TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

Antti Larjo

Computational Methods for Modelling and Analysing Biological Networks

Julkaisu 1289 • Publication 1289

Tampere 2015

Tampereen teknillinen yliopisto. Julkaisu 1289
Tampere University of Technology. Publication 1289

Antti Larjo

Computational Methods for Modelling and Analysing Biological Networks

Thesis for the degree of Doctor of Science in Technology to be presented with due permission for public examination and criticism in Sähkötalo Building, Auditorium S2, at Tampere University of Technology, on the 27th of March 2015, at 12 noon.

Tampereen teknillinen yliopisto - Tampere University of Technology
Tampere 2015

ISBN 978-952-15-3483-6 (printed)
ISBN 978-952-15-3490-4 (PDF)
ISSN 1459-2045

Reviewers
Dr. Tero Aittokallio
Dr. Florence d'Alché-Buc

Opponent
Dr. Jens Lagergren

Abstract

The main theme of this thesis is modelling and analysis of biological networks. Measurement data from biological systems is being produced at such a pace that it is impossible to make use of it without computational models and inference algorithms. The methods and models presented here aim at allowing to extract relevant relationships from the masses of data and formulating complex biological hypotheses that can be studied via simulation.

The problem of learning the structure of a popular method class, Bayesian networks, from measurement data is investigated in this thesis, and an improvement to the standard method is presented that facilitates finding the correct network structure. Furthermore, this thesis studies active learning, where the structure inference algorithm can itself suggest measurements to be made. Active learning is applied to realistic scenarios with measured datasets and an active learning method that can deal with heterogeneous data types is presented.

Another focus of this thesis is on analysing networks whose structure is known. The utility of a standard method for selecting beneficial mutations in metabolic networks is evaluated in the context of engineering the network to produce a desired substance at a higher rate than normally. Metabolic network modelling is also used in conjunction with a simulation of a biochemical network controlling bacterial movement in a state-based and executable framework that can integrate different submodels. This combined model is then used to simulate the behaviour of a population of bacteria.

In summary, this thesis presents improvements on methods for learning network structures, evaluates the utility of an analysis method for identifying suitable mutations for producing a substance of interest, and introduces a state-based modelling framework capable of integrating several submodels.

Preface

This thesis is the result of work done mainly at the Department of Signal Processing, Tampere University of Technology, and also partly when working at Department of Information and Computer Science at Aalto University School of Science, and during a short stay at Microsoft Research, Cambridge, UK. Financial support from Academy of Finland's Centre of Excellence in Molecular Systems Immunology and Physiology Research (SyMMyS), Emil Aaltonen Foundation, TISE graduate school, the Finnish Cultural Foundation, and Tuula and Yrjö Neuvo Foundation is gratefully acknowledged.

I acknowledge gratefully my indebtedness to Prof. Olli Yli-Harja for providing the possibility of working on this subject and supervising this thesis. I also record my deep appreciation to my other thesis supervisor, Prof. Harri Lähdesmäki, for his insightful comments and being a hard-working and inspiring example. The members of the Computational Systems Biology research group are also acknowledged for creating a great, albeit occasionally distracting atmosphere. A special mention for making it more fun to go to work goes to former cellmates Timo Erkkilä, Pekka Ruusuvuori, Jenni Seppälä and Tarmo Äijö. Also, extra credit is due to Antti Ylipää, Matti Nykter and Virpi Kivinen, for numerous lunch and coffee breaks. I am also grateful to Tommi Aho, Ville Santala, Prof. Matti Karp and Dr. Hillel Kugler for collaboration.

I also want to thank the secretaries of the department, and in particular Virve Larmila, without whom the department would undoubtedly not exist.

Above all, I express my sincerest gratitude to my wonderful family and my parents and brother.

List of Abbreviations

ABCP	algorithm for blocking competing pathways
BN	Bayesian network
CW	clockwise
CCW	counter-clockwise
DNA	deoxyribonucleic acid
EM	elementary mode
EP	extreme pathway
FBA	flux balance analysis
GEM	genome-scale model
GRN	gene regulatory network
MFA	metabolic flux analysis
MCMC	Markov chain Monte Carlo
mRNA	mature RNA
nt	nucleotide
ODE	ordinary differential equation
qPCR	quantitative real-time PCR
RNA	ribonucleic acid
SNP	single-nucleotide polymorphism
UML	Unified Modeling Language

Contents

List of Included Publications

This thesis is a compound thesis consisting of the following 5 publications

I **A. Larjo and H. Lähdesmäki**, "Using multi-step proposal distribution for improved MCMC convergence in Bayesian network structure learning," *submitted to EURASIP Journal on Bioinformatics and Systems Biology*, 2014.

II **A. Larjo, H. Lähdesmäki, M. Facciotti, N. Baliga, I. Shmulevich, and O. Yli-Harja**, "Active learning of Bayesian network structure in a realistic setting," in *Fifth International Workshop on Computational Systems Biology (WCSB 2008)*, Leipzig, Germany, June 11-13, 2008, pp. 85-88.

III **A. Larjo and H. Lähdesmäki**, "Active learning for Bayesian network models of biological networks using structure priors," in *IEEE International Workshop on Genomic Signal Processing and Statistics*, Houston, TX, USA, November 17-19, 2013, pp. 78-81.

IV **J.J. Seppälä*, A. Larjo*, T. Aho, O. Yli-Harja, M.T. Karp, and V. Santala**, "Prospecting hydrogen production of Escherichia coli by metabolic network modeling," *International Journal of Hydrogen Energy*, vol. 38, no. 27, pp. 11780-11789, 2013.

V **H. Kugler*, A. Larjo*, and D. Harel**, "Biocharts: A visual formalism for complex biological systems," *Journal of the Royal Society Interface*, vol. 7, no. 48, pp. 1015–1024, 2010.

* denotes equal contribution.

Author's Contributions to the Publications

The author of this thesis contributed to the included publications as follows:

In Publications I and III the author designed and implemented the methods, derived the mathematical proofs, ran simulations, and wrote the manuscript. In Publication II the author implemented the methods, did all the simulations and wrote the manuscript.

In Publication IV the author performed all metabolic simulations except for the ABCP method, and wrote the corresponding sections of the manuscript. This publication has also been part of the thesis of J. Seppälä, who did the biological work for the manuscript.

In publication V the author designed and implemented the model system and did all the modeling, simulation and analysis work. H. Kugler supervised the study. The manuscript was written by A. Larjo and H. Kugler.

Chapter 1

Introduction

Models have an essential role in may fields of science (and not least in biology (Mogilner et al., 2006)). They aim at representing the system being studied as accurately as possible and represent crystallizations of the knowledge gained by studying the system. Models themselves can be studied and simulated, allowing, e.g., quantitative testing of hypotheses, producing predictions, and interpreting measurement data.

As an example, a scientist may have a hypothesis about a complex system and if the hypothesis can be formalized as a model, then comparing the simulation results to new measurements either lends support to the hypothesis or suggests something needs to be fixed. This validation of hypotheses can become impossible to perform without the help of computational models and simulations that can be run on computers. The reason is the growing complexity of the hypotheses, which is happening in many fields of science thanks to increasing capability to measure more variables at the same time with greater accuracy. This is one of the reasons computational modelling has lately become a central part of research in many fields.

Another task impossible to perform "manually" or "by eye" is finding patterns or structure in big datasets that would allow gaining knowledge for example about which entities function together and what are their causal relationships. By using machine learning and suitable computational models as well as analysis and simulation methods, even huge datasets consisting of heterogeneous data types can be sifted through and meaningful relationships extracted.

Biology has traditionally been a hypothesis-driven science but it can be argued that lately it has become more and more data-driven because of the recent whole-cell or genome-wide measurements, together with machine learning approaches. It has also been argued that, largely due to improved computational and measurement techniques, there has been a paradigm shift form the traditional reductionism towards a holistic approach called systems biology.

In the last decades, network models have become extremely popular (Barabasi,

2002; Barabasi and Oltvai, 2004). This growth in interest is because people are studying larger systems of interacting entities that are naturally modelled as networks, and to a large extent the network models and their development is what has in fact allowed studying such systems. At least one enabling reason is the ability brought with increased computing power to learn and analyse larger models. Networks can be found in almost every aspect of life and they have been extensively studied using methods from physics, control theory, and graph theory, becoming something dubbed “network science" (Lewis, 2011). Plenty of attention has been paid to the topological properties of networks, such as degree distributions, network motifs and modules (Zhu et al., 2007).

Identifying the network structure of an underlying system from measurement data is of great interest in many areas. This process is called inference, structure learning or reverse-engineering of the networks. In this thesis, the inference problem is studied in the field of biological sciences where data suitable for this purpose is nowadays ample. It can actually be argued that the huge progress in generating measurement data has not been completely followed by development of computational analysis and modelling methodologies in biosciences. Yet, there is a pressing need to understand how cellular networks are built and how they function as they govern the cellular activities, and problems in their ability to function properly can cause for example trouble for immune system and elicit diseases such as cancer.

The model class used in this thesis for inference of network structure is Bayesian networks and they are based on two sets of methods that have become very popular lately. Bayesian networks are a model class with roots in Bayesian methods and statistics (Pearl, 1985; Eddy, 2004; Beaumont and Rannala, 2004). Bayesian methods are increasingly popular, some seeing them even as a new paradigm for statistics. Inarguably, they are often able to solve tricky problems, but they do bring about conceptual and even philosophical issues that are somewhat debated (Lindley, 2000).

The other set of methods that has in the last decades become popular is Markov chain Monte Carlo (MCMC), which is a class of estimation methods that rely on heavy sampling and can in practice be only done with computers (Diaconis, 2009). In fact, MCMC methods (and in general Monte Carlo methods) have been able to overcome many computational problems in Bayesian methods, together with increasing computing capability. Consequently, the big increase in interest in Bayesian methods since the 1980s was likely mainly due to discovery of Markov Chain Monte Carlo methods.

For networks whose structure is known, it is of interest to analyse and simulate them, e.g., to examine behaviour that can arise due to different conditions and to predict changes in phenotypes resulting from modifications to the network structure. An example of a model class where such analysis is routinely done is metabolic networks, whose structure can be for some organisms (especially

bacteria) well known. Metabolic networks are also studied in this thesis.

1.1 Objectives of the thesis

The objective of the thesis is to present computational methods for inference of network structures and to simulate biological networks for which the structure is already known.

Although Bayesian networks are a theoretically sound and justified method for modelling gene regulatory networks, as well as protein-protein interaction networks, inferring the Bayesian network structure from experimental data can be challenging. For all but the smallest networks one must resort to Markov chain Monte Carlo methods but a major problem with them is difficulty in convergence. To alleviate this problem, Publication I investigates a new way to propose transitions in the MCMC chain that is shown to increase rate of convergence by escaping local maxima.

Another problem inherent in the inference of network structure is how to make sure that the learnt relationships are causal and not mere non-causal relationships (like correlations). Causality can be separated from correlation by introducing interventions but they can be costly to perform and selecting them in non-optimal way might waste resources. Thus, selecting interventions in the most beneficial way is an important problem and methods called active learning try to perform this. Publication II looks at the performance of one such method in structure inference of Bayesian networks while Publication III studies how to combine more than one data type in active learning.

Metabolic engineering is a field of growing importance as there is an increasing need to modify and design biological organisms to produce beneficial substances and get rid of harmful ones. Traditional methods for this rely on biological experiments and more or less luck. Model-based selection of modifications has the potential to make the process much faster and at least considerably narrow down the choices that need to be tested. Publication IV looks at this problem in the context of trying to find knock-out mutations for increased hydrogen production.

Publication V investigates the integration of more than one model, which is important as the submodels by themselves are not always able to explain the observed behaviour. Another aspect is the utilization of executable models, which is done using a framework where modelled systems are described with states and transitions between them. The lower-level functionality is encoded using a programming language so that the whole model is directly compilable to a computer executable program.

1.2 Outline of the thesis

Chapter 2 describes how modelling of certain biological networks can be done. It also introduces chemotaxis as an example system and briefly describes the stochastic simulator and chemotactic model used in Publication V.

Chapter 3 presents the theory needed in Bayesian network structure learning, including definition of Bayesian networks and Markov chain Monte Carlo methods, which are the necessary basis for Publication I. Active learning in context of Bayesian networks is also shortly presented in Chapter 3 as this is the subject of Publications II and III.

In Chapter 4, methods for analysing metabolic networks under steady-state are presented. These so called constraint-based methods include flux balance analysis, which is applied in Publications IV and V.

Chapter 5 summarizes the results of Publications I-V. Finally, Chapter 6 contains concluding remarks with some possible future directions, and the included publications follow after the list of references.

Chapter 2

Models of biological networks

Development of a biological system and its responses to stimuli depend on a complex interplay between hundreds or thousands of molecules and these actions need to take place in a coordinated and robust manner while ensuring that often subtle signals from environment are taken into account. Due to the interactory nature, biological systems are commonly described as networks where nodes represent the entities and edges the interactions between them. The emergence of high-throughput measurement technologies has enabled identification of network components and their interactions in large-scale and spurred the development of inference and modelling methods.

Although entities of a certain type do not work in isolation, one is usually restricted to concentrate on only some subsystems because of for example limited measurement data and modelling difficulties. Networks of these subsystem types include, e.g., transcription factor binding, protein-protein interaction, protein phosphorylation, metabolic interaction and genetic interaction networks (Zhu et al., 2007). Due to the diverse nature of molecules and interactions in these different subsystems, the models used to describe these systems and methods to analyse and simulate them are often different. This chapter shortly reviews the network types and methods used to model the subsystems that are encountered in the publications that are part of this thesis.

2.1 Gene regulatory networks and signalling networks

The states of genes can influence other genes, creating networks and cascades. The interaction is (always) indirect, mediated via the downstream products of a gene. These can be for example RNA molecules or proteins that can bind the DNA sequence controlling the expression of another gene (or also its own in autoregulation) or for example inhibit the mRNA produced by another gene. Therefore, genes are meta-level entities in gene regulatory network (GRN) modelling and the levels of their expression results (mRNA molecules) are modelled

instead. Transcription factor-binding networks are one type of gene regulatory networks, where mediation of a gene state is done via its produced protein that binds the DNA regions controlling expressions of other genes.

Recent developments of measurement techniques allow huge genomic datasets to be produced. Some of the notable advancements relevant for inference of gene regulatory networks include high-throughput gene expression measurements capable of measuring the expression states of basically all genes at a time, which can be done using, e.g., cDNA microarrays (Schena et al., 1995) or RNA-seq (Mortazavi et al., 2008). Technologies to investigate which genes are being affected by a protein of interest include chromatin immunoprecipitation (ChIP) followed by either microarray measurement (ChIP-chip) (Ren et al., 2000) or DNA sequencing (ChIP-seq) (Johnson et al., 2007), both of which measure the genomic binding sites of a protein.

Modelling GRNs has been under intense research and several different models and methods to infer them from experimental data have been developed and used (Bansal et al., 2007; Karlebach and Shamir, 2008; Noor et al., 2013). Different models include Boolean networks (Kauffman, 1969), probabilistic Boolean networks (Shmulevich et al., 2002), Bayesian networks (Friedman et al., 2000), dynamic Bayesian networks (Ghahramani, 1998; Murphy and Mian, 1999), state-space models (Wu et al., 2004; Quach et al., 2007), rule-based simulations (Meyers and Friedland, 1984), information theoretic methods (Margolin et al., 2006; Zhao et al., 2006), ordinary and partial differential equations (De Jong, 2002), and Gaussian processes (Äijö and Lähdesmäki, 2009).

Cellular information processing and responses to environmental stimuli are often implemented in the cells via signalling networks, which consist of interacting signalling molecules. These are usually proteins and their interactions can cause changes in the states of phosphorylation, conformations, and physical locations of the molecules. Measuring signalling protein expression as well as modification state levels can also be done in a high-throughput fashion using for example multi-color flow cytometry, and many of the methods used for GRN inference and modelling can and have been used also in context of signalling networks (Sachs et al., 2005).

In this thesis (Publications I, II and III) Bayesian networks are used for the purpose of modelling biological networks, even though they are applicable to much wider set of problems than just biological ones. Bayesian networks are dealt with in Chapter 3.

2.2 Metabolic networks

Metabolism of a cell is the totality of processes responsible for converting molecules (called metabolites) into another molecules and producing energy and material for growth and sustenance. Metabolism consists of a set of reactions

that step-by-step break, modify, and construct new metabolites. These reactions are mostly made possible by proteins called enzymes that are produced using information from the genes of an organism. The set of consecutive metabolic reactions and the molecules acting as substrates and products is intuitively modelled as pathways and networks. These network models can in the most simplistic case be reconstructed by identifying the enzyme-coding genes for several organisms owing to the ability to easily and cheaply sequence complete genomes (Covert et al., 2001a).

However, the models produced this way mainly contain information only about the structure. A detailed view of a cellular process like metabolism requires also understanding its dynamics and regulation. The problem is that kinetic and regulatory information are very often unknown as measuring them is much more difficult and costly. Still, considerably accurate models of metabolism have been built based on network structure (Covert et al., 2001a; Feist et al., 2009), since the structure is a prerequisite for kinetic and regulatory models and sets limits to the behaviour of the system. Another aspect is that biological systems often attain a constant or a steady state, at least under certain environmental conditions, and, even without whole-cell dynamic information, biologically meaningful results can still be achieved based on only structural analysis.

