Use of prior knowledge in biological systems modelling

Reshetova, P.V.

Publication date
2017
Document Version
Other version
License
Other

Chapter 1
Introduction

An enormous amount of biological knowledge has been generated by the scientific community and is available from a large number of biological databases, scientific literature, and domain experts. This knowledge is actively used to define new hypotheses and to validate new findings, but it may also be included in computational modelling and high-throughput data analysis as prior knowledge in order to improve the analysis or guide it towards meaningful solutions. However, the use of prior knowledge is not always straightforward and may additionally be hampered by its incompleteness. Moreover, the use of prior knowledge may bias the results towards known biology thereby preventing new findings. In this thesis we explored the use of prior knowledge in data-driven and knowledge-driven modelling approaches for high-throughput data analysis and biological systems modelling. In the first part of this thesis we reviewed methods that incorporate prior knowledge in statistical models for the analysis of high-throughput transcriptomics and metabolomics data (Chapter 2). We highlighted characteristics and differences of this methods and the type of prior knowledge that was used. In the second part of this thesis we used prior knowledge to model two biological systems. First, we used sparse prior knowledge to build a network-based model of a multi-organ genistein elimination pathway that can assist in the design of new experiments (Chapter 3). Secondly, we developed a mathematical model of B-cell affinity maturation based on incomplete prior knowledge about the selection of high affinity B cells. Here, we used prior knowledge to simplify the mathematical description of the B-cell selection process to avoid excessive model complexity. We showed that despite this simplification the model generated valuable insights in the affinity distribution among (un)expanded subclones (Chapter 4). Further, the model was used to identify changes in B-cell lineage trees during affinity maturation (Chapter 5). In summary, we explored possibilities to facilitate high-throughput data analysis with prior knowledge , and demonstrated the use of prior knowledge in biological systems modelling.

1.1. Modelling approaches

The modelling of biological systems is an essential part of nowadays research in systems biology. It aims to abstract a biological system in a statistical or mathematical modelling framework and to subsequently apply computational methods to determine (emerging) properties of the system. The work in this thesis aims to show how the modelling of biological systems can benefit from prior knowledge, and also how prior knowledge can be incorporated in the modelling procedure. We considered three types of models that either incorporate prior knowledge directly in the model computation or have been constructed by using prior knowledge:

- Statistical models for high-throughput data analysis;

- Network-based models of biological systems;
- Mathematical models.

Statistical models aim to find and quantify relationships between the variables in a dataset. These models may or may not assume a distribution of the variables under investigation. An example of a statistical model is a (linear) regression model that describes a relationship between one or more explanatory variables and a dependent variable. Other examples of statistical models include component models such as principal component analysis [1] or cluster methods such as k-means clustering [2] that also aim to find associations between objects (e.g. samples) and variables (e.g. genes). However, these statistical models generally do not explain the precise nature of relationships between variables in terms of biological processes. Hence, they are phenomenological. In contrast, mathematical models, such as differential equations, use equations to specify a mechanistic model, that is, the nature of the relationships between, for example, genes is explicitly specified. Moreover, the parameters in such model have biological definitions. Network-models such as Petri nets [3], Bayesian networks [4], and Boolean networks [5, 6] live between the phenomenological and mechanistic models. These models do not necessarily include a full mechanistic description of the biological system but specify relationships between objects in the model more explicitly then is the case in statistical models. Models reviewed or used in our research involve various statistical models, Petri nets, and ordinary differential equations.

Modelling approaches may further be divided in two categories: data-driven and knowledge-driven (Figure 1.1). Data-driven approaches include statistical models and network-based models to analyse high-throughput experimental data such as coming from transcriptomics and metabolomics studies in which many genes and metabolites, respectively, are measured. In this type of modelling one is generally interested in identifying (linear) relationships or correlations between the variables and, therefore, the models do not use any *a priori* known facts about the modelled biological system. However, due to the high data dimensionality (many variables are measured compared to the number of samples), these models may reveal chance correlations (i.e. spurious correlations, which are found for a specific data set but have no biological relevance). As a solution, incorporation of prior knowledge has been suggested, which may also improve the interpretability of the results by focusing them on known biology. For example, prior knowledge has been used to softly penalise the minimization function in a principal component based method in order to find principal components that are partially defined by already known information [7]. We reviewed a range of methods that followed this strategy in Chapter 2 [8].

