LABORATOIRE DE BIOLOGIE ET MODELISATION DE LA CELLULE

ENS - CNRS UCBL UMR 5239 - INSERM U 1210
46, allée d'Italie - 69364 Lyon cedex 07 – France

Tél : +33 4 72 72 81 71 - http://www.ens-lyon.fr/LBMC

Graph-Based non-linear dimension reduction for single cell transcriptomics

Recent technological advances in massively parallel sequencing and high-throughput cell biology technologies now give us the ability to describe population of cells with high dimensional molecular features. The so-called single-cell transcriptomic technology allows us to study cell-to-cell variability within a biological sample and investigate new questions like intra-tissue heterogeneity. Like many contemporary scientific fields, single-cell genomics raises new mathematical and computational challenges that are inherent to the massive production of large, high-resolution datasets that are complex and high-dimensional. In particular, unsupervised analysis is mandatory for researchers to handle the complexity of modern data, and machine learning methods known as dimensionality reduction have become a standard to reduce the size and complexity of data. Embedding high dimensional data into spaces with fewer dimensions is a central problem of machine learning, with the core motivation to preserve the intrinsic structure of the original data by keeping similar data points close and dissimilar data points distant in the low-dimensional space. In the literature, methods have been proposed, linear (like PCA) and nonlinear (Isomap, Locally Linear Embedding, Laplacian Eigenmaps). Among non-linear methods, tSNE and UMAP have been the most successful in proposing new representations that respect the local complex geometry of single-cell datasets. These techniques are now routinely incorporated in most analysis pipelines. They consist in embedding the original dataset in a 2D space by preserving the non linear dissimilarity between points thanks to a Kullback Leibler divergence between distances in the original and in the embedded space. In a recent work we proposed a unifying statistical and probabilistic framework that encompasses many non linear embedding methods, thanks to the coupling of random graphs that govern the proximities of observations. Our model provides a probabilistic interpretation of most used methods like tSNE and UMAP, by showing how they rely on specific prior hypothesis on the underlying connectivity structure. Our first results concern simple graph priors like bernoulli or fixed degree distribution, and the project is to benefit from our framework to generalize those methods to more complex topologies. Two research directions would be to consider very general priors, like scale-free networks or stochastic block model priors, that would have the advantage to introduce some clustering information in the model, which could constitute a powerful extension of our first framework to perform non linear dimension reduction and clustering at the same time.

The candidate will be co-supervised by Franck Picard (CNRS, ENS Lyon) Thibault Espinasse (Institut Camille Jordan, Lyon) and Julien Chiquet (INRA Saclay), experts in computational statistics and statistical learning. The candidate will work at the ENS de Lyon, in an interdisciplinary environment, between mathematics, computer science and biology. Moreover, the candidate will benefit from the SingleStatOmics ANR project that gathers an interdisciplinary consortium in machine learning / IA dedicated to single cell genomics, with experts in machine learning, optimal transport and statistics.References :

- A Probabilistic Graph Coupling View of Dimension Reduction, van Assel, Hugues and Espinasse, Thibault and Chiquet, Julien and Picard, Franck, Neurips 2022
- L.J.P. van der Maaten and G.E. Hinton. Journal of Machine Learning Research 9(Nov):2579-2605, 2008
- L. McInnes and J. Healy and J. Melville, 2020, arxiv 1802.03426
- Regev A, Teichmann SA, Lander ES, et al. The Human Cell Atlas. Elife. 2017; 6:e27041.
- Lähnemann, D., Köster, J., Szczurek, E. et al. Eleven grand challenges in single-cell data science. Genome Biol 21, 31 (2020).