PhD

PROGRAM IN TRANSLATIONAL AND MOLECULAR MEDICINE

DIMET

UNIVERSITY OF MILANO-BICOCCA
SCHOOL OF MEDICINE AND SCHOOL OF SCIENCE

Continuous time Bayesian networks for gene networks reconstruction

Coordinator: Prof. Andrea Biondi
Tutor: Prof. Andrea Biondi
Co-Tutor: Prof. Fabio Stella.

Dr. Enzo ACERBI
Matr. No. 715223

XXVI CYCLE
ACADEMIC YEAR
2012-2013

Contents

Chapter 1

Introduction: Computational Reconstruction of Biochemical Networks

The contents of this introductory chapter have been published in:

Enzo Acerbi, James Decraene, and Alexandre Gouaillard. "*Computational reconstruction of biochemical networks*". Proceedings of the 15th International Conference on Information Fusion. IEEE, 2012.

However, some post-publication additions and improvements have been made. For this reason, there are some slight differences between the contents of this chapter and the related published manuscript.

1.1 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.

1.2 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 reconstruction and simulation are now being widely employed to further our understanding of complex biological systems [1].

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 first step is to define and reconstruct *in silico* the underlying mechanisms governing the system's fat. This phase 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 (methodology known in Bayesian statistic as "expert elicitation" [2, 3]). All these heterogeneous data sources must then be combined and integrated in a consistent and biologically meaningful way using computational methods. Once the model is obtained, it can be used to perform simulation studies which can potentially offer 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 [4].

From above it is clear that model's reconstruction is the first key phase from which the outcome of the subsequent simulation studies strictly depends: an incorrect/biased model will lead to incorrect/biased simulation results. For this reason, after the model's reconstruction phase, a rigorous validation process must be performed in order to establish how reliable the inferred system is. This can typically be done by comparing the obtained model with existing literature evidence or, in some cases, verifying if well-know biological mechanisms can be reproduced from the reconstructed model. The topic of model validation is not specifically discussed in this chapter.

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 [5]. 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.

1.3 Background

1.3.1 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/extracellular 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. 1.1). 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 the activation of that molecule). Signalling transductions may occur within minutes up to hours.

- *Genetic Regulatory Networks* (GRNs), also referred to 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 involves 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 then a thousand reactions [6].

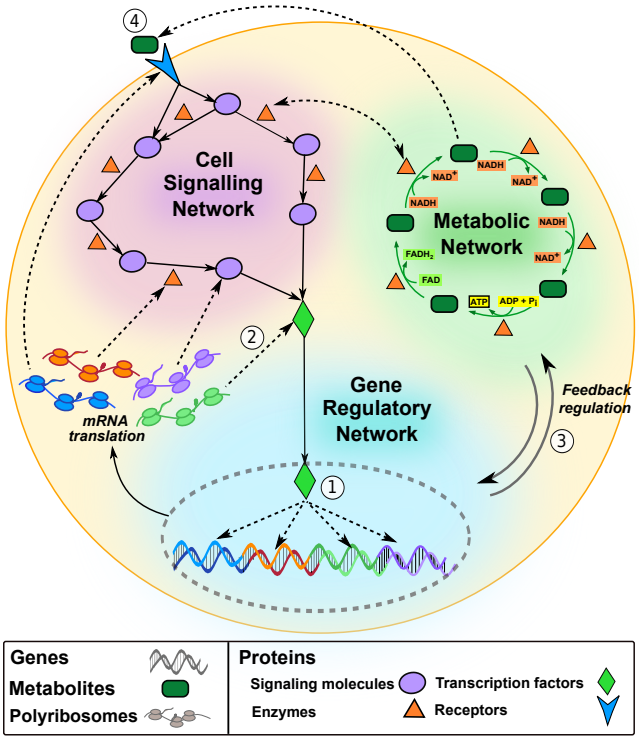
4
Cell
Signalling
Network
NADH
NAD+
NADH
NAD+
Metabolic
Network
NADH
FADH2
NAD+
FAD
ATP
ADP + Pi
2
mRNA
translation
Gene
Regulatory
Network
Feedback
regulation
3
1
Genes
Metabolites
Polyribosomes
Proteins
Signaling molecules
Transcription factors
Enzymes
Receptors

FIGURE 1.1

FIGURE 1.1: Schematic representation of interconnected and overlapping biochemical networks. Transcription factors (TFs) are proteins able to transfer into the nucleus and bind to specific genes (DNA sub-sequences) to positevily or negatively regulate the transcription of genes into mRNAs. These mRNAs subsequently carry protein translation instructions to the ribosomes, where protein synthesis occurs. The figure illustrates how specific genes, differentiated by multiple colors, support the translation of different types of proteins, and also highlights some interesting interactions identified as follows: 1. TFs binding to genes control the transcription of signalling entities (e.g. receptors, signalling proteins, TFs) as well as signalling and metabolic catalysts including enzymes. 2. Through the production of TFs, genes can cross-regulate the expression of other genes. 3. A negative-positive regulatory feedback mechanism allows metabolic networks to interact with gene regulatory networks to control the production of enzymes that modulate the production of metabolites. 4. Metabolites bind to receptors that elicit a signalling cascade to control the activation of TFs.

Metabolic pathways are essentially characterized by a flow of matter whereas signalling and regulatory networks are defined as flows of information [7]. Despite being regarded as generic *networks*, the structures detailed above can potentially include information about the localization of their consituent 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. 1.1), and the comprehensive mapping of these networks remains incomplete.

1.3.2 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.

Model reconstruction

Model reconstruction, also known as model "reverse-engineering" or computational inference, is an important research area in system biology [8]. 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* [9–13]. Specifically, computational reconstruction methods are needed when current knowledge of the studied system is incomplete. 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 (or simply: reconstructed model).

Each reconstructed model will depend both on data and on some modeling assumption. For instance, ordinary differential equations assume a population to be uniformly distributed over the physical space and dynamic Bayesian networks assume the system to evolve at discrete and evenly spaced points in time. Particular attention has to be made regarding these modeling assumptions, as a wrong initial hypothesis can lead to biased results. However, modeling assumptions are necessary to reduce the complexity of the reconstruction problem, which would be otherwise intractable.

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 analysis. Simulation can lead to the prediction and discovery of new entities and interactions, which is further described in the next section.

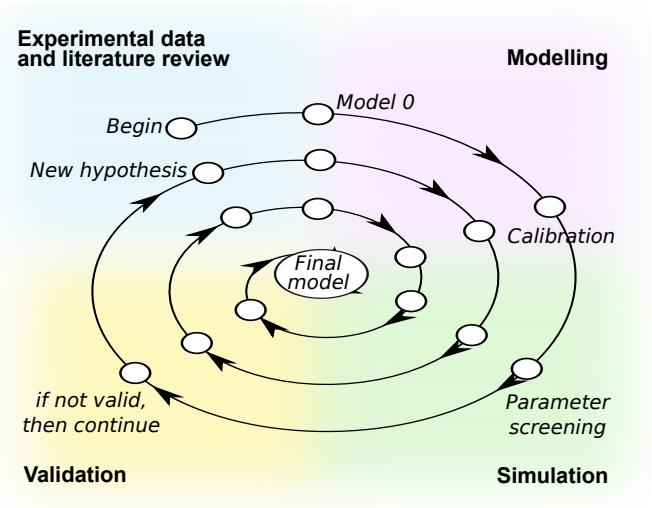

FIGURE 1.2: The iterative model reconstruction process. Four main sub-processes are distinguished: data generation, modelling, simulation and validation.

Simulation

Mathematical simulations (i.e. through ordinary/stochastic differential equations) allows to investigate how quantities related to biological entities, such as gene expression levels or protein phosphorylation levels, vary and influence each other over time. Mathematical simulations have now been successfully applied to elucidate the dynamics of gene networks [14], signaling networks [15] and metabolic networks [16]. It is important to notice that running a simulation always requires the previous specification of a model (mathematical or algorithmic).

The key benefit of 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. 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. As simplest example, let's consider the situation where are investigating the interaction between species A and B through ordinary differential equations (reaction kinetics). Failing to find the reaction parameters able to fit the data, is a strong indicator

that a third specie C, that was not originally considered, may be involved in the reaction.

- *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 1.2. In a descriptive analysis of data, relationships and correlations between variables are examined: statistical predictions can be provided but these are interpolated from data already acquired. Although the terminology in not universally agreed, this descriptive analysis of data is often referred to as bioinformatic analysis. Modeling and simulation have the additional advantage of being able to generate new (simulated) data from which new hypothesis may emerge. Model reconstruction and computer simulation are thus well placed to exploit the large datasets resulting from bioinformatic studies. Simulation 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.

1.4 Computational reconstruction

1.4.1 Overview

To date, no precise standard has been established for the reconstruction of biochemical network models. Every model is built to investigate a certain phenomena of interest and every model is built with some specific purposes; the model's granularity (level of detail) has not necessarily to be the highest possible. Instead, the appropriate model's granularity is the one incorporating the minimum set of entities sufficient to capture/reproduce the features of the biological phenomena which are relevant to the modeling purposes. 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 and the modeling purposes. Entities and related reactions which remain constant during the experiment are typically ignored. However, before excluding such variables, one must carefully consider that a variable may appear to be constant

as result of the specific subspace selection (e.g. the variable may be constant because of the influence of other variables which do change, but that have not been included in the analysis).

It is important to notice that an accurate reconstruction process cannot rely solely on observational/experimental data obtained from a single experimental approach with a defined set of starting conditions. Indeed, this would only allow for a type of analysis aimed to uncover association/correlation between variables. Association relationships are informative, for example they allow to model the past and future behavior of variables and to model the impact of obtaining new evidence. However, the ultimate purpose of every reconstruction effort is to uncover causal relationships. At this purpose, data resulting from multiple initial experimental conditions or *perturbations* is necessary.

Manipulation over variables, also known as interventional data (i.e. gene knock-outs), is extremely effective in uncovering causal relationships. For example, when working with probabilistic graphical models such as Bayesian Networks, with interventional data we are able to break the symmetries within equivalence classes of graphs (permitting different posterior probabilities among an equivalent graphs class). Therefore, through the usage of interventional data is possible to determine the direction of causal relationships [17–19]. However, it is important to notice that under certain conditions,

causal relationships can also be uncovered from observational data alone [20].

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.

As mentioned, the validity of the model must be established through the use of model predictions and/or additional targeted experimental data. A model can also be considered valid when it provides a better (more plausible) biological explanation for the data. Computational model checking techniques [21, 22] 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. 1.2.

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.

1.4.2 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 [23] but formats and protocols are still far from unified. The meta-database Pathguide [24] is a powerful gateway that provides access to the most commonly used databases including Reactome [25], KEGG [26] and wikiPathways [27], (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 [11], data integration is a time-consuming and difficult procedure that requires domain experts. No reliable automated techniques for data integration exist at present. This

		Experimental Technique	Static/Time course	Type of Data
High-Troughput Wet-lab Experiments	Next Generation Sequencing	RNA Sequencing	Static/Time course	Transcriptomics
		DNA Sequencing	Static	Genomics
		ChIP Sequencing	Static/Time course	Genomics
	Microarray	Protein Microarray	Static/Time course	Proteomics
		Expression Microarray	Static/Time course	Transcriptomics
		Genotyping Microarray	Static	Genomics
	Others	Mass Spectrometry (MS)	Static	Proteomics and Metabolomics
		Yeast 2 Hybrid system (Y2H)	Static	Proteomics
		Nuclear Magnetic Resonance (MRN)	Static	Metabolomics
Targeted Wet-lab Experiments		Reporter Gene Assay (Luciferase)	Static/Time course	Gene expression
		Northen Blotting	Static/Time course	Gene expression
		Quantitative Polymerase Chain Reaction (qPCR)	Static/Time course	Gene expression
		Flow Cytometry (FCM)	Static/Time course	Protein Expression
		Immunoprecipitation (IP)	Static	Protein Interactions
		Various Microscopy Techniques (e.g Atomic Force)	Static/Time course	Protein Expression
		Western Blotting	Static/Time course	Protein Expression

FIGURE 1.3: Overview of high-throughput and targeted experimental laboratory techniques. The suffix *omics* characterizes words such*genomics*, *proteomics*, *transcriptomics* or *metabolomics*, referring to the whole amount of large-scale data coming from high-throughput technologies. The property "Static/Time course" indicates whether an experimental techniques can provide longitudinal data. For *omics* data coming from high-throughput experiments, obtaining dynamic data is more difficult.

FIGURE 1.4: Box plot of number of positions sent per iteration using this scheme

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 [9–13].

1.4.3 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 [28]. 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 [29]. 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.

1.4.4 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 [30]. In large scale networks, the number of entities and reactions involved is dramatically more important. Such networks are characterized by the presence of multiple crosstalking pathways and negative-positive feedback loops which pose further challenges for

the reconstruction/data fitting process [31]. The system scale also affects the data requirements. As outlined previously, despite significant progresses 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 1.4.5) 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 [32] where 909 species and 752 reactions were reconstructed.

1.4.5 Computational techniques

Ordinary Differential Equations

Ordinary Differential Equations (ODEs) [33, 34] 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) [35].

Bayesian Networks

Bayesian networks (BNs), also referred to as beliefs networks [36, 37], are probabilistic graphical models where nodes are random variables and edges represent probabilistic 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. A connection can be made between the concept of probabilistic dependency around which BNs are built and the notion of direct causal influence [38], making BNs suitable for a causal interpretation of the phenomena under investigation. Dynamic Bayesian Networks (DBNs) [39] extend classic BNs to allow for a discrete representation of time, which enables the modelling of feedback loops.

Flux Balance Analysis

Flux balance analysis (FBA) [40, 41] 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 transduction processes [42, 43]. 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.

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 [44]. 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 [45, 46]. 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 [47].

Agent-based Models

Agent-based Models (ABMs) [48, 49] 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

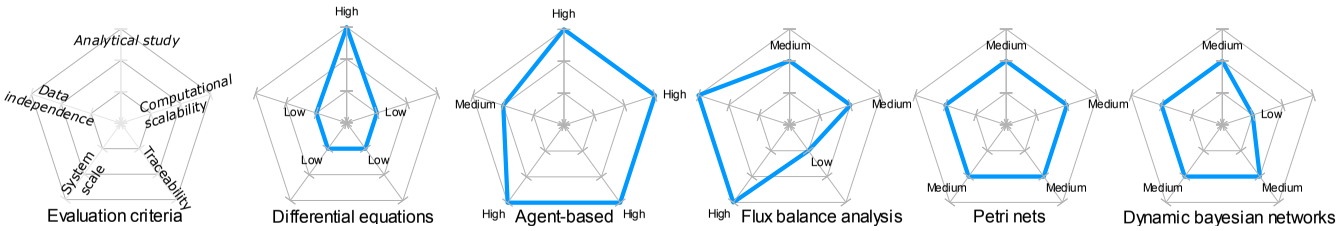

FIGURE 1.5: Different parameters were chosen to evaluate five major reconstruction approaches. "Data independence" describes the amount of data/knowledge required, where "low" independence would indicate that large detailed datasets are required, e.g. using high-throughput technologies and including reaction kinetics. In contrast, "high" data independence indicates that relatively few details about the system are necessary e.g. using biochemical static maps which can be found in the literature. "Analytical study" set to "low" indicates that the analysis is limited to the qualitative study of causal relationships, while "high" indicates the capacity to provide a fully detailed quantitative analysis. "System scale" is "low" for approaches suitable only for small scale systems such as a single signalling pathway, whereas "high" distinguishes approaches able to scale up to large-scale systems and able to model complex phenomena like cross-talk between different pathways. "Low" "Computational scalability" indicates that the reconstruction method requires significant computing resources, while "high" means that the approach is computationally cheap. Finally "traceability" set to "low" means that the technique's resolution is limited at the molecular population level while "high" means that atomic elements can be traced independently. It should be noted that approaches whose graphs have bigger areas do not necessarily include the advantages of the ones with smaller areas.

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 [50]. Moreover, ABMs are significantly less computationally expensive than PDEs. Differential equations and ABMs are complementary approaches that can be combined together [51]. 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.

1.4.6 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 1.5 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 [48] when using ABMs and Petri nets. However, such

fined-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 [52] 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 [53, 54].

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.

1.5 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.

1.6 Scope

The rest of this thesis is organized as follows. In Chapter 3 continuous time Bayesian networks are proposed as new approach for the gene network reconstruction problem. The methods is compared with two state-of-the-art approaches under several conditions on simulated and experimental data. Continuous time Bayesian networks are also applied to the problem of elucidating the regulatory interactions governing the T helper 17 cell differentiation process. In Chapter 3 conclusions of the work and future perspectives are discussed.

1.7 References

[1] Jasmin Fisher and Thomas A. Henzinger. Executable cell biology. *Nature Biotechnology*, 25(11):1239–1249, November

2007. ISSN 1087-0156. doi: 10.1038/nbt1356.

[2] Nancy J Cooke. Varieties of knowledge elicitation techniques. *International Journal of Human-Computer Studies*, 41(6): 801–849, 1994.

[3] Seiya Imoto, Tomoyuki Higuchi, Takao Goto, Kousuke Tashiro, Satoru Kuhara, and Satoru Miyano. Combining microarrays and biological knowledge for estimating gene networks via bayesian networks. *Journal of Bioinformatics and Computational Biology*, 2(01):77–98, 2004.

[4] SA Sandersius, CJ Weijer, and TJ Newman. Emergent cell and tissue dynamics from subcellular modeling of active biomechanical processes. *Physical Biology*, 8:045007, 2011.

[5] A.M. Feist, M.J. Herrgård, I. Thiele, J.L. Reed, and B.Ø. Palsson. Reconstruction of biochemical networks in microorganisms. *Nature Reviews Microbiology*, 7(2):129–143, 2008.

[6] E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, and A.L. Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551, 2002.

[7] Edda Klipp and Wolfram Liebermeister. Mathematical modeling of intracellular signaling pathways. *BMC Neuroscience*, 7(Suppl 1):S10+, 2006. ISSN 1471-2202. doi: 10.1186/1471-2202-7-S1-S10.

[8] B. Palsson. *Systems biology: properties of reconstructed networks*. Cambridge Univ Pr, 2006.

[9] Andrew R. Joyce and Bernhard O. Palsson. The model organism as a system: integrating 'omics' data sets. *Nat Rev Mol Cell Biol*, 7(3):198–210, March 2006. ISSN 1471-0072. doi: 10.1038/nrm1857.

[10] P. Tieri, A. De La Fuente, A. Termanini, and C. Franceschi. Integrating omics data for signaling pathways, interactome reconstruction, and functional analysis. *Methods Mol. Biol*, 719:415–433, 2011.

[11] M.E. Adriaens, M. Jaillard, A. Waagmeester, S.L.M. Coort, A.R. Pico, and C.T.A. Evelo. The public road to high-quality curated biological pathways. *Drug discovery today*, 13(19-20):856–862, 2008.

[12] A. Bauer-Mehren, L.I. Furlong, and F. Sanz. Pathway databases and tools for their exploitation: benefits, current limitations and challenges. *Molecular systems biology*, 5(1), 2009.

[13] J.L. Gardy, D.J. Lynn, F.S.L. Brinkman, and R.E.W. Hancock. Enabling a systems biology approach to immunology: focus on innate immunity. *Trends in immunology*, 30(6):249–262, 2009.

[14] Ting Chen, Hongyu L He, George M Church, et al. Modeling gene expression with differential equations. In *Pacific symposium on biocomputing*, volume 4, page 4, 1999.

[15] Bree B Aldridge, John M Burke, Douglas A Lauffenburger, and Peter K Sorger. Physicochemical modelling of cell signalling pathways. *Nature cell biology*, 8(11):1195–1203, 2006.

[16] Ralf Steuer, Thilo Gross, Joachim Selbig, and Bernd Blasius. Structural kinetic modeling of metabolic networks. *Proceedings of the National Academy of Sciences*, 103(32):11868–11873, 2006.