These so called constraint-based methods that mostly use only the structure of the metabolic networks are used in Publications IV and V and are discussed in Chapter 4.

2.3 Biochemical reaction networks

Simulation of (bio-)chemical reaction systems can be performed in several different ways (Andrews and Arkin, 2006). Perhaps the most traditional method is to use ordinary differential equations (ODEs). They are a deterministic means of modelling and approximating (bio-)chemical reality by ignoring the discreteness of molecules and assuming the reaction volume to be homogeneous and well-stirred. Extension to take into account spatiality can also be achieved by making the concentrations depend both on time and position, causing the time dependence of the molecules to be governed by partial differential equations.

To make the models and simulations more realistic, the stochastic nature of the system as well as the fact that quantities of molecules are integer values needs to be taken into consideration. One of the popular algorithms is Gillespie algorithm (Gillespie, 1976; Gillespie, 1977). This is not a spatial simulation but instead assumes reaction volume to be small enough so that substances are well-mixed by diffusion. Spatiality can however be allowed for by dividing into small subvolumes and simulating reactions within and between them (Elf and Ehrenberg, 2004).

The biochemical simulation scheme used in Publication V is that imple-

mented in the program StochSim (Morton-Firth and Bray, 1998; Le Novere and Shimizu, 2001), which is a mesoscopic-scale stochastic simulator, whose intracellular simulations are non-spatial, assuming fast enough diffusion of substances within a cell. The basic functioning of StochSim is such that first the length of time-slice is chosen by the most rapid reaction and then, within each time slice, two objects (molecules) are chosen randomly and whether a reaction between them happens is determined by a look-up table that tells the probabilities of reactions happening. The main advantage of the algorithm is its capability of handling multi-state molecules, such as a receptor with different methylation or phosphorylation states that can affect the way it functions, which in Gillespie simulation would require multiple (pseudo-)molecules. StochSim is also able to model changes taking place much faster than chemical reactions, such as ligand binding or conformational changes, by changing such "fast flagged" states according to a probability (that can depend for example on concentrations of other substances) and only after that continue with the selected two species.

Many other simulation systems exist, including Smoldyn (Andrews and Bray, 2004; Andrews, 2012), VCell (Loew and Schaff, 2001; Slepchenko and Loew, 2010), Moleculizer (Lok and Brent, 2005), and AgentCell (Emonet et al., 2005). Of these at least the last one has been used to model *E. coli* chemotaxis that is also the system modelled in Publication V.

2.3.1 Bacterial chemotaxis

The movement of a bacterium to a beneficial direction is made possible by its ability to sense gradients in its environment. There are several different stimuli that can affect the movement, for example light (phototaxis) or temperature (thermotaxis). Movement guided by chemical stimulus is called chemotaxis (Wadhams and Armitage, 2004; Armitage, 1999) and its direction can be either towards a higher concentration of a substance (positive chemotaxis) or away from it (negative chemotaxis). The sensing is done in a temporal fashion during movement since the diameter of most bacteria are likely too small for sensing gradients across their diameter (Adler, 1975), though not necessarily in all cases (Thar and Kühl, 2003). Bacterial chemotaxis is being studied since it plays an important part for example in pathogenicity and formation of biofilms. Bacterial chemotaxis also serves as an important and well-characterized model system.

Movement of many swimming bacteria consist of repeated straight runs followed by rapid changes (called tumbling) to another direction. The frequency of tumbling is controlled by the chemosensory system so that the more favourable the conditions for the bacteria are, the more infrequent the tumbling.

The best-studied case of chemotaxis is the movement of *Escherichia coli* bacteria (Adler, 1966; Berg and Brown, 1972) and is also used as a model system in Publication V. The helical semi-rigid filaments (called flagella) or the bacterium are each rotated by its own motor, in either clockwise (CW) or counter-clockwise

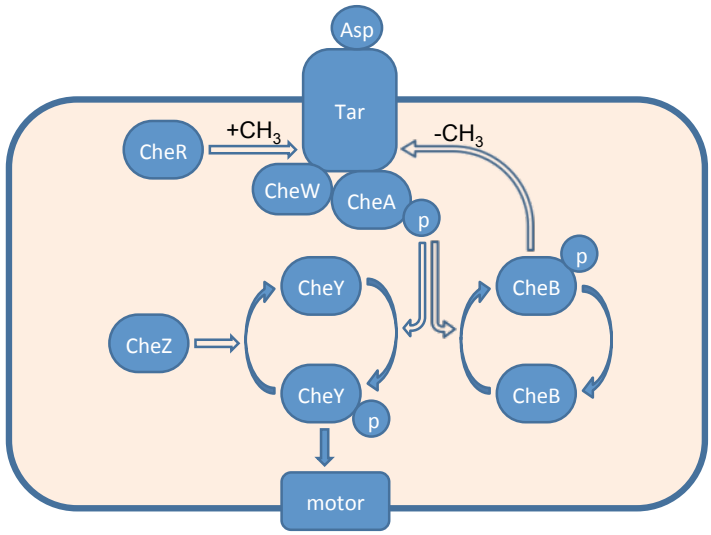


Figure 2.1: Simplified chemotactic network of *E. coli*. The blue border denotes cell wall and shaded area intracellular space. Filled arrows denote reactions and empty arrows regulation. Figure from Publication V.

(CCW) direction. If the flagella all rotate counter-clockwise, they form a bundle and cause forward movement of the cell. However, clockwise rotation of a single or several motors cause tumbling and the result is a change in direction of the cell when all motors return to counter-clockwise rotation.

The structural and biochemical details for the network of chemotactic proteins controlling the direction of rotation of the motors in *E. coli* are thought to be completely known (Wadhams and Armitage, 2004). A simplified picture of this network is shown in Figure 2.1 for the example ligand molecule aspartate (Asp). Explained briefly, the functioning of the network is such that the binding of extracellular Asp to the receptor complex (Tar) decreases the autophosphorylation of protein CheA, which in turn causes reduced amount of CheY-P. The level of CheY-P bound to the motors controls the direction of rotation so that less binding of CheY-P increases counter-clockwise rotation, thus producing longer straight runs. CheZ is a protein increasing spontaneous dephosphorylation of CheY-P and results in more rapid clearance of the CheY-P signal.

An important part in the behaviour of the system is adaptation, which is acquired by the binding of ligand causing reduced amount of CheB-P and consequently heightened methylation level (methyl groups being introduced by CheR) of the Tar complex, which has the effect of increasing CheA autophosphorylation. This feedback allows the system to adapt to varying levels of the ligand. When the ligand dissociates from the receptor, the effect is opposite.

Many *in silico* models have been built on the accumulated knowledge on *E. coli* chemotaxis, with the first one being (Bray et al., 1993), and they have

been observed to capture the main chemotactic behaviour very well. Thus, choosing such a very well developed model (implemented and presented in Morton-Firth et al. (1999)) as a part of our model in Publication V allows us to concentrate on other issues of the modelling and may also hint that if the results do not conform with experiments then the problem is most probably in the other parts of our model. Comprehensive reviews on modelling bacterial chemotaxis include (Tindall et al., 2008b) and (Tindall et al., 2008a).

Chemotactic assays

The most used chemotactic assays include capillary assay tubes (Adler, 1966) and swarm plates. Capillary assay tube is a cylinder filled with a nutritious medium and the other end of the tube is placed into a medium containing a population of bacteria. If the tube contains a nutrient favoured by the bacteria, they start moving into the tube, all the while consuming the nutrient substrate and thus creating a concentration gradient towards the other end of the tube. This gradient moves with the bacteria and the resulting population behaviour is a band of bacteria travelling along the tube. Depending on the medium and bacteria, more than one bands can be formed.

In swarm plates a population of bacteria is positioned to a small area on a plate containing a medium with chemoattractant(s). Population behaviour similar to tube assays is observed as expanding concentric rings, resulting from the same phenomenon of self-created concentration gradients. Also more complex behaviour is observed for many bacteria, such as formation of patterns. These are supposedly created by chemoattractants excreted and sensed by the cells moving themselves.

These phenomena derive from interplay between chemotaxis and metabolism in that the metabolic activity causes gradients of chemoattractants to form and thus creates an excitation of the chemotactic network. On the other hand, chemotactic behaviour drives the organism to environments with varying substrate profiles, thus having an effect on metabolic behaviour. This "communication" between the metabolic and chemotactic networks was one of the aspects modelled in Publication V.

2.4 Executable models

Biochemical pathways presented in many textbooks and publications are rather easy to comprehend but this is thanks to them covering a very limited scope (often in a simplified way). One feature of such models that can be considered a major shortcoming is that the exact meanings of symbols can vary from one presentation to another. When trying to depict larger and more complex systems or processes with numerous interconnections and (auto)regulatory loops, it

quickly becomes impossible to understand how the system behaves without the help of computer simulations.

These facts clearly call for a modelling formalism defining unambiguous meanings for the model entities and interactions, and also allows the system depiction to be read by a computer and simulated. Additional advantage of such a formalism is that the models described using it are exchangeable and thus easily shared between people studying the same system. The formal depiction should also be such that it is easily converted between a view that is comprehensible and easy to modify for humans and a format readable by computers.

Ways to formalize the notation of biological networks have been presented, for example process diagrams (Kitano et al., 2005) and molecular interaction maps (Kohn et al., 2006), which are included (with some modifications) as sublanguages in the community-developed standard visual language called Systems Biology Graphical Notation (SBGN) (Le Novere et al., 2009). These enable sharing of models using for example BioPAX (Demir et al., 2010) but usually lack the information about how to simulate the model, which is often essential in order to understand the dynamic processes.

Systems Biology Markup Language (SBML) (Hucka et al., 2003) and CellML (Lloyd et al., 2004) are also widely used as a means of exchanging and defining models and can incorporate information needed for the simulation of the model, for example as differential equation formulas of reaction rates. In models described using e.g. SBML, the simulation software is not predetermined but the user is free to select any simulator that can read in SBML models. In some sense not being restricted to a single simulator or algorithm is an advantage, but one problem with this scenario is that simulation results are not guaranteed to be exactly equivalent when the same model is simulated with two different simulators. Discrepancies can be for example due to different ODE solvers and/or parameters, all of which may not be explicitly stated in the model description.

A distinction can be made between computational and mathematical models (Fisher and Henzinger, 2007), the latter of which are the kind of "normal" models consisting of, e.g., ODEs and requiring an algorithm for simulating them. Computational models are built from entities (called state machines (Fisher and Henzinger, 2007) that can also be small computer programs) having different states and the state changes are dependent on defined events, such as interactions with other entities. Composition of such entities constitutes a reactive system, the mathematical analysis of which can be impossible. However, the model determines the sequence of steps or instructions that can be executed on a computer and its behaviour observed. Examples of computational models include Boolean networks, Petri nets (Murata, 1989) and interacting state machine models such as statecharts (Harel, 1987).

Thinking and modelling of a biological system by means of computational models may be more natural in some cases, partly due to more explicitly stat-

ing causal relationships (i.e., a state changes to another given a certain event). Further motivation for using a state-based modelling approach is that many biological systems and their subparts can be characterized in separate states. Simple examples are the activity of a gene (on / off) or, in the context of Publication V, the swimming state of bacterium (running / tumbling) and direction of rotation of a motor (CW / CCW).

The approach dubbed Biocharts and illustrated in Publication V is well suited to modelling and simulating systems on different levels of detail as it is a hybrid modelling framework based on object-oriented version of statecharts (Harel, 1987; Harel and Gery, 1996), which allows modelling the high-level behaviour in state-based manner, combined with a well-defined language suitable for describing the lower-level behaviour of the parts of the system, which need not be state-based. Due to using statecharts, the visual formalism at higher-level is closely related to notation used in software engineering (like UML). More formal definitions and some further developments of Biocharts can be found in Kugler (2013).

The Biocharts framework is modular and allows easy reusability and modification of individual parts. This is because each relevant (sub)system (environment, bacterium, motor, etc.) was represented with a class. This also enables the multiplicities of objects (like the number of bacteria) to be changed easily by creating or deleting instances of the classes even during runtime. The internal function of the classes can be divided into states whenever it is sensible. One idea is that it should be possible for a biologist to take part in the modelling effort easily in defining states, transitions between them, and for example also in defining lower-level modelling in particular if it is described in a diagrammatic language.

In Publication V Biocharts were used to model the chemotaxis of *E. coli*, taking into account simultaneously the chemotactic signalling network, metabolic network and environment models to mimic the situation in capillary assay tubes. A descriptive part of the model is shown in Figure 2.2, which shows a part of the statechart modelling the different states of a bacteria and its flagella. Each of these (sub)states can contain a lower-level model (or just code), for example the Growth state under Metabolism triggers an FBA simulation.

Modelling using the same kind of framework has been done for other systems, like stem cell population dynamics in *C. elegans* (Setty et al., 2012) and pancreatic organogenesis (Setty et al., 2008).

2.5 Measurement techniques

Several different measurement techniques can be employed to obtain data from biological systems and then used, e.g., for constructing the models and evaluating simulation results by comparing to data. All of these experimental techniques

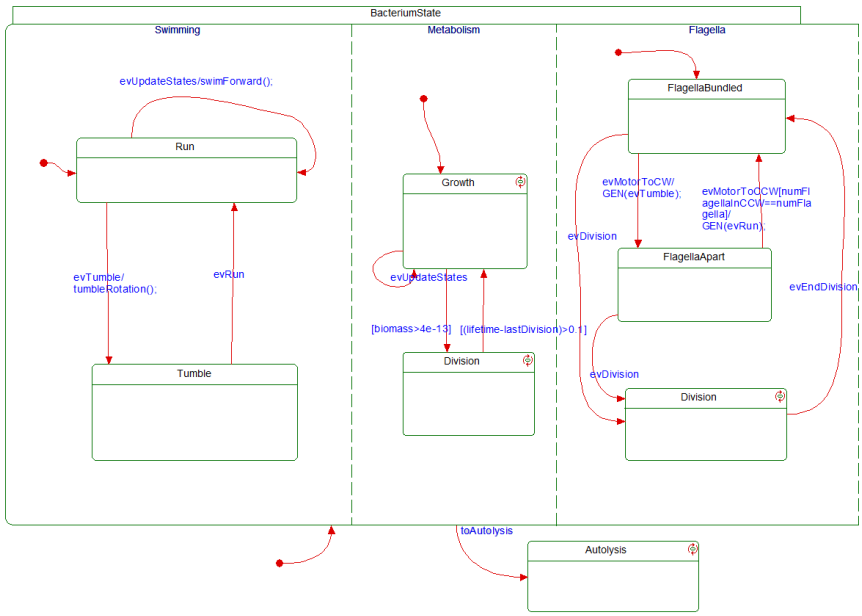


Figure 2.2: Part of the statechart of a bacterium used in modelling in Publication V. The dashed lines divide the bacterium state into substates. Possible transitions between states are shown with arrows and they can have conditions, which are stated within square brackets.

have their specific sources of errors and biases.

For building models of gene regulatory networks, perhaps the most essential measurements concern the states of genes under different conditions, timepoints or perturbations. A method for accurate quantification of mRNA levels is reverse transcription quantitative real-time PCR (RT-qPCR) measurement. However, since it is quite tedious and the interrogated genes need to be known beforehand when designing the required primer sequences, qPCR can only be performed for a very limited number of genes. Quantitative PCR is also not devoid of sources of errors, a major one being that the measured values need to be compared to either normalizing genes or DNA standards and for example variations in the normalizing genes between conditions can skew the results. It has also been noted that there are disagreements in the estimates obtained from different qPCR-based assays (SEQC/MAQC-III Consortium, 2014).

Microarrays (Schena et al., 1995; Pease et al., 1994) are a method for measuring the expression states of thousands of genes simultaneously. A simplified description of how microarrays work is that they contain short (about 25-60 nucleotide long) oligonucleotides, called probes, designed to target genes of interest. Labelled transcripts from the organism under study are then allowed to bind (hybridize) to these probes. Exciting the arrays with laser and mea-

suring the intensities then gives higher values for probes of genes having higher expression.

The probes on microarrays need to be specifically designed to target genes of interest. Although arrays capturing all the annotated genes are available for several different organisms, this requirement of pre-specified design is one of the problems of microarray technology, complicating studies of uncharacterised organisms and missing those genes that are not included in the array. Other major problems and sources of potential biases and errors in microarrays range from issues in manufacturing and hybridization (such as malformed or missing spots on array or targets cross-hybridizing several probes) to experimental design, imaging (for example strong and uneven background signal), data analysis (such as inadequate normalizations or batch-effect corrections) and statistical inference (Allison et al., 2006). Gene expression microarray data was used in Publication II.

Lately, RNA-seq (Mortazavi et al., 2008) has been replacing microarrays as the preferred assay for gene expression. The technique is based on massively parallel sequencing, where the mRNA molecules are extracted from the cells of the studied organism and from a subset of them a stretch of nucleotides is read from each. These sequences (called reads) vary depending on the used technology but are usually around 50-150nt long and thus do not cover the whole transcripts of most (in particular protein coding) genes. These short reads are then mapped to the genome or transcriptome (also called alignment) in order to identify the genes where they likely originated from. The quantities of reads aligning to each gene are then used as starting values for expression estimation algorithms.

