Single-cell transcription group sequencing and the application of artificial intelligence in developmental biology.

Le Yang[1]

[1]Biology technology Westa college Westsouth university, Chongqing, 400715, China

Abstract. In the past two or three years, genome sequencing technology has been rapidly developed. Large-scale sequencing projects such as the Human Genome Project and the Cancer Genome Project have been launched one after another. Up to now, due to the emergence and research of artificial intelligence, it has brought us many possibilities. The purpose of this article is to use artificial intelligence to help single-cell transcription sequencing as much as possible. Based on the idea of Euclid algorithm, an improved K-means algorithm is proposed, which to a certain extent avoids the phenomenon of clustering results falling into local solutions, and reduces the appearance of the original K-means algorithm due to the use of error squares criterion function. In the case of dividing large clusters, the simulation experiment results show that the improved K-means algorithm is better than the original algorithm and has better stability.

1 Introduction

Due to the reprogramming of genome and epigenome and DNA replication errors during cell division and differentiation, different genomes, transcriptome and epigenome will appear at the single cell level [1]. The development and improvement of single cell transcriptome sequencing (scrnaseq) technology can be used to detect cell heterogeneity and transcriptome analysis of trace samples, so as to understand the complexity of eukaryotic transcriptome genes, single nucleotide polymorphism (SNP), copy number variation of single cell genome, genomic structure variation of single cell genome, gene expression level and gene fusion, Alternative splicing and DNA methylation status contribute to a comprehensive understanding of disease and life processes [2].

The first step of single cell sequencing (SCS) is to culture colonies or tissues infected by single cell isolated pathogenic microorganisms. The commonly used single cell separation techniques include microfluidic method, continuous dilution method, micromanipulation method, fluorescence activated cell sorting method and laser capture microdissection method [3]. This is a conventional method used before. Later, with the continuous improvement and development of the second and third generation sequencing technology, single cell transcriptome sequencing technology has been widely used in various fields such as tumor genome project and human genome project [4-5]. These sequencing projects use conventional transcriptome analysis methods to analyze the mixed samples of millions of cells. The results are the average values obtained by a large number of cell sequencing analysis, or reflect the main cell data, but ignore the differences in gene expression among heterogeneous single cells, which is not conducive to the tracking of cellular pathological process and the study of biodiversity [6].

Artificial intelligence is a branch of computer science [7]. It tries to understand the nature of intelligence and produces a new type of intelligent machine that can respond in a way similar to human intelligence [8]. Research in this field includes robotics, language recognition, image recognition, natural language processing and expert systems. Since the birth of artificial intelligence, the theory and technology are becoming more and more mature, and the application field is also expanding. It can be imagined that in the future, the science and technology products brought by artificial intelligence will become "containers" for human beings [9]. With the emergence and development of single cell sequencing technology, the research on the regulation and difference of biological development has been raised to the single cell level. With the help of artificial intelligence (AI) algorithm, the accuracy of single cell sequencing data analysis is improved, and the breakthrough understanding of individual development and disease occurrence is further deepened [10].

2 Algorithms are established and optimized

1.1. Machine learning algorithm -k-Means clustering analysis.

When the single cell transcriptome was classified, it was treated according to the situation. When only the index data of single cell transcriptome are collected, cluster analysis can be used to complete the classification. At the same time, the classification of single cell transcriptome can be determined, and the risk of single cell transcriptome can be studied by supervised methods.

The algorithm mainly calculates the similarity between each sample and centroid, so as to select the nearest centroid. There are many functions and methods to measure the similarity

For each class J, recalculate the heart:

$$U_j = \frac{\sum_{i=1}^{m} 1\{C^{(i)} = j\} X^{(i)}}{\sum_{i=1}^{m} 1\{C^{(i)} = j\}} \quad (1)$$

European distance is the most commonly used method of calculating vector distance in mathematics, and the mathematical formula for two-dimensional spatial Euthic distance is:

$$\mathrm{d} = \sqrt{(\mathrm{x}_1 - \mathrm{x}_2)^2 + (\mathrm{y}_1 - \mathrm{y}_2)^2} \quad (2)$$

Where (x_1, y_1) and (x_2, y_2) are two points in space.

Manhattan distance is also known as city block distance, in two-dimensional space, the Manhattan distance between two points formula is:

$$\mathrm{d} = |x_1 - x_2| + |y_1 - y_2| \quad (3)$$

During K-means calculation, the first step is to set the K-value, which is the number of clusters in the final dataset. Second, the center point of the initial data cluster is randomly generated. For the initial mass, the algorithm will traverse for the first time, classify all the data for the first time, and select the k-class data point closest to each heart in the K-quality heart as the clustering result for the first iteration. Then, define a new heart point for the next cluster. The new center of mass is the central point of each category in the previous clustering results. With the new heart points, clustering can be repeated, and new clustering can be calculated using the new heart, which can then be iterative.