[17] K. Sachs, O. Perez, D. Pe'er, D.A. Lauffenburger, and G.P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523, 2005.

[18] D. Pe'er, A. Regev, G. Elidan, and N. Friedman. Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, 17(suppl 1):S215, 2001.

[19] I. Pournara and L. Wernisch. Reconstruction of gene networks using bayesian learning and manipulation experiments. *Bioinformatics*, 20(17):2934–2942, 2004.

[20] Judea Pearl. *Causality: models, reasoning and inference*, volume 29. Cambridge Univ Press, 2000.

[21] M. Kwiatkowska, G. Norman, and D. Parker. Using probabilistic model checking in systems biology. *ACM SIGMETRICS Performance Evaluation Review*, 35(4):14–21, 2008.

[22] S. Jha, E. Clarke, C. Langmead, A. Legay, A. Platzer, and P. Zuliani. A bayesian approach to model checking biological systems. In *Computational Methods in Systems Biology*, pages 218–234. Springer, 2009.

[23] M.Y. Galperin. The molecular biology database collection: 2008 update. *Nucleic acids research*, 36(suppl 1):D2–D4, 2008.

[24] G.D. Bader, M.P. Cary, and C. Sander. Pathguide: a pathway resource list. *Nucleic Acids Research*, 34(suppl 1):D504–D506, 2006.

[25] G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, GR Gopinath, GR Wu, L. Matthews, et al. Reactome: a knowledgebase of biological pathways. *Nucleic acids research*, 33(suppl 1):D428–D432, 2005.

[26] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 27(1):29, 1999.

[27] A.R. Pico, T. Kelder, M.P. Van Iersel, K. Hanspers, B.R. Conklin, and C. Evelo. Wikipathways: pathway editing for the people. *PLoS biology*, 6(7):e184, 2008.

[28] Daniel R. Hyduke and Bernhard Palsson. Towards genome-scale signalling-network reconstructions. *Nat Rev Genet*, 11(4):297–307, February 2010. ISSN 1471-0056. doi: 10.1038/nrg2750.

[29] Jason A. Papin, Tony Hunter, Bernhard O. Palsson, and Shankar Subramaniam. Reconstruction of cellular signalling networks and analysis of their properties. *Nature Reviews Molecular Cell Biology*, 6(2):99–111, February 2005. ISSN 1471-0072. doi: 10.1038/nrm1570.

[30] Y. Zhou, J. Liepe, X. Sheng, M.P.H. Stumpf, and C. Barnes. Gpu accelerated biochemical network simulation. *Bioinformatics*, 27(6):874, 2011.

[31] M.A. Schwartz and V. Baron. Interactions between mitogenic stimuli, or, a thousand and one connections. *Current opinion in cell biology*, 11(2):197–202, 1999.

[32] F. Li, I. Thiele, N. Jamshidi, and B.Ø. Palsson. Identification of potential pathway mediation targets in toll-like receptor signaling. *PLoS computational biology*, 5(2):e1000292, 2009.

[33] M. Helmy, J. Gohda, J. Inoue, M. Tomita, M. Tsuchiya, and K. Selvarajoo. Predicting novel features of toll-like receptor 3 signaling in macrophages. *PLoS One*, 4(3):e4661, 2009.

[34] K. Selvarajoo, Y. Takada, J. Gohda, M. Helmy, S. Akira, M. Tomita, M. Tsuchiya, J. Inoue, and K. Matsuo. Signaling flux redistribution at toll-like receptor pathway junctions. *PLoS One*, 3(10):e3430, 2008.

[35] M. Pogson, R. Smallwood, E. Qwarnstrom, and M. Holcombe. Formal agent-based modelling of intracellular chemical interactions. *Biosystems*, 85(1):37–45, July 2006. ISSN 03032647. doi: 10.1016/j.biosystems.2006.02.004.

[36] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems:Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[37] Finn V. Jensen. *Bayesian Networks and Decision Graphs*. Information Science and Statistics. Springer, corrected edition, July 2002. ISBN 0387952594.

[38] Judea Pearl, Thomas Verma, et al. *A theory of inferred causation*. Morgan Kaufmann San Mateo, CA, 1991.

[39] K.P. Murphy. *Dynamic bayesian networks: representation, inference and learning*. PhD thesis, University of California, 2002.

[40] Jeffrey D. Orth, Ines Thiele, and Bernhard Palsson. What is flux balance analysis? *Nature Biotechnology*, 28(3):245–248, March 2010. ISSN 1087-0156. doi: 10.1038/nbt.1614.

[41] K. Raman and N. Chandra. Flux balance analysis of biological systems: applications and challenges. *Briefings in bioinformatics*, 10(4):435–449, 2009.

[42] J.M. Lee, E.P. Gianchandani, J.A. Eddy, and J.A. Papin. Dynamic analysis of integrated signaling, metabolic, and regulatory networks. *PLoS computational biology*, 4(5): e1000086, 2008.

[43] M.W. Covert, N. Xiao, T.J. Chen, and J.R. Karr. Integrating metabolic, transcriptional regulatory and signal transduction models in escherichia coli. *Bioinformatics*, 24(18): 2044–2050, 2008.

[44] W. Reisig. Petri nets. *Modeling in Systems Biology*, pages 37–56, 2011.

[45] Claudine Chaouiya. Petri net modelling of biological networks. *Brief Bioinform*, 8(4):bbm029–219, July 2007. doi: 10.1093/bib/bbm029.

[46] Dong-Yup Lee, Ralf Zimmer, and Sang-Yup Lee. Knowledge representation model for systems-level analysis of signal transduction networks. *Genome Inform Ser Workshop Genome Inform*, 15(2):234–243, 2004. doi: 10.1.1.80.7078.

[47] David Gilbert, Hendrik Fuss, Xu Gu, Richard Orton, Steve Robinson, Vladislav Vyshemirsky, Mary J. Kurth, C. Stephen Downes, and Werner Dubitzky. Computational methodologies for modelling, analysis and simulation of signalling networks. *Brief Bioinform*, 7(4):339–353, December 2006. ISSN 1467-5463. doi: 10.1093/bib/bbl043.

[48] Amy L. Bauer, Catherine A. A. Beauchemin, and Alan S. Perelson. Agent-based modeling of host–pathogen systems: The successes and challenges. *Information Sciences*, 179(10):1379–1389, April 2009. ISSN 00200255. doi: 10.1016/j.ins.2008.11.012.

[49] Xu Dong, Panagiota T. Foteinou, Steven E. Calvano, Stephen F. Lowry, and Ioannis P. Androulakis. Agent-based modeling of endotoxin-induced acute inflammatory response in human blood leukocytes. *PLoS ONE*, 5(2): e9249, 02 2010. doi: 10.1371/journal.pone.0009249.

[50] Mark Pogson, Mike Holcombe, Rod Smallwood, and Eva Qwarnstrom. Introducing Spatial Information into Predictive NF-kB Modelling - An Agent-Based Approach. *PLoS ONE*, 3(6):e2367+, June 2008. doi: 10.1371/journal.pone.0002367.

[51] G.V. Bobashev, MD Goedecke, F. Yu, and J.M. Epstein. A hybrid epidemic model: combining the advantages of agent-based and equation-based approaches. In *Simulation Conference, 2007 Winter*, pages 1532–1537. IEEE, 2007.

[52] D.M. Chickering, D. Geiger, and D. Heckerman. Learning bayesian networks is np-hard. *Microsoft Research*, pages 94–17, 1994.

[53] C.G. Moles, P. Mendes, and J.R. Banga. Parameter estimation in biochemical pathways: a comparison of global optimization methods. *Genome research*, 13(11):2467–2474, 2003.

[54] M. Ashyraliyev, Y. Fomekong-Nanfack, J.A. Kaandorp, and J.G. Blom. Systems biology: parameter estimation for biochemical models. *FEBS Journal*, 276(4):886–902, 2009.

Chapter 2

Gene network inference using continuous time Bayesian networks: a comparative study and application to Th17 cell differentiation

A portion of the contents of this chapter has been published in:

Enzo Acerbi, Fabio Stella."*Continuous Time Bayesian Networks for Gene Network Reconstruction: a Comparative Study on Time Course Data.*" Proceedings of the 10th International Symposium on Bioinformatics Research and Applications. 2014.

An extended manuscripts matching the whole content of this chapter has recently been *submitted* for publication to a peer reviewed

journal: Enzo Acerbi, Teresa Zelante, Viping Narang, Fabio Stella. "*Continuous time Bayesian networks for gene network inference: a comparative study and application to Th17 cell differentiation.*".

2.1 Abstract

Understanding regulatory interactions between genes in the form of gene networks is an open challenge in molecular and computational biology. Dynamic aspects of gene networks are typically investigated by measuring system variables at multiple time points. Current state-of-the-art computational approaches for reconstructing gene networks directly build on such data, making a strong assumption that the system under investigation evolves in a synchronous fashion at discrete points in time. However, nowadays omics data are being generated with increasing time course granularity. Thus, modellers now have the possibility to represent the system as evolving in continuous time and improve the models' expressiveness.

In this work continuous time Bayesian networks are proposed as a new approach for gene regulatory network reconstruction from time course expression data. Their performance has been compared to two state-of-the-art methods: dynamic Bayesian networks and

Granger causality analysis. On simulated data methods's comparison was carried out for networks of increasing dimension, for measurements taken at different time granularity densities and for measurements evenly vs. unevenly spaced over time. Continuous time Bayesian networks outperformed the other methods in terms of the accuracy of regulatory interactions learnt from data for all network dimensions. Furthermore, their performance degraded smoothly as the dimension of the network increased. Continuous time Bayesian network were significantly better than dynamic Bayesian networks for all time granularities tested and better than Granger causality for dense time series. Both continuous time Bayesian networks and Granger causality performed robustly for unevenly spaced time series, continuous time Bayesian networks and Granger causality did not show a significant loss of performance compared to the evenly spaced case, while the same did not hold true for dynamic Bayesian networks. The comparison included the IRMA experimental datasets which confirmed the effectiveness of the proposed method. Continuous time Bayesian networks were then applied to elucidate the regulatory mechanisms controlling murine T helper 17 (Th17) cell differentiation and were found to be effective in discovering well-known regulatory mechanisms as well as new plausible biological insights.

We suggest continuous time Bayesian networks as an effective approach for gene network reconstruction from time course expression

data. The method is effective on networks of both small and big dimensions and is particularly feasible when the measurements are not evenly distributed over time. Reconstruction of the murine Th17 cell differentiation network using continuous time Bayesian networks revealed several autocrine loops suggesting that Th17 cells may be auto regulating their own differentiation process.

2.2 Background

In response to internal and external stimuli, a cell modifies its transcriptional state through the activation of multiple regulatory interactions that take place over time and which include complex mechanisms such as regulation chains, auto-regulations and feedback loops. Understanding gene regulatory networks (GRNs) is of extreme relevance in molecular biology and represents an open challenge for computational sciences. The task of uncovering the underlying causal structure of these cellular dynamics is referred to as gene network reconstruction or *(network)* "*reverse-engineering*".

Reconstruction of gene regulatory networks from time course expression data is an active area of research [1, 2]. In recent years, the granularity and length of time course data made available by omics technologies has been constantly increasing. This offers a

chance for a deep study of the dynamic evolution of regulatory networks [3] and calls for computational approaches able to effectively exploit the dynamic nature of data. In fact, most of the state-of-the-art methodologies for gene network reconstruction have been conceived before the advent of omic technologies and may not be always suitable for the new types and magnitudes of data.

A number of approaches have been applied to the GRNs reconstruction problem. Boolean networks [4] have been widely applied but are now giving way to more sophisticated approaches. Probabilistic graphical models such as Bayesian Networks [5] were shown to be powerful tools for solving the GRN reconstruction problem [6] and they led to significant discoveries [7]. When richer time course measurements started to be made available, Dynamic Bayesian networks (DBNs) [8] gained more and more relevance in the field, and today are largely applied with many variations and optimizations proposed. Other probabilistic approaches are state space models [9] and probabilistic boolean networks [10]; however it has been shown that the latter are outperformed by DBNs for GRN reconstruction problems [11]. Other approaches are ordinary differential equations (ODEs) [12, 13] which tend to became infeasible as the dimension of the network increases. Information-theoretic algorithms such as ARACNE [14] led to interesting discoveries [15], as well as evolutionary algorithms, which are reviewed in [16]. Finally, Granger causality (GC) [17, 18] is a robust method for analysing time course

data; since its early introduction it is successfully applied to a multitude of domains such as economics, neuroscience and biology. Exhaustive reviews of the existing network reconstruction approaches can be found in [19–23].

Dynamic aspects of regulatory networks are investigated by measuring the system variables at multiple time points (e.g. through gene expression microarray or mRNA sequencing). This approach is the result of technological constraints of the experimental techniques which only allow for measurements of "snapshots" of the system at multiple time points. In such situation the risk of missing important pieces of information is high if the sample rate is not adequately chosen or not fine enough (issue known as temporal aggregation bias). While this issue is currently unavoidable, when computationally analyzing these time course datasets it can be advantageous to separate the way the time course data is experimentally obtained from the way the time is represented in the computational model. Current state-of-the-art approaches described above directly build on "snapshot-like" data, making the strong assumption that the system under investigation evolves in a synchronous fashion at fixed points in time. Even when only discrete time data is available, modeling the system as continuously evolving over time represents a conceptually more correct/natural approximation and improves model expressiveness [24]. Nowadays, the finely grained time course data made available by high throughput technologies

make this continuous time representation feasible. It is also relevant to note that time course data are often unevenly spaced (measurements are not taken at equal width intervals). In such situations a continuous time model is preferable as it makes the analysis independent of the data sampling intervals.

In this paper continuous time Bayesian networks (CTBNs) [25] are proposed as a new approach for GRN reconstruction from time course data. In a CTBN variables can evolve continuously over time as a function of a continuous time conditional Markov process while the efficient factored state representation derives from the theory of Bayesian networks. Such setting brings many advantages to the description of the temporal aspect of a system, some of them directly relevant to the GRN reconstruction task. Firstly, the structural learning problem for CTBNs can be solved locally and in polynomial time with respect to the dimension of the dataset once the maximum number of regulators for each gene is set. This feature suits well regulatory networks, which are systems characterized by a large number of variables (genes) and where genes are typically regulated only by a limited number of other genes [26]. The second advantage is that CTBNs can naturally handle variables evolving at different time granularities. Gene networks are characterized by the presence of both regulatory interactions which happen quickly, e.g. within minutes from a given triggering event, as well as interactions which take place at a slower pace, e.g. within

hours or days. To reconstruct such regulatory networks, one may want to integrate data coming from experiments measuring genes whose state evolve at different rates. In such a context, CTBNs is naturally able to learn the overall causal network by combining data coming from different time granularities. The third advantage is that once the network structure and parameters have been inferred, through inference CTBNs can answer queries directly involving the quantification of the temporal aspects such as *"for how long does gene X have to remain up-regulated to have an effect on the regulation on gene Y?"* and in presence of partial evidence such as *"What is the most probable state for gene X at time t given that I observed that gene Y was up-regulated from time t - α to t - β?"*. With their graphical representation of causal relations, CTBNs also provide an intuitive and meaningful level of abstraction of dynamic regulatory process which can help a molecular biologist to gain a better understanding of the studied systems. Finally, CTBNs conserve all of the advantages which are characteristic of probabilistic graphical models and which make them suitable for the analysis of biological networks [27].

The effectiveness of CTBNs for GRN reconstruction is verified through a comparison with two state-of-the-art models, namely DBNs and GC, in the case where no *a priori* knowledge about the system is

available. Both DBNs and GC do not implement a direct representation of time. DBNs are built on the observational model assumption, with time slices representing the status of the system at evenly spaced time points. Hence if data samples are not collected at fixed width intervals one must either choose a time granularity equal to the smallest time interval between measurements or bias the data by imposing a uniform time granularity: in the first case the computational cost may increase dramatically while the second solution can lead to biased results. Moreover, due to the presence of intra-slice arcs for which the acyclicity constraint must be respected, learning DBNs is a NP-hard problem. GC implements a type of analysis based on an autoregressive model aimed to test if knowledge about the past values of a variable can help predicting the future value of another variable. GC has a great historical and current relevance when faced with the task of inferring causal relations from time series data. Its simplicity, flexibility and effectiveness made it broadly applied. However, almost all GC tests assume that the time intervals between measurements are fixed, exposing to the risk of obtaining biased results if this assumption is not verified. DBNs and GC were also directly compared for the reconstruction of gene networks in [28]: the authors showed that when the length of the time course is smaller than a given threshold, DBNs tend to outperform GC while vice-versa when the length of the time course is

greater than a threshold. CTBNs theoretically overcome the limitations associated with the discrete-time assumptions of both DBNs and GC. Therefore, we had reasons to believe that CTBNs would show advantages over DBNs and GC when applied to the problem of gene network reconstruction.

The analysis and comparisons performed here are based on an extensive and robust set of numerical experiments run on simulated time course data and include a test on an experimental dataset as well. The study with simulated data has been conducted on networks of 10, 20, 50 and 100 genes in order to investigate how the approaches perform on systems of increasing dimensions; the networks were extracted from the known transcriptional networks of two different organisms: *E. coli* and *S. cerevisiae*. To ensure robustness, the performance is not calculated on a single network instance but it is estimated by the average value computed over a set of 10 randomly sampled network instances of the same dimension.

We then investigated the methods' performances with respect to different time course granularities (11, 21 and 31 time points) while keeping the overall time duration of the experiment fixed. Finally, we investigated how the methods perform when the measurements are collected at unevenly spaced time points. For a robust comparison we evaluated the performance on 10 different random time point

instances. Our comparative investigation also included an experimental dataset as well: a 5 genes regulatory network synthetically constructed in the yeast *S. cerevisiae* (IRMA network) [29] which provided rich time course expression data and a gold standard for accurate benchmarking. In the second part of this work, we applied CTBNs for the reconstruction of the regulatory network responsible for murine T helper 17 (Th17) cell differentiation, testing their ability to confirm known regulatory interactions and generate new plausible biological insights.

2.3 Methods

2.3.1 Continuous time Bayesian networks

CTBNs cannot be considered a direct extension of DBNs, but a direct comparison naturally arises and helps to better understand the differences between the two approaches. DBNs model dynamic systems without representing time explicitly. They discretize time to represent a dynamic system through several time slices. In [25] the authors pointed out that "*since DBNs slice time into fixed increments, one must always propagate the joint distribution over the variables at the same rate*". Therefore, if the system consists of processes which evolve at different time granularities and/or the

observations are unevenly spaced in time, the inference process may become computationally intractable. CTBNs overcome the limitations of DBNs by explicitly representing temporal dynamics and thus allow us to recover the probability distribution over time when specific events occur. CTBNs have been used to discover intrusion in computers [30], to analyse the reliability of dynamic systems [31], for learning social network dynamics [32] and to model cardiogenic heart failure [33]. However, CTBNs have never been applied to the analysis of molecular data.

A continuous time Bayesian network (CTBN) is a probabilistic graphical model, whose nodes are associated with random variables and whose state evolves in continuous time. The evolution of each variable is conditioned on the state of its parents in the graph associated with the CTBN model. A CTBN consists of two main components: *i)* an initial probability distribution and *ii)* the dynamics which rule the evolution over time of the probability distribution associated with the CTBN.

