BIOINFORMATICS APPLICATIONS NOTE

Vol. 26 no. 7 2010, pages 987–989
doi:10.1093/bioinformatics/btq071

Data and text mining

Advance Access publication February 19, 2010

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

Jens Hollunder[1,2,*,†], Maik Friedel[3,†], Martin Kuiper[1,2] and Thomas Wilhelm[4,*]

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

Associate Editor: Alex Bateman

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: jehol@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 (Friedel *et al*., 2009), BLAST (Altschul *et al*., 1990) is a widely used and regularly improved standard (Cameron *et al*., 2004). However, there is no such standard for analyses of non-sequential data. Hundreds of papers and books for gene expression cluster analysis have been published since 1998 (Eisen *et al*., 1998). Biclustering, the identification of common patterns for a subset of genes in a subset of conditions, was first adapted to expression analysis a decade ago (Cheng and Church, 2000). It is becoming more and more important. Two corresponding toolboxes containing a number of algorithms have been developed (Barkow *et al*., 2006; Kaiser and Leisch, 2008). Indeed, biclustering, also called co-clustering (Madeira and Oliveira, 2004), two-way clustering (Kaiser and Leisch, 2008) or subspace clustering (Liu and Wang, 2003), is an important data mining technique with numerous applications, in bioinformatics and beyond. It has been used in recommendation systems and targeted marketing in e-commerce, information retrieval and text mining, dimensionality reduction in databases, and in analyses of electoral data, nutritional data and currency exchange. It was found that this is 'only a small fraction of the potential applications' (Madeira and Oliveira, 2004).

We have developed a complementary approach for pattern identification in non-sequence data with a largely overlapping range of applications, called discovery of all significant substructures (DASS) (Hollunder *et al*., 2007a). It is applicable to any data that can be represented by sets containing elements. DASS comprises two parts (i) identification of closed sets (cs; a set is 'closed' if there exists no superset with the same frequency; for simple sets a cs is equivalent to a clique in a corresponding bipartite graph) and (ii) evaluation of the statistical significance of cs. Importantly, DASS works for simple sets (each element unique per set) and multi-sets (may contain elements more than once). This feature sets it apart from standard data mining methods for identification of meaningful subsets, such as APRIORI (Agrawal and Srikant, 1994) for the identification of frequent (sub)sets and CHARM (Zaki and Hsiao, 2002) for the identification of frequent cs. In contrast to these algorithms that iterate through the elements, DASS-cs, the algorithm to identify cs, iterates through sets. The more modular and hierarchical the data are, the higher the advantage of this approach (Hollunder *et al*., 2007a). However, in cases of very many sets and few elements, other algorithms might be better suited. Interestingly, data can be transformed for efficient use of DASS-cs also in such cases (cf. description of calculation mode in www.ifr.ac.uk/dass/gui).

Here, we present the newly developed tool DASS-GUI, enabling easy usage of all DASS algorithms. It works in two modes: the calculation mode (Fig. 1a) for calculation of cs and corresponding *P*-values (using, among others, the DASS algorithms, Hollunder *et al*., 2007a), and the analysis mode (Fig. 1b), allowing additional filtering, calculation of cs hierarchy, calculation of means and standard deviations of different numerical features, enrichment

