CHAPTER 9

An Application of Unsupervised learning methods to Proteomic Data from Colon Cancer

Nairanjana Dasgupta[1]
Department of Statistics,
Washington State University,
Pullman, WA 99164-3144, US
E-mail: dasgupta@wsu.edu

Yibing Oliver Chen[1], **Rashmita Basu**[1]
Washington State University, Department of Statistics,
Pullman, WA, US
and
Sayed S. Daoud[2]
Washington State University, College of Pharmacy,
Pullman, WA, US

ABSTRACT: Clustering approaches are a useful tool to identify pattern and groups in data sets. The clustering technique has become increasingly popular in the analysis of genetic/proteomic data mainly because of easy availability in most statistical software packages. Biologists often use clustering methods to identify novel classes and groups and call them "class discoveries". In this paper we studied the agreement between different clustering approaches. classification, hierarchical as well as model-based techniques. We used a proteomic data for colon cancer as an illustration of the methods. Based on the different proteins observed (using the number of peptides for each protein), we used hierarchical (using Jaccard and Euclidean distances), *k*-means, and model based clustering methods to group patients. Our findings suggest little agreement between the different techniques in terms of clustering. The four methods agreed for only 34 out of the 92 proteins. Most agreement was between the two hierarchical methods we used. This indicates that clustering is an exploratory tool and should be used with caution in the science of discoveries.

9.1 INTRODUCTION TO CLUSTERING ANALYSIS

In the analysis of genetic and proteomic data there has been a major emphasis on unsupervised learning or clustering. The general premise of clustering is that we have multiple *k* variables on n observations and we would like to use the similarity of the k variables to determine how the n observations are grouped in g groups. Clustering is an EXPLORATORY statistical technique used to break up a data set into smaller groups or "clusters" with the idea that objects within a cluster are similar and objects in different clusters are different. Clustering is called "unsupervised learning" by computer scientists and "class discovery" by micro-array biologists. Although a large body of literature exists on this topic [1-5], to name a few, a very few formal theories have developed about clustering analysis. Intuitively the idea of clustering is: *cluster the internal cohesion and external isolation*. For clustering we need two mathematical measures:

Distance: the distance metric used to measure the distance between two points

Linkage: condensation of each group of observations into a single representative point.

9.1.1 DIFFERENT DISTANCE MATRICES AND LINKAGES IN CLUSTERING

Let $x = (x_1, ..., x_n)$ and $y = (y_1, ..., y_n)$, (z_{ij} be the standardized coefficients) and $\Theta = Var(x - y)$, then we can define some of the common distance metrics used in clustering:

1. L_1 (Manhattan): $d_1(x, y) = \sum | x_i - y_i |$
2. L_2 (Euclidean, ruler distance): $d_2(x, y) = [\sum (x_i - y_i)^2]^{1/2} = [(x_i - y_2)(x_i - y_2)]^{1/2}$
3. Standardized ruler-distance $[(z_{i1} - z_{i2})(z_{i1} - z_{i2})]^{1/2}$
4. Mahalanobis Distance: $[(x_i - y_i)\Theta^{-1}(x_i - y_i)]^{1/2}$
5. Correlation distance: $1 - r$, where r is the Pearson correlation coefficient.
6. Jaccard's distance (used for binary data): the Jaccard coefficient measures the overlap that two binary data sets share. Let M_{11} represents the total number of attributes where both have a value of 1. M_{01} and M_{10} represents the total number of attributes where the one is 0 and the other is 1. M_{00} represents the total number of attributes both have a value of 0. The Jaccard Distance, J is given as $(M_{10} + M_{01})/(M_{01} + M_{10} + M_{11})$.

Some of the common linkage methods used are: *Average Linkage*: the distance between two groups of points is the average of all pairwise distances.

Median Linkage: the distance between two groups of points is the median of all pairwise distances.

Centroid Linkage: the distance between two groups of points is the distance between the centroids of the two groups.

Single Linkage: the distance between two-groups is the smallest of all pairwise distances.

Complete Linkage: the distance between two-groups is the largest of all pairwise distances.