Knowledge-driven approaches comprise methods that use known biological facts to define the model structure (e.g. equations and parameters in a mathematical model, or the topology in a network-based model) and, therefore, depend on literature, information from public biological databases, and/or knowledge provided by domain experts. Knowledge-driven approaches include mathematical models based on ordinary differential equations, and network-based models such as Petri nets, Boolean, and Bayesian networks. Both mathematical and network-based models are widely used to model relations between molecules or cells. Usually, network-based methods are preferred if quan-

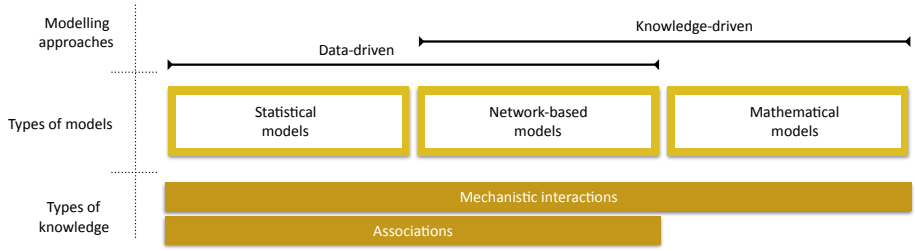

Figure 1.1: Conceptual representation of data-driven and knowledge-driven modelling approaches. Three types of models are considered, which are shown in the relation to two types of prior knowledge. Among data-driven modelling approaches we consider only models that incorporate prior knowledge.

titative values for parameters, such as required for differential equations, are difficult to obtain. While knowledge-driven approaches construct a model solely based on *a priori* available knowledge they generally use (public)experimental data at a second phase for parametrization and validation.

It is important to note that network-based models are successfully employed in both knowledge and data-driven approaches, which is indicated by the overlap of these approaches in Figure 1.1. For example, using a knowledge-driven approach Schlatter and co-authors have built a literature-based Boolean model to study a set of pathways involved in apoptosis [9]. In contrast, Sahoo and co-authors used a data-driven approach to construct a genome-wide Boolean model of gene pairs from a comprehensive set of microarray experiments [10].

The reviewed types of models have different requirements to the used prior knowledge (Figure 1.1). Mathematical models require detailed specifications of biological mechanisms (mechanistic interactions) to construct equations. In addition, some parameters in these models also have to be based on *a priori* biological knowledge (while other may be estimated from experimental data). For network-based models the same is true although, in general, less detailed information is required and the model also contains fewer parameters that do not necessarily have to reflect biological values. Statistical models and network-based models (used in the data-driven approaches) are less demanding and have capability to incorporate a wide variety of prior knowledge. Prior knowledge may come from associations between biological entities such as the regulation of gene expression by transcription factors, or the co-expression of genes. Associations may also come from pre-defined groups of biological entities such as gene products (defined in the gene ontology; GO)[11] that participate in the same pathway. Associations do not present mechanistic details and can not be incorporated in mathematical models.

Statistical models for high-throughput data analysis

High-throughput experimental techniques characterize each samples by thousands of variables (e.g., genes, proteins, metabolites) and potentially allow to reconstruct the underlying gene, protein, and metabolic networks involved in biological processes. This attractive property secured a place for high-throughput experiments in biomedi-

cal research. Consequently, a new subfield of bioinformatics has emerged that applies network-based and statistical models to the data. However, these methods have to handle specific features of the high-throughput data. Particularly, these methods do not always distinguish between biologically significant correlations and chance correlations and consequently lead to spurious findings. Moreover, biological differences among technical and biological replicates in high-throughput experimental data may be not the primary interest but may also significantly challenge the data analysis [8]. To prevent spurious findings, the use of prior knowledge has been suggested to restrict or guide the statistical modelling. We dedicated a chapter of this thesis to give a fairly broad overview of such methods in transcriptomics and metabolomics (Chapter 2).

