Journal of Proteomics & Bioinformatics - Open Access

Research Article JPB/Vol.2/May 2009

CIT: A Cluster Identification Tool based on Biclustering and Hierarchical Clustering

Tabinda Hussain[1], Ammara Mazhar[1], Ammad-ud-din[2], Asif Mir[1]*

[1]Department of Biosciences, COMSATS Institue of Information Technology (CIIT), Chak Shazad Campus, Islamabad-44000, PAKISTAN

[2]Department of Bioinformatics, Qauid-i-Azam University, Islamabad-44000, PAKISTAN

*Corresponding author: Asif Mir, Department of Biosciences, COMSATS Institue of Information Technology (CIIT), Chak Shazad Campus, Islamabad-44000, PAKISTAN, E-mail: mir77uspk@gmail.com; Tel: 92-323-5022292

Received May 10, 2009; Accepted May 16, 2009; Published May 16, 2009

Citation: Hussain T, Mazhar A, Din AU, Mir A (2009) *CIT*: A Cluster Identification Tool based on Biclustering and Hierarchical Clustering. J Proteomics Bioinform 2: 222-225. doi:10.4172/jpb.1000080

Abstract

Cluster analysis is one of the most popular techniques applied in microarray data studies. Thousands of genes can be analyzed within minutes if cluster analysis is embedded in a computational tool. With such modern technologies, it has now become easier to find practical manifestations of microarray data in the fields of pharmacogenomics, cancer genetics and biological network construction. With this project work, we have developed a cluster identifying tool, i.e. CIT which is based on two different clustering methodologies namely; Biclustering and Hierarchical Clustering. We intend to embrace new possibilities in CIT in future, like; dendogram view, interactive outputs etc.

Keywords: Gene expression data; Microarray; Dendogram; Clustering; Hierarchical clustering; Biclustering

Abbreviations

HC:	Hierarchical Clustering;
CC:	Cheng and Church;
SAMBA:	Statistical-Algorithmic Method for Bicluster Analysis;
AHC:	Agglomerative Hierarchical Clustering;

Background

Rapid advances in genome-scale sequencing has led to immense increase in the amount of biological information .Simply visualizing this kind of data which is widely called gene expression data or simply expression data is challenging and extracting biologically relevant knowledge is harder still (Eisen et al., 1998)

By knowing groups of genes that are expressed in a similar fashion through a biological process, biologists are able to infer gene function and gene regulation mechanisms (Quackenbush, 2001; Slonim, 2002). Since these data consist of expression profiles of thousands of genes, their analysis cannot be carried out manually, making necessary the application of computational methods which are included under the domain of microarray data analysis techniques.

Microarray Data Analysis

Microarrays and high-throughput sequencing methods can be used to measure the expression of thousands of genes in a biological sample in a few days. A natural follow-up to such experiments is organizing and inferring useful information from this data [Risques et al., 2008]. Microarray technology is although a powerful technique but it relies heavily on the availability of computational methods which

help in the array design, microarray image analysis, storage of microarray data and lastly the comparison of expression profiles to achieve functional interpretation of groups of genes (which were studied in the initial experiment) [Tamames et al., 2002]. We are presenting a cluster analysis tool named CIT which can perform gene expression data analysis.

Implementation

CIT (Cluster Identification Tool) has been made to perform cluster analysis on genes based on two different methods, namely; Biclustering and Hierarchical Clustering. A brief overview of both algorithms and how they are implemented in this tool is as under:

Hierarchical clustering builds a cluster hierarchy or, in other words, a tree of clusters, also known as a dendrogram. Every cluster node contains child clusters; sibling clusters partition the points covered by their common parent. The type of clustering that we have used is called Agglomerative Hierarchical Clustering (AHC). AHC, agglomerative approach is the one where each entity/gene is taken as a single cluster and at each step the cluster is expanded.

Biclustering algorithms do not belong to traditional data-mining techniques. Simple clustering methods can be applied to either the rows or the columns of the data matrix. Contrarily a more focused version of clustering is 'biclustering'; where simultaneous row and column clustering takes place. A bicluster (or a module) is a subset of the genes exhibiting consistent patterns over a subset of the conditions.

