In the box labeled Search cPath, enter a gene name (e.g., `p53`) and click on the Search button. Cytoscape will produce a network similar to the one shown in Figure 8.13.8 (shown with the JGraph radial layout).

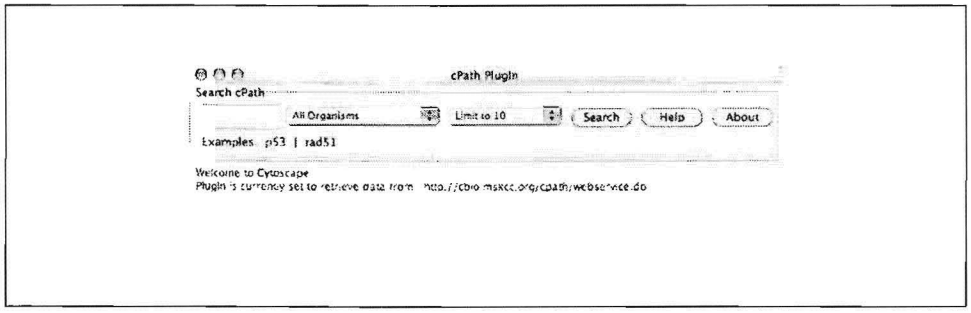

Figure 8.13.7 The cPath Cytoscape plug-in searches the MINT and IntAct databases to automatically import network data into Cytoscape.

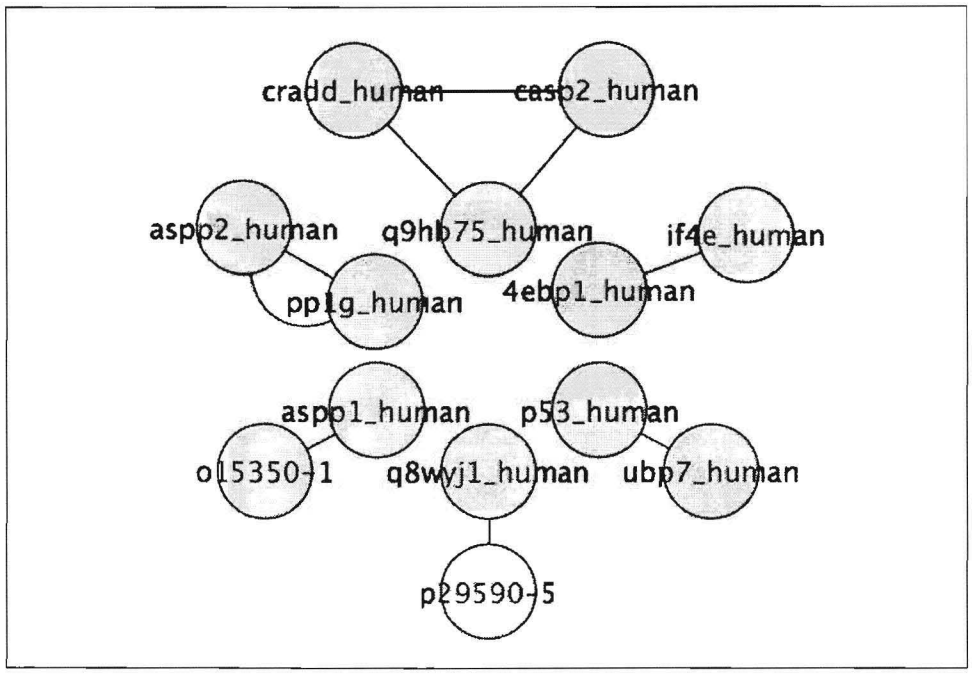

Figure 8.13.8 A sample Cytoscape network created using the cPath plug-in to search for p53 in Homo sapiens. The JGraph radial layout was applied.

The maximum number of records is set to Limit to 10 by default. While the default setting is useful for exploratory queries with a single gene of interest, a larger number of records must typically be retrieved to achieve connectivity between a set of genes of interest. Note that the number of interactions retrieved may be greater than the limit set, because many database records contain more than one protein interaction. In these cases, all proteins in the interaction are connected to each other, up to an internal threshold set in the cPath plug-in.

4. To obtain all interactions for this gene set, select No Limit, remembering that a higher limit will result in a longer download time. The Cytoscape canvas will show a protein interaction network with proteins (nodes) arranged in a grid, connected by retrieved interactions (edges).

   cPath searches can include other attributes, such as diseases (e.g., lymphoma) and biological processes (e.g., apoptosis). Search terms can also be combined using the `AND` *and* `OR` *operations (e.g.,* `p53 AND apoptosis`*).*

5. By default, Cytoscape displays networks with 10,000 or fewer nodes because large networks take a long time to draw. For larger networks, request a view by right-clicking on the network label in the Network Tree Viewer and select Create View from the pop-up menu.