RNA-seq does not suffer from the same reliance on gene annotations to start with as microarrays and allows new transcripts to be identified and even transcriptomes to be built. It can also be more robust to for example single-nucleotide polymorphisms (SNPs) than microarrays. However, there are several potential sources of biases in the analyses, starting with the construction of sequencing libraries (where, e.g., PCR amplification artifacts and differing GC content can produce biased counts) and alignment of the reads to the genome (problems of reads aligning to several locations or sequences having polymorphisms compared to the reference genome) (Fang and Cui, 2011). RNA-seq has been found quite reliable in quantifying relative expressions of genes even between different platforms but quantification of absolute expression is not yet very accurate (SEQC/MAQC-III Consortium, 2014).

The problems and error sources listed above for large-scale RNA quantification are mostly shared by a technique for measuring protein-DNA interactions, namely ChIP-chip (Ren et al., 2000) for array-based and ChIP-seq (Johnson et al., 2007) for sequencing-based platforms. Protein-DNA interactions are of great interest for example in trying to find regulatory interactions between genes. In addition to the potential sources of errors listed above, one notable factor caus-

ing uncertainty is due to the imperfect specificity of the antibody used to target the protein of interest.

Several cellular parameters, including for example protein expressions and enzyme activities, can be measured using flow cytometry. The technique is based on labelling certain cellular substances (such as the studied signalling proteins) with different fluorescent dyes, then making the cells pass laser(s) and detector(s) one at a time, thus measuring the fluorescence signal for each individual cell separately. The fluorescent labels are attached to antibodies targeting desired proteins on the surface of or inside the cell and therefore the applicability of the technique is restricted by the availability of suitable antibodies. Potential sources of errors in flow cytometry include for example overlapping emission spectra of labels and possibility of two instead of one cell passing the laser together. Flow cytometry data was used in Publications I–III.

Most, if not all, of the measurement methods are in addition affected by many other sources of errors, such as batch effects due to the person performing the experiments, different dates, and different batches of reagents, which all need to be taken into consideration in the analyses. Furthermore, replicates, which are essential for reliable statistical analysis, are often very hard to obtain for various reasons.

Chapter 3

Bayesian networks

Probabilistic graphical models represent a set of random variables and the probabilistic relationships between them using a graph-based representation. This allows the joint distribution to be described in a compact manner by encoding the independence structure as well as the factorization of the distribution. Bayesian networks (Pearl, 1985) are a class of probabilistic graphical models with a directed and acyclic graph structure. Bayesian networks have also been called with other names, such as belief networks, probabilistic networks, influence diagrams, causal networks, and probabilistic graphical models. The Bayesian network (BN) representation was presented already in 1921 by Wright (1921) and has also been appearing for example in Good (1961), Rousseau (1968), and Cooper (1984). BNs have been used as expert systems coding uncertain knowledge from experts and more recently they have largely been constructed from data.

The applicability and usability of BNs has seen a dramatic rise with the availability of cheap and efficient computing resources. BNs are a versatile model and their applications range from inference of genetic/cellular networks (Friedman, 2004; Segal et al., 2002), protein signalling networks (Sachs et al., 2005), protein-protein interaction networks (Jansen et al., 2003), predicting protein-protein interaction sites (Bradford et al., 2006) and numerous other applications (Heckerman et al., 1995b).

Similarly as with other graphical models, the graph-representation is perhaps the most attractive aspect of BNs. Being an intuitive visualization of the system structure, it is also easy to see (and formulate) some assumptions and make inferences based on the graph. From a mathematical point of view, graphs are invaluable in coding joint probability distributions efficiently.

Bayesian networks are often used to learn causal relationships between physical entities (Pearl, 2009). As the network structure of the system of interest is in many cases unknown, the aim is to infer it based on experimental data. Given a causal model it is then possible to predict results of interventions and explore ways in which to change the state of the system, for example from a disease state

to a healthy one. The ability of BNs to make use of interventional data to find correct edge directions (i.e. causal relationships) has also been found to set them apart from some other models having lower accuracy (Werhli et al., 2006).

Bayesian networks have several additional benefits that make them a very interesting model. Probabilistic models are a good choice for modelling stochastic systems or measurements, and Bayesian methodology introduces some extra advantages. For example, in many domains there is prior (expert) knowledge about the system and using it in conjunction with data in learning BNs is naturally handled via priors of the network. This aspect of utilizing priors as well as the ability to mix nodes of different types (discrete/continuous, varying dimensions and parameters) also allows combining data from different domains, such as using gene expression data together with putative promoter elements to learn gene regulatory network structure (Tamada et al., 2003; Troyanskaya et al., 2003; Myers et al., 2005; Segal et al., 2002; Hartemink et al., 2002) and other data types (Werhli and Husmeier, 2007). BNs can also handle incomplete data, where datapoints may be missing.

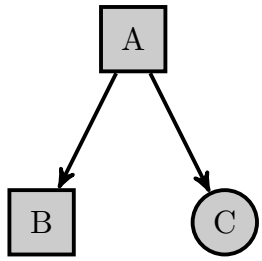
As a drawback, applicability of BNs is restricted in many areas by the computational complexity. Another major setback is the acyclicity requirement, which represents an obvious limitation for the usefulness of BNs as many real-life systems include feedback loops and self regulation. One way to circumvent this restriction is to use dynamic Bayesian networks (DBNs) (Friedman et al., 1998), however, their applicability is diminished by the requirement to have time-series data, although approaches to learn DBNs with static data have been presented (Lähdesmäki and Shmulevich, 2008).

In order to allow easier applicability to real-world problems, tools have been developed for BN reconstruction and simulation, such as Bayes Net Toolbox (Murphy, 2001b), BDAGL (Eaton and Murphy, 2007c), Banjo (Smith et al., 2006), BNFinder (Wilczyński and Dojer, 2009; Dojer et al., 2013), and many others (Murphy, 2013).

3.1 Definition of BNs

Bayesian networks (Pearl, 1985; Heckerman, 1998; Husmeier, 2005) represent joint probability distributions in a semi-graphical way. First, a Bayesian network includes a graph structure that describes the dependencies between a set of random variables. Second, for each random variable (represented by a node in the graph) there is a conditional probability distribution defining the relationships between its state and the states of its parent nodes. See Figure 3.1 for an example.

Formally, Bayesian network defines a joint probability distribution and consists of a pair (G, θ), where G is a directed acyclic graph (DAG) whose n nodes represent the set of random variables $\mathcal{X} = \{X_1, ..., X_n\}$ and its edges give a



A	$P(B=b_1\|A)$	$P(B=b_2\|A)$
a_1	b_{11}	b_{21}
a_2	b_{12}	b_{22}
a_3	b_{13}	b_{23}

A	$P(C\|A)$
a_1	$N(\mu_1, \sigma_1^2)$
a_2	$N(\mu_2, \sigma_2^2)$
a_3	$N(\mu_3, \sigma_3^2)$

Figure 3.1: Example Bayesian network graph structure. Square nodes represent discrete random variables and circle nodes stand for continuous random variables. Values of $P(B|A)$ are given in a conditional probability table and each row forms a discrete probability distribution, whereas conditional densities $P(C|A)$ are normal distributions characterized by mean μ_i and variance σ_i^2.

graphical representation of the conditional independencies between these random variables, namely each node X_i is conditionally independent of its non-descendants given the values of its parents in G. Parameter set θ defines the conditional probability distributions of the variables in $\mathcal{X}$. Based on G one can get the factorization of the joint distribution over $\mathcal{X}$ as

$$P\left(X_1, ..., X_n|G, \theta\right) = \prod_{i=1}^{n} P\left(X_i|\mathrm{Pa}_G(X_i), \theta_i\right), \tag{3.1}$$

where $\mathrm{Pa}_G(X_i)$ is the set of parents of node X_i in G, while θ_i denotes the parameters for the distribution of X_i conditional on its parents. Probability distributions that factorize according to the DAG are said to respect the directed factorization property and BNs can be used to model such distributions.

Learning the network parameters (i.e. those of the conditional distributions) can be performed for example by maximum likelihood estimation, where parameters θ maximizing $P(D|\theta)$ are searched for and D is the data. Incomplete datasets can be handled by using for example expectation-maximization (EM) (Dempster et al., 1977) or sampling methods such as Gibbs sampling or other Monte Carlo methods.

Inference is conceptually simple in Bayesian networks: One just calculates the conditional probability for a node of interest, e.g., $P(X|Y)$, where Y can be a set of nodes. In practice, however, computing this is not usually easy, although many efficient methods for it exist (Cowell et al., 1999).

3.2 Learning the structure of BNs

In many cases the network structure of the system that generated the observed data is unknown. Bayesian networks can be used to find the most probable network structure with or without any prior knowledge about the domain. It must be noted that in strict sense BNs are restricted to acyclic networks but even in cyclic cases they can reveal many true interactions, in particular if one is not restricting to a single DAG structure but uses (a sample from) the posterior distribution.

Searching for the structure that most probably generated the data is performed by trying to find the DAG G that maximizes the posterior probability given the data D

$$P\left(G|D\right) = \frac{P\left(D|G\right)P(G)}{P(D)} \propto P\left(D|G\right)P(G), \tag{3.2}$$

where $P(G)$ is the prior probability of G,

$$P(D) = \sum_{G' \in \mathcal{G}_n} P\left(D|G'\right)P(G') \tag{3.3}$$

is the probability of data (also called evidence), $\mathcal{G}_n$ is the set of all possible DAG structures with n nodes, and

$$P\left(D|G\right) = \int_{\theta} P\left(D|G,\theta\right)P\left(\theta|G\right)d\theta \tag{3.4}$$

is the marginal likelihood. As $P(D)$ is constant when searching for maximizing G, it can be dropped from Equation 3.2, which is then often called the score function in learning BN structure.

For certain choices of probability distributions and parameter priors, it is possible to arrive at a closed form solution for the marginal likelihood. The two main cases are multinomial distributions with (independent) Dirichlet priors (Cooper and Herskovits, 1992; Heckerman et al., 1995a) and Gaussian distributions with normal-Wishart priors (Geiger and Heckerman, 1998). It is therefore tempting to choose to use discrete-valued data and BNs having multinomial conditional probability distributions. This of course necessitates discretization of data in many practical cases, which on one hand can be seen to cause quantization noise but on the other hand can reduce measurement noise, especially in case it is known that the observables have quantized states, which however is not true for, e.g., genes in general (Hartemink, 2001).

Using uniform Dirichlet parameter priors $P\left(\theta|G\right)$ (as Dirichlet distribution is the conjugate prior of multinomials), Equation 3.4 becomes (Heckerman et al., 1995a)

$$P(D|G) = \prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij}+N_{ij})}\prod_{k=1}^{r_i}\frac{\Gamma(\alpha_{ijk}+N_{ijk})}{\Gamma(\alpha_{ijk})}, \tag{3.5}$$

where N_{ijk} is the number of times the configuration $(X_i = V_{i,k}, Pa_G(X_i) = j)$ occurs in data D when the set of r_i values that variable X_i can take is $\{V_{i,1}, V_{i,2}, ..., V_{i,r_i}\}$, α_{ijk} are hyperparameters (a.k.a. pseudo-counts) of the Dirichlet distributions, $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$ and $\alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk}$, and q_i is the number of different parent configurations.

For many other choices of parameter distributions the marginalization in Equation 3.4 is impractical, and sampling or approximation methods need to be used. An example is Bayesian information criterion (BIC), which approximates the logarithm of the score as $-2 \ln p(D|\hat{\theta}, M) + k \ln N_D$, where $\hat{\theta}$ is the (maximum likelihood) estimate for parameter values, k is the number of parameters in the model and N_D is the number of data points.

The structure priors $P(G)$ can be used to include information about the network structure gathered from other data sources, either as expert knowledge or based on previously measured data (Bernard and Hartemink, 2005; Mukherjee and Speed, 2008). When such information is not available, the usual choices are uniform prior or priors that penalize for growing complexity (Friedman and Koller, 2003).

For data points where the value of one or more variables has been perturbed (also called "clamped") to have certain values, the above equations can be used as such. However, to allow this, the edges that end in the clamped nodes need to be removed first and the parameters of the clamped nodes are not updated (Cooper and Yoo, 1999). This works unless there are hidden nodes, in which case Pearl's "do calculus" (Pearl, 2009) must be used. Another assumption is that the interventions are ideal, i.e. that the values of nodes are indeed what they were set to be. In reality some perturbations might not work totally as intended, which has been taken into account in some models (Eaton and Murphy, 2007b).

In some (especially biological) applications, the amount of learning data can be very restricted. If the dataset size is small relative to the network size, suboptimal models explain the data almost equally well as the optimal one and as a solution Friedman and Koller (2003) suggest using frequently appearing features instead of whole structures. This is also sensible if the underlying system is not necessarily completely following a DAG structure and thus selecting the strongest features allows us to extract some information about the system. The estimated probabilities of features are calculated as

$$P(f|D) = \sum_{G \in \mathcal{G}_n} P(G|D) I_f(G), \tag{3.6}$$

where I_f is an indicator function, i.e. $I_f(G) = 1$ if graph G contains the wanted feature f and $I_f(G) = 0$ otherwise. An example of features are edges, whose estimated probabilities can be used to, e.g., derive a network with high confidence edges taking, say, all edges with $P(edge|D) > 0.5$.

The space of different DAGs grows super-exponentially with number of nodes, see Table 3.1. Exhaustive evaluation of Equation 3.2 (as well as Equation 3.3 or

number of nodes	number of DAGs
1	1
2	3
3	25
4	543
5	29281
6	3781503
7	1.1×10^9
8	7.8×10^{11}
9	1.2×10^{15}
10	4.2×10^{18}

Table 3.1: Number of different directed acyclic graphs (DAG) as a function of the number of nodes (Robinson, 1977).

3.6) is therefore practically impossible already when n is greater than about 7. It is therefore necessary to use heuristic or sampling methods for structure learning, e.g. Markov chain Monte Carlo that is discussed in the next section. Some other possibilities for learning include greedy search and simulated annealing. There are also developments for structure learning that can yield the optimal structure in less than super-exponential time (Koivisto and Sood, 2004; Eaton and Murphy, 2007a).

3.3 Markov Chain Monte Carlo

Only chains with discrete states y_i from a finite set $\mathcal{Y}$ are considered here. Let a sequence of random variables $Y_1, Y_2, Y_3, \ldots$ obey the Markov property

$$P(Y_{n+1} = y_{n+1} | Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n) = P(Y_{n+1} = y_{n+1} | Y_n = y_n) \tag{3.7}$$

i.e. what happens next depends only on the current state. Such a stochastic process is called a (discrete-time) Markov chain of first order. The transitions between consecutive states are determined by a *transition matrix* $K_n(Y_n = i, Y_{n+1} = j)$ that obeys $K_n(Y_n = i, Y_{n+1} = j) \geq 0$ and $\sum_j K_n(Y_n = i, Y_{n+1} = j) = 1$ for all $i, j \in \mathcal{Y}$. Thus, each element of K_n represents a probability of transition, i.e., the chain moves from i to j with probability $P(Y_{n+1} = j | Y_n = i) = K_n(Y_n = i, Y_{n+1} = j)$. Here we consider only *time-homogeneous* Markov chains, where

$$P(Y_{n+1} = j | Y_n = i) = P(Y_n = j | Y_{n-1} = i) = K(i, j) \tag{3.8}$$

for all n. Consequently, probabilities of transitions of length m are given by the mth power of matrix K.

A state j is said to be *accessible* from state i if there exists an $m \geq 1$ so that $P(Y_{n+m} = j|Y_n = i) > 0$. The state j *communicates* with state i if both i is accessible from j and j is accessible from i. A Markov chain is called *irreducible* if all its possible states communicate with each other, thus it is possible to get from any state to any other state.

The period of a state i is defined as

$$m = \gcd \{m \in \mathbb{Z} : P(Y_{n+m} = i|Y_n = i) = K^m(i, j) > 0\}, \tag{3.9}$$

where gcd is greatest common divisor, and a state is said to be *aperiodic* if $m = 1$. The chain is aperiodic if all its states are aperiodic. For an irreducible chain a single aperiodic state suffices to make the whole chain aperiodic. Furthermore, a finite-state irreducible and aperiodic Markov chain is called *ergodic*.

The *stationary distribution* of a Markov chain is a probability distribution π over the state-space that satisfies

$$\sum_{i \in \mathcal{Y}} \pi(i)K(i, j) = \pi(j). \tag{3.10}$$

This means that if the chain is in a stationary distribution then it will stay there, therefore being called also equilibrium distribution.

The fundamental theorem of Markov chains states that an ergodic Markov chain has a unique stationary distribution π and that

$$\pi(j) = \lim_{m \to \infty} K^m(i, j) \tag{3.11}$$

for all i and j. In words this says that when running the chain long enough, it will converge to the stationary distribution and the probabilities $\pi(j)$ are independent of the initial states i.

Equation 3.10 can also be written as $\pi K = \pi$, where π is a row vector. Computing a stationary distribution can thus be done by simply solving π from the system of equations $\pi K = \pi$. However, this becomes practically impossible when the state-space is large.

The practical usability of this is that if the cardinality of the state-space, $|\mathcal{Y}|$, is very large and if it is hard to sample directly from π, then it is still possible to achieve a good estimate of π if transitions between states are easily done using $K(i, j)$. This is the main reason why MCMC methods have become hugely popular. It is also sufficient to know π only up to a normalizing constant.

The question then is how to devise a Markov chain that has the distribution of interest as its stationary distribution. The way to sample from π is described by the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970), which we briefly review here given the application of sampling from the posterior distribution of BN structures $P(G|D)$ (Madigan and York, 1995).