Challenges using prior knowledge in high-throughput data analysis

Incorporation of prior knowledge in statistical analysis of high-throughput data guides the analysis towards known biological relationships and thereby reduces the detection of spurious relationships among variables. The improved separation of biologically relevant variation from the noise in the data could potentially lead to enhanced discovery of new biology. However, the amount, nature and quality of prior knowledge may drastically influence the quality of the resulting models and their ability to assist in exploring the system. Moreover, prior knowledge incorporated in the analysis may even bias the results towards the expected biology and thus leading to false positive detection of the expected relationships suggested by the prior knowledge, but not present in the data. Thus, there is a delicate balance between data-driven and knowledge-driven analysis. Unfortunately, there are no guidelines or a credible unified indicator that can help with this. We will discuss this topic in more details in Chapter 2 of this thesis.

Another important issue with prior knowledge is so-called negative prior knowledge [12]. While positive prior knowledge refers to known interactions (of any nature) between two biological entities, negative prior knowledge would reflect truly non-existing interaction between two entities. However, such non-existing interactions are generally not explicitly specified in literature or public databases making it hard to distinguish between interactions that truly do not exist and interactions that remain to be discovered and described. Non-existing interactions may be included in modelling approaches by defining which entities do not interact and therefore any identified correlations between them can be considered as spurious.

Another issue with prior knowledge and methods that incorporate prior knowledge is the evaluation of its added value. Currently, this problem has been addressed only by few authors. For example, in the work of Tian and co-authors “random” knowledge has been considered to evaluate the prior knowledge influence on the inference of gene interaction networks from high-throughput genomic data [13]. This allowed them to conclude that using real prior knowledge in their method was robust to false positive interactions between genes. Other research forwarded a general framework to investigate the relevance of different prior knowledge sources (such as databases, literature, and *a priori* gene co-expression experiments) for inferring gene interaction networks [14]. In our opinion, the evaluation of the added value of prior knowledge and the evaluation of the relevance of different prior knowledge sources requires a well-thought-of universal test framework including appropriate (synthetic) datasets. This would also greatly improve

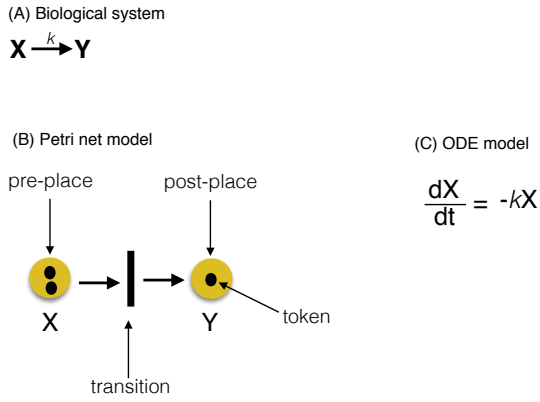

Figure 1.2: Example of (A) a biological system (chemical reaction): molecule X is converted to molecule Y with rate k. (B) Petri net model describing this chemical reaction. This Petri net contains two places and one transition. Firing of the transition moves token(s) from "pre-place" X to "post-place" Y. (C) ODEs model corresponding to the chemical reaction. The concentration change dX (and therefore dY) in time is proportional to the concentration X.

application and development of statistical approaches that include prior knowledge [8].

