Tensor Decomposition Discriminates Tissues Using scATAC-seq

Y-h. Taguchi
Department of Physics, Chuo University
Tokyo 112-8551, Japan
Email: tag@granular.com

Turki Turki
Department of Computer Science, King Abdulaziz University
Jeddah, 21589, Saudi Arabia
Email: tturki@kau.edu.sa
San Francisco, California 96678–2391

***Abstract*—ATAC-seq is a powerful tool to measure the landscape structure of a chromosome. scATAC-seq is a recently updated version of ATAC-seq performed in a single cell.**

The problem of scATAC-seq is data sparsity and most of the genomic sites are inaccessible.

Here, tensor decomposition (TD) was used to fill in missing values. In this study, TD was applied to massive scATAC-seq data and it generated averaging within 200 bp intervals and this number can be as high as 13,627,618. Currently, no other methods can deal with large sparse matrices.

***Index Terms*—scATAC-seq, tensor decomposition, large sparse matrices,**

I. Introduction

ATAC-seq [1] is a powerful tool to profile chromatin accessibility. Especially, scATAC-seq [2] can profile chromatin accessibility genome-wide within individual cells. Although scRNA-seq is an effective tool, the data need additional information to interpret its results. scATAC-seq data can be analyzed in two ways: 1) the scATAC-seq data coupled to scRNA-seq data [3], or 2) tracing for potential factors that bind to identified open chromatin regions from scATAC-seq data. This prevents us from understanding scATAC-seq as it is.

For example, Satpathy et al. [4] tried to find cell type-specific cis- and trans-regulatory elements, mapping of disease-associated enhancer activity, and reconstruction of trajectories of differentiation from progenitors to diverse and rare immune cell types. Giansanti et al. [5] made use of a predefined set of genomic regions to interpret the scATAC-seq. Buenrostro et al. [6] also tried to find an association with transfactors and cis elements. Although these are only a few examples, it is obvious that they need massive external information to interpret the scATAC-seq results.

In this paper, we applied tensor decomposition(TD) [7] to the scATAC-seq data set and found that the low-dimensional embedding obtained by UMAP applied to that obtained by TD is highly tissue-specific. This can open numerous avenues of research to make use of scATAC-seq data without additional biological information.

II. Materials and Methods

Figure 1 shows the flow chart of the analysis.

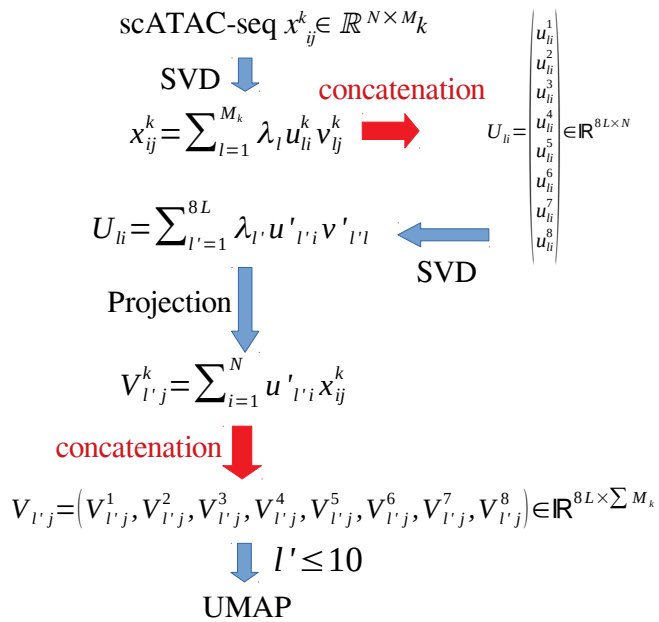

Fig. 1. Flow chart of the analysis

A. scATAC-seq profiles

The scATAC-seq data set we analyzed in this study is obtained from Gene Expression Omnibus (GEO) [8], the ID is GSE167050. We used eight samples from GSM5091379 to GSM5091386 of four mice tissues (CTX, MGE, CGE and LGE) with two replicates each (Table I). Each of the eight samples is associated with three files, barcodes, features, and matrices, which correspond to single cells, genomic coordinates, and scATAC-seq values, respectively.

