Jackknife distances for clustering time–course gene expression data

Theresa Scharl[1], Friedrich Leisch[2]
Department of Statistics and Probability Theory, Vienna University of Technology,
Wiedner Hauptstraße 8-10/1071, A-1040 Wien, Austria; Theresa.Scharl@ci.tuwien.ac.at[1]
Department of Statistics, University of Munich,
Ludwigstraße 33, D-80539 München, Germany; Friedrich.Leisch@stat.uni-muenchen.de[2]

Abstract

Clustering time–course gene expression data is a common tool to find co–regulated genes and groups of genes with similar temporal or spatial expression patterns. The distance measure used for clustering has major impact on the properties of the resulting clusters. As technical problems can easily distort the microarray data there is a need for distance measures which are able to deal with outliers. Here we present new so-called "Jackknife" distance measures which can handle outlier time points. In a simulation study on a publicly available dataset from yeast the utility of such distance measures is investigated.

Keywords: Cluster Analysis, Time–course Microarray Data, Distance Measures, R.

1 Introduction

The interpretation of enormous amounts of data from microarrays has been a challenging task in statistics and bioinformatics for the past few years. One possible approach to deal with the complexity of the data is cluster analysis which has been widely applied for example for grouping tissue samples in cancer research (e.g. Pollard and van der Laan, 2005; Thomas et al., 2001) or for clustering time–course gene expression data. Time–course microarray experiments make it possible to look at the gene expression of thousands of genes at several time points simultaneously. Genes with similar expression pattern are co–expressed genes which are likely to be co–regulated. Hence clustering gene expression patterns may help to find groups of co–regulated genes or to identify common temporal or spatial expression patterns. Finally cluster results can suggest functional pathways and interactions between genes (Eisen et al., 1998; Tavazoie et al., 1999; Ben–Dor et al., 1999).

The distance measure used has major impact on the resulting clusters (Gentleman et al., 2005). The properties of different distance measures have to be investigated to be able to answer biological questions more precisely. A comparison of different distance measures which are commonly used in the context of clustering time–course microarray data was done in Scharl and Leisch (2006). In this paper we want to investigate new distance measures for clustering time-course gene expression data which are robust against outlier variables. There are several algorithms which are able to deal with outlier observations. Partitioning around medoids described in Kaufman and Rousseeuw (1990) is a more robust version of k–means for arbitrary distance measures. Trimmed K-means (Cuesta-Albertos et al., 1997) is a robust version of the original algorithm. All these algorithms can handle outliers in the data points. Our goal is to identify outliers in the variables. We want to be robust against outliers in the time points as technical problems like dust or a scratch on the slide can easily distort the microarray data. The Jackknife correlation which can handle one outlier time point was introduced by Heyer et al. (1999). Here we want to extend this promising approach to further distance measures and investigate the properties of Jackknife correlation as well as Jackknife versions of Euclidean, Manhattan and Maximum distance.

The different distance measures are compared in a simulation study using two cluster algorithms, stochastic QT-Clust (Scharl and Leisch, 2006) and the well–known k–means algorithm. For that purpose two evaluation criteria are chosen. To investigate the stability of a method and the agreement between partitions pairwise comparisons of cluster results are computed using the adjusted Rand Index (Hubert and Arabie, 1985). As a measure of the quality of a partition the sum of within cluster distances is observed. All algorithms used are implemented in R (**http://www.r-project.org**, R Development Core Team, 2006) package `flexclust` (Leisch, 2006) available from CRAN (**http://cran.r-project.org**). `flexclust` is a flexible toolbox for clustering which allows to try out various distance measures with only minimal programming effort.

In this simulation study a publicly available dataset from yeast was used, the seventeen time point mitotic cell cycle data (Cho et al., 1998) available at **http://genome-www.stanford.edu**. This dataset was preprocessed adapting the instructions given by Heyer et al. (1999). After rescaling the data genes that were expressed at very low levels and did not vary significantly over the time points were removed. This procedure yields gene expression data on $G = 2090$ genes (observations) for $T = 17$ time points (variables). As time point 10 was reported to be an outlier variable the simulations were conducted on the 17 time point dataset as well as on a dataset with time point 10 removed to investigate the functionality of the Jackknife distance measures.

