Journal of Information Technology Impact
Vol. 9, No. 3, pp. 135-144, 2009

Clustering Methods for Statistical Analysis of Genome Databases

Jayanthi Ranjan [1]
Institute of Management Technology
India

Abstract

Clustering techniques find interesting and previously unknown patterns in large-scale data embedded in a large multi dimensional space and are applied to a wide variety of problems like customer segmentation, Biology, data mining techniques, machine Learning and geographical information systems. Clustering algorithms are used efficiently to scale up with the dimensionality of the data sets and the data base size. Hierarchical clustering methods in particular are widely used to find patterns in multi dimensional data. Since clustering is an unsupervised learning technique, fewer or greater numbers of clusters may be desired. A key step in the analysis of gene expression data is the identification of groups of genes that are similar in nature. The developments of micro array technologies provide a powerful tool by which the expression patterns of thousands of genes can be monitored simultaneously. This paper describes the major statistical approaches in hierarchical clustering and compares the linkage methods that are used in gene expression data along with experimental results.

Keywords: Agglomerative techniques, statistical analysis, Dendrogram, micro array data.

Introduction

Clustering has been extensively studied in statistics, machine learning, pattern recognition and image processing (Kaufman & Rousseeuew, 1989). Clustering techniques find patterns previously unknown in large-scale data, embedded in a large, multi dimensional space. Efficient representation of the detected clusters is as important as cluster detection and improves its usability. Most of the earlier works in statistics operate and find clusters in the whole data space. The outputs of these algorithms are very sensitive to the input parameters. The scalability of these algorithms with the database size is important as their scalability with their dimensionality of the data sets. Noise present with data makes cluster detection harder. A wide range of techniques has been applied for clustering gene expression data (Eisen et. al, 1998). Examples include hierarchical clustering, adaptive resonance theory, self-organizing map, *k*-means, graph-theoretic approaches and growing cell structures network. However, most of the above-mentioned clustering algorithms are heuristically motivated, and the issues of determining the "correct" number of clusters and choosing a "good" clustering algorithm are not yet rigorously solved. Clustering gene expression data using hierarchical clustering and Self Organized Maps has been very popular among the bioinformatics community. Typically this will involve data processing

using various statistical techniques to identify the patterns. In addition, data needs to be packaged, presented, archived, and compared with other types of information.

Background

The most frequently used analysis on gene expression data is clustering. It is an exploratory technique for gene expression data as it groups similar objects together and allows the biologist to identify meaningful relationships between the objects. There are numerous algorithms and programs associated with clustering like hierarchical methods, self-organized maps, k-means, and model-based approaches (Ranjan, 2005). We in this paper focus on hierarchical agglomerative clustering. There are many similarity measures that can be applied to calculate the similarity and dissimilarity between a pair of objects. We have chosen the most popular Euclidian distance and correlation co-efficient. High correlation implies high similarity (Yeung, 2003). Euclidian distance is a dissimilarity measure, high distance means low similarity. Most clustering algorithms uses similarity matrix as input and produces group of objects similar to each other as output. The output of hierarchical algorithms is usually in form of dendrogram. The cluster similarity can be computer from the similarity matrix or on the data matrix (Shannon, 2003) . The Euclidean (and squared Euclidean) distances are usually computed from raw data, and not from standardized data.

The rest of the sections are organized as follows:

In section 2.1 and 2.2 we discussion the statistical analysis of micro array data and hierarchical clustering. Section 3 introduces the concept of micro array data and the data sets used. In section 4, we discuss the experimental results. Section 5 discusses the linkage method's results in detail using several software and online web servers. Finally we conclude in section6 with our remarks.

Statistical Analysis.

The statistical analysis of micro array data is probably the most difficult problem associated with the use of clustering. The aim is to apply standard statistical approaches to determine gene expression, thus enabling the extraction of significant biological information from noise and variability. Statisticians are experiencing with handling data involving a limited number of variables, but a large number of samples. A number of different methods have been explored. The output of statistical analysis of a micro array experiment is usually a large data spreadsheet or dendrogram filled with numbers related to the signal intensity for each element on the chip. Further analysis is required to identify groups of elements that are similarly regulated across the biological samples under study. A variety of mathematical and statistical procedures have been developed that partition elements or samples into groups, or clusters, with maximum similarity, and thus enabling the identification of elements.

