TUCS

Charmi Panchal

Qualitative Methods for Modeling Biochemical Systems and Datasets

The Logicome and the Reaction Systems Approaches

Turku Centre *for* Computer Science

TUCS Dissertations
No 229, February 2018

Qualitative Methods for Modeling Biochemical Systems and Datasets

The Logicome and the Reaction Systems Approaches

Charmi Panchal

To be presented, with the permission of the Faculty of Science and Engineering of Åbo Akademi University, for public criticism in Auditorium XX on February 12, 2018, at 13:00.

Åbo Akademi
Faculty of Science and Engineering
Vattenborgsvägen 3
20500 Åbo

2018

Supervisor

Professor Ion Petre
Faculty of Science and Enginering
Åbo Akademi University
Vattenborgsvägen 3 20500 Turku
Finland

Co-supervisor

Dr. Vladimir Rogojin
Faculty of Science and Enginering
Åbo Akademi University
Vattenborgsvägen 3 20500 Turku
Finland

Reviewers

Associate Professor Sergey Verlan
LACL, Départment Informatique
UFR Sciences et Technologie
Université Paris Est Créteil Val de Marne
61, av. Général de Gaulle, 94010 Créteil
France

Professor Agustín Riscos-Núñez
Department of Computer Science and Artificial Intelligence
University of Seville
E1.43. E.T.S. Ingeniería Informática
Avda. Reina Mercedes s/n, 41012, Sevilla
Spain

Opponent

Associate Professor Sergey Verlan
LACL, Départment Informatique
UFR Sciences et Technologie
Université Paris Est Créteil Val de Marne
61, av. Général de Gaulle, 94010 Créteil
France

ISBN 978-952-12-3671-6
ISSN 1239-1883

Abstract

In our everyday life we use a number of complex systems that consist of many closely interconnected components. None of the individual components possess a property of the whole system but when they come together they give rise to special properties which are called emergent properties. A similar scenario one may observe in biological systems. There are many interconnected entities such as genes, proteins, and metabolites involved in biological systems. Through their interactions with one another and also with the environment, they exhibit a number of observable characteristics. In order to understand the complexity in biological processes, it is required to understand not just how individual entities function but also how they interact with one another.

The molecular activities involved in biological processes very often remain difficult to understand due to their complex structure. We address the issue in this thesis with the focus on development and demonstration of qualitative approaches, to gain useful insights into several characteristics and dynamics lying within biological phenomena.

The first part of the thesis presents the development of logic-based approaches aka *logicome*, where we use simple heuristics and logical operations to interpret complex scenarios. We apply logicome approaches to capture high-level understanding in terms of mathematical logic of the biological phenomena under study. We demonstrated the logicome approach on two case-studies: (i) the numerical model of Epidermal Growth Factor Receptor (EGFR) signaling pathway (ii) the microarray datasets of Head and Neck/Oral squamous-cell carcinoma (HNOSCC). The logicome proposed for the EGFR signaling pathway investigates activation dependencies within the key species whereas the logicome for HNOSCC microarray datasets produces boolean signatures using the representative genes.

The second part of the thesis presents so-called reaction systems, a nature inspired qualitative modeling framework which functions based on two main principles: threshold principle and no permanency principle. The interactive processes within the reaction systems framework are controlled through two main mechanisms: facilitation and inhibition. We developed reaction systems models which are built on the simple concepts of set theory. Our models demonstrated the feasibility and expressive power of the reaction systems framework as a versatile modeling framework for several dynamics that typically emerged through the traditional quantitative modeling framework. We show reaction systems models to be natural correspondents of models of known dynamic systems such as kinetic models of self-assembly of intermediate filaments, and dynamic models of systems exhibiting behavior such as bi-stability, multi-stability and period doubling bifurcation.

The doctoral thesis is set out to develop and demonstrate the potential role that qualitative approaches play in understanding complex behaviors which are typically observed in biological systems. The hypotheses derived with our approaches are well consistent with the literature findings and the results obtained in other modeling frameworks. We therefore expect that our approaches can be efficient at providing new biological findings for case-studies with intractable complex details.

Sammanfattning

I vårt dagliga liv använder vi ett flertal komplexa system som består av många tätt hopkopplade komponenter. Ingen av de individuella komponenterna har en egenskap som gäller hela systemet, men tillsammans ger de upphov till speciella egenskaper som kallas framväxande egenskaper. Ett liknande scenario kan observeras i biologiska system. Det finns många hopkopplade entiteter, såsom gener, proteiner och metaboliter, involverade i biologiska system. Genom deras interaktioner med varandra och med omgivningen upp-visar de ett flertal observerbara kännetecken. För att kunna förstå komplexiteten i biologiska processer måste man förstå inte bara hur individuella entiteter fungerar men också hur de interagerar med varandra.

De molekylära aktiviteterna involverade i biologiska processer förblir väldigt ofta svårförstådda på grund av sin komplexa struktur. I denna avhandling behandlar vi detta problem med fokus på utveckling och demonstration av kvalitativa förhållningssätt, för att få användbara inblickar i en stor mängd kännetecken och dynamik som ligger inom biologiska fenomen.

Den första delen av avhandlingen presenterar utvecklingen av logikbaserade förhållningssätt, Logicome, som fångar förståelse på hög nivå i form av matematisk logik gällande de biologiska fenomen som studeras. Vi demonstrerade Logicome-förhållningssättet på två fallstudier: (i) den numeriska modellen av Epidermal Growth Factor Receptor(EGFR)-signalväg (ii) Mikromatris-datamängden Head and Neck/Oral squamous-cell carcinoma (HNOSCC). Den Logicome som föreslogs för EGFR-signalvägen undersöker aktivationsberoenden inom nyckelarten medan den Logicome som föreslogs för HNOSCC-mikromatris-datamängden producerar booleska signaturer med hjälp de representativa generna.

Den andra delen av avhandlingen presenterar Reaktionssystem, ett kvalitativt modelleringsramverk inspirerat av naturen. Reaktionssystemramverket fungerar baserat på två principer: tröskelprincipen och principen om ingen beständighet. De interaktiva processerna inom reaktionssystemramverket kontrolleras genom två huvudsakliga mekanismer: underlättande och hämning. Vi utvecklade reaktionssystem-modeller för att demonstrera genomförbarheten och den uttrycksfulla kraften hos reaktionssystemramverket. Vi presenterar reaktionssystem som ett mångsidigt modeller-

ingsramverk för en stor mängd dynamik som typiskt framträder genom de traditionella kvantitativa modelleringsramverken. Reaktionssystem-modellerna är byggda på de enkla mängdlärokoncepten. Vi visar att reaktionssystem-modeller är naturliga motsvarigheter till modeller av kända dynamiska system såsom kinetiska modeller av självsammansättning av mellanliggande filament, och dynamiska modeller av system som uppvisar beteenden såsom bi-stabilitet, multi-stabilitet och perioddubbleringsbifurkation.

Doktorsavhandlingen har som målsättning att utveckla och demonstrera den potentiella roll som kvalitativa förhållningssätt kan spela när det gäller att förstå komplexa beteenden som typiskt observeras i biologiska system. Hypoteserna härledda med våra förhållningssätt är konsekventa med rön inom litteraturen och de resultat som har erhållits med andra modelleringsramverk. Därför förväntar vi oss att våra förhållningssätt kan vara effektiva när det gäller att bestå nya biologiska rön till fallstudier med svårbehandlade, komplexa detaljer.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor Professor Ion Petre, Leader of the Computational Biomodeling Laboratory, for the incredible mentorship. I am indebted to him for the invaluable help, advice, encouragement and patient guidance. I was exceptionally fortunate to have a mentor who not only cared about my work but also provided constant support and frequent insights throughout my time as his student.

An immense gratitude goes to my co-supervisor Dr. Vladimir Rogojin for many useful advice. He has always been extremely supportive and generous with his considerable amount of time that he spent to give me constructive feedback for the improvement of my thesis.

With a special mention to Professor Sergey Verlan, I am very thankful to him for agreeing to serve as a reviewer of my dissertation and act as the opponent at my doctoral defence. I thank him also for providing crucial comments and critical reviews to improve the quality of the thesis. I am grateful to Professor Agustín Riscos-Núñez who kindly agreed to be a reviewer of the thesis, and provided me with encouraging suggestions for the thesis. I am very grateful to both of them for allocating time and efforts to carefully reading and providing their valuable remarks.

The scientific articles constitute the doctoral thesis and it would not have been possible without my co-authors. I have had the privilege to work with Dr. Sepinoud Azimi, Dr. Eugen Czeizler, Dr. Andzej Mizera and Dr. Vladimir Rogozin. I express my sincere gratitude for their collaboration and for making this thesis possible. Indeed, I enjoyed working with them and learned many things from their knowledge and experience.

I am very grateful to Professor Ralph-Johan Back and the Software Construction Laboratory for the various support, encouragement, and assistance that I received, especially at the early stage of my studies.

During my time at Computation Biomodeling Laboratory in Åbo Akademi, I have always been supported by former and present members in the laboratory. I thank each and everyone of them for the consistent scientific assistance, regular Combio seminars and refreshing lab excursions. I extend my heartfelt gratitude to my wonderful friend Diana-Elena Gratie for always being ready to provide enormous support and encouragement, especially during the stressful situations. I am deeply grateful to Bogdan Iancu, Sepinoud Azimi, Vladimir Rogojin, Eugen Czeizler, and Mikhail Barash for creating a nice and friendly atmosphere in which I could enjoy my doctoral studies. Further, I extend my special thanks to Krishna Kanhaiya, Usman Sanwal, Dwitiya Tyagi, Utku Kapucu and Victor Popescu for many fruitful discussions, enjoyable and memorable moments.

I express my appreciation to Åbo Akademi University and TUCS (Turku Centre for Computer Science) for the excellent infrastructure as well as a pleasant working environment that I was provided during my studies. I sincerely thank the administrative and technical personnel at the department and at TUCS for their assistance with all the substantial administrative and technical matters, especially I thank Christel Engblom, Minna Carla, Nina Hutholm, Tove Österroos, Susanne Ramstedt, Karl Rönnholm and Tomi Suovuo. I thank TUCS, Otto A. Malm donationsfond, Oskar Öflunds Stiftelse, Finnish Cultural Foundation, and the Rector of Åbo Akademi for providing me financial support throughout my doctoral studies. I thank TUCS, Åbo Akademi, Akademi of Finland and CompLifeSci grants for providing the funding which gave me the opportunities to attend conferences, workshops, and summer schools and meet so many interesting people.

Finishing the doctoral work would have been way more difficult if it was not for the genuine friendship and support provided by the other members of the department. Especially I want to express my gratitude to Maryam Kamali, Mojgan Kamali, Mats Neovius, Irum Rauf, Natalia Díaz Rodríguez, and Inna Vistbakka, with whom I spent moments of pleasure and leisure that helped me to overcome stressful times. I am truly grateful to them for always being available to provide prompt help whenever I needed it. I deeply appreciate my friends Faezeh Siavashi, Sergey Ostroumov, Yuliya Prokhorova and Sara Seppola for their thoughtfulness and words of encouragement. I am thankful to Petter Sandvik and Annamari Soini for their help with the Swedish version of the abstract.

I feel very fortunate to have Mikolaj Olszewski (Miki) and Marta Olszewska as my close friends. You always make me feel like at home with family. I can't thank you enough for the friendship, words of comfort, joyous moments and strength that you always gave me. Furthermore, I would like to thank Linda Mannila. Linda, I will be forever grateful for all the positive influence you had on me. Thank you for your concern, care and useful advice.

My life in Turku would not have been rich and joyful without the amazing friends outside the university. I am deeply thankful to Hannele Svanström and Ignė Stalmokaitė, for making my life in Turku a truly enjoyable and memorable. I am grateful to Säde Kankare, Eero Kankare, Rebecca Brückner, Pia Brückner and other neighbors for their help and lovely moments that we spent together. I thank Indian community in Turku, especially Srinidhi Bellur Ravindra, Dwitiya Tyagi, Bishwesvar Pratap Singh, Debangana Baruah, Jimit Sanghvi, and Chamundeeswari Sanghvi (Chamu), for keeping my stay full of pleasurable and excellent times. Further, I thank Diti Desai for her care, support and being such a wonderful friend. I express my sincere gratitude to Dr. Abhay Bulsari for his moral support and guiding me into a better path with his valuable lessons, and inspiration. Thank you for your generosity and for making Turku a home away from home. I am very grateful for being a part of the Academic Badminton Club that kept me active and helped me relieve stress.

I am glad to have met many friends from all around the world during my stay in Lappeenranta, I deeply appreciate all of you for being always there for me. You all made my stay in Finland very memorable and joyous. I am grateful to Sakari Kiiskinen and his family for being so kind and loving. My special thanks go to Seija and Reijo Turunen for creating Finland as a second home and introducing a wonderful and fascinating life in Finland to me. I thank all of my Indian friends from Lappeenranta for making my stay in Finland even more cheerful.

I will always be obliged to all of my teachers from India, who helped me get through many struggles, especially I thank Ms. Pritika Kashyap for her enormous efforts to grow confidence in myself. I'm extremely grateful to my friends in India with whom I am still enjoying the friendship since my school time. Thank you, Gayatri Dhaneshwar and Ankita Solanki for being my friends during all stages of life and keep bringing good old memories.

Above all, I am grateful to my dear family for their unconditional love, all-embracing support, encouragements, and prayers. I express my deep gratefulness to my cousin sister Nilam Panchal for always motivating me and giving me additional strength. The heartfelt appreciation goes to my other cousins who have been special and loyal friends from the same family. I am beyond lucky to have my dear brother Jay Panchal here in Finland. My heartfelt thanks to him for his love and being with me throughout both tough and happy times. Thanks for making my life a true festival. No words can describe the boundless trust, patience, and love that I always receive from my parents Devendra Panchal and Tilottama Panchal. I am one of the privileged ones to have you as parents in my life and cannot thank you enough for the sacrifices that you made to support all my pursuits. The acknowledgment would not be complete if I did not mention Soham Shah, Gayatri Shah, and Mahesh Shah. Thank you for gracing my life with your love, support, and encouragement. Thank you Soham, for cultivating positive feelings and changing my life without even trying.

List of original publications

1. Charmi Panchal, Sepinoud Azimi, and Ion Petre. Generating the Logicome of a Biological Network. In: María Botón-Fernández, Carlos Martín-Vide, Sergio Santander-Jiménez, Miguel A. Vega-Rodríguez (eds). Algorithms for Computational Biology. Lecture Notes in Computer Science, volume 9702, pp 38-49. Springer International Publishing, 2016.

2. Charmi Panchal, and Vladimir Rogojin. Generating the Logicome from Microarray Data. In Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2017 IEEE Conference on, pp. 1-8. IEEE, 2017.

3. Sepinoud Azimi, Charmi Panchal, Eugen Czeizler, and Ion Petre. Reaction systems models for the self-assembly of intermediate filaments. Annals of University of Bucharest,62:9–24, Editura Universităţii din Bucureşti, 2015.

4. Sepinoud Azimi, Charmi Panchal, Andrzej Mizera and Ion Petre. Multi-Stability, Limit Cycles, and Period-Doubling Bifurcation with Reaction Systems. To appear in International Journal of Foundations of Computer Science, 2018.

Contents

Chapter 1

Introduction

Most systems in biology are often recognized as “complex systems” which implies systems composed of multiple components that may interact with each other. Computer science has been increasingly successful in exploring and interpreting the dynamical features of the complex operations within the biological components [62, 73]. In pace with discovering ever growing incredible amounts of information embedded in a biological system, a wide plethora of novel methodologies contributed by computer science have progressed towards directly modeling and analyzing a biological system in a simple and intuitive manner. The methodologies are not only limited to modeling and analysis but also produce sophisticated executable bio-models which help comprehend the processes at a level useful for prediction and may aid into reducing the cost of expensive *in vivo* and *in vitro* biological experiments. Novel concepts and well-designed tools facilitate the understanding of complex multi-domain biological phenomena and extract meaningful insights from them [15, 43, 69]. The computer models and simulations have brought remarkable transformations in biological research and were progressively successful in demonstrating causality mechanisms between multi-scale biological processes [108].

The computer models of biological systems can be investigated from both qualitative and quantitative perspectives [92, 67, 32]. Quantitative modeling deals with numerical models and requires numerical setups, like reaction rates, rate laws, initial concentrations, etc.[24, 6, 94, 54, 2]. Qualitative modeling operates within a discrete domain of a system where the variables of the system may take one of a few discrete values. Also, the qualitative modeling approaches have less computational requirements as well as much less initial data requirements [92, 93, 64, 100, 5].

The research reported in the present doctoral thesis concentrates on developing and utilizing different methodologies arising from computer science that extract simplified representations of the characteristics of the biological phenomena. The thesis contains two parts, each reporting a research direction we were involved in.

In the first part, we used heuristic strategies expressed in terms of Boolean logic to develop methodologies to infer qualitative models. Just as terms like genome, proteome, metabolome, etc. that study genes, proteins and metabolites respectively, we introduced the term *logicome* that extracts static description of a biological phenomenon. The logicome approach presents external/internal characteristics of a biological phenomenon on a high abstraction level using a logic-based formal framework.

The logic-based approaches have successfully captured the most salient properties of the biological systems and have been applied to model a wide range of biological processes (for instance, differentiation of T-helper cells, gene and molecular networks, the fission yeast cell cycle network) [3, 67, 25].

Logic-based approaches are also introduced for learning causal relationships and interactions between the components in biological systems where the system is very complex and contains uncertainty (e.g., measurement error, experimental error, noisy data, and incomplete data). For example, authors in [45, 89] demonstrate the ability of logical models to discover the structure and build verified predictive models of the biochemical systems. Logical modelling (often Boolean modelling) has shown to be capable of inferring regulatory mechanisms, performing structural analysis of cellular networks and demonstrating the dynamics of the biochemical networks [105, 25, 4, 72, 92, 77].

The logicome approach represents the characteristics of the phenomenon through a simplified model represented with the Boolean formalism. By using this approach, we perform following case studies: biochemical/molecular network of Epidermal Growth Factor Recepter (EGFR) signaling pathway, given as a set of reactions and microarray datasets of Head and Neck/Oral squamous-cell carcinoma (HNOSCC). The logicome outcomes for EGFR signaling pathway provide logical description of the activation dependencies of the user selected key elements within the pathway, whereas the logicome outcomes for HNOSCC microarray datasets identify regulations of significant genes and provide boolean classifier for each sample groups.

In the second part, we used the qualitative framework of *Reaction Systems* (RS in short), which was introduced by A. Ehrenfeucht and G. Rozenberg, as a framework to formalise the interactions between biochemical reactions [35, 37]. The reaction systems approach relies on the notions of set theory. A reaction in reaction systems is characterized by its set of reactants, set of inhibitors and set of products. In the reaction systems framework, the interactions between the biochemical reactions are controlled through two main mechanisms: *facilitation* and *inhibition*. Intuitively, a reaction is successfully enabled if all of its reactants are present and all of its inhibitors are absent. Moreover, the two main assumptions that make the reaction systems framework different from other modeling frameworks are: *threshold assumption* and *no permanency assumption*. In the *threshold assumption*, it is assumed that if an element is present, it is available in a sufficient amount for the reactions to take place. In the *no permanency assumption*, it is assumed that the element ceases to exist in the system, if it is not sustained by any reaction.

Each of the products will be present after the reaction has successfully taken place. The research on reaction systems has proven to be very promising and growing in different directions. The original motivation of reaction systems was nature inspired processes which came from biology, but the research on reaction systems has been very versatile in theoretical and biomodeling aspects. The diverse applications of reaction systems as novel models of computation can be found in e.g. [12, 11, 74, 41, 91, 85, 90].

We consider several case studies that demonstrate the natural correspondence of reaction systems to quantitative modeling frameworks. With reaction systems, we build models for several dynamical systems and through its interactive process we reproduced a similar behavior as observed in dynamical models. For this, we consider a kinetic model of self-assembly of intermediate filaments, and several dynamical systems exhibiting behavior such as bi-stability, multi-stability and period doubling bifurcation.

The above mentioned research directions are discussed in the subsequent chapters. In Chapter 2, we introduce the *logicome* approach and present the methodologies involved in it. We also give brief biological backgrounds of the investigated case studies with logicome approaches. In Chapter 3, we introduce the reaction systems framework and discuss its applicability. We further present our case studies modeled with reaction systems and demonstrate the dynamics of reaction systems models in correspondence with the quantitative models. In Chapter 4, we list the original research contribution and briefly discuss the contribution of each publication. Finally in Chapter 5, we conclude with a perspective for future research directions.

Chapter 2

The Logicome Approach

There exists a number of modeling and computational approaches to elucidate high-level understanding of complex molecular, cellular and organ level biological processes [56, 103, 20]. The insights gained from such studies aid the development of new predictive models that complement conventional experimental approaches. We define the term *logicome* as a metaphor to depict the static snapshot of the phenomenon under study in terms of logical expressions. The main emphasis of the logicome approach is on unveiling the interplay between the selected key elements and describe the observable characteristics of the phenomenon under study in terms of Boolean expressions. In this chapter, we discuss the methods involved in the logicome approach and demonstrate them on two case studies.

2.1 Background

The activities of different components within a complex multi-component biological system are not independent of each other. It is possible to understand the complex relationships between these components with a method that facilitates systematic functional analysis of the system. The most common approaches for analyzing and modeling biological systems are ordinary differential equations (ODEs) [86], Bayesian networks [107], Petri nets [21], linear programming [81] and agent-based model [101]. In addition to these approaches, modeling with Boolean Networks (BNs) [60, 25] has also proven to be successful in investigating large and complicated biological systems [7, 39].

Moreover, there has been an enormous contribution towards developing simplified methodologies and tools that facilitate construction and analysis of models of biological processes with different levels of abstraction [27, 59, 6, 92, 40, 49, 53, 42]. The level of abstraction depends on the availability of the experimental data of the model such as model components, interactions, kinetic information. However, such kind of experimental information is often lacking in the published models. The wide array of qualitative approaches is developed to describe processes of high-throughput cell biology when precise experimental information is not available. [3, 102, 99, 29].

Many studies have demonstrated the ability of logical approaches to extract structural features and functional analysis of cellular signaling or regulatory networks (see for example [63, 89]). The study in [93] introduced scalable qualitative approach for building and analyzing executable models of biological systems. In [46], authors proposed a tool incorporating a rich modeling language and semantics of logic programming to explore a family of feasible logic models of signal transduction.

In the logicome approach, the main goal is to derive a Boolean network/formulation that elucidates the emergent properties and behaviors of the model under study. The selected components of the model are reduced to discrete values: “0” and “1”. In other words, the components take discrete values which represent qualitative levels of activity and inactivity, or up-regulation and down-regulation. We used following two logicome approaches to represent the dynamics underlying the signaling pathway model and to identify the signatures for categories within the data driven models.

- Model-based logicome: in this approach, the input is a numerical model of EGFR signaling pathway and the outcome is a qualitative model in terms of Boolean network. The nodes of the Boolean network are the selected key species of the pathway model. The network analyses the activation dependencies within the selected key species.

- Data-based logicome: in this approach, the input is microarray data and the outcome is a set of Boolean signatures inferred from the microarray data for the sample groups, where Boolean variables represent expression levels of some selected representative genes. Intuitively, this method comprehends the data in terms of boolean formulas.

2.2 Case studies

We applied the logicome approach to the model of Epidermal Growth Factor Receptor (EGFR) signaling pathway and to the microarray data sets of head and neck/oral squamous cell carcinoma (HNOSCC). In this section, we highlight some basic features of the EGFR signaling pathway and of microarray datasets.

The signaling in cellular components originates from the extra-cellular domain via receptors situated on the cellular surface. For instance, intracellular signaling pathways (cascades) usually originate from those receptors. These signaling cascades regulate many important cellular processes, such as cell differentiation, division and proliferation [110]. The Epidermal Growth Factor Receptors (EGFRs) are a family of receptor tyrosine kinases extensively studied in several types of cancer and have a major contribution in fundamental cellular processes [109]. Due to their potential involvement in numerous tumour cell responses, the EGFR signaling pathway has been the target of effective cancer therapies [96, 78].

The EGFR is an extra-cellular receptor for EGF (Epidermal Growth Factor) and the formation of the signaling cascade begins with the binding of EGF with the receptor. The signaling cascade initiates critical molecular events such as dimerization and phosphorylation of its intracellular species, which control various cellular responses [61, 78]. The signaling pathways typically are composed of several conserved functional domains, also called modules. The signal propagation activates dynamics of several modules within the intracellular domain, that exhibit intertwined control mechanisms like positive and negative feedback loops, which also play a key role in preserving the stability of the signal [47]. The modules interact with each other through their common species, which we agree to call "interface species" [104]. The activity of each interface species depends on other interface species. The discovery of such dependencies helps to analyze the overall control mechanisms within the EGFR signaling pathway.

2.2.1 Cell signaling pathway

The numerical model of the modularised EGFR signaling pathway that we used is adapted from the previous studies reported in [52, 104, 95], which also analyze the same pathway. We focus on the dynamics of the interface species within the EGFR signaling pathway. The abstract schema depicted in Figure 2.1 illustrates the modularized EGFR signaling pathway with modules and interface species that we focus on. Figure 2.1 is based on the biochemical map given in [104] that depicts reactions involved in the EGFR signaling pathways. We analyze this pathway by identifying the activation dependencies associated with the interface species. In [82], we treat interface species as "key elements" of the pathway.

The interface species that we focus on are: (EGF-EGFR*)2-GAP, (EGF-EGFR*)2-GAP-Grb2-Sos, (EGF-EGFR*)2-GAP-Shc*-Grb2-Sos, Ras-GTP, Ras-GTP*, MEK-PP, Raf* and ERK-PP. The end result of the signaling cascade is the phosphorylated ERK protein (ERK-PP) that regulates several cellular proteins and nuclear transcription factors essential for cellular responses [61].

The quantitative models of biochemical and cellular systems are available at several freely-accessible public repositories [66, 97, 70]. We used the BioModels Database [68] to obtain a numerical model of the EGFR signaling pathway represented in SBML (Systems Biology Markup Language).

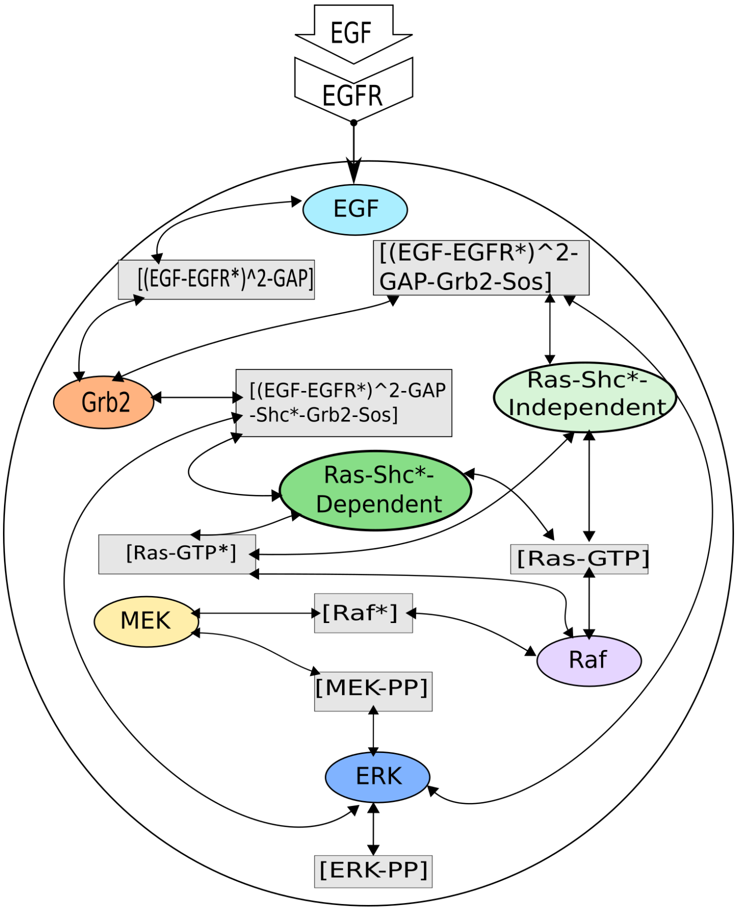


Figure 2.1: The abstract schema representing links between modules and interface species that we focus on in the EGFR signaling pathway: the ovals represent modules while the grey boxes represent interface species. The figure is based on the biochemical map given in [104] that depicts reactions involved in the EGFR signaling pathways.

2.2.2 DNA microarray datasets

The microarray provides a natural platform for systematic and comprehensive analysis of the genome. At present, there exist several types of microarrays such as DNA microarrays, protein microarrays, tissue microarrays, cellular microarrays (transfection microarrays), chemical compound microarrays, antibody microarrays and carbohydrate arrays [75] to address various biological questions. The microarray technology has been successfully applied in various fields including but not limited to the functional analysis of different cellular processes, neuroscience, ecology and evolution [50, 44, 87, 71].

The processes mediating the phenotype of the cells (e.g. tumor and normal) are encoded in molecular units called “genes”. The gene expression is a process by which a cell responds to its changing environment. In other words, specific genes possessed by the cell are expressed and are involved in the production of a functional product such as a protein. This process is measured at two levels, as shown in Figure 2.2: transcription and translation. During the transcription level, the information stored in the DNA sequence of genes is transferred into an RNA copy of a gene sequence, referred to a messenger RNA (mRNA) sequence. At the translation level, the information represented by the mRNA sequence is used as a template for the synthesis of proteins.

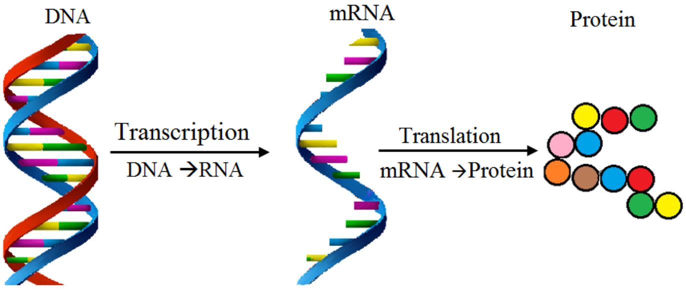


Figure 2.2: The central dogma of gene expression: transcription and translation [23].

In this work, we used expression microarrays, also known as DNA microarrays, that measure thousands of gene expression patterns simultaneously. The DNA microarrays quantify expressions of the genes present in the cell and the expression values are based on the measurement of mRNA translated into the end functional products, i.e. proteins.

The genomic high-throughput technologies, such as the DNA microarrays, are rapidly evolving with numerous applications in gene expression, genotyping, resequencing, mutation analysis, drug discovery and pharmacogenomics [94]. The DNA microarrays is a powerful tool for studying complex changes in the expression level patterns of thousands of genes simultaneously and classifying the biological entities (e.g. classifying tumor samples) based on these patterns. A DNA microarray is a slide imprinted with an ordered array of thousands of tiny spots, with each spot representing a known DNA sequence or gene. The DNA molecules on each slide act as gene probes (short section of a gene) which are a set of oligonucleotides complementary to fragments of the corresponding genes.

The microarray data are available at several microarray repositories, such as Gene Expression Omnibus (GEO), storing a vast amount of microarray data [31]. The microarray dataset contains expression values of probes corresponding to 20,000-40,000 genes for hundreds of samples. The microarray is scanned to measure expression of each gene printed on the slide through a hybridization process [80]. Following the hybridization process, the genes are labeled according to their expression levels.

For this work, we have used RNA expression data for samples from nine microarray datasets of Head and Neck/Oral squamous cell carcinoma (HNOSCC) downloaded from GEO.

2.3 Methods

The goal of the logicome approaches is to describe the structure of biological phenomena and extract the specific behavior of the selected components of the phenomena under study. Logicome approaches comprise two directions: *model-based logicome* and *data-based logicome*. The model-based logicome approach analyzes a Boolean model derived from the numerical model of the cell signaling pathway. The data-based logicome approach describes large datasets using boolean signatures. In this section, we present steps involved in both logicome approaches.

2.3.1 Model-based logicome

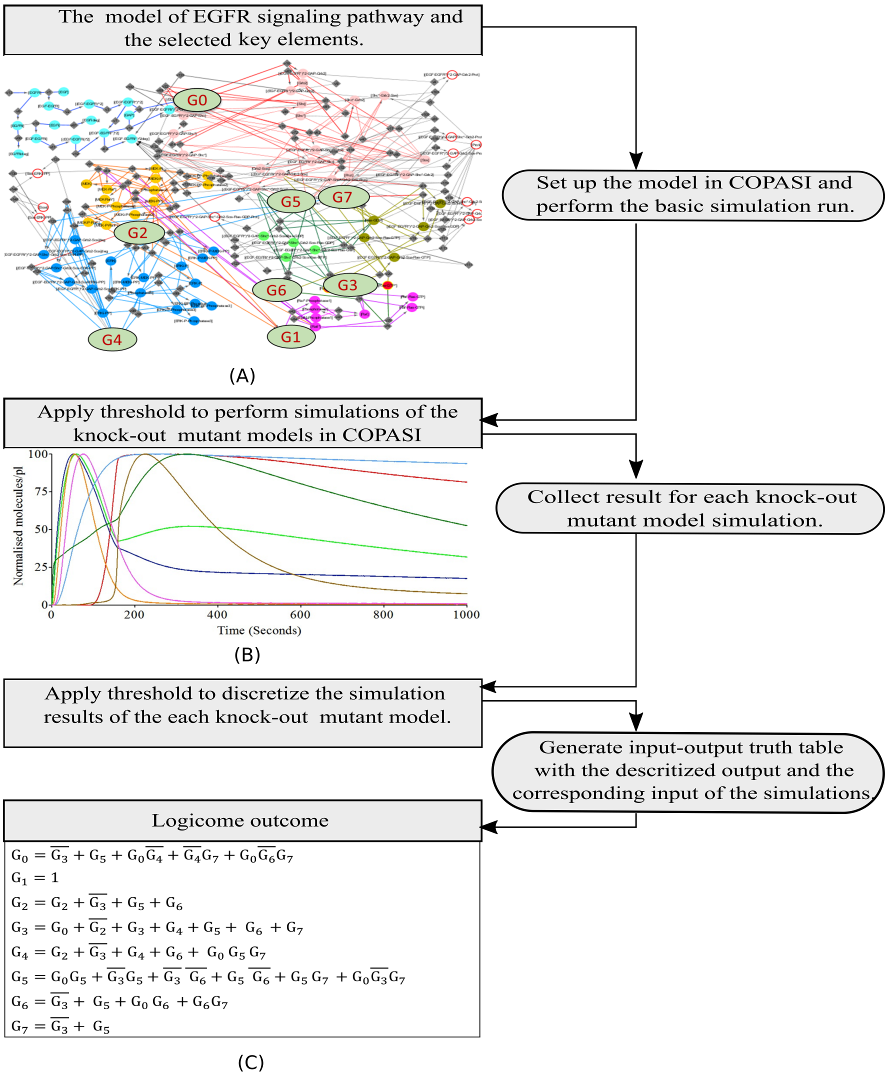
In our studies, the model-based logicome captures the main functionalities lying within the EGFR signaling pathway model. The methodology is summarized in Figure 2.3.

The steps involved in the model-based logicome include:

1. Setting up the model in the modeling framework
2. Threshold-based discretization
3. Simulation of the knock-out mutants of the model
4. Deriving the logicome outcome

The logicome outcome derived in this method represents a numerical model of the EGFR signaling pathway in terms of Boolean networks. We discuss each of these steps below.

