Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Software**

Mfuzz: A software package for soft clustering of microarray data

Lokesh Kumar[1, 2, 3] and Matthias Futschik[1*]

[1]Institute of Medical Informatics and Biometry, Charite, Humboldt University, Invalidenstra Beta e 43, 10115 Berlin, Germany; [2]Department of Systems Biology, Keio University, Yamagata 997-0035, Japan; [3]Department of Biotechnology, Indian Institute of Technology, Guwahati - 781039, India; Matthias E. Futschik* - Email: m.futschik@staff.hu-berlin.de; Phone: 049 30 2093 9106; * Corresponding author

received April 12, 2007; accepted May 01, 2007; published online May 20, 2007

Abstract:
For the analysis of microarray data, clustering techniques are frequently used. Most of such methods are based on hard clustering of data wherein one gene (or sample) is assigned to exactly one cluster. Hard clustering, however, suffers from several drawbacks such as sensitivity to noise and information loss. In contrast, soft clustering methods can assign a gene to several clusters. They can overcome shortcomings of conventional hard clustering techniques and offer further advantages. Thus, we constructed an R package termed Mfuzz implementing soft clustering tools for microarray data analysis. The additional package Mfuzzgui provides a convenient TclTk based graphical user interface.

Keywords: gene expression; soft clustering; software

Availability: The R package Mfuzz and Mfuzzgui are available at http://itb1.biologie.hu-berlin.de/~futschik/software/R/Mfuzz/index.html. Their distribution is subject to GPL version 2 license.

Background:
Clustering methods are popular tools in data analysis. They can be used to reveal hidden-patterns (clusters of objects in large complex data sets). Most clustering methods assign one object to exactly one cluster. **[1]** While this so-called hard clustering approach is suitable for a variety of applications, it may be insufficient for microarray data analysis. Here, the detected clusters of co-expressed genes indicate co-regulation. However, genes are frequently not regulated in a simple 'on'--'off' manner, but instead their expression levels are tightly regulated by a number of fine-tuned transcriptional mechanisms. This is reflected in expression data sets generated in microarray experiments. It is a common observation that many genes show expression profiles similar to several cluster patterns. **[2, 3]**

Ideally, clustering methods for microarray analysis should be capable of dealing with this complexity in an adequate manner. They should not only differentiate how closely a gene follows the main expression pattern of a cluster, but they should also be capable to assign genes to several clusters if their expression patterns are similar.

Soft clustering can provide these favourable capacities. Recently we have shown that applying soft clustering to microarray data analysis leads to i) more adequate clusters with information-rich structures, ii) increased noise-robustness and iii) and improved identification of regulatory sequence motifs. **[4]**

Methodology:
Soft clustering has been implemented using the fuzzy c-means algorithm. **[5]** It is based on the iterative optimization of an objective function to minimize the variation of objects within clusters. Poorly clustered objects have decreased influence on the resulting clusters making the clustering process less sensitive to noise. Notably this is a valuable characteristic of fuzzy c-means method as microarray data tends to be inherently noisy. As a result, fuzzy c-means produces gradual membership values μ_{ij} of a gene i between 0 and 1 indicating the degree of membership of this gene for cluster j. This strongly contrasts hard clustering e.g. the commonly used k-means clustering that generates only membership values μ_{ij} of either 0 or 1. Thus, soft clustering can effectively reflect the strength of a gene's association with a cluster. Obtaining gradual membership values allows the definition of cluster cores of tightly co-expressed genes. Moreover, as soft clustering displays more noise robustness, the commonly used procedure of filtering genes to reduce noise in microarray data can be avoided and loss of the potentially important information can be prevented. **[4]**

Software input
Like most other clustering software, the Mfuzz package requires as input the data to be clustered and the setting of clustering parameters.

Microarray expression data can be entered either as simple table or as Bioconductor (i.e. exprSet) object. Whereas the table format is an easy and sufficient way to handle data for most experiments, Bioconductor data objects can be used for more complex experimental designs. **[6]** The format for tables is the same as for the standard clustering software Cluster **[7]**, so that users can easily use both software packages without reformatting their input.

Further, the number of clusters and the so-called fuzzification parameter m have to be chosen. By variation of both parameters, users can probe the stability of obtained clusters as well as the global clustering structure **[4]**

Software output

As basic output, the partition matrix is supplied containing the complete set of membership values. This information can be used to define cluster cores consisting of highly correlated genes and to improve the subsequent detection of regulatory mechanism. **[4]** Results of the cluster analysis can be either further processed within the Bioconductor framework or stored in simple table format.

Several functions serve the visualization of the results such as internal or global cluster structures. Figure 1 shows some examples of the graphical output.