The Metropolis-Hastings algorithm starts from an initial structure and then iterates a given number of times using the following transition matrix

$$K(G_i, G_j) = \begin{cases} Q(G_j|G_i) & \text{if } G_i \neq G_j, A(G_i, G_j) \geq 1 \\ Q(G_j|G_i)A(G_i, G_j) & \text{if } G_i \neq G_j, A(G_i, G_j) < 1 \\ Q(G_j|G_i) + \sum_{G_k : A(G_i, G_k) < 1} Q(G_k|G_i)(1 - A(G_i, G_k)) & \text{if } G_i = G_j \end{cases} \tag{3.12}$$

where $Q()$ is a so called *proposal distribution* that proposes for example a transition from G_i to G_j with probability $Q(G_j|G_i)$, and

$$A(G_i, G_j) = \frac{P(D|G_j)P(G_j)Q(G_i|G_j)}{P(D|G_i)P(G_i)Q(G_j|G_i)} \tag{3.13}$$

is called acceptance ratio, based on which the proposed moves are accepted with probability 1 if $A(G_i, G_j) \geq 1$ and with probability $A(G_i, G_j)$ if it is less than 1, otherwise the chain stays at G_i.

It can be shown that the transition matrix of Equation 3.12 together with the acceptance ratio of Equation 3.13 satisfies a condition called *detailed balance*, which guarantees that the stationary distribution of the chain is the desired posterior distribution $P(G|D)$. To prevent sampling from mostly low-probability areas of the state-space, the chain is usually allowed to run for a long time (burn-in phase) after which the sampling is done (sampling phase). Based on the dimensions and complexities of the state-spaces, these phases might need to be very long.

3.3.1 Convergence

Even though an ergodic MCMC chain is guaranteed to converge to the stationary distribution, it is guaranteed to do so only when running infinitely long. In practical cases it is thus crucial to assess if the chain has converged and whether the obtained sample is representative enough. There exists no way to be sure whether the chain is really converged but several methods representing necessary though not sufficient criteria for assessing this have been presented (Cowles and Carlin, 1996).

In case of BN structure learning one frequently used heuristic indicator for convergence is the similarity of edge posterior probabilities (Equation 3.6) that have been calculated from two or more independent chains. These posterior estimates need to be close to each other if the chains have converged to the same area, which can easily be visually examined for example by plotting them against each other and observing whether there are noticeable deviations from the diagonal.

3.3.2 Proposal distribution

Based on Equations 3.12 and 3.13 it is evident that, in addition to the burn-in and sampling phase lengths, the proposal distribution is the only adjustable part in the Metropolis-Hastings algorithm. In cases of continuous state-spaces, the movements may often be tuned more easily and adaptive schemes, where the proposal distribution changes during the running of the chain, can be used (Haario et al., 2006). In case of more complex state-spaces, such as the space of DAGs, the movements can be more involved. In DAG space the basic movements consist of adding, deleting or reversing an edge between a given pair of nodes, while adhering to the acyclicity constraint. The convergence of MCMC in structure learning of BNs with this simple proposal distribution is often slow in convergence and tends to get trapped in local maxima (Castelo and Kočka, 2003; Friedman and Koller, 2003). Improvements to the proposal distributions in context of Bayesian network structure learning have therefore been developed (Castelo and Kočka, 2003; Moore and Wong, 2003; Grzegorczyk and Husmeier, 2008)

In Publication I, a modification is presented that allows efficient movements to larger than one-step neighbourhoods in the DAG space. An issue when constructing the proposal distribution stems from the fact that, in order to calculate the acceptance probability in Equation 3.13, the value of $\frac{Q(G_i|G_j)}{Q(G_j|G_i)}$ (called Hastings ratio) must be determined and to do this the sizes of the neighbourhoods are needed in case of the usual uniform proposal distribution. For neighbourhoods consisting of structures that differ by only a single-edge modification this is not a big problem, as an efficient algorithm for movements taking into account the acyclicity requirement has been proposed (Giudici and Castelo, 2003). However, the sizes of neighbourhoods grow super-exponentially and consequently, for larger than single-edge neighbourhoods, their evaluation and acyclicity checks quickly become computationally very demanding.

As presented in Publication I, instead of evaluating whole neighbourhoods, one can use consecutive single-edge modification steps. Formally, let $Q^t(k|i,\mathbf{r})$ be a proposal distribution from structure i to k that can be decomposed into t (here $t>1$) independent (sub)distributions Q_j, $j=1,\ldots,t$, and $\mathbf{r}=(r_1,r_2,\ldots,r_{t-1})$ is a tuple of intermediate structures so that the whole move is $i \to r_1 \to \cdots \to r_{t-1} \to k$. Note that, for simplicity, we have here marked structures by only their indices, e.g., structure G_i is denoted by i. Now the probability of proposing a move from i to k is

$$\begin{aligned} Q^t\left(k|i,\mathbf{r}\right) &= Q_1\left(r_1|i\right) Q_2\left(r_2|r_1\right) \cdots Q_t\left(k|r_{t-1}\right) \\ &= Q_1\left(r_1|i\right) Q_t\left(k|r_{t-1}\right) \prod_{j=2}^{t-1} Q_j\left(r_j|r_{j-1}\right) \end{aligned} \tag{3.14}$$

and each subdistribution $Q_j\left(r_j|r_{j-1}\right)$ is the standard uniform probability distri-

bution over the single-edge modification neighbourhood of r_{j-1}, i.e. $Q_j\left(r_j|r_{j-1}\right) = \frac{1}{q(r_{j-1})}$, where $q(\cdot)$ is a function giving the number of structures differing by a single-edge modification from DAG r_{j-1}.

It is shown in Publication I that in this case the Hastings ratio becomes

$$\frac{Q^t\left(i|k\right)}{Q^t\left(k|i\right)} = \frac{q(i)}{q(k)}. \tag{3.15}$$

Thus, it suffices to evaluate only the starting and ending DAG neighbourhoods, which creates only minor increase in computational requirements. It is also shown that such composite proposal distributions are ergodic, better than the basic single-edge proposal distribution in terms of convergence, and allow for construction of adaptable proposal distributions.

Other modifications to BN structure learning

Developments where the order of the nodes is first learned have also been presented (Friedman and Koller, 2003; Ellis and Wong, 2008; Niinimäki and Koivisto, 2013). Searching in the space of equivalence classes has also been suggested (Chickering, 2002). The method of Koivisto and Sood (2004), where dynamic programming is employed to calculate the posterior probabilities of all BNs in exponential time, and variations of this like (Eaton and Murphy, 2007a), are also important contributions in the field. Even though such methods can lead to improvements, e.g., in the convergence of the chains, using informative structural priors in conjunction with them may be tricky.

3.4 Active learning

A problem in learning BN structures is that factorization of the joint probability using Equation 3.1 can lead to the exactly same result for several different networks. In other words, there can be more than one network coding the same independence assumptions between variables. The set of network structures that produce the same factorization is called an *equivalence class*. An example is given in Figure 3.2. The equivalency of networks can easily be checked graphically since two DAGs are equivalent if they have the same structure ignoring edge directions (called skeleton of the graph) and the same v-structures (a node with two non-adjacent parents) (Verma and Pearl, 1991). Equivalence classes are in fact a manifestation of the well-known saying that correlation does not imply causation.

Equivalence classes lead to the problem that in learning the network structure several different DAGs are score-equivalent and, consequently, one is able to only distinguish between equivalence classes but not the DAGs inside them. However, as shown in Figure 3.2, one can break these classes by making interventions, i.e.

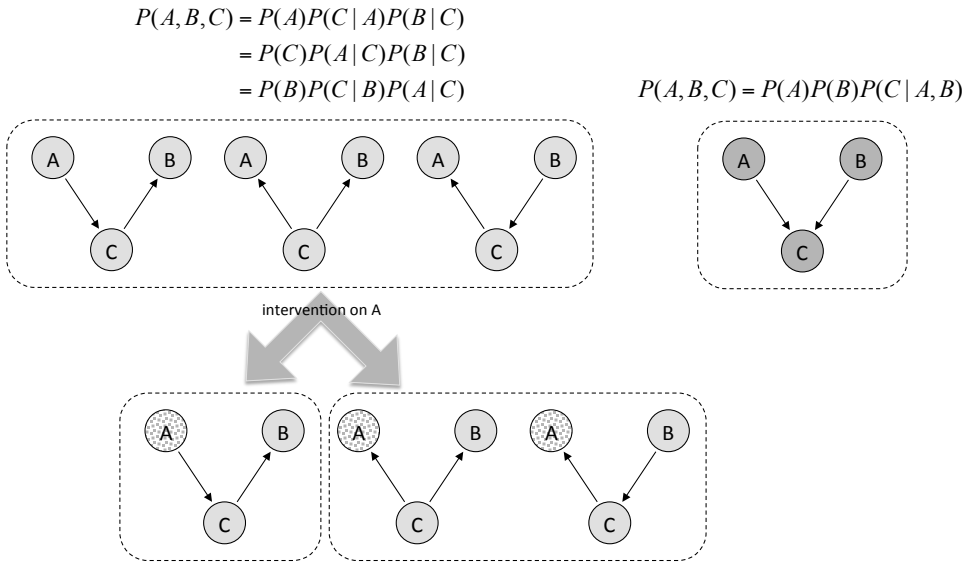


Figure 3.2: Factorization of the three of the four BNs on the upper row result in the same join probability distribution and belong to the same equivalence class. Intervening with node A by setting it to a given known state allows breaking an equivalence class into two smaller classes (of which the other is a singleton). Modified from (Husmeier, 2005).

setting values for certain nodes and then observing how those interventions affect the network state. Another way of making a difference between networks in an equivalence class is via structure priors.

The selection of measurements to perform in order to learn the network structure as efficiently as possible is an important problem. Instead of selecting the interventions to be done totally randomly, one can try to select a set of nodes to be intervened so that the expected benefit for the learning process is maximized and consequently the cost of experiments is minimized.

Active learning (Settles, 2009) aims at selecting the measurements in the most beneficial way and can therefore be also called optimal experimental design. It is a special case of semi-supervised machine learning, meaning it is capable of making use of both unlabelled and labelled data. Instead of the user or teacher deciding and providing the algorithm with labelled samples, active learning algorithm itself queries the user for labelled data points. Active learning methods have also been presented in context of BN structure learning (Murphy, 2001a; Tong and Koller, 2001; Pournara and Wernisch, 2004).

In general the problem can be formulated as selecting the action a^* giving

the maximal expected utility, as defined by function $F(a)$, i.e.

$$a^* = \arg\max_{a \in A} F(a), \tag{3.16}$$

where the set A consists of all the possible actions. The actions can be for example perturbations of nodes in the network before measurement, as in Publication II. In this case the expected utility was defined as

$$F(a) = \sum_{G \in \mathcal{G}} \sum_{y \in \mathcal{Y}_{G,a}} P\left(y|G,a,D\right) P\left(G|D\right) U\left(G,a,y,D\right), \tag{3.17}$$

where $\mathcal{G}$ is the set of possible DAGs and $\mathcal{Y}_{G,a}$ denotes the set of possible observations that G can produce given that perturbation a has been made. For the utility function $U(G,a,y,D)$ we use (assuming equivalent cost for each intervention) $\log P\left(G|a,y,D\right)$ (Murphy, 2001a). Due to the super-exponentially growing number of DAGs and the large amount of different possible states $\mathcal{Y}_{G,a}$ the network can take, computational complexity increases so that in trying to select the action maximizing Equation 3.16, one must use sampling instead of exhaustive evaluation with networks having more than about 6 nodes.

Publication II presents and benchmarks this active learning algorithm in two realistic cases with actual experimental data. The first dataset consists of expression measurements of 7 transcription factor genes in *Halobacterium salinarum*, including both wild-type measurements as well as over-expressions. The second one contains flow cytometry measurements of 11 signalling network proteins, of which for 5 the dataset includes perturbation measurements (Sachs et al., 2005). Active learning is found to perform on average more efficiently than randomly selecting the measurements (or interventions) to perform, see Figure 3.3 for an example. However, the performance is not quite as good as with simulated data, which has most often been used to assess the methods. Reasons for lower average performance, although still better than when selecting perturbations randomly, are likely at least factors outside the model affecting the measured nodes and cyclic regulatory relationships not captured by BNs.

Publication III discusses how to apply active learning in a scenario with heterogeneous datatypes by means of using structure priors. There can often exist multiple types of measurement data from the same biological system and including those in the inference of network structure can be done with BNs, by incorporating part of the data through likelihoods and the rest via structure priors. We concentrated on a biologically motivated setting, where one datatype contains measurements of the states of network nodes, e.g. gene expression measurements, whereas the other datatype can be used to measure the probabilities of outward-edges for nodes, such as done in for example ChIP-seq measurements.

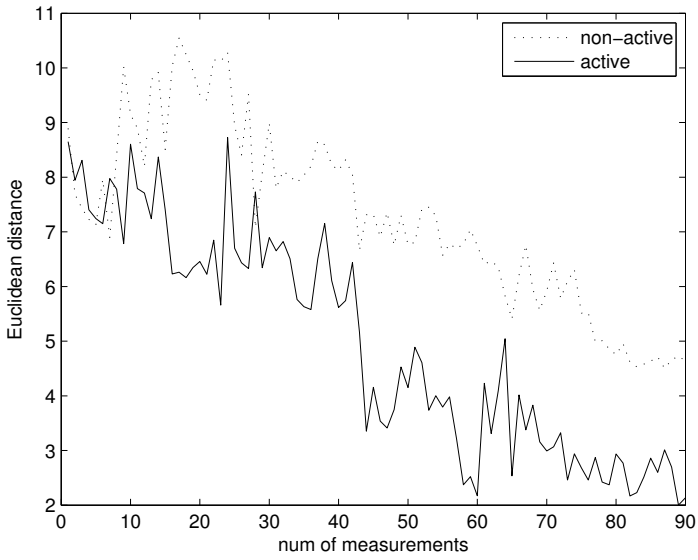


Figure 3.3: Comparing performance of randomly selecting the next experiment to selecting using active learning. Test case is a signalling network dataset from (Sachs et al., 2005). Euclidean distances between the current estimates of edge posterior probabilities (calculated with data gathered so far) and the "true" posterior probabilities (calculated from the whole dataset with big samples from MCMC chains) are shown as function of number of interventional measurements selected so far. Results are averaged over four runs. Figure from Publication II

In this case the posterior probability of a graph structure g can be written as

$$P(g|D,B) = \frac{P(D,B|g)P(g)}{P(D,B)} = \frac{P(D|g)P(B|g)P(g)}{\sum_{g'} P(D|g')P(B|g')P(g')}$$
$$= \frac{P(D|g)P(g|B)}{\sum_{g'} P(D|g')P(g'|B)},$$

where D is the dataset representing expression-type measurements and B is the adjacency probability matrix containing the outward-edge dataset, for which the measurements are included by converting p-values to probabilities of edges and setting $B(i,j) = P(X_i \rightarrow X_j)$, as discussed in Publication III. In the case of utilizing active learning for selecting new data points to be measured for B, the set of actions A in Equation 3.16 consists of the nodes whose outward-edges can be measured, and two different functions $F(a)$ are presented in Publication III. Both are shown to achieve better expected learning performance than randomly selecting the nodes for measurement.

It should be noted that active learning is not guaranteed to deterministically outperform random selecting as the whole framework is stochastic. In fact, even

worse performance can at times be observed, even though on average active learning yields beneficial results.

Chapter 4

Modeling metabolic networks

Metabolites interact with each other, being able to combine into new metabolites, exchange molecules, or break apart. This process of interactions is naturally described as a network, where nodes represent metabolites and edges the interactions between them. Constructing such descriptions or models of metabolism (often called *reconstructions* or genome-scale models (GEMs) of metabolism) usually starts by identifying genes that code for known metabolic enzymes in the genome of the organism being studied. Finding out the functions of genes is initially done by homology search and sometimes followed by manual or experimental curation. The rapidly reduced cost of genome sequencing has greatly lowered the threshold of obtaining such annotated (even though often just draft) genomes.

Given a list of enzymes assumed to be present in the cell, one has implicitly also a list of metabolites and possible interactions between them. However, these lists are still most likely incomplete, as even the most studied genomes are not thoroughly annotated, and the lists are also prone to contain erroneous items, due to for example inadequacy of homology to resolve functional proteins and independence of some reactions (like transfer/exchange) on enzymes. As an example of a well-curated but still incompletely annotated organism one might consider human and its metabolic reconstruction, which still in the latest version approximately doubled the numbers of both metabolites and reactions (Thiele et al., 2013).

Knowledge of the reactions catalyzed by the set of enzymes allows addition of stoichiometric coefficients to the model, revealing how many molecules of each metabolite take part in one interaction step. Some reactions are also thermodynamically feasible in only other direction and this directionality information should also be included.

The well-known central dogma of molecular biology (Crick, 1958) states that the information flow is directed from gene to RNA to protein. Based on this, rules can be generated for reactions in a reconstruction stating whether the

enzyme (which is a protein) needed for the reaction to take place is available within the cell. Thus, in the most basic setting, if gene G codes for RNA R, which is transcribed to enzyme E, then from expression of G it can be deduced that the reaction catalyzed by E can take place.

