Mapper for contact maps

October 4, 2021

Abstract

The goal of this project is to analyze a data set of single-cell Hi-C contact maps. An Hi-C contact map is a pairwise distance matrix that encodes how chromatin is folded in the nucleus of a cell: each row and column of the matrix represents a small DNA window, and each entry in the matrix is the spatial distance between these windows in the nucleus. This project aims at recovering the topological structure of the cell cycle (a loop) using Mappers computed on contact maps, processed with Stratum-adjusted Correlation Coefficients (SCC).

- Download the data set.

  https://drive.google.com/drive/folders/1bjEPIgchjLdDnoAO78Z-2VHhrvEyeImi?usp=sharing

  This folder contains the contact maps encoded in sparse COO matrices from the `SciPy` module. The rows and columns are small DNA windows, and the chromosome they belong to is encoded in the file ”chromosomes.txt”. The feature file containing the cell info is ”features.txt”.

- Take a look at the SCC article

  https://drive.google.com/file/d/1gREDKaqsonPAKTw_AiY5aNju4L-Vu8qM/view?usp=sharing,

  in particular ”2D mean filter smoothing”, ”Stratification by distance” and ”Stratum-adjusted correlation coefficient (SCC)” in ”Methods” section, and implement a function that takes as inputs two contact maps, and outputs the SCC.

- Implement your own Mapper algorithm. Alternatively, compile the `mapper` branch of `Gudhi`:

  https://github.com/MathieuCarriere/gudhi/tree/mapper

- Compute the Mapper of this data set using the eigenfunctions of a Kernel PCA run on the pairwise SCC matrix as filters. The data set is large, so you might want to subsample it. Try various Mapper and SCC parameters, and interpret their influence on the Mapper shape.

- Find a set of parameters for which the cell cycle can be detected as a big loop in the Mapper. Prove this loop indeed represents the cell cycle by coloring the Mapper nodes with various markers correlated with the cell cycle, such as ”mean insu”, ”f near band”, ”f mitotic band”, ”repli score” in provided feature file.

- Quantify the cell cycle statistical robustness (whether it is an artifact of computation or not) by bootstrapping the data: subsample the data set (with replacement) many times, and count the number of runs in which the cell cycle is detected.

- Compare and discuss your results with directly running dimensionality reduction on the raw contact maps.