OBTAIN A BIOLOGICAL PATHWAY FROM THE REACTOME DATABASE

SUPPORT PROTOCOL 4

The Reactome database (*UNIT 8.7*; Joshi-Tope et al., 2005) is a biological pathway database containing curated human information, along with inferred orthologous pathways in a number of other species. It provides pathways in a number of formats (see Table 8.13.1), including BioPAX (*http://www.biopax.org*).

Necessary Resources

See Basic Protocol

1. Launch Cytoscape as in the Basic Protocol, step 1, and go to the Reactome home page at *http://reactome.org*. The default species displayed in the reaction map is *Homo sapiens*.
2. Click on the drop-down list immediately above the map to change species if necessary.
3. Select a pathway by clicking on its image in the reaction map or the labels underneath. A summary page will appear.
4. Scroll down to the bottom and click the link marked [BioPAX] to download a Reactome file (extension `.owl`) containing the pathway data.
5. Continue to Support Protocol 5 (Load an existing network data file) to import the `.owl` file.

LOAD AN EXISTING NETWORK DATA FILE

SUPPORT PROTOCOL 5

This protocol provides different procedures for a number of file formats.

Necessary Resources

See Basic Protocol

1. Launch Cytoscape as in Basic Protocol, step 1.

2a. *To open a Cytoscape session file (* `.cys`*):* Go to File→Open. Select the session file and click Open.

2b. *To open a text or Excel file:* Go to File→Import. . .→Network from Table (Text/MS Excel). Select the appropriate file using the Select File button, and define the importing options and data columns, with the help of the preview at the bottom of the dialog box.

Since free-format tables contain user-defined columns instead of a standard format, a preview window is provided to indicate how Cytoscape will interpret the input data. Once a file is selected, the first few lines of the file contents will be shown.

If Cytoscape is not parsing files correctly, it may be necessary to change the advanced settings. Check the box marked Show Text File Import Options to display these settings.

Drop-down menu lists are available for specifying the columns containing Source Nodes (purple), Interaction/Edge Type, (red), and Target Nodes (orange). The preview will color-code each column accordingly; blue is used to indicate columns that will be interpreted as edge attributes.

Note that node attributes must be imported separately. Any data columns that are not to be loaded into Cytoscape can be disabled by clicking on the header (Column X, where X is the column number) in the preview. A Reload button is provided to refresh the preview after making any changes (Fig. 8.13.9).

Edge attributes can be subdelimited within a column by right-clicking on the column header and selecting the List option as the Attribute Data Type. For example, this might be used to indicate PubMed records relevant to the interactions. Select or enter the appropriate List Delimiter and click OK. Note that this sub-delimiter must be different from the delimiter used to separate columns.

For Excel users: Only single-sheet workbooks are currently supported.

2c. *To open a different supported network file type (SIF, GML, XGMML, SBML, PSI-MI, BioPAX; see Table 8.13.1)*: Go to File→Import→Network (Multiple File Types). Select the appropriate file and click Open.

