Full Length Research Paper

ClustPK: A windows-based cluster analysis tool

Masood ur Rehman Kayani, Umair Shahzad Alam, Farida Anjum and Asif Mir*

Department of Biosciences,COMSATS Institute of Information Technology,Bio-Physics Block, Chak Shahzad Campus,Islamabad-44000, Pakistan.

There is a great need to develop analytical methodologies to analyze and exploit the information contained in gene expression data obtained from microarray-based experiments. Because of large number of genes and complexity of biological networks, clustering is a useful exploratory technique for analysis of such data. Different data analysis techniques and algorithms have been developed which are used to cluster the gene expression data. Various tools have been developed that implement these algorithms. Clusters of co-expressed genes provide useful basis for further investigation of gene function, regulation and their possible involvement in causing different diseases. ClustPK has been developed using C# .NET and implementing *k-means* and PCA algorithms. Analysis of microarray data using the already existing tools is difficult and the results are also hard to be analyzed. While, ClustPK is an easy-to-use and user friendly tool that provides the easy visualization and analysis of the results obtained from either *k-means* or PCA.

Key words: Microarray, gene expression, data sets, cluster analysis, k-means, principle component analysis.

INTRODUCTION

Genes are responsible for the functionality of various cellular components and are also important for the phenotype of an organism. Expression of genes under different conditions has a different effect on cell proliferation, differentiation and on various other cellular processes (Leung and Cavalieri, 2003; Ahmed, 2002). Traditional methods of analyzing the gene expression are either too time consuming, difficult to automate or analyze one mRNA at a time. Multiple mRNAs can be analyzed by using the newly developed techniques such as microarrays (Ahmed, 2002; Jiang et al., 2004).

Microarray technology is very powerful and comes with enormous benefits. Use of this technology can provide an opportunity to identify new drug targets and finding out what effects are produced by a drug on the expression of a gene (Brazma et al., 2001; Eisen et al., 1998). It has also been used for the prediction of function of various genes that may be involved in causing diseases under certain conditions (Burgess, 2001; Garaziar et al., 2006).

Microarray has also been used for expression profiling of immune cells (Ambrosio et al., 2005; Subaramanya et al., 2003).

After performing a microarray experiment, data analysis is required for the interpretation of results. This analysis involves multiple steps including acquiring image of the microarray, image analysis, data preprocessing and normalization (Brazma et al., 2001). The processed data is represented as a gene expression matrix with rows representing the genes included in the experiment and columns representing the conditions under which genes were studied. This matrix can be manipulated for cluster analysis (Brazma et al., 2001). Cluster analysis is a technique that clusters genes into different groups on the basis of similarities in their expression under different conditions. Genes present in one group have high similarity with each other and highly dissimilar with genes present in a different group. Computational tools that are based on clustering algorithms (including *k-means*, Principle Component Analysis, Hierarchical Clustering and Biclustering) are required to achieve the clustering of genes (Jiang et al., 2004).Various tools have been developed for performing the cluster analysis using different algorithms. Most of the existing softwares do not provide an easy interface for their usage and the analysis of a dataset as well. For overcoming these limitations, ClustPK has been developed.

---

*Corresponding author. E-mail: asif_mir@comsats.edu.pk. Tel: +92-323-5022292.

