Abstract

Dimensionality reduction is a crucial step in analyzing high-dimensional biological data, such as gene expression profiles from the Connectivity Map (CMap) dataset. Selecting an appropriate dimensionality reduction method can significantly impact the preservation of biological relationships and the interpretability of the data such as understanding of heterogeneity of cells among samples. Despite the existence of various techniques, there is a lack of systematic evaluations comparing their effectiveness in biological contexts.

This study aims to benchmark 30 different dimensionality reduction methods on subsets of CMap data. By assessing the preservation of biological similarity, clustering accuracy, and overall data structure, we seek to identify the most suitable methods for analyzing complex genomic datasets in 5 themes.