Published online 14 May 2008 *Nucleic Acids Research, 2008, Vol. 36, Web Server issue* ***W315–W319***
doi:10.1093/nar/gkn265

AMIC@: All MIcroarray Clusterings @ once

Filippo Geraci, Marco Pellegrini* and M. Elena Renda

Istituto di Informatica e Telematica del C.N.R., Via Moruzzi 1, Pisa, Italy

Received February 18, 2008; Revised April 17, 2008; Accepted April 21, 2008

ABSTRACT

The *AMIC@* Web Server offers a light-weight multi-method clustering engine for microarray gene-expression data. *AMIC@* is a highly interactive tool that stresses user-friendliness and robustness by adopting AJAX technology, thus allowing an effective interleaved execution of different clustering algorithms and inspection of results. Among the salient features *AMIC@* offers, there are: (i) automatic file format detection, (ii) suggestions on the number of clusters using a variant of the *stability-based* method of Tibshirani *et al.* (iii) intuitive visual inspection of the data via heatmaps and (iv) measurements of the clustering quality using *cluster homogeneity*. Large data sets can be processed efficiently by selecting algorithms (such as FPF-SB and *k*-Boost), specifically designed for this purpose. In case of very large data sets, the user can opt for a batch-mode use of the system by means of the *Clustering wizard* that runs all algorithms at once and delivers the results via email. *AMIC@* is freely available and open to all users with no login requirement at the following URL http://bioalgo.iit.cnr.it/amica.

INTRODUCTION

Microarray technology for profiling genes according to their expression levels has became a popular tool in modern biological research. It finds applications in a wide range of tasks, such as tissue classification, detection of metabolic networks and drug discovery. The amount of available data is increasing by the day and microarray data analysis has become a fundamental task, in order to extract knowledge from the data in an effective and timely manner. From this point of view, *clustering* plays an important role in the unsupervised analysis of microarray data, and many efforts have been exerted by researchers in the design of accurate *ad hoc* clustering algorithms. Due to the heterogeneity of microarray data, none of the currently available clustering algorithms has shown consistently better accuracy than the others; thus, biologists may have to try many different clustering softwares and compare outcomes in order to establish which algorithm is the best candidate for their specific task. This exploratory activity can be a burden, unless an integrated tool is adopted that: (i) allows running several methods simultaneously with a variety of parameters; (ii) offers a uniform visual interface for inspecting the raw data and the clustered results and (iii) gives the possibility of downloading the results for further offline analysis and investigations.

Inspecting the available free web tools for microarray gene-expression data analysis, we found out that there are:

- comprehensive systems that offer a full suite of tools from storing raw data, filtering and preprocessing, data clustering and classification, to linking with annotation ontologies (e.g. Gene Ontology) (1), some of them placing less emphasis on offering many clustering methods (2–6);
- systems devoted to microarray analysis for specific tasks [metabolic network inference (7), biclustering in (8), interspecies analysis (9), transcription factor-binding sites detection (10)];
- systems that, though devoted to the analysis of microarray data, do not offer true clustering analysis [e.g. Array Pipe (11), Expressyourself (12), RACE (13)].

A common characteristic of these tools is that they usually are quite rich in functionalities but also complex, requiring the user to climb a steep learning curve before they can be mastered. They can be inadequate if the main interest is performing comparisons among several clustering results, with less emphasis on pre- or post-processing operation. For these reasons, we developed *AMIC@—All MIcroarray Clustering @ once* (pronounced [a:meeka:]), a web application aiming at providing users with a common *user-friendly* interface to a wide range of microarray gene-expression data clustering algorithms. *AMIC@* is really simple and intuitive to use, and allows to:

- load and cluster data from microarray gene-expression experiments (in standard ASCII tab or space delimited form);
- get a hint on the number of clusters for the given dataset;
- run several algorithms (and different configurations of them) on the same data set;

*To whom correspondence should be addressed. Tel: +39 050 3152410; Fax: +39 050 3152593; Email: marco.pellegrini@iit.cnr.it

- visualize the raw data matrix and all the resulting clusterings online, by means of heatmaps;
- visualize the expression level of each heatmap cell;
- view, for each cluster in the resulting clustering, its *homogeneity value* (14), and the average homogeneity of the whole clustering;
- download the outcome of the clustering activities as a standard clustered data files (CDT), for further, offline investigation and,
- receive via email the clustering results by means of a clustering wizard, that allows to simultaneously execute all the algorithms on the given data set.

