BioGrapher Version 4.21 BETA for WinXP and Mac OS 10.5
(Workbook for Excel 2003, 2004, 2007)

May 29, 2008

*****USE AT YOUR OWN RISK****

------>PLEASE TAKE A MOMENT TO READ THE "READ ME FIRST FOR INSTALL" DOCUMENT FOR INSTALLATION INSTRUCTIONS AND KNOWN ISSUES--GraphViz calls will not work and graphs will not be drawn unless the GraphViz packages are installed in the right locations in the directory tree! All the files and executables needed by BioGrapher (including the AT&T GraphViz binaries) for WinXP/PC are now placed in a single self-contained BioGrapher FOLDER. In general, this folder should have been placed in a directory for which the user has full access (including write) privileges. BioGrapher works well off a flash drive. However, on a Mac running OS 10.5, the latest version 2.18 of the AT&T GraphViz package must have been installed (done when the BioGrapher installer script is executed), and this installation places the GraphViz binaries in /usr/local/bin and uses Mac system libraries (including XCODE libraries) in /usr/lib. It is also possible to use a special version of the BioGrapher software for older Mac OS (10.3.9 through 10.4.x) if you are able to download and access the older 2.16 version of GraphViz. See installation instructions for details.

----->This workbook uses MACROS, whose use MUST be enabled via the standard TOOLS--->MACRO--->SECURITY Excel menu item in Excel 2003 (or in Excel 2007 through the menu bar/ribbon alert that is displayed) when the workbook is first launched. In addition, the user must select "Enable Macros" from the security dialog box that pops up each time the workbook is opened when the "medium" security setting is specified. Finally, note that the custom menu items installed for this workbook will not work after the user enters a value into a cell until the user hits the enter key or clicks on a different blank cell.

This work was originally funded in part by the BioQUEST Consortium (project BEDROCK) at Beloit College (PI Professor John Jungck) and is currently supported by NSF-DUE-CCLI Grant #0633651.

USE AT YOUR OWN RISK!

PURPOSE: A simple Excel-based workbook with worksheets as a front end for the AT&T GraphViz Graph Layout software suite. BioGrapher enhances Excel-based tools developed in the Chemistry and Biology Departments at Beloit College to allow for convenient visualization of graphs and graphical connections that are important in systems and computational biology, molecular biology, biochemistry, and genetics.

DESCRIPTION: The Excel spreadsheet allows for graphs (currently UNDIRECTED graphs and trees ONLY) to be specified in the standard adjacency matrix notation (also in NEWICK notation for trees, as well as in a condensed NODE LIST notation, see separate section at the end of this document)--diagonal elements of a square matrix represent nodes, non-zero [usually 1] off-diagonal elements represent connected edges. Since undirected graphs can be represented by a symmetric matrix, only the upper half of the matrix needs to be entered as 1's and 0's. The rest of the graph can be automatically filled in by invoking 'Complete Undirected' (see below). Once data are entered or imported (see below) into the spreadsheet, visualization of the graph in a number of different layout styles, each of which can provide unique insights into the nature of the graph and connections

between nodes, is achieved using a form window that provides access to some very powerful graphical layout routines (GraphViz software) available from AT&T in the public domain under the Common Public License (CPL).

ORGANIZATION of Excel Workbook: The workbook has ***seven worksheets***, whose functions are described below. In addition, it has two special menus (CALCULATE and GRAPHICAL VISUALIZATION) that are added to the Excel menu bar and must be used for all operations on the spreadsheets, INCLUDING LOADING AND SAVING DATA in comma separated value (CSV) Excel and text-compatible format.

---- > It is important that names and the order of the worksheets not be changed, else the code will fail!

1. DATA MATRIX (or NODE LIST, as appropriate): FIRST ENTER THE ORDER OF THE MATRIX in cell (1, 1) of WORKSHEET "INFO," and then enter the adjacency matrix here as 1's and 0's, starting with cell (1, 1). Immediately after the last column for a particular row (node), it is possible to specify a label for the node (optional, else row numbers are used as labels by default), followed by another optional column where a background COLOR (red, blue, green, yellow, cyan, orange, or magenta) for a node label text box can be specified. Only half the symmetric matrix for undirected graphs needs be entered, and the rest of the graph can be completed by invoking the 'Complete Matrix (Undirected)' menu item of the CALCULATE menu.

NEW FEATURE---->If a **NODE LIST** representation of the graph, which is merely a list of node labels, one node per row, with node labels of nodes connected to a given node (row) placed in columns, is to be used, you must enter the number of nodes in cell (1, 1) of worksheet "INFO" first, along with the keyword "LIST" in cell (1, 2), i.e., row 1 and column 2, of worksheet "INFO." On the other hand, note that cell (1, 2) is blank, i.e., has no keyword, if an adjacency matrix is to be specified. Sorry, no colors can be specified for nodes in a node list.

