

A Heuristic Approach to Scoring Gene Clustering Algorithms

Longde Yin

Dept. of Computer Science & Engineering
University of Connecticut
Storrs, CT06269, USA
yin@engr.uconn.edu

Chun-Hsi Huang

Dept. of Computer Science & Engineering
University of Connecticut
Storrs, CT06269, USA
huang@engr.uconn.edu

Abstract

In the past decades, many clustering algorithms have been proposed for the analysis of gene expression data, but little guidance is available to help choose among them. Given the same data set, different clustering algorithms can potentially generate very different clusters. A biologist with a gene expression data set is faced with the problem of choosing an appropriate clustering algorithm for his or her data set. In this paper, we present a new tool that allows the similarity analysis of clusters generated by different algorithms. This tool may: (1) improve the quality of the data analysis results, (2) support the prediction of the number of relevant clusters in the Microarray datasets, and (3) provide cross-reference between different algorithms. The software tool can also be used to analyze cluster similarities from other biomedical data. We demonstrate the use of this tool with gene expression data of Leukaemia and Sporulation.

Keywords: Clustering algorithms, Gene expression, Microarray, Cluster Similarity Analysis.

1. Introduction

Recent advances of the DNA Microarray technology allow monitoring gene expression profiles of thousands of genes simultaneously^[1]. However, the analysis and handling of such fast growing data is becoming the major bottleneck in the utilization of the technology. Clustering analysis is one of the most effective methods for analyzing such gene expression data^[2, 14].

A lot of clustering methods have been proposed for the analysis of gene expression data, but little guidance is available to help choose among them. Assessing the clustering results and interpreting the clusters found are as important as generating the clusters^[13, 22]. Given the same data set, different clustering algorithms can potentially generate very different clusters. A biologist with a gene expression data set is faced with the problem of choosing an appropriate clustering algorithm for his or her data set. Our paper provides a data mining tool, *Cluster Diff*, which allows the similarity analysis of clusters generated by different algorithms. This tool may: (1) improve the quality of the data analysis results, (2) support the prediction of the number of relevant clusters in the microarray datasets, and (3) provide cross-reference between different algorithms. The software tool can also be used to analyze cluster similarities from other biomedical data.

In this paper we first introduce this software tool, then, as an application example, we apply this software to analyze the clusters generated by K-means^[5, 19, 20], Cluster Identification via Connectivity Kernels (CLICK)^[23], and Self-Organizing Map (SOM)^[8, 9].

The remainder of the paper is organized as follows. In Section 2 we present the software tool in detail. Section 3 describes clustering results from three major clustering algorithms: K-means, CLICK, and SOM using two different datasets Leukaemia gene expression data and Sporulation data. A comparative study is presented in Section 4. Conclusions are presented in Section 5

2. Software Tool Overview

There are many clustering algorithms proposed in the last several decades, but little guidance is available to help choose among them. For example, they lack

facilities for estimating the optimal number of clusters, as well as components for evaluating the quality of the clusters obtained. In this section, we present a software tool that offers cluster similarity analysis methods for DNA microarray data analysis.

We present a new tool, *Cluster Diff*, which allows the similarity analysis of clusters generated by different algorithms. This tool may: (1) improve the quality of the data analysis results, (2) support the prediction of the number of relevant clusters in the microarray datasets, and (3) provide cross-reference between different algorithms. The software tool can also be used to analyze cluster similarities from other biomedical data.

2.1. Software Introduction

The software allows working with two datasets each time. The Main Window (panel) (Figure 1.) contains the file, view, and help.

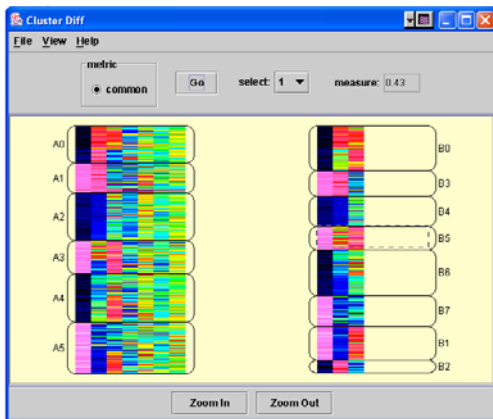


Figure 1. Screenshot of the main window

In Figure 1, the left group (A) has 6 clusters, from A0 to A5; the right group (B) has 8 clusters, from B0 to B7.

