Comparative Research of Different Dimension Reduction Methods Combined with RWR Network Smoothing in Single Cell RNA-seq Data

Xuesong Xiao, Pengchao Ye, Wenbin Ye and Guoli Ji*
Department of Automation, Xiamen University, Xiamen, Fujian 361005, China
*Corresponding author e-mail: glji@xmu.edu.cn

Abstract. Single-cell RNA sequencing (scRNA-seq) has been an important inspiration for the study of biomolecules through its reveal of cell heterogeneity. However, due to the low capture efficiency and frequent drop-out events in the single-cell sequencing process, the scRNA-seq data often has high sparsity and random missing values, which brings great difficulties to the subsequent analysis. The network propagation method based on random walk with restart (RWR) effectively fills in the missing values in the scRNA-seq data and reduces noise by referring to the prior information of gene interaction. Dimensionality reduction is also a commonly used pre-processing method for high-dimensional and sparse scRNA-seq data, which can be combined with the RWR-based data imputation to achieve noise reduction and feature extraction of scRNA-seq data. This article compares the performance of the commonly used single-cell data dimension reduction methods combined with the RWR network smoothing in different type of scRNA-seq data sets, and analyzes their applicability and stability.

1. Introduction
Single-cell RNA-seq sequencing makes it possible to study the gene expression information of a single cell, and brings new ideas to the fields of zoology, botany, human immunology, and molecular biology. Traditional sequencing methods show average information at the multi-cell level, while sequencing at the single-cell level can completely reflect the genome and transcriptome status of different cells in the same cell population [1]. However, due to technical factors, interferences such as data loss and noise often appear in scRNA-seq data, which brings great difficulties to its downstream analysis [2]. The application of random walks in network propagation makes it possible to solve this problem, which is used to realize the ranking and correlation analysis of important nodes in the network. Random walk with restart (RWR) optimizes the iterative process of the algorithm by adding probability parameters, which makes it not likely for the random walk to fall into a local optimum and improve its performance [3]. Based on RWR, the missing genes in scRNA-seq data can be imputed using the inter-gene association information contained in the gene interaction network [4]. Our previous research [5] has shown that RWR achieved outstanding performance in scRNA-seq network smoothing compared with other popular imputation methods (scImpute [6], SAVER [7], MAGIC [8]). However, the scRNA-seq data itself has high dimension and sparseness, so it is often necessary to perform dimension reduction on subsequent analysis. Based on the principles and scope of the currently used dimension reduction methods, we compared the dimension reduction effects of PCA [9], t-SNE [10], Isomap [11], SCRL [12] on diverse scRNA-seq datasets imputed by RWR. By comparing
the visualization and clustering indicators, we analyze the applicability and stability of the scRNA-seq data imputation based on RWR to the dimension reduction methods.

2. ScRNA-seq data imputation based on RWR
   The gene interaction network contains information of function associations between genes. For a gene interaction network, its adjacency matrix $A$ shows the score value of the correlation between gene nodes. $D$ is a diagonal matrix to ensure convergence, and $\alpha$ is the restart probability, which determines the probability that each iteration process starts from the initial value [3]. For each cell sample in the scRNA-seq expression matrix, $p_0$ is a vector composed of the original gene expression level of the cell in the scRNA-seq matrix, and the random walking process is performed according to the following recursive relationship:

$$ p_t = \alpha p_0 + (1 - \alpha)AD^{-1}p_{t-1} $$  \hspace{1cm} (1)

For each iteration $t$, the random walk may start with a gene in the initial vector, and there is a certain possibility to walk to the neighbor nodes of the t-1 gene. The updated expression value in the walk depends on the current weight of the gene and the sum of the weights of the gene and all neighboring nodes at t-1 times:

$$ p_k(v) = \sum_{u \in N(v)} p_{k-1}(u)w(u,v) $$  \hspace{1cm} (2)

When the iteration is completed, $p$ is a stable and smooth vector that can reflect the global topology of the gene network.

3. Evaluation of dimension reduction performance with RWR network smoothing
   We have summarized the dimension reduction methods commonly used in data analysis situations, and selected representative ones from different modeling ideas for performance comparison.

3.1. Dimension reduction method
   Based on the modeling principles of different dimension reduction methods, we have selected representative ones including PCA in linear analysis, t-SNE in non-linear mapping, Isomap in manifold learning, and SCRL in network embedding. These methods are used to reduce the dimension of the scRNA-seq matrix network-smoothed by RWR, and the performance analysis is completed based on the experimental results.

3.1.1. PCA
   Principal Components Analysis (PCA) is a linear multivariate statistical method [9]. Its main idea is to construct a set of orthogonal features that may have a correlation, that is, principal components, so as to map high-dimensional data to low-dimensional space.

3.1.2. T-SNE
   T-distributed Stochastic Neighbor Embedding (t-SNE) converts the similarity between samples into conditional probability [10]. The similarity of the data in the original space is represented by the Gaussian joint distribution, and the similarity in the embedded space is represented by the student t-distribution.

