Use of prior knowledge in biological systems modelling

Reshetova, P.V.

Publication date
2017
Document Version
Other version
License
Other

Chapter 6
Discussion

Systems biology is a multi-disciplinary rapidly developing research field that focuses on complex (dynamic) non-linear interactions within biological systems such as biological networks. It generally involves a combination of wet-lab experiments and computational approaches. Experimental data is integrated in statistical or mathematical models to generate testable hypotheses, to predict the system's behaviour, and to facilitate discovery and description of the system's properties. A range of approaches towards the modelling of biological networks have been developed and according to Stelling and co-authors may be divided into three categories [159]. The first category involves methods based on interactions between network components only. These methods are often explorative and based on statistical approaches applied to genome-wide omics data such as discussed in Chapter 2 of this thesis. Examples include the construction of co-expression networks [160] and protein-protein interaction networks [161]. A second category consists of constrained based methods that aim to include information such as reaction stoichiometry and reaction reversibility. Flux Balance Analysis is an example from the second category [84]. We did not consider this type of modelling in our research. Finally, there are methods that include detailed interaction mechanisms, for example, ordinary differential equations (ODEs). ODEs are typically used to model the kinetics of metabolic networks [83] or cellular mechanisms such as discussed in Chapter 4 and 5 of this thesis. Interaction-based and constrained-based methods are static methods that do not require detailed parameters of the modeled network. In contrast, dynamic models such as represented by ODEs generally require such information to be applied successfully. Moreover, static models provide a qualitative description of the system dynamics in comparison to the quantitative results of ODEs. In our research we also considered network-based models (Petri nets) such as discussed in Chapter 3. These models can either be implemented as static models or as dynamic models.

All the modelling frameworks mentioned here can make use of prior biological knowledge either to define the model's topology and parameters in knowledge-driven modelling approaches or to guide the modelling process in data-driven approaches by directly incorporating the prior knowledge in the modelling method. In our research we demonstrated several approaches to accomplish this.

6.1. Prior knowledge in statistical models

In chapter 2 we reviewed more than twenty high-throughput data analysis methods in transcriptomics and metabolomics that incorporating prior knowledge to restrict or guide the statistical modelling. We highlighted features and differences of the methods and the type of prior knowledge that was used. However, it is extremely difficult to compare different methods without a proper framework which would allow a fair compar-

ison and would further facilitate understanding of how prior knowledge influences the results. Such framework could be based on an appropriate synthetic datasets. For example, a prototype of such test framework can be based on the Dialogue on Reverse Engineering Assessment and Methods (DREAM) project. DREAM aims to provide a framework for a fair and rigorous testing of various gene network inference methods. DREAM suggests to infer a gene network from simulated gene expression data without sharing the details of the kinetic model that was used to generate the data [162]. Various error tests allow to measure a performance of every method and to show the difficulties faced by each of the applied methods. For example, if two genes are co-regulated by a single transcription factor their transcription levels correlate. This correlation may lead to a false prediction of an interaction between the two co-regulated nodes. Analysis of such network motifs helps to reveal systematic prediction errors of the tested methods. Further, based on the used synthetic network, prior knowledge may be generated, perhaps with variable percentage of missing or false positive gene interaction links. This would allow both to compare different methods that incorporate prior knowledge in network inference and to assess the added value of the prior knowledge in each method.

Despite that synthetic data certainly may offer a valuable basis for a validation framework, wet-lab experiments remain highly desirable. Few projects present an example of a possible experimental system where an artificial network has been constructed and incorporated into a cell. For example, a synthetic gene network in yeast as a part of a test framework for systems biology approaches has been presented by Cantone and co-authors [163]. Another example is a synthetic gene network integrated in human kidney cells by Kang and co-authors [164]. These examples present gene networks with known interactions and thus suitable to generate wet-lab experimental data for further use in a validation framework.