Complications with this simple model of gene-enzyme connection arise in many fronts. First, the relationship between gene expression and protein levels is far from being simple (Vogel and Marcotte, 2012). Another level of complexity arises in eukaryotic cells, where a single gene can produce multiple transcripts, which in turn can each produce multiple different proteins as depicted in Figure 4.1. In addition, many enzymes are protein complexes being built from multiple (different) proteins. Therefore the reconstructions may contain gene-protein-reaction association rules that state in Boolean formalism the required states of genes in order for the reaction to be able to occur. This information about the gene-reaction associations is essential in using metabolic models for predicting phenotypic effects in for example knock-out mutants.

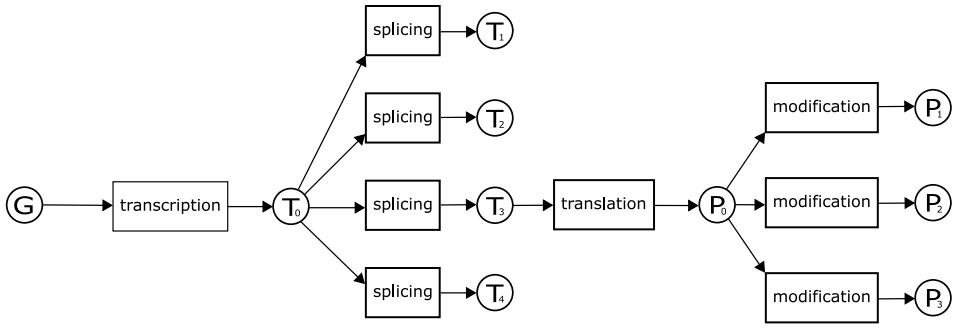


Figure 4.1: The synthesis principle of proteins in an eukaryotic cell. G denotes a gene, T_i are the different splice variants and P_i the resulting proteins after post-translational modifications.

The above description of the process of building (called reconstruction of) metabolic networks is oversimplified. However, the way to make metabolic reconstructions is quite established (Feist et al., 2009; Thiele and Palsson, 2010) and many advanced tools for the task exist (Karp et al., 2010; Thorleifsson and Thiele, 2011; Schellenberger et al., 2011; Pabinger et al., 2014).

Since the first genome-scale metabolic reconstruction in Edwards and Palsson (1999), over a hundred published reconstructions have appeared (Feist et al., 2009; InSilicoOrganisms, 2009). Some of the most studied and well-developed ones include *E. coli* (Feist et al., 2007; Orth et al., 2011) and *S. cerevisiae* (Mo et al., 2009). Also for some widely-studied organisms, the reconstructions have recently been developed and improved in a community-driven effort, such as for human (Thiele et al., 2013) and yeast (Herrgård et al., 2008). However, investigation of the reconstructions reveals that there is still much room for im-

provement, for example in that many model mainly only the primary metabolism (Monk et al., 2014).

4.1 Constraint-based models

The structure of metabolic networks can be studied using graph theoretic methods but here we are more interested in the flow (or, more precisely, fluxes) of metabolites through the network. Even though kinetic information for the majority of reactions is unknown and absent from the models described above, the constraints set by the models (such as structural, stoichiometric and thermodynamic ones) limit the allowed functioning of the metabolic network and studying the resulting flux space is insightful and can yield accurate predictions. Studying the models built given the structure and constraints is called constraint-based modelling (Bordbar et al., 2014). In the following, we shortly review basic concepts in characterization of the set of all possible flux distributions, which has also been called metabolic pathway analysis (Schilling et al., 1999). See also (Schuster et al., 2002) for a good description of the matters discussed below.

In a system of m metabolites and r possible reactions between these metabolites, the reaction rates at a given time t can be described as a vector $\mathbf{v}(t) = [v_1(t), v_2(t), \ldots, v_r(t)]^T$, where $v_i(t)$ is the reaction rate of the ith reaction. Vector $\mathbf{v}$ is often called flux distribution (we drop the notation t wherever it is better for readability).

Each of the reactions is either reversible or irreversible. The direction is defined by thermodynamic constraints so that even though a reaction might in reality be able to proceed in both directions, it is still classified as irreversible if under the prevailing conditions it can only happen in the other direction. Based on these, a flux distribution $\mathbf{v}$ can be divided into subvectors $\mathbf{v}_{rev}$ and $\mathbf{v}_{irr}$ containing the reversible and irreversible reactions, respectively, and now we can write $\mathbf{v} = \begin{bmatrix} \mathbf{v}_{rev} \\ \mathbf{v}_{irr} \end{bmatrix}$. The directions of the irreversible reactions are defined to satisfy

$$\mathbf{v}_{irr} \geq 0. \tag{4.1}$$

System boundary modelling can be done in at least two ways, as shown in Figure 4.2. First, it is possible to define some reactions as transfer reactions capable of bringing a metabolite to the system or removing it from there (i.e. they are one-ended edges across the system boundary). The other way is to categorize metabolites to be either *internal* or *external*. In the first category all the reactions having influence on the concentration of the metabolite are included in the studied model, while in the latter one not all the reactions with an effect on the metabolite are included in the model and their concentration is assumed constant (sometimes called source or sink metabolites).

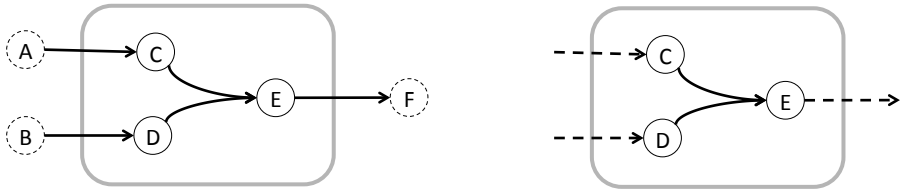


Figure 4.2: Two alternative views for modelling system boundaries. In the left figure the dashed nodes represent external metabolites and the nodes with solid lines are internal metabolites. In the figure on the right hand-side dashed lines denote exchange or transfer reactions. The grey lines depict the system boundaries.

The network structure and stoichiometric coefficients (which define the quantity of a molecule taking part in a reaction) of the system can be described by an $m \times r$ stoichiometric matrix S. Now, we can define the system to be in *steady-state* if all the metabolites i are being consumed at the same rate as they are being produced, i.e.

$$\frac{dX_i}{dt} = \sum_j S_{i,j} v_j = 0, \tag{4.2}$$

where X_i is the concentration of metabolite i. Thus, for the whole system

$$S\mathbf{v} = \mathbf{0}, \tag{4.3}$$

which can be attained (at least approximately) by many biochemical systems in suitable stable conditions, such as in some chemostat cultivations.

Equations 4.3 and 4.1 together define the allowed part of the flux space where the system can be given the constraints. If we first look at Equation 4.3, we see that the vectors $\mathbf{v}$ satisfying it form a null-space (also called kernel) of S, i.e. $K = null(S) = \{\mathbf{v} \in \mathbb{R}^r : S\mathbf{v} = \mathbf{0}\}$.

If all the reactions in the network are reversible, then the limits of the metabolic capabilities can be defined as the linear basis of the null-space. Usually this is not the case, and the irreversibility constraints of Equation 4.1 restrict the null-space into an allowed sub-space, namely every element in $\mathbf{v}_{irr}$ defines a half-space in the null-space. The result is a convex polyhedral cone, which is referred to as the *flux cone*. This is illustrated in Figure 4.3.

In addition to the above constraints, the flux cone can also be defined by the vectors that form its edges and these vectors are called *extreme rays* or *generating vectors* (Gagneur and Klamt, 2004). They unambiguously define the cone, much like a set of basis vectors, except that they are not necessarily linearly independent.

There are two different but closely related approaches for identifying fluxes that are close to being generating vectors. These are elementary modes (EM) (Schuster et al., 2000) and extreme pathways (EP) (Schilling et al., 1999;

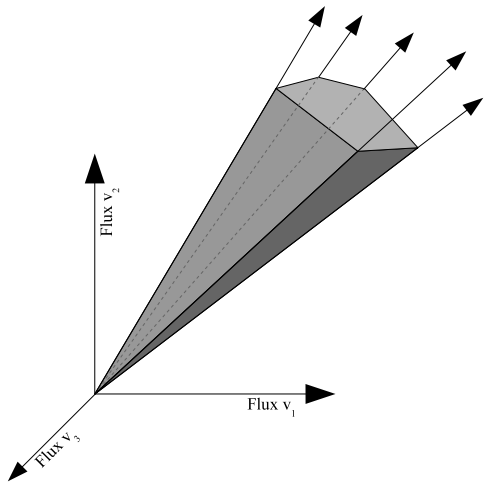


Figure 4.3: A flux cone in the space of three fluxes. The arrows forming the edges of the cone are the generating vectors.

Schilling et al., 2000). See for example (Klamt and Stelling, 2003) for a review and comparison of both.

Intuitively, both EMs and EPs represent a minimal set of fluxes (or pathways) whose superposition can be used to achieve any feasible steady-state of the network. For example in case of EMs

$$\mathbf{v} = \sum_j \alpha_j \mathbf{e}_j, \tag{4.4}$$

where $\alpha_j \in \mathbb{R}, \alpha_j \geq 0$, and $\mathbf{e}_j$ is the jth EM of the network in consideration. This property means that the EMs give the limits of the metabolic network and work as its basis (although not linearly independent). An example of a network and its elementary modes is shown in Figure 4.4.

Both EM and EP methods are seriously limited by combinatorial explosion, which has resulted them being mostly of theoretical interest while flux balance analysis, which is presented in the next subsection, is the main tool used. However, there are also developments that characterize the flux cone much more efficiently (Larhlimi and Bockmayr, 2009; Rezola et al., 2011).

Some methods aim at characterizing the flux distributions based on experimental data, such as measurements of nutrient uptake rates or isotope labelling data. Such methods can be called Metabolic Flux Analysis (MFA) (Wiechert, 2001) or Metabolic Network Analysis (MNA) (Christensen and Nielsen, 2000). Including such data basically further constraints the flux space and thus enables the estimation of intracellular fluxes in greater detail. The possibility of obtaining such data is limited, e.g., due to difficulties in labeling the substrates

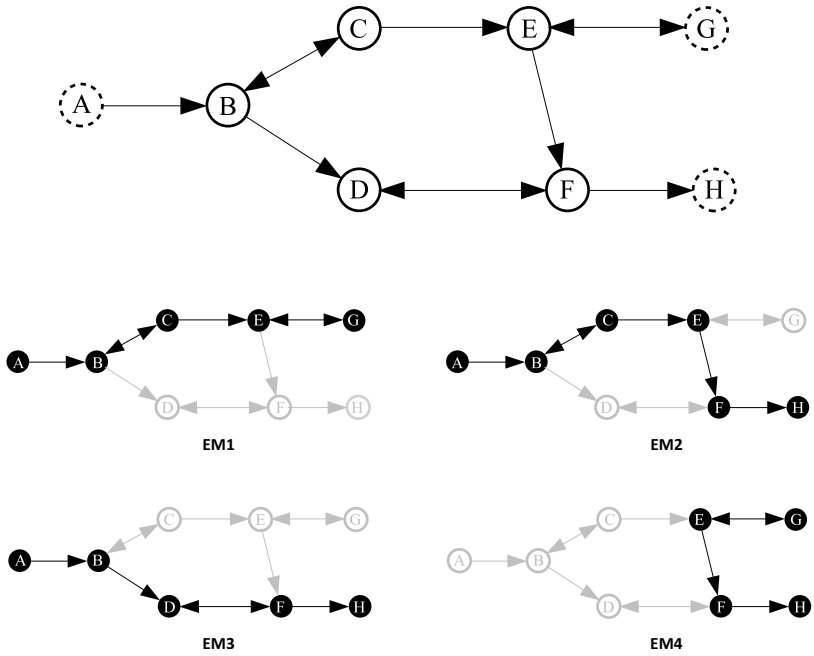


Figure 4.4: An example network on top. Below are its four elementary modes denoted in black. Metabolites A, G and H are external.

and in addition mass spectrometry or nuclear magnetic resonance techniques are required for the measurements.

As the majority of biochemical reactions are mediated by enzymes, which are proteins coded by genes, it is natural that information about genes and their expression states is utilized in metabolic network modelling. The information about connection between genes and reactions is often existing as result of the reconstruction of metabolic network being based on identified enzyme-coding genes, as described above. However, factors complicating this task are that gene expression values cannot in general be taken as surrogates for protein concentration or expression values as this connection has been shown to be only moderate (Greenbaum et al., 2003; Vogel and Marcotte, 2012) and that transcriptional regulation has been found insufficient to explain flux changes (Chubukov et al., 2013).

Thus, even though the effect of gene regulation is reduction in the space of steady-state solutions (Covert et al., 2001b; Covert and Palsson, 2003) and including this extra information in the modelling should in principle make the results more accurate (and indeed has, e.g., (Covert et al., 2001b; Covert et al., 2004; Lerman et al., 2013)), the combination of metabolic and gene models is not easy. For a survey on the methods integrating transcriptomic data to constraint-based models see, e.g., (Machado and Herrgård, 2014).

4.1.1 Flux balance analysis

The goal of flux balance analysis (FBA) (Varma and Palsson, 1994; Kauffman et al., 2003; Raman and Chandra, 2009) is to find a particular optimal flux distribution given the same constraints as discussed above, i.e. the result is a single point in the flux cone. The optimal solution $\mathbf{v}$ is one that minimizes (or maximizes) an objective function. This can be stated as a linear programming problem where $\mathbf{v}$ is subject to the mass balance (steady-state) constraint $\mathbf{Sv} = \mathbf{0}$ and constraints on the magnitudes of reactions $\alpha_i \leq v_i \leq \beta_i$. The objective function is defined as

$$Z = \sum_i c_i \cdot v_i = \mathbf{c}^T \mathbf{v}, \tag{4.5}$$

where vector $\mathbf{c}$ selects the desired set of fluxes and gives them the weights for optimization. For example, if one wants to find the maximal cellular growth rate, then $\mathbf{c}$ is defined so that it contains each of the requirements of growth in correct relation to others, which can be found out experimentally for example by studying the composition of biomass.

The use of an objective function can be justified from the principle that evolution drives organisms to function optimally. Then the selection of objective function should follow biological reality as well as possible in order to make the results meaningful. Finding such an aim for a complex cell type in a complicated and evolving environment can be close to impossible. For a simple organism, such as a bacterium, the main aim is often to grow as fast as possible. Then, a sensible objective to be maximized is the growth function and results of FBA using this have been found to closely follow experimental results (Edwards et al., 2001).

However, there are conditions and phenotypes where simple growth objective function clearly does not hold. Several different objective functions can be tried in such conditions (Schuetz et al., 2007). It has also been noted that, for example in predicting knock-outs, the straight-forward maximization of the objective function (even if it is correct) might not be ideal due to, e.g., limited evolutionary time (Segre et al., 2002). Also, it is unlikely that the metabolism of a cell can be characterized sufficiently using only the formalism of FBA together with a suitably chosen objective function, as for example variations in gene activities often have effects on metabolism.

Problems with FBA might arise with the alternate optimal solutions, i.e. the possible existence of infinitely many solutions resulting the exactly same optimal value of the objective function (Mahadevan and Schilling, 2003). This can cause problems in situations where something else is of interest than the function being maximized in FBA. For example, in trying to predict hydrogen yield, the maximization of biomass might be possible when hydrogen gets any value within some given constraints and is thus not uniquely determined. An approach in which this is often solved is to first run FBA maximizing biomass production, then fix biomass to this value and use FBA again with objective function being

on the flux of interest.

Flux balance analysis (and constraint-based methods in general) have been shown to perform well when compared to experimental results, see e.g. (Edwards et al., 2001; Famili et al., 2003; Feist et al., 2007) and correct prediction rates of more than 80% have been achieved (Duarte et al., 2004) when predicting viability in knock-out experiments. However, the accuracy of predictions is highly dependent on the coverage and quality of the metabolic reconstruction and how well the phenomenon at hand can be approximated given the assumptions of FBA.

Computationally FBA is easy to solve as it is a simple optimization problem solvable by linear programming. Thus, as a method it is easily implemented and solved on most of the current programming languages (using an external solver if needed). There are also software tools developed that can perform FBA analysis (including FBA-SimVis (Grafahrend-Belau et al., 2009), CycSim (Le Fèvre et al., 2009), BioMet Toolbox (Cvijovic et al., 2010), COBRA toolbox (Schellenberger et al., 2011), and others (Lakshmanan et al., 2014)), some of which also offer visualizations and other analysis possibilities.

Flux balance analysis has been applied for already quite long, e.g., in modelling (Jørgensen et al., 1995; Edwards et al., 2001), predicting drug targets (Folger et al., 2011; Oberhardt et al., 2013), and metabolic engineering (Liao et al., 1996) due to the possibility to aid in redesigning the network for overproduction of a desired substance (Burgard et al., 2003; Pharkya et al., 2004; Bro et al., 2006).

In Publication IV, the accuracy of predicting yields of a specific metabolite (hydrogen) in *E. coli* was studied in anaerobic batch-cultures that are suitable for screening beneficial mutations but challenge in particular the steady-state assumption of FBA as well as the used metabolic reconstruction, which was developed for and tested on mostly aerobic conditions (Feist et al., 2007). The data against which the predictions were compared was obtained by gas chromatography and optical density measurements. Although the performance of FBA in predicting viability of mutants is also in this case rather good (~78%), predicting the yields was found to be much more difficult, with only 33% of the cases with predicted increase in hydrogen production yielding increase experimentally. Reasons for the encountered discrepancies were likely due to imperfect reconstruction and in particular its unsuitability to model anaerobic conditions. Also the differences in alternate metabolic intake (glucose vs. galactose) were found to be not captured by the model, where modelling the phosphotransferase system and its regulation would be needed. Despite the differences in predicted yields, computational predictions are valuable in screening out cases that show no change so that they can be omitted in further investigation.

