VOLUME-8, ISSUE-12, DECEMBER-2019 • PRINT ISSN No. 2277 - 8160 • DOI : 10.36106/gjra

Original Research Paper

Engineering

ANALYSIS OF DIFFERENTIAL GENE EXPRESSION LEVELS BASED ON RNA-SEQ

Wenyan Jiang Tiangong University, No. 399 Binshui Road, Xiqing District.

ABSTRACT In the research of life science, bioinformatics has become the main direction of current research, mainly in the aspects of gene expression analysis, genomics and so on. With the development of the RNA-Seq technique, we have made a breakthrough in the biotechnology, and on the basis of the expression level of the RNA-Seq data, the data is getting bigger and more and more and more, and the accuracy of the gene expression level is improved by the method of clustering. In this paper, four clustering algorithms are introduced, and the clustering algorithm with relatively good performance is compared.

KEYWORDS : RNA-Seq, clustering, K-mean, Louvain

1.INTRODUCTION

In recent years, low-cost, high-throughput, digital RNA-Seq technology has gradually replaced traditional gene chip technology, and its widespread application has played an important role in genomics and gene expression [1]. The RNA-Seq technique is mainly used for sequencing the cDNA fragments which are reversely transcribed by the mRNA, and a large number of reading segment data is obtained for researching the expression degree of the gene. With the development of sequencing technology, how to analyze these gene expression values and process RNA-Seq data is the content of this paper.

In biocomputing, genes will have different levels of gene expression in different situations [2]. The common regulation of genes will affect the traits of organisms, so cluster analysis can cluster genes into different classes according to different expression patterns. By clustering genes with similar functions into the same class, the unknown biological functions of the genes can be obtained. In this paper, the application of some clustering algorithms in RNA-Seq is introduced.

First, we need to understand the application of RNA-Seq technology. RNA-Seq, also known as full transcription shotgun sequencing, is the next generation sequencing technology to sequence the transcriptome of biological samples [3]. This technology extracts all the transcriptional RNA from the samples, and then through reverse transcription to cDNA for sequencing, so as to estimate the expression form and level of RNA, and understand and analyze the gene expression. The resulting data is getting larger and more diverse, so on this basis, the processing of data through data mining technology also came into being.

Clustering [4] is to divide the data set in the sample into different classes according to the internal connection, and make the same class have great similarity. There are four types of clustering algorithm: distributed, structured, density and graph structured. Distributed clustering [5] is a kind of clustering that can be generated at one time, and the representative algorithm is K-means clustering. Structural clustering uses aggregators that have been successfully classified, which can be calculated from top to bottom or from bottom to top, which represents hierarchical clustering. The density clustering algorithm is used for categories of arbitrary shape features, and these categories are considered as areas in the data set that are larger than a certain threshold. The graph structured clustering algorithm only returns one solution, which can reduce the running cost, and the representative algorithm is Louvain algorithm.

2. Algorithm research method

2.1 K-mean algorithm

K-mean algorithm is a relatively simple algorithm in clustering, which can be applied to many fields. Clustering cells with the K-mean algorithm is also a successful method. Given a set of data ($x_1, x_2, \ldots x_n$), divide the n data into k sets so that the variance within each class is minimized. The k initial points are randomly selected from the training data as the initial points of the class, the Euclidean distance from each point to the center point is calculated, and the mean value of the data points in each class is calculated as the new center point. If there is no change relative to the original center point or the change value is less than the threshold, then the algorithm ends, otherwise the algorithm is repeated. The key of K-mean algorithm is the choice of k value and initial point. The value range of k value is [-1, 1]. The larger the value of k, the better the result.