2. GRAPH: This worksheet displays the graph, laid out using the options invoked in the SHOW GRAPH menu item of the GRAPHICAL VISUALIZATION menu. The graph is a vector object and can be displayed and printed on any scale without the "jaggies." It can also be copied and pasted into other worksheets, spreadsheets, or MSWord. Note that you can zoom in and out using CONTROL Z and entering a value!

3. GRAPH PROPERTY: This worksheet is reserved for display of numerical and graphical results of the computation of various mathematical properties of the graph using the CALCULATE menu. See the description of the CALCULATE menu for more details. Note that the name of this worksheet will change depending on properties computed. These calculations work only with adjacency matrices and WILL NOT WORK WITH NODE LISTS.

4. INFO: This special worksheet is reserved for saving temporary global data and for use in future versions of BioGrapher. IN PARTICULAR, A KEY VALUE THAT MUST FIRST BE ENTERED INTO THIS WORKSHEET BEFORE ENTERING THE ADJACENCY MATRIX IN THE "DATA MATRIX" WORKSHEET IS THE ORDER OF THE MATRIX (in cells(1,1)). IN ADDITION, THE KEYWORD "LIST" MUST BE ENTERED INTO CELL (1, 2) (First Row, Second column) if the NODE LIST

representation of the graph is being entered or used. Finally, the keyword "NEWICK" must be entered into cell (1, 3) if the NEWICK representation of the tree where anchor nodes are not displayed is to be displayed. Note that a separate menu item that allows for direct entry of a graph using the standard Newick representation with nodes names separated by commas and anchor nodes indicated by parentheses is also provided. BioGrapher will automatically convert such a representation into a node list and store it in Sheet 1, which will then be labeled NODE LIST.

5. PIXEL MATRIX: This worksheet displays a "pixel" representation of entered data (black squares are ones and white squares are zeroes), and is automatically updated to reflect changes in the data matrix. Again, you can zoom in and out using CONTROL Z and entering a value.

6. FOR USER (1): Can be used be used as a scratch worksheet.

7. FOR USER (2): Can be used as a scratch worksheet.

--- > IT IS IMPORTANT THAT ORDER AND NAMES OF WORKSHEETS NOT BE CHANGED—THE WORKBOOK IS APPROPRIATELY PROTECTED!

All manipulations of data are carried out using **TWO ENHANCED AND SPECIAL MENUs** that has been added to the standard Excel menu bar. [In Excel 2007, use the ADD INs standard menu tab to display these menus.] Note that the menus are visible only when you click on a blank cell and NOT when you have a chart or drawing/shape selected! Here is a brief description of the menus and various menu items:

+++GRAPHICAL VISUALIZATION+++

DRAW GRAPH will draw a graph of the data using a specified layout (method of representing the graph's topology) from the ATT GraphViz suite. The default layout is the circular layout. The spreadsheet remembers the last layout style selected and uses it for subsequent displays. Invoking this menu item displays a form with the various choices, and choices for font sizes for node (vertex) labels. Note that computation of the layout may take some time (even minutes!) depending on the size and complexity (connectivity) of the graph to be displayed, especially for circular layout. The displayed graph can be conveniently zoomed using Control-Z from the keyboard. Also note that the displayed graph is a (vector--no jaggies at any resolution) shape object that can be copied and pasted into other worksheets and into Word documents! Finally, you can select from a number of fixed font sizes, provided as radio button options. However, clicking on the ADVANCED button in the form will also bring up more options, including manual entry of a specific font size (currently restricted to sizes from3 pt to 23 pt.), changing label text color to blue (default black), using plain text instead of the default bold, and finally, drawing thicker lines.

DRAW GRAPH FROM NEWICK NOTATION allows the user to enter a TREE description that conforms to Newick specifications into a user form, and draws a graph using the TREE layout. The user-entered expression is saved in the form, while a NODE LIST is generated and stored in Sheet 1, which is relabeled NODE LIST, while the number of nodes and the LIST keyword are automatically entered into the INFO sheet. By default, the anchor nodes are hidden in the displayed graph, but can be explicitly shown as labels by removing the NEWICK keyword from cell (1, 3) of the INFO sheet and then (re)DRAW GRAPH, the menu item described previously.