2 Methods

2.1 Distance Measures

Four distance measures were chosen which are commonly used in the context of clustering time–course microarray data (see for example Chipman et al., 2003; Jiang et al., 2004; Sheng et al., 2005; Gentleman et al., 2005). Here three geometric distances and "1–correlation" distance are used. One of the most commonly used methods to measure the distance between two data objects is *Euclidean distance* which is given by

$$d_{xy} = \sqrt{\sum_{i=1}^{T}(x_i - y_i)^2},$$

where x and y are T–dimensional vectors and T is the number of time points in the experiment. *Manhattan distance*

$$d_{xy} = \sum_{i=1}^{T}|x_i - y_i|$$

is more robust to outlies than Euclidean distance. Both Euclidean and Manhattan distance yield clusters with a certain band width which can vary from one time point to the next. *Maximum distance*

$$d_{xy} = \max|x_i - y_i|$$

looks at the maximum differences between time points and yields clusters of a fixed band width.

If one is interested in the relative changes of gene expression a *correlation*–based distance measure is more appropriate as correlation is invariant to location and scale. The dissimilarity between two gene profiles can be defined as

$$d_{xy} = 1 - \rho_{xy} = 1 - \frac{\sum_{i=1}^{T}(x_i - \bar{x})(y_i - \bar{y})}{[\sum_{i=1}^{T}(x_i - \bar{x})^2]^{1/2}[\sum_{i=1}^{T}(y_i - \bar{y})^2]^{1/2}}$$

where ρ_{xy} is the Pearson sample correlation coefficient. This distance measure removes changes in the average or range of the expression level from one gene to the next. Both strongly positively correlated as well as negatively correlated genes are considered co–expressed.

A possible problem using these distance measures for clustering time–course gene expression data is that single outlier variables can completely change the expression pattern of certain genes. Outliers at special time points are very common in microarray experiments as technical problems like dust or a scratch on the slide can easily distort the data. In such a case these outlier variables can lead to unwanted correlations between genes and to incorrect assignment to clusters. There is a need for distance measures which are robust against outlier variables. The idea of Jackknife (Efron, 1982) distance measures is not to exclude the whole observation for such a gene but rather one or several variables. We want to introduce so–called "Jackknife" distance measures which can handle one outlier time point. The *Jackknife correlation* was first used by Heyer et al. (1999) to cluster gene expression data. It is defined as

$$d_{xy} = 1 - \min(\rho_{xy}^{(1)}, \rho_{xy}^{(2)}, \ldots, \rho_{xy}^{(T)})$$

where $\rho_{xy}^{(t)}$ is the correlation of pair x,y computed with the tth time point deleted.

Now we want to extend this concept and introduce robust versions of the three geometric distance measures Euclidean, Manhattan and Maximum distance. Jackknife Euclidean distance is defined as

$$d_{xy} = \min(d_{xy}^{(1)}, d_{xy}^{(2)}, \ldots, d_{xy}^{(T)})$$

where $d_{xy}^{(t)}$ is the Euclidean distance of pair x,y computed with the tth time point deleted. Jackknife Manhattan distance and Jackknife Maximum distance can be defined in the same way.

Some example clusters can be seen in Figures 1, 2 and 3. The clusters of the genes YDL223C, YDR044W and YDR006C are observed for Maximum distance, "$1 - correlation$" distance and Euclidean distance and their Jackknife versions using two different cluster algorithms. Figure 1 shows gene expression profiles with a clear peak at time point 10. Figure 2 shows genes with high activity at time point 3 and Figure 3 contains genes with high gene expression at time point 3 as well as time point 10. As Figure 3 shows that allowing for outlier variables genes can fall into different clusters. Using Jackknife distance measures the clusters of gene YDR006C have a stronger peak at time point 3 and a new peak at time point 11 instead of time point 10.

2.2 Cluster Algorithms

In the literature numerous cluster algorithms for clustering gene expression profiles are proposed (e.g. De Smet et al., 2002; Luan and Li, 2003; Bickel, 2003; Erdal et al., 2004; Serban and Wasserman, 2005; Zeng and García–Frías, 2006). In this simulation study our focus is on the influence of the distance measure used for two different cluster algorithms. Both are implemented in R package `flexclust`, a flexible toolbox for clustering which makes it easy for data analysts to try out a variety of distance measures. New distance measures can easily be implemented as new distance functions. One cluster algorithm is the well–known k-means algorithm where the number of clusters can be chosen arbitrarily. The second one is stochastic QT–Clust (Leisch, 2006; Scharl and Leisch, 2006) which is the stochastic approximation of a quality-based cluster algorithm specially designed to identify co–expressed genes (Heyer et al., 1999). Here the number of clusters is controlled by the quality of a cluster as it is hard to investigate the true number of clusters for gene expression data (Bryan, 2003).

