RESEARCH ARTICLE

Application of Dimension Reduction Methods to High-Dimensional Single-Cell 3D Genomic Contact Data

Zilin Wang[1,*], Ping Zhang[1,2], Weicheng Sun[1] and Dongxu Li[2]

[1] College of Informatics, Huazhong Agricultural University, Wuhan 430070, China
[2] School of Computer, BaoJi University of Arts and Sciences, Baoji 721016, China

Abstract

The volume and complexity of data in various fields, particularly in biology, are increasing exponentially, posing a challenge to existing analytical methods, which often struggle with high-dimensional data such as single-cell Hi-C data. To address this issue, we employ unsupervised methods, specifically Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE), to reduce data dimensions for visualization. Furthermore, we assess the information retention of the decomposed components using a Linear Discriminant Analysis (LDA) classifier model. Our findings indicate that these dimensionality reduction techniques effectively capture and present information not readily apparent in the original high-dimensional data, facilitating the visualization and interpretation of complex biological data. The LDA classifier's performance suggests that PCA and t-SNE maintain critical information necessary for accurate classification. In conclusion, our study demonstrates that PCA and t-SNE are powerful tools for visualizing and analyzing high-dimensional biological data, enabling researchers to gain new insights and understandings that are challenging to achieve with traditional approaches.

Academic Editor:
Jinchao Chen

Submitted: 13 December 2023
Accepted: 28 January 2024
Published: 12 February 2024

Vol. 2, **No.** 1, 2024.
10.62762/TIOT.2024.186430

***Corresponding author:**
✉ Zilin Wang
WZLin3288@finmail.com

Keywords: Dimensionality reduction, Single-cell Hi-C, PCA, t-SNE, LDA.

Citation
Wang, Z., Zhang, P., Sun, W., & Li, D. (2024). Application of Dimension Reduction Methods to High-Dimensional Single-Cell 3D Genomic Contact Data. IECE Transactions on Pattern Recognition and Intelligent Systems, 2(1), 20–25.

1 Introduction

Dimensional disaster is a widely faced problem in data science realm. In the field of biology, there are a lot of high-dimensional data with large samples that are difficult to be processed by ordinary methods. The dimension reduction method using machine learning can enable us to quickly find the variation and features among biological samples among the biology data mountain.

Dimension reduction is a widely used feature extraction method especially in computer vision. There are multiple applications on extracting low-dimensional features from images, such as Eigen Face, Fisher Face and handwritten dimension reduction as well as other feature selection methods based on support vector machine or deep learning. Hence, we decide to apply those method on biology data.

The contact matrix of single-cell Hi-C data can regard as image, which contains three-dimensional structure information of the chromosome (Fig. 1). The positions of the horizontal and vertical axes of the matrix represent the two contact sites on the chromosome, and the value of the matrix represents the interaction frequency of the two sites. The interaction matrix can describe the three-dimensional structure of cell chromosome to some extent.

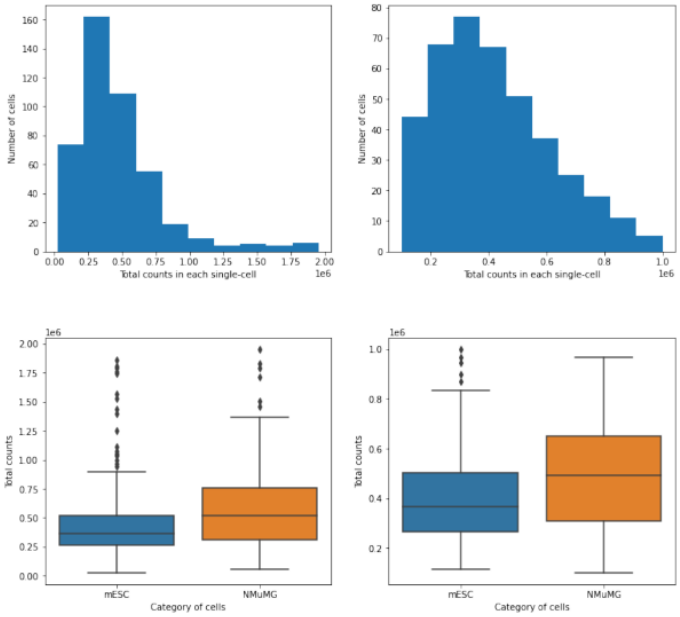

Figure 1. The quality control of single-cell Hi-C data. The left two figures show the total counts of each single-cell before quality control, the right two show after quality control.

