Computational Reconstruction of Biochemical Networks

Enzo Acerbi$^{1,2}$, James Decraene$^1$, and Alexandre Gouaillard$^1$

$^1$Singapore Immunology Network (SIgN), A*STAR, Singapore
$^2$Doctoral School in Translational and Molecular Medicine (DIMET), University of Milano-Bicocca, Italy

{enzo_acerbi, james_decraene, alexandre_gouaillard}@immunol.a-star.edu.sg

Abstract—Biochemical networks are hierarchical complex systems involving many heterogeneous molecular species and intricate mechanisms such as crosstalks between different pathways and emergent dynamic behaviour. Computational modelling and simulation have proved to be powerful new approaches to the investigation of such complex systems. Modelling and simulation initially require the reconstruction in silico of the biochemical system in question using experimental datasets and complementary sources. While all reconstruction projects are to some extent unique, they can all be characterized by specific research questions, data/knowledge requirements, computational expertise, etc. To date, no single approach can be applied successfully to all biochemical reconstruction efforts. Moreover, no guidelines have yet been proposed to guide investigator through this process. Here we attempt to address this gap by providing a comprehensive overview of the reconstruction methods commonly applied to biochemical networks. We evaluate the principal methods of computational reconstruction with regards to data availability and type, target system scale, research/study aims and computational requirements.

I. INTRODUCTION

Biochemical networks are characterized by a high degree of heterogeneity and connectivity: many different molecular species are involved, including genes, transcription factors, proteins and ions) which each interact with each others in multitude of disparate ways. These relatively simple local biochemical interactions lead to the emergence of complex behaviour at the system level. Moreover, these networks can be organized in a hierarchical manner whereby each network layer corresponds to specific metabolic, regulatory or signalling functions. Computational modelling and simulations are now being widely employed to further our understanding of complex biological systems [1]. This in silico approach offers cost efficiency, traceability and predictability power benefits. An investigator may employ a computational model to assess ‘what-if?’ scenarios that would otherwise prove difficult to realise in vivo/vitro due to technological or financial constraints. Computational models are effective tools that can be used to mimic and simulate complex emergent dynamics [2]. In order to build these computational models, detailed and reliable biochemical datasets are required to support the reconstruction of the model. To fully understand the dynamics of complex biochemical systems, the underlying mechanisms governing the system’s fate must be defined and reconstructed in silico, which depends on comprehensive experimental datasets and thorough analytical studies. Large-scale “system reconstructions” require different kinds of data either generated by modern high-throughput technologies, such as micro-arrays and ChIP-Seq, or extracted from the literature. Moreover, insights can also be provided by domain experts. All these heterogeneous data sources must then be combined and integrated in a consistent and biologically meaningful way using computational methods. To date, few studies have attempted to provide a holistic overview of the computational methods used to integrate and reconstruct biochemical models while considering data availability and type, target system scale and research or analytical aims [3]. This issue is addressed in the current paper, which is structured as follows: First a brief description of biochemical networks and computational modelling is introduced. Next, the process of model reconstruction is presented. Then the different evaluation criteria follow accompanied by an overview of the computational reconstruction techniques. Finally, an evaluation of the reconstruction techniques according to defined evaluation criteria is conducted.

II. BACKGROUND

A. Biochemical networks

Biochemical networks can be categorized into three principal types of interlinked sub-network: signalling, regulatory and metabolic networks:

- **Cell signalling networks**, also referred as CSNs, are comprised of signalling pathways, transduction pathways and signalling transductions. Specific chemical reactions or “signalling events” (such as phosphorylations or ubiquitinations) cascade and propagate throughout the cell to process internal and external stimuli to trigger appropriate cellular responses. In other words, signalling networks can be regarded as signal processing networks that transform input signals (e.g. intra/extra-cellular stimuli sensed by receptors) into appropriate outputs signals (e.g. triggering the production of specific cytokines).

Signalling molecules include a variety of proteins, receptors and enzymes which interact with each others to induce the signalling cascade (see Fig. 1). The functional state of these molecules may change as a result of this signal processing. For example, protein phosphorylation
during cell signalling can induce the activation of that molecule). Signalling transductions may occur within minutes up to hours.