Ward's Method: it is an Analysis of Variance (ANOVA) based method aimed to minimize the loss associated with each grouping.

9.1.2 Types of Clustering

Clustering is broadly divided into Hierarchical and Non-hierarchical methods. In non-hierarchical clustering we partition the data into a pre-specified number *g* of mutually exclusive and exhaustive groups. We iteratively reallocate the observations to clusters until some criterion is met, e.g. minimize within cluster sums of squares. One major issue is that we need to know the seeds and the number of clusters to start off with. The advantage to this type of clustering is this is optimal for certain criteria and that all observations are automatically assigned to clusters without any subjectivity. However, the disadvantage is that the prior knowledge of how many clusters is crucial and is often computationally time consuming.

Hierarchical clustering methods produce a tree or dendrogram thus the method avoids specifying how many clusters are appropriate by providing a partition for each g obtained from cutting the tree at some level. The tree can be built in two distinct ways bottom-up: agglomerative clustering and top-down: divisive clustering. The main advantage is it is a visual method and it is faster computationally. Most biologists (especially Geneticists) use this method often as there is a lot of software that produce different graphics like dendograms. However, the dendogram should be interpreted with care, as each branch of the dendogram is really like a mobile and can rotate, without altering the mathematical structure of the tree. Neighboring nodes are “close” ONLY if they lie on the same branch. It has been proposed one should slice the tree and look at the clusters produced therein. However, where to cut the tree is subjective and there is no consensus about this. Deciding, “how many clusters” has been researched and a number of authors have provided different techniques to aid this decision process. These are statistics based on within- and between-clusters matrices of sums-of-squares and cross-products (30 methods reviewed by [6]), Average silhouette [7], Graph theory [8] and Model-based methods: EM algorithm for Gaussian mixtures, Fraley & Raftery [9-11] and [12]. Recently there have been more interest in re-sampling based methods and the following are a few methods that have been suggested using bootstrapping methods for estimating the number of clusters. These include Gap statistic [13], WADP [14]), Clest [15] and Boostrap [16].

Simplistically, clustering cannot fail. That is, every clustering method will return clusters, whether the data are organized in groups or not. Clustering helps to group / order information and is a visualization tool for learning about the data. However, clustering results do not provide any kind of “proof”. However, recently more model based methods are being used following [9]. Here they suggest that the *n* observations are really normal mixtures coming from the *g* groups with a mean vector and a Variance-Covaraince Matrix. The k groups are identified by their parameters using the EM algorithm to estimate the variance structure and the number of groups *g*. The method to detect the number of groups is based on BIC (Bayes Information Criterion). Model Based Clustering also identifies the shape and structure of the Variance-Covariance matrix.

In this paper we use partitioning, hierarchical and model based technique to look at group assignment for disease stage for the for colon cancer patients. We focus on the k-nearest neighbor for the partitioning method, hierarchical model with Euclidean Distance and Jaccard distance and model based clustering, which we discuss in Section 3. In Section 2 we provide a background for the data and describe the data set used in this manuscript. In Section 4 we delineate our results from clustering. Section 5 includes our discussion.

9.2 DATA BACKGROUND AND DESCRIPTION

Colon cancer is a very common cancer and is the third most commonly diagnosed cancer in the world. It is more common in the developed world and it has been estimated that more than half of the people who die of colorectal cancer live in a developed region of the world. In 2008, 1.23 million new cases of colorectal cancer were clinically diagnosed, and that this type of cancer killed more than 600,000 people. While curable in the early stages this cancer does not have many symptoms and is often left untreated when it enters the lymph nodes. In Colon cancer increased sensitivity to chemotherapeutic agents is associated with the lack of the protein p21 [17-18]. Thus there was a need to study the molecular stress responses of p21 following cell damage. In the study described in Lee and Daoud (2009) [19] they compared the protein expression change in subcellular fractions (nucleus, cytoplasm and mitochondria) of the human colon carcinoma cell line HCT-116 and an isogenic p21 knock-out HCT-116 p21—after treatment with Topoisomerase I inhibitor topotecan (TPT), which is often used for chemotherapy in Ovarian cancers. The data collection and description are provided in [19]. In the present study we are interested in seeing how the different proteins cluster together. Our data comprise of the number of peptides present after proteomic analysis for each protein under the 12 conditions (3 subcellular fractions, p21 and knockout, TPT and without). For our analysis we used 92 of the proteins identified for the 12 conditions for clustering and we are interested in the number of peptides identified in each condition.

