W596–W599 *Nucleic Acids Research, 2005, Vol. 33, Web Server issue*
doi:10.1093/nar/gki469

GEMS: a web server for biclustering analysis of expression data

Chang-Jiun Wu[1] and Simon Kasif[1,2,*]

[1]Program in Bioinformatics and [2]Department of Biomedical Engineering, Boston University, Boston, MA 02215, USA

Received February 14, 2005; Revised March 30, 2005; Accepted April 13, 2005

ABSTRACT

The advent of microarray technology has revolutionized the search for genes that are differentially expressed across a range of cell types or experimental conditions. Traditional clustering methods, such as hierarchical clustering, are often difficult to deploy effectively since genes rarely exhibit similar expression pattern across a wide range of conditions. Biclustering of gene expression data (also called co-clustering or two-way clustering) is a non-trivial but promising methodology for the identification of gene groups that show a coherent expression profile across a subset of conditions. Thus, biclustering is a natural methodology as a screen for genes that are functionally related, participate in the same pathways, affected by the same drug or pathological condition, or genes that form modules that are potentially co-regulated by a small group of transcription factors. We have developed a web-enabled service called GEMS (Gene Expression Mining Server) for biclustering microarray data. Users may upload expression data and specify a set of criteria. GEMS then performs bicluster mining based on a Gibbs sampling paradigm. The web server provides a flexible and an useful platform for the discovery of co-expressed and potentially co-regulated gene modules. GEMS is an open source software and is available at http://genomics10.bu.edu/terrence/gems/.

INTRODUCTION

Advances in high-throughput microarray technology facilitate the profiling of the transcriptional level of genes on a genome-wide scale (1,2). Transcriptome data can be gathered efficiently from a large number of samples in different conditions. A major challenge in functional genomics is the elucidation of genes that are differentially expressed across a variety of cell types in a range of experimental conditions (3). Clustering analysis is a first step to understanding the functional roles of genes since genes with similar expression profiles are potentially functionally related and are often co-regulated. Traditional clustering methods, such as hierarchical (4) and *K*-means (5) approaches, consider the similarity of genes over all conditions. However, genes rarely exhibit similar expression patterns across a wide range of conditions. It is not uncommon that a group of genes is co-expressed only in particular conditions, but exhibits independent expression levels in others. In particular, genes that are involved in the same pathway (e.g. early immediate response genes in signaling pathways) are often activated in response to specific stimuli that are present only in certain conditions (6).

Heuristic biclustering algorithms have been reported (7), such as Cheng and Church (8), coupled two-way clustering (9), plaid model (10), SPLASH (11), SAMBA (12), xMotif (13) and Gibbs sampling (14). Although some of them provide software resources for a download, there are only few online biclustering servers so far (9). In this paper, we present a web server GEMS (Gene Expression Mining Server) (15) that is based on a biclustering heuristic algorithm motivated by a Gibbs sampling paradigm. This server aims to provide a valuable resource in the field of microarray and functional genomics research, with all its relative simplicity, flexibility and functionality.

One of the main advantages of GEMS as compared with other tools for clustering and biclustering is the fact that GEMS identifies clusters of genes whose mRNA abundance is constrained to be in a certain range. The majority of clustering algorithms, such as *K*-means, identify a set of samples whose expression profile is sufficiently similar. Consequently, the clusters of samples produced by these algorithms have an ellipsoid or arbitrary 'shape'. An example is shown in Figure 1a. In particular, the similarity is defined by an additive function integrated over all dimensions, but for one or more genes the range (or variance) of expression can be substantial. GEMS aims to identify sets of samples that are restricted to a particular range for each gene, forming axis-parallel hyper-rectangles (ap-bicluster) in the space defined by the selected

*To whom correspondence should be addressed. Tel: +1 617 358 1845; Fax: +1 617 353 6766; Email: kasif@bu.edu

