Bogdan Iancu

Quantitative Refinement of Reaction-Based Biomodels

Turku Centre *for* Computer Science

TUCS Dissertations
No 199, June 2015

Quantitative Refinement of Reaction-Based Biomodels

Bogdan Iancu

To be presented, with the permission of the Faculty of Science and Engineering of Åbo Akademi University, for public criticism in Auditorium Gamma on June 25, 2015, at 12 noon.

Åbo Akademi University
Department of Computer Science
Joukahaisenkatu 3-5A, FIN-20520 Turku, Finland

2015

Supervisor

Prof. Ion Petre
Department of Computer Science
Åbo Akademi University
Joukahaisenkatu, 3-5A, FIN-20520 Turku
Finland

Reviewers

Prof. Giancarlo Mauri
University of Milano-Bicocca
Dipartimento di Informatica, Sistemistica e Comunicazione
Edificio U14, viale Sarca 336
20126 Milan
Italy

Dr. Russell Harmer
Centre National de la Recherche Scientifique
Bureau GN1 Sud 355, 46 allée d'Italie
69364 Lyon Cedex 07
France

Opponent

Prof. Giancarlo Mauri
University of Milano-Bicocca
Dipartimento di Informatica, Sistemistica e Comunicazione
Edificio U14, viale Sarca 336
20126 Milan
Italy

ISBN 978-952-12-3230-5
ISSN 1239-1883

Abstract

In the field of molecular biology, scientists adopted for decades a reductionist perspective in their inquiries, being predominantly concerned with the intricate mechanistic details of subcellular regulatory systems. However, integrative thinking was still applied at a smaller scale in molecular biology to understand the underlying processes of cellular behaviour for at least half a century. It was not until the genomic revolution at the end of the previous century that we required model building to account for systemic properties of cellular activity. Our system-level understanding of cellular function is to this day hindered by drastic limitations in our capability of predicting cellular behaviour to reflect system dynamics and system structures. To this end, *systems biology* aims for a system-level understanding of functional intra- and inter-cellular activity.

Modern biology brings about a high volume of data, whose comprehension we cannot even aim for in the absence of computational support. Computational modelling, hence, bridges modern biology to computer science, enabling a number of assets, which prove to be invaluable in the analysis of complex biological systems, such as: a rigorous characterization of the system structure, simulation techniques, perturbations analysis, etc. Computational biomodels augmented in size considerably in the past years, major contributions being made towards the simulation and analysis of large-scale models, starting with signalling pathways and culminating with whole-cell models, tissue-level models, organ models and full-scale patient models. The simulation and analysis of models of such complexity very often requires, in fact, the integration of various sub-models, entwined at different levels of resolution and whose organization spans over several levels of hierarchy.

This thesis revolves around the concept of quantitative model refinement in relation to the process of model building in computational systems biology. The thesis proposes a sound computational framework for the stepwise augmentation of a biomodel. One starts with an abstract, high-level representation of a biological phenomenon, which is materialised into an initial model that is validated against a set of existing data. Consequently, the model is refined to include more details regarding its species and/or reactions. The framework is employed in the development of two models,

one for the heat shock response in eukaryotes and the second for the ErbB signalling pathway. The thesis spans over several formalisms used in computational systems biology, inherently quantitative: reaction-network models, rule-based models and Petri net models, as well as a recent formalism intrinsically qualitative: reaction systems. The choice of modelling formalism is, however, determined by the nature of the question the modeler aims to answer. Quantitative model refinement turns out to be not only essential in the model development cycle, but also beneficial for the compilation of large-scale models, whose development requires the integration of several sub-models across various levels of resolution and underlying formal representations.

Sammanfattning

Inom molekylärbiologi har forskare i årtionden använt sig av ett reduktionistiskt perspektiv i sina undersökningar, med fokus främst på de invecklade mekaniska detaljerna i subcellulära regleringssystem. Redan i ett halvt sekel har dock integrativt tänkande varit i användning i mindre skala för att bättre förstå de processer som ligger bakom cellulärt beteende. Först i och med den genomiska revolutionen vid slutet av det föregående århundradet krävdes att modelleringen skulle beakta cellulär aktivitet som ett system. Vår förståelse av cellulära funktioner på systemnivå är än idag begränsad av vår oförmåga att effektivt förutspå cellulärt beteende i relation till systemdynamik och systemstrukturer. Det man inom systembiologi strävar efter är en förestående av funktionell intra- och intercellulär aktivitet på systemnivå.

Modern biologi framkallar stora mängder information som vi i brist på beräkningsstöd inte har möjlighet att hantera. Beräkningsmodellering förenar därmed modern biologi med datavetenskap. Detta ger tillgång till ett antal fördelar som visar sig vara ovärderliga för analysen av komplexa biologiska system, däribland rigorös karaktärisering av systemstrukturer, simuleringstekniker, störningsanalys, etc. De senaste åren har beräkningsbiologiska modeller växt betydligt och viktiga insatser har gjorts gällande simulering och analys av storskaliga modeller, från signaltransduktionsvägar till modeller på cell-, vävnads- och organnivå, samt fullskaliga patientmodeller. För simulering och analys av så pass invecklade modeller krävs att flera modeller av olika delstrukturer integreras och sammankopplas, i varierande upplösning, och vars organisation sträcker sig över flera nivåer av hierarkin.

Denna avhandling behandlar förbättringen av kvantitativa modeller i förhållande till modelleringsprocesser inom beräknande systembiologi. I denna avhandling föreslås ett grundligt beräknande ramverk för gradvis förbättring av en biologisk modell. Man utgår ifrån en biologisk företeelse, framställd på hög abstraktionsnivå, som formas till en första modellversion och bekräftas med hjälp av existerande data. Modellen blir följaktligen förbättrad och innehåller ett större antal detaljer angående modellens art och/eller reaktioner. Ramverket används för utvecklingen av två sorters modeller; en för eukaryotisk värmechockresponsen och en för ErbB-signaltransduktionsvägen. Avhandlingen inkluderar flera kvantitativa formalis-

mer som används inom beräknande systembiologi: modeller av reaktionsnätverk, petrinäts-modeller, samt en ny kvalitativ formalism: reaktionssystem. Valet av modelleringsformalism beror dock på vilken typ av fråga modelleraren ämnar besvara. Kvantitativ modellförbättring visar sig inte endast vara central under modellens utvecklingsprocess, utan är även till nytta vid sammanställningen av storskaliga modeller, vilka kräver att man integrerar flera modeller av delstrukturer i varierande upplösning och underliggande formella framställningar.

Acknowledgements

As I was growing up, I have always been interested in facts, details, and predictions of events, even without being aware of it. Even now, as an adult, I have an analytical personality. Scientific research came then as a great opportunity to tackle at least partially some aspects of life, which I was trying to understand. Being a computer engineer and always interested in biology and medicine, computational systems biology opened up a new door towards a fascinating domain that I started to enjoy from the very beginning. Consequently, I would like to express my sincere and deep gratitude to my supervisor, Prof. Ion Petre, who has introduced me to this whole new world. I cannot even express by words the excitement of working under his supervision. I am extremely thankful to Prof. Ion Petre for inspiring me every step of the way with his passion for science, with his energy and his devotion for work. I wish to thank him for the tremendous support I have received since joining his research group and even from before, for all the wise advice he had given me and for the patience he had in supervising a young researcher who was still trying to find his way in the scientific community. I am grateful to have him as my supervisor during my doctoral studies, he has always been someone I knew I could rely on and I could trust and the Computational Biomodelling Laboratory became a second family.

I am very thankful to Prof. Giancarlo Mauri who kindly agreed to be a reviewer of my doctoral dissertation and who accepted to act as the opponent at my doctoral defence. I am very grateful to Dr. Russell Harmer for accepting to be a reviewer for my doctoral thesis. I would like to thank both of them for taking the time and making the effort to carefully read my thesis and for their insightful, valuable and helpful comments.

I am very thankful to Dr. Elena Czeizler and Dr. Eugen Czeizler, who have been very close collaborators during my first years of doctoral studies. I am especially thankful to Dr. Elena Czeizler who has been involved in the early supervision of my doctoral studies. I am very grateful to both of them for the encouragement and support they gave me since we first met. I am thankful for all scientific discussions, valuable comments and advice that have been materialised into a couple of scientific publications. Your dedication for science is truly inspiring.

Another person that has had a great impact on my doctoral studies was Dr. Vladimir Rogojin. I would like to express my gratitude for his support, our fruitful collaborations during a few courses I had the opportunity to act as a course assistant for and for all the great times we had playing computer games or talking about life in general and virtual life in particular.

I would like to thank Prof. Adina Magda Florea from the Faculty of Automatic Control and Computer Science, University Politehnica of Bucharest, whose lectures on artificial intelligence have been of great value for me to understand what I wanted to work on. I am also very grateful for her support regarding my visit in the Computational Biomodelling Laboratory to write my diploma thesis.

I am grateful to Dr. Luigia Petre for very interesting and inspiring discussions on action systems, which made me think about my research from a completely new perspective.

I would like to express my gratitude for all my co-authors and people I have collaborated with during my PhD studies: Sepinoud Azimi, Elena Czeizler, Eugen Czeizler, Diana-Elena Gratie, Cristian Gratie, Charmi Panchal, Ion Petre, Luigia Petre, Tolou Shadbahr.

I am grateful for the financial support I have received for my doctoral studies and for my research from Turku Centre for Computer Science, Åbo Akademi, Åbo Akademi Foundation, the Rector of Åbo Akademi, Centre for International Mobility, and Academy of Finland.

As I mentioned above, the Computational Biomodelling Laboratory has become my second family during my doctoral studies. I would like to express again my deep gratitude to Prof. Ion Petre, who put so much effort into creating a balanced and friendly environment, that all lab members have always enjoyed. I am thankful to Dr. Elena Czeizler and Dr. Eugen Czeizler who have been supporting me since I first came to Turku as a visiting student and who have been dear friends to me from the very beginning. I would like to thank Dr. Andrzej Mizera and Dr. Vladimir Rogojin who put a great deal of attention and care during my first year while I was still getting accommodated to a new type of environment, both at work and at home. I am thankful to Sepinoud Azimi and Diana-Elena Gratie, my two *COMBIO*-sisters, I am very grateful for all support you gave me all these years and for all discussions we had, either scientific or regarding the philosophy of life in general. You have been a great positive influence both in my research, but also in my personal life. I am grateful to Dr. Cristian Gratie who has been of great support, ever since we were students in Bucharest; you have never failed to be there for me when I needed and I thank you for that; your scientific curiosity is very exciting, inspiring and also confusing at times. I would also like to thank Charmi Panchal for all the positive thoughts and for the long discussions we had either about biochemical networks or yoga. I would also like to thank Tolou Shadbahr, who has always been so keen on

learning; I have discovered many things about myself while working together and I am grateful for that. I would like to thank Mikhail Barash, Nebiat Ibssa, Krishna Kanhaiya, Dr. Chang Li, Romina Paţurcă, Fatimah Shokri, Dwitiya Tyagi, Muhammad Usman for making these past years so pleasant in the lab.

My life at the department would have not been as pleasant without the joyful discussions with Dr. Maryam Kamali, Dr. Mats Neovius, Dr. Marta Olszewska, Dr. Mikołaj Olszewski, Dr. Viorel Preoteasa, Dr. Dorina Rajanen.

I would like to express my deep gratitude to Åbo Akademi University and Turku Centre for Computer Science for the great infrastructure and working environment I have been provided with. I am very thankful to the administrative staff at both the Department of Computer Science at Åbo Akademi University and Turku Centre for Computer Science for their tremendous help with organizational and administrative matters. I am especially thankful to Christel Engblom, Nina Hultholm, Irmeli Laine, Tove Österroos, Susanne Ramstedt and Tomi Suovuo.

I am thankful to Anna Skult for the translation of the abstract of this thesis into Swedish and to Sanna Laurila and Jenni Vasara for proofreading it.

I would like to express my deep gratitude to many old friends that have been close to me even if miles away during these years: Monica Dobre, Elena Popa, Oana Stanciu, Ana-Maria Stroe, Alexandra Zota. I would like to thank Monica for believing in me, for being of tremendous support for so many years already and for being someone I could always count on. I thank Elena for always being there for me; you are my oldest friend and it's amazing we managed to stay friends living so far apart for so many years. I would like to express my gratitude to Oana and Ana-Maria for their wise advice, for inspiring me to improve every day and for all interesting talks we had over the years. Many thanks go to Alexandra who was always on the *Bogdan-team*. It has meant the world to me to have you all in my life. Moreover, I would very much like to thank all friends I made in Finland during my doctoral studies. I would like to express my gratefulness to Sepinoud Azimi, Luca Bersanini, Matthew Broad, Elena Czeizler, Eugen Czeizler, Erla Elíasdóttir, Diana-Elena Gratie, Maria Ganeva, Daria Gritsenko, Natalia Díaz Rodriguez, Daniela Karlsson, Sanna Laurila, Elina Mehtälä, Jutta Myrttinen, Charmi Panchal, Ion Petre, Luigia Petre, Joanna Pylvänäinen, Dorina Rajanen, Vladimir Rogojin, Sirkku Ruokkeinen, Hannele Svanström, Victor Sifontes, Anna Skult, Soile Tirri, Jenni Vasara, Eveliina Viitaniemi, Emrah Yildir, Turo Ylitalo. I would like to thank Luca for his joyfulness that gets you entertained in most difficult times. I am grateful for the wonderful long tea sessions with Daria and Erla that always end up in some philosophical discussion about the human kind or the universe. I would like

to thank Maria for always being there for me when I needed a friend by my side. I thank Natalia for being such a positive friend and for being the *friendship-hub* of early days in Turku. I want to express my gratitude to Sanna, Anna and Jutta for being such good friends, for making me part of *the Zoo* and for always being there with open arms and doors. I would like to thank Eveliina, Sirkku and Elina for being close and crazy friends. I thank Jenni for being a comrade in arms, a close friend that has always had a shoulder to complain on. All of you have made my spare time incredibly pleasant, good times even better and difficult times easier to bear with, and I am grateful I have you all in my life.

My extreme gratitude goes to my family, without the support of whom writing this thesis would have been impossible. I cannot express how grateful I am to my parents, Nina Iancu and Valeriu Iancu, for being always by my side and on my side, for believing in me and for supporting me in life choices they didn't always agree with completely. Thank you for giving me the freedom to take chances, making mistakes or succeeding, but allowing me to find my own path. I express my deep gratitude to my grandparents, Eugenia Cîrlănaru and Virgil Cîrlănaru, my aunt and uncle, Magdalena Teodoriu and Mihai Teodoriu, and my cousins Lavinia Nucu, Andreea Teodoroiu, Iuliana Teodoroiu and Casiana Teodoroiu. I would like to thank you all for the love, kindness and support I have received over the years.

Bogdan Iancu Turku, May 2015

Contents

List of original publications

We list below the papers included in this thesis as follows:

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

2. Diana-Elena Gratie, Bogdan Iancu and Ion Petre. ODE analysis of biological systems. In: *Formal Methods for Dynamical Systems*, Editors: Marco Bernardo, Erik de Vink, Alessandra di Pierro and Herbert Wiklicky, LNCS 7938, pages 29–62, Springer, 2013.

3. Diana-Elena Gratie, Bogdan Iancu, Sepinoud Azimi and Ion Petre. Quantitative model refinement in four different frameworks, with applications to the heat shock response. In: *From Action System to Distributed Systems*, Taylor & Francis Group, Editors: Luigia Petre, Emil Sekerinski, *Accepted, To appear*, 2015.

4. Bogdan Iancu, Cristian Gratie and Ion Petre. Refinement-based modelling of the ErbB signalling pathway – Extended abstract. In: Annals of University of Bucharest LXI, pages 7–14, The Bucharest University Press, 2014.

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

Chapter 1

Introduction to modelling in computational systems biology

Biology has been governed for decades by reductionism. The reductionist approach has brought about major contributions to biology, enabling exceptional progress towards the understanding of biological phenomena. However, the genomic revolution that characterized the end of the last century has made possible the evolution of traditional molecular biology towards systems biology, see [66].

Biological systems are highly complex. Traditionally, molecular biology has been concerned with the detailed study of individual actors, such as genes, proteins, cells, etc., and the identification of their functions, but a biological function can hardly be associated with just one individual cellular component. Individual scattered information regarding diverse cellular components are insufficient to determine a full understanding of complex biological phenomena, as most of them are the effect of complex interactions between various cellular components, see [5, 59]. An insightful metaphorical presentation of the challenge of understanding complex biological systems behaviour from the isolated study of individual components is shown in [51]; the problem is described there in terms of fixing a transistor radio through studying/characterizing its resistors, capacitors, and transistors.

A system-level understanding of biological systems often focuses on a few central properties such as: system structures or system dynamics. System structures concern the study of gene interactions, protein-protein interactions, biochemical pathways, etc. System dynamics refers to system evolution over time under diverse conditions, which can be pursued through metabolic analysis, sensitivity analysis, bifurcation analysis, etc. The control method refers to the identification of mechanisms that regulate the cell

state evolution and their potential manipulation so as to avoid various cellular impairments induced by potential malfunctions. The design method concerns detecting methods to alter and/or construct biological models with a set of given properties, see [46].

Systems biology aims at an integrative study of biological systems and at an understanding of the interactions among their components. A biological system is regarded in this context as a large number of distinct components, with diverse functionalities, which interact non-linearly with one another and generate complex, but coherent behaviours, see [45, 59]. There are two main directions that systems biology is driven by: the attempt to understand the interactions among cellular components and the functional organization of the system of interest, and the functionality preservation across several levels of organization, at different levels of resolution, see [67].

Modern biology includes research sub-areas such as *theoretical biology*, *mathematical biology*, *computational biology*, *algorithmic biology*, *systems biology*, *synthetic biology*. Several of the prefixes used in naming these branches are imported by biology from computer science, showing the deep impact computer science has on biology. The same can be seen through some of the concepts biologists use in their everyday work: *network*, *component*, *robustness*, *efficiency*, *control*, *regulation*, *synchronization*, *concurrency*, etc.

Computational modelling is one of the proposals that computer science makes to biology (and not only). Computational modelling introduces numerous benefits: an explicit description of the underlying components of a given system, the availability of simulation methods, perturbation analysis (which allows for a close examination of every component of the system). Computational modelling also allows for the exploration of systems which are not feasible to investigate empirically and the construction of systems whose components can be refined, making it very suitable for managing high throughput simulations, see [59].

There are two major approaches to computational modelling: data-driven modelling and hypothesis-driven modelling. In data-driven modelling, the focus is on finding an(a) (computational) explanation/model for a collection of data; the modeller has no a-priori bias/preference for any model and will adopt one based on its ability to explain the existing data. In hypothesis-driven modelling, the focus is on a given mechanism/model for a biological process and on finding a suitable numerical setup for the model that would fit the existing data. Both approaches are commonly used in systems biology, sometimes even within the same project, and both bring a number of benefits. In this thesis, we focused only on hypothesis-driven modelling and on techniques for algorithmically augmenting a given model, while preserving its data fit.

The modelling cycle which lies at the core of systems biology relies on a few steps, starting with the acquisition of high throughput data, developing

an abstract model of the biological system, which undergoes the iterative process of hypothesis generation, experimental design, experimental analysis and model refinement, see [46, 59]. Initially, a biochemical model representing the reaction network corresponding to the underlying biological process is compiled. Subsequently, the associated mathematical model is set up, by generating the dynamics of the system and identifying values for free parameters that fit given experimental data. The mathematical model obtained is then validated against other experimental data or qualitative properties proven empirically. The model is consequently used for novel hypothesis generation, which entails new experimental research. The results obtained as such can be used for the refinement of the model, see [59].

Building a large-scale model, for instance at tissue-level, but even modelling an entire cell, represents in itself an arduous challenge. Large-scale modelling requires taking into account numerous biological characteristics and model predictions have to be validated against biological and/or clinical data, which are frequently insufficient. Simulation of large-scale models, for instance in the case of tissue or organ modelling, frequently compels the integration of numerous models, organized on several levels of hierarchy, across various orders of magnitude regarding the scale or diverse qualitative properties (e.g. biochemical networks, intercellular communications, etc). This line of work has received much attention through the 2013 Nobel prize in Chemistry awarded to Martin Karplus, Michael Levitt and Arieh Warshel "for the development of multi-scale models for complex chemical systems".

Computational biomodels have expanded considerably in the last few years. For example, a model for the *ErbB* signalling pathway was proposed in 2009, see [13], consisting of 499 species and 828 reactions; we consider this pathway as one of the case-studies of this thesis. A whole-cell model was also proposed in 2012 in [43], that included 1900 parameters and was based on the scanning of over 900 publications. The *process* of building large models remains to this day highly non-standard and to some extent ad-hoc. This thesis can be seen as a proposal towards a standardized process of model building. Our proposal is to start from a relatively small model on a high-level of abstraction, seen as a "*blueprint*" of the bioprocess of interest. After validating this initial model against existing data, the model can be augmented in a step-wise manner to include more details about its species and processes. The main technique we focus on is quantitative model refinement, which addresses the problem of numerical setup of the refined model, so that the numerical fit of the original model is preserved.

The structure of this thesis is the following: Chapter 2 focuses on computational biomodelling systems biology and introduces a few modelling aproaches, such as: reaction-based models, rule-based models, Petri nets and reaction systems. In Chapter 3 we discuss quantitative model refinement as a technique for model construction and model augmentation. Chapter

4 introduces two case-studies that have been used for the implementation of quantitative model refinement. In Chapter 5, we list the original contributions of all articles included in this thesis. In Chapter 6, we discuss implications of the thesis and future perspectives.

Chapter 2

Modelling formalisms in systems biology

We discuss in this chapter a few modelling approaches: reaction networks, rule-based models, ODE-based models, Petri nets, and reaction systems. Our selection was based on the approaches we used in the publications included in this thesis. The selection is clearly non-exhaustive and several other popular modelling approaches exist.

2.1 Reaction network models

A reaction-based model can be described through a set of chemical reactions detailing on the main interactions among its species. A reaction can be either *reversible* or *irreversible*. We consider model M to be a pair $M = (\sum, \mathcal{R})$, where $\sum = \{S_1, \ldots, S_m\}$ is the set of species and $\mathcal{R} = \{R_1, R_2, \cdots, R_n\}$ is the set of reactions. Any reversible reaction can be divided into two irreversible reactions accounting separately for its left-to-right and right-to-left directions. Thus, a reaction network comprising reversible reactions can be rewritten as a network involving only irreversible reactions. We will discuss this in more detail later.

An *irreversible* reaction can be presented as a rewriting rule of the following form:

$$R_j \ : \ c_{1,j}S_1 + c_{2,j}S_2 + \ldots + c_{m,j}S_m \xrightarrow{k_j} c'_{1,j}S_1 + c'_{2,j}S_2 + \ldots + c'_{m,j}S_m, \quad (2.1)$$

where $k_j \geq 0$ is a non-negative integer representing the *kinetic rate constant* and $c_{1,j}, \ldots, c_{m,j}, c'_{1,j}, \ldots, c'_{m,j} \geq 0$ are non-negative integers specifying the *stoichiometry* of reaction R_j.

We note that, even though, we list here all species both on the left- and on the right-hand side of the reaction, in practice most of the coefficients

are zero. Indeed, chemical reactants are most often of molecularity at most 2. A model can also include reactions of higher molecularity, describing a complex biochemical process on a higher level of abstraction. The species appearing on the left-hand side of a reaction are called *substrates* and the species occurring on the right-hand side of a reaction are referred to as *products*.