Hierarchical Clustering

The hierarchical clustering techniques create a hierarchy of clusters from small to big since clustering is an unsupervised learning technique. Hence depending on the particular application

of the clustering, fewer or greater numbers of clusters may be desired. With a hierarchy of clusters defined it is possible to choose the number of clusters that are desired and required. Any clustering algorithm that ends up with as many clusters as there are records has not helped the user understand the data better. Exactly how many clusters should be formed is a matter of interpretation. The advantage of hierarchical clustering methods is that they allow the end user to choose from either many clusters or only a few. Hierarchical clustering has the advantage over non-hierarchical techniques in that the clusters are defined solely by the data (not by the users predetermining the number of clusters) and that the number of clusters can be increased or decreased by simple moving up and down the hierarchy (Ranjan, 2005).

All the hierarchical methods often work either in a top down manner,(by repeatedly partitioning the set of elements) or in a bottom up manner. More information on clustering techniques is available in Hartigan (1975), Dopazo (2002) . Hierarchical clustering methods are represented usually by a dendrogram. We focus on top down manner, which is called agglomeration. In agglomerative clustering technique we start with as many clusters where each cluster contains just one record. The clusters that are nearest to each other are merged together to form the next largest cluster. This merging is continued until a hierarchy of clusters is built with just a single cluster containing all the records at the top of the hierarchy. There are several agglomerative procedures that are probably most widely used in hierarchical clustering. They produce a series of partitions of the data: the first consists of n single-member 'clusters'; the last consists of a single group containing all n individuals (Everett, 2003). At each stage the methods fuse individuals or groups of individuals that are closest or most similar (single linkage). Differences between the methods arise because of the different ways of defining the distance (or similarity) between an individual and a group containing several individuals or between two groups of individuals (Everett, 2003). The standard agglomerative hierarchical clustering methods are Single Linkage, Complete Linkage, Average Linkage, Centroid Linkage, Median Linkage, and Ward's Method. A detailed description of these methods can be found in (Everett, 2003). Researchers and scientists have developed several clustering software packages using agglomerative methods for gene expression data. The information can be found in Appendix I.

Micro Array Data

Micro arrays exploit the preferential binding of complementary single stranded nucleic acid sequences. A micro array is typically a glass slide, onto which DNA molecules are attached at fixed locations (Shannon, 2003). There may be tens of thousands of spots on an array each containing a huge number of identical DNA molecules. Micro array technology makes use of the sequences recourses created by the genome projects and other sequencing efforts (Ranjan & Ahson, 2004). For instance, they allow comparisons of gene expression between normal and diseased (cancerous) cells. There are several names for this technology-DNA micro arrays, DNA arrays, DNA chips and gene chips. The technology is very new; methodologies are still evolving. Micro array techniques find utility in Gene discovery, Gene regulation, Drug discovery and toxicology, Diagnosis and identification of patterns of gene expression that define disease states and they may represent prognostic indicators. The micro array technology is rapidly developing therefore it is natural that currently there are no established standards for micro array experiments and how the raw data can be processed. The repository for gene expression data is being

developed at NCBI - National Center for Biotechnology Information in US. Array Express is storing at all the information, the details of which is called Minimum Information About A Micro array Experiment (MIAME) defined by the Micro array Gene Expression Database (MGED) consortium. We have taken - acute myeloid leukemia (AML) and lymphoblastic leukemia (ALL) data at the cancer genomic center www.broad.mit.edu. We have also collected data sets fromhttp://sdmc.lit.org.sg/GEDatasets/Datasets.html#BreastCancer, http:// www.genomebiology.com/2003/4/5/r34/suppl/, http://www.columbia.edu/~xy56/project.htm www.expression.washington.edu/public. In recent years there has been an explosion in the rate of acquisition of bio medical data (Piatesky-Shapiro & Tamayo, 2005). The major challenges in micro array data mining includes gene selection, classification of deceases and predicting outcomes and finding new biological classes or refining the existing ones (Piatesky-Shapiro & Tamayo, 2005).