9.3 A BRIEF STATISTICAL REVIEW OF THE METHODS USED

9.3.1 K-means Clustering

K-means is a very common partitioning technique due to MacQueen (1967) [20]. In this method we compare the distances of each observation from the mean vectors of the g proposed clusters and the observation is assigned to the cluster with the “nearest” mean vector. The distances are then recomputed and reassignments made and the process continued till all subjects are assigned to a cluster. Any of the distances described in Section 3 can be used. In R the algorithm Hartigan and Wong (1979) [21] is used by default.

9.3.2 Hierarchical Clustering (Ward’s Method)

Hierarchical cluster analysis uses distances for the n objects being clustered. Initially, each object is assigned to its own cluster and then iteratively, at each stage the two most similar clusters are joined, continuing joining until there is just a single cluster. At each stage distances between clusters are recomputed by the Lance.Williams dissimilarity update formula according to the particular clustering method being used. A number of different linkage methods are provided in Section 1. We used the Ward's minimum variance method which aims at finding compact, spherical clusters. We used both the Euclidean Distance and the Jaccard distance (in Section 1) in our analysis.

3.3 Model Based Clustering

This technique is popular from Fraley and Rafterys (2000) [10] manuscript. Essentially, it is a generalization of the K-means method. This method assumes that the data come from a model with mixture distributions and the method uses the EM algorithm to identify the model. The model that is recovered from the data then defines clusters and an assignment of objects to clusters. Let θ denote the parameters of the model and Maximum Likelihood method is used to maximize the likelihood given the data to identify the parameters. More generally, the maximum likelihood criterion is to select the parameters θ that maximize the log-likelihood of generating the data D:

$$\hat{\theta} = ArgMax(L(D\,|\,\theta))$$

$L(D\,|\,\theta)$ is the objective function that measures the goodness of the clustering. We use EM algorithm to identify the parameters and pick the estimate of θ that has the largest $L(D\,|\,\theta)$. Model Based Clustering also is able to look at various forms of the $\sum$ matrix. The forms it can identify are given in Table 1 of [10].

9.4 RESULTS FROM CLUSTERING

We first ran a SCREE plot to understand the true dimension of the data. The plot indicates 4 clusters appear to be sufficient as the Eigen value of 1 corresponds to 4 components.

9.4.1 K-Means Method

The plot for the *k*-means clustering is given in Figure 1. It shows that Group 1 is well separated from Group 4, but there are some overlaps in the other groups. We present the results in Table 1 with the proteins and the group memberships in the column K means.

9.4.2 Hierarchical clustering (Euclidean Distance)

We then focus on using hierarchical clustering (Ward's Linkage) using the Euclidean Distance. We present our dendogram in Figure 2. Again here we have one large cluster and the rest of them smaller. The proteins and the cluster membership are given in Table 1 under the column Hierarchical - Euclidean.

9.4.3 Hierarchical Clustering using Jaccard's Distance

Using this model with the Jaccard distance (suitable for binary data and count data) we have the dendogram given in Figure 3. Here we have a larger clusters as well, but not as large as k-means or the Euclidean distance. Table 1 shows the proteins with their class memberships under the column Hierarchical- Jaccard.

9.4.4 Model Based Clustering:

Using model based clustering we find that the best model with the BIC criterion is an elliposidal multivariate normal model. The BIC plot is given in Figure 4. The class memberships for the proteins are given in Table 1.

