Abstract

Advancements in high-throughput technologies have facilitated cost-efficient large-scale productions of biological data in numerous labs worldwide. The production and accumulation of data have led to a big data era – an era where researchers have unprecedented opportunities for learning about biology via bioinformatic analysis of the various and large datasets. But, with the new possibilities, comes a range of new challenges that need sophisticated solutions.

Single-cell RNA sequencing (scRNA-seq) data are a quintessential example of a big data type that depends on bioinformatic solutions. Though, many scRNA-seq-specific analytical methods have been invented to process and analyze the data, it has so far been an unexplored endeavor to identify mechanisms that drive the observed single-cell development trajectories. Another challenge that has emerged in the wake of the extensive data productions relates to the thousands of nucleotide variants that have been identified in recent years. The mechanisms by which they affect cellular dynamics are vastly unknown, but important to identify, as they may affect many cellular processes. It is a general bioinformatic challenge to discover patterns that can explain observed biological phenomena. But, with the extensive biological data available it is now possible to synergistically combine different data types and use these to propose new approaches that can predict the outcomes of complicated endogenous signaling pathways.

This thesis presents three manuscripts that introduce new bioinformatic methods and approaches, which seek to address the above-raised challenges. All manuscripts make use of techniques and concepts central to "big data analytics in bioinformatics". The first manuscript presents Scellnetor; Single-cell Network Profiler for Extraction of Systems Biology Patterns from scRNA-seq Trajectories, which is a novel clustering tool for scRNA-seq data. Scellnetor is implemented as an interactive webtool that allows researchers to select and compare single-cell development trajectories. We show that Scellnetor is able to find connected gene subnetworks essential for elucidating differences between distinct cellular development courses.

The second manuscript presents DeepCLIP; a convolutional LSTM (long short-term memory) neural network for analysis of binding preferences of RNA-binding proteins (RBPs). We demonstrate that DeepCLIP produces binding predictions and binding profiles that correlate strongly with *in vitro* and *in vivo* experiments and are sensitive to the effects of nucleotide variants on RBP binding affinity.

The third manuscript explores the potential role of cGMP-dependent protein kinase (PKG) as a reducer of damaging reactive oxygen species (ROS) formation post-stroke. After establishing a crosstalk between PKG- and ROS-signaling, we investigate it by developing a new approach to simulate the downstream transcriptional regulations that takes place upon kinase activity. We predict how expression of transcription factors that regulate gene expression of core ROS-forming enzymes is changed as a result of PKG-activity.

In conclusion, this thesis presents bioinformatic work that proposes solutions to the above-outlined challenges that have arisen with the advent of the big data era. The scientific contributions of this thesis include novel biological insights, new bioinformatic methods and freely available tools that can be readily applied to biological data. Additionally, the thesis includes suggestions to how the presented work might be improved by any researcher who wishes to extend it.