The algorithm used in CIT to perform biclustering is the Statistical-Algorithmic Method for Bicluster Analysis (SAMBA) [Tanay et al., 2002]. SAMBA is incorporated in the cluster analysis tool called Expander [Shamir et al., 2005]. Using a statistical model for the data, normalization is done by translating the gene expression matrix to a weighted bipartite graph.

Results

The objective of making the proposed cluster analysis tool is to outline the behaviour of genes in biological processes. In addition, the need for making cluster analysis tools is due to the large amounts of data generated by whole-genome expression profiling, aided by the advent of

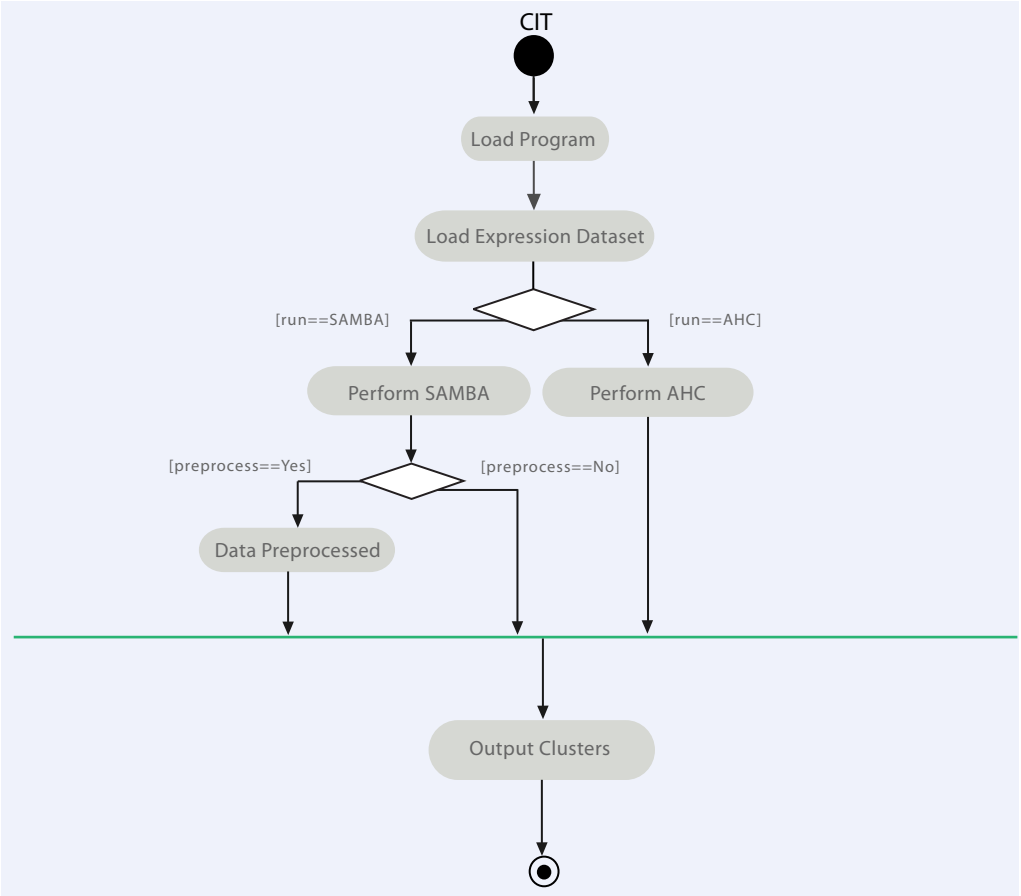

Figure 1: The figure is illustrating the Activity diagram of CIT, it show how one step follows the other in CIT.