The *stoichiometric coefficient* $n_{i,j}$ corresponding to chemical species S_i of reaction R_j is defined as $n_{i,j} = c'_{i,j} - c_{i,j}$. The stoichiometric coefficients are collected in a *stoichiometric matrix* $\mathbf{N} = (n_{i,j})_{m \times n}$. The (i, j) entry of the matrix represents the stoichiometric coefficient of S_i in reaction R_j. Depending on the sign of the stoichiometric coefficient corresponding to a given species in a reaction, a species is either *consumed* ($n_{i,j} > 0$) or *produced* ($n_{i,j} > 0$).

Given an irreversible reaction R_j, the sum $\sum_{i=1}^{m} c_{i,j}$ represents the *molecularity* of reaction R_j. In this thesis, we only consider reactions with a molecularity of at most two, i.e. *unimolecular* and *bimolecular* reactions. Termolecular reactions are very rare, due to the high improbability of having three molecular entities simultaneously colliding and forming a correct configuration that leads to the constitution of a molecular complex, see [41, 64]. A molecularity greater than three for an elementary reaction is unattainable, since a number of molecules greater than three cannot concomitantly collide, see [55].

A reversible *reaction* $R_j \in \mathcal{R}$ is written as follows:

$$R_j \ : \ c_{1,j}S_1 + c_{2,j}S_2 + \ldots + c_{m,j}S_m \underset{k_j^-}{\overset{k_j^+}{\rightleftarrows}} c'_{1,j}S_1 + c'_{2,j}S_2 + \ldots + c'_{m,j}S_m, \quad (2.2)$$

where $k_j^+, k_j^- \geq 0$ are the *kinetic rate constants* of the reaction and $c_{1,j}, \ldots, c_{m,j}$, $c'_{1,j}, \ldots, c'_{m,j} \geq 0$ are defined as before.

A reversible reaction can be separated into a pair of distinct irreversible reactions as follows:

$$R_j^{(fwd)} \ : \ c_{1,j}S_1 + c_{2,j}S_2 + \ldots + c_{m,j}S_m \xrightarrow{k_j^+} c'_{1,j}S_1 + c'_{2,j}S_2 + \ldots + c'_{m,j}S_m;$$

$$R_j^{(bwd)} \ : \ c'_{1,j}S_1 + c'_{2,j}S_2 + \ldots + c'_{m,j}S_m \xrightarrow{k_j^-} c_{1,j}S_1 + c_{2,j}S_2 + \ldots + c_{m,j}S_m.$$

Example 1 *We will use the following example as a running example throughout the introduction of this thesis. We consider a transcription factor* trans*, the promoter region it binds to,* prom*, and the protein* prot *encoded by the gene promoted by* prom*. We also include in our example the interaction between* prot *and* trans *:* prot *may bind to* trans*, even in the case when* trans *is bound to* prom*; in this case,* trans *will unbind from* prom *and instead, bind*

to prot. *These interactions are described in the reaction network model in Table 2.1.*

Table 2.1: The running example of the introduction to the thesis.

$$\mathsf{trans} + \mathsf{prom} \underset{k_1^-}{\overset{k_1^+}{\longleftrightarrow}} \mathsf{trans{:}\ prom}$$
$$\mathsf{trans{:}\ prom} \xrightarrow{k_2} \mathsf{trans{:}\ prom} + \mathsf{prot}$$
$$\mathsf{trans} + \mathsf{prot} \underset{k_3^-}{\overset{k_3^+}{\longleftrightarrow}} \mathsf{trans{:}\ prot}$$
$$\mathsf{trans{:}\ prom} + \mathsf{prot} \xrightarrow{k_4} \mathsf{trans{:}\ prot} + \mathsf{prom}$$

As it will become clear in Section 4.1, this example is inspired by the heat shock response model, one of the case-studies of this thesis.

The stoichiometric matrix N of this model is shown below. The rows correspond to species trans, prom, trans: prom, prot, trans: prot, in this order. The columns correspond to the reactions of the model, in the order listed in Table 2.1.

$$[N] = \begin{bmatrix} -1 & 0 & -1 & 0 \\ -1 & 0 & 0 & 1 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & -1 \\ 0 & 0 & 1 & 1 \end{bmatrix}.$$

2.1.1 ODE-based models

System dynamics can be described through the time-dependent evolution of all its species concentrations at a given moment in time. ODEs are a common choice for the modelling of deterministic kinetics of reaction-based models. We discuss other modelling formalisms later in this chapter. Given a reaction-based model M depicted as above, the main variables describing the system of ODEs are the species concentrations $\{[S_1], \cdots, [S_m]\}$ and the reactions rates $\{\nu_1, \cdots, \nu_m\}$. We associate a real function $[X] : \mathbb{R}_+ \to \mathbb{R}_+$ to every species X, with the interpretation that $X[t]$ is the concentration at time t.

Species levels can be either expressed in particle numbers (the number of molecules of a chemical species S_i in a volume V) or the number of moles of a chemical species per volume V. We denote the number of molecules of a chemical species S_i by $\#S_i$. The number of moles and molecules of a

species S_i obey the following relation:

$$\#S_i = [S_i] \cdot N_A,$$

where N_A is the Avogadro number, $N_A \approx 6.02214179 \cdot 10^{23}$particles/mol.

2.1.2 The law of mass-action

To associate an ODE-based model to a reaction network, one uses a so-called kinetic rate law. This allows the modeller to associate in a standard way a reaction rate, i.e., a real function, to every reaction in the model. This, in turn, will give the rate of change of every species in the model, i.e., the system of ODEs describing the concentration of the model's species. We discuss in the next subsections the two most widely used kinetic rate laws: mass-action and Michaelis-Menten.

Consider the amount of each chemical species, S_i, given as a *concentration*, $[S_i]$, measured in molars, changing continuously. Each species S_i is associated a function $[S_i] : \mathbb{R}_+ \to \mathbb{R}_+$ expressing the evolution of the chemical species concentration over time. The instantaneous reaction rate is proportional to the probability of the collision between reactants (in general two reactants, see the previous section for details). Assuming homogeneity of the environment and constant temperature, this probability is in turn proportional to the concentration of the colliding reactants raised to the power of their respective molecularity, see [49]. This kinetic law is known as the law of mass-action and was initially introduced by Guldberg and Waage in the 19th century, see [29, 30].

Example 2 *Given a simple reaction of the form below:*

$$S_1 + S_2 \underset{k^-}{\overset{k^+}{\rightleftarrows}} 2P, \tag{2.3}$$

the reaction rate is defined as:

$$v = v_+ - v_- = k^+[S_1][S_2] - k^-[P]^2,$$

where v_+ is the forward reaction rate, v_- is the backward reaction rate, k^+ is the kinetic rate constant corresponding to the forward direction, and k^- is the kinetic rate constant corresponding to the backward direction. Considering now the forward direction of reaction (2.3)*, the molecularity of each substrate is* 1*; considering now the backward direction for the same reaction, the molecularity is* 2*. Assuming the measure unit for time is seconds (s), and for the concentration is molars (M), then the reaction rate's unit is $M \cdot s^{-1}$. In consequence, the measure unit for rate constants for monomolecular reactions (e.g. $S \to \emptyset$) is s^{-1}, while for bimolecular reactions, the measure unit for the rate constant is $M \cdot s^{-1}$.*

Given a reversible reaction R_j of the form (2.2), conforming to the law of mass-action, the reaction rate is given by:

$$v = v_+ - v_- = k_j^+ \prod_{i=1}^{m} [S_i]^{c_{i,j}} - k_j^- \prod_{i=1}^{m} [S_i]^{c'_{i,j}}.$$

The corresponding system of ODEs is given below:

$$\frac{d[S_i]}{dt} = n_{i,j} v = (c'_{i,j} - c_{i,j}) \left(k_j^+ \prod_{l=1}^{m} [S_l]^{c_{l,j}} - k_j^- \prod_{l=1}^{m} [S_l]^{c'_{l,j}} \right), 1 \leq l \leq m.$$

For reversible reactions, the substrate to product ratio at steady state (i.e., when $v_+ = v_-$) is defined as the equilibrium constant, K_{eq}:

$$K_{eq} = \frac{k_j^+}{k_j^-} = \frac{\prod_{i=1}^{m} [S_i]_{eq}^{c'_{i,j}}}{\prod_{i=1}^{m} [S_i]_{eq}^{c_{i,j}}},$$

where $[S_i]_{eq}$ represents the equilibrium concentration of species S_i.

Given a species S, the time course can be calculated by integrating the corresponding ODE. Take, for example, a simple decay reaction: $S \xrightarrow{k}$; the time dynamics of this reaction is given by the ODE: $d[S]/dt = -k[S]$. The integration of this ODE over the time interval [0, t) leads to the analytical solution below:

$$\int_{S_0}^{S} d[S]/dt = -\int_{t=0}^{t} k dt \Rightarrow [S](t) = S_0 e^{-kt},$$

where S_0 is the initial concentration. Calculating the analytical solution for complex systems is hardly possible. However, one can search for numerical methods which would give a satisfactory numerical approximation.

Example 3 *The mass-action rates $v_1 - v_4$ of our running example are the following:*

$$\begin{aligned}
v_1 &= {k_1}^+[\mathsf{trans}][\mathsf{prom}] - {k_1}^-[\mathsf{trans{:}\ prom}];\\
v_2 &= k_2[\mathsf{trans{:}\ prom}];\\
v_3 &= {k_3}^+[\mathsf{trans}][\mathsf{prot}] - {k_3}^-[\mathsf{trans{:}\ prot}];\\
v_4 &= k_4[\mathsf{trans{:}\ prom}][\mathsf{prot}].
\end{aligned}$$

The mass-action ODE model of this example is the following:

$$\begin{aligned}
\frac{d[\mathsf{trans}]}{dt} &= -\,{k_1}^{+}[\mathsf{trans}][\mathsf{prom}] + {k_1}^{-}[\mathsf{trans{:}\ prom}] - {k_3}^{+}[\mathsf{trans}][\mathsf{prot}] \\
&\quad + {k_3}^{-}[\mathsf{trans{:}\ prot}]; \qquad (2.4)\\
\frac{d[\mathsf{prom}]}{dt} &= -\,{k_1}^{+}[\mathsf{trans}][\mathsf{prom}] + {k_1}^{-}[\mathsf{trans{:}\ prom}] \\
&\quad + k_4[\mathsf{trans{:}\ prom}][\mathsf{prot}]; \qquad (2.5)\\
\frac{d[\mathsf{trans{:}\ prom}]}{dt} &= {k_1}^{+}[\mathsf{trans}][\mathsf{prom}] - {k_1}^{-}[\mathsf{trans{:}\ prom}] \\
&\quad - k_4[\mathsf{trans{:}\ prom}][\mathsf{prot}]; \qquad (2.6)\\
\frac{d[\mathsf{prot}]}{dt} &= k_2[\mathsf{trans{:}\ prom}] - {k_3}^{+}[\mathsf{trans}][\mathsf{prot}] + {k_3}^{-}[\mathsf{trans{:}\ prot}] \\
&\quad - k_4[\mathsf{trans{:}\ prom}][\mathsf{prot}]; \qquad (2.7)\\
\frac{d[\mathsf{trans{:}\ prot}]}{dt} &= {k_3}^{+}[\mathsf{trans}][\mathsf{prot}] - {k_3}^{-}[\mathsf{trans{:}\ prot}] \\
&\quad + k_4[\mathsf{trans{:}\ prom}][\mathsf{prot}]. \qquad (2.8)
\end{aligned}$$

2.1.3 Michaelis-Menten kinetics

Enzyme kinetics applies to a certain class of chemical reactions called enzymatic reactions, i.e. reactions catalysed by enzymes. An enzyme reaction in its general form as described in [10] can be written as follows:

$$S + E \underset{k_b}{\overset{k_f}{\rightleftarrows}} S:E \overset{k_{ir}}{\to} P + E, \qquad (2.9)$$

where E denotes the enzyme, S the substrate , $S:E$ the substrate-enzyme complex, and P the product. Reaction (2.9) represents the formation of the substrate-enzyme complex through the reversible binding of enzyme E to substrate S and the irreversible conversion of the substrate-enzyme complex into product P through the release of the enzyme.

The system can be described by the following irreversible reactions:

$$S + E \overset{k_f}{\to} S:E, \quad S:E \overset{k_b}{\to} S + E, \quad S:E \overset{k_{ir}}{\to} P + E. \qquad (2.10)$$

Following mass-action kinetics, the corresponding system of ODEs de-

scribing the dynamics of the system (2.10) can be written as follows:

$$\frac{d[S]}{dt} = -k_f[S][E] + k_b[S:E]; \tag{2.11}$$

$$\frac{d[E]}{dt} = -k_f[S][E] + k_b[S:E] + k_{ir}[S:E]; \tag{2.12}$$

$$\frac{d[S:E]}{dt} = k_f[S][E] - k_b[S:E] - k_{ir}[S:E]; \tag{2.13}$$

$$\frac{d[P]}{dt} = k_{ir}[S:E]. \tag{2.14}$$

The system (2.11) - (2.14) has been subjected to a number of assumptions, for example, the *quasi-equilibrium* between the free enzyme E and the compound $S:E$, see [52]. We assume the kinetic constants k_b, k_f to be much greater than k_{ir} ($k_b, k_f \gg k_{ir}$), i.e. $[S:E]$ is, in fact, negligible compared to $[S]$ and $[P]$, taking into account that the substrate-enzyme complex concentration is reduced.

Moreover, the quasi-equilibrium assumption has been refined (see [10]) to considering the system will eventually come about a *quasi-steady state* of $S:E$, a state for which the concentration of substrate-enzyme complex is constant; the assumption is valid only for $S_0 \gg E_0$. This leads to:

$$d[S:E]/dt = 0, \text{ i.e., } k_f[S][E] - k_b[S:E] - k_{ir}[S:E] = 0. \tag{2.15}$$

The right hand side of (2.12) is the complement of the right hand side of (2.13). Adding them we obtain: $d[E]/dt + d[S:E]/dt = 0$. That implies:

$$[E] + [S:E] = E_{tot}, \text{or equivalently } [E] = E_{tot} - [S:E], \tag{2.16}$$

where constant E_{tot} describes the total enzyme amount.

Recalling the *quasi-steady state* assumption and (2.16), equation (2.15) can be rewritten as follows:

$$\begin{aligned} &k_f[S]E_{tot} = k_f[S][S:E] + k_b[S:E] + k_{ir}[S:E], \text{ i.e.,} \\ &[S:E] = \frac{k_f[S]E_{tot}}{k_f[S] + k_b + k_{ir}}, \text{ i.e.,} \\ &[S:E] = \frac{[S]E_{tot}}{[S] + \frac{k_b + k_{ir}}{k_f}} \end{aligned} \tag{2.17}$$

Bringing (2.17) in (2.14), we obtain the following:

$$\frac{d[P]}{dt} = \frac{k_{ir}[S]E_{tot}}{[S] + \frac{k_b + k_{ir}}{k_f}}. \tag{2.18}$$

The Michaelis-Menten equation states that the reaction rate v of product synthesis is proportional to the substrate concentration, $[S]$, and the

maximum rate the system can reach for saturated values of $[S]$ (V_{max}) and inversely proportional to the sum of the Michaelis constant K_m and the substrate concentration:

$$v = \frac{V_{max}[S]}{[S] + K_m}, \tag{2.19}$$

The Michaelis constant K_m represents the substrate concentration for the half-maximal reaction rate. Introducing the parameters of (2.19) in (2.18), one attains the relation between Michaelis-Menten kinetics and mass-action inferred kinetics of an enzymatic reaction:

$$V_{max} = k_{ir}E_{tot}, \qquad K_m = \frac{k_b + k_{ir}}{k_f}.$$

Assuming *quasi-equilibrium*, then k_{ir}/k_f is negligible, leading to $K_m \cong k_b/k_f$. For a more detailed description of Michaelis-Menten kinetics, we refer the reader to [48].

Reversible Michaelis-Menten kinetics. Consider the extension of the Michaelis-Menten equation to its reversible form, allowing for the catalysation of the reaction by the enzyme in both directions, as follows:

$$S + E \underset{k_2}{\overset{k_1}{\rightleftarrows}} X \underset{k_4}{\overset{k_3}{\rightleftarrows}} P + E, \tag{2.20}$$

where S and P are substrates, E is the enzyme, and X represents the intermediary enzyme-substrate compound. The set of irreversible reactions corresponding to the Michaelis-Menten equation above are:

$$\begin{aligned} &S + E \overset{k_1}{\rightarrow} X;\\ &X \overset{k_2}{\rightarrow} S + E;\\ &X \overset{k_3}{\rightarrow} P + E;\\ &P + E \overset{k_4}{\rightarrow} X. \end{aligned} \tag{2.21}$$

The system of ODEs describing the dynamics of the model, following the principle of mass-action (2.21) is described as follows:

$$\frac{d[S]}{dt} = -k_1[S][E] + k_2[X]; \tag{2.22}$$

$$\frac{d[E]}{dt} = -k_1[S][E] + k_2[X] + k_3[X] - k_4[P][E]; \tag{2.23}$$

$$\frac{d[X]}{dt} = k_1[S][E] - k_2[X] - k_3[X] + k_4[P][E]; \tag{2.24}$$

$$\frac{d[P]}{dt} = k_3[X] - k_4[P][E]. \tag{2.25}$$

Addition of (2.23) and (2.24) results into the following equation:

$$\frac{d[E]}{dt} + \frac{d[X]}{dt} = 0 \Rightarrow [E] + [X] = E_{tot}.$$

The *quasi-steady state* condition, $d[X]/dt = 0$, i.e., $k_1[S](E_{tot} - [X]) - [X](k_2 + k_3) + k_4[P](E_{tot} - [X]) = 0$, leads to the following equation:

$$[X] = \frac{k_1[S]E_{tot} + k_4[P]E_{tot}}{k_1[S] + k_4[P] + k_2 + k_3}. \tag{2.26}$$

Introducing (2.26) into equation (2.25), we obtain:

$$v = \frac{k_1k_3[S]E_{tot} - k_2k_4[P]E_{tot}}{k_1[S] + k_4[P] + k_2 + k_3} = \frac{k_3E_{tot}\frac{k_1[S]}{k_2+k_3} - k_2E_{tot}\frac{k_4[P]}{k_2+k_3}}{1 + \frac{k_1[S]}{k_2+k_3} + \frac{k_4[P]}{k_2+k_3}} =$$

$$\frac{\frac{V_{fw}}{K_{mS}}[S] - \frac{V_{bw}}{K_{mP}}[P]}{1 + \frac{[S]}{K_{mS}} + \frac{[P]}{K_{mP}}},$$

where $K_{mS} = (k_2 + k_3)/k_1$ and $K_{mP} = (k_2 + k_3)/k_4$ are the Michaelis-Menten constants (i.e. for half-maximal forward and backward rate) for the substrate and product, respectively, and $V_{fw}(V_{bw})$ represents the maximal rate in forward (resp. backward) direction. An exact solution for the reversible Michaelis-Menten mechanism is available in [53]. Details about reversible Michaelis-Menten kinetics can be found in [31].

2.1.4 Analysis of reaction-based models

Steady-state analysis

Steady states (*equilibrium points*) are points in the phase plan for which all state variables are constant throughout the system's dynamics, i.e. for reaction-based models, the species concentrations are constant when initial values correspond to steady-state values. A system is at steady state if it maintains its equilibrium unless any external perturbations arise. There are several types of steady-states: *stable* (the system eventually returns to the state), *asymptotically stable* (the system is stable and adjacent initial conditions tend to this state as time approaches infinity), *unstable* (the system leaves the state) etc.

The behaviour of a reaction-based model at steady state is given by the system of equations below:

$$\frac{d[S]}{dt} = 0, \tag{2.27}$$

where $[S] = ([S_1], [S_2], ..., [S_m])^T$ represents the vector of concentrations of species. Given $v = (v_1, v_2, ..., v_n)^T$, the vector of reaction fluxes, the system of equations (2.27) becomes:

$$Nv = 0. \tag{2.28}$$

Vector v can be determined solving the corresponding system of equations with variables $[S_1], [S_2], \ldots, [S_m]$. The solutions of the system are non-trivial (not all variables 0) if $\text{rank}(N) < n$. We recall n is the number of reactions in the system. But this means matrix M has at least two linearly dependent columns (if $\text{rank}(N) < n$). This dependency can be reflected through a so-called *kernel matrix* K, such that:

$$NK = 0, \tag{2.29}$$

where K has $c = n - \text{rank}(N)$ columns.

The kernel matrix K is not unique. For instance, one can obtain another matrix K' through a multiplication $K' = KQ$, where Q has dimensions $[n - \text{rank}(N)] \times [n - \text{rank}(N)]$. Given that K is a solution of Equation (2.29), K' is also a solution. For details, we refer the reader to [48].

Mass conservation In biochemical systems, checking whether the total amount of a certain species or association of species is constant over time or not can have significant implications. The total amount of a given species, in this case, refers to the sum over the free species as well as the species that are contained within other compounds. This section introduces briefly the mass conservation relations and their importance in relation to reaction-based systems. A more detailed discussion regarding mass-conservation relations can be found in [34].

Mass conservation relations are very often the first type of analysis one engages, for the reason that they offer a primary perception towards the dynamics of the model, reducing concomitantly the number of free variables of the model. Formally, a mass-conservation relation represents the linear combination of species concentrations which is constant over time:

$$g^T S = C, \tag{2.30}$$

where g is a vector of constant elements, S is the vector of species concentrations, and C is a constant. The existence of mass-conservation relations implies that some rows of the stoichiometric matrix are linearly dependent, i.e.:

$$g^T N = 0^T. \tag{2.31}$$

Equations (2.30) and (2.31) are equivalent. By derivation of the former equation and then allow for $d[S]/dt = Nv$, we obtain:

$$(g^T S)' = g^T \dot{S} = g^T N v = 0.$$

Vector g is not unique, there can be more linearly independent vectors satisfying (2.31), each accounting for a mass-conservation relation. The number of mass-conservation relations is given by $m -$ rank(N), where m is the

number of species in the system. The set of vectors g characterizing the mass-conservation relations are collected in a *conservation matrix* G , see [34], which satisfies the following relation:

$$GN = 0.$$

As a consequence, G^T is the kernel matrix for N^T. A conservation matrix G is not unique and can be determined employing the Gauss algorithm (any alternative matrix $G' = PG$, where P is any non-singular matrix of the appropriate dimensions, constitutes a valid conservation matrix).

Sensitivity analysis *Sensitivity analysis* represents a technique used to determine the extent to which infinitesimal changes of some independent parameters of a model affect the parameters dependent on them. We mention here two types of sensitivity analysis extensively used: *local sensitivity analysis* and *global sensitivity analysis*. To employ local sensitivity analysis, one varies solely one parameter alone within a small interval around a certain value. Global sensitivity analysis, however, requires the variation of all parameters simultaneously, spanning the entire space of each parameter. In this thesis, we only present the former.

Let the system below be the system of ODEs of a reaction-based model, expressed as a function of concentrations of all species and all parameter values:

$$\frac{d[S_i]}{dt} = f_i([S_1], [S_2], ..., [S_m], \kappa), \tag{2.32}$$

where $\kappa = (k_1, k_2, ..., k_n)^T$ is the vector of rate constants (assuming only n irreversible reactions). Consider $\mathcal{S}(t, \kappa) = ([S_1](t, \kappa), [S_2](t, \kappa), ..., [S_m](t, \kappa))^T$ to be the solution of Equation (2.32) relative to κ, named *sensitivity matrix*. The matrix entries are given by the partial derivatives $\partial[S_i]/\partial k_j$.