There are also several (free of charge or for payment) software suites that can be downloaded, installed and executed on the client end to perform microarray analysis. Among the free of charge ones, we may cite TM4 (MeV) (15), a downloadable application for microarray analysis that offers a wide range of options for data filtering, data normalization, clustering, classification and miscellaneous statistical analysis. Mev v4.1 has the capability of storing and loading micro-array data on the local host, as well as the partial results of the analysis, for later reuse and provides several visualization tools for data inspection.

METHODS

AMIC@ is a web-based application for performing multiple microarray data clustering. The web server currently supports five clustering algorithms: *k-means* (16), *SOM* (Self Organizing Maps) (16), *HAC* (Hierarchical Agglomerative Clustering) (16), *FPF-SB* (Furthest-Point-First Stability-Based) (17) and *k-Boost* (18).

While *k*-means, SOM and HAC are standard algorithms, well-known and broadly used in bioinformatics, FPF-SB and *k*-Boost are two novel algorithms for microarray data clustering proposed by our group. FPF-SB and *k*-Boost have an appreciable property: they automatically suggest to the user a reasonable number of clusters, computed with a variant of the stability-based technique proposed in (19). Each of the supported algorithms can be run with six different distance metrics: *Euclidean distance*, *Pearson correlation coefficient*, *City-block distance*, *Cosine similarity*, *Spearman's rank correlation*, and *Kendall's similarity*. Furthermore, each algorithm admits a certain number of customizable parameters that can be set, as specified in details in Web Application description section.

AMIC@ is based on the AJAX technology (Asynchronous JavaScript and XML), providing the user with an interface resembling the one of a classical stand-alone application, but working inside a web browser. The advantage of this technology is to overcome the inherent limitations of standard web pages, for example, by avoiding to reload the entire page for each server request. Moreover, the user has no need to install or configure specific software. Furthermore, *AMIC@* architecture allows multiple, parallel operations, making it possible, for instance, inspecting a clustering result, while another clustering algorithm is running. All the algorithms have been implemented in Python, using, for *k*-means, SOM and HAC some of the Pycluster (http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/software.htm) and BioPython libraries [see the Biopython Project http://biopython.org (16)].

WEB APPLICATION DESCRIPTION

The first steps required to use *AMIC@* are: file uploading (I) and data filtering (II); then access is granted to the main clustering page (III).

In *AMIC@*'s upload page (phase I), it is possible to upload an ASCII file up to 10 MB in size containing microarray gene expression data (space or tab delimited). *AMIC@* also accepts zipped data files (the compression ratio for gene-expression data in this format is roughly about 2, thus data sets up to 20 MB can be managed by *AMIC@*). For the purpose of offering a quick demo, we give the option of uploading three predefined data sets.

It is worth nothing that *AMIC@*, differently from the majority of online available tools, does not require the user to specify a file format beforehand, or to declare the columns as header or data. In fact, *AMIC@* tries to interpret the uploaded file and automatically distinguish label columns, columns containing data and comment columns. If the upload phase is successful, *AMIC@* returns a page asking for confirmation of the guess made; in the same page, it is possible to do a rough form of feature selection, by filtering out some of the experiments (phase II).

During data uploading, the system detects missing data, if any, and reports this fact to the user. Many different techniques for estimating/replacing missing values have been presented in literature. We implemented three standard methods, broadly used in practice: (i) set the missing values to zero, (ii) set the missing values to the row average value and (iii) *k*-nearest-neighbor [http://helix-web.stanford.edu/pubs/impute/; (20)], both weighted and unweighted, with a user-selected value of *k* varying form 1 to 12.

After the upload and filtering phases, *AMIC@* is ready for clustering (phase III). The main page is a tabbed page, where each tab corresponds to the parameter setting page of a clustering algorithm (Figure 1). The raw matrix data is visualized, and in the meanwhile *AMIC@* automatically starts determining the number of clusters to suggest (this is a feature of *k*-Boost and FPF-SB algorithms) for the given data set. The suggested number of clusters (*k*) can be accepted or not for running the chosen algorithms.

