scDCCA: Deep contrastive clustering for single-cell RNA-seq data based on auto-encoder network

2024.11

Abstract

- **Task:** Cell clustering in scRNA-seq analysis.
- **Technical Challenges:**
  - High dimensionality, noise, and significant sparsity of scRNA-seq data.
  - Limitations of previous methods.
- **Challenge Key Insights:**
  - Intrinsic properties of cells.
  - Relationship among cells.
- **Technical Contributions:**
  - Denoising Zero-Inflated Negative Binomial (ZINB) model-based auto-encoder.
  - Dual contrastive learning module for pairwise proximity of cells.
  - Joint feature learning and clustering in an end-to-end manner.
- **Experiment:**
  - Outperform 8 methods on 14 real datasets.
  - Metrics: Accuracy, generalizability, scalability, and efficiency.
  - Cell visualization and biological analysis.

Introduction: Background and Challenges

- **Task and Application:**
  - Characterize cell types in scRNA-seq.
- **Technical Challenges for Previous Methods:**
  - **Methods without prior information:**
    - SIMLR: Multiple kernel.
    - Seurat: Louvain algorithm on KNN.
    - SHARP: Ensemble random projection-based.
    - scHFC: Fuzzy C Mean and Gath-Geva algorithms.
  - **Limitations of these methods:**
    - Dimension reduction (e.g., PCA) may lose vital information.
    - Similarity matrices fail to reflect cell similarity.
    - High computational and time cost.

Introduction: Advanced Methods and Limitations

- **Deep Clustering Methods:**
  - scGMAI, scCCESS: Auto-encoder, minimize MSE.
  - DCA: Auto-encoder with ZINB.
  - scDeepCluster: Combines DCA and DEC.
- Focus on data only, neglecting cell relationships.
- **Cell Relations Considered Methods:**
  - **Graph-based methods:**
    - GraphSCC, scDSC, scGNN.
    - Limitation: Mixed clustering results.
  - **Contrast-sc:**
    - Ignores the characteristics of the data itself.
    - Impact clustering performance as dividing into 2 stages: embedding and clustering

Introduction: Technical Contributions of scDCCA

- Dual contrastive learning module to acquire pairwise cell proximity.
- Denoising ZINB auto-encoder for intrinsic feature representation.
- End-to-end training with clustering optimization.

Method

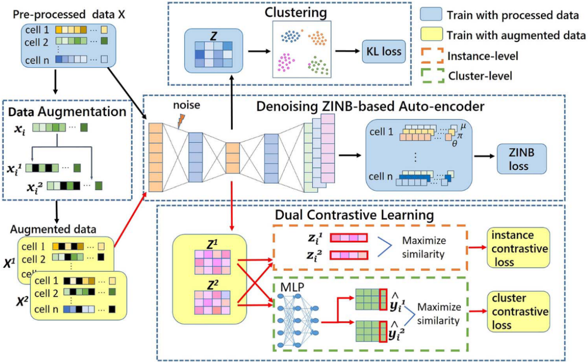

Figure: scDCCA

Method Overview

- **Four Components:**
  - Data augmentation for contrastive learning.
  - Denoising ZINB-based auto-encoder.
  - Dual contrastive learning module.
  - Clustering module.
- **Training Process:**
  - Pre-train with auto-encoder to capture clustering-friendly embeddings.

$$L_{pre-train} = \min \left(L_{ZINB} + \alpha L_{ins} \right)$$

  - Fine-tune the embeddings with the cluster-level contrastive loss (Lcc) and cluster data with Kullback–Leibler (KL) loss (Ldec) to achieve intra-cluster compactness and inter-cluster separation.

$$L_{train} = \min \left(L_{ZINB} + \beta L_{cc} + \gamma L_{dec} \right)$$

- **Data Processing:**
  - Log normalization and gene selection using SCANPY.

Denoising Auto-Encoder Embedding Module

- **Corruption with Gaussian noise:** Add random Gaussian noise to the input to enhance feature robustness.
- **Encoder-Decoder:** The encoder maps input to a latent space, and the decoder reconstructs it. This learns essential features.
- **Loss Function:** Minimize reconstruction loss using the ZINB loss with three parameters: mean, dispersion, and coefficient:

$$L = -\log P(\text{data}|\text{model parameters}).$$

$$L_{ZINB} = \sum_{ij} -\log\left(ZINB\left(x_{ij}|\pi_{ij}, \mu_{ij}, \theta_{ij}\right)\right)$$

Data Augmentation Module

- **Purpose:** Enhance contrastive learning by diversifying training data without labeled examples.
- **Positive/Negative Pairs:** Construct positive pairs (similar) and negative pairs (dissimilar).
- **Augmentation:** For each input cell x_i, create two augmented views x_i^1 and x_i^2 by randomly masking genes.

Dual Contrastive Learning Module

- **Cell Relationships:** Learn relationships among cells to improve intra-cluster compactness and inter-cluster separation.
- **Loss in Latent Space:** Compute losses at both instance and cluster levels in the ZINB latent space to optimize clustering.