For stochastic QT–Clust one has to define the quality of clusters which is given by the maximum diameter of the clusters and the minimal number of points that form a single cluster. This means that the number of clusters

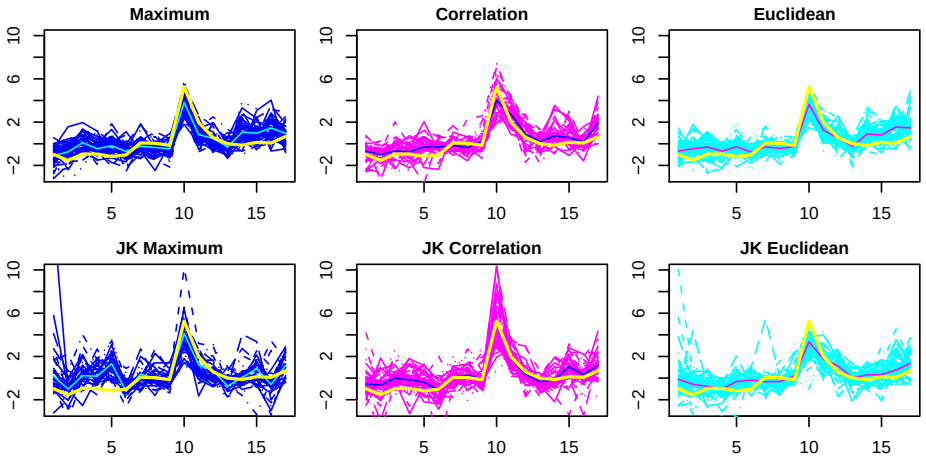

Figure 1: The clusters of gene YDL223C (yellow) for Maximum, "$1 - correlation$" and Euclidean distance and their Jackknife versions using k–means algorithm. Time points are shown on the x–axis and gene expression is shown on the y–axis.

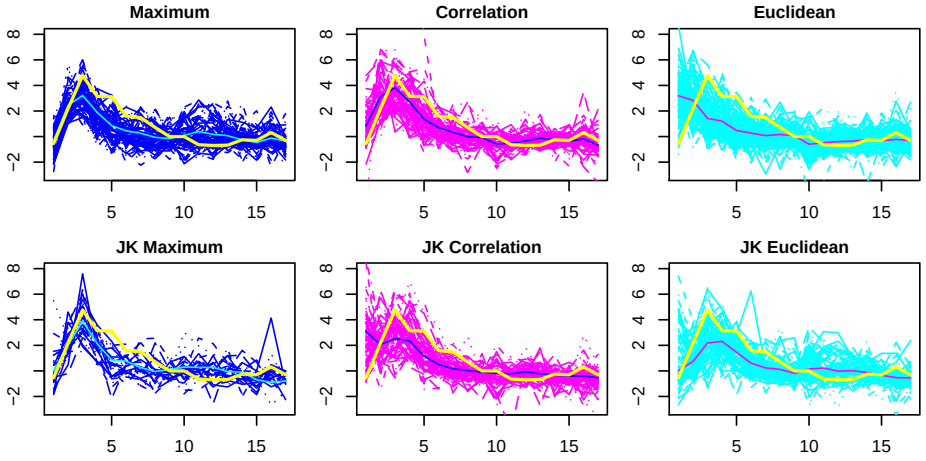

Figure 2: The clusters of gene YDR044W (yellow) for Maximum, "$1 - correlation$" and Euclidean distance and their Jackknife versions using QT–CLUST algorithm.

