BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 11 2011, pages 1585–1586
doi:10.1093/bioinformatics/btr179

Systems biology

Advance Access publication April 15, 2011

datPAV—an online processing, analysis and visualization tool for exploratory investigation of experimental data

Ambarish Biswas[1], Raghuraj Rao[1], Shivshankar Umashankar[1,2], Kalyan C. Mynampati[1], Sheela Reuben[1], Gauri Parab[2] and Sanjay Swarup[1,2,3,*]

[1]Singapore-Delft Water Alliance, National University of Singapore, Singapore 117576, [2]Metabolites Biology Laboratory, Department of Biological Sciences, National University of Singapore, Singapore 117543 and [3]NUS Environmental Research Institute (NERI), #02-01, T-Lab Building (TL), National University of Singapore, Singapore 117411

Associate Editor: Trey Ideker

ABSTRACT

Summary: Data processing, analysis and visualization (datPAV) is an exploratory tool that allows experimentalist to quickly assess the general characteristics of the data. This platform-independent software is designed as a generic tool to process and visualize data matrices. This tool explores organization of the data, detect errors and support basic statistical analyses. Processed data can be reused whereby different step-by-step data processing/analysis workflows can be created to carry out detailed investigation. The visualization option provides publication-ready graphics. Applications of this tool are demonstrated at the web site for three cases of metabolomics, environmental and hydrodynamic data analysis.

Availability: datPAV is available free for academic use at http://www.sdwa.nus.edu.sg/datPAV/.

Contact: sanjay@nus.edu.sg

Received and revised on February 16, 2011; accepted on March 29, 2011

1 INTRODUCTION

It is vital to all the domains of scientific investigations to design systematic experiments and extract relevant knowledge from the data collected. Preliminary investigation of the data is required to remove the errors such as instrument noise or misplaced samples or incorrect data entry, all of which lead to serious misinformation during the analysis step. Exploratory data analysis uses visualization tools to get overview of data and to transform variables/data for further analysis (Borcard *et al*., 2011). For example, systematic sampling or measurement errors can be detected by observing the correlation between the variables or by viewing the nature of the profile for the standard used, during the experiment. A number of application-specific tools are available for handling and processing of experimental datasets (de la Nava *et al*., 2003; Lavine *et al*., 1993; Stadler *et al*., 2006; Stanimirova *et al*., 2004) and for graphical visualization (Militky and Meloun, 1993; Ong and Lee, 1996). However, data preconditioning poses different challenges than what these tools are designed to address. Spreadsheet software such as MS Excel (Palocsay *et al*., 2009) or MINITAB (Harvill, 1993) requires users to individually select the data in the columns and separately perform desired calculations by entering the formulae. They involve repetitions of tasks every time a new column or a data sheet has to be analyzed. Higher level analysis software such as MATLAB, SPSS and SAS require special training or programming skills and are expensive commercial packages. For a quick and easy examination of data, we need a simple data exploration software, which can provide options for basic processing and quick data visualization. In this article, we present the description of a platform-independent data exploratory software called datPAV (data processing, analysis and visualization).

*To whom correspondence should be addressed.

2 METHODS AND IMPLEMENTATION

2.1 Computational model and user interface

This tool features an interactive graphical user interface and harnesses the power of algorithms designed using efficient web programming techniques. Programs in datPAV are developed on Red Hat Enterprise Linux Server Release 5.5 (Tikanga) using GNU Plot, R, CGI and advanced Perl programming (all of them are open source or freewares). This version of the datPAV software is completely compatible with the operating systems—Windows XP, Windows Vista, Windows 7, Linux and Macintosh. datPAV has been tested to be compatible with the web browsers—Firefox, Internet Explorer 6, 7 & 8, Google Chrome and Safari. The tool is best viewed in 'Firefox' or 'IE 8'.

2.2 Data processing, analysis and output

datPAV has a user-friendly three-step procedure (Step 1–3) for data exploration involving data input, processing and visualization which can be performed easily. The software has four theme tabs for general data analysis, omics data, hydrodynamics data and environmental data (Fig. 1). The input data can be any file that stores tabular data, including flat files (either CSV or tab-delimited text files) which can be uploaded to datPAV server using standard file selection interaction.

