BIOINFORMATICS APPLICATIONS NOTE Vol. 19 no. 1 2003
Pages 157–158

Microarray data assembler

Ramswamy Anbazhagan

Departments of Pathology and Oncology, Johns Hopkins University School of Medicine, Baltimore MD 21231, USA

Received on April 29, 2002; revised and accepted on June 20, 2002

ABSTRACT

Summary: Large volumes of microarray data are generated and deposited in public databases. Most of this data is in the form of tab-delimited text files or Excel spreadsheets. Combining data from several of these files to reanalyze these data sets is time consuming.*Microarray Data Assembler* is specifically designed to simplify this task. The program can list files and data sources, convert selected text files into Excel files and assemble data across multiple Excel worksheets and workbooks. This program thus makes data assembling easy, saves time and helps avoid manual error.

Availability: The program is freely available for non-profit use, via email request from the author, after signing a Material Transfer Agreement with Johns Hopkins University.

Contact: anba@jhmi.edu

INTRODUCTION

Large volumes of microarray data have been generated in the past couple of years and further increase in such data is expected in the future. A challenging task for the biologists and statisticians is to reanalyze these data from various sources to confirm the validity of the original analysis and to develop new methods of analysis to discover new information from the existing data. Most of the microarray data is currently being deposited in the form of tab-delimited text files or in Excel spreadsheets. Various programs available for microarray data analysis handle data input in different ways; some require the user to assemble the data in a single file before loading them into the program. Furthermore, assembling the desired data into a single file makes it convenient for analysis by using other statistical programs. Thus one of the major tasks in preparing the data for analysis is to assemble the required data in a single file. At present, there is no dedicated and freely available program for this purpose. Microarray Data Assembler is specifically designed to accomplish this task. This program can list files and data sources, convert selected text files into Excel files and assemble data across multiple Excel worksheets and workbooks.

PROGRAM OVERVIEW

The program file, *mda.xls*, contains a built-in program that performs all the functions required for assembling microarray data. The program works with Microsoft Excel 2000 under Windows platform and needs no special installation. The program can be started by opening the program file *mda.xls*. When this file is opened, a dialogue box appears asking whether to enable or disable macros. Click on the *Enable Macros* button. When the file opens, a new menu, *Microrarray Data Assembler*, is added to the program menu bar. This menu has five items: *Find Folder Path, List Files, Convert Text Files, List Data Source, and Assemble Data*. These menu items can be executed to perform various functions of the program.

The workbook opened by the program contains five worksheets. They are *Folder Path Info, Folder and File List, Data Source List, Assembled Data Source*, and *Assembled Data*. The first two sheets are meant to help the user convert tab-delimited text files into Excel files. The sheet *Folder Path Info* shows a simple form with information regarding the path of the folder, which contains the data files. The form also displays a field for entering file extension to specify the file type, which should be used for listing. A third field in this form provides an option to indicate whether subfolders are to be included while listing files. The program automatically displays the path of the folder in which *mda.xls* file is located as the default data folder path. If the data files are located in some other folder, execute the command *Find Folder Path* from the *Microarray Data Assembler* menu and select the data folder. This will enter the path of the selected folder in cell B1 of *Folder Path Info* sheet. Users may enter any file extension such as txt, xls, doc, etc. in cell B2 to list corresponding file types or an asterisk (*) to list all the file types. To include subfolders while listing files, enter ‘yes’ in cell B3. If ‘no’ is entered here, only the files in the main folder will be listed. After entering this information, execute the command *List Files* from the *Microarray Data Assembler* menu. The programs displays a list of all the selected file types in the *Folder and File List* sheet including their path name. The second column in this sheet is meant for marking the files that are to be

converted to Excel files. Desired text files can be marked with any alphanumerical character(s). After marking the required files, execute the command *Convert Text Files* from the *Microarray Data Assembler* menu. The program converts all the marked text files into Excel files and also lists them in the third column. The program can convert only text files into Excel files. If other file types are marked for conversion, they are ignored. By default the program is set up to convert text files in which data fields are delimited by tabs and double quotation marks (“ ”) are used as text qualifiers, since this is the most common format used for dumping data as text files.

Once all the required text files are converted to Excel files, the next step is to prepare a list of all the data sources in these files. To accomplish this, execute the command, *List Files*, again and list all the Excel files in *Folder and File List sheet*. You may use ‘xls’ as the file extension option to selectively list Excel files, since only these files will be used for assembling data. From the list of displayed files, mark the desired files for assembling data using alphanumerical character(s) as before. After marking the desired files, execute the command *List Data Sources* from the *Microarray DataAssembler* menu. This command prepares a list of all the data sources from the marked files and displays it in *Data Source List* sheet. The program displays the folder name, file name, worksheet name and the data column header for every data source in the selected files. The content of the first row of the data column is listed as its column header.

To assemble the data from these data sources, mark all the desired data sources using alphanumerical character(s) in the fifth column of this sheet and then execute the command *Assemble Data* from the *Microarray Data Assembler* menu. When this command is executed the program combines all the marked data sources and displays them in the *Assembled Data* sheet. The program also displays a separate list of assembled data sources in the *Assembled Data Source* sheet. Data assembled in this way can be easily saved again as tab-delimited text file by using built-in Excel command, if necessary. Detailed instructions and figures of screen shots of this program are available at http://astor.som.jhmi.edu/∼anba/mda.htm.

This program is very useful for assembling microarray data. Up to 256 data sources can be assembled in an Excel worksheet with up to 65 000 data points in each data source. The program is very easy to use, saves a lot of time and also avoids manual error while assembling microarray data.

ACKNNOWLEDGEMENTS

Supported by a grant (CA88843) from the National Cancer Institute.