The simplest approach in calculating the local sensitivity of the concentrations relies on the *finite-difference method*, using the finite difference approximation. Consider the change of the j-th parameter, k_j, with the amount δk_j at time point t_1, given that all other parameters remain unmodified. The newly obtained matrix $[S]$ can be determined making use of the change between the original and the perturbed solution, see Equation (2.33). The method has to be run $n + 1$ times, n times for changing each parameter separately and one for the initial values of the parameters:

$$\frac{\partial[S](t_2)}{\partial k_j(t_1)} = \frac{[S](t2, k_j + \delta k_j) - [S](t_2, k_j)}{\delta k_j}, 1 \leq j \leq n. \tag{2.33}$$

We obtain the solutions for the differential equations for the sensitivity coefficients $\partial[S_i]/\partial k_j$ differentiating Equation (2.32). In this manner, we

obtain the set of sensitivity equations below:

$$\frac{d}{dt}\frac{\partial[S]}{\partial k_j} = \mathcal{J}\frac{\partial[S]}{\delta k_j} + \frac{\partial f}{\partial k_j}, 1 \leq j \leq n,$$

where $\mathcal{J}$ represents the Jacobian for Equation (2.32). For details on the mathematical derivation of the result above, we refer the reader to [68].

Example 4 *To search for the steady-state of our running example we consider the system of non-linear equations obtained by equaling to* 0 *all ODEs in Example 3. We know that there exists a non-trivial solution because the rank of the matrix* N *in Example 1 is less than* 4*. The fourth row of the matrix is linearly dependent on rows* 1 *and* 3 *(in fact, the rank of the matrix is* 3*). Finding the steady state of the model can be done by applying numerical approximation methods to a fully specified numerical setup of the model.*

To find all mass-conservation relations of our example we solve the following system of linear equations: $g^T N = 0$*, i.e.:*

$$\begin{aligned} -g_1 - g_2 + g_3 &= 0; \\ g_4 &= 0; \\ -g_1 - g_4 + g_5 &= 0; \\ g_2 - g_3 - g_4 + g_5 &= 0. \end{aligned}$$

The solution of this system of equations is:

$$\{(g_1, g_2, g_1 + g_2, 0, g_1) | g_1, g_2 \in \mathbb{N}\}. \tag{2.34}$$

The fact that this is a vectorial space of dimension 2 *shows that the model has two mass-conservation relations (this could also be deduced based on the observation that* $rank(N) = 3$*). The solution above gives the following mass-conservation relations (obtained for* $g_1 = 0, g_2 = 1$*, respectively):*

$$\begin{aligned} &[\mathsf{trans}] + [\mathsf{trans{:}\ prom}] + [\mathsf{trans{:}\ prot}] = const.; \\ &[\mathsf{prom}] + [\mathsf{trans{:}\ prom}] = const. \end{aligned}$$

The two relations can be interpreted based on the biological intuition we gave for our running example as the conservation of (the 3 *forms of) the transcription factor and as the conservation of (the* 2 *forms of) the promoter.*

2.2 Rule-based models

Cellular response to environmental signals is highly determined by intricated protein-protein interactions networks, which identify signals from the environment and convert them into appropriate responses. Protein-protein

interactions potentially produce a massive amount of chemical species, which in turn interact extensively with one another.

The combinatorial complexity of these interactions arises from two fundamental reasons: post-translational modifications (acetylation, phosphorylation, ubiquitination, etc), which emerge due to catalytic interactions among proteins, and the formation of heterogeneous protein complexes due to non-covalent protein interactions, see [22, 15]. A species (e.g. a protein) can accomodate multiple modification sites and can interact with multiple binding partners. The amplitude of this combinatorial complexity can be described, for instance, by considering the following scenario. Given a protein P, which can be post-translationally modified at n sites, the number of possible states for P alone is 2^n. Protein P can potentially bind to other proteins on m independent sites, which generates a number of 2^m interactions, corresponding to 2^m possible bound-states. This complexity propagates through the entire network.

Rule-based models appeared as a response to the need of handling major combinatorial aspects in protein-protein interaction networks, signalling pathways, etc. Within a rule-based modelling framework, model specification relies on the identification of molecules of interest, their components (i.e. a post-translational modification site) and possible states of the considered components, see [15].

Rule-based languages rely on rules to describe system dynamics. *Rules* generate chemical reactions by defining classes of reactions characterizing certain types of interactions between molecules or complexes of molecules. A rule defines *group rules*, which describe the interactions between components and/or states of the molecules through regular expressions. Rules specify patterns which identify the reactants among a set of chemical species. Each rule enables transformations from reactants to products through a rate law.

System representation can be assimilated at two different levels: one is an algebraic, text-based representation of the system, while the other is a graph-based representation.

Within a text-based representation of the system, chemical species are indicated through strings. By convention, an individual species, also referred to as *single-state species* is represented through a simple name, for instance, a capital letter. A *molecule* is regarded as a set of given components that constitute an entity (a polypeptide chain, a multimeric protein, etc). A more complex chemical species, called *multi-state species* is characterized through a *molecule string*, which consists of the name of the molecule and an ordered list of indices. A molecule string can account either for a certain species or a group of species containing a certain molecule.

A molecule can have several components. A component is a part of the molecule such as: a post-translational modification, a motif, a catalytic domain, a bio-molecular recognition domain, etc. Some components can be in

various possible states. These states enact diverse conformations or alternate forms of components. The states can be specified through *attributes*, which are assigned specific values. For a detailed discussion about the text-based representation of the system, we refer the reader to [22].

In the past years, various research groups developed languages for the specification of rule-based models, which provide formal frameworks for the representation of biochemical networks that are able to correlate traditional biology and mathematical modelling, such as: Bionetgen ([8, 24]) or Kappa ([18]). We discuss in the following sections a graph-based representation for rule-based models along with a description language for the characterization of rule-based models, Bionetgen.

2.2.1 A graph-based representation of rule-based models

A graph-based formalisation uses graphs and graph-rewriting rules for the representation of biochemical networks. We will adopt here the conventions and definitions in [7] and [23].

A graph, in this context, consists of nodes, which have associated labels, and undirected edges that connect them. Nodes represent *components* (such as binding sites, sites for post-translational modifications or domains of proteins, etc.), that can admit multiple *states* used to depict bonds, conformational modifications, post-translational modifications (e.g., affected by a post-translational modification, unaffected, etc.). Edges are used to represent bonds connecting components.

A *molecular-entity graph* consists in a triple $G = (V_G, E_G, A_G)$, where V_G represents a set of *labelled attributed* nodes (e.g., components such as: sites or domains of proteins, etc.), E_G is a set of unlabelled, undirected edges, which identify bonds between components and A_G is an optional set of attributes. Node labelling is not unique. Identical labels on distinct nodes specify similar components. Edges indicate intramolecular or intermolecular bonding. The molecular entity graph admits a unique label.

A *complex graph* is defined as a set of interconnected molecular-entity graphs.

A *chemical species graph* is defined as a molecular-entity graph with all attributes being assigned specific values, or as a complex graph, whose constituent molecular-entity graphs have all attributes assigned specific values.

A *pattern graph* $P = (V_P, E_P)$ consists in a set of molecular-entity graphs and/or complex graphs, which are not necessarily interconnected. The molecular entity graphs or complex graphs that a given pattern graph P consists of, can each be associated a set of variable attributes. Connectivity of the graphs in P is established through an interface, which partitions the nodes in three categories: nodes representing components which cannot be bound to external components, nodes representing components that must

be bound to other external components and nodes for components that can be bound either to components inside the pattern graph or outside of it.

A pattern graph G contains a chemical species represented by graph H if and only if there exists a subgraph of H isomorphic to G which is consistent with the pattern graph's interface and preserves the attributes of components, molecular entities and complexes, see [7].

Rules are, in this context, graph rewriting rules, which generate reactions identifying classes of chemical transformations of reactants into products. Each rule comprises a set of pattern graphs corresponding to reactants and one corresponding to products. A rate law is associated with the transformation. A rule is applied to a set of chemical species, identifying sets of reactant group graphs and product group graphs. A rule is applied for all combinations of reactants. Rule instantiation consists in the substitution of the subgraphs corresponding to the chemical species matched by pattern graphs defining reactants with the chemical species matched by pattern graphs defining products. For more details regarding the graph-based representation, we refer the reader to [7] and [23].

All reactions generated by a rule comply with the same rate law, though different reactions can be characterized by distinct rate constants, some of which can compel the multiplication by diverse factors, see [22].

2.2.2 A description language for the characterization of rule-based models

Explicit elementary reaction kinetics for large biochemical networks has several limitations such as: the ambiguity of key role elements in molecular interactions, the combinatorial complexity due to post-translational modifications, protein-protein interactions, etc., the massive computational effort invested in the simulation of such networks and the arduous attempts of reusability of the networks, see [62].

Bionetgen is a rule-based modelling framework and a description language for the development and simulation of large-scale biochemical networks, aiming at a compact representation of biochemical information (e.g. interactions of molecular domains, non-covalent protein interactions, etc).

A Bionetgen input file comprises definitions for molecular species, rules, kinetic rate constants, initial concentrations and simulation commands. For example, molecule types encode diverse functional *attributes* that are expressed through *components*. This structural characterization emulates actual biological entities with hierarchical substructures, for instance, proteins. Components can be found in multiple states. Molecule bonding can be expressed for components with or without states, depending on the biological context. A component can only take part in a single reaction at any time. Bionetgen implements, through molecule bonding, a structured representa-

tion of any complex. Intricate configurations are not associated with explicit singular names. The namespace for complex configurations reduces to the number of molecules and their components and not that of the number of all configurations, which can turn out to be infinite. The name *species* indicates a particular configuration of a specific molecule, see [24, 62].

Reactions are regarded as graph-transformations. The transformations implemented are the following: forming a bond, breaking a bond, changing component states, creating and destroying molecules. A reaction in Bionetgen is regarded as a succession of various of the previously mentioned transformations. A rule can incorporate information regarding a set of reactions with identical kinetics which share a common syntactic substructure by ignoring the components that do no influence all reactions in the set. Thus, a rule is a generator of reactions, each of which has a substructure that matches exactly the rule. Nevertheless, the specification of a rule-based model relies on the the identification of autonomous interactions between molecules. *Reactant patterns* are complexes that are not completely specified. Patterns select species that identify a certain substructure that corresponds to the rule. The product of a reaction is attained by species selection through patterns and application of the aforementioned transformations, see [24, 62].

Observables define sums of concentrations of species of interest. There are two types of observables: species-observables and molecules-observables. Species-observables represent the unweighted sum of the concentration of species defined by a specific pattern. Additionally, the molecules-observable evaluates the concentration of the species conforming to the number of matches to a particular given pattern, see [24, 62].

Symmetries and multiplicities

Every reaction generated by a rule is associated with a kinetic rate constant, which consists in the product of the rate constant indicated by the rule and the multiplicity of that specific reaction, which depends on the symmetry of the reaction. The multiplicity of the reaction in turn depends on the number of paths there are from reactants to products. A multiplicity of 1 denotes a singular path from reactants to products, see [22].

Consider the dimerization of molecule A: $2A \rightarrow A_2$. Consider the existence of two types of isoforms A_1 and A_2. The traditional reaction-based representation of the dimerization results in the formation of two types of

dimers, homo- and hetero-dimers, as follows:

$$A_1 + A_2 \rightarrow A_1 : A_2$$
$$A_1 + A_1 \rightarrow A_1 : A_1$$
$$A_2 + A_2 \rightarrow A_2 : A_2$$

Even though the binding affinities are indistinguishable for both homo- and hetero-dimers, the homo-dimerization would come about at half the rate of hetero-dimerization due to the *symmetry effect*, the symmetry in choosing a homo-dimerizing partner. Take for instance the dimerization reaction: $A + A \xrightarrow{k} A : A$. Bionetgen identifies symmetrical reactions and automatically applies the symmetry factor (in the example above: 1/2), see [62]. Hence, the kinetic rate constants are expressed as below:

$$A_1 + A_2 \xrightarrow{k} A_1 : A_2$$
$$A_1 + A_1 \xrightarrow{k/2} A_1 : A_1$$
$$A_2 + A_2 \xrightarrow{k/2} A_2 : A_2$$

Multivalent interactions bring about an *effect of multiplicity*. The multiplicity of a reaction is initially 1 and it increases during the process of model generation when identical instances of the same reaction are generated. Take, for example, molecule A which has two binding-indistinguishable sites for binding molecule B. The traditional reaction-based representation consists of the following reactions:

$$A + B \rightarrow A : B$$
$$A : B + B \rightarrow A : B_2$$

The kinetic rate constant of the first reaction is double than the one of the second reaction due to the number of free binding sites of the A molecule. Bionetgen allows for the multiplicity factor automatically when generating the network. This is usually stated as follows:

$$A + B \xrightarrow{2k} A : B$$
$$A : B + B \xrightarrow{k} A : B_2$$

Multiplicity can originate in rules which generate multiple instances of the same reaction. The total effect over the network is additive. Given, for instance, a process of dissociation of dimers. The rule can be represented as follows:

$$A : A \xrightarrow{k} A + A$$

Taking into account the types complexes that could possibly be generated by dimerization, homodimers and heterodimers respectively, the corresponding reactions in Bionetgen are the following:

$$A_1 : A_2 \xrightarrow{k} A_1 + A_2$$
$$A_1 : A_1 \xrightarrow{k/2} A_1 + A_1$$
$$A_2 : A_2 \xrightarrow{k/2} A_2 + A_2$$

During simulation, the total effect of these three reactions over the network is cumulative.

Example 5 *A rule-based model in Bionetgen for our running example is given in Table 5. We assume here a binding site t for* trans*, a binding site r for* prom *and a binding site p for* prot*.*

Table 2.2: A Bionetgen model for the reaction network in Example 1.

$\mathsf{trans}(t) + \mathsf{prom}(r) \leftrightarrow \mathsf{trans}(t!1).\,\mathsf{prom}(r!1)$
$\mathsf{trans}(t!1).\,\mathsf{prom}(r!1) \rightarrow \mathsf{trans}(t!1).\,\mathsf{prom}(r!1) + \mathsf{prot}(p)$
$\mathsf{trans}(t) + \mathsf{prot}(p) \leftrightarrow \mathsf{trans}(t!1).\,\mathsf{prot}(p!1)$
$\mathsf{trans}(t!1).\,\mathsf{prom}(r!1) + \mathsf{prot}(p) \rightarrow \mathsf{trans}(t!1).\,\mathsf{prot}(p!1) + \mathsf{prom}(r)$

2.3 Petri Net models

Petri nets, defined in [58], appeared as a solution to modelling systems with concurrent processes. At the core of this formalism lies the idea of concurrency through competition on resources. Many types of Petri nets have been investigated: non-deterministic, synchronous, asynchronous, stochastic, continuous, hybrid, etc. Key notions in Petri nets are those of places, transitions and tokens. Concurrency has been implemented by classifying the vertices in two categories: places (which can be regarded as passive) and

transitions (which can be regarded as active). The applications to modelling chemical processes has been proposed by Petri himself, see [50].

Biochemical networks can be regarded as bipartite systems, since they can be abstracted to being composed of two different types of components, the species and the interactions among them. They are also intrinsically concurrent, given that various interactions could arise autonomously and in parallel with one another, see [33].

A qualitative description of a biochemical network can be essentially described by its topology, a bipartite graph, whose arcs encode the stoichiometry of diverse reactions.

The above qualitative description can be extended to a quantitative representation, which relies on introducing *tokens* in places to account either for the number of molecules or concentration of a certain species. Such a quantitative representation makes the distinction between two types of Petri nets: stochastic, characterized by a discrete state space, discrete values for the species and a probabilistic distribution of the reaction rates, and continuous, represented by a continuous state space and deterministic reaction rates, describing the global behaviour of a species through concentrations, see [33].

2.3.1 Petri nets – basic notions

A Petri net is a directed bipartite weighted graph, with two types of nodes: places and transitions. It can be formally defined ([19, 33, 50]) as a quadruple $\mathcal{PN} = (P, T, f, m_0)$, where P and T are finite, non-empty, disjoint sets, representing the set of places and transitions of the net, respectively. Places and transitions are connected through weighted directed arcs, defined by $f : ((P \times T) \cup (T \times P)) \to \mathbb{N}$. The initial marking m_0 is defined by $m_0 : P \to \mathbb{N}_0$.

Given a place p and a marking m, the number of tokens place p holds in marking m is defined by the mapping $m : P \to \mathbb{N}_0$. A place p with $m(p) = 0$ is called *clean*(or *unmarked*). Accordingly, a set of places is also called *clean* if all places are clean, or marked otherwise, see [50].

Considering a vertex $x \in P \cup T$, its preset is defined as: $\bullet x = \{y \in P \cup T | f(y, x) \neq 0\}$ and its postset as: $x\bullet = \{y \in P \cup T | f(x, y) \neq 0\}$. This classification draws distinction then between 4 types of vertices, [33]:

- the preplaces of transition t, denoted as $\bullet t$;
- the postplaces of transition t, denoted as $t\bullet$;
- the pretransitions of place p, denoted as $\bullet p$;
- the posttransitions of place p, denoted as $p\bullet$.

Petri net dynamics

Given a Petri net defined as above, $\mathcal{PN} = (P, T, f, m_0)$, a transition $t \in T$ is *enabled* at a marking m, denoted as $m[t\rangle$ if $\forall p \in {\bullet}t : m(p) \geq f(p,t)$; otherwise, transition t is *disabled*. An enabled transition may *fire*. As a consequence of firing, when transition t fires at marking m, the system reaches a new marking m', denoted as $m[t\rangle m'$, such that:

$$\forall p \in P : m'(p) = m(p) - f(p,t) + f(t,p).$$

The behaviour of the net is given by all possible partially ordered firing sequences or all possible completely ordered firing sequences, see [33].

Each marking is given by the current token dispersal in all places $m \in \mathbb{N}_0{}^{|P|}$. Given a marking m of a Petri net $\mathcal{PN}$, its *set of reachable markings*, denoted as $[m\rangle$, consists of all markings that can be reached from marking m as the result of any firing sequence of arbitrary length. The set of markings reachable from the initial marking, denoted by $[m_0\rangle$, represents the state space of the system, see [33].

Given a reaction network, its Petri net representation is given as follows: model reactants are encoded by places (passive system components) and the reactions by transitions (active system components). Places and transitions are linked through directed edges, which model abstract relationships between components. Edges can also be assigned weights encoding stoichiometric information, whereby the default weight of 1, corresponding to a stoichiometry of 1 is often omitted.

Analysis techniques

Qualitative analysis can bring about new information regarding the constituents of a network and the connections among them, which can be used in the prediction of its behaviour.

A qualitative overview can also verify the concordance between the specification of a network as a Petri net and its conceptual structure, see [50]. A main concern in the analysis of a net is verifying general behavioural properties regardless of its particular functionality, see [33]. The most important three properties are the following:

- *boundedness*: place p of Petri net $\mathcal{PN}$ is *bounded* (*k-bounded*), if $\exists\ k \in \mathbb{N}_0$ such that $\forall m \in [m_0\rangle$: $m(p) \leq k$. Additionally, Petri net $\mathcal{PN}$ is *k-bounded* if all its places are k-bounded, see [33].

- *liveness*: Transition t is defined as *dead* in marking m if, given any marking m' reachable from m, t is not enabled in marking m'. Transition t is defined as *live* if, given any marking reachable from the initial marking m_0, t is not dead in the marking under consideration. Petri

net marking m of $\mathcal{PN}$ is defined as *dead* if it enables no transitions in $\mathcal{PN}$. Petri net $\mathcal{PN}$ is defined as $deadlock - free$ if dead markings are not reachable and *live* if all transitions are live. For a more detailed discussion see [33, 50].

- *reversibility*: A Petri net is *reversibile* if its initial marking m_0 is reachable from any reachable marking: $\forall m \in [m_0\rangle : m_0 \in [m\rangle$, see [33, 50].

The behavioural properties of a Petri net can be also tackled by the analysis of its corresponding *incidence matrix*. The incidence matrix C of a Petri net $\mathcal{PN} = (P, T, f, m_0)$ is an integer $(|P| \times |T|)$-matrix, whose entries $c_{i,j}$ express the token change at place p_i generated by the firing of transition t_j, see [33, 50]. Consequently, the entry for a pre-place of a given transition t (which is not also a post-place) is negative and for a post-place (which is not a pre-place) is positive and their absolute values match the respective arc multiplicities. The entry for a place which is a pre-place as well as a post-place for a transition t is given by the difference of arc multiplicities, see [33].

In this context, one is interested in the non-trivial solutions of the homogenous systems of linear equations, see [33, 50]:

$$x \cdot \mathsf{C} = 0 \quad \text{and} \quad \mathsf{C} \cdot y = 0. \tag{2.35}$$

The non-trivial non-negative integer solutions to these homogenous systems of linear equations are place vectors, called *P-invariants*, or transition vectors, called *T-invariants*, respectively. A P-invariant represents a set of places across the network for which the weighted sum of tokens is fixed, regardless of any combination of firings.

Given a P-invariant $w = (w_1, \cdots, w_{|P|})$, then it holds that:

$$w_1 m^{(1)} + \cdots + w_{|P|} m^{|P|} = w_1 m_0^{(1)} + \cdots + w_{|P|} m_0^{|P|},$$

for all $m = (m^{(1)}, \cdots, m^{|P|})$ reachable from m_0.

Respectively, the multiplication of a T-invariant and a row of matrix C corresponding to a place of the network gives 0. Therefore, the T-invariant has no effect on the marking of the place under consideration. For more details, we refer to [33].

By definition ([33, 50]), the set of vertices corresponding to the entries included in the invariant is called *the support of the invariant*. Given invariant x, the support of x, denoted as $supp(x)$ is:

$$supp(x) = \{x_i \in P \parallel x_i \in T \mid x_i \neq 0\}.$$

An invariant x is *minimal*, if its support is minimal, i.e., there is no invariant z, such that $supp(z) \subset supp(x)$ and the greatest common divisor of all non-zero elements of x is 1.

A Petri net $\mathcal{PN}$ is *covered by P-invariants* iff for all places $p_i \in P$, there is a P-invariant x, such that $p_i \in supp(x)$. Analogously, a Petri net is *covered by T-invariants* iff $\forall$ transitions $t_j \in T$, $\exists$ T-invariant y, such that $t_j \in supp(y)$.

2.3.2 Quantitative Petri nets

Qualitative Petri net analysis gives the potential behaviour of the network under consideration regardless of time restrictions. However, a more complex analysis of a Petri net requires quantitative analysis. We discuss it briefly in the following section.

Stochastic Petri nets

A stochastic Petri net considers only a discrete number of tokens distributed across the network. However, in contrast to the time-free class of Petri nets, each transition t of the net has a corresponding firing rate, representing the waiting time associated with the firing of that particular transition. The firing rate is a random variable $X_t \in [0, +\infty)$, specified as a probability distribution. The system is described accordingly by a discrete state space, see [19, 33].

A stochastic Petri net is defined([19, 33]) as a pair as follows:

$$\mathcal{SPN} = \{\mathcal{PN}, \upsilon\},$$

where $\mathcal{PN}$ is a Petri net as described above and υ is a stochastic hazard function which defines, for every transition, a transition rate dependent on the current marking: $\upsilon_t : {\mathbb{N}_0}^{|\bullet t|} \to \mathbb{R}_+, \forall t \in T$.

The firing rates of a transition t is exponentially distributed and is described by a parameter λ, which is dependent on the marking and which determines the behaviour of the respective transition: $\lambda_t(m)$.

