Prediction of Host-Pathogen RNA Interaction From RNA Sequences and Dual RNA-seq Data using Variational Autoencoders and Supervised Machine Learning Methods

1. Abstract

Eukaryotic host cells are prone to infection by diverse agents, from viruses to bacteria to eukaryotic parasites such as fungi and protozoa [1]. During the course of bacterial infection, bacteria and host eukaryotic cells engage in a complex interplay as they negotiate their respective survival and defense strategies [2]. Understanding the interactions between host cells and bacteria pathogens is essential for improving therapeutics' diagnostic and development.

RNA sequencing (RNA-seq) enables the analysis of differentially expressed genes (DEGs) in infected cells. However, RNA-seq is limited to analyzing either pathogen or host cell after their physical separation [3]. Recently developed dual RNA-Seq technology can simultaneously capture host and bacterial transcriptomes from an infected cell without disturbing the complex host-bacteria interactions [2]. However, dual RNA-Seq, as a high-throughput technology, creates high-dimensional data which is prone to biological and technical noise. Moreover, visualization of high dimensional data is a fundamental problem in dual RNA-seq data analysis since it can lead to misinterpreting the biological features [4]. Subsequently, dimension reduction is a crucial step prior to analyzing dual RNA-seq data.

There are several traditional dimension-reduction methods used for RNA-seq data, of which the most commonly used are principal component analysis (PCA) [5] and t-distributed stochastic neighbor embedding (t-SNE) [6]. However, PCA is mainly limited to linear dimensions and utilizes data with normal distribution, which may not be appropriate for RNA-seq datasets [7]. Although researchers widely use the t-SNE method for data visualization, it may not be a practical solution for dimensionality reduction [8]. Recently, multiple studies have established the utility of variational autoencoder (VAE) [9] for generating meaningful latent features from scRNA-seq data and have indicated that VAE yielded features that retained biological patterns and enabled more accurate classifiers for prediction tasks compared to PCA or t-SNE [10, 11].

With the advent of machine learning (ML), ML algorithms have been widely used for biological studies, including biological-image analysis, cancer study, as well as gene discovery. Several single-cell RNA-seq studies [12–15] have demonstrated that machine learning and deep learning approaches can assist with the identification of DEGs that are missed by traditional RNA-seq data analysis techniques. However, current studies of dual RNA-seq analysis are limited to traditional bioinformatics approaches [16–18].

In this research, we will apply VAE to encode dual RNA-seq expression data and compare the performance of VAE to commonly used dimensionality reduction methods, including PCA, t-SNE, ZIFA [19], and UMAP [20] in classification task, predicting interacting host and bacterial RNAs. The performance of the dimensionality reduction methods will be compared in terms of area under the precision recall curve (AUPRC) achieved by ML models trained with the generated features and RNA sequences.

References

[1] A. J. Westermann, S. A. Gorski, and J. Vogel. "Dual RNA-seq of pathogen and host". In: *Nature Reviews Microbiology* 10.9 (2012), pp. 618–630. DOI: 10.1038/nrmicro2852. URL: www.doi.org/10.1038/nrmicro2852.

[2] J. W. Marsh, R. J. Hayward, A. C. Shetty, et al. "Bioinformatic analysis of bacteria and host cell dual RNA-sequencing experiments". In: *Briefings in Bioinformatics* (2017). DOI: 10.1093/bib/bbx043. URL: www.doi.org/10.1093/bib/bbx043.

[3] A. J. Westermann, K. U. Förstner, F. Amman, et al. "Dual RNA-seq unveils noncoding RNA functions in host–pathogen interactions". In: *Nature* 529.7587 (2016), pp. 496–501. DOI: 10.1038/nature16547. URL: www.doi.org/10.1038/nature16547.

[4] K. R. Moon, D. van Dijk, Z. Wang, S. Gigante, D. B. Burkhardt, W. S. Chen, K. Yim, A. van den Elzen, M. J. Hirn, R. R. Coifman, N. B. Ivanova, G. Wolf, and S. Krishnaswamy. "Visualizing structure and transitions in high-dimensional biological data". In: *Nature Biotechnology* 37.12 (2019), pp. 1482–1492. DOI: 10.1038/s41587-019-0336-3. URL: https://doi.org/10.1038%2Fs41587-019-0336-3.

[5] S. Wold, K. Esbensen, and P. Geladi. "Principal component analysis". In: *Chemometrics and Intelligent Laboratory Systems* 2.1 (1987). Proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists, pp. 37–52. ISSN: 0169-7439. DOI: https://doi.org/10.1016/0169-7439(87)80084-9. URL: https://www.sciencedirect.com/science/article/pii/0169743987800849.

