Inferring cell state in disease progression from gene expression data using tractable deep probabilistic models

Abstract
In this cross-disciplinary project spanning machine learning, statistics, and biomedicine, the student will study a large class of tractable probabilistic models (TPMs), including probabilistic circuits[1] (PCs) and normalising flows[2] (NFs). PCs and NFs are expressive deep generative models as they encode several layers of latent variables into large graphs with potentially millions of connections and parameters. These models allow for exact probabilistic inference in time linear in the size of the model, which is necessary when dealing with thousands of gene expression variables across multiple cell types and conditions. The biomedical aim of this project is to identify and compare cell (sub)types and states in normal and disease tissue at large-scale, e.g. human cell atlases involving hundreds to millions of cells across multiple tissues and/or disease progression. This project is suitable for a candidate with a strong mathematical and/or machine learning background, as well as biomedical motivation. The project is a cross-disciplinary project, and will be supervised across 3 institutes: Institute of Genetics and Cancer and School of Informatics (Khamseh [quantitative biology]), School of Informatics (Vergari [deep probabilistic models]) and School of Mathematics (Beentjes [mathematical statistics]).

Introduction

Technological developments in high-throughput molecular phenotyping of single cells in the past ten years, such as single cell RNA sequencing (scRNA-seq)[3], allow researchers to construct disease biomarkers by capturing molecular changes in thousands to millions of cells. In order to attribute disease to cell type, and molecular features of cells to disease state, we need to define and distinguish cell types, sub-types and states. The ongoing Human Cell Atlas[4] project has taken a step in this direction, by seeking to define all human cell types and their molecular features, most often gene expression, within a multidimensional 'cell space'. Typing of cells is easiest when their lineages are well separated, and hardest when they are distinguished only by state (such as cell cycle phase, level of maturity, or response to stimulus) or spatial location. Approaches commonly used in the literature, including clustering, group cells based on proximity in expression space, thus yielding cell type definitions at relatively low-resolution.

Instead, much resolution can be gained by quantifying higher-order (i.e., beyond pairwise) dependencies across genes that are then interpreted as co-ordinated signatures leading to various cell states and types. Building on our previous theory work[5,6], we developed Stator[7], a novel cluster-free method that finely resolves cell types, subtypes and states among cells whose transcriptomes appear homogeneous upon clustering. Stator takes advantage of lowly-expressed as well as not-expressed genes, and can identify rare biological states (~0.2% of 10k single cells). The approach: (i) applies a model-free estimator of higher-order interactions to quantify expression dependencies amongst n-tuples of genes (beyond pair-wise), (ii) extracts significantly deviating combinatorial gene signatures (tuples) driving these higher-order gene dependencies, and finally (iii) combines tuples into Stator states when they commonly co-occur in the same cell. Typically, Stator labels a cell not just by type and sub-type but also by biological state, for example an immature interneuron in G2/M cell cycle phase.

Research Challenge

Stator is based on statistical techniques that are robust, but has a memory and size bottleneck: currently we can only analyse 1000 highly variable genes across 30,000 cells. This size of data is appropriate for single experiments, but it does not scale to multi-tissue-type atlases. Pilot

experiments indicate that normalising flows are successful at quantifying 3rd and 4th order dependencies using GPUs, when training the model on 20,000 scRNA-seq cells, at a fraction of the time of Stator. The project will begin by familiarising the student with training PCs and NFs on scRNA-seq data, evaluating higher-order interactions, quantifying uncertainties on these gene signatures and annotating cell types and states to benchmark their success in this biomedical application end-to-end. If successful, the model will be easily scalable to hundreds of thousands to millions of cells, solving the challenge of applying Stator to atlas-scale data. The best models will then be applied to disease progression scRNA-seq data in collaboration with industry partners.

As part of this PhD project, we will also develop and/or apply techniques in (causal) representation learning to extract relevant molecular features from multi-modal molecular data (e.g. scRNA-seq, scATAC-seq and imaging) for experimental hypothesis generation.

Data and Methodology

The starting data is an scRNA-seq developmental mouse brain, containing 10,000-40,000 cells with a variable number of genes (1000 or more, depending on the success of the methodology). The data has already been quality controlled and its biology has been well-understood, in order to allow for proper benchmarking of the deep probabilistic model above with the current state-of-the-art statistical techniques. Once the method is proven to be successful, it will be applied in various contexts, such as a pan-cancer cell state atlas and/or liver disease progression from normal to non-alcoholic fatty liver disease (NAFLD) to non-alcoholic steatohepatitis (NASH).

RRI/Ethical Considerations

All data used will be publicly available. The software will be made publicly available on GitHub.

Expected Outcome & Impact

Methodological outcome: Demonstrating the capabilities of deep tractable models in quantifying higher-order dependencies amongst a large number of variables.

Biomedical outcome: Creation of a large-scale human cell state atlas of disease progression which can be used by researchers in academia and industry.

The role of the external partner(s) in the project

The liver disease progression scRNA-seq datasets, that will be used to create the cell state atlas, have been gathered in discussion with the external partners as being relevant for R&D at Novo Nordisk. The researchers involved also have interest and are working on development of similar ML methodologies in this space.

References

1. Choi Y, et al., Probabilistic Circuits: A Unifying Framework for Tractable Probabilistic Models, http://starai.cs.ucla.edu/papers/ProbCirc20.pdf
2. Papamakarios G, et al., Normalizing Flows for Probabilistic Modeling and Inference, https://jmlr.org/papers/volume22/19-1028/19-1028.pdf

3. Haque A, et al., A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications, https://doi.org/10.1186/s13073-017-0467-4
4. Regev A, et al., The Human Cell Atlas, https://doi.org/10.7554/eLife.27041
5. Cossu G, et al., Machine learning determination of dynamical parameters: The Ising model case, https://doi.org/10.1103/PhysRevB.100.064304
6. Beentjes S, et al., Higher-order interactions in statistical physics and machine learning: A model-independent solution to the inverse problem at equilibrium, https://doi.org/10.1103/PhysRevE.102.053314
7. Jansma A, et al., High order expression dependencies finely resolve cryptic states and subtypes in single cell data, https://doi.org/10.1101/2023.12.18.572232