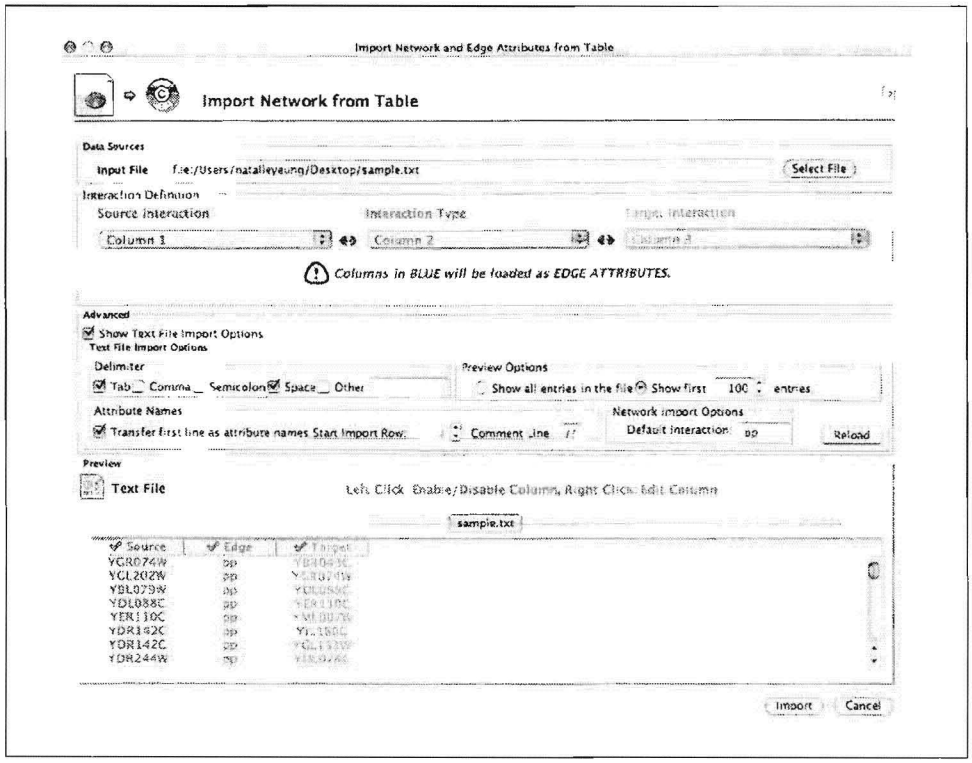

Figure 8.13.9 The window that appears when importing an Excel or delimited text network file.

2d. *To open files from the local hard drive:* Select Local Data Source Type (this is the default) and choose the file using the Select button. Selecting Remote Data Source Type allows files to be loaded from the Internet by typing in the URL or using Cytoscape bookmarks.

The directory in which Cytoscape is installed contains a folder called sampleData. This folder holds a number of example files containing published experimental data. References for these data are available in the Cytoscape user manual (accessed from the Help menu online at http://www.cytoscape.org/cgi-bin/moin.cgi/Cytoscape_User_Manual; included in the Cytoscape installation directory).

INTEGRATE EXPRESSION DATA

ALTERNATE PROTOCOL

Cytoscape offers the ability to combine network data with expression data, which can provide information about network dynamics over time or across different experimental conditions. This alternate protocol outlines the process of loading expression data and then visualizing it on an existing network.

In order to import attribute files or expression data into Cytoscape, the gene or protein identifier in the file must exactly match the corresponding Cytoscape node ID (or other Cytoscape attribute that has been previously loaded). If no matching identifiers are present, additional identifiers can be created using external online ID mapping services such as Synergizer (*http://llama.med.harvard.edu/cgi/synergizer/translate*), provided by the Roth laboratory at Harvard University.

Necessary Resources *(also see Basic Protocol)*

Files

Network files, downloaded (see Basic Protocol, step 3)
Expression data files, created locally: currently supported expression data file formats include Excel spreadsheets and delimited text (tab, comma, or space delimiters), along with standard file extensions such as `.mrna` and `.pvals` (see Figs. 8.13.10 and 8.13.11; also see the Expression Data chapter of the Cytoscape user manual)

NOTE: To use this protocol as a tutorial, go to the Cytoscape/sampleData folder to select `galFiltered.sif` as the network file and `galExpData.pvals` as the expression data file.

1. Launch Cytoscape, then load and layout a network (see the Basic Protocol, steps 1 to 8).

For standard file formats

2a. Load an expression data file by going to the drop-down list and selecting the attribute that is the same in both the network and expression data files. Click Import.

3a. A status window will appear showing the number of experimental conditions found and information on significance values (if found in the file). Click the Close button.

For nonstandard file formats (e.g., text and Excel)

2b. Load an expression data file by using the File→Import→Attribute from Table (text/MS Excel . . . option).

3b. This will pull up a window similar in operation to the one used to import text and Excel network files (see Basic Protocol, step 3d). Be sure that the values in the column labeled Key (blue) exactly match those of a column in the network file. Click the Close button.