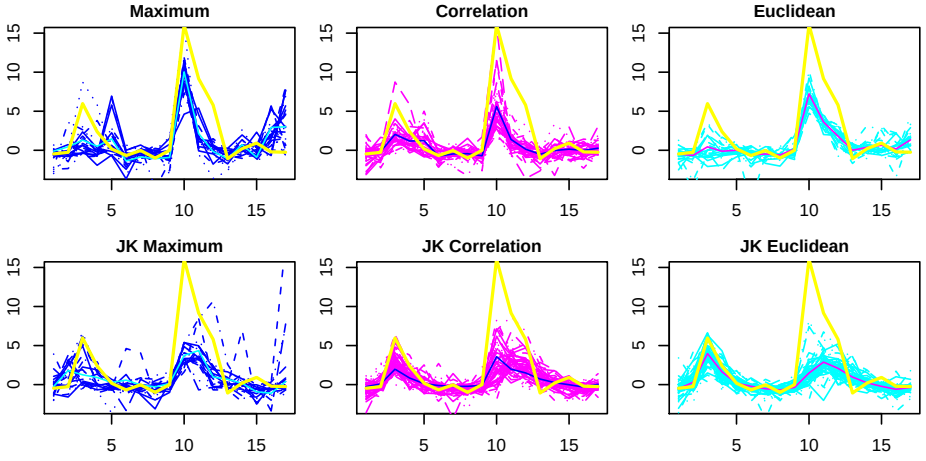

Figure 3: The clusters of gene YDR006C (yellow) for Maximum, "$1 - correlation$" and Euclidean distance and their Jackknife versions using k–means algorithm.

can only be controlled indirectly through these two parameters. Additionally the number `ntry` of times steps 1 to 4 should be run in each step has to be defined.

1. Start with a randomly chosen center.
2. Iteratively add the gene that minimizes the increase in cluster diameter.
3. Continue until no gene can be added without surpassing the diameter threshold.
4. Repeat from 1. for `ntry` $-$ 1 further centers.
5. Select the largest candidate cluster and remove the genes it contains from further consideration.
6. Goto 1. on the smaller dataset.
7. Stop when the largest remaining cluster has fewer than some prespecified number of elements.

Stochastic QT–Clust is robust as outlier observations will not be added to any cluster as opposed to k-means where every gene will be part of a cluster. If `ntry` is equal to the number of genes G the original QT–Clust algorithm (Heyer et al., 1999) is obtained. Stochastic QT–Clust speeds up the procedure and we get different local maxima of the objective function. The original algorithm will always converge in the same local optimum. As shown in Scharl and Leisch (2006) smaller values of `ntry` yield smaller sum of within cluster distances for all distance measures used. Therefore `ntry` $= 5$ is used for all simulations.

3 Simulation

3.1 Simulation Design

The goal of this simulation study is to compare the four classical distance measures to the Jackknife distance measures using both k–means and stochastic QT–Clust. For k–means the following procedure is used

1. draw 100 bootstrap samples from the original data,
2. cluster them into 50 clusters using each of the distance measures, and
3. compare the obtained results using the sum of within cluster distances and the adjusted Rand index.

The sum of within cluster distances is computed as a measure of the quality of a partition. It is given by

$$W = \sum_{j=1}^{k} \sum_{x \in X_j} d(x, c_j)$$

where x is an element of cluster X_j, d is the distance measure and c_j is the center of cluster j. The adjusted Rand index (Hubert and Arabie, 1985) is computed to be able to make pairwise comparisons between partitions. A Rand index of 1 corresponds to identical partitions and a Rand index of 0 corresponds to agreement by chance given cluster size. The Rand index is used as a measure of stability and reproducibility of the resulting clusters and the agreement between partitions.

The number of clusters for k–means is chosen arbitrarily. As biologists prefer to work with smaller groups of genes to be able to take a closer look at the resulting clusters the genes were grouped into 50 clusters for k–means. For QT–Clust 100 replicates of the algorithm were computed for each distance measure on the original data because the algorithm has no prediction step. It was tried to find an appropriate radius for QT–Clust to get a similar numbers of clusters like in k–means.The number of clusters of QT–Clust is changing with the diameter of the clusters and the minimal number of points that for a single cluster. Therefore the number of clusters of QT–Clust is varying between distance measures and even between replicates of QT–Clust using the same distance measure.

3.2 Stability of the Resulting Clusters

As researchers want to know how reliable the resulting clusters are the stability of the cluster results is investigated. For that purpose all consecutive pairwise comparisons of cluster results are computed using the adjusted Rand index. 100 replicates on the original data for QT–Clust and 100 bootstrap samples of k–means are used. Boxplots of the Rand index of all consecutive pairwise comparisons are shown in Figure 4.

It can be seen that the reproducibility of cluster results is in general very high for k–means. The Rand index for consecutive comparisons within distance measures is over 0.95 for all distance measures. For QT–Clust the stability of cluster results is smaller that for k–means. Here the stability is highest for Jackknife "$1 - correlation$" distance followed by Jackknife Maximum distance. The reproducibility of cluster results using Manhattan and Euclidean distance is much lower on this dataset. For QT–Clust it can be clearly seen that for all distance measures the stability of the Jackknife version of the distance measure is higher than the stability of the original distance measure.