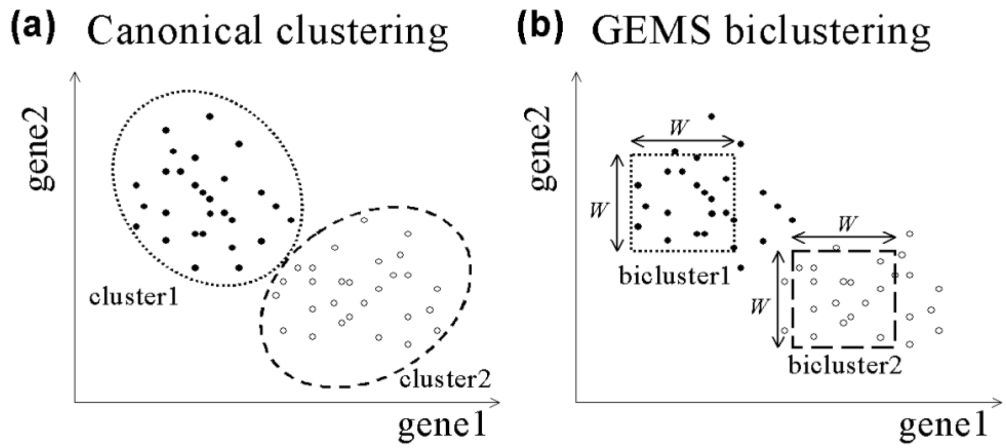

Figure 1. Examples of the shape of clusters or biclusters. (**a**) Algorithms based on similarity in overall gene expression produce ellipsoid or arbitrary shape clusters. (**b**) GEMS sets a width constraint on a subset of genes and produces axis-parallel hyper-rectangular biclusters.

genes (Figure 1b). An important aspect of the biclusters identified by GEMS is that by thresholding the expression values for each gene, a simple Boolean rule describing the cluster may be obtained. The rule is a conjunction of conditions. An example of such a conjunctive bicluster is a set of samples where gene A is upregulated, gene B is downregulated and gene C is upregulated. Traditional clustering techniques do not necessarily provide such a natural qualitative description.

OVERVIEW OF THE SERVICE

Given a microarray dataset collected from a set of patients, different tissue types or a set of conditions, a researcher may want to glean one or more biclusters from this dataset, estimate the statistical significance of each bicluster and save the extracted submatrices into array files with the same format. After the GEMS server receives an array file uploaded by the user, the first step is to preprocess the data. Expression values of mRNAs below some threshold may be considered as random noise rather than a signal. Users can set a threshold value to filter out the genes that have low-expression values in most of the samples. Different genes typically have a wide range of expression levels and variances; therefore, it is frequently desirable to normalize the expression values into a fixed range or equal variance. The GEMS server includes an elementary preprocessing facility to normalize the data at the request of the user.

The second step in the pipeline is the deployment of a sampling algorithm to find a subset of samples corresponding to a maximal subset of genes. We provide a statistical motivation for this approach and a workflow chart describing the methods in Supplementary Material. Two parameters, the size constraint α and width constraint W, are applied to define an acceptable bicluster. The α parameter sets a lower bound for the fraction of samples included in a bicluster. In particular, we insist that each bicluster contains at least αS samples where S is the total number of samples in the microarray study. The typical value of α is problem-dependent, e.g. corresponds to the size of the known pathways implicated in a particular disease or process. The W parameter constrains the expression range of each gene included in the bicluster. Recall that our biclusters are axis parallel thus $W = 0.1$ means that maximal length of the side of any side of the hyper-rectangle is 0.1. Biologically, this constraint implies that for any gene included in a bicluster: the difference between the maximum and the minimum values of the samples included in the bicluster is at most $W = 0.1$.

The Gibbs sampling is a stochastic simulation process (16), which repeatedly samples from a distribution defined over biclusters aiming to identify an optimal bicluster, i.e. a bicluster that satisfies the user-defined parameters and has the maximum number of genes. The longer the sampling process, the higher is the probability that the result will approximate the global optimum. The GEMS users can choose either faster execution with a shorter lag period or seek to obtain a 'better' result sacrificing response time. The program can also detect multiple biclusters, and users have an option to ask the server for unique biclusters. Earlier extracted biclusters can be masked to avoid overlapping between biclusters and to speed up the searching process. Three different masking methods can be selected: (i) masking early extracted gene clusters, (ii) masking previously detected sample subsets and (iii) masking selected genes on selected samples in earlier biclusters.

The final computational step performed by GEMS uses a local search step to refine the bicluster. More details about the algorithms and the user-defined parameters can be found in Supplementary Material. After completion of the above three steps, the GEMS server sends an email to notify the user and provides a website address where the results can be downloaded.

IMPLEMENTATION

The GEMS web service was developed using dynamic CGI scripts in Perl language, and is available at http://genomics10.bu.edu/terrence/gems/. The core GEMS program is implemented in C++. The server is currently running on a machine with dual Intel Pentium III 900 MHz, 2 GB of RAM, Redhat Linux version 2.4.20 and an Apache web server. The core program of GEMS can be downloaded as a standalone application on a local machine, which works successfully in both Linux/Unix and Windows platforms.

INPUT

The GEMS web interface requires users to input their email address and upload microarray expression data in a tab-delimited plain text file. The format of the expression data file is similar to the commonly used formats in many gene expression datasets. The first column contains a unique identifier name for each gene and the second column contains the descriptive text about each gene. The following columns contain gene expression data, where one column is allocated for each microarray sample. The first row starting with the third column contains names of samples. The following rows contain the expression data with one row per gene. The current version of the web server accepts up to 50 000 genes and 512 samples. Users have to specify the width constraint and size constraint used to search for the biclusters. The GEMS server generates one bicluster for every query by default, but users can choose to request multiple biclusters and select a method to mask earlier results.