Comparing with the face image, the sample amount and categories of single-cell data is fewer, while each single-cell data contains more information. In additional, Hi-C data is apparently different from images such as human faces and handwritten characters. And it is difficult for human eyes to distinguish obvious features between single-cell Hi-C matrices. Therefore, there is a great demand for machine learning algorithm [1] to classify and recognize the variation between those biology samples. Otherwise, the interaction matrix of single-cell data can reach very high dimensions depending on the resolution we choose, so it is necessary to apply dimension reduce function in primary step when start to analysis scHIC data.

2 Related Work

At present, single-cell sequencing technology is becoming more and more mature, and more high-dimensional single-cell sequencing data are applied in data analysis, such as single-cell RNA-seq, ATAC-seq, Chip-seq, etc. Since Nagano pioneered single-cell HIC data in 2013 [1], single-cell Hi-C data had become more and more common.

Due to the high throughput Hi-C data and 2D contact matrix of Hi-C data, variety of transformations methods for measuring the quality of and reproducibility of Hi-C experiments were developed previously. HiCRep [2], GenomeDISCO [3], HiC-Spector [4], and QuASAR-Rep [5] measure reproducibility and compute pairwise similarities between Hi-C matrices. And those methods can also use in single-cell data to measure the contact matrix.

As for the single-cell data, Ren's group use SnapHi-C identify loop domain and intercellular variation in scHIC data [6]. Richard apply Explicit-PCA reveal the dominant motion of genome 3D structure [7]. Zhou use scHiCluster on different types of cells and find out domain-like structures (TLSs) in single-cell data [8]. Lu and Feng apply unsupervised embedding method handle scHiC data and distinguish them according cell cycle [9]. Michael use Bayesian estimation and priori information in bulk Hi-C to infer 3D structures of chromosomes from single- cell Hi-C [1].

3 Methodology

3.1 Data Preparation

Single-cell Hi-C data used in this study were derived from sequencing data of mouse embryonic stem cells (mESC) and normal epithelia mouse mammary gland cells (NMUMG) [11]. According to the previous research, single-cell Hi-C data are inherently variable, and there are differences between different cells and different cell states, and between different types of cells [2]. In addition, due to technical limitations [17–20], the current single-cell HI-C library construction method can only capture less than 1% of the interactions in a single cell. Therefore, due to the sparsity of data and the randomness of sampling results, there are extensive noises and a large number of missing in scHIC data, so we need to carry out quality control, noise reduction and smoothing processing on the interaction matrix.

In this experiment, we first apply the quality control on mESC and NMUMG cells. And descriptive statistical analysis was carried out to eliminate cells with too little or too much interactive or with other obvious problems. Gaussian blur and 2D Mean Filter are used to smooth the matrix to solve the sparsity problem. In order to reduce the noise caused by the experiment, we try HiCNorm, ICE, Knight-Ruiz Matrix-Balancing algorithms [12–14] which is commonly used in Hi-C matrix normalization.