3.3 Quality of the Partitions

As a measure of the quality of a partition the sum of within cluster distances is used. It was computed for each distance measure for all of the 100 replicates on the original data using QT–Clust and 100 bootstrap samples of k–means. Figure 5 shows boxplots of the sum of within cluster distances for all distance measures.

The sum of within cluster distances is not directly comparable between distance measures. But Figure 5 shows that using the Jackknife version of any of the four distance measures leads to a smaller sum of within cluster distances for both QT–Clust and k–means. This

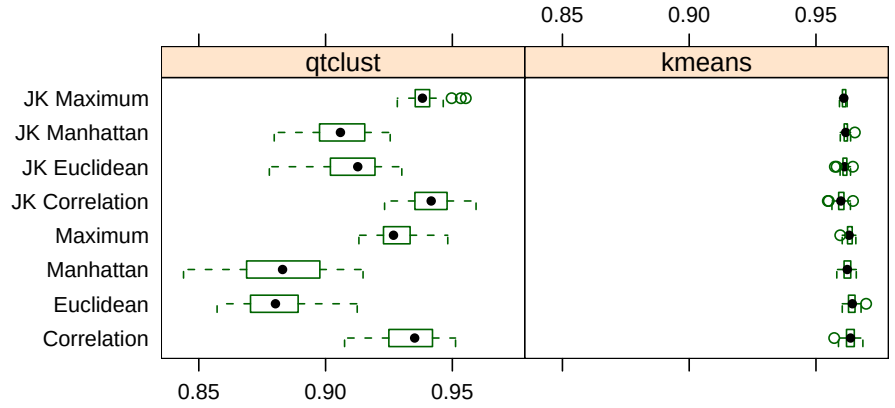

Figure 4: Boxplots of the Rand indices for all pairwise comparisons of 100 replicates of QT–Clust (left panel) and 100 bootstrap samples of k–means (right panel) within different distance measures.

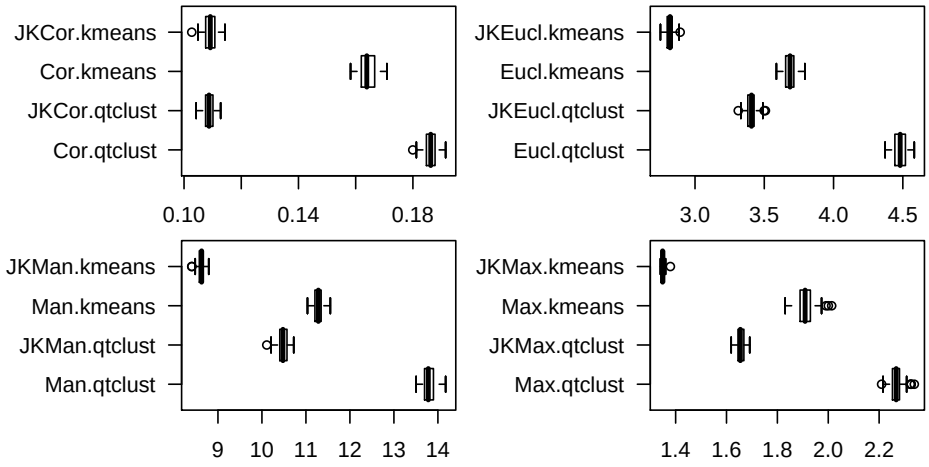

Figure 5: Boxplots of the sum of within cluster distances for 100 replicates of QT–Clust and 100 bootstrap samples of k–means for the different distance measures.

means that the similarity between genes within a cluster is higher using Jackknife distance measures. Additionally the sum of within cluster distances is smaller for k–means than for QT–Clust for all distance measures.

3.4 Comparing Different Distance Measures

Now we want to investigate the differences between cluster results of different distance measures. The agreement between partitions is computed using the Rand index. The results are shown in Figure 6.

For QT–Clust the agreement between "$1-correlation$" distance and Maximum distance and their Jackknife versions is very high as well as the agreement between these Jackknife versions. The results obtained using Manhattan and Euclidean distance disagree most.