B. Pre-processing scATAC-seq profile

The values stored in the matrix files are averaged over 200 bp intervals, which are supposed to correspond to the length of one wrap of chromatin and linker . Since this results in a sparse matrix, it is stored in a sparse matrix format having columns and rows equivalent to the number of single cells and total number of intervals, which is up to 13,627,618. This value is approximately equal to 12.5 million, which is calculated by

TABLE I
NUMBER OF SINGLE CELLS INCLUDED IN THE FILES ANALYZED IN THIS STUDY. CTX: CORTEX, MGE: MEDIAL GANGLIONIC EMINENCE, CGE: CAUDAL GANGLIONIC EMINENCE, LGE: LATERAL GANGLIONIC EMINENCE

Tissues	CTX1	MGE1	CGE1	LGE1	CTX2	MGE2	CGE2	LGE2	total
number of single cells (M_k)	4108	6845	4013	6577	4946	3465	4530	4769	39253

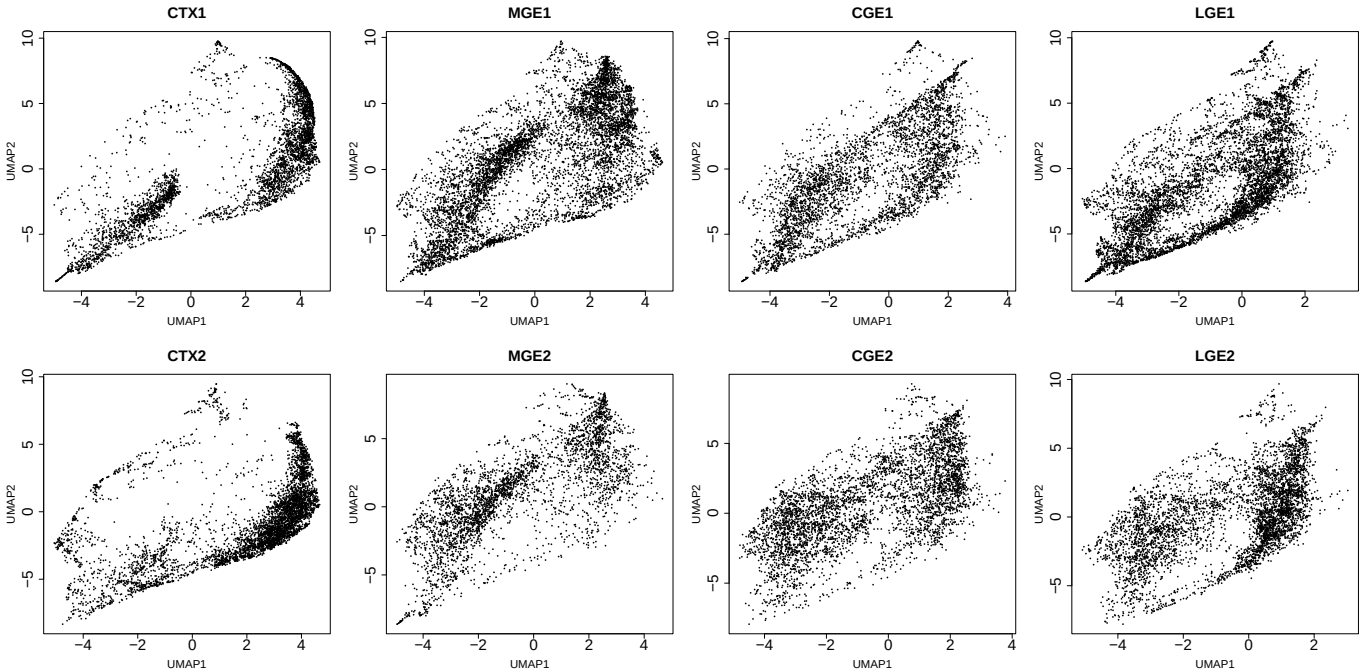

Fig. 2. UMAP embedding of eight samples analyzed in this study (see Table I)

dividing the total number of mice genome bps to the interval length, i.e. 2.5 billion by 200.

