BIOINFORMATICS APPLICATIONS NOTE
Vol. 24 no. 13 2008, pages 1556–1558
doi:10.1093/bioinformatics/btn217

Data and text mining

DAnTE: a statistical tool for quantitative analysis of -omics data

Ashoka D. Polpitiya, Wei-Jun Qian, Navdeep Jaitly, Vladislav A. Petyuk, Joshua N. Adkins, David G. Camp II, Gordon A. Anderson and Richard D. Smith*

Pacific Northwest National Laboratory, Richland, WA 99352, USA

Received on February 22, 2008; revised on April 9, 2008; accepted on April 29, 2008

Advance Access publication May 3, 2008

Associate Editor: John Quackenbush

ABSTRACT

Summary: Data Analysis Tool Extension (DAnTE) is a statistical tool designed to address challenges associated with quantitative bottom-up, shotgun proteomics data. This tool has also been demonstrated for microarray data and can easily be extended to other high-throughput data types. DAnTE features selected normalization methods, missing value imputation algorithms, peptide-to-protein rollup methods, an extensive array of plotting functions and a comprehensive hypothesis-testing scheme that can handle unbalanced data and random effects. The graphical user interface (GUI) is designed to be very intuitive and user friendly.

Availability: DAnTE may be downloaded free of charge at http://omics.pnl.gov/software/

Contact: rds@pnl.gov or proteomics@pnl.gov

Supplementary information: An example dataset with instructions on how to perform a series of analysis steps is available at http://omics.pnl.gov/software/

1 INTRODUCTION

Although a number of tools are available for high-throughput microarray data processing (Gentleman *et al*., 2004; Saeed *et al*., 2003), the data from LC–MS based quantitative bottom-up proteomics measurements (i.e. label-free approaches, stable isotope labeling methods, spectral counting approaches and the Accurate Mass and Time Tag method) pose different challenges than what these tools are designed to address. One of the major issues associated with proteomics data is often the extent of missing values that is largely due to the larger number of species near the threshold for detection and leads to unbalanced datasets. In addition, proteomics data involves another level of grouping or 'rollup' information to map peptides to proteins. Peptide abundances are often used to infer the corresponding protein abundances.

Developed to address the issues common to proteomics data, Data Analysis Tool Extension (DanTE) is readily extendable. Though the target application is high-throughput proteomics, DAnTE has also been successfully demonstrated for microarray data analysis and can readily be applied to other forms of high-throughput 'omics' data that bears similar characteristics (e.g. metabolomics data). A screenshot of the DAnTE user interface is illustrated in Figure 1.

2 DESCRIPTION

2.1 Dependencies

The graphical user interface (GUI) of DAnTE is implemented using the C# language, and the core algorithms are implemented in the open source R statistical environment (R Development Core Team, 2008). DAnTE runs on a Microsoft WindowsXP platform within a .NET 2.0 framework. The connectivity between R and the C#/.NET environment is achieved by using the open source R(D)COM server application (Baier and Neuwirth, 2007). This unique choice of environments makes DAnTE a very user friendly software tool, even though it cannot integrate into the popular Bioconductor (Gentleman *et al*., 2004) project.

2.2 Application features

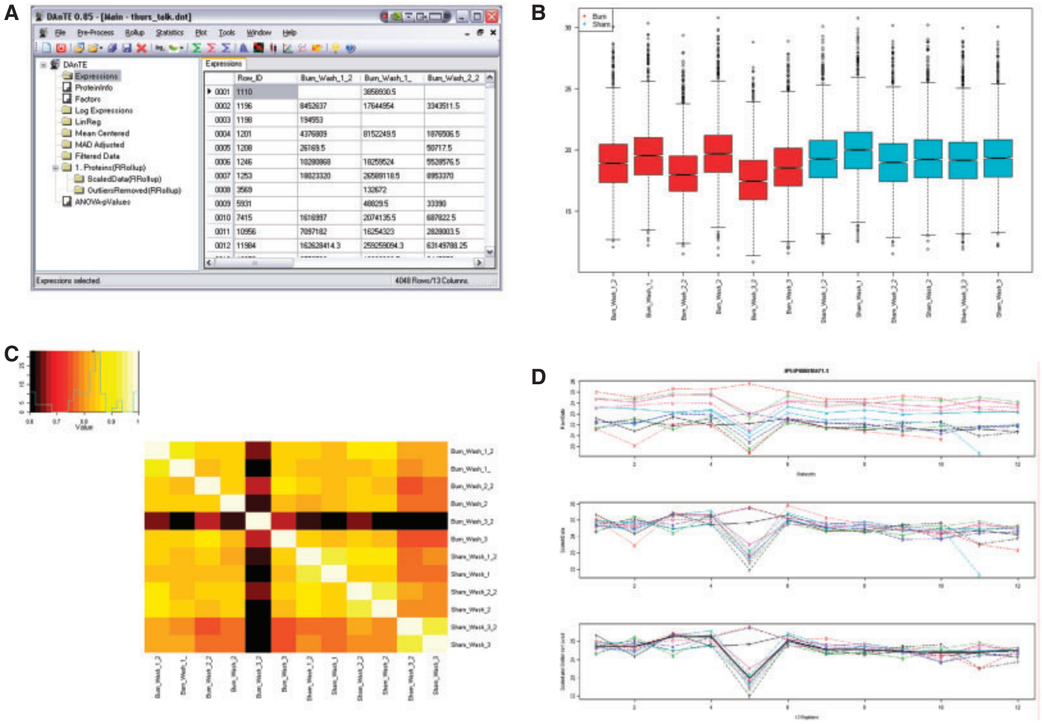
2.2.1 Data loading The input data to DAnTE can be any file that stores tabular data, including flat files (either CSV or tab-delimited text files) and Microsoft Excel files. A unique feature of the data loading mechanism is that it preserves peptide-to-protein mapping information for use later in plotting peptides that belong to a particular protein, as well as in the peptides-to-protein rollup methods. In addition, DAnTE can also process SEQUEST (Eng *et al*., 1994) results and create spectral count tables.