For k-means the impact of the distance measure used is much smaller on this dataset. The agreement between cluster results using different distance measures is very high for this cluster algorithm. Here the partitions agree most for the Jackknife versions of Maximum distance and "$1 - correlation$" distance and disagree most for Maximum distance and "$1 - correlation$" distance.

3.5 Excluding one Outlier Time Point

As time point 10 was reported not to be dependable (Heyer et al., 1999) the simulations were repeated on a smaller dataset with this time point deleted. Again we would like to compare the partitions using the Rand index. First the cluster results obtained on each dataset are compared for all distance measures. The results are shown in Figure 7. For k–means the agreement between results on the full dataset and on the dataset with one time point removed is again very high (over 0.95). For QT–Clust the pattern of agreement is very similar to Figure 4. This means that the cluster results agree between the two datasets and excluding the outlier time point does not change the results significantly.

Finally we want to find out how good the Jackknife distance measures work and how frequently the outlier time point 16 is detected and excluded from calculating the distance between two gene expression profiles. For that purpose the original dataset using Jackknife distance measures is compared to the 16 time point dataset using the original distance measures (see Figure 8). It was found that on this dataset the agreement between cluster results for Jackknife distance measures on the original dataset and original distance measures on the dataset with one outlier time point deleted is very high for Maximum distance and "$1 - correlation$" distance using QT–Clust. Therefore Jackknife distance measures are applicable when outliers at special time points distort the gene expression data.

As the example clusters in Figure 2 show time point 10 is not the most extreme observation for all genes. For those genes other observations than time point 10 will be detected by the Jackknife distance measures.

4 Summary and Outlook

The distance measure used for clustering time–course gene expression data has major impact on the resulting clusters. In this area four commonly used distance measures are Euclidean distance, Manhattan distance, Maximum distance and "$1 - correlation$" distance. In microarray experiments outliers can easily distort the data and there is need for distance measures that are robust against outlier variables. For that purpose the concept of the so–called "Jackknife" correlation (Heyer et al., 1999) was extended for Euclidean, Manhattan and Maximum distance. In a simulation study on the yeast cell–cycle dataset (Cho et al., 1998) the properties of these distance measures were investigated using two cluster algorithm, the classical k–means and stochastic QT–Clust. For k–means clusters on 100 bootstrap samples were computed and for stochastic QT–Clust 100 replicates were generated using all eight distance measures. In order to test the functionality of the Jackknife distance measures the simulations were run on the original dataset and on a smaller dataset with the outlier time point deleted.

Cluster results were compared using the Rand index and the sum of within cluster distances. It was found that the sum of within cluster distances is smaller for the Jackknife versions of the distance measures. This might be an indicator that Jackknife distance measures lead to tighter clusters with more similar gene expression profiles. Additionally it could be shown that the partitions on the original dataset using Jackknife distance measures agree very good with the partitions on the smaller dataset using the classical distance measures.

As expected the different methods used have major impact on the resulting clusters on this dataset. The method of choice depends on the data and on the biological questions asked. There is no single distance measure that is always relevant. In any case we found that the influence of the distance measure is much higher for QT–Clust than for k–means. There are still open questions for the future like how the results change when the dataset changes slightly. Additionally it would be interesting to allow more than one outlier time point but this is computationally very intensive. Further simulations on different datasets need to be conducted. So far we already did the same simulations on a dataset from E.Coli and found similar results.

Acknowledgements

This work was supported by the Austrian K_{ind}/K_{net} Center of Biopharmaceutical Technology (ACBT) and the Austrian Science Foundation (FWF) under grant P17382.

References

Ben–Dor A., Shamir R., and Yakhini Z. (1999), "Clustering gene expression patterns", *Journal of Computational Biology*, **6(3–4)**, 281–297.