3.1.3. Isomap
   Isomap (Isometric Feature Mapping) is a kind of manifold learning, an unsupervised algorithm for dimension reduction of non-linear data [11]. Isomap creates a map by connecting each sample to its nearest neighbours, then reduces the dimension while trying to preserve the geodesic distance between the samples.

3.1.4. SCRL
   SCRL (Single Cell Representation Learning) performs representation learning on scRNA-seq data based on a large-scale network embedding model [12]. SCRL constructs a cell-gene bipartite network, uses the joint distribution probability of low-dimensional space to fit the empirical distribution of the original space, and constructs an objective function through KL divergence and optimizes it:

$$ O_{ca} = -\sum_{(i,j) \in E_{ca}} y_{ij} \log p_j(a_j|c_i) $$  \hspace{1cm} (3)
3.2. Data and Experiment

We selected three scRNA-seq datasets obtained from different tissues of mice and humans by different sequencing technologies. The detailed information of the datasets is shown in Table 1.

| Dataset       | Species | Gene | Cell | Cell type | Expression Type |
|---------------|---------|------|------|-----------|-----------------|
| Usoskin [13]  | Mouse   | 25334| 622  | 4         | RPM             |
| Li [14]       | Human   | 55182| 561  | 7         | FPKM            |
| Patel [15]    | Human   | 5948 | 430  | 5         | TPM             |

All the three scRNA-seq datasets are smoothed by RWR using a fusion network containing 7,055,494 interactions between 19,404 genes. The network is constructed based on information entropy and combines four networks: HIPPIE, HumnaNet, FunCoup, and STRING to complement the information of multiple gene relationships [16]. We compared the effects of different algorithms on visualization, internal indicator DBI (Davies-Bouldin index), and external indicator NMI (Normalized Mutual Information) of clustering performance. The restart probability $\alpha$ of RWR in the experiment was set to 0.5. For visualization, both PCA and t-SNE with default settings reduced the data to two dimensions. Isomap and SCRL first reduced the data to 200 dimensions according to common parameter settings [12] and then mapped them to 2D through PCA. And then we calculated internal indicator DBI and external indicator NMI, ARI, Purity based on K-means clustering to evaluate the performance between different dimension reduction methods combined with RWR.

4. Results

4.1. Visualization

Figure 1 shows the 2D visualization of three smoothed scRNA-seq data sets based on different dimension reduction algorithms. And the colors represent the true cell types by previous study.

As can be seen from Figure 1, PCA focuses more on linear data structures, and the processing power for non-linear complex structures is the worst of all. Isomap cannot outline clear data structures and category relationships. Both t-SNE and SCRL perform well in retaining the original data features, and can clearly show the boundaries between different sample categories. The data distribution characteristics described by SCRL are more clear, and the samples within the category are more concentrated. However, it is worth noting that in the data set Li, the data points of the category Green show two subclasses in the SCRL dimensionality reduction results, indicating that SCRL has a poor resolving power for the subclasses.

4.2. Internal indicator

DBI measures the degree of concentration within a category and the distance between categories [17]. A smaller DBI means that the categories are better distinguished.

| Methods | Usoskin | Li | Patel |
|---------|---------|----|-------|
| PCA     | 3.23    | 5.31| 2.57  |
| t-SNE   | 9.81    | 4.02| 1.41  |
| Isomap  | 2.87    | 3.91| 4.30  |
| SCRL    | 3.90    | 2.56| 2.57  |

As is shown in Table 2, among the DBI indicators of the three datasets, Isomap is the best for Usoskin, t-SNE is the best for Patel, and SCRL is the best for Li. SCRL is steadily ranked at medium...
or optimal level for the all data sets. The other three methods have unstable performance in different data sets, and each of them has the worst performance in individual data sets.

Figure 1. Visualization after dimension reduction

4.3. External indicators after clustering
As a widely used clustering method, K-means [18] was applied to the RWR smoothed and dimension-reduced data. Then we calculated external indicators NMI [19], ARI [20] and Purity [21] to evaluate the clustering effect of different methods. For the three indicators, a larger value means better clustering performance.

Table 3. External indicators after clustering

| Methods | Usoskin | Li | Patel |
|---------|---------|----|-------|
|         | NMI     | ARI | Purity | NMI     | ARI | Purity | NMI     | ARI | Purity |
| PCA     | 0.118   | 0.637 | 0.532 | 0.200   | 0.655 | 0.421 | 0.112   | 0.321 | 0.337 |
| t-SNE   | 0.065   | 0.645 | 0.445 | 0.327   | 0.821 | 0.656 | 0.485   | 0.868 | 0.784 |
| Isomap  | 0.245   | 0.677 | 0.55  | 0.266   | 0.752 | 0.529 | 0.081   | 0.516 | 0.453 |
| SCRL    | 0.629   | 0.779 | 0.706 | 0.703   | 0.923 | 0.818 | 0.435   | 0.852 | 0.786 |

As shown in Table 3, all the three indicators show similar results. SCRL achieved the best clustering effect on both Usoskin and Li. For Patel, t-SNE and SCRL achieved the best results in
individual indicators, but the performance of t-SNE in the three indicators is slightly better than SCRL. Generally, SCRL achieved higher or comparable clustering performance than other methods, and the performance of PCA and Isomap on clustering is not good for all data sets.