2.2.2 Factor definitions Factors are used to capture the fixed and random effects in experimental design. For example, the biological condition is a fixed effect factor, while a list of liquid chromatography (LC) columns used to separate the samples can be treated as a random effect. This information is vital in normalization, imputation and hypothesis testing methods in DAnTE. Factors can either be declared once the data is loaded or be loaded from a flat file.

2.2.3 Investigative plots Various statistical plots, including histograms, box plots, correlation diagrams and MA (or R-I: ratio-intensity) plots can be plotted in DAnTE. These plots help the user evaluate reproducibility within the study set and single-out problematic datasets so that they can be excluded from further analysis.

2.2.4 Data normalization As normalization is arguably the most important step in downstream data analysis, DAnTE employs several normalization methods that have been successfully tested for both proteomics data (Callister *et al*., 2006) and microarray genomics data (Quackenbush, 2002; Smyth *et al*., 2003).

*To whom correspondence should be addressed.



DAnTE 0.85 - [Main - thurs_talk.dnt]

File Pre-Process Rollup Statistics Plot Tools Window Help

DAnTE
Expressions
ProteinInfo
Factors
Log Expressions
LinReg
Mean Centered
MAD Adjusted
Filtered Data
1. Proteins(RRollup)
ScaledData(RRollup)
OutliersRemoved(RRollup)
ANOVA-pValues

Expressions

	Row_ID	Burn_Wash_1_2	Burn_Wash_1_	Burn_Wash_2_2
0001	1110		3858930.5	
0002	1196	8452637	17644954	3343511.5
0003	1198	194553		
0004	1201	4376809	8152249.5	1876506.5
0005	1208	26169.5		50717.5
0006	1246	10280868	18259524	5528576.5
0007	1253	18023320	26589118.5	8953370
0008	3569		132672	
0009	5931		48829.5	33390
0010	7415	1616997	2074135.5	687822.5
0011	10956	7097182	16254323	2828003.5
0012	11984	162628414.3	259259094.3	63149788.25

Expressions selected. 4048 Rows/13 Columns

Fig. 1. Representative screen shots from DAnTE. (**A**) Data grid and the navigation panel on the left; (**B**) box plot of log transformed data; (**C**) a correlation heatmap of a set of data showing a possible outlier dataset; (**D**) peptides-to-protein rollup results from RRollup method (panels from top to bottom: raw data; scaled data; median profile shown as a thick black line after outliers removed).

Among them are a robust linear regression method, lowess method and a quantile normalization method. In addition, global intensity adjustment based on median absolute deviation (MAD) and central tendency adjustment methods are also available.

2.2.5 Missing value imputation Incomplete datasets due to missing values are common with high-throughput proteomics. As imputing these values is a much-debated topic (Troyanskaya *et al*., 2001), DAnTE offers several simple methods, as well as some advanced algorithms to chose from. The simple methods allow the user to fill in missing values with either the dataset mean/median or with a pre-chosen constant. Advanced methods include filling in with a row mean based on a user-defined factor, K-nearest neighbor imputation (KNNimpute), and singular value decomposition-based imputation (SVDimpute).