C. Application of singular value decomposition

To obtain a low-dimensional embedding of individual kth samples to be reformatted as a tensor, we applied singular value decomposition (SVD) to marix, $x^k_{ij} \in \mathbb{R}^{N \times M_k}$ ($N =$ 13,627,618 and M_ks are in Table I) as

$$x^k_{ij} = \sum_{\ell=1}^{M_k} \lambda_\ell u^k_{\ell i} v^k_{\ell j} \tag{1}$$

where $u^k_{\ell i} \in \mathbb{R}^{M_k \times N}$ and $v^k_{\ell j} \in \mathbb{R}^{M_k \times M_k}$ are orthogonal singular value matrices. Then we get concatenated matrix $U_{\ell i} \in \mathbb{R}^{8L \times N}$ of $u_{\ell i}$ where $L < \min(M_k)$ is the used number of principal components. SVD is applied to $U_{\ell i}$ and we get

$$U_{\ell i} = \sum_{\ell'=1}^{8L} \lambda_{\ell'} u'_{\ell' i} v'_{\ell' \ell} \tag{2}$$

where $u'_{\ell' i} \in \mathbb{R}^{8L \times N}$ and $v_{\ell' \ell} \in \mathbb{R}^{8L \times L}$ are orthogonal singular value matrices.

Low-dimensional embedding can be obtained as

$$V^k_{\ell' j} = \sum_{i=1}^{N} u'_{\ell' i} x^k_{ij} \in \mathbb{R}^{8L \times M_k}. \tag{3}$$

and we get concatenated matrix $V_{\ell' j} \in \mathbb{R}^{8L \times \sum_{k=1}^{8} M_k}$

D. Sparse matrix format and singular value decomposition

Storing x^k_{ij} and $U_{\ell i}$ in a sparse matrix format allows us to deal with a matrix, which has one dimension N greater than 10 million. It also enables us to apply SVD to these using the irlba package [9] in R [10], which is adapted for large sparse matrices.

E. UMAP

UMAP was applied to concatenated matrix, $V_{\ell' j}$, $\ell' \leq L$, by umap [11] function in R with the option `umap.defaults$n_neighbors <-30` (other options remain as default).

F. Estimation of distribution of single cells in UMAP

To quantify the distribution of single cells in UMAP, we divided the complete region into 10×10 regions. The region $S_{IJ}, 1 \leq I, J \leq 10$ is $\{(x, y) | (I-1)\Delta x \leq x \leq I\Delta x, (J-1)\Delta y \leq y \leq J\Delta y\}$ where $\Delta x = \frac{\max(x_j) - \min(x_j)}{10}$ and $\Delta y = \frac{\max(y_j) - \min(y_j)}{10}$ and x_j and y_j are coordinates of jth cell in UMAP embedding. The number of single cells, N_{IJ}, in the region S_{IJ} is equal to the number of $(x_j, y_j) \in S_{IJ}$.