6.2. Prior knowledge to model genistein elimination pathway with Petri nets

Petri nets are used to study the dynamics of biological systems (for a review see [165]). Similar to ODEs, Petri nets provide a formal mathematical framework for the analysis of biological systems. Various extensions for Petri nets have been developed to qualitatively or quantitatively model networks. Basic (also referred as original or time-less) Petri nets require only the topological structure and stoichiometry of the studied network. While they provide some insight in possible dynamics the result of the analysis is qualitative [77, 166]. Further two examples are based on the assumption that kinetic parameters are less important than network topology and therefore topology based models are able to provide information about system dynamics, thus providing quantitative insights. A method of Ruths and co-workers uses this assumption to develop a strategy for non-parametric Petri net modeling and execution that uses token distribution and sampling to reproduce the dynamics of cellular signaling networks [81]. A method of Kuffner and co-workers is based on fuzzy logic and provides a very good estimation of gene regulatory networks from gene expression data *in silico* [19]. The authors argued that their Petri net extension provides a simpler discrete modelling system compared to more detailed ODEs. Examples of Petri nets that aim to quantitatively model networks comprise

Stochastic Petri nets [16], Time Petri Nets [17], and Hybrid Functional Petri Nets [18]. Similar to ODEs they require kinetic parameters of the system.

However, so far, it seems to be ignored that even in the absence of kinetic parameters Petri nets may directly be converted to differential equations if all the kinetic parameters can be obtained through parameter estimation procedures. Using the same prior knowledge, our Petri net of the genistein elimination pathway can also be modelled with ODEs:

$$
\begin{aligned}
\frac{dG_{GL}}{dt} &= -(F1+F7)G_{GL} + F11G_L \\
\frac{dG_{GE}}{dt} &= -(F2+F30+F31)G_{GE} + F1G_{GL} \\
\frac{dG_L}{dt} &= -(F3+F11)G_L + F2G_{GE} + F6G_{VB} \\
\frac{dG_A}{dt} &= -F4G_A + F3G_L \\
\frac{dG_O}{dt} &= -(F5+F8)G_O + F4G_A \\
\frac{dG_{VB}}{dt} &= -F6G_{VB} + F5G_O \\
\frac{dGG_{GL}}{dt} &= -(F17+F16)GG_{GL} + F20GG_L \\
\frac{dGG_{GE}}{dt} &= -F12GG_{GE} + F16GG_{GL} + F30G_{GE} \\
\frac{dGG_L}{dt} &= -(F13+F20)GG_L + F12GG_{GE} + F19GG_{VB} \\
\frac{dGG_A}{dt} &= -F14GG_A + F13GG_L \\
\frac{dGG_O}{dt} &= -(F18+F15)GG_O + F14GG_A \\
\frac{dGG_{VB}}{dt} &= -F19GG_{VB} + F15GG_O \\
\frac{dS_{GL}}{dt} &= -(F25+F26)S_{GL} + F29S_L \\
\frac{dS_{GE}}{dt} &= -F21S_{GE} + F31G_{GE} + F25S_{GL} \\
\frac{dS_L}{dt} &= -(F22+F29)S_L + F12S_{GE} + F28S_{VB} \\
\frac{dS_A}{dt} &= -F23S_A + F22S_L \\
\frac{dS_O}{dt} &= -(F24+F27)S_O + F23S_A \\
\frac{dS_{VB}}{dt} &= -F28S_{VB} + F24S_O
\end{aligned}
$$

All parameters in this ODE model can then be estimated through parameter estimation based on experimental data as was done with the Petri net model (Chapter 3). Then the question is what would be the advantage of the Petri net model over the ODE based model? Also if both methods allow the dynamic analysis then what would be the difference between results of two methods? Additional work is required, which would compare genistein elimination Petri net and ODE based models. This work hopefully would lead to a grounded advice when to choose Petri nets over ODEs and vice versa in the situation of lack of the exact topology and kinetic parameters.

6.3. Prior knowledge to model B-cell affinity maturation with differential equations

The B-cell affinity maturation has been intensively studied for a few decades, however, its precise mechanism still remains to be elucidated. For modelling a Germinal Centre Reaction (GCR) in Chapter 4 and 5 the known details were not enough to set equations that would precisely describe the crucial B-cell selection mechanisms. For example, following cell division and somatic hypermutation, B-cells are programmed to undergo apoptosis unless they receive survival signals through interactions with the antigen and T follicular helper cells. These selection mechanisms impose competition between the B-cell subclones, which is assumed to be based on their relative BCR affinities. The precise mechanisms involved in B-cell competition is unknown. To avoid assumptions due to the lack of knowledge and to avoid an overly complex model, antigene and T follicular helper survival signals were modelled with a general sigmoidal function. This function converts relative B-cell affinities to a signal strength between 0 and 1. Despite this simplification, the model successfully generated valuable insights in B-cell affinity distribution after affinity maturation.