The time required for a transition to fire is given by a random variable X_t with the probability density function:

$$f_{X_t}(\tau) = \lambda_t(m) \cdot e^{-\lambda_t(m)\cdot\tau}, \tau \geq 0.$$

The semantics of $\mathcal{SPN}$ is characterized by a continuous time Markov chain. A stochastic Petri net describing a biochemical process can be described by giving a more explicit stochastic hazard function. An example to this end is the *stochastic mass-action hazard function*, where tokens represent molecules ([33]):

$$h_t = c_t \cdot \prod_{p \in \bullet t} \binom{m(p)}{f(p,t)},$$

where c_t is a constant specific to the transition under consideration and $m(p)$ is the number of tokens at the preplace of the current transition.

Continuous Petri nets

A continuous Petri net considers the marking at a place as a *token value*, which corresponds to a real positive number and it accounts for the concentration level of the species represented by that particular place.

A continuous Petri net is defined [19, 33] as a tuple $\mathcal{CPN} = (P, T, f, v, m_0)$, where P and T are finite, non-empty, disjoint sets, representing the set of continuous places and continuous transitions of the net, respectively. Function v is defined as follows:

$$v_t : \mathbb{N}_0{}^{|\bullet t|} \to \mathbb{R}_+, \forall t \in T.$$

Places and transitions are connected through weighted directed arcs, defined by:

$$f : ((P \times T) \cup (T \times P)) \to \mathbb{R}_{\geq 0}.$$

The initial marking m_0 is defined by $m_0 : P \to \mathbb{R}_{\geq 0}$.

A continuous marking of $\mathcal{CPN}$ is a place vector, defined as: $m \in (\mathbb{R}_{\geq 0})^{|P|}$. Then $m(p)$ represents the marking at place p, given by a real number as opposed to an integer number in a stochastic Petri net. Continuous transition t is enabled at marking m, if:

$$\forall p \in \bullet t : m(p) > 0.$$

The semantics for a continuous Petri net is characterized by a system of differential equations, depicting the continuous change in time of the token value of a specific place p given by a continuous rise in the flow corresponding to its pre-transitions and the reduction in the flow corresponding to its post-transitions [33], as follows:

$$\frac{dm(p)}{dt} = \sum_{t \in \bullet p} f(t,p) \cdot v(t) - \sum_{t \in \bullet p} f(p,t) \cdot v(t).$$

Example 6 *The Petri net model of our running example is illustrated in Figure 2.1.*

2.4 Reaction systems

Reaction systems are a formal framework designed for the representation and qualitative analysis of biochemical reaction networks. The formal framework for modelling reactions relies on two mechanisms: *facilitation* and *inhibition*, see [11, 21].

A reaction system consists in a finite set of reactions. Each reaction is described by a finite set of reactants (necessary for the reaction to occur), a

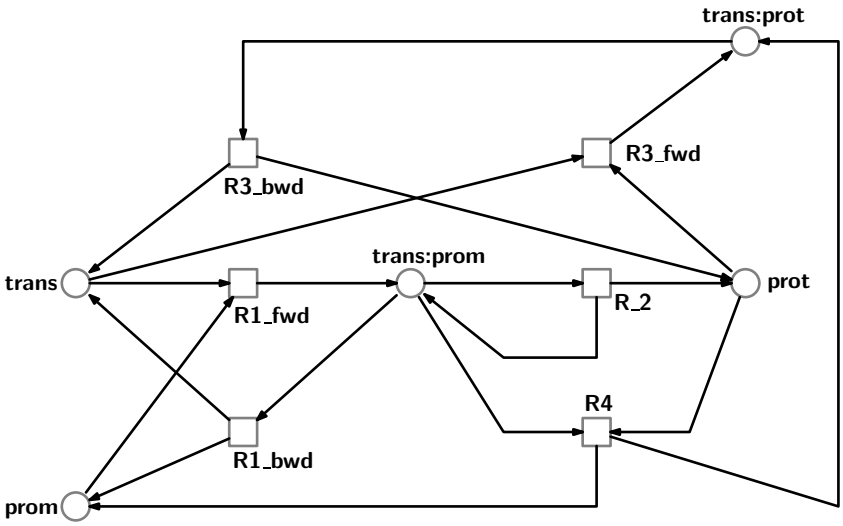


Figure 2.1: The Petri net model corresponding to the reaction network in Example 1.

finite set of inhibitors (which preclude the reaction from taking place) and a finite set of products (representing the results of the reaction occurrence). The reunion of the set of reactants and the set of inhibitors is defined as the set of resources of a reaction.

Three crucial presumptions that reaction systems rely on distinguish them from all other biomodelling frameworks:

- *The threshold assumption:* the presence of a resource ensures the existence of a "sufficient amount"of that particular resource, precluding any possible conflict between various reactions which consume the resource;

- *No permanency:* any entity that is not produced by at least one of the reactions enabled in the current state dissipates;

- The model is always an open system, where the contribution of the environment is explicit.

By definition, see [21], a reaction is characterized by a triplet $a = (R, I, P)$, where R, I, P are finite, non-empty sets representing the sets of reactants, inhibitors and products, respectively, and $R \cap I = \emptyset$. For a set S satisfying $R, I, P \subseteq S$, then a is a reaction in S. The set of reactions in S is denoted by $rac(S)$.

Given a set of reactions A and a finite set T with $a \in A$, then the *result* of a on T ([21]), denoted $res_a(T)$, is defined as:

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

Consequently, the *result of A on T* ([21]), denoted as $res_A(T)$, is defined as follows:

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

A *reaction system* is therefore defined as an ordered pair $\mathcal{A} = (S, A)$, so that S is a finite set and $A \subseteq rac(S)$. The set S is called the *background set* of A. Given reaction system $\mathcal{A} = (S, A)$ and a set $T \subseteq S$, the result of $\mathcal{A}$ on T, $res_{\mathcal{A}}(T)$, is:

$$res_{\mathcal{A}}(T) = res_A(T).$$

The formalisation for the dynamic behaviour of a reaction system is given by the concept of *interactive process*. By definition ([21]), an interactive process in $\mathcal{A}$ is a pair $\pi = (\gamma, \delta)$, with $\gamma = C_0, C_1, \ldots, C_n, \delta = D_1, D_2, \ldots, D_n \subseteq S$, $n \geq 1$, such that:

$$\begin{cases} D_1 = res_{\mathcal{A}}(C_0) \\ D_i = res_{\mathcal{A}}(C_{i-1} \cup D_{i-1}), & \text{for each } 1 < i \leq n. \end{cases}$$

Sequence γ is the *context (or interaction) sequence* of π, $con(\pi)$, and sequence δ is the *result sequence* of π, $res(\pi)$. The *state sequence* of π is defined as $\tau = W_0, W_1, \ldots, W_n$, with $W_i = C_i \cup D_i$, for all $i \in \{0, \ldots, n\}$ and $W_0 = C_0$. W_0 is called the *initial state* of π, $\text{init}(\pi)$, and W_n is called the *final state* of π, $\text{fst}(\pi)$.

Interactive process π is *context-independent* if for all $i \in \{0, \ldots, n\}$, $C_i \subseteq D_i$. The context sequence promotes the idea that the behaviour of the system is influenced by the environment through context. Context-independent sequences depend only on their initial states and lengths. Moreover, in a context-independent sequence, $\tau = W_0, W_1, \ldots, W_i, W_{i+1}, \ldots, W_n$, during the transition $W_i \to W_{i+1}$, all elements of $W_i - res_{\mathcal{A}}(W_i)$ dissipate.

Chapter 3

Quantitative model refinement

Model refinement has been extensively investigated in the context of software engineering, the field it originated from. We briefly introduce as follows the main concepts revolving around model refinement in software engineering. Subsequently, we discuss the context of model refinement in systems biology, in particular with respect to rule-based modelling. Then we introduce the quantitative model refinement framework and discuss how to set the parameters of a given model in our framework so that it preserves previously obtained fits.

3.1 Refinement in software engineering

The problem of *stepwise refinement* has been greatly studied in the field of software engineering, particularly in the context of parallel computing. *Refinement calculus* was defined as a logical framework for the analysis of computer programs, which tackles two aspects: the *correctness* of programs relative to a given specification and the *refinement* of a program assuring its correctness preservation, see [3].

In software engineering, the concept of refinement and, more specifically, the framework of refinement calculus, are closely related to the *algebra of contracts*. Computer programs and their specifications can be regarded as *contracts* between separate *agents*. Agents can be defined as entities that are able to make free choices, reason for which their behaviour is considered *nondeterministic*. Agents can be users, programmers, processes of a computer system etc. and, consequently, the nondeterminism originates in *information hiding*, see [3]. Information hiding refers to the segregation of decisions made by a specific agent. In software engineering, the concept of information hiding is to some extent equivalent to the concept of *encapsula-*

tion, see [9]. However, the scope is to protect the information regarding an agent's behaviour so as to hold open the prospects for a further extension of its behaviour. An agent can change the environment through actions and choose between various lines of action. This behaviour is regulated by contracts, since a contract can impose the order for the actions to be performed. Therefore, refinement is defined as relation between contracts, see [3].

Back et al. in [3] argue that classic programming constructs, conditional statements, iteration, recursion etc., can be depicted as contracts. In this context, two notions are essential: the *assertion* and the *assumption*. An assertion is defined as the condition an agent is required to satisfy in a certain state. An agent continues to complete the contract if the assertion holds in the given state. Contrarily, if the assertion does not hold, the contract is *breached*. Assumptions, however, are relevant in the context of contracts between two entities. Each entity makes assumptions which restrict the engagement of the entity in the contract.

An agent is considered to satisfy its contract S to establish a given postcondition q in initial state σ if it reaches a final state in which q is established (without breaching the contract) or the agent is exempted from satisfying the contract by an assumption that was breached, see [3]. This description is denoted by:

$$\sigma\{|S|\}q. \tag{3.1}$$

The notation for a contract statement will be here assimilated with the one used above for a contract. Given two contract statements corresponding to an agent, denoted as S and S', S' is said to be "as good as" S if S' can establish any condition that S establishes, and S is *refined* by S', denoted as: $S \sqsubseteq S'$. Formally, $S \sqsubseteq S'$ holds if:

$$\sigma\{|S|\}q \Rightarrow \sigma\{|S'|\}q, \forall\sigma \text{ and } q.$$

The specification of program statements is in general attained by providing a precondition and a postcondition that the implementation must fulfill. Then, given a contract statement S and two predicates accounting for the precondition p and postcondition q, the correctness is defined([3]) as follows:

Contract statement S is *correct* relative to p and q, denoted as $p\{|S|\}q$, if for any σ that satisfies p, $\sigma\{|S|\}q$ holds.

In other words, for any initial state satisfying the precondition, the agent under consideration can choose a sequence of events that either establishes the postcondition q or that resides in at least one of its assumptions being breached.

In software engineering, at large, the general purpose is building programs that are correct in regards to a given specification and this represents the *programming problem*. The framework of refinement calculus is meant for the study of this problem.

3.1.1 Stepwise refinement

Stepwise refinement aroused as a fundamental application of the refinement calculus from the need to ensure that the derivation of a program corresponds to its original specification. At the heart of stepwise refinement lies a high-level specification of the program requirements, which emerges into the implementation of the program. This implementation, however, consists of various implementation statements, which might describe only partially the specification. The elements of the specification uncovered by an implementation statement can be however implemented by other implementation statements, which might further be incomplete relative to the entire set of specifications. This process lasts until the entire set of specifications was covered by the implementation and resides into an executable program. This is a *top-down* approach in the program development cycle and it has been associated with the concept of *program transformation*, which can make changes in the program by applying certain transformation while preserving program correctness, see [3].

The refinement between two contract statements preserves correctness:

If $S \sqsubseteq S'$ and $p\{|S|\}q$ holds, then $p\{|S'|\}q$ holds, $\forall$ precondition p and postcondition q.

This leads to the formalization of stepwise refinement. Given an initial statement S_0, assuming it satisfies $p\{|S_0|\}q$, one can derive a sequence of refinements of the form:

$$S_0 \sqsubseteq S_1 \sqsubseteq \cdots \sqsubseteq S_n.$$

Taken into account that the refinement relation is transitive, we obtain that $S_0 \sqsubseteq S_n$ and given that, by definition, refinement preserves correctness, then $p\{|S_n|\}q$ holds as well. So, S_n is derived from the original statement S_0 and it satisfies the original requirement. For more details, we refer the reader to [3].

3.2 Refinement in systems biology

Model refinement plays a central role in the model development cycle. Starting from an already fit model, integrating new components within the model, having modules eliminated or substituted for others, adding species and/or reactions to the model, induces the alteration of its fit. Very often the newly attained model has a significantly larger number of parameters compared to the model of reference. Even for fairly small models, parameter estimation proves to be computationally-intensive, hence dissipative resource-wise. An alternative to engaging into a re-estimation of all parameters of the newly attained model is an iterative approach. Such an approach consists in the

construction of a model, which ensures quantitative model fit preservation at every step of the development. This method is called *quantitative model refinement* and it has been previously investigated in [16, 17, 22, 25]. In rule-based modelling, the focus is mostly on encapsulating data refinement as an inherent property of agents, implementing data refinement through the concept of agent resolution. This thesis is concerned with a novel framework for quantitative model refinement, whereby we focus on refinement as a method to expand an already fit model to include more information about its reactants and/or reactions, setting the parameters of the refined model so that model fit is preserved.

Example 7 *We refined our running example considering the phosphorylation status of a specific site of the transcription factor* trans*. To keep the example simple, we will assume that site to be binding site* t*. The effect of phosphorylation on our model is assumed to be only in the second reaction, which will only be active in case the transcription factor is phosphorylated on site* t*. The refined reaction network is given in Table 3.1.*

Table 3.1: The refined running example.

$$
\begin{aligned}
&\mathsf{trans_u} + \mathsf{prom} \leftrightarrow \mathsf{trans_u}\colon \mathsf{prom}\\
&\mathsf{trans_p} + \mathsf{prom} \leftrightarrow \mathsf{trans_p}\colon \mathsf{prom}\\
&\mathsf{trans_p}\colon \mathsf{prom} \rightarrow \mathsf{trans_p}\colon \mathsf{prom} + \mathsf{prot}\\
&\mathsf{trans_p} + \mathsf{prot} \leftrightarrow \mathsf{trans_p}\colon \mathsf{prot}\\
&\mathsf{trans_u} + \mathsf{prot} \leftrightarrow \mathsf{trans_u}\colon \mathsf{prot}\\
&\mathsf{trans_p}\colon \mathsf{prom} + \mathsf{prot} \rightarrow \mathsf{trans_p}\colon \mathsf{prot} + \mathsf{prom}\\
&\mathsf{trans_u}\colon \mathsf{prom} + \mathsf{prot} \rightarrow \mathsf{trans_u}\colon \mathsf{prot} + \mathsf{prom}
\end{aligned}
$$

3.2.1 Refinement in rule-based modelling

Rule-based modelling tackles the problem of combinatorial explosion of biomodels accounting for protein-protein interactions, posttranslational modifications, etc., through agent- and rule-resolution. The concept of *agent resolution* relies on hierarchical agents which allow for the production of new agents from existing ones making minor changes in the specification of their sites, see [32]. A rule represents the formalization of a biological process. A set composed of generic rules can be scaled using different types of agents: generic or concrete. Another option for representing the system at a different level of resolution is *rule instantiation* or rule refinement, which consists in refining a given rule to allow only particular states corresponding to certain descendants of the rule, see [32].

Example 8 *To build a rule-based model for our refined reaction network in Example 7, we assume that site t of* trans *will have two specific internal states: t_u (standing for unphosphorylated) and t_p (standing for phosphorylated). The Bionetgen model we built in this case is given in Table 3.2.*

$\mathsf{trans}(t) + \mathsf{prom}(r) \leftrightarrow \mathsf{trans}(t!1).\,\mathsf{prom}(r!1)$
$\mathsf{trans}(t_p!1).\,\mathsf{prom}(r!1) \rightarrow \mathsf{trans}(t_p!1).\,\mathsf{prom}(r!1) + \mathsf{prot}(p)$
$\mathsf{trans}(t) + \mathsf{prot}(p) \leftrightarrow \mathsf{trans}(t!1).\,\mathsf{prot}(p!1)$
$\mathsf{trans}(t!1).\,\mathsf{prom}(r!1) + \mathsf{prot}(p) \rightarrow \mathsf{trans}(t!1).\,\mathsf{prot}(p!1) + \mathsf{prom}(r)$

Table 3.2: A Bionetgen model for the reaction network in Example 7.

3.2.2 Refinement of reaction-based models

Data refinement of reaction-based models

We introduced the quantitative refinement of reaction-based models in [40] and further extended it in [28]. A reaction-based model can be refined to integrate additional information regarding its reactants and/or reactions. Various types of refinement can be employed, relying upon the entities to be refined, for instance the reactants or the reactions of the model. If the modeler is specifically interested in data, the model can be refined by including more information regarding the internal states and/or attributes of certain variables. This type of refinement is called *data refinement* and it essentially consists in the substitution of one species for its subspecies. The focus here lies on the behavioural variations of the model's subspecies. Another type of refinement concerns the reactions of the model, where a generic reaction describing a specific process is replaced by a set of reactions specifying the transitional steps of the process. The latter type of refinement is called *process refinement*.

Consider model $M = (\sum, \mathcal{R})$, where $\sum = \{S_1, \ldots, S_m\}$ is the set of species and $\mathcal{R} = \{R_1, R_2, \cdots, R_n\}$ is the set of reactions of the form (2.1). It can be refined to distinguish between various subspecies of any species, say S_1. The distinction between the subspecies is very often drawn by post-translational modifications such as acetylation, phosphorylation, sumoylation, etc. All previously mentioned subspecies of S_1 take part in all reactions S_1 engaged in, conceivably obeying a different kinetic setup. Given model M and species S_1, substituting subspecies $B_1, \ldots, B_l$ for species S_1 in M leads to attaining a new model M_R, comprising species $\{S_2', S_3', \ldots, S_m'\} \cup \{B_1, \ldots, B_l\}$, for some $l \geq 2$, where variables S_j', $2 \leq j \leq m$ from M_R,

coincide with S_j from model M and $B_1, \ldots, B_l$ substitute for species S_1 in M_R. Furthermore, each reaction R_j of M is replaced in the new model M_R by all possible reactions $R_{i,j}$ of the following form:

$$R_{i,j} \ : \ (t^i_{1,j}B_1 + \ldots + t^i_{l,j}B_l) + c_{2,j}S'_2 + \ldots + c_{m,j}S'_m \xrightarrow{k_{i,j}} (t'^i_{1,j}B_1 + \ldots + t'^i_{l,j}B_l) + c'_{2,j}S'_2 + \ldots + c'_{m,j}S'_m,$$

where $k_{i,j}$ is the kinetic rate constant of $R_{i,j}$ and $(t^i_{1,j}, \ldots, t^i_{l,j}, t'^i_{1,j}, \ldots, t'^i_{l,j})$ are all possible nonnegative integers so that $t^i_{1,j} + \ldots + t^i_{l,j} = c_{1,j}$ and $t'^i_{1,j} + \ldots + t'^i_{l,j} = c'_{1,j}$. Model M_R is said to be a *data refinement of model M on variable S_1* if and only if the following conditions are fulfilled:

$$[S_j](t) = [S'_j](t), \tag{3.2}$$

$$[S_1](t) = [B_1](t) + \ldots + [B_l](t), \tag{3.3}$$

for all $2 \leq j \leq m$, $t \geq 0$. Fulfilling these conditions depends on the numerical setup of model M_R, i.e., on the kinetic constants of its reactions (both those adopted from the basic model, as well as those newly introduced in the construction) and on the initial concentrations of its species.

The refined model, M_R comprises $m + l - 1$ reactants. The number of reactions in M_R replacing reaction $R_{i,j}$ of M is given by the number of non-negative integer solutions of system of equations below:

$$t^i_{1,j} + \ldots + t^i_{l,j} = c_{1,j}; \tag{3.4}$$

$$t'^i_{1,j} + \ldots + t'^i_{l,j} = c'_{1,j}, \tag{3.5}$$

over the independent unknowns $t_{k,j}, t'_{k,j}, 1 \leq k \leq l$.

The number of solutions of equation (3.4) is given by the *multinomial coefficient* "l multichooses $c_{1,j}$", see [26]:

$$\left(\!\!\binom{l}{c_{1,j}}\!\!\right) = \binom{l + c_{1,j} - 1}{c_{1,j}} = \frac{(l + c_{1,j} - 1)!}{c_{1,j}!(l-1)!}.$$

The equations in the system are independent, resulting in a number of solutions given by $\left(\!\!\binom{l}{c_{1,j}}\!\!\right)\left(\!\!\binom{l}{c'_{1,j}}\!\!\right)$, representing the augmentation in the number of reactions from M to M_R. The number of new free kinetic parameters in M_R is given by $\left(\!\!\binom{l}{c_{1,j}}\!\!\right)$, the number of all possible combinations among reactants $B_1, ..., B_l$. Any reaction with a stoichiometry greater than two has been disregarded in [40] due to biokinetics, for more details see [54]. If $c_{1,j} = 1$, M_R witnesses a linear increase in the parameter space (with $\left(\!\!\binom{l}{1}\!\!\right) = l$ new kinetic parameters). If $c_{1,j} = 2$, M_R witnesses a quadratic increase in the parameter space (with $\left(\!\!\binom{l}{2}\!\!\right) = \frac{l(l+1)}{2}$ new parameters).

From atomic to complex species

The refinement of a species in a given model leads to cascading refinements, whose dynamics can be manipulated by partitioning the set of species in two categories: *atomic* and *complex*, see [28]. Given model M, described in Section 2.1, a complex species consists of two or more atomic species. Let Γ_M be the set of atomic species $\Gamma_M = \{S_1, \ldots, S_l\}$ and Δ_M the set of the complex species $\Delta_M = \{S_{l+1}, \ldots, S_m\}$. Each complex species $CS \in \Delta_M$ is associated with a multiset $\sigma(CS)$, detailing on its atomic constituents:

$$\sigma(CS) = \{m_1^{CS} S_1, \ldots, m_l^{CS} S_l\},$$

where $m_1^{CS}, \ldots, m_l^{CS} \in \mathbb{N}$ and $\sum_{i=1}^{l} m_i^{CS} \geq 2$. If $m_i^{CS} \neq 0$, then complex species CS *contains* atomic species S_i with multiplicity m_i^{CS}. Each atomic species $S_i, 1 \leq i \leq l$, associated with a set of complex species containing S_i, is denoted as $\mathcal{C}_M(S_i)$:

$$\mathcal{C}_M(S_i) = \{CS \in \{S_{l+1}, \ldots, S_m\} \mid m_i^{CS} \neq 0\}.$$

Taken into consideration the notations before, we are only interested, in this formulation, in the refinement of a single atomic species, say S_1. Refinement of species S_1 results into the refinement of all complexes of in $\mathcal{C}_M(S_1)$. Given a species $X \in \Sigma_M \setminus \big(\{S_1\} \cup \mathcal{C}_M(S_1)\big)$ which is not to be refined, its substitution in the refined model M_R will be X_R.

Species S_1 is replaced by $\{S_1^1, \ldots, S_1^\rho\}$ in the refined model and, consequently, species $CS \in \mathcal{C}_M(S_1)$ with $\sigma(CS) = \{m_1 S_1, \ldots, m_l S_l\}$ is replaced by $\mathcal{R}(CS) = \{CS^1, \ldots, CS^\mu\}$.

Refined species CS^i are complex species with $\sigma(CS^i)$ of the form

$$\sigma(CS^i) = \{\tau_1 S_1^1, \ldots, \tau_\rho S_1^\rho, m_2 S_2^R, \ldots, m_l S_l^R\},$$

where $\tau_1, \ldots, \tau_\rho \in \mathbb{N}$ and $\tau_1 + \cdots + \tau_\rho = m_1$.