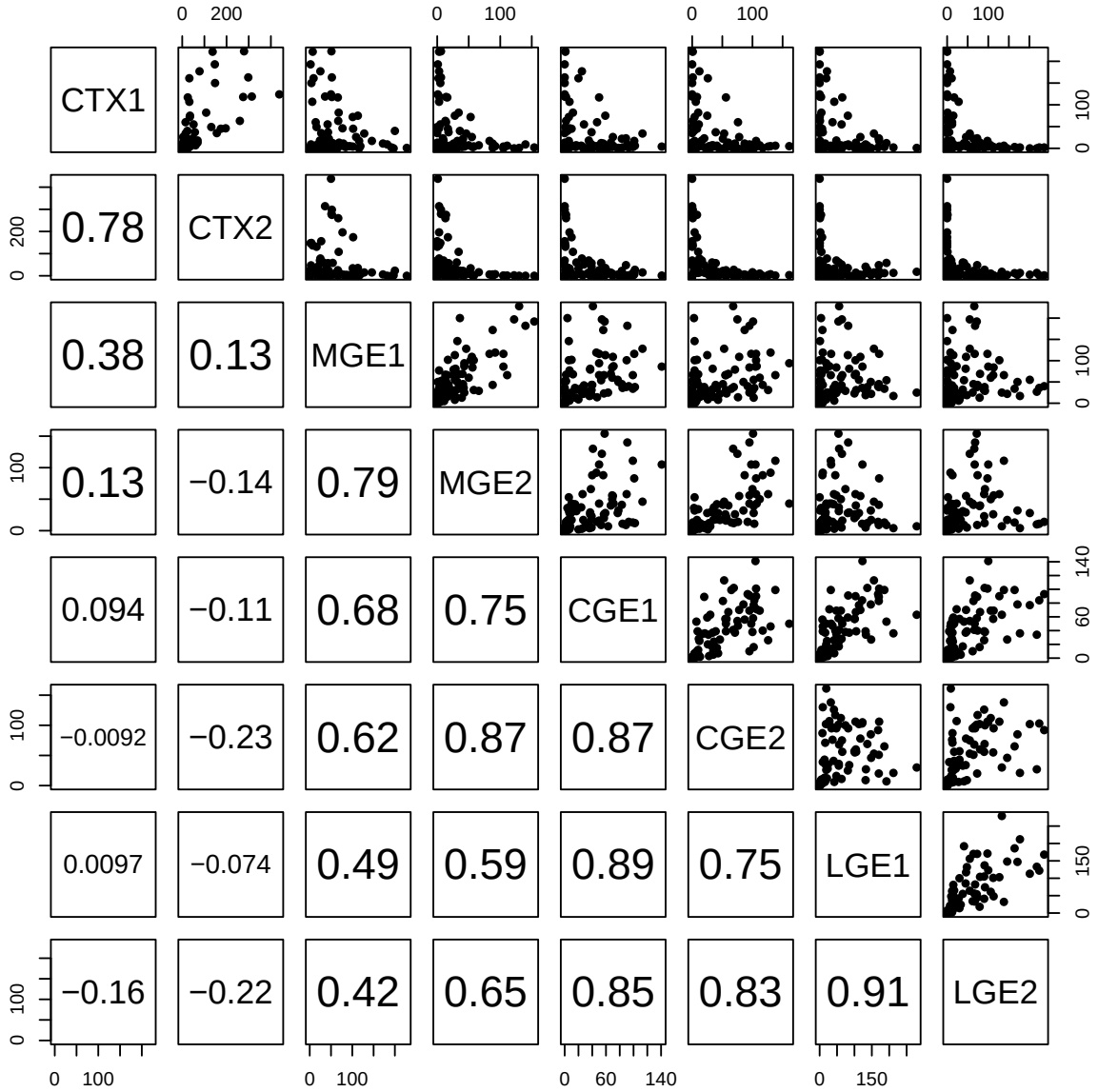

Fig. 3. Upper: scatter plot of the number of single cells, $N_{IJ}, 1 \leq I, J \leq 10$, in one of 10×10 regions, S_{IJ}. Lower: Spearman's correlation coefficients of N_{IJ} between eight samples.

G. UPGMA

UPGMA was performed with `hclust` function in R using option method='average'.