Definition 2.1. (Continuous time Bayesian network). [25]. Let $\mathbf{X}$ be a set of random variables $X_1, X_2, ..., X_N$. Each X_n has a finite domain of values $Val(X_n) = \{x_1, x_2, ..., x_{I(n)}\}$. A continuous time Bayesian network B over $\mathbf{X}$ consists of two components: the first is an initial distribution $P^0_{\mathbf{X}}$, specified as a Bayesian network $\mathcal{B}$ over $\mathbf{X}$. The second is a continuous transition model, specified as:

- a directed (possibly cyclic) graph $\mathcal{G}$ whose nodes are $X_1, X_2, ..., X_N$; $Par_{\mathcal{G}}(X_n)$, often abbreviated $\mathbf{U}_n$, denotes the parent set of X_n in $\mathcal{G}$.
- a conditional intensity matrix, $\mathbf{Q}_{X_n}^{Par_{\mathcal{G}}(X_n)}$, for each variable $X_n \in \mathbf{X}$.

Given the random variable X_n, the *conditional intensity matrix* (CIM) $\mathbf{Q}_{X_n}^{Par(X_n)} = \mathbf{Q}_{X_n|\mathbf{U_n}}$ consists of a set of intensity matrices, one intensity matrix

$$\mathbf{Q}_{X_n|\mathbf{u}} = \begin{bmatrix} -q_{x_1|\mathbf{u}} & q_{x_1x_2|\mathbf{u}} & . & q_{x_1x_{I(n)}|\mathbf{u}} \\ q_{x_2x_1|\mathbf{u}} & -q_{x_2|\mathbf{u}} & . & q_{x_2x_{I(n)}|\mathbf{u}} \\ . & . & . & . \\ q_{x_{I(n)}x_1|\mathbf{u}} & q_{x_{I(n)}x_2|\mathbf{u}} & . & -q_{x_{I(n)}|\mathbf{u}} \end{bmatrix},$$

for each instantiation $\mathbf{u}$ of the parents $\mathbf{U_n}$ of node X_n, where $q_{x_i|\mathbf{u}} = \sum_{x_j \neq x_i} q_{x_ix_j|\mathbf{u}}$ is the rate of leaving state x_i for a specific instantiation $\mathbf{u}$ of $\mathbf{U_n}$, while $q_{x_ix_j|\mathbf{u}}$ is the rate of arriving to state x_j from state x_i for a specific instantiation $\mathbf{u}$ of $\mathbf{U_n}$. Matrix $\mathbf{Q}_{X_n|\mathbf{U_n}}$ can equivalently be summarized by using two types of parameters, $q_{x_i|\mathbf{u}}$ which is associated with each state x_i of the variable X_n when its parents are set to $\mathbf{u}$, and $\theta_{x_ix_j|\mathbf{u}} = \frac{q_{x_ix_j|\mathbf{u}}}{q_{x_i|\mathbf{u}}}$ which represents the probability of transitioning from state x_i to state x_j, when it is known that the transition occurs at a given instant in time and parents $\mathbf{U_n}$ are set to $\mathbf{u}$.

Learning the structure of a CTBN from a data set $\mathcal{D}$ consists of finding the structure $\mathcal{G}$ which maximizes the *Bayesian score* [34]:

$$\mathbf{score}_B(\mathcal{G} : \mathcal{D}) = \ln P(\mathcal{D}|\mathcal{G}) + \ln P(\mathcal{G}). \tag{2.1}$$

Efficiency of the search algorithm for finding the optimal structure $\mathcal{G}^*$ is significantly increased if we assume *structure modularity* and *parameter modularity*. The prior over the network structure $P(\mathcal{G})$ satisfies the structure modularity property if $P(\mathcal{G}) = \prod_{n=1}^{N} P(Par(X_n) = Par_{\mathcal{G}}(X_n))$, while the prior over parameters satisfies the parameter modularity property, if for any pair of structures $\mathcal{G}$ and $\mathcal{G}$' such that $Par_{\mathcal{G}}(X) = Par_{\mathcal{G}'}(X)$ we have that $P(\mathbf{q_X}, \theta_{\mathbf{X}}|\mathcal{G}) = P(\mathbf{q_X}, \theta_{\mathbf{X}}|\mathcal{G}')$. In [34] the authors combined parameter modularity, parameter independence, local parameter independence and assumed a Dirichlet prior over θ parameters and a beta prior over q parameters to obtain the following expression of the Bayesian score for a CTBN B:

$$\mathbf{score}_B(\mathcal{G} : \mathcal{D}) = \sum_{n=1}^{N} FamScore(X_n, Par_{\mathcal{G}}(X_n) : \mathcal{D}) \tag{2.2}$$

where

$$\begin{aligned} FamScore(X_n, Par_{\mathcal{G}}(X_n) : \mathcal{D}) = \\ \ln P(Par(X_n) = Par_{\mathcal{G}}(X_n)) + \\ \ln MargL^q(X_n, \mathbf{U}_n : \mathcal{D}) + \\ \ln MargL^\theta(X_n, \mathbf{U}_n : \mathcal{D}). \end{aligned} \tag{2.3}$$

According to [34] $MargL^q(X_n, \mathbf{U}_n : \mathcal{D})$ can be written as follows:

$$\prod_{\mathbf{u}} \prod_{x} \frac{\Gamma\left(\alpha_{x|\mathbf{u}} + M[x|\mathbf{u}] + 1\right) \tau_{x|\mathbf{u}}^{\alpha_{x|\mathbf{u}}+1}}{\Gamma\left(\alpha_{x|\mathbf{u}} + 1\right) \left(\tau_{x|\mathbf{u}} + T[x|\mathbf{u}]\right)^{\alpha_{x|\mathbf{u}}+M[x|\mathbf{u}]+1}} \tag{2.4}$$

while $MargL^\theta(X_n, \mathbf{U}_n : \mathcal{D})$ can be written as follows:

$$\prod_{\mathbf{u}} \prod_{x} \frac{\Gamma\left(\alpha_{x|\mathbf{u}}\right)}{\Gamma\left(\alpha_{x|\mathbf{u}} + M[x|\mathbf{u}]\right)} \times \prod_{x' \neq x} \frac{\Gamma\left(\alpha_{xx'|\mathbf{u}} + M[x, x'|\mathbf{u}]\right)}{\Gamma\left(\alpha_{xx'|\mathbf{u}}\right)}. \tag{2.5}$$

where Γ is the Gamma function, $M[x, x'|\mathbf{u}]$ represents the count of transitions from state x to state x' for the node X_n when the state of its parents $\mathbf{U_n}$ is set to $\mathbf{u}$, while $T[x|\mathbf{u}]$ is the time spent in state x by the variable X_n when the state of its parents $\mathbf{U_n}$ is set to $\mathbf{u}$. Furthermore, $M[x|\mathbf{u}] = \sum_{x' \neq x} M[x, x'|\mathbf{u}]$, $\alpha_{x|\mathbf{u}}$ and $\tau_{x|\mathbf{u}}$ are hyperparameters over the CTBN's q parameters while $\alpha_{xx'|\mathbf{u}}$ are hyperparameters over the CTBN's θ parameters. However, $Par(\mathcal{G})$ does not grow with the amount of data. Therefore, the significant terms of $FamScore(X_n, Par_{\mathcal{G}}(X_n) : \mathcal{D})$ are $MargL^q(X_n, \mathbf{U}_n : \mathcal{D})$

and $MargL^{\theta}(X_n, \mathbf{U}_n : \mathcal{D})$. Thus, given a dataset $\mathcal{D}$, the optimal CTBN's structure is selected by solving the following problem:

$$\max_{\mathcal{G}\in\mathbf{G}} \sum_{n=1}^{N} \ln MargL^{q}(X_n, \mathbf{U}_n : \mathcal{D}) + \ln MargL^{\theta}(X_n, \mathbf{U}_n : \mathcal{D}), \tag{2.6}$$

where $\mathbf{G} = \{\mathbf{U}_n \in \mathbf{X} : n = 1, ..., N\}$ represents all possible choices of parent set $\mathbf{U}_n$ for each node X_n, $n = 1, ..., N$. Optimization problem (2.6) is over the space $\mathbf{G}$ of possible CTBN structures, which is significantly simpler than that of BNs and DBNs. Indeed, learning optimal BN's structure is NP-hard even when the maximum number of parents for each node is limited, while the same does not hold true in the context of CTBNs. In fact, in CTBN all edges are across time and represent the effect of the current value of one variable to the next value of other variables. Therefore, no acyclicity constraints arise, and it is possible to optimize the parent set $\mathbf{U}_n$ for each variable X_n, $n = 1, ..., N$, independently. In [34] the authors proved that, if the maximum number of parents is restricted to k, then learning the optimal CTBN's structure is polynomial in the number of nodes N. However, we usually do not want to exhaustively enumerate all possible parent sets $\mathbf{U}_n$ for each variable X_n, $n = 1, ..., N$. In this case we resort to *greedy hill-climbing* search by using operators that add/delete edges to the CTBN structure $\mathcal{G}$. It is worthwhile to mention that family scores of different variables do

not interact. Therefore, the *greedy hill-climbing* search on CTBNs can be performed separately on each variable X_n, thus making the overall search process much more efficient than on BNs and DBNs.

2.3.1.1 CTBNs for gene network reconstruction

In a CTBN the amount of time that a gene spends in a given state before switching to a different state plays a central role. This is a key point since the duration of a regulatory interaction is known to be relevant. For example, Th17 cells tend to became pathogenic when the production of Il17a remains protracted for a long time. When cells become pathogenic, the regulatory interactions are different compared to the non-pathogenic phenotype. From this, it is clear how the duration of a certain regulatory event can trigger different reactions. The learned structure of a CTBN provides an intuitive and meaningful level of abstraction of the evolution of regulatory process over time. For instance, a transcription factor which works as permanent hub during the whole process will most likely be at the top of the inferred network hierarchy and characterized by a high degree of outgoing arcs. On the other hand, transcription factors which act only during some time intervals will likely appear at an intermediate level, with both incoming and outgoing connections. Intuitively, genes which are only regulated (i.e. cytokines) will be leaf nodes with only incoming arcs. In the learned network,

arcs are directed but do not encode information about positive or negative regulation. A direct arc between two genes implies a direct causal relation (regulation) between the pair. Longer paths between two nodes suggest that the influence of one gene on the other pass through a regulatory chain involving intermediate genes. Even if not displayed in the networks, auto regulation interactions, interaction directions (positive/negative) and relative timings are encoded within the conditional intensity matrices (CIMs) associated with each node. Let's consider the following example consisting of a small network of 3 genes and shown in Figure 2.1. The three variables are binary, for example the gene A can be in either the status a_0 = normally expressed or a_1 = over expressed. The set of CIMs below describes the full dynamic behavior of the system. Specifically, each CIM describes the expected times of transition of a node conditioned to the current state of its parents. Here, we assume the time unit being equal to one minute. If the gene C is normally expressed and both its parents A and B are currently over expressed, then its transient behaviors is described by the CIM $\mathbf{Q}_{C|a_1,b_1}$, which is telling us that the gene C is expected to switch from normally expressed to over expressed in $1/0.7 = 1.43$ minutes.

$$\mathbf{Q}_{A|c_0} = \begin{bmatrix} \text{-0.1} & \text{0.1} \\ \text{0.2} & \text{-0.2} \end{bmatrix} \quad \mathbf{Q}_{A|c_1} = \begin{bmatrix} \text{-0.5} & \text{0.5} \\ \text{0.1} & \text{-0.1} \end{bmatrix}$$

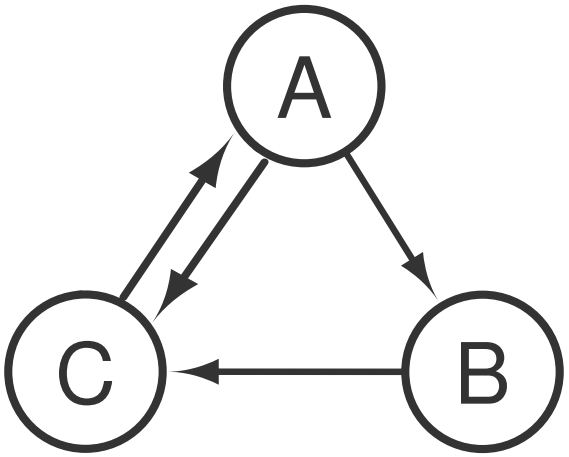

FIGURE 2.1: A simple gene network.

$$\mathbf{Q}_{B|a_0} = \begin{bmatrix} \text{-0.1} & 0.1 \\ 0.2 & \text{-0.2} \end{bmatrix} \quad \mathbf{Q}_{B|a_1} = \begin{bmatrix} \text{-0.5} & 0.5 \\ 0.1 & \text{-0.1} \end{bmatrix}$$

$$\mathbf{Q}_{C|a_0,b_0} = \begin{bmatrix} \text{-0.1} & 0.1 \\ 0.2 & \text{-0.2} \end{bmatrix} \quad \mathbf{Q}_{C|a_0,b_1} = \begin{bmatrix} \text{-0.5} & 0.5 \\ 0.1 & \text{-0.1} \end{bmatrix}$$

$$\mathbf{Q}_{C|a_1,b_0} = \begin{bmatrix} \text{-0.5} & 0.5 \\ 0.1 & \text{-0.1} \end{bmatrix} \quad \mathbf{Q}_{C|a_1,b_1} = \begin{bmatrix} \text{-0.7} & 0.7 \\ 0.1 & \text{-0.1} \end{bmatrix}$$

From this CIM $\mathbf{Q}_{C|a_1,b_1}$, the probability distribution over the possible states of C can be propagated forward to any continuous point in time, by calculating:

$$exp((\mathbf{Q}_{C|a_1,b_1}) \cdot \Delta t) \tag{2.7}$$

Where exp is the matrix exponential and Δt is the difference between the last known state for the parents of C and the time t for which we want to calculate the probability distribution of C. CIMs are learned together with the graph structure and represent the basis for the inference task, which is not directly investigated in this work.

2.3.2 Dynamic Bayesian networks

The definition of DBN has necessarily to start from the definition of a Bayesian network. A Bayesian network (BN) is a graphical model consisting of two components - a causal graph (qualitative component) which encodes conditional dependence and independence relationships between the variables (nodes), and a set of conditional probability tables (CPTs) (quantitative component) quantifying how strong is the influence of one variable on the others. More formally:

Definition 2.2. (Bayesian Network). [35] A BN consists of:

- A set of random variables (nodes) and a set of oriented arcs connecting the random variables which form a Direct Acyclic Graph (DAG).

- A finite set of mutually exclusive states associated with each random variable.

- For each random variable X with parents $Y_1, ..., Y_n$ a CPT encoding the probability $P = (X|Y_1, ..., Y_n)$. In other words, the CPT quantifies the effect of the parents $Y_1, ..., Y_n$ on X. If X has no parents, X is associated with an unconditional probability table, that is $P(X)$.

Exploiting the concept of conditional independence, a BN represents the joint probability distribution over a set of random variables in a compact way by factorizing it into a product of conditional distributions contained in the CPTs associated with each node in the graph.

Learning a BN involves:

- Parameter learning: learning of the conditional probability distributions.

- Structural learning: learning of the qualitative component of the network, e.g. the relations of conditional independence between variables.

- The goal of the learning phase is the finding of the structure and the parameters which best describe the initial data.

Bayesian networks are a static model since variables cannot change their state over time. Dynamic Bayesian networks (DBNs) [8] extend BNs by introducing a temporal dimension to represent dynamic systems. DBNs represent the state of the system through "snapshots" or 'time slices" of the system at each time point, where each "time slice" is a traditional BN.

In a DBN a random variable X_i can assume different values, one for each time point t: a "trajectory" is an assignment of values to each variable $X_i^{(t)}$ for each time t. A number of assumptions are made in order to keep this representation tractable [36].

The first assumption is discretization of time into time slices where system's measurements are assumed to be collected at regularly spaced time intervals. According to this assumption, we can reparametrize the joint probability distribution (using the chain rule) in the following way:

$$P(X^{(0)}, ..., X^{(t)}) = \prod_{t=1}^{T} P(X^{(t+1)}|X^{(0:t)}) \tag{2.8}$$

From equation (2.8) is it clear how the distribution over the trajectories is calculated as the product of the conditional distributions of the variables in each time slice given their values in the preceding ones.

The second assumption is the Markovian assumption that the state of X at the future time $t+1$ is independent from its past given its present, i.e, for every $t \geq 0$,

$$(X^{(t+1)} \perp X^{(0:(t-1))} | X^{(t)}) \tag{2.9}$$

Equation (2.8) can now be represented compactly as:

$$P(X^{(0)}, ..., X^{(t)}) = \prod_{t=1}^{T} P(X^{(t+1)} | X^{(t)}) \tag{2.10}$$

We can now formally define a DBN.

Definition 2.3. (Dynamic Bayesian network) [36]. A dynamic Bayesian network is a pair $(B_0, B_{\rightarrow})$. B_0 is Bayesian network over a set of random variables $X_1 ... X_n$ and represents the initial distribution over the states. $B_{\rightarrow}$ is a 2-timeslice Bayesian network (2-TBNs) which provides a transition model from the timeslice t to timeslice t+1. For any desired time span $T \geq 0$, the distribution over $X^{0:T}$ is defined as an "unrolled" Bayesian network, where for any $i = 1...n$:

- the structure and conditional probability distributions of $X_i^{(0)}$ are the same for those for X_i in B_0.
- the structure and conditional probability distribution of $X_i^{(t)}$ for $t > 0$ are the same as those for X_i' in $B_{\rightarrow}$

Is it therefore clear that a DBN represents the state of a system at different time points, but does not implement an explicit representation of time. A DBN for example cannot be queried to obtain a distribution over when a specific event takes place.

One of the most popular approaches for structural learning of a dynamic Bayesian network is to find the graph structure which maximizes the Bayesian information criterion (BIC) [37], which for a DBN is defined as follows:

$$\log P(\mathcal{D}|\theta) - \frac{d}{2}\log N \tag{2.11}$$

where θ are the estimated parameters of the structure, d is the number of parameters and N is dimensionality of the data. In equation 2.11 $\log P(\mathcal{D}|\theta)$ describes how well the graph predicts the data while $(d/2) \cdot logN$ keeps the model's complexity under control by penalizing the addition of edges to the graph. As it can be noted, the BIC does not depend on any *a priori* information. A survey of the structural learning algorithms for DBNs can be found in [38].