Abbreviations: C#; C Sharp, **PC;** Principle Component, **PCA;** Principle Component Analysis.
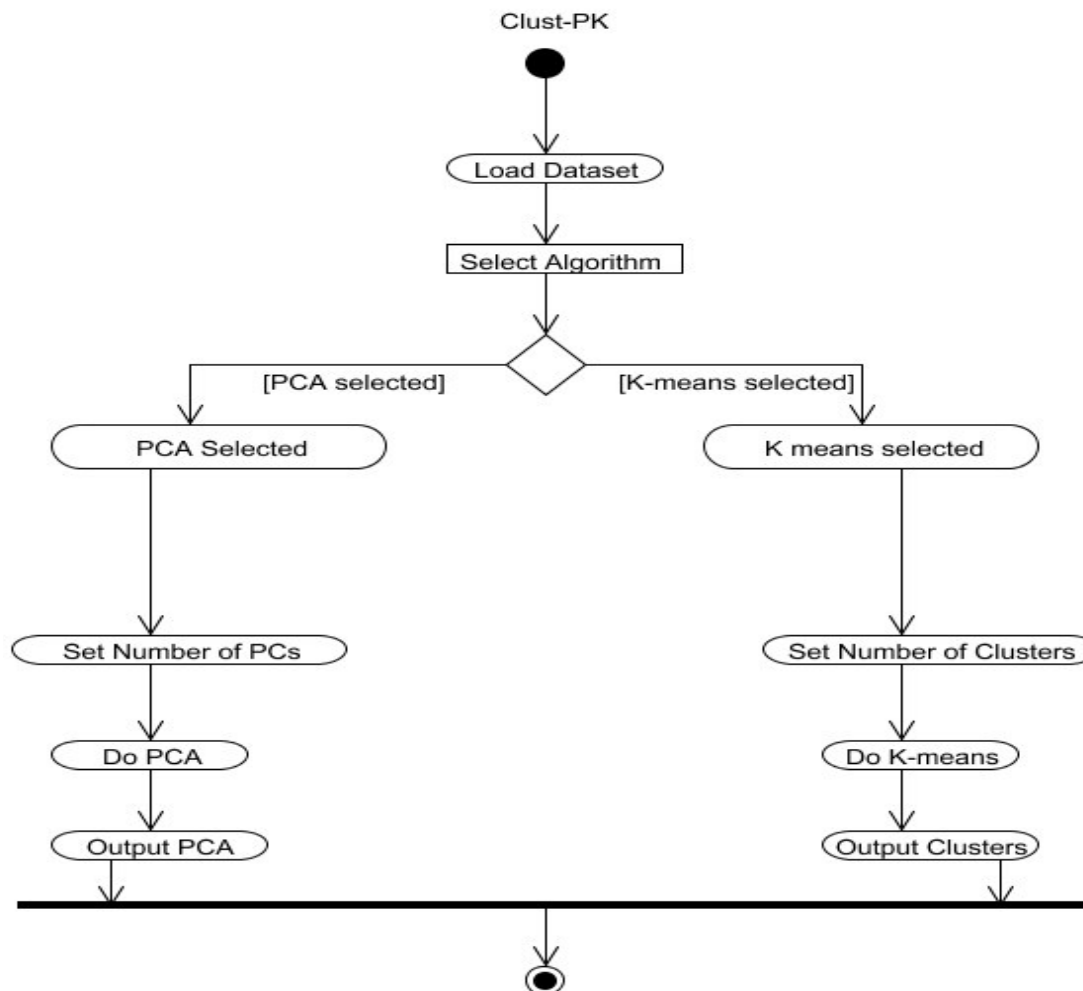
Figure 1. Activity Diagram of ClustPK. A user begins by uploading the dataset to ClustPK and then proceeds with the selection of an appropriate algorithm. In case *k-means* is selected, he/she will be prompted to input the number of clusters that is, *k* then *k-means* is performed and finally clusters are output. In case PCA is selected, he/she will be asked to enter the number of PCs after which PCA will be performed and finally output will be provided.

Program description

ClustPK has been developed using C# .NET (C-sharp dot net) and has an implementation of the following two clustering algorithms:1. *k-means* that groups the given genes into *k* number of clusters.Here *k* is a positive integer value that should be at least 2 and less than the total number of genes in an experiment. This number is selected by the user and must be input before performing *k-means*. 2. Principle Component Analysis (PCA) is a data-dimension reduction technique and identifies patterns in a given dataset and expresses them in a way that highlights their similarities and differences. Figure 1 is the activity diagram of ClustPK and shows different steps required to perform clustering using ClustPK.

Program interface

Heat map

Heat map (Figure 2) is generated immediately after the user inputs

a valid dataset to ClustPK. Heat map is a graphical representation of the expression values present in the dataset. Different expression values are represented by different colors.

Expression view

Expression view shows a graphical representation of the results after the input of a valid dataset and performing clustering using any of the two algorithms. ClustPK displays the results as a bargraph for *k-means* (Figure 3A) and for PCA, a scatter plot is displayed (Figure 3B).