Experiment

- **Comparison:** scDCCA outperforms 8 state-of-the-art methods.
- **Batch size effect:**
  - Suitable batch size for each dataset is related to its sample size.
- **Ablation Studies:**
  - Instance-level and cluster-level contrastive learning both critical.
- **Biological Analysis:**
  - Cell annotation, KEGG pathway, and DEG analysis.
- **Robustness:** down-sampling and change ratio of masked genes
- **Scalability and Efficiency:**
  - Handles large datasets.
  - Reasonable runtime and memory usage.

Experiment: Comparisons on 3 metrics

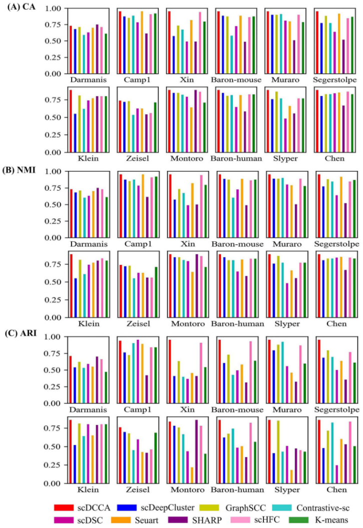

Figure: Clustering performance

Experiment: Comparisons with cell visualization

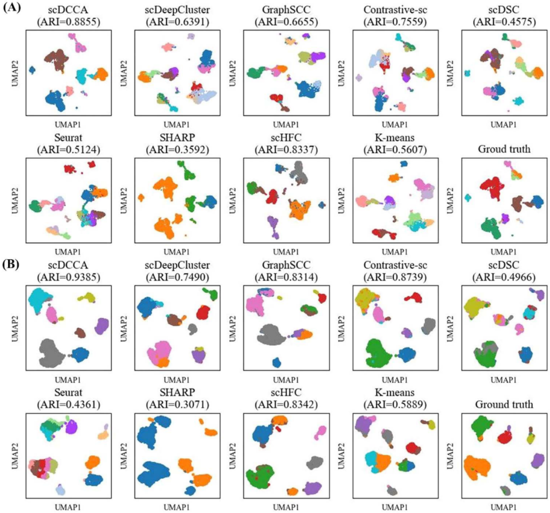

Figure: Cell visualization

Experiment: Batch size effect

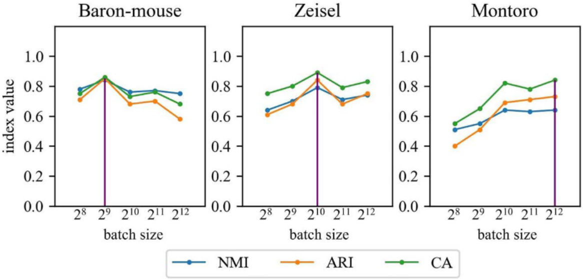

Figure: Batch size effect

Experiment: Ablation Study

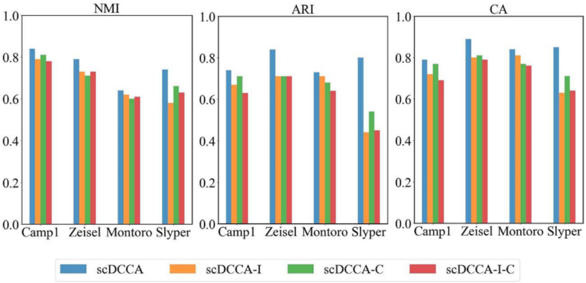

Figure: Ablation Study

Experiment: Robustness

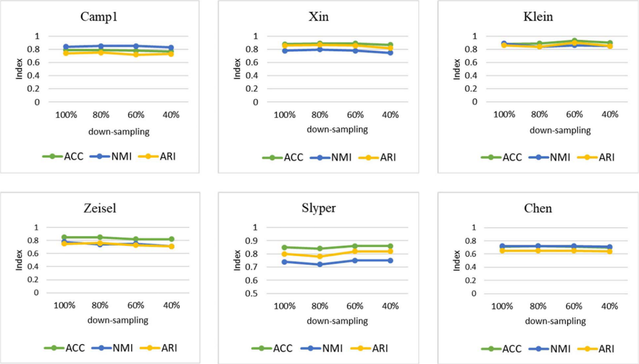

Figure: Robustness experiment

Experiment: Biological analysis

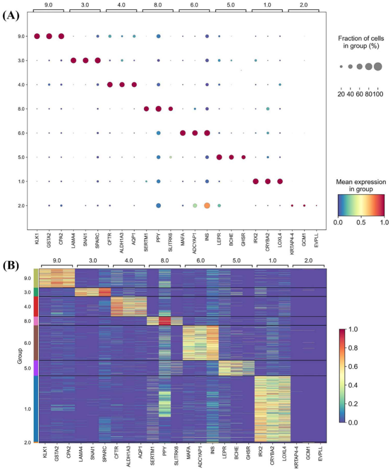

Figure: Biological analysis

Limitations

- Only uses genes as features, ignoring gene relationships.
- Future work:
  - Integrate gene relationships.
  - Explore GNN for higher-order structural information.