PeakTrajectory

PeakTrajectory is a Python package for inferring trajectories from single-cell ATAC-seq data. It has been adapted from the GeneTrajectory package to work with peak accessibility data instead of gene expression data.

Overview

This package provides computational tools to:

1. Select variable peaks from single-cell ATAC-seq data
2. Compute peak-to-peak distances using Earth Mover's Distance (EMD)
3. Construct peak trajectories via diffusion maps
4. Visualize peak trajectories in 2D and 3D
5. Add peak bin scores to an AnnData object

The core methodology uses optimal transport to calculate distances between peaks based on their accessibility patterns across cells.

Installation

Unset

```
# Clone the repository
git clone https://github.com/PCMGF-Limited/BMCS4575_Project.git
cd BMCS4575_Project

# Set up the environment
bash setup_env.sh

# Install the package in development mode
pip install -e .
```

Usage Example

Here's a basic example of how to use PeakTrajectory:

Python

```python
import scanpy as sc
import numpy as np
import pandas as pd
from peak_trajectory import extract_peak_trajectory,
add_peak_bin_score, coarse_grain
from peak_trajectory.plot import plot_peak_trajectory_3d,
plot_peak_trajectory_umap

# Load and preprocess your ATAC-seq data
adata = sc.read_h5ad('scanpy_objects/pbmc_10k_v3.h5ad')

# Select variable peaks
from peak_trajectory.coarse_grain import select_top_peaks
variable_peaks = select_top_peaks(adata, n_top_peaks=5000)

# Compute peak-peak distances (this is computationally intensive)
# You can use the provided command-line functions for large
datasets

# Extract peak trajectories
peak_embedding = pd.DataFrame(...)  # Peak embedding from
diffusion maps
dist_mat = np.array(...)  # Peak-peak distance matrix
peak_names = adata.var_names[variable_peaks].tolist()

peak_trajectories = extract_peak_trajectory(
    peak_embedding=peak_embedding,
    dist_mat=dist_mat,
    peak_names=peak_names,
    t_list=[3, 3, 3],
    dims=5,
    k=10,
    quantile=0.02
)

# Visualize the trajectories
```

```
plot_peak_trajectory_3d(peak_trajectories)

# Add peak bin scores to the AnnData object
add_peak_bin_score(adata, peak_trajectories, n_bins=5)

# Visualize trajectory scores on UMAP
plot_peak_trajectory_umap(adata, trajectory='Trajectory1')
```

Project Structure

- **data_processing/**: Scripts for processing raw ATAC-seq and RNA-seq data
    - `pbmc_10k_atac.py`, `pbmc_10k_atac.sh`: Process 10k PBMC ATAC-seq data
    - `pbmc_10k_v3.py`, `pbmc_10k_v3.sh`: Process 10k PBMC RNA-seq data
- **peak_trajectory/**: Core package functionality
    - `add_peak_bin_score.py`: Add peak bin scores to AnnData
    - `coarse_grain.py`: Reduce number of cells via metacell aggregation
    - `compute_peak_distance_cmd.py`: Calculate peak-peak EMD matrix
    - `diffusion_map.py`: Run diffusion maps
    - `extract_peak_trajectory.py`: Extract peak trajectories
    - `get_graph_distance.py`: Compute cell-cell graph distances
    - `peak_distance_shared.py`: Core EMD calculation functions
    - `run_dm.py`: Run diffusion maps on AnnData object
    - **plot/**: Plotting functions for trajectory visualization
    - **util/**: Utility functions
    - **widgets/**: Interactive Jupyter widgets
- **tests/**: Test suite for package functionality
- **notebooks/**
    - `1-Gene-Expression-Processing.ipynb`
    - `2-Chromatin-Accessibility-Processing.ipynb`
    - `3-Multimodal-Omics-Data-Integration.ipynb`
    - `4-Gene_Covariance_Benchmark.ipynb`
    - `5-Peak_Trajectories.ipynb`
    - `6-Cell_Trajectories.ipynb`

Data Sources

The package includes scripts to download and process 10x Genomics PBMC data:

- 10k PBMCs, 3' v3 (scRNA-seq)
- 10k PBMCs, ATAC v1 (scATAC-seq)

These datasets can be downloaded using the provided shell scripts.

Dependencies

PeakTrajectory makes use of the following packages:

- scanpy
- anndata
- mudata
- numpy
- pandas
- scipy
- scikit-learn
- igraph
- pot (Python Optimal Transport)
- matplotlib
- seaborn
- tqdm

To simply replicate our conda environment, run `conda env create -f conda_env.yaml`

License

Acknowledgments

This package is adapted from the GeneTrajectory package developed by the Kluger Lab.