- **Genetic Regulatory Networks** (GRNs), also referred to as Transcriptional Regulatory Networks, are responsible for regulating the expression of genes (substrings of nucleic acids) encoded within the DNA. Transcription factors are proteins that bind to specific genes to positively or negatively regulate transcription into mRNA. Since transcription factors are themselves proteins produced by genes, it follows that genes themselves can regulate the expression of other genes, thus resulting in a regulatory network. Through these regulatory interactions, the cell can modify its genetic transcriptional state in response to internal and external stimuli (e.g., activated by some signalling cascades). GRNs are thus similar to CSNs but involve genes as the key interacting entities. Regulatory interactions also scale from minutes to hours.

- **Metabolic networks**, also referred to as metabolic pathways, are the most well-studied biochemical networks. These networks describe the core chemical reactions that support the creation/destruction of molecular species. These chemical reactions typically mediate energy harvesting (obtaining ATP molecules) or construct molecular species (consuming ATP molecules) as necessary for the host organism to grow and survive. Metabolic pathways convert organic compounds into other chemicals through chain reactions catalysed by enzymes. Raw material from the environment is required to enable these reactions. In addition to CSNs and GRNs, metabolic reactions occurring within a cell, determine its physiological and biochemical properties. Metabolic reactions may occur on a scale of milliseconds to seconds. Large scale metabolic networks contain hundreds of metabolites and support more than a thousand reactions [4].

Metabolic pathways are essentially characterized by a flow of matter whereas signalling and regulatory networks are defined as flows of information [5]. Despite being regarded as generic networks, the structures detailed above can potentially include information about the localization of their constituent entities (cellular components), which is a feature rarely accounted for when modelling other types of networks. In sum, CSNs, GRNs and metabolic networks are highly correlated and overlapping (see Fig. 1), and the comprehensive mapping of these networks remains incomplete.

**B. From experimental data to computer simulations**

We first introduce the process of model reconstruction using experimental data. We then briefly describe the corresponding field of computer simulation where reconstructed models of biochemical networks are executed.

1) **Model reconstruction**: Model reconstruction, also known as model “reverse-engineering” or computational inference, is an important research area in system biology [6]. This approach aims at to define the intricate mechanisms that underpin biochemical networks by using a systemic approach that integrates the various ‘omics’ data in silico [7], [8], [9], [10], [11]. Specifically, computational reconstruction methods are needed when current knowledge of the studied system is incomplete. While a significant amount of knowledge and data are available in the literature and in various online databases [12], [13], these sources are often characterized by inconsistencies and conflicting reports. Reconstruction approaches can aid the resolution of discrepancies between datasets, leading to clarification, validation and unification of knowledge. Initially, reconstruction requires the identification of the molecular species and interactions that are involved in the network by computational integration of conventional experimental data in a biologically consistent way. This step is inter-disciplinary, since both biologists and computer scientists must collaborate effectively to integrate these data successfully. The principal outcome of reconstruction is the ‘executable’ computational model. The investigator can then execute or ‘run the model’ to observe the dynamics of the simulated networks. For instance,
when executing a quantitative mathematical model, one may compute and observe the level/state of expression of specific molecular species over time. This contrasts with reports available in the literature, which are non-executable and provide only a static description of the network in question. Once an initial model describing the entities and their interactions has been inferred from experimental data and integrated with information from the literature, computer simulations can be run that enable further analytical studies. Simulation can lead to the prediction and discovery of new entities and interactions, which is further described in the next section.

![Diagram of the iterative model reconstruction process](image)

**Fig. 2.** The iterative model reconstruction process. Four main sub-processes are distinguished: data generation, modelling, simulation and validation.

2) Simulation: Executing computational models (running a simulation) is a well established technique in computer science which has now been successfully applied to various different disciplines including engineering, physics and economics [14], [15]. The key benefit of computer simulations is their ability to generate predictions or forecasts about system dynamics given an initial set of conditions. In other words, simulations can be used to explore system dynamics which have not yet been observed in laboratory-based experiments. Additional benefits are identified as follows:

- **Cost benefit:** simulations of a model can be run under many different initial conditions e.g. inhibiting some molecular species, varying initial concentrations, adding an interaction between two species, thus enabling the exploration of system dynamics in a more exhaustive fashion. In contrast, systematic explorations in conventional laboratory-based settings are currently limited by financial, methodological and time constraints.