3 Model building

3.1 Cosine Association Analysis Algorithm.

Cosine similarity, also known as cosine similarity, is assessed by calculating the angle cosine values of the two vectors. Suppose a and b are two different vectors, and the rest of the string similarity is calculated as:

$$\cos\theta = \frac{a \cdot b}{|a||b|} \quad (4)$$

If the coordinates of a-b are (x1, Y1), (x2, y2), then the algorithm formula can be rewritten to:

$$\cos\theta = \frac{x_1 x_2 + y_1 y_2}{\sqrt{x_1^2 + y_1^2}\sqrt{x_2^2 + y_2^2}} \quad (5)$$

Using the cosine association algorithm to represent the correlation of single-cell transcription groups, the algorithm formula can be rewritten to: .

$$\gamma(C_i, C_j) = \frac{C_i \cdot C_j}{|C_i||C_j|} \quad (6)$$

Assuming a single-cell transcription $\mathrm{C_i}$ group, $\mathrm{C_j}$expressed as points on the coordinate system, can be further derived from: .

$$\gamma(C_i, C_j) = \frac{\sum_{k=1}^{k} c_i^k c_j^k}{\sqrt{\sum_{k=1}^{k} (c_i^k)^2}\sqrt{\sum_{k=1}^{k} (c_j^k)^2}} \quad (7)$$

- The model describes the correlation between the sequencing of its single-cell transcription group in developmental biology.
- Then through the algorithm to simulate its reasonable application, modeling came to a conclusion.

3.2 Selection of subjects.

Before this data mining experiment, it is very important to use the data set used in the test algorithm, and different algorithms also need different data sets to support it, so as to get better classification accuracy. Therefore, the data sources used in this paper are two sets of single cell transcriptome in the laboratory. In order to get more effective data, the download data should be preprocessed first, and the suitable data preprocessing method can greatly improve the accuracy of data mining. So we all preprocess it and analyze the results. The experimental results are mentioned below.

4 Evaluation results

4.1 Effects of various algorithms on the application of single-cell transcription groups in developmental biology.

Table 1. Single-cell transcription group 1 predicts classification results.

Single-cell transcription group 1	Euclid	Cosine classification	Improved K-means
The accuracy of the training dataset	70.12%	85.11%	97.27
Check accuracy	83%	97.87%	100%
Standard error	0.1631	0.0122	0.0012

Table 2. Single-cell transcription group 2 predicts classification results.

Single-cell transcription group 2 .	Euclid.	Cosine classification	Improved K-means
The accuracy of the training dataset	74.27%	86.39%	94.17
Check accuracy	83.17%	98.21%	99.35%
Standard error	0.1693	0.1025	0.0461

According to the data described in Table 1, Table 2, the results of the analysis of single-cell transcriptome 1 and 2 prediction are that according to the data of all the above students, after different data preprocessing operations, the accuracy of Euclid's algorithm The lowest, the SMOTE processing of the data when applying cosine classification, and the calculation of the optimal parameters at the same time, the results obtained are satisfactory but not stable, and the improved K-means combination obtains a better prediction accuracy rate in predicting the current data set And more stable.

The current relatively novel classification algorithm is the most commonly used method in data mining, which is similar to the discrimination mentioned in the classic diversification statistics. In classification problems, the dependent variable is generally a categorical variable. If the classification problem to be solved does not meet this condition, the continuous variable needs to be discretized into a categorical variable. Under normal circumstances, discriminant analysis can directly solve common classification problems, but if the independent variables contain more categorical variables, then discriminant analysis is no longer applicable. We can try some methods in data mining to solve this classification problem.

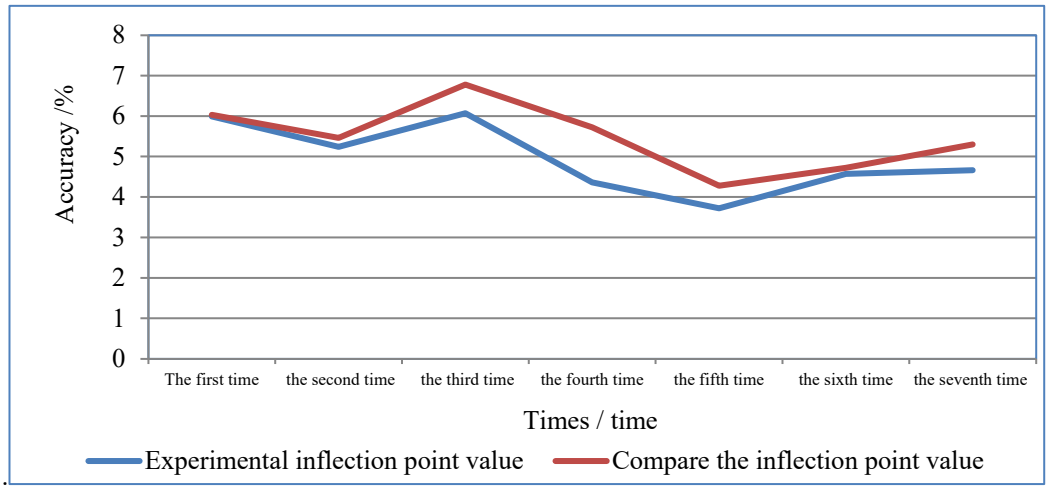

Figure 1. Inflection point graph based on the prediction accuracy of the Euclid algorithm.

