Big data analytics in bioinformatics

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.

Dansk resumé

Fremskridt indenfor high-throughput-teknologier har faciliteret omkostnings-effektiv generering af biologiske data i adskillige laboratorier verden over. Produktionen og akkumuleringen af biologiske data har ført os til en big data æra – en æra, hvor forskere har nye muligheder for at lære om biologi via bioinformatiske analyser af de forskellige og store datasæt. Men med de nye muligheder følger en række a nye udfordringer, der kræver sofistikerede løsninger.

Single-cell RNA sequencing (scRNA-seq) data er klasseeksemplet på en datatype, der er afhængig af bioinformatisk analyse. Selvom mange scRNA-seq-specifikke analysemetoder er blevet opfundet til at processere of analysere dataene, så har det indtil videre været et stort set uudforsket foretagende at finde mekanismer, der driver de observerede single-cell udviklingsbaner. En anden udfordring, der er dukket op i kølvandet på de omfattende dataproduktioner, vedrører de tusinder af nukleotidvarianter, der er blevet identificeret i de senere år. De mekanismer, hvormed de påvirker nedstrøms genprodukter og cellulære pathways, er stort set ukendte, men alligevel vigtige at identificere, da de kan påvirke og ændre mange cellulære processer. Det er en generel bioinformatisk udfordring at opdage mønstre, der kan forklare observerede biologiske fænomener. Men med de omfattende tilgængelige biologiske data er det nu muligt at synergistisk kombinere forskellige datatyper og bruge disse til at foreslå nye tilgange, der kan forudsige resultaterne af komplicerede endogene signalveje.

Denne afhandling præsenterer tre manuskripter, som introducerer nye bioinformatiske metoder og tilgange, der alle søger at tackle de ovennævnte udfordringer. Alle manuskripter gør brug af teknikker og koncepter, der er centrale for ”big data analytics in bioinformatics”. Det første manuskript præsenterer Scellnetor; Single-cell Network Profiler for Extraction of Systems Biology Patterns from scRNA-seq Trajectories, som er et nyt clustering-værktøj til scRNA-seq-data. Scellnetor er implementeret som et interaktivt webtool, der giver forskere mulighed for at vælge og sammenligne single-cell development trajectories. Vi viser at Scellnetor kan finde gen-subnetworks, der er essentielle for at belyse forskelle mellem forskellige cellulære udviklingsforløb.

Det andet manuskript præsenterer DeepCLIP; et convolutional LSTM (long short-term memory) neuralt netværk til analyse af bindingspræferencer for RNA-bindende proteiner (RBP'er). Vi viser at DeepCLIP producerer outputs, der korrelerer stærkt med *in vitro*- og *in vivo*-eksperimenter, og som er sensitive over for virkningerne af nukleotidvarianter på RBP-bindingsaffinitet.

Det tredje manuskript undersøger den potentielle rolle for cGMP-dependent protein kinase (PKG) som en reducer af den af skadelige reaktiv ilt-art (ROS) dannelse post-stroke. Efter at have etableret en crosstalk mellem PKG- og ROS-signalering, undersøger vi den ved at udvikle en ny tilgang til at simulere de downstream transkriptionelle reguleringer, der finder sted ved kinaseaktivitet. Vi forudsiger, hvordan ekspression af transkriptionsfaktorer, der regulerer

genekspression af core-ROS-dannende enzymer, ændres som et resultat af PKG-aktivitet.

Sammenfattende præsenterer denne afhandling bioinformatisk arbejde, der foreslår løsninger på ovennævnte udfordringer, der er opstået med fremkomsten af big data-æraen. Denne afhandlings videnskabelige bidrag inkluderer ny biologisk indsigt, nye bioinformatiske metoder og frit tilgængelige værktøjer, der let kan anvendes på biologiske data. Desuden indeholder afhandlingen forslag til, hvordan det præsenterede arbejde kan forbedres af enhver forsker, der ønsker at udvide det.