- **Hypothesis validation and generation:** model simulations can facilitate the validation of a new hypothesis. For instance, we may consider a hypothetical interaction between species $A$ and $B$ where a third species may also be involved. Through mathematical inference, we can establish that the behaviour of $A$ is not sufficient to explain the behaviour of $B$, thus the model may suggest the presence of a third intermediary protein involved in the interaction between $A$ and $B$. Similar in silico validations (or computational screening) can better inform biologists wishing to conduct laboratory-based experiments aiming to validate their hypotheses. It is also possible that computer simulations may refute a given hypothesis, leading to the formulation of novel interpretations of data. This iterative feedback between biologists and computer scientists can therefore lead to the refinement of hypotheses and may ultimately lead to new scientific discoveries.

- **Overcoming technological constraints:** computer simulations enable investigators to conduct experiments that would be hard or impossible to carry out in the laboratory. The technological limitations of wet-lab experiments impose limits on observable data. For example, flow cytometry may only measure a limited number (typically less than 20) of different molecular species in parallel, whereas computer simulations may describe the dynamics of much more molecular species simultaneously.

The model reconstruction and simulation processes are part of an iterative method which is summarized in Figure 2. It should be noted that model reconstruction differs significantly from the field of bioinformatics, which consists of a deep and accurate descriptive analysis of data wherein relationships and correlations between variables are examined. Bioinformatics can provide statistical predictions but these are interpolated from data already acquired. Although bioinformatic tools can describe the statistical properties of the target system, this approach cannot robustly predict the behaviour of unknown or hidden entities and interactions. Model reconstruction and computer simulation are thus well placed to exploit the large datasets resulting from bioinformatic studies. Systems biology and bioinformatics approaches are best applied in combination to more effectively study complex biological networks. This paper focuses on the computational tools available to reconstruct models from experimental data, which is presented in the next section.

### III. Computational Reconstruction

#### A. Overview

To date, no precise standard has been established for the reconstruction of biochemical network models. It is important to emphasize that a good model is not necessarily the most comprehensive, accurate and detailed model, but incorporates the minimum set of entities sufficient to capture or reproduce the phenomena of interest. A reasonable overall strategy for model reconstruction can therefore be formulated: one should attempt to reduce the set of entities and interactions involved to include only those potentially affecting the phenomenon of interest. Entities and related reactions which remain constant during the experiment can be ignored. Indeed, increasing the model specifications may dramatically alter the system dynamics, which would be counter-productive [16], and will
limit the benefits of performing modelling and simulation. The process of model reconstruction can be detailed as follows:

1) Determine the research question being examined.
2) Determine which entities/reactions to include in the system. In this regards, addressing the following questions is key:
   - What is the system scope, i.e., what are the boundaries of the studied system?
   - What is the system granularity, i.e., what is the level of detail/resolution necessary to study the underlying biochemical processes of interest.
   - What are the system behavioural properties being investigated?
   - What data are available?
3) Refine the entity and reaction lists using preliminary experimental data. Through these initial experiments, one may further clarify which species and interactions play significant roles in the system dynamics of interest. Entities identified as negligible can be removed.
4) Represent the list of entities and interactions using a suitable mathematical or computational formalism.
5) Evaluate and validate the model content using various mathematical methods and biological experiments.

The validity of the model must be established through the use of model predictions and additional targeted experimental data. Furthermore, computational model checking techniques [17], [18] may assist in the validation process. If the model can be successfully validated, it can then generate reliable system predictions which can be used as a rapid screening tool to more efficiently direct future experiments. In contrast, if the model cannot be validated with the available data, then the results may suggest possible refinements to be examined in the next iteration of the model development. Indeed, a failure in the validation process is often fruitful: it may represent the first step towards uncovering novel entities, reactions and/or mechanisms that then stimulate the generation of new hypotheses. This iterative reconstruction process is summarized in Fig. 2.

When reconstructing a biochemical model, several considerations must be addressed to identify the appropriate methods to employ. These considerations, namely data availability, data types, system scale and research questions being examined are presented in the remainder of this section.