When the user clicks on the OK button in the form where the Newick expression is entered/imported, a dialog box is displayed to permit the optional specification of a

user-selected node as the root node. The name of the node must be entered exactly (case sensitive) as it was displayed in the Newick specification. An easy way to specify a node is to copy (drag and select with CONTROL-C on BOTH Mac and PC) a node name in the form before clicking on OK and then paste into the dialog box using CONTROL-V on a PC and COMMAND-V on a Mac. *(There is no mistake in this description, and non-standard [for a Mac] use of Control-C in form to copy is a Mac Excel bug!)*

Note that the user form displayed by invoking the NEWICK NOTATION menu item also has options that permit import and export of NEWICK format text files.

LOAD MATRIX DATA allows for data representing a graph in the adjacency matrix notation described above to be imported into the spreadsheet. The data have usually been SAVED (see below) in CSV format previously after data entry in BioGrapher, although the data could also have been generated by another program or have been manually typed in as a standard file. The only constraint is that the file must follow the specific labeling format (look at a sample file using a word processor), and must conform to standard Excel CSV (comma separated values) format in terms of delimiters and must have a .CSV extension, i.e., the file name must be something like yourgraphdata.CSV

SAVE MATRIX DATA saves previously loaded or entered data in a CSV format compatible with BioGrapher and readable by the LOAD DATA menu item. YOU MUST SAVE DATA PRIOR TO QUITTING BioGrapher. IF YOU DO NOT, ALL DATA, GRAPHS, COMPUTATIONS, ETC., WILL BE LOST! CURRENTLY, ALL YOU CAN SAVE IS THE ADJACENCY MATRIX/NODE LIST IN CSV FORMAT. NO GRAPHS AND OTHER COMPUTATIONS ARE SAVED. This has been done intentionally, since the BioGrapher workbook is intended for use by students in a computer laboratory. In order to prevent the proliferation of multiple altered versions of the workbook filled with unknown data, the standard Excel workbook save option has been disabled. In effect, whenever BioGrapher is opened, it will be initialized to a known, clean state with no data. However, saved data can always be imported, followed by subsequent visualizations and computations, by using the LOAD menu item (see above).

LOAD NODE LIST loads previously saved data specifying a graph in the node list format described above. Note that this file format is incompatible with the matrix format. If you attempt to read a matrix in with this menu item, you will get an error message.

SAVE NODE LIST saves data for a connection graph specified in the node list format. Note that if the current data represent an adjacency matrix (flagged by the absence of keyword LIST in cell (1, 2) of the INFO sheet (sheet 4), you will not be allowed to save the data using his menu item and will get an error message.

CLEAR CELLS clears ALL CELLS, GRAPHS, AND CHARTS IN ALL WORKSHEETS!

PRINT GRAPH displays (in the WinXP version) the Excel PRINT PREVIEW dialog window with the current graphical layout automatically sized to print as a single page. Of course, all options for printing (including destination printer) can be changed from this preview window! Since the graph is drawn as a (vector) shape object, it can be scaled and printed without the "jaggies." On a Mac, the graph is directly printed. Make sure you have the right printer specified through the standard print setup menu before you print, since no dialog box with options is presented!

Note that "PRINT GRAPH" DOES NOT print graphs displayed by clustering coefficient calculations (see below). You must select and print those graphs using the standard Excel print menu.

HELP and INFO displays this screen. (Text file BioGrapherHelp.txt, in BioGrapher distribution folder, along with pdf printable version BioGrapherHelp.pdf)

ABOUT displays version and other information about BioGrapher.

+++CALCULATE MENU+++

These routines are not completely checked or error-free! Use at your own risk!

Shortest Path: Computes the shortest path between nodes using a version of Floyd's algorithm, implemented in VBA by Michael H. Cole, (mcole@ie.montana.edu, http://logistica-design.blogspot.com/2006/02/floyds-shortest-path-algorithm.html).

Also computes the DIAMETER AND THE CHARACTERISTIC PATH LENGTH (as defined by Watts.)

The results are displayed in the GRAPH PROPERTY worksheet, whose name changes to SHORTEST PATH.

Clustering Coefficient: Calculates the value for each node as well as the mean value for the graph, as defined by Watts. It displays the results in columns showing the node ID number, neighborhood size (how many other nodes directly connected), and clustering coefficient for each node. A header shows the mean clustering coefficient for the graph.

In addition, graphs showing the neighborhood distribution (number of nodes for a given neighborhood number v/s neighborhood number) and the log-log plot of this distribution (with a fitted trendline), useful for "SMALL WORLD" connectivity tests!

Complete Matrix (Undirected): Automatically completes the lower half of the (symmetric) adjacency matrix if you enter only half the matrix, i.e., a set of connections in one direction.

Flip 1s and 0s: Flips the graph, yielding the complementary graph.