Results

We have implemented several agglomerative approaches and performed the comparison study by evaluating the clustering results using the data sets. The data are arranged in tables where rows represent all genes and the columns represent the individual expression values obtained in each DNA array. We studied the performance of agglomerative algorithms on our data sets. For assessing the cluster quality, we obtained desired number of clusters from the dendrogram by cutting the merging process. For efficient visualization of the patterns extraction from the micro array data sets, we used different tools like HCE (HCE Server (2005), GEPAS- Gene Expression Pattern Analysis Suite (Dopazo, 2002), European Bio-Informatics Institute's Expression Profiler (Kapushesky, 2004), Clusfavor software (Peterson, 2002) SPSS, S-Plus (S-Plus, 2005), and Eisen's *Tree view* (Eisen et. al, 1998). We are interested in comparing the same type of data on agglomerative different methods.

We found our quantitative measures of cluster quality to be positively correlated with external standards of cluster quality. Once we have our agglomerative hierarchy of clustering, the best clustering can be chosen via domain-dependent preferences. One general method is by looking at the dendrogram for clusters, and pick the clustering when all (or most) such clusters have formed as the "natural" clustering; other less subjective methods also exist, which use heuristics such as ensuring that the within-cluster variance is less than the between-clusters variance. We found that hierarchical clustering has the advantage that the clusters are defined solely by the data (not by the users predetermining the number of clusters). In hierarchical clustering, the number of clusters can be increased or decreased by simple moving up and down the hierarchy (Kaufman & Rousseeuew, 1989). We evaluated the performance of the various clustering methods to obtain hierarchical solutions using the available clustering tools from S-Plus, Eisen's *Cluster*, European Bioinformatics Institute's expression Profiler, GEPAS-Gene Expression Pattern Analysis Suite, Hierarchical Clustering Explorer and Rousseau & Kaufman's *Web clusters* for single linkage, complete linkage, and average linkage methods. The first step we did in analyzing micro array data is to filter out genes that are not expressed or do not show variation across samples. This usually reduces the data set by 3000– 5000 genes. We have filtered genes using *Eisen's Cluster* (Eisen et. al, 1998).

Table 1. Agglomeration schedule of Clusters using Median Linkage(SPSS)

	Cluster Combined		Co-efficients	Stage Cluster First Appears		Next Stage
Stage	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	6	25	.163	0	0	4
2	13	26	.185	0	0	11
3	12	24	.200	0	0	4
4	6	12	.242	1	3	8
5	1	11	.244	0	0	7
6	18	19	.251	0	0	12
7	1	3	.324	5	0	9
8	6	28	.368	4	0	10
9	1	9	.396	7	0	14
10	6	23	.464	8	0	14
11	8	13	.478	0	2	16
12	18	29	.509	6	0	15
13	2	17	.529	0	0	17
14	1	6	.656	9	10	15
15	1	18	.687	14	12	16
16	1	8	.717	15	11	17
17	1	2	1.074	16	13	27
18	15	20	1.138	0	0	22
19	7	14	1.787	0	0	26
20	4	16	2.234	0	0	21
21	4	21	1.753	20	0	23
22	10	15	2.418	0	18	23
23	4	10	2.410	21	22	24
24	4	22	2.215	23	0	25
25	4	5	4.673	24	0	26
26	4	7	5.592	25	19	27
27	1	4	9.561	17	26	28
28	1	27	32.905	27	0	0

Discussion

We preprocessed the patterns by transforming the scale, handling the missing values, removing of flat gene expression patterns and the standardizing the remaining patterns. However these steps are optional and depend on the actual data set. All the algorithms to a certain extent impose their own structure of visualization. Of course there is no single best clustering procedure (Han & Kamber, 2000). Single linkage method's optimality is the oldest and simplest among all. Its useful property is that monotone transformations on the dissimilarities do not change the clustering (Han & Kamber, 2000). Its nearest neighborhood property even works for large sets of data. Figure 1 shows Dendrogram for Median linkage and represents a hierarchical agglomerative clustering of classes and genes derived using Squared Euclidian Distance. The results of the median linkage method produced agglomerative clusters as shown in Figure 1. We found that cluster centers are less sensitive to outliers. The median link method is invariant under monotone transformation of the distances like single linkage and complete linkage.(where as in average linkage, centroid methods the results show to be variant).