OUTPUT

The biclustering results can be queried by users on the GEMS website. For each submitted task, GEMS will generate a report containing the parameter setting, the number of biclusters extracted, the number of samples and genes as well as the permutation *P*-value for every bicluster. For each bicluster extracted, four files will be generated and packed in a zip file, including one expression matrix file, one heatmap image file and two index files indicating which samples and which genes are selected in the biclusters. The matrix and the index output files are tab-delimited plain text files, which can easily be imported into other software, such as the statistical package R (17), for further analysis.

An example query result is illustrated in Figure 2. The array file comes from the T-matrix cDNA microarray data of the NCI 60 Cancer Cell Lines (18) containing the expression profiles of 1375 genes. GEMS projects all expression values to a range from 0 to 1 and detects biclusters with at least 10 samples and 10 genes. We use a width constraint $W = 0.1$ to limit the expression range of genes, and three biclusters are detected.

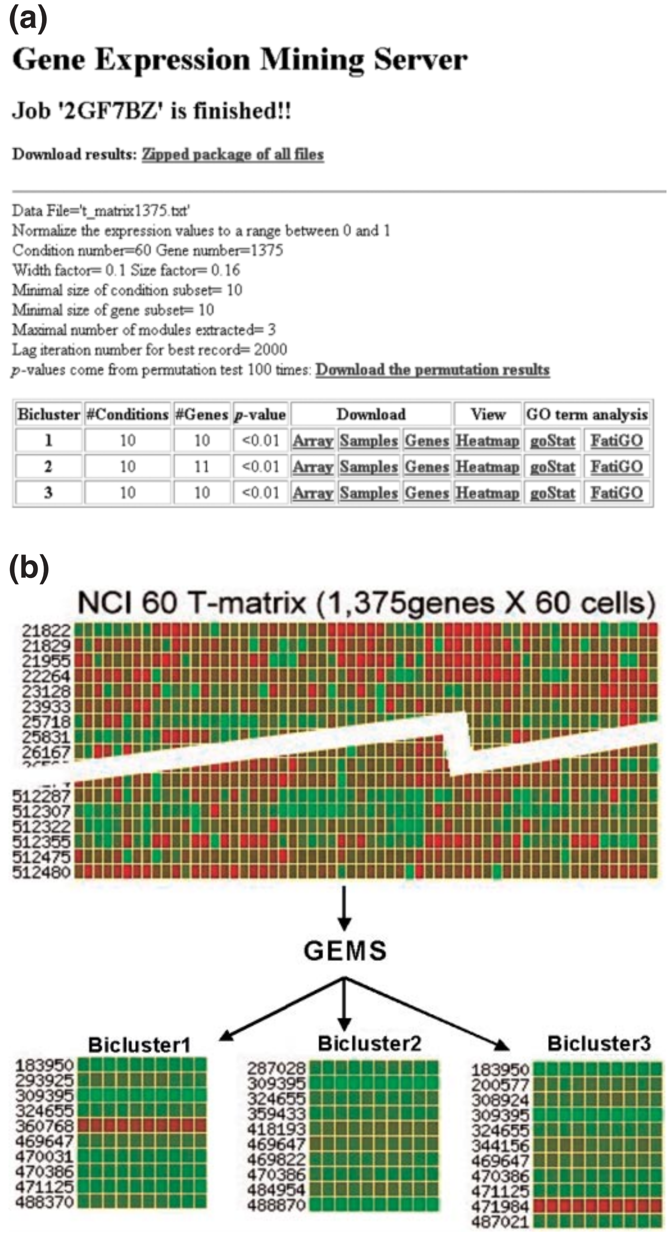

Figure 2. Illustration of GEMS output using NCI60 cDNA expression data as an example. (**a**) Three biclusters are detected: the numbers of genes in the biclusters are 10, 11 and 10, respectively. (**b**) The heatmaps of original T-matrix cDNA expression dataset (truncated) and three extracted biclusters. The expression values of every gene in the biclusters are consistent across the subset of samples.

DISCUSSION

The web service of GEMS is a user-friendly interface for biclustering analysis of microarray expression data that aims to identify locally conserved biclusters. Each bicluster can be considered as a 'hypothesis generator' for future follow-up by the researchers using the system. In particular, the typical follow-up steps might include a literature search to see if pairs of genes included in the biclusters have been previously found to be associated with each other (e.g. as co-factors or part of the same protein–protein interaction network). In addition, the information produced by GEMS can be integrated with other sources of information. For instance, we can integrate the gene sets produced by GEMS with information on protein–protein networks to produce more reliable functional gene annotation (19).