2.3.3 Granger causality

The Granger causality test was firstly conceived for the economic domain [17] and is based on a linear vector autoregressive model

(VAR). The intuitive idea behind it is that an effect never happens before its cause and translated into the GRN domain it can be explained as follows. Suppose we have a sequence of time measurements for the genes X and Y. X is said to Granger cause Y if the autoregressive model of Y is more accurate when based on the past values of both X and Y rather than Y alone. The accuracy of the prediction is measured through the variance of the prediction error. Let's suppose to have bivariate linear autoregressive model, for the variables X and Y defined as:

$$X(t) = \sum_{j=1}^{p} A_{xx,j} X(t-j) + \sum_{j=1}^{p} A_{xy,j} Y(t-j) + \epsilon_x(\mathrm{t}) \tag{2.12}$$

$$Y(t) = \sum_{j=1}^{p} A_{yx,j} X(t-j) + \sum_{j=1}^{p} A_{yy,j} Y(t-j) + \epsilon_y(\mathrm{t}) \tag{2.13}$$

Where p indicates the model's order, e.g. the number of past observations of the time series to incorporate in the autoregressive model. The impact that each one of these observations has on the predicted values of X and Y is contained in the matrix A. ϵ represents the prediction error for the time series (residuals). Considering the first equation, for Y to Granger cause X the variance of ϵ_x must be smaller than the variance of ϵ_x when the Y term is removed from the equation. This original GC formulation is meant to uncover causal relationships among two variables; in multivariate systems a pairwise analysis of this kind applied to all possible pairs of variables is limited in the type of causal relationships that can be uncovered. For this reason, this concept was extended [18, 39] to the analysis of multivariate data by introducing the concept of conditional GC. Suppose we have a system with three variables, X, Y and Z. Intuitively, the multivariate linear autoregressive model for the variable X can be written as:

$$X(t) = \sum_{j=1}^{p} A_{xx,j} X(t-j) + \sum_{j=1}^{p} A_{xy,j} Y(t-j) + \sum_{j=1}^{p} A_{xz,j} Z(t-j) + \epsilon_x(\mathrm{t}) \tag{2.14}$$

In the equation above, Y Granger causes X if the variance ϵ_x is smaller than what it would be when the Y term is removed from the equation. VAR models have the undeniable advantage of being well-understood and widely applied in many disciplines such as neurosciences, economics and biology. In this work GC, like in almost the totality of its applications and theoretical investigations, is considered in its formulation which assumes the observations to be taken at regular and fixed time intervals. As it is underlined in [40], the Granger causality test can be sensitive to the sampling frequency of the time series, with the risk of the results of the test being biased. Many theoretical efforts have been made to extend this formulation to enable it to directly accommodate time. However, most of the contributions remain theoretical and not much investigation has been performed about adequate test statistics [41]. GC is usually applied in its linear version. However, gene expression data is known to contain non-linear features. Many extensions of GC to the non-linear case have been proposed. Hiemstra and Jones [42] investigated a nonparametric test for both linear and non-linear Granger causality in the economic domain (HJ test), resulting in their method being used in a number of subsequent works. However, Diks and Panchenko [43] more recently showed that the HJ test has a tendency to detect spurious non-linear GC. Among other

alternatives proposed to deal with nonlinearities are kernel methods [44], with many kernels being proposed and the Gaussian being one of the most common. Non-linear extensions of GC have to deal with the issue of overfitting, which makes the statistical interpretation of the results less clear [45]. Moreover, it is known that different non-linear transformations lead to different results of the GC test [46]. A recent study [47] showed that for Gaussian distributed variables, non-linear GC approaches cannot account for any extra information in the data, because a stationary Gaussian autoregressive process is necessarily linear. For these reasons, in this study GC is considered in its linear approximation, which has been found to work well on systems characterized by a large number of variables.

2.4 Results

2.4.1 Simulated data

Simulated datasets are important for benchmarking the accuracy of gene regulatory network reconstruction as the true network structure is known *a priori*, which is seldom the case with real biological datasets. In this section simulated time course datasets have been used to benchmark the accuracy network reconstruction with GC, DBNs and CTBNs.

The datasets datasets were generated by the same methodology as was used in the DREAM4 competition [48], extracting subnetworks from the known *in vivo* gene networks of *E. coli* [49] and *S. Cerevisiae*. Subnetworks were extracted by randomly choosing a seed node and progressively adding nodes with the greedy neighbor selection procedure, which maximizes the modularity and is able to preserve the functional building blocks of the full network [50].

To ensure robustness, our studies are not based on one single network instance, but are always based on a set of 10 different networks instances. The reconstruction algorithms are tested under several conditions: for increasing number of nodes in the network (network dimension), for different time points densities in the dataset (time course granularity) and for datasets with time measurements not evenly but unevenly distributed (randomly spaced). The accuracy of network reconstruction was measured using the F_1 measure for binary classification which is defined as:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

where $precision = \frac{\text{true positive arcs}}{\text{true positive arcs + false positive arcs}}$

and $recall = \frac{\text{true positive arcs}}{\text{true positive arcs + false negative arcs}}$.

in statistic the *recall* is referred to as *sensitivity* and the *precision* as *positive predicted* value.

Benchmarking for increasing network dimension

The first step of our analysis on simulated data consisted in studying how the three methods perform when faced with the task of reconstructing gene networks of different dimensions. From the known *in vivo* gene regulatory network structures of *E. coli* [49] and *S. cerevisiae* we randomly extracted sets of 10 networks consisting of 10, 20, 50 and 100 genes for both organisms. For the sake of brevity, the sets of 10 networks consisting of 10, 20, 50 and 100 genes will be referred to as 10-NETs, 20-NETs, 50-NETs and 100-NETs respectively. Statistical analysis of the complexity of the extracted network structures is provided in Figure 2.2.

The generated dataset consists of 21 evenly spaced time points. This dataset aims to simulate the amount of data that high-throughput techniques will soon generate while maintaining a realistic time course magnitude: expression microarray experiments repeated with this many time points are today possible. On the other hand, the dataset is still unrealistically rich in terms of number of perturbations and replicates. Such a comprehensive dataset is however necessary to fairly compare the analyzed methods.

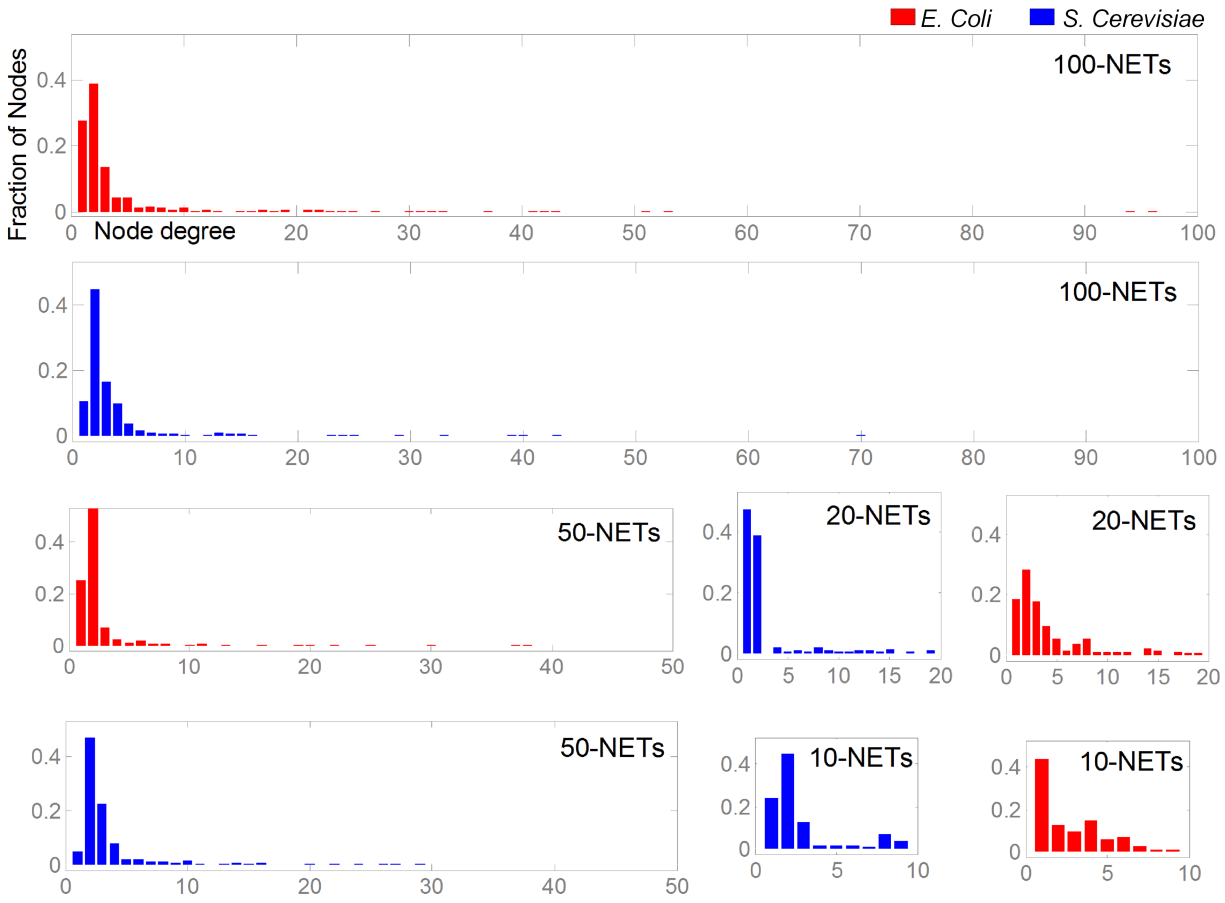

FIGURE 2.2: Degree distribution (in-degree plus out-degree) of nodes in *E. coli* (red) and *S. cerevisiae* (blue) for 10-NETs, 20-NETs, 50-NETs and 100-NETs. Each distribution is obtained from the data of all 10 sampled network instances. X-axis has been shifted up for better visibility. The distribution shows the presence of both large and intermediate hubs indicating that the networks are non-trivial.

Prior to learning, we performed an empirical *optimization* of the model parameters for the three methods; for CTBNs and DBNs this included experimentally establishing the optimum number of discretization levels. More details can be found in a dedicated section at this end of the document.

Results on *E. coli* dataset are summarized in Table 2.1 (top), where aggregate F_1 values are calculated as the arithmetic mean over the sets of 10 sampled network instances, and Figure 2.3A, where the individual F_1 values obtained by the methods on the 10 sampled network instances are represented through boxplots. For 50-NETs and 100-NETs learning with DBNs became computationally intractable therefore the corresponding results are not available. It can be concluded that the reconstructed network structures were the most accurate for CTBNs which outperformed DBNs and GC for 10-NETs, 20-NETs, 50-NETs and 100-NETs in terms of the mean F_1 values. A paired t-test confirmed that the F_1 values for CTBNs were significantly higher than for DBNs and GC in all tested network dimensions (p-value cutoff 0.05). Moreover CTBNs were robust with respect to the increasing network dimension: their performance smoothly degraded as the number of nodes of the network increased. Indeed, the difference between mean F_1 values for CTBNs and GC increased progressively with the network's dimension. GC outperformed DBNs on 10-NETs (0.13 mean F_1 gap)

while on 20-NETs GC were only marginally more accurate than DBNs with a limited mean F_1 difference of 0.02.

Results on *S. cerevisiae* dataset shown in Table 2.1 (bottom) and Figure 2.3B reaffirmed the same conclusions even more emphatically. CTBNs outperformed DBNs and GC for all network dimensions, with the mean F_1 difference between CTBNs and GC increasing from 0.17 for 10-NETs up to 0.29 for 100-NETs. Interestingly, on this dataset DBNs outperformed GC (+0.04 mean F_1 on 10-Nets, +0.17 mean F_1 on 20-NETs). The paired t-test confirmed the significant superiority of CTBNs in all cases over both DBNs and GC. DBNs were significantly better than GC on 20-NETs.

As a negative test we also simulated a *random* reconstruction method which starts with an empty graph and randomly adds edges to it. As expected, this random algorithm had low precision while its recall stabilized around 0.50. As shown in Table 2.1 the performances of the three methods were always better than the random algorithm, confirming their effectiveness.

Benchmarking for increasing time course granularity

The second set of tests are conceived to compare the network reconstruction algorithms with time course datasets of increasing time

granularity. Although the overall duration of the simulated experiment was kept fixed, measurements were collected at increasing frequencies (11, 21 and 31) of evenly spaced time points. As in the previous section, datasets were generated for both *E. coli* and *S. cerevisiae*. The network dimension was kept constant at 20 nodes as this was seen in the previous section to represent a good trade-off between network complexity and computational cost.

Results on *E. coli* are shown in Table 2.1 (top) and Figure 2.4A. Looking at the aggregate F_1 values calculated as the arithmetic average over the sets of 10 network instances (Table 2.1 (top)) it can be observed that GC appeared to perform consistently, achieving mean F_1 values of 0.50, 0.49 and 0.47 for granularities 11, 21 and 31 respectively, whereas both DBNs and CTBNs achieved their peak performance for a time granularity of 21. DBNs performed poorly (mean F_1 0.26) for a low time granularity of 11, best for granularity 21 (mean F_1 0.47) and achieved a slightly lower accuracy for granularity 31 (mean F_1 0.40). CTBNs achieved a slightly lower accuracy than GC for time granularity 11 (mean F_1 0.47), achieved the overall best performance for time granularity 21 (mean F_1 0.57) and had a slightly lower accuracy for granularity 31 (mean F_1 0.54). A paired t-test over the F_1 values concluded that CTBNs performed significantly better than DBNs for all time course granularities (p-value) and also better than GC (p-value) with the exception of time

courses of granularity 11. Finally, GC proved to be significantly better than DBNs for granularity 11 while no statistically significant difference emerged between the two for higher time granularities. The three methods share the trend of reconstruction accuracy initially increasing from time granularity 11 to 21, reaching a peak at 21 and then decreasing for granularity 31: this behavior could be explained by the fact that the *optimal* number of discretization levels has been empirically established for time granularity 21 data and subsequently applied to time granularity 11 and 31 data. The discretization level applied to granularity 31 data may be therefore *suboptimal*.

Results on *S. cerevisiae* are shown in Table 2.1 (bottom) and Figure 2.4B. GC performed well on time courses of granularity 11 achieving a mean F_1 of 0.57; however, the drop of effectiveness for granularities 21 and 31 was clear with mean F_1 values of 0.41 and 0.42 respectively. CTBNs were always the most accurate achieving mean F_1 values of 0.60, 0.70 and 0.60 for the three time course densities. Again, DBNs performed poorly for granularity 11 (mean F_1 0.32, with a -0.28 gap from CTBNs), while better for more finely grained data (0.58 and 0.48 mean F_1). With the exception of granularity 11, DBNs outperformed GC, which is the opposite of what we observed for *E. coli* datasets. A paired t-test concluded CTBNs significantly outperformed DBNs for all time granularities and GC for granularities 21 and 31. Interestingly, is possible to prove the superiority of

GC over DBNs for granularity 11, while vice-versa for granularity 21.

It has to be noted that, due to the computational cost of an empirical *optimization* of the model parameters, the CTBNs' hyper-parameters α and τ for time courses of granularity 11 and 31 have not been found. Instead, the optimal parameters found on granularity 21 data were maintained. Consequently, the results achieved by CTBNs are to be considered sub-optimal.

Benchmarking evenly vs. unevenly spaced time measurements

The third step of our analysis on simulated data consisted in evaluating how the performance of the three methods changes when the time measurement are not evenly spaced over time but randomly sampled. This is a typical scenario in wet-lab experiments.

For the purpose of the test, 10 different random time point instances were sampled and used to generate 10 unevenly distributed time course datasets; test were run on the set of 20-NETs of the the organism *E. coli*. We repeated the numerical experiments for time courses of granularity of 11, 21 and 31 (keeping the 10 random time point instances consistent).

Results are shown in Figure 2.5 and are consistent for all the three time course granularities (panels A, B, C). For all the network instances, the minimum F_1 value achieved by DBNs among the 10 unevenly (randomly) sampled time points instances is always smaller than the minimum F_1 value achieved by CTBNs on the same 10 unevenly sampled time points instances. Furthermore, the maximum F_1 value achieved by DBNs on the same samples is always smaller than the maximum achieved by CTBNs, for all network instances and time course granularities. The result is clear, showing that CTBNs are always preferable to DBNs when the time course data is not evenly spaced. CTBNs and GC performed similarly with respect to both minimum and maximum F_1 values (for all granularities). GC was better than DBN with respect to both minimum and maximum F_1 values (for all granularities), with only a few cases for which DBNs resulted to be preferable.

2.4.2 Synthetic gene network in *S. cerevisiae*

Due to the current lack of reliable large scale gold standards, *in vivo* evaluation is a critical point for GRN reconstruction methods which often rely on less quantifiable evaluations such as comparison with existing literature and/or information available in public databases. The benchmarking of CTBNs was performed on a small but *certified network*: a network consisting of five genes synthetically constructed

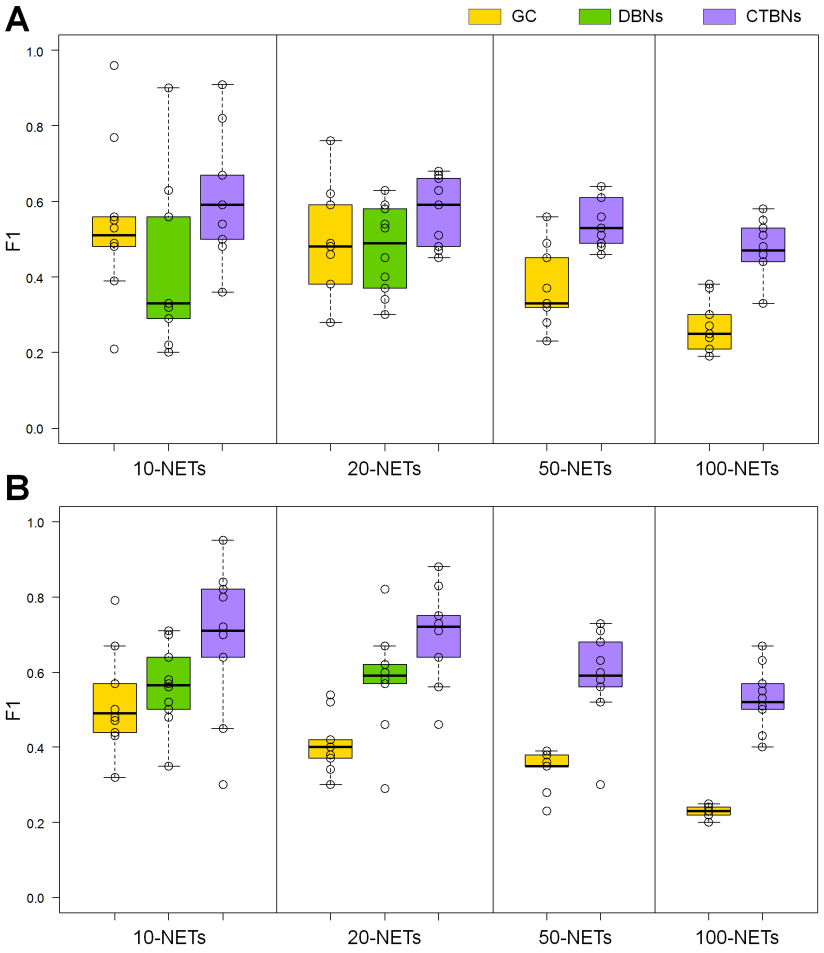

FIGURE 2.3: Performance comparison of CTBNs, DBNs and GC on simulated data for different network dimensions. Organism *E.coli* (A) and *S. cerevisiae* (B). Boxplots represents the F_1 values obtained on the 10 sampled network instances of each dimensions, which are also plotted individually as circles.

in the yeast *S. cerevisiae* [29] and shown in Figure 2.6 was used. This network, despite its small size, contains a representative set

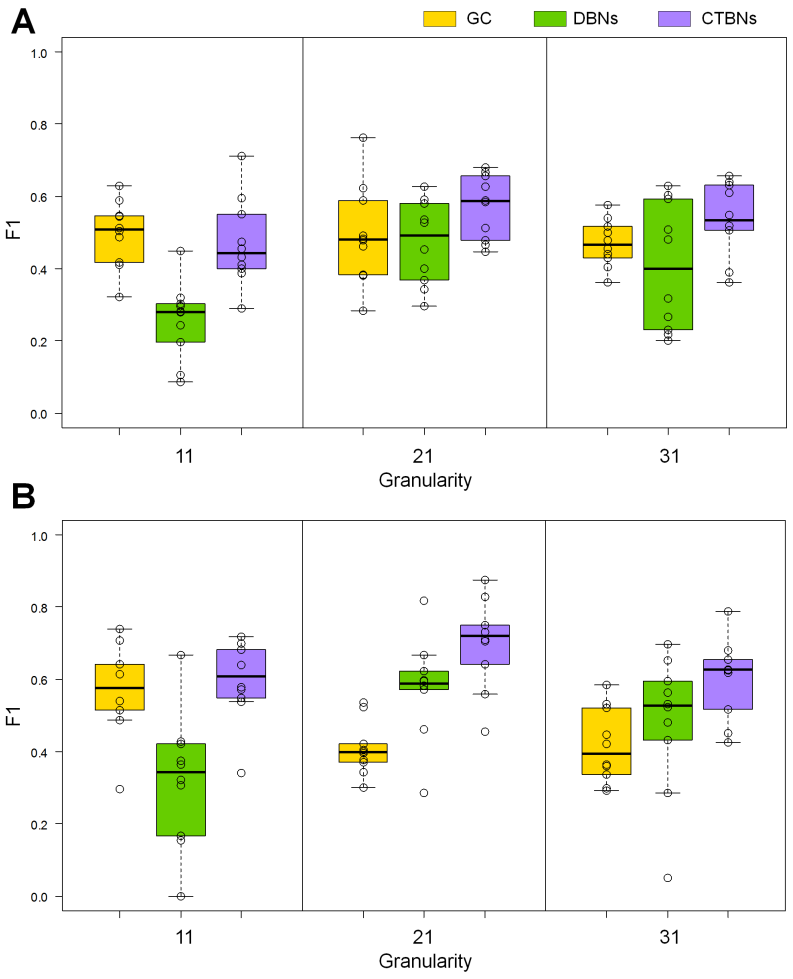

FIGURE 2.4: Performance comparison of CTBNs, DBNs and GC on simulated data for different time granularities on 20NETs, organism *E. coli* (A) and *S. cerevisiae* (B). The set of 20NETs does not change, what changes is the granularity of the time course data generated from the networks. Boxplots represents the F_1 values obtained on the 10 sampled network instances of each dimensions, which are also plotted individually as circles.