Clustering algorithms

Here, for each supported algorithm, we report a short description of its parameters:

- *k*-Boost does not require any compulsory parameter. It is a heuristic for the *k*-center problem conjugated with a variant of the stability-based technique for determining the number of clusters for the given data set. Nevertheless, the user has the possibility to

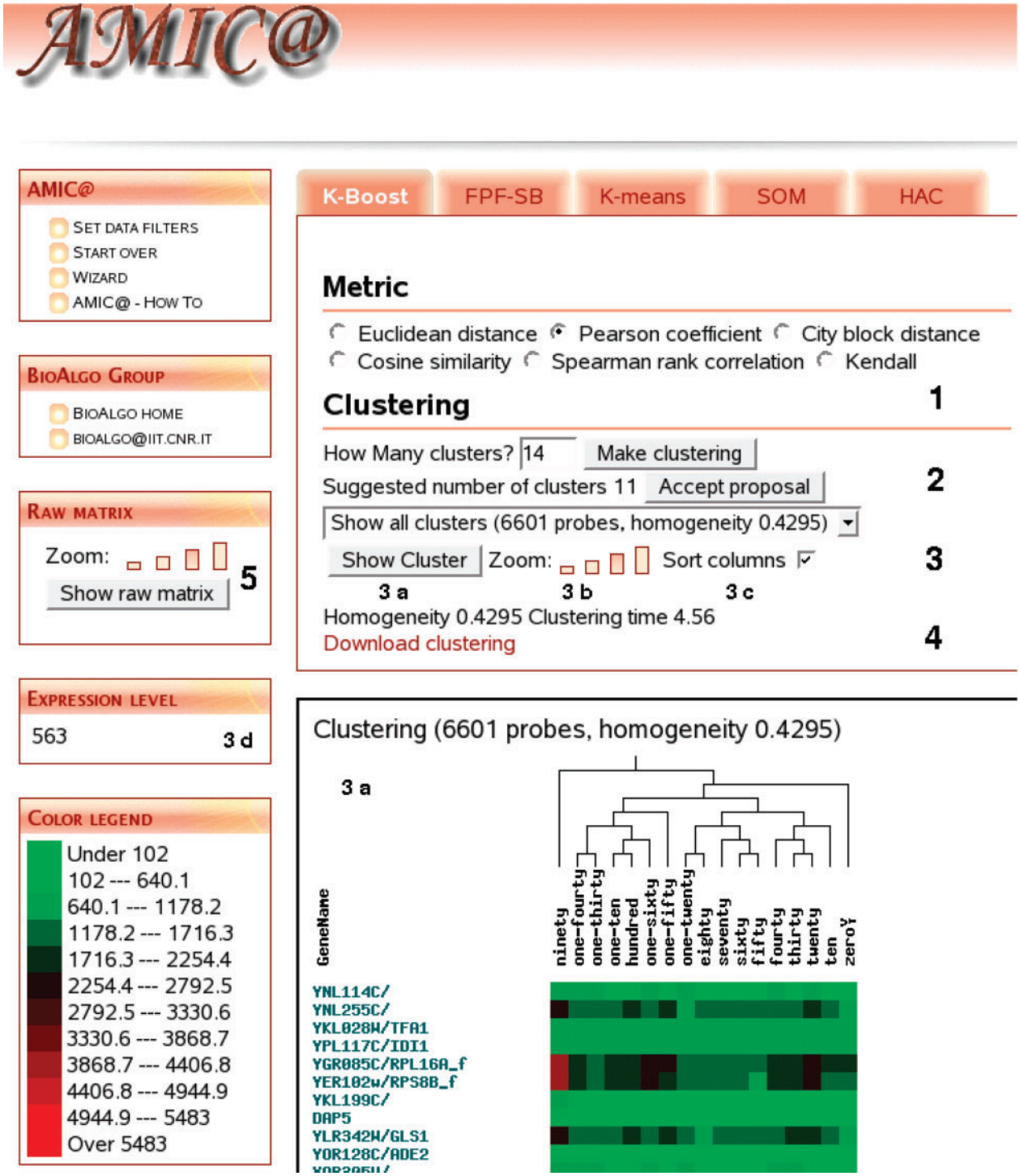

Figure 1. Cluster visualization. (1) Choose the metric; (2) Accept the number of clusters proposed or set another value; (3) Run the selected algorithm and visualize the computed clustering; For each cluster and the whole clustering it is possible to: (3a) view its heatmap representation, plus the number of probes it contains and its homogeneity value; (3b) zoom the image representing the heatmap, in and out; (3c) enable/disable the column reordering; (3d) visualize the expression level of each cell by passing the mouse over the cell; (4) Download the result as a file; (5) Visualize (and zoom) the raw matrix.

provide a different number of clusters with respect to the one proposed by *k*-Boost.