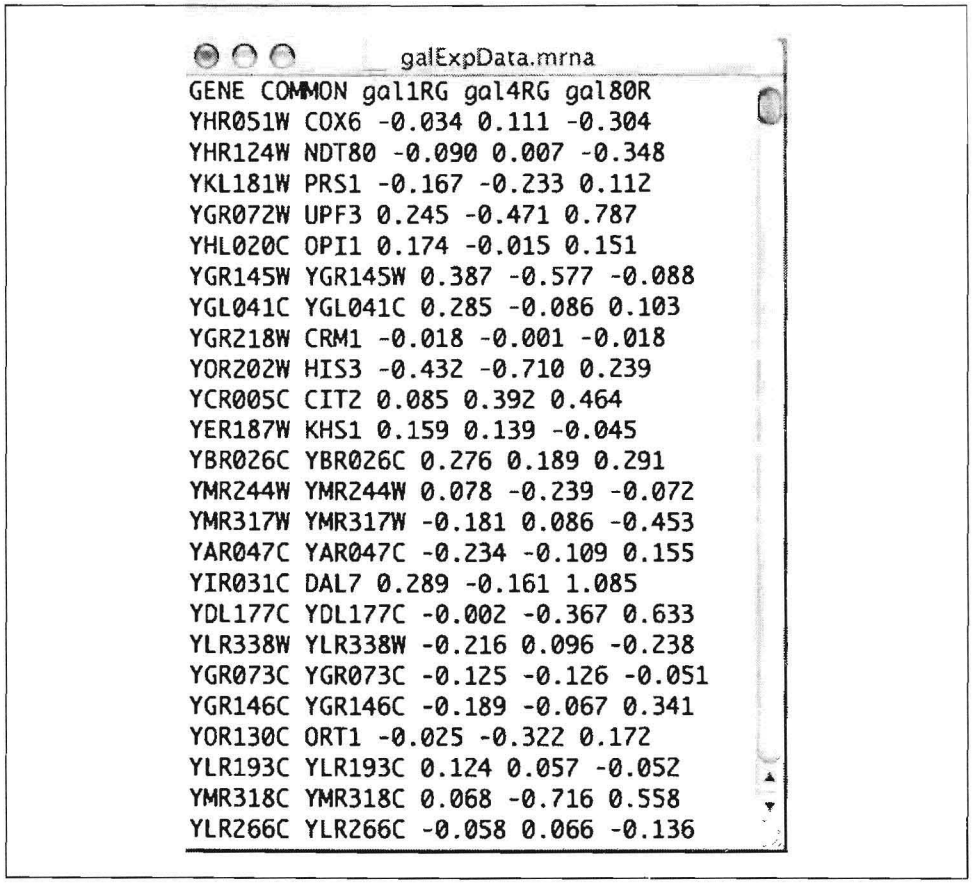

Figure 8.13.10 The first few lines of `galExpData.mrna`, a sample expression data file. The first row is a header row. The first column contains gene names, and the second has the common names for each gene, followed by expression level data from three experimental conditions. The first column is mapped to the node IDs in the network unless otherwise specified.

4. View the expression data by going to the Node Attribute Browser tab in the Data Panel and displaying the experimental conditions of interest (see Basic Protocol, step 9).
5. Open the VizMapper and copy the default visual style (see Basic Protocol, steps 12 to 15).
6. Define a node color gradient that corresponds to experimental expression data to create multiple mappings for visualizing multiple data attributes (see Basic Protocol, steps 12 to 15, for more detail):
   a. Select Node Color.
   b. Define the Map Attribute value as one of the experimental conditions (e.g., `Gal80RGexp` in the `galExpData.pvals` sample file).
   c. Select Continuous Mapping as the Mapping Type.
   d. Double-click on the white rectangle next to Graphical View to open the Color Gradient Mapper. This dialog is used to define the points where colors will change.
   e. Click the Add button twice to create the first two boundary points. Additional clicks will add boundary points that will show up as overlapping triangles at the right of the scale.