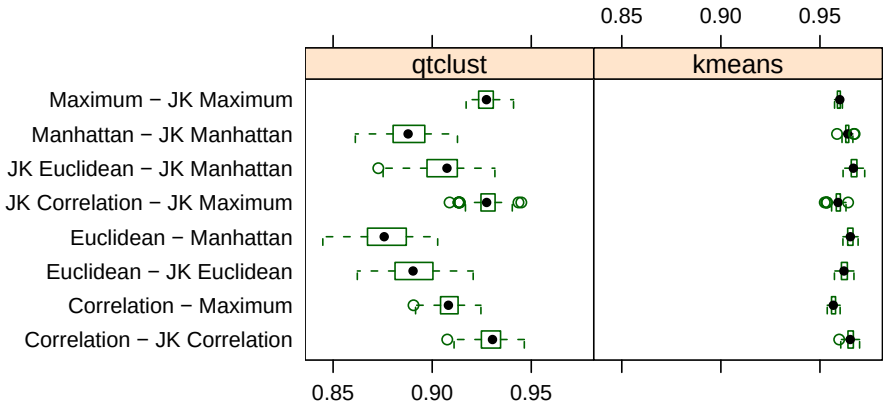

Figure 6: Rand index for pairwise comparison of 100 replicates of QT–Clust (left panel) and 100 bootstrap samples of k–means (right panel) between different distance measures.

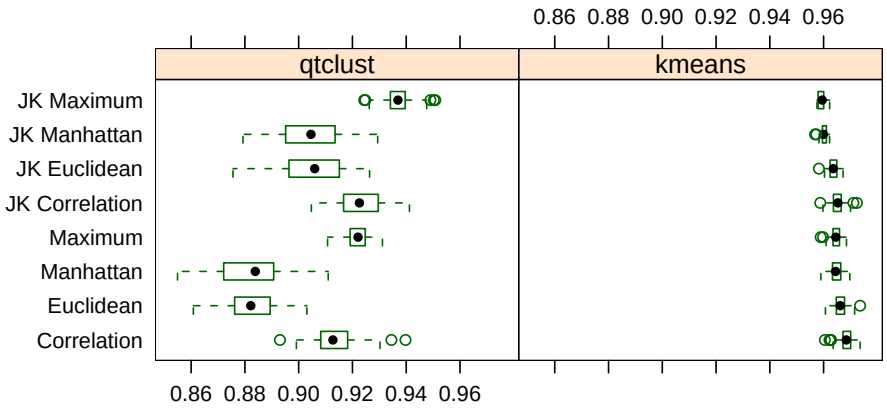

Figure 7: Rand index for pairwise comparison between the full dataset and the 16 time point dataset for QT–Clust (left panel) and k–means (right panel) using different distance measures.

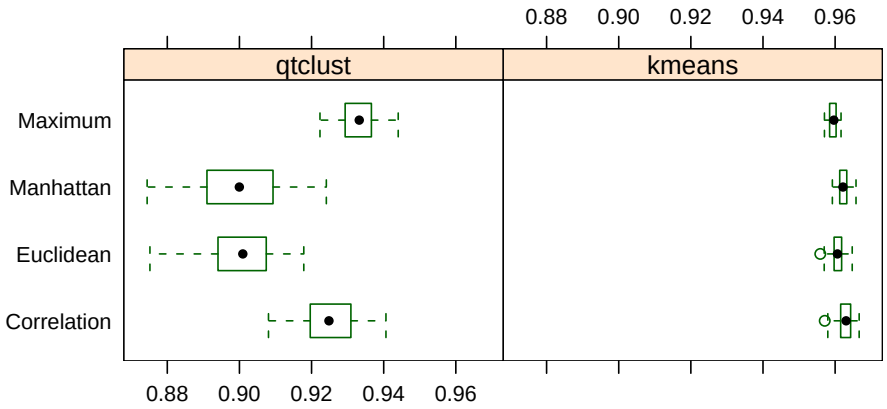

Figure 8: Rand index for pairwise comparison between the full dataset using Jackknife distance measures and the 16 time point dataset using the original distance measures for QT–Clust (left panel) and k–means (right panel).

Bickel D.R. (2003), "Robust cluster analysis of microarray gene expression data with the number of clusters determined biologically", *Bioinformatics*, **19(7)**, 818–824.

Bryan J. (2004), "Problems in gene clustering based on gene expression data", *Journal of Multivariate Analysis*, **90(1)**, 44–66.

Chipman, H., Hastie T.J., and Tibshirani R. (2003), "Clustering microarray data." In Terence P. Speed, editor, *Statistical Analysis of Gene Expression Microarray Data*, chapter 4, pages 159–200, Chapman & Hall/CRC Press.

Cho R.J., Campbell M.J. , Winzeler E.A., Steinmetz L., Conway A., Wodicka L., Wolfsberg T.G., Gabrielian A.E., Landsman D., Lockhart D.J., and Davis R.W. (1998), "A genome-wide transcriptional analysis of the mitotic cell cycle", *Molecular Cell*, **2(1)**, 65–73.