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

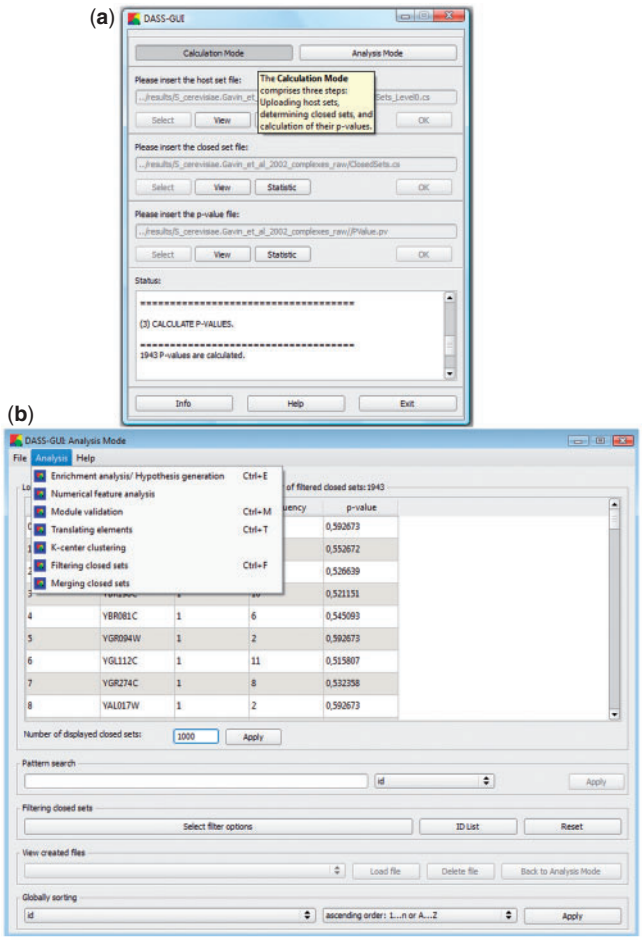

Fig. 1. DASS-GUI: (**a**) the calculation mode and (**b**) the analysis mode.

analysis, module validation, analysis of additional numerical data, easy handling of synonymous names, clustering and merging (allowing, for instance, identification of non-complete bicliques). 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 (Grahne 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]. 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 (Goncalves *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 tooltips 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

Agrawal,R. and Srikant,R. (1994) Fast algorithms for mining association rules. In *Proceedings of 20th International Conference on Very Large Data Bases*, pp. 487–499.

Altschul,S.F. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Barkow,S. *et al*. (2006) BicAT: a biclustering analysis toolbox. *Bioinformatics*, **22**, 1282–1283.

Beyer,A. *et al*. (2006) Integrated assessment and prediction of transcription factor binding. *PLoS Comput. Biol.*, **2**, 615–626.

Cameron,M. *et al*. (2004) Improved gapped alignment in BLAST. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **1**, 116–129.

Cheng,Y. and Church,G.M. (2000) Biclustering of expression data. In *Proceedings of 8th International Conf erence on Intelligent Systems for Molecular Biology (ISMB'00)*, pp. 93–103.

Eisen,M.B. *et al*. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci.* USA, **95**, 14863–14868.

Friedel,M. *et al*. (2009) DiProGB: the dinucleotide properties genome browser. *Bioinformatics*, **25**, 2603–2604.

Goncalves,J.P. *et al*. (2009) BiGGEsTS: integrated environment for biclustering analysis of time series expression data. *BMC Res. Notes*, **2**, 124.

Grahne,G. and Zhu,J. (2003) Efficiently using prefix-trees in mining frequent itemsets. In Goethals,B. and Zaki,M.J. (eds). *Proceedings of the ICDM 2003 Workshop on Frequent Itemset Mining Implementations, FIMI'03*, pp. 125–134.

Hollunder,J. *et al*. (2005) Identification and characterization of protein subcomplexes in yeast. *Proteomics*, **5**, 2082–2089.

Hollunder,J. *et al*. (2007a) DASS: efficient discovery and p-value calculation of substructures in unordered data. *Bioinformatics*, **23**, 77–83.

Hollunder,J. *et al*. (2007b) Protein subcomplexes – molecular machines with highly specialized functions. *IEEE Trans. Nanobioscience*, **6**, 86–93.

Kaiser,S. and Leisch,F. (2008) A toolbox for bicluster analysis in R. In Brito,P. (ed.). Compstat 2008 – *Proceedings in Computational Statistics*. Physica Verlag, Heidelberg. Available at http://epub.ub.uni-muenchen.de/3293/ (last accessed date November 10, 2009).

Liu,J. and Wang,W. (2003) OP-Cluster: clustering by tendency in high dimensional space. *Proceedings of 3rd IEEE Int ernational Conference on Data Mining*, pp. 187–194.

Madeira,S.C. and Oliveira,A.L. (2004) Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **1**, 24–45.

Tanay,A. *et al*. (2002) Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, **18**, S136–S144.

Uno,T. *et al*. (2003) LCM: an efficient algorithm for enumerating frequent closed itemsets. In Goethals,B., Zaki,M.J. (eds). *Proceedings of the ICDM 2003 Workshop on Frequent Itemset Mining Implementations FIMI'03*.

Zaki,M.J. and Xiao,W. (2002) CHARM: an efficient algorithm for closed itemset mining. In *Proceedings of the 2nd SIAM International Conference on Data Mining (SDM 2002)*, pp. 457–473.