9.5 CONCLUSION AND DISCUSSION

In Table 1 we give the proteins and the clusters they belonged to. As it is seen that out of the 92 proteins used in the three methods that we used only 34 proteins were similarly clustered by all four methods. This is disturbing since this indicates in 2/3 of the situations there were differences in the techniques used for clustering. Comparing them pairwise the results follow:

Jaccard and Euclidean: 63 agreements

Jaccard and Kmeans: 51 agreements

Jaccard and Model Based: 57 agreements

Euclidean and Kmeans: 57 agreements

Euclidean and Model Based: 58 agreements

Kmeans and Model Based: 40 agreements

out of a total of 92 proteins. This indicates the most agreement among the two hierarchical methods and least between the model based and the *k*-means technique.

We used several other methods of clustering for the peptide data and find the clusters were almost always different and it mattered which method was used. As a matter of fact different distance metrics (different linkages) in hierarchical clustering does produce different clusters and we do not include all the different ones for space consideration. This is disturbing as in general class membership is not known and this means every different clustering method can result in different results. While this has been shown before, clustering is still widely used and interpreted in terms of class discovery by geneticists. Hence, this manuscript is intended to add to the word of caution in using clustering methods.

Our main statistical finding is that clustering, in spite of all the research in the topic, is still an exploratory method and results from clustering should be interpreted with care. Using a single clustering method to propose class discovery is a very dangerous trend and if anything, clustering should be used for exploratory purpose and several methods tried to look at general trends in the clusters.

References

[1] Everitt, B., Cluster Analysis, London: Heinemann Educ. Books, 1974.

[2] Hartigan, J. A., Clustering Algorithms, New York: Wiley, 1975.

[3] Anderberg, M. R., Cluster Analysis for Applications, Academic Press: New York, 1973.

[4] Gordon, A. D., Classification. Second Edition, London: Chapman and Hall / CRC, 1999.

[5] Murtagh, F., “`Multidimensional Clustering Algorithms”’, in COMPSTAT Lectures 4, Wuerzburg: Physica-Verlag (for algorithmic details of algorithms used), 1985.

[6] Milligan G., and Cooper M., An examination of procedures for determining the number of clusters in a data set, Psychometrika, Vol 50, 2 (159-179),1985.

[7] Kaufman, L., and Rousseeuw, P, Finding Groups in Data: An Introduction to Cluster Analysis, John Wiley & Sons, Inc., 1990.

[8] Ben-Dor, A., Shamir, R, and Yakhini, Z., Clustering Gene Expression Patterns, Journal of Computational Biology, 6(3-4): 281-297, 1999.

[9] Fraley, C., and Raftery, A.E., Model-Based Clustering, Discriminant Analysis, and Density Estimation, Journal of the American Statistical Association, 97(458): 611-631, 2002.

[10] Fraley, C., and Raftery, A.E., MCLUST Version 3 for R: Normal Mixture Modeling and Model-Based Clustering, Technical Report no. 504, Department of Statistics, University of Washington, 2006.

[11] Fraley, C., and Raftery, A.E., How Many Clusters ? Which Clustering Method? Answers via Model-Based Cluster Analysis: Computer Journal, Volume 41, Issue8, Pp. 578-588, 1998.

[12] McLachlan G. and Krishnan T, The EM algorithm and extensions. Second Edition, John Wiley and Sons., 2008.

[13] Tibshirani R., Estimating the number of clusters in a data set via the gap statistic, Journal of the Royal Statistical Society: Series B, Volume 63, Issue 2, pp 411–423, 2001.

[14] Bittner M. et al, Molecular classification of cutaneous malignant melanoma by gene expression profiling, Nature, 536-540, 2000.

[15] Dudoit, S., Fridlyand, J., and Speed, T., Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data, Journal of the American Statistical Association, 97(457): 77-87, 2002.

[16] Van der Laan, M., Dudoit, S., and Pollard, K., Augmentation Procedures for Control of the Generalized Family-Wise Error Rate and Tail Probabilities for the Proportion of False Positives, Statistical Applications in Genetics and Molecular Biology, Vol. 3: Iss. 1, Article 15, 2004.

[17] Bartek J., and Lukas J., Pathways governing GI/S transition and their response to DNA damage, FEEBS Lett.,490, 117-122, 2001.