2.2.6 Peptide-to-protein rollup In most proteomics methods, peptide measurements are rolled up to corresponding protein abundances. Ideally, all peptides from a single protein should have similar abundances that manifest as similar signal intensities; however, in reality many factors, such as digestion efficiency, electrospray ionization efficiency, etc., can affect the identifications and abundances or signal intensities of peptides. In the RRollup method available in DAnTE, peptides that originate from the same protein are first scaled on the basis of a chosen reference peptide in order to bring all peptide profiles across biological conditions to the same level and then averaged to obtain the protein abundance. During scaling, the peptide with the most observations is chosen as the reference peptide and its total abundance across datasets is used as a tiebreaker. In the ZRollup method, a scaling method similar to *z*-scores (except that medians instead of means from peptide profiles across biological conditions are used) is applied first to peptides that originate from a single protein and then the scaled pepetides are averaged to obtain relative protein abundance. In both RRollup and Zrollup methods, outlying peptide values are excluded from protein abundance calculations, using a Grubb's outlier test (Grubbs, 1969). In the third QRollup method, peptides are selected on the basis of a user selected abundance cutoff value, and protein abundance is calculated as the average of these selected peptides.

2.2.7 Analytical algorithms DAnTE offers several well-characterized algorithms to further explore patterns in the data. Traditional principal component analysis (Jolliffe, 2002) and associated scores and loadings plots can be useful as an unsupervised way of finding the principal variation in the data. In contrast, the partial least squares method (Wold *et al*., 1984) available in DAnTE can be used as a discrimination procedure whereby the grouping information is assigned using factors. Hierarchical and *k*-means clustering methods on features/samples are also available as part of the heat map plotting function.

2.2.8 Hypothesis testing A comprehensive ANOVA scheme for unbalanced studies that uses marginal sums of squares (Fox, 1997) and mixed models (Pinheiro and Bates, 2000) is included in DAnTE. The user can also test for interactions among factors in a multi-way analysis of variance (ANOVA). The *q*-values are also calculated along with the *p*-values in order to control the false discovery rate in multiple testing (Storey, 2002). In addition, DAnTE can check whether the data follows a normal distribution by employing the Shapiro–Wilks test and features two non-parametric hypothesis tests (Wilcoxon rank sum test and Kruskal–Wallis test) when the normality assumption fails to hold.

3 SUMMARY

DAnTE is designed as a complete downstream analysis tool that incorporates a host of algorithms for large-scale bottom-up proteomics data. This tool features an interactive GUI interface and harnesses the power of R statistical environment; its uniqueness lies in its ability to handle incomplete data and to roll peptides up to proteins. Though designed specifically for analyzing proteomics data, DAnTE performs equally well on genomics microarray data.

ACKNOWLEDGEMENTS

The authors thank Joel Pounds, Susan Varnum, and Kim Hixson for their many suggestions and extensive testing; and Thomas O. Metz for data and support for early methods development.

Funding: Portions of this research were supported by the National Institute of General Medical Sciences (NIGMS, Large Scale Collaborative Research Grants U54 GM-62119-02), the NIH National Center for Research Resources (RR18522), Laboratory Directed Research and Development (LDRD) program (W.-J.Q.) at Pacific Northwest National Laboratory (PNNL) and the National Institute of Allergy and Infectious Diseases NIH/DHHS (through interagency agreement Y1-AI-4894-01). Work was performed at PNNL in the Environmental Molecular Sciences Laboratory, a national scientific user facility sponsored by the US Department of Energy (DOE) Office of Biological and Environmental Research. PNNL is operated by Battelle for the DOE under contract DE-AC05-76RLO-1830.

Conflict of Interest: none declared.

REFERENCES

Baier,T. and Neuwirth,E. (2007) R (D)COM Server V2.01. Available at http://sunsite.univie.ac.at/rcom/ (last accessed date May 23, 2008).

Callister,S.J. *et al*. (2006) Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. *J. Proteome Res.*, **5**, 277–286.

Eng,J.K. *et al*. (1994) An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database. *J. Am. Soc. Mass Spectrom.*, **5**, 976–989.

Fox,J. (1997) *Applied Regression Analysis, Linear Models, and Related Methods*. Sage Publications, Thousand Oaks, CA.

Gentleman,R.C. *et al*. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.

Grubbs,F. (1969) Procedures for detecting outlying observations in samples. *Technometrics*, **11**, 1–21.

Jolliffe,I.T. (2002) *Principal Component Analysis*. Springer, New York.

Pinheiro,J.C. and Bates,D.M. (2000) *Mixed-Effects Models in S and S-PLUS*. Springer, New York.

Quackenbush,J. (2002) Microarray data normalization and transformation. *Nat. Genet.*, **32**(Suppl.), 496–501.

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

Saeed,A.I. *et al*. (2003) TM4: a free, open-source system for microarray data management and analysis. *Biotechniques*, **34**, 374–378.

Smyth,G.K. *et al*. (2003) Statistical issues in cDNA microarray data analysis. *Methods Mol. Biol.*, **224**, 111–136.

Storey,J.D. (2002) A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **64**, 479–498.

Troyanskaya,O. *et al*. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.

Wold,S. *et al*. (1984) Modeling data tables by principal components and pls – class patterns and quantitative predictive relations. *Analusis*, **12**, 477–485.