- *Setting up the model in the modeling framework:* As the first step of the method, the SBML model of the EGFR signaling pathway model is implemented in the modeling framework of COPASI. Before performing knock-out mutant simulations, the unperturbed model is run


The model of EGFR signaling pathway and the selected key elements.
G0
G5
G7
G2
G6
G3
G4
G1
(A)
Set up the model in COPASI and perform the basic simulation run.
Apply threshold to perform simulations of the knock-out mutant models in COPASI
Normalised molecules/pl
100
75
50
25
0
0
200
400
600
800
1000
Time (Seconds)
(B)
Collect result for each knock-out mutant model simulation.
Apply threshold to discretize the simulation results of the each knock-out mutant model.
Generate input-output truth table with the descritized output and the corresponding input of the simulations.
Logicome outcome
(C)

$$G_0 = \overline{G_3} + G_5 + G_0\overline{G_4} + \overline{G_4}G_7 + G_0\overline{G_6}G_7$$
$$G_1 = 1$$
$$G_2 = G_2 + \overline{G_3} + G_5 + G_6$$
$$G_3 = G_0 + \overline{G_2} + G_3 + G_4 + G_5 + G_6 + G_7$$
$$G_4 = G_2 + \overline{G_3} + G_4 + G_6 + G_0 G_5 G_7$$
$$G_5 = G_0G_5 + \overline{G_3}G_5 + \overline{G_3}\,\overline{G_6} + G_5\overline{G_6} + G_5 G_7 + G_0\overline{G_3}G_7$$
$$G_6 = \overline{G_3} + G_5 + G_0 G_6 + G_6G_7$$
$$G_7 = \overline{G_3} + G_5$$

Figure 2.3: Model-based logicome:(A) Model of EGFR signaling pathway and the selected key elements (interface species) highlighted with green ovals (B) Illustration of COPASI simulation for a knock-out mutant model (C) Logicome outcome derived in terms of a Boolean network using the tool LogicFriday[1]

and the simulation results of the selected key elements are collected. We call this simulation run the *basic simulation run*. As mentioned in Section 2.2.1, we focus on the selected interface species and consider these key species as the key elements of the model.

- *Threshold-based discretization:* The discretization is performed on the selected interface species. Discretization is performed in the following two stages:

  (i)the initial state of interface species are set ON/OFF in the beginning of knock-out mutant simulation. For an interface species species S, the initial state S_{init} is set as follows:

  $$S_{init} = \begin{cases} p\% \text{ of } max_b(S), & \text{if } S \text{ is set to } ON \\ 0 & \text{if } S \text{ is set to } OFF. \end{cases}$$

  where $max_b(S)$ is the maximum value of S in the basic simulation run and p is parameter that modeler chooses to calculate the initial value.

  (ii) the interface species are labeled to active/inactive in the results collected for the knock-out simulation.

  $$S = \begin{cases} 1, & \text{if } max(S) \geq max_b(S) \\ 0, & \text{otherwise.} \end{cases}$$

  where $max(S)$ is the maximum value of S in the knock-out model simulation.

- *Knock-out mutant simulation:* For the selected n interface species, there are 2^n knock-out mutant models generated by setting the interface species to ON/OFF in all the possible combinations. Each knock-out mutant model is simulated and results are collected.

- *The logicome outcome:* The outcomes generated with all the knock-out mutant simulations are collected and the corresponding logicome is derived. In the logicome approach, the level of abstraction depends on the selected number of the interface species. The logicome outcome unveils the activation mechanisms between the selected interface species. The outcome depends on parameters such as the choice of threshold parameters and the initial values of the species in the ODE model.

In the model-based logicome approach, the signaling pathway model written in SBML (Systems Biology Mark-up Language) is used. The model is divided into highly connected modules consisting of tightly interconnected nodes/elements. In the model-based logicome approach, the complex details of the model are hidden with the threshold-based discretization. The logical functions are inferred from the simplified model.

The logical functions in the model-based logicome are derived from the complete truth table with n inputs and n outputs where n is the selected number of interface species. The n inputs represent all the knock-in/outs for all the interface species, and outputs represent their resulting activation states after the simulation.

The EGFR signaling pathway has an intrinsic capacity to maintain certain crucial mechanisms when faced with the perturbation [8]. The pathway contains multiple components with similar functions and effects. These multiple components are also called redundant components. These functional redundancies in the pathway are associated with signal sustainability and provide compensatory activity in the case of component failure. The presence of redundancies ensures the biological robustness (ability to preserve outcome) of the pathway [106, 8]. As we reported in [82], the logicome outcome identifies several control mechanisms including redundancies and self-regulation emerging out of the complex interplay between the interface species of the model.

In [82], we also derived the logicome outcome when it is not straightforward to set certain ON/OFF combination into the numerical model. In such a situation, the outcome is derived from the discretized data with the assumption that several knock-out mutants are not available.

The model-based logicome approach has been robust to incomplete data for our case study. In other words, the logicome approach was able to produce an almost similar outcome even when some portion of the data was missing.

The model-based logicome approach allows modeler to focus on the selected key elements and to choose a appropreate threshold depending upon the network under study to descritise the selected key elements. The final outcome generated with the model-based logicome approach captures an abstract snap-shot of the detailed network. The outcome provides the activation dependancies within the selected key elements for the given initial activation status.

2.3.2 Data-based logicome

The data-based logicome approach demonstrates the signature inference from the microarray datasets. We used microarray datasets of head and neck/oral squamous cell carcinoma (HNOSCC) as a case study. The extracted signatures are Boolean formulations and represent a simple characterization of various cancer and normal samples.

The logicome outcome derived in this approach is a Boolean signature built with the selected set of genes, which classify the sample into one of several known groups. The methodology is summarized in Figure 2.4.

The main steps involved in the data-based logicome methodology are: (i) data collection (ii) data preprocessing (iii) multinomial logistic approach to find the representative subsets (iv) extracting Boolean signature. In the following we discuss the main steps in details.

- *Data collection:* The microarray datasets of HNOSCC were retrieved from Gene Expression Omnibus (GEO), the public repository for a wide range of high-throughput experimental data. The data in GEO data series are arranged in tab-delimited tables and describe related samples from a study, the overall study aim and design. Each data series is assigned with unique accession numbers with the prefix GSE. In GEO data series, the samples are organized into meaningful datasets and arranged by a common attribute. The datasets are represented in a row-column array format, with GEO samples listed in columns and probes listed in rows.

- *Data preprocessing:* The obtained datasets are normalized with Robust Multi-Array Average (RMA) method [55] and consists of more than 50,000 probe sets for different genes. The measurements expressed by the probe sets often make the dataset large and complicated, which require enormous efforts to analyze and obtain valid results. For these reasons, it is required to process the dataset before the actual analysis to find Boolean signature. In our approach, the steps involved in preprocessing are: (i)deriving gene expression matrices (ii)selecting a set of significant genes (iii)re-scaling gene expression datasets to the same level (iv) removing similar samples between different groups. In the following we describe each step in details:

  - *Deriving gene expression matrices:* The expression values of probe sets are summarized into gene expression by combining multiple measurements of probe sets on the same gene. In our datasets, to transform probe expressions into gene expressions we used the median, calculated from the expression values of the probe sets mapped to the same gene. After all the probes mapped to the respective genes, we obtained a gene expression matrix of approximately 25,000 rows represented by gene symbols, and columns by the samples distributed into control and cancer groups.

  - *Selecting a set of significant genes:* In order to find the significant genes from the datasets, we performed differential expression analysis [30]. The differential gene expression analysis was performed with the web-based tool GEO2R [14]. The differential expression analysis investigated the genes within the microarrays whose expression levels change between two sample groups. This analysis identified which genes are increased in expression (up-regulated) or decreased in expression (down-regulated) between the control and cancer groups.

As a result of differential expression analysis in GEO2R with the method of Benjamini-Hochberg (False discovery rate) [88], we obtained a list of genes ordered by their p-values.

The purpose of differential expression analysis is to find those genes which show difference in expression between groups, thereby signifying their involvement in some sample group (cancer or control) of interest. To identify the differentially expressed genes, two hypotheses are tested: null hypothesis H_0 and the alternate hypothesis H_1.

- H_0 = for a gene g, no real difference exists between the expression values in the control and cancer groups

- H_1 = for a gene g, a difference exists between the expression values in the control and cancer groups

If we reject H_0, then gene g has different expressions under the two groups, and so is differentially expressed. The p-value measures probability of observing expression values for a gene g, under the null hypothesis. A small p-value indicates that there is a small chance of obtaining expression values of a gene g with no real difference. In our case-study, by small we mean 0.05.

From the list of genes ordered by their p-values, we selected the genes with a p-value $\leq$ 0.05. We performed this for all the data series and extracted the common genes significant in all the datasets.

- *Re-scaling gene expression datasets to the same level:* The GEO data series are first submitted to the GEO platforms by various scientific communities and are generated through a diverse range of technologies. The data series derived from a different GEO platform may not be in uniform standards and may involve disparities in scale. Before combining the datasets extracted from different data series, it is necessary to rescale individually in order to improve the numerical stability and reliability of the results. The Boolean signatures are derived for the datasets rescaled with the following approach: for each dataset D, for each sample S from D and for every gene G, we re-scaled its expression value $x_{G,S}$ associated to S to $z_{G,S}$ as follows:

$$z_{G,S} = \frac{x_{G,S} - m_{G,D}}{R_{G,D}} \tag{2.1}$$

where $m_{G,D}$ is the mean value of expressions of gene G for all the samples from dataset D, and $R_{G,D}$ is the range or interquartile range of expressions of gene G among all the samples of D.

For the full range we have $R_{G,D} = MAX_{G,D} - MIN_{G,D}$, where $MAX_{G,D}$ ($MIN_{G,D}$) is the highest (the lowest, respectively) expression level for a gene G among all the samples from dataset D. For the interquartile range we have $R_{G,D} = Q_3 - Q_1$, where Q_3 is the third quartile and Q_1 is the first quartile.

  - *Removing similar samples between different groups:* The rescaled data are combined into one dataset and used for further processing and analysis. The combined dataset contains expression values of significant genes for the samples divided into different groups derived from 9 different dataseries. The combined dataset has four different groups: Oral tongue squamous cell carcinoma (OTSCC), Oral squamous cell carcinoma of the oral cavity and oropharynx (OSCC), Head and neck squamous cell carcinoma (HNSCC), and Normal/control samples.

    The groups are checked for similarities using the measure of Euclidean distance between the group pairs. The identical samples are removed.

- *Multinomial logistic approach to find the representative subsets :*

  After the similarity check, the dataset is subjected to a machine learning method that derives the subsets of significant genes to construct Boolean signatures. From the set of significant genes, all the possible subsets of size three or larger are generated. For each subset, the expression data are partitioned into training and validation in the ratio of 60:40. We choose the multinomial logistic regression method to train the model on the training data and collect accuracies of the model using the validation data. The subsets giving maximum accuracy in the individual size group are collected. From the collected subsets, the subset with minimal size having accuracy $\geq 70\%$ is picked to derive the Boolean signature. This step of the method is executed multiple times and minimal subsets are collected for deriving the Boolean signature.

- *Extracting Boolean signature:* The gene expression data extracted for the minimal size subset collected in the previous step were subjected to discretization by some measure. The discretization reduces the large domain of numerical values to a nominal scale and provides concise data representation. As a threshold for discretization, we used the value of median calculated from the expression values of a gene across the samples. For a gene, the expression value below the median is replaced with "0"(down-regulated) and "1" (up-regulated) otherwise. The Boolean signature is deduced from the discretized gene expression matrix. In the discretized gene expression matrix, the frequencies of Boolean vectors representing the samples, are calculated within a group. In each group, the Boolean vector with the maximum frequency $Pmax$ and the Boolean vectors with the frequencies belonging in the interval $[max(0.25, 0.75 * Pmax), Pmax]$ are the representatives for that group. The disjunctive normal form constructed from the selected representative Boolean vectors forms the Boolean signature for the group.

The logicome outcome derived as a set of Boolean signatures represents the most occurring patterns in the respective groups of samples. The Boolean signatures are combinatorial patterns built from the selected gene and the signatures not only classify the samples in the respective group but also explain further the properties of each group. As we reported in [83], the logicome outcomes agree well with the literature findings. Apart from verified Boolean signatures, we recognized some combinatorial patterns which are yet to be biologically validated.

The amount of data regularly deposited into the public repositories is increasing, hence it becomes more difficult to analyze the data and extract specific information. The data-based logicome approach aims to provide a methodology where the size of the data is reduced by subseting the data of significant genes and complexity is reduced through the discretization of the extracted data. The proposed method aims to provide the general research community with a better understanding of the nature of the data and derive meaningful simplified explanation out of the large domain of data in terms of Boolean signatures.

In [83], it is shown that the Boolean signatures predict the combination of the upregulated/downregulated genes required to accurately classify the sample in the specific group. The obtained results are in good agreement with the literature findings.

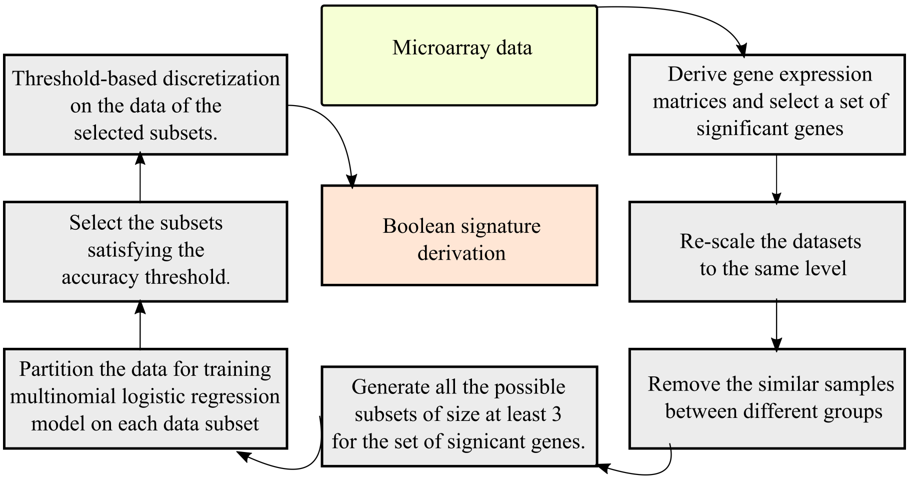

Microarray data
Derive gene expression matrices and select a set of significant genes
Re-scale the datasets to the same level
Remove the similar samples between different groups
Generate all the possible subsets of size at least 3 for the set of signicant genes.
Partition the data for training multinomial logistic regression model on each data subset
Select the subsets satisfying the accuracy threshold.
Threshold-based discretization on the data of the selected subsets.
Boolean signature derivation

Figure 2.4: The data-based logicome approach. The detailed outline of the methodology is depicted in the figure presented in [83].

Chapter 3

Reaction Systems

Gaining a precise understanding of the individual mechanisms that drive processes inside living cells is an active field of research and has inspired researchers to explore more towards solving computationally difficult problems. For example, see [19, 98, 111, 57]. In [58], authors discussed a number of research fields, such as cellular automata, membrane computing, neural computation, evolutionary computation, and swarm intelligence that attempt to investigate the natural phenomena in terms of information processing and connect the phenomena with the computing paradigms. In 2004, A. Ehrenfeucht and G. Rozenberg [35] introduced a formal framework of Reaction Systems (RS) with the purpose of investigating the interactions between the biochemical reactions inside the living cell. In this chapter, we discuss the reaction systems framework, its applicability to formalize different dynamical systems and our reaction systems models demonstrating the dynamics of the corresponding quantitative models.

3.1 Reaction Systems framework

Reaction system is a qualitative modelling framework where reactions are formalized using simple set-theory notions. The basic notions and properties of the reaction systems framework were first introduced in [35, 37]. Many studies in the domain of reaction systems have investigated properties of RS framework to formalize computer science and biology oriented problems, see for example [22, 28, 48, 64, 91, 90, 13, 84].

The functioning of reaction systems is based on the mechanisms of facilitation and inhibition: a reaction is enabled only if all its reactants needed to facilitate the reaction are present and all the inhibitors of the reaction are absent from the environment. Based on this mechanism a reaction is formalized as a triplet: $a = (R_a, I_a, P_a)$, where the sets R_a, I_a, P_a stand for finite non-empty sets of *reactants*, *inhibitors*, *products* of a, respectively with

$R_a \cap I_a = \emptyset$. The set $R_a \cup I_a$ contains the entities that directly influence a either as reactants or as inhibitors.

A *reaction system* is defined as an ordered pair $\mathcal{A} = (S, A)$, where A is a finite set of reactions and S is a finite background set of entities influencing reactions of A. If R_a, I_a, $P_a \subseteq S$, then a is a reaction in S, and $rac(S)$ is the set of all reactions with $R_a, I_a, P_a \subseteq S$ and $A \subseteq rac(S)$.

The formal model of reaction systems was first introduced in [37], which works based on two fundamental assumptions: (i) the threshold principle, and (ii) the no-permanency principle. The threshold principle assumes that either a resource is available and in such case, it is available enough, or it is absent. This means that, there is no counting involved in the interactions taking place in the RS framework. The principle of no permanency assumes that an object (molecule or entity) ceases to exit from the environment unless it is produced or sustained by a reaction. Thus, there is an immediate decay of a resource [34].

3.2 Applicability of reaction systems

In this section, we highlight key research topics emphasizing the special properties and notions for extending the reaction systems framework to successfully formalize and to execute different dynamic systems. The research topic mentioned here is the motivation for us to present reaction systems as a qualitative counterpart to the quantitative modeling.

The original motivation behind introducing the reaction systems framework was to model interactions between biochemical reactions. The overall interactions in a reaction systems model are driven by the context elements provided by the external environment. The reaction systems create their own well-defined structure of interactive processes which makes them different from other abstract formalisms [22].

Since there is no counting in the reaction systems framework, it is a qualitative model. There have been considerable efforts made by the research community to develop different notions of reaction systems, for making the reaction systems framework able to address certain situations emerging from dynamic systems. For example, in [38] the authors present reaction systems with measurements involving time as a measurement function. Also, reaction systems models dealing with more quantitative aspects of processes in living cells are discussed in [16, 32, 33]. The work in [36] presents notions of modules in order to obtain the evolutionary sequence of events typically seen in biochemical developmental processes. The study of interactive processes of reaction systems with duration is reported in [18] where, by introducing duration, the “immediate decay” property of the reaction systems model is relaxed. The research on static/structural and dynamic cause-effect relationships in reaction systems is systematically conducted in [17].

Next, we present some of the efforts in building a connection between computer science and biologically oriented phenomena, through reaction systems models. In [22], the authors explore the reaction systems modeling framework to formalize regulation of gene expression in *lac* operon, a genetic regulatory network involved in the metabolism of carbon sources in *E.coli* bacterial cells and also, demonstrate an implementation of reaction systems as an interactive version of the tower of Hanoi algorithm. The dynamics produced from the reaction system model built for the regulation of gene expression in *lac* operon is in correspondence with the response of cells to changing natural environmental conditions. The interactive process generated to solve the puzzle of tower of Hanoi solves the puzzle by producing correct sequences of states and movements. Also, in [22] and [37] the authors demonstrate the translation between Boolean functions and reaction systems model.

The work reported in [10] presents the reaction systems model built for heat shock response and shows the correlation between qualitative behavior emerging from the reaction systems model and the quantitative behavior emerging from the corresponding ODE model. The detailed study on mass conservation revealed from the internal structure of the reaction systems is presented in [9], which also introduces an automating tool for executing reaction systems models.

The biological phenomenon can be studied faithfully when the complexity and the less relevant information of the phenomenon is reduced according to the need. That is achieved with the reaction systems framework with its capability of supplying and absorbing the objects at certain fixed steps.

3.3 From quantitative models to Reaction Systems models

In this section, we describe how we construct the reaction systems models that exhibit the complex dynamics closely resembling the dynamics of the corresponding quantitative models. We demonstrate dynamics of self-assembly of intermediate filaments and period doubling cascade with the interaction processes emerging from reaction systems models.

3.3.1 Reaction systems model for self-assembly of intermediate filaments

We translated the molecular model of self-assembly of intermediate filaments presented in [24] into a Reaction Systems-based (RS-based) model. In Table 3.1, we list the reactions of the molecular model of self-assembly of inter-

Table 3.1: The direct translation of the biochemical reactions of the basic model adopted from [24], to a reaction system $\mathcal{A} = (S, A)$ where $S = \{\mathsf{T}, \mathsf{O}, \mathsf{H}, \mathsf{U}, \mathsf{F}, \mathsf{d_l}\}$.

Reaction in the chemical network	Reaction in the reaction system	
$2\,\mathsf{T} \to \mathsf{O}$	$(\{\mathsf{T}\}, \{\mathsf{d_l}\}, \{\mathsf{O}\})$	(3.1)
$2\,\mathsf{O} \to \mathsf{H}$	$(\{\mathsf{O}\}, \{\mathsf{d_l}\}, \{\mathsf{H}\})$	(3.2)
$2\,\mathsf{H} \to \mathsf{U}$	$(\{\mathsf{H}\}, \{\mathsf{d_l}\}, \{\mathsf{U}\})$	(3.3)
$2\,\mathsf{U} \to \mathsf{F}$	$(\{\mathsf{U}\}, \{\mathsf{d_l}\}, \{\mathsf{F}\})$	(3.4)
$\mathsf{F} + \mathsf{T} \to \mathsf{F}$	$(\{\mathsf{F}, \mathsf{T}\}, \{\mathsf{d_l}\}, \{\mathsf{F}\})$	(3.5)
$\mathsf{F} + \mathsf{U} \to \mathsf{F}$	$(\{\mathsf{F}, \mathsf{U}\}, \{\mathsf{d_l}\}, \{\mathsf{F}\})$	(3.6)
$2\,\mathsf{F} \to \mathsf{F}$	$(\{\mathsf{F}\}, \{\mathsf{d_l}\}, \{\mathsf{F}\})$	(3.7)

Table 3.2: An interactive process for the basic RS model in Table 3.1. The interactive process enters a loop after the third state from which every state contains all species of the system.

State	C_i	D_i	W_i
0	$\{\mathsf{T}\}$	$\emptyset$	$\{\mathsf{T}\}$
1	$\{\mathsf{T}\}$	$\{\mathsf{O}\}$	$\{\mathsf{T}, \mathsf{O}\}$
2	$\{\mathsf{T}\}$	$\{\mathsf{O}, \mathsf{H}\}$	$\{\mathsf{T}, \mathsf{O}, \mathsf{H}\}$
3	$\{\mathsf{T}\}$	$\{\mathsf{O}, \mathsf{H}, \mathsf{U}\}$	$\{\mathsf{T}, \mathsf{O}, \mathsf{H}, \mathsf{U}\}$
4	$\{\mathsf{T}\}$	$\{\mathsf{O}, \mathsf{H}, \mathsf{U}, \mathsf{F}\}$	$\{\mathsf{T}, \mathsf{O}, \mathsf{H}, \mathsf{U}, \mathsf{F}\}$
5	$\{\mathsf{T}\}$	$\{\mathsf{O}, \mathsf{H}, \mathsf{U}, \mathsf{F}\}$	$\{\mathsf{T}, \mathsf{O}, \mathsf{H}, \mathsf{U}, \mathsf{F}\}$

mediate filaments and the corresponding reactions in the RS-based model. The model in its basic form, its refined version and the refined model in different variants can be found in [11]. The intermediate filaments are one of the three types of protein filaments that together with the attached proteins sustain the mechanical strength of the cell, control the shape of the cell and drive/guide the cellular movement [65]. In the initial stage, the vimentine proteins associate laterally to form dimers and then subsequently form tetramers denoted by (T). The rapid lateral association of tetramers yield short filaments called *unit length filaments* denoted by (U). The longitudinal association of unit length filaments forms an elongated filament denoted by (F) and in the next level, the elongated filament (F) elongates with other filaments and itself. The model illustrated in Table 3.1 focuses on the two phases of the assembly: (i) formation of unit length filaments represented through the reactions (3.1), (3.2), (3.3), and (ii) longitudinal annealing of unit length filaments and elongated and further grown filaments represented through the reactions (3.4), (3.5), (3.6), (3.7).

The interactive process presented in Table 3.2 portrays the dynamics of the self-assembly of intermediate filaments represented by the molecular model in Table 3.1. The molecular model shows that the tetramers are always present from the beginning and in the reaction systems model the presence of tetramers in the environment is maintained through the context.

3.3.2 Reaction systems model producing complex dynamics

In [12], we aimed to demonstrate the reaction systems framework as a natural correspondence to sophisticated quantitative modeling concepts such as multi-stability, limit cycles and bifurcation. The reaction systems model exhibiting bi-stability, limit cycles, and period doubling cascade are presented in [12]. The RS models for multi-stability or limit cycles are explicitly constructed as correspondents of quantitative reaction-based models. The model representing period doubling cascade is built on the foundation of the binary counter RS model introduced by [37].

Table 3.3: Reaction systems model for period doubling bifurcation

Set of reactions	Set of reactions with n=3
$a_{10} = (\{e_1\}, \{e_0, t\}, e_1)$ $a_{ij} = (\{e_i\}, \{e_j, t, 1, \ldots, i-1\}, \{e_i\})$ for all i, j such that $1 \leq j < i \leq n$	$a_{10} = (\{e_1\}, \{e_0, t\}, e_1)$ (3.8) $a_{21} = (\{e_2\}, \{e_1, t, 1\}, e_2)$ (3.9) $a_{31} = (\{e_3\}, \{e_1, t, 1, 2\}, e_3)$(3.10) $a_{32} = (\{e_3\}, \{e_2, t, 1, 2\}, e_3)$(3.11)
$b_1 = (\{e_0\}, \{e_1, t\}, \{e_1\})$ $b_i = (\{e_0, \ldots, e_{i-1}\}, \{e_i, 1, \ldots, i-1, t\}, \{e_i\})$ for all i such that $2 \leq i \leq n$	$b_1 = (\{e_0\}, \{e_1, t\}, e_1)$ (3.12) $b_2 = (\{e_0, e_1\}, \{e_2, 1, t\}, e_2)$ (3.13) $b_3 = (\{e_0, e_1, e_2\}, \{e_3, 1, 2, t\}, e_3)$ (3.14)
$r_1 = (\{e_0\}, \{t\}, \{e_0\})$ $r_2 = (\{e_0, t\}, \{e_1, \ldots, e_n\}, \{e_0\})$	$r_1 = (\{e_0\}, \{t\}, e_0)$ (3.15) $r_2 = (\{e_0, t\}, \{e_1, e_2, e_3\}, e_0)$ (3.16)
$l = (\{t\}, \{e_0\}, e_0)$	$l = (\{t\}, \{e_0\}, e_0)$ (3.17)
$q_i = (\{e_0, \ldots, e_i, i\}, \{t\}, \{t\})$ for all i such that $1 \leq i \leq n$	$q_1 = (\{e_0, e_1, 1\}, \{t\}, t)$ (3.18) $q_2 = (\{e_0, e_1, e_2, 2\}, \{t\}, t)$ (3.19) $q_3 = (\{e_0, e_1, e_2, e_3, 3\}, \{t\}, t)$ (3.20)
$s_i = (\{i\}, \{t\}, \{i\})$ for all i such that $1 \leq i \leq n$	$s_1 = (\{1\}, \{t\}, 1)$ (3.21) $s_2 = (\{2\}, \{t\}, 2)$ (3.22) $s_3 = (\{3\}, \{t\}, 3)$ (3.23)

As an illustration, we present a period doubling cascade visualized with

Table 3.4: The interaction Process of the reaction systems model for Period Doubling (n = 3)

State	C	D	W	Applicable reactions from Table 3.3
0	$e_0, 1$	$\emptyset$	$e_0, 1$	3.15, 3.12, 3.21
1	$\emptyset$	$e_0, e_1, 1$	$e_0, e_1, 1$	3.18
2	$\emptyset$	t	t	3.17
3	2	e_0	$e_0, 2$	3.12, 3.15, 3.22
4	$\emptyset$	$e_0, e_1, 2$	$e_0, e_1, 2$	3.13, 3.15, 3.22
5	$\emptyset$	$e_0, e_2, 2$	$e_0, e_2, 2$	3.9, 3.12, 3.15, 3.22
6	$\emptyset$	$e_0, e_1, e_2, 2$	$e_0, e_1, e_2, 2$	3.19
7	$\emptyset$	t	t	3.17
8	3	e_0	$e_0, 3$	3.12, 3.15, 3.23
9	$\emptyset$	$e_0, e_1, 3$	$e_0, e_1, 3$	3.15, 3.13, 3.23
10	$\emptyset$	$e_0, e_2, 3$	$e_0, e_2, 3$	3.9, 3.12, 3.15, 3.23
11	$\emptyset$	$e_0, e_1, e_2, 3$	$e_0, e_1, e_2, 3$	3.14, 3.15, 3.23
12	$\emptyset$	$e_0, e_3, 3$	$e_0, e_3, 3$	3.10, 3.12, 3.15, 3.23
13	$\emptyset$	$e_0, e_1, e_3, 3$	$e_0, e_1, e_3, 3$	3.11, 3.13, 3.15, 3.23
14	$\emptyset$	$e_0, e_2, e_3, 3$	$e_0, e_2, e_3, 3$	3.9, 3.10, 3.12, 3.15, 3.23,
15	$\emptyset$	$e_0, e_1, e_2, e_3, 3$	$e_0, e_1, e_2, e_3, 3$	3.20
16	$\emptyset$	t	t	3.17
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

a reaction system framework. In period doubling bifurcations, the system becomes unstable as the parameter value increases. The instability is indicated by the regular periodic orbits with all the periods k, 2k, 4k, 8k, . . . so that it contains precisely one orbit of period k. For a graphical illustration we refer the figures depicting period doubling cascade presented in [12].

Table 3.3 presents the reaction systems model demonstrating the period doubling behavior and the corresponding set of reactions for the periods $n \in \{1, 2, 3\}$. The interactive process for switching subsequently from period 1 to period 2 and from period 2 to period 3 is presented in Table 3.4. In order to switch from one period to another, the new period is introduced by the context which is denoted by n. In the RS model for period doubling, e_i represents 1 on the position 2^i where $0 \leq i \leq n$ and $\{e_0, n\}$ indicate the initial state of the model. For the initial state $\{e_0, n\}$, the model enters into an orbit of length 2^n and the model introduces t (denotes termination of the current period) when the count has reached $2^n - 1$.

Chapter 4

Summaries of the Included Articles

4.1 Paper 1: Generating the Logicome of a Biological Network

- *Charmi Panchal, Sepinoud Azimi, and Ion Petre. Generating the Logicome of a Biological Network. In: María Botón-Fernández, Carlos Martín-Vide, Sergio Santander-Jiménez, Miguel A. Vega-Rodríguez (eds). Algorithms for Computational Biology. Lecture Notes in Computer Science, volume 9702. Springer International Publishing, 2016.*

Article [82] presents a methodology that infers activation dependencies between the selected key nodes from a biological network. These dependencies are derived as logical formulations obtained as Boolean network. For this study we begin with the modularized ODE-based model of the EGFR (Epidermal Growth Factor Receptor) pathway [79, 78] and a set of key nodes that play a significant role in the pathway.

The model is simulated for all the possible knock-out mutants generated by making the key nodes active/inactive in all possible combinations. The knock-out simulations are performed in COPASI [51] and the results of each simulation are incorporated with threshold criteria. Depending upon the choice of the threshold, the discretization step translates data into “1” (active) and “0” (inactive). We derived Boolean networks inferring the relationships between the selected nodes and analyzed them for different choices of threshold. Also, we derived results for the case when several knock-out mutants are not available. The collected logicome outcomes are compared and conclusions are derived from the overall analysis.

Final conclusions derived from the results suggest that the logicome approach allows the modeler to focus on the selected key nodes while abstracting away from the rest of the network and obtain a high-level understanding of the functionalities within the network even in the case of unavailability of some portion of the data. The method's outcome depends on both the numerical setup of the basic model and the choice of the threshold value. The method is practical as long as the number of key nodes n is such that it is possible to run 2^n simulations.

4.2 Paper 2: Generating the Logicome from Microarray Data

- *Charmi Panchal, and Vladimir Rogojin. Generating the Logicome from Microarray Data. In Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2017 IEEE Conference on, pp. 1-8. IEEE, 2017.*

In [83], we follow the research intuition presented in 4.1. This article presents a method for building a data-driven logicome: the method for building a set of small Boolean expressions as classifiers for disjoint groups within the microarray datasets. A method requires to choose a set of significant genes from the considered microarray datasets. We choose them using differential gene expression analysis performed in web interface GEO2R [14]. We proceed with the selected significant genes and the respective subset of the gene expression data from the large datasets. The data preprocessing is performed with R package GEOquery [26]. The method employs a machine learning approach on pre-processed data and finds minimal subsets of significant genes that maintain the accuracy threshold for classification. The expression matrix of minimal subsets of significant genes is discretized and reduced to “1” (up-regulated) and “0” (down-regulated). In the matrix, a row is viewed as a binary combination of genes and the frequency of occurrence of each binary combination in every group is calculated. We employ probability and coverage threshold, and select the representative combinations. The final outcome of the method is Boolean signature (logical formula as disjunctive normal form) constructed from the representative Boolean combinations.

We conclude that outcomes from the typical machine learning and statistics approaches usually provide predictions and classifications but lack the information about the internal structure of the system under studies and relations between its components. Our method captures the most representative patterns in the input datasets for each of the clusters/categories/groups and generates small and simple Boolean classifiers for them. These signatures can be utilized to elaborate the properties of each category.

4.3 Paper 3: Reaction Systems Models for the Self-Assembly of Intermediate Filaments

- *Sepinoud Azimi, Charmi Panchal, Eugen Czeizler, and Ion Petre. Reaction systems models for the self-assembly of intermediate filaments. Annals of University of Bucharest,62:9–24, Editura Universităţii din Bucureşti, 2015.*

Article [11] demonstrates the expressivity of the reaction systems as a modelling framework that is able to capture dynamics of the self-assembly process of intermediate filaments. We built reaction systems models for self-assembly of intermediate filaments based on the molecular model presented in [24] and [76]. We have presented both basic and refined models of intermediate filaments using formalism of reaction systems. We compared the dynamics of the reaction systems model with the corresponding ODE-based model. Besides basic and refined models, we present different versions of the RS-based model to control the length of the filaments produced within the system. We conclude that the reaction systems framework is a good qualitative counterpart to quantitative modeling frameworks such as ODE. The Reaction systems framework is a simple set theory-based framework and with this we could produce similar behavior as that is produced with the ODE framework.

4.4 Paper 4: Multi-Stability, Limit Cycles, and Period-Doubling Bifurcation with Reaction Systems

- *Sepinoud Azimi, Charmi Panchal, Andrzej Mizera and Ion Petre. Multi-Stability, Limit Cycles, and Period-Doubling Bifurcation with Reaction Systems. To appear in International Journal of Foundations of Computer Science, 2018.*

Article [12] introduces notions of Reaction Systems (RS) reproducing dynamical behaviors such as multi-stability, limit cycles and period-doubling bifurcations. The systems with multi-stability have multiple distinct steady states, whereas a system with limit cycle behavior exhibits sustained oscillations or a closed curve converging to a steady state. The period doubling cascade is a series of period doubling bifurcations in which a small change in the parameter value causes the system to switch to a new behavior with a doubling of period. In quantitative models, the dynamics is controlled through the numerical parameters, whereas in the reaction systems models such dynamics are represented through the choice of the context sequence.