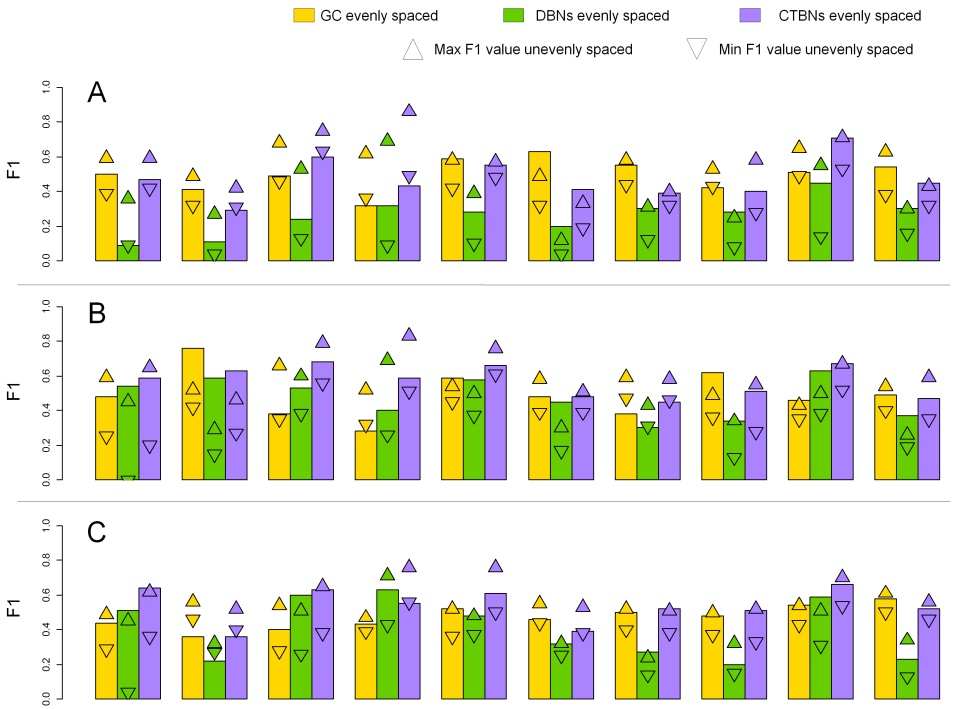

FIGURE 2.5: Performance comparison of CTBNs, DBNs and GC on simulated data for evenly vs. unevenly spaced time points on 20NETs, organism *E.coli*. Panel A, B and C refer to the time course data with granularity 11, 21 and 31 respectively. Bars represent the F_1 achieved by the method when the time points were evenly spaced. For the unevenly spaced case, we report the minimum (down triangle) and the maximum (up triangle) F_1 values obtained by the method over the set of 10 unevenly sampled time points instances. The sampled time points are consistent among the three methods.

of interconnections including regulator chains and feedback loops. The dynamic behaviour of the network was studied by shifting cells from glucose to galactose and vice-versa, and collecting samples every 20 min up to 280 min for the switch-on and every 10 min up to 190 min for the switch-off. 4 and 5 biological replicates were analyzed respectively, gene expression levels were measured through RT-PCR. The authors also made available some interventional data obtained by over expressing each of the five genes in cells grown in either glucose or galactose; however, since only steady-state data was generated for this perturbational experiments, the benchmark was performed on time course unperturbed data alone. On the *S. cerevisiae* experimental dataset the results were coherent with those obtained on simulated datasets: CTBNs outperformed DBNs and GC. A graphical representation of the true network compared with the ones inferred by DBNs, GC and CTBNs is provided in Figure 2.6. CTBNs achieved both the maximum value of true positives (5) and the minimum value of false negatives (3) while all the three methods made exactly one false positive prediction each.

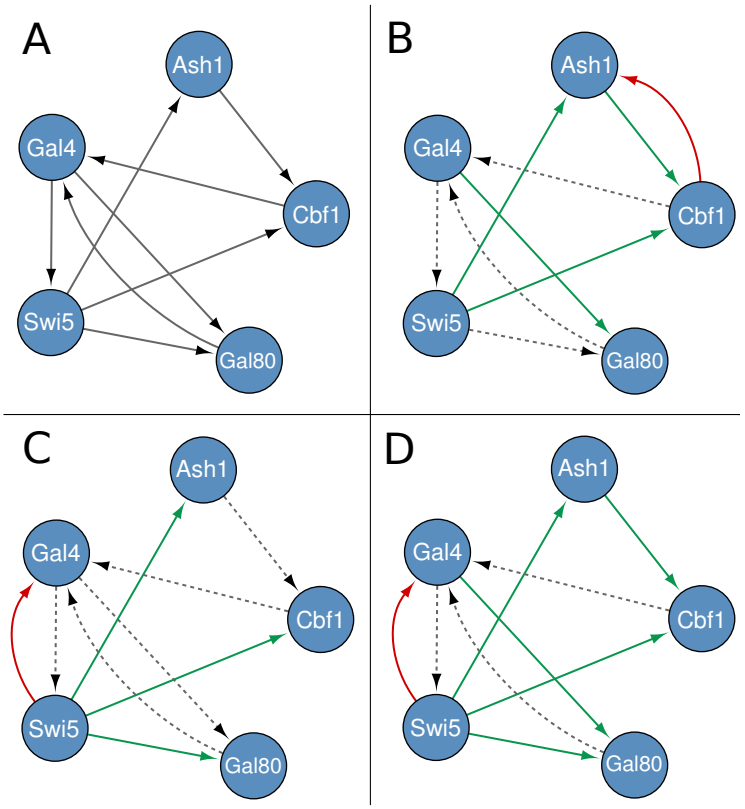

FIGURE 2.6: Performance comparison on *S. cerevisiae* experimental data. True network (A), network inferred by GC (B), DBNs (C), CTBNs (D). Green arcs represent true positives, red arcs false positives and dotted lines false negatives.

2.4.3 Elucidating the regulatory network responsible for murine Th17 differentiation using CTBNs

Gene regulatory networks have been described extensively in the regulation of immune response but more importantly in the control

of inflammation. Inflammation is a multifaceted cellular response critical for the protection of the host against different types of injuries as infections. However, the dark side of the inflammatory process is represented by tissue damage since inflammatory responses react against self-tissues. Precise regulation of gene expression is extremely important in the context of inflammation for host survival under its own immune activation. In particular, gene regulation of inflammatory cellular differentiation appears essential for fine-tuning of the entire inflammatory response. At the onset of chronic inflammation, Th17 cellular response is of particular interest. Th17 cells produce well-known soluble molecules such as IL17A, IL17F and IL21 which are important for neutrophil recruitment, infection clearance and delivery of antimicrobial peptides. Fine tuning of Th17 cells differentiation program appears to be pivotal for proper control of over exuberant inflammatory processes in the vertebrate immune system. While some key regulators of the Th17 differentiation are known, a large portion of the regulatory mechanisms controlling this process remains unclear.

Naive T cells (or Th0) can be polarized to differentiate into one of the T helper phenotypes (such as Th1, Th2, or Th17) by exposing them to various polarizing cytokines. The external signals through cytokines drive different regulatory pathways within the cells and gene regulatory networks involving master regulator transcription

genes dictate the final differentiation status. Th0 cells can be programmed to undergo differentiation into the Th17 phenotype by activating transcription factors such as Stat3 and RORγt through soluble molecules such as IL6, TGFβ, IL1β. Furthermore, stabilization of the Th17 phenotype requires the activation of IL23R receptor through the innate cytokine IL23 [51].

Here, structure learning of CTBNs is applied to elucidate the gene regulatory network controlling differentiation of murine naive Th0 to the Th17 phenotype. Data for this study is derived from a recently published time course microarray experiment [52] resulting in transcriptional profiles obtained during murine Th17 differentiation. The microarray measurements were taken with Th0 cells cultured in the presence or absence of polarizing cytokines IL6 and TGFβ1 in two biological replicates. Measurements were taken at 18 time points unevenly spanned over the first 72 hours following induction. Furthermore, separate measurements were taken involving perturbation with the stabilizing innate cytokine IL23 50h from the start of polarization. This dataset is one of the longest and more finely grained time course data ever generated in the T helper differentiation context, with a total of 58 gene expression microarray samples. In order to keep the results interpretable, the analysis was restricted to the representative set of 275 genes individuated by the authors [52] as reflecting as many aspects of the differentiation program as possible. The bioinformatic analysis of raw data and

the data discretization process allowed to further narrow down this set to 186 genes (excluding genes whose fold-changes levels resulted to be constant among all the time points). More details about the pre-processing steps can be found at the end of the document. Since the goal of this study is to investigate mechanisms which are characteristic of the IL6+TGFβ1 type and not those regulatory fluctuations which take place independently of the differentiation process (in both Th0 and IL6+TGFβ1 cells), fold-change values of IL6+TGFβ1 versus Th0 were used as input data for the learning algorithm.

Two separate networks have been learned: the first one using unperturbed time course series (from fold changes IL6+TGFβ1 vs. Th0), the second one using the time course series with the addition of the Il23 cytokine into the culture (from fold changes IL6+TGFβ1+IL23 vs. Th0+IL23). In order to evaluate which mechanisms are relevant to the stabilization of the phenotype, we looked at differences among the two networks. If the perturbations would have been of the type of gene knock-outs and/or gene knock downs, the correct way to proceed would have been to learn one single network from both the unperturbed and perturbed data. Here, the perturbation is of stabilizing nature, e.g. it enhances differentiation process through the activation of additional regulatory mechanism and the inhibition of others. For simplicity, from now on we will refer to

the first network as IL6+TGFβ1 network and to the second one as IL23 network.

While a few attempts have been recently made to elucidate the molecular mechanisms of the Th17 stabilization following the addition of IL23 [53, 54], the validation of the network dynamic is still open to debate. Consequently, the interpretation and validation of results is more difficult on the IL23 network than on IL6+TGFβ1. For this reason, a large part of the discussion and quantitative validation of the results refers to the IL6+TGFβ1 network, while only main differences and specific interesting mechanisms that emerged in the IL23 network are discussed.

Network validation in absence of gold-standard

CTBNs bring to light the interactions happening in between densely sampled time slices, resulting in a detailed description of all the regulatory steps taking place over the 72 hours differentiation process. Due to the lack of biological analysis with this level of detail, validation through literature evidence of the inferred network is non trivial. Indeed, literature evidence of gene interactions are often derived from studies based on static or coarsely grained measurements. As a consequence, what emerges from such studies can be incomplete since the known set of interactions may represent only

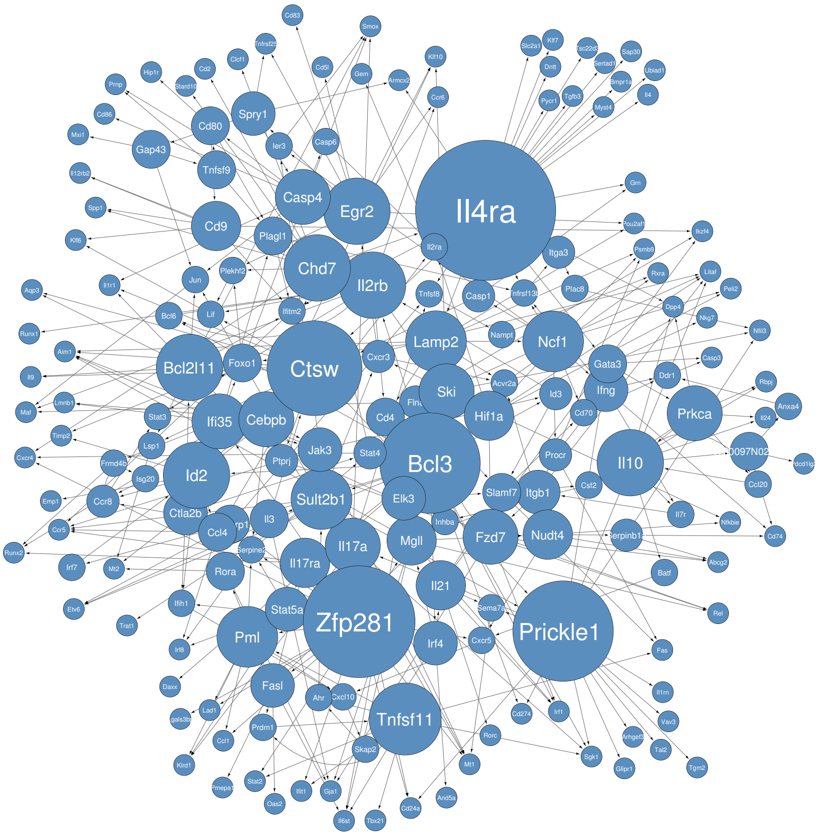

FIGURE 2.7: IL6+TGFβ1 inferred network. Node sizes are proportional to the number of outgoing arcs.

a subset of all the interactions that are taking place. For this reason, a validation approach that tries to enumerate how many of the predicted direct interaction are known is not a reliable one. On the other hand, it is known that when considering networks encoding temporal interactions like in the case of CTBNs, the graph can allow cycles. In this situation, the presence of an incorrectly

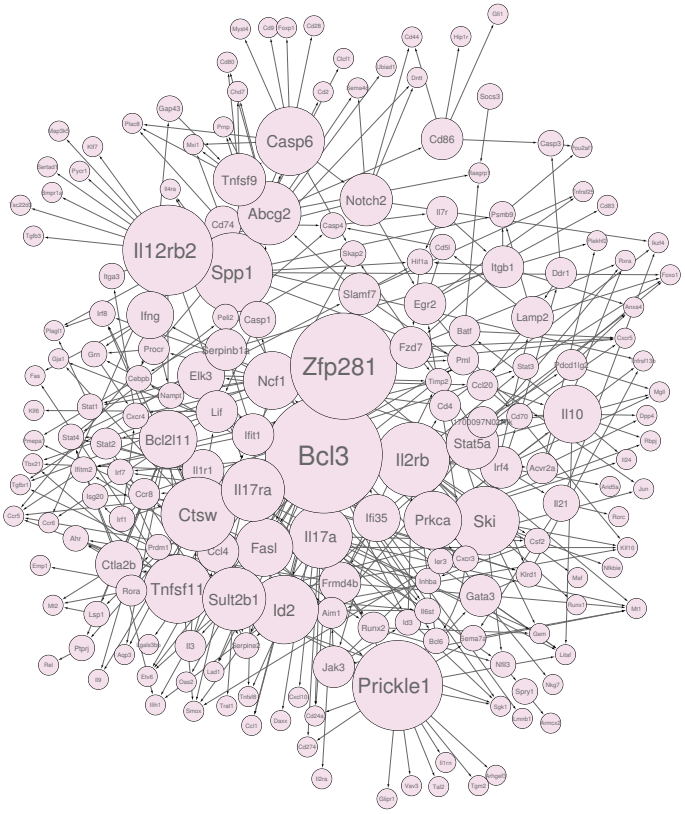

FIGURE 2.8: IL23 inferred network. Node sizes are proportional to the number of outgoing arcs.

inferred arc at some point of the network (something likely to happen) creates a large number of additional paths connecting genes. For this reason, a validation approach which tries to find a pathway between genes known to be related could lead to biased results, where incorrectly inferred arcs paradoxically lead to a bigger number of true positives. It is clear that the benchmarking of CTBNs in the absence of a gold-standard cannot be performed in a purely quantitative way, but it has to be complemented with a reasoned

biological interpretation of the network.

Quantitative validation of the IL6+TGFβ1 inferred network

The IL6+TGFβ1 network inferred from data is shown in Figure 2.7. The graph is characterized by 186 nodes connected by 365 arcs. For 67 of these arcs solid literature evidence has been found. Only direct known relations were considered while known relations separated by one or more unknown intermediary nodes were not included in this statistic. A list of these known interactions together with related PubMed IDs is provided in Table 2.2. Among the listed arcs, 14 appeared in the predicted IL6+TGFβ1 network with a reverse orientation compared to the literature. This is a well known problem with reconstructing networks referred to as model *non-identifiability*, which arises when given the data, it is not possible to recover (learn) a unique set of parameters. Instead, in such situations we have multiple sets of parameters settings that are indistinguishable given the data [36]. The *non-identifiability* of a model can be due to data scarcity (and/or lack of interventional data) or presence of hidden variables. Given that we are examining a subset of genes, both hypothesis are possible. For these reasons, the inverted interactions were considered as valid.

An additional assessment of the validity of the inferred network was performed by looking at the leaf nodes (nodes with no children) and the root nodes (nodes with no parents).

In the temporal network semantic leaf nodes are associated with final products (cytokines in our case). In the inferred IL6+TGFβ1 many of the leaf nodes represented soluble immune mediators, which usually characterize the cells at final steps of their differentiation processes. These include the cytokines *Il4*, *Il9*, *Il24*, *Il1rn*, *Clcf1* and *Tgfb3*, cytokine signal transducer *Il6st* which is shared by many cytokines, cytokine receptors such as *Il12rb2*, *Il1r1*, chemokines such as *Ccl1*, and chemokine receptors such as *Ccr5*, *Ccr6*, *Cxcr4*. Among leaf nodes we also found clusters of differentiations such as *Cd2*, *Cd24*, *Cd274*, *Cd86* which represent a clear marker of the final steps in acquisition of the terminal Th17 phenotype. Furthermore, apoptosis markers like *Casp3*, *Fas*, *Daxx*, *Vav3*, *Trat1*, *Tnfrsf25*, *Tgm2*, *Sertad1* together with programmed cell death 1 ligand 2 (*Pdcd1lg2*) which follow T cell activation and exhaustion were correctly associated with leaf nodes. Transcription factors regulators of late phases of the differentiation processes as *Tbet*, *Runx2*, *Runx1*, *Rorc*, *Maf*, all responsible of the final steps for the definition of the Th17 cell phenotype, are correctly placed at the end of the chain. Finally, *Sgk1* is a recently discovered marker identifying the Th17 pathogenic phenotype, acquired by T cell at the late phases of the

T cell polarization [55]; in our *Sgk1* network is correctly represented as leaf node.

Conversely, root nodes are associated with molecules at the beginning of the cascade. Two root nodes were observed at the top of the network structure and both appear to be correctly so with their role in initiating the differentiation cascade. One of them is Filamin A (*Flna*), an actin binding and signal mediator scaffolding protein, required for T cell activation in response to TCR activation, also known as "signal1" [56]. Same applies to *Bcl3*, which is known to be activated in response to initial TCR activation [57]. The role of *Bcl3* is discussed more in details in the next paragraphs as new interesting insight related to its role emerged from the network.