Cuesta-Albertos J.A., Gordaliza A., and Matran C. (1997), "Trimmed k-Means: An Attempt to Robustify Quantizers", *The Annals of Statistics*, **25**, 553–576.

De Smet F., Mathys J., Marchal K., Thijs G., De Moor B., and Moreau Y. (2002), "Adaptive quality-based clustering of gene expression profiles", *Bioinformatics*, **18(5)**, 735–746.

Efron B. (1982), *The Jackknife, the Bootstrap, and Other Resampling Plans*, CBMS–NSF Regional Conference Series in Applied Mathematics, Philadelphia: Society for Industrial and Applied Mathematics (SIAM).

Eisen M.B., Spellman P.T., Brown P.O., and Botstein B. (1998), "Cluster analysis and display of genome–wide expression patterns", *Proc. Natl. Acad. Sci. USA*, **95**, 14863–14868.

Erdal S., Ozturk O., Armbruster D., Ferhatosmanoglu H., and Ray W.C. (2004), "A Time Series Analysis of Microarray Data", *Proceedings of the Fourth IEEE Symposium on Bioinformatics and Bioengineering (BIBE 04)*, Los Alamitos, CA, USA.

Gentleman R., Ding B., Dudoit S., and Ibrahim J. (2005), "Distance Measures in DNA Microarray Data Analysis" In Robert Gentleman, Vincent J. Carey, Wolfgang Huber, Rafael A. Irizarry, Sandrine Dudoit, editors, *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, chapter 12, pages 189–208, Springer.

Heyer L.J., Kruglyak S., and Yooseph S. (1999), "Exploring expression data: Identification and analysis of coexpressed genes", *Genome Research*, **9**, 1106–1115.

Hubert L. and Arabie P. (1985),. "Comparing partitions", *Journal of Classification*, **2**, 193–218.

Jiang D., Tang C., and Zhang A. (2004), "Cluster Analysis for Gene Expression Data: A Survey", *IEEE Transactions on Knowledge and Data Engineering*, **16(11)**, 1370–1386.

Kaufman, L. and Rousseeuw P.J. (1990), *Finding Groups in Data*, John Wiley and Sons, Inc., New York, USA.

Leisch F. (2006), "A toolbox for k-centroids cluster analysis", *Computational Statistics and Data Analysis*, **51(2)**, 526-544.

Luan Y. and Li H. (2003), "Clustering of time-course gene expression data using a mixed-effects model with B-splines", *Bioinformatics*, **19(4)**, 474–482.

Pollard K.S. and van der Laan M.J. (2005), "Cluster Analysis of Genomic Data", In Robert Gentleman, Vincent J. Carey, Wolfgang Huber, Rafael A. Irizarry, Sandrine Dudoit, editors, *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, chapter 13, pages 209–228, Springer.

R Development Core Team (2006), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, URL **http://www.R-project.org**.

Scharl T. and Leisch F. (2006), "The stochastic qt–clust algorithm: evaluation of stability and variance on time– course microarray data" In Alfredo Rizzi and Maurizio Vichi, editors, *Compstat 2006 – Proceedings in Computational Statistics*, pages 1015–1022, Physica Verlag, Heidelberg, Germany, ISBN 3-7908-1708-2.

Serban N. and Wasserman L. (2005), "CATS: Clustering After Transformation and Smoothing", *Journal of the American Statistical Association*, **100(471)**, 990–999.

Sheng Q., Moreau Y., De Smet F., Marchal K., and De Moor, B. (2005), "Advances in cluster analysis of microarray data." In Francisco Azuaje and Joaquin Dopazo, editors, *Data Analysis and Visualization in Genomics and Proteomics*, John Wiley & Sons, Ltd, ISBN 0-470-09439-7.

Tavazoie S., Hughes J.D., Campbell M.J., Cho R.J., and Church G.M. (1999), "Systematic determination of genetic network architecture", *Nature Genetics*, **22(3)**, 213–5.

Thomas J.G., Olson J.M., Tapscott S.J., and Zhao L.P. (2001), "An Efficient and Robust Statistical Modeling Approach to Discover Differentially Expressed Genes Using Genomic Expression Profiles", *Genome Research*, **11**, 1227–1236.

Zeng Y. and García–Frías J. (2006), "A Novel HMM–Based Clustering Algorithm for the Analysis of Gene Expression Time–Course Data", *Computational Statistics and Data Analysis*, **50**, 2472–2494.