III. Result

After obtaining concatenated matrix $V_{\ell' j}$ as described in the Methods section, UMAP was applied to $V_{\ell' j} \in \mathbb{R}^{L \times \sum_{k=1}^{8} M_k}$, i.e., the first L dimensions of $V_{\ell' j}$ (in this case, $L = 10$). Figure 2 shows the two-dimensional embedding of eight samples in Table I. As observed in investigated coordinates, the distribution of single cells in eight samples fully overlapped with one another. Nevertheless, the distribution of single cells seems to be somewhat tissue specific. To quantify the similarity of distributions, we divided the whole region into 10×10 regions and counted the number of cells in the individual regions. Figure 3 shows the scatter plots and correlation coefficients of N_{IJ} between eight samples. The correlation coefficients between similar tissues were generally high. Whereas in few cases, correlation coefficients between different tissues were also high. To see if correlation coefficients were useful for classifying samples, we applied UPGMA (unweighted pair group method with arithmetic mean) to negatively signed correlation coefficients (Fig. 4). It is obvious that similar tissues were paired in the clustering. In addition to this,

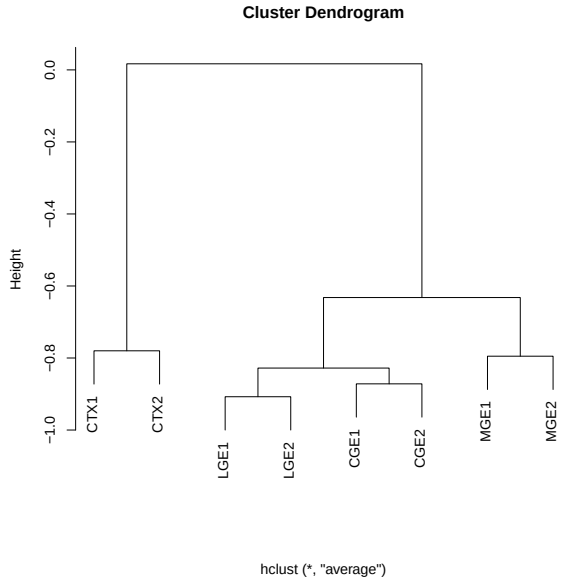

Fig. 4. Hierarchical clustering by UPGMA of eight samples (Table I) using coordinates obtained by UMAP. Distance is represented by negatively signed correlation coefficients in Fig. 3

two CTX samples were clustered together, apart from six ganglionic eminence samples (LGE, CGE, and MGE); which is also biologically reasonable.

IV. Discussion

Although tensor decomposition can generate the feature that can cluster samples properly, if other methods can do this, the proposed method is still efficient and unique in terms of tensor representation. Currently, there are very few tools to process scATAC-seq data with only matrix data. For example, although the extended data in Fig. 1 of the past study [12] summarizes ten *de facto* standard methods that can deal with a scATAC-seq data set, no methods can process the scATAC-seq data set with only matrix data. We also tried some methods [13], [14], [15] not included in the above list. No tool could successfully process the x_{ij}^k efficiently, it is possible because of large N. Thus, the proposed method is the only method that can deal with a data set of this size. The reason why the proposed method can handle this large N is that it stores the data set in a sparse matrix format. SVD performs using the tool adapted for the sparse matrix format, and we do not need to process dense format. Thus, the proposed method can deal with huge data sets.

V. Conclusion

In this paper, we applied Tensor Decomposition to an scATAC-seq data set and the obtained embedding can be used for UMAP, following which the embedded material obtained by UMAP can differentiate tissues from which the scATAC sequence was retrieved. TD can deal with large sparse data sets generated by approximately 200 bp intervals, as these can be stored in a sparse matrix format. The large size of these data sets cannot be processed by any other methods. The proposed method is the only method that can deal with high-resolution native scATAC-seq data sets.

Acknowledgment

This study was supported by KAKENHI 20K12067.