It is noted that modifications to FBA that may be better suited for knock-out (or perturbed) mutants have been presented. One approach, called minimization

of metabolic adjustment (MOMA) (Segre et al., 2002), does not calculate the optimum value for knock-out mutants but instead a projection of the wild-type optimal value to the feasible space of the knock-out mutant, which in general is not the same as the optimal result obtained via FBA. A closely related approach, named regulatory on/off minimization (ROOM) (Shlomi et al., 2005), minimizes the number of significant alterations of fluxes in the knock-out mutant when compared to the wild-type. Both of these approaches are based on the idea that a knock-out mutant has not had enough time to evolve into optimality, as required by FBA, but instead the flux configuration of the mutant strain lies close to the wild-type one without in general being optimal. Investigating the performance of these models in the same scenario would be interesting in the future.

In Publication V, FBA was used as a part of a model aiming to simulate the interplay between chemotactic movement of a population of *E. coli* cells and substrate concentration in the environment. FBA was used to model the metabolic state of each individual bacterium of the population given the concentration of the substrates in the environment. As the environmental state and consequently the metabolic state of the bacteria evolve over a time-course, the modelling was done by simply dividing the time into intervals and assuming (pseudo-)steady-state within each interval, which is sometimes called dynamic flux balance analysis (Mahadevan et al., 2002). The state of the environment is then updated based on the uptake rates of the bacteria.

Chapter 5

Summary of the results

Identification of interacting entities, such as genes or proteins, and the networks they form is one of the key targets in many biological studies. These networks can be inferred and modelled using several different formalisms, one of which is Bayesian networks. Learning the structure of BNs gets complicated because of the huge model-space even at low numbers of nodes in the network. Therefore it is often done by using some way of approximating or sampling, such as generating a sample from the posterior distribution of the space of network structures as done with Markov Chain Monte Carlo.

One of the main problems related to MCMC is the rate of convergence to the actual target distribution. This can be to some extent tackled with modified proposal distributions but an issue when constructing them in the context of Bayesian networks is that their structure is restricted to directed acyclic graphs (DAGs).

In Publication I, a modification to MCMC for BN structure learning is presented that allows efficient movements to larger than the standard single-step neighbourhoods in the DAG space, which is achieved by using composite proposal distributions consisting of one or more single-step transitions. The method is compared to the standard one that proposes only single-step transitions using a big dataset of about 5,400 datapoints, each containing the states of 11 signalling network proteins measured using flow-cytometry. The standard method is noted to have serious difficulties in convergence while the method presented in Publication I is shown to improve convergence by avoiding the MCMC chains getting stuck to local maxima and increasing computational demands only very little. The possibility to tune the proportions of different length transitions is also demonstrated and it is shown that a mixture of for example one- and two-step transitions seems to produce the best outcome in this application. Thus, the method increases the possibility of finding the correct network structure for a given biological system.

As the causal relationships are usually of ultimate interest in studying bio-

logical systems, it is necessary to introduce interventions to the studied systems. Bayesian networks can readily handle interventional data together with observational data but they can also be used to predict which interventions would be beneficial to perform in order to learn the structure of the network with maximum efficiency. This is called active learning and Publication II investigates how such methods perform in comparison with normal, "non-active" learning, with two different sets of real measurement data. One dataset consists of expression measurements from seven transcription factor genes of a bacterium, and the other dataset is the same signalling protein dataset as used in Publication I. Both test cases show that active learning does on average produce better results in the sense that fewer measurements are needed to get to the same accuracy of results.

However, in these cases with real data the improvements of active versus random learning can be relatively modest, whereas the gap between the two has been wider for most studies where simulated datasets have been used. One reason why the improvement is not so dramatic in the presented cases is that the real underlying biological systems are usually not acyclic and thus cannot fully be modelled using Bayesian networks, whereas the simulated datasets are usually generated from a BN. Another reason is that although the simulated data also has a certain level of uncertainties, the noise is likely much more prevalent in actual measurements.

Publication III presents an active learning method for Bayesian networks, where the measurement data subject to active learning selection is incorporated via structure priors. In addition, other data without active learning selection is included through likelihoods as normally. Since the structure priors define beliefs in different structures and therefore edges in the network, the most natural measurements to be included using them are for example such that somehow measure more directly the interactions between entities, denoted in Publication III as binding data.

In the first test case with simulated gene regulatory networks, binding data would correspond to, e.g., ChIP-seq data, which can be used to quantify binding of the product of a gene to the promoter (or another regulatory element) of a target gene. In the other test case the same signalling protein network as previously was used, in which case the real binding data could come from some way of quantifying protein-protein interactions. Due to lack of suitable measurement datasets, the binding measurements were simulated in Publication III. Again, active learning is shown on average to outperform randomly selecting the measurements to be done. Taken together, Publications II and III demonstrate that active learning methods are potentially very useful in biological research because of their ability to reduce the amount of required experiments, which are often tedious and expensive to carry out.

A computational network model for one of the best-known and most widely-

used bacterium, *Escherichia coli*, is used in Publication IV to predict effects of single gene knock-outs on the amount of hydrogen produced by the bacterium. The used model was a well-developed version based on whole-genome data and the predictions of the metabolic network performance for all different knock-outs available were made using flux balance analysis. Predictions were also made with two other methods: (1) based on an algorithm looking at reactions that compete for precursors of hydrogen and (2) selecting knock-outs manually.

The knock-outs giving highest yields of hydrogen were selected and the corresponding mutants were then ordered and cultivated, and the measured biomass and hydrogen productions were compared to the predicted values. Viability, as determined by the ability to produce biomass, was correct in about 78% of cases. However, increased hydrogen production was observed in only 33% of the cases predicted to have increased yield. One reason for the relatively bad prediction accuracy is that the metabolic model is primarily developed and tested under aerobic conditions whereas the cultivations were in anaerobic conditions. Another reason is that batch cultures do not necessarily attain a steady-state, which is the assumption underlying FBA. Even if the cells at some point in the culture are in a steady-state, the end-point measurements made in Publication IV represent a mixture of all the states and thus might not give a correct picture of the steady-state performance.

Part of the discrepancies between measurements and predictions - in particular for viability predictions, i.e. formation of biomass - can also be explained by difficulties in defining the growth media completely in FBA since the used casamino acids included in the growth medium contain several of the essential amino acids. The metabolic reconstructions are usually not tested and developed in such complex growth media and thus might not correctly take into account the uptakes and usage routes of extracellular amino acids, which may be the reason behind some of the erroneous predictions. The use of metabolic models for making predictions to guide in engineering organisms for production of substances is anyhow a useful tool in evaluating potential usefulness of the mutations and in particular the viability of the mutants.

In Publication V, the simulation of metabolic network of *E. coli* is combined with a model for its chemotactic network and environmental substrate concentration. For a single bacterium, one submodel simulates the chemotactic network response as a function of the extracellular aspartate concentration and the resulting changes in the swimming behaviour, and the other submodel simulates the effect of metabolism on both the external aspartate concentration and growth of the bacterium using flux balance analysis. Even though the bacteria are independent, they depend on each other indirectly by influencing their shared environment by creating concentration gradients which redirect the movements of the (sub)populations.

The modelling is done in a state-based and executable framework, which is

convenient in that for example states of bacteria (running/tumbling) are naturally handled and that the executable components can be easily run independently and created or destroyed dynamically. In Publication V, the modelled population sizes were small (~100 bacteria) and further development is clearly needed.

Chapter 6

Conclusions

Inference of structure of many networks is a task needed in several fields and applications. For example, advanced methods for measuring activities of all the genes of an organism have become available within the last decades. Such datasets in effect represent measurements of thousands of variables at a time and trying to, for example, find the network structure explaining the data is not a trivial problem. In Publication I, an improvement to one specific inference problem has been presented, which improves the performance of the standard method for learning the structure of a Bayesian network. Besides increased rate of convergence, the method is also flexible and allows, e.g., adjustments to be done based on the dataset. The potential next steps could include studying adaptive MCMC schemes for BN structure learning.

Limiting the use of static Bayesian networks is the requirement for acyclicity of the structure, which in biological systems is clearly not justified due to the ubiquitous feed-back loops that have been observed. However, even though one BN structure might not be the biologically correct one, it is still possible to find the interactions with high posterior probabilities. In fact, often one is not necessarily so interested in the maximum a posteriori network but instead it is more useful to calculate posterior probabilities over certain features based on a sample from the posterior distribution, which is the rationale behind MCMC and thus also Publication I. The usability of Bayesian networks is also restricted by the huge size of structure-space, making them unusable for genome scale analyses or necessitating other modifications, such as dividing the problem into manageable-sized subparts, to be used.

In order to guarantee that the inferred relationships between entities of the network are causal, it is necessary to introduce interventions to the system and measure the effects of them. Selecting those interventions and/or measurements that maximally increase our knowledge about the network structure, therefore also making inference as fast and efficient as possible, is a problem tackled by active learning methods. As presented in Publications II and III, active learning

improves learning on average.

For many complex systems it is generally not possible to get enough information using only one data type and several different "views" into the system are instead used to measure various aspects. Therefore, the ability to integrate several different data types into the same inference framework is very beneficial, as done in Publication III. In the same publication, the ability of active learning in the context of Bayesian networks is shown to improve efficiency of structure learning also when the data is included via priors. Further study on the subject could include investigating cost functions that allow suggesting measurements from more than one data type by balancing the cost of performing a measurement and expected benefit from receiving the result of the measurement. Moreover, improvements in the computational efficiency of the methods are necessary in order to allow applications to larger systems. The method could also be updated to utilize the calculation of exact edge posterior probabilities (Koivisto and Sood, 2004).

Biological organisms are increasingly being used in producing substances of interest as well as for degrading unwanted substances. As all wanted metabolic capabilities are rarely available in a single organism or the desired end-products might be consumed by the organism itself, metabolic engineering needs to be performed to modify existing cells. Simplest forms of such modifications are knock-outs and computationally predicting their effect on metabolism would be extremely beneficial. As presented in Publication IV, the state of the methods does not completely allow doing this, highlighting the difficulty of capturing biological complexity in a model. Still, modelling offers an invaluable way to narrow down the list of possibilities in deciding which modifications to make, even though there is plenty of room for improvements.

In the future, it could be interesting to study the predictability under different growth conditions (such as chemostat cultures) and using for example dynamic flux balance analysis given that time-course measurements would be available. Performing the experiments in a much simpler growth medium would also be sensible, which probably would allow the models to fit the experimental data better. Besides FBA, the utility of other methods, such as MOMA (Segre et al., 2002) and ROOM (Shlomi et al., 2005), should in addition be evaluated. None of the methods has been reported to produce best results consistently, so a reasonable test could be to use them all in a kind of ensemble approach to select the most promising candidate mutations. The performance of the methods for predicting more than single-gene deletions or additions would be interesting to study, as usually modifying only single gene might not give considerable improvements.

Often biological systems are studied not as a whole due to the very high complexity. Instead, focus is more on certain subsystems, for which highly developed and accurate models may have been built. An example of such separate

subsystems are metabolism and chemotaxis in *E. coli*. Integration of submodels is of importance because certain responses and behaviours can only arise as the combined action of all of them. The chemotaxis model presented in Publication V is able to integrate both metabolic and signalling network models into an executable model that can give rise to behaviour at population-level as a result of their interplay via the state of environment. The presented approach, Biocharts, also allows easily producing predictions and further hypotheses as the model is basically a program and its subprogram parameters are easily modified. Including some recent advances in stochastic models of chemotaxis has been done after the publication, which allows simulations of much larger populations of bacteria, thus getting closer to numbers in real biological assays.

In the future it would be interesting and rather easy to extend the model for scenarios with multiple substrates for which the bacteria have varying preferences, thereby generating several subpopulations as also shown in experiments. However, limitations set by the incapabilities of metabolic and chemotactic models correctly taking into account uptakes and effects of other substrates is likely a major problem.

Computational analysis and mathematical modelling of biological phenomena are arguably needed more than ever nowadays, and their importance is likely to increase in the future. All the models suffer from being incomplete and are to varying extent unable to make accurate predictions, necessitating further developments. This thesis has addressed some of the issues related to learning and simulation of the models, and has hopefully taken some small steps in the path towards increased applicability of the models.

Bibliography

Adler, J. (1966), "Chemotaxis in Bacteria", *Science*, 153 (3737): 708–716.

Adler, J. (1975), "Chemotaxis in Bacteria", *Annual Review of Biochemistry*, 44 (1): 341–356.

Äijö, T. and Lähdesmäki, H. (2009), "Learning Gene Regulatory Networks from Gene Expression Measurements Using Non-Parametric Molecular Kinetics", *Bioinformatics*, 25 (22): 2937–2944.

Allison, D. B., Cui, X., Page, G. P., and Sabripour, M. (2006), "Microarray Data Analysis: From Disarray to Consolidation and Consensus", *Nature Reviews Genetics*, 7 (1): 55–65.

Andrews, S. S. (2012), "Spatial and Stochastic Cellular Modeling with the Smoldyn Simulator", in: *Bacterial Molecular Networks*, Springer: pp. 519–542.

Andrews, S. S. and Arkin, A. P. (2006), "Simulating Cell Biology", *Current Biology*, 16 (14): R523–R527.

Andrews, S. S. and Bray, D. (2004), "Stochastic Simulation of Chemical Reactions with Spatial Resolution and Single Molecule Detail", *Physical Biology*, 1 (3): 137–151.

Armitage, J. P. (1999), "Bacterial Tactic Responses", *Advances in Microbial Physiology*, 41: 229–289.

Bansal, M., Belcastro, V., Ambesi-Impiombato, A., and Bernardo, D. di (2007), "How to Infer Gene Networks from Expression Profiles", *Molecular Systems Biology*, 3: 78.

Barabasi, A.-L. (2002), *Linked: How Everything Is Connected to Everything Else and What It Means*, Plume Editors.

Barabasi, A.-L. and Oltvai, Z. N. (2004), "Network Biology: Understanding the Cell's Functional Organization", *Nature Reviews Genetics*, 5 (2): 101–113.

Beaumont, M. A. and Rannala, B. (2004), "The Bayesian Revolution in Genetics", *Nature Reviews Genetics*, 5 (4): 251–261.

Berg, H. C. and Brown, D. A. (1972), "Chemotaxis in *Escherichia coli* Analysed by Three-Dimensional Tracking", *Nature*, 239 (5374): 500–504.

Bernard, A. and Hartemink, A. J. (2005), "Informative Structure Priors: Joint Learning of Dynamic Regulatory Networks from Multiple Types of Data.", in: *Pacific Symposium on Biocomputing (PSB05)*, ed. by R. B. Altman, A. K. Dunker, L. Hunter, T. A. Jung, and T. E. Klein, vol. 10: pp. 459–470.

Bordbar, A., Monk, J. M., King, Z. A., and Palsson, B. O. (2014), "Constraint-Based Models Predict Metabolic and Associated Cellular Functions", *Nature Reviews Genetics*, 15 (2): 107–120.

Bradford, J. R., Needham, C. J., Bulpitt, A. J., and Westhead, D. R. (2006), "Insights into Protein-Protein Interfaces Using a Bayesian Network Prediction Method", *Journal of Molecular Biology*, 362 (2): 365–86.

Bray, D., Bourret, R. B., and Simon, M. I. (1993), "Computer Simulation of the Phosphorylation Cascade Controlling Bacterial Chemotaxis.", *Molecular Biology of the Cell*, 4 (5): 469–482.

Bro, C., Regenberg, B., Förster, J., and Nielsen, J. (2006), "In Silico Aided Metabolic Engineering of *Saccharomyces cerevisiae* for Improved Bioethanol Production", *Metabolic Engineering*, 8 (2): 102–111.

Burgard, A. P., Pharkya, P., and Maranas, C. D. (2003), "Optknock: A Bilevel Programming Framework for Identifying Gene Knockout Strategies for Microbial Strain Optimization", *Biotechnology and Bioengineering*, 84 (6): 647–657.

Castelo, R. and Kočka, T. (2003), "On Inclusion-Driven Learning of Bayesian Networks", *The Journal of Machine Learning Research*, 4: 527–574.

Chickering, D. M. (2002), "Learning Equivalence Classes of Bayesian-Network Structures", *The Journal of Machine Learning Research*, 2: 445–498.

Christensen, B. and Nielsen, J. (2000), "Metabolic Network Analysis", in: *Bioanalysis and Biosensors for Bioprocess Monitoring*, Springer: pp. 209–231.

Chubukov, V., Uhr, M., Le Chat, L., Kleijn, R. J., Jules, M., Link, H., Aymerich, S., Stelling, J., and Sauer, U. (2013), "Transcriptional Regulation Is Insufficient to Explain Substrate-Induced Flux Changes in *Bacillus subtilis*", *Molecular Systems Biology*, 9: 709.

Cooper, G. F. (1984), *NESTOR: A Computer-Based Medical Diagnostic Aid That Integrates Causal and Probabilistic Knowledge*. PhD thesis, Stanford, CA: Calif. Univ. Stanford.

Cooper, G. F. and Herskovits, E. (1992), "A Bayesian Method for the Induction of Probabilistic Networks from Data", *Machine Learning*, 9 (4): 309–347.