Analysis view

For further analysis of a single cluster/PC, user can select that cluster/PC from the analysis menu. For *k-means*, a list of genes and a graph (showing the range of expression values in selected cluster) appears in the Analysis view (Figure 3C). For PCA, a list of final values obtained after performing statistical and linear

Figure 2. Heat map of the loaded dataset. Differences in expression values are indicated by different colors.
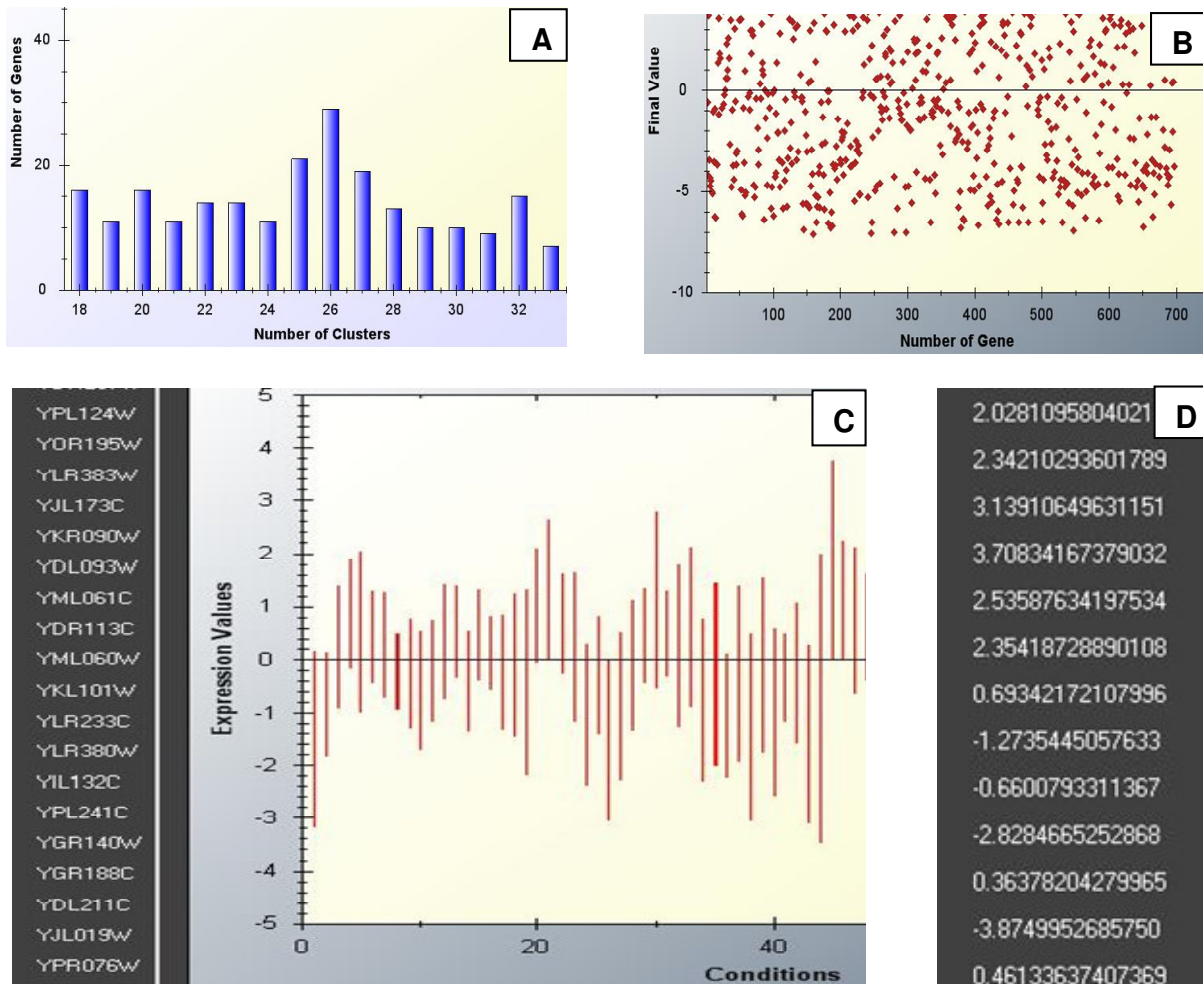

Figure 3. Clustering Session in ClustPK. (A) Bar graph shows the results of *k-means* that is, number of clusters is shown on x-axis while number of genes in a cluster is shown on y-axis by height of bar. (B) Scatter plot of PCA. (C) Analysis view of a cluster that is, for a selected cluster in *k-means*, genes included that cluster and their expression value ranges are plotted. (D) Analysis view of a PC shows the final transformed data values in algebraic calculations is shown in the analysis view (Figure 3D).

