BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 9 2011, pages 1330–1331
doi:10.1093/bioinformatics/btr115

*Gene expression*

Advance Access publication March 3, 2011

# SC$^2$ATmd: a tool for integration of the figure of merit with cluster analysis for gene expression data

Amy L. Olex[1,*] and Jacquelyn S. Fetrow[1,2]

[1]Department of Computer Science and [2]Department of Physics, Wake Forest University, Winston-Salem, NC 27109, USA

Associate Editor: Trey Ideker

**ABSTRACT**

**Summary:** Standard and Consensus Clustering Analysis Tool for Microarray Data (SC$^2$ATmd) is a MATLAB-implemented application specifically designed for the exploration of microarray gene expression data via clustering. Implementation of two versions of the clustering validation method figure of merit allows for performance comparisons between different clustering algorithms, and tailors the cluster analysis process to the varying characteristics of each dataset. Along with standard clustering algorithms this application also offers a consensus clustering method that can generate reproducible clusters across replicate experiments or different clustering algorithms. This application was designed specifically for the analysis of gene expression data, but may be used with any numerical data as long as it is in the right format.

**Availability:** SC$^2$ATmd may be freely downloaded from http://www.compbiosci.wfu.edu/tools.htm.

**Contact:** olexal@wfu.edu

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

Received and revised on January 3, 2011; accepted on February 27, 2011

## 1 INTRODUCTION

Today's technological advancements have provided researchers with an abundance of information, especially in the field of molecular biology. High-throughput technologies, such as microarrays, are capable of generating overwhelming amounts of biological data. Due to the large scale of these datasets, non-computational analysis is unrealistic; therefore, new computational tools and techniques are needed for the efficient organization and analysis of these data.

Clustering is an important exploratory tool that aids in the analysis and organization of biological data by dividing the dataset into smaller, more manageable groups based on a chosen definition of similarity. Many freely available applications, such as the Cluster and Tree View program suite (Eisen *et al.*, 1998), Hierarchical Clustering Explorer v3.0 (Seo and Shneiderman, 2002) and MultiExperiment Viewer v4.0 (Saeed *et al.*, 2006), implement several standard clustering techniques with a multitude of options; however, selection of the most appropriate method for each dataset can be difficult, and clustering results from standard clustering algorithms are not always reproducible across replicate experiments.

To aid users in the selection of clustering options, SC$^2$ATmd integrates two versions of the figure of merit (FOM) (Olex *et al.*, 2007; Yeung *et al.*, 2001) with standard clustering techniques. This creates a virtually seamless process that tailors the clustering analysis according to the individual characteristics of each dataset. SC$^2$ATmd also implements consensus clustering to address the issue of cluster reproducibility (Monti *et al.*, 2003; Swift *et al.*, 2004). Consensus clustering has the ability to identify robust clusters across multiple clustering algorithms and/or replicate experiments. Other features of SC$^2$ATmd include portable tab-delimited output files, the customization of figures with access to a variety of image file formats and implementation of a cluster mapping routine which is a novel technique that defines one clustering solution in terms of another.

SC$^2$ATmd was designed to aid scientific researchers in the analysis of microarray data, and has already proven to be an invaluable tool in this area. Currently, SC$^2$ATmd is being used in the analysis of several large time-course microarray experiments including the study of phenylpropanoid signaling in *Arabidopsis thaliana*, the early effects of osteoarthritis, and the process of dendritic cell maturation (Olex *et al.*, 2010). SC$^2$ATmd has also proven to be an effective teaching tool with its user-friendly graphical user interface (GUI) and intuitive output that is easy for students to understand.

## 2 APPLICATION DESCRIPTION

SC$^2$ATmd is specifically designed for the cluster analysis of time-course gene expression data; however, with correctly formatted input files it can be used with any type of numerical data. Two versions of SC$^2$ATmd are available. For non-MATLAB users, a standalone application with a GUI is provided. For advanced MATLAB users the MATLAB-dependent version allows access to each m-file directly in addition to the GUI. This allows a greater flexibility in the input, output and overall use of the program. MATLAB .fig output files provide the ability to customize figures within the MATLAB environment (not available for the standalone version). The ability to export to a variety of image formats such as JPEG, BMP, EPS, PDF and others is included with both versions of the application.

### 2.1 FOM analysis

The FOM analysis enables users to compare the performance of various standard clustering methods on their dataset to determine which method generates the most homogeneous clusters. It is

*To whom correspondence should be addressed.

has been shown that highly homogeneous clusters, with respect to the similarity measure employed, form biologically relevant groups of genes that have the potential to reveal functional and regulatory relationships (Jiang *et al*., 2004; Swift *et al*., 2004). The FOM analysis quantitatively determines which clustering algorithm generates clusters with the highest homogeneity on a dataset-by-dataset basis; it also suggests the ideal number of clusters that should be used in the formal analysis (Supplementary Fig. 1). The FOM (Yeung *et al*., 2001) and cFOM (Olex *et al*., 2007), a variation of FOM tailored for use with Pearson's correlation coefficient, are implemented in this application.