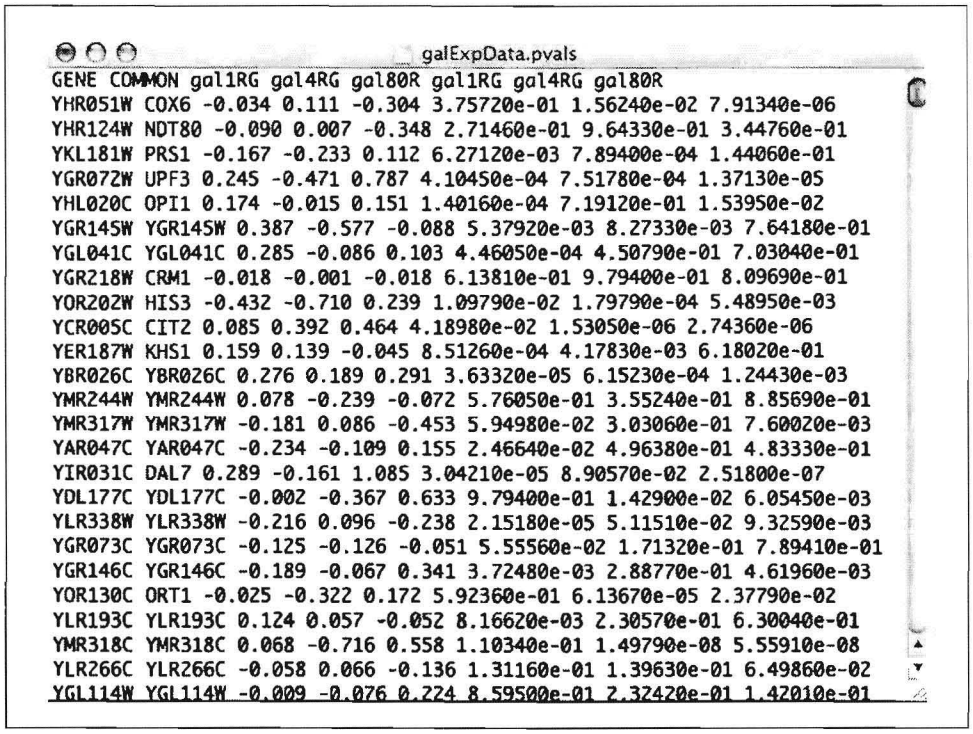

Figure 8.13.11 The first few lines of `galExpData.pvals`, an expression data file included in the `Cytoscape/sampleData directory`. The first row is a header row. The first column contains gene names, and the second has the common names for each gene. The next three columns contain expression level data from three experimental conditions. The last three columns contain the significance or *p*-values associated with each piece of experimental data. Note that the *p*-value columns must contain exactly the same headers in the same order as the data columns in order for Cytoscape to associate the *p*-values with the data.

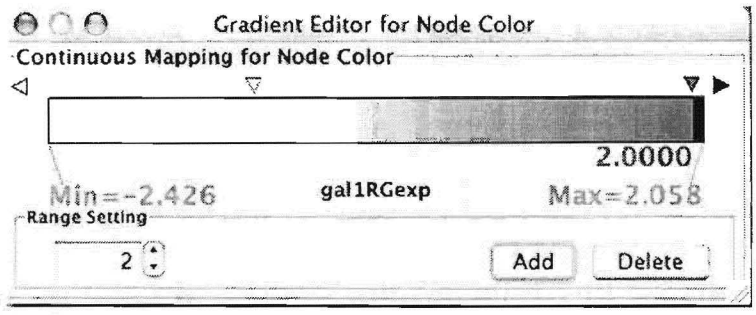

Figure 8.13.12 An example color gradient created to map node color to the `gal80RGexp` condition in the `galExpData.pvals` sample file.

f. Click and drag each triangle to define the boundaries between colors on the scale, or type the desired value in the Range Setting box. The values shown on the scale correspond to the existing values for the experimental condition that has been chosen. To delete boundary points, use the Delete button, but note that at least two boundary points must exist (not including the two default extremes) in order to create a gradient.

g. Define the color gradients between boundary points by double-clicking on the triangles at each of the endpoints, in turn, to open a color palette. Select a color

and click OK. Nodes with expression levels of this value will be colored with the selected color.

Nodes that have expression levels in between the two boundary points will be rendered with a color in between the two boundary colors. For example, if the lower boundary point is white and the upper boundary point is red, nodes with expression values in between the two points will be colored pink, with darker shades of pink indicating higher expression values. When the boundary colors are set, the colors of the nodes will be updated immediately (see Fig. 8.13.12).

GUIDELINES FOR UNDERSTANDING RESULTS

The protocols provided here can stand alone as methods for analyzing biological networks and also serve as a starting point for more in-depth analysis using various Cytoscape analysis plug-ins. Plug-ins can be downloaded for use directly from Cytoscape (via Plug-in → Manage Plug-ins) or online at *http://cytoscape.org/plugins2.php*.

The Basic Protocol, which produces a two-dimensional network, can be used to infer certain biological properties based on topology. For instance, critical genes and proteins tend to be hubs (nodes connected to many other nodes) or part of the shortest path through the network between two other nodes (Yu et al., 2007). Plug-ins such as PeSca and ShortestPath implement shortest path algorithms for use in Cytoscape. Additional plug-ins are available for creating networks, e.g., the Agilent Literature Search plug-in, which extracts relationships about given genes or proteins automatically from multiple online sources, including PubMed (Vailaya et al., 2005).

Certain network data formats include explicit nodes denoting modules or complexes, e.g., the BioPAX Reactome networks. For networks without this information represented, it is possible to infer complexes by searching for groups of nodes with a high degree of internal connectivity (interactions amongst themselves) compared to external connectivity (interactions with nodes outside the group). Putative complexes can be identified visually, or automatically using Cytoscape's MCODE plug-in (Bader and Hogue, 2003).