According to the data of the inflection point diagram of the improved K-means prediction accuracy rate according to Figure 1, in the early stage of the current algorithm construction, the algorithm is not stable in the prediction of single-cell transcriptome sequencing, but according to the above seven experiments, we The prediction results of each experiment are compared and analyzed with the actual results respectively. The purpose is to improve the efficiency of the current single-cell transcriptome sequencing of the prediction model, improve the accuracy of the prediction data, and make it more accurate.

4.2 Analysis of algorithms used for single-cell transcriptome sequencing

From a biological point of view, cells are heterogeneous. A cell group usually contains different cell

subpopulations, such as neurons and glial cells collected from brain samples; in addition, different states of the same cell type, such as stimulated and unstimulated T cells, can also be seen. From a mathematical point of view, de novo recognition of cell groups is an unsupervised clustering problem. At present, several mature schemes have been applied to single cell. The possibility of dividing a large number of cells into k groups is unimaginable, so we can not consider all the possible clustering cases, but should seek the optimal solution. The quality of clustering depends on the similarity comparison between intra group and inter group. Different indicators make different assumptions on the basic segments of data. K-means is a common clustering algorithm for single cell analysis, which is usually used after feature selection and dimension reduction. It computes faster by iteratively assigning cells to the nearest cluster center (or "centroid") and recalculating the centroid of the cluster. However, K-means needs to specify the number of clusters in advance and provide a random starting position for each cluster. It needs to run several times to check the robustness of these parameters. These results can be passed to SC3 for combination. One disadvantage of K-means is that it first assumes a predetermined number of circles of equal size. If the hypothesis is not met, then K-means will recognize many adjacent clusters along the differentiation trajectory, merging rare cells with common cell types. Therefore, we need to carry on the restriction analysis of other algorithms, so we introduce Euclidean algorithm, and finally get the better data as shown in Figure 1.

5 Conclusions.

To sum up, single cell is small, easy to destroy, sensitive to the external environment, and the internal and external components of the battery are unstable. Although it is difficult to separate, many kinds of single cell separation techniques can meet the requirements of isolation and extraction of different types of cells, and obtain the complete transcription expression profile at the single cell level, so as to obtain the ideal single cell to the maximum extent. It can deeply analyze the heterogeneity and gene expression network between different single cells; high-throughput, efficient and low-cost high-throughput sequencing technology can directly sequence RNA, discover new transcripts, and directly sequence RNA, so as to obtain more accurate results. To sum up, the wide application of sequencing technology also promotes the improvement and update of sequencing methods and data analysis algorithms. The improvement of sequencing flux is a typical example of increasing application demand and promoting the development of sequencing technology. With the development of AI and interdisciplinary integration, the update of single cell sequencing technology and data analysis methods will help to make more breakthroughs in developmental biology. With the continuous development of single cell transcription sequencing technology and artificial intelligence and the continuous improvement of various technical links, a great breakthrough has been made in developmental biology. Finally, great achievements will be made in the field of medical and health care, which will become an indispensable technology for people to treat diseases and explore life sciences.

References

1. Gao S, Yan L, Wang R, et al. 2018, Tracing the temporal-spatial transcriptome landscapes of the human fetal digestive tract using-single cell RNA-sequencing. Nature Cell Biology, 20 (6): 721.
2. Li S, Sun J, Allesøe R, et al. 2019, RNase H-dependent PCR-enabled T-cell receptor sequencing for highlyly specific and efficient target sequencing for T-cell receptor mRNA for single-cell and repert analysiss. Nature Protocols, 14 (8): 2571.
3. Zhang G L, Pan L L, Huang T, et al. 2019, the transcriptome difference between colorectal tumor and normal tissues revealed by single-cell sequencing. Journal of Cancer, 10 (23): 5883-5890.
4. Johnson H E, Toettcher J E .2018, Illuminating developmental biology with cellular optogenetics. Current Opinion in Biotechnology, 52:42-48.
5. Ma G, Wang T, Korhonen P K, et al. 2020, Elucidating the molecular and developmental biology of parasitic nematodes: Move to a multiomics paradigm. Advances in Parasitology, 108:175-229.
6. Corfitsen H T, Drago A. 2020, Enriched developmental biology molecular pathways on impact antipsychotics-induced weight gain. Pharmacogenetics and Genomics, 30 (1): 9-20.
7. Hutson M. 2018, An intelligence faces reproducibility crisis. ence, 359 (6377): 725-726.
8. Liu R, Yang B, Zio E, et al. 2018, An intelligence for fault diagnosis of rotating machinery: A review. Mechanical Systems and Signal Processing, 108 (AUG.) :33-47.
9. Yan Y, Zhu J. 2019, Sali detection based on superpixel correlation and cosine window filtering. Multimedia Tools and Applications, 78 (15): 21205-21221.
10. Jin M, Liu W, Xing W. 2019, A Robust Visual Tracker Based on DCF Algorithm. International journal of software engineering and knowledge engineering, 29 (11/12): 1819-1834.