The reaction systems model expressing multi-stability is built as a correspondent of an ODE-based model of a minimal bi-stable system operating in two distinct stable steady states. The RS model representing bi-stability takes a system from one state to another state with a low or a high signal provided through context. The reaction systems model reproducing limit cycle behavior is built through a small modification in the RS model of bi-stability. The RS model of limit cycles corresponds well to the phenomenon of limit cycles where the model may either have a steady state or it may eventually cycle between states. The reaction systems model for a period doubling cascade is built through the modification of the RS model performing a binary counter. The RS model of period doubling bifurcation facilitates the system to transit from one period to the other one where the period i is labeled with a binary number between 0 and $2^i - 1$.

We conclude that reaction systems are a natural correspondence to several quantitative modeling concepts and they provide transparent causality between events and an explicit formulation of the mechanisms responsible for triggering events. We reason about the clear advantage of reaction systems as a modeling framework alongside traditional modeling frameworks.

Chapter 5

Conclusion and Future work

The research presented in this doctoral thesis comprises developing and utilizing different methodologies originating from computer science. The central objective of these methodologies is to produce discrete logic-based and formal models that reproduce the known high-level knowledge and the complex dynamics, lying within the biologically motivated models. We studied several biological phenomena to describe them using qualitative modeling and study the complex relationships lying within the phenomena. In particular, the presented research has been divided into two parts.

In the first part, we developed logicome methodologies to infer qualitative models from a given biological phenomenon. The logicome methodologies allow modeler to derive static snapshot of the detailed phenomena using the selected set of key elements. This static snapshot is obtained in the form of logical formula that provides high-level understanding of the functionality or characteristics of the phenomena. The inferred qualitative models leverage the construction of predictive tools in cases of incomplete information in the case study. The logicome methods aim to reproduce the high-level knowledge of the biological phenomenon without portraying complex details of the phenomenon. In a logicome, the model is reduced to its simplified version with model abstraction. The model abstraction is determined by the number of key/significant components selected by the modeler. The method focuses on the selected components and the outcome produced as logical formulas. The outcomes produced with logicome approaches are small and easily interpretable predictive models with the binarized selected components as input. The obtained logic models are a good qualitative approximation of characteristics observed in the biological phenomena. The outcomes are not only straightforward and easy to understand, but also in good agreement with the literature based knowledge.

As case studies for the logicome approach, we have chosen two biological phenomena: (1) the epidermal growth factor receptor (EGFR) signaling pathway (2) microarray gene expression data of head and neck/oral squamous cell carcinoma (HNOSCC). In their simplest form, the logicome methods allow the modeler to select the components from the case study under consideration and permit the selected components to be in one of the discrete states: “1” (symbolizes ON or active or up-regulated) or “0” (symbolizes OFF or inactive or down-regulated). For instance, for the EGFR signaling pathway, we use interface species as the selected key elements and assign the element to *active or inactive*, and for HNOSCC microarrray data, we use differential expression analysis to select significant genes and translate their expression values to *up-regulated or down-regulated.*

In [82], we introduced the logicome methodology (model-based logicome) for the numerical model of EGFR signaling pathway. The main steps involved in the methodology are: selection of key elements, discretization, knock-out mutant simulations and logicome outcome. The logicome outcome produces high-level understanding of the pathway functionality and complements the numerical model. The outcome is formulated as a boolean network model which solely focuses on the selected key elements and describes activation conditions between the elements. Furthermore, the boolean network model allows global analysis of the model dynamics and compensates for the lack of model data. The efficiency of model-based logicome approach depends on numerical set-up of the model, choice of the key elements and choice of the threshold. Among many possible steps for an advancement of the current model based logicome approach, a step towards going from boolean logic to many valued logic would enhance the flexibility of the methodology. Moreover, involving different discretization techniques could possibility further improve the efficiency of the method.

In [83], we introduced the logicome methodology (data-based logicome) for the HNOSCC microarray data. The main steps involved in the methodology are: selecting the significant genes, data preprocessing (reducing probe expression to gene expressions), finding the minimal subset of significant genes (applying multinomial logistic regression), discretization, and deriving the logicome outcome. In this case, the logicome outcome is a set of logical rules which determine categorization of the samples into control or various types of cancer groups. The set of logical rules represent most occurring patterns in the respective sample groups. This logical rules can further aid in development of accurate detailed models. In the present data-based logicome approach, it is remained to measure the performance of the method when different classification methods are used. Moreover the potential of the method would further improve with the implementation on more publicly available “gold standard” cancer-related microarray datases.

The outcomes produced with model-based and data-based logicome approaches provide more abstract, systematic and objective description of the case-studies. The logicome serves as global blueprint for the detailed case-studies built from different units. The logicome methodologies aim at providing realistic snap-shot of the phenomena using the main players within the phenomena. This snap-shot is not only useful in overcoming the difficulties caused by massive amount of details but also helps in accurate analysis and prediction of the system's behavior. The main appeal of logicome approaches over other logic based approaches is its relative ability to derive a high-level comprehensible logic based models and produce a description that help understanding system in a simple term.

In the second part, we used the qualitative modeling framework of reaction systems that successfully investigates dynamic behavior of the biologically inspired case studies.

The reaction systems were first introduced in [37] as a formal model with the objective to formalize the interactions between biochemical reactions. The interactions are influenced by two main mechanisms: facilitation and inhibition. The main potentials of the reaction systems framework are its expressive power and flexibility, to give qualitative aspects to most multidisciplinary quantitative problems that range from computer science to biology.

we demonstrated that the sophisticated behavior such as period doubling, multi-stability, and limit cycles can be obtained through small models based only on the elementary tools we have in reaction systems. This is in contrast with the usual way such behavior is demonstrated, through numerical setup and ODE-based models, and it demonstrates that the fundamental source for such behavior is the structure of the interactions in the model, rather than its numerical setup.

In [11], we presented reaction systems models for self-assembly of intermediate filaments. We start with the ODE-based models and analyze the dynamics of the self-assembly process under several modifications. In the reaction systems framework, we built models that are equally versatile and produce the same behavior as that of ODE models. In [12], we presented notions as well as models exhibiting system level behavior such as multi-stability, limit cycles and period doubling bifurcation.

The clear advantage of modeling with reaction systems is the causalities between the reactions are directly visible, meaning, it is easy to understand how entities participating in the reactions influence each other and it is possible to gain detailed insights into the structure of the system under study.

A major challenge in examining biological case studies is to shift from descriptive narratives towards comprehensible explanations of general mechanisms and processes. In the efforts to address this challenge, we propose the above-mentioned research directions that capture, investigate and give a clear view of the underlying mechanisms by abstracting away the complicated details of the phenomenon. As a part of future research directions, we suggest to develop more powerful logic-based methods which permit components to be in more than two discrete states. With such a model, it would be feasible to investigate intermediate states of the input/output components and capture several aspects of their intermediate connections. This can aid in constructing more powerful and reliable models with *ad-hoc* knowledge of the phenomenon. It would be interesting to expand our research on the multi-scale biological phenomena which can be explained precisely by observing and integrating responses at different resolutions. The modeling efforts that aim at exploring such multi-scale systems incorporate different modeling techniques including logic-based formalisms. There are many efforts presenting the feasibility of reaction systems as a modeling framework for multidisciplinary applications. The possible research directions in the field of reaction systems is to investigate fundamental structural properties of even larger and diverse case studies. Moreover, several potential challenges can be addressed by utilizing the expressive power of reaction systems with formalizing notions that facilitate building reaction systems models for any given case study.

Bibliography

[1] LogicFriday. `http://sontrak.com/downloads.html`.

[2] A quantitative kinetic model for the in vitro assembly of intermediate filaments from tetrameric vimentin. *The Journal of Biological Chemistry*, 282(25):18563–72, June 2007.

[3] Wassim Abou-Jaoudé, Pauline Traynard, Pedro T Monteiro, Julio Saez-Rodriguez, Tomáš Helikar, Denis Thieffry, and Claudine Chaouiya. Logical modeling and dynamical analysis of cellular networks. *Frontiers in genetics*, 7, 2016.

[4] Tatsuya Akutsu, Satoru Miyano, and Satoru Kuhara. Identification of genetic networks from a small number of gene expression patterns under the boolean network model. In *Pacific Symposium on Biocomputing*, volume 4, pages 17–28, 1999.

[5] Tatsuya Akutsu, Satoru Miyano, and Satoru Kuhara. Inferring qualitative relations in genetic networks and metabolic pathways. *Bioinformatics*, 16(8):727–734, 2000.

[6] Réka Albert, James J. Collins, and Leon Glass. Introduction to focus issue: quantitative approaches to genetic networks. *Chaos*, 23(2):025001, 2013.

[7] Réka. Albert and Hans G. Othmer. The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in drosophila melanogaster. *Journal of Theoretical Biology*, 223(1):1–18, 2003.

[8] Ido Amit, Ron Wides, and Yosef Yarden. Evolvable signaling networks of receptor tyrosine kinases: relevance of robustness to malignancy and to cancer therapy. *Molecular Systems Biology*, 3(1):151, 2007.

[9] Sepinoud Azimi, Cristian Gratie, Sergiu Ivanov, and Ion Petre. Dependency graphs and mass conservation in reaction systems. *Theoretical Computer Science*, 598:23–39, 2015.

[10] Sepinoud Azimi, Bogdan Iancu, and Ion Petre. Reaction system models for the heat shock response. *Fundamenta Informaticae*, 131(3):299–312, 2014.

[11] Sepinoud Azimi, Charmi Panchal, Eugen Czeizler, and Ion Petre. Reaction systems models for the self-assembly of intermediate filaments. *Annals of University of Bucharest*, 62(2):9–24, 2015.

[12] Sepinoud Azimi, Charmi Panchal, Andrzej Mizera, and Ion Petre. Multi-stability, limit cycles, and period-doubling bifurcation with reaction systems. *TUCS Technical Reports*, 1167:20–25, 2016.

[13] Roberto Barbuti, Roberta Gori, Francesca Levi, and Paolo Milazzo. Investigating dynamic causalities in reaction systems. *Theoretical Computer Science*, 623:114–145, 2016.

[14] Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Michelle Holko, et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research*, 41(D1):D991–D995, 2013.

[15] Ezio Bartocci and Pietro Lió. Computational modeling, formal analysis, and tools for systems biology. *PLoS computational biology*, 12(1):e1004591, 2016.

[16] Robert Brijder, Andrzej Ehrenfeucht, Michael Main, and Grzegorz Rozenberg. A tour of reaction systems. *International Journal of Foundations of Computer Science*, 22(07):1499–1517, 2011.

[17] Robert Brijder, Andrzej Ehrenfeucht, and Grzegorz Rozenberg. A note on causalities in reaction systems. *Electronic Communications of the EASST*, 30, 2010.

[18] Robert Brijder, Andrzej Ehrenfeucht, and Grzegorz Rozenberg. Reaction systems with duration. In *Computation, Cooperation, and Life - Essays Dedicated to Gheorghe Paun on the Occasion of His 60th Birthday*, pages 191–202. Springer, 2011.

[19] Cris Calude and Gheorghe Paun. *Computing with cells and atoms: an introduction to quantum, DNA and membrane computing.* CRC Press, 2000.

[20] Filippo Castiglione, Francesco Pappalardo, Carlo Bianca, Giulia Russo, and Santo Motta. Modeling biology spanning different scales: an open challenge. *BioMed Research International*, 2014, 2014.

[21] Claudine Chaouiya. Petri net modelling of biological networks. *Briefings in Bioinformatics*, 8(4):210–219, 2007.

[22] Luca Corolli, Carlo Maj, Fabrizio Marini, Daniela Besozzi, and Giancarlo Mauri. An excursion in reaction systems: From computer science to biology. *Theoretical Computer Science*, 454:95–108, 2012.

[23] Francis Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

[24] Eugen Czeizler, Andrzej Mizera, Elena Czeizler, Ralph-Johan Back, John E Eriksson, and Ion Petre. Quantitative analysis of the self-assembly strategies of intermediate filaments from tetrameric vimentin. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 9(3):885–898, 2012.

[25] Maria I. Davidich and Stefan Bornholdt. Boolean network model predicts cell cycle sequence of fission yeast. *PloS one*, 3(2):e1672, 2008.

[26] Sean Davis and Paul S Meltzer. Geoquery: a bridge between the gene expression omnibus (geo) and bioconductor. *Bioinformatics*, 23(14):1846–1847, 2007.

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

[28] Alberto Dennunzio, Enrico Formenti, and Luca Manzoni. Reaction systems and extremal combinatorics properties. *Theoretical Computer Science*, 598:138–149, 2015.

[29] Robin Donaldson, Carolyn Talcott, Merrill Knapp, and Muffy Calder. Understanding signalling networks as collections of signal transduction pathways. In *Proceedings of the 8th International Conference on Computational Methods in Systems Biology*, pages 86–95. ACM, 2010.

[30] Friederike Dündar, Luce Skrabanek, and Paul Zumbo. Introduction to differential gene expression analysis using rna-seq. *Applied Bioinformatics Core/Weill Cornell Medical College*, pages 1–67, 2015.

[31] Ron Edgar, Michael Domrachev, and Alex E. Lash. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210, 2002.

[32] Andrzej Ehrenfeucht, Jetty Kleijn, Maciej Koutny, and Grzegorz Rozenberg. *Qualitative and quantitative aspects of a model for processes inspired by the functioning of the living cell.* Wiley Online Library, 2012.

[33] Andrzej Ehrenfeucht, Michael Main, and Grzegorz Rozenberg. Functions defined by reaction systems. *International Journal of Foundations of Computer Science*, 22(01):167–178, 2011.

[34] Andrzej Ehrenfeucht, Ion Petre, and Grzegorz Rozenberg. Reaction systems: A model of computation inspired by the functioning of the living cell. In *The Role of Theory in Computer Science - Essays Dedicated to Janusz Brzozowski*, pages 1–32, 2017.

[35] Andrzej Ehrenfeucht and Grzegorz Rozenberg. Basic notions of reaction systems. In *Developments in Language Theory, 8th International Conference, DLT 2004, Auckland, New Zealand, December 13-17, 2004, Proceedings*, pages 27–29. Springe, 2004.

[36] Andrzej Ehrenfeucht and Grzegorz Rozenberg. Events and modules in reaction systems. *Theoretical Computer Science*, 376(1-2):3–16, 2007.

[37] Andrzej Ehrenfeucht and Grzegorz Rozenberg. Reaction systems. *Fundamenta Informaticae*, 75(1):263–280, 2007.

[38] Andrzej Ehrenfeucht and Grzegorz Rozenberg. Reaction systems: a formal framework for processes based on biochemical interactions. *Electronic Communications of the EASST*, 26, 2010.

[39] Carlos Espinosa-Soto, Pablo Padilla-Longoria, and Elena R. Alvarez-Buylla. A gene regulatory network model for cell-fate determination during arabidopsis thaliana flower development that is robust and recovers experimental gene expression profiles. *The Plant Cell*, 16(11):2923–2939, 2004.

[40] Jasmin Fisher and Nir Piterman. The executable pathway to biological networks. *Briefings in Functional Genomics*, 9(1):79–92, 2010.

[41] Enrico Formenti, Luca Manzoni, and Antonio E. Porreca. Cycles and global attractors of reaction systems. In Helmut Jürgensen, Juhani Karhumäki, and Alexander Okhotin, editors, *Descriptional Complexity of Formal Systems*, volume 8614 of *Lecture Notes in Computer Science*, pages 114–125. Springer, 2014.

[42] Steven Gay, Sylvain Soliman, and François Fages. A graphical method for reducing and relating models in systems biology. *Bioinformatics*, 26(18):i575–i581, 2010.

[43] Samik Ghosh, Yukiko Matsuoka, Yoshiyuki Asai, Kun-Yi Hsin, and Hiroaki Kitano. Software for systems biology: from tools to integrated platforms. *Nature Reviews Genetics*, 12(12):821, 2011.

[44] Greg Gibson. Microarrays in ecology and evolution: a preview. *Molecular Ecology*, 11(1):17–24, 2002.

[45] Leon Glass and Stuart A. Kauffman. The logical analysis of continuous, non-linear biochemical control networks. *Journal of Theoretical Biology*, 39(1):103–129, 1973.

[46] Carito Guziolowski, Santiago Videla, Federica Eduati, Sven Thiele, Thomas Cokelaer, Anne Siegel, and Julio Saez-Rodriguez. Exhaustively characterizing feasible logic models of a signaling network using answer set programming. *Bioinformatics*, 29(18):2320–2326, 2013.

[47] John D Haley and William John Gullick. *EGFR signaling networks in cancer therapy.* Springer Science & Business Media, 2009.

[48] Mika Hirvensalo. On probabilistic and quantum reaction systems. *Theoretical Computer Science*, 429:134–143, 2012.

[49] William S. Hlavacek, James R. Faeder, Michael L. Blinov, Richard G. Posner, Michael Hucka, and Walter Fontana. Rules for modeling signal-transduction systems. *Sci STKE*, 2006(344):re6, 2006.

[50] Jörg D. Hoheisel. Microarray technology: beyond transcript profiling and genotype analysis. *Nature Reviews Genetics*, 7(3):200–210, 2006.

[51] Stefan Hoops, Sven Sahle, Ralph Gauges, Christine Lee, Jürgen Pahle, Natalia Simus, Mudita Singhal, Liang Xu, Pedro Mendes, and Ursula Kummer. COPASI – a complex pathway simulator. *Bioinformatics*, 22(24):3067–3074, 2006.

[52] Jorrit J. Hornberg, Bernd Binder, Frank J. Bruggeman, Birgit Schoeberl, Reinhart Heinrich, and Hans V. Westerhoff. Control of MAPK signalling: from complexity to what really matters. *Oncogene*, 24(36):5533–42, 2005.

[53] C Anthony Hunt, Glen EP Ropella, Sunwoo Park, and Jesse Engelberg. Dichotomies between computational and mathematical models. *Nature Biotechnology*, 26(7):737–738, 2008.

[54] Bogdan Iancu, Diana-Elena Gratie, Sepinoud Azimi, and Ion Petre. On the implementation of quantitative model refinement. In *Algorithms for Computational Biology*, pages 95–106. Springer, 2014.

[55] Rafael A Irizarry, Bridget Hobbs, Francois Collin, Yasmin D Beazer-Barclay, Kristen J Antonellis, Uwe Scherf, and Terence P Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003.

[56] Zhiwei Ji, Ke Yan, Wenyang Li, Haigen Hu, and Xiaoliang Zhu. Mathematical and computational modeling in complex biological systems. *BioMed research international*, 2017, 2017.

[57] Kenneth De Jong. Evolutionary computation. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(1):52–56, 2009.

[58] Lila Kari and Grzegorz Rozenberg. The many facets of natural computing. *Communications of the ACM*, 51(10):72–83, 2008.

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

[60] Stuart A. Kauffman. Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of Theoretical Biology*, 22(3):437–467, 1969.

[61] Boris N. Kholodenko. Cell-signalling dynamics in time and space. *Nature Reviews Molecular Cell Biology*, 7(3):165–176, 2006.

[62] Hiroaki Kitano. Perspectives on systems biology. *New Generation Computing*, 18(3):199–216, 2000.

[63] Steffen Klamt, Julio Saez-Rodriguez, Jonathan A Lindquist, Luca Simeoni, and Ernst D Gilles. A methodology for the structural and functional analysis of signaling and regulatory networks. *BMC Bioinformatics*, 7(1):56, 2006.

[64] Jetty Kleijn and Maciej Koutny. Membrane systems with qualitative evolution rules. *Fundamenta Informaticae*, 110(1):217–230, 2011.

[65] Elias Lazarides. Intermediate filaments as mechanical integrators of cellular space. *Nature*, 283:249–255, 1980.

[66] Nicolas Le Novère. Model storage, exchange and integration. *BMC Neuroscience*, 7(1):S11, 2006.

[67] Nicolas Le Novere. Quantitative and logic modelling of molecular and gene networks. *Nature Reviews Genetics*, 16(3):146–158, 2015.

[68] Nicolas Le Novere, Benjamin Bornstein, Alexander Broicher, Melanie Courtot, Marco Donizelli, Harish Dharuri, Lu Li, Herbert Sauro, Maria Schilstra, Bruce Shapiro, Jacky L. Snoep, and Michael Hucka. Biomodels database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Research*, 34(Database issue):D689–D691, 2006.

[69] Gregory Linshiz, Alex Goldberg, Tania Konry, and Nathan J Hillson. The fusion of biology, computer science, and engineering: towards efficient and successful synthetic biology. *Perspectives in Biology and Medicine*, 55(4):503–520, 2012.

[70] Catherine M. Lloyd, Matt D.B. Halstead, and Poul F. Nielsen. Cellml: its future, present and past. *Progress in Biophysics and Molecular Biology*, 85(2):433–450, 2004.

[71] Zheng Luo and Daniel H. Geschwind. Microarray applications in neuroscience. *Neurobiology of Disease*, 8(2):183–193, 2001.

[72] Shawn Martin, Zhaoduo Zhang, Anthony Martino, and Jean-Loup Faulon. Boolean dynamics of genetic regulatory networks inferred from microarray time series data. *Bioinformatics*, 23(7):866–874, 2007.

[73] Tom Melham. Modelling, abstraction, and computation in systems biology: A view from computer science. *Progress in Biophysics and Molecular Biology*, 111(2):129–136, 2013.

[74] Artur Męski, Wojciech Penczek, and Grzegorz Rozenberg. Model checking temporal properties of reaction systems. *Information Sciences*, 313:22–42, 2015.

[75] Melissa B. Miller and Yi-Wei Tang. Basic concepts of microarrays and potential applications in clinical microbiology. *Clinical Microbiology Reviews*, 22(4):611–633, 2009.

[76] Andrzej Mizera, Eugen Czeizler, and Ion Petre. Self-assembly models of variable resolution. In *Corrado Priami, Ion Petre, Erik de Vink (Eds.) Transactions on Computational Systems Biology XIV. Lecture Notes in Computer Science*, pages 181–203. Springer, Berlin, Heidelberg, 2012.

[77] Melody K. Morris, Julio Saez-Rodriguez, Peter K. Sorger, and Douglas A. Lauffenburger. Logic-based models for the analysis of cell signalling networks. *Biochemistry*, 49(15):3216–3224, 2010.

[78] Nicola Normanno, Antonella De Luca, Caterina Bianco, Luigi Strizzi, Mario Mancino, Monica R Maiello, Adele Carotenuto, Gianfranco De Feo, Francesco Caponigro, and David S Salomon. Epidermal growth factor receptor (EGFR) signaling in cancer. *Gene*, 366(1):2–16, 2006.

[79] Kanae Oda, Yukiko Matsuoka, Akira Funahashi, and Hiroaki Kitano. A comprehensive pathway map of epidermal growth factor receptor signaling. *Molecular Systems Biology*, 1(1), 2005.

[80] David T. Okou, Karyn Meltz Steinberg, Christina Middle, David J Cutler, Thomas J. Albert, and Michael E. Zwick. Microarray-based genomic selection for high-throughput resequencing. *Nature Methods*, 4(11):907–909, 2007.

[81] Jeffrey D. Orth, Ines Thiele, and Bernhard Ø. Palsson. What is flux balance analysis? *Nature Biotechnology*, 28(3):245–248, 2010.

[82] Charmi Panchal, Sepinoud Azimi, and Ion Petre. Generating the logicome of a biological network. In *María Botón-Fernández, Carlos Martín-Vide, Sergio Santander-Jiménez, Miguel A. Vega-Rodríguez (Eds.) Algorithms for Computational Biology 2016. Lecture Notes in Computer Science*, pages 38–49. Springer International Publishing, 2016.

[83] Charmi Panchal and Vladimir Rogojin. Generating the logicome from microarray data. In *2017 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 1–8. IEEE, 2017.

[84] Giovanni Pardini, Roberto Barbuti, Andrea Maggiolo-Schettini, Paolo Milazzo, and Simone Tini. Compositional semantics and behavioural equivalences for reaction systems with restriction. *Theoretical Computer Science*, 551:1–21, 2014.

[85] Gheorghe Păun, Mario J. Pérez-Jiménez, and Grzegorz Rozenberg. Bridging membrane and reaction systems–further results and research topics. *Fundamenta Informaticae*, 127(1):99–114, 2013.

[86] Huiming Peng, Weiling Zhao, Hua Tan, Zhiwei Ji, Jingsong Li, King Li, and Xiaobo Zhou. Prediction of treatment efficacy for prostate cancer using a mathematical model. *Scientific Reports*, 6:21599, 2016.

[87] Jose M. Ranz and Carlos A. Machado. Uncovering evolutionary patterns of gene expression using microarrays. *Trends in Ecology & Evolution*, 21(1):29–37, 2006.

[88] Anat Reiner, Daniel Yekutieli, and Yoav Benjamini. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, 19(3):368–375, 2003.

[89] Julio Saez-Rodriguez, Luca Simeoni, Jonathan A. Lindquist, Rebecca Hemenway, Ursula Bommhardt, Boerge Arndt, Utz-Uwe Haus, Robert Weismantel, Ernst D. Gilles, Steffen Klamt, et al. A logical model provides insights into T cell receptor signaling. *PLoS Computational Biology*, 3(8):e163, 2007.

[90] Arto Salomaa. Functions and sequences generated by reaction systems. *Theoretical Computer Science*, 466:87–96, 2012.

[91] Arto Salomaa. Applications of the chinese remainder theorem to reaction systems with duration. *Theoretical Computer Science*, 2015.

[92] Regina Samaga and Steffen Klamt. Modeling approaches for qualitative and semi-quantitative analysis of cellular signaling networks. *Cell Communication and Signaling*, 11(1):43, 2013.

[93] Marc A. Schaub, Thomas A. Henzinger, and Jasmin Fisher. Qualitative networks: a symbolic approach to analyze biological signaling networks. *BMC Systems Biology*, 1(1):4, 2007.

[94] Mark Schena, Dari Shalon, Ronald W. Davis, and Patrick O. Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467, 1995.

[95] Birgit Schoeberl, Claudia Eichler-Jonsson, Dieter Gilles Ernst, and Gertraud Müller. Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. *Nature Biotechnology*, 20(118):370 – 375, 2002.

[96] Parthasarathy Seshacharyulu, Moorthy P. Ponnusamy, Dhanya Haridas, Maneesh Jain, Apar K. Ganti, and Surinder K. Batra. Targeting the EGFR signaling pathway in cancer therapy. *Expert Opinion on Therapeutic Targets*, 16(1):15–31, 2012.

[97] Jacky L. Snoep and Brett G. Olivier. Java web simulation (jws); a web based database of kinetic models. *Molecular Biology Reports*, 29(1-2):259–263, 2002.

[98] Zoltan Szallasi, Jörg Stelling, and Vipul Periwal. *System Modeling in Cell Biology From Concepts to Nuts and Bolts*. The MIT Press, 2006.

[99] Carolyn Talcott. The pathway logic formal modeling system: Diverse views of a formal representation of signal transduction. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1468–1476. IEEE, 2016.

[100] D. Thieffry and R. Thomas. Qualitative analys is of gene networks. In *Biocomputing'98: Proceedings of the Pacific Symposium*, page 77. World Scientific, 1997.

[101] Xuming Tong, Jinghang Chen, Hongyu Miao, Tingting Li, and Le Zhang. Development of an agent-based model (abm) to simulate the immune system and integration of a regression method to estimate

the key abm parameters by fitting the experimental data. *PloS one*, 10(11):e0141295, 2015.

[102] Santiago Videla, Irina Konokotina, Leonidas G. Alexopoulos, Julio Saez-Rodriguez, Torsten Schaub, Anne Siegel, and Carito Guziolowski. Designing experiments to discriminate families of logic models. *Frontiers in Bioengineering and Biotechnology*, 3(131), 2015.

[103] Joseph Walpole, Jason A. Papin, and Shayn M. Peirce. Multiscale computational models of complex biological systems. *Annual Review of Biomedical Engineering*, 15:137–154, 2013.

[104] Dennis YQ Wang, Luca Cardelli, Andrew Phillips, Nir Piterman, and Jasmin Fisher. Computational modeling of the EGFR network elucidates control mechanisms regulating signal dynamics. *BMC Systems Biology*, 3(1):1–18, 2009.

[105] Rui-Sheng Wang, Assieh Saadatpour, and Reka Albert. Boolean modeling in systems biology: an overview of methodology and applications. *Physical Biology*, 9(5):055001, 2012.

[106] James M. Whitacre. Biological robustness: paradigms, mechanisms, and systems principles. *Frontiers in Genetics*, 3(67), 2012.

[107] Darren J. Wilkinson. Bayesian methods in bioinformatics and computational systems biology. *Briefings in Bioinformatics*, 8(2):109–116, 2007.

[108] John C. Wooley, Herbert S. Lin, National Research Council, et al. Computational modeling and simulation as enablers for biological discovery. 2005.

[109] Yosef Yarden. The EGFR family and its ligands in human cancer: signalling mechanisms and therapeutic opportunities. *European Journal of Cancer*, 37(4):3–8, 2001.

[110] Yosef Yarden and Ben-Zion Shilo. Snapshot: EGFR signaling pathway. *Cell*, 131(5):1018.e1—-1018.e2, 2007.

[111] Hector Zenil. A behavioural foundation for natural computing and a programmability test. In *Dodig-Crnkovic, Gordana and Giovagnoli, Raffaela (Eds.) Computing Nature: Turing Centenary Perspective*, pages 87–113. Springer Berlin Heidelberg, 2013.

Paper I

Generating the Logicome of a Biological Network

Charmi Panchal, Sepinoud Azimi, and Ion Petre

In María Botón-Fernández, Carlos Martín-Vide, Sergio Santander-Jiménez, Miguel A. Vega-Rodríguez (Eds). Algorithms for Computational Biology. Lecture Notes in Computer Science. Vol 9702. Springer International Publishing Switzerland 2016. The publication is available at Springer via https://doi.org/10.1007/978-3-319-38827-4_4

Generating the Logicome of a Biological Network

Charmi Panchal, Sepinoud Azimi, and Ion Petre(✉)

Computational Biomodeling Laboratory, Turku Centre for Computer Science,
Åbo Akademi University, Agora, Domkyrkotorget 3, 20500 Åbo, Finland
{cpanchal,sazimi,ipetre}@abo.fi

Abstract. There has been much progress in recent years towards building larger and larger computational models for biochemical networks, driven by advances both in high throughput data techniques, and in computational modeling and simulation. Such models are often given as unstructured lists of species and interactions between them, making it very difficult to understand the *logicome* of the network, i.e. the logical connections describing the activation of its key nodes. The problem we are addressing here is to predict whether these key nodes will get activated at any point during a fixed time interval (even transiently), depending on their initial activation status. We solve the problem in terms of a Boolean network over the key nodes, that we call the logicome of the biochemical network. The main advantage of the logicome is that it allows the modeler to focus on a well-chosen small set of key nodes, while abstracting away from the rest of the model, seen as biochemical implementation details of the model. We validate our results by showing that the interpretation of the obtained logicome is in line with literature-based knowledge of the EGFR signalling pathway.

Keywords: Biomodeling · Boolean network · Logicome · EGFR pathway · ODE models

1 Introduction

One of the central topics of interest in systems biology is to identify the functionalities of a living cell and to understand how the huge number of interactions within a cell facilitate such functionalities. The set of complex and involved interactions lead to obtaining a large number of collected experimental data as well as complex networks. These broad sources of information can prove to be very useful in providing a realistic life picture of the phenomenon under study, but can also make it difficult to analyze the system and can cause inaccuracy in predicting the system's behavior. Identifying the main players within a network and understanding how they activate each other can help to overcome these difficulties.

There have been many studies on the logical modelling of biological networks; for example, [4–6,30] discuss the correspondence between Boolean networks and ODEs; for an introduction to Boolean networks and ODEs we refer

M. Botón-Fernández et al. (Eds.): AlCoB 2016, LNBI 9702, pp. 38–49, 2016.
DOI: 10.1007/978-3-319-38827-4_4

to [13,14] respectively. Fuzzy logic was used in [19] to yield the logical models corresponding to the biological networks. As a different approach, [27] build the Boolean logic models by training a literature-based prior knowledge network against biochemical data. These studies mainly proposed approaches where the full understanding of the biological aspects of the phenomenon under study was crucial and the goal was to obtain a mathematical model reproducing that understanding. Our study goes in the reverse direction: it starts from an existing mathematical model and aims to obtain an abstract, high-level understanding of the functionality of the biological network underlying the model. Our goal is to obtain a logical description of the activation conditions between the key nodes of the network; even in the case when one starts from a detailed biological model going towards the mathematical model, our reverse engineering approach brings a new higher-level understanding of the functionality of the biological model we started from. The result of our approach is formulated as a Boolean network whose nodes are the key species we focus on; we coin the term *logicome* to name this network.

Extracting a Boolean network model from a given ODE-based model is a well-studied topic with many different solutions, see, e.g., [30] for a recent new solution and a good overview of the topic. Typically, the Boolean network model is seen as a companion of the ODE-based model, compensating for the lack of detailed kinetic-level data for the model, or allowing for alternative global analysis of model dynamics, such as attractor- or multi-stability- analysis, see [30]. A key step going from an ODE model to its corresponding Boolean network model is the discretization scheme allowing to replace continuous variables with their corresponding 0/1 variables. This is typically done by sampling the numerical integration of the continuous variables at different time points and by discretizing their values at those points. This leads to the dynamics of the Boolean model being interpreted in terms of discrete time series reflecting the behavior of the original ODE model. Our approach is coarser: we aim to capture the activation of the key nodes of the model over the whole time interval (to be thought of as much larger than those involved in the discretization of ODE models). This includes capturing the transient activation of a node over that interval, even if at the extremities of the interval the node may be inactive. The result is a Boolean network that accompanies the starting ODE model in terms of describing asynchronous cause-effect relationships among its key nodes over a fixed time interval.

As a case study we focus on the *EGFR (epidermal growth factor receptor) signaling pathway.* Epidermal growth factors are key players in cell proliferation, survival, migration and differentiation. EGFR signaling also has a major role in EGFR-dependent signal transduction, see [29]. Therefore, understanding their behavior is crucial in any cancer related studies, see [20]. For more information on EGFR signaling pathways we refer to [2,29,32].

This paper is organized as follows. In Sect. 2, we present our methodology to infer the logicome of biochemical networks. In Sect. 3, we introduce the case study we used in this paper. In Sect. 4 we present the results of applying the method to the case study and analyze the produced results and finally we conclude with some discussions in Sect. 5. All the models and data files used in this paper can be found at: http://combio.abo.fi/research/logicome-models-2/.

2 Methodology

In this section we present our method to infer the logicome of an ODE-based model. The steps are described in a generic way – their detailed implementation is up to the modeler and it depends on the case study. In the next section we discuss one particular way in which we used this method in the case of the EGFR pathway.

Step 1 – Setup. We start with an ODE model for a biochemical network. We assume also to have a set of "key nodes" whose influences over each others' activation we aim to capture. The choice of the key nodes from among the variables of the ODE model depends on the modeler and on the network under study.

Step 2 – Discretization. To be able to describe the logicome of a network in terms of Boolean network, we need to translate continuous simulation data to a Boolean, "on/off"-based language. Therefore, as the second step we incorporate a discretization algorithm into our method. Many discretization methods exist, see for example [18,26]. In this study our discretization step is based on a threshold-based approach in which we assign "1" to a species if at any time during the simulation its value is above a given threshold, and "0" otherwise. The precise choice of the threshold depends on the network under study.

