A novel discretization method for processing digital gene expression profiles

Jibin Qu, Jinxia Zhang, Chenyang Huang†
Institute of Agricultural
Resources and Regional Planning,
Chinese Academy of
Agricultural Sciences.
Beijing 100081, China
Email: qujibin@amss.ac.cn
zhangjinxia@caas.cn
huangchenyang@caas.cn

Baogui Xie†
Mycological Research center
of Fujian Agricultural
and Forestry University.
Fuzhou Fujian 350002, China
Email: mrcfafu@163.com

Yong Wang, Xiang-Sun Zhang
Institute of Applied Mathematics,
Academy of Mathematics
and Systems Science,
Chinese Academy of Sciences.
Beijing 100190 China
Email: ywang@amss.ac.cn
zxs@amt.ac.cn

†: corresponding authors

***Abstract*—Discretization serves as an important preprocessing step for analyzing gene expression data and many algorithms have been proposed. However, most of the discretization methods were designed for microarrays. As a new technology, digital gene expression (DGE) profiles can overcome the limitation of microarrays and were applied in a widely range. In this paper, we proposed a novel discretization method for DGE data and the validations in a time-series gene expression dataset proved the efficiency of our method.**

I. Introduction

Gene expression data measure the concentration of mRNAs in given conditions and depict the quantitative characters of gene products in transcription. Microarray was the first high-throughout approach to measure the gene expression in a genome-wide scale [1] . This technique has been widely used in many sub-fields of molecular biology and produces massive amount of data [2][3]. However, microarray has its inherent limitations which result in some drawback of the data [4].

Digital gene expression (DGE) profiles are the tag-based gene expression profiling with next-generation sequencing techniques [5]. This approach counts the number of times of cDNA fragments from a corresponding transcript in a given sample, so the output is digital, rather than analog [6]. DGE profiles have significant advantages over microarray for many functional genomic applications[4].

Gene expression data are formed as matrix whose rows are genes (or transcripts) and columns are conditions. When analyzing the gene expression data, discretization is a useful preprocessing step to lower the dimensions of the data matrix [7][8][9]. Discretization, also called symbolization, means transformation of the raw data matrix into a symbol matrix. There are limited number of symbols and each symbol represents values in a certain scale in the raw matrix. Most of the current discretization methods were designed for microarray. Because of the different characters (e.g. scale, distribution, and others) between the two data types, it is inappropriate to apply these approaches directly to DGE data.

In this paper, we propose a novel discretization method to process gene expression data by DGE. Firstly we introduce the procedure of our method and the dataset used. Then some indices [10] are selected to compare the effect of k-means clustering algorithm on different discreted matrices with each other. To validate our method for feature-based clustering [11], we use a simple symbol clustering algorithm on the discreted matrices by different discretization methods and evaluate the clustering results by a new defined index. This validity index considers the intra-cluster diameter, inter-cluster distance, and number of included genes. Our method demonstrates better performance and can be applied to the RNA-seq data.

II. Material and Method

A. Material

We generated a time-series gene expression data by DGE: the transcriptomic profiles of the growth and development of *Volvariella volvacea*. The number of unambiguous clean tags for each gene was calculated as the gene expression value and then normalized to the number of transcripts per million clean tags (TPM), which was a standard indicator [12]. This dataset contained 6971 genes with six time points which represented six key stages in the life cycle, that was: mycelia, primordia, button, egg, elongation, and mature stages. The raw sequencing data of digital gene expression (DGE) of mycelia were submitted to Gene Expression Omnibus (GEO) database with association NO.GSE43019 [13].

In many studies, genes with low expression values in each stage were considered to be no information for further analysis. So we just selected genes with standard deviation of expression values in all stages larger than 20. There were 963 genes left after the filtering.

B. The discretization method

There were three steps for our discretization method.

Step 1: fitting.

978-1-4799-1389-3/13/$31.00 ©2013 IEEE

Firstly the distribution of the expression values was fitted. Here we assumed the raw data followed an exponential distribution, and the single parameter was estimated (see Figure 1).

