DASS-GUI: a user interface for identification and analysis of significant patterns in non-sequential data

Jens Hollunder1,2,*,†, Maik Friedel3,†, Martin Kuiper1,2 and Thomas Wilhelm4,*

1Department of Plant Systems Biology, VIB, 2Department of Molecular Genetics, Ghent University, Technologiepark 927, B-9052 Gent, Belgium, 3Biocomputing Group, Leibniz Institute for Age Research – Fritz Lipmann Institute, Beutenbergstrasse 11, 07745 Jena, Germany and 4Theoretical Systems Biology, Institute of Food Research, Norwich Research Park, Colney, Norwich NR4 7UA, UK

ABSTRACT

Summary: Many large ‘omics’ datasets have been published and many more are expected in the near future. New analysis methods are needed for best exploitation. We have developed a graphical user interface (GUI) for easy data analysis. Our discovery of all significant substructures (DASS) approach elucidates the underlying modularity, a typical feature of complex biological data. It is related to biclustering and other data mining approaches. Importantly, DASS-GUI also allows handling of multi-sets and calculation of statistical significances. DASS-GUI contains tools for further analysis of the identified patterns: analysis of the pattern hierarchy, enrichment analysis, module validation, analysis of additional numerical data, easy handling of synonymous names, clustering, filtering and merging. Different export options allow easy usage of additional tools such as Cytoscape.

Availability: Source code, pre-compiled binaries for different systems, a comprehensive tutorial, case studies and many additional datasets are freely available at http://www.ifr.ac.uk/dass/gui/. DASS-GUI is implemented in Qt.

Contact: jeholl@psb.vib-ugent.be; thomas.wilhelm@bbsrc.ac.uk

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on December 21, 2009; revised on February 12, 2010; accepted on February 17, 2010

‘Data mining is the process of extracting patterns from data. Data mining is becoming an increasingly important tool to transform these data into information’ (en.wikipedia.org/wiki/Data_mining). Numerous pattern discovery tools have been developed in all fields of science and beyond. In bioinformatics, analysis of sequence data is most prominent. Although the requirement of new tools remains under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

© The Author(s) 2010. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
data); (iii) shuffle-binomial (improved shuffling model exploiting computational shuffling of the dataset working for medium sized shuffling (also called random permutation test, the straightforward $P$-value calculation, applicable only to very small datasets); (ii) DASS-pv [the algorithms of Hollunder et al. (2007a), working for large data]. The first three models work for all four cases (single, multi, unique, ambiguous). DASS-pv can only handle the two ambiguous cases. Nevertheless, DASS-pv helps closing the gap of methods for the analysis of statistical significances of cs and biclusters (Madeira and Oliveira, 2004). BiGGEsTS, a GUI for bicluster analysis of time series gene expression data, also contains specific calculations of significances (Gonçalves et al., 2009), but the only corresponding general approach we are aware of is the SAMBA algorithm (Tanay et al., 2002). It is based on the analogy between biclusters and cliques in bipartite graphs, so it cannot handle multi-sets. The statistical evaluation of such cs is unique to DASS-pv. A comprehensive tutorial of DASS-GUI is available (see availability); many toolkits facilitate usage.

We have already applied our DASS approach for the analysis of protein complexes (Hollunder et al., 2005, 2007b), multi-domain proteins (Hollunder et al., 2007a) and transcription factor binding sites (Beyer et al., 2006; Hollunder et al., 2007a). There are many more applications, in biology and beyond. We are already working on corresponding analyses of genomics, transcriptomics, proteomics and metabolomics data. New interesting and feasible problems are expected to be found in future.

Funding: Human Frontier Science Program Grant (to J.H. and M.K.); Leibniz Graduate School for Aging and Age-related Diseases (to M.F.); Biotechnology and Biological Sciences Research Council Core Strategic Grant for the Institute of Food Research (to T.W.).

Conflict of Interest: none declared.

REFERENCES


Fig. 1. DASS-GUI: (a) the calculation mode and (b) the analysis mode.

analysis, module validation, analysis of additional numerical data, easy handling of synonymous names, clustering and merging (allowing, for instance, identification of non-complete biclues). Different export options allow easy usage of additional tools such as Cytoscape.

The calculation mode allows identification of cs (all or sufficiently dissimilar ones, according to pre-selected size and frequency) and calculation of the statistical significance of cs. Three algorithms have been implemented for cs identification: DASS-cs [as already presented in Hollunder et al. (2007a), but with important additional features, such as similarity pruning], LCM (Uno et al., 2003) and FPclose (Grahné and Zhu, 2003). Only DASS-cs can handle single and multi-sets. This distinction is also important for the calculation of statistical significances. Together with a second distinction, unique [each (host)set is unique] or ambiguous [same (host)set might occur more than once in the dataset], DASS-GUI considers four different cases for significance calculations: single-unique, single-ambiguous, multi-unique and multi-ambiguous. For each of these cases, different models can be used: (i) permutation (exact $P$-value calculation, applicable only to very small datasets); (ii) shuffling (also called random permutation test, the straightforward computational shuffling of the dataset working for medium sized data); (iii) shuffle-binomial (improved shuffling model exploiting the corresponding complete random distribution, assuming binomial distribution); and (iv) DASS-pv [the algorithms of Hollunder et al. (2007a), working for large data].