In each cluster, the column represents division of the microarray data, and the row represents the gene's profile. For example, in Figure 1, the group A has 7 divisions; the group B has 3 divisions.

The score is the measurement of similarity. The maximum number is 1.00 that means the profiles of these two clusters have similar trends. That is to say the most genes in the two clusters are similar. If the score is 0.00, two clusters are not matched.

The output has multiple visualizations. From button View, you may check different options to get different views.

2.2. Data Source and Data Format

This tool uses the textual tab-delimited data files. The format is similar to the Stanford tab-delimited format (<http://genome-www5.stanford.edu/microarray/help/formats.shtml>) except that you should put tab [cluster] and [/cluster] between a cluster dataset. An example of the described format is shown in Table 1.

Table 1. Input data file format

[cluster]			
YKR007G	-0.16	0.12	-0.1
YER067A	-0.17	0.16	0.18
YBH291C	-0.45	-0.11	-0.58
[/cluster]			
[cluster]			
YPL184C	-0.76	-0.61	-0.36
YTR075W	-0.78	-0.53	0.84
YCR059S	-0.17	0.24	0.15
[/cluster]			

3. Clustering Algorithms and Analysis

Clustering methods, which determine the natural sub-groups in a data set, have some advantages over other methods, because no previous knowledge is necessary for clustering analysis [2, 14]. Several clustering algorithms have been proposed in past few decades [2, 3, 10, 11, 16]. In this section, we briefly describe three such methods, including the K-means clustering methods, the Cluster Identification via Connectivity Kernels (CLICK), and the Self-Organizing Map (SOM) neural networks.

3.1. Clustering Algorithms

3.1.1 K-means Clustering Algorithm

The k-means clustering algorithm [5, 20] is a popular form of cluster analysis. The basic idea is that you start with a collection of items (e.g. genes) and some chosen number of clusters (k) you want to find. The items are initially randomly assigned to a cluster. The k-means clustering proceeds by repeated application of a two-

step process where (1) the mean vector for all items in each cluster is computed; (2) items are reassigned to the cluster whose center is closest to the item. The k-means clustering algorithm is repeated many times, each time starting from a different initial clustering. The sum of distances within the clusters is used to compare different clustering solutions. The clustering solution with the smallest sum of within-cluster distances is saved. The parameters that control k-means clustering are the number of clusters (k) and the number of trials.

The k-means clustering algorithm should be repeated with more trials. If the optimal solution is found many times, the solution that has been found is likely to have the lowest possible within-cluster sum of distances. We can then assume that the k-means clustering procedure has then found the overall optimal clustering solution^[4].

3.1.2 CLICK Clustering Algorithm

CLICK-Cluster Identification via Connectivity Kernels^[23] is a clustering algorithm which is applicable to gene expression analysis as well as to other biological applications. The algorithm utilizes graph-theoretic and statistical techniques to identify tight groups of highly similar elements (kernels), which are likely to belong to the same true cluster. Several heuristic procedures are then used to expand the kernels into the full clustering. The algorithm does not make any prior assumptions on the number or the structure of the clusters. At the heart of the structure of the cluster, there is a process of recursively partitioning a weighted graph into components using minimum cut computations. The edge weights and the stopping criterion of the recursion are assigned probabilistic meaning, which give the algorithm higher accuracy.

3.1.3 Self-Organizing Map (SOM) Neural Network

SOM^[8, 9] is a neural network with a number of nodes or neurons. Usually the configuration of these nodes is rectangular or hexagonal^[15, 21]. The nodes have an associated vector of the same length of the input data. All nodes have initial random values and these reference vectors are adjusted during the training process. After the network is stable, these reference vectors are used to group the genes based on the closeness of the genes to the reference vectors.

During the training stage, the strength of the updating the reference vectors depends on their distances to the winner vector, which is the closest vector to a randomly selected gene. The training length, the training rate, and the size of the updating

neighborhood can be customized. Usually the training is performed in two phases: the first one is the ordering phase (strong training rate and large updating radius) and the last one is the fine-tuning phase (long training length with a weak training rate and a smaller radius).

3.2. Clustering Analysis

The purpose of our study is to compare the clusters generated by above three clustering methods.

3.2.1. Software for Clustering Analysis

The software we use for clustering analysis includes the following:

(1) EXPANDER (EXpression Analyzer and DisplayER)^[24]: This is a java-based tool for analysis of gene expression data. We use it for CLICK clustering.