Some notes and caveats about the Newick notation: The ability to convert Newick notation to trees was recently added. Note that node labels must have standard alphanumeric characters (no punctuation and special characters). In addition, blanks, including trailing blanks, in node labels are converted to underscores. While the Newick parser checks for matching parentheses, it can fail to detect missing commas and may produce inaccurate graphs. Additionally, duplicate node names can be problematic. Each node name should be unique! Try experimenting with all the graphical layout options provided, not just the tree layout. Note that by default, Newick notation suppresses anchor node display, but these nodes (labeled Anchor1 through AnchorN) are stored as valid nodes with a list of their connecting nodes, shown in a NODE LIST that is automatically generated and displayed as a sheet. In addition to the NEWICK expression (which can be exported and imported as a text file) the NODE LIST can be saved in standard CSV format and can also be displayed using a graph layout such that the anchor nodes are explicitly displayed----> Just delete the NEWICK keyword from the INFO sheet and (re)DRAW the graph.

MISCELLANEOUS HINTS AND TIPS

Use CONTROL Z and enter a number to zoom in or out of any of the sheets and displays. This is especially useful when looking at large graph layouts and their associated “pixel” matrix representations. However, note that (re)rendering the graphs can be slow, depending on the graphics memory and processor speed of the computer being used!

Need to see your graphical layout on the same sheet as the data? You can copy and paste the layout (and resize it) wherever you want! You can also copy and paste the layouts into a MSWORD document by using the PASTE SPECIAL menu in MSWORD to paste the layout as an enhanced graphics meta file or drawing.

Label your files with meaningful names. Node lists should have a '_LIST' suffix after the rest of the file name since their format is different from the adjacency matrix representation, even though they are also Excel CSV files.

Newick expression text files are NOT CSV files; they are ordinary text files with extension .txt, and once again you want to consider tagging a '_NEWICK' suffix to those file names.

You can cut/copy and paste the high-quality graph drawing wherever you wish, including on another spreadsheet in the current workbook, a different Excel workbook, or in a Word document.

Irritated by the fact that node labels are placed in the last column to the right of the screen, and can be off the screen for large matrices? Use the SPLIT PANES feature of the VIEW menu of Excel after selecting the last (labels) column. The window will be divided into two separate regions, with a separate horizontal scroll-bar for all columns BEFORE the label column, which now begins a separate scrolling region.

Specify COLOR for each node by placing a standard color name in the column immediately after the column reserved for labels. Valid colors are red, blue, green, yellow, cyan, orange, or magenta. You need to select the check box that enables colors on the GRAPH form that is displayed each time you generate a graphical layout.

In general, you are best off by deciding on a matrix order or number of entries (rows) in a node list before you begin entering the data. If you attempt to enter data into a blank DATA MATRIX/NODE LIST spreadsheet , it will prompt you to enter this number. You will not be allowed to enter data until you specify a valid number here, which can be changed later in the first cell of the INFO worksheet. In general, you should not change or delete this key value, since the consequences are unpredictable. Our routines rely on this number (rather than smart detection of the blank cell region, which is very buggy in Excel) for all subsequent computations, calculations, and layout, and will ignore numbers placed in cells that are outside the bounds specified by this number. Also, once you have specified this number, if you try entering a number in cells outside the specified bounds, our software will complain and delete your entries!

Software producing weird results/errors? Invoking the “CLEAR CELLS” menu item will reset everything, but you will lose all the data entered!

Once again, make sure you install/use the BioGrapher folder in a location where you have full access privileges, including write privileges, since BioGrapher writes to special files in this folder! This is particularly problematic on Macs (where the Applications folder may be read-only for general users and guests) and on public machines where all folders except a shared folder may be protected!

+++BIBLIOGRAPHY+++

1. "Small Worlds: The Dynamics of Networks between Order and Randomness." Duncan J. Watts, Princeton University Press (1999).

2. For a description of the DOT graphical layout language and implementation of the AT&T GraphViz* package, see http://www.graphviz.org/Documentation.php.

*The original Mac GUI version (GraphViz 2.16), by Glen Low, may (possibly) be downloaded from http://www.pixelglow.com/graphviz/download/. This Excel workbook (Mac version 216) can call the package and pass it the input file to be processed, as well as retrieve the output file for display in Excel.

3. For a mathematical definition of the clustering coefficient, see http://en.wikipedia.org/wiki/Clustering_coefficient. For small world networks, also see http://en.wikipedia.org/wiki/Small-world_network.

4. For a good discussion of shortest path algorithms, see http://www-unix.mcs.anl.gov/dbpp/text/node35.html

5. For NEWICK tree specifications, see http://evolution.genetics.washington.edu/phylip/newicktree.html

Comments and constructive criticism are welcome ---> email to ramav@beloit.edu.