B. Data and knowledge

Data collection is essential for reconstructing biochemical networks. Recent developments in high-throughput technologies have lead to the generation of dataset that can facilitate model reconstruction. However, it is still common for reconstruction of large-scale and fined-grained models of biochemical networks to be limited by lack of data.

Most online databases provide data using an exchangeable computer-readable format (e.g. PS-MI, BIOPAX, SBML) which enable the use of several analytical tools. There are a growing number of public repositories offering biochemical networks data [12] but formats and protocols are still far from unified. The meta-database Pathguide [13] is a powerful gateway that provides access to the most commonly used databases including Reactome [19], KEGG [20] and wikiPathways [21], (currently linked to 190 databases coming from the scientific literature and/or from high-throughput experiments). However, none of the above databases is truly comprehensive, thus integration of data from these different sources must be conducted first. Due to the concurrent existence of different formats and the variety of datasets that describe different biochemical properties [9], data integration is a time-consuming and difficult procedure that requires domain experts. No reliable automated techniques for data integration exist at present. This manual “compare-and-combine” process also includes the integration of information databases extracted from the literature. Detailed descriptions of omics data integration techniques, together with reviews of pathways databases can be found in [7], [8], [9], [10], [11].

An accurate reconstruction process cannot rely solely on observational/experimental data obtained from a single experimental approach with a defined set of starting conditions. Instead, data resulting from multiple initial experimental conditions, assay perturbations and interventions is necessary. For instance, when working with probabilistic graphical models such as Bayesian Networks, the usage of non-interventional data would only provide a partially non-directed graph of connections. By breaking the symmetries within equivalence classes of graphs (permitting different posterior probabilities among an equivalent graphs class) interventional data is able to determine the direction of causal relations [22], [23], [24]. Several experimental techniques can readily perform such targeted interventions (e.g. gene knock-out).

C. Analytical study

The suitability of the reconstruction and modelling methods used may vary depending on the aims of the analysis. Two main types of analytical study are distinguished as follows:

Qualitative analysis: The qualitative approach focuses on the structure and function of the biochemical entities [25]. This approach requires the modeller to consider “cause and effect” rather than rates of change. The basic entity is the state machine, which relates the different qualitative configurations “states” to one another. An algorithm is then used to simulate the dynamics of the modelled biochemical system. Some modelling techniques such as Petri Nets or Bayesian Networks are more suitable to study chains of events and explore the topological characteristics of biochemical networks (e.g. identify which signal transduction pathways can result from an environmental perturbation) but they will not provide exact details on entities’ concentrations and reactions. Qualitative analysis also includes network structural analysis where the network topology properties (e.g. scale-free networks) are examined [26]. Qualitative methods have often been used for Gene Regulatory Networks.

Quantitative analysis: Quantitative analysis, also know as dynamic model analysis, is based on the use of transfer functions, e.g. equations, that describe a relationship between
cellular entities and how the quantities of those species change over time. These dynamic quantities, including molecular concentrations, can then be described in an accurate manner. Quantitative models such as ODEs are well established with a strong mathematical background. However, quantitative analyses require an exhaustive set of precise parameters to be specified, e.g., reaction rate kinetics. These approaches are thus difficult to apply when the number of variables is high. Indeed, the required parameters are often not available in the literature, and must therefore be estimated based on expensive lab experiments. When confronted with a high number of unknown parameters, is generally not possible to find a unique solution: a potentially infinite number of solutions may fit the given target time-series data. In such cases, it is necessary to reduce the solution search space by integrating as much a priori knowledge as possible. Traditionally quantitative approaches have been applied to the study of metabolic networks.

