Biologically Refined Imputation via Diffusion for Gene Expression

Keywords: Single-cell RNA sequencing (scRNA-seq), Imputation, Foundational model

Motivation: Machine learning has become an essential tool in healthcare for analyzing complex transcriptomic data, where challenges such as sparsity, high dimensionality, and interpretability limit traditional methods. A key application domain is Single-cell RNA sequencing (scRNA-seq), a powerful technology that allows scientists to analyze the gene expression of individual cells. scRNA-seq captures the transcriptional heterogeneity within tissues and facilitates the identification of cell-type-specific expression patterns relevant to biological development and disease. scRNA-seq data is typically represented as gene-cell expression matrix, where rows denote genes and columns represent individual cells. Each element in this matrix reflects the expression level of a particular gene within a specific cell. A fundamental challenge in scRNA-seq analysis is the substantial sparsity of the data, with zero expression values comprising up to more than 90% of the gene-cell expression matrix[1]. This poses significant challenges for accurately capturing gene–gene and cell–cell relationships, which are critical components of downstream biological analyses such as cell type annotation, cell clustering, and gene network inference. Most of the existing approaches fail to preserve these relationships, largely because they inadequately integrate biological context during the imputation process.

Method: We propose BRIDGE (Biologically Refined Imputation via Diffusion for Gene Expression). This novel two-stage framework integrates external biological knowledge in conjunction with data-driven representation learning to impute the missing data. In the first stage, BRIDGE enhances gene-gene similarity by computing gene similarity matrix from the gene-cell expression matrix and then refining it using curated gene sets from resources (MSigDB[2]). This biologically informed filtering restricts diffusion to functionally relevant gene pairs, preserving biologically meaningful structures while reducing the propagation of noise. In the second stage, BRIDGE leverages cell embeddings from pretrained scGPT[3], a foundational model for single-cell transcriptomic data. These embeddings capture rich, biologically informed representations of cellular states. The cell embeddings are then used to construct a cell similarity matrix, which guides the subsequent cell-wise feature propagation. By combining gene-wise propagation constrained by biologically validated gene sets with cell-wise refinement guided by pretrained cell representations, BRIDGE generates an imputed gene–cell expression matrix that better preserves functional relationships and improves performance in cell clustering task.

Evaluation and Results: We evaluate our proposed method, BRIDGE, on two benchmark scRNA-seq datasets: Human Innate T Cell (93.05% sparsity) and Human Lung (94.49% sparsity), focusing on cell clustering tasks. Clustering performance was quantitatively assessed using six metrics: Rand Index (RI), Adjusted Rand Index (ARI), Mutual Information (MI), Normalized Mutual Information (NMI), Adjusted Mutual Information (AMI), and Fowlkes-Mallows Index (FMI). In these experiments, BRIDGE consistently outperforms the baseline method scBFP[4] (p -value < 0.05). To further assess robustness, we introduced additional sparsity by masking non-zero gene expression values, progressively increasing sparsity levels up to 98%. BRIDGE maintained strong performance under these extreme conditions, demonstrating its effectiveness in preserving biological signals even with highly sparse data.

Conclusion: Our work underscores the potential of combining domain knowledge and representation learning to advance scRNA-seq imputation. By situating the imputation process within a biologically meaningful context, BRIDGE not only improves accuracy but also enhances interpretability, laying the foundation for more reliable downstream single-cell analyses.

References

1. Jiang R, Sun T, Song D, et al. Statistics or biology: the zero-inflation controversy about scRNA-seq data. Genome biology 2022;23(1):31.
2. Liberzon A, Subramanian A, Pinchback R, et al. Molecular signatures database (MSigDB) 3.0. Bioinformatics 2011;27(12):1739–1740.
3. Cui H, Wang C, Maan H, et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. Nature methods 2024;21(8):1470–1480.
4. Lee J, Yun S, Kim Y, et al. Single-cell RNA sequencing data imputation using bi-level feature propagation. Briefings in Bioinformatics 2024;25(3):bbae209.