Cooper, G. F. and Yoo, C. (1999), "Causal Discovery from a Mixture of Experimental and Observational Data", in: *UAI'99 Proceedings of the Fifteenth Uncertainty in Artificial Intelligence*, Morgan Kaufmann, CA: pp. 116–125.

Covert, M. W. and Palsson, B. O. (2003), "Constraints-Based Models: Regulation of Gene Expression Reduces the Steady-State Solution Space", *Journal of Theoretical Biology*, 221 (3): 309–325.

Covert, M. W., Schilling, C. H., Famili, I., Edwards, J. S., Goryanin, I. I., Selkov, E., and Palsson, B. O. (2001a), "Metabolic Modeling of Microbial Strains *in silico*", *Trends in Biochemical Sciences*, 26 (3): 179–186.

Covert, M. W., Schilling, C. H., and Palsson, B. (2001b), "Regulation of Gene Expression in Flux Balance Models of Metabolism", *Journal of Theoretical Biology*, 213 (1): 73–88.

Covert, M. W., Knight, E. M., Reed, J. L., Herrgard, M. J., and Palsson, B. O. (2004), "Integrating High-Throughput and Computational Data Elucidates Bacterial Networks", *Nature*, 429 (6987): 92–96.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. (1999), *Probabilistic Networks and Expert Systems*, New York: Springer.

Cowles, M. K. and Carlin, B. P. (1996), "Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review", *Journal of the American Statistical Association*, 91 (434): 883–904.

Crick, F. H. (1958), "On Protein Synthesis", in: *Symposia of the Society for Experimental Biology*, vol. 12: p. 138.

Cvijovic, M., Olivares-Hernández, R., Agren, R., Dahr, N., Vongsangnak, W., Nookaew, I., Patil, K. R., and Nielsen, J. (2010), "BioMet Toolbox: Genome-Wide Analysis of Metabolism", *Nucleic Acids Research*, 38 (suppl 2): W144–W149.

De Jong, H. (2002), "Modeling and Simulation of Genetic Regulatory Systems: A Literature Review", *Journal of Computational Biology*, 9 (1): 67–103.

Demir, E., Cary, M. P., Paley, S., Fukuda, K., Lemer, C., Vastrik, I., Wu, G., D'Eustachio, P., Schaefer, C., Luciano, J., et al. (2010), "The BioPAX Community Standard for Pathway Data Sharing", *Nature Biotechnology*, 28 (9): 935–942.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood from Incomplete Data Via the EM Algorithm", *Journal of the Royal Statistical Society. Series B (Methodological)*, 39 (1): 1–38.

Diaconis, P. (2009), "The Markov Chain Monte Carlo Revolution", *Bulletin of the American Mathematical Society*, 46 (2): 179–205.

Dojer, N., Bednarz, P., Podsiadlo, A., and Wilczynski, B. (2013), "BNFinder2: Faster Bayesian Network Learning and Bayesian Classification", *Bioinformatics*, 29 (16): 2068–70.

Duarte, N. C., Herrgård, M. J., and Palsson, B. Ø. (2004), "Reconstruction and Validation of *Saccharomyces cerevisiae* IND750, a Fully Compartmentalized Genome-Scale Metabolic Model", *Genome Research*, 14 (7): 1298–1309.

Eaton, D. and Murphy, K. (2007a), "Bayesian Structure Learning Using Dynamic Programming and MCMC", in: *UAI 2007, Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, Vancouver, BC, Canada: Morgan Kaufmann, CA: pp. 101–108.

Eaton, D. and Murphy, K. (2007b), "Exact Bayesian Structure Learning from Uncertain Interventions", in: *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, San Juan, Puerto Rico: pp. 107–114.

Eaton, D. and Murphy, K. (2007c), *BDAGL: Bayesian DAG Learning*, `http://www.cs.ubc.ca/~murphyk/Software/BDAGL/`.

Eddy, S. R. (2004), "What Is Bayesian Statistics?", *Nature Biotechnology*, 22 (9): 1177–1178.

Edwards, J. S., Ibarra, R. U., and Palsson, B. O. (2001), "*In silico* Predictions of *Escherichia coli* Metabolic Capabilities Are Consistent with Experimental Data", *Nature Biotechnology*, 19 (2): 125–130.

Edwards, J. S. and Palsson, B. O. (1999), "Systems Properties of the *Haemophilus influenzae* Rd Metabolic Genotype", *Journal of Biological Chemistry*, 274 (25): 17410–17416.

Elf, J. and Ehrenberg, M. (2004), "Spontaneous Separation of Bi-Stable Biochemical Systems into Spatial Domains of Opposite Phases", *Systems Biology*, 1 (2): 230–236.

Ellis, B. and Wong, W. H. (2008), "Learning Causal Bayesian Network Structures from Experimental Data", *Journal of the American Statistical Association*, 103 (482): 778–789.

Emonet, T., Macal, C. M., North, M. J., Wickersham, C. E., and Cluzel, P. (2005), "AgentCell: A Digital Single-Cell Assay for Bacterial Chemotaxis", *Bioinformatics*, 21 (11): 2714–2721.

Famili, I., Förster, J., Nielsen, J., and Palsson, B. O. (2003), "*Saccharomyces cerevisiae* Phenotypes Can Be Predicted by Using Constraint-Based Analysis of a Genome-Scale Reconstructed Metabolic Network", *Proceedings of the National Academy of Sciences*, 100 (23): 13134–13139.

Fang, Z. and Cui, X. (2011), "Design and Validation Issues in RNA-seq Experiments", *Briefings in Bioinformatics*, bbr004.

Feist, A. M., Henry, C. S., Reed, J. L., Krummenacker, M., Joyce, A. R., Karp, P. D., Broadbelt, L. J., Hatzimanikatis, V., and Palsson, B. Ø. (2007), "A Genome-Scale Metabolic Reconstruction for *Escherichia coli* K-12 MG1655 That Accounts for 1260 ORFs and Thermodynamic Information", *Molecular Systems Biology*, 3: 121.

Feist, A. M., Herrgård, M. J., Thiele, I., Reed, J. L., and Palsson, B. Ø. (2009), "Reconstruction of Biochemical Networks in Microorganisms", *Nature Reviews Microbiology*, 7 (2): 129–143.

Fisher, J. and Henzinger, T. A. (2007), "Executable Cell Biology", *Nature Biotechnology*, 25 (11): 1239–1249.

Folger, O., Jerby, L., Frezza, C., Gottlieb, E., Ruppin, E., and Shlomi, T. (2011), "Predicting Selective Drug Targets in Cancer Through Metabolic Networks", *Molecular Systems Biology*, 7: 501.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000), "Using Bayesian Networks to Analyze Expression Data", *Journal of Computational Biology*, 7 (3-4): 601–620.

Friedman, N. (2004), "Inferring Cellular Networks Using Probabilistic Graphical Models", *Science*, 303 (5659): 799–805.

Friedman, N. and Koller, D. (2003), "Being Bayesian About Network Structure – a Bayesian Approach to Structure Discovery in Bayesian Networks", *Machine Learning*, 50 (1-2): 95–125.

Friedman, N., Murphy, K., and Russell, S. (1998), "Learning the Structure of Dynamic Probabilistic Networks", in: *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, UAI'98, Madison, Wisconsin: Morgan Kaufmann Publishers Inc.: pp. 139–147.

Gagneur, J. and Klamt, S. (2004), "Computation of Elementary Modes: A Unifying Framework and the New Binary Approach", *BMC Bioinformatics*, 5 (1): 175.

Geiger, D. and Heckerman, D. (1998), "A Characterization of the Bivariate Wishart Distribution", *Probability and Mathematical Statistics*, 18 (1): 119–131.

Ghahramani, Z. (1998), "Learning Dynamic Bayesian Networks", in: *Adaptive Processing of Sequences and Data Structures*, Springer: pp. 168–197.

Gillespie, D. T. (1976), "A General Method for Numerically Simulating the Stochastic Time Evolution of Coupled Chemical Reactions", *Journal of Computational Physics*, 22 (4): 403–434.

Gillespie, D. T. (1977), "Exact Stochastic Simulation of Coupled Chemical Reactions", *The Journal of Physical Chemistry*, 81 (25): 2340–2361.

Giudici, P. and Castelo, R. (2003), "Improving Markov Chain Monte Carlo Model Search for Data Mining", *Machine Learning*, 50 (1-2): 127–158.

Good, I. J. (1961), "A Causal Calculus (I)", *The British Journal for the Philosophy of Science*, 11 (44): 305–318.

Grafahrend-Belau, E., Klukas, C., Junker, B. H., and Schreiber, F. (2009), "FBA-SimVis: Interactive Visualization of Constraint-Based Metabolic Models", *Bioinformatics*, 25 (20): 2755–2757.

Greenbaum, D., Colangelo, C., Williams, K., and Gerstein, M. (2003), "Comparing Protein Abundance and mRNA Expression Levels on a Genomic Scale", *Genome Biology*, 4 (9): 117.

Grzegorczyk, M. and Husmeier, D. (2008), "Improving the Structure MCMC Sampler for Bayesian Networks by Introducing a New Edge Reversal Move", *Machine Learning*, 71 (2-3): 265–305.

Haario, H., Laine, M., Mira, A., and Saksman, E. (2006), "DRAM: Efficient Adaptive MCMC", *Statistics and Computing*, 16: 339–354.

Harel, D. (1987), "Statecharts: A Visual Formalism for Complex Systems", *Science of Computer Programming*, 8 (3): 231–274.

Harel, D. and Gery, E. (1996), "Executable Object Modeling with Statecharts", in: *Proceedings of the 18th International Conference on Software Engineering*, IEEE Computer Society: pp. 246–257.

Hartemink, A. J., Gifford, D. K., Jaakkola, T. S., and Young, R. A. (2002), "Combining Location and Expression Data for Principled Discovery of Genetic Regulatory Network Models", in: *Pacific Symposium on Biocomputing*, vol. 7: pp. 437–449.

Hartemink, A. J. (2001), *Principled Computational Methods for the Validation and Discovery of Genetic Regulatory Networks*. PhD thesis, MIT.

Hastings, W. K. (1970), "Monte Carlo Sampling Methods Using Markov Chains and Their Applications", *Biometrika*, 57 (1): 97–109.

Heckerman, D. (1998), "A Tutorial on Learning with Bayesian Networks", English, in: *Learning in Graphical Models*, ed. by M. Jordan, vol. 89, NATO ASI Series, Springer Netherlands: pp. 301–354.

Heckerman, D., Geiger, D., and Chickering, D. M. (1995a), "Learning Bayesian Networks: The Combination of Knowledge and Statistical Data", *Machine Learning*, 20 (3): 197–243.

Heckerman, D., Mamdani, A., and Wellman, M. P. (1995b), "Real-World Applications of Bayesian Networks", *Communications of the ACM*, 38 (3): 24–26.

Herrgård, M. J., Swainston, N., Dobson, P., Dunn, W. B., Arga, K. Y., Arvas, M., Blüthgen, N., Borger, S., Costenoble, R., Heinemann, M., et al. (2008), "A Consensus Yeast Metabolic Network Reconstruction Obtained from a Community Approach to Systems Biology", *Nature Biotechnology*, 26 (10): 1155–1160.

Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., Arkin, A. P., Bornstein, B. J., Bray, D., Cornish-Bowden, A., et al. (2003), "The Systems Biology Markup Language (SBML): A Medium for Representation and Exchange of Biochemical Network Models", *Bioinformatics*, 19 (4): 524–531.

Husmeier, D. (2005), *Probabilistic Modeling in Bioinformatics and Medical Informatics*. In: ed. by D. Husmeier, R. Dybowski, and S. Roberts, Springer Berlin Heidelberg, chap. Introduction to Learning Bayesian Networks from Data: pp. 17–57.

InSilicoOrganisms, *Available Predictive Genome-Scale Metabolic Network Reconstructions*, `http://gcrg.ucsd.edu/InSilicoOrganisms/OtherOrganisms`.

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., Emili, A., Snyder, M., Greenblatt, J. F., and Gerstein, M. (2003), "A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data", *Science*, 302 (5644): 449–53.

Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007), "Genome-Wide Mapping of in Vivo Protein-DNA Interactions", *Science*, 316 (5830): 1497–1502.

Jørgensen, H., Nielsen, J., Villadsen, J., and Møllgaard, H. (1995), "Metabolic Flux Distributions in *Penicillium chrysogenum* During Fed-Batch Cultivations", *Biotechnology and Bioengineering*, 46 (2): 117–131.

Karlebach, G. and Shamir, R. (2008), "Modelling and Analysis of Gene Regulatory Networks", *Nature Reviews Molecular Cell Biology*, 9 (10): 770–780.

Karp, P. D., Paley, S. M., Krummenacker, M., Latendresse, M., Dale, J. M., Lee, T. J., Kaipa, P., Gilham, F., Spaulding, A., Popescu, L., et al. (2010), "Pathway Tools Version 13.0: Integrated Software for Pathway/genome Informatics and Systems Biology", *Briefings in Bioinformatics*, 11 (1): 40–79.

Kauffman, K. J., Prakash, P., and Edwards, J. S. (2003), "Advances in Flux Balance Analysis", *Current Opinion in Biotechnology*, 14 (5): 491–496.

Kauffman, S. A. (1969), "Metabolic Stability and Epigenesis in Randomly Constructed Genetic Nets", *Journal of Theoretical Biology*, 22 (3): 437–467.

Kitano, H., Funahashi, A., Matsuoka, Y., and Oda, K. (2005), "Using Process Diagrams for the Graphical Representation of Biological Networks", *Nature Biotechnology*, 23 (8): 961–966.

Klamt, S. and Stelling, J. (2003), "Two Approaches for Metabolic Pathway Analysis?", *TRENDS in Biotechnology*, 21 (2): 64–69.

Kohn, K. W., Aladjem, M. I., Weinstein, J. N., and Pommier, Y. (2006), "Molecular Interaction Maps of Bioregulatory Networks: A General Rubric for Systems Biology", *Molecular Biology of the Cell*, 17 (1): 1–13.

Koivisto, M. and Sood, K. (2004), "Exact Bayesian Structure Discovery in Bayesian Networks", *Journal of Machine Learning Research*, 5: 549–573.

Kugler, H. (2013), "Biocharts: Unifying Biological Hypotheses with Models and Experiments", in: *EScience (eScience), 2013 IEEE 9th International Conference On*, IEEE: pp. 317–325.

Lähdesmäki, H. and Shmulevich, I. (2008), "Learning the Structure of Dynamic Bayesian Networks from Time Series and Steady State Measurements", *Machine Learning*, 71 (2-3): 185–217.

Lakshmanan, M., Koh, G., Chung, B. K., and Lee, D.-Y. (2014), "Software Applications for Flux Balance Analysis", *Briefings in Bioinformatics*, 15 (1): 108–122.

Larhlimi, A. and Bockmayr, A. (2009), "A New Constraint-Based Description of the Steady-State Flux Cone of Metabolic Networks", *Discrete Applied Mathematics*, 157 (10): 2257–2266.

Le Fèvre, F., Smidtas, S., Combe, C, Durot, M, d'Alché-Buc, F., and Schachter, V (2009), "CycSim—an Online Tool for Exploring and Experimenting with Genome-Scale Metabolic Models", *Bioinformatics*, 25 (15): 1987–1988.

Le Novere, N. and Shimizu, T. S. (2001), "STOCHSIM: Modelling of Stochastic Biomolecular Processes", *Bioinformatics*, 17 (6): 575–576.

Le Novere, N., Hucka, M., Mi, H., Moodie, S., Schreiber, F., Sorokin, A., Demir, E., Wegner, K., Aladjem, M. I., Wimalaratne, S. M., et al. (2009), "The Systems Biology Graphical Notation", *Nature Biotechnology*, 27 (8): 735–741.

Lerman, J. A., Chang, R. L., Hyduke, D. R., Palsson, B. Ø., et al. (2013), "Genome-Scale Models of Metabolism and Gene Expression Extend and Refine Growth Phenotype Prediction", *Molecular Systems Biology*, 9: 693.

Lewis, T. G. (2011), *Network Science: Theory and Applications*, John Wiley & Sons.

Liao, J. C., Hou, S.-Y., and Chao, Y.-P. (1996), "Pathway Analysis, Engineering, and Physiological Considerations for Redirecting Central Metabolism", *Biotechnology and Bioengineering*, 52 (1): 129–140.

Lindley, D. V. (2000), "The Philosophy of Statistics", *Journal of the Royal Statistical Society: Series D*, 49 (3): 293–337.

Lloyd, C. M., Halstead, M. D., and Nielsen, P. F. (2004), "CellML: Its Future, Present and Past", *Progress in Biophysics and Molecular Biology*, 85 (2): 433–450.

Loew, L. M. and Schaff, J. C. (2001), "The Virtual Cell: A Software Environment for Computational Cell Biology", *TRENDS in Biotechnology*, 19 (10): 401–406.

Lok, L. and Brent, R. (2005), "Automatic Generation of Cellular Reaction Networks with Moleculizer 1.0", *Nature Biotechnology*, 23 (1): 131–136.

Machado, D. and Herrgård, M. (2014), "Systematic Evaluation of Methods for Integration of Transcriptomic Data into Constraint-Based Models of Metabolism", *PLoS Computational Biology*, 10 (4): e1003580.

Madigan, D. and York, J. (1995), "Bayesian Graphical Models for Discrete Data", *International Statistical Review*, 63 (2): 215–232.

Mahadevan, R and Schilling, C. (2003), "The Effects of Alternate Optimal Solutions in Constraint-Based Genome-Scale Metabolic Models", *Metabolic Engineering*, 5 (4): 264–276.

Mahadevan, R., Edwards, J. S., and Doyle III, F. J. (2002), "Dynamic Flux Balance Analysis of Diauxic Growth in *Escherichia coli*", *Biophysical Journal*, 83 (3): 1331–1340.

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R. D., and Califano, A. (2006), "ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context", *BMC Bioinformatics*, 7 (Suppl 1): S7.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953), "Equation of State Calculations by Fast Computing Machines", *The Journal of Chemical Physics*, 21 (6): 1087–1092.

Meyers, S. and Friedland, P. (1984), "Knowledge-Based Simulation of Genetic Regulation in Bacteriophage Lambda", *Nucleic Acids Research*, 12 (1Part1): 1–9.

Mo, M. L., Palsson, B. Ø., and Herrgård, M. J. (2009), "Connecting Extracellular Metabolomic Measurements to Intracellular Flux States in Yeast", *BMC Systems Biology*, 3 (1): 37.

Mogilner, A., Wollman, R., and Marshall, W. F. (2006), "Quantitative Modeling in Cell Biology: What Is It Good For?", *Developmental Cell*, 11 (3): 279 –287.

Monk, J., Nogales, J., and Palsson, B. O. (2014), "Optimizing Genome-Scale Network Reconstructions", *Nature Biotechnology*, 32 (5): 447–452.

Moore, A. W. and Wong, W.-K. (2003), "Optimal Reinsertion: A New Search Operator for Accelerated and More Accurate Bayesian Network Structure Learning", in: *Proceedings of the Twentieth International Conference on Machine Learning*, ed. by T. Fawcett and N. Mishra, AAAI Press: pp. 552–559.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008), "Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq", *Nature Methods*, 5 (7): 621–628.

Morton-Firth, C. J. and Bray, D. (1998), "Predicting Temporal Fluctuations in an Intracellular Signalling Pathway", *Journal of Theoretical Biology*, 192 (1): 117–128.

Morton-Firth, C. J., Shimizu, T. S., and Bray, D. (1999), "A Free-Energy-Based Stochastic Simulation of the Tar Receptor Complex", *Journal of Molecular Biology*, 286 (4): 1059–1074.

Mukherjee, S. and Speed, T. P. (2008), "Network Inference Using Informative Priors", *Proceedings of the National Academy of Sciences*, 105 (38): 14313–14318.

Murata, T. (1989), "Petri Nets: Properties, Analysis and Applications", *Proceedings of the IEEE*, 77 (4): 541–580.

Murphy, K. (2001a), *Active Learning of Causal Bayes Net Structure*, tech. rep., Department of Computer Science, U.C. Berkeley.

Murphy, K. (2001b), "The Bayes Net Toolbox for Matlab", *Computing Science and Statistics*, 33 (2): 1024–1034.

Murphy, K. (2013), *Software Packages for Graphical Models*, `http://people.cs.ubc.ca/~murphyk/Software/bnsoft.html`.

Murphy, K. and Mian, S. (1999), *Modelling Gene Expression Data Using Dynamic Bayesian Networks*, tech. rep. 104, University of California, Berkeley, CA: Computer Science Division.

Myers, C. L., Robson, D., Wible, A., Hibbs, M. A., Chiriac, C., Theesfeld, C. L., Dolinski, K., and Troyanskaya, O. G. (2005), "Discovery of Biological Networks from Diverse Functional Genomic Data", *Genome Biology*, 6 (13): R114.

Niinimäki, T. and Koivisto, M. (2013), "Annealed Importance Sampling for Structure Learning in Bayesian Networks", in: *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, AAAI Press: pp. 1579–1585.

Noor, A., Serpedin, E., Nounou, M., Nounou, H., Mohamed, N., and Chouchane, L. (2013), "An Overview of the Statistical Methods Used for Inferring Gene Regulatory Networks and Protein-Protein Interaction Networks", *Advances in Bioinformatics*, 2013.

Oberhardt, M. A., Yizhak, K., and Ruppin, E. (2013), "Metabolically Re-Modeling the Drug Pipeline", *Current Opinion in Pharmacology*, 13 (5): 778–785.

Orth, J. D., Conrad, T. M., Na, J., Lerman, J. A., Nam, H., Feist, A. M., and Palsson, B. Ø. (2011), "A Comprehensive Genome-Scale Reconstruction of *Escherichia coli* Metabolism", *Molecular Systems Biology*, 7: 535.

Pabinger, S., Snajder, R., Hardiman, T., Willi, M., Dander, A., and Trajanoski, Z. (2014), "MEMOSys 2.0: An Update of the Bioinformatics Database for Genome-Scale Models and Genomic Data", *Database*, 2014: bau004.

Pearl, J. (1985), "Bayesian Networks: A Model of Self-Activated Memory for Evidential Reasoning", in: *Proceedings of the 7th Conference of the Cognitive Science Society, University of California, Irvine*: pp. 329–334.

Pearl, J. (2009), *Causality: Models, Reasoning and Inference*, 2nd edition, Cambridge University Press.

Pease, A. C., Solas, D., Sullivan, E. J., Cronin, M. T., Holmes, C. P., and Fodor, S. (1994), "Light-generated Oligonucleotide Arrays for Rapid DNA Sequence Analysis", *Proceedings of the National Academy of Sciences*, 91 (11): 5022–5026.

Pharkya, P., Burgard, A. P., and Maranas, C. D. (2004), "OptStrain: A Computational Framework for Redesign of Microbial Production Systems", *Genome Research*, 14 (11): 2367–2376.

Pournara, I. and Wernisch, L. (2004), "Reconstruction of Gene Networks Using Bayesian Learning and Manipulation Experiments", *Bioinformatics*, 20 (17): 2934–2942.

Quach, M., Brunel, N., and d'Alché-Buc, F. (2007), "Estimating Parameters and Hidden Variables in Non-Linear State-Space Models Based on ODEs for Biological Networks Inference", *Bioinformatics*, 23 (23): 3209–3216.

Raman, K. and Chandra, N. (2009), "Flux Balance Analysis of Biological Systems: Applications and Challenges", *Briefings in Bioinformatics*, 10 (4): 435–449.

Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., et al. (2000), "Genome-Wide Location and Function of DNA Binding Proteins", *Science*, 290 (5500): 2306–2309.

Rezola, A., Figueiredo, L. F. de, Brock, M, Pey, J., Podhorski, A., Wittmann, C., Schuster, S., Bockmayr, A., and Planes, F. J. (2011), "Exploring Metabolic Pathways in Genome-Scale Networks Via Generating Flux Modes", *Bioinformatics*, 27 (4): 534–540.

Robinson, R. W. (1977), "Counting Unlabeled Acyclic Digraphs", in: *Proceedings of the Fifth Australian Conference on Combinatorial Mathematics*, ed. by C. Little, Springer, Berlin.

Rousseau, W. F. (1968), *A Method for Computing Probabilities in Complex Situations*. Tech. rep., Center for Systems Research, Stanford University.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. (2005), "Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data", *Science*, 308 (5721): 523–529.

Schellenberger, J., Que, R., Fleming, R. M., Thiele, I., Orth, J. D., Feist, A. M., Zielinski, D. C., Bordbar, A., Lewis, N. E., Rahmanian, S., et al. (2011), "Quantitative Prediction of Cellular Metabolism with Constraint-Based Models: The COBRA Toolbox V2. 0", *Nature Protocols*, 6 (9): 1290–1307.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995), "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray", *Science*, 270 (5235): 467–470.

Schilling, C. H., Letscher, D., and Palsson, B. Ø. (2000), "Theory for the Systemic Definition of Metabolic Pathways and Their Use in Interpreting Metabolic Function from a Pathway-Oriented Perspective", *Journal of Theoretical Biology*, 203 (3): 229–248.

Schilling, C. H., Schuster, S., Palsson, B. O., and Heinrich, R. (1999), "Metabolic Pathway Analysis: Basic Concepts and Scientific Applications in the Post-Genomic Era", *Biotechnology Progress*, 15 (3): 296–303.

Schuetz, R., Kuepfer, L., and Sauer, U. (2007), "Systematic Evaluation of Objective Functions for Predicting Intracellular Fluxes in *Escherichia coli*", *Molecular Systems Biology*, 3: 119.

Schuster, S., Fell, D. A., and Dandekar, T. (2000), "A General Definition of Metabolic Pathways Useful for Systematic Organization and Analysis of Complex Metabolic Networks", *Nature Biotechnology*, 18 (3): 326–332.

Schuster, S., Hilgetag, C., Woods, J. H., and Fell, D. A. (2002), "Reaction Routes in Biochemical Reaction Systems: Algebraic Properties, Validated Calculation Procedure and Example from Nucleotide Metabolism", *Journal of Mathematical Biology*, 45: 153–181.

Segal, E., Barash, Y., Simon, I., Friedman, N., and Koller, D. (2002), "From Promoter Sequence to Expression: A Probabilistic Framework", in: *Proceedings of the Sixth Annual International Conference on Computational Biology*, ACM: pp. 263–272.

Segre, D., Vitkup, D., and Church, G. M. (2002), "Analysis of Optimality in Natural and Perturbed Metabolic Networks", *Proceedings of the National Academy of Sciences*, 99 (23): 15112–15117.

SEQC/MAQC-III Consortium (2014), "A Comprehensive Assessment of RNA-seq Accuracy, Reproducibility and Information Content by the Sequencing Quality Control Consortium", *Nature Biotechnology*, 32 (9): 903–914.

Settles, B. (2009), *Active Learning Literature Survey*, Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Setty, Y., Cohen, I. R., Dor, Y., and Harel, D. (2008), "Four-Dimensional Realistic Modeling of Pancreatic Organogenesis", *Proceedings of the National Academy of Sciences*, 105 (51): 20374–20379.

Setty, Y., Dalfó, D., Korta, D. Z., Hubbard, E. J. A., and Kugler, H. (2012), "A Model of Stem Cell Population Dynamics: In Silico Analysis and in Vivo Validation", *Development*, 139: 47–56.

Shlomi, T., Berkman, O., and Ruppin, E. (2005), "Regulatory On/Off Minimization of Metabolic Flux Changes After Genetic Perturbations", *Proceedings of the National Academy of Sciences*, 102 (21): 7695–7700.

Shmulevich, I., Dougherty, E. R., Kim, S., and Zhang, W. (2002), "Probabilistic Boolean Networks: A Rule-Based Uncertainty Model for Gene Regulatory Networks", *Bioinformatics*, 18 (2): 261–274.

Slepchenko, B. M. and Loew, L. M. (2010), "Use of Virtual Cell in Studies of Cellular Dynamics", *International Review of Cell and Molecular Biology*, 283: 1–56.

Smith, V. A., Yu, J., Smulders, T. V., Hartemink, A. J., and Jarvis, E. D. (2006), "Computational Inference of Neural Information Flow Networks", *PLoS Computational Biology*, 2 (11): e161.

Tamada, Y., Kim, S., Bannai, H., Imoto, S., Tashiro, K., Kuhara, S., and Miyano, S. (2003), "Estimating Gene Networks from Gene Expression Data by Combining Bayesian Network Model with Promoter Element Detection", *Bioinformatics*, 19 Suppl 2: ii227–ii236.

Thar, R. and Kühl, M. (2003), "Bacteria Are Not Too Small for Spatial Sensing of Chemical Gradients: An Experimental Evidence", *Proceedings of the National Academy of Sciences*, 100 (10): 5748–5753.

Thiele, I. and Palsson, B. Ø. (2010), "A Protocol for Generating a High-Quality Genome-Scale Metabolic Reconstruction", *Nature Protocols*, 5 (1): 93–121.

Thiele, I., Swainston, N., Fleming, R. M., Hoppe, A., Sahoo, S., Aurich, M. K., Haraldsdottir, H., Mo, M. L., Rolfsson, O., Stobbe, M. D., et al. (2013), "A Community-Driven Global Reconstruction of Human Metabolism", *Nature Biotechnology*, 31 (5): 419–425.

Thorleifsson, S. G. and Thiele, I. (2011), "rBioNet: A COBRA Toolbox Extension for Reconstructing High-Quality Biochemical Networks", *Bioinformatics*, 27 (14): 2009–2010.

Tindall, M. J., Maini, P. K., Porter, S. L., and Armitage, J. P. (2008a), "Overview of Mathematical Approaches Used to Model Bacterial Chemotaxis II: Bacterial Populations", *Bulletin of Mathematical Biology*, 70 (6): 1570–1607.

Tindall, M., Porter, S., Maini, P., Gaglia, G, and Armitage, J. (2008b), "Overview of Mathematical Approaches Used to Model Bacterial Chemotaxis I: The Single Cell", *Bulletin of Mathematical Biology*, 70 (6): 1525–1569.

Tong, S. and Koller, D. (2001), "Active Learning for Structure in Bayesian Networks", in: *IJCAI'01 Proceedings of the 17th International Joint Conference on Artificial Intelligence*, vol. 2, Seattle, Washington, USA: Morgan Kaufmann: pp. 863–869.

Troyanskaya, O. G., Dolinski, K., Owen, A. B., Altman, R. B., and Botstein, D. (2003), "A Bayesian Framework for Combining Heterogeneous Data Sources for Gene Function Prediction (in Saccharomyces Cerevisiae)", *Proceedings of the National Academy of Sciences*, 100 (14): 8348–53.

Varma, A. and Palsson, B. O. (1994), "Metabolic Flux Balancing: Basic Concepts, Scientific and Practical Use", *Nature Biotechnology*, 12 (10): 994–998.

Verma, T. and Pearl, J. (1991), "Equivalence and Synthesis of Causal Models", in: *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, UAI '90, New York, NY, USA: Elsevier Science Inc.: pp. 255–270.

Vogel, C. and Marcotte, E. M. (2012), "Insights into the Regulation of Protein Abundance from Proteomic and Transcriptomic Analyses", *Nature Reviews Genetics*, 13 (4): 227–232.

Wadhams, G. H. and Armitage, J. P. (2004), "Making Sense of It All: Bacterial Chemotaxis", *Nature Reviews Molecular Cell Biology*, 5 (12): 1024–1037.

Werhli, A. V. and Husmeier, D. (2007), "Reconstructing Gene Regulatory Networks with Bayesian Networks by Combining Expression Data with Mul-

tiple Sources of Prior Knowledge", *Statistical Applications in Genetics and Molecular Biology*, 6 (1).

Werhli, A. V., Grzegorczyk, M., and Husmeier, D. (2006), "Comparative Evaluation of Reverse Engineering Gene Regulatory Networks with Relevance Networks, Graphical Gaussian Models and Bayesian Networks", *Bioinformatics*, 22 (20): 2523–2531.

Wiechert, W. (2001), "^{13}C Metabolic Flux Analysis", *Metabolic Engineering*, 3 (3): 195–206.

Wilczyński, B. and Dojer, N. (2009), "BNFinder: Exact and Efficient Method for Learning Bayesian Networks", *Bioinformatics*, 25 (2): 286–7.

Wright, S. (1921), "Correlation and Causation", *Journal of Agricultural Research*, 20 (7): 557–585.

Wu, F.-X., Zhang, W.-J., and Kusalik, A. J. (2004), "Modeling Gene Expression from Microarray Expression Data with State-Space Equations.", in: *Pacific Symposium on Biocomputing*, vol. 9: pp. 581–592.

Zhao, W., Serpedin, E., and Dougherty, E. R. (2006), "Inferring Gene Regulatory Networks from Time Series Data Using the Minimum Description Length Principle", *Bioinformatics*, 22 (17): 2129–2135.

Zhu, X., Gerstein, M., and Snyder, M. (2007), "Getting Connected: Analysis and Principles of Biological Networks", *Genes & Development*, 21 (9): 1010–1024.

Publications

Publication I

A. Larjo and H. Lähdesmäki, "Using multi-step proposal distribution for improved MCMC convergence in Bayesian network structure learning," *submitted to EURASIP Journal on Bioinformatics and Systems Biology*, 2014.

Publication II

A. Larjo, H. Lähdesmäki, M. Facciotti, N. Baliga, I. Shmulevich, and O. Yli-Harja, "Active learning of Bayesian network structure in a realistic setting," in *Fifth International Workshop on Computational Systems Biology (WCSB 2008)*, Leipzig, Germany, June 11-13, 2008, pp. 85-88.

Publication III

A. Larjo and H. Lähdesmäki, "Active learning for Bayesian network models of biological networks using structure priors," in *IEEE International Workshop on Genomic Signal Processing and Statistics*, Houston, TX, USA, November 17-19, 2013, pp. 78-81.

Publication IV

J.J. Seppälä*, A. Larjo*, T. Aho, O. Yli-Harja, M.T. Karp, and V. Santala, "Prospecting hydrogen production of Escherichia coli by metabolic network modeling," *International Journal of Hydrogen Energy*, vol. 38, no. 27, pp. 11780-11789, 2013.

Publication V

H. Kugler*, A. Larjo*, and D. Harel, "Biocharts: A visual formalism for complex biological systems," *Journal of the Royal Society Interface*, vol. 7, no. 48, pp. 1015–1024, 2010.

Tampereen teknillinen yliopisto
PL 527
33101 Tampere

Tampere University of Technology
P.O.B. 527
FI-33101 Tampere, Finland

ISBN 978-952-15-3483-6
ISSN 1459-2045