Saving project

During each session of *k-means* or PCA, ClustPK generates different intermediate files. For *k-means*, files for the 1-0 matrix, Euclidean distance and Gene ID are generated. For PCA, these files include the files for covariance matrix, Eigen vector, Feature Vector and Final transformed data. The graphs generated in a session can also be saved and printed.

DISCUSSION

Microarray analysis is widely used for studying gene expression data. Clustering is used for grouping the given objects into distinct groups so that the objects within one group have high similarity and objects in separate groups are more dissimilar. An object refers to a gene or an experimental condition.

algebraic calculations is shown in the analysis view (Figure 3D).Clustering techniques are applied on the microarray datasets using different available software tools. ClustPK applies two clustering algorithms that is, *k-means* and PCA on the microarray gene expression datasets. *k-means* clusters the objects into a predefined number of clusters that is, *k*. Principle Component Analysis (PCA) is used to reduce the dimensions of a given dataset and can also be used for clustering microarray gene expression dataset. ClustPK, a user friendly tool, provides the utility to visualize and analyze the results of *k-means* or PCA. The results are visualized as a fully zoom-able and pan-able graph. Designed tool provides a gene list for a selected cluster and the final transformed values along with their respective gene ID for a selected PC as analysis of results.

ClustPK can be used to analyze only one type of microarray dataset that is, the dataset in a text format only. Future developments of ClustPK include: (1) analysis of more than one format of microarray datasets that is, CEL, GPR, CHP formats, (2) clustering of microarray datasets using biological networks that is, metabolic network, gene networks or any other type of network and (3) annotation of clusters.

System Requirements and Availability

ClustPK has been developed using C# language and based on the .NET technology. The main requirements for running this software include:

a) Windows XP/ Windows Vista.
b) .NET framework 2.0 or higher.
c) 512 MB of RAM (minimum).

For installation of .NET framework, windows installer 3.0 or later is required. Windows installer and .NET framework can be downloaded from Microsoft's website. ClustPK is freely-available software and can be downloaded from our web site (http://www.bioinformaticshub.com/softwares/ClustPK.rar).

REFERENCES

Ahmed EF (2002). Molecular techniques for studying gene expression in carcinogenesis. J. Env. Sci. Health 20: 77-116.

Ambrosio DC, Gatta L, Bonin S (2005). The future of microarray technology: networking the genome search. *Allergy* 60: 1219-1226.

Brazma A, Hingamp P, Quakenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball AC, Causton CH, Gaasterland T, Glenisson P, Holstege CPF, Kim FI, Markowitz V, Matese CJ, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M (2001). Minimum information about a microarray experiment (MIAMI)-toward standards for microarray data. Nature Genetics 29: 365-371.

Burgess KJ. Special Technical Review (2001). Gene expression studies using microarrays. Clin. Exper. Pharmacol, Physiol. 28: 321-332.

Eisen MB, Spellman PT, Brown P, Botstein D (1998). Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. 95: 14863-14868.

Garaziar J, Rementeria A, Porwollik S (2006). DNA microarray technology: a new tool for the epidemiological typing of bacterial pathogens. *FEMS* Immun. Med. Microbiol. 47: 178-189.

Jiang D, Tang C, Zhang A (2004). Cluster Analysis for Gene Expression Data: A Survey. *IEEE* Transactions on Knowledge and Data Engineering 16: 1370-1386.

Leung FY, Cavalieri D (2003). Fundamentals of cDNA microarray data analysis. Trends in Genetics 19 : 649-659.

Subaramanya DR, Lucchese G, Kanduc D, Sinha AA (2003). Clinical applications of DNA microarray analysis. J. Exp. Therap. Oncol. 3: 297-304.