Step 3 – Simulation. We simulate all possible knock-out mutants; in other words, all models where the key species are turned on/off in all possible combinations. We then apply to each simulation result the discretization step to obtain the Boolean results corresponding to each mutant. In this way we produce a truth table describing the output of each simulation as a Boolean function with the key nodes as its Boolean variables. Translating the input Boolean values of the key nodes to absolute numerical values to be used in the simulation can be done in several different ways, depending on the case study. For example, the 0 value for a Boolean key node may be translated to value 0 for the corresponding variable(s) in the knock-out mutant, while value 1 may be translated to the threshold value chosen for that variable in Step 2. The other, non-key nodes get the same initial values as in the original model.

Step 4 – Logicome generation. In this step we generate the logicome corresponding to the given biochemical network from the produced truth table in the previous step. Different algorithms can be used to implement this step, see for example [1,11,16,21]. In this paper we use the *Logic Friday* tool which incorporates the *Espresso algorithm* proposed in [21].

3 Case-Study: The EGFR Pathway

We focus in this paper on a signaling network that is strongly associated with the development of cancer processes: the *EGFR signaling pathway*. In the following subsections we provide a brief biological background and some computational details of this model.

3.1 Biological Background

The epidermal growth factor receptor (EGFR) pathway regulates several important cellular processes including cell proliferation, survival, differentiation and development, see [20]. Because of its association with the various types of cancer processes, this pathway is a widely investigated signal transduction system. The EGFR pathway can be seen as a union of several smaller pathways, also called *modules*, see [3,31]. The proteins situated at the intersection between these modules are called *interface species*. The analysis presented in [10] identifies the locations of oncogenes and essential components of the EGFR signaling cascade that define most of the interface regions. Our model is adopted from [31] that uses the model originally presented in [28] and implements it in the stochastic pi-calculus language together with the results identified by [10]. We follow the approach of [31] and their modularization of the EGFR signaling pathway in the following 7 modules: EGF, Grb2, Ras-Shc-Dependent /Independent, Raf, MEK, and ERK. These modules communicate with each other through the following 8 *interface species*: (EGF-EGFR*)2-GAP, (EGF-EGFR*)2-GAP-Grb2-Sos, (EGF-EGFR*)2-GAP-Shc*-Grb2-Sos, Ras-GTP, Ras-GTP*, MEK-PP, Raf* and ERK-PP. We adopt these interface species as the key nodes in our approach.

We briefly describe the functionality of the EGFR pathway focusing mainly on the signal propagation within the interface species, as suggested in [10]; the modules of the pathway are considered as black-boxes communicating to each other through the interface species. The EGFR is situated on the extracellular surface of the cell and signal transduction begins upon binding of ligand EGF (epidermal growth factor) to EGFR. The EGF-bounded receptor induces dimerization and autophosphorylation of several members of intracellular domains, which leads to the recruiting of several cytoplasmic enzymes and adaptor proteins. This initiates to the activation of two principal pathways, one Shc-dependent and another Shc-independent, that play a significant role in the activation of downstream signaling processes like hydrolyzation of Ras-GDP and activation of Ras-GTP that follows by dissociation of Ras-GTP from the receptor complex. Further dissociation of Ras-GTP makes it inactive and promotes the intrinsic activity of Ras protein regulated by the GTPase activating protein (GAP) that is involved in several crucial cellular processes see [10,24]. It is assumed that the dissociated Ras-GTP molecule causes phosphorylation of the Raf protein that in-turn double phosphorylates MEK (turning it to MEK-PP) and ERK (turning it to ERK-PP) proteins. The final result of the signaling cascade is the double phosphorylated ERK-PP that further regulates a number of transcription factors and essential proteins for cell differentiation and growth.

A systematic analysis of control mechanisms (including positive/negative feedback loops) underlying EGFR pathway are presented in [10,31]. We aim to represent the functional relationships associated with the interface species through a Boolean network – the *logicome* of the EGFR signaling pathway.

3.2 Mathematical Model, Simulation and Discretization

We associated a mass-action ODE-based model, see [8,14], to the reaction based model of [10]. Each of the 103 variable molecular species of the model in [10] gets a variable in our mathematical model. We wrote the reaction-based model of the EGFR pathway in the COPASI software, see [9], and used its feature to automatically generate the mass-action-based system of ordinary differential equations associated to the model. We call the resulting model our *basic model.*

Following the approach of [31], we simulated in COPASI this model for an EGF stimulus of 4981 molecules/pl which is enough to phosphorylate 50000 EGF-receptors. The simulation was run for 6000 s and the time series results of each interface species were collected.

For our method we are interested in analyzing all knock-out mutants where the interface species are active/inactive in all possible combinations. In the knock-out mutants the initial values of the inactive interface species are set to the value 0, while the active interface species are set to a specific threshold value of 1 % of that species' maximum value in the simulation of the basic model up to 6000 s. Since we considered 8 interface species, we have $256 = 2^8$ knock-out mutant simulations.

3.3 Generating the Logicome

Each knock-out mutant can be seen as a particular truth assignment over the 8 Boolean variables standing for the interface species. The results of the 256 knock-out simulations were discretized as follows.

Collecting the outputs of all knock-out mutants can be done in the form of a Boolean function with 8 inputs and 8 outputs.

We used the *LogicFriday* software to generate the Boolean function associated to the EGFR pathway based on the Boolean table collected above. We then used the 5 types of Boolean gates illustrated in Fig. 1 to generate the logicome associated to the EGFR signaling pathway.

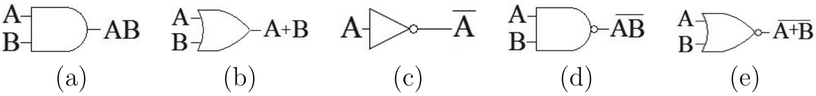


Fig. 1. The Boolean gates for the logical outcome: (a) AND : AB, (b) OR : $A + B$, (c) NOT : $\overline{A}$, (d) NAND: $\overline{AB}$, (e) NOR : $\overline{A + B}$, where we denote the negation of A with $\overline{A}$, the disjunction of A and B with $A + B$, and the conjunction of A and B with AB.

4 Results

The interface species are denoted in the logicome as the nodes of the Boolean network in the way explained in Table 1. The Boolean functions generated as the result of the steps described in Sect. 3 are shown in Table 2. We repeated the same experiment where we set the initial values of the active key nodes to 10 % (rather than 1 %) of their maximum value in the simulation of the basic model; the corresponding Boolean formulation is presented in Table 3.

Table 1. The notation used for the interface species in the Boolean network.

Node	Interface species
G_0	(EGF-EGFR*)2-GAP
G_1	Raf*
G_2	MEK-PP
G_3	Ras-GTP*
G_4	ERK-PP
G_5	(EGF-EGFR*)2-GAP-Shc*-Grb2-Sos
G_6	Ras-GTP
G_7	(EGF-EGFR*)2-GAP-Grb2-Sos

Table 2 shows G_1 as getting activated in all knock-out models and thus, being set to constant 1. This means that for all combinations of active/inactive key nodes (even those where G_1 is initialized as inactive), G_1 gets eventually activated in the time interval $[0, 6000]$ sec. This can be interpreted as G_1 being insensitive to (relatively) small changes in the levels of the other key nodes; indeed, all the key nodes are 0 in the basic model, leading to activation of G_1; setting the initial values of the key nodes to 1 % of their maximum level in the basic model does not change the situation. This result also suggests that in the case of small perturbations in the initial values of key nodes, the activation of G_1 is driven by other factors, outside the set of key nodes. The situation is different if we look into bigger changes in the initial values of the key nodes, e.g., setting them to 10 % of their maximum values in the basic model; as shown in Table 3, G_1 is in this case non-constant and influencing the behavior of G_6. In Table 3, we observe that the activation of G_1 depends on the key nodes G_3, G_5 and G_6 – this is consistent with the results reported in [25].

Another interesting observation of the logicome in Table 2 is that all key nodes get activated in the case of G_3 starts inactive and G_5 starts active. The same observation is found in the results obtained for the threshold of 10 %, see Table 3, and even for 20 % and 30 % see Tables 4 and 5. This is consistent with the observation of [7,10,23,31] about the role played by the shc*-dependent component (denoted by G_5) and the Ras subfamily protein (denoted by G_3) in the activation of several pathway components, including all of our key nodes.

Table 2. The Boolean functions describing the logicome of the EGFR signaling pathway for the threshold of 1 %. An overline over a variable's name denotes its negation, the plus denotes disjunction, while the concatenation of two variables denotes their conjunction.

Boolean functions
$G_0 := \overline{G}_3 + G_5 + G_0\overline{G}_4 + \overline{G}_4G_7 + G_0\overline{G}_6G_7;$
$G_1 := 1;$
$G_2 := G_2 + \overline{G}_3 + G_5 + G_6;$
$G_3 := G_0 + \overline{G}_2 + G_3 + G_4 + G_5 + G_6 + G_7;$
$G_4 := G_2 + \overline{G}_3 + G_4 + G_6 + G_0G_5G_7;$
$G_5 := G_0G_5 + \overline{G}_3G_5 + \overline{G}_3\overline{G}_6 + G_5\overline{G}_6 + G_5G_7 + G_0\overline{G}_3G_7;$
$G_6 := \overline{G}_3 + G_5 + G_0G_6 + G_6G_7;$
$G_7 := \overline{G}_3 + G_5$

Table 3. The Boolean functions describing the logicome of the EGFR signaling pathway for the threshold of 10 %.

Boolean functions
$G_0 := G_5 + G_0\overline{G}_3\overline{G}_4 + \overline{G}_3\overline{G}_4\overline{G}_6 + G_0\overline{G}_3G_7 + \overline{G}_3\overline{G}_4G_7;$
$G_1 := \overline{G}_3 + G_5 + G_6;$
$G_2 := G_2 + \overline{G}_3 + G_5 + G_6;$
$G_3 := G_0 + \overline{G}_2 + G_3 + G_5 + G_6 + G_7;$
$G_4 := G_2 + \overline{G}_3 + G_4 + G_6 + G_0G_5G_7;$
$G_5 := G_0G_5 + \overline{G}_3G_5 + \overline{G}_3\overline{G}_6 + G_5\overline{G}_6 + G_5G_7 + G_0\overline{G}_3G_7;$
$G_6 := G_5 + G_0\overline{G}_3 + \overline{G}_1\overline{G}_3 + G_0G_6 + \overline{G}_3G_6 + \overline{G}_3G_7 + G_6G_7;$
$G_7 := \overline{G}_3G_5 + \overline{G}_3\overline{G_6} + \overline{G}_3G_7 + G_0G_5\overline{G}_6 + G_0G_5G_7 + G_5\overline{G}_6G_7$

It is also interesting to note that the EGFR signaling pathway has an internal mechanism for compensating the potential failure of G_5 by G_7. Based on [7,10,31], G_0 mediates the activation of both G_5 and G_7; in case G_5 fails while G_3 remains inactive then G_7 gets activated and this is enough to activate all key nodes. This is seen in Table 3, if $G_0 = \overline{G}_3 = \overline{G}_5 = G_7 = 1$, then all key nodes get activated.

4.1 Sensitivity to the Numerical Setup of the Model

To investigate the sensitivity of our method to changes in the numerical setups of the underlying ODE model, we re-ran all simulations for different values of EGF and EGFR. We first experimented with different concentrations of EGF stimulus keeping the same EGFR concentration of 50000 molecules and then with different concentrations of EGFR keeping the same EGF stimulus of 4981 molecules. We observe that the obtained logicomes are almost identical to the previous result

Table 4. The Boolean functions describing the logicome of the EGFR signaling pathway for the threshold of 20 %.

Boolean functions
$G_0 := G_5 + G_0\overline{G}_3\overline{G}_4 + \overline{G}_3\overline{G}_4\overline{G}_6 + G_0\overline{G}_3G_7 + \overline{G}_3\overline{G}_4G_7;$
$G1 := \overline{G}_3 + G_5 + G_6;$
$G_2 := G_2 + \overline{G}_3 + G_5 + G_6;$
$G_3 := G_0 + \overline{G}_2 + G_3 + G_5 + G_6 + G_7;$
$G_4 := G_2 + \overline{G}_3 + G_4 + G_6 + G_0G_5G_7;$
$G_5 := G_0G_5 + \overline{G}_3G_5 + \overline{G}_3\overline{G}_6 + G_5\overline{G}_6 + G_5G_7 + G_0\overline{G}_3G_7;$
$G_6 := G_5 + G_0\overline{G}_3 + \overline{G}_1\overline{G}_3 + G_0G_6 + \overline{G}_3G_6 + \overline{G}_3G_7 + G_6G_7;$
$G_7 := \overline{G}_3G_5 + \overline{G}_3\overline{G}_6 + \overline{G}_3G_7 + G_0G_5G_7 + G_5\overline{G}_6G_7$

Table 5. The Boolean functions describing the logicome of the EGFR signaling pathway for the threshold of 30 %.

Boolean functions
$G_0 := G_5 + G_0\overline{G}_3\overline{G}_4 + \overline{G}_3\overline{G}_4\overline{G}_6 + G_0\overline{G}_3G_7 + \overline{G}_3\overline{G}_4G_7;$
$G_1 := \overline{G}_3 + G_5 + G_6;$
$G_2 := G_2 + \overline{G}_3 + G_5 + G_6;$
$G_3 := G_0 + G_3 + G_5 + G_6 + G_7 + \overline{G}_1\overline{G}_2 + \overline{G}_2G_4;$
$\mathrm{G}_4 := G_2 + \overline{G}_3 + G_4 + G_6 + G_0G_5G_7;$
$G_5 := G_0G_5 + \overline{G}_3G_5 + \overline{G}_3\overline{G}_6 + G_5\overline{G}_6 + G_5G_7 + G_0\overline{G}_3G_7;$
$G_6 := G_5 + G_0\overline{G}_3 + \overline{G}_1\overline{G}_3 + G_0G_6 + \overline{G}_3G_6 + \overline{G}_3G_7 + G_6G_7;$
$G_7 := \overline{G}_3G_5 + \overline{G}_3\overline{G}_6 + \overline{G}_3G_7 + G_0G_5G_7 + G_5\overline{G}_6G_7$

presented in Table 2. To investigate the sensitivity of our method to different threshold criteria, we repeated the experiments above with a threshold value of 30 % of each interface species' maximum value. By comparing results, we note that the logicome results obtained with the threshold value of 10 %, 20 %, and 30 % (see Tables 3, 4, and 5) are much more complex than the previous one.

4.2 Incomplete Availability of the Knock-Out Mutants

In the way we described our method in Sects. 2 and 3, we implicitly assume the full availability of the simulation results of all knock-out mutant models. We considered the case when the data on several knock-out mutants is in fact not available and compared the results to the case when all data is available. We considered the simulations results of only 186 knock-out mutants and assumed that the data on the other 70 knock-out mutants is unavailable. We used the threshold value of 1 % and the numerical setups of EGF and EGFR as 4981 and 50000 molecules, respectively.

Table 6. The Boolean functions associated with the logicome of the model where the data of 70 knockout mutants are not available. The result is almost identical to that in Table 2 where all data was available, showing that the method in this case was robust to missing data.

Boolean functions
$G_0 := \overline{G}_3 + G_5 + G_0\overline{G}_4 + \overline{G}_4G_7 + G_0\overline{G}_6G_7;$
$G_1 := 1;$
$G_2 := G_2 + \overline{G}_3 + G_5 + G_6;$
$G_3 := \overline{G}_2 + G_3 + G_4 + G_5 + G_6 + G_7;$
$G_4 := G_2 + \overline{G}_3 + G_4 + G_6 + G_0G_5G_7;$
$G_5 := G_0G_5 + \overline{G}_3G_5 + \overline{G}_3\overline{G}_6 + G_5\overline{G}_6 + G_5G_7 + G_0\overline{G}_3G_7;$
$G_6 := \overline{G}_3 + G_5 + G_0G_6 + G_6G_7;$
$G_7 := \overline{G}_3 + G_5$

The result obtained in this case is shown in the Table 6 and it is almost the same as the result in Table 2 obtained by using the full data. This shows that in this case the logicome extraction method was robust to the missing data; this may of course be different for other models and for other missing data.

5 Discussion

We propose in this article an addition to the rich field of logic modeling of biological networks, see, e.g., [4,15,19]. We start from a mathematical model of the network, taking advantage of the growing availability of mathematical models. The logicome approach proposed in this article allows the modeler to focus on a selected set of key nodes, important for the network under study, while abstracting away from the rest of the network; the output is a description of their influence on each other (even transient) activation over a fixed time interval.

The bottom-up modeling approaches (e.g., large-scale modeling [17], automatic knowledge extraction [22], data-driven network construction [12], etc.) have been very popular due to their ability to provide a very detailed picture, to explain the data, and to reproduce the behaviour of the phenomenon under study. The logicome is a companion to such detailed models; it gives a more abstract, systematic and objective description of the functionalities of the model. This is especially relevant in the case of big models built from many different sub-models and for which a full global "blueprint" does not exist. The logicome aims to be such a blueprint, deduced a-posteriori, based on an existing detailed view of the model.

The output of the logicome approach depends on the numerical setup of the method: both on the numerical setup of the basic mathematical model, and on the choice of the threshold values in the discretization step. This is natural since the method is dependent on the numerical ODE-based simulations of the basic model and of the knock-out mutants; this suggests choosing an already well-fitted and -validated model for the network under study. The choice of the threshold value is in fact a decision on how a species of the model can be labeled as 'active'; we suggested using a percentage of the maximum value reached by that species in the simulation of the basic model, but other choices may also be appropriate depending on the case study.

The computational efficiency of the method is dependent on the number of key nodes selected in the analysis: with more key nodes selected, exponentially more knock-out mutant models should be analyzed. Eliminating some of the knock-out mutants is possible, and the result of the method will be in this case an only-partial description of the logical dependencies between the key nodes. On the other hand, the method scales up very well in the size of the basic model: as long as the ODE-based models may be simulated efficiently, the method will be practical; this means that networks with thousands of nodes may be analyzed, as long as the number of key nodes n is so that it remains practical to run 2^n simulations.

References

1. Akutsu, T., Miyano, S., Kuhara, S.: Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. In: Somogyi, R., Kitano, H. (eds.) Pacific Symposium on Biocomputing, vol. 4, pp. 17–28. Citeseer (1999)
2. Britton, D., Hutcheson, I.R., Knowlden, J.M., Barrow, D., Giles, M., McClelland, R.A., Gee, J.M., Nicholson, R.I.: Bidirectional cross talk between ERα and EGFR signalling pathways regulates tamoxifen-resistant growth. Breast Cancer Res. Treat. **96**(2), 131–146 (2006)
3. Bruggeman, F.J., Westerhoff, H.V., Hoek, J.B., Kholodenko, B.N.: Modular response analysis of cellular regulatory networks. J. Theor. Biol. **218**(4), 507–520 (2002)
4. Chaves, M., Sontag, E.D., Albert, R.: Methods of robustness analysis for Boolean models of gene control networks. IEEE Proc. Syst. Biol. **153**(4), 154–167 (2006)
5. Davidich, M.I., Bornholdt, S.: Boolean network model predicts cell cycle sequence of fission yeast. PloS ONE **3**(2), e1672 (2008)
6. Glass, L., Kauffman, S.A.: The logical analysis of continuous, non-linear biochemical control networks. J. Theor. Biol. **39**(1), 103–129 (1973)
7. Gong, Y., Zhao, X.: Shc-dependent pathway is redundant but dominant in mapk cascade activation by egf receptors: a modeling inference. FEBS Lett. **554**(3), 467–472 (2003)
8. Gratie, D.-E., Iancu, B., Petre, I.: ODE analysis of biological systems. In: Bernardo, M., de Vink, E., Di Pierro, A., Wiklicky, H. (eds.) SFM 2013. LNCS, vol. 7938, pp. 29–62. Springer, Heidelberg (2013)
9. Hoops, S., Sahle, S., Gauges, R., Lee, C., Pahle, J., Simus, N., Singhal, M., Xu, L., Mendes, P., Kummer, U.: COPASI - a complex pathway simulator. Bioinformatics **22**(24), 3067–3074 (2006)

10. Hornberg, J.J., Binder, B., Bruggeman, F.J., Schoeberl, B., Heinrich, R., Westerhoff, H.V.: Control of MAPK signalling: from complexity to what really matters. Oncogene **24**(36), 5533–5542 (2005)
11. Hwa, H.R.: A method for generating prime implicants of a Boolean expression. IEEE Trans. Comput. **23**(6), 637–641 (1974)
12. Janes, K.A., Yaffe, M.B.: Data-driven modelling of signal-transduction networks. Nat. Rev. Mol. Cell Biol. **7**(11), 820–828 (2006)
13. Kauffman, S.: Homeostasis and differentiation in random genetic control networks. Nature **224**, 177–178 (1969)
14. Klipp, E., Herwig, R., Kowald, A., Wierling, C., Lehrach, H.: Systems Biology in Practice: Concepts, Implementation and Application. Wiley, Weinheim (2008)
15. Le Novere, N.: Quantitative and logic modelling of molecular and gene networks. Nat. Rev. Genet. **16**(3), 146–158 (2015)
16. Liang, S., Fuhrman, S., Somogyi, R.: Reveal, a general reverse engineering algorithm for inference of genetic network architectures. In: Bryant, B., Milosavljevic, A., Somogyi, R. (eds.)Pacific Symposium on Biocomputing, vol. 3, pp. 18–29. Citeseer (1998)
17. Macklin, D.N., Ruggero, N.A., Covert, M.W.: The future of whole-cell modeling. Curr. Opin. Biotechnol. **28**, 111–115 (2014)
18. Martin, S., Zhang, Z., Martino, A., Faulon, J.: Boolean dynamics of genetic regulatory networks inferred from microarray time series data. Bioinformatics **23**(7), 866–874 (2007)
19. Morris, M.K., Saez-Rodriguez, J., Sorger, P.K., Lauffenburger, D.A.: Logic-based models for the analysis of cell signalling networks. Biochemistry **49**(15), 3216–3224 (2010)
20. Oda, K., Matsuoka, Y., Funahashi, A., Kitano, H.: A comprehensive pathway map of epidermal growth factor receptor signaling. Curr. Opin. Biotechnol. **1**(1), 1–17 (2005)
21. Pantel, P., Pennacchiotti, M.: Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In: Carpuat, M., Duh, K. (eds.) Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, pp. 113–120 (2006)
22. Pitkänen, E., Jouhten, P., Hou, J., Syed, M.F., Blomberg, P., Kludas, J., Oja, M., Holm, L., Penttilä, M., Rousu, J., Arvas, M.: Comparative genome-scale reconstruction of gapless metabolic networks for present and ancestral species. PLoS Comput. Biol. **10**(2), 1–12 (2014)
23. Rajalingam, K., Schreck, R., Rapp, U.R., Albert, V.: Ras oncogenes and their downstream targets. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research **1773**(8), 1177–1195 (2007)
24. Rajasekharan, S., Raman, T.: Ras and ras mutations in cancer. Cent. Eur. J. Biol. **8**(7), 609–624 (2013)
25. Roskoski, R.: Raf protein-serine/threonine kinases: structure and regulation. Biochem. Biophys. Res. Commun. **399**(3), 313–317 (2010)
26. Saez-Rodriguez, J., Alexopoulos, L.G., Epperlein, J., Samaga, R., Lauffenburger, D.A., Klamt, S., Sorger, P.K.: Discrete logic modelling as a means to link protein signalling networks with functional analysis of mammalian signal transduction. Mol. Syst. Biol. **5**(1), 331 (2009)
27. Saez-Rodriguez, J., Alexopoulos, L.G., Zhang, M., Morris, M.K., Lauffenburger, D.A., Sorger, P.K.: Comparing signaling networks between normal and transformed hepatocytes using discrete logical models. Cancer Res. **71**(16), 5400–5411 (2011)

28. Schoeberl, B., Eichler-Jonsson, C., Gilles, E.D., Müller, G.: Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. Nat. Biotechnol. **20**(118), 370–375 (2002)
29. Sebastian, S., Settleman, J., Reshkin, S.J., Azzariti, A., Bellizzi, A., Paradiso, A.: The complexity of targeting EGFR signalling in cancer: from expression to turnover. Biochimica et Biophysica Acta (BBA)-Reviews on Cancer **1766**(1), 120–139 (2006)
30. Stötzel, C., Röblitz, S., Siebert, H.: Complementing ODE-based system analysis using Boolean networks derived from an Euler-like transformation. PLoS ONE **10**(10), e0140954 (2015)
31. Wang, D.Y., Cardelli, L., Phillips, A., Piterman, N., Fisher, J.: Computational modeling of the EGFR network elucidates control mechanisms regulating signal dynamics. BMC Syst. Biol. **3**(1), 1–18 (2009)
32. Yarden, Y.: The EGFR family and its ligands in human cancer: signalling mechanisms and therapeutic opportunities. Eur. J. Cancer **37**(4), 3–8 (2001)

Paper II

Generating the Logicome from Microarray Data

Charmi Panchal and Vladimir Rogojin

Charmi Panchal, and Vladimir Rogojin. Generating the Logicome from Microarray Data. In Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2017 IEEE Conference on, pp. 1-8. IEEE, 2017.

Generating the Logicome from Microarray Data

Charmi Panchal* and Vladimir Rogojin†
*†*Computational Biomodeling Laboratory*
Turku Centre for Computer Science
Åbo Akademi University
Agora, Domkyrkotorget 3, FIN-20500 Åbo.
**cpanchal@abo.fi*, † *vrogojin@abo.fi*

Abstract—**The advances in complex statistics and machine learning methods lead to the development of powerful classifiers that can be used to recognize cellular states (such as gene expression profiles) that are associated to a number of gene-scale expressed diseases, for instance, cancer. However, the data-driven models built by means of learning from datasets in a number of cases represent "black boxes" that cannot be easily analyzed and understood. In this article, we suggest a method for building a *data-driven logicome*. I.e., the method for building a set of small boolean expressions as classifiers for disjoint groups of samples from a microarray dataset. We validate our method on the microarray dataset of head and neck/oral squamous cell carcinoma, where our boolean signature presented a set of gene activity/inactivity combinations that are characteristic for various cancer sub-types and normal samples. Our findings correlate well with the literature.**

1. Introduction

The Microarray is one of the most important and widely used experimental developments in biotechnologies in recent years. It contains the high throughput genome-wide expression profiles, that allow monitoring of expression levels in cells for thousands of genes simultaneously and conduct large-scale quantitative assessments of gene expression. The microarrays are the most important tools for discovering key insights from the massive quantities of gene expression such as: identification of biomarkers, classification of cancer subtypes, predicting response to therapy and understanding the mechanisms involved in the disease progression [1]–[3].

We develop a method for deriving classifiers with clear internal structure based on boolean logic for samples from various microarray datasets. We agree to call these classifiers *boolean signatures*. Here, a boolean signature is a boolean formula in disjunctive normal form. We focus on microarray datasets for head and neck/oral cancer as the case study and build boolean signatures to distinguish between different subtypes of cancer and healthy cells.

Nowadays, there has been already collected an overwhelming amount of diverse genome and molecular-scale information as well as clinical data on cancers and other genetic-related diseases. On one hand, abundance of the data helps to increase our understanding about the disease. On the other hand, large amount of unstructured data represents a great challenge to be interpreted and comprehended. Complex statistics and machine learning methods provide means to build data-driven models that can be used as classifiers that recognize biological or clinical samples to belong to certain disjoint groups, like disease or healthy cell-lines. However, in a number of cases those models are constructed without comprehensible internal structure that can be understood. In other words, the data-driven models built by means of machine learning methods will tell what samples belong to what groups, but will not always tell exactly why. The goal of our work is to provide a method that can "answer" in a simple manner why a sample should belong to a certain group. The answer will be provided in terms of boolean logic.

Systems biology deals with understanding of the functionalities of a living cell and of deviations in cellular functions that lead to diseases. For instance, there have been many studies on constructing Boolean logic models for various biological phenomena. For example, in [4], [5] there was established correspondence between Boolean networks and ODE-based models. Some data-driven Boolean logic model building methods were suggested in [6].

Some of the studies from above mainly focus on approaches where the full understanding of the biological aspects of the phenomenon of interest is required. However, even though highly detailed models can provide a realistic life picture, sometimes, it can be difficult to analyze and reason about the large models. Hereby, studies in [7] aimed at obtaining a higher-abstraction level of understanding of biological systems starting from existing "larger" models. In particular, the goal of that work was in deriving a simple logical description of the activation conditions between the "key nodes" of a bio-model under study. As the result, [7] presented a method for translating a highly detailed large biological model in form of a bio-molecular (signaling pathway) pathway into a relatively small Boolean network (so-called *logicome*) representing activation relations between the key nodes as logic relations. A biological model presented in the form of the logicome should be easier to comprehend and reason about.

In this article, we advance further with the idea from [7]

978-1-4673-8988-4/17/$31.00 ©2017 IEEE

of developing high-level comprehensible Boolean logic based models for biological phenomena. Here, we develop a simple method for deriving logical relations between key/significant elements associated to certain groups of samples from microarray gene expression data. We call this *logicome* derived for microarray datasets. Those relations should provide a simple explanation in terms of logical formulas in disjunctive normal form.

We have chosen the Head and Neck/Oral Squamous Cell Carcinoma (HNOSCC) microarray datasets from studies in [8] as case-studies. The Head and Neck/Oral Squamous Cell Carcinoma (HNOSCC) is the most common cancer world-wide, and the important risk factors are tobacco and alcohol consumption [9]. The head and neck/oral cancers are categorized by the tissues of the head or neck from which they originate. The HNOSCC usually originates from squamous cells that are located inside the mouth, the nose, and the throat (for example, in the paranasal sinuses, Salivary glands, nasal cavity, oral cavity, pharynx, oropharynx and larynx) [8], [10]. We have selected four sample groups from [8]: oral tongue squamous cell carcinoma, samples in squamous cell carcinoma of the oral cavity and oropharynx, head and neck squamous cell carcinoma (locations of pharynx and larynx), and normal cell lines.

In our case studies, we apply our method to a subset of differentialy expressed genes between normal and cancer cells. We have performed differential gene expression analysis by means of GEO2R, an R-based web application [11].

We have derived for each of these groups boolean formulas representing their characteristic patterns of gene expression profiles. We have validated our findings against well known gene expression patterns associated to head and neck/oral cancer [12]–[25] which do not contradict the discovered boolean signatures associated to cancer cell lines. In the same time, we have derived a number of gene expression patterns that have not been discovered so far.

In our methodology we employ multinomial logistic regression to find small subsets of genes for which we derive boolean signature for all the four groups.

2. Methodology

In this section we describe methods that are used in our study. Firstly, we present a formal definition of the classification method used in this paper, i.e., *multinomial logistic regression (MLR)*, then we describe our approach for deriving a unique Boolean expression (*boolean signature*) corresponding to each group (cluster/class/category) of samples.

2.1. Multinomial Logistic Regression

Multinomial logistic regression (MLR) is a classification method used to measure the relationship between a group distributed dependent variable and one or more independent variables. When building an MLR model, it is assumed that the groups are mutually exclusive, i.e., a sample belongs to exactly one group, for more information see [26].

The multinomial logistic regression is a simple extension of the binomial logistic regression. It is used when the outcome variable or the dependent variable has more than two nominal (unordered) groups. The multinomial logistic regression is often considered as most effective and reliable way to obtain the probability of group membership which are calculated with the maximum likelihood estimation approach [27]. The advantage of using multinomial logistic regression over other alternatives are identified in [28]. Particularly, we find MLR as most attractive and robust tool for analysing microarray data since it does not assume multivariate normality for the predictor variables, variance homoscedasticity or linearity for the independent variables and allows independent variables to be continuous, discrete or dummy [29], [30].

2.1.1. Accuracy of The Model. For any given classifier and any given sample Y, there are four possible classification outcomes:

- if Y belongs to cluster C and it is classified as such, we denote it as a *true positive (TP)*,
- if Y belongs to cluster C and it is classified in a different cluster, we denote it as a *false negative (FN)*,
- if Y does not belong to cluster C and it is classified as such, we denote it as a *true negative (TN)*,
- if Y does not belong to cluster C and it is classified in C, we denote it as a *false positive (FP)*.

Accuracy are calculated in terms of TP, TN, FN and FP.

$$\text{Accuracy} = \frac{\text{TN + TP}}{\text{TN+TP+FN+FP}}$$

Accuracy is the proportion of true results, whether it means belonging to the right cluster or not belonging to the wrong cluster. For more detailed information we refer to [31], [32]

2.2. Inferring Boolean Signatures

In the following, we describe the algorithm 2.1 that gives minimal size subset from the set of predictor variables (genes) G and the algorithm 2.2 that is applied on the selected minimal size subset to derive a boolean signature for each group.

2.2.1. Reducing the set of predictor variables. In order to generate the signature to be as simple as possible yet accurate, our goal here is to reduce the size of G in such a way that the accuracy of MLR is not compromised.

Algorithm 2.1. Let us consider the multinomial logistic regression (MLR) model for gene expression matrix M, A its predictive accuracy, and G the set of predictor variables and T an accuracy threshold and l is the threshold for the size of the subsets. Let us take the following steps:

Step 1 Enumerate all $2^{|G|}$ subsets of G.

Step 2 For all $S \subseteq G$, train its corresponding MLR model $\mathcal{M}_S$, and calculate its predictive accuracy denoted by A_S, where $|S| \geq l, 1 \leq l \leq |G|$.

Step 3 Collect set $\mathcal{G}_m$ of subsets from G where $\mathcal{G}_m = \{S \subseteq G | |S| = m$ and $A_S = max(A_{S_m} | S_m \subseteq G$ and $|S_m| = m)\}$. In other words, we select all subsets from G of size m which have the maximal accuracy among all the subsets of size m.

Step 4 Output a subset S^{min} of minimal size such that $S^{min} \in \{\mathcal{G}_m \mid l \leq m \leq |G|$ and $A_{S^{min}} \geq T\}$. In other words, we select a subset from all $\mathcal{G}_m$, $l \leq m \leq |G|$ of minimal size whose accuracy is not below the threshold T.

2.2.2. Boolean signature. In our approach the minimal size subset obtained from the algorithm 2.1, is further analyzed to derive Boolean signature for each group.

The boolean signature is derived as follows:

Algorithm 2.2. Let MB be the binarized gene expression matrix of expression matrix M generated for subset of genes $S \subseteq G$. Let $C = \{C_1, \ldots, C_k\}$ be the set of disjoint groups of samples from M. Let Pr be the probability threshold and $covg$ be the coverage threshold. The probability threshold Pr for a binary values combination frequency is the lower border for combinations to be considered as "frequent". The $covg$ threshold for binary values combination frequency indicates the border below which we consider binary values combination as "insignificant". We recall here, that for each group we select its frequent (defined by Pr) significant (defined by $covg$) binary values combination that we use to derive the disjunctive normal form:

Step 1 Consider set T_S of all the binary values combinations of genes from S in MB, where $S \subseteq G$.

Step 2 **Frequency of occurrence:** For each combination of binary values from $c_j \in T_S$, count the number of its occurrences in every group $C_i \in C$, divide it by $|C_i|$, denote it by $N^i_{c_j}$ where $1 \leq j \leq 2^{|S|}$. Intuitively, $N^i_{c_j}$ denotes the frequency of occurrence of combination c_j in group C_i.

Step 3 **Maximal frequency of occurrence:** Find $N^i_{max} = max\{N^i_{c_1}, N^i_{c_2} \ldots N^i_{c_{2^{|S|}}}\}$. In other words, N^i_{max} is the frequency of the most occurring combination in group C_i.