The Alternate Protocol superimposes expression data on a network, which can result in some interesting biological insights. Combining expression and interaction data is a procedure sometimes performed to find causative disease agents when comparing control and case samples for clinical studies. While the causative agents might not exhibit dramatic expression changes themselves, one can often see significant and coordinated variation of expression (co-expression) in genes regulated by the causative agents. Using the network as a visual aid to find common neighbors of co-expressed genes is therefore an effective method of finding possible causative agents. This process can be automated by the Active Modules plug-in, which finds active regions of a network across multiple experimental molecular profile measurements (Ideker et al., 2002).

More generally, a plausible biological explanation for co-expression of genes or proteins is functional relatedness. This is especially true in prokaryotes, where functionally-related genes may be organized into the same operons in the genome. Genes involved in a complex can exhibit just-in-time assembly, where one highly regulated critical gene controls the overall activity of the entire complex (de Lichtenberg et al., 2005). Comparing different expression patterns across experimental conditions can also reveal different mechanisms that cause the same end result. The BiNGO plug-in finds significantly over-represented Gene Ontology terms annotated to the genes of interest. This helps identify functions enriched in a set of genes, including sets of genes that are co-expressed (Maere et al., 2005). Additional Cytoscape tutorials explaining these uses in more detail are available from the Cytoscape Web site (*http://www.cytoscape.org*).

COMMENTARY

Background Information

Biological network visualization is an important tool in systems biology. While traditional reductionist biology focuses on a single gene or protein, systems biology focuses on the interplay of multiple genes or proteins: how they form regulated subsystems and how changes in experimental conditions affect subsystem behavior. While systems biology can include mathematical modeling of network dynamics, network visualization is arguably the most common method of modeling systems; it does not require detailed measurement of subsystem dynamics and can suggest information about gene function and impact of gene loss or transcriptional repression. A typical biological pathway presents enough complexity that it is difficult for the human mind to process new observations in the context of the whole pathway. Visualization offers a straightforward mechanism to assess the new observations and existing data together.

Biological network data comes in two major forms: curated pathways and interaction networks. Curated pathways describe sequences of intermolecular interactions that yield some measurable result. Examples include converting organic compounds into energy (*metabolic pathways*); transmitting an extracellular signal into the nucleus, resulting in transcription (*signal transduction pathways*); or transcribing a set of genes after production of the necessary transcription factors (*regulatory networks*). Curated pathway repositories contain descriptions of pathways, derived from a combination of the literature and experimental verification. Major pathway repositories include KEGG (Wixon and Kell, 2000), Reactome (Joshi-Tope et al., 2005), and BioCyc (Krummenacker et al., 2005); additional repositories are listed in Pathguide (Bader et al., 2006). These are rich sources of information, describing the context and consequences of each interaction, but they are limited in coverage. In general, they describe basic cellular processes that are highly conserved between organisms and certain processes involved in well studied diseases. In other instances, where the data is sparser, protein interaction networks can be a useful alternative.

Protein interaction networks contain nodes representing proteins and edges representing experimentally measured interactions between the proteins. Interactions are potential associations; they may occur in a cell if the proteins are both present and in the correct modification states. Major interaction data repositories include IntAct (Hermjakob et al., 2004), MINT (*UNIT 8.5*; Zanzoni et al., 2002), BIND (*UNIT 8.9*; Bader et al., 2001), DIP (Xenarios et al., 2002), and HPRD (Peri et al., 2004); additional repositories are listed in Pathguide (Bader et al., 2006).

The most common type of interaction data is measured using the yeast two-hybrid method, a genetic technique for detecting pairs of proteins that can interact. This technique has been adapted for high-throughput use and now represents the majority of interaction data (Hermjakob et al., 2004).

Other interaction data comes from biochemical purification experiments (e.g., co-immunopreciptation, pull-down, and tap-tagging assays), which have also been used in high-throughput studies. While these assays report interactions that may occur in the cell, they do not report which of the proteins were in direct physical contact. Rather, they find a set of proteins that likely represent a population of complexes. Thus, for such data, interaction is interpreted as membership in the same complex. The contrast between these two types of interaction data illustrate why different types of interactions demand slightly different interpretation. Thus, when analyzing interaction networks, it is useful to distinguish the varying interaction types.

Another element of this protocol is coloring nodes according to expression data. First of all, this provides a visual indication of what portions of the network might be produced, indicating where an interaction might occur in a protein interaction network or where there might be a missing element in a pathway. Expression data can provide further information on network dynamics. For example, when several genes are part of the same complex, the complex might not be active until all genes are expressed (de Lichtenberg et al., 2005). Finally, there are cases where functionally-related proteins are produced from co-expressed genes. Prokaryotic genomes contain operons, sections of DNA that contain genes and are transcribed together as a unit, and genes in the same operon tend to be functionally-related. Yet even in eukaryotes, genes that are co-expressed in multiple species and experimental contexts tend to be functionally-related (Stuart et al., 2003).

