Homework 2
Due: Midnight, Mar 6st, 2022

Description:
1. We are going to use what we learned in the class to integrate real-world single-cell RNA-seq and single-cell ATAC-seq data. You will need to integrate the single-cell RNA-seq and single-cell ATAC-seq data you used for homework 1.
2. We are going to use the graph partition methods we learned to analyze real-world protein-protein interaction networks.

Write everything in a jupyter notebook. Write your code in the "code" section in jupyter notebook and your answers to the questions in the "Markdown" section. You can use python package to do the following tasks.

1. Data preparation (They are the data you used for homework 1):
1.1 Download the single-cell RNA-seq data from the GEO website:
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE126074
You need to download the following three files

Supplementary file	Size	Download	File type/resource
GSE126074_AdBrainCortex_SNAREseq_cDNA.barcodes.tsv.gz	48.3 Kb	(ftp)(http)	TSV
GSE126074_AdBrainCortex_SNAREseq_cDNA.counts.mtx.gz	28.3 Mb	(ftp)(http)	MTX
GSE126074_AdBrainCortex_SNAREseq_cDNA.genes.tsv.gz	93.7 Kb	(ftp)(http)	TSV

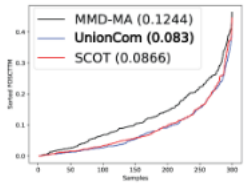
Hint: you can use scanpy.read_mtx to read in the data.
1.2 Download the single-cell ATAC-seq data from the GEO website:
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE126074
You need to download the following three files

GSE126074_AdBrainCortex_SNAREseq_chromatin.barcodes.tsv.gz	54.1 Kb	(ftp)(http)	TSV
GSE126074_AdBrainCortex_SNAREseq_chromatin.counts.mtx.gz	77.7 Mb	(ftp)(http)	MTX
GSE126074_AdBrainCortex_SNAREseq_chromatin.peaks.tsv.gz	1.9 Mb	(ftp)(http)	TSV

Hint: you can use scanpy.read_mtx to read in the data.

2. Integration (40points):
2.1 Use MMD-MA to integrate the data above.
2.2 Use SCOT to integrate the data above. How to use SCOT? Following the tutorial and examples in https://rsinghlab.github.io/SCOT/ .
2.3 Use FOSCTTM score the evaluate the performance. Here is how you should compute the FOSCTTM score: or each domain, we compute the Euclidean distances between a fixed sample

point and all the data points in the other domain. Next, we use these distances to compute the fraction of samples that are closer to the fixed sample than its true match. Finally, we average these values for all the samples in both domains. For perfect alignment, all samples would be closest to their true match, yielding an average FOSCTTM of zero. Therefore, a lower average FOSCTTM corresponds to better alignment performance. You are supposed to draw a figure similar as following.



2.4 Comment on the performance of the MMD-MA and SCOT.

3. Data preparation (10 points)
Download S. cerevisiae physical protein-protein interactions from https://dip.doe-mbi.ucla.edu/ . Generate a graph to represent protein-protein interactions and present it using its adjacency matrix A_1.

4. Graph partition (30 points)
4.1 Use mincut to partition A_1 into 50 clusters.
4.2 Use Normalized cut to partition A_1 into 50 clusters.
4.3 Use Modularity to partition A_1 into around 50 clusters.

5. Evaluation the partition performance (20 points)
Use Gene Ontology enrichment analysis to evaluate the graph partition results.
http://pantherdb.org/
5.1 Check how many clusters out of 50 are enriched in GO-slim biology for mincut results.
5.2 Check how many clusters out of 50 are enriched in GO-slim biology for Normalized cut results.
5.3 Check how many clusters are enriched in GO-slim biology for Modularity results.