As for the resolution, we use the 1 MB to initialize the matrix due to the sparse, which means each bin of the contact matrix contains counts within 1 MB

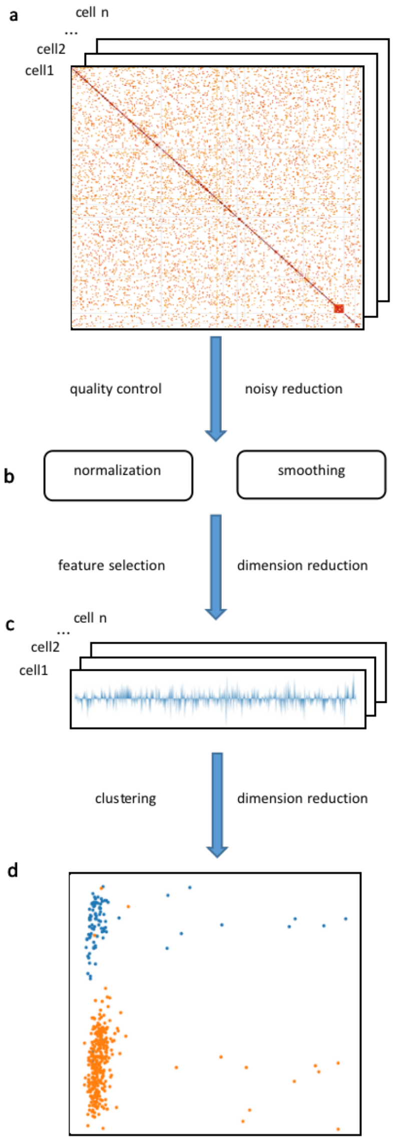

Figure 2. Workflow of single-cell Hi-C data dimension reduction. (a) the single-cell Hi-C contact matrix on one chromosome, horizontal and vertical axis represent the contact loci, (b) normalized and smoothed convert matrix, (c) cell embedding vectors, (d) clustering image in low dimension.

scope. After data Preparation, we attain a 2735×2735 dimension global converted contact matrix for each single-cell. Fig. 2(a) shows local feature of raw contact matrix.

3.2 Matrix Embedding

Since two-dimensional matrix data cannot be directly used for data analysis, many scholars have proposed great ideas on embedding of Hi-C matrix, such as Yang's HiCRep [2], Ursu's GenomeDISCO [3] and Yan's Hic-Spector [4]. These methods provided some algorithm to process Hi-C contact matrix based on biology and statistic principle, which can measure the similarity of Hi-C contact matrix in reasonable ways and accomplish dimensional reduction [16].

In this paper, we focus on the machine learning dimension-reduction methods, so we decide to reference the processing technique in image data, that is, transform the matrix into vector format or use unsupervised dimension reduction algorithm to carry out subsequent analysis.

3.3 Dimensionality Reduction

After the preprocess, now we have processed data sample set $D = \{(X_1, Y_1), (X_2, Y_2), \ldots, (x_m, y_m)\}$, where any sample x_i is the vector representation of the sample, and the sample label $y_i \in \{0, 1\}$, where label 0 represents mESC cells and label 1 represents NMuMG cells. At the resolution of 1 MB, the dimension $n \approx 7,000,000$ for each X_i, and increased by a factor of N^2 for each n-fold increase in resolution. We will use the following dimensional-reduction algorithms to project the data to low dimensional.

3.3.1 PCA

Principal Component Analysis (PCA) is one of the most commonly used methods in unsupervised feature selection. It does not need to use labels and retain the internal characteristics of the samples.

For a sample $N = (x_1, x_2, \ldots, x_m)$ that you want to map to a lower dimensional space, μ is the mean value of the vector, and we need to project these samples onto the hyperplane of dimension d using PCA. An orthogonal matrix W of $n \times d$ is needed, and each sample after projection $y_i = W^T x_i$. This matrix is an orthonormal matrix according to the properties of PCA $W = \{w_1, w_2, \ldots, w_d\}$. According to the principle of PCA, to make the projection vectors have the maximum variance or minimum square error, we obtain the equation:

$$\arg\max L(W) = \sum_{i=1}^{m} W^T x_i x_i^T W$$
$$\text{s.t. } W^T W = I$$

We can calculate the projection matrix W for the first d largest eigenvector of the $n \times n$ matrix XX^T components, then according to the projection formula $y_i = W^T x_i$ each sample can be projected to the new d dimensional space.