The suite of basic statistical programs in datPAV include normalization (global and column), scaling (pareto and mean centering), distribution of data, averaging, noise removal, moving average filter, correlation (variable, auto- and cross-correlation), linear regression, fold change and *t*-test. These programs help to determine the distribution of data, establish correlations, perform fold-change analysis and fit a relation between variables using linear regression techniques. The function of 'averaging' or 'grouping' is required for experiments that involve replicated measurements. The 'variable correlation' tool can be used to explore the experimental consistency or instrument reliability by visualizing correlations between the replicates in the data. This could further help in establishing systematic errors or identifying

Fig. 1. datPAV user interface shows options to select the tabs for the different themes. Processes and visualizations can be selected using drop-down menu. Work history appears at the right-hand panel. Graphical outputs from the various programs in datPAV are shown at the bottom.

outliers. 'Fold-change' analysis can be used for controlled experiments to compare the variation between different trials. 'Filtering (moving average)' is exclusively designed for time series data common to environmental or hydrodynamic studies. In addition, simple data transformation programs for sorting, transposing and creating new columns with desired formulae are included.

datPAV has been designed to provide a range of visualization tools such as scatter plot, scatter plot matrix, box plot, profiles, multi-scale plot, bar chart (simple, multiple and stacked), heat map, Venn diagram and pie-chart. datPAV can plot the results of any process in Step 2 (Processing) using any of the available visualization schemes in Step 3. In order to facilitate visualization suitable for different processes, datPAV provides user-friendly suggestions. Users can develop workflow involving 'heat map' plots that show correlations between variables as a way to detect the outlier columns or systematic errors in data structure along with dependencies between the variables. Plots of 'stacked bar chart' and 'pie chart' can be used to visualize the distribution of data in the selected column.

Other features of this tool include a personalized login and workspace for users to retrieve previous analyses. Workflow history allows users to track and retrieve data which can be reused. Graphics editing capabilities are provided for captions, labels and graphical resolution.

3 CONCLUSIONS

datPAV is designed as an efficient, free online software for generalized exploratory data analysis. Data analysis and visualization process is highly simplified in three steps. The software provides tools and useful tips for important data processing/analyses techniques combined with several visualization programs. The fast and easy-to-use features of datPAV are especially suited for experimentalists for quick preconditioning of their datasets.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the support and contributions of the Singapore-Delft Water Alliance (SDWA). The research presented in this work was carried out as part of the SDWA's Aquatic Science Centre@ Ulu Pandan programme.

Funding: Singapore-Delft Water Alliance (WBS No: R-264-001-002-272).

Conflict of Interest: none declared.

REFERENCES

Borcard,D. *et al.* (2011) Exploratory data analysis. In Gentleman,R. *et al.* (eds), *Numerical Ecology with R*. Springer, New York, pp. 9–30.

de la Nava,J.G. *et al.* (2003) Engene: the processing and exploratory analysis of gene expression data. *Bioinformatics,* **19**, 657–658.

Harvill,J.L. (1993) MINITAB: statistical software, release 7.2 SUN-4 version. *Chemom. Intell. Lab. Syst.*, **18,** 111–112.

Lavine,B.K. (1993) SCAN: software for chemometric analysis. *Chemom. Intell. Lab. Syst*., **20**, 93–94.

Militký,J and Meloun,M. (1993) Some graphical aids for univariate exploratory data analysis. *Anal. Chim. Acta*, **277**, 215–221.

Ong,H.L. and Lee,H.Y. (1996) Software report: Winviz—a visual data analysis tool. *Comput. Graph.*, **20**, 83–84.

Palocsay,S.W. *et al.* (2010) Utilizing and teaching data tools in Excel for exploratory analysis. *J. Bus. Res*., **63**, 191–206.

R Development Core Team (2008) R: a language and environment for statistical computing. R Foundaton for Statistical Computing, Vienna, Austria. Available at http://www.R-project.org (last accessed date 2011).

Stadler,M. *et al.* (2006) Web-based tools for data analysis and quality assurance on a life-history trait database of plants of Northwest Europe. *Environ. Modell. Softw.*, **21**, 1536–1543.

Stanimirova,I. *et al.* (2004) STATIS, a three-way method for data analysis. Application to environmental data. *Chemom. Intell. Lab. Syst*., **73**, 219–233.