### 2.2 Standard cluster analysis

Standard clustering is performed in two steps: (i) the selected clustering method, hierarchical agglomerative clustering (HAC) or k-means, is used to cluster all the data once; (ii) each generated cluster from step 1 is re-clustered using HAC. The second hierarchical re-clustering organizes the genes so that the most similar gene profiles within each cluster group together in the resulting heat map figure (Supplementary Fig. 2).

The standard clustering analysis implements HAC differently in the first clustering step than most other applications, so it deserves further explanation. Generally, HAC is implemented so that the entire hierarchical tree is generated from the bottom up; dissection into subtrees (to obtain clusters) is left to user discretion. However, the FOM analysis requires the number of clusters to be specified in advance, thus traditional hierarchical clustering is not compatible because it does not specify discrete clusters. In SC$^2$ATmd, the HAC algorithm is instructed to stop building the hierarchical tree at the appropriate level to generate the specified number of discrete clusters for the FOM algorithm. This allows the FOM to identify the optimal number of subtrees (clusters) that best represent the characteristics of the entire dataset. Output includes a global hierarchical tree that relates all subtrees; hence the entire tree can be reconstructed if desired.

### 2.3 Consensus cluster analysis

The consensus clustering analysis is based on the algorithm described by Monti *et al.* (2003) where robust clusters are identified by means of a bootstrapping technique. The consensus clustering algorithm implemented in SC$^2$ATmd expands that of Monti *et al.* by allowing the user to identify robust clusters—subgroups of genes that are consistently clustered together—across multiple clustering algorithms, similarity measures, replicate experiments or any combination of the three. The option to import and use custom preclustered data for the identification of consensus clusters provides the user with unlimited flexibility in what types of consensus clusters are extracted.

### 2.4 Cluster mapping

Cluster mapping is an analysis technique that describes one clustering solution in terms of another. Each clustering technique will represent the data in a different way, elucidating different biological characteristics. Cluster mapping can identify relationships between different clustering solutions of the same data. For example, a comparison between the effects of different similarity measures can be made where the same data are clustered using Euclidean distance and then reclustered using Pearson's correlation coefficient. A mapping of the first solution to the second will give an idea of how many Pearson clusters are located in one Euclidean cluster or vice versa. Interesting information about the composition of clusters generated under each method can be extracted from this comparison.

## 3 IMPLEMENTATION

All source files were implemented using MATLAB version 7.0.4.365 (R14) Service Pack 2. The GUI was designed and implemented using GUIDE in MATLAB. To create the standalone version it was compiled into an executable using the MATLAB compiler. Currently, the standalone is only compatible with a Windows OS, but the MATLAB version may be run on any platform that has MATLAB installed. Adobe Acrobat Reader and the MATLAB Component Runtime v72 (provided with the distribution) are required for proper execution of the standalone application. The current implementation provides two clustering methods, *k*-means and HAC, but can be expanded to include more in the future. Supplementary Table 1 summarizes the file hierarchy of the entire program, and screen shots of the application interface can be found in Supplementary Figures 3–8.

## ACKNOWLEDGEMENTS

We thank William Turkett and David John for a critical review of earlier versions of this manuscript.

*Funding*: National Science Foundation and National Institute of General Medical Sciences Program in Mathematical Biology (grant number R01-GM075304 to J.F.); the National Institutes of Health (grant number 1R21-AI082474 to Elizabeth M. Hiltbold and J.F.); and the Wake Forest University Cross-Campus Collaborative Research Support Fund grant (to J.F. and Elizabeth M. Hiltbold).

*Conflict of Interest*: none declared.

## REFERENCES

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

Jiang,D. *et al.* (2004) Cluster analysis for gene expression data: a survey. *IEEE Trans. Knowl. Data Eng.*, **16**, 1370–1386.

Monti,S. *et al.* (2003) Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Mach. Learn.*, **52**, 91–118.

Olex,A.L. *et al.* (2007) Additional limitations of the clustering validation method figure of merit. In *45th ACM Southeast Annual Conference*. ACM, Winston-Salem, NC, pp. 238–243.

Olex,A.L. *et al.* (2010) Dynamics of dendritic cell maturation are identified through a novel filtering strategy applied to biological time-course microarray replicates, *BMC Immunol.*, **11**, 41.

Saeed,A.I. *et al.* (2006) TM4 microarray software suite. *Methods Enzymol.*, **411**, 134–193.

Seo,J. and Shneiderman,B. (2002) Interactively exploring hierarchical clustering results. *IEEE Comput.*, **35**, 80–86.

Swift,S. *et al.* (2004) Consensus clustering and functional interpretation of gene-expression data. *Genome Biol.*, **5**, R94.

Yeung,K.Y. *et al.* (2001) Validating clustering for gene expression data. *Bioinformatics*, **17**, 309–318.