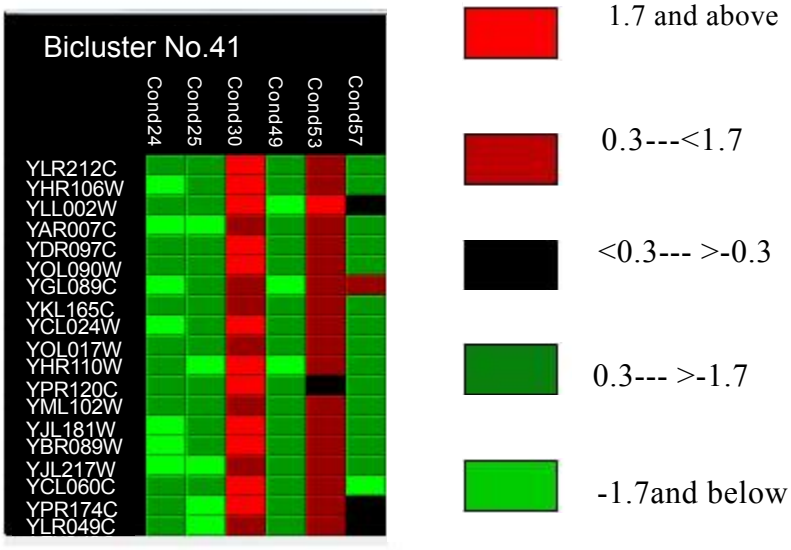

Figure 2: The figure is illustrating the expression profile of a bicluster in the form of heatmap. The red colors are showing the up-regulation of gene expression while the green colors signify the down regulation in expression. The top row has conditions and the first column has genes.

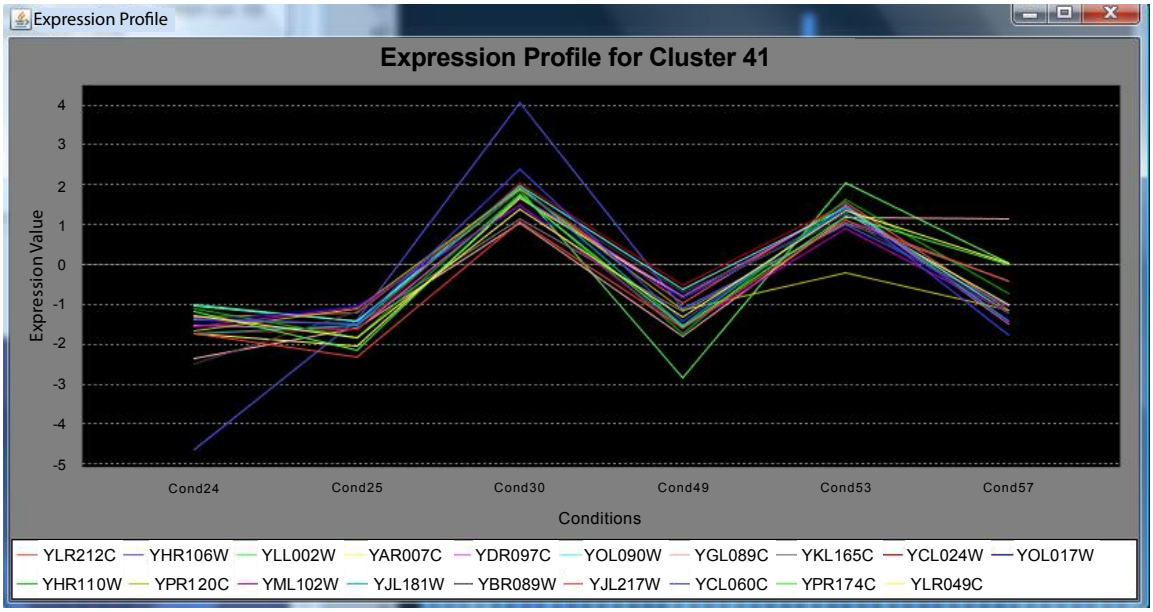

Figure 3: The figure is illustrating the expression profile of a bicluster. The expression pattern is spanned over a set of conditions (x-axis) on which the Bicluster is based. The y-axis has the expression values varying from negative to positive values.

microarray technology, which needs to be interpreted to construct biological networks [Hughes et al., 2000; Zhu et al., 2007]. The way clustering is performed in CIT is illustrated in the following steps. (Refer to Figure.1).

Data Normalization

The tool will take a microarray dataset as its input and pre-process the loaded dataset, if required, and then analyze this data through the selected algorithm. Normalizing the data is also called as data pre-processing. Cluster analysis tools frequently incorporate options to pre-process the data. This helps in bringing the data in a standardized range. The normalization applied in CIT is 'Statistical Normalization with mean 0 and variance 1'.