The refined model M_R consists of atomic species $\Gamma_{M_R} = \{S_1^1, \ldots, S_1^\rho\} \cup \{S_2^R, \ldots, S_l^R\}$ and of complex species $\Delta_{M_R} = \bigcup_{CS \in \mathcal{C}_M(S_1)} \mathcal{R}(CS) \bigcup \{CS^R \mid CS \in \Delta_M \setminus \mathcal{C}_M(S_1)\}$.

For every reaction R_j in M, S_1 and the species in $\mathcal{C}_M(S_1)$ are replaced by their corresponding refined subspecies, in all possible combinations. Given species X to be refined into the set of species $\{X^1, \ldots, X^\nu\}$, consider that it has stoichiometric coefficient c in reaction R, then it will be replaced in the refinement of R by $c^1 X^1 + \ldots + c^\nu X^\nu$, where $c^1, \ldots, c^\nu \in \mathbb{N}$, $c^1 + \ldots + c^\nu = s$. Reaction R is replaced with all possible reactions $\bar{R}$ of the form:

$$\bar{R} \ : \big(c_1^1 S_1^1 + \ldots + c_1^\rho S_1^\rho\big) + \big(c_2 S_2^R + \ldots + c_l S_l^R\big) + \sum_{S_i \in \mathcal{C}(S_1)} \sum_{C \in \mathcal{R}(S_i)} c_{S_i}^C C + \sum_{S_i \in \Delta \setminus \mathcal{C}_M(S_1)} c_i S_i^R$$
$$\xrightarrow{k_{\bar{r}}} \big(c_1'^1 S_1^1 + \ldots + c_1'^\rho S_1^\rho\big) + \big(c_2' S_2^R + \ldots + c_l' S_l^R\big) + \sum_{S_i \in \mathcal{C}(S_1)} \sum_{C \in \mathcal{R}(S_i)} c_{S_i}'^C C + \sum_{S_i \in \Delta \setminus \mathcal{C}_M(S_1)} c_i' S_i^R,$$

where:

- $c_1^1 + \ldots + c_1^\rho = c_1,\ c_1^1, \ldots, c_1^\rho \in \mathbb{N}$;
- $c_1'^1 + \ldots + c_1'^\rho = c_1',\ c_1'^1, \ldots, c_1'^\rho \in \mathbb{N}$;
- $\sum_{C \in \mathcal{R}(S_i)} c_{S_i}^C = c_i,\ c_{S_i}^C \in \mathbb{N},\ \forall S_i \in \mathcal{C}(S_1)$;
- $\sum_{C \in \mathcal{R}(S_i)} c_{S_i}'^C = c_i',\ c_{S_i}'^C \in \mathbb{N},\ \forall S_i \in \mathcal{C}(S_1)$.

Fit-preserving refinement

Model M_R is a *data refinement of model M on variable S_1* iff the conditions below are satisfied $\forall t \geq 0$:

$$[S_1](t) = [S_1^1](t) + \ldots + [S_1^\rho](t); \tag{3.6}$$
$$[S_i](t) = \sum_{C \in \mathcal{R}(S_i)} [C](t), \forall S_i \in \mathcal{C}(S_1); \tag{3.7}$$
$$[S_i](t) = [S_i^R](t), \forall S_i \in \Sigma_M \setminus \{S_1\} \cup \mathcal{C}_M(S_1). \tag{3.8}$$

The setup of the model, the kinetic parameters and the initial concentrations of the species of the the initial model influence the fulfillment of these conditions.

Extrapolating from Section 2.1, a reaction network consisting of m species and n reactions, $m, n \in \mathbb{N}^*$, can be refined to include more details regarding its species and/or reactions. However, the refinement of one atomic species alone propagates through the whole network, making a distinction for all complex species containing the atomic species to be refined. To extend on the data refinement in Section 3.2.2, we consider here a different formulation for refinement, all species S_i are refined into $\{B_{i1}, \ldots, B_{ip_i}\}$, where $p_i \geq 1$, where $p_i = 1$ indicates that the corresponding species are in fact not refined.

We obtain in this manner a new model $M_R{}'$, consisting of species $\{B_{iq} \mid 1 \leq i \leq m, 1 \leq q \leq p_i\}$. Given a species S_i, all its subspecies engage in all

reactions S_i participated in, assumably obeying a different kinetic law than that of the ancestor species. A reaction of the form:

$$r_j : \sum_{i=1}^{m} c_{ji} S_i \xrightarrow{k_j} \sum_{i=1}^{m} d_{ji} S_i, \ \textit{for } 1 \leq j \leq n \ . \tag{3.9}$$

is replaced in the refined model M_R by all possible reactions $r_j^{(l)}$ of the following form:

$$r_j^{(l)} \ : \ \sum_{q=1}^{p_1} c_{j1q}^{(l)} B_{1q} + \ldots + \sum_{q=1}^{p_m} c_{jmq}^{(l)} B_{mq} \xrightarrow{k_j^{(l)}} \sum_{q=1}^{p_1} d_{j1q}^{(l)} B_{1q} + \ldots + \sum_{q=1}^{p_m} d_{jmq}^{(l)} B_{mq}, \tag{3.10}$$

where $k_j^{(l)}$ is the kinetic rate constant of reaction $r_j^{(l)}$ and coefficients $c_{jiq}^{(l)}$ and $d_{jiq}^{(l)}$ are non-negative integers such that $c_{ji1}^{(l)} + \ldots + c_{jip_i}^{(l)} = c_{ji}$ and $d_{ji1}^{(l)} + \ldots + d_{jip_i}^{(l)} = d_{ji}$.

Considering model $M_R{}'$ defined above, the aim is to find values for its kinetic rate constants which typify identical dynamics to ancestor species, as follows:

$$[S_i](t) = [B_{i1}](t) + \ldots + [B_{ip_i}](t) \tag{3.11}$$

for all $1 \leq i \leq m$ and $t \geq 0$. The aforementioned constraints characterize the *fit-preserving refinement*, as defined in [27]. The framework specifies a set of constraints which ensure that (3.11) is satisfied. A formal definition of these constraints requires some preliminary notations: we collect the stoichiometric coefficients of the original reactions r_j into vectors $\boldsymbol{c_j}$ and $\boldsymbol{d_j}$, respectively. Furthermore, we assemble the stoichiometric coefficients of the refined reactions $r_j^{(l)}$ in vectors $\boldsymbol{c_j^{(l)}}$ and $\boldsymbol{d_j^{(l)}}$. Herewith, the sufficient conditions for $M_R{}'$ to be a fit-preserving refinement of M, as introduced in [27], are the following:

$$\sum_{l \ s.t. \ \boldsymbol{c_j^{(l)}} = \boldsymbol{c_j^{(s)}}} k_j^{(l)} = \binom{\boldsymbol{c_j}}{\boldsymbol{c_j^{(s)}}} k_j \ , \tag{3.12}$$

for any reaction r_j and any chosen left-hand side $\boldsymbol{c_j^{(s)}}$ of corresponding refined reactions. The sum is defined over all refined reactions that have a communal left-hand side. The formula above is based on the generalized binomial coefficient for vectors $\boldsymbol{x}, \boldsymbol{y}$ shown below:

$$\binom{\boldsymbol{x}}{\boldsymbol{y}} = \frac{\prod_i x_i!}{\prod_j y_j!}, \ .$$

In practice, we divided the value inferred from the constraint to the actual number of reactions with an identical selected left-hand side.

The fit-preserving refinement method described above offers a sufficient condition for setting the kinetic rate constants to obtain a structurally refined reaction network, whose dynamics is preserved from that of the original model. However, it is yet uncertain whether the condition is also necessary in this formulation.

This method for setting the kinetic parameters of the refined model is very flexible, given that it allows for an infinite number of assignments for the kinetic parameters, based on the constraint that the refined parameters depend on the values of their counter-parts only to some extent, particularly they add up to certain values dependent on the original model.

Also, fit-preserving refinement can be integrated with partial information on the kinetic parameters of the refined model, allowing for construction of models which have some of their kinetic parameters automatically set. This could be an essential step in improving the performance of parameter estimation methods.

One practical use of fit-preserving refinement is the construction of large quantitative models, allowing to capture diverse level of details for certain biological processes. In the next chapter, we introduce two case studies, the heat shock response and the ErbB signalling pathway. We refined the models describing these processes and describe the behaviour of the refined models.

Chapter 4

Case studies

4.1 The heat shock response

The eukaryotic heat shock response is an evolutionarily conserved regulatory network that controls the cellular response to proteotoxicity originating in various environmental and physiological stressors, such as elevated temperatures. It acts as a defense mechanism, which is crucial for the adaptation and survival of the cell. Elevated temperatures have deleterious effects on proteins, which can unfold, misfold or aggregate, impairing protein homeostasis and eventually inducing apoptosis (controlled cell death). Protein homeostasis is promoted by the synthesis of molecular chaperones. They assist the cell in the process of recovery and minimize protein damage, which would otherwise be lethal.

4.1.1 The basic reaction network model

We consider here the initial molecular model for the heat shock response, introduced in [57]. Heat shock proteins (hsp's) play a key role in the recovery of protein homeostasis, assisting misfolded proteins (mfp's) in the process of refolding (chaperoning) and inducing the degradation of acutely damaged proteins, thus promoting cell survival. Due to the affinity towards misfolded proteins, heat shock proteins sequester misfolded proteins, leading to the formation of hsp: mfp complexes, providing an environment favorable to the recovery of their native conformation (prot). The transactivation of the hsp-encoding genes regulates the response to heat stress. Gene transcription is activated by specific proteins, called heat shock factors (hsf's). In the absence of environmental stressors, such as elevated temperatures, heat shock factors can be found extensively in a monomeric conformation, predominantly bound to heat shock proteins (hsp: hsf). Heat stress induces protein misfolding and breakage of hsp: hsf complexes, which results into the release of hsf's. Consequent to heat stress and upon release, heat shock

factors dimerize (hsf_2) and, subsequently, trimerize (hsf_3), reaching a DNA binding competent conformation, see [57, 63]. In this conformation, they bind heat shock elements (hse), i.e. regulatory upstream promoter elements of the hsp-encoding genes. Consequently, DNA binding enables the transcription and translation of the hsp-encoding gene, inducing hsp synthesis at a significantly higher rate.

Once the amount of heat shock proteins reaches a sufficient level for the cell to withstand the heat stress, hsp synthesis is turned off due to the excess of heat shock proteins. For this purpose, heat shock proteins sequestrate free hsf's, break dimers and trimers and actuate DNA unbinding, constituting $\mathsf{hsp}\colon\mathsf{hsf}$ complexes. As a consequence, hsf trimers production is hindered.

When the temperature rises, proteins start to misfold, driving heat shock proteins away from heat shock factors, breaking $\mathsf{hsp}\colon\mathsf{hsf}$ complexes. Free hsf's begin the process of trimerization, hence promoting hsp synthesis. The complete list of reactions is shown in Table 4.1

Table 4.1: The molecular model for the eukaryotic heat shock response proposed in [57]

Reaction	Description
$2\,\mathsf{hsf} \rightleftarrows \mathsf{hsf}_2$	Dimerization (1)
$\mathsf{hsf} + \mathsf{hsf}_2 \rightleftarrows \mathsf{hsf}_3$	Trimerization (2)
$\mathsf{hsf}_3 + \mathsf{hse} \rightleftarrows \mathsf{hsf}_3\colon\mathsf{hse}$	DNA binding (3)
$\mathsf{hsf}_3\colon\mathsf{hse} \to \mathsf{hsf}_3\colon\mathsf{hse} + \mathsf{hsp}$	hsp synthesis (4)
$\mathsf{hsp} + \mathsf{hsf} \rightleftarrows \mathsf{hsp}\colon\mathsf{hsf}$	hsf sequestration (5)
$\mathsf{hsp} + \mathsf{hsf}_2 \to \mathsf{hsp}\colon\mathsf{hsf} + \mathsf{hsf}$	Dimer dissipation (6)
$\mathsf{hsp} + \mathsf{hsf}_3 \to \mathsf{hsp}\colon\mathsf{hsf} + 2\,\mathsf{hsf}$	Trimer dissipation (7)
$\mathsf{hsp} + \mathsf{hsf}_3\colon\mathsf{hse} \to \mathsf{hsp}\colon\mathsf{hsf} + 2\,\mathsf{hsf} + \mathsf{hse}$	DNA unbinding (8)
$\mathsf{hsp} \to \emptyset$	hsp degradation (9)
$\mathsf{prot} \to \mathsf{mfp}$	Protein misfolding (10)
$\mathsf{hsp} + \mathsf{mfp} \rightleftarrows \mathsf{hsp}\colon\mathsf{mfp}$	mfp sequestration (11)
$\mathsf{hsp}\colon\mathsf{mfp} \to \mathsf{hsp} + \mathsf{prot}$	Protein refolding (12)

The molecular model in [57] includes the following three mass-conservation relations, for the total amount of hsf, the total amount of proteins (except for hsp and hsf) and for the total amount of hse:

- $[\mathsf{hsf}] + 2[\mathsf{hsf}_2] + 3[\mathsf{hsf}_3] + 3[\mathsf{hsf}_3\colon\mathsf{hse}] + [\mathsf{hsp}\colon\mathsf{hsf}] = C_1$,
- $[\mathsf{prot}] + [\mathsf{mfp}] + [\mathsf{hsp}\colon\mathsf{mfp}] = C_2$,
- $[\mathsf{hse}] + [\mathsf{hsf}_3\colon\mathsf{hse}] = C_3$,

where C_1, C_2 and C_3 are constants.

4.1.2 The basic mathematical model

The model in Table 4.1 was derived assuming all reactions follow the principle of mass action. This resulted into a system of ODEs. The rate for the reaction accounting for protein misfolding in Table 4.1 is given by the formula:

$$\varphi(T) = (1 - \frac{0.4}{e^{T-37}}) \cdot 1.4^{T-37} \cdot 1.45 \cdot 10^{-5} s^{-1},$$

where T is the ambient temperature, expressed in $^{\circ}C$, in accordance to [56]. Each species S in the molecular model is associated with a continuous, time-dependent function $[S](t)$, representing the concentration of the species respectively. The dynamics for the refined heat shock response model is described in Table 4.2.

Table 4.2: The system of ODE's associated with the basic model proposed in [57].

$$
\begin{aligned}
d[\mathsf{hsf}]/dt &= -2k_1^+[\mathsf{hsf}]^2 + 2k_1^-[\mathsf{hsf}_2] - k_2^+[\mathsf{hsf}][\mathsf{hsf}_2] + k_2^-[\mathsf{hsf}_3] \\
&\quad -k_5^+[\mathsf{hsf}][\mathsf{hsp}] + k_5^-[\mathsf{hsp{:}\ hsf}] + k_6[\mathsf{hsf}_2][\mathsf{hsp}] \\
&\quad +2k_7[\mathsf{hsf}_3][\mathsf{hsp}] + 2k_8[\mathsf{hsf_3{:}\ hse}][\mathsf{hsp}]; \\
d[\mathsf{hsf}_2]/dt &= k_1^+[\mathsf{hsf}]^2 - k_1^-[\mathsf{hsf}_2] - k_2^+[\mathsf{hsf}][\mathsf{hsf}_2] + k_2^-[\mathsf{hsf}_3] \\
&\quad -k_6[\mathsf{hsf}_2][\mathsf{hsp}]; \\
d[\mathsf{hsf}_3]/dt &= k_2^+[\mathsf{hsf}][\mathsf{hsf}_2] - k_2^-[\mathsf{hsf}_3] - k_3^+[\mathsf{hsf}_3][\mathsf{hse}] + k_3^-[\mathsf{hsf_3{:}\ hse}] \\
&\quad -k_7[\mathsf{hsf}_3][\mathsf{hsp}]; \\
d[\mathsf{hse}]/dt &= -k_3^+[\mathsf{hsf}_3][\mathsf{hse}] + k_3^-[\mathsf{hsf_3{:}\ hse}] + k_8[\mathsf{hsf_3{:}\ hse}][\mathsf{hsp}]; \\
d[\mathsf{hsf_3{:}\ hse}]/dt &= k_3^+[\mathsf{hsf}_3][\mathsf{hse}] - k_3^-[\mathsf{hsf_3{:}\ hse}] - k_8[\mathsf{hsf_3{:}\ hse}][\mathsf{hsp}]; \\
d[\mathsf{hsp}]/dt &= k_4[\mathsf{hsf_3{:}\ hse}] - k_5^+[\mathsf{hsf}][\mathsf{hsp}] + k_5^-[\mathsf{hsp{:}\ hsf}] - k_6[\mathsf{hsf}_2][\mathsf{hsp}] \\
&\quad -k_7[\mathsf{hsf}_3][\mathsf{hsp}] - k_8[\mathsf{hsf_3{:}\ hse}][\mathsf{hsp}] - k_{11}^+[\mathsf{hsp}][\mathsf{mfp}] \\
&\quad +(k_{11}^- + k_{12})[\mathsf{hsp{:}\ mfp}] - k_9[\mathsf{hsp}]; \\
d[\mathsf{hsp{:}\ hsf}]/dt &= k_5^+[\mathsf{hsf}][\mathsf{hsp}] - k_5^-[\mathsf{hsp{:}\ hsf}] + k_6[\mathsf{hsf}_2][\mathsf{hsp}] \\
&\quad +k_7[\mathsf{hsf}_3][\mathsf{hsp}] + k_8[\mathsf{hsf_3{:}\ hse}][\mathsf{hsp}]; \\
d[\mathsf{mfp}]/dt &= \varphi(T)[\mathsf{prot}] - k_{11}^+[\mathsf{hsp}][\mathsf{mfp}] + k_{11}^-[\mathsf{hsp{:}\ mfp}]; \\
d[\mathsf{hsp{:}\ mfp}]/dt &= k_{11}^+[\mathsf{hsp}][\mathsf{mfp}] - (k_{11}^- + k_{12})[\mathsf{hsp{:}\ mfp}]; \\
d[\mathsf{prot}]/dt &= -\varphi(T)[\mathsf{prot}] + k_{12}[\mathsf{hsp{:}\ mfp}].
\end{aligned}
$$

The kinetic rate constants and the initial values of reactants were estimated in [57] to satisfy the conditions below:

(i) the system is in a steady state at $37^{\circ}C$ since the model should not exhibit any response in the absence of the heat stress, i.e., at $37^{\circ}C$;

(ii) the numerical predictions for $[\mathsf{hsf_3{:}\ hse}](t)$ should confirm the experimental data from [47] at $42^{\circ}C$;

(iii) the numerical prediction of the model for $[\mathsf{hsp}](t)$ should be in accordance with the data in [57] at $42^{\circ}C$.

The numerical setup of the basic model can be found in [57]. The estimation of parameters relied on experimental data on DNA binding in HeLa cells at $42°C$ from [47]. The model should be in a steady-state at $37°C$. There are 17 independent parameters in the basic model and 10 initial conditions. Taking into account the 3 conservation relations shown in the previous section leads to only 7 initial conditions to be specified. That gives in total 17 independent parameters to be estimated. The model was fit against experimental data on DNA binding from [47]. For more details about the parameter estimation in COPASI ([36]) and the attained results for the parameters, we refer the reader to [57]. The prediction of the basic model for the heat shock response regarding the concentration of $\mathsf{hsf_3{:}\,hse}$ correlated to the experimental data on DNA binding from [47] is shown in Figure 4.1.

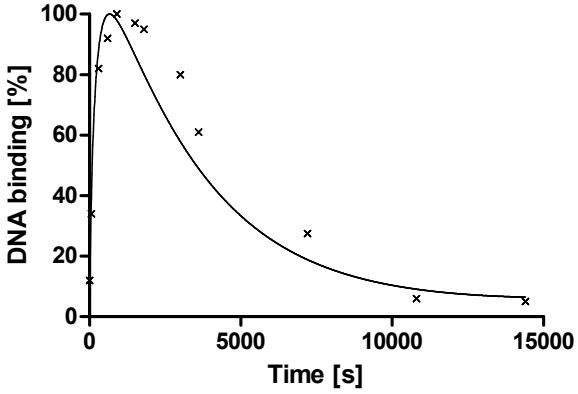


Figure 4.1: The dynamic behaviour of $\mathsf{hsf_3{:}\,hse}$ in the model exhibiting best fit. The continuous line represents the model prediction and the crossed points indicate the experimental data of [47]. The plot was originally published in [57].

Model validation and properties

The basic model for the heat shock response indicates a very low rate for protein misfolding at $37°C$ and a high rate for protein folding, in accordance with the findings of [4] and [42]. Furthermore, the model was able to predict a transient increase in the level of hsf trimers, in compliance with the results in [35].

A second validation test considered in [57] consisted in the application of the heat shock in two consecutive phases. The second heat shock was applied after a *recovery period*, when hsp's had reached a maximal level. In this scenario, the model predicted a considerably diminished response for

the second heat shock, which was to be expected given a high level of hsp as a consequence of the first heat shock. This confirmed the findings of [56].

A subsequent validation test consisted in applying a heat shock of $43°C$ and comparing the results with those reported in [60]. Contrary to the results in [60], the model in [57] predicted a prolonged transactivation for DNA binding. This prediction is also supported by experimental data from [1]. Moreover, removal of the heat shock at $42°C$, at the peak of the response, resulted into an accelerated attenuation phase, which is in accordance with the findings of [60].

Alternative numerical model fits and model identifiability

Model identifiability is one major problem a modeler witnesses in the process of model fitting. It refers to the uniqueness of the sets of parameters that bring about good model fits, all satisfying a set of required conditions. Various different sets of parameters might fit experimental data, but not satisfy diverse numerical properties of the model. In other words, different numerical setups for a given model could reside into fairly satisfying model fits, but discordant interpretations for the data under consideration. For the clarification of such difficult problems, the model requires supplementary data, especially referring to the conflicting behaviour between models [57].

We aimed at finding alternative numerical model fits for the basic heat shock response model in [57]. We repeated the parameter estimation procedure from the very beginning. We obtained different values for the parameters which fit well the experimental data. From among a few hundreds of model fits to experimental data, we selected only 29 that reached a satisfactory score after parameter estimation (where we considered a satisfactory score for parameter estimation in COPASI to be a score less than 100). These 29 models have been subjected to another validation test which consisted in checking that the initial values obtained after parameter estimation are steady states for the model at $37°C$. We have subsequently examined whether these 7 newly obtained model fits satisfy the other two validation tests: first, by applying the heat shock in two successive steps, and second, by applying a heat shock of $43°C$. Even though some of the aforementioned fits pass either of the tests, none of them pass both validation tests. Clearly, this does not prove the model in [57] is uniquely identifiable. However, it shows that fitting the model to the experimental data and obtaining a set of parameters that pass all validation tests is an arduous task.

4.1.3 The acetylation-refined reaction network model

The role of protein acetylation within the heat shock response

Heat shock factors (hsf's) play a crucial role in cell adaptation and survival, protecting the cell from protein damage originating in environmental and physiological stressors, for instance protein misfolding. Acetylation of human heat shock factors has been confirmed to have a pervasive significance within the attenuation of the heat shock response and the DNA binding activity, i.e. the higher the acetylation level is, the higher is the rate of DNA undbinding. For more details, we refer the reader to [65].

The process of protein acetylation consists in the substitution of an acetyl group for a hydrogen atom in a chemical compound. Acetylation can occur either as a co-translational modification at the α-amino group of the N-terminal (N-terminal acetylation) ([38]) or as a post-translational modification at the ϵ-amino group on lysine residues (lysine acetylation) ([14]).

N-terminal acetylation is an irreversible process. Despite extensive research invested in N-terminal acetylation in the past decades, its entire functional role is still not completely understood. Recent studies, however, suggest N-terminal acetylation plays the role of a degradation signal and acknowledge it as a central player in cell survival, see [2].