Topological properties and hub nodes of the IL6+TGFβ1 inferred network

From a topological point of view, the sparsity of the learned causal structure (186 nodes, 365 arcs) is appreciable. From a theoretical point of view, given that the number of variables under study is several order of magnitude bigger than the data sample size, network sparsity is something that reconstruction methods seek for [58]. A network densely connected may indicate that the learning algorithm is failing in identifying true causal relations. Furthermore, sparsity has been shown to be a feature of regulatory networks [26, 59].

Even considering that the number of potential arcs was limited by the maximum number of parents allowed per node, which was set to 5, the learned network with 365 interactions (arcs) connecting 186 nodes remains way below such threshold. Another topological feature of the network which emerged is the presence of a few hub nodes regulating a vast number of other genes together and signs of naturally occurring modularity. Both of these features are well-known characteristics of gene networks. Interestingly modularity has been shown to be a characteristic of static gene networks but so far modularity has not been studied as a characteristic of networks evolving over time.

A major hub node in the network is *Il4ra*, the receptor of the cytokine *Il4*, shown in figure 2.8A. Its role in Th2 differentiation is well known, but more interestingly, its preeminent role in regulating Th17 differentiation is a subject of current investigation. Importantly, an inherited polymorphism of *Il4ra* seems to control the ability of the human immune system to regulate the magnitude of Il17 production [60]. Thus a central role of *Il4ra* in negative regulation of Th17 differentiation is expected [61].

Other major hub nodes include Cathepsin W (*Ctsw*), *Bcl3*, *Zfp281*, *Il4Ra*, *Prickle1* and *Tnfsf11*. Among these *Bcl3* and *Tnfsf11* are known to have a significant influence on Th17 differentiation. *Bcl3*,

a member of *IkB* family of proteins, is an essential negative regulator of Toll-like receptor-induced responses and inhibitor of NFkB. Reduced *Bcl3* expression has been associated with Crohn's disease [62] which is known to be mediated by Th17 chronic expansion. *Bcl3* has an inhibitory role in regulating IL17 release [63]. Indeed, *Bcl3-/-* mice develop autoimmune diabetes with increased Th17 type cytokine expression. Therefore, *Bcl3* is correctly inferred as hub node. *Tnfsf11* alias *Rankl* is known to be a marker of pathogenic Th17 cells in inflammation and therefore its status as hub in the network is correct [64]. *Ctsw* is a member of the peptidase C1 family, a cysteine lysosomal proteinase that plays a crucial role in the turnover of intracellular proteins as antigens and has a specific function in the mechanism or regulation of $CD8^+$ T-cell cytolytic activity [65]. However its role in Th17 differentiation is presently unknown. Similarly the role of *Zfp281*, a zinc finger transcription factor required in embryonic stems cells for pluripotency [66], and *Prickle1*, a nuclear receptor which is a negative regulator of Wnt/beta-catenin signaling pathway, in Th17 differentiation is yet unknown.

Impact of IL23 addition on the differentiation process

As mentioned, by looking at differences between IL6+TGFβ1 and IL23 networks we can analyses the impact that the addition of the

IL23 cytokine has on the differentiation process. Significant differences emerged between the two networks (IL23 network shown in Figure 2.8). 165 arcs that were present in the IL6+TGFβ1 network disappeared in the IL23 network while 173 new arcs appeared, confirming the widespread impact that IL23 treatment has on the regulatory interactions taking place in the cells [52].

It is interesting to observe how the hub nodes in the IL6+TGFβ1 network are affected by IL23 perturbation. Considering that the IL23 perturbation represents a positive impulse in Th17 differentiation, it is expected that the IL23 network will not contain hubs that represent a negative regulation of the Th17 differentiation process. This is the case with *Il4ra*, which loses all its outgoing connections and its status as a hub in the IL23 network. On the other hand, IL23 network is expected have hub nodes which positively regulate the Th17 phenotype. Some newly introduced hubs in the IL23 network include *Il12rb2* and *Il2rb*, both of which are well known for being positive regulators and hubs of the phenomenon [67–69]. *Il2rb* is known to strongly influence the regulation of Th17 differentiation depending on the levels of *Il2* [70]. Another hub node, *Spp1* [71], is particularly interesting because while *Spp1* is known to increase Th17 differentiation, its direct relation with IL23 is still unproven.

Some specific well-known regulatory mechanisms emerged both in

the IL6+TGFβ1 and IL23 networks together with the new biological insights which can be derived from them are discussed in the next section.

2.5 Discussion

2.5.1 Comparative study

For the first time continuous time Bayesian networks (CTBNs) were applied to the gene regulatory network reconstruction task from gene expression time course data. A comparison with two state-of-the-art methods, i.e. dynamic Bayesian networks (DBNs) and Granger causality analysis (GC), was conducted. Method's performances were analyzed in three different directions: for networks of increasing dimension, for time course data of increasing granularity and for evenly versus unevenly spaced time course data.

CTBNs achieved the highest value of the F_1 measure for all network dimensions and both *E. coli* and *S. cerevisiae*. Furthermore, they suffered from a limited and smooth loss of performance with respect to the networks of increasing size. This suggests that if applied to networks larger than those analyzed in this paper, CTBNs can still effectively help to uncover the causal structure of the regulatory

network. These aspects makes CTBNs a good candidate for solving the reconstruction of regulatory networks, which are systems characterized by a large number of variables.

CTBNs were the best performing approach when the time course granularity was sufficiently fine (21 and 31 time points in our experiments), while for coarser granularities (11 time points) CTBNs and GC performed analogously. DBNs performed poorly in the granularity 11 case, showing a big gap from CTBNs and GC on both organisms. The result of CTBNs for granularity 11 was unexpected: probabilistic approaches tend to require a big amount of data in order to be effective.

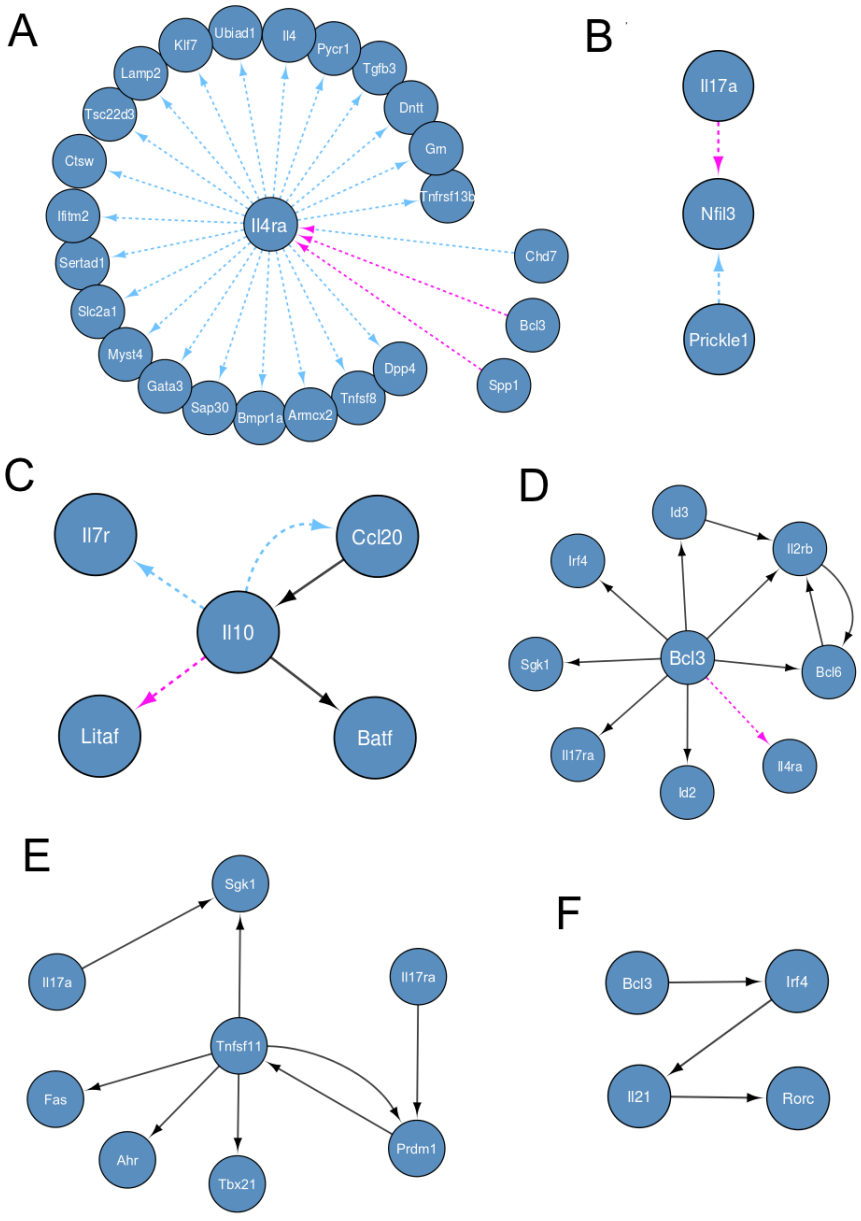
A
Klf7
Ubiad1
Il4
Pycr1
Tgfb3
Dntt
Gm
Tnfrsf13b
Chd7
Bcl3
Spp1
Dpp4
Tnfsf8
Armcx2
Bmpr1a
Sap30
Gata3
Myst4
Slc2a1
Sertad1
Ifitm2
Ctsw
Tsc22d3
Lamp2
Il4ra
B
Il17a
Nfil3
Prickle1
C
Il7r
Ccl20
Il10
Litaf
Batf
D
Id3
Irf4
Il2rb
Sgk1
Bcl3
Bcl6
Il17ra
Id2
Il4ra
E
Sgk1
Il17a
Il17ra
Tnfsf11
Fas
Ahr
Tbx21
Prdm1
F
Bcl3
Irf4
Il21
Rorc

FIGURE 2.8: Some selected interesting known and novel regulatory mechanisms that emerged from the inferred IL6+TGFβ1 and IL23 networks. Light-blue arcs are specific to the IL6+TGFβ1 network, while pink arcs are specific to the IL23 network. Black arcs are present in both networks.

Thanks to their explicit representation of the time, CTBNs were always preferable to DBNs when the time points are not evenly spaced: the worst case in terms of F_1 value that one can obtain when learning a network from unevenly sampled data (over 10 random samples) is always better than the worst case obtainable when learning with DBNs. The same favorable situation for CTBNs applies to the best cases. Considerations made for CTBNs over DBNs applies to GC over DBNs as well, while CTBNs and GC showed a similar behavior in response to unevenly spaced data. The poor performance of DBNs on unevenly spaced data is due to the observational model assumption on which is built their representation of the time: variables are assumed to evolve at fixed increments; when that is not the case, time points are treated as evenly spaced with consequent introduction of incorrect information in the model. On the other hand, the good performance of GC on unevenly spaced time course data is surprising; in order to understand the exact reason why GC does not suffers significantly further studies are required. This emerged feature of both CTBNs and GC is particularly relevant to the gene network reconstruction problem. Indeed, time course data are rarely collected at regular time intervals while the

most common scenario is to have time measurements more densely sampled during some specific phases of the studied phenomenon and coarsely sampled during other phases.

In accordance with what was shown in [28], DBNs and GC were found to perform similarly. In particular, it was not possible to determine if one of these methods was definitively better than the other: for simulated data, GC performed better than DBNs on *E. coli* (Figure 2.3A) while on *S. cerevisiae* DBNs performed better than GC (Figure 2.3B). However, when tested on coarsely grained time course data DBNs showed a net loss of performances on both *E. coli* and *S. cerevisiae*, remaining way below the level of accuracy achieved by GC. This result is in contrast with [28] where the authors showed that when the length of the time course is smaller than a given threshold, DBNs outperform GC while vice-versa when the length of the time course is greater than the threshold. However, their test was performed on a 5 genes network and the authors themselves stated that the results of the test could have changed on networks of bigger dimension.

The simulated time course dataset that we used for the analysis is at present unrealistically rich in terms of number of perturbations and replicates. However, continuous improvement in experimental technologies will soon allow researchers to reach this level in the near future. When tested on a real experimental dataset of limited

dimension and with no interventional data available, CTBNs still achieved the best performance. This result suggest that CTBNs can perform well also on dataset of small dimension and that they could be suitable for the reconstruction of other types of biological networks as well, such as signaling cascades, where direct manipulation and measurement of the individual members of the cascade are difficult.

2.5.2 Biological insights emerged from application of CTBNs to Th17 cell differentiation

As follows we discuss some well-known regulatory mechanisms emerged both in the IL6+TGFβ1 and IL23 networks together with the new biological insights which can be derived from them. For specific direct interactions which are said to be known in the literature, the corresponding reference is omitted in the text but included in Table 2.2.

Negative regulator Il4ra is suppressed upon IL23 addition

As described previously, IL4RA, which mediates a negative role on the Th17 differentiation process, loses its role as a hub node upon IL23 perturbation (Figure 2.8A). Thus the negative role exerted by *Il4* on Th17 differentiation is suppressed. On the other hand *Bcl3*

and *Spp1* are seen to target *Il4ra* in the IL23 network. Since *Bcl3* and *Spp1* are known to regulate both activation and proliferation of T cells and Th17 differentiation the interaction between *Bcl3*, *Spp1* and IL23 as suggested by the model is highly plausible.

IL23 activates an autocrine loop involving Nfil3

Nfil3 is a basic leucin zipper transcription factor, known to regulate NK cell differentiation processes and development of NK progenitors [72]. Recently, it has been found that *Nfil3* is required to control the Th17 phenotype by binding the *Rorc* promoter gene and repressing its expression [73]. *Nfil3* is regulated by the circadian clock, which determines the Th17 ability to release Il17a. The interruption of the normal circadian clock reduces *Nfil3* expression leading to a disregulated Th17 with higher *Il17a* expression and occurrence of various inflammatory diseases [73]. The perturbation with IL23 leads interestingly to a change in the *Nfil3* gene interactions: in the IL6+TGFβ1 network *Nfil3* appear regulated by *Prickle1* (Figure 2.8B), whose function is still unknown for Th17 differentiation. In the IL23 network, *Nfil3* is regulated by *Il17*. This further underlines the importance of the activation, by IL23 cytokine, of an autocrine loop mediated by Il17. This mechanism is currently unknown and in light of this result may be worth a biological validation.

The role of Il10 in Th17 cell differentiation

IL10 is a very well known cytokine, which represents a strong immunoregulator of inflammatory processes. Thus, it is not surprising that in this regulatory network *Il10* represents one of the minor hubs. In particular, the network highlights an interaction/loop already extensively described in literature between *Ccl20* ligand and *Il10* (Figure 2.8C). *Il10* is know to be highly expressed in Th17 cells, furthermore the interaction with *Batf* is known as well. A correlation between levels of *Il10* and *Il7r* is also described in T cells. Interestingly, IL23 perturbation here shows that IL23 eliminates this last interaction favoring a new one between *Il10* and lipopolysaccharide-induced TNF-alpha factor (*Litaf*), a DNA-binding protein that mediates the TNF-alpha expression binding to the promoter of the TNF-alpha gene. *Litaf* may be then important to delineate the Th17 pathogenic phenotype, which is achieved thanks to the addition of IL23 in the culture and regulated by *Il10* during Th17 differentiation (Figure 2.8C). Furthermore, in the IL23 network the loop between *Ccl20* and *Il10* does not appear anymore, which is worth of investigation to better understand the function of *Ccl20* in Th17 differentiation.

Bcl3 may play a key role in balancing positive and negative markers of Th17 cells

The IL6+TGFβ1 network shows a central role of *Bcl3*. An interesting new interaction between *Bcl3* and *Id3*, a transcription factor involved in T cell development, is suggested (Figure 7D). *Bcl3* is also seen to interact with *Bcl6* and *Il2rb*. All of these genes are known to be negative regulators of Th17 differentiation [74, 75]. In particular, the transcriptional repressor protein *Bcl6* regulates T cell differentiation by repressing Th17 responses and promoting follicular Th cell responses [74]. Interestingly, *Bcl3*, which also interacts with *Il4ra* upon IL23 addition, interacts in normal conditions (IL6+TGFβ1 network) also with *Irf4*, *Sgk1*, *Il17ra* and *Id2*, which are all known as being phenotypic markers of Th17 pathogenic cells [76]. This suggests a crucial role of *Bcl3* in Th17 differentiation since it appears to be able to interact and probably affect the balance between positive and negative markers of Th17 cells (Figure 2.8D). Also, *Bcl3* is revealed by the network as a very important regulator of the final Th17 program. *Bcl3* indeed regulates a chain in the network upon IL23 addition (Figure 2.8F). The interaction between *Il21* and *Rorc* is extensively known as well as the interaction between *Irf4* and *Il21*. The whole chain seems then to be regulated by *Bcl3*, which as shown before (Figure 2.8D) is able to

regulate other Th17 differetiation markers. Finally, *Rorc* is correctly placed at the end of the chain as it represents a marker of final differentiated Th17 cells.

Prdm1 and Tnfsf11 regulation loop may play a key role in balancing Th17 pathogenic and non pathogenic cells

The IL6+TGFβ1 network highlights a known interaction between *Tnfsf11* alias *Rankl* and *Prdm1*, alias *Blimp1* (B lymphocyte–induced maturation protein-1) (Figure 2.8E). *Tnfsf11* is known to be a marker of pathogenic Th17 cells in inflammation whereas *Prdm1* binds to the *Il17a* gene and acts as repressor of *Il17a* expression [77]. The network highlights a loop between *Tnfsf11* and *Prdm1* genes, suggesting an inter-regulation between the two. Interestingly, this interaction is known in other cell types but not in Th17. The negative feedback loop between the inhibitory transcription factor *Prdm1* and *Tnfsf11* may indicate a balancing mechanism between pathogenic and non pathogenic Th17 cells with *Prdm1* acting as a negative regulator of pathogenic Th17 cells characterized by high expression of *Tnfsf11*.Furthermore, the regulatory chain between *Il17ra*, *Prdm1* and *Tnfsf11* suggests a negative regulation of *Prdm1* on *Tnfsf11* in response to *Il17a*. This is significant considering that *Il4ra* is also hub, which highlights the importance of cytokine autocrine loops in Th17 differentiation. In other words, this shows

that as in many others systems, Th17 cells autoregulate their differentiation. Finally, according to the prediction, *Tnfsf11* might represents a master regulator of phenotipic markers of Th17 differentiated phenotype since the network underlines its regulation on *Tbx21*, *Ahr*, *Fas*, and *Sgk1*. This last consideration is worth of further investigations since the regulator of finally differentiated pathogenic Th17 cells is not known.

Il17a directly regulates Salt-sensing kinase Sgk1

One of the genes controlled by *Tnfsf11* is the salt-sensing kinase *Sgk1* (Figure 2.8E), which has recently been described as a marker of pathogenic Th17 cells [52]. It has been shown recently that environmental factors promote and stabilize Th17 cells and affect their pathogenic role in autoimmune diseases. Sodium chloride has recently been found to drive experimental autoimmune encephalomyelitis (EAE) disease by the induction of pathogenic Th17, thus linking sodium salt intake as an environmental factor influencing the development of autoimmune diseases. In the model proposed in [52], *Sgk1* has been found to be an essential node downstream *Il23* signaling in Th17 differentiation and stabilization. Our network reconstruction confirms the relevance of *Sgk1* node as it is

controlled exclusively and directly by three main hubs (*Bcl3*, *Tnfsf11*, *Prickle1*) and *Il17a* in the IL6+TGFβ1 as well as IL23 network. This represents new information that *Sgk1* in reality would be independent of *Il23* signaling but dependent on *Il17* itself (Figure 2.8E). Interestingly, the regulation of *Sgk1* also occurs through the receptor of *Il17* (*Il17ra*), through the regulatory chain involving *Prdm1* and *Tnfsf11*, strengthening the theory that *Sgk1* depends on *Il17* and suggesting once again the existence of an autocrine loop in the regulation of *Sgk1*.