5. Conclusion
Single-cell sequencing has brought the research and analysis of genetic information to a more detailed level, which undoubtedly has a pioneering significance for biology, environment and medicine. RWR-based network smooth brings prospects to scRNA-seq data preprocessing through its superiority in noise reduction. Dimension reduction plays an important role in extracting key information of the data and reducing the amount of calculation, especially for high-dimensional scRNA-seq data. Through the performance evaluation of different dimension reduction methods combined with RWR network smoothing, we can find that the effect of dimensionality reduction on RWR smoothed data is related to multiple factors.

RWR network smoothing effectively avoids the effects of noise on scRNA-seq data analysis. Under this prerequisite, the preservation of the spatial structure and feature distribution determines the performance of the dimension reduction methods. PCA, as an analysis method focused on mining linear structures, appears weak for more complex non-linear data. t-SNE has undeniably excellent performance in 2-3 dimensional visualization, but has a poor ability to retain the original data structure when processing high-dimensional sparse data. The effectiveness of Isomap depends more on the manifold structure of the data, which determines its limitations and fluctuations in the description of novel data structures. SCRL has the best and stable performance for scRNA-seq, but it also consumes the most memory and time compared to other methods. SCRL based on the network embedding model preserves the topological structure of the data by fitting the similarity distribution between samples. Therefore, it has a promoting effect when cooperating with the RWR network propagation in accordance with the network topology.

In addition, network smoothing based on RWR may also have potential limitations, which will also affect the performance of downstream analysis tools. It cannot locate the missing data with absolute precision., and the prior knowledge from the gene interaction network may bring redundant information. When combining it with other data analysis tools, the accuracy of data imputation, the adaptation between models, and the running cost should all be taken into consideration. The selection of personalized downstream tools based on the structural characteristics of different data is also a topic worthy of research. At the same time, since the current gene interaction network is constructed based on bulk RNA sequencing data, its effectiveness on scRNA-Seq data remains to be investigated. In the subsequent study of single cell, it is also possible to construct a gene network from the single-cell level by integrating high-quality scRNA-seq data. Targeted network selection and elimination of redundant information from gene networks also contribute to more efficient data imputation.

Acknowledgments
This work was supported by the National Natural Science Foundation of China [grant numbers 61573296, 61871463], and Natural Science Foundation of Fujian Province of China [grant number 2017J01068].

References
[1] Abdallah, B.Y., et al., Single cell heterogeneity. 2015. 12(23): p. 3640-3649.
[2] Grun, D., L. Kester, and A. van Oudenaarden, Validation of noise models for single-cell transcriptomics. Nature Methods, 2014. 11(6): p. 637-+.
[3] Duc-Hau, L., Random walk with restart: A powerful network propagation algorithm in Bioinformatics field. 2017 4th Nafosted Conference on Information and Computer Science (Nics), 2017: p. 242-247.
[4] Ronen, J. and A. Akalin, netSmooth: Network-smoothing based imputation for single cell RNA-seq. F1000Research, 2018. 7: p. 8-8.
Ye, W., et al., scNPF: an integrative framework assisted by network propagation and network fusion for preprocessing of single-cell RNA-seq data. Bmc Genomics, 2019. 20.

Li, W.V. and J.J. Li, An accurate and robust imputation method scImpute for single-cell RNA-seq data. Nature Communications, 2018. 9.

Huang, M., et al., SAVER: gene expression recovery for single-cell RNA sequencing. Nature Methods, 2018. 15(7): p. 539-+

van Dijk, D., et al., Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. Cell, 2018. 174(3): p. 716-+

Wold, S., et al., Principal component analysis. 1987. 2(1): p. 37-52.

van der Maaten, L. and G. Hinton, Visualizing Data using t-SNE. Journal of Machine Learning Research, 2008. 9: p. 2579-2605.

Balasubramanian, M. and E.L. Schwartz, The Isomap algorithm and topological stability. Science, 2002. 295(5552).

Li, X., et al., Network embedding-based representation learning for single cell RNA-seq data. Nucleic Acids Research, 2017. 45(19).

Usoskin, D., et al., Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. Nature Neuroscience, 2015. 18(1): p. 145-+

Li, H., et al., Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. Nature Genetics, 2017. 49(5): p. 708-+

Patel, A.P., et al., Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science, 2014. 344(6190): p. 1396-1401.

Yang, F., et al., The integration of weighted gene association networks based on information entropy. Plos One, 2017. 12(12).

Davies, D.L. and D.W.J.I.T.P.A.M.I. Bouldin, A Cluster Separation Measure. 1979. PAMI-1(2): p. 224-227.

Jain, A.K., Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 2010. 31(8): p. 651-666.

Cover, T.M. and J.A. Thomas, Elements of information theory. 2003.

Hubert, L. and P. Arabie, Comparing partitions. Journal of Classification, 1985. 2(1): p. 193-218.

Kim, H. and H. Park, Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. Bioinformatics, 2007. 23(12): p. 1495-1502.