Lysine acetylation is a reversible post-translational modification for cellular regulation in eukaryotes, which plays a central role in the regulation of gene expression, by modifying histone tails through histone acetyltransferases or histone deacetylases, see [14] for more details.

The acetylation-refined model

The acetylation-refined model was developed so as to include the acetylation of heat shock factors. The refined model comprises two subtypes of heat shock factors: one subtype accounts for the acetylation of the lysine residue ($K80$) of hsf's and one subtype for the non-acetylated hsf's. Therefore, the $\mathsf{hsf_3}\colon \mathsf{hse}$ complex, for instance, will be refined into 4 subtypes according to the status of every hsf molecule, taking into account the symmetry of the molecule: $\mathsf{rhsf_3}\colon \mathsf{rhse}$, $\mathsf{rhsf_3}^{(1)}\colon \mathsf{rhse}$, $\mathsf{rhsf_3}^{(2)}\colon \mathsf{rhse}$, $\mathsf{rhsf_3}^{(3)}\colon \mathsf{rhse}$. The refined model relies on the following data refinements:

$$\mathsf{hsf} \to \{\mathsf{rhsf}, \mathsf{rhsf}^{(1)}\};$$
$$\mathsf{hsf_2} \to \{\mathsf{rhsf_2}, \mathsf{rhsf_2}^{(1)}, \mathsf{rhsf_2}^{(2)}\};$$
$$\mathsf{hsf_3} \to \{\mathsf{rhsf_3}, \mathsf{rhsf_3}^{(1)}, \mathsf{rhsf_3}^{(2)}, \mathsf{rhsf_3}^{(3)}\};$$
$$\mathsf{hsf_3}\colon \mathsf{hse} \to \{\mathsf{rhsf_3}\colon \mathsf{rhse}, \mathsf{rhsf_3}^{(1)}\colon \mathsf{rhse}, \mathsf{rhsf_3}^{(2)}\colon \mathsf{rhse}, \mathsf{rhsf_3}^{(3)}\colon \mathsf{rhse}\};$$
$$\mathsf{hsp}\colon \mathsf{hsf} \to \{\mathsf{rhsp}\colon \mathsf{rhsf}, \mathsf{rhsp}\colon \mathsf{rhsf}^{(1)}\}.$$

The refined model in [40] consists of 20 species, 55 irreversibile reactions and 55 kinetic parameters, compared to 10 species, 17 irreversible reactions and 17 kinetic parameters in the basic model of [57]. The complete list of reactions can be found in Table 4.3.

Table 4.3: The list of reactions for the refined model that includes the acetylation status of hsf. For an irreversible reaction q_i, r_i denotes its kinetic rate constant. For a reversible reaction q_i, r_i^+ and r_i^- denote the kinetic rate constants of its 'left-to-right' and 'right-to-left' directions, resp.

Reaction	Reaction number	Kinetic rate constants
$2\,\mathsf{rhsf} \leftrightarrows \mathsf{rhsf}_2$	$[q_1]$	r_1^+, r_1^-
$\mathsf{rhsf} + \mathsf{rhsf}^{(1)} \leftrightarrow \mathsf{rhsf}_2{}^{(1)}$	$[q_2]$	r_2^+, r_2^-
$2\mathsf{rhsf}^{(1)} \leftrightarrows \mathsf{rhsf}_2{}^{(2)}$	$[q_3]$	r_3^+, r_3^-
$\mathsf{rhsf} + \mathsf{rhsf}_2 \leftrightarrows \mathsf{rhsf}_3$	$[q_4]$	r_4^+, r_4^-
$\mathsf{rhsf}^{(1)} + \mathsf{rhsf}_2 \leftrightarrows \mathsf{rhsf}_3{}^{(1)}$	$[q_5]$	r_5^+, r_5^-
$\mathsf{rhsf} + \mathsf{rhsf}_2{}^{(1)} \leftrightarrows \mathsf{rhsf}_3{}^{(1)}$	$[q_6]$	r_6^+, r_6^-
$\mathsf{rhsf}^{(1)} + \mathsf{rhsf}_2{}^{(1)} \leftrightarrows \mathsf{rhsf}_3{}^{(2)}$	$[q_7]$	r_7^+, r_7^-
$\mathsf{rhsf} + \mathsf{rhsf}_2{}^{(2)} \leftrightarrows \mathsf{rhsf}_3{}^{(2)}$	$[q_8]$	r_8^+, r_8^-
$\mathsf{rhsf}^{(1)} + \mathsf{rhsf}_2{}^{(2)} \leftrightarrows \mathsf{rhsf}_3{}^{(3)}$	$[q_9]$	r_9^+, r_9^-
$\mathsf{rhsf}_3 + \mathsf{rhse} \leftrightarrows \mathsf{rhsf}_3\mathsf{:rhse}$	$[q_{10}]$	r_{10}^+, r_{10}^-
$\mathsf{rhsf}_3{}^{(1)} + \mathsf{rhse} \leftrightarrows \mathsf{rhsf}_3{}^{(1)}\mathsf{:rhse}$	$[q_{11}]$	r_{11}^+, r_{11}^-
$\mathsf{rhsf}_3{}^{(2)} + \mathsf{rhse} \leftrightarrows \mathsf{rhsf}_3{}^{(2)}\mathsf{:rhse}$	$[q_{12}]$	r_{12}^+, r_{12}^-
$\mathsf{rhsf}_3{}^{(3)} + \mathsf{rhse} \leftrightarrows \mathsf{rhsf}_3{}^{(3)}\mathsf{:rhse}$	$[q_{13}]$	r_{13}^+, r_{13}^-
$\mathsf{rhsf}_3\mathsf{:rhse} \rightarrow \mathsf{rhsf}_3\mathsf{:rhse} + \mathsf{rhsp}$	$[q_{14}]$	r_{14}
$\mathsf{rhsf}_3{}^{(1)}\mathsf{:rhse} \rightarrow \mathsf{rhsf}_3{}^{(1)}\mathsf{:rhse} + \mathsf{rhsp}$	$[q_{15}]$	r_{15}
$\mathsf{rhsf}_3{}^{(2)}\mathsf{:rhse} \rightarrow \mathsf{rhsf}_3{}^{(2)}\mathsf{:rhse} + \mathsf{rhsp}$	$[q_{16}]$	r_{16}
$\mathsf{rhsf}_3{}^{(3)}\mathsf{:rhse} \rightarrow \mathsf{rhsf}_3{}^{(3)}\mathsf{:rhse} + \mathsf{rhsp}$	$[q_{17}]$	r_{17}
$\mathsf{rhsp} + \mathsf{rhsf} \leftrightarrows \mathsf{rhsp:rhsf}$	$[q_{18}]$	r_{18}^+, r_{18}^-
$\mathsf{rhsp} + \mathsf{rhsf}^{(1)} \leftrightarrows \mathsf{rhsp:rhsf}^{(1)}$	$[q_{19}]$	r_{19}^+, r_{19}^-
$\mathsf{rhsp} + \mathsf{rhsf}_2 \rightarrow \mathsf{rhsp:rhsf} + \mathsf{rhsf}$	$[q_{20}]$	r_{20}
$\mathsf{rhsp} + \mathsf{rhsf}_2{}^{(1)} \rightarrow \mathsf{rhsp:rhsf} + \mathsf{rhsf}^{(1)}$	$[q_{21}]$	r_{21}
$\mathsf{rhsp} + \mathsf{rhsf}_2{}^{(1)} \rightarrow \mathsf{rhsp:rhsf}^{(1)} + \mathsf{rhsf}$	$[q_{22}]$	r_{22}
$\mathsf{rhsp} + \mathsf{rhsf}_2{}^{(2)} \rightarrow \mathsf{rhsp:rhsf}^{(1)} + \mathsf{rhsf}^{(1)}$	$[q_{23}]$	r_{23}
$\mathsf{rhsp} + \mathsf{rhsf}_3 \rightarrow \mathsf{rhsp:rhsf} + 2 * \mathsf{rhsf}$	$[q_{24}]$	r_{24}
$\mathsf{rhsp} + \mathsf{rhsf}_3{}^{(1)} \rightarrow \mathsf{rhsp:rhsf} + \mathsf{rhsf}^{(1)} + \mathsf{rhsf}$	$[q_{25}]$	r_{25}
$\mathsf{rhsp} + \mathsf{rhsf}_3{}^{(1)} \rightarrow \mathsf{rhsp:rhsf}^{(1)} + 2 * \mathsf{rhsf}$	$[q_{26}]$	r_{26}
$\mathsf{rhsp} + \mathsf{rhsf}_3{}^{(2)} \rightarrow \mathsf{rhsp:rhsf} + 2\mathsf{rhsf}^{(1)}$	$[q_{27}]$	r_{27}
$\mathsf{rhsp} + \mathsf{rhsf}_3{}^{(2)} \rightarrow \mathsf{rhsp:rhsf}^{(1)} + \mathsf{rhsf}^{(1)} + \mathsf{rhsf}$	$[q_{28}]$	r_{28}
$\mathsf{rhsp} + \mathsf{rhsf}_3{}^{(3)} \rightarrow \mathsf{rhsp:rhsf}^{(1)} + 2\mathsf{rhsf}^{(1)}$	$[q_{29}]$	r_{29}

Table 4.3: The list of reactions for the refined model - Continued

$\mathsf{rhsp} + \mathsf{rhsf_3}\colon \mathsf{rhse} \to \mathsf{rhsp}\colon \mathsf{rhsf} + 2\,\mathsf{rhsf} + \mathsf{rhse}$	$[q_{30}]$	r_{30}
$\mathsf{rhsp} + \mathsf{rhsf_3}^{(1)}\colon \mathsf{rhse} \to \mathsf{rhsp}\colon \mathsf{rhsf}^{(1)} + 2\,\mathsf{rhsf} + \mathsf{rhse}$	$[q_{31}]$	r_{31}
$\mathsf{rhsp} + \mathsf{rhsf_3}^{(1)}\colon \mathsf{rhse} \to \mathsf{rhsp}\colon \mathsf{rhsf} + \mathsf{rhsf}^{(1)} + \mathsf{rhsf} + \mathsf{rhse}$	$[q_{32}]$	r_{32}
$\mathsf{rhsp} + \mathsf{rhsf_3}^{(2)}\colon \mathsf{rhse} \to \mathsf{rhsp}\colon \mathsf{rhsf}^{(1)} + \mathsf{rhsf}^{(1)} + \mathsf{rhsf} + \mathsf{rhse}$	$[q_{33}]$	r_{33}
$\mathsf{rhsp} + \mathsf{rhsf_3}^{(2)}\colon \mathsf{rhse} \to \mathsf{rhsp}\colon \mathsf{rhsf} + 2\mathsf{rhsf}^{(1)} + \mathsf{rhse}$	$[q_{34}]$	r_{34}
$\mathsf{rhsp} + \mathsf{rhsf_3}^{(3)}\colon \mathsf{rhse} \to \mathsf{rhsp}\colon \mathsf{rhsf}^{(1)} + 2\mathsf{rhsf}^{(1)} + \mathsf{rhse}$	$[q_{35}]$	r_{35}
$\mathsf{rhsp} \to \emptyset$	$[q_{36}]$	r_{36}
$\mathsf{rprot} \to \mathsf{rmfp}$	$[q_{37}]$	r_{37}
$\mathsf{rhsp} + \mathsf{rmfp} \leftrightarrows \mathsf{rhsp}\colon \mathsf{rmfp}$	$[q_{38}]$	r_{38}^{+}, r_{38}^{-}
$\mathsf{rhsp}\colon \mathsf{rmfp} \to \mathsf{rhsp} + \mathsf{rprot}$	$[q_{39}]$	r_{39}

4.1.4 The acetylation-refined mathematical model

The refined heat shock response model has been attained through a procedure of data refinement, in such a way that the refinement conditions are met. The data refinement relations give the initial setup of the refined variables as follows:

$$[\mathsf{hsf}](0) = [\mathsf{rhsf}](0) + [\mathsf{rhsf}^{(1)}](0);$$
$$[\mathsf{hsf_2}](0) = [\mathsf{rhsf_2}](0) + [\mathsf{rhsf_2}^{(1)}] + [\mathsf{rhsf_2}^{(2)}](0);$$
$$[\mathsf{hsf_3}](0) = [\mathsf{rhsf_3}](0) + [\mathsf{rhsf_3}^{(1)}](0) + [\mathsf{rhsf_3}^{(2)}](0) + [\mathsf{rhsf_3}^{(3)}](0);$$
$$[\mathsf{hsp}\colon \mathsf{hsf}](0) = [\mathsf{hsp}\colon \mathsf{rhsf}](0) + [\mathsf{rhsp}\colon \mathsf{rhsf}^{(1)}](0);$$
$$[\mathsf{hsf_3}\colon \mathsf{hse}](0) = [\mathsf{rhsf_3}\colon \mathsf{rhse}](0) + [\mathsf{rhsf_3}^{(1)}\colon \mathsf{rhse}](0) + [\mathsf{rhsf_3}^{(2)}\colon \mathsf{rhse}](0) + + [\mathsf{rhsf_3}^{(3)}\colon \mathsf{rhse}](0);$$
$$[\mathsf{hsp}](0) = [\mathsf{rhsp}](0);$$
$$[\mathsf{hsp}\colon \mathsf{mfp}](0) = [\mathsf{rhsp}\colon \mathsf{rmfp}](0);$$
$$[\mathsf{mfp}](0) = [\mathsf{rmfp}](0);$$
$$[\mathsf{prot}](0) = [\mathsf{rprot}](0);$$
$$[\mathsf{hse}](0) = [\mathsf{rhse}](0).$$

The molecular model and the system of the mass-action based ODEs for the refined model is shown in [40]. The model refined as such consists of 20

species, 55 irreversibile reactions and 55 kinetic parameters.

As expected, the final model obeys the following three mass-conservation relations, which are the refinements of the mass-conservation relations satisfied by the basic model:

- $([\mathsf{rhsf}]+[\mathsf{rhsf}^{(1)}])+2\cdot([\mathsf{rhsf_2}]+[\mathsf{rhsf_2}^{(1)}]+[\mathsf{rhsf_2}^{(2)}])+3([\mathsf{rhsf_3}]+[\mathsf{rhsf_3}^{(1)}]+[\mathsf{rhsf_3}^{(2)}]+[\mathsf{rhsf_3}^{(3)}])+3\cdot([\mathsf{rhsf_3}\colon\mathsf{rhse}]+[\mathsf{rhsf_3}^{(1)}\colon\mathsf{rhse}]+[\mathsf{rhsf_3}^{(2)}\colon\mathsf{rhse}]+[\mathsf{rhsf_3}^{(3)}\colon\mathsf{rhse}])+([\mathsf{hsp}\colon\mathsf{rhsf}]+[\mathsf{rhsp}\colon\mathsf{rhsf}^{(1)}])=C_1$
- $[\mathsf{rprot}]+[\mathsf{rmfp}]+[\mathsf{rhsp}\colon\mathsf{rmfp}]=C_2$
- $([\mathsf{rhse}]+[\mathsf{rhsf_3}\colon\mathsf{rhse}])+([\mathsf{rhsf_3}^{(1)}\colon\mathsf{rhse}]+[\mathsf{rhsf_3}^{(2)}\colon\mathsf{rhse}]+[\mathsf{rhsf_3}^{(3)}\colon\mathsf{rhse}])=C_3$,

where C_1, C_2 and C_3 are the mass constants considered in the basic model.

The dynamics for the refined heat shock response molecular model is described in [40].

4.2 The ErbB signalling pathway

The ErbB signalling network is an evolutionary ancient pathway, which plays a major role in the regulation of diverse cellular responses such as survival, growth, differentiation, motility, etc. Moreover, aberrant ErbB signalling has been correlated with the progression of cancers in humans [6, 13, 61]. The ErbB signalling network involves multiple cellular ligands (e.g. EGF,HRG), the four receptor tyrosine kinases (ErbB1 (EGFR), ErbB2 (HER2 /NEU), ErbB3, ErbB4), cytoplasmic adaptors (Shc, Grb2, etc), scaffolds, enzymes, etc, see [6]. The signalling process is triggered by homo- and hetero-dimerization of the ligand-bound receptors, which brings about the auto-phosphorylation of the tyrosine residues in the cytoplasmic receptor domain. Signal activation leads to the recruitment of cytoplasmic adaptor, scaffold and enzymatic proteins to the membrane initiating a signalling cascade [6, 44]. Recruitment is succeeded by internalization and dissociation of signalling complexes, resulting in the activation of Ras-GTP through two pathways, one of which is Shc-dependent and the other Shc-independent. Ras-GTP plays the role of a switch which activates the MAPK cascade through the Raf, MEK and Erk-1/2 kinases, [6].

4.2.1 The initial ErbB signalling pathway model

The initial model, introduced in [37], is a reaction-based model of the EGFR(ErbB1)-signalling pathway and consists of 148 reactions, 103 reactants and 90 kinetic rate constants. The model is a revised version of models in [61] and [44]. The epidermal growth factor (EGF) binds to the epidermal

growth factor receptor EGFR (ErbB1) inducing the dimerization of EGFR and an accelerated activation of its tyrosine kinase, succeeded by the auto-phosphorylation of the tyrosine residues in the cytoplasmic receptor domain, see [44]. The auto-phosphorylation brings about the propagation of the signal through two different pathways. One of them commences with the binding of the Src protein to the dimerized, phosphorylated, ligand-bound receptor, succeeded by binding to the Grb2 protein. The other pathway involves an unmediated binding of the Grb2 protein. Both pathways involve the recruitment of Sos to the membrane through the association with the Grb2-bound EGFR containing complex. The model in [37] describes also an intricate process of internalization, as well as a detailed process of degradation of several complexes. Sos recruitment induces the association of Sos with Ras, which activates Ras as a consequence of Ras-GTP formation. Subsequently, Ras-GTP is inactivated as a result of a dissociation from the receptor complex involving protein GAP. The kinase reponsible for the phosphorylation of protein Raf is hitherto unindentified, however, the model in [37] assumes Raf is phosphorylated by a free Ras-GTP molecule. Phosphorylated Raf can phosphorylate the MEK kinase. Double-phosphorylated MEK can, in turn, phosphorylate ERK, see [37]. The initial model accounts for a negative feedback loop from dual phosphorylated ERK kinase to protein Sos, which induces unbinding of Grb2-Sos from the receptor complex. For more details, we refer the reader to [12] and [20]. Protein isoform specificity is not accounted for in the model. The absence of stimulus EGF brings about a stable steady state for the initial model of [37]. The model identifies two pools of doubly phosphorylated ERK (ERK-PP), one of which is located in the cytoplasm and one related to the internalized receptor. The model depicted in [37] comprises 13 chemical processes: EGFR activation, Shc, Grb2, Sos recruitment, Ras activation and inactivation, Raf activation by phosphorylation and its dephosphorylation, phosphorylation and dephosphorylation of MEK, ERK dephosphorylation, negative feedback from ERK to Sos, internalization of receptor complexes and degradations reaction. For a more detailed description of the initial model, we refer the reader to [37]. In this section, we focus on expanding the EGFR signalling pathway model from [37] through fit-preserving data refinement. To this end, we consider all four members of the ErbB family, namely ErbB1-4 receptors: ErbB1 (EGFR), ErbB2 (HER2), ErbB3, ErbB4, along with two ligands: EGF and HRG.

4.2.2 The refined ErbB signalling pathway model

We consider two classes of species within our refined model: *atomic* or *complex*. By definition, see [28], atomic species are self-contained, i.e. their structure is autonomous, not reliant on any other species. The structure of a complex species inherently consists of at least two bound atomic species.

All species mentioned above, the four members of the ErbB family, as well as the two ligands are atomic species. No other atomic species in the model presented in [37] is refined. However, all complex species in [37] comprising ErbB1 (EGFR) and/or EGF are refined to include all four members of the ErbB family and both types of ligands: EGF and HRG, as well as all dimer and receptor-ligand binding combinations. Formally, the refined model comprises the data refinements below:

$$\begin{aligned}&\mathsf{EGFR} \rightarrow \{\mathsf{ErbB1}, \mathsf{ErbB2}, \mathsf{ErbB3}, \mathsf{ErbB4}\};\\ &\mathsf{EGF} \rightarrow \{\mathsf{EGF}, \mathsf{HRG}\}.\end{aligned}$$

Signalling is activated through binding of a ligand (in our refined model: EGF or HRG) to a receptor (in our refined model: ErbB1, ErbB2, ErbB3, ErbB4). The process of receptor activation is initiated by the following reaction:

$$\mathsf{EGF} + \mathsf{EGFR} \underset{k^-_{lb}}{\overset{k^+_{lb}}{\longleftrightarrow}} \mathsf{EGF\text{-}EGFR}, \tag{4.1}$$

which is refined to include both ligands ($L_i, i = \overline{1,2}$)and all four receptors ($R_j, j = \overline{1,4}$) to the following reactions:

$$L_i + R_j \underset{k^-_{i,j}}{\overset{k^+_{i,j}}{\longleftrightarrow}} L_i - R_j, \text{ for all } i = \overline{1,2}, j = \overline{1,4}, \tag{4.2}$$

where $L_i \in \{\mathsf{EGF}, \mathsf{HRG}\}$ and $R_j \in \{\mathsf{ErbB1}$,$\mathsf{ErbB2}$, $\mathsf{ErbB3}$, $\mathsf{ErbB4}\}$.

Our aim is to set the kinetic rate constants for the refined model so that the sufficient conditions (3.12) for our model to be a fit-preserving refinement of the model in [37] are fulfilled. For instance, given the ligand-binding reaction (4.2), the kinetic rate constants for the refined model, conforming to (3.12) are the following: $k^-_{i,j} = k^-_{lb}$ and $k^+_{i,j} = k^+_{lb}$, for all $i = \overline{1,2}$, $j = \overline{1,4}$.

The reaction of ligand-binding of the receptor activation process precedes a dimerization reaction, whose products are in turn involved in an extensive number of reactions down the signalling pathway. Given the dimerization reaction

$$2\,\mathsf{EGF\text{-}EGFR} \underset{k^-_d}{\overset{k^+_d}{\longleftrightarrow}} (\mathsf{EGF\text{-}EGFR})^2,$$

in the basic model of [37]), the refinement takes into account all combinations of monomers on the *left-hand side* of reactions, thus generating a number of 64 reactions (since we have 8 types of monomers) to account for

dimerization in the refined model. The dimerization reaction is refined into the following reactions:

$$(\mathsf{L\text{-}R})_i + (\mathsf{L\text{-}R})_j \underset{k^-_{i,j,l}}{\overset{k^+_{i,j,l}}{\longleftrightarrow}} (\mathsf{L\text{-}R})_l, \tag{4.3}$$

where $j, l = \overline{1,8}$ and $1 \leq i \leq j$, $\{(\mathsf{L\text{-}R})_m \,|m \in \{i,j,l\}\} \in \{\mathsf{EGF-ErbBp}, \mathsf{HRG-ErbBq}|p, q = \overline{1,4}\}$ represent the ligand-bound receptors in the refined model. We take into consideration only homo-dimers.