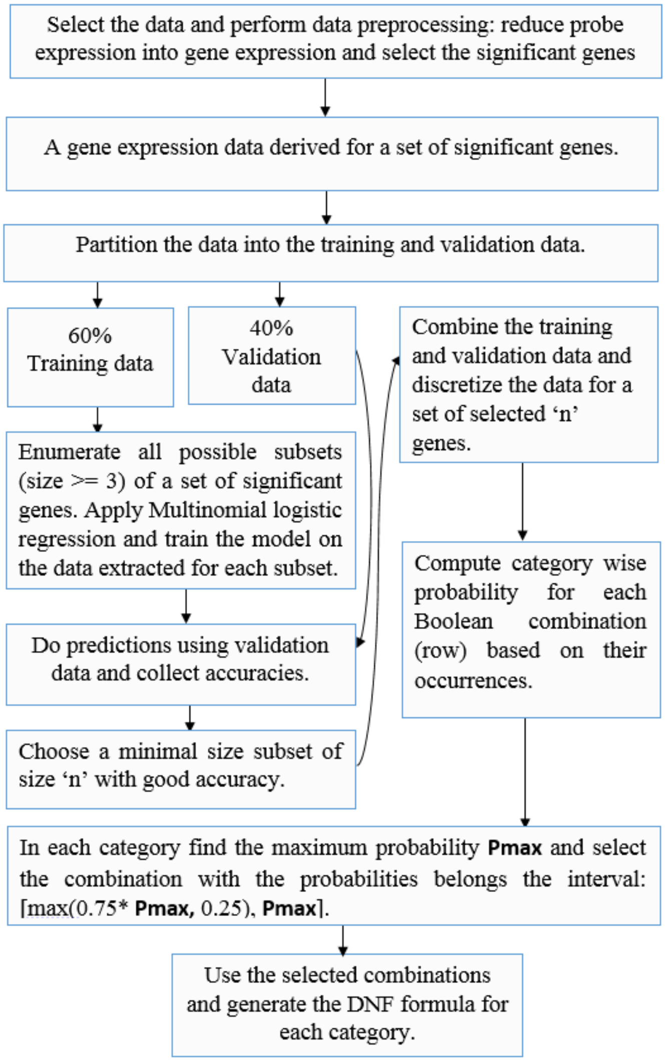
Step 4 **Representative combinations for a group:** For $C_i \in C$, $1 \leq i \leq k$, find the set $\mathcal{C}^i = \{c_j \in T_S \mid \max(Pr * N^i_{max}, covg) \leq N^i_{c_j} \leq N^i_{max}\}$, $1 \leq j \leq |T_S|$. In other words, here we select the representative combinations for a group C_i, those combinations are significant enough ($N^i_{c_j} \geq covg$) and are frequent ($N^i_{c_j} \geq Pr * N^i_{max}$) in C_i.

Step 5 **Deriving boolean signature:** For every $c_j \in \mathcal{C}^i$, where $c_j = (b_{g1}, b_{g2}, \ldots, b_{g|S|})$ and $b_{gl} \in \{0, 1\}$ is a binarized expression value for a gene $g_l \in S$ where $1 \leq l \leq |S|$, we construct the conjunction of gene variables associated to combination c_j as follows: $B_{ij} = (\bigwedge_{b_{gl}==1} g_l) \wedge (\bigwedge_{b_{gl}==0} \neg g_l)$. For the set of representative combinations $\mathcal{C}^i$ we construct the disjunctive normal form (boolean signature) BC_i as follows: $BC_i = \bigvee_{c_j \in \mathcal{C}^i} B_{ij}$. I.e., BC_i is the boolean signature of group C_i in the disjunctive normal form.

Step 6 **OUTPUT:** Output (C_i, B_i), for every $1 \leq i \leq k$.

The outline of our methodology is presented in the Figure 1.

Figure 1: Outline of the methodology


Select the data and perform data preprocessing: reduce probe expression into gene expression and select the significant genes
A gene expression data derived for a set of significant genes.
Partition the data into the training and validation data.
60% Training data
40% Validation data
Combine the training and validation data and discretize the data for a set of selected 'n' genes.
Enumerate all possible subsets (size >= 3) of a set of significant genes. Apply Multinomial logistic regression and train the model on the data extracted for each subset.
Compute category wise probability for each Boolean combination (row) based on their occurrences.
Do predictions using validation data and collect accuracies.
Choose a minimal size subset of size 'n' with good accuracy.
In each category find the maximum probability Pmax and select the combination with the probabilities belongs the interval: [max(0.75* Pmax, 0.25), Pmax].
Use the selected combinations and generate the DNF formula for each category.

3. Case studies

We use nine microarray data series of head and neck/oral squamous cell carcinoma (HNOSCC) from studies in [8] that are available at Gene Expression Omnibus (GEO) database [11]. We explain the pre-processing of the microarray data and obtaining the gene expression matrix, that we use to derive boolean expressions associated with various groups of HNOSCC and with non-tumor cells.

3.1. Samples

Studies in [8] consider 9 data series from GEO database [11] with 675 samples in total: *GSE6791*, *GSE9844*, *GSE30784*, *GSE31056*, *GSE2379*, *GSE3524*, *GSE6631*, *GSE13601* and *GSE23036*. In [8], due to an unsupervised learning method the samples were split in 22 groups out of which we have selected the following 4 groups with 509 samples for our studies: 58 samples in oral tongue squamous cell carcinoma (OTSCC), 189 samples in squamous cell carcinoma of the oral cavity and oropharynx (OSCC), 98 samples in head and neck squamous cell carcinoma (HNSCC), and 164 normal/control samples.

3.2. Microarray data

We get the microarray data from GEO in form of normalized probe signals as sample data matrices (probe expression matrices). In a sample data matrix, the rows correspond to probes and the columns correspond to samples. The data series that we have selected for our studies contain genome-wide gene expression profiling of head and neck/oral squamous cell carcinoma (HNOSCC) that was measured by Affymetrix platform [33], [34]. The probe signals in these series were normalized through Robust Multi-array Average technique, [35], GeneChip RMA (GCRMA) [36], and Microarray Suite version 5.0 (MAS 5.0, Affymetrix, Inc.), [37].

3.3. Data Preprocessing

For our case study, we performed differential gene expression analysis on the 9 data series and selected 11 “significant genes” for further analysis. We have generated gene expression matrix for the “significant genes”.

3.3.1. Gene expression matrix. We have processed the selected GEO data series in our study by using *Bioconductor GEOquery R Library* [34], [38]. We have transformed normalized probe measurements into gene expression levels as follows:

- We have found mappings between sets of probes and their associated genes for the respective affimetrix platforms by using an online web-tool *DAVID* (Database for Annotation, Visualization, and Integrated Discovery) [39], *GPL* (GEO platform record) [11] and *Affymetrix Human Genome U133 2.0 Array* annotation data (*hgu133plus2*).
- We have considered the expression level for a gene to be the median of the gene’s associated probes.

In the result, we have generated the gene expression matrix of approximately 25000 rows represented by gene symbols and columns by samples.

3.3.2. Selecting a set of differentially expressed genes. We have selected differentially expressed probes between control and all cancer samples for each data series separately by means of *GEO2R* web-tool [11] (https://www.ncbi.nlm.nih.gov/geo/info/geo2r.html). The differential gene/probe expression analysis was performed as in [40]. The result of differential gene/probe expression analysis for each data series is collected as a table of probes ranked by their p-values. We have selected those probes that satisfy the threshold of p-value ≤ 0.05. Then, for each dataset, we have selected those genes that correspond to the probes with the p-value not exceeding the threschold 0.05. We call these selected genes “significant genes”. Finally, we took the intersection of significant genes from all the datasets and considered it for our boolean signatures construction method. In particular, in the intersection we have the following significant genes: *MAL, LAPTM4B, HPGD, KRT4, EXT1, AIM1, SPP1, MYO10, MYO1B, MMP1, MGST2.*

3.3.3. Re-scaling gene expression datasets to the same level. In order to perform further analysis based on a number of different datasets, we need to bring the gene expression signals from them to the same scale. We performed rescaling by means of full range and interquartile range (with the interquartile range-based rescaling, we also take into account the outliers that may occur in the datasets). We perfomed the same analysis for the results of each of the two rescaling methods separately.

In more details, for each dataset D, for each sample S from D and for every gene G, we have re-scaled its expression value $x_{G,S}$ associated to S to $z_{G,S}$ as follows:

$$z_{G,S} = \frac{x_{G,S} - m_{G,D}}{R_{G,D}} \tag{1}$$

where $m_{G,D}$ is the mean value of expressions of gene G for all the samples from dataset D, and $R_{G,D}$ is the range/interquartile range of expressions of gene G among all the samples of D. For the full range we have $R_{G,D} = MAX_{G,D} - MIN_{G,D}$, where $MAX_{G,D}$ ($MIN_{G,D}$) is the highest (the lowest, respectively) expression level for a gene G among all the samples from dataset D. For the interquartile range we have $R_{G,D} = Q_3 - Q_1$, where Q_3 is the third quartile and Q_1 is the first quartile.

3.3.4. Removing similar samples between different groups. The goal of our methodology is to generate unique “boolean signatures” for different sample groups. Hereby, in order to increase the accuracy of our method, we need to make sure that the samples in the groups are “different enough". We filter out those samples that have "near identical" gene expression profiles but belong to different groups. We regard samples as vectors defined by their corresponding gene expression profiles and employ Euclidean distance in vector space as the closeness measurement between samples. We define a minimal distance threshold ϵ under which we consider samples to be "near identical". We calculate ϵ as follows:

$$\epsilon = C \times MAX_{NORM},$$

where MAX_{NORM} is a maximal norm among all the vectors in our studies, and C is a constant, which we have fixed in our studies to be $C = 0.01$. We did not find any "near identical" samples between different groups according to this criteria in our gene expression matrix.

The processed data and source code (in R) are available at Github (https://github.com/cpanchal/Dataset_Logicome.git)

3.4. Data analysis and results

We analysed the preprocessed gene expression data extracted for the significant genes as follows:

- Randomly partitioned the data into training and validation set with the ratio of $60 : 40$.
- Enumerated all the possible subsets (size ≥ 3) of a set of significant genes and extracted the data for each subset. We trained the MLR model on this data and collected the predictive accuracies using the validation data.
- Collected the subset of genes with maximum accuracies from the subsets of each size.
- From the collected subsets, picked a minimal size subset with accuracy $\geq 70\%$. That step rendered us a subset of genes of size 3.
- Applied threshold based discretization to the data for the rendered subset of genes (i.e, produce binarized gene expression by replacing expression values of genes with 1 if above threshold, 0 otherwise.)
- We derived "boolean signatures" in the disjunctive normal form for each group in terms of the selected minimal size subset of genes.

In our method the training and validation data are partitioned randomly, hence we re-ran the algorithm multiple times and the results are collected for each run. The boolean signatures derived using the resultant subsets are listed in the Tables 1 and 2. The Table 1 contains boolean signatures obtained from the datasets rescaled with the full range, and the Table 2 contains boolean signatures obtained from the datasets rescaled with the interquartile range. The Table 3 represents consistently up-regulated and down-regulated genes observed in the boolean signatures in the Tables 1 and 2.

The boolean formulas in the Tables 1 and 2 identify the groups Normal, OSCC, OTSCC and HNSCC with different combination of genes.

The down-regulated genes are denoted with '$^{-}$' and genes without '$^{-}$' are upregulated.

Table 1: Subsets and boolean formulations for each group: '$\wedge$' denotes conjunction, '$\vee$' denotes disjunction and '$^{-}$' denotes negation . Results obtained based on rescaling datasets with respect to full range.

No.	Subset	Boolean formula
1	(KRT4, MYO10, HPGD)	$Normal =$ $KRT4 \wedge HPGD$ $OSCC =$ $\overline{KRT4} \wedge (\overline{MYO10} \vee \overline{HPGD})$ $OTSCC =$ $\overline{KRT4} \wedge MYO10 \wedge \overline{HPGD}$ $HNSCC =$ $MYO10 \wedge \overline{HPGD}$
2	(KRT4, MAL, MMP1)	$Normal =$ $KRT4 \wedge MAL \wedge \overline{MMP1}$ $OSCC =$ $\overline{KRT4} \wedge \overline{MAL} \wedge MMP1$ $OTSCC =$ $\overline{KRT4} \wedge \overline{MAL} \wedge MMP1$ $HNSCC =$ $\overline{KRT4} \wedge \overline{MAL} \wedge MMP1$
3	(AIM1, SPP1, MMP1)	$Normal =$ $= AIM1 \wedge \overline{SPP1} \wedge \overline{MMP1}$ $OSCC =$ $MMP1 \wedge (\overline{AIM1} \wedge SPP1 \vee$ $AIM1 \wedge \overline{SPP1})$ $OTSCC =$ $MMP1 \wedge (\overline{AIM1} \wedge SPP1 \vee$ $AIM1 \wedge \overline{SPP1})$ $HNSCC =$ $\overline{AIM1} \wedge SPP1$
4	(KRT4, HPGD, SPP1)	$Normal =$ $KRT4 \wedge HPGD \wedge \overline{SPP1}$ $OSCC =$ $\overline{KRT4} \wedge \overline{HPGD} \wedge SPP1$ $OTSCC =$ $\overline{HPGD} \wedge \overline{SPP1}$ $HNSCC =$ $\overline{HPGD} \wedge SPP1$
5	(MGST2, AIM1, MMP1)	$Normal =$ $MGST2 \wedge AIM1 \wedge \overline{MMP1}$ $OSCC =$ $\overline{MGST2} \wedge MMP1$ $OTSCC =$ $MMP1 \wedge (MGST2 \wedge AIM1 \vee$ $\overline{MGST2} \wedge \overline{AIM1})$ $HNSCC =$ $\overline{MGST2} \wedge \overline{AIM1} \wedge MMP1$
6	(HPGD, MYO1B, MMP1)	$Normal =$ $HPGD \wedge \overline{MYO1B} \wedge \overline{MMP1}$ $OSCC =$ $\overline{HPGD} \wedge MYO1B \wedge MMP1$ $OTSCC =$ $\overline{HPGD} \wedge MMP1$ $HNSCC =$ $\overline{HPGD} \wedge MYO1B \wedge MMP1$

Table 2: Subsets and boolean formulations for each group: '∧' denotes conjunction, '∨' denotes disjunction and '¯' denotes negation. Results obtained based on re-scaling datasets with respect to interquartile range.

No.	Subset	Boolean formula
1	(MYO10,MAL,SPP1)	$Normal = \overline{MYO10} \wedge MAL \wedge \overline{SPP1}$ $OSCC = MYO10 \wedge \overline{MAL} \wedge SPP1$ $OTSCC = MYO10 \wedge \overline{MAL}$ $HNSCC = MYO10 \wedge \overline{MAL} \wedge SPP1$
2	(KRT4, MAL,MMP1)	$Normal = KRT4 \wedge MAL \wedge \overline{MMP1}$ $OSCC = \overline{KRT4} \wedge \overline{MAL} \wedge MMP1$ $OTSCC = \overline{KRT4} \wedge \overline{MAL} \wedge MMP1$ $HNSCC = \overline{KRT4} \wedge \overline{MAL} \wedge MMP1$
3	(MAL, LAPTM4B, SPP1)	$Normal = MAL \wedge \overline{LAPTM4B} \wedge \overline{SPP1}$ $OSCC = \overline{MAL} \wedge LAPTM4B \wedge SPP1$ $OTSCC = \overline{MAL} \wedge LAPTM4B$ $HNSCC = \overline{MAL} \wedge LAPTM4B \wedge SPP1$
4	(MAL, MYO1B, MMP1)	$Normal = MAL \wedge \overline{MYO1B} \wedge \overline{MMP1}$ $OSCC = \overline{MAL} \wedge MYO1B \wedge MMP1$ $OTSCC = \overline{MAL} \wedge MYO1B \wedge MMP1$ $HNSCC = \overline{MAL} \wedge MYO1B \wedge MMP1$
5	(MGST2,MAL,SPP1)	$Normal = MGST2 \wedge MAL \wedge \overline{SPP1}$ $OSCC = \overline{MGST2} \wedge \overline{MAL} \wedge SPP1$ $OTSCC = \overline{MGST2} \wedge \overline{MAL} \wedge SPP1$ $HNSCC = \overline{MGST2} \wedge \overline{MAL}$
6	(MAL,MYO1B,SPP1)	$Normal = MAL \wedge \overline{MYO1B} \wedge \overline{SPP1}$ $OSCC = \overline{MAL} \wedge MYO1B \wedge SPP1$ $OTSCC = \overline{MAL} \wedge MYO1B$ $HNSCC = \overline{MAL} \wedge MYO1B \wedge SPP1$
7	(MAL,SPP1,MMP1)	$Normal = MAL \wedge \overline{SPP1} \wedge \overline{MMP1}$ $OSCC = \overline{MAL} \wedge SPP1 \wedge MMP1$ $OTSCC = \overline{MAL} \wedge MMP1$ $HNSCC = \overline{MAL} \wedge SPP1 \wedge MMP1$
8	(KRT4,MAL,SPP1)	$Normal = KRT4 \wedge MAL \wedge \overline{SPP1}$ $OSCC = \overline{KRT4} \wedge \overline{MAL} \wedge SPP1$ $OTSCC = \overline{KRT4} \wedge \overline{MAL}$ $HNSCC = \overline{KRT4} \wedge \overline{MAL} \wedge SPP1$
9	(EXT1,MAL,SPP1)	$Normal = \overline{EXT1} \wedge MAL \wedge \overline{SPP1}$ $OSCC = EXT1 \wedge \overline{MAL} \wedge SPP1$ $OTSCC = EXT1 \wedge \overline{MAL}$ $HNSCC = EXT1 \wedge \overline{MAL} \wedge SPP1$
10	(MAL,AIM1,SPP1)	$Normal = MAL \wedge AIM1 \wedge \overline{SPP1}$ $OSCC = \overline{MAL} \wedge \overline{AIM1} \wedge SPP1$ $OTSCC = \overline{MAL} \wedge \overline{AIM1} \wedge SPP1$ $HNSCC = \overline{MAL} \wedge \overline{AIM1} \wedge SPP1$
11	(MYO10,MAL,MMP1)	$Normal = \overline{MYO10} \wedge MAL \wedge \overline{MMP1}$ $OSCC = MYO10 \wedge \overline{MAL} \wedge MMP1$ $OTSCC = MYO10 \wedge \overline{MAL} \wedge MMP1$ $HNSCC = MYO10 \wedge \overline{MAL} \wedge MMP1$

Table 3: Down-regulated and Up-regulated genes observed in each group. This observation is based on the boolean signatures presented in the Tables 2 and 1.

Group	Up-regulated genes	Down-regulated genes
Normal	HPGD, MGST2, AIM1 KRT4, MAL	MYO1B, MMP1, SPP1, MYO10,EXT1, LAPTM4B
OSCC	MMP1, MYO1B, SPP1 MYO10,EXT1, LAPTM4B	MGST2, KRT4, MAL, AIM1 MYO10, HPGD
OTSCC	MMP1,MYO1B,SPP1 MYO10,EXT1, LAPTM4B	HPGD, MAL, KRT4 AIM1
HNSCC	MYO10, MYO1B, MMP1, SPP1,EXT1,LAPTM4B	HPGD, MAL, KRT4, AIM1, MGST2

We can see from the boolean signatures, the subsets (KRT4, MAL, MMP1),(MAL, MYO1B, MMP1), (MAL, AIM1, SPP1), and (MYO10, MAL, MMP1) identify all the cancer groups by the same formula. The boolean signatures produced with these subsets clearly distinguish cancer and normal groups.

Moreover the genes in (KRT4, MAL, MMP1) appear frequently in the boolean signatures for all the cancer groups with MMP1 being up-regulated and (KRT4, MAL) being donw-regulated. These genes are the most promising and relevant genes for identifying HNOSCC tumor cells that is also inline with the study reported on cancer-specific genes in [12], [14], [15]. Also these findings are inline with the studies reported in [12] that also reports genes in (KRT4, MAL, MMP1) among the remarkable predictive bio-markers for identifying HNOSCC tumor cells.

We can see from the Table 3 that in all the three cancer groups genes MYO10, MYO1B, MMP1, SPP1, EXT1, LAPTM4B are up-regulated, whereas the genes HPGD, MAL, KRT4, and AIM1 are down-regulated. The boolean signatures presenting up-regulated (MYO10, MYO1B,MMP1,SPP1) and down-regulated HPGD, in the groups OSCC, OTSCC and HNSCC are confirmed by the results reported in [18], [21]–[23] that also reports these genes as significant predictor and the top most up-regulated (down-regulated) genes.

The boolean signatures presenting genes KRT4, HPGD, MAL and AIM1 as consistently down-regulated agree well with the work reported in [24], [25] and [41] for the groups HNSCC, OSCC and OTSCC. The gene MGST2 inolved as down regulated genes in the boolean signatures for the groups OSCC and HNSCC is confirmed in the studies [19] and [20].

Besides the verified results reported in the Tables 1 and 2, we discover some combinations of genes for tumor groups that remain to be validated experimentally. The up-regulated gene LAPTM4B is identified in various types of tumors, however it is less known in the tumor groups of HNOSCC. In our results in the Table 2 we found boolean signatures involving the up-regulated LAPTM4B together with the promising bio-marker genes MAL and SPP1 i.e. the subset (MAL, LAPTM4B, SPP1). These boolean signatures distinguish groups: Normal, (OSCC, HNSCC) and OTSCC. The signature represents the gene LAPTM4B equally significant bio-marker for tumor groups of HNOSCC.

We discover for OSCC in the boolean signature for the 1st subset in the Table 1, the possibility for the gene MYO10 to be down regulated. Similar observation for OSCC is found in the boolean signature for the 3rd subset where the genes SPP1 and AIM1 are found down-regulated and up-regulated respectively. Moreover in the boolean signature for the 5th subset in the Table 1, the result shows the possibility of both genes MGST2 and AIM1 to be up-regulated for OTSCC.

The boolean signatures produced with our methodology elucidate the comprehensive description/knowledge lying within the high-throughput micro-array datasets and unveil the gene expression profiles associated with the normal and cancer groups and precisely explain the associated connections.

4. Discussion

In this work we develop a methodology that reproduces the high-level knowledge lying within the complex gene expression datasets and presents the knowledge in the form of boolean formulation that can be easily viewed and understood. The boolean formulations represent so-called boolean signatures for the disjoint groups of samples within the micro datasets of HNOSCC. Through boolean signature, our method identifies regulations of significant genes and provides combinatorial patterns for each group of samples which are supported well with the literature findings.

Here, we propose a continuation of the direction initiated in [7], where logicome building methods are suggested to be a companion to the bottom-up modeling approaches. In [7], the authors suggested a way to generate a higher-level representation of a network model in terms of boolean logic relations between key nodes of the network. That logicome approach should allow the modeler to concentrate on a selected set of significant network nodes and relations between them, while abstracting from the rest of the model. Also, due to the fact that machine learning and statistical approaches usually do not provide information about the internal structure of the system under studies and relations between its components, but, rather act as "oracles" generating predictions and classifications, we decided to come up with a complementing approach.

We present here a simple method for deriving boolean classifiers (signatures) for all the groups of samples as small boolean expressions in disjunctive normal form. Those signatures represent most occurring patterns in the respective sample groups and can be based on to reason further about the properties of each group. In the same time, our modeling method is not meant for deriving highly detailed models from microarray data that can be used for accurate simulations. We rather suggest here a way to understand better the observed data in simple terms, that can aid in further efforts of building accurate complex models for the phenomena under studies.

As a continuation of this research, we plan to use a number of other classification methods (such as, Classification and Regression Tree(CART) [42] and Naïve Bayes Classification [43]) for the selection of subsets of significant genes. To measure the performance of the method, we will take into account different measures which are capable of evaluating binary classification (such as, Kappa [44] and Mattew's correlation coefficient [45]). We are planning to test our method with more publicly available "gold standard" cancer-related microarray datases (for instance, from the Cancer Genome Atlas [46]). Our goal is to produce a well-performing logicome method which is capable of handling several constrains without compromising the classification accuracy.

Acknowledgments

This work was partially funded by the Academy of Finland (project 267915), the Otto A. Malm foundation and the Oskar Öflunds stiftelse foundation. We thank Professor Ion Petre (Computational Biomodeling Laboratory, Åbo Akademi University, Turku, Finland) for his instructions in the project and Dr. Sepinoud Azimi (Computational Biomodeling Laboratory, Åbo Akademi University and Turku Centre for Computer Science) for her suggestions to the methodology of this study.

References

[1] A. Perez-Diez, A. Morgun, and N. Shulzhenko, "Microarrays for cancer diagnosis and classification," *Microarray Technology and Cancer Gene Profiling*, pp. 74–85, 2007.

[2] N. Ramachandran, S. Srivastava, and J. LaBaer, "Applications of protein microarrays for biomarker discovery," *Proteomics-Clinical Applications*, vol. 2, no. 10-11, pp. 1444–1459, 2008.

[3] A. L. Tarca, R. Romero, and S. Draghici, "Analysis of microarray experiments of gene expression profiling," *American journal of obstetrics and gynecology*, vol. 195, no. 2, pp. 373–388, 2006.

[4] M. Chaves, E. Sontag, and R. Albert, "Methods of robustness analysis for boolean models of gene control networks," *IEE Proceedings - Systems Biology*, vol. 153, no. 4, p. 154, 2006.

[5] L. Glass and S. A. Kauffman, "The logical analysis of continuous, non-linear biochemical control networks," *Journal of Theoretical Biology*, vol. 39, no. 1, pp. 103–129, April 1973.

[6] M. Anthony and P. L. Hammer, "A boolean measure of similarity," *Discrete Applied Mathematics*, vol. 154, no. 16, pp. 2242–2246, 2006.

[7] C. Panchal, S. Azimi, and I. Petre, "Generating the logicome of a biological network," in *Algorithms for Computational Biology*. Springer Nature, 2016, pp. 38–49.

[8] L. De Cecco, M. Nicolau, M. Giannoccaro, M. G. Daidone, P. Bossi, L. Locati, L. Licitra, and S. Canevari, "Head and neck cancer subtypes with biological and clinical relevance: Meta-analysis of gene-expression data," *Oncotarget*, vol. 6, no. 11, pp. 9627–42, 2015.

[9] A. Jemal, R. Siegel, E. Ward, Y. Hao, J. Xu, T. Murray, and M. J. Thun, "Cancer statistics, 2008," *CA: A cancer journal for clinicians*, vol. 58, no. 2, pp. 71–96, 2008.

[10] J. Massano, F. S. Regateiro, G. Januário, and A. Ferreira, "Oral squamous cell carcinoma: review of prognostic and predictive factors," *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology*, vol. 102, no. 1, pp. 67–76, 2006.

[11] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko *et al.*, "NCBI GEO: archive for functional genomics data setsupdate," *Nucleic acids research*, vol. 41, no. D1, pp. D991–D995, 2013.

[12] B. Lallemant, A. Evrard, C. Combescure, H. Chapuis, G. Chambon, C. Raynal, C. Reynaud, O. Sabra, D. Joubert, F. Hollande *et al.*, "Clinical relevance of nine transcriptional molecular markers for the diagnosis of head and neck squamous cell carcinoma in tissue and saliva rinse," *BMC cancer*, vol. 9, no. 1, p. 370, 2009.

[13] M. E. Whipple, E. Mendez, D. G. Farwell, S. N. Agoff, and C. Chen, "A genomic predictor of oral squamous cell carcinoma," *The Laryngoscope*, vol. 114, no. 8, pp. 1346–1354, 2004.

[14] K.-Y. Kim, X. Zhang, and I.-H. Cha, "Identification of human papillomavirus status specific biomarker in head and neck cancer," *Head & neck*, vol. 37, no. 9, pp. 1310–1318, 2015.

[15] K. Dahiya and R. Dhankhar, "Updated overview of current biomarkers in head and neck carcinoma," *World Journal of Methodology*, vol. 6, no. 1, pp. 77–86, 2016.

[16] P. Choi and C. Chen, "Genetic expression profiles and biologic pathway alterations in head and neck squamous cell carcinoma," *Cancer*, vol. 104, no. 6, pp. 1113–1128, 2005.

[17] G. A. Jeon, J.-S. Lee, V. Patel, J. S. Gutkind, S. S. Thorgeirsson, E. C. Kim, I.-S. Chu, P. Amornphimoltham, and M. H. Park, "Global gene expression profiles of human head and neck squamous carcinoma cell lines," *International journal of cancer*, vol. 112, no. 2, pp. 249–258, 2004.

[18] S. V. Thangaraj, V. Shyamsundar, A. Krishnamurthy, P. Ramani, K. Ganesan, M. Muthuswami, and V. Ramshankar, "Molecular portrait of oral tongue squamous cell carcinoma shown by integrative meta-analysis of expression profiles with validations," *PloS One*, vol. 11, no. 6, p. e0156582, 2016.

[19] A. Shukla, A. Singh, and R. Srivastava, "Oral submucous fibrosis: an update on etiology and pathogenesis-a review," *Rama Univ J Dent Sci*, vol. 2, pp. 24–33, 2015.

[20] J.-z. Li, H.-y. Pan, J.-w. Zheng, X.-j. Zhou, P. Zhang, W.-t. Chen, and Z.-y. Zhang, "Benzo (a) pyrene induced tumorigenesity of human immortalized oral epithelial cells: transcription profiling," *Chinese Medical Journal (English Edition)*, vol. 121, no. 19, p. 1882, 2008.

[21] C. Chen, E. Méndez, J. Houck, W. Fan, P. Lohavanichbutr, D. Doody, B. Yueh, N. D. Futran, M. Upton, D. G. Farwell *et al.*, "Gene expression profiling identifies genes predictive of oral squamous cell carcinoma," *Cancer Epidemiology and Prevention Biomarkers*, vol. 17, no. 8, pp. 2152–2162, 2008.

[22] G. Ohmura, T. Tsujikawa, T. Yaguchi, N. Kawamura, S. Mikami, J. Sugiyama, K. Nakamura, A. Kobayashi, T. Iwata, H. Nakano *et al.*, "Aberrant myosin 1b expression promotes cell migration and lymph node metastasis of HNSCC," *Molecular Cancer Research*, vol. 13, no. 4, pp. 721–731, 2015.

[23] G. A. Toruner, C. Ulger, M. Alkan, A. T. Galante, J. Rinaggio, R. Wilk, B. Tian, P. Soteropoulos, M. R. Hameed, M. N. Schwalb *et al.*, "Association between gene expression profile and tumor invasion in oral squamous cell carcinoma," *Cancer genetics and cytogenetics*, vol. 154, no. 1, pp. 27–35, 2004.

[24] M. Kuriakose, W. Chen, Z. He, A. Sikora, P. Zhang, Z. Zhang, W. Qiu, D. Hsu, C. McMunn-Coffran, S. Brown *et al.*, "Selection and validation of differentially expressed genes in head and neck cancer," *Cellular and molecular life sciences*, vol. 61, no. 11, pp. 1372–1383, 2004.

[25] H. Ye, T. Yu, S. Temam, B. L. Ziober, J. Wang, J. L. Schwartz, L. Mao, D. T. Wong, and X. Zhou, "Transcriptomic dissection of tongue squamous cell carcinoma," *BMC genomics*, vol. 9, no. 1, p. 69, 2008.

[26] H. Jr, D. W, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*. John Wiley & Sons, 2013, vol. 398.

[27] B. Madhu, N. Ashok, and S. Balasubramanian, "A multinomial logistic regression analysis to study the influence of residence and socioeconomic status on breast cancer incidences in southern karnataka," *Int. J. Math. Stat. Invention*, vol. 2, no. 5, pp. 01–8, 2014.

[28] B. G. Tabachnick, L. S. Fidell, and S. J. Osterlind, "Using multivariate statistics," 2001.

[29] F. J. Janzen and H. S. Stern, "Logistic regression for empirical studies of multivariate selection," *Evolution*, pp. 1564–1571, 1998.

[30] B. Chen, X. Chen, B. Li, Z. He, H. Cao, and G. Cai, "Reliability estimation for cutting tools based on logistic regression model using vibration signals," *Mechanical Systems and Signal Processing*, vol. 25, no. 7, pp. 2526–2537, 2011.

[31] W. Zhu, N. Zeng, N. Wang *et al.*, "Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS implementations," *NESUG proceedings: health care and life sciences, Baltimore, Maryland*, pp. 1–9, 2010.

[32] M. H. Zweig and G. Campbell, "Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine." *Clinical chemistry*, vol. 39, no. 4, pp. 561–577, 1993.

[33] L. Gautier, M. Møller, L. Friis-Hansen, and S. Knudsen, "Alternative mapping of probes to genes for affymetrix chips," *BMC bioinformatics*, vol. 5, no. 1, p. 1, 2004.

[34] S. Davis and P. S. Meltzer, "GEOquery: a bridge between the gene expression omnibus (GEO) and bioconductor," *Bioinformatics*, vol. 23, no. 14, pp. 1846–1847, 2007.

[35] R. A. Irizarry, "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics*, vol. 4, no. 2, pp. 249–264, April 2003.

[36] Z. J. Wu and R. Irizarry, "Description of gcrma package," 2010.

[37] J. A. Berger, S. Hautaniemi, A.-K. Järvinen, H. Edgren, S. K. Mitra, and J. Astola, "Optimized lowess normalization parameter selection for dna microarray data," *BMC bioinformatics*, vol. 5, no. 1, p. 194, 2004.

[38] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry *et al.*, "Bioconductor: open software development for computational biology and bioinformatics," *Genome biology*, vol. 5, no. 10, p. 1, 2004.

[39] G. Dennis, B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane, and R. A. Lempicki, "David: database for annotation, visualization, and integrated discovery," *Genome biology*, vol. 4, no. 9, p. R60, 2003.

[40] F. Dundar, L. Skrabanek, and P. Zumbo, "Introduction to differential gene expression analysis using RNA-seq," *Applied Bioinformatics Core/Weill Cornell Medical College*, pp. 1–67, 2015.

[41] M. M. Kim, A. L. Carvalho, C. Jeronimo, R. Henrique, W. M. Koch, D. Sidransky, and J. Califano, "Identification of aim1 as a hypermethylation-inactivated tumor suppressor gene in HNSCC," *Otolaryngology-Head and Neck Surgery*, vol. 131, no. 2, pp. P173–P174, 2004.

[42] M. A. Razi and K. Athappilly, "A comparative predictive analysis of neural networks (nns), nonlinear regression and classification and regression tree (CART) models," *Expert Systems with Applications*, vol. 29, no. 1, pp. 65–74, 2005.

[43] K. M. Leung, "Naïve bayesian classifier," *Polytechnic University Department of Computer Science/Finance and Risk Engineering*, 2007.

[44] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochemia medica*, vol. 22, no. 3, pp. 276–282, 2012.

[45] B. W. Matthews, "Comparison of the predicted and observed secondary structure of t4 phage lysozyme," *Biochimica et Biophysica Acta (BBA)-Protein Structure*, vol. 405, no. 2, pp. 442–451, 1975.

[46] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J. M. Stuart, C. G. A. R. Network *et al.*, "The cancer genome atlas pan-cancer analysis project," *Nature genetics*, vol. 45, no. 10, pp. 1113–1120, 2013.

Paper III

Reaction systems models for the self-assembly of intermediate filaments.

Sepinoud Azimi, Charmi Panchal, Eugen Czeizler, and Ion Petre.

Originally published in Annals of University of Bucharest. Vol 62, pp: 9-24. Editura Universităţii Bucureşti, 2015.