D. System scale

What is referred to here as the “system scale” is the number of biochemical entities involved in the target system. The scale of the system to be reconstructed is critical as this directly influences the data/knowledge requirements and the selection of suitable modelling/reconstruction techniques. The number of entities and associated reactions affects the number of parameters to be estimated during the reconstruction process. The system scale therefore affects the difficulty of the model reconstruction which can be regarded as an optimization process. The search space, or design space of model candidates grows exponentially with the system scale. In addition to computational difficulties during the reconstruction process, the large number of parameters to be optimized may also impede the modelling phase. For instance, numerical simulations of differential equations involving hundreds of

entities are time-consuming and would require advanced high performance computing facilities [27]. In large scale networks, the number of entities and reactions involved is dramatically more important. Such networks are characterized by the presence of multiple crosslinking pathways and negative-positive feedback loops which pose further challenges for the reconstruction/data fitting process [28]. The system scale also affects the data requirements. As outlined previously, despite significant progress in “omics” technologies, conventional laboratory datasets remain comparatively limited in scope. Even “high resolution” or fined-grained experimental datasets, where many species are simultaneously monitored in real time, e.g., using flow cytometry or real-time PCR, are still restricted to a few species (commonly less than 20 molecular species). When reconstructing genome-scale networks, it is therefore likely that high resolution datasets will not be available. To address this problem, techniques such as Flux Balance Analysis (described in Section III-E3) exploit stoichiometric matrices without requiring detailed chemical kinetics data. Large scale reconstruction has been successfully conducted on metabolic and regulatory networks, whilst the reconstruction of large-scale signalling networks is still a nascent endeavour. To date, the largest signalling network reconstruction was performed in [29] where 909 species and 752 reactions were reconstructed.

E. Computational techniques

1) Ordinary Differential Equations: Ordinary Differential Equations (ODEs) [30], [31] provide an aggregate and quantitative description of the cellular entities. Due to this aggregated view of the chemical entities, limited information can be derived with regards to possible deviation in system dynamics. ODEs assume a homogeneous composition of the system where entities are uniformly well mixed and distributed.
over the reaction space. This approach may therefore not be suitable where spatial effects are important. Partial Differential Equations (PDEs) have been successfully applied to address this need. ODEs also assume that a large quantity of molecules is involved, in which case the law of mass action can be considered. ODE-based approaches appear to be limited when considering systems where the number of entities involved is small (where statistical fluctuations may significantly affect the system dynamics) [32].

2) Bayesian Networks: Bayesian networks (BNs), also referred to as beliefs networks [33], [34], are probabilistic graphical models where nodes are random variables and edges represent conditional dependencies. “Beliefs” about values of random variables are expressed as probability distributions which can be estimated from data, and these can be updated as new evidence is provided. BNs are able to handle noisy and incomplete data, which is a common situation when working with biological data. Furthermore, BNs permit the easy introduction of a priori knowledge into the model and can successfully accommodate hidden variables. The structure of the network can be learnt from data, which makes BNs suitable for biochemical network reconstruction. Dynamic Bayesian Networks (DBNs) [35] extend classic BNs to allow for a discrete representation of time, which enables the modelling of feedback loops.

3) Flux Balance Analysis: Flux balance analysis (FBA) [36], [37] is a constraint-based formalism that has been largely applied to the modelling of metabolic networks; recently this technique has been combined with other approaches to model regulatory and transaction processes [38], [39]. FBA assumes that the biochemical system in question is being studied under homeostatic conditions. When modelling a metabolic network using FBA, the total concentrations of metabolites in the system are assumed to remain relatively stable over time: the reconstruction problem is reduced to the balancing of fluxes within the system. FBA is based on the use of reaction matrices which contain the stoichiometric coefficients of each reaction. Finally, since FBA does not rely on reaction kinetic parameters, it cannot predict species concentrations.

4) Petri Nets: Petri nets provide a well-established and constantly growing semi-quantitative computational modelling technique. This graph-based technique (weighted, directed and bipartite) is well suited for the analysis of distributed systems [40]. The principal elements are nodes and arcs which are used to model biological compartments, molecular species and interactions. Biochemical networks are characterized by non-deterministic behaviours and a high degree of concurrency, which Petri nets can handle [41], [42]. Indeed Petri nets are able to model uncertainty through devising stochastic transition rules. A comprehensive review of Petri nets as applied to biochemical network modelling can be found in [43].