Network-based models of biological systems

A variety of network-based models have been developed including Petri nets [3], Boolean [5, 6] and Bayesian networks [4]. These models represent biological systems as a graph and allow to naturally resemble biological pathways. Network-based approaches allow to model a system in a range of abstraction levels depending on available prior knowledge or on the model purpose. In Chapter 3 of this thesis Petri nets were used to study the dynamics of a multi-organ genistein elimination pathway and to assist in the design of new experiments. Petri nets are bipartite graphs with two types of nodes: places and transitions [15, 3]. In Chapter 3 Petri net places represent molecules (metabolites) and transitions represent interactions between connected nodes (enzymatic reactions or metabolite transport). In addition, places contain tokens that represent the relative concentration of the corresponding molecule (Figure 1.2). During a simulation several steps are executed. At each step a transition 'fires' and tokens are moved from pre-transition places to post-transition places. A set of firing rules (heuristics) define which transition fires at each step and, in combination with topology of the network, allows to reproduce the dynamic behaviour of a real system. The unique feature of Petri nets is that various firing rules may be assigned to transitions to represent system dynamics in a desirable level of details and abstraction. This resulted in a wide range of Petri nets based methods such as Stochastic Petri nets [16], Time Petri Nets [17], Hybrid Functional Petri Nets [18] and Petri nets that incorporate Fuzzy logic [19] and even ODEs [20] providing a wide modelling capability.

Together with Petri nets, Boolean and Bayesian networks are widely used to model biological systems. Although Petri nets were used in this thesis, Boolean and Bayesian network-based approaches also play an important role in biological systems modelling.

To model biological pathways both approaches represent the biological system as a graph in which every node represents a molecule (e.g., protein, gene, and metabolite) and every edge represents a defined interaction between two molecules. A variety of interaction types may be represented: directional and undirectional, signed (inhibition/activation) and unsigned, a physical binding (e.g., binding of regulatory molecules), or correlations of gene expression. The simplest representation of biological systems is provided by Boolean network models [6]. They represent qualitative behaviour and may be used to model biological systems with no or sparse prior knowledge about quantitative parameters. Despite their simplicity Boolean networks have been widely used to study various signalling pathways and their properties [21, 22, 9, 23]. In Boolean models, each node has a Boolean state (0 or 1). Each edge holds a rule from the set of AND/OR/NOT values and this defines the interaction between two nodes. By changing initial node states or edge rules different hypotheses may be investigated. However, because Boolean networks only hold binary values they are strongly limited in the representation of continuous values and time. If a lower abstraction level is needed to model a system then Bayesian networks may be employed. Instead of a set of AND/OR/NOT rules Bayesian networks assign probabilities to node interactions. The probabilistic representation of relationships in a model is believed to be suitable to handle biological and experimental noise in high-throughput data and allows to combine Bayesian networks with analysis of high-throughput microarray data [24, 25, 4, 26, 27]. Further research led to the development of discrete multi-state models, which allow to assign multiple states to nodes or values between 0 and 1 and, therefore, allow to model sensitivities or concentrations of molecules [28]. Finally, discrete statements are transformed to continuous values of input and output in fuzzy logic models [29]. The time limitations of Boolean networks have been overcome by advanced approaches such as continuous or mixed discrete continuous Boolean networks [30, 31], and probabilistic Boolean networks [32, 33]. However, a lower abstraction level (i.e., more detail) and therefore a higher complexity of models require higher computational power as well as more parameters to be estimated.

Mathematical models of biological systems

Ordinary differential equations (ODEs) provide one often used framework to specify mathematical models of biological systems. To specify ODEs prior knowledge about the biological system is used to define the dynamics and relations between all biological entities involved (e.g. genes, metabolites, cells). Relations in an ODE may, for example, represent chemical reactions or cell differentiations (Figure 1.2). Solutions of the resulting set of equations describe or predict the temporal or spatial dynamic behaviour of the system. Because differential equations allow to represent non-linear behaviour they are widely used to model nontrivial dynamics like limit-cycle oscillations and multi-stability [34]. Moreover, while most sets of differential equations cannot be solved analytically, numerical methods are well developed and supported by various computational tools. Together the possibility to model a broad spectrum of biological systems behaviour and availability of various computational tools promote the use of differential equations. Therefore, we choose differential equations to model the complex behaviour of B-cell affinity maturation in Chapter 4. The resulting model allowed us to follow a large set of

B-cell subclones individually and, consequently, to follow affinity change in the context of B-cell lineage trees (Chapter 5). As a result, the model expanded our understanding of B-cell repertoire sequencing experimental data.