[18] Bunz F., Hwang P.M., Torrance C., Waldham T., Zhang Y., Dillehay L., Williams J., Lenguar C., Kinzler K.W. and Vogelstein B., Disruption of p53 in human cancer cells alters response to therapeutic agents. J. Clin. Invest, 57, 1-12, 1999.

[19] Lee K. and Daoud S.S., Subcellular Proteomics for Topotecan-Induced Stress Response, The Open Proteomics Journal, 2, 30-39, 2009.

[20] MacQueen, J., Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, eds L. M. Le Cam & J. Neyman, 1, pp. 281–297, Berkeley, CA: University of California Press, 1967

[21] Hartigan, J. A., and Wong, M. A., A K-means clustering algorithm, Applied Statistics 28, 100–108, 1979.

Figure 9.1: Plot of the *k*-means grouping

Figure 9.2: Dendogram from Hierarchical Clustering with Euclidean Distance with Ward's Linkage

Figure 9.3: Dendogram from Hierarchical Clustering with Jaccard Distance and Ward's linkage

Figure 9.4: BIC plot for deciding on model from model based clustering

Table 9.1: Cluster Assignments For Each Protein By Method and Their Overall Agreement Status

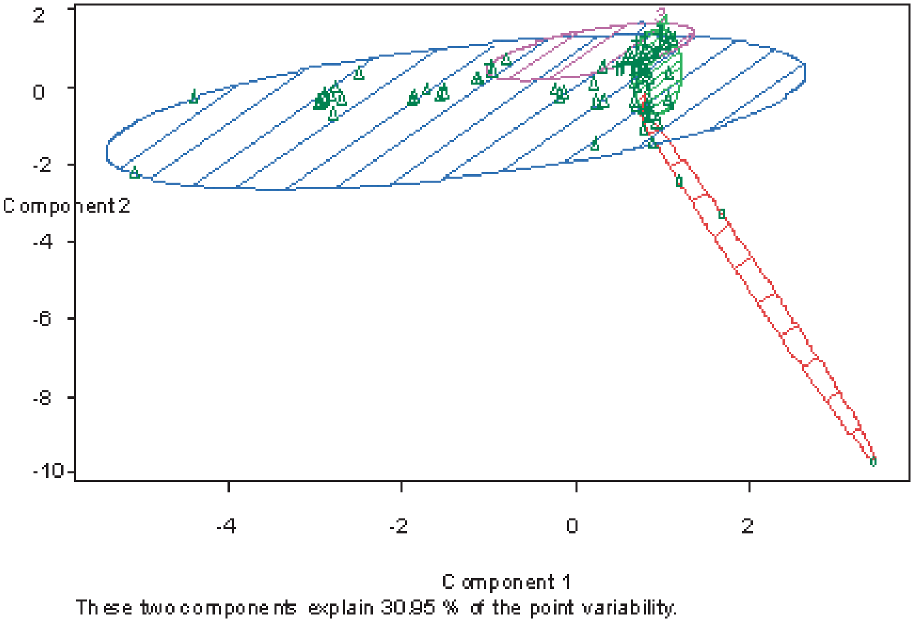
Protein	*Hierarchical (Euclidean)*	*Hierarchical (Jaccard)*	*Kmeans*	*Model Based*	*Agreement among all 4 methods*
O43169	1	1	1	1	*agree*
O60814	2	2	2	3	*disagree*
O75489	2	2	2	2	*agree*
P00505	2	2	2	2	*agree*
P04350	2	3	3	3	*disagree*
P04843	2	2	2	2	*agree*
P05388	3	3	3	2	*disagree*
P06733	2	2	2	2	*agree*
P06748	2	4	2	4	*disagree*
P07195	3	3	3	2	*disagree*
P07195	3	3	3	2	*disagree*
P08574	1	1	1	2	*disagree*
P08758	1	1	4	2	*disagree*
P11940	2	4	2	4	*disagree*
P12235	2	3	3	2	*disagree*
P12236	1	1	4	2	*disagree*
P12236	1	1	4	2	*disagree*
P12277	4	4	2	4	*disagree*
P12814	1	2	1	2	*disagree*