[6] L. van der Maaten and G. Hinton. "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9.86 (2008), pp. 2579–2605. URL: http://jmlr.org/papers/v9/vandermaaten08a.html.

[7] T. S. Andrews and M. Hemberg. "Identifying cell populations with scRNASeq". In: *Molecular Aspects of Medicine* 59 (2018). The emerging field of single-cell analysis, pp. 114–122. ISSN: 0098-2997. DOI: https://doi.org/10.1016/j.mam.2017.07.002. URL: https://www.sciencedirect.com/science/article/pii/S0098299717300493.

[8] E. Lin, S. Mukherjee, and S. Kannan. "A deep adversarial variational autoencoder model for dimensionality reduction in single-cell RNA sequencing analysis". In: *BMC Bioinformatics* 21.1 (2020). DOI: 10.1186/s12859-020-3401-5. URL: www.doi.org/10.1186/s12859-020-3401-5.

[9] C. Doersch. *Tutorial on Variational Autoencoders*. arXiv: 1606.05908. 2021. eprint: 1606.05908.

[10] Q. Wei and S. A. Ramsey. "Predicting chemotherapy response using a variational autoencoder approach". In: *BMC Bioinformatics* 22.1 (2021). DOI: 10.1186/s12859-021-04339-6. URL: www.doi.org/10.1186/s12859-021-04339-6.

[11] R. Umarov, Y. Li, and E. Arner. "DeepCellState: An autoencoder-based framework for predicting cell type specific transcriptional states induced by drug treatment". In: *PLOS Computational Biology* 17.10 (2021). Ed. by M. Rattray, e1009465. DOI: 10.1371/journal.pcbi.1009465. URL: www.doi.org/10.1371/journal.pcbi.1009465.

[12] L. Wang, Y. Xi, S. Sung, et al. "RNA-seq assistant: machine learning based methods to identify more transcriptional regulated genes". In: *BMC Genomics* 19.1 (2018). DOI: 10.1186/s12864-018-4932-2. URL: www.doi.org/10.1186/s12864-018-4932-2.

[13] G. Chen, B. Ning, and T. Shi. "Single-Cell RNA-Seq Technologies and Related Computational Data Analysis". In: *Frontiers in Genetics* 10 (2019). DOI: 10.3389/fgene.2019.00317. URL: www.doi.org/10.3389%2Ffgene.2019.00317.

[14] R. Petegrosso, Z. Li, and R. Kuang. "Machine learning and statistical methods for clustering single-cell RNA-sequencing data". In: *Briefings in Bioinformatics* 21.4 (2019), pp. 1209–1223. DOI: 10.1093/bib/bbz063. URL: www.doi.org/10.1093/bib/bbz063.

[15] M. T. Hira, M. A. Razzaque, C. Angione, et al. "Integrated multi-omics analysis of ovarian cancer using variational autoencoders". In: *Scientific Reports* 11.1 (2021). DOI: 10.1038/s41598-021-85285-4. URL: www.doi.org/10.1038/s41598-021-85285-4.

[16] A. D'Mello, A. N. Riegler, E. Martínez, et al. "An in vivo atlas of host–pathogen transcriptomes during Streptococcus pneumoniae colonization and disease". In: *Proceedings of the National Academy of Sciences* 117.52 (2020), pp. 33507–33518. DOI: 10.1073/pnas.2010428117. URL: www.doi.org/10.1073/pnas.2010428117.

[17] A. J. Westermann, L. Barquist, and J. Vogel. "Resolving host–pathogen interactions by dual RNA-seq". In: *PLOS Pathogens* 13.2 (2017). Ed. by J. B. Bliska, e1006033. DOI: 10.1371/journal.ppat.1006033. URL: www.doi.org/10.1371/journal.ppat.1006033.

[18] R. J. Hayward, M. S. Humphrys, W. M. Huston, et al. "Dual RNA-seq analysis of in vitro infection multiplicity and RNA depletion methods in Chlamydia-infected epithelial cells". In: *Scientific Reports* 11.1 (2021). DOI: 10.1038/s41598-021-89921-x. URL: www.doi.org/10.1038/s41598-021-89921-x.

[19] E. Pierson and C. Yau. "ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis". In: *Genome Biology* 16.1 (2015). DOI: 10.1186/s13059-015-0805-z. URL: https://doi.org/10.1186%2Fs13059-015-0805-z.

[20] L. McInnes, J. Healy, N. Saul, and L. Großberger. "UMAP: Uniform Manifold Approximation and Projection". In: *Journal of Open Source Software* 3.29 (2018), p. 861. DOI: 10.21105/joss.00861. URL: https://doi.org/10.21105/joss.00861.