References

[1] F. C. Grandi, H. Modi, L. Kampman, and M. R. Corces, "Chromatin accessibility profiling by ATAC-seq," *Nature Protocols*, vol. 17, no. 6, pp. 1518–1552, Apr. 2022. [Online]. Available: https://doi.org/10.1038/s41596-022-00692-9

[2] S. Baek and I. Lee, "Single-cell ATAC sequencing analysis: From data preprocessing to hypothesis generation," *Computational and Structural Biotechnology Journal*, vol. 18, pp. 1429–1439, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2001037020303019

[3] T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck, Y. Hao, M. Stoeckius, P. Smibert, and R. Satija, "Comprehensive integration of single-cell data," *Cell*, vol. 177, no. 7, pp. 1888–1902.e21, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0092867419305598

[4] A. T. Satpathy, J. M. Granja, K. E. Yost, Y. Qi, F. Meschi, G. P. McDermott, B. N. Olsen, M. R. Mumbach, S. E. Pierce, M. R. Corces, P. Shah, J. C. Bell, D. Jhutty, C. M. Nemec, J. Wang, L. Wang, Y. Yin, P. G. Giresi, A. L. S. Chang, G. X. Y. Zheng, W. J. Greenleaf, and H. Y. Chang, "Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion," *Nat Biotechnol*, vol. 37, no. 8, pp. 925–936, 08 2019.

[5] V. Giansanti, M. Tang, and D. Cittaro, "Fast analysis of scatac-seq data using a predefined set of genomic regions [version 2; peer review: 2 approved]," *F1000Research*, vol. 9, no. 199, 2020.

[6] J. D. Buenrostro, B. Wu, U. M. Litzenburger, D. Ruff, M. L. Gonzales, M. P. Snyder, H. Y. Chang, and W. J. Greenleaf, "Single-cell chromatin accessibility reveals principles of regulatory variation," *Nature*, vol. 523, no. 7561, pp. 486–490, Jul 2015.

[7] Y.-H. Taguchi, *Unsupervised Feature Extraction Applied to Bioinformatics*. Springer International Publishing, 2020. [Online]. Available: https://doi.org/10.1007/978-3-030-22456-1

[8] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis, and A. Soboleva, "NCBI GEO: archive for functional genomics data sets-update," *Nucleic Acids Research*, vol. 41, no. D1, pp. D991–D995, 11 2012. [Online]. Available: https://doi.org/10.1093/nar/gks1193

[9] J. Baglama, L. Reichel, and B. W. Lewis, *irlba: Fast Truncated Singular Value Decomposition and Principal Components Analysis for Large Dense and Sparse Matrices*, 2021, r package version 2.3.5. [Online]. Available: https://CRAN.R-project.org/package=irlba

[10] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/

[11] T. Konopka, *umap: Uniform Manifold Approximation and Projection*, 2022, r package version 0.2.8.0. [Online]. Available: https://CRAN.R-project.org/package=umap

[12] J. M. Granja, M. R. Corces, S. E. Pierce, S. T. Bagdatli, H. Choudhry, H. Y. Chang, and W. J. Greenleaf, "ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis," *Nat Genet*, vol. 53, no. 3, pp. 403–411, 03 2021.

[13] L. Xiong, K. Xu, K. Tian, Y. Shao, L. Tang, G. Gao, M. Zhang, T. Jiang, and Q. C. Zhang, "SCALE method for single-cell ATAC-seq analysis via latent feature extraction," *Nat Commun*, vol. 10, no. 1, p. 4576, 10 2019.

[14] Z. Li, C. Kuppe, S. Ziegler, M. Cheng, N. Kabgani, S. Menzel, M. Zenke, R. Kramann, and I. G. Costa, "Chromatin-accessibility estimation from single-cell ATAC-seq data with scOpen," *Nat Commun*, vol. 12, no. 1, p. 6386, 11 2021.

[15] W. Kopp, A. Akalin, and U. Ohler, "Simultaneous dimensionality reduction and integration for single-cell ATAC-seq data using deep learning," *Nat Mach Intell*, vol. 4, no. 2, pp. 162–168, Feb. 2022.