Reaction Systems Models for the Self-Assembly of Intermediate Filaments

SEPINOUD AZIMI AND CHARMI PANCHAL
EUGEN CZEIZLER AND ION PETRE

Abstract - Reaction systems are a recent addition to the spectrum of computational modeling frameworks. We construct in this paper several reaction systems models for the self-assembly of vimentin tetramers into intermediate filaments. We demonstrate that reaction systems are a versatile modeling framework, able to capture several aspects of the dynamics of the self-assembly of intermediate filaments using only simple, set-theoretical-based concepts.

Key words and phrases : Reaction systems; modeling; intermediate filaments; self-assembly.

1 Introduction

Reaction systems were introduced in [9] as a new modelling framework inspired by the functioning/bio-energetics of the living cell. It differs drastically from the traditional modelling frameworks (such as ordinary differential equations (ODEs), stochastic processes, Boolean networks, state machines) in focusing on reactions and in having the environment as an integral part of the model. Reactions in reaction systems are enabled through promotion and inhibition, see [4]. Reaction systems are based on two main principles. The first one, called the *threshold principle* makes reaction systems a qualitative framework by stating that if available, a resource (reactant) is available abundantly. In other words, reactions can not limit each other through a quantitative competition on resources. The second principle, called the *no permanency principle*, states that a resource or reactant vanishes unless sustained by a reaction. That is to say, the next state of a reaction system is only obtained from the output of the reaction enabled in the previous state plus the contribution of the environment.

Studies on reaction systems have been quite diverse, see for example [1, 3, 5, 10, 11, 15, 19, 22–25]. They can be categorized into two main streams. While the first direction is concerned with more theoretical aspects of reaction systems (e.g., [8,10,11,24,25]), the second direction has taken a more practical spin, mainly regarding reaction systems as a platform to do bio-modelling (e.g., [1, 3, 5]).

In this paper we follow the second line of research and focus on the expressivity of reaction systems as a modelling framework. We consider a case-study on the self-assembly of intermediate filaments from vimentin tetramers. We start from the molecular models and the ODE-based analysis in [6]. We build several different reaction systems models based on the molecular model of [6] that capture the self-assembly process at two levels of resolution: one where no distinction is made between filaments of different lengths (resolution 0) and one where the model distinguishes between unit-length, short and long filaments (resolution 2).

2 Preliminaries

In this section we recall some of the basic definitions of reaction systems that we need throughout the paper.

Reactions are the building blocks of a reaction system. Intuitively, a reaction is triggered if all resources needed for the reaction are available in the environment and there exist no species that inhibit the reaction. In this case, the reaction transforms the set of resources to the set of products. This intuition is formally captured in the definition of a reaction in reaction systems framework as follows, see [4, 9] for more details.

Definition 2.1 *[9] A* reaction *is a tuple $a = (R_a, I_a, P_a)$, where R_a, I_a and P_a are finite, non-empty sets and $R_a \cap I_a = \emptyset$. The sets R_a, I_a and P_a are called the set of* reactants, inhibitors *and* products *of a, respectively. We say a is a reaction over set S, if R_a, I_a, $P_a \subseteq S$. We denote the set of reactions in S by $\mathsf{rac}(S)$.*

We next define the result of applying a reaction and a set of reactions on a given set. In this definition the result of applying a number of reactions to a set is the collective result of applying each reaction to the set independently. Indeed this is true because by *threshold assumption* there is no competition for the resources between different reactions and hence running a reaction does not inhibit enabling any other one.

Definition 2.2 *[9] Let A be a set of reactions, $a \in A$ and T a set.*

(i) The result *of a on T, denoted by $res_a(T)$, is*

$$res_a(T) = \begin{cases} P_a, & \text{if } R_a \subseteq T \text{ and } I_a \cap T = \emptyset \\ \emptyset, & \text{otherwise.} \end{cases}$$

(ii) The result of A on T, *denoted by $res_A(T)$, is*

$$res_A(T) = \bigcup_{a \in A} res_a(T).$$

A reaction system *(RS in short) is defined as an ordered pair $\mathcal{A} = (S, A)$, where S is a finite set and $A \subseteq \mathsf{rac}(S)$. Set S is called the* background set *of $\mathcal{A}$.*

To capture the dynamics of a given reaction system the notion of *interactive process* has been introduced. In what follows we present a formal definition of such a process.

Definition 2.3 *[9] For a given reaction system $\mathcal{A}$, an* interactive process *in $\mathcal{A}$ is a pair $\pi = (\gamma, \delta)$, where $\gamma = C_0, C_1, ..., C_n, \delta = D_1, D_2, ..., D_n \subseteq S$, $n \geq 1$, with $D_1 = \text{res}_{\mathcal{A}}(C_0)$ and, for every $1 < i \leq n$, $D_i = \text{res}_{\mathcal{A}}(C_{i-1} \cup D_{i-1})$.*

The sequences γ and δ are called the context sequence *and the* result sequence *of π, respectively. The sequence $\tau = W_0, W_1, ..., W_n$ is the* state sequence *of π, where $W_0 = C_0$ and $W_i = C_i \cup D_i$, for all $i \in \{0, ..., n\}$. W_0 is called the* initial state *of π.*

Next we recall the definition of a reaction system's *steady state*, introduced in [3].

Definition 2.4 *[3] Let $\mathcal{A} = (S, A)$ be a reaction system and $C \subseteq S$. We say that $D \subseteq S$ is a* steady state *of $\mathcal{A}$ for C if $\text{res}_{\mathcal{A}}(C \cup D) = D$.*

3 A model for self-assembly of intermediate filaments

In this section we describe the *in-vitro* assembly principles of vimentin filaments, as a representative for the class of intermediate filaments proteins. Based on the recent studies in [6] and [21] we present both a basic and a refined molecular model for vimentin assembly, from the level of first stable subunits till the emerging of mature filaments.

Intermediate filaments (IFs) are one of the three types of protein filaments inside the eukaryotic cell [26]. Together with *microtubules* and *actin filaments*, they form the *cytoskeleton*, which is a complex network of filaments with active role in a number of cellular processes, including sustaining the mechanical integrity of the cell, controlling its shape, but also facilitating the intracellular transport [20].

IF subunits are α-helical rods which assemble by both lateral and end-to-end interactions into rope-like filaments [13]. The length of these filaments ranges from hundreds of nm long to micro-meter values, while their width (when in mature state) is preserved at 11 nm. We are particularly interested in the (*in-vitro*) assembly of IF generated from human vimentin proteins; one of the several types of eukaryotic IF proteins [14]. In the case of vimentin-based IFs the *in-vitro* assembly process follows four stages.

The first assembly stage of vimentin intermediate filaments is the fast lateral association of monomers into dimers and subsequently into tetramers (denoted as T). Tetramers are the first stable filament subunits, as both monomers and dimers are not chemically stable. Moreover, for the case of *in-vitro* assembly, the process can be blocked/freezed after tetramer formation, and the system can be initialized starting from this level. Thus, when modelling the *in-vitro* assembly this first stage is generally omitted, and the process is assumed to start from a mono-populated system of tetrameric subunits.

The second phase of the vimentin IF assembly consists of further lateral associations: two tetramers join to form an octamer (denoted as O), two octamers join to form a hexadecamer (denoted as H), while two hexadecamer join to form a unit length filament (ULF). ULFs are the basic unit of the mature filament structure, as from now on, the formation and elongation of the filaments is performed only through end-to-end association reactions.

The third stage represents the formation and elongation of filaments from individual ULF's. Here, on one hand we have elongation reactions, when filament complexes are enlarged by one ULF at a time, and merger reactions on the other, where two longer filaments join by end-to-end interactions and form a longer complex. Depending on the number k of constituent ULF's within one filament, we can differentiate between the emergent assemblies based on their "size" k.

The final assembly stage represents a radial compaction of the filaments from a ULF diameter of about 15 nm to a filament diameter of about 11 nm, see [13] for details. However, from a modeler point of view, we can consider the assembly process as complete after the first three stages above, as the radial compaction does not modify the ULF per filament ratio.

3.1 Basic and refined molecular models

A common problem in modelling self-assembly systems is dealing with the combinatorial explosion of all different emergent assemblies as possible distinct species. In the case of the IF model above, this translates into the problem of representing and reasoning about all the emergent filaments of size 1, 2, 3, etc. Depending on the level of details the modeler chooses to describe the assembly process, there might be a number of models which could describe such a process. In our study we concentrate over two such detail levels, thus generating two models for IF assembly; we call them the *basic* and the *refined model*.

In the basic model, we consider the ULF's as elementary filaments (generically denoted as F), and we do not differentiate between them and other filaments. The molecular model of this basic representation is presented in Table 1(a). On the refined model, in order to be able to capture the formation and dynamics of short vs long emerging filaments we differentiate between ULFs (filaments of size 1, denoted as $\mathsf{F_u}$), filaments of size 2 (short filaments, denoted as $\mathsf{F_s}$), and filaments of size 3 or more (long filaments, denoted as $\mathsf{F_l}$). The molecular model of this second refined system is presented in Table 1(b).

The above basic model, as well as a refined version differentiating between ULFs and the remaining filaments, were introduced and analyzed in [6] in correlation with experimental results for *in-vitro* vimentin self-assembly taken from [18]. ODE mathematical models based on mass action kinetics formulations were derived, numerically fitted, and validated using data from [18]. A generic method of quantitative model refinement was introduced and discussed in [12] and [21].

Table 1: The molecular models of (a) the basic and (b) the refined representations of the IF assembly process.

(a) Basic model		(b) Refined model	
$2\,\mathsf{T} \rightarrow \mathsf{O}$	(1)	$2\,\mathsf{T} \rightarrow \mathsf{O}$	(5)
$2\,\mathsf{O} \rightarrow \mathsf{H}$	(2)	$2\,\mathsf{O} \rightarrow \mathsf{H}$	(6)
$2\,\mathsf{H} \rightarrow \mathsf{F}$	(3)	$2\,\mathsf{H} \rightarrow \mathsf{F_u}$	(7)
$2\,\mathsf{F} \rightarrow \mathsf{F}$	(4)	$2\,\mathsf{F_u} \rightarrow \mathsf{F_s}$	(8)
		$\mathsf{F_u} + \mathsf{F_s} \rightarrow \mathsf{F_l}$	(9)
		$\mathsf{F_u} + \mathsf{F_l} \rightarrow \mathsf{F_l}$	(10)
		$2\,\mathsf{F_s} \rightarrow \mathsf{F_l}$	(11)
		$\mathsf{F_s} + \mathsf{F_l} \rightarrow \mathsf{F_l}$	(12)
		$2\,\mathsf{F_l} \rightarrow \mathsf{F_l}$	(13)

Based on those methods we can estimate the kinetic rate constants of the current refined model in Table 1(b), in order to obtain a perfect fit-preserving refinement of the models in [6].

3.2 Variants of the kinetic model

The previously introduced basic and refined models were numerically fitted in [6] in order to corroborate the experimental data of vimentin assembly reported in [18]. However, we show in the following that by modifying the kinetic rate constants of these reactions we can generate several setups with different overall behaviours. These behaviours could be differentiated by measuring the average length of the emerging filament populations. The formula used for generating these measurements was introduced in [18] and updated in [6]. The formula can be easily extended for any level of refinement, including the one considered in our models.

In Figure 1 we present the average filament length over time for the model of [6], as well as for two different kinetic setups of the same model. The plot in the case of the original model is shown in Figure 1 with a solid line. We modified the model of [6] by inhibiting reactions (11) and (12); the result is shown with a dashed line in Figure 1. In another modification of the model in [6], we inhibited reactions (9) and (11); the result is shown with a dotted line in Figure 1. As it can be seen from Figure 1, the three kinetic setups of the model in [6] yield very different results. In the original setup of [6], the model favors the formation of (fewer and fewer) long filaments. The other two setups favour the formation of both short and long filaments, and that of only short filaments, respectively.

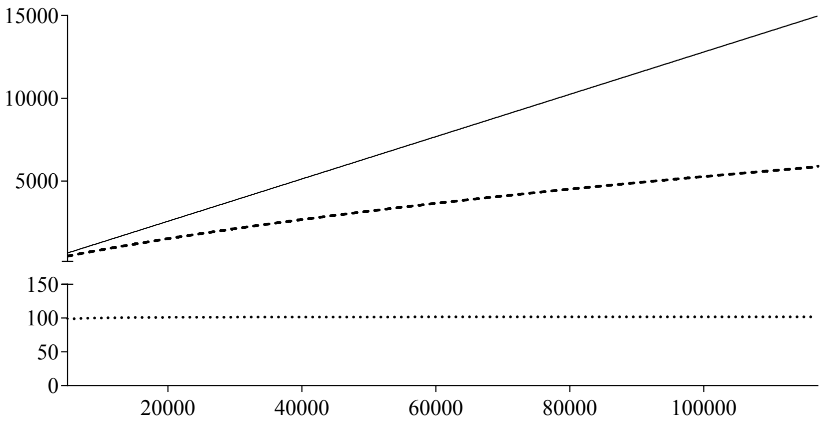


Figure 1: The plot of the time evolution of the average filament length in three different setups. The solid line plot represents the original model of [6], favouring the formation of long filaments. The dashed line plots represents a variant of the model where both short and long filaments are formed. The dotted line plot represents a variant of the model where only short filaments are formed. The three models have the same set of reactions and only differ in their kinetic setups.

4 Reaction systems models

In this section we present the corresponding RS-based models for self-assembly of intermediate filaments introduced in Section 3.

4.1 Basic reaction systems model

First we build an RS-based model for the basic self-assembly of intermediate filaments introduced in Table 1(a). To formulate the reactions of the corresponding RS, every reaction of type $A_1 + A_2 \to B$ in Table 1(a) is translated to a reaction $(\{A_1, A_2\}, \{\mathsf{d_I}\}, \{B\})$ in the corresponding reaction system. The dummy variable, $\mathsf{d_I}$, is only used to respect the constraint that the set of inhibitors of all RS reactions should be non-empty, see [9]. Note that the coefficients of the species in the chemical reaction do not play a role here since we are translating a quantitative framework to a qualitative one and by *threshold assumption* we know that existence of a reactant in the environment implies the abundance of it as well. The RS-based model for the basic self-assembly of intermediate filaments is presented in Table 2.

We are interested in analysing the dynamics of the RS-based model and compare it with the corresponding properties of the quantitative ODE-based model for IF assembly, discussed in Section 3. To construct an environment in which tetramers are always present, we consider $\{\mathsf{T}\}$ as the given context in every step of the interactive process. This corresponds to the ODE models having an initial

Table 2: The direct translation of the biochemical reactions of the basic model to a reaction system $\mathcal{A} = (S, A)$ where $S = \{\mathsf{T}, \mathsf{O}, \mathsf{H}, \mathsf{F}, \mathsf{d_I}\}$.

Reaction in the chemical network	Reaction in the reaction system	
$2\,\mathsf{T} \to \mathsf{O}$	$(\{\mathsf{T}\}, \{\mathsf{d_I}\}, \{\mathsf{O}\})$	(14)
$2\,\mathsf{O} \to \mathsf{H}$	$(\{\mathsf{O}\}, \{\mathsf{d_I}\}, \{\mathsf{H}\})$	(15)
$2\,\mathsf{H} \to \mathsf{F}$	$(\{\mathsf{H}\}, \{\mathsf{d_I}\}, \{\mathsf{F}\})$	(16)
$2\,\mathsf{F} \to \mathsf{F}$	$(\{\mathsf{F}\}, \{\mathsf{d_I}\}, \{\mathsf{F}\})$	(17)

large pool of tetramers that are assembling into IFs. The interactive process is illustrated in Table 3. The result state sequences of the examples of this paper were obtained by using the reaction system simulator proposed in [2]. The simulator can be reached at [27]. The interactive process thus obtained shows that the model enters into a steady state, see [3], where tetramers, octamers, hexadecamers and filaments are present.

Table 3: An interactive process for the basic RS model. The interactive process enters a loop after the second state from which every state contains all species of the system.

State	C_i	D_i	W_i
0	$\{\mathsf{T}\}$	$\emptyset$	$\{\mathsf{T}\}$
1	$\{\mathsf{T}\}$	$\{\mathsf{O}\}$	$\{\mathsf{T}, \mathsf{O}\}$
2	$\{\mathsf{T}\}$	$\{\mathsf{O}, \mathsf{H}\}$	$\{\mathsf{T}, \mathsf{O}, \mathsf{H}\}$
3	$\{\mathsf{T}\}$	$\{\mathsf{O}, \mathsf{H}, \mathsf{F}\}$	$\{\mathsf{T}, \mathsf{O}, \mathsf{H}, \mathsf{F}\}$
4	$\{\mathsf{T}\}$	$\{\mathsf{O}, \mathsf{H}, \mathsf{F}\}$	$\{\mathsf{T}, \mathsf{O}, \mathsf{H}, \mathsf{F}\}$

4.2 Refined reaction systems model

In this section we are taking a step further to increase the level of resolution of our model. To do so, we apply the proposed refinement approach of Table 1(b) and modify our basic model to fit this new information. The methodology for translating the chemical reaction network to the RS-based model does not change with the new setup. The obtained RS-based model is presented in Table 4. The species $\mathsf{F_u}$, $\mathsf{F_s}$ and $\mathsf{F_l}$ correspond to *unit length filament*, *short filament* and *long filament* respectively.

Similarly as in the case of the basic model, we analyze the dynamics of the refined model by running an interactive process with a constant $\{\mathsf{T}\}$ as the given context in every step. The interactive process is presented in Table 5. We conclude

Table 4: The direct translation of the biochemical reactions of the refined model to a reaction system $\mathcal{A}' = (S', A')$ where $S' = \{\mathsf{T}, \mathsf{O}, \mathsf{H}, \mathsf{F_u}, \mathsf{F_s}, \mathsf{F_l}, \mathsf{d_l}\}$.

Reaction in the chemical network	Reaction in the reaction system	
$2\,\mathsf{T} \to \mathsf{O}$	$(\{\mathsf{T}\}, \{\mathsf{d_l}\}, \{\mathsf{O}\})$	(18)
$2\,\mathsf{O} \to \mathsf{H}$	$(\{\mathsf{O}\}, \{\mathsf{d_l}\}, \{\mathsf{H}\})$	(19)
$2\,\mathsf{H} \to \mathsf{F_u}$	$(\{\mathsf{H}\}, \{\mathsf{d_l}\}, \{\mathsf{F_u}\})$	(20)
$2\,\mathsf{F_u} \to \mathsf{F_s}$	$(\{\mathsf{F_u}\}, \{\mathsf{d_l}\}, \{\mathsf{F_s}\})$	(21)
$\mathsf{F_u} + \mathsf{F_s} \to \mathsf{F_l}$	$(\{\mathsf{F_u}, \mathsf{F_s}\}, \{\mathsf{d_l}\}, \{\mathsf{F_l}\})$	(22)
$2\,\mathsf{F_s} \to \mathsf{F_l}$	$(\{\mathsf{F_s}\}, \{\mathsf{d_l}\}, \{\mathsf{F_l}\})$	(23)
$\mathsf{F_u} + \mathsf{F_l} \to \mathsf{F_l}$	$(\{\mathsf{F_u}, \mathsf{F_l}\}, \{\mathsf{d_l}\}, \{\mathsf{F_l}\})$	(24)
$\mathsf{F_s} + \mathsf{F_l} \to \mathsf{F_l}$	$(\{\mathsf{F_s}, \mathsf{F_l}\}, \{\mathsf{d_l}\}, \{\mathsf{F_l}\})$	(25)
$2\,\mathsf{F_l} \to \mathsf{F_l}$	$(\{\mathsf{F_l}\}, \{\mathsf{d_l}\}, \{\mathsf{F_l}\})$	(26)

that, similarly as in the case of the basic model, we reach a steady state where tetramers, octamers, hexadecamers, and the three types of filaments are present. This is of course consistent with this model being a refinement of the basic model.

Table 5: An interactive process for the refined RS model. The interactive process enters a loop after the fourth state from which every state contains all species of the system.

State	C_i	D_i	W_i
0	$\{\mathsf{T}\}$	$\emptyset$	$\{\mathsf{T}\}$
1	$\{\mathsf{T}\}$	$\{\mathsf{O}\}$	$\{\mathsf{T}, \mathsf{O}\}$
2	$\{\mathsf{T}\}$	$\{\mathsf{O}, \mathsf{H}\}$	$\{\mathsf{T}, \mathsf{O}, \mathsf{H}\}$
3	$\{\mathsf{T}\}$	$\{\mathsf{O}, \mathsf{H}, \mathsf{F_u}\}$	$\{\mathsf{T}, \mathsf{O}, \mathsf{H}, \mathsf{F_u}\}$
4	$\{\mathsf{T}\}$	$\{\mathsf{O}, \mathsf{H}, \mathsf{F_u}, \mathsf{F_s}\}$	$\{\mathsf{T}, \mathsf{O}, \mathsf{H}, \mathsf{F_u}, \mathsf{F_s}\}$
5	$\{\mathsf{T}\}$	$\{\mathsf{O}, \mathsf{H}, \mathsf{F_u}, \mathsf{F_s}, \mathsf{F_l}\}$	$\{\mathsf{T}, \mathsf{O}, \mathsf{H}, \mathsf{F_u}, \mathsf{F_s}, \mathsf{F_l}\}$
6	$\{\mathsf{T}\}$	$\{\mathsf{O}, \mathsf{H}, \mathsf{F_u}, \mathsf{F_s}, \mathsf{F_l}\}$	$\{\mathsf{T}, \mathsf{O}, \mathsf{H}, \mathsf{F_u}, \mathsf{F_s}, \mathsf{F_l}\}$

4.3 Variants of the refined reaction systems model

We consider in this section the corresponding reaction systems of the setup considered in Section 3.2 for the ODE-based model: we focus on modifying our RS model to control the length of the filaments produced within the system. Three different variants of the RS-based model proposed in Section 3 are presented. In the first variant, we are interested in having only long filaments in the result state, whereas in the second one we expect to produce only short filaments. The last variation is responsive to what is asked by the modeler through the context of the system, i.e. reaction system produces only short filaments if short is part of the context and long filaments if long is included in the context. If the context contains neither short nor long, the filaments of all lengths are present in the result state. In what follows we discuss each of these variations separately.

Variant one - only long filaments. We have modified the reaction system presented in Table 4 to produce only short filaments by adding $\{\mathsf{F_l}\}$ to the set of inhibitors of reactions (21) and (22). In this way, we effectively favour reaction (23) over reactions (21) and (22) and the unit length filaments immediately get to elongate the existing long filaments rather than short filaments. The corresponding RS model is presented in Table 6. The efficiency of the reaction system structure can be observed through the steps of the interactive process illustrated in Table 7: in the steady state there are no short filaments.

Variant two - only short filaments. In this case we modified the reaction system of Table 4 by removing reactions (22) and (23). The corresponding reaction system is presented in Table 6. On one hand, if no long filaments are introduced by the context, the system produces only short filaments. On the other hand, if long filaments are added by the context, they would get extended recurrently. Both cases can be observed in the interactive process shown in Table 8.

Table 6: The list of reactions of the reaction systems corresponding to the first two variants of the refined model. For both reaction systems the background set is $\{\mathsf{T}, \mathsf{O}, \mathsf{H}, \mathsf{F_u}, \mathsf{F_s}, \mathsf{F_l}, \mathsf{d_l}\}$.

Reactions in variant one		Reactions in variant two	
$(\{\mathsf{T}\}, \{\mathsf{d_l}\}, \{\mathsf{O}\})$	(27)	$(\{\mathsf{T}\}, \{\mathsf{d_l}\}, \{\mathsf{O}\})$	(36)
$(\{\mathsf{O}\}, \{\mathsf{d_l}\}, \{\mathsf{H}\})$	(28)	$(\{\mathsf{O}\}, \{\mathsf{d_l}\}, \{\mathsf{H}\})$	(37)
$(\{\mathsf{H}\}, \{\mathsf{d_l}\}, \{\mathsf{F_u}\})$	(29)	$(\{\mathsf{H}\}, \{\mathsf{d_l}\}, \{\mathsf{F_u}\})$	(38)
$(\{\mathsf{F_u}\}, \{\mathsf{F_l}\}, \{\mathsf{F_s}\})$	(30)	$(\{\mathsf{F_u}\}, \{\mathsf{d_l}\}, \{\mathsf{F_s}\})$	(39)
$(\{\mathsf{F_u}, \mathsf{F_s}\}, \{\mathsf{F_l}\}, \{\mathsf{F_l}\})$	(31)	$(\{\mathsf{F_u}, \mathsf{F_l}\}, \{\mathsf{d_l}\}, \{\mathsf{F_l}\})$	(40)
$(\{\mathsf{F_s}\}, \{\mathsf{d_l}\}, \{\mathsf{F_l}\})$	(32)	$(\{\mathsf{F_s}, \mathsf{F_l}\}, \{\mathsf{d_l}\}, \{\mathsf{F_l}\})$	(41)
$(\{\mathsf{F_u}, \mathsf{F_l}\}, \{\mathsf{d_l}\}, \{\mathsf{F_l}\})$	(33)	$(\{\mathsf{F_l}\}, \{\mathsf{d_l}\}, \{\mathsf{F_l}\})$	(42)
$(\{\mathsf{F_s}, \mathsf{F_l}\}, \{\mathsf{d_l}\}, \{\mathsf{F_l}\})$	(34)		
$(\{\mathsf{F_l}\}, \{\mathsf{d_l}\}, \{\mathsf{F_l}\})$	(35)		

Table 7: An interactive process for the first variant of the refined RS model. The interactive process enters a loop after the fifth state from which no short filament is produced.

State	C_i	D_i	W_i
0	$\{\mathsf{T}\}$	$\emptyset$	$\{\mathsf{T}\}$
1	$\{\mathsf{T}\}$	$\{\mathsf{O}\}$	$\{\mathsf{T},\mathsf{O}\}$
2	$\{\mathsf{T}\}$	$\{\mathsf{O},\mathsf{H}\}$	$\{\mathsf{T},\mathsf{O},\mathsf{H}\}$
3	$\{\mathsf{T}\}$	$\{\mathsf{O},\mathsf{H},\mathsf{F_u}\}$	$\{\mathsf{T},\mathsf{O},\mathsf{H},\mathsf{F_u}\}$
4	$\{\mathsf{T}\}$	$\{\mathsf{O},\mathsf{H},\mathsf{F_u},\mathsf{F_s}\}$	$\{\mathsf{T},\mathsf{O},\mathsf{H},\mathsf{F_u},\mathsf{F_s}\}$
5	$\{\mathsf{T}\}$	$\{\mathsf{O},\mathsf{H},\mathsf{F_u},\mathsf{F_s},\mathsf{F_l}\}$	$\{\mathsf{T},\mathsf{O},\mathsf{H},\mathsf{F_u},\mathsf{F_s},\mathsf{F_l}\}$
6	$\{\mathsf{T}\}$	$\{\mathsf{O},\mathsf{H},\mathsf{F_u},\mathsf{F_l}\}$	$\{\mathsf{T},\mathsf{O},\mathsf{H},\mathsf{F_u},\mathsf{F_l}\}$
7	$\{\mathsf{T}\}$	$\{\mathsf{O},\mathsf{H},\mathsf{F_u},\mathsf{F_l}\}$	$\{\mathsf{T},\mathsf{O},\mathsf{H},\mathsf{F_u},\mathsf{F_l}\}$

Table 8: An interactive process for the second variant of the refined RS model. The interactive process enters a loop after the third state from which no long filament is produced until $\mathsf{F_l}$ is introduced to the fifth state which triggers the production of long filaments in the consecutive states.

State	C_i	D_i	W_i
0	$\{\mathsf{T}\}$	$\emptyset$	$\{\mathsf{T}\}$
1	$\{\mathsf{T}\}$	$\{\mathsf{O}\}$	$\{\mathsf{T},\mathsf{O}\}$
2	$\{\mathsf{T}\}$	$\{\mathsf{O},\mathsf{H}\}$	$\{\mathsf{T},\mathsf{O},\mathsf{H}\}$
3	$\{\mathsf{T}\}$	$\{\mathsf{O},\mathsf{H},\mathsf{F_u}\}$	$\{\mathsf{T},\mathsf{O},\mathsf{H},\mathsf{F_u}\}$
4	$\{\mathsf{T}\}$	$\{\mathsf{O},\mathsf{H},\mathsf{F_u},\mathsf{F_s}\}$	$\{\mathsf{T},\mathsf{O},\mathsf{H},\mathsf{F_u},\mathsf{F_s}\}$
5	$\{\mathsf{T},\mathsf{F_l}\}$	$\{\mathsf{O},\mathsf{H},\mathsf{F_u},\mathsf{F_s}\}$	$\{\mathsf{T},\mathsf{O},\mathsf{H},\mathsf{F_u},\mathsf{F_s},\mathsf{F_l}\}$
6	$\{\mathsf{T}\}$	$\{\mathsf{O},\mathsf{H},\mathsf{F_u},\mathsf{F_s},\mathsf{F_l}\}$	$\{\mathsf{T},\mathsf{O},\mathsf{H},\mathsf{F_u},\mathsf{F_s},\mathsf{F_l}\}$
7	$\{\mathsf{T}\}$	$\{\mathsf{O},\mathsf{H},\mathsf{F_u},\mathsf{F_s},\mathsf{F_l}\}$	$\{\mathsf{T},\mathsf{O},\mathsf{H},\mathsf{F_u},\mathsf{F_s},\mathsf{F_l}\}$

Variant three - context dependent. To fit the requirements of the final variation, the reaction system of Table 4 is modified as follows:

- The background set is extended by adding three species, i.e. short, long and $\mathsf{F_{long}}$, to set S;
- $\mathsf{F_{long}}$ is added to the inhibitor set of reaction (21) to prevent producing short filaments whenever the context asks for the long ones;
- short is added to the inhibitor set of reactions (22), (23), (24), (25) and (26) to prevent producing long filaments whenever the context asks for the short ones and
- the set of reactions is extended by adding $(\{\mathsf{F_l}, \mathsf{long}\}, \{\mathsf{d_l}\}, \{\mathsf{F_{long}}\})$ to signal the start of long filament production to the system.

The corresponding reaction system is presented in Table 9, while its interactive process and the steady state with constant context $\{\mathsf{T}, \mathsf{short}\}$ and with constant context $\{\mathsf{T}, \mathsf{long}\}$ is described in Table 10.

Table 9: The list of reactions of the last variant of the refined model for reaction system $\mathcal{A}'' = (S'', A'')$ where $S'' = \{\mathsf{T}, \mathsf{O}, \mathsf{H}, \mathsf{F_u}, \mathsf{F_s}, \mathsf{F_l}, \mathsf{long}, \mathsf{short}, \mathsf{F_{long}}, \mathsf{d_l}\}$.

Reaction		Reaction	
$(\{\mathsf{T}\}, \{\mathsf{d_l}\}, \{\mathsf{O}\})$	(43)	$(\{\mathsf{F_s}\}, \{\mathsf{short}\}, \{\mathsf{F_l}\})$	(48)
$(\{\mathsf{O}\}, \{\mathsf{d_l}\}, \{\mathsf{H}\})$	(44)	$(\{\mathsf{F_u}, \mathsf{F_l}\}, \{\mathsf{short}\}, \{\mathsf{F_l}\})$	(49)
$(\{\mathsf{H}\}, \{\mathsf{d_l}\}, \{\mathsf{F_u}\})$	(45)	$(\{\mathsf{F_s}, \mathsf{F_l}\}, \{\mathsf{short}\}, \{\mathsf{F_l}\})$	(50)
$(\{\mathsf{F_u}\}, \{\mathsf{F_{long}}\}, \{\mathsf{F_s}\})$	(46)	$(\{\mathsf{F_l}\}, \{\mathsf{short}\}, \{\mathsf{F_l}\})$	(51)
$(\{\mathsf{F_u}, \mathsf{F_s}\}, \{\mathsf{short}\}, \{\mathsf{F_l}\})$	(47)	$(\{\mathsf{F_l}, \mathsf{long}\}, \{\mathsf{d_l}\}, \{\mathsf{F_{long}}\})$	(52)

5 Discussion

We investigated in this paper the expressivity of the reaction systems framework as a modelling formalism for biology. Our case study, on protein self-assembly, involves intermediate filaments of arbitrary size. This made the case study a good choice to test the expressivity of modelling with reaction systems. We showed that our reaction systems model is a good qualitative counterpart to quantitative modelling with ODE. Both types of models could demonstrate, albeit with a different language and tools, the formation of intermediate filaments from vimentin tetramers. In the case of ODE models we showed that by disabling various reactions, we could observe different outputs of the self assembly model. Thus, while

Table 10: An interactive process for the third variant of the refined RS model. In this example short is added to the context and the interactive process enters a loop after the third state from which no long filaments are produced. Next, long is added to the context and the interactive process produces only long filaments up to state 9. Once short is added again to the context in state 10, the system only produces short filaments.

State	C_i	D_i	W_i
0	$\{\mathsf{T}, \mathsf{short}\}$	$\emptyset$	$\{\mathsf{T}\}$
1	$\{\mathsf{T}, \mathsf{short}\}$	$\{\mathsf{O}\}$	$\{\mathsf{T}, \mathsf{O}\}$
2	$\{\mathsf{T}, \mathsf{short}\}$	$\{\mathsf{O}, \mathsf{H}\}$	$\{\mathsf{T}, \mathsf{O}, \mathsf{H}\}$
3	$\{\mathsf{T}, \mathsf{short}\}$	$\{\mathsf{O}, \mathsf{H}, \mathsf{F}_\mathsf{u}\}$	$\{\mathsf{T}, \mathsf{O}, \mathsf{H}, \mathsf{F}_\mathsf{u}\}$
4	$\{\mathsf{T}, \mathsf{short}\}$	$\{\mathsf{O}, \mathsf{H}, \mathsf{F}_\mathsf{u}, \mathsf{F}_\mathsf{s}\}$	$\{\mathsf{T}, \mathsf{O}, \mathsf{H}, \mathsf{F}_\mathsf{u}, \mathsf{F}_\mathsf{s}\}$
5	$\{\mathsf{T}, \mathsf{short}\}$	$\{\mathsf{O}, \mathsf{H}, \mathsf{F}_\mathsf{u}, \mathsf{F}_\mathsf{s}\}$	$\{\mathsf{T}, \mathsf{O}, \mathsf{H}, \mathsf{F}_\mathsf{u}, \mathsf{F}_\mathsf{s}\}$
6	$\{\mathsf{T}, \mathsf{long}\}$	$\{\mathsf{O}, \mathsf{H}, \mathsf{F}_\mathsf{u}, \mathsf{F}_\mathsf{s}, \mathsf{F}_\mathsf{l}, \mathsf{F}_\mathsf{long}\}$	$\{\mathsf{T}, \mathsf{O}, \mathsf{H}, \mathsf{F}_\mathsf{u}, \mathsf{F}_\mathsf{s}, \mathsf{F}_\mathsf{l}, \mathsf{F}_\mathsf{long}\}$
7	$\{\mathsf{T}, \mathsf{long}\}$	$\{\mathsf{O}, \mathsf{H}, \mathsf{F}_\mathsf{u}, \mathsf{F}_\mathsf{l}, \mathsf{F}_\mathsf{long}\}$	$\{\mathsf{T}, \mathsf{O}, \mathsf{H}, \mathsf{F}_\mathsf{u}, \mathsf{F}_\mathsf{l}, \mathsf{F}_\mathsf{long}\}$
8	$\{\mathsf{T}, \mathsf{long}\}$	$\{\mathsf{O}, \mathsf{H}, \mathsf{F}_\mathsf{u}, \mathsf{F}_\mathsf{l}, \mathsf{F}_\mathsf{long}\}$	$\{\mathsf{T}, \mathsf{O}, \mathsf{H}, \mathsf{F}_\mathsf{u}, \mathsf{F}_\mathsf{l}, \mathsf{F}_\mathsf{long}\}$
9	$\{\mathsf{T}, \mathsf{short}\}$	$\{\mathsf{O}, \mathsf{H}, \mathsf{F}_\mathsf{u}\}$	$\{\mathsf{T}, \mathsf{O}, \mathsf{H}, \mathsf{F}_\mathsf{u}\}$
10	$\{\mathsf{T}, \mathsf{short}\}$	$\{\mathsf{O}, \mathsf{H}, \mathsf{F}_\mathsf{u}, \mathsf{F}_\mathsf{s}\}$	$\{\mathsf{T}, \mathsf{O}, \mathsf{H}, \mathsf{F}_\mathsf{u}, \mathsf{F}_\mathsf{s}\}$
11	$\{\mathsf{T}, \mathsf{short}\}$	$\{\mathsf{O}, \mathsf{H}, \mathsf{F}_\mathsf{u}, \mathsf{F}_\mathsf{s}\}$	$\{\mathsf{T}, \mathsf{O}, \mathsf{H}, \mathsf{F}_\mathsf{u}, \mathsf{F}_\mathsf{s}\}$

in the original setup the system favors the formation of long filaments, two altered variants were generated in which we produced either short filaments or a combination of short and long filaments as the output. We showed that reaction systems are equally versatile and we built three versions of the model demonstrating the same behaviour as that of the ODE models: only short, only long and both (short and long filaments). Moreover, our final reaction systems model is able to switch its preferred output target between short and long filaments, depending on the trigger coming from the environment.

