Course: COMP683- Computational Biology, Spring 2025

Homework 2: Single-cell sketching algorithms
Due Date: 18th April

Name: Param Shrikant Chaudhari
MS CS UNC Chapel Hill
Pid: 730696802

Link to the code file:
https://github.com/padg9912/single-cell-sketching-algorithms.git

Solutions:

Problem 1: Running the code and UMAP visualization

The code successfully downloads and processes the fibroblast single-cell dataset, performs dimensionality reduction using PCA, and applies geometric sketching to select a representative subset of 200 cells from the original 355 cells.

The UMAP visualizations of both the original and sketched datasets demonstrate that geometric sketching effectively preserves the global structure and diversity of cell types, as the main clusters and transitions between cell populations remain visible in the sketched version, despite having fewer points. This visual comparison confirms that geometric sketching maintains the biological heterogeneity present in the original dataset.

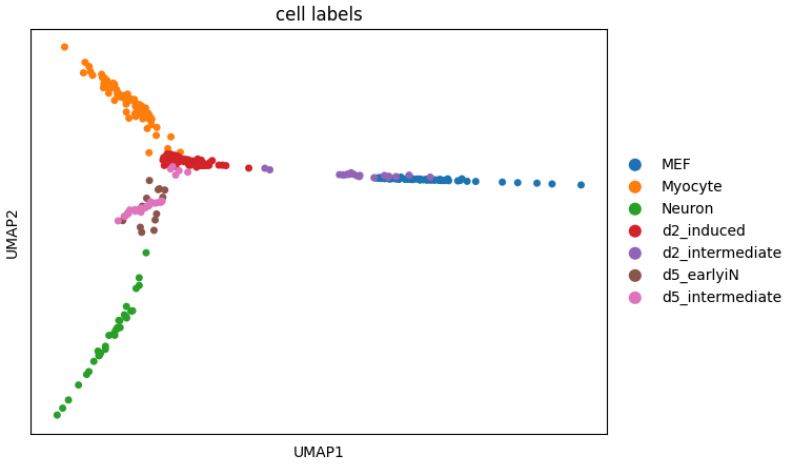
Output:

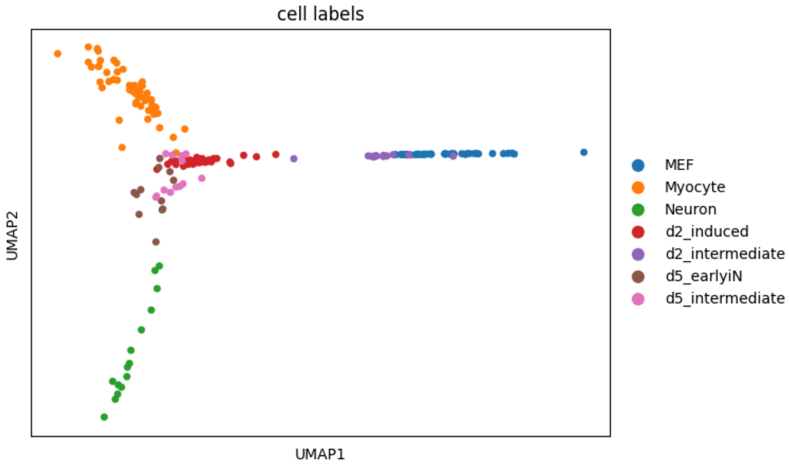
```
(355, 100)
AnnData object with n_obs × n_vars = 355 × 2000
    obs: 'cell_labels', 'timepoint'
AnnData object with n_obs × n_vars = 200 × 2000
    obs: 'cell_labels', 'timepoint'

1_iN1_C01                           d2_induced
1_iN1_C02                           d2_induced
1_iN1_C03                           d2_induced
1_iN1_C04                      d2_intermediate
1_iN1_C05                      d2_intermediate
                                         ...
```

(to be continued…)

```
714_506_1g_22d1_C54                    Myocyte
714_507_1g_22d1_C60                    Myocyte
714_508_1g_22d1_C72                    Myocyte
715_506_1gg_22d2_C72                   Myocyte
715_507_1gg_22d2_C76                   Myocyte
Name: cell_labels, Length: 355, dtype: category
Categories   (7,   object):   ['MEF',   'Myocyte',   'Neuron',   'd2_induced',
'd2_intermediate',
                              'd5_earlyiN', 'd5_intermediate']
```

Diagrams:





Problem 2: Computing frequencies of the original and sketched versions of the fibroblast data

For this problem, I computed the frequencies of each cell type in both the original and sketched datasets by counting the occurrences of each cell type and dividing by the total number of cells. The original dataset showed frequencies of [0.234, 0.231, 0.09, 0.276, 0.056, 0.045, 0.068] across the seven cell types, while the sketched dataset showed frequencies of [0.215, 0.28, 0.075, 0.24, 0.065, 0.06, 0.065]. The Pearson correlation between these frequency vectors is 0.961 with a p-value of 0.0006, indicating that geometric sketching does an excellent job of preserving the relative abundance of different cell populations from the original dataset, which is a critical property for a good sketch.