- FPF-SB does not require any (compulsory) parameter. As for *k*-Boost, the user can accept the number of clusters proposed by the algorithm, or chose another value.
- *k*-means requires the number of clusters and the number of iterations. The initial centroids are randomly selected among all the input points.
- SOM requires the size of the grid of the SOM and the number of epoques for learning. Input data are also

used for training and each node is initialized to a small random value.
- HAC can be run without a specific number of clusters (thus obtaining the whole hierarchy) or with a specified number of clusters (thus obtaining a partition). It is also possible to select the aggregation criterion to use among four standard linkage methods: single, complete, average, centroid.

When the chosen clustering algorithm has finished to run, *AMIC@* returns information about the whole clustering (running time and average homogeneity), and about each cluster (size and homogeneity), prompting an interface to visualize and download the result (see Visualization and download section for details).

Visualization and download

The visualization of the results is by means of heatmaps, since they are considered as one of the most effective visualization formats (21). When the clustering has been completed, it is possible to visualize, upon user request (Figure 1, 3a), either the single clusters, or the whole clustering, as heatmaps. In the latter case, the clusters have been sorted in decreasing homogeneity order and each cluster is highlighted by assigning different colors to the header data. By default, the heatmap columns (samples) are reordered according to the dendogram obtained by clustering the columns with HAC, thus pulling together samples with similar behavior across the genes, and the column dendogram is showed. The user can also visualize the heatmap without sample reordering (Figure 1, 3c). It is possible to visualize the expression level of each cell (Figure 1, 3d). In order to make the visualization of large data sets more comfortable, it is also provided with a zooming function (both for raw matrix and for clustering result) (Figure 1, 3b), and the clustering image is showed in a new window, on user request.

The clustering results can be downloaded as a tab delimited data file, known as clustered data file (CDT), useful for further offline investigation and analysis. In case of HAC algorithm, if it has been executed without specifying the number of clusters, the resulting file will contain a representation of the gene dendogram.

Clustering wizard

AMIC@ also provides the *Clustering Wizard*, in which the parameters for all supported algorithms can be set simultaneously, and an email address provided. *AMIC@* will send the results to the user by email when finished. This feature is very useful when clustering requires a long time to be performed because the data file is large or the chosen algorithm is slow, or when the user wants to get all the results at once.

CONCLUSIONS

The *AMIC@* web service provides a simple and uniform interface to several (currently five) clustering algorithms for gene-expression data from microarray experiments. It is a highly interactive tool, supporting automatic file format detection, intuitive visual inspection, hints on setting critical clustering parameters and output quality measurements. Large data sets can be processed efficiently by selecting algorithms (such as FPF-SB and *k*-Boost), specifically designed for this purpose. *AMIC@* stresses user-friendliness, efficiency and robustness by adopting AJAX technology.

Furthermore, *AMIC@* gives the possibility to run a Clustering Wizard to obtain via email *all the clusterings at once*. We plan to expand in time *AMIC@*, with new clustering algorithms, new visualization features, and new formats for data uploading.

ACKNOWLEDGEMENTS

We wish to thank the anonymous referees whose comments and suggestions helped in improving *AMIC@*. Funding to pay the Open Access publication charges for this article was provided by Consiglio Nazionale delle Ricerche of Italy (C.N.R.).

Conflict of interest statement. None declared.