Building a reaction systems model with a qualitative behaviour "similar" to the quantitative behaviour of an ODE model is in general a difficult problem. Starting from a chemical reaction network, an ODE model can be built using standard kinetic principles such as the law of mass action, whereas building a reaction systems model can be quite intricate, see, e.g. [3]. In the case of the intermediate filaments, building the reaction systems models was straightforward because the chemical reaction network we started from contained no feedback loops.

We built our reaction systems models in two steps: we first built a basic model that did not distinguish between the filaments and then we refined it to a more detailed one that distinguished between unit-length, short, and long filaments. The technique of building models by adding details step by step in called *model refinement* and it is currently being investigated in connection to several different modelling frameworks, e.g., rule-based models [7], ODE models [16], and Petri net models [17]. Exploring in more details the refinement of reaction systems models seems an interesting research topic and we plan to return to it.

Acknowledgements

The authors acknowledge the support of Academy of Finland through projects 267915 and 272559. Charmi Panchal acknowledges the support from Otto A. Malm foundation and from Oskar Öflunds stiftelse foundation.

References

[1] Sepinoud Azimi, Cristian Gratie, Sergiu Ivanov, Luca Manzoni, Ion Petre, and Antonio E. Porreca. Complexity of model checking for reaction systems. Technical Report 1122, Turku Centre for Computer Science, 2014.

[2] Sepinoud Azimi, Cristian Gratie, Sergiu Ivanov, and Ion Petre. Dependency graphs and mass conservation in reaction systems. Technical Report 1123, 2014.

[3] Sepinoud Azimi, Bogdan Iancu, and Ion Petre. Reaction system models for the heat shock response. *Fundamenta Informaticae*, 131(3):299–312, 2014.

[4] Robert Brijder, Andrzej Ehrenfeucht, Michael Main, and Grzegorz Rozenberg. A tour of reaction systems. *International Journal of Foundations of Computer Science*, 22(07):1499–1517, 2011.

[5] Luca Corolli, Carlo Maj, Fabrizio Marini, Daniela Besozzi, and Giancarlo Mauri. An excursion in reaction systems: From computer science to biology. *Theoretical Computer Science*, 454:95–108, 2012.

[6] Eugen Czeizler, Andrzej Mizera, Elena Czeizler, Ralph-Johan Back, John E Eriksson, and Ion Petre. Quantitative analysis of the self-assembly strategies of intermediate filaments from tetrameric vimentin. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 9(3):885–898, 2012.

[7] Vincent Danos, Jerome Feret, Walter Fontana, Russ Harmer, and Jean Krivine. Rule-based modelling and model perturbation. *Transactions on Computational Systems Biology XI*, pages 116–137, 2009.

[8] Andrzej Ehrenfeucht, Michael Main, and Grzegorz Rozenberg. Functions defined by reaction systems. *International Journal of Foundations of Computer Science*, 22(01):167–178, 2011.

[9] Andrzej Ehrenfeucht and Grzegorz Rozenberg. Reaction systems. *Fundamenta Informaticae*, 75(1):263–280, 2007.

[10] Enrico Formenti, Luca Manzoni, and Antonio E. Porreca. Cycles and global attractors of reaction systems. In Helmut Jürgensen, Juhani Karhumäki, and Alexander Okhotin, editors, *Descriptional Complexity of Formal Systems*, volume 8614 of *Lecture Notes in Computer Science*, pages 114–125. Springer, 2014.

[11] Enrico Formenti, Luca Manzoni, and Antonio E. Porreca. Fixed points and attractors of reaction systems. In Arnold Beckmann, Erzsébet Csuhaj-Varjú, and Klaus Meer, editors, *Language, Life, Limits, 10th Conference on Computability in Europe, CiE 2014*, volume 8493 of *Lecture Notes in Computer Science*, pages 194–203. Springer, 2014.

[12] Cristian Gratie and Ion Petre. Fit-preserving data refinement of mass-action reaction networks. In *Language, Life, Limits*, pages 204–213. Springer, 2014.

[13] Harald Herrmann and Ueli Aebi. Intermediate filaments: molecular structure, assembly mechanism, and integration into functionally distinct intracellular scaffolds. *Annual review of biochemistry*, 73(1):749–789, 2004.

[14] Harald Herrmann, Markus Häner, Monika Brettel, Nam-On Ku, and Ueli Aebi. Characterization of distinct early assembly units of different intermediate filament proteins. *Journal of molecular biology*, 286(5):1403–1420, 1999.

[15] Mika Hirvensalo. On probabilistic and quantum reaction systems. *Theoretical Computer Science*, 429:134–143, 2012.

[16] Bogdan Iancu, Elena Czeizler, Eugen Czeizler, and Ion Petre. Quantitative refinement of reaction models. *International Journal of Unconventional Computing*, 8(5-6):529–550, 2012.

[17] Bogdan Iancu, Diana-Elena Gratie, Sepinoud Azimi, and Ion Petre. On the implementation of quantitative model refinement. In *Algorithms for Computational Biology*, pages 95–106. Springer, 2014.

[18] Robert Kirmse, Stephanie Portet, Norbert Mücke, Ueli Aebi, Harald Herrmann, and Jörg Langowski. A quantitative kinetic model for the in vitro assembly of intermediate filaments from tetrameric vimentin. *The Journal of biological chemistry*, 282(25):18563–72, June 2007.

[19] Jetty Kleijn and Maciej Koutny. Membrane systems with qualitative evolution rules. *Fundamenta Informaticae*, 110(1):217–230, 2011.

[20] Elias Lazarides. Intermediate filaments as mechanical integrators of cellular space. *Nature*, 283(5744):249–256, 1980.

[21] Andrzej Mizera, Eugen Czeizler, and Ion Petre. Transactions on computational systems biology xiv. chapter Self-assembly Models of Variable Resolution, pages 181–203. Springer-Verlag, Berlin, Heidelberg, 2012.

[22] Fumiya Okubo, Satoshi Kobayashi, and Takashi Yokomori. Reaction automata. *Theoretical Computer Science*, 429:247–257, 2012.

[23] Gheorghe Păun, Mario J Pérez-Jiménez, and Grzegorz Rozenberg. Bridging membrane and reaction systems–further results and research topics. *Fundamenta Informaticae*, 127(1):99–114, 2013.

[24] Arto Salomaa. Functions and sequences generated by reaction systems. *Theoretical Computer Science*, 466:87–96, 2012.

[25] Arto Salomaa. Functional constructions between reaction systems and propositional logic. *International Journal of Foundations of Computer Science*, 24(1):147–159, 2013.

[26] Manfred Schliwa. The cytoskeleton. cell biology monographs, vol. 13, 1986.

[27] The web interface of the reaction system simulator. http://combio.abo.fi/research/reaction-systems/reaction-system-simulator.

Sepinoud Azimi
Computational Biomodeling Laboratory
Turku Centre for Computer Science
Åbo Akademi University
Joukahaisenkatu 3-5A, 20520 Turku, Finland
E-mail: sazimi@abo.fi

Charmi Panchal
Computational Biomodeling Laboratory
Turku Centre for Computer Science
Åbo Akademi University
Joukahaisenkatu 3-5A, 20520 Turku, Finland
E-mail: cpanchal@abo.fi

Eugen Czeizler
Computational Biomodeling Laboratory
Turku Centre for Computer Science
Åbo Akademi University
Joukahaisenkatu 3-5A, 20520 Turku, Finland
and
Department of Computer Science
Aalto University
PO Box 15400, 00076 Aalto, Finland
E-mail: eczeizle@abo.fi

Ion Petre
Computational Biomodeling Laboratory
Turku Centre for Computer Science
Åbo Akademi University
Joukahaisenkatu 3-5A, 20520 Turku, Finland
E-mail: ipetre@abo.fi

Paper IV

Multi-Stability, Limit Cycles, and Period-Doubling Bifurcation with Reaction Systems.

Sepinoud Azimi, Charmi Panchal, Andrzej Mizera and Ion Petre

Sepinoud Azimi, Charmi Panchal, Andrzej Mizera and Ion Petre. Multi-Stability, Limit Cycles, and Period-Doubling Bifurcation with Reaction Systems. (To appear in International Journal of Foundations of Computer Science), 2018.

International Journal of Foundations of Computer Science
© World Scientific Publishing Company

Multi-Stability, Limit Cycles, and Period-Doubling Bifurcation with Reaction Systems

Sepinoud Azimi

Computational Biomodeling Laboratory
Åbo Akademi University and Turku Centre for Computer Science
Turku 20500, Finland
sepinoud.azimi@abo.fi

Charmi Panchal

Computational Biomodeling Laboratory
Åbo Akademi University and Turku Centre for Computer Science
Turku 20500, Finland
charmi.panchal@abo.fi

Andrzej Mizera

Université du Luxembourg
L-1359, Luxembourg
andrzej.mizera@uni.lu

Ion Petre

Computational Biomodeling Laboratory
Åbo Akademi University and Turku Centre for Computer Science
Turku 20500, Finland
ion.petre@abo.fi

Quantitative models may exhibit sophisticated behaviour that includes having multiple steady states, bistability, limit cycles, and period-doubling bifurcation. Such behaviour is typically driven by the numerical dynamics of the model, where the values of various numerical parameters play the crucial role. We introduce in this paper natural correspondents of these concepts to reaction systems modelling, a framework based on elementary set theoretical, forbidding/enforcing-based mechanisms. We construct several reaction systems models exhibiting these properties.

Keywords: Qualitative models; bistability; limit cycle; period-doubling bifurcation; steady state; reaction systems.

1. Introduction

During the recent years a shift in the focus of molecular biology is observed. It is progressively moving from the determination of novel cellular components (e.g., transcription factors, genes, receptors) and the recognition of their individual functions, to the comprehension of how ensembles of cellular components operate in a concerted manner in order to receive, transmit, and process various stimuli into system-level, complex physiological responses, see, e.g., $[23, 9, 29]$. The molecular machinery that underlies the regulation of complex cellular phenomena such as proliferation, differentiation, and apoptosis, is being progressively uncovered and characterised, see, e.g., $[24, 26, 35, 11]$. As our knowledge about the components and modules necessary for proper functioning of a cell is constantly growing, the resulting biological models become increasingly complex. As a consequence, they become difficult or even impossible to intuit. Sketched-out drawings, flow charts, and other forms of static diagrams used sometimes by biologists become insufficient to identify and analyse system-level functionalities and their characteristics. They are undergoing a transformation from purely static representation of biological knowledge into dynamical computational models, which can provide insights into the functioning of the systems. Analytical and predictive power of computational modelling and formal reasoning becomes more and more essential for our understanding of biology, in particular the comprehension of how compositions of cellular components lead to various, common types of emergent behaviour.

The analysis and understanding of these behaviours are typically performed in the realm of quantitative models, in particular models based on ordinary differential equations. In such models, the characteristics of a system are usually generated through a quantitative interplay between the well-chosen numerical values of the kinetic constants and of the initial concentrations of the variables. This is to some extent unsatisfactory, being governed by numerical setups that say nothing about the structure (nature) of the system under study. Therefore, our insight into the causes of and the mechanisms driving these behaviours remains on a basic level of detail. We address the problem on an elementary level here, by adopting *reaction systems* as our modeling framework and we continue in this paper the line of research initiated in $[6, 5, 4]$ to introduce to the framework of reaction systems the formalization of several notions of central interest in biomodeling such as mass-conservation, steady state, periodicity, elementary fluxes, invariants, stationary process, multi-stability, bifurcation. Our definitions of these notions for reaction systems aim to be a natural correspondent of their usual definitions in quantitative frameworks, see, e.g. $[25]$. This line of research provides biomodellers with a set of basic modelling tools and concepts which in turn facilitates building and analysing a biomodel with reaction systems.

In this article we introduce the definitions of multi-stability, mono-stability and periodic reaction systems as natural correspondents of their counterpart concepts in dynamical systems. In connection to them, we also discuss the concepts of bistability, limit cycles, and bifurcations. Several studies have employed other qualitative frameworks, e.g., based on Boolean logic, to demonstrate the above mentioned behaviours, see for example $[3, 19, 27]$. They are introduced here for the first time to reaction systems using only elementary set-theoretical operations. Our underlying idea is that a dynamical system's trajectories correspond to its counterpart reaction system's interactive processes: the initial conditions of a trajectory are mirrored into the initial state of the corresponding interactive process and its numerical setup is represented through a constant context sequence. For a trajectory incorporating changes in its initial numerical setup, we will consider as a counterpart an interactive process with a non-constant context sequence (such as is the case with period-doubling behaviour).

Bistability is an example of a system-level characteristic property recurring in the description of various cellular systems ([2, 32]). Bistable systems are ones that toggle between two alternative stable states. They are considered to impose switch-like biochemical behaviour. There exists a number of reviews in the literature that present theoretical and experimental advances that cast light on what is needed for a biological system to exhibit bistability, e.g., [22, 34, 32].

Limit cycle oscillation is another system level behaviour of interest for this study. The usefulness of limit cycles in describing periodic biological and ecological phenomena (such as the Lotka-Volterra system [20]) makes them a compelling subject to study. In a dynamical system with limit cycle, all stable periodic trajectories are attracted to a unique unstable steady state [16].

Period-doubling bifurcation is another interesting system level behaviour which is well connected to chaotic behaviour in nature. This mode of deterministic chaos is a common pattern in living organisms, see [17]. In a system with a period-doubling bifurcation, a slight change in the system’s parameters, makes an initially stable cycle of length k unstable, and produces a new stable cycle of length $2k$, see [30].

The article is structured as follows. In Section 2 we introduce a few basic definitions of reaction systems. In Section 3- 5 we introduce the notions of *multi-stable*, *mono-stable* and *periodic* reaction systems and provide some examples having these properties. We conclude with a brief discussion in Section 6.

2. Reaction systems

Reaction systems, introduced in [13], is a qualitative framework inspired by the functioning of the living cells. There are only two main regulation mechanisms, *facilitation* and *inhibition*, in reaction systems, that drive the interactions between reactions. Intuitively a reaction is enabled when all components needed to facilitate the reaction are present and all components which inhibit such a facilitation are absent from the environment. Based on this intuition a reaction is formalised as a triplet: its reactants, its inhibitors, and its product set.

In the world of reaction systems, reactions are the pivotal ingredients and it is reactions that lead the transformation of the system from one state to the other. This modelling approach provides a causal insight for the modeller and facilitates a better understanding of the cause-effect relationships of the reactions and consequently of the model as a whole. In contrast, in traditional modelling approaches one mainly deals with the outcome of a process and not with the process itself. Another point that makes reaction systems an interesting tool for modelling is its qualitative nature and how it deals with the phenomena under study only through the facilitation and inhibition mechanisms. There are two main assumptions considered in the reaction systems framework:

- *The threshold assumption.* It is assumed that either an element is present in the environment in abundance or it is absent from it. This implies that there is no counting in (the basic formulation of) the reaction systems framework and as a result, reaction systems are qualitative, rather than quantitative;

- *The no permanency assumption.* It is assumed that an element vanishes from the environment if no reaction is triggered to preserve it. This follows the basic energetics of the living cells, where all the different components are actively supported through energy, i.e., through various cellular reactions.

The two main assumptions of the reaction systems framework yield a very different modelling framework than in traditional ODE-based modelling. For example, concurrency on resources between different reactions is described in reaction systems through facilitators and inhibitors, rather than through a competition driven by the numerical values of kinetic constants as in ODE-based model. This provides a deeper and more explicit understanding of the phenomenon under study. We refer to [6] and [7] for two biological models implemented in reaction systems including a comparison with the corresponding ODE-based models.

We recall some basic definitions of reaction systems. For details we refer to [13].

A *reaction* is a triplet of non-empty, finite sets: $a = (R_a, I_a, P_a)$, where $R_a \cap I_a = \emptyset$. The sets R_a, I_a, P_a stand for the set of *reactants*, *inhibitors*, *products* of a, respectively. Given a set S, if R_a, I_a, $P_a \subseteq S$, then a is a reaction in S. The set of reactions in S is denoted by $\mathit{rac}(S)$.

Let A be a finite set of reactions, T a finite set, and $a \in A$.

(i) The *result* of a on T, denoted $res_a(T)$, is

$$res_a(T) = \begin{cases} P_a, & \text{if } R_a \subseteq T \text{ and } I_a \cap T = \emptyset \\ \emptyset, & \text{otherwise.} \end{cases}$$

(ii) The *result of* A *on* T, denoted $res_A(T)$, is

$$res_A(T) = \bigcup_{a \in A} res_a(T).$$

A *reaction system* (RS in short) is defined as an ordered pair $\mathcal{A} = (S, A)$, where S is a finite set and $A \subseteq \mathit{rac}(S)$. The set S is called the *background* (set) of $\mathcal{A}$.

Let $\mathcal{A}$ be an RS. An *interactive process* in $\mathcal{A}$ is a pair $\pi = (\gamma, \delta)$, where $\gamma = C_0, C_1, ..., C_n, \delta = D_0, D_1, D_2, ..., D_n$, $n \geq 1$, $C_i, D_i \subseteq S$, for all $1 \leq i \leq n$, with

- $D_0 = \emptyset$ and,
- $D_i = \mathit{res}_{\mathcal{A}}(C_{i-1} \cup D_{i-1})$, for all $1 \leq i \leq n$.

The sequence γ is the *context sequence* of $\pi = (\gamma, \delta)$. The *state sequence* of π is $\tau = W_0, W_1, ..., W_n$, where $W_i = C_i \cup D_i$, for all $0 \leq i \leq n$. We say that the state sequence τ is *generated* by the context sequence γ in $\mathcal{A}$. The *initial state* of π is $W_0 = C_0$ and its *final state* is W_n. We say γ is a *constant context sequence* over $C \subseteq S$ if $C_1 = C_2 = \ldots = C_n = C$. (Note that we allow the first context set C_0 to be different than the other sets in the sequence so that we can have interactive processes with a constant context sequence start from different initial states.) We say γ is *an empty context sequence* if $C_1 = C_2 = \ldots = C_n = \emptyset$.

We say that a non-empty set $D \subset S$ is a *steady state* of $\mathcal{A}$ for context set C if $res_{\mathcal{A}}(C \cup D) = D$, see [6] and [4]. Note that the background set cannot be a steady state: $res_{\mathcal{A}}(S) = \emptyset$, since each reaction has a non-empty inhibitor set.

Note that the interactive processes of a reaction system can also be represented through a path in the state transition diagram, where the nodes are the subsets of S and the transitions are $U \to V$, labeled by C, with $U, V, C \subseteq S$, where $res_{\mathcal{A}}(U \cup C) = V$. In this context we may say that an interactive process *leads* to a state (or to a cycle) if the corresponding path in the state transition diagram reaches that state (or that cycle).

3. Multi-stable Reaction Systems

In this section we introduce the notion of *multi-stable reaction systems* and discuss in some details the explicit construction of a model with bistability.

Definition 1. *We say that* $\mathcal{A} = (S, A)$ *is a* multi-stable reaction system *for context* $C \subset S$ *if there exist distinct* $W_1, W_2 \subset S$ *such that* W_1 *and* W_2 *are steady states for context* C *in* $\mathcal{A}$.

It is easy to see that there exists an RS with no steady state. Such an example consists of the following two reactions: $(\{a\}, \{b\}, \{b\})$, $(\{b\}, \{a\}, \{a\})$.

The following result proves the existence of reaction systems with any fixed number of steady states. This is a slight adaptation of some results of [8].

Lemma 2. *For any finite set* S *of cardinality* $n \geq 2$ *and any* $0 \leq k \leq 2^n - 2$*, there exists an RS* $\mathcal{A} = (S, A)$ *with exactly* k *steady states.*

Proof. Let $S = \{s_1, \ldots, s_n\}$ and $0 \leq k \leq 2^n$.

It is easy to see that the RS with S as the background set and the following set of reactions has no steady state: $(\{s_i\}, \{s_{i+1}\}, \{s_{i+1}\})$, for all $1 \leq i \leq n-1$, $(\{s_n\}, \{s_1\}, \{s_1\})$.

Consider now the case $k \geq 1$. We construct an RS having k steady states for the empty context. We denote $t_1, t_2, \ldots, t_{2^n-2}$ the subsets of S (in some fixed but arbitrary ordering), other than the empty set and the full set; the latter two are denoted t_{2^n-1} and t_{2^n}, respectively.

We construct a state transition diagram T with nodes $\{t_1, \ldots, t_{2^n}\}$ as follows. We add self loops on nodes $t_1, \ldots, t_k$ and edges from all other nodes to node t_1. We label all the edges with empty sets. It is shown in [8] that any finite state transition diagram F can be translated to an RS $\mathcal{F}$ in such a way that the transition sequences of F are in a one-to-one correspondence with a certain kind of interactive processes of $\mathcal{F}$. Following the construction in [8] we consider the following reactions:

- $(t_i, S \setminus t_i, t_i)$, for all $1 \leq i \leq k$;
- $(t_j, S \setminus t_j, t_1)$, for all $k < j \leq 2^n - 2$.

Clearly, this RS has exactly k steady states for the empty context: $t_1, \ldots, t_k$. □

Bistable systems have the capacity to operate in two distinct modes in a stable manner. Typically, the system can switch from one stable mode to the other in response to a specific external input. Mathematically, these bistable systems are usually described by models that exhibit two distinct stable steady states [10]. Bistability is a recurrent motif in biology; for instance, in the well known lac operon in the bacteria Escherichia coli, a group of genes are repressed in the presence of glucose and transcribed in the combined absence of glucose and presence of lactose [21, 31].

The smallest chemical reaction system with bistability is presented in [37]. It consists of the minimal number of reactants, reactions, and terms in the associated system of ordinary differential equations (ODEs). Its set of chemical reactions are in Table 1 and an illustration of its behaviour is in Fig. 1. As it can be seen, for lower levels of the input signal S, the system has only one *base-level* steady state $x = 0$. As the level of S increases, the system undergoes a saddle-node bifurcation, which renders the system bistable. This behaviour is observable in Fig. 1 at $S = 4$: beyond that point the system has one more stable steady state in addition to $x = 0$ (as well as another unstable steady state). The process can be reversed: for a high level of the signal strength, the system is bistable. As the signal decreases and reaches the lower saddle-node bifurcation point, the drastic jump to the lower steady-state will occur.

In the remaining of this section we are constructing a counterpart in the reaction systems framework of the example in [37]. The existence of an RS with bistability follows already from Lemma 2. Our example has the additional benefit of being explicitly constructed as a correspondent of a quantitative reaction-based model, rather than as a direct encoding of a desired behaviour. This example will also be used as the basis of another model, to be built in the next section, with the mono-stability property.

Table 1: The chemical reaction bistable model of [37] and the corresponding RS.

Reactions in chemical reaction netweork	Reactions in bistable RS
$S + y \xrightarrow{k_1} 2x$	$(\{S, y\}, \{\mathrm{d_I}\}, \{x\})$
$2x \xrightarrow{k_2} x + y$	$(\{x\}, \{\mathrm{d_I}\}, \{y\})$
$x + y \xrightarrow{k_3} y + P$	$(\{x, y\}, \{S\}, \{y\})$
$x \xrightarrow{k_4} P$	No correspondent reaction in RS

In the background set we introduce variables x, y corresponding to the variables with the same names of the ODE-based model. We also introduce variables s, S to distinguish between low and high external signal, allowing for capturing the bistability switch. We are ignoring variable P in our RS model. Additionally, we use (as usual in RS modeling) a dummy inhibitor variable $\mathrm{d_I}$.

The first reaction of the model in [37], $S + y \rightarrow 2x$, is translated into the RS reaction

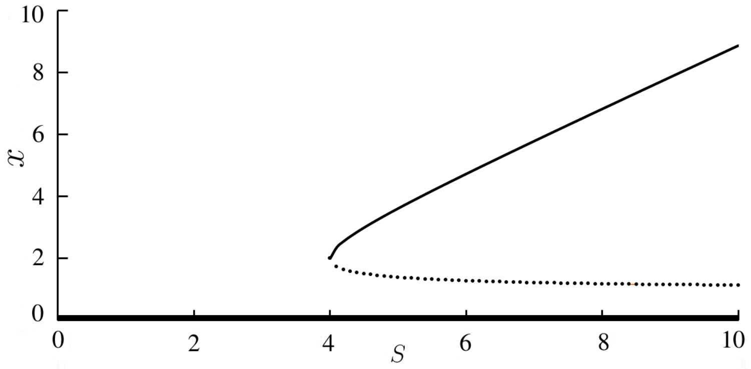


Fig. 1: Bistability. The figure shows two stable steady states (solid lines) and one unstable state (dotted line).

$(\{S, y\}, \{d_I\}, \{x\})$.

For the second reaction, $2x \to x + y$, we do not include x in the product set to take into account the multiplicities of x on the left- and on the right-hand side of the reaction. We translate this reaction to $(\{x\}, \{d_I\}, \{y\})$.

The third reaction, $x + y \to y + P$, essentially implements the degradation of x (when P is ignored) in the presence of x and y. Because of the threshold assumption in RS, degrading x can only be observed in our RS model if the first reaction of the RS model (that produces x) is not enabled at the same time. With this intuition in mind, we introduce S as an inhibitor of the third reaction of our RS model: $(\{x, y\}, \{S\}, \{y\})$.

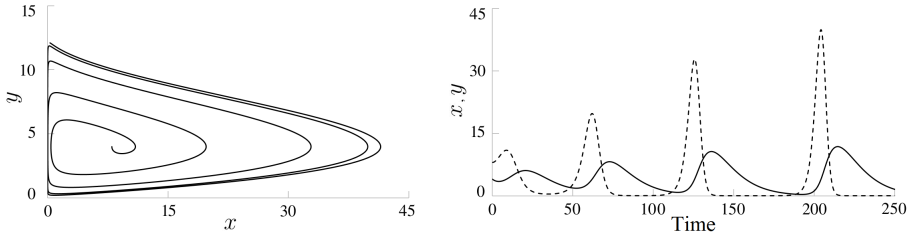
The resulting RS reactions are presented in Table 1.

The behaviour of the RS model corresponds well to that of the chemical reaction model. For example, it is clear from the set of reactions that an interactive process with a constant context sequence over $\{s\}$ takes the system from any state to state $\{s\}$, showing that the model has in this case only one steady state, similarly as the chemical reaction model has with low levels of the external signal. Also, an interactive process with a constant context sequence over $\{S\}$ leads to either state $\{S\}$ or to state $\{x, y, S\}$, showing that the model has in this case two stable steady states, again similar as the chemical reaction model. Furthermore, note that the system cycles between states $\{x, S\}$ and $\{y, S\}$ with a constant context sequence over $\{S\}$; this can be interpreted as corresponding to the third unstable steady state of the chemical reaction model under high levels of the external signal.

4. Mono-stability and Limit Cycles for Reaction Systems

In this section we introduce the notions of *mono-stability* and *limit cycle* for reaction systems, and construct an example of an RS with limit cycle as a slight modification of the example in the previous section.

Being able to explain the oscillatory phenomena in biological systems makes the limit



(a)

(b)

Fig. 2: Limit cycle behaviour: (a) closed loop; (b) periodic oscillations.

cycle one of the most interesting kinetic behaviours [12]. A cycle with length greater than one in the phase space is called a *limit cycle* (a cycle of length one is a steady state). It is known that in a dynamical system with limit cycle behaviour, there is a *unique* (unstable) steady state surrounded by stable periodic trajectories that converge to the steady state [16]. Limit cycles have proven to be useful in describing periodic processes in nature, e.g. the Lotka-Volterra system [20]. That is why finding such trajectories is an interesting subject to study [19]. Fig. 2 shows the typical behaviour of systems with limit cycle.

We introduce now mono-stability and limit cycle for reaction systems.

Definition 3. *We say that $\mathcal{A} = (S, A)$ is a* mono-stable *reaction system for context $C \subset S$ if there exists a* unique *$W \subset S$ such that W is a steady state for context C in $\mathcal{A}$.*

Note that for such an RS, an interactive process with an initial state other than W and with a (long-enough) constant context sequence over C eventually leads to either the empty state, to W itself, or to a cycle of length greater than one. The special case where there is only one such cycle leads to the concept of *limit cycle*. We formulate a definition for reaction systems with limit cycle as follows.

Definition 4. *We say that $\mathcal{A} = (S, A)$ is a reaction system with* limit cycle *for context $C \subset S$ if:*

- *there exists only one steady state W for context C in $\mathcal{A}$;*
- *there exist $W_1, \ldots, W_n \subset S$, $n > 1$, such that $\mathrm{res}_{\mathcal{A}}(C \cup W_i) = W_{i+1}$ for every $1 \leq i < n$ and $\mathrm{res}_{\mathcal{A}}(C \cup W_n) = W_1$, and*
- *any interactive process with a constant context sequence over C and with an initial state other than W will eventually reach a state from $\{W_1, \ldots, W_n\}$.*

It is known that a small modification of the numerical parameters of a dynamical system may make it switch from a bistable behaviour to one with a limit cycle, see [14]. Following this suggestion we build an RS with limit cycle through a small modification of the RS of Table 1. The only modification we make to it is to add an inhibitor to its first reaction; the result is presented in Table 2. Indeed, this RS has a limit cycle for context $\{S\}$: the interactive processes with a constant context sequence over $\{S\}$, the model may either reach state $\{S\}$ (a steady state for context $\{S\}$), or it may eventually cycle between states $\{x, S\}$ and $\{y, S\}$.

Table 2: An RS model with limit cycle for context $\{S\}$ over the background set $\{s, S, x, y\}$.

List of reactions
$(\{S, y\}, \{x\}, \{x\})$
$(\{x\}, \{\mathrm{d_I}\}, \{y\}$
$(\{x, y\}, \{S\}, \{x\})$

5. Periodic Reaction Systems

In this section we introduce the notion of *periodic reaction system* and provide an example of a dynamical system with such behaviour.

Definition 5. *We say that $\mathcal{A} = (S, A)$ is a* periodic *reaction system for context $C \subset S$ if for every $W \subset S$, there is a constant context sequence $\gamma = W, C, \ldots, C$ such that the interactive process $\pi = (\gamma, \delta)$ leads to cycles of length greater than one.*

An interesting addition to the periodic behaviour is the *period-doubling bifurcation*. A period-doubling is a bifurcation in which a modification of a parameter value causes the system to switch to a new behaviour where the period of the system is twice as large as the original one. A period-doubling cascade is a sequence of doublings of the period. For details on period-doubling bifurcation we refer to [30, 28]. Understanding the period-doubling behaviour is of utmost important since it facilitates the better explaining, and possibly controlling, the chaotic phenomena occurring in nature, see for example [15] and [36].

Fig. 3 illustrates a period-doubling bifurcation for the discrete dynamical system $x_{n+1} = r - x_n^2$ where x_0 and r belong to the intervals $[-2, 2]$ and $[0, 2]$ respectively. As it can be seen from Fig. 3, the period of the system doubles as r increases. This behaviour is indicative of the onset of chaos, see [18, 33, 1].

The cascade of period-doubling can be viewed as a binary counter with adjustable length, i.e., for every $1 \leq i \leq n$, the period i is of length 2^i and each state of the period i is labeled with a binary number between 0 and $2^i - 1$ as depicted in Fig. 4. We use this intuition in building an RS with period-doubling behaviour. In this model the change from period i to period j of the system is induced by having j introduced into the system by the context. As the foundation of our model we use the binary counter RS model introduced in [13]. We add a few new reactions to the system to control the length of the counter as well as to facilitate the transition from one period to the other. The model is constructed as follows.

Our RS model will be $\mathcal{A} = (S, A)$ with $S = \{e_0, \ldots, e_n, t, 1, \ldots, n\}$, where e_0 denotes the start of the counting, e_i represents 1 on the ith binary position for all $1 \leq i \leq n$, t is

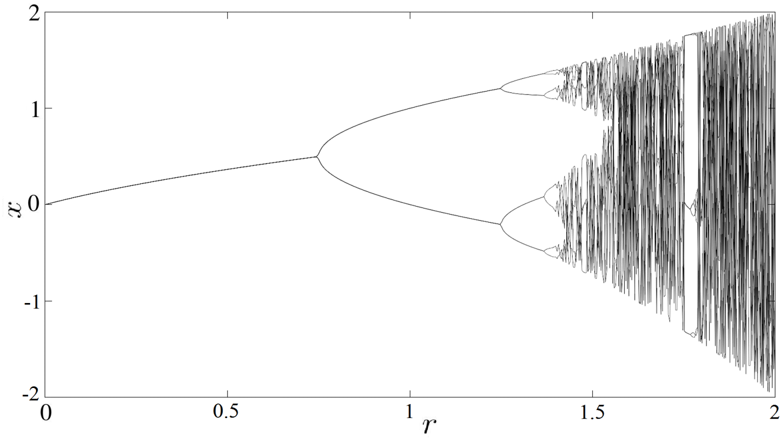


Fig. 3: The discrete dynamical system $x_{n+1} = r - x_n^2$ where $x_0 = 0$ exhibits a period-doubling bifurcation. The plot shows the attractors of this dynamical system for different values of r.

the trigger for ending the counting process, and $1 \leq k \leq n$ presents the current counting threshold, that indicates $2^k - 1$ as the upper bound for the current counter. The set of reactions A is defined as:

- $a_{10} = (\{e_1\}, \{e_0, t\}, \{e_1\})$,
- $a_{ij} = (\{e_i\}, \{e_j, t, 1, \ldots, i-1\}, \{e_i\})$, for all i, j such that $1 \leq j < i \leq n$,
- $b_1 = (\{e_0\}, \{e_1, t\}, \{e_1\})$,
- $b_i = (\{e_0, \ldots, e_{i-1}\}, \{e_i, 1, \ldots, i-1, t\}, \{e_i\})$, for all i such that $2 \leq i \leq n$,
- $r_1 = (\{e_0\}, \{t\}, \{e_0\})$,
- $r_2 = (\{e_0, t\}, \{e_1, \ldots, e_n\}, \{e_0\})$,
- $l = (\{t\}, \{e_0\}, \{e_0\})$,
- $q_i = (\{e_0, \ldots, e_i, i\}, \{t\}, \{t\})$, for all i such that $1 \leq i \leq n$,
- $s_i = (\{i\}, \{t, i+1, \ldots, n\}, \{i\})$, for all i such that $1 \leq i \leq n$.