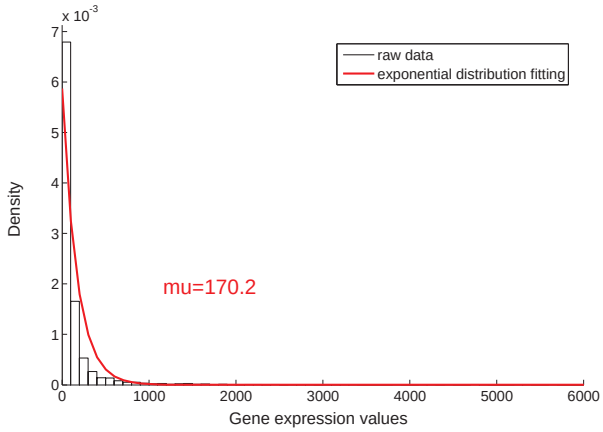

Fig. 1. The exponential distribution fitting (the red line) of the gene expression values in raw data. The estimated parameter is $\mu = 170.2$.

Step two: partition.

Based on the estimated distribution, we then partitioned the confidence interval into K_1 sub-intervals equally (with the same length). The expression values in a certain sub-interval were replaced by the mean value of this sub-interval. Here we usually selected an enough large K_1.

Step three: merge.

In this step, we merged the K_1 mean values to K clusters by the hierarchical clustering algorithm. So values of the sub-intervals whose mean values were assigned to the $ith(i = 1, 2, ..., K)$ clusters and were marked with the same symbol $i, (i = 1, 2, ..., K)$. A larger K_1 could maintain the robustness of the hierarchical clustering. K was the only parameters of our method.

The flowchart of our method was in Figure 2.

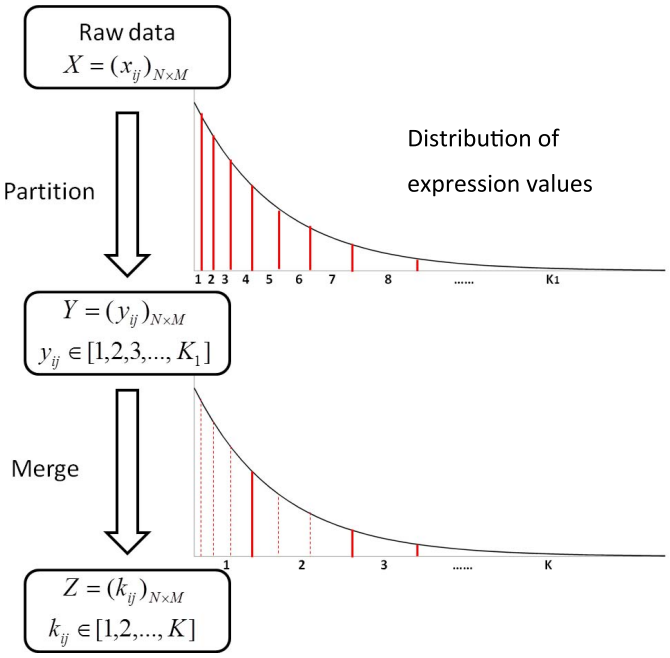

Fig. 2. The flowchart of our method. Z is our target matrix.

III. Result

A. Validation by K-means clustering

To validate our discretization method, we utilized the K-means clustering algorithm to cluster on the discretized symbol matrices, and then compared with the clustering results by some standard discretization methods. These procedures were as follows.

(i) Equal frequency discretization (EFD)[14]: this method divided the interval of expression values into k sub-intervals so that each sub-interval contained approximately the same number of expression values.

(ii) K-means discretization (KD)[15]: divided the interval of expression values of a particular gene into k sub-intervals by K-means clustering such that adjacent values were classified into the same sub-interval.

(iii) Column K-means discretization (Cokmeans)[8]: divided the interval of expression values of genes at a particular time point by K-means clustering so that adjacent expression values at the same time point were classified into the same sub-interval.

(iv) Bidirectional K-means discretization (Bikmeans)[8]: for this method, both K-means and Cokmeans were implemented with parameter $k + 1$, giving every expression value two discretized value. If the product of the two values was equal to or greater than x^2 and less than $(x + 1)^2$, the final discretized value of this expression value was x, where x was a positive integer ranging from 1 to k. Finally, expression values were divided into k sub-intervals.

There were many validity indices used to measure the quality of clustering results. Here four famous indices were selected to evaluate different discretization methods, that is, Calinski-Harabasz index (C-H index)[16], Davies-Bouldin index (D-B index)[17], Dunn index[18] and Silhouette index[19]. Among them smaller D-B index meant better results while the larger the better for other three indices.