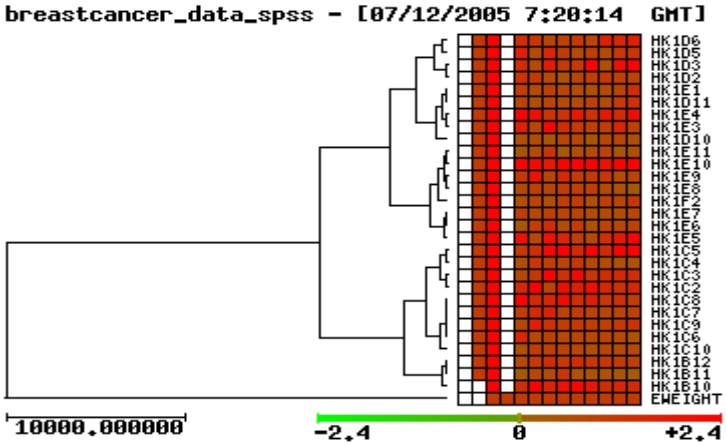
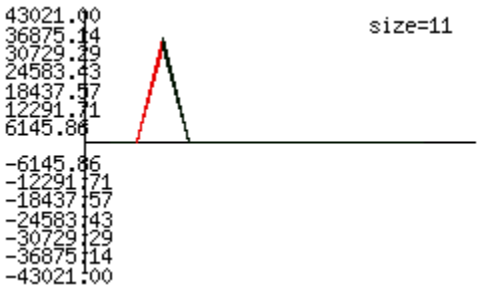


Figure 1. Dendrogram pertaining to breast cancer data.

It was difficult to spot a structure in a data set by merely looking at its dissimilarity matrix. The complete linkage method produced compact clusters since they had small diameter. We tried to investigate the effectiveness of clustering gene expression data. Since genes are clustered, the experimental conditions are the variables.

We have investigated the linkage algorithms that would detect the number of clusters in the gene expression data we used for the experiments. Prior to clustering, the data were normalized to have zero mean and unit variance. As described before, the number of clusters at which the tree-based index reaches its highest peak can be used as an indicator of optimal number of clusters (Figure 2). The results Expression Profiler (Kapushesky, 2004) indicated the distance of 2763.090 when the cluster size is 11. To evaluate the consistency of the clustering, the number of common genes in two clusters for different number of initial nodes was investigated. The results, which are represented in Figure 2 and Figure 3, show the number of genes obtained using UPGMA and the

correlation between the genes. The results using UPGMA indicated robustness in cluster quality (Figure 4).