For the dimerization reaction, whose refined version is detailed in 4.3, we set the kinetic rate constants conforming to (3.12) as follows:

$$k^+_{i,j,l} = \begin{cases} 0, & \text{if } l \neq i; j \\ k^+_d, & \text{otherwise.} \end{cases} \tag{4.4}$$

$$k^-_{i,j,l} = \begin{cases} 0, & \text{if } l \neq i; j \\ \frac{k^-_d}{8}, & \text{otherwise.} \end{cases} \tag{4.5}$$

The kinetic rate constants set in 4.4 and 4.5 are set so as to comply with the following relations, obtained based on (3.12). We chose a symmetrical solution for the kinetic rate constants, so that no refined reaction is favoured. We list the relations below:

$$\sum_{i=1}^{8} k^+_{i,j,l} = \begin{cases} k^+_d, & \text{if } i = j \\ 2k^+_d, & \text{if } i < j \end{cases} \tag{4.6}$$

$$\sum_{\substack{\{i,j\} \\ 1 \leq i \leq j \leq 8}} = k^-_d. \tag{4.7}$$

The receptor activation process comprises a receptor production reaction, which we indicate below:

$$\xrightarrow{k_p} \mathsf{EGFR} \tag{4.8}$$

This receptor production reaction is refined into four corresponding reactions:

$$\xrightarrow{k_i} R_i, i = \overline{1,4}, \tag{4.9}$$

with $R_i \in \{\mathsf{ErbB1}, \mathsf{ErbB2}, \mathsf{ErbB3}, \mathsf{ErbB4}\}$. The kinetic rate constants for the receptor production reaction are set, in accordance to (3.12), as follows:

$$k_i = \frac{k_p}{4}. \tag{4.10}$$

All reactions involving reactants which contain an EGF-EGFR dimer in their structure or its internalized counterpart refine into 8 different reactions, each of which involving one of the types of dimer-derived refined complex.

Complex species in the basic model of [37], comprising species EGFR and/or EGF, are refined to include all receptor-ligand binding combinations. For instance, given species $(\mathsf{EGF\text{-}EGFR^*})^2\text{-}\mathsf{AC}$, where AC stands for a chain of bound atomic species (e.g. GAP-Grb2-Sos-Ras-GDP-Prot), species $(\mathsf{EGF\text{-}EGFR^*})^2\text{-}\mathsf{AC}$ is refined into the following subspecies:

$$(\mathsf{EGF\text{-}EGFR^*})^2\text{-}\mathsf{AC} \rightarrow \{((\mathsf{L\text{-}R})_i{}^*)^2 - AC\}, \tag{4.11}$$

where $i = \overline{1,8}$ and $(\mathsf{L\text{-}R})_i \in \{\mathsf{EGF-ErbBp}, \mathsf{HRG-ErbBq}\,|p, q = \overline{1,4}\}$ and the "$*$" character illustrates the phosphorylation status of the respective ErbB molecule. Species refinement in turn reflects into the refinement of all reactions which involve species to be refined. In the examples above, all reactions involving species that comprise EGF and/or EGFR were refined subsequently.

The basic model in [37] consists of a number of 103 species and 148 reactions, while the refined model comprises a number of 421 species involved in 928 reactions. We refined the reactions of the basic model in [37] following the approach shown in Section 3.2.2. The refined model can be found in [39].

The numerical setup of the refined model

We address, in this section, the numerical setup of the refined model for the ErbB signalling pathway in [37]. We determine the numerical values for the refined model so that relation(3.11) is satisfied. The initial values for the refined variables are set conforming to the data refinement relations, the sum of the concentrations all subspecies of a given species is identical to the concentration of the species itself. For example, given the complex species introduced as above, the setup of the kinetic initial values:

$$[(\mathsf{EGF\text{-}EGFR^*})^2\text{-}\mathsf{AC}](0) = \sum_{i=1}^{8}[((\mathsf{L\text{-}R})_i{}^*)^2 - AC](0),$$

with $(\mathsf{L\text{-}R})_i \in \{\mathsf{EGF-ErbBp}, \mathsf{HRG-ErbBq}\,|p, q = \overline{1,4}\}$ and the "$*$" character illustrates the phosphorylation status of the respective ErbB molecule.

The basic model in [37] consists of a number of 103 species and 148 reactions, while the refined model comprises a number of 421 species involved in 928 reactions. We refined the reactions of the basic model in [37] using the fit-preserving refinement method introduced in Section 3.2.2 and we obtained a model consisting of over 900 reactions, which preserves model fit by construction. This methodology, however, is not only applicable to this case study in particular, but it extends to other reaction-based mass-action

models as a pilar in the model construction cycle on which other methods can rely, for instance parameter estimation.

Chapter 5

Original research contribution

We list, in this chapter, the original contribution of each paper included in this thesis as follows:

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

    - We introduced the concepts of (quantitative) data refinement and process refinement of a reaction-based biomodel;
    - The heat shock response model, introduced in [57], was refined to include the role of protein acetylation of the heat shock factor. The first stage of the refinement consisted in data refinement, which produced a substantial increase in the number of kinetic parameters and variables, but taking into account our methodology, all numerical properties of the model were preserved with a minimal computational effort.
    - A second stage of the refinement consisted in process refinement, we replaced in the data-refined model the reactions corresponding to hsp-mediated unbinding of trimers with a set of reactions accounting for the unbinding of trimers in the absence of hsp. In the latter case, most of the parameters were fixed in the previous stage of data refinement, the only kinetic rates left to be estimated were for the newly introduced reactions. In order to estimate these parameters, we fitted the model to experimental data on DNA binding and gene transcription levels.

2. Diana-Elena Gratie, Bogdan Iancu and Ion Petre. ODE analysis of biological systems. In: *Formal Methods for Dynamical Systems*, Edi-

tors: Marco Bernardo, Erik de Vink, Alessandra di Pierro and Herbert Wiklicky, LNCS 7938, pages 29–62, Springer, 2013.

- We discussed basic basic computational, numerical techniques for used in modelling reaction-based models. We discussed basic notions for model building : reaction-based models, ODE-based models, kinetic laws, parameter estimation, model identifiability model refinement.
- We reviewed a few analysis techniques for the computational analysis of ODE-based models: steady state analysis, mass-conservation relations, sensitivity analysis. We considered as a case-study a model for the heat shock response in eukaryotes.

3. Diana-Elena Gratie, Bogdan Iancu, Sepinoud Azimi and Ion Petre. Quantitative model refinement in four different frameworks, with applications to the heat shock response. In: *From Action System to Distributed Systems*, Taylor & Francis Group, Editors: Luigia Petre, Emil Sekerinski, *Accepted, To appear*, 2015.

   - We investigated and discussed the implementation of quantitative model refinement in four different frameworks: reaction-based models, rule-based modelling, Petri nets and guarded command languages.
   - We used as a case study a model for the eukaryotic heat shock response, which we refined in all the above frameworks to include the acetylation of the heat shock factor.
   - We discussed the the implementation of the refinement so as to avoid the combinatorial explosion of the refined model as a consequence of introducing a post-translational modification, namely the acetylation.
   - We discussed the versatility of quantitative model refinement and concluded that it is a cross-platform framework, emphasizing its implementability under different modelling paradigms.

4. Bogdan Iancu, Cristian Gratie and Ion Petre. Refinement-based modelling of the ErbB signalling pathway – Extended abstract. In: Annals of University of Bucharest LXI, pages 7-14, The Bucharest University Press, 2014.

   - We discussed here the advantages of fit-preserving data refinement in the cycle of model building.
   - The paper focused on the implementation of fit-preserving data refinement on a model for the ErbB signalling pathway, extended

to include more types of receptor and ligands. The refined model ensured a good fit by construction.

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

    - In this paper, we discussed the aspect that reactions systems can capture the fundamental characteristics of an ODE-based model.
    - We built a reaction system model of the heat shock response, whose qualitative behaviour captures the essential characteristic of the corresponding ODE-based model. To this end, we introduced a new concept, the dominance graph, to express the competition on resources of the ODE-based model.
    - The paper concluded drawing a comparison between the expressivity of reactions systems and that of ODE-based models.

Chapter 6

Conclusions and future perspectives

Contemporary biology produces massive amounts of data, whose interpretation and analysis necessitate computational tools. Computational systems biology has a holistic perspective as opposed to traditional biology: we can only aim to understand the intricate mechanisms behind highly complex diseases, such as cancer, neurodegenerative diseases, etc., by looking at a biological phenomenon in its entirety. A hierarchical system-level representation of the mechanisms involved turns out to be essential in the effort of understanding complex biological systems.

The research presented in this thesis aims to tackle one fundamental problem that systems biology has been challenged by, that of modelling at different levels of resolution. Formal methods proved to be essential in a system-level representation of biological phenomena. To this end, this thesis investigates a sound computational framework for quantitative model refinement, addressing a major challenge computational systems biology faces: given a model, fit and validated at a certain level of resolution, how can one variate its level of resolution, essentially increasing the level of detail, while ensuring its fit is preserved? We exploit our framework in the refinement of two biomodels depicting two biochemical networks, expressed as reaction-based networks, which have been of great interest in the past decade: the heat shock response in eukaryotes and the ErbB signalling pathway. However, the applicability of the quantitative refinement framework is not limited to the aforementioned case studies, but it extends to the refinement of other biological mechanisms and beyond.

In this thesis, we address issues referring to a few formalisms in computational systems biology, quantitative in nature: reaction-network models, rule-based models and Petri net models, along with a novel formalism, qualitative in nature, reactions systems. We review the major advantages and

challenges that are associated with each of the aforementioned formalisms. The decision regarding the choice of a particular modelling framework depends on the problem at hand. But the massive amount of data produced lately converges to an eventual integration of different frameworks across scales and, moreover, moves towards compiling models across various biological networks: signalling, regulatory, metabolic, etc. The completion of such an integration would benefit greatly from the use of quantitative model refinement as a technique for model expansion.

The thesis revolves around the concept of quantitative model refinement, providing first an overview of refinement in software engineering and previous work in computational systems biology, in connection with rule-based modelling and Petri nets. We subsequently derive a framework for the quantitative refinement of reaction-based biomodels, consistent with the law of mass-action, referring to the refinement of a species to include more details about its subspecies and also refinement of reactions. Consequently, we discuss the extension of the framework to distinguish between atomic and complex species. Ultimately, we present an extension of the quantitative refinement framework, called fit-preserving refinement, which builds upon the quantitative refinement framework to provide a methodology, which takes into account all possible combinations of substrates and products a reaction can be refined into. We chose two case studies that have been widely analyzed in the past decades, but whose intricate mechanisms are not yet fully understood: the heat shock response and the ErbB signalling pathway. For the heat shock response model, we started with an initial model of 17 irreversible reactions and we have refined it to include the acetylation of the heat shock factor at its $K80$ residue, resulting into a model of 54 irreversible reactions and subsequently performed process refinement to reflect the unbinding of heat shock factor in a trimer conformation in the absence of heat shock proteins. We performed another refinement on a model of the ErbB signalling pathway to include four types of receptors and two type of ligands, starting with a model of 148 reactions and compiling a refined model of 928 reactions. The refinement proved to be crucial in the expansion of the aforementioned models and setting the kinetic rate constants of the refined model to assure model fit preservation avoiding performing parameter estimation, which is computationally intensive and time consuming.

Quantitative model refinement is essential in compiling large scale models, where the transition from a level of resolution to another is topical in certain circumstances, the methodology providing a solid foundation for the initialization of kinetic rate constants, aiming at parameter estimation on a significantly reduced space of parameters. The scalability of quantitative model refinement does not seem to extend to genome scale models, however, it can be used for the the development of large-scale modules, which are integrated within larger hierarchical structures, allowing a transfer among

different levels of resolution. There are also other aspects that can be taken into account: for instance, since the solution for choosing the parameters of the refined model is not unique, we can examine whether, given a different setup for the kinetic rate constants of the refined model, there are other models which preserve model fit withal, but simultaneously exhibit other properties worthwhile to inspect. Another aspect that is notable to mention is that the current framework is limited to mass-action models, an extension to other kinetic laws and, also, integrating models over several different kinetic laws would be beneficial to compile large-scale models, assumably organized across different modules.

Bibliography

[1] K. Abravaya, B. Phillips, and R.I. Morimoto. Attenuation of the heat shock response in HeLa cells is mediated by the release of bound heat shock transcription factor and is modulated by changes in growth and in heat shock temperatures. *Genes & development*, 5(11):2117–2127, 1991.

[2] T. Arnesen. Towards a functional understanding of protein N-terminal acetylation. *PLoS Biology*, 9(5):e1001074, 2011.

[3] R.J. Back and J. von Wright. *Refinement Calculus*. Springer (New York), 1998.

[4] R.M. Ballew, J. Sabelko, and M. Gruebele. Direct observation of fast protein folding: the initial collapse of apomyoglobin. *Proceedings of the National Academy of Sciences*, 93(12):5759–5764, 1996.

[5] A.L. Barabasi and Z. N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004.

[6] M.R. Birtwistle, M. Hatakeyama, N. Yumoto, B.A. Ogunnaike, J.B. Hoek, and B.N. Kholodenko. Ligand-dependent responses of the ErbB signaling network: experimental and modeling analyses. *Molecular systems biology*, 3(1), 2007.

[7] M. Blinov, J. Yang, J. Faeder, and W. Hlavacek. Graph theory for rule-based modeling of biochemical networks. *Transactions on Computational Systems Biology VII*, pages 89–106, 2006.

[8] M.L. Blinov, J.R. Faeder, B. Goldstein, and W.S. Hlavacek. Bionetgen: software for rule-based modeling of signal transduction based on the interactions of molecular domains. *Bioinformatics*, 20(17):3289–3291, 2004.

[9] G. Booch, R. Maksimchuk, M. Engle, B. Young, J. Conallen, and K. Houston. *Object-oriented Analysis and Design with Applications, Third Edition*. Addison-Wesley Professional, third edition, 2007.

[10] G.E. Briggs and J.B.S. Haldane. A note on the kinetics of enzyme action. *Biochemical Journal*, 19(2):338–339, 1925.

[11] R. Brijder, A. Ehrenfeucht, M. Main, and G. Rozenberg. A tour of reaction systems. *International Journal of Foundations of Computer Science*, 22(07):1499–1517, 2011.

[12] L. Buday, P.H. Warne, and J. Downward. Downregulation of the Ras activation pathway by MAP kinase phosphorylation of Sos. *Oncogene*, 11(7):1327–1331, 1995.

[13] W.W. Chen, B. Schoeberl, P.J. Jasper, M. Niepel, U.B. Nielsen, D.A. Lauffenburger, and P. Sorger. Input–output behavior of ErbB signaling pathways as revealed by a mass action model trained against dynamic data. *Molecular systems biology*, 5(1), 2009.

[14] C. Choudhary, C. Kumar, F. Gnad, M.L. Nielsen, M. Rehman, T.C. Walther, J.V. Olsen, and M. Mann. Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science*, 325(5942):834–840, 2009.

[15] V. Danos, J. Feret, W. Fontana, R. Harmer, and J. Krivine. Rule-based modelling of cellular signalling. In *CONCUR 2007–Concurrency Theory*, pages 17–41. Springer, 2007.

[16] V. Danos, J. Feret, W. Fontana, R. Harmer, and J. Krivine. Rule-based modelling, symmetries, refinements. In *Formal Methods in Systems Biology*, pages 103–122. Springer, 2008.

[17] V. Danos, J. Feret, W. Fontana, R. Harmer, and J. Krivine. Rule-based modelling and model perturbation. In *Transactions on Computational Systems Biology XI*, pages 116–137. Springer, 2009.

[18] V. Danos and C. Laneve. Graphs for core molecular biology. In *Computational Methods in Systems Biology*, pages 34–46. Springer, 2003.

[19] R. David and H. Alla. *Discrete, continuous, and hybrid Petri nets*. Springer, 2010.

[20] C. Dong, S.B. Waters, K.H. Holt, and J.E. Pessin. Sos phosphorylation and disassociation of the Grb2-SOS complex by the ERK and JNK signaling pathways. *Journal of Biological Chemistry*, 271(11):6328–6332, 1996.

[21] A. Ehrenfeucht and G. Rozenberg. Reaction systems. *Fundamenta Informaticae*, 75(1):263–280, 2007.

[22] J.R. Faeder, M.L. Blinov, B. Goldstein, and W.S. Hlavacek. Rule-based modeling of biochemical networks. *Complexity*, 10(4):22–41, 2005.

[23] J.R. Faeder, M.L. Blinov, and W.S. Hlavacek. Graphical rule-based representation of signal-transduction networks. In *Proceedings of the 2005 ACM symposium on Applied computing*, pages 133–140. ACM, 2005.

[24] J.R. Faeder, M.L. Blinov, and W.S. Hlavacek. Rule-based modeling of biochemical systems with BioNetGen. In *Systems Biology*, pages 113–167. Springer, 2009.

[25] J. Feret, J. Krivine, and R. Harmer. Rule based modeling and model refinement. *Elements of Computational Systems Biology*, 8:83, 2010.

[26] I.M. Gessel and R.P. Stanley. Algebraic enumeration. *Handbook of combinatorics*, 2:1021–1061, 1995.

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

[28] D.-E. Gratie, B. Iancu, S. Azimi, and I. Petre. Quantitative model refinement in four different frameworks, with applications to the heat shock response. Technical Report 1067, TUCS, 2013.

[29] C.M. Guldberg and P. Waage. Studies concerning affinity. *CM Forhandlinger: Videnskabs-Selskabet i Christiana*, 35:1864, 1864.

[30] C.M. Guldberg and P. Waage. Concerning chemical affinity. *Erdmanns Journal für practische Chemie*, 127:69–114, 1879.

[31] J.B.S. Haldane. Enzymes. 1930, 1965.

[32] R. Harmer. Rule-based modelling and tunable resolution. *EPTCS*, 9:65–72, 2009.

[33] M. Heiner, D. Gilbert, and R. Donaldson. Petri nets for systems and synthetic biology. In *Formal methods for computational systems biology*, pages 215–264. Springer, 2008.

[34] R. Heinrich and S. Schuster. *The regulation of cellular systems*, volume 416. Chapman & Hall New York, 1996.

[35] C.I. Holmberg, S.E.F. Tran, J.E. Eriksson, and L. Sistonen. Multisite phosphorylation provides sophisticated regulation of transcription factors. *Trends in biochemical sciences*, 27(12):619–627, 2002.

[36] S. Hoops, S. Sahle, R. Gauges, C. Lee, J. Pahle, N. Simus, M. Singhal, L. Xu, P. Mendes, and U. Kummer. Copasi - a complex pathway simulator. *Bioinformatics*, 22(24):3067–3074, 2006.

[37] J.J. Hornberg, B. Binder, F.J. Bruggeman, B. Schoeberl, R. Heinrich, and H.V. Westerhoff. Control of MAPK signalling: from complexity to what really matters. *Oncogene*, 24(36):5533–5542, 2005.

[38] C.S. Hwang, A. Shemorry, and A. Varshavsky. N-terminal acetylation of cellular proteins creates specific degradation signals. *Science*, 327(5968):973–977, 2010.

[39] B. Iancu. The Copasi implementation of the fit-preserving refinement of the ErbB signalling pathway. `http://combio.abo.fi/research/quantitative-model-refinement/refinement-of-the-egfr-signalling-pathway/`, 2013.

[40] B. Iancu, El. Czeizler, Eu. Czeizler, and I. Petre. Quantitative refinement of reaction models. *International Journal of Unconventional Computing*, 8(5-6):529–550, 2012.

[41] R. A. Jackson. *Mechanisms in organic reactions*, volume 23. Royal Society of Chemistry, 2004.

[42] C.M. Jones, E.R. Henry, Y. Hu, C.K. Chan, S.D. Luck, A. Bhuyan, H. Roder, J. Hofrichter, and W.A. Eaton. Fast events in protein folding initiated by nanosecond laser photolysis. *Proceedings of the National Academy of Sciences*, 90(24):11860–11864, 1993.

[43] J.R. Karr, JayoditaC. Sanghvi, D.N. Macklin, M.V. Gutschow, J.M. Jacobs, B. Bolival Jr., N. Assad-Garcia, J.I. Glass, and M.W. Covert. A whole-cell computational model predicts phenotype from genotype. *Cell*, 150(2):389 – 401, 2012.

[44] B.N. Kholodenko, O.V. Demin, G. Moehren, and J.B. Hoek. Quantification of short term signaling by the epidermal growth factor receptor. *Journal of Biological Chemistry*, 274(42):30169–30181, 1999.

[45] H. Kitano. Computational systems biology. *Nature*, 420(6912):206–210, 2002.

[46] H. Kitano. Systems biology: a brief overview. *Science*, 295(5560):1662–1664, 2002.

[47] M.P. Kline and R.I. Morimoto. Repression of the heat shock factor 1 transcriptional activation domain is modulated by constitutive phosphorylation. *Molecular and cellular biology*, 17(4):2107–2115, 1997.

[48] E. Klipp, R. Herwig, A. Kowald, C. Wierling, and H. Lehrach. *Systems biology in practice: concepts, implementation and application*. Wiley-Vch, 2005.

[49] E. Klipp, R. Herwig, A. Kowald, C. Wierling, and H. Lehrach. *Systems biology in practice: concepts, implementation and application*. John Wiley & Sons, 2008.

[50] I. Koch, W. Reisig, and F. Schreiber. *Modeling in systems biology: the Petri Net approach*. Springer-Verlag New York, Inc., 2010.

[51] Y. Lazebnik. Can a biologist fix a radio? – or, what I learned while studying apoptosis. *Cancer cell*, 2(3):179–182, 2002.

[52] L. Michaelis and M. L. Menten. Die kinetik der invertinwirkung. *Biochem. Z.*, 49(333-369):352, 1913.

[53] W.G. Miller and R.A. Alberty. Kinetics of the reversible Michaelis-Menten mechanism and the applicability of the steady-state approximation. *Journal of the American Chemical Society*, 80(19):5146–5151, 1958.

[54] D. L. Nelson and M. M. Cox. Lehninger principles of biochemistry. *Worth Publishers*, 2000.

[55] D.L. Nelson, A.L. Lehninger, and M.M. Cox. *Lehninger principles of biochemistry*. Macmillan, 2008.

[56] A. Peper, C.A. Grimbergen, J.A.E. Spaan, J.E.M. Souren, and R. Van Wijk. A mathematical model of the hsp70 regulation in the cell. *International Journal of Hyperthermia*, 14(1):97–124, 1998.

[57] I. Petre, A. Mizera, C.L. Hyder, A. Meinander, A. Mikhailov, R.I. Morimoto, L. Sistonen, J.E. Eriksson, and R.J. Back. A simple mass-action model for the eukaryotic heat shock response and its mathematical validation. *Natural Computing*, 10(1):595–612, 2011.

[58] C. A. Petri. *Kommunikation mit Automaten*. PhD thesis, Bonn: Institut für Instrumentelle Mathematik, 1962.

[59] K. Raman and N. Chandra. Systems biology. *Resonance*, 15(2):131–153, 2010.

[60] T.R. Rieger, R.I. Morimoto, and V. Hatzimanikatis. Mathematical modeling of the eukaryotic heat-shock response: Dynamics of the hsp70 promoter. *Biophysical journal*, 88(3):1646–1658, 2005.

[61] B. Schoeberl, C. Eichler-Jonsson, E.D. Gilles, and G. Müller. Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. *Nature biotechnology*, 20(4):370–375, 2002.

[62] J.A.P. Sekar and J.R. Faeder. Rule-based modeling of signal transduction: A primer. In *Computational Modeling of Signaling Networks*, pages 139–218. Springer, 2012.

[63] Y. Shi, D.D. Mosser, and R.I. Morimoto. Molecular chaperones as HSF1-specific transcriptional repressors. *Genes & development*, 12(5):654–666, 1998.

[64] N. J. Tro. *Chemistry: a molecular approach*. Prentice Hall, 2010.

[65] S.D. Westerheide, J. Anckar, S.M. Stevens, L. Sistonen, and R.I. Morimoto. Stress-inducible regulation of heat shock factor 1 by the deacetylase SIRT1. *Science*, 323(5917):1063–1066, 2009.

[66] H. V. Westerhoff and B. O. Palsson. The evolution of molecular biology into systems biology. *Nature Biotechnology*, 22(10):1249–1252, 2004.

