Abstract

Molecular biology explores the flow of information from DNA to cell function and seeks to understand complex biological processes on a molecular level. Driven by huge advancements in the field of omics technologies over the last few decades, we can now study biomolecules such as DNA, RNA, proteins, and metabolites on a large scale. With applications in diagnostic tools, therapeutic interventions, genetically modified organisms, and personalized medicine, molecular biology and omics research contribute to our quality of life in numerous ways.

The word "omics" encompasses fields as genomics, transcriptomics, proteomics, metabolomics, ... depending on the specific type of biomolecule studied. Each omic modality relies on high-throughput infrastructure to generate large-scale data sets, and advanced bioinformatics to unravel the molecular mechanisms underlying physiological processes, pathogenesis, and environmental responses.

A single omics experiment results in a complex, large, high-dimensional, and noisy data set, but provides a wealth of information. As manual exploration of these data sets has become cumbersome, the challenge today lies in detecting the useful information using computational tools. Throughout this dissertation, we inspect mass spectrometry imaging (MSI), multiplexed immunohistochemistry (mIHC), and chromatin conformation capture (3C) data sets. For all of these modalities, computational analysis can be very complex due to low signal-to-noise ratios, complicated normalization approaches, and crucial information often being hidden in the smaller details.

Tackling these challenges of omics data analysis, we researched the use of non-negative matrix factorization (NMF) for the unsupervised dimensionality reduction and clustering of the data sets in an explainable manner. NMF is an optimization algorithm which aims to approximate a non-negative data set by a number of linearly additive components, represented as a product of two low rank matrices. It requires that both the data and the factor matrices only contain non-negative numbers, which is often the case for omics data.

In the first part of this thesis, we focus solely on the analysis of MSI data. MSI studies proteins and metabolites in their micro-environment. It is an untargeted, label-free technology that measures the spatial distribution of thousands of molecules present in a slice of tissue. This makes MSI an essential tool for biomarker discovery and the study of heterogeneous cancers. For the analysis of MSI data, we started with a mathematical comparison of the performance of four different NMF-related methods for the analysis of multiple MSI data sets. We provided detailed information on the mathematical implications of each method and demonstrated how NMF approaches that assume Poisson noise, a known property of MSI data, perform better. Secondly, we inspected the Mapper algorithm, a topological data analysis approach, to visualize MSI data of mice pancreas. Using the optimal parameter settings, we discovered a more detailed structure in the data compared to uniform manifold approximation and projection (UMAP). More specifically, we detected two clusters within the islets of Langerhans that were not detected with UMAP. In the final chapter of the first part, we introduced an extension of NMF tailored for the analysis of MSI data. Using a hierarchical implementation of NMF, we gained more information from the small and convoluted signals. This new approach allowed us to identify the two clusters in the islets and see differences in abundance between the two types of insulin correlating with these clusters.

In the second part of the thesis, we extend the knowledge on NMF collected from MSI data to data sets obtained with other omics modalities. In mIHC, multiple protein markers are measured in-situ using fluorescence signals. While this is a targeted approach with a limited number of molecules measured, mIHC offers a much higher spatial resolution than MSI up to (sub)cellular level. Here, we presented a downstream analysis pipeline based on the combination of NMF and a trained classifier for the fast clustering and automatic labelling of human melanoma samples obtained with MILAN technology. Due to NMF's ability to handle large data sets, the proposed pipeline obtained better computational efficiency than previous analyses of this data set. Finally, we move to genomics data analysis. More specifically, we inspected the three-dimensional structure of the chromatin in single cells obtained from developing mouse brain. We applied multiple non-negative matrix and tensor factorization approaches to find the correlation between cell-types and connected parts of the chromatin. Unique chromatin contacts for a certain cell-type could indicate an essential protein encoding region, which could then be linked to the corresponding transcripts. Unfortunately, more advanced tools are required to uncover the true signals, since the high presence of noise misdirects traditional analyses. To this end, we highlight how factorization techniques can improve results and can be extended to multi-omic analyses.