Protein	*Hierarchical (Euclidean)*	*Hierarchical (Jaccard)*	*Kmeans*	*Model Based*	*Agreement among all 4 methods*
P12956	2	4	1	2	*disagree*
P 17844	1	1	1	1	*agree*
P17987	2	2	2	2	*agree*
P18077	2	2	2	2	*agree*
P18621	2	4	2	4	*disagree*
P21333	4	4	2	4	*disagree*
P23396	3	3	3	2	*disagree*
P26641	2	2	2	2	*agree*
P26641	2	2	2	2	*agree*
P29692	2	2	2	2	*agree*
P31930	2	2	2	2	*agree*
P31943	2	2	2	2	*agree*
P31948	2	4	2	4	*disagree*
P32969	2	3	4	2	*disagree*
P35580	2	3	3	2	*disagree*
P36873	2	2	2	2	*agree*
P40429	2	2	2	2	*agree*
P46776	4	4	3	4	*disagree*
P46779	4	4	2	4	*disagree*
P46782	3	3	2	2	*disagree*
P47914	2	2	2	2	*agree*
P48643	4	4	2	4	*disagree*
P49411	2	2	2	2	*agree*

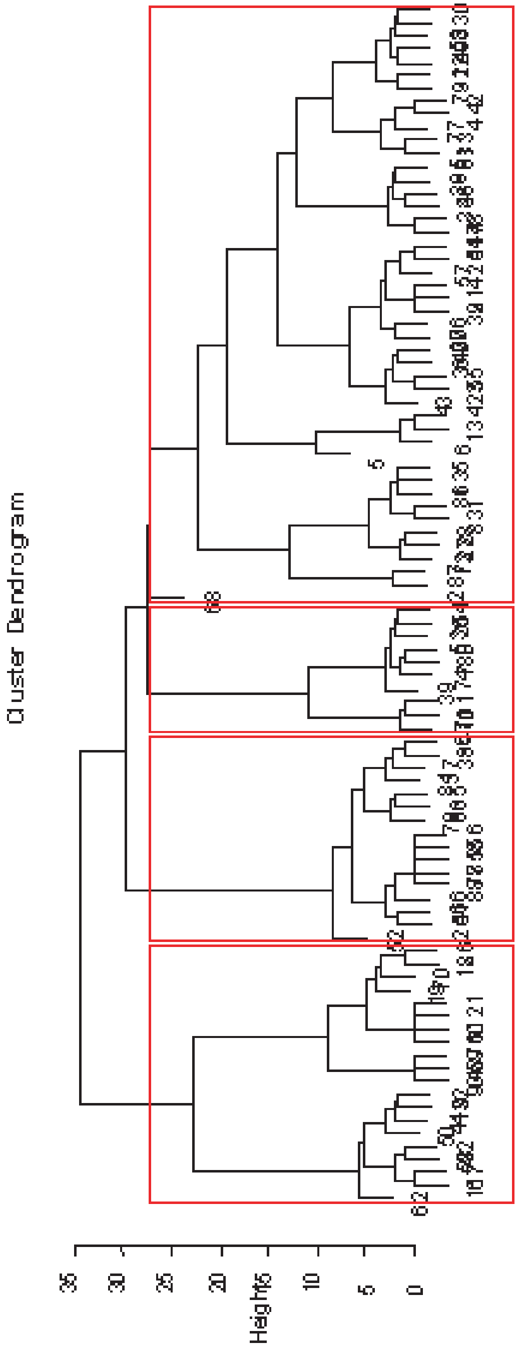
Protein	*Hierarchical (Euclidean)*	*Hierarchical (Jaccard)*	*Kmeans*	*Model Based*	*Agreement among all 4 methods*
P50914	2	2	2	2	*agree*
P51149	1	1	4	2	*disagree*
P54652	1	1	1	1	*agree*
P55072	2	3	4	2	*disagree*
P60842	2	3	2	2	*disagree*
P60866	2	3	2	2	*disagree*
P60953	2	2	2	2	*agree*
P61006	1	1	4	2	*disagree*
P61158	2	4	2	4	*disagree*
P61254	4	2	2	3	*disagree*
P61313	3	3	4	2	*disagree*
P61353	3	3	2	2	*disagree*
P61978	2	2	2	2	*agree*
P61981	4	4	2	4	*disagree*
P62081	2	4	2	4	*disagree*
P62280	2	2	2	2	*agree*
P62304	1	1	4	2	*disagree*
P62306	1	1	1	1	*agree*
P62491	2	2	2	2	*agree*
P62736	1	1	4	2	*disagree*
P62750	4	4	2	4	*disagree*
P62753	4	4	2	4	*disagree*
P62899	4	4	2	4	*disagree*