Using a function-oriented labeling scheme, it is also possible to classify the selected genes and samples into functional categories of interest. We are adding this ability to the server that in particular will provide an enrichment score of the genes in a particular functional category among the genes selected for the bicluster (20,21). In the future studies, the interface will be updated and enhanced with a number of other features that include a functional enrichment test, a semi-supervised learning schema and a probabilistic version of biclustering that are currently in development. Each update will be highlighted and reported on the web page.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We thank Yutao Fu and T. M. Murali for their help in developing the GEMS algorithm. We also thank Megon J. Walker and Mike Schaffer for their valuable suggestions. This work is supported in part by NSF grants DBI-0239435 and ITR-048715, NHGRI grant #1R33HG002850-01A1 and NIH grant U54 LM008748. Funding to pay the Open Access publication charges for this article was provided by NSF grant ITR-048715.

Conflict of interest statement. None declared.

REFERENCES

1. Mootha,V.K., Lindgren,C.M., Eriksson,K.F., Subramanian,A., Sihag,S., Lehar,J., Puigserver,P., Carlsson,E., Ridderstrale,M., Laurila,E. *et al.* (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genet.*, **34**, 267–273.
2. Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J.P., Coller,H., Loh,M.L., Downing,J.R., Caligiuri,M.A. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.
3. Su,Y., Murali,T.M., Pavlovic,V., Schaffer,M. and Kasif,S. (2003) RankGene: identification of diagnostic genes based on expression data. *Bioinformatics*, **19**, 1578–1579.
4. Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.
5. Herwig,R., Poustka,A.J., Muller,C., Bull,C., Lehrach,H. and O'Brien,J. (1999) Large-scale clustering of cDNA-fingerprinting data. *Genome Res.*, **9**, 1093–1105.
6. Tullai,J.W., Schaffer,M.E., Mullenbrock,S., Kasif,S. and Cooper,G.M. (2004) Identification of transcription factor binding sites upstream of human genes regulated by the phosphatidylinositol 3-kinase and MEK/ERK signaling pathways. *J. Biol. Chem.*, **279**, 20167–20177.
7. Madeira,S.C. and Oliveira,A.L. (2004) Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM TCBB*, **1**, 24–45.
8. Cheng,Y. and Church,G.M. (2000) Biclustering of expression data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 93–103.
9. Getz,G., Levine,E. and Domany,E. (2000) Coupled two-way clustering analysis of gene microarray data. *Proc. Natl Acad. Sci. USA*, **97**, 12079–12084.
10. Lazzeroni,L. and Owen,A. (2002) Plaid models for gene expression data. *Stat. Sin.*, **12**, 61–86.
11. Califano,A., Stolovitzky,G. and Tu,Y. (2000) Analysis of gene expression microarrays for phenotype classification. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 75–85.
12. Tanay,A., Sharan,R. and Shamir,R. (2002) Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, **18**, 136–144.
13. Murali,T.M. and Kasif,S. (2003) Extracting conserved gene expression motifs from gene expression data. *Pac. Symp. Biocomput.*, **2003**, 77–88.
14. Sheng,Q., Moreau,Y. and De Moor,B. (2003) Biclustering microarray data by Gibbs sampling. *Bioinformatics*, **19** (Suppl. 2), II196–II205.
15. Wu,C.J., Fu,Y., Murali,T.M. and Kasif,S. (2004) Gene expression module discovery using Gibbs sampling. *Genome Inform. Ser. Workshop Genome Inform.*, **15**, 239–248.
16. Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F. and Wootton,J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.
17. Ihaka,R. and Gentleman,R. (1996) R: a language for data analysis and graphics. *J. Comput. Graph. Stat.*, **5**, 299–314.
18. Scherf,U., Ross,D.T., Waltham,M., Smith,L.H., Lee,J.K., Tanabe,L., Kohn,K.W., Reinhold,W.C., Myers,T.G., Andrews,D.T. *et al.* (2000) A gene expression database for the molecular pharmacology of cancer. *Nature Genet.*, **24**, 236–244.
19. Karaoz,U., Murali,T.M., Letovsky,S., Zheng,Y., Ding,C., Cantor,C.R. and Kasif,S. (2004) Whole-genome annotation by using evidence integration in functional-linkage networks. *Proc. Natl Acad. Sci. USA*, **101**, 2888–2893.
20. Beissbarth,T. and Speed,T.P. (2004) GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, **20**, 1464–1465.
21. Al-Shahrour,F., Diaz-Uriarte,R. and Dopazo,J. (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, **20**, 578–580.