3.3.2 t-SNE

The next method is t-distributed Stochastic Neighbor Embedding (t-SNE). Both t-SNE and PCA are unsupervised dimension reduction methods, which do not need sample labels. And t-SNE is also a commonly used method for dimension-reducing clustering, which is further developed based on LLE and SNE.

Its principle is using affine transformation to map data to a probability distribution in low dimension. The distance within samples in high-dimensional space transformed into a conditional probability to represent the similarity between point and point, then reconstruct the probability distribution of these points in low dimensional space. make the probability distribution between the low and high dimension as close as possible. In other words, the most similar points are clustered together and the least similar points are moved away.

In order to make the projection vector closer to the original distribution and retain the distance and local characteristics, the solution of t-SNE involves using t distribution as a probability distribution function, KL divergence of before and after projection as the loss function and updating parameters by means of gradient descent. After a certain number of iterations or achieve a certain indicator, we can obtain a character representation in specified dimension.

3.3.3 LDA

Linear Discriminant Analysis (LDA) is a supervised dimension reduction method and can also be used for classification. The principle of LDA is similar to PCA, which is to find an n×d projection matrix W to project high-dimensional data. However, the purpose of LDA is that the projection points of each category of data should be as close as possible, while the distance between the data center points of different categories should be as far as possible.

In the binary classification problem, the samples can only reduce to 1 dimension. So we need to find a vector rather than a matrix to project the samples onto a linear space. The projected formular is $y = w^T x$, that is, for any sample x_i, its projection on the line is $w^T x_i$. The projection center of the two types of samples $(j = 0, 1)$ is μ_j, and the covariance matrix of the two types of samples is Σ_j. At this time, we can define the intra-class divergence matrix S_w and the inter-class divergence matrix S_b in the case of dichotomization problem:

$$S_w = \Sigma_0 + \Sigma_1$$

$$S_b = (\mu_0 - \mu_1)(\mu_0 - \mu_1)^T$$

Then we can define the loss function below, and use maximum likelihood estimation to calculate the project vector.

$$\arg\max J(w) = \frac{w^T S_b w}{w^T S_w w}$$

In the binary classification problem, LDA calculates the matrix S_w and the matrix S_b, and the final solution projection vector is the maximum eigenvector of matrix $S_w^{-1} S_b$. Once we obtain the vector, we can use it for projection and classification.

As for the multiple classification and decomposition problem, we need to calculate the corresponding $S_w^{-1} S_b$ matrix and its first several maximum eigenvalues and eigenvectors instead of the projection vector.

3.4 Cell Classification

In We use PCA or t-SNE on single-cell reduce the data to low dimensions then use LDA for training. The projection vector is reserved as LDA classifier to identify the new input samples. We use the LDA classifier and CV to measure the embedding result.

4 Experiments

In the dimension reduction problem, we try to use PCA and t-SNE reduce the single-cell Hi-C data to low dimension. And we supposed to the variance between single-cell. As we expected, we discover that using the principal components (PC) can distinguish the two cell types directly. The PC1 and PC3 explain 57 percent and 6 percent variance, respectively. And they were finely separated into two parts according their cell types. So did the 2d t-SNE result demonstrate the separation of two cell types.

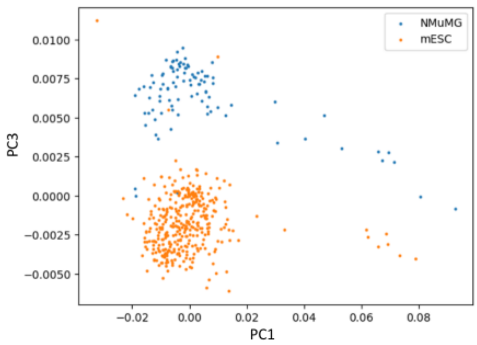

Figure 3. PCA after embedding.