Note that Mfuzz is not restricted to microarray data analysis, but has recently also successfully applied to examine protein phosphorrylation time series. **[8]**

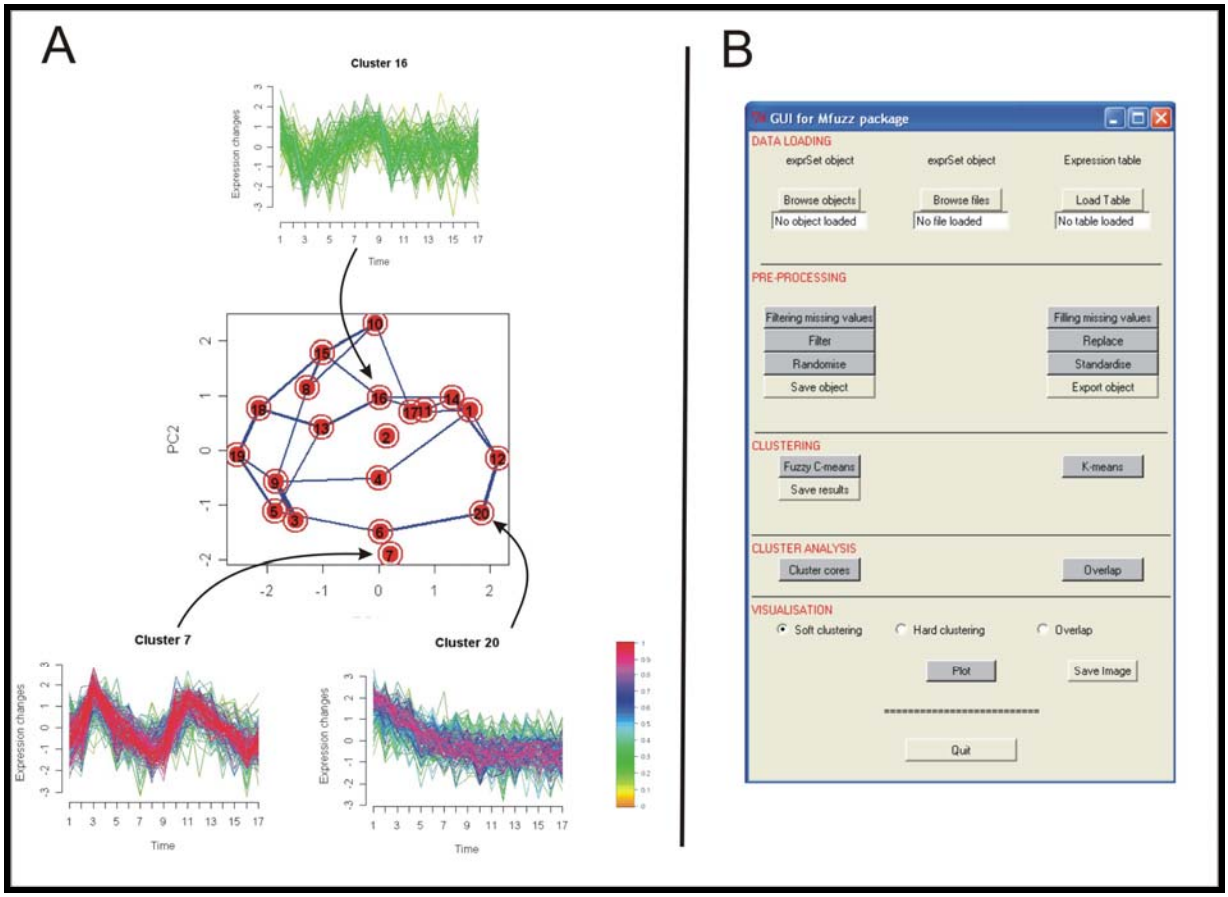

Figure 1: A) Examples for visualization of clustering results produced by Mfuzz. Graphs show the temporal expression pattern during the yeast cell cycle (top and lower panels) and the global clustering structure (central panels). Membership values are color-encoded with red shades denoting high membership values and green shades denoting low membership values of genes. Using this color scheme, clusters with a large core of tightly co-regulated genes (e.g. cluster 7) can be easily distinguished from week or noisy clusters (e.g. cluster16). The central panel shows the principal components of the clusters. Lines between clusters indicate their overlap i.e. how many genes they share. B) Graphical user interface implemented in the Mfuzzgui package. Its outline follows the standard steps of cluster analyses of microarray data: Data loading and pre-processing, clustering, examination and visualization of results

Caveat & Future development:

Mfuzz and Mfuzzgui are R packages. R is a statistical programming language and is freely available open-software. **[9]** Both developed packages follow conventions of the Bioconductor platform. **[6]** The graphical user interface implemented in Mfuzzgui demands an existing installation of Tcl/Tk. For convenience, we supply scripts for automatic installation of the software packages. Additionally, scripts are provided for a direct start of the packages enhancing their stand-alone character. Future versions will include extended export options such as automatically generated HTML pages reporting the results of the clustering analysis.

Acknowledgement:

Lokesh Kumar was supported by the SFB 618 grant of the Deutsche Forschungsgemeinschaft. We would like to thank

Hanspeter Herzel for his assistance of the project and B. Carlisle for critical reading of the manuscript.

References:

[01] A. K. Jain, *et al.*, *ACM Computing Surveys*, 31:264 (1999)
[02] R. J. Cho, *et al.*, *Mol Cell*, 2:65 (1998) [PMID: 9702192]
[03] S. Chu, *et al.*, *Science*, 282:699 (1998) [PMID: 9784122]
[04] M. E. Futschik and B. Carlisle, *J Bioinform Comput Biol.*, 3:965 (2005) [PMID: 16078370]
[05] R. Hathaway and J. Bezdek, *Pattern Recognition*, 19:477 (1985)
[06] http://www.bioconductor.org
[07] http://rana.lbl.gov/EisenSoftware.htm
[08] J. V. Olsen *et al.*, *Cell*, 127:635 (2006) [PMID: 17081983]
[09] http://www.r-project.org

Edited by P. Kangueane
Citation: Kumar & Futschik, Bioinformation 2(1): 5-7 (2007)