(2) GEPS (Gene Expression Pattern Analysis Suite)^[7]: It includes following servers:

a) K-means Server: This interface performs a Partitioning Clustering algorithm. The number of clusters k is specified by the user.

b) SOM Server: This is an interface to SOM package. The map is plotted with SomPlot. The resulting clusters can be extracted to continue with the analysis.

3.2.2. The Data Set and Data Preprocessing

Data Source

- (1) Leukaemia dataset* (7129 genes, 38 samples)
- (2) Sporulation dataset** (6116 genes, 7 samples)

We experiment with a subset of the Leukaemia dataset and a subset of Sporulation dataset. Both datasets are obtained using an Affymetrix hybridization array.

Data Preprocessing

We randomly select 500 genes from each dataset and save them as in plain text files, respectively. Then we formatted them as EXPANDER and GEPS required.

These two preprocessed data sets are used for comparing the algorithms.

* The original data and experimental methods are available at <http://www.genome.wi.mit.edu/MPR>

**The original data and experimental methods are available at <http://genome-www.stanford.edu>

3.3. Clustering Results Comparison

3.3.1 Clustering with Leukaemia dataset

Dataset: 500-gene Leukaemia

Test condition

Test condition for CLICK Algorithm

Default homogeneity

Test condition for K-means Algorithm

(1) K value: 4

(2) Distance function: Pearson correlation coefficient

Test condition for Self Organizing Map (SOM):

(1) 2 * 2 hexagonal lattices (This will result in 4 clusters)

(2) Number of trials: 20

Clustering results

The clustering results of CLICK, SOM, and K-means are shown in Figure 2, each of which includes the profile of a cluster and the profiles of the genes in that cluster.

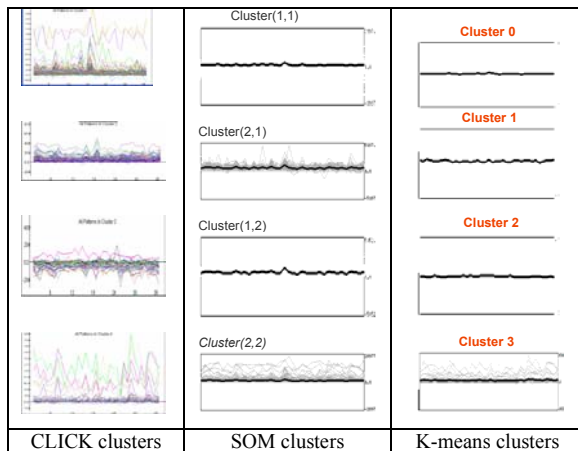


Figure 2. Clustering results

3.3.2 Clustering with Sporulation dataset

Dataset: 500-gene Sporulation

Test condition

Test condition for CLICK Algorithm

Default homogeneity

Test condition for K-means Algorithm

(1) K value: 6

(2) Distance function: Pearson correlation coefficient

Test condition for Self Organizing Map (SOM):

(1) 2 * 3 hexagonal lattices Number of trials: 20

Clustering results

The clustering results of CLICK, SOM, and K-means are shown in Figure 3, each of which includes the profile of a cluster and the profiles of the genes in that cluster.

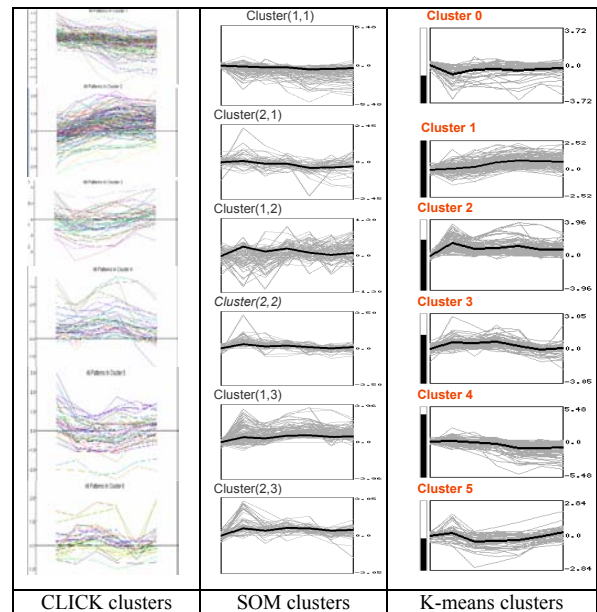


Figure 3. Clustering results

4. Comparison of the Clustering Methods

The Self-Organizing Map (SOM) is a popular unsupervised neural network algorithm. It is very efficient in handling large datasets such as gene expressive data. The SOM algorithm is also robust even when the data set is noisy [25]. So we chose it as the target clustering algorithm for this study.