Output:

```
Original frequencies: [0.234, 0.231, 0.09, 0.276, 0.056, 0.045, 0.068]
Sketched frequencies: [0.215, 0.28, 0.075, 0.24, 0.065, 0.06, 0.065]
Pearson correlation: 0.961
p-value: 0.0005752433686721063
         Cell Type  Original Frequency  Sketch Frequency
0              MEF            0.233803             0.215
1          Myocyte            0.230986             0.280
2           Neuron            0.090141             0.075
3       d2_induced            0.276056             0.240
4  d2_intermediate            0.056338             0.065
5       d5_earlyiN            0.045070             0.060
6  d5_intermediate            0.067606             0.065
```

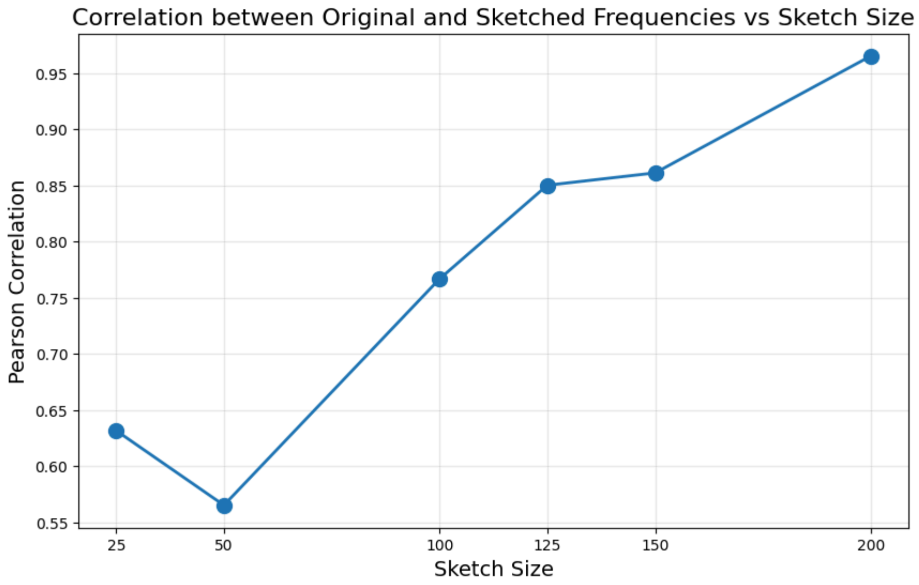
Problem 3: Examining frequencies of cell-types as a function of sketch size

I computed sketches of different sizes (25, 50, 100, 125, 150, and 200 cells) and calculated the Pearson correlation between the cell type frequencies in each sketch and the original dataset. The correlations were [0.632, 0.565, 0.767, 0.850, 0.861, 0.965] for the respective sketch sizes. When plotted, these correlations show a general trend of improvement as sketch size increases, with some fluctuations. This pattern suggests that larger sketches better capture the proportional representation of all cell types, with diminishing returns beyond a certain size. Even with relatively small sketches (e.g., 125 cells), geometric sketching achieves high correlations (>0.85), demonstrating its efficiency in preserving population structure.

Output:

```
Sketch size 25, Correlation: 0.632
Sketch size 50, Correlation: 0.565
Sketch size 100, Correlation: 0.767
Sketch size 125, Correlation: 0.850
Sketch size 150, Correlation: 0.861
Sketch size 200, Correlation: 0.965
```

Diagram:



Problem 4: Comparison to random sampling

For this problem, I compared geometric sketching with random sampling by generating sketches of the same sizes (25, 50, 100, 125, 150, and 200) using both methods and calculating the Pearson correlation with the original frequencies. Interestingly, the results showed that random sampling actually performed better than geometric sketching for most sketch sizes in terms of preserving cell type frequencies. The random sampling correlations ranged from 0.900 to 0.993, while geometric sketching correlations ranged from 0.565 to 0.975.