CLUSTER_0001: HK1C10 HK1C9 HK1C8 HK1C7 HK1C6 HK1C5 HK1C4 HK1C3 HK1C2 HK1B12 HK1B11

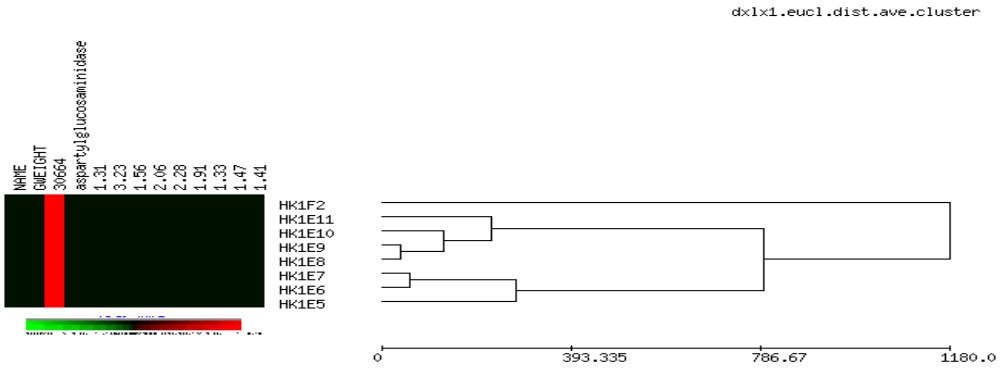


Figure 2. Visualization for the Complete Linkage method

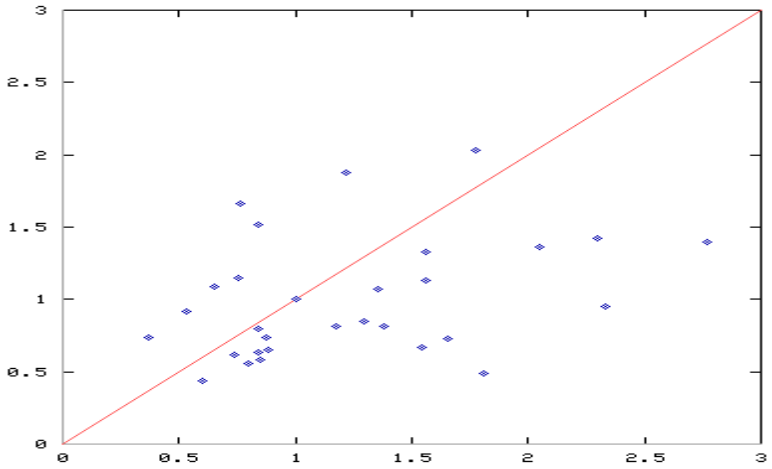


Figure 3. Correlation coefficient between genes, the pointers representing blue colors consist of genes against the experiments.

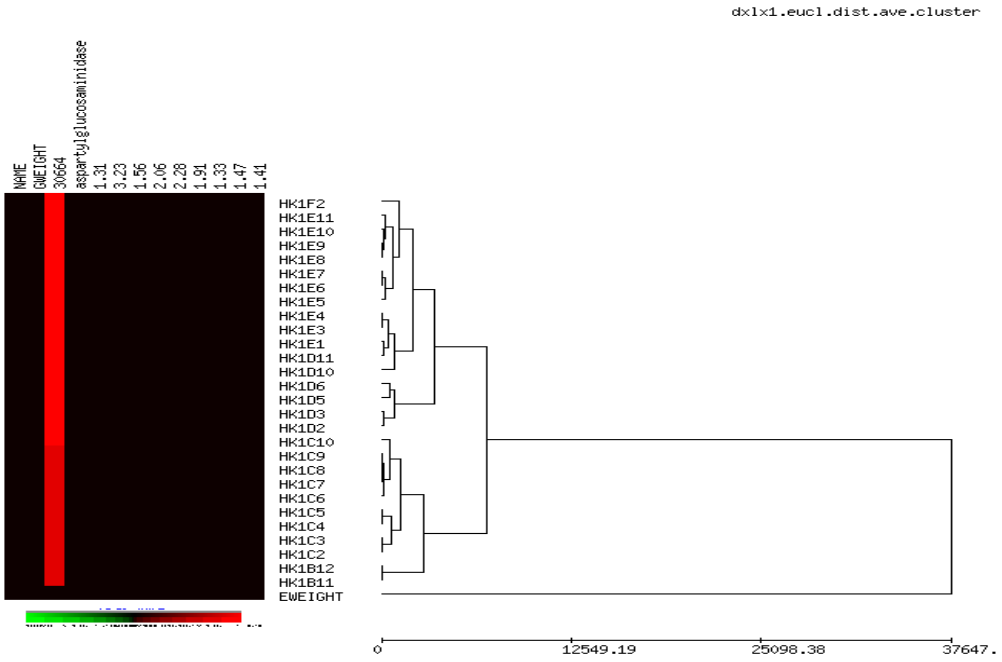


Figure 4. Hierarchical Clustering results from expression Profiler using UPGMA

The distance matrix is recalculated to include the distance between genes not clustered and the new cluster formed by the two genes. Many algorithms follow the series of merging of number of genes to produce hierarchical clustering of data. Variations between the algorithms can lead to different dendrograms and hence different clusters. It should always be noted that different authors define average clustering in different ways. Average operates by iteratively merging the genes or gene clusters with the smallest distance between them followed by an updating of the distance matrix. Single linkage calculates the distances between each gene in the new cluster to each of the genes in another cluster and takes the smallest distance. Complete linkage uses the largest distance of all these distances as the distance between the clusters. We have found the average linkage algorithm generally works well with standardized micro array data in contrast to single linkage that generally performs poorly.