4.1 Comparison with Leukaemia dataset

4.1.1 CLICK vs. SOM

Both clustered data files from CLICK and SOM with 500-genes Leukaemia, after formatting as Section 2.2, were loaded to the *cluster diff* for the cluster similarity analysis. The result is shown in Figure 4.

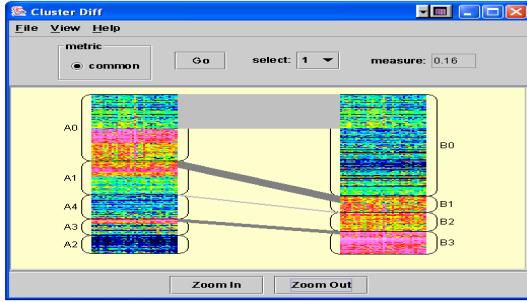


Figure 4. CLICK vs. SOM (Leukaemia)

For detail cluster similarity analysis, we input a pair of clusters each time, one by CLICK and one by SOM, The scores are summarized in Table 2.

Table 2. Cluster similarity analysis(Leukaemia) results (CLICK vs. SOM)

	SOM11	SOM12	SOM21	SOM22
CLICK0	0.26	0.09	0.14	0.20
CLICK1	0.15	0.15	0.11	0.08
CLICK2	0.18	0.01	0.00	0.00
CLICK3	0.09	0.03	0.07	0.09
CLICK4	0.20	0.03	0.03	0.01

4.1.2 K-means vs. SOM

Both clustered data files from K-means and SOM with 500-genes Leukaemia, after formatting as Section 2.2, were loaded to the *cluster diff* for the cluster similarity analysis. The result is shown in Figure 5.

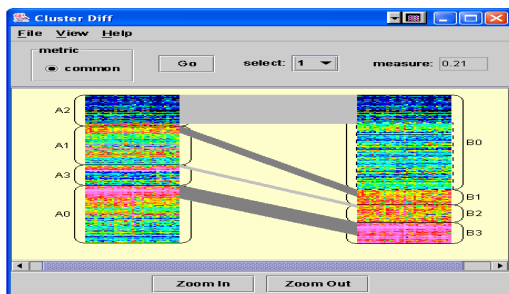


Figure 5. K-means vs. SOM (Leukaemia)

For detail cluster similarity analysis, we input a pair of clusters each time, one by K-means and one by SOM. The scores are summarized in Table 3.

Table 3. Cluster similarity analysis(Leukaemia) results (K-means vs. SOM)

	SOM11	SOM12	SOM21	SOM22
Kmeans0	0.24	0.09	0.13	0.20
Kmeans1	0.20	0.15	0.11	0.06
Kmeans2	0.30	0.02	0.01	0.01
Kmeans3	0.13	0.03	0.07	0.11

4.2 Comparison with Sporulation dataset

4.2.1 CLICK vs. SOM

Both clustered data files from CLICK and SOM with 500-genes Sporulation, after formatting as Section 2.2, were loaded to the *cluster diff* for the cluster similarity analysis. The result is shown in Figure 6.

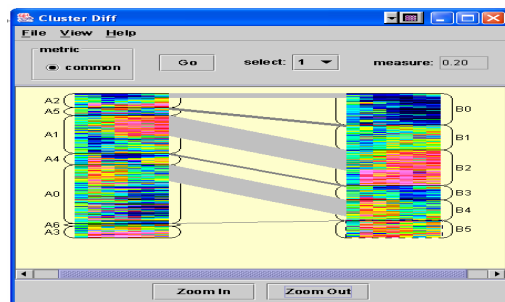


Figure 6. CLICK vs. SOM (Sporulation)

For detail cluster similarity analysis, we input a pair of clusters each time, one by CLICK and one by SOM, The scores are summarized in Table 4.