6.4. Databases as stores of prior knowledge

There are many sources of knowledge that may be used as prior knowledge in the analysis of biological data and systems. Perhaps the most widely used sources are expert domain knowledge and traditional (low-throughput) experiments that are very precise and reliable. However, knowledge from an expert can only be used in data analysis if we capture his knowledge in a computer accessible format, which is time consuming and requires significant effort for a stable collaboration and to ensure that the domain expert also benefits from such effort. A store of knowledge with easy access may facilitate collaboration with domain experts. Moreover, a database may help to structure knowledge from a large range of interdisciplinary studies, which otherwise would be too comprehensive and complex to be absorbed by one mind.

Several technologies have been suggested to support biological databases varying from saving data in a comma separated file to complex object-oriented databases [167, 168]. One of the most widely used technologies to store biological data is a relational database [169]. Some authors emphasize that relational databases provide a straightforward way to think about biological knowledge in the information way and allow to facilitate collaborations between experts and to cover large areas of knowledge

[170]. Many relational biological databases with a variety of content, purpose, and technical characteristics have been created [171].

Recently, a new promising approach, i.e., the Resource Description Framework (RDF) technology has been suggested [172]. A large effort to standardize RDF has been taken by W3 community (https://w3.org/RDF/), which greatly facilitates the use of the technology. Moreover, RDF does not require a fixed list of biological entities and relations as is required by other standards. The format flexibility combined with the standardization effort makes it easier to create biological data stores, to share and integrate data among various fields and to promote database evaluation over time [173, 174]. As a result, this technology has been successfully applied in systems biology to create databases that facilitate the decision making procedure and experimental data analysis. Venkatesan and co-authors created the Gene eXpression Knowledge Base (GeXKB) - a database that contains integrated knowledge about gene expression regulation [175]. The flexibility of the technology allowed to integrate the database and experimental gene expression data to explore new potential candidates for regulatory network extensions. Willemsen and co-authors used the technology to create a comprehensive knowledge base which focuses on four key peroxisome pathways and several related genetic disorders in humans [176]. The authors particularly have focused on creating a general framework to construct biomedical knowledge bases from scattered resources and on its visualization aspects.

Our research would greatly benefit from a knowledge base containing details of genistein elimination pathway (Chapter 3) or B-cell maturation process (Chapter 4). Considering the capacity of knowledge bases to organize and share data, serve as an expert communication platform, and to facilitate knowledge visualisation, the creation of such a knowledge base may be advised as a first step in biological systems modelling.

6.5. Biomedical text mining as a source of prior knowledge

Without doubts, scholarly documents provide the biggest source of scientific knowledge: millions of scholarly documents are currently available from literature databases (e.g. PubMed) and other repositories [177]. Such huge amount of data makes the manual extraction of relevant information extremely time consuming. Text mining provides a solution to this problem. In general, this scientific field is called natural language processing (NLP). The main idea behind NLP is to define and recognize terms used in the domain of interest, to create a collection of these terms together with annotation describing their meaning, and to use this collection to analyse a large corpus of text looking for co-occurrence among terms identified in the text. Further, co-occurrence is used in various statistical tests to find relations among terms and consequently suggest a relation among corresponding biological entities. Text mining may be used as a stand alone tool to discover relationships among biological terms or to create biological databases for various purposes [178, 179, 180, 181]. Moreover, text mining has been used to create advanced search engines that aim to speed up and facilitate literature search for specific topics [182, 183]. Use of such search engines would greatly improve time spend on knowledge gathering during modelling process. Worthy to notice that biomedical sources used in text mining are not limited by scientific literature in biology, medicine, and chemistry. Information from medical internet communities and patient records also

contain valuable knowledge as well and some interesting applications already have been suggested (e.g., [184, 185, 186, 187]).