Several types of differential equations have been developed and applied in biological systems modelling. Widely used are mathematical models based on ordinary differential equations (ODEs) [35, 36]. They can be used to describe time dependent concentration or signal changes. To also model spatial dynamics partial differential equations (PDEs) may be used [37]. For example, Smith and co-authors used PDEs to reflect protein diffusion through the cytosol [38]. Another feature of biological systems is that stochastic processes inherent or external to the system may affect their behaviour. Individual differences in hormone levels or diet of subjects as well as small differences in experimental procedures like temperature may produce variations in experimental results. To address this dynamic variations stochastic differential equations (SDEs) have been applied [39]. For example, Chen and co-authors used SDEs to address stochasticity in gene regulations by transcription factors [39].

Despite their capability to model and analyse the dynamic behaviour of biological systems differential equations are not always the first choice. Particularly, the model output may strongly depend on its topology and corresponding parameter values and, therefore, a good knowledge about the biological system is required, as well as precise quantitative measurements for all parameters involved. Sometimes parameter values can be obtained from scientific literature or public databases. However, these parameter values may have been obtained under different experimental conditions that do not exactly match requirements for the new model. Alternatively, experiments to directly measure the parameter values may be conducted but this is usually time consuming if many parameters are needed, or may even be impossible. In such cases parameters may sometimes be estimated from experimental (time series) data that match the output of the model. However, parameters may be non-identifiable if the experimental data is insufficient to unambiguously determine all parameters [40, 41]. Consequently, for large and complex biological systems the parameterization of the model may become challenging and computationally expensive.

1.2. Scope and outline of the thesis

This thesis explored the use of prior knowledge in modelling approaches for high-throughput data analysis and biological systems modelling. Because the main interest was in how prior knowledge might be used in the analysis of biological systems in general, no restrictions on the specific types of models were specified. Therefore, three types of models for various applications were used. Firstly, the use of prior knowledge in data driven statistical models of high-throughput data was reviewed. Secondly, a Petri net model of human genistein elimination pathway was build based on sparse prior knowledge. Finally, ODEs were used to model B cells affinity maturation during an immune response using prior knowledge about affinity maturation to construct and eventually simplify the mathematical model.

Use of prior knowledge in the analysis of high-throughput data is an emerging field which is hoped to boost our understanding of biological systems on a big scale. The research started by a review of more than twenty high-throughput data analysis methods

in Chapter 2. The set includes methods of high-throughput data analysis in transcriptomics and metabolomics that use prior knowledge to define or to estimate statistical model parameters. Because prior knowledge may be incomplete, incorrect, or may hide new discoveries, specific attention was paid to the problem of balancing experimental data and prior knowledge. It was concluded that for further understanding of the influence of prior knowledge on the analysis, and for comparing different methods to incorporate prior knowledge, a well-defined test framework is required.

In Chapter 3 of this thesis very scarce and incomplete knowledge about a complex multi-organ genistein elimination pathway that comprises several concurrent routes was used. A Petri net based model was suggested that relied on topology alone to reconstruct metabolite concentration profiles. Furthermore, this model was used as an experimental design tool to propose new metabolite measurements to more precisely infer the relative contributions of concurrent elimination routes, and to improve the reconstruction of concentration profiles of all metabolites in this pathway. Overall it was shown that a Petri net model based on scarce prior knowledge may be used to assist in the design of future experiments to complete missing knowledge.

The second application of biological systems modelling aimed for better understanding of high-throughput B-cell repertoire RNA sequencing data. A mathematical model was developed, based on ordinary differential equations, to simulate B-cell affinity maturation during an immune response to determine the affinity distribution in (un)expanded B-cell subclones (Chapter 4) and to study the evolution of B-cell lineage trees during affinity maturation (Chapter 5). In this case available prior knowledge was used to define a simplified representation of the B-cell competition process (survival and positive selection). The simplification avoids an overcomplicated model but sufficient realistic to allow further interpretation of repertoire sequencing experiments.

This thesis is closed with Chapter 6 where some open issues are discussed and future research opportunities are suggested that could move forward the analysis of biological systems and experimental data with the use of prior knowledge.