2.2 Hierarchical clustering algorithm method

Hierarchical clustering algorithm is a kind of structured clustering method, which constructs hierarchical clustering tree by calculating the similarity between different categories of data points. The creation of hierarchical clustering tree is divided into two ways: top-down splitting and bottom-up merging. Hierarchical clustering can be divided into two types: splitting and cohesion [6]. Splitting is a top-down idea. At the beginning, all data are regarded as a class until there is only one sample in the class. Cohesion is a bottom-up idea [7]. At the beginning, each sample is regarded as a different class, and the latest pair of classes are combined repeatedly until all the data belong to the same class. There are three methods for determining the distance by the agglomeration hierarchical clustering algorithm: Single Linkage, Complete Linkage, Average Linkage. The distance determined by the first method is the distance when the sample points in the two classes are closest. The distance determined by the second method is the distance when the sample points in the two classes are farthest, which may make the two similar classes unable to be combined because the extreme values are too far. The last method is to calculate the mean value of all distances between samples. Although this method has a large amount of calculation, the results are reasonable.

2.3 Graph structure clustering algorithm

Lougain [8] algorithm is a heuristic clustering algorithm based on optimization. The implementation of this algorithm mainly consists of two steps. Firstly, the nodes in the network are traversed continuously, and all the neighbors are traversed for each node. The income generated by adding neighbors to the node is calculated, and the neighbor who gets the maximum income is selected to join the group. If there is no income or the income is negative, the node remains in the group. Repeat the process until all the nodes no longer change. Secondly, the group formed in the first stage is folded, each combination is combined and the super node is taken as a super node to reconstruct the network, and the weight value among the new nodes is calculated. The two steps are iterated

until the algorithm ends.

Compared with other algorithms, this algorithm is relatively easy to implement, and the results are unsupervised. This algorithm is very fast and the complexity of typical sparse data is linear. After several iterations, the data will become less sharply, so the time of this algorithm is mainly consumed in the first iteration.

3.CONCLUSIONS

These traditional clustering methods are based on heuristic algorithm, which is difficult to compare the advantages and disadvantages of various algorithms. These algorithms can get better clustering results when applied to RNA-Seq data. These methods basically use Poisson distribution to model RNA-Seq data, but there are some differences between these data and Poisson model [9][10], and there is no specific principle to determine the optimal number of classes, so it is also a problem to be solved in the future to propose a better algorithm for clustering.

REFERENCES:

[1] Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics[J]. Nature reviews genetics, 2009, 10(1): 57.
[2] Pachter L. Models for transcript quantification from RNA-Seq[J]. arXiv preprint arXiv:1104.3889, 2011.
[3] Marioni J C, Mason C E, Mane S M, et al. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays[J]. Genome research, 2008, 18(9): 1509-1517.
[4] Tyagi A, Sharma S. Implementation Of ROCK Clustering Algorithm For The Optimization Of Query Searching Time[J]. International Journal on Computer Science and Engineering, 2012, 4(5): 809.
[5] Zhang T, Ramakrishnan R, Livny M. BIRCH: an efficient data clustering method for very large databases[C]//ACM Sigmod Record. ACM, 1996, 25(2): 103-114.
[6] Tamayo P, Slonim D, Mesirov J, et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation[J]. Proceedings of the National Academy of Sciences, 1999, 96(6): 2907-2912.
[7] Guha S, Rastogi R, Shim K. ROCK: A robust clustering algorithm for categorical attributes[J]. Information systems, 2000, 25(5): 345-366.
[8] Blondel V D, Guillaume J L, Lambiotte R, et al. Fast unfolding of communities in large networks[J]. Journal of statistical mechanics: theory and experiment, 2008, 2008(10): P10008.
[9] Witten D M. CLA SSIFICATION AND CLUST ERING OF SEQUENCING DATA USING A POISSON MODEL[J]. Annals of Applied Statistics, 2011, 5(4): 2493 -2518.
[10] Wang N, Wang Y, Han H, et a l. A b i-Poisson model for c lustering gene expression profiles by RNA-Seq[J]. Brie fings in Bio informatics, 2013, 15(4): 534-541.