Altogether, biological network visualization is highly useful for integrating multiple data types in the context of known biological

processes. While biological network visualization has been discussed in this unit, Cytoscape is capable of handling any type of network. As long as the data can be represented as sets of nodes and edges, Cytoscape can display the data as a network. For example, the StructureViz Cytoscape plug-in allows the user to compare related protein structures under the Chimera protein structure viewer, while a Cytoscape network relates the protein structure(s) to others in the same structural family (Morris et al., 2007).

Critical Parameters and Troubleshooting

Out of memory errors

Symptoms: Cytoscape behaves strangely. Java null pointer exception error messages may appear, or there will be no reported error but the expected action does not occur.

Possible causes: This type of problem will occur when Cytoscape tries to analyze very large networks or when a number of other applications are also running on the computer.

Remedies: Make more memory available to Cytoscape by closing unnecessary networks and applications, rebooting the computer, or increasing Cytoscape's memory allocation on the computer (see *http://cytoscape.org/cgi-bin/moin.cgi/How_to_increase_memory_for_Cytoscape* for details).

Data integration errors

Symptom: Expression or attribute data files are not properly integrated with the loaded network.

Possible causes: The gene identifier columns that synchronize the two files do not match exactly, or the files may not be in the correct format.

Remedies: Use the Node or Edge Attribute tabs (see Basic Protocol, step 12) to check that the network identifiers exactly match the identifiers in the expression or attribute data file. To determine the correct format of an attribute or expression file, see the Web sites provided in Table 8.13.1.

Large networks

Symptoms: The network loads without an automatically generated view, or the dataset is so large that effective analysis is difficult.

Cause: The loaded network is very large.

Remedies: Cytoscape can create views for large networks (see Basic Protocol, step 7), and child networks can also be created (see Basic Protocol, step 9) to create a smaller and more manageable network.

Acknowledgments

Cytoscape is developed through an ongoing collaboration between the University of California at San Diego, the University of Toronto, the Institute for Systems Biology, Memorial Sloan-Kettering Cancer Center, Institut Pasteur, Agilent Technologies, and the University of California at San Francisco. We gratefully acknowledge the contributions of many Cytoscape developers: Nada Amin, Mark Anderson, Iliana Avila-Campilo, Richard Bonneau, Ethan Cerami, Rowan Christmas, Michael Creech, Benjamin Gross, Kristina Hanspers, Larissa Kamenkovich, Ryan Kelley, Sarah Killcoyne, Nerius Landys, Samad Lotia, Andrew Markiel, John Morris, Keiichiro Ono, Owen Ozier, Alexander R. Pico, Paul Shannon, Robert Sheridan, Aditya Vailaya, Jonathan Wang, Peng-Liang Wang, Chris Workman, and principal investigators Annette Adler, Bruce R. Conklin, Leroy Hood, Trey Ideker, Chris Sander, Ilya Schmulevich, Benno Schwikowski, and Guy J. Warner.

Many research groups have developed plug-ins to Cytoscape and provided them for download free of charge from *http://www.cytoscape.org*. These plug-ins represent key contributions to the overall utility of Cytoscape, and we gratefully thank the authors for their contributions. Thanks to Vuk Pavlovic for editing help.

Funding for Cytoscape is provided by the U.S. National Institute of General Medical Sciences of the National Institutes of Health under award number GM070743-01. Corporate funding is provided through a contract from Unilever PLC. Cytoscape contributions by G.D.B. were funded in part by Genome Canada through the Ontario Genomics Institute.

Literature Cited

Bader, G.D. and Hogue, C.W. 2003. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 4:2 (*http://www.biomedcentral.com/1471-2105/4/2*).

Bader, G., Donaldson, I., Wolting, C., Ouellette, B., Pawson, T., and Hogue, C. 2001. BIND—The Biomolecular Interaction Network Database. *Nucleic Acids Res.* 29:242-245.

Bader, G.D., Cary, M.P., and Sander, C. 2006. Pathguide: A pathway resource list. *Nucleic Acids Res.* 34:D504- D506.

Cerami, E.G., Bader, G.D., Gross, B.E., and Sander, C. 2006. cPath: Open source software for collecting, storing, and querying biological pathways. *BMC Bioinformatics* 7:497 (*http://www.biomedcentral.com/1471-2105/7/497*).