Protein	*Hierarchical (Euclidean)*	*Hierarchical (Jaccard)*	*Kmeans*	*Model Based*	*Agreement among all 4 methods*
P62913	4	4	2	4	*disagree*
P63104	3	3	3	2	*disagree*
P68032	2	2	2	3	*disagree*
P68366	2	3	3	2	*disagree*
P78371	1	3	1	2	*disagree*
P84090	1	1	1	1	*agree*
P84243	2	2	2	2	*agree*
Q00325	2	2	2	2	*agree*
Q02543	3	3	3	2	*disagree*
Q02790	4	4	2	4	*disagree*
Q02878	2	2	2	2	*agree*
Q03135	2	2	2	2	*agree*
Q04917	4	4	2	4	*disagree*
Q05639	4	4	1	4	*disagree*
Q06830	3	3	3	2	*disagree*
Q07955	2	2	2	2	*agree*
Q12906	1	1	4	2	*disagree*
Q13162	2	2	2	2	*agree*
Q14152	4	4	2	4	*disagree*
Q15149	2	3	2	2	*disagree*
Q16695	1	1	1	2	*disagree*
Q8N257	2	2	2	3	*disagree*

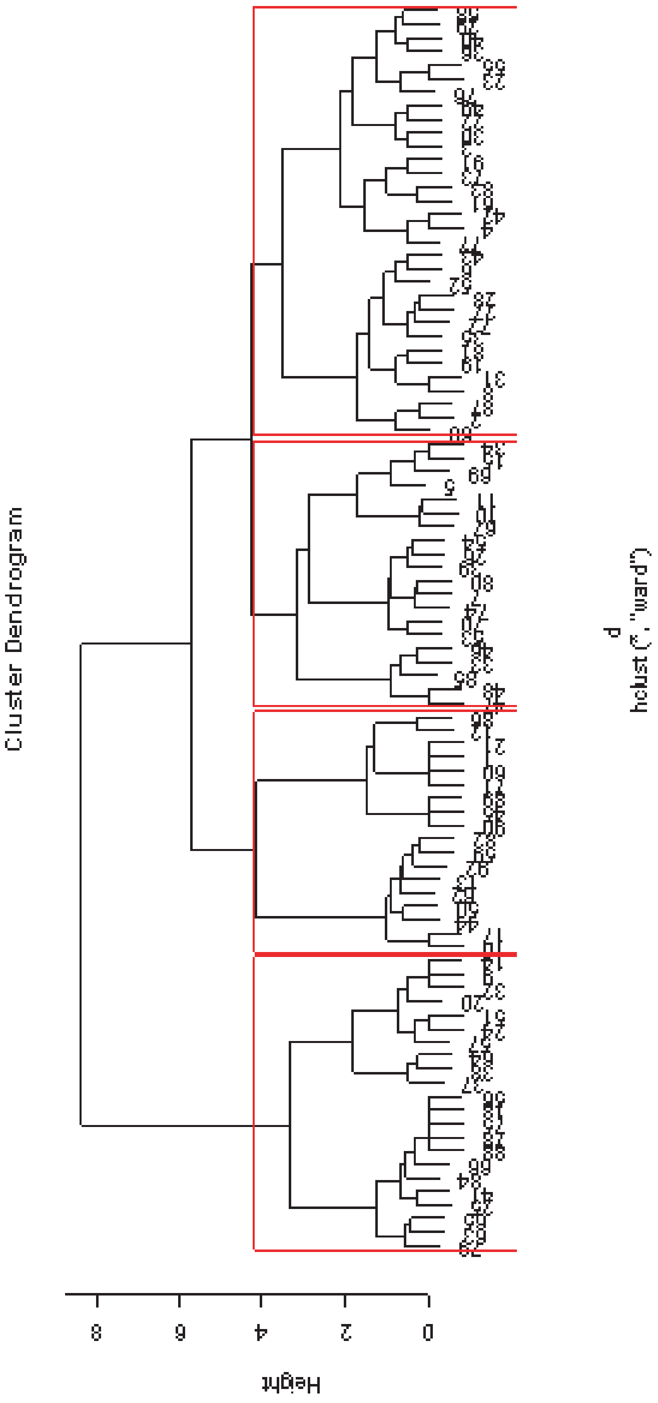
Protein	*Hierarchical (Euclidean)*	*Hierarchical (Jaccard)*	*Kmeans*	*Model Based*	*Agreement among all 4 methods*
Q92598	4	4	2	4	*disagree*
Q96A08	1	1	1	1	*agree*
Q99867	1	1	1	1	*agree*
Q9BQE3	2	2	2	2	*agree*
Q9NR30	1	1	4	2	*disagree*