All the considered discretization methods had the common parameter, the number of symbols K. It was a user-defined parameter. We assigned the K value of EFD and Cokmeans from 3 to 9. Because our data contained only six column, the K value of KD was given from 3 to 6 and the K value of Bikmeans was from 2 to 5. The indices under all parameters were calculated and we selected the best values for every index of every discretization method to compare with each other. The number of clusters for the K-means clustering algorithm was assigned from 2 to 20.

Figure 3 shows the comparation of various validity indices for K-means clustering on matrices discreted by different discretization methods. It is clear that for all K values of C-H index, D-B index and Silhouette index and most K values of Dunn index, our method (the green line) outperforms other discretization methods. Furthermore, we compared the

978-1-4799-1389-3/13/$31.00 ©2013 IEEE

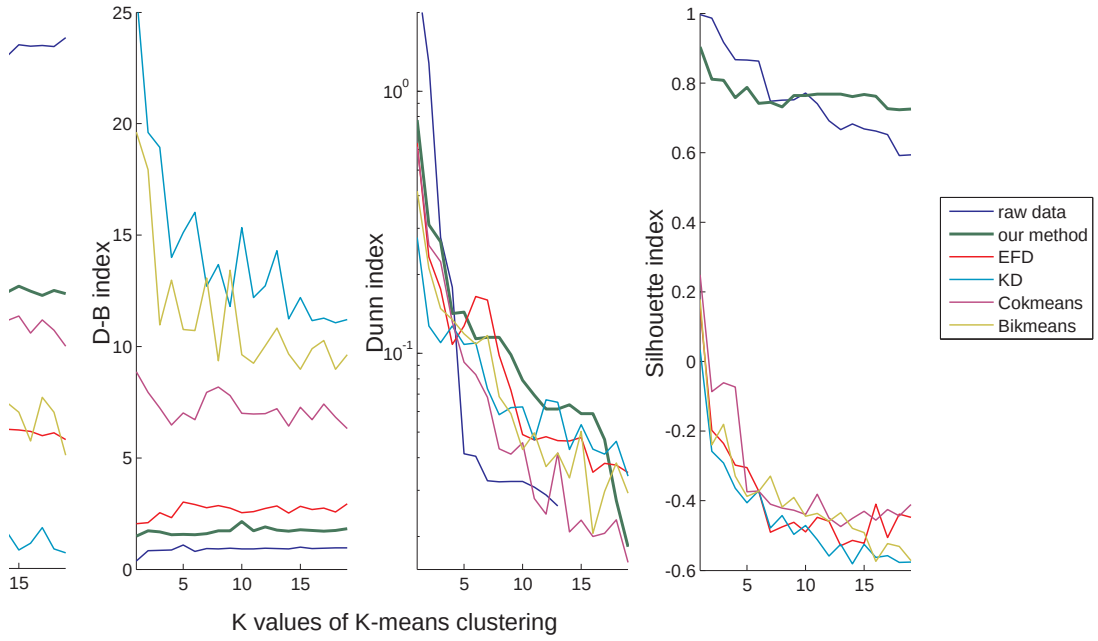

Fig. 3. Comparation of various validity indices for K-means clustering on matrices discretized by different discretization methods. K is from 2 to 20. The green line represents results by our method. The blue line represents clustering results for raw data.

clustering result on our discretized matrix and the clustering directly on the raw data. The results show that our method could outperform the clustering on the raw data (the blue line) in some cases (Dunn index and Silhouette index for some values of K).

B. *Validation by a symbol clustering method*

To further validate our method, we applied a simple symbol clustering algorithm on the discretized matrices. This algorithm selected genes with the same symbol list in all time points to the same cluster. Not all genes might be assigned a class label. Based on this algorithm we considered three aspects to evaluate the clustering result: the intra-cluster diameter, inter-cluster distance, and the number of included genes. The intra-cluster diameter for each cluster was defined as the average distance between every sample of this cluster and the cluster center. The inter-cluster distance for each cluster was defined as the distance between the center of this cluster and the center of the nearest cluster. The number of included genes for each cluster was the size of the cluster. A better clustering result would have smaller diameters, larger distances between clusters and include as many genes as possible. To meet the requirement, we defined a new index called Q-index as follows:

$$Q - index = \sum_{i=1\ldots K} \frac{distance(i) \times N(i)}{diameter(i)} \qquad (1)$$

K is the number of clusters in the clustering result, $diameter(i)$ is the diameter of the ith cluster, $distance(i)$ is the distance of the ith cluster to the nearest cluster, $N(i)$ is the included genes of the ith cluster.