Conclusion

The statistical analysis of micro array data is probably the most difficult problem associated with the use of these techniques. We have applied standard statistical agglomerative approaches to determine gene expression and gene expression alteration, thus enabling the extraction of significant biological information. Statisticians are experienced with a large number of samples instead of variables, Micro arrays produces thousands of variables from a small number of samples to handle the large samples. We have investigated the linkage algorithms that would

detect the number of clusters in the gene expression data. The linkage methods are usually considered together because of their easy implementation with the same computer program. In this paper, DNA micro array data from samples of primary breast tumors were analyzed using statistical clustering analysis to evaluate the patterns of interactions of groups of genes. Through Hierarchical clustering using linkage methods, we compared the gene expression data. The methods used are very general and applies to any data type providing that they can be coded as a series of numbers and that a computable measure of similarity between data items can be used. These improvements of neighborhoods correlate well with the improvements of overall clustering solutions, which suggest that constrained agglomerative schemes benefit from starting with purer neighborhoods and hence lead to clustering solutions with better quality.

References

Dopazo, J. (2002). Microarray data processing and analysis. In S.M. Lin & K.F. Johnson (Eds.), *Microarray data analysis II* (pp. 43-63). Kluwer Academic Publishers.

Eisen , M. B., Spellman, P. T., Brown, P.O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences* 95, 14863-68.

Everett, B. S., Landau, S., & Leese, M. (2003). *Cluster Analysis* (4th ed.). Oxford University Press, 55-89.

Han, J., Kamber, M. (2000). *Data Mining: Concepts and Techniques* (1st ed.). New York: Morgan-Kaufman.

Hartigan, J. A. (1975). *Clustering Algorithms* (4th Ed.). New York: John Wiley & Sons.

HCE Server (2005). Hierarchical Clustering Explorer HCE web server and software. Retrieved from www.cs.umd.edu/hcil/hce

Kapushesky, H., Kemmeren, P., Culhane, A.C., Durinck,S., Ihmels, J., Körner, C., Kull, M., Torrente, Sarkans, U., Vilo and Brazma.., A. (2004). Expression Profiler: next generation—an online platform for analysis of micro array data. *Nucleic Acids Research*, 32 (Web Server issue): W465-W470. Retrieved September 18, 2007, from http://www.ebi.ac.uk/expressionprofiler/

Kaufman, L., & Rousseeuew, P. J. (1989). *Finding groups in data: An introduction to cluster analysis*. In Wiley Series in *Probability and Statistics* (pp. 125-160). New York: John Wiley & Sons.

Peterson, L. E. (2002, Jun 24). CLUSFAVOR 5.0: Hierarchical cluster and principal-component analysis of micro array-based transcriptional profiles. *Genome Biology, 3*(7).

Piatesky-Shapiro, G., & Tamayo, P. (2005). Micro Array Data Mining: facing the Challenges. *SIGKDD Explorations, 5*(2), 1-5.

Ranjan, J., & Ahson, S. I. (2004). Efficient Agglomerative Clustering Methods for Microarray Data on Brest Cancer Outcome, Proceedings of 6th International Conference on Cognitive Systems (ICCS), Delhi India.

Ranjan, J., & Ahson, S.I. (2004). Statistical Agglomerative Clustering Methods for Gene Expression Data – A Comparison. *Bioinformatics India, 2*(3), 79-85.

Ranjan, J. (2005). *Some statistical methods in clustering techniques for large databases* (Doctoral dissertation). Jamia Millia Islamia, Central University Library, New Delhi.

Shannon, W., Culverhouse, R., & Duncan, J. (2003) Analyzing Micro Array data using cluster Analysis. *Pharmacogenomics, 4*(1), 41-51.

S Plus Statistical Software (2005). Retrieved from http://www.insightful.com/products/splus/default.asp

Yeung, K. Y., Medvedovic, M., & Bumgarner, R. E. (2003). Clustering Gene Expression Data with Repeated Measurements, Retrieved from *Genome Biology*, http://genomebiology.com /2003/4/5/R34

Appendix I

List of online web tools servers/ used.

1. www.ihone.cuhk.edu.hk/nb400559/arraysoft_mining.html
2. www.gepas.bioinfo.cnio.es
3. www.ebi.ac.uk/expressionprofiler/
4. www.condor.bcm.tmc.edu/genepi/clusfavor.html

---

[1] Dr. Jayanthi Ranjan is a professor of information technology and systems at the Institute of Management Technology, India. He can be eached at: Institute of Management Technology, Raj Nagar, Ghaziabad, India. E-mail: jranjan@imt.edu