5) Agent-based Models: Agent-based Models (ABMs) [44], [45] are a relatively intuitive approach where systems are described as a set of concurrent entities (or “agent”) combined with behavioural rules determining the interactions between the agents. ABMs can capture the stochastic nature of biochemical networks through the use of probability-based interactions. Moreover, agent interactions can be asynchronous, with individual agents responding independently to incoming environmental signals. An exact matching between agents and biochemical entities is feasible: ABMs can treat each molecule as a single identifiable and traceable agent. From the simple agent-level behaviours may raise complex emergent behaviours at the system level (e.g. time delays). In contrast to ODEs, ABMs may assume spatial heterogeneity. This may result in a more accurate approximation of biochemical conditions, which are often characterized by heterogeneous spatial distribution of their components [46]. Moreover, ABMs are significantly less computationally expensive than PDEs. Differential equations and ABMs are complementary approaches that can be combined together [47]. A drawback of ABMs lies with the reproducibility of agent-based experiments, since no ABM standards have been established and no central data repository is available yet.

F. Evaluation

In this section, we provide an overall qualitative assessment of computational techniques to assist in the selection of the most suitable approach according to the evaluation criteria described in the previous section (data and knowledge requirements, analytical study aims, system scale). In addition, we consider the traceability and computational requirements as potentially important factors. Traceability is the level of resolution desired: atomic level, molecule level, molecular species level, etc. Computational scalability determines the computational requirements needed to conduct the in silico simulations. Figure 4 provides a graphical overview of the main reconstruction techniques. It can be observed that a universal optimum reconstruction approach does not exist and that every situation must be evaluated independently.

Among the different evaluation criteria introduced earlier, it is apparent that data availability is the primary constraint when conducting model reconstruction. Due to difficulties in collecting sufficient data, quantitative techniques such as ODEs are limited, since they require detailed and hard-to-get information concerning reaction kinetics. Thus ODE-based approaches tend to be more suitable for relatively small systems (typically including fewer than 50 entities). ABMs and Petri nets offer more flexibility than ODEs, since the modeller is not constrained to a particular model resolution. Here, the data requirements depend on the resolution of the model as chosen by the modeller. If a high-resolution or highly detailed model is wanted, then these techniques would require the specification or inference of a large number of parameters, similarly to when using ODEs. As a consequence, a significant amount of experimental data may be needed to support model validation and refinement [44] when using ABMs and Petri nets. However, such fine-grained experimental datasets are difficult to obtain from web-lab experiments. With respect to data availability, FBA only requires the specification of constraints (flux rates) instead of detailed parameters such as
concentration and kinetics. This approach has clear benefits in terms of data and computational requirements. Although, Bayesian networks can handle partial and noisy data, the structural learning of BNs is computationally challenging [48] and this method is limited when addressing the reconstruction of large scale biochemical networks. Moreover, this technique still requires large amounts of experimental and interventional data to uncover hidden molecular correlations. With FBA, the size of the stoichiometric matrix (the principal data requirement) is typically very large, which may result in an undetermined system wherein too many solutions satisfy the flux balancing problem. However, the size of the solution space can then be reduced by specifying search constraints. In terms of traceability, the methods of differential equation-based, FBA and Bayesian networks are limited as they consider molecular species as a whole in an aggregated manner. In contrast, agent-based and Petri models may consider the fine grained behaviour of individual molecules. This property is desirable when considering stochastic systems involving few entities wherein statistical fluctuations may strongly affect the dynamics.

Despite the use of different formalizations to describe the system, each of these separate techniques subsequently requires an optimization method to fit the data with the model: this second phase, also referred to as parameter estimation or parameter fitting, reduce the initial problem to a parameter optimization task, which can be tackled by many different search algorithms [49], [50].

It has also been shown that a combination of these techniques can be used to integrate multi-source data (metabolomic, regulatory, transduction), see [46], [47] for recent works on integrated model reconstruction approaches.

IV. CONCLUSION

An overview of computational reconstruction methods for biochemical networks has been presented, focusing on important features such as data and computational requirements, analytical study type and system scale. An overall reconstruction methodology has been outlined, together with an evaluation of the main reconstruction approaches. Although, no optimum reconstruction formalism can be identified, a guideline for the selection of a suitable approach accommodating various conditions has been provided.

ACKNOWLEDGEMENTS

The authors would like to thank Teresa Zelante, Alicia Wong Yoke Wei, Francesca Zolezzi, Elena Viganò and Fabio Stella for their valuable help. Neil McCarthy of Insight Editing London for manuscript editing.

REFERENCES