We applied the symbol clustering algorithm to discretized matrices by different discretization methods with the same parameters as above. Q-index was calculated to compare the effects of discretization (see Figure 4). Our method (the blue curve) clearly outperformed others in all cases.

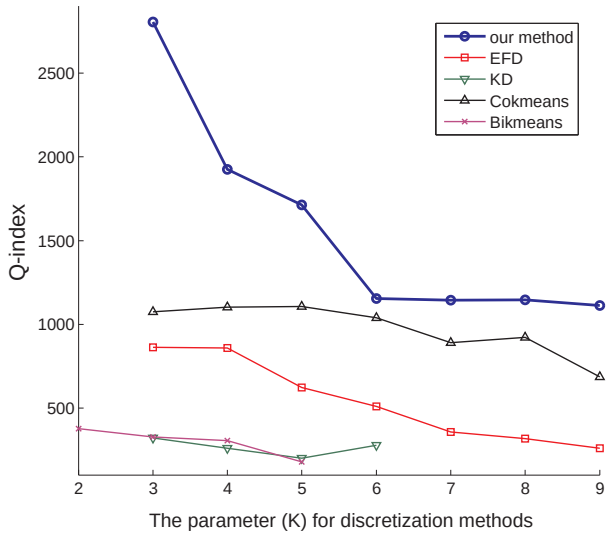

Fig. 4. Q-index for the symbol clustering results on different discreted matrices. The blue curve represents our method.

Here the clustering result with $K = 5$ by our discretization preprocessing was drawn as heatmap in Figure 5. Six clusters with the number of genes larger than 10 were shown, five of them had distinct expression patterns.

978-1-4799-1389-3/13/$31.00 ©2013 IEEE

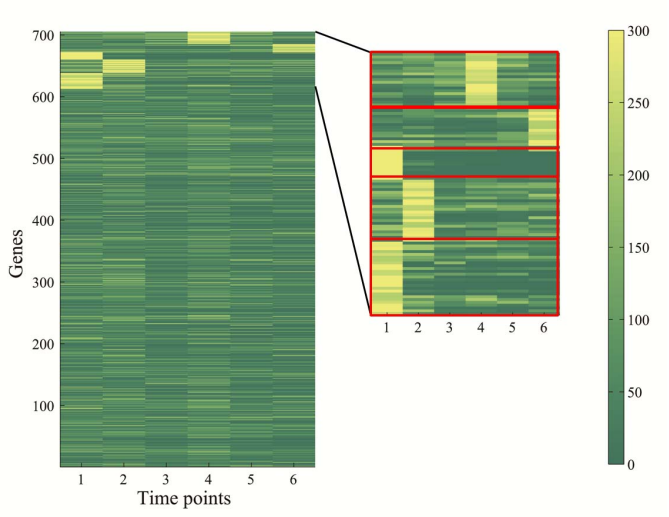

Fig. 5. The heatmap of clustering result by our discretization preprocessing with $K = 5$. Clusters with gene number larger than 10 are shown. Five clusters have obvious and specific expression patterns.

IV. Discussion and Conclusion

Discretization is a useful technology to preprocess the microarray data. This process could reduce the experimental errors and increase the robustness and accuracy for further analysis. Nowadays the next-generation sequencing technique has higher accuracy and repeatability which overcomes microarray, but the new data show extreme non-uniform and infinite range. The discretization methods appropriate to microarray are in pressing need to deep-sequencing data like DGE and RNA-seq.

In this paper we propose a novel discretization method for DGE profiles considering the distribution of data and compare with other discretization methods. The comparison was carried out by compare the clustering results on the discretized matrices in two aspects. The first is K-means clustering which is not a feature-based clustering and do not need discretization. Our method surpasses all of the other discretization methods and could even beat the clustering result on raw data in some cases. This means that our discretization method can maintain the variation of the raw data in a simplified mode. The other clustering algorithm is a simple feature-based clustering method and designed just for discretized matrix. The clustering result by our method has obvious advantages over the others. In fact, we can adjust the clustering result by the number of symbols in discreted matrix (the parameter K). A larger K means clusters with higher expression consistency and less number of genes. It turns out to be easy to detect genes with similar expression patterns based on our discretization method, especially by the simplest algorithm.

To evaluate the feature-based clustering, we defined a new validity index. Besides the diameter and inter-cluster distance, we also consider the number of included genes. It is important and needed to find correlations among genes during clustering. The Q-index may be useful to evaluate the clustering results with partial gene set.