As for the cell type classification problem, we first apply PCA on each contact matrix and remain several principal components, then we using another LDA on each PC to find out the first-class represent on each sample. We preformed 5-fold CV and found

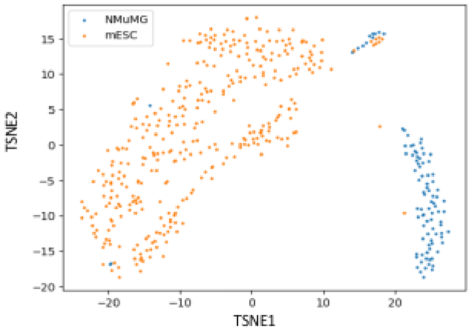

Figure 4. t-SNE after embedding.

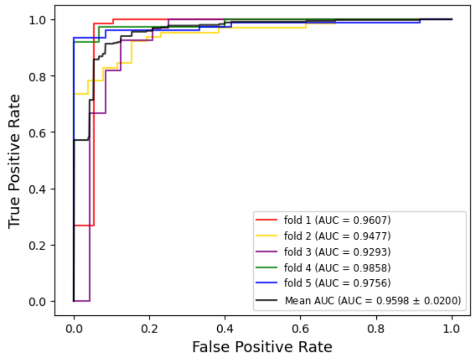

Figure 5. The ROC curve of PC1 for 5-fold CV.

PC1 achieve the highest AUC (average 0.95) in all the principal components.

Table 1. 5-fold CV of first 5 principals.

Principal Components	AUC (%)	ACC (%)
PC1	**95.98**	**92.12**
PC2	89.59	88.60
PC3	89.07	88.35
PC4	83.94	81.35
PC5	82.54	80.60

5 Conclusion

Our results show the difference between cell types and distinguish them in a decent method. However, deeper analysis needs to take more factors into consideration. And there are also many noises in Hi-C data, such as random ligation noise and genomic distance noise. We only consider the data structure and not too much biological feature. In other area, there are sharply increased large amount of high-dimensional complicated data like single cell Hi-C contact matrix, which are hard to analysis them directly. Therefore, some combined machine learning methods need to be used in this field to help people figure out the difference or similarity between those samples. In our further research we can discuss multiple classification problem basing on more cell types, pattern recognition on more specific genome structure.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgement

This work was supported by Name of funding agency under Grant 16JK1048.

References

[1] Rosenthal, M., Bryner, D., Huffer, F., Evans, S., Srivastava, A., & Neretti, N. (2019). Bayesian estimation of three-dimensional chromosomal structure from single-cell Hi-C Data. *Journal of Computational Biology*, 26(11), 1191–1202. [CrossRef]

[2] Yang, T., Zhang, F., Yardımci, G. G., Song, F., Hardison, R. C., Noble, W. S., Yue, F., & Li, Q. (2017). HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. *Genome Research*, 27(11), 1939–1949. [CrossRef]

[3] Ursu, O., Boley, N., Taranova, M., Wang, Y. R., Yardimci, G. G., Stafford Noble, W., & Kundaje, A. (2018). GenomeDISCO: a concordance score for chromosome conformation capture experiments using random walks on contact map graphs. *Bioinformatics*, 34(16), 2701-2707. [CrossRef]

[4] Yan, K. K., Yardımcı, G. G., Yan, C., Noble, W. S., & Gerstein, M. (2017). HiC-spector: a matrix library for spectral and reproducibility analysis of Hi-C contact maps. Bioinformatics, 33(14), 2199-2201. [CrossRef]

[5] Sauria, M. E., & Taylor, J. (2017). QuASAR: quality assessment of spatial arrangement reproducibility in Hi-C data. *BioRxiv*, 204438. [CrossRef]

[6] Yu, M., Abnousi, A., Zhang, Y., Li, G., Lee, L., Chen, Z., ... & Hu, M. (2020). Snaphic: a computational pipeline to map chromatin contacts from single cell hi-c data. *BioRxiv*, 2020-12.[CrossRef]