Table 4. Cluster similarity analysis(Sporulation) results (CLICK vs. SOM)

	SOM12	SOM13	SOM21	SOM22	SOM23	SOM11
CLICK0	0.03	0.00	0.02	0.01	0.03	0.03
CLICK1	0.07	0.04	0.11	0.26	0.14	0.22
CLICK2	0.17	0.39	0.03	0.00	0.02	0.08
CLICK3	0.10	0.05	0.09	0.01	0.05	0.10
CLICK4	0.08	0.15	0.01	0.04	0.06	0.00
CLICK5	0.02	0.04	0.07	0.06	0.07	0.09
CLICK6	0.07	0.02	0.04	0.04	0.05	0.05

4.2.2 K-means vs. SOM

Both clustered data files from K-means and SOM with 500-genes Sporulation, after formatting as Section 2.2, were loaded to the *cluster diff* for the cluster similarity analysis. The result is shown in Figure x.

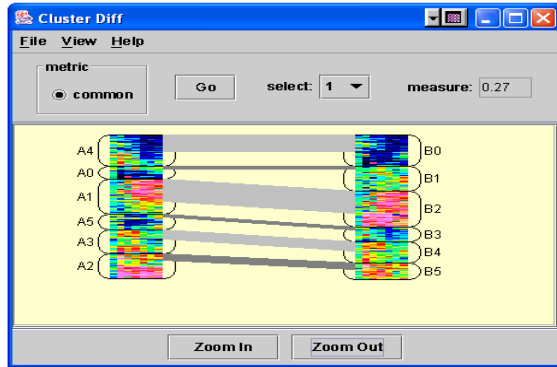


Figure 7. K-means vs. SOM (Sporulation)

For detail cluster similarity analysis, we input a pair of clusters each time, one by CLICK and one by SOM, The scores are summarized in Table 5.

Table 5. Cluster similarity analysis (Sporulation) results (K-means vs. SOM)

	SOM12	SOM13	SOM21	SOM22	SOM23	SOM11
Kmeans0	0.10	0.00	0.06	0.01	0.00	0.20
Kmeans1	0.17	0.46	0.02	0.00	0.04	0.02
Kmeans2	0.05	0.25	0.00	0.09	0.19	0.00
Kmeans3	0.10	0.01	0.05	0.28	0.23	0.01
Kmeans4	0.05	0.00	0.18	0.11	0.00	0.20
Kmeans5	0.09	0.01	0.13	0.08	0.02	0.14

4.3 Comparison Results:

From the tables in Section 4.1 and 4.2, we can find that most SOM clusters match the K-means clusters (or CLICK) well and vice versa. An example of good match is Kmeans1 with SOM13 (0.46), and CLICK2 with SOM13(0.39) in Sporulation dataset (see Figure 8.). The profiles of these three clusters have similar trends, meaning that the most genes in these three clusters are similar.

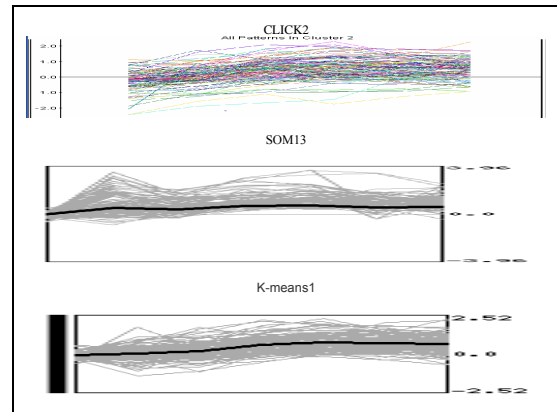


Figure 8. Example of a good match

The average scores for both CLICK and K-means algorithms are summarized in Table 6.

Table 6. Clustering Method Comparison analysis results (CLICK vs.K-means)

Dataset	CLICK vs. SOM	K-means vs. SOM
Leukaemia	0.16	0.21
Sporulation	0.20	0.27

The case study results indicate that the clusters generated by the CLICK, K-means or SOM algorithms are comparable. Most clusters match the SOM clusters well and vice versa. Given a target clustering algorithm (or clusters), the tool can efficiently determining the closest matching from a set of clustering algorithms.

5. Conclusions

In this paper, we present a new data mining tool, *Cluster Diff*, which allows the similarity analysis of clusters generated by different algorithms. This tool may: (1) improve the quality of the data analysis results, (2) support the prediction of the number of relevant clusters in the microarray datasets, and (3) provide cross-reference between different algorithms. The software tool can also be used to analyze cluster similarities from other biomedical data. This software tool may significantly support gene expression data analyses.