2.6 Conclusions and Future Works

The encouraging results achieved in this investigation suggest that structural learning of CTBNs should be considered as a new reliable gene network reconstruction method when time course expression data is available; results indicate that CTBNs would be particularly suitable for the learning of large networks and when the time measurements are not collected at evenly spaced time points.

CTBNs assume the duration of the events to be a random variable exponentially distributed. The exponential distribution has the characteristics of being *"memoryless"*. CTBNs can be extended to the modeling of systems with memory by introducing hidden

nodes/states and representing the system through a mixture of exponential distributions. The application of this extension to the gene networks domain is relevant and remains to be explored. Another key aspect to be investigated is the inference task, which would allow for a deeper analysis of the dynamic aspect of the reconstructed gene network, such as answering queries directly involving the time.

CTBNs helped elucidate the regulatory network responsible for murine Th17 differentiation, confirming well-known regulatory interactions and main regulators, as well as formulating new biological hyphothesis. Apart for a number of new potential regulators, the network inferred by CTBNs highlighted the presence of several autocrine loops through which Th17 could be autoregulating their own differentiation process. The relevance of this insight comes from the fact that, while self-regulating mechanisms are known to exist in other cell lines, their existence in Th17 has not emerged yet. Wet-lab experiments aimed at validating this hypothesis are now required.

2.7 Details of numerical experiments

2.7.1 Simulated data generation

The simulated dataset was generated with the help of Gene Net Weaver tool [50, 78] which has previously been used to generate datasets for network inference challenges of the international Dialogue for Reverse Engineering Assessments and Methods (DREAM) competition [48]. The tool allows extraction of subnetworks from known *in vivo* gene regulatory network structures of *E. coli* [49] and *S. cerevisiae* [79] endowing them with dynamic models of regulation. When extracting the 10-NETs and 20-NETs, no constraint on the minimum number of regulators (i.e. nodes that have at least one outgoing link in the full network) to include was specified, while for the 50-NETs and 100-NETs this parameter was set to 10 and 20 respectively. This choice on 50-NETs and 100-NETs was made to avoid the generation of networks characterized by a large number of leaf nodes and thus with a too simple structure. No constraint was set on the maximum number of parents allowed per node.

Given each extracted network structure, Gene Net Weaver combines ordinary and stochastic differential equations to generate the corresponding dataset. Perturbations are applied to the first half of the time series and removed from the second part, showing how

the system reacts and then goes back to the wild type state. The multiplicative constant of the white noise term in the stochastic differential equations was set to 0.05 as in DREAM4. Finally, all expression values were normalized by dividing them by the maximum mRNA concentration of the related dataset.

2.7.2 Parameter optimization and data discretization for simulated data

Prior to run the tests on simulated data, an empirical *optimization* of the model parameters for the three methods was run; for CTBNs and DBNs this included experimentally establishing the optimum number of discretization levels. Here all the steps aimed to individuate the best configurations for the three methods are described. It is important to notice that with the term *optimization* we do not refer to the optimization of an objective function, but to a set of independent numerical experiments where the structural learning is run for different values of the model's parameters. The *optimal* parameters are considered the ones for which the algorithms achieve the highest values of the F_1 measure.

For CTBNs, *optimization* experiments were run on the 10-NETs and 20-NETs, where the required learning time was still feasible. The optimal parameter values found were subsequently applied to

the 50-NETs and 100-NETs. Because the Bayesian models cannot handle continuous data, a discretization was applied. Discretization of continuous data is known to be a critical task: too few bins (levels) of discretization lead to a loss of important information, while when increasing the number of bins it is known that the required amount of data and computational resources increases as well. To find the optimal number of bins, tests with data discretized into 3, 4, 5, 6 and 7 equal width bins were performed. Best performances were obtained when using 5 equal width bins. It is worthwhile to notice that discretization intervals were chosen individually for each variable (gene) based on the max and min value of expression levels of each variable among the whole set of data generated. In order to preserve the significance and comparability of the results, one needs to keep track of the discretization intervals applied to each variable. An analysis on the importance of the discretization strategy can be found in [5]. Regarding the hyperparameters α and τ, introduced in section Methods, best values were found to be 0.01 and 5 respectively. Because of the local nature of the learning, the optimal hyperparameters values found on 10-NETs and 20-NETs are expected to be optimal for 50-NETs and 100-NETs as well. Indeed, separate *optimization* process on 10-NETs and 20-NETs returned the same optimal values. The computational nature of the exact structural learning problem lent itself to greedy learning.

However, preliminary tests on the 10-NETs returned the same results for both exhaustive and greedy learning, although it cannot be established whether the exhaustive learning on the larger networks would have returned better results. The last parameter investigated was the maximum number of parents allowed for each node: since the greater this value is, the longer is the computational time required, sequential tests with an increasing value of this parameter were run. Interestingly, it was observed that CTBNs were never able to detect more than 3 parents per node even when the true networks contain nodes with a number of parents greater than 3.

For DBNs parameter *optimization* on the number of discretization bins was re-run and results confirmed that what is optimum for CTBNs may not be the best option for learning with DBNs. Indeed, results indicated 3 as optimum number of discretization bins for DBNs. Discretization intervals were selected individually for each variable as it was done for CTBNs. Model selection has been performed by using the BIC criterion [37], which reduces the chance of overfitting. Analogously to what observed with CTBNs, DBNs were never able to detect more than 3 parents per node. Experiments with 50-NETs and 100-NETs are not shown because the problem became intractable.

For GC analysis no discretization was required since the approach

can handle continuous data. Best value for the model order parameter, i.e. the number of past observations to incorporate into the regression model, was discovered to be equal to 1. Covariance stationary (CS) is a requirement for the GC to be applied. Data resulted to be CS according to the ADF criterium [80], but not according to KPSS [81]. However when differencing was applied to correct this condition, data interpretation may have become more complicate and in fact performances resulted to be significantly worse; as a consequence no differencing has been applied. Pre-processing steps of detrending and demeaning have been applied as well. Analysis was based on the conditional GC test. After performing the GC analysis and obtaining the matrix of magnitudes of GC interactions, the statistically significant set of interactions has been individuated. The best results were observed with a significance cut-off of 0.01 and a Bonferroni multiple test correction.

Parameter *optimization* was run also with respect to the synthetically reconstructed yeast dataset. Optimal number of bins resulted to be 3 for DBNs and CTBNs, while the maximum number of parents was set to 5. Optimal prior values for CTBNs were equal to those on simulated data. Learning criteria for DBNs was set to BIC. For GC all the pre-processing steps listed for the simulated data were applied, finding a p-value cutoff of 0.05 with an approximation of the False Discovery Rate (FDR) correction being the best performing one.

2.7.3 Bioinformatic analysis and data pre processing for murine Th17 data

The microarray raw data for the 275 genes indicated by [52] were analyzed using the Bioconductor package for Affymetrix platform, with annotation chip mouse430a2. Quantile normalization and log2 conversion were performed using RMA. Fold-change values were obtained separately for different biological replicates, assuming the fold-change values being equal to 0 at time point 0. Data was corrected to have mean 0 and standard deviation 1. Supposing X to be the fold-change values, noise and random fluctuations in the data resulted to be heavy for $X < 1.2$ and $X >$ -1.2; as a consequence, X was discretized into 3 different levels: $X < -1.2$, $-1.2 \leq X \leq 1.2$, $X > 1.2$. Genes whose fold-changes levels after discretization resulted to be constant among all the time points were excluded from the analysis.

Software and tools

Experiments were run using: for CTBNs the CTBN Matlab Toolbox developed at the MAD (Models and Algorithms for Data and text mining) Lab of the University of Milano-Bicocca, for DBNs the

Bayesian Net toolbox of Murphy [82] version 1.07, for GC the toolbox for Granger causal connectivity analysis (GCCA) [83] version v2.9.

Author's contributions

EA and FS conceived the study, EA realized all the numerical experiments under the supervision of FS and with relevant suggestions from VN, EA and TZ performed the biological interpretation of the Th17 dataset, EA wrote the manuscript with contributions from TZ, VN and FS.

Acknowledgements

The authors would like to thank Michael Poidinger, Federico M. Stefanini, Alessandro la Torraca, Francesca Zolezzi, Francesca Cordero, Andrea Canazza for their valuable help and Yeser Amer for making available the CTBN Matlab Toolbox.

2.8 References

[1] Chao Sima, Jianping Hua, Sungwon Jung, et al. Inference of gene regulatory networks using time-series data: a survey.

Current genomics, 10(6):416, 2009.

[2] Feng He, Rudi Balling, and An-Ping Zeng. Reverse engineering and verification of gene networks: principles, assumptions, and limitations of present methods and future perspectives. *Journal of biotechnology*, 144(3):190–203, 2009.

[3] Martin G Grigorov. Analysis of time course omics datasets. In *Bioinformatics for Omics Data*, pages 153–172. Springer, 2011.

[4] Harri Lähdesmäki et al. On learning gene regulatory networks under the boolean network model. *Machine Learning*, 52 (1-2):147–167, 2003.

[5] Nir Friedman et al. Using bayesian networks to analyze expression data. *Journal of computational biology*, 7(3-4): 601–620, 2000.

[6] Jun Zhu, Matthew C Wiener, Chunsheng Zhang, Arthur Fridman, Eric Minch, Pek Y Lum, Jeffrey R Sachs, and Eric E Schadt. Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations. *PLoS computational biology*, 3(4):e69, 2007.

[7] Karen Sachs et al. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721): 523–529, 2005.

[8] T Dean and K Kanazawa. A model for reasoning about persistence and causation. *Computational intelligence*, 5(2): 142–150, 1989.

[9] Osamu Hirose et al. Statistical inference of transcriptional module-based gene networks from time course gene expression profiles by using state space models. *Bioinformatics*, 24(7):932–942, 2008.

[10] Ilya Shmulevich et al. Probabilistic boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, 18(2):261–274, 2002.

[11] Peng Li et al. Comparison of probabilistic boolean network and dynamic bayesian network approaches for inferring gene regulatory networks. *BMC bioinformatics*, 8(Suppl 7):S13, 2007.

[12] Hidde De Jong. Modeling and simulation of genetic regulatory systems: a literature review. *Journal of computational biology*, 9(1):67–103, 2002.

[13] Mukesh Bansal, Giusy Della Gatta, and Diego Di Bernardo. Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinformatics*, 22(7):815–822, 2006.

[14] Adam Margolin et al. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics*, 7(Suppl 1):S7, 2006.

[15] Katia Basso et al. Reverse engineering of regulatory networks in human b cells. *Nature genetics*, 37(4):382–390, 2005.

[16] Alina Sîrbu et al. Comparison of evolutionary algorithms in gene regulatory network model inference. *BMC bioinformatics*, 11(1):59, 2010.

[17] Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

[18] Mingzhou Ding et al. Granger causality: Basic theory and application to neuroscience. *Handbook of time series analysis*, page 437, 2006.

[19] Enzo Acerbi et al. Computational reconstruction of biochemical networks. In *Information Fusion (FUSION), 2012 15th International Conference on*, pages 1134–1141. IEEE, 2012.

[20] Guy Karlebach et al. Modelling and analysis of gene regulatory networks. *Nature Reviews Molecular Cell Biology*, 9(10): 770–780, 2008.

[21] Adam M Feist et al. Reconstruction of biochemical networks in microorganisms. *Nature Reviews Microbiology*, 7(2):129–143, 2008.

[22] Mukesh Bansal et al. How to infer gene networks from expression profiles. *Molecular systems biology*, 3(1), 2007.

[23] Daniel Marbach, James C Costello, Robert Küffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison, Manolis Kellis, James J Collins, Gustavo Stolovitzky, et al. Wisdom of crowds for robust gene network inference. *Nature methods*, 2012.

[24] Keunkwan Ryu. Analysis of a continuous-time proportional hazard model using discrete duration data. *Econometric reviews*, 14(3):299–313, 1995.

[25] U Nodelman et al. Continuous time bayesian networks. In *Proc. of the 18th Conf. on Uncertainty in Artificial Intelligence*, pages 378–387. Morgan Kaufmann Publishers Inc., 2002.

[26] Denis Thieffry, Araceli M Huerta, Ernesto Pérez-Rueda, and Julio Collado-Vides. From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in escherichia coli. *Bioessays*, 20(5):433–440, 1998.

[27] Nir Friedman. Inferring cellular networks using probabilistic graphical models. *Science*, 303(5659):799–805, 2004.

[28] Cunlu Zou et al. Granger causality vs. dynamic bayesian network inference: a comparative study. *BMC bioinformatics*, 10(1):122, 2009.

[29] Irene Cantone et al. A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches. *Cell*, 137(1):172–181, 2009.

[30] J Xu et al. Continuous time bayesian networks for host level network intrusion detection. *Machine Learning and Knowledge Discovery in Databases*, pages 613–627, 2008.

[31] H Boudali et al. A continuous-time bayesian network reliability modeling, and analysis framework. *Reliability, IEEE Trans. on*, 55(1):86–97, 2006.

[32] Y Fan et al. Learning continuous-time social network dynamics. In *Proc. of the 25th Conf. on Uncertainty in Artificial Intelligence*, pages 161–168. AUAI Press, 2009.

[33] E Gatti et al. A continuous time bayesian network model for cardiogenic heart failure. *Flexible Services and Manufacturing Journal*, pages 1–20, 2011.

[34] U Nodelman et al. Learning continuous time bayesian networks. In *Proc. of the 19th Conf. on Uncertainty in Artificial Intelligence*, pages 451–458. Morgan Kaufmann Publishers Inc., 2002.

[35] Thomas Dyhre Nielsen and FINN VERNER JENSEN. *Bayesian networks and decision graphs*. Springer, 2009.

[36] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[37] Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[38] Timo JT Koski and John Noble. A review of bayesian networks and structure learning. *Mathematica Applicanda*, 40(1):51–103, 2012.

[39] John F Geweke. Measures of conditional linear dependence and feedback between time series. *Journal of the American Statistical Association*, 79(388):907–915, 1984.

[40] Simon Kwok. A nonparametric test of granger causality in continuous time. 2012.

[41] J Roderick McCrorie and Marcus J Chambers. Granger causality and the sampling of economic processes. *Journal of econometrics*, 132(2):311–336, 2006.

[42] Craig Hiemstra and Jonathan D Jones. Testing for linear and nonlinear granger causality in the stock price-volume relation. *The Journal of Finance*, 49(5):1639–1664, 1994.

[43] Cees Diks and Valentyn Panchenko. A new statistic and practical guidelines for nonparametric granger causality testing. *Journal of Economic Dynamics and Control*, 30(9):1647–1669, 2006.

[44] Nicola Ancona, Daniele Marinazzo, and Sebastiano Stramaglia. Radial basis function approach to nonlinear granger causality of time series. *Physical Review E*, 70(5):056221, 2004.

[45] Daniele Marinazzo, Wei Liao, Huafu Chen, and Sebastiano Stramaglia. Nonlinear connectivity by granger causality. *Neuroimage*, 58(2):330–338, 2011.

[46] David L Roberts and Stephen Nord. Causality tests and functional form sensitivity. *Applied Economics*, 17(1):135–141, 1985.

[47] Lionel Barnett, Adam B Barrett, and Anil K Seth. Granger causality and transfer entropy are equivalent for gaussian variables. *Physical review letters*, 103(23):238701, 2009.

[48] Gustavo Stolovitzky et al. Dialogue on reverse-engineering assessment and methods. *Annals of the New York Academy of Sciences*, 1115(1):1–22, 2007.

[49] Socorro Gama-Castro, Verónica Jiménez-Jacinto, Martín Peralta-Gil, Alberto Santos-Zavaleta, Mónica I Peñaloza-Spinola, Bruno Contreras-Moreira, Juan Segura-Salazar, Luis Muñiz-Rascado, Irma Martínez-Flores, Heladia Salgado, et al. Regulondb (version 6.0): gene regulation model

of escherichia coli k-12 beyond transcription, active (experimental) annotated promoters and textpresso navigation. *Nucleic acids research*, 36(suppl 1):D120–D124, 2008.

[50] Daniel Marbach, Thomas Schaffter, Claudio Mattiussi, and Dario Floreano. Generating realistic in silico gene networks for performance assessment of reverse engineering methods. *Journal of Computational Biology*, 16(2):229–239, 2009.

[51] Dan R Littman and Alexander Y Rudensky. Th17 and regulatory t cells in mediating and restraining inflammation. *Cell*, 140(6):845–858, 2010.

[52] Nir Yosef, Alex K Shalek, Jellert T Gaublomme, Hulin Jin, Youjin Lee, Amit Awasthi, Chuan Wu, Katarzyna Karwacz, Sheng Xiao, Marsela Jorgolli, et al. Dynamic regulatory network controlling th17 cell differentiation. *Nature*, 2013.

[53] Chuan Wu, Nir Yosef, Theresa Thalhamer, Chen Zhu, Sheng Xiao, Yasuhiro Kishi, Aviv Regev, and Vijay K Kuchroo. Induction of pathogenic th17 cells by inducible salt-sensing kinase sgk1. *Nature*, 2013.

[54] Markus Kleinewietfeld, Arndt Manzel, Jens Titze, Heda Kvakan, Nir Yosef, Ralf A Linker, Dominik N Muller, and David A Hafler. Sodium chloride drives autoimmune disease by the induction of pathogenic th17 cells. *Nature*, 2013.

[55] Maria Ciofani, Aviv Madar, Carolina Galan, MacLean Sellars, Kieran Mace, Florencia Pauli, Ashish Agarwal, Wendy Huang, Christopher N Parkurst, Michael Muratet, et al. A validated regulatory network for th17 cell specification. *Cell*, 151(2):289–303, 2012.

[56] Keitaro Hayashi and Amnon Altman. Filamin a is required for t cell activation mediated by protein kinase c-θ. *The Journal of Immunology*, 177(3):1721–1728, 2006.

[57] Baosheng Ge, Olga Li, Phillip Wilder, Angie Rizzino, and Timothy W McKeithan. Nf-κb regulates bcl3 transcription in t lymphocytes through an intronic enhancer. *The Journal of Immunology*, 171(8):4210–4218, 2003.

[58] George Michailidis and Florence d'Alché Buc. Autoregressive models for gene regulatory network inference: Sparsity, stability and causality issues. *Mathematical biosciences*, 246 (2):326–334, 2013.

[59] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[60] Susan K Wallis, Laura A Cooney, Judith L Endres, Min Jie Lee, Jennifer Ryu, Emily C Somers, David A Fox, et al. A polymorphism in the interleukin-4 receptor affects the ability of interleukin-4 to regulate th17 cells: a possible immunoregulatory mechanism for genetic control of the severity of rheumatoid arthritis. *Arthritis Res Ther*, 13(1):R15, 2011.

[61] Jan Piet van Hamburg, Marjolein JW De Bruijn, Claudia Ribeiro de Almeida, Marloes van Zwam, Marjan van Meurs, Edwin de Haas, Louis Boon, Janneke N Samsom, and Rudi W Hendriks. Enforced expression of gata3 allows differentiation of il-17-producing cells, but constrains th17-mediated pathology. *European journal of immunology*, 38 (9):2573–2586, 2008.

[62] Shashi Bala, Alexander Tang, Donna Catalano, Jan Petrasek, Odette Taha, Karen Kodys, and Gyongyi Szabo. Induction of bcl-3 by acute binge alcohol results in toll-like receptor 4/lps tolerance. *Journal of leukocyte biology*, 92(3):611–620, 2012.

[63] Qingguo Ruan, Shi-Jun Zheng, Scott Palmer, Ruaidhri J Carmody, and Youhai H Chen. Roles of bcl-3 in the pathogenesis of murine type 1 diabetes. *Diabetes*, 59(10):2549–2557, 2010.

