Software implementations allowing new approaches toward data analysis, modeling and curation of biological knowledge for Systems Medicine

John Zobolas

Last updated: 25 April, 2021

Contents

Preface

Abstract

Cancer is one the most prevalent human diseases. There has been considerable effort from the scientific community to understand the mechanisms behind this disease and search for treatments that promise a better quality of life for patients. In order to accomplish this goal, Biology and Medicine have joined forces with Computer Sciences, using the power of Computational modeling, Mathematics, Machine Learning and Statistics. This interdisciplinary effort to address the cancer problem, constitutes the basis upon which this thesis was formed. We present several contributions to this effort, consisting of software, data analyses and mathematical investigations, which have enabled the more efficient curation of biological knowledge, the use of computer models to prioritize drug treatments for cancer and the derivation of molecular mechanistic insights from the simulation results.

In order to build a computational model of a biological system such as a cancer cell, we first need a way to describe the structure of such a system. A common network-based approach provides an elegant representation of such a structure, where molecular entities such as proteins and genes are connected to each other via causal interactions, which in turn determine cellular behavior and the functional properties of the cell as an integrated system of individual components. These interactions form the Prior Knowledge Network (PKN), which serves as the basic building block for most computational biological models. Nonetheless, several challenges exist, even at this early stage of the modeling process.

The first problem is that biological information by its very nature is largely complex, and therefore its formalization to a structured, computable form for use in modeling applications, demands extra attention. The translation of scientific knowledge from publications into such a computable form is achieved with the use of specialized software tools and is the main responsibility of biocurators. In order to help biocurators be more efficient in their annotation tasks, we proposed the Visual Syntax Method (VSM) as an alternative approach for general-purpose knowledge formalization. In particular, we implemented a user-interface software component (VSM-box) that enables curators to

annotate any type of information, no matter its complexity, and translate it into an intuitive, flexible sentence-like format. This software was used to build a prototype curation interface (CausalBuilder) for the annotation of molecular causal interactions, which constitute the cornerstone of a model's PKN.

The second problem concerns the availability and ease of access of causal molecular interaction data for modeling or other scientific endeavors. A standard format for the representation of such signaling information was developed (causalTAB), and we supported the export of the causality statements from causalBuilder's interface to this format. But there exist several other molecular interaction databases that could update their data to fit the new causalTAB standard. PSICQUIC is a web-service platform that was initially built so that users can conveniently fetch in a standard way molecular interaction data from different sources. We extended PSICQUIC to incorporate the new causalTAB format, so that causality-enriched information generated by our curation prototype tool or from other data providers could be shared through a common channel.

A third major problem arose during the design process of the VSM-box and it's application, causalBuilder. Behind the scenes, the curator interface has to communicate with a large number of diverse biological data resources, each with its own online API service that provides access to the respective data. In order to present to the user the available terms that pertain to a specific annotation of interest, a uniform way to query all these resources was needed. This prompted us to build the Unified Biological Dictionaries (UBDs), a software suite that provides a unified gateway for life science data, helping users retrieve the right query terms. In addition, curators sometimes come across new knowledge that is not yet available through the standard authoritative resources. To address this related problem, we connected UBDs with PubDictionaries, an online resource of simple dictionaries, allowing curators to publicly create and share ad-hoc terms, and further use them as annotations in VSM-based applications.

After the signaling information has been curated and the causal interactions assembled to form the PKN, we then need to specify the mathematical equations of the cancer cell model. This allows us to describe and analyze its dynamical behavior subject to external stimuli, such as drug perturbations. The modeling approaches can in general range from qualitative to quantitative and in this work we focused on Boolean modeling, where signaling components are assigned either an active or inactive state. An automated computational pipeline was developed to produce an ensemble of Boolean models from a PKN, calibrated to a specific cancer cell signaling phenotype. These models are then analyzed to suggest possible synergistic drug combinations and the results are compared with experimental findings, where all possible combinations are tested in a high-throughput screen

setup. We demonstrated that our pipeline could prioritize specific drug combinations, reducing the number of drugs that need to be tested in experiments, before a viable treatment is found for a patient. Moreover, several analyses indicated that our models can be used to derive mechanistic insights about the diseased model and generate novel biological hypotheses. Lastly, we showed the significance of the PKN quality, where even small modifications to the cancer signaling network could severely affect our pipeline's drug prediction performance.

To exploit the range of parameterizations present in the Boolean models produced by our pipeline, we devised several strategies to split and compare the different models in a dedicated R package (emba). This supplementary effort allowed us to find potential biomarkers, which are nodes whose state is decisive for the global behavior of the models and can indicate parts of the PKN that are responsible for a drug combination to be synergistic. Additionally, we noticed particular patterns in the way specific equations always correspond to specific signaling states in our models, so we more deeply investigated the influence of the choice of parameterization on the output behavior of these nodes. This led us to propose a list of Boolean function metrics that can assist modelers in choosing more appropriate equations, meaning those that are consistent with the regulatory information present in the PKN and whose expected output better matches experimental observations. Finally, results from a study of diverse Boolean functions indicated that these also exhibit diverse output behaviors, with some being highly biased towards specific Boolean outcomes while others depending more on the ratio between positive and negative regulators, as these are derived from the two distinct types of causal interactions present in the model's PKN.