6. Acknowledgements

We would like to thank Dr. Dong-Guk Shin and Dr. Jae-guon Nam at the Univ. of Connecticut for providing the software for cluster similarity analysis in this work.

7. Reference

- [1]. Robin L. Stears, et al., Trends in Microarray analysis. Nature medicine, Volume 9, 140-145, January 2003
- [2]. Marco F. Ramoni, et al., Cluster analysis of gene expression dynamics. PNAS, Vol.99, July 2002
- [3]. M.B.Eisen, P.T.Spellman, P.O.Brown, D.Botstein. Cluster analysis and display of genome-wide expression patterns Proc. Natl. Acad. Sci., 95:14863-14868, 1998
- [4]. <http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/manual/KMeans.html>
- [5]. J. B. MacQueen (1967): "Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability", Berkeley, University of California Press, 1:281-297
- [6]. A. Brazma, J. Vilo. Gene expression data analysis, FEBS Letters, 480, 17-24, 2000
- [7]. Vaquerizas, J.M., Conde, L., Yankilevich, P., Cabezon, A., Minguez, P., Diaz-Uriarte, R., Al-Shahrour, F., Herrero, J & Dopazo, J. Gepas an experiment-oriented pipeline for the analysis of microarray gene expression data. Nucleic Acids Research 33 (Web Server issue):W616-W620, 2005
- [8]. Teuvo Kohonen, The self-organizing map. Neurocomputing 21: 1-6. 1998
- [9] Kohonen, T. Self-Organizing Maps, Springer, Berlin. 1995
- [10]. Tamayo, P. Dmitrovsky, E., et al. Interpreting patterns of gene expression with self-organizing maps: Methods and applications to hematopoietic differentiation, Proc. Nat. Acad. Sci 96, 2907-2912, 1999
- [11] Herrero, J. & Dopazo, J. Combining hierarchical clustering and self-organizing maps for exploratory analysis of gene expression patterns. Journal of Proteome Research, 1(5):467-470. 2002.
- [12]. Botstein, D. & Brown, P., Exploring the new world of the genome with DNA microarrays, Nature Genetics (supp.) 21, 33-7. 1999
- [13]. Jain, A.K. and Dubes, R.C. Algorithms for Clustering Data. Prentice Hall, Englewood Cliffs, NJ. 1988.
- [14]. Everitt, B., Cluster analysis, Halstead, New York. 1980
- [15]. Kohonen, T., Self-Organization and Associative Memory (3rd edition), springer-Verlag, Berlin. 1989.
- [16]. Dopazo, J. Microarray Data Processing and Analysis. Microarray data analysis II. Kluwer Academic. Publ. Pp. 43-63. 2002
- [17]. Dopazo, J., Zanders, E., Dragoni, I., Amphlett, G. and Falciani, F. Methods and approaches in the analysis of gene expression data J. Immunol. Meth. 250, 93-112. 2001
- [18]. <http://titan.biotech.uiuc.edu/cs491jh/slides/cs491jh-QJ.ppt>
- [19]. K. Alsabti, S. Ranka, and V. Singh, An Efficient k-means Clustering Algorithm, Proc. First Workshop High Performance Data Mining, Mar. 1998.
- [20]. http://www.clustan.com/k-means_critique.html , March, 2001
- [21]. T. Kanungo, D.M. Mount, N.S. Netanyahu, C. Piatko, R. Silverman, and A.Y. Wu, The Analysis of a Simple k-means Clustering Algorithm, Proc. 16th Ann. ACM Symp. Computational Geometry, pp. 100-109, June 2000.
- [22]. K.Y. Yeung, W.L.Ruzzo, et al. Validating clustering for gene expression data. Bioinformatics Vol 17, p 309-318, 2001
- [23]. R. Sharan and R. Shamir. CLICK: a Clustering Algorithm with Applications to Gene Expression Analysis. In Proc. 8th International Conference on Intelligent Systems for Molecular Biology (ISMB'00), pages 307-316, AAAI Press, Menlo Park, CA, 2000
- [24]. <http://www.cs.tau.ac.il/~rshamir/expander/ver2Help.html#Biclustering>
- [25]. Paul Mangiameli, Shaw K. Chen and David West, "A Comparison of SOM neural network and hierarchical clustering methods," European Journal of Operational Research, Vol. 93, Issue 2, 6, pp. 402-417, 1996