REFERENCES

1. Montaner,D., Tárraga,J., Huerta-Cepas,J., Burguet-Castell,J., Vaquerizas,J.M., Conde,L., Minguez,P., Vera,J. Mukherjee,S., Valls,J. *et al.* (2006) Next station in microarray data analysis: Gepas. *Nucleic Acids Res.*, **34(Web-Server-Issue)**, 486–491.
2. Coessens,B., Thijs,G., Aerts,S., Marchal,K., Smet,F.D., Engelen,K., Glenisson,P., Moreau,Y., Mathys,J. and Moor,B.D. (2003) Inclusive: a web portal and service registry for microarray and regulatory sequence analysis. *Nucleic Acids Res.*, **31**, 3468–3470.
3. Grant,J.D., Somers,L.A., Zhang,Y., Manion,F.J., Bidaut,G. and Ochs,M.F. (2004) Fgdp: functional genomics data pipeline for automated, multiple microarray data analyses. *Bioinformatics*, **20**, 282–283.
4. Kapushesky,M., Kemmeren,P., Culhane,A.C., Durinck,S., Ihmels,J., Körner,C., Kull,M., Torrente,A., Sarkans,U., Vilo,J. *et al.* (2004) Expression profiler: next generation – an online platform for analysis of microarray data. *Nucleic Acids Res.*, **32(Web-Server-Issue)**, 465–470.
5. Rainer,J., Sanchez-Cabo,F., Stocker,G., Sturn,A. and Trajanoski,Z. (2006) Carmaweb: comprehensive r- and bioconductor-based web service for microarray data analysis. *Nucleic Acids Res.*, **34(Web-Server-Issue)**, 498–503.
6. Romualdi,C., Vitulo,N., Favero,M.D. and Lanfranchi,G. (2005) Midaw: a web tool for statistical analysis of microarray data. *Nucleic Acids Res.*, **33(Web-Server-Issue)**, 644–649.
7. Aburatani,S., Goto,K., Saito,S., Toh,H. and Horimoto,K. (2005) Asian: a web server for inferring a regulatory network framework from gene expression profiles. *Nucleic Acids Res.*, **33(Web-Server-Issue)**, 659–664.
8. Wu,C.-J. and Kasif,S. (2005) Gems: a web server for biclustering analysis of expression data. *Nucleic Acids Res.*, **33(Web-Server-Issue)**, 596–599.
9. Lu,Y., He,X. and Zhong,S. (2007) Cross-species microarray analysis with the oscar system suggests an insr->pax6->nqo1 neuroprotective pathway in aging and Alzheimer's disease. *Nucleic Acids Res.*, **35(Web-Server-Issue)**, 105–114.
10. Knudsen,S., Workman,C.T., Sicheritz-Ponten,T. and Friis,C. (2003) Genepublisher: automated analysis of dna microarray data. *Nucleic Acids Res.*, **31**, 3471–3476.
11. Hokamp,K., Roche,F.M., Acab,M., Rousseau,M.-E., Kuo,B., Goode,D., Aeschliman,D., Bryan,J., Babiuk,L.A., Hancock,R.E.W. *et al.* (2004) Arraypipe: a flexible processing pipeline for microarray data. *Nucleic Acids Res.*, **32(Web-Server-Issue)**, 457–459.

12. Luscombe,N.M., Royce,T.E., Bertone,P., Echols,N., Horak,C.E., Chang,J.T., Snyder,M. and Gerstein,M. (2003) Expressyourself: a modular platform for processing and visualizing microarray data. *Nucleic Acids Res.*, **31**, 3477–3482.
13. Psarros,M., Heber,S., Sick,M., Thoppae,G., Harshman,K. and Sick,B. (2005) *RACE*: remote analysis computation for gene expression data. *Nucleic Acids Res.*, **33(Web-Server-Issue)**, 638–643.
14. Sharan,R., Maron-Katz,A. and Shamir,R. (2003) Click and expander: a system for clustering and visualizing gene expression data. *Bioinformatics*, **19**, 1787–1799.
15. Saeed,A.I., Sharov,V., White,J., Li,J., Liang,W., Bhagabati,N., Braisted,J., Klapa,M., Currier,T., Thiagarajan,M. *et al.* (2003) Tm4: a free, open-source system for microarray data management and analysis. *Biotechniques*, **34**, 374–378.
16. de Hoon,M.J.L., Imoto,S., Nolan,J. and Miyano,S. (2004) Open source clustering software. *Bioinformatics*, **20**, 1453–1454.
17. Geraci,F., Leoncini,M., Montangero,M., Pellegrini,M. and Renda,M.E. (2007) Pfp-sb: a scalable algorithm for microarray gene expression data clustering. In *Proc. HCI Int. 2007. Lecture Notes in Comp. Sci.*, **4561**, 606–615.
18. Geraci,F., Leoncini,M., Montangero,M,. Pellegrini,M. and Renda,M.E. (2007) *K-boost: A Scalable Algorithm for High Quality Clustering of Microarray Gene Expression Data TR IIT-2007-015*, Istituto di Informatica e Telematica del CNR, Pisa, Italy. http://www.iit.cnr.it/staff/marco.pellegrini/.
19. Tibshirani,R., Walther,G., Botstein,D. and Brown,P. (2005) Cluster validation by prediction strength. *J. Comp. Graph. Stat.*, **14**, 511–528.
20. Troyanskaya,O., Cantor,M., Sherlock,G., Brown,P., Hastie,T., Tibshirani,R., Botstein,D. and Altman,R.B. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.
21. Saraiya,P., North,C. and Duca,K. (2004) An evaluation of microarray visualization tools for biological insight. In *10th IEEE Symp. Inform. Visual. (INFOVIS)*, 1–8.