RNA-seq technique has clear advantages over all existing approaches for mapping and quantifying transcriptomes [20]. Our method aims to discrete RNA-seq data just by fitting a proper distribution. Actually the DGE data and RNA-seq data have similar distribution, the exponentially distributed assumption may be still valid.

Acknowledgment

This work was supported by Research foundation of CAAS (302-3) and China Agriculture Research System (No. CARS24) .

References

[1] Schena M, Shalon D, Davis RW, Brown PO, *Quantitative monitoring of gene expression patterns with a complementary DNA microarray.* Science 1995, 270(5235):467-470.

[2] Yamato E, Tashiro F, Miyazaki J, *Microarray analysis of novel candidate genes responsible for glucose-stimulated insulin secretion in mouse pancreatic β cell line MIN6.* PLoS One 2013, 8(4):e61211.

[3] Kapushesky M, Adamusiak T, Burdett T, Culhane A, et al, *Gene Expression Atlas update–a value-added database of microarray and sequencing-based functional genomics experiments.* Nucleic Acids Research 2012, 40(Database issue): D1077-1081.

[4] Shendure J, *The beginning of the end for microarrays?* Nature Methods 2008, 5(7): 585-587.

[5] Audic S, Claverie JM, *The significance of digital gene expression profiles.* Genome Research 1997, 7(10): 986-995.

[6] Asmann YW, Klee EW, Thompson EA, Perez EA, et al, *3' tag digital gene expression profiling of human brain and universal reference RNA using Illumina Genome Analyzer.* BMC Genomics 2009, 10: 531.

[7] Mahanta P, Ahmed HA, Kalita JK, Bhattacharyya DK, *Discretization in gene expression data analysis: a selected survey.* Proceedings of the Second International Conference on Computational Science, Engineering and Information Technology 2012, 69-75.

[8] Li Y, Liu L, Bai X, Cai H, et al, *Comparative study of discretization methods of microarray data for inferring transcriptional regulatory networks.* BMC Bioinformatics 2010, 11:520.

[9] Li GJ, Ma Q, Tang HB, Paterson AH, Xu Y, *QUBIC: a qualitative biclustering algorithm for analyses of gene expression data.* Nucleic Acids Research 2009, 37(15): e101.

[10] Wang KJ, Wang BJ, Peng LQ, *CVAP: Validation for cluster analyses.* Data Science Journal 2009, 8: 88-93.

[11] Androulakis IP, Yang E, Almon RR, *Analysis of time-series gene expression data: methods, challenges, and opportunities* Annual Review of Biomedical Engineering 2007, 9:205C228

[12] 't Hoen PAC, Ariyurek Y, Thygesen HH, Vreugdenhil E, et al, *Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms.* Nucleic Acids Research 2008, 36: 141.

[13] Chen BZ, Gui F, Xie BG, Deng YJ, et al, *Composition and expression of genes encoding carbohydrate-active enzymes in the straw-degrading mushroom Volvariella volvacea.* PLoS One 2013, 8(3):e58780.

[14] Dougherty J, Kohavi R, Sahami M, *Supervised and unsupervised discretization of continuous features.* Proceedings of the Twelfth International Conference on Machine Learning 1995, 194-202.

[15] MacQueen JB, *Some methods for classification and analysis of multivariate observations.* In Proc of the fifth Berkeley Symposium on Mathematical Statistics and Probability 1967(1):281-297.

[16] Dudoit S, Fridlyand J, *A prediction-based resampling method for estimating the number of clusters in a dataset.* Genome Biology 2002, 3(7): 0036, 1-21.

[17] Dimitriadou E, Dolnicar S, Weingessel A, *An examination of indexes for determining the number of cluster in binary data sets.* Psychometrika 2002, 67(1): 137-160.

[18] Halkidi M, Batistakis Y, Vazirgiannis M, *On Clustering Validation Techniques.* Intelligent Information Systems Journal 2001, 17(2-3): 107-145.

978-1-4799-1389-3/13/$31.00 ©2013 IEEE

[19] Chen G, Jaradat SA, Banerjee N, Tanaka TS, et al, *Evaluation and comparison of clustering algorithms in anglyzing ES cell gene expression data.* Statistica Sinica 2002, 12: 241-262.

[20] Wang Z, Gerstein M, Snyder M, *RNA-Seq: a revolutionary tool for transcriptomics.* Nature Reviews Genetics 2009, 10(1):57-63.

978-1-4799-1389-3/13/$31.00 ©2013 IEEE