[67] O. Wolkenhauer and M. Mesarović. Feedback dynamics and cell function: why systems biology is called systems biology. *Molecular BioSystems*, 1(1):14–16, 2005.

[68] Z. Zi. Sensitivity analysis approaches applied to systems biology models. *Systems Biology, IET*, 5(6):336 –346, 2011.

Publication I

Quantitative refinement of reaction models

Bogdan Iancu, Elena Czeizler, Eugen Czeizler and Ion Petre

Originally published in: *International Journal of Unconventional Computing*, 8(5-6), 529–550, 2012.

A preliminary shorter version of this article was published in:

Elena Czeizler, Eugen Czeizler, Bogdan Iancu and Ion Petre, *Quantitative model refinement as a solution to the combinatorial size explosion of biomodels*, Electronic Notes in Theoretical Computer Science, 284: 35-53, 2012, Proceedings of the 2nd International Workshop on Static Analysis and Systems Biology (SASB 2011).

Int. Journ. of Unconventional Computing, Vol. 8, pp. 529–550
Reprints available directly from the publisher
Photocopying permitted by license only

Quantitative Refinement of Reaction Models

BOGDAN IANCU[1], ELENA CZEIZLER[2], EUGEN CZEIZLER[3] AND ION PETRE[1]

[1]*Department of Information Technologies, Åbo Akademi University, FI-20520 Turku, Finland*
E-mail: Bogdan.Iancu@abo.fi, Ion.Petre@abo.fi
[2]*Department of Computer Science and Engineering, Aalto University School of Science, FI-00076 Aalto, Finland*
E-mail: Elena.Czeizler@aalto.fi
[3]*Department of Information and Computer Science, Aalto University School of Science, FI-00076 Aalto, Finland*
E-mail: Eugen.Czeizler@aalto.fi

Received: August 8, 2012. Accepted: October 22, 2012.

One approach to modelling complex biological systems is to start from an abstract representation of the biological process and then to incorporate more details regarding its reactions or reactants through an iterative refinement process. The refinement should be done so as to ensure the preservation of the numerical properties of the model, such as its numerical fit and validation. Such approaches are well established in software engineering: starting from a formal specification of the system, one refines it step-by-step towards an implementation that is guaranteed to satisfy a number of logical properties. We introduce here the concepts of (quantitative) *data refinement* and *process refinement* of a biomolecular, reaction-based model. We choose as a case study a recently proposed model for the heat shock response and refine it to include some details of its acetylation-induced control. Although the refinement process produces a substantial increase in the number of kinetic parameters and variables, the methodology we propose preserves all the numerical properties of the model with a minimal computational effort.

Keywords: Quantitative model refinement; biomodeling; quantitative analysis; heat shock response; acetylation; data refinement; process refinement.

1 INTRODUCTION

When developing system-level models for large biological systems, such as regulatory networks, signaling pathways, or metabolic pathways, one often

starts with an abstraction of the biological phenomena, incorporating a rather small number of biochemical reactions that illustrate the main mechanisms of the considered process. Moreover, the considered reactions can be themselves abstractions of some particular subprocesses, melting in one step a sequence of several biochemical reactions. This abstract biochemical network is then translated into a mathematical model comprising the mathematical equations illustrating the dynamic behaviour of the system. To achieve this, one has to select first a suitable kinetic law (e.g., mass-action law, Michaelis-Menten kinetics, Hill equations, etc.) which is then used to derive the mathematical equations describing the system's evolution. The numerical setup required for this mathematical model can be either attained from literature or acquired from experimental data by means of diverse model fit techniques.

After constructing this abstract model, the process of model development continues with a succession of steps including hypothesis generation, experimental design and analysis, and model refinement [1, 10]. For instance, one might be interested in refining some of the abstractions that were incorporated in the initial model. One approach for this would be to reiterate the whole development routine in order to include the proposed modifications. However, since that would involve refitting and revalidating the model, this method is squandering resource-wise [2]. An alternative to this approach, not thoroughly inspected so far, would be refining the model gradually, ensuring at the same time that the quantitative model fit and validation are preserved. That is, the numerical setup of the refined model should be derived from that of the initial one, see [14].

The focus of this paper is the model refining step from the above model development process. We analyse here both *data refinement* and *process refinement* of a model. In the first case, the model is refined by replacing some of its species with several types of their subspecies. This situation appears, for instance, when we want to include more details about the post-translational modifications of proteins, e.g., acetylation, phosphorylation, or ubiquitination. This leads to the replacement of a given protein P by its variants that indicate its status with respect to the post-translational modifications, e.g., whether P was acetylated or not. This substitution requires also a refinement of all the reactions involving that particular protein. In the latter type of refinement, the model is enhanced by including more details about a given subprocess, i.e., by adding more biochemical reactions to the model. Then, we present a method that derives the parameter values of the refined model from those of the initial model. This way we ensure the preservation of the systemic properties previously obtained, e.g., the numerical fit. Moreover, we also took into account the biological considerations from [22] and we validated the refined model according to the total level of acetylation.

The paper is organized as follows. We start by introducing the model refinement procedure making sure that we preserve the numerical properties of the system, i.e., numerical fit and validation. To illustrate this technique, we then consider as a case study the role of protein acetylation within the eukaryotic heat shock response. Specifically, we start with a recently introduced model for the heat shock response, see [17], and we refine it to include the acetylation of some specific proteins. Thereafter, we proceed with the process refinement of the model, including its numerical fit and validation conforming to the observations in [22].

A short version of this paper has been presented in [4]. In the short version, we skipped all details of how the acetylation of the heat shock factor contributes to the modulation of the heat shock response. In this extended version, we add this part to our model through a second refinement cycle, incorporating the biological mechanisms recently demonstrated in [22].

2 METHODS

Various types of refinement can be implemented on a given model. For instance, when focussing on a model's data, one could refine some particular species replacing them by several of their subspecies. The model refined as such would depict behavioural variations in the subspecies. We call this *data refinement*: some of the variables of the model are replaced (or *refined*) to include more details, e.g. about some of their internal state or attributes. One could also concentrate on the model reactions and refine the model by substituting the intermediary steps of a certain process for the process itself. We call the latter type of refinement *process refinement*.

The formal refinement problem originated in the field of software engineering from the need to formally prove the correspondence between a system's original specifications and its final implementation. Quantitative model refinement has been studied, in systems biology, particularly within the framework of rule-based modelling, see [5, 8, 15]. *Data refinement*, as defined above, is already contained in this framework through the concept of agent resolution, see [8]. The main refinement technique considered within this formalism is rule refinement, a technique for refining the set of rules so that the systemic properties of the model are preserved [15].

Irrespective of the kind of refinement preferred, the refinement procedure ought to preserve the numerical properties of the model, i.e. quantitative fit and validation. Consequently, we introduce a method to refine model data so as to preserve the formerly attained numerical properties of the model.

Consider that a model M consists of a list of m species $\Sigma = \{A_1, A_2, \ldots, A_m\}$ and n reactions r_i, $1 \leq i \leq n$, given in the form of rewriting rules:

$$r_i \ : \ S_{i,1}A_1 + S_{i,2}A_2 + \ldots + S_{i,m}A_m \xrightarrow{k_i} S'_{i,1}A_1 + S'_{i,2}A_2 + \ldots + S'_{i,m}A_m,$$

where $S_{i,1}, \ldots, S_{i,m}, S'_{i,1}, \ldots, S'_{i,m} \geq 0$ are some integers called *the stoichiometric coefficients* of r_i and $k_i \geq 0$ is *the kinetic rate constant* of r_i.

One can formally specify a model M in various ways, with respect to continuous or discrete variables, deterministic or non-deterministic evolution etc. We preferred, for this study, a continuous, mass-action formulation for the model. We associated to each variable A_i, $1 \leq i \leq m$, a time-dependent function $[A_i] : \mathbb{R}_+ \to \mathbb{R}_+$. Specifically, for each species A_i, we expressed the concentration at a specific moment in time t by $[A_i](t)$. Hence, the evolution of the system is characterized by a mass-action-based system of differential equations, see [12].

Particularly, the model M follows the subsequent system of differential equations:

$$\frac{d[A_j]}{dt} = -\sum_{i=1}^{n}(k_i S_{i,j} \prod_{k=1}^{m}[A_k]^{S_{i,k}}) + \sum_{i=1}^{n}(k_i S'_{i,j} \prod_{k=1}^{m}[A_k]^{S_{i,k}}), \quad 1 \leq j \leq m. \tag{1}$$

The following lemma pertains to the existence and uniqueness of solutions of systems of ODEs derived following the mass action law.

Lemma 1. *[9] Given a molecular model and its associated system of ODEs derived based on the principle of mass action, for any fixed initial condition, there exists an interval of the form* $[0, x)$, $x \in \mathbb{R}_+ \cup \{+\infty\}$ *and a solution* ϕ *such that any other solution is a restriction of* ϕ.

Consider now that the model M is refined by drawing the distinction between several subspecies of A_1. These subspecies could be various types of A_1, with different biochemical conformations, for instance, produced by some post-translational modifications etc. Each of the aforementioned subspecies engages in all reactions A_1 participated in (in model M), probably with different kinetics. If we substitute the species A_1 by subspecies $B_1, \ldots, B_l$, then we will obtain a new model M_R, whose species are denoted by the new variables $\{A'_2, A'_3, \ldots, A'_m\} \cup \{B_1, \ldots, B_l\}$, for some $l \geq 2$. Variables A'_i, $2 \leq i \leq m$, correspond to species A_i from model M, whereas $B_1, \ldots, B_l$ substitute the species A_1 in model M_R. Furthermore, each reaction r_i from M is replaced in M_R by a reaction r'_i as follows:

$$r'_i \ : \ (T_{i,1}B_1 + \ldots + T_{i,l}B_l) + S_{i,2}A'_2 + \ldots + S_{i,m}A'_m \xrightarrow{k'_i} (T'_{i,1}B_1 + \ldots + T'_{i,l}B_l) + S'_{i,2}A'_2 + \ldots + S'_{i,m}A'_m,$$

with k'_i its kinetic rate constant, and $T_{i,1}, \ldots, T_{i,l}, T'_{i,1}, \ldots, T'_{i,l}$ nonnegative integers such that $T_{i,1} + \ldots + T_{i,l} = S_{i,1}$ and $T'_{i,1} + \ldots + T'_{i,l} = S'_{i,1}$.

We say now that *the model M_R is a data refinement of M on variable A_1* if and only if the following two conditions are satisfied:

$$[A_i](t) = [A'_i](t), \text{ for all } 2 \leq i \leq m, \tag{2}$$
$$[A_1](t) = [B_1](t) + \ldots + [B_l](t), \text{ for all } t \geq 0. \tag{3}$$

The refined model M_R comprises $m + l - 1$ species while M consists of m species. Hence, M_R follows a linear-growth in the size of its data set. The number of reaction substitutions in M_R for the reaction r_i of M is given by the number of non-negative integer solutions of the following system of equations:

$$\begin{aligned} T_{i,1} + T_{i,2} + \ldots + T_{i,l} &= S_{i,1}; \\ T'_{i,1} + T'_{i,2} + \ldots + T'_{i,l} &= S'_{i,1}; \end{aligned}$$

over the independent unknowns $T_{i,j}, T'_{i,j}, 1 \leq j \leq l$. The number of solutions of the first equation is the *multinomial coefficient* "l multichooses $S_{i,1}$", see [6]:

$$\left(\!\!\binom{l}{S_{i,1}}\!\!\right) = \binom{l + S_{i,1} - 1}{S_{i,1}} = \frac{(l + S_{i,1} - 1)!}{S_{i,1}!(l-1)!}.$$

Taking into account that both equations in the system are independent, the number of solutions of the system is $\left(\!\!\binom{l}{S_{i,1}}\!\!\right) \cdot \left(\!\!\binom{l}{S'_{i,1}}\!\!\right)$. This represents the increase in the number of reactions in the model refinement from M to M_R. In terms of kinetic parameters, M_R will have $\left(\!\!\binom{l}{S_{i,1}}\!\!\right)$ new free parameters, given by the number of possible combinations of $B_1, \ldots, B_l$ as reactants.

Based on biokinetics arguments, see [16], we assumed that $S_{i,1} \leq 2$: any reaction with a ternary (or higher) stoichiometric coefficient would be so slow that its effects may be ignored. If $S_{i,1} = 1$, then M_R will gain $\left(\!\!\binom{l}{1}\!\!\right) = l$ new kinetic parameters, i.e., a linear growth in the parameter space. If $S_{i,1} = 2$, M_R will have $\left(\!\!\binom{l}{2}\!\!\right) = l(l+1)/2$ new parameters, i.e., a quadratic increase in the parameter space.

The values of some of the new kinetic parameters of M_R could be obtained from literature or they could be empirically estimated. The remaining parameters, whose values are not known and cannot be *directly* determined, ought to be calculated using a computational technique in such a way that (2) and (3) hold.

Repeating the parameter estimation process for a model whose parameter space may grow quadratically at every stage of the refinement, is prodigal resource-wise. Furthermore, any intermediate fit of the model would not be preserved in the following refinement step, which makes such an approach senseless.

We propose in this paper an approach through which we can consistently set values for the unknown parameters of the refined model so that the relations (2) and (3) are satisfied. Some of the possible choices for the kinetic parameters of the refined model could be discarded immediately as not viable; for instance, those for which all reactions comprising $B_2, B_3, \ldots,$ or B_l had their kinetic rate constants set to 0. This type of choice would simply translate the model M to M_R by renaming all its variables. If any information about the numerical values of some of the parameters exists, then it will be integrated in the model. For the remaining parameters, we take the approach in which the refined subspecies $B_1, \ldots, B_l$ of species A_1 cannot be set apart by the kinetics of the reactions they are involved in; in other words, our setting of the numerical values of the parameters is symmetric in $B_1, \ldots, B_l$.

Example 1. *We consider the Lotka-Volterra model [13, 21], composed of the species A (the prey) and B (the predator), and the following reactions:*

$$A \xrightarrow{k_1} 2A, \qquad A + B \xrightarrow{k_2} 2B, \qquad B \xrightarrow{k_3} \emptyset.$$

The set of mass-action based ODEs describing the dynamics of this system are:

$$d[A]/dt = k_1[A] - k_2[A][B] \qquad d[B]/dt = k_2[A][B] - k_3[B].$$

After refining the Lotka-Volterra model on variable A into subspecies A_1 and A_2, we obtain the following set of reactions:

$$\begin{array}{lll} A_1 \xrightarrow{r_1} A_1 + A_1, & A_1 \xrightarrow{r_2} A_1 + A_2, & A_1 \xrightarrow{r_3} A_2 + A_2, \\ A_2 \xrightarrow{r_4} A_1 + A_1, & A_2 \xrightarrow{r_5} A_1 + A_2, & A_2 \xrightarrow{r_6} A_2 + A_2, \\ A_1 + B' \xrightarrow{r_7} 2B', & A_2 + B' \xrightarrow{r_8} 2B', & B' \xrightarrow{r_9} \phi. \end{array}$$

This leads to the following system of ODEs describing the dynamics of the variables A_1, A_2, and B':

$$\begin{aligned} d[A_1]/dt &= (r_1 - r_3)[A_1] + (2r_4 + r_5)[A_2] - r_7[A_1][B'], \\ d[A_2]/dt &= (r_2 + 2r_3)[A_1] + (-r_4 + r_6)[A_2] - r_8[A_2][B'], \\ d[B']/dt &= r_7[A_1][B'] + r_8[A_2][B'] - r_9[B']. \end{aligned}$$

Thus, $d([A_1]+[A_2])/dt = (r_1 - r_3 + r_2 + 2r_3)[A_1] + (2r_4 + r_5 - r_4 + r_6)[A_2] - (r_7 + r_8)[A_1][B']$. *Consequently, if we choose* $r_1 = r_2 = r_3 = r_4 = r_5 = r_6 = k_1/3$, $r_7 = r_8 = k_2$ *and* $r_9 = k_3$ *then we obtain*

$$d([A_1]+[A_2])/dt = k_1([A_1]+[A_2]) - k_2([A_1]+[A_2])[B'],$$
$$d[B']/dt = k_2([A_1]+[A_2])[B'] - k_3[B'],$$

which is identical to the initial system up to a renaming of variables where $[A]$ *is replaced by* $[A_1]+[A_2]$ *and* $[B]$ *is replaced by* $[B']$. *For any* $x_0 \geq 0$, *if we set the initial values of the variables* A_1, A_2, *and* B' *such that* $[A](x_0) = [A_1](x_0) + [A_2](x_0)$ *and* $[B](x_0) = [B'](x_0)$, *it follows from Lemma 1 that there exists an interval of the form* $[0, x)$, $x \in \mathbb{R}_+ \cup \{+\infty\}$ *including* x_0 *such that* $[A](t) = [A_1](t) + [A_2](t)$ *and* $[B](t) = [B'](t)$, *for all* $t \in [0, x)$. *That is, the second model is indeed a quantitative refinement of the initial one (on the interval* $[0, x)$*).*

3 MODELS

3.1 The eukaryotic heat shock response: molecular model

The eukaryotic heat shock response is a highly conserved defence mechanism, crucial to cell survival, that protects cells from various environmental stressors, such as elevated temperatures. Exposure of cells to elevated temperatures, due to its proteotoxic effects, causes protein misfolding and constitution of large conglomerates that ultimately lead to controlled cell death (apoptosis). Cell survival is induced by a defence mechanism that resumes protein homeostasis, i.e. an equilibrium between the synthesis, folding and degradation of proteins.

We describe below the molecular model for the heat shock response proposed in [18]. The eukaryotic cell reacts to elevated temperatures by promoting the synthesis of some well conserved proteins, called heat shock proteins (hsp), which assist the misfolded proteins in their refolding process (process known as chaperoning) and also induce the degradation of severely damaged proteins. The transactivation of the hsp-encoding genes controls the heat shock reponse. In eukaryotes, gene transcription is regulated by heat shock transcription factors (hsf's), which undergo a conversion from a monomeric to a trimeric DNA binding competent conformation. Hsf trimers have a high affinity towards the promoter site of the hsp-encoding gene, called heat shock element (hse). Thus, they bind to the heat shock elements and form hsf_3: hse complexes. Subsequently, hsp synthesis is activated. Removal of the heat stress leads to the repression of hsp synthesis [17, 18]. Heat shock proteins sequestrate heat shock factors, constituting hsp: hsf complexes , hence compelling hsf's to trimerize and subsequently bind to the promoter site of the

$2\,\mathsf{hsf} \leftrightarrows \mathsf{hsf}_2$	$\mathsf{hsp} + \mathsf{hsf}_3 \rightarrow \mathsf{hsp}\colon \mathsf{hsf} + 2\,\mathsf{hsf}$
$\mathsf{hsf} + \mathsf{hsf}_2 \leftrightarrows \mathsf{hsf}_3$	$\mathsf{hsp} + \mathsf{hsf}_3\colon \mathsf{hse} \rightarrow \mathsf{hsp}\colon \mathsf{hsf} + 2\,\mathsf{hsf} + \mathsf{hse}$
$\mathsf{hsf}_3 + \mathsf{hse} \leftrightarrows \mathsf{hsf}_3\colon \mathsf{hse}$	$\mathsf{hsp} \rightarrow \emptyset$
$\mathsf{hsf}_3\colon \mathsf{hse} \rightarrow \mathsf{hsf}_3\colon \mathsf{hse} + \mathsf{hsp}$	$\mathsf{prot} \rightarrow \mathsf{mfp}$
$\mathsf{hsp} + \mathsf{hsf} \leftrightarrows \mathsf{hsp}\colon \mathsf{hsf}$	$\mathsf{hsp} + \mathsf{mfp} \leftrightarrows \mathsf{hsp}\colon \mathsf{mfp}$
$\mathsf{hsp} + \mathsf{hsf}_2 \rightarrow \mathsf{hsp}\colon \mathsf{hsf} + \mathsf{hsf}$	$\mathsf{hsp}\colon \mathsf{mfp} \rightarrow \mathsf{hsp} + \mathsf{prot}$

TABLE 1
The molecular model for the eukaryotic heat shock response proposed in [17].

heat shock gene. Once the temperature is elevated, some proteins prot start to misfold and cause the hsp: hsf complexes to break down. This rapidly actuates the heat shock response once heat shock factors are free and get in their DNA-binding competent state. The reactions of the molecular model in [17] are presented in Table 1.

3.2 Mathematical model

We consider a mathematical model formulated as a mass-action based system of ODEs, see [12], associated to the molecular model in Table 1. The system of ODEs is shown in Table 2; the reader is referred to [17] for a more detailed description of the model.

Based on [17], all kinetic rate constants and initial values of all reactants were estimated so that they satisfy the following conditions:

(i) the system is in a steady state for a temperature of $37^{\circ}C$ due to the fact that the model should not express any response in the absence of the heat stress;

(ii) the numerical predictions of the model for $[\mathsf{hsf}_3\colon \mathsf{hse}](t)$ should agree with the experimental data from [11], for a temperature of $42^{\circ}C$;

(iii) the numerical prediction of the model for $[\mathsf{hsp}](t)$ should confirm the data obtained in [17] through a de-novo fluorescent reporter-based experiment, for a temperature of $42^{\circ}C$.

3.3 The role of protein acetylation within the eukaryotic heat shock response

Heat shock factors (hsf's), i.e., the transcription factors for the hsp-encoding genes, are crucial for cell survival, protecting the cell from environmental stressors that cause protein alteration, such as protein misfolding. Recent studies show that protein acetylation is highly relevant to regulating the heat shock response, see [22]. Protein acetylation resides in replacing a hydrogen atom by an acetyl group in a chemical compound. It can arise either at the alpha-amino group of the amino-terminus or on the lysine residues at

the ε-amino group, see [7]. N-terminal acetylation is an irreversible reaction, hence the proteins preserve their acetylation state until degradation. Although N-terminal acetylation in eukaryotes is an extensive process, its biological role has not been yet accurately clarified [19]. Nonetheless, lysine acetylation is a reversible reaction that can posttranslationally modify the positive charge of lysine and, thus, vary the role of a protein. Post-translational ε-amino lysine acetylation plays an important role in gene regulation due to the changes it brings to the charge of histone tails, neutralizing the positive charge of histones. Lysine acetylation is known to attenuate DNA-binding affinity, see [3, 20].

4 RESULTS

4.1 Data refinement of the model

In this section, we focus on augmenting the heat shock model proposed in [17] by including the acetylation of hsf and the effect it exhibits on the response. We refined all compounds comprising hsf so as to include two subtypes of hsf: one having the $K80$ residue acetylated and one having it not acetylated. Therefore, given species hsp: hsf in the initial model, it will be refined into two subspecies, with respect to the acetylation status of its hsf constituent. For hsf_2, hsf_3 and hsf_3: hse, we perform the refinement counting the number of hsf constituents of the complex that have the K80 residue acetylated. Subsequently, we carried out the following data refinements:

$$\mathrm{hsf} \to \{\mathrm{rhsf}, \mathrm{rhsf}^{(1)}\};$$
$$\mathrm{hsf}_2 \to \{\mathrm{rhsf}_2, \mathrm{rhsf}_2{}^{(1)}, \mathrm{rhsf}_2{}^{(2)}\};$$
$$\mathrm{hsf}_3 \to \{\mathrm{rhsf}_3, \mathrm{rhsf}_3{}^{(1)}, \mathrm{rhsf}_3{}^{(2)}, \mathrm{rhsf}_3{}^{(3)}\};$$
$$\mathrm{hsf}_3\mathrm{:hse} \to \{\mathrm{rhsf}_3\mathrm{:rhse}, \mathrