Model-checking based Approaches to Parameter Estimation of Gene Regulatory Networks

Andrzej Mizera*, Jun Pang*†, Qixia Yuan*

* Faculty of Science, Technology and Communication, University of Luxembourg
† Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg
6 rue Richard Coudenhove-Kalergi, L-1359 Luxembourg
Emails: firstname.lastname@uni.lu

***Abstract*—The expression of genes is a fundamental process in living cells, both eukaryotic and prokaryotic. The regulation of gene expression is achieved via sophisticated networks of interactions between DNA, RNA, proteins, and small chemical compounds. The qualitative and quantitative characterisation of interactions between genes is one of the major current research targets in systems biology. In this PhD research project, we view gene regulatory networks as Markov chains, resulting from popular formalisation frameworks such as Dynamic Bayesian Networks and Probabilistic Boolean Networks. This will allow us to reason about both the structure and strength of gene interactions. Our goal is to develop new algorithms and tools, which are tailored for the modelling and analysis of gene regulatory networks, by exploring model checking techniques that have been developed and widely used in computer science. More specifically, we will combine model checking techniques with sampling and optimisation methods from the literature to derive new techniques to solve the parameter estimation problem of Markov models of gene regulatory networks.**

I. Introduction

Systems biology is a new, emerging and rapidly developing, multidisciplinary research field. The topics associated with systems biology attract interest of researchers having their background in a wide range of field of expertise, e.g., biology, chemistry, computer science, mathematics, physics or engineering. Systems biology aims to study biological systems from a holistic perspective, with the goal to provide a comprehensive, system-level understanding of cellular behaviour. The research in this field involves identification, modelling and analysis of biochemical networks (e.g., metabolic pathways, regulatory networks or signal transduction networks), in close linkage to experiments with the focus on understanding the system's structure and dynamics. Such comprehensive approach enables the capturing of complex properties of a system such as robustness, emergence or adaptation, which are ubiquitous features of biological systems [1].

Computer science plays a prominent role in the field of systems biology. One of the main reasons is that the key concepts in systems biology, such as component, network, robustness, efficiency, regulation, control, signalling, synchronisation, parallelism, etc., have been studied for a long time in computer science (albeit from different perspectives). A key contribution brought to systems biology by computer science is the formal means for manipulation, analysis, and reasoning about system-level concepts and structures. For example, formal system specifications, control design, mathematical modelling belong to the mainstream techniques utilised in systems biology [2]. Over the last decade concepts and approaches from computer science, and software engineering in particular, have started to penetrate the field of systems biology in an increasing pace. In this project, we focus on the application of model-checking, which is a mathematically-based technique for the specification, development and verification of computer systems, to the analysis of biological systems. More specifically, our goal in this project is to develop and apply model-checking algorithms and tools which are tailored for the modelling and analysis of biological systems.

Literature review. The expression of genes is a fundamental process in living cells, both eukaryotic and prokaryotic. It is central to the control of cellular processes, such as cell differentiation, cell division, and the response of a cell to environmental signals. The regulation of gene expression is achieved via sophisticated networks of interactions between DNA, RNA, proteins, and small chemical compounds. These interactions form positive and negative feedback loops. Their orchestrated operation leads to a dynamic, complex behaviour of the system. Obtaining an intuitive understanding of this behaviour is usually infeasible. Hence, utilisation of formal methods and computer tools for modelling and simulation is indispensable for acquiring insight into the dynamics of the system. Particularly, the qualitative and quantitative characterisation of interactions between genes is one of the major current research targets in systems biology. A number of formalisms has been employed to study genetic regulatory systems. These formalisms include Directed Graphs, Bayesian Networks, Boolean Networks and Probabilistic Boolean Networks, ordinary and partial differential equations, qualitative differential equations, stochastic equations, and rule-based formalisms [3], [4].

The frameworks of Dynamic Bayesian Networks and Probabilistic Boolean Networks are broadly applied to represent gene regulatory networks [5], [6], [7], [8], [9], [10], [11]. In both cases the resulting network can be recast as a Markov chain. This allows the reasoning on both the structure and strength of gene interactions. The dynamics of the resulting

system can be investigated by applying techniques originating from the theory of Markov chains. One class of standard techniques relies on Monte Carlo simulations. However, simulation covers only a subset of the behaviour of the investigated system as covering the complete behaviour would be extremely costly in terms of processing time and computational demand. In contrast, model-checking [12] can provide a formal validation by exploring all possible behaviours of system models and grant higher precision. Due to efficient algorithms and state-space reduction techniques, model-checking can analyse large systems exhibiting complex behaviours, and this process is typically supported by computer tools. As biological systems usually have complicated stochastic behaviours, the stochastic model-checking approach has been applied for the analysis of biological systems, e.g., see [13], [14], [15], many of which use the model checker PRISM [16].

A Markov model contains a number of parameters, which values need to be determined so that the dynamics of the model is in accordance with experimental results. There exist a number of methods for parameter estimation for Markov models describing networks of biochemical reactions. These methods either make use of the moment closure techniques to derive equations for the moments of the solution of the master equation (e.g., [17], [18]) or focus on the solution of the master equation (e.g., [19]). In the former case certain assumptions regarding the distribution of the Markov chain at each time point are made, in the latter case the methods are built upon various approximations, e.g., some form of state space truncation. Various classes of parameter inference methods can be distinguished. For example, inference methods that maximise the likelihood of the discrete-time observations of system states (e.g., [20], [19], [21], [22]), methods for inference given complete data (e.g., [23]), likelihood-free methods for observations with measurement errors (e.g., [24]), etc. Usually the inference techniques are based on Bayesian inference and make use of Markov chain Monte Carlo algorithms (e.g., [25]).

Structure of the paper. In the following, for this PhD research project we discuss its hypothesis and objectives (Section II), its innovation and originality (Section III), and our methodology to achieve our goals of the project (Section IV).

II. Hypothesis and Research Objectives

The formalisms of Probabilistic Boolean Networks (PBNs) and Dynamic Bayesian Networks (DBNs) are well-established frameworks for modelling gene regulatory networks. The dynamics of these networks is given by the underlying Markov chain. We assume that the possible interactions, i.e., a prior knowledge on the structure of the network, is given from literature studies. This is a common assumption as the model structure can often be identified from the literature, though model identification itself is another important and long-standing research problem in systems biology. However, the relative strengths of interactions are often unknown and are treated as model parameters. By investigating the models underlying Markov chain and by fitting its steady-state distribution to steady-state experimental measurements, we aim to infer the relative strengths of the given interactions between the nodes of the network. In particular, we aim to combine model-checking techniques with sampling and optimisation methods from the literature for parameter estimation of gene regulatory networks modelled as Markov chains.

III. Innovation and Originality

The novelty of this PhD project comes from the fact that we model gene regulatory networks as Markov chains and analyse them through model-checking techniques.

We consider Markov chains as models of gene regulatory networks. For example, the qualitative behaviour of gene regulatory networks is often analysed in the well-established framework of Probabilistic Boolean Networks ([9]) or Dynamic Bayesian Networks, which dynamics is governed by a Markov chain. We plan to combine parameter estimation techniques with Markov chain analysis techniques for inferring model parameters, e.g., relative strengths of interactions between genes. The main task is to find the parameter values of the models that lead to Markov chain which steady-state distributions correspond with the steady-state experimental read-outs. However, obtaining the steady-state distribution of a Markov chain poses a challenge. The originality of our approach is to classify the resulting models with respect to the size of the underlying Markov chain state space and consider different methods for different classes: (1) In the case of *small gene regulatory networks*, numerical model checking techniques, e.g., available in PRISM, allow us to obtain the steady-state distribution in an accurate and exact way. (2) For *medium-size gene regulatory networks*, we will apply statistical model checking which combines sampling methods such as perfect simulation and statistical hypothesis testing [26] to study the steady-state properties of Markov chains. (3) Finally, for *large gene regulatory networks*, we aim to perform pre-analysis of the structure of the underlying Markov chain and utilise from the structure when obtaining the steady-state distribution with stochastic model checking.

IV. Methodology

In this project we focus on biological models of gene regulatory networks in the form of Markov chains and investigate the role of model-checking algorithms when estimating parameters of such models. We first discuss the type of experimental data required for our research and briefly present our methodology.

Experimental data. The existing parameter estimation methods for Markov models require high-resolution time-course measurements. Therefore, we will focus on the use of wet-lab results on steady-state levels, which currently can be obtained more easily than time-course data of required quality. The number of steady-state measurements may, however, be insufficient to identify all the parameters of the model. To address this difficulty, we will design our new methods by taking into account (1) steady-state experimental data of the so called knockout mutant models [27], and (2) qualitative

and/or quantitative information derived from existing time-course measurements, e.g., available at the Gene Expression Omnibus database (see http://www.ncbi.nlm.nih.gov/geo/).

Our methodology. The core idea of our approach is to exploit the capabilities of stochastic model checking to efficiently compute steady-state distributions of Markov chains and combine them with existing sampling and optimisation techniques. This will allow us to search for parameter values that give rise to a Markov chain with its steady-state distribution best fitting the distribution given by the experimental data. The goodness of the fit will be evaluated by means of a cost function, e.g., the Kullback-Leibler divergence. It can be perceived as corresponding to the method of [20], where instead of time-course data the steady-state data are used and maximisation of likelihood is replaced with minimisation of the distance between two probability distributions.

As obtaining the steady-state distribution of a Markov chain (MC) is the main challenge of our research, we will develop different parameter estimation methods for gene regulatory networks of different size. Namely, we will consider

- small networks with $\leq$ 20 nodes leading to Markov chains of the size of not more than 2^{20} states,
- medium-size networks with 20 to 50 nodes, and
- large networks with ≥ 50 nodes.

This will allow us to achieve a good balance between the efficiency (in terms of computational overhead) and effectiveness (in terms of quality of the estimated parameters) for the developed methods.

More specifically, to infer strengths of interactions between genes of small regulatory networks, the strengths will be treated as parameters and inference will be performed by fitting the parameters to experimental data such that the steady-state distribution of the underlying MC will be in agreement with the experimental steady-state data. Numerical model checking techniques will be utilised to obtain the steady-state distribution of the MC in an accurate and exact way. This will lead us to high quality estimates of the model parameters. For medium-size gene regulatory networks, we plan to apply statistical model checking which combines sampling methods and statistical hypothesis testing to study the steady-state properties of the underlying Markov chains. This approach will enable us to obtain the steady-state distribution with high confidence. Sampling methods such as perfect simulation will be explored. In this case, issues of (fast) convergence to steady-state distribution and its efficient estimation will be relevant. For the analysis of Markov chains underlying models of large gene regulatory networks, we aim to pre-analyse the structure of the Markov chains and utilise such information when obtaining the steady-state distribution with stochastic model checking.

V. Conclusion

In this paper, we describe our PhD research project, where we view gene regulatory networks as Markov chains, allowing us to reason about both the structure and strength of gene interactions. We aim to explore model checking techniques with sampling and optimisation methods from the literature to derive new techniques to solve the parameter estimation problem of Markov models of gene regulatory networks. Currently, the project has already started and our preliminary results from the implemented sampling methods seem promising. We will continue to report on new results from this project.

References

[1] H. Kitano, "Systems biology: A brief overview," *Science*, vol. 295, no. 5560, pp. 1662–1664, 2002.

[2] A. Mizera, "Methods for construction and analysis of computational models in systems biology: applications to the modelling of the heat shock response and the self-assembly of intermediate filaments," Ph.D. dissertation, Åbo Akademi University (TUCS), 2011.

[3] E. Klipp, R. Herwig, A. Kowald, C. Wierling, and H. Lehrach, *Systems Biology in Practice. Concepts, Implementation and Application*. Wiley-VCH, 2005.

[4] H. de Jong, "Modeling and simulation of genetic regulatory systems: A literature review," *Journal of Computational Biology*, vol. 9, no. 1, pp. 67–103, 2002.

[5] N. Dojer, A. Gambin, A. Mizera, B. Wilczyński, and J. Tiuryn, "Applying dynamic Bayesian networks to perturbed gene expression data," *BMC Bioinformatics*, vol. 7, p. 249, 2006.

[6] C. Needham, I. Manfield, A. Bulpitt, P. Gilmartin, and D. Westhead, "From gene expression to gene regulatory networks in arabidopsis thaliana," *BMC Systems Biology*, vol. 3(85), pp. 1–17, 2009.

[7] F. Ferrazzi, F. Engel, E. Wu, A. Moseman, I. Kohane, R. Bellazzi, and M. Ramoni, "Inferring cell cycle feedback regulation from gene expression data," *Journal of Biomedical Informatics*, vol. 44, pp. 565–575, 2011.

[8] H. Yu, S. Zhu, B. Zhou, H. Xue, and J. Han, "Inferring causal relationships among different histone modifications and gene expression," *Genome Research*, vol. 18, pp. 1314–1324, 2008.

[9] I. Shmulevich and E. R. Dougherty, *Probabilistic Boolean Networks: The Modeling and Control of Gene Regulatory Networks*. SIAM, 2010.

[10] Q. Yuan, J. Pang, S. Mauw, P. Trairatphisan, M. Wiesinger, and T. Sauter, "A study of the PDGF signaling pathway with PRISM," in *Proc. 3rd Workshop on Computational Models for Cell Processes*, ser. EPTCS, vol. 67, 2011, pp. 65–81.

[11] P. Trairatphisan, A. Mizera, J. Pang, A. A. Tantar, J. Schneider, and T. Sauter, "Recent development and biomedical applications of probabilistic boolean networks," *Cell Communication and Signalling*, vol. 11, p. 46, 2013.

[12] C. Baier and J.-P. Katoen, *Principles in Model Checking*. MIT Press, 2008.

[13] J. Heath, M. Kwiatkowska, G. Norman, D. Parker, and O. Tymchyshyn, "Probabilistic model checking of complex biological pathways," *Theoretical Computer Science*, vol. 319, no. 3, pp. 239–257, 2008.

[14] D. Bosnacki, T. Pronk, and E. de Vink, "In silico modelling and analysis of ribosome kinetics and aa-tRNA competition," *Transactions on Computational Systems Biology XI*, vol. 5750, pp. 69–89, 2009.

[15] Q. Yuan, P. Trairatphisan, J. Pang, S. Mauw, M. Wiesinger, and T. Sauter, "Probabilistic model checking of the PDGF signaling pathway," *Transactions on Computational Systems Biology XIV*, vol. 7625, pp. 151–180, 2012.

[16] M. Z. Kwiatkowska, G. Norman, and D. Parker, *Symbolic Systems Biology*. Jones and Bartlett, 2010, ch. Probabilistic Model Checking for Systems Biology, pp. 31–59.

[17] S. Engblom, "Computing the moments of high dimensional solutions of the master equation," *Applied Mathematics and Computation*, vol. 180, no. 2, pp. 498–515, 2006.

[18] A. Singh and J. P. Hespanha, "Lognormal moment closures for biochemical reactions," in *Proceedings of the 45th IEEE Conference on Decision and Control*, San Diego, CA, USA, December 13-15 2006, pp. 2063–2068.

[19] A. Andreychenko, L. Mikeev, D. Spieler, and V. Wolf, "Approximate maximum likelihood estimation for stochastic chemical kinetics," *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2012, p. 9, 2012.

[20] ——, "Parameter identification for Markov models of biochemical reactions," in *Proc. 23rd Conference on Computer Aided Verification*, ser. LNCS, vol. 6806. Springer, 2011, pp. 83–98.

[21] J. Barnat, L. Brim, A. Krejci, A. Streck, D. Safránek, M. Vejnar, and T. Vejpustek, "On parameter synthesis by parallel model checking," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 3, pp. 693–705, 2012.

[22] L. Brim, M. Ceska, S. Drazan, and D. Safránek, "Exploring parameter space of stochastic biochemical systems using quantitative model checking," in *Proc. 25th Conference on Computer Aided Verification*, ser. LNCS, vol. 8044. Springer, 2013, pp. 107–123.

[23] R. J. Boys, D. J. Wilkinson, and T. B. L. Kirkwood, "Bayesian inference for a discretely observed stochastic kinetic model," *Statistics and Computing*, vol. 18, no. 2, pp. 125–135, 2008.

[24] D. J. Wilkinson, *Stochastic Modelling for Systems Biology*. Boca Raton, FL, USA: Chapman & Hall/CRC, 2012.

[25] L. Kaderali, E. Dazert, U. Zeuge, M. Frese, and R. Bartenschlager, "Reconstructing signaling pathways from RNAi data using probabilistic boolean threshold networks," *Bioinformatics*, vol. 25, no. 17, pp. 2229–2235, 2009.

[26] D. E. Rabih and N. Pekergin, "Statistical model checking using perfect simulation," in *Proc. 7th Conference on Automated Technology for Verification and Analysis*, ser. LNCS, vol. 5799. Springer, 2009, pp. 120–134.

[27] A. Mizera, J. Pang, T. Sauter, and P. Trairatphisan, "A balancing act: Parameter estimation for biological models with steady-state measurements," in *Proc. 11th Conference on Computational Methods in Systems Biology*, ser. LNBI, vol. 8130. Springer, 2013, pp. 253–254.