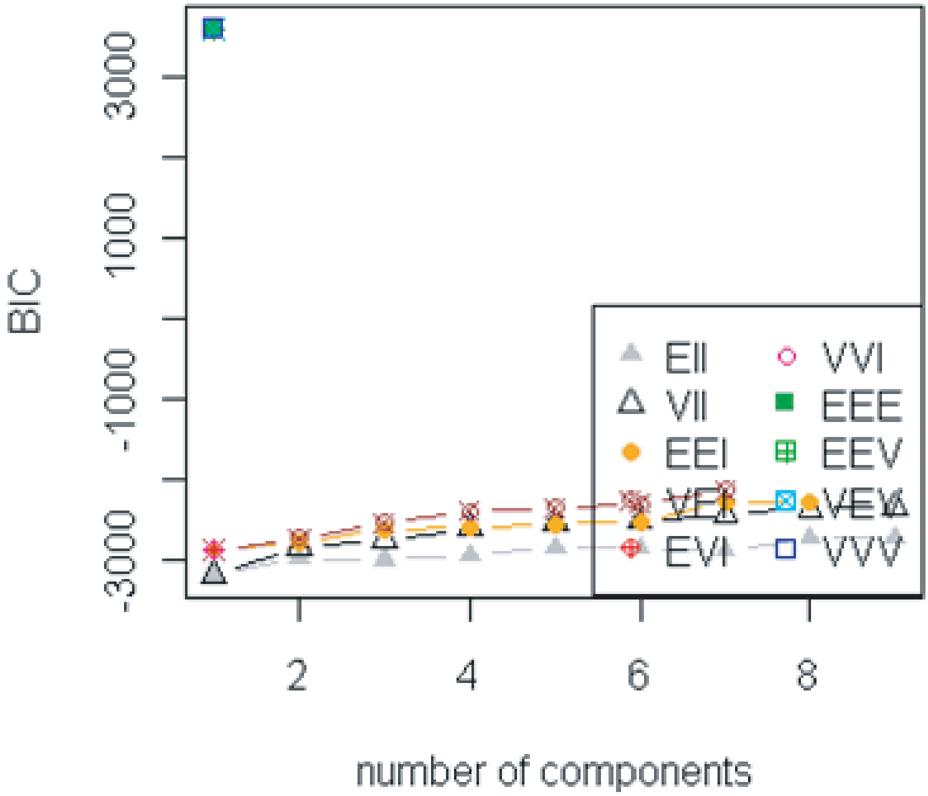
Please
send me all Figures again
all Figure are not clear



Please
send me all Figures again
all Figure are not clear



Please
send me all Figures again
all Figure are not clear



Please
send me all Figures again
all Figure are not clear