Christie, K.R., Weng, S., Balakrishnan, R., Costanzo, M.C., Dolinski, K., Dwight, S.S., Engel, S.R., Feierbach, B., Fisk, D.G., Hirschman, J.E., Hong, E.L., Issel-Tarver, L., Nash, R., Sethuraman, A., Starr, B., Theesfeld, C.L., Andrada, R., Binkley, G., Dong, Q., Lane, C., Schroeder, M., Botstein, D., and Cherry, J.M. 2004. Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms. *Nucleic Acids Res.* 32:D311-D314.

de Lichtenberg, U., Jensen, L.J., Brunak, S., and Bork, P. 2005. Dynamic complex formation during the yeast cell cycle. *Science* 307:724-727.

Garcia, O., Saveanu, C., Cline, M., Fromont-Racine, M., Jacquier, A., Schwikowski, B., and Aittokallio, T. 2007. GOlorize: A Cytoscape plug-in for network visualization with Gene Ontology-based layout and coloring. *Bioinformatics* 23:394-396.

Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., Margalit, H., Armstrong, J., Bairoch, A., Cesareni, G., Sherman, D., and Apweiler, R. 2004. IntAct: An open source molecular interaction database. *Nucleic Acids Res.* 32:D452-D455.

Ideker, T., Ozier, O., Schwikowski, B., and Siegel, A.F. 2002. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* 18:S233-S240.

Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G.R., Wu, G.R., Matthews, L., Lewis, S., Birney, E., and Stein, L. 2005. Reactome: A knowledgebase of biological pathways. *Nucleic Acids Res.* 33:D428-D432.

Krummenacker, M., Paley, S., Mueller, L., Yan, T., and Karp, P.D. 2005. Querying and computing with BioCyc databases. *Bioinformatics* 21:3454-3455.

Maere, S., Heymans, K., and Kuiper, M. 2005. BiNGO: A Cytoscape plug-in to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* 21:3448-3449.

Morris, J.H., Huang, C.C., Babbitt, P.C., and Ferrin, T.E. 2007. structureViz: Linking Cytoscape and UCSF Chimera. *Bioinformatics* 23:2345-2347.

Peri, S., Navarro, J.D., Kristiansen, T.Z., Amanchy, R., Surendranath, V., Muthusamy, B., Gandhi, T.K., Chandrika, K.N., Deshpande, N., Suresh, S., et al. 2004. Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res.* 32:D497-D501.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. 2003. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13:2498-2504.

Stuart, J.M., Segal, E., Koller, D., and Kim, S.K. 2003. A gene-coexpression network for global discovery of conserved genetic modules. *Science* 302:249-255.

Vailaya, A., Bluvas, P., Kincaid, R., Kuchinsky, A., Creech, M., and Adler, A. 2005. An architecture for biological information extraction and representation. *Bioinformatics* 21:430-438.

Wixon, J. and Kell, D. 2000. The Kyoto encyclopedia of genes and genomes—KEGG. *Yeast* 17:48-55.

Xenarios, I., Salwínski, L., Duan, X., Higney, P., Kim, S.M., and Eisenberg, D. 2002. DIP, the Database of Interacting Proteins: A research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* 30:303-305.

Yu, H., Kim, P.M., Sprecher, E., Trifonov, V., and Gerstein, M. 2007. The importance of bottlenecks in protein networks: Correlation with gene essentiality and expression dynamics. *PLoS Comput. Biol.* 3:e59 (*http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.0030059*).

Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G., Helmer-Citterich, M., and Cesareni, G. 2002. MINT: A Molecular INTeraction database. *FEBS Lett.* 513:135-140.

Key References

Bader et al., 2006. See above.

Pathguide provides an extensive list of electronic pathway resources, both public and private, along with references and URLs for each.

Shannon et al., 2003. See above.

This article provides further background on Cytoscape, and the questions that it was first developed to address.

Internet Resources

http://www.cytoscape.org

The home page of the Cytoscape project contains download links, the latest manual, plug-ins, online tutorials, and links to the Cytoscape discussion forums and project development wiki.

http://java.sun.com

This is the central Internet resource for Sun Java, with download links, documentation, and software development packages. Java must be installed for Cytoscape to run. Most computers already have Java installed.

http://www.yeastgenome.org

The Saccharomyces Genome Database, available at this site, contains a wealth of yeast genomic and experimental data, tools, and resources for the study of yeast; in particular, the SGD maintains a large database of yeast interaction data, and provides this data in formats including Cytoscape SIF format.

http://www.reactome.org
Reactome provides curated pathway data for many of the key pathways in humans, and over twenty other organisms.

http://llama.med.harvard.edu/cgi/synergizer/translate
The Synergizer offers an effective, usable solution to one of the most frequent and frustrating problems in computational molecular biology: identifier mapping.