[64] Noriko Komatsu, Kazuo Okamoto, Shinichiro Sawa, Tomoki Nakashima, Masatsugu Oh-hora, Tatsuhiko Kodama, Sakae Tanaka, Jeffrey A Bluestone, and Hiroshi Takayanagi. Pathogenic conversion of foxp3+ t cells into th17 cells in autoimmune arthritis. *Nature medicine*, 20(1): 62–68, 2014.

[65] Christina Stoeckle, Cécile Gouttefangeas, Michael Hammer, Ekkehard Weber, Arthur Melms, and Eva Tolosa. Cathepsin w expressed exclusively in cd8¡ sup¿+¡/sup¿ t cells and nk cells, is secreted during target cell killing but is not essential for cytotoxicity in human ctls. *Experimental hematology*, 37(2):266–275, 2009.

[66] Zheng-Xu Wang, Christina Hui-Leng Teh, Caroline Man-Yee Chan, Ci Chu, Michael Rossbach, Galih Kunarso, Tahira Bee Allapitchay, Kee Yew Wong, and Lawrence W

Stanton. The transcription factor zfp281 controls embryonic stem cell pluripotency by direct activation and repression of target genes. *Stem Cells*, 26(11):2791–2799, 2008.

[67] Maria H Lexberg, Annegret Taubner, Inka Albrecht, Inga Lepenies, Anne Richter, Thomas Kamradt, Andreas Radbruch, and Hyun-Dong Chang. Ifn-γ and il-12 synergize to convert in vivo generated th17 into th1/th17 cells. *European journal of immunology*, 40(11):3017–3027, 2010.

[68] David Bending, Stephen Newland, Alena Krejčí, Jenny M Phillips, Sarah Bray, and Anne Cooke. Epigenetic changes at il12rb2 and tbx21 in relation to plasticity behavior of th17 cells. *The Journal of Immunology*, 186(6):3373–3382, 2011.

[69] Myew-Ling Toh, Masanori Kawashima, Arnaud Hot, Philippe Miossec, and Pierre Miossec. Role of il-17 in the th1 systemic defects in rheumatoid arthritis through selective il-12rβ2 inhibition. *Annals of the rheumatic diseases*, 69(8): 1562–1567, 2010.

[70] Shane E Russell, Anne C Moore, Padraic G Fallon, and Patrick T Walsh. Soluble il-2rα (scd25) exacerbates autoimmunity and enhances the development of th17 responses in mice. *PloS one*, 7(10):e47748, 2012.

[71] Ming Shan, Xiaoyi Yuan, Li-zhen Song, Luz Roberts, Nazanin Zarinkamar, Alexander Seryshev, Yiqun Zhang, Susan Hilsenbeck, Seon-Hee Chang, Chen Dong, et al. Cigarette smoke induction of osteopontin (spp1) mediates th17 inflammation in human and experimental emphysema. *Science translational medicine*, 4(117):117ra9–117ra9, 2012.

[72] Victoria Male, Ilaria Nisoli, Tomasz Kostrzewski, David SJ Allan, James R Carlyle, Graham M Lord, Andreas Wack, and Hugh JM Brady. The transcription factor e4bp4/nfil3 controls commitment to the nk lineage and directly regulates eomes and id2 expression. *The Journal of experimental medicine*, 211(4):635–642, 2014.

[73] Xiaofei Yu, Darcy Rollins, Kelly A Ruhn, Jeremy J Stubblefield, Carla B Green, Masaki Kashiwada, Paul B Rothman, Joseph S Takahashi, and Lora V Hooper. Th17 cell differentiation is regulated by the circadian clock. *science*, 342 (6159):727–730, 2013.

[74] Arpita Mondal, Deepali Sawant, and Alexander L Dent. Transcriptional repressor bcl6 controls th17 responses by controlling gene expression in both t cells and macrophages. *The Journal of Immunology*, 184(8):4123–4132, 2010.

[75] Takashi Maruyama, Jun Li, Jose P Vaque, Joanne E Konkel, Weifeng Wang, Baojun Zhang, Pin Zhang, Brian F Zamarron, Dongyang Yu, Yuntao Wu, et al. Control of the differentiation of regulatory t cells and th17 cells by the dna-binding inhibitor id3. *Nature immunology*, 12(1):86–95, 2011.

[76] Yen-Yu Lin, Mary E Jones-Mason, Makoto Inoue, Anna Lasorella, Antonio Iavarone, Qi-Jing Li, Mari L Shinohara, and Yuan Zhuang. Transcriptional regulator id2 is required for the cd4 t cell immune response in the development of experimental autoimmune encephalomyelitis. *The Journal of Immunology*, 189(3):1400–1405, 2012.

[77] Soofia Salehi, Rashmi Bankoti, Luciana Benevides, Jessica Willen, Michael Couse, Joao S Silva, Deepti Dhall, Eric Meffre, Stephan Targan, and Gislâine A Martins. B lymphocyte–induced maturation protein-1 contributes to intestinal mucosa homeostasis by limiting the number of il-17–producing cd4+ t cells. *The Journal of Immunology*, 189(12):5682–5693, 2012.

[78] Thomas Schaffter et al. Genenetweaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27(16):2263–2270, 2011.

[79] S Balaji, M Madan Babu, Lakshminarayan M Iyer, Nicholas M Luscombe, and L Aravind. Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. *Journal of molecular biology*, 360(1):213–227, 2006.

[80] James D Hamilton et al. Autoregressive conditional heteroskedasticity and changes in regime. *Journal of Econometrics*, 64(1):307–333, 1994.

[81] Denis Kwiatkowski et al. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of econometrics*, 54(1):159–178, 1992.

[82] Kevin Murphy et al. The bayes net toolbox for matlab. *Computing science and statistics*, 33(2):1024–1034, 2001.

[83] Anil K Seth et al. A matlab toolbox for granger causal connectivity analysis. *Journal of neuroscience methods*, 186 (2):262, 2010.

Tables

Method	NETs size	Mean precision	Mean recall	**Mean** F_1	F_1 SEM
GC	10	0.46	0.68	**0.54**	6.40E-02
	20	0.40	0.70	**0.49**	4.33E-02
	50	0.24	0.82	**0.37**	3.23E-02
	100	0.16	0.82	**0.27**	2.13E-02
DBNs	10	0.90	0.29	**0.41**	6.90E-02
	20	0.55	0.42	**0.47**	3.66E-02
CTBNs	10	0.66	0.58	**0.61**	5.13E-02
	20	0.72	0.48	**0.57**	2.79E-02
	50	0.53	0.57	**0.54**	1.95E-02
	100	0.45	0.51	**0.48**	2.28E-02
Random	10	0.16	0.55	**0.24**	2.12E-02
	20	0.11	0.51	**0.18**	1.68E-02
	50	0.03	0.49	**0.06**	4.35E-03
	100	0.02	0.50	**0.04**	1.15E-03

Method	NETs size	Mean precision	Mean recall	**Mean** F_1	F_1 SEM
GC	10	0.42	0.75	**0.52**	4.18E-02
	20	0.28	0.81	**0.41**	2.32E-02
	50	0.22	0.78	**0.34**	1.58E-02
	100	0.14	0.80	**0.23**	5.24E-03
DBNs	10	0.62	0.53	**0.56**	3.40E-02
	20	0.60	0.57	**0.58**	4.31E-02
CTBNs	10	0.95	0.58	**0.69**	6.08E-02
	20	0.72	0.70	**0.70**	3.86E-02
	50	0.64	0.56	**0.59**	3.84E-02
	100	0.56	0.51	**0.53**	2.65E-02
Random	10	0.18	0.59	**0.27**	2.10E-02
	20	0.07	0.49	**0.12**	1.27E-02
	50	0.05	0.50	**0.08**	4.88E-03
	100	0.02	0.50	**0.05**	2.63E-03

TABLE 2.1: Performance comparison of CTBNs, DBNs and GC on simulated data for different network dimensions. Organism *E.coli* (top) and *S. cerevisiae* (bottom). Aggregate F_1, *precision* and *recall* values are calculated as the arithmetic mean over the sets of 10 sampled network instances, the standard error of the F_1 mean (SEM) is also shown. See also Figure 2.3

Source	Target	PubMed ID
Il17a	Klrd1	21911461
Il17a	Sgk1	23467085
Il17a	Cd24	19830744
Il21	Rorc	19682929
Stat3	Foxo1	22761423
Irf4	Il21	24430438
Il2rb	Runx1	21292764
Fasl	Rora	19119024
Il10	Ccl20	11244051
Il10	Il7r	18401464
Il10	Rbpj	22933629
Il10	Il24	24130510
Il10	Batf	22992523
Il10	Csf2	24222115
Prkca	Il10	9278292
Stat3	Foxo1	22761423
Foxo1	Smox	22761592
Jun	Maf	22001828
Il4ra	Il4	11918534
Il4ra	Cd30l	11918534
Il4ra	Tgfb3	8601720
Il4ra	Gata3	18792410

Hif1a	Il2ra	23183047
Stat5a	Cxcr5	22318729
Stat5a	Irf8	18342552
Tnfsf11	Prdm1	20133620
Ahr	Tnfsf11	18396263
Egr2	Spry1	21826097
Stat4	Tgfbr1	19808254
Il21	Irf1	19617351
Gata3	Nkg7	19805038
Cebpb	Jak3	12794134
Ifng	Cd74	11009094
Tnfsf8	Nampt	11719441
Csf2	Inhba	12456957
Ccl4	Ccr5	11278962
Bcl3	Irf1	16306601
Bcl3	Id2	22580608
Ncf1	Ifng	15557642
Prdm1	Tnfsf11	20133620
Prkca	Csf2	15661932
Tnfsf11	Fas	12171919
Rora	Mt1h	17666523
Cd80	Cd9	9686645
Elk3	Hif1a	20427288

Foxo1	Timp2	18277385
Bcl3	Il2rb	20235165
Bcl3	Il6st	12969979
Casp1	Tgfbr1	10096572
Ifng	Il7r	18250439
Il2rb	Stat3	9192639
Bcl6	Il2rb	19307668
Ccl20	Il10	20720211
Rora	Stat4	12912921 *
Lamp2	Foxo1	16492665 *
Il2rb	Bcl6	19307668 *
Gap43	Jun	22920255 *
Ctla2b	Stat4	15153495 *
Bcl3	Bcl6	23589612 *
Bcl2l11	Jun	11301023 *
Bcl2l11	Lsp1	23446150 *
Cd9	Spp1	24412090 *
Cxcr5	Cxcl10	22349504 *
Ccl4	Irf8	23853600 *
Ccr8	Stat3	20064451 *
Stat4	Tgfbr1	15879087 *
Sult2b1	Jun	18277385 *

TABLE 2.2: List of direct interactions in the IL6+TGFβ1 network for which literature evidence has been found, together with related PubMed IDs.

Method	Time course granularity	Mean precision	Mean recall	**Mean** F_1	F_1 SEM
GC	11	0.43	0.61	**0.50**	2.88E-02
	21	0.40	0.70	**0.49**	3.35E-02
	31	0.35	0.75	**0.47**	3.80E-02
DBNs	11	0.84	0.15	**0.26**	4.33E-02
	21	0.55	0.42	**0.47**	3.66E-02
	31	0.68	0.30	**0.40**	2.79E-02
CTBNs	11	0.70	0.36	**0.47**	2.05E-02
	21	0.72	0.48	**0.57**	5.54E-02
	31	0.59	0.51	**0.54**	3.23E-02

Method	Time course granularity	Mean precision	Mean recall	**Mean** F_1	F_1 SEM
GC	11	0.47	0.76	**0.57**	4.05E-02
	21	0.28	0.81	**0.41**	5.78E-02
	31	0.29	0.80	**0.42**	3.56E-02
DBNs	11	0.76	0.21	**0.32**	2.32E-02
	21	0.60	0.57	**0.58**	4.31E-02
	31	0.63	0.40	**0.48**	3.86E-02
CTBNs	11	0.60	0.53	**0.60**	3.25E-02
	21	0.72	0.70	**0.70**	6.03E-02
	31	0.56	0.67	**0.60**	3.48E-02

TABLE 2.1: Performance comparison of CTBNs, DBNs and GC on simulated data for different time granularities on 20NETs, organism *E.coli* (top) and *S. cerevisiae* (bottom). Aggregate F_1, *precision* and *recall* values are calculated as the arithmetic mean over the sets of 10 sampled network instances, the standard error of the F_1 mean (SEM) is also shown. See also Figure 2.4.

Chapter 3

Summary, Conclusions and Future Perspectives

Understanding the dynamics of gene regulation is of utmost importance in molecular and translational medicine: deciphering the regulatory mechanisms underlying a certain phenomenon/pathology allows us to formulate targeted molecular therapies. Regulatory interactions between genes are represented as a gene network. While wet lab experiments can validate regulatory interactions between a few genes at a time, computational methods can reconstruct the regulatory network for tens or hundreds of genes at the same time from time course gene expression data. However, reconstruction of gene networks is not a trivial task. Gene networks usually have a complex topology including features such as regulation chains,

auto-regulations and feedback loops which are difficult to reconstruct algorithmically. Furthermore, the number of genes involved is usually in the order of hundreds or thousands while expression data samples are typically several orders of magnitude smaller. Despite the complexity of the problem, several new biological discoveries have been made thanks to the many reconstruction algorithms proposed during the last years. However, current state-of-the-art approaches are still far from being considered standard procedure and gene network reconstruction remains an open and hot research topic.

In recent years, the amount of omic data generated from high-throughput techniques has been constantly increasing; in particular, what is interesting is that time course gene expression dataset are becoming richer (longer and more densely sampled) offering an unprecedented opportunity to gain a better understanding of regulatory network's dynamics. Most of the state-of-the-art methodologies for gene network reconstruction were conceived before the advent of high throughput omic technologies and are not always suitable for this new magnitude of time course data.

My main contribution to the field of molecular and translational medicine consist in the proposal of continuous time Bayesian networks (CTBNs) [1] as a new approach for solving gene network reconstruction problems from time course gene expression data. In

a CTBN variables can evolve continuously over time as a function of a continuous time conditional Markov process while the efficient factored state representation derives from the theory of Bayesian networks. Such setting brings many advantages to the description of the temporal aspect of a system, some of them directly relevant to the GRN reconstruction task.

Firstly, the structural learning problem for CTBNs can be solved locally and in polynomial time with respect to the dimension of the dataset once the maximum number of regulators for each gene is set. This feature suits well regulatory networks, which are systems characterized by a large number of variables (genes) and where genes are typically regulated only by a limited number of other genes [2].

The second advantage is that CTBNs can naturally handle variables evolving at different time granularities. Gene networks are characterized by the presence of both regulatory interactions which happen quickly, e.g. within minutes from a given triggering event, as well as interactions which take place at a slower pace, e.g. within hours or days. To reconstruct such regulatory networks, one may want to integrate data coming from experiments measuring genes whose state evolve at different rates. In such a context, CTBNs is naturally able to learn the overall causal network by combining data coming from different time granularities.

The third advantage is that once the network structure and parameters have been inferred, through inference CTBNs can answer queries directly involving the quantification of the temporal aspects such as *"for how long does gene X have to remain up-regulated to have an effect on the regulation on gene Y?"* and in presence of partial evidence such as *"What is the most probable state for gene X at time t given that I observed that gene Y was up-regulated from time t - α to t - β?"*. With their graphical representation of causal relations, CTBNs also provide an intuitive and meaningful level of abstraction of dynamic regulatory process which can help a molecular biologist to gain a better understanding of the studied systems. Finally, CTBNs conserve all of the advantages which are characteristic of probabilistic graphical models and which make them suitable for the analysis of biological networks [3].

I tested the effectiveness of CTBNs for gene network reconstruction through an extensive comparison with two state-of-the-art approaches (dynamic Bayesian networks and Granger causality), both on simulated and experimental data. On simulated data methods's comparison was carried out for networks of increasing dimension, for measurements taken at different time granularity densities and for measurements evenly vs. unevenly spaced over time. Continuous time Bayesian networks outperformed the other methods in terms of the accuracy of regulatory interactions learnt from data

for all network dimensions. Furthermore, their performance degraded smoothly as the dimension of the network increased. Continuous time Bayesian network were significantly better than dynamic Bayesian networks for all time granularities tested and better than Granger causality for dense time series. Both continuous time Bayesian networks and Granger causality performed robustly for unevenly spaced time series, continuous time Bayesian networks and Granger causality did not show a significant loss of performance compared to the evenly spaced case, while the same did not hold true for dynamic Bayesian networks. The comparison on the experimental datasets confirmed the effectiveness of the proposed method. The positive results achieved suggest that structural learning of CTBNs should be considered as a new reliable gene network reconstruction method when time course expression data is available; results indicate that CTBNs would be particularly suitable for the learning of large networks and when the time measurements are not collected at evenly spaced time points.

I also applied CTBN for the elucidation of the regulatory network responsible for murine Th17 differentiation. The inferred network confirmed well-known regulatory interactions and main regulators. Furthermore, new biological hyphothesis could be derived from the model. Apart for a number of new potential regulators, the network inferred by CTBNs highlighted the presence of several autocrine

loops through which Th17 could be autoregulating their own differentiation process. The relevance of this insight comes from the fact that, while self-regulating mechanisms are known to exist in other cell lines, their existence in Th17 has not emerged yet. Wet-lab experiments aimed at validating this hypothesis are now required.

The encouraging results achieved in this study has opened up a plethora of directions which merit further investigation. CTBNs assume that the duration of the events (state change of a gene) is a random variable which is exponentially distributed. The exponential distribution has the characteristics of being *"memoryless"*. CTBNs can be extended to the modeling of systems with memory by introducing hidden nodes/states and representing the system through a mixture of exponential distributions. The application of this extension to the gene networks domain is relevant and remains to be explored. Another key aspect to be investigated is the inference task, which would allow for a deeper analysis of the dynamic aspect of the reconstructed gene network, such as answering queries directly involving the time. In this work, the gene network has been reconstructed from gene expression data only. While this was necessary for the comparison of CTBNs with other state-of-the-art approaches, it is known that expression data alone does not contain all the necessary information for a correct reconstruction (i.e. post-translational modifications or protein degradation also affect regulatory mechanisms but are not accounted for in this work). For

this purpose, adapting CTBNs to learn gene networks from multiple types of data (i.e. expression and proteomics data) would also represent a natural extension of the work presented here. Another important research direction is the elicitation of *a priori* knowledge about the system [4], which can lead to more accurate predictions.

3.1 References

[1] U Nodelman et al. Continuous time bayesian networks. In *Proc. of the 18th Conf. on Uncertainty in Artificial Intelligence*, pages 378–387. Morgan Kaufmann Publishers Inc., 2002.

[2] Denis Thieffry, Araceli M Huerta, Ernesto Pérez-Rueda, and Julio Collado-Vides. From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in escherichia coli. *Bioessays*, 20(5):433–440, 1998.

[3] Nir Friedman. Inferring cellular networks using probabilistic graphical models. *Science*, 303(5659):799–805, 2004.

[4] Federico M Stefanini. Chain graph models to elicit the structure of a bayesian network. *The Scientific World Journal*, 2014, 2014.

Chapter 4

Publications

- Enzo Acerbi, James Decraene, and Alexandre Gouaillard. "*Computational reconstruction of biochemical networks*". Proceedings of the 15th International Conference on Information Fusion. IEEE, 2012.
- Enzo Acerbi, Fabio Stella."*Continuous Time Bayesian Networks for Gene Network Reconstruction: a Comparative Study on Time Course Data.*" Proceedings of the 10th International Symposium on Bioinformatics Research and Applications. 2014.
- Enzo Acerbi, Teresa Zelante, Viping Narang, Fabio Stella. "*Continuous time Bayesian networks for gene network inference: a comparative study and application to Th17 cell differentiation.*" *Submitted* to peer reviewed journal.

Acknowledgements

I am grateful to Professor Fabio Stella, who has been the *de facto* supervisor of this thesis. His dedication and professionalism are an example for me to follow.