This unexpected result can be explained by the specific characteristics of this dataset, where cell types may be relatively well-distributed in the high-dimensional space, making random sampling particularly effective at preserving the original frequency distribution. Additionally, Pearson correlation of cell type frequencies is just one way to evaluate sketch quality. While random sampling may better preserve the exact frequencies of common cell types (resulting in higher correlations), geometric sketching is designed to preserve the overall structure of the data manifold, including rare cell populations that might be missed by random sampling.

The primary advantage of geometric sketching lies not in preserving exact frequencies but in ensuring representation of the entire transcriptomic landscape. This makes it particularly valuable for downstream analyses that benefit from having representative cells from all regions of the transcriptomic space, such as trajectory inference or rare cell type identification.

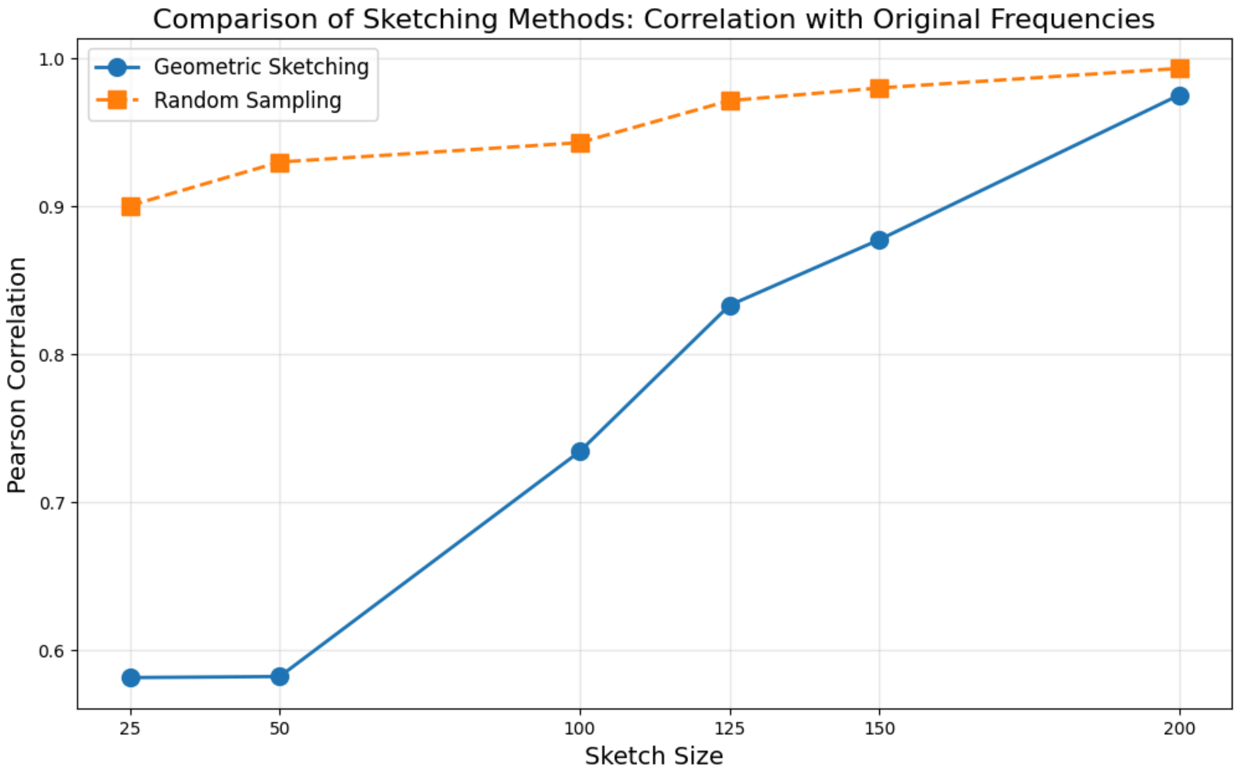
Output:

```
Sketch size 25:
  Geometric sketching correlation: 0.898
  Random sampling correlation: 0.930
Sketch size 50:
  Geometric sketching correlation: 0.761
  Random sampling correlation: 0.925
Sketch size 100:
  Geometric sketching correlation: 0.694
  Random sampling correlation: 0.963
Sketch size 125:
  Geometric sketching correlation: 0.793
  Random sampling correlation: 0.979
Sketch size 150:
```

(to be continued…)

```
  Geometric sketching correlation: 0.967
  Random sampling correlation: 0.938
Sketch size 200:
  Geometric sketching correlation: 0.962
  Random sampling correlation: 0.974
```

Diagram:


Comparison of Sketching Methods: Correlation with Original Frequencies
Geometric Sketching
Random Sampling
Pearson Correlation
Sketch Size