Selection of Clustering Method

After the normalization or pre-processing step the user will specify algorithm that he wants to apply on the normalized dataset. After the analysis has been performed by either AHC or SAMBA the results will be displayed in the form of charts, graphs or tables to outline the clusters.

Visualization of Clusters

Initially when the dataset has been loaded the gene expression matrix can be viewed as a heat map (Refer to Figure 2) which is in the form a coloured map mimicking the pattern of fluorescence from a microarray chip. Bright red signifies up-regulation of expression while green indicates down regulation in expression. After the clustering has been performed, the Bicluster is shown as a heat map and a line graph while the AHC clusters are shown in the form of bar charts and line graphs (Refer to Figure. 3).

Conclusion

Clustering is the classification of objects into different groups, the grouping of gene expression data is usually carried out with cluster analysis. Traditionally clustering techniques are divided in two categories, namely hierarchical and partitional. Biclustering constructs a subset of genes exhibiting consistent pattern over a subset of conditions. Using the techniques significant biclusters and clusters are generated in an unsupervised manner.

This cluster analysis tool, CIT, has the ability to pre-process the data if required by user. The pre-processing method implemented by CIT is statistical normalization with mean 0 and variance 1, which is a commonly used efficient data normalization technique. The tool uses two diverse approaches to perform clustering. Simple clustering is carried out with the popular Agglomerative hierarchical clustering (AHC) algorithm. CIT can also perform biclustering with SAMBA (Statistical-Algorithmic Method for Bicluster Analysis), SAMBA is a relatively newer method in the field of Biclustering as compared to CC algorithm [Cheng and Church, 2000] which is commonly used for biclustering. SAMBA has improved performance and can handle datasets with thousands of conditions profiled over large no. of genes.

So far CIT is compatible to run in Windows Operating System only, it cannot read data from file formats that are other than the .txt format and does not generate dendogram in the output of AHC. In the future we intend to update our tool with innovations like, a tree view of AHC and multiple functionalities with interactive outputs.

System Requirements and Availability

For access to CIT, Contact us at: 123@gmail.com

Operating System: Windows XP and higher
Programming Language: Java
Runtime Environment: JRE 5 and higher

References

1. Cheng Y, Church GM (2000) Biclustering of expression data. In *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology*. (ISMB'00): 93–103. » CrossRef » Pubmed » Google Scholar

2. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 98: 14863–14868. » CrossRef » Pubmed » Google Scholar

3. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, et al. (2000) Functional discovery via a compendium of expression profiles. Cell 102: 109-126. » CrossRef » Pubmed » Google Scholar

4. Quackenbush J (2001) Computational analysis of cDNA microarray data. Nature Review on Genetics 6: 418-428. » CrossRef » Pubmed » Google Scholar

5. Risques RA, Rondeau G, Judex M, McClelland M, Welsh J (2008) Assessment of gene expression in many samples using vertical arrays. Nucleic Acids Research, Advance Access *(in printing)*: 1-9.

6. Shamir R, Maron-Katz A, Tanay A, Linhart C, Steinfeld I, et al. (2005) *EXPANDER* – an integrative program suite for microarray data analysis. BMC Bioinformatics 6: 232.» CrossRef » Pubmed » Google Scholar

7. Slonim D (2002) From patterns to pathways: Gene expression data analysis comes of age. Nature Genetics 32: 502-508. » Pubmed » Google Scholar

8. Tamames J, Clark D, Herrero J, Dopazo J, Blaschke C, et al. (2002) Bioinformatics methods for the analysis of expression arrays: data clustering and information extraction. Journal of Biotechnology 98: 269-283. » Pubmed » Google Scholar

9. Tanay A, Sharan R, Shamir R (2002) Discovering statistically significant biclusters in gene expression data. Bioinformatics 18: 136-144. » Pubmed » Google Scholar

10. Zhu X, Gerstein M, Snyder M (2007) Getting connected: analysis and principles of biological networks. Genes and Development 21: 1010-1024. » CrossRef » Pubmed » Google Scholar