[7] Lindsay, R. J., Pham, B., Shen, T., & McCord, R. P. (2018). Characterizing the 3D structure and dynamics of chromosomes and proteins in a common contact matrix framework. *Nucleic acids research*, 46(16), 8143-8152. [CrossRef]

[8] Zhou, J., Ma, J., Chen, Y., Cheng, C., Bao, B., Peng, J., ... & Ecker, J. R. (2019). Robust single-cell Hi-C clustering by convolution-and random-walk–based imputation. *Proceedings of the National Academy of Sciences*, 116(28), 14011-14018. [CrossRef]

[9] Liu, J., Lin, D., Yardımcı, G. G., & Noble, W. S. (2018). Unsupervised embedding of single-cell Hi-C data. *Bioinformatics*, 34(13), i96-i104. [CrossRef]

[10] Lee, D. S., Luo, C., Zhou, J., Chandran, S., Rivkin, A., Bartlett, A., ... & Ecker, J. R. (2019).

Simultaneous profiling of 3D genome structure and DNA methylation in single human cells. *Nature methods*, 16(10), 999-1006.[CrossRef]

[11] Imakaev, M., Fudenberg, G., McCord, R. P., Naumova, N., Goloborodko, A., Lajoie, B. R., ... & Mirny, L. A. (2012). Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature methods*, 9(10), 999-1003. [CrossRef]

[12] Hu, M., Deng, K., Selvaraj, S., Qin, Z., Ren, B., & Liu, J. S. (2012). HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics*, 28(23), 3131-3133. [CrossRef]

[13] Knight, P. A., & Ruiz, D. (2013). A fast algorithm for matrix balancing. *IMA Journal of Numerical Analysis*, 33(3), 1029-1047.[CrossRef]

[14] Y. Hua & X. Wang (2023). Forest Fire Assessment and Analysisin Liangshan, Sichuan Province Based on Remote Sensing. *IECE Transactions on Internet of Things*, 1(1), 15-21. [CrossRef]

[15] Yardımcı, G. G., Ozadam, H., Sauria, M. E., Ursu, O., Yan, K. K., Yang, T., ... & Noble, W. S. (2019). Measuring the reproducibility and quality of Hi-C data. *Genome biology*, 20, 1-19. [CrossRef]

[16] Li, Y., & Cao, J. (2023). Adaptive Binary Particle Swarm Optimization for WSN Node Optimal Deployment Algorithm. *IECE Transactions on Internet of Things*, 1(1), 1-8. [CrossRef]

[17] Wang, N., Fang, F., & Feng, M. (2014, May). Multi-objective optimal analysis of comfort and energy management for intelligent buildings. In *The 26th Chinese control and decision conference (2014 CCDC)* (pp. 2783-2788). IEEE.

[18] Lv, Y., Fang, F. A. N. G., Yang, T., & Romero, C. E. (2020). An early fault detection method for induced draft fans based on MSET with informative memory matrix selection. *ISA transactions*, 102, 325-334. [CrossRef]

[19] Fang, F. A. N. G., Tan, W., & Liu, J. Z. (2005). Tuning of coordinated controllers for boiler-turbine units. *Acta Automatica Sinica*, 31(2), 291-296.

[20] Fang, F., Jizhen, L., & Wen, T. (2004). Nonlinear internal model control for the boiler-turbine coordinate systems of power unit. *PROCEEDINGS-CHINESE SOCIETY OF ELECTRICAL ENGINEERING*, 24(4), 195-199.

Zilin Wang Currently studying in the first year of BS in Huazhong Agricultural University. The research direction is bioinformatics.

Ping Zhang Currently studying in the first year of PhD in Huazhong Agricultural University. He is a lecturer in Baoji University of Arts and Sciences. His current research interests include bioinformatics, machine learning and graph neural network.

Weicheng Sun Currently studying in the first year of BS in Huazhong Agricultural University. The research direction is bioinformatics, machine learning and graph neural network.

Dongxu Li Graduated with B.S at the Department of Computer of Baoji University of Arts and Sciences from 2018 to 2022. His current research interests include machine learning and computer vision.