Reactions $a_{10}, a_{ij}, b_1, b_i, r_1, l$ and q_i for all i, j such that $1 \leq j < i \leq n$, are adopted from the RS of [13] with small modifications respecting the newly introduced counting upper bound in our study. This set of reactions is responsible for the binary counting as well as ending the process whenever trigger $\{t\}$ is introduced in the system.

Reaction a_{10} guarantees that if e_0 is not present, then the incrementing process is not performed, while it takes place otherwise.

For all $\{i, j\}$ such that $1 \leq j < i \leq n$, the reaction a_{ij} produces e_i as long as e_j and ending trigger t are not present and the counter threshold is greater than i. Thus, if a binary number has 1 on the ith position and 0 on some jth position, $j < i$, its successor still has 1 on the ith position.

For each $2 \leq i \leq n$, the reaction b_i produces e_i if e_i is not present while all of

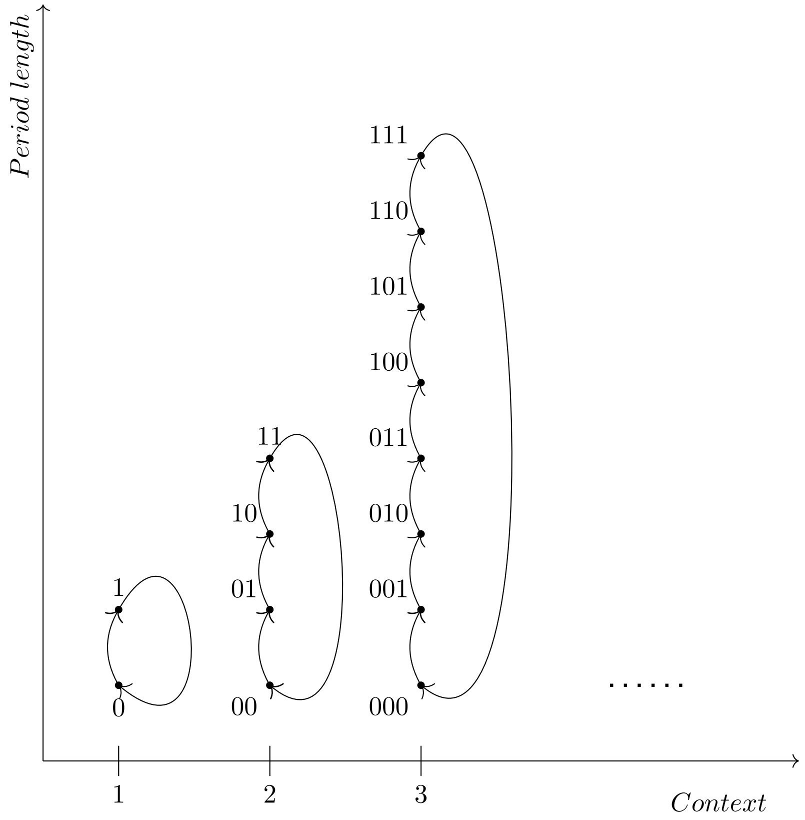


Fig. 4: Period-doubling cascade illustrated as a binary counter.

$e_0, \ldots, e_{i-1}$ are present. Thus, b_i inserts 1 to the ith position when 1 is added to a number that has 0 on position i and 1 on each position smaller than i, the ending trigger is not available and the counter threshold is greater than i.

Reaction l starts the counting by adding e_0 to the current state when the trigger t is introduced.

Reactions q_i, for all $1 \leq i \leq n$, stop the counter when the binary number has reached its threshold.

Reaction r_1 keeps e_0 in the system as long as t is not added to the system. Reaction r_2 resets the system.

Reactions s_i, for all $1 \leq i \leq n$, are responsible for preserving the current period length and for switching from one period length to another. Note that for switching from one period to the other, the binary length i of the new period needs to be introduced into the system through the context.

To better illustrate the behaviour of this RS, we provide an example here. The state sequence corresponding to the initial state $\{e_0, 3\}$ with an empty context sequence is:

$\{e_0, 3\}$, $\{e_0, e_1, 3\}$, $\{3, e_0, e_2\}$, $\{3, e_0, e_1, e_2\}$, $\{3, e_0, e_3\}$, $\{3, e_0, e_1, e_3\}$, $\{3, e_0, e_2, e_3\}$, $\{3, e_0, e_1, e_2, e_3\}$, $\{3, e_0, t\}$, which, based on our defined notions, translates to the following binary sequence: $000, 001, 010, 011, 100, 101, 110, 111, 000$. This sequence represents the period of length 2^3 in our period-doubling RS. Note that any other initial state $\{e_0, k\}$ would enter a cycle with period of length 2^k.

6. Discussion

We continued in this paper the line of research initiated in $[6, 5, 4]$ to bring to the framework of reaction systems natural correspondents of various quantitative modelling concepts such as mass-conservation, steady state, periodicity, elementary fluxes, invariants, stationary processes, multi-stability, bifurcation. The aim of this line of research is to provide a biomodeller with a set of basic modelling tools and concepts to serve her in building and analysing a biomodel with reaction systems. There are clear advantages in using reaction systems as a modelling framework alongside traditional (both quantitative and qualitative) modelling frameworks; to mention only two: the transparent causality between events taking place in a system, and the explicit formulation of the mechanisms responsible for triggering a reaction, in terms of facilitation and inhibition.

To demonstrate that the definitions we introduced in this paper are natural with respect to the similar concepts in dynamical systems, we built several RS models as natural correspondents of known dynamical systems and showed that their status with respect to multi-stability or limit cycle is preserved.

Acknowledgments

This work was partially funded by the Academy of Finland (project 267915), the Otto A. Malm foundation and the Oskar Öflunds stiftelse foundation.

References

[1] Eyad H. Abed, Helen Wang, and RC Chen. Stabilization of period doubling bifurcation and implications for control of chaos. In *Decision and Control, Proceedings of the 31st IEEE Conference on*, pages 2119–2124. IEEE, 1992.

[2] David Angeli, James E. Ferrell Jr, and Eduardo D. Sontag. Detection of multistability, bifurcations, and hysteresis in a large class of biological positive-feedback systems. *PNAS*, 101(7):1822—1827, 2004.

[3] Julio Aracena, Luis Gómez, and Lilian Salinas. Limit cycles and update digraphs in boolean networks. *Discrete Applied Mathematics*, 161(1):1–12, 2013.

[4] Sepinoud Azimi, Cristian Gratie, Sergiu Ivanov, Luca Manzoni, Ion Petre, and Antonio E Porreca. Complexity of model checking for reaction systems. *Theoretical Computer Science*, 623:103 – 113, 2016.

[5] Sepinoud Azimi, Cristian Gratie, Sergiu Ivanov, and Ion Petre. Dependency graphs and mass conservation in reaction systems. *Theoretical Computer Science*, 598:23 – 39, 2015.

[6] Sepinoud Azimi, Bogdan Iancu, and Ion Petre. Reaction system models for the heat shock response. *Fundamenta Informaticae*, 131(3-4):299–312, 2014.

[7] Sepinoud Azimi, Charmi Panchal, Eugen Czeizler, and Ion Petre. Reaction systems models for the self-assembly of intermediate filaments. *Ann. Univ. Buchar*, 62:9 – 24, 2015.

[8] Robert Brijder, Andrzej Ehrenfeucht, Michael Main, and Grzegorz Rozenberg. A tour of reaction systems. *International Journal of Foundations of Computer Science*, 22(07):1499–1517, 2011.

[9] Mariajose Castellanos, David B Wilson, and Michael L Shuler. A modular minimal cell model: purine and pyrimidine transport and metabolism. *Proceedings of the National Academy of Sciences of the United States of America*, 101(17):6681–6686, 2004.

[10] Madalena Chaves, Thomas Eissing, and Frank Allgower. Bistable biological systems: A characterization through local compact input-to-state stability. *IEEE Transactions on Automatic Control*, 53(Special Issue):87–100, 2008.

[11] Sara Checa and Patrick J Prendergast. A mechanobiological model for tissue differentiation that includes angiogenesis: a lattice-based modeling approach. *Annals of biomedical engineering*, 37(1):129–145, 2009.

[12] Attila Császár, Laszlo Jicsinszky, and Tamás Turányi. Generation of model reactions leading to limit cycle behavior. *Reaction Kinetics and Catalysis Letters*, 18(1-2):65–71, 1982.

[13] Andrzej Ehrenfeucht and Grzegorz Rozenberg. Reaction systems. *Fundamenta Informaticae*, 75(1):263–280, 2007.

[14] Alexandra Erbach, Frithjof Lutscher, and Gunog Seo. Bistability and limit cycles in generalist predator–prey dynamics. *Ecological Complexity*, 14:48–55, 2013.

[15] Alan Garfinkel. Controlling cardiac chaos. *Science*, 1992.

[16] Hannah Gay. *The Silwood Circle: A history of ecology and the making of scientific careers in late twentieth-century Britain*. World Scientific, 2013.

[17] Torben Geest, Curtis G Steinmetz, Raima Larter, and Lars F Olsen. Period-doubling bifurcations and chaos in an enzyme reaction. *The Journal of Physical Chemistry*, 96(14):5678–5680, 1992.

[18] Michael R Guevara and Leon Glass. Phase locking, period doubling bifurcations and chaos in a mathematical model of a periodically driven oscillator: a theory for the entrainment of biological oscillators and the generation of cardiac dysrhythmias. *Journal of mathematical biology*, 14(1):1–23, 1982.

[19] Franziska Hinkelmann and Reinhard Laubenbacher. Boolean models of bistable biological systems. *arXiv preprint arXiv:0912.2089*, 2009.

[20] Josef Hofbauer and Josef So. Multiple limit cycles for three dimensional lotka-volterra equations. *Applied Mathematics Letters*, 7(6):65–70, 1994.

[21] François Jacob and Jacques Monod. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of molecular biology*, 3(3):318–356, 1961.

[22] James E Ferrell Jr. Self-perpetuating states in signal transduction: positive feedback, double-negative feedback and bistability. *Current Opinion in Cell Biology*, 14(2):140–148, 2002.

[23] Jonathan R Karr, Jayodita C Sanghvi, Derek N Macklin, Miriam V Gutschow, Jared M Jacobs, Benjamin Bolival, Nacyra Assad-Garcia, John I Glass, and Markus W Covert. A whole-cell computational model predicts phenotype from genotype. *Cell*, 150(2):389–401, 2012.

[24] Daniel J Kelly and Patrick Prendergast. Mechano-regulation of stem cell differentiation and tissue regeneration in osteochondral defects. *Journal of biomechanics*, 38(7):1413–1422, 2005.

[25] Edda Klipp, Ralf Herwig, Axel Kowald, Christoph Wierling, and Hans Lehrach. *Systems biology in practice: concepts, implementation and application*. John Wiley & Sons, 2008.

[26] Pamela K Kreeger and Douglas A Lauffenburger. Cancer systems biology: a network modeling perspective. *Carcinogenesis*, 31(1):2–8, 2010.

[27] Chris J Kuhlman, Henning S Mortveit, David Murrugarra, and VS Kumar. Bifurcations in boolean networks. *arXiv preprint arXiv:1108.2974*, 2011.

[28] Yuri A Kuznetsov. *Elements of applied bifurcation theory*, volume 112. Springer Science & Business Media, 2013.

[29] Jong Min Lee, Erwin P Gianchandani, James A Eddy, and Jason A Papin. Dynamic analysis of

integrated signaling, metabolic, and regulatory networks. *PLoS Comput Biol*, 4(5):e1000086, 2008.

[30] Robert M May. Simple mathematical models with very complicated dynamics. *Nature*, 261(5560):459–467, 1976.

[31] Ertugrul M Ozbudak, Mukund Thattai, Han N Lim, Boris I Shraiman, and Alexander Van Oudenaarden. Multistability in the lactose utilization network of escherichia coli. *Nature*, 427(6976):737–740, 2004.

[32] Joseph R Pomerening. Uncovering mechanisms of bistability in biological systems. *Current Opinion in Biotechnology*, 19:381—388, 2008.

[33] Thomas Simpson, Jia-Ming Liu, Athanasios Gavrielides, Vassilios Kovanis, and Paul Alsing. Period-doubling cascades and chaos in a semiconductor laser with optical injection. *Physical review A*, 51(5):4181, 1995.

[34] John J Tyson, Katherine C Chen, and Bela Novak. Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. *Current Opinion in Cell Biology*, 15(2):221–231, 2003.

[35] Dawn Walker, Grant Hill, Steven Wood, Rod Smallwood, and Jenny Southgate. Agent-based computational modeling of wounded epithelial cell monolayers. *IEEE transactions on nanobioscience*, 3(3):153–163, 2004.

[36] Xin Wang. Period-doublings to chaos in a simple neural network: An analytical proof. *Complex Systems*, 5(4):425–44, 1991.

[37] Thomas Wilhelm. The smallest chemical reaction system with bistability. *BMC Systems Biology*, 3(1):1, 2009.

Turku Centre for Computer Science
TUCS Dissertations

1. **Marjo Lipponen**, On Primitive Solutions of the Post Correspondence Problem
2. **Timo Käkölä**, Dual Information Systems in Hyperknowledge Organizations
3. **Ville Leppänen**, Studies on the Realization of PRAM
4. **Cunsheng Ding**, Cryptographic Counter Generators
5. **Sami Viitanen**, Some New Global Optimization Algorithms
6. **Tapio Salakoski**, Representative Classification of Protein Structures
7. **Thomas Långbacka**, An Interactive Environment Supporting the Development of Formally Correct Programs
8. **Thomas Finne**, A Decision Support System for Improving Information Security
9. **Valeria Mihalache**, Cooperation, Communication, Control. Investigations on Grammar Systems.
10. **Marina Waldén**, Formal Reasoning About Distributed Algorithms
11. **Tero Laihonen**, Estimates on the Covering Radius When the Dual Distance is Known
12. **Lucian Ilie**, Decision Problems on Orders of Words
13. **Jukkapekka Hekanaho**, An Evolutionary Approach to Concept Learning
14. **Jouni Järvinen**, Knowledge Representation and Rough Sets
15. **Tomi Pasanen**, In-Place Algorithms for Sorting Problems
16. **Mika Johnsson**, Operational and Tactical Level Optimization in Printed Circuit Board Assembly
17. **Mats Aspnäs**, Multiprocessor Architecture and Programming: The Hathi-2 System
18. **Anna Mikhajlova**, Ensuring Correctness of Object and Component Systems
19. **Vesa Torvinen**, Construction and Evaluation of the Labour Game Method
20. **Jorma Boberg**, Cluster Analysis. A Mathematical Approach with Applications to Protein Structures
21. **Leonid Mikhajlov**, Software Reuse Mechanisms and Techniques: Safety Versus Flexibility
22. **Timo Kaukoranta**, Iterative and Hierarchical Methods for Codebook Generation in Vector Quantization
23. **Gábor Magyar**, On Solution Approaches for Some Industrially Motivated Combinatorial Optimization Problems
24. **Linas Laibinis**, Mechanised Formal Reasoning About Modular Programs
25. **Shuhua Liu**, Improving Executive Support in Strategic Scanning with Software Agent Systems
26. **Jaakko Järvi**, New Techniques in Generic Programming – C++ is more Intentional than Intended
27. **Jan-Christian Lehtinen**, Reproducing Kernel Splines in the Analysis of Medical Data
28. **Martin Büchi**, Safe Language Mechanisms for Modularization and Concurrency
29. **Elena Troubitsyna**, Stepwise Development of Dependable Systems
30. **Janne Näppi**, Computer-Assisted Diagnosis of Breast Calcifications
31. **Jianming Liang**, Dynamic Chest Images Analysis
32. **Tiberiu Seceleanu**, Systematic Design of Synchronous Digital Circuits
33. **Tero Aittokallio**, Characterization and Modelling of the Cardiorespiratory System in Sleep-Disordered Breathing
34. **Ivan Porres**, Modeling and Analyzing Software Behavior in UML
35. **Mauno Rönkkö**, Stepwise Development of Hybrid Systems
36. **Jouni Smed**, Production Planning in Printed Circuit Board Assembly
37. **Vesa Halava**, The Post Correspondence Problem for Market Morphisms
38. **Ion Petre**, Commutation Problems on Sets of Words and Formal Power Series
39. **Vladimir Kvassov**, Information Technology and the Productivity of Managerial Work
40. **Frank Tétard**, Managers, Fragmentation of Working Time, and Information Systems

41. **Jan Manuch**, Defect Theorems and Infinite Words
42. **Kalle Ranto**, Z_4-Goethals Codes, Decoding and Designs
43. **Arto Lepistö**, On Relations Between Local and Global Periodicity
44. **Mika Hirvensalo**, Studies on Boolean Functions Related to Quantum Computing
45. **Pentti Virtanen**, Measuring and Improving Component-Based Software Development
46. **Adekunle Okunoye**, Knowledge Management and Global Diversity – A Framework to Support Organisations in Developing Countries
47. **Antonina Kloptchenko**, Text Mining Based on the Prototype Matching Method
48. **Juha Kivijärvi**, Optimization Methods for Clustering
49. **Rimvydas Rukšėnas**, Formal Development of Concurrent Components
50. **Dirk Nowotka**, Periodicity and Unbordered Factors of Words
51. **Attila Gyenesei**, Discovering Frequent Fuzzy Patterns in Relations of Quantitative Attributes
52. **Petteri Kaitovaara**, Packaging of IT Services – Conceptual and Empirical Studies
53. **Petri Rosendahl**, Niho Type Cross-Correlation Functions and Related Equations
54. **Péter Majlender**, A Normative Approach to Possibility Theory and Soft Decision Support
55. **Seppo Virtanen**, A Framework for Rapid Design and Evaluation of Protocol Processors
56. **Tomas Eklund**, The Self-Organizing Map in Financial Benchmarking
57. **Mikael Collan**, Giga-Investments: Modelling the Valuation of Very Large Industrial Real Investments
58. **Dag Björklund**, A Kernel Language for Unified Code Synthesis
59. **Shengnan Han**, Understanding User Adoption of Mobile Technology: Focusing on Physicians in Finland
60. **Irina Georgescu**, Rational Choice and Revealed Preference: A Fuzzy Approach
61. **Ping Yan**, Limit Cycles for Generalized Liénard-Type and Lotka-Volterra Systems
62. **Joonas Lehtinen**, Coding of Wavelet-Transformed Images
63. **Tommi Meskanen,** On the NTRU Cryptosystem
64. **Saeed Salehi**, Varieties of Tree Languages
65. **Jukka Arvo**, Efficient Algorithms for Hardware-Accelerated Shadow Computation
66. **Mika Hirvikorpi**, On the Tactical Level Production Planning in Flexible Manufacturing Systems
67. **Adrian Costea**, Computational Intelligence Methods for Quantitative Data Mining
68. **Cristina Seceleanu**, A Methodology for Constructing Correct Reactive Systems
69. **Luigia Petre**, Modeling with Action Systems
70. **Lu Yan**, Systematic Design of Ubiquitous Systems
71. **Mehran Gomari**, On the Generalization Ability of Bayesian Neural Networks
72. **Ville Harkke**, Knowledge Freedom for Medical Professionals – An Evaluation Study of a Mobile Information System for Physicians in Finland
73. **Marius Cosmin Codrea**, Pattern Analysis of Chlorophyll Fluorescence Signals
74. **Aiying Rong**, Cogeneration Planning Under the Deregulated Power Market and Emissions Trading Scheme
75. **Chihab BenMoussa**, Supporting the Sales Force through Mobile Information and Communication Technologies: Focusing on the Pharmaceutical Sales Force
76. **Jussi Salmi**, Improving Data Analysis in Proteomics
77. **Orieta Celiku**, Mechanized Reasoning for Dually-Nondeterministic and Probabilistic Programs
78. **Kaj-Mikael Björk**, Supply Chain Efficiency with Some Forest Industry Improvements
79. **Viorel Preoteasa**, Program Variables – The Core of Mechanical Reasoning about Imperative Programs
80. **Jonne Poikonen**, Absolute Value Extraction and Order Statistic Filtering for a Mixed-Mode Array Image Processor
81. **Luka Milovanov**, Agile Software Development in an Academic Environment
82. **Francisco Augusto Alcaraz Garcia**, Real Options, Default Risk and Soft Applications
83. **Kai K. Kimppa**, Problems with the Justification of Intellectual Property Rights in Relation to Software and Other Digitally Distributable Media
84. **Dragoş Truşcan**, Model Driven Development of Programmable Architectures
85. **Eugen Czeizler**, The Inverse Neighborhood Problem and Applications of Welch Sets in Automata Theory

86. **Sanna Ranto**, Identifying and Locating-Dominating Codes in Binary Hamming Spaces
87. **Tuomas Hakkarainen**, On the Computation of the Class Numbers of Real Abelian Fields
88. **Elena Czeizler**, Intricacies of Word Equations
89. **Marcus Alanen**, A Metamodeling Framework for Software Engineering
90. **Filip Ginter**, Towards Information Extraction in the Biomedical Domain: Methods and Resources
91. **Jarkko Paavola**, Signature Ensembles and Receiver Structures for Oversaturated Synchronous DS-CDMA Systems
92. **Arho Virkki**, The Human Respiratory System: Modelling, Analysis and Control
93. **Olli Luoma**, Efficient Methods for Storing and Querying XML Data with Relational Databases
94. **Dubravka Ilić**, Formal Reasoning about Dependability in Model-Driven Development
95. **Kim Solin**, Abstract Algebra of Program Refinement
96. **Tomi Westerlund**, Time Aware Modelling and Analysis of Systems-on-Chip
97. **Kalle Saari**, On the Frequency and Periodicity of Infinite Words
98. **Tomi Kärki**, Similarity Relations on Words: Relational Codes and Periods
99. **Markus M. Mäkelä**, Essays on Software Product Development: A Strategic Management Viewpoint
100. **Roope Vehkalahti**, Class Field Theoretic Methods in the Design of Lattice Signal Constellations
101. **Anne-Maria Ernvall-Hytönen**, On Short Exponential Sums Involving Fourier Coefficients of Holomorphic Cusp Forms
102. **Chang Li**, Parallelism and Complexity in Gene Assembly
103. **Tapio Pahikkala**, New Kernel Functions and Learning Methods for Text and Data Mining
104. **Denis Shestakov**, Search Interfaces on the Web: Querying and Characterizing
105. **Sampo Pyysalo**, A Dependency Parsing Approach to Biomedical Text Mining
106. **Anna Sell**, Mobile Digital Calendars in Knowledge Work
107. **Dorina Marghescu**, Evaluating Multidimensional Visualization Techniques in Data Mining Tasks
108. **Tero Säntti**, A Co-Processor Approach for Efficient Java Execution in Embedded Systems
109. **Kari Salonen**, Setup Optimization in High-Mix Surface Mount PCB Assembly
110. **Pontus Boström**, Formal Design and Verification of Systems Using Domain-Specific Languages
111. **Camilla J. Hollanti**, Order-Theoretic Mehtods for Space-Time Coding: Symmetric and Asymmetric Designs
112. **Heidi Himmanen**, On Transmission System Design for Wireless Broadcasting
113. **Sébastien Lafond**, Simulation of Embedded Systems for Energy Consumption Estimation
114. **Evgeni Tsivtsivadze**, Learning Preferences with Kernel-Based Methods
115. **Petri Salmela**, On Commutation and Conjugacy of Rational Languages and the Fixed Point Method
116. **Siamak Taati**, Conservation Laws in Cellular Automata
117. **Vladimir Rogojin**, Gene Assembly in Stichotrichous Ciliates: Elementary Operations, Parallelism and Computation
118. **Alexey Dudkov**, Chip and Signature Interleaving in DS CDMA Systems
119. **Janne Savela**, Role of Selected Spectral Attributes in the Perception of Synthetic Vowels
120. **Kristian Nybom**, Low-Density Parity-Check Codes for Wireless Datacast Networks
121. **Johanna Tuominen**, Formal Power Analysis of Systems-on-Chip
122. **Teijo Lehtonen**, On Fault Tolerance Methods for Networks-on-Chip
123. **Eeva Suvitie**, On Inner Products Involving Holomorphic Cusp Forms and Maass Forms
124. **Linda Mannila**, Teaching Mathematics and Programming – New Approaches with Empirical Evaluation
125. **Hanna Suominen**, Machine Learning and Clinical Text: Supporting Health Information Flow
126. **Tuomo Saarni**, Segmental Durations of Speech
127. **Johannes Eriksson**, Tool-Supported Invariant-Based Programming

128. **Tero Jokela**, Design and Analysis of Forward Error Control Coding and Signaling for Guaranteeing QoS in Wireless Broadcast Systems
129. **Ville Lukkarila**, On Undecidable Dynamical Properties of Reversible One-Dimensional Cellular Automata
130. **Qaisar Ahmad Malik**, Combining Model-Based Testing and Stepwise Formal Development
131. **Mikko-Jussi Laakso**, Promoting Programming Learning: Engagement, Automatic Assessment with Immediate Feedback in Visualizations
132. **Riikka Vuokko**, A Practice Perspective on Organizational Implementation of Information Technology
133. **Jeanette Heidenberg**, Towards Increased Productivity and Quality in Software Development Using Agile, Lean and Collaborative Approaches
134. **Yong Liu**, Solving the Puzzle of Mobile Learning Adoption
135. **Stina Ojala**, Towards an Integrative Information Society: Studies on Individuality in Speech and Sign
136. **Matteo Brunelli**, Some Advances in Mathematical Models for Preference Relations
137. **Ville Junnila**, On Identifying and Locating-Dominating Codes
138. **Andrzej 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.
139. **Csaba Ráduly-Baka**, Algorithmic Solutions for Combinatorial Problems in Resource Management of Manufacturing Environments
140. **Jari Kyngäs**, Solving Challenging Real-World Scheduling Problems
141. **Arho Suominen**, Notes on Emerging Technologies
142. **József Mezei**, A Quantitative View on Fuzzy Numbers
143. **Marta Olszewska**, On the Impact of Rigorous Approaches on the Quality of Development
144. **Antti Airola**, Kernel-Based Ranking: Methods for Learning and Performace Estimation
145. **Aleksi Saarela**, Word Equations and Related Topics: Independence, Decidability and Characterizations
146. **Lasse Bergroth**, Kahden merkkijonon pisimmän yhteisen alijonon ongelma ja sen ratkaiseminen
147. **Thomas Canhao Xu**, Hardware/Software Co-Design for Multicore Architectures
148. **Tuomas Mäkilä**, Software Development Process Modeling – Developers Perspective to Contemporary Modeling Techniques
149. **Shahrokh Nikou**, Opening the Black-Box of IT Artifacts: Looking into Mobile Service Characteristics and Individual Perception
150. **Alessandro Buoni**, Fraud Detection in the Banking Sector: A Multi-Agent Approach
151. **Mats Neovius**, Trustworthy Context Dependency in Ubiquitous Systems
152. **Fredrik Degerlund**, Scheduling of Guarded Command Based Models
153. **Amir-Mohammad Rahmani-Sane**, Exploration and Design of Power-Efficient Networked Many-Core Systems
154. **Ville Rantala**, On Dynamic Monitoring Methods for Networks-on-Chip
155. **Mikko Pelto**, On Identifying and Locating-Dominating Codes in the Infinite King Grid
156. **Anton Tarasyuk**, Formal Development and Quantitative Verification of Dependable Systems
157. **Muhammad Mohsin Saleemi**, Towards Combining Interactive Mobile TV and Smart Spaces: Architectures, Tools and Application Development
158. **Tommi J. M. Lehtinen**, Numbers and Languages
159. **Peter Sarlin**, Mapping Financial Stability
160. **Alexander Wei Yin**, On Energy Efficient Computing Platforms
161. **Mikołaj Olszewski**, Scaling Up Stepwise Feature Introduction to Construction of Large Software Systems
162. **Maryam Kamali**, Reusable Formal Architectures for Networked Systems
163. **Zhiyuan Yao**, Visual Customer Segmentation and Behavior Analysis – A SOM-Based Approach
164. **Timo Jolivet**, Combinatorics of Pisot Substitutions
165. **Rajeev Kumar Kanth**, Analysis and Life Cycle Assessment of Printed Antennas for Sustainable Wireless Systems
166. **Khalid Latif**, Design Space Exploration for MPSoC Architectures

167. **Bo Yang**, Towards Optimal Application Mapping for Energy-Efficient Many-Core Platforms
168. **Ali Hanzala Khan**, Consistency of UML Based Designs Using Ontology Reasoners
169. **Sonja Leskinen**, m-Equine: IS Support for the Horse Industry
170. **Fareed Ahmed Jokhio**, Video Transcoding in a Distributed Cloud Computing Environment
171. **Moazzam Fareed Niazi**, A Model-Based Development and Verification Framework for Distributed System-on-Chip Architecture
172. **Mari Huova**, Combinatorics on Words: New Aspects on Avoidability, Defect Effect, Equations and Palindromes
173. **Ville Timonen**, Scalable Algorithms for Height Field Illumination
174. **Henri Korvela**, Virtual Communities – A Virtual Treasure Trove for End-User Developers
175. **Kameswar Rao Vaddina**, Thermal-Aware Networked Many-Core Systems
176. **Janne Lahtiranta**, New and Emerging Challenges of the ICT-Mediated Health and Well-Being Services
177. **Irum Rauf**, Design and Validation of Stateful Composite RESTful Web Services
178. **Jari Björne**, Biomedical Event Extraction with Machine Learning
179. **Katri Haverinen**, Natural Language Processing Resources for Finnish: Corpus Development in the General and Clinical Domains
180. **Ville Salo**, Subshifts with Simple Cellular Automata
181. **Johan Ersfolk**, Scheduling Dynamic Dataflow Graphs
182. **Hongyan Liu**, On Advancing Business Intelligence in the Electricity Retail Market
183. **Adnan Ashraf**, Cost-Efficient Virtual Machine Management: Provisioning, Admission Control, and Consolidation
184. **Muhammad Nazrul Islam**, Design and Evaluation of Web Interface Signs to Improve Web Usability: A Semiotic Framework
185. **Johannes Tuikkala**, Algorithmic Techniques in Gene Expression Processing: From Imputation to Visualization
186. **Natalia Díaz Rodríguez**, Semantic and Fuzzy Modelling for Human Behaviour Recognition in Smart Spaces. A Case Study on Ambient Assisted Living
187. **Mikko Pänkäälä**, Potential and Challenges of Analog Reconfigurable Computation in Modern and Future CMOS
188. **Sami Hyrynsalmi**, Letters from the War of Ecosystems – An Analysis of Independent Software Vendors in Mobile Application Marketplaces
189. **Seppo Pulkkinen**, Efficient Optimization Algorithms for Nonlinear Data Analysis
190. **Sami Pyöttiälä**, Optimization and Measuring Techniques for Collect-and-Place Machines in Printed Circuit Board Industry
191. **Syed Mohammad Asad Hassan Jafri**, Virtual Runtime Application Partitions for Resource Management in Massively Parallel Architectures
192. **Toni Ernvall**, On Distributed Storage Codes
193. **Yuliya Prokhorova**, Rigorous Development of Safety-Critical Systems
194. **Olli Lahdenoja**, Local Binary Patterns in Focal-Plane Processing – Analysis and Applications
195. **Annika H. Holmbom**, Visual Analytics for Behavioral and Niche Market Segmentation
196. **Sergey Ostroumov**, Agent-Based Management System for Many-Core Platforms: Rigorous Design and Efficient Implementation
197. **Espen Suenson**, How Computer Programmers Work – Understanding Software Development in Practise
198. **Tuomas Poikela**, Readout Architectures for Hybrid Pixel Detector Readout Chips
199. **Bogdan Iancu**, Quantitative Refinement of Reaction-Based Biomodels
200. **Ilkka Törmä**, Structural and Computational Existence Results for Multidimensional Subshifts
201. **Sebastian Okser**, Scalable Feature Selection Applications for Genome-Wide Association Studies of Complex Diseases
202. **Fredrik Abbors**, Model-Based Testing of Software Systems: Functionality and Performance
203. **Inna Pereverzeva**, Formal Development of Resilient Distributed Systems
204. **Mikhail Barash**, Defining Contexts in Context-Free Grammars
205. **Sepinoud Azimi**, Computational Models for and from Biology: Simple Gene Assembly and Reaction Systems
206. **Petter Sandvik**, Formal Modelling for Digital Media Distribution

207. **Jongyun Moon**, Hydrogen Sensor Application of Anodic Titanium Oxide Nanostructures
208. **Simon Holmbacka**, Energy Aware Software for Many-Core Systems
209. **Charalampos Zinoviadis**, Hierarchy and Expansiveness in Two-Dimensional Subshifts of Finite Type
210. **Mika Murtojärvi**, Efficient Algorithms for Coastal Geographic Problems
211. **Sami Mäkelä**, Cohesion Metrics for Improving Software Quality
212. **Eyal Eshet**, Examining Human-Centered Design Practice in the Mobile Apps Era
213. **Jetro Vesti**, Rich Words and Balanced Words
214. **Jarkko Peltomäki**, Privileged Words and Sturmian Words
215. **Fahimeh Farahnakian**, Energy and Performance Management of Virtual Machines: Provisioning, Placement and Consolidation
216. **Diana-Elena Gratie**, Refinement of Biomodels Using Petri Nets
217. **Harri Merisaari**, Algorithmic Analysis Techniques for Molecular Imaging
218. **Stefan Grönroos**, Efficient and Low-Cost Software Defined Radio on Commodity Hardware
219. **Noora Nieminen**, Garbling Schemes and Applications
220. **Ville Taajamaa**, O-CDIO: Engineering Education Framework with Embedded Design Thinking Methods
221. **Johannes Holvitie**, Technical Debt in Software Development – Examining Premises and Overcoming Implementation for Efficient Management
222. **Tewodros Deneke**, Proactive Management of Video Transcoding Services
223. **Kashif Javed**, Model-Driven Development and Verification of Fault Tolerant Systems
224. **Pekka Naula**, Sparse Predictive Modeling – A Cost-Effective Perspective
225. **Antti Hakkala**, On Security and Privacy for Networked Information Society – Observations and Solutions for Security Engineering and Trust Building in Advanced Societal Processes
226. **Anne-Maarit Majanoja**, Selective Outsourcing in Global IT Services – Operational Level Challenges and Opportunities
227. **Samuel Rönnqvist**, Knowledge-Lean Text Mining
228. **Mohammad-Hashem Hahgbayan**, Energy-Efficient and Reliable Computing in Dark Silicon Era
229. **Charmi Panchal**, Qualitative Methods for Modeling Biochemical Systems and Datasets: The Logicome and the Reaction Systems Approaches

TURKU
CENTRE *for*
COMPUTER
SCIENCE

http://www. tucs.fi
tucs@abo.fi

University of Turku
Faculty of Science and Engineering
- Department of Future Technologies
- Department of Mathematics and Statistics

Turku School of Economics
- Institute of Information Systems Science

Åbo Akademi University
Faculty of Science and Engineering
- Computer Engineering
- Computer Science

Faculty of Social Sciences, Business and Economics
- Information Systems

ISBN 978-952-12-3671-6
ISSN 1239-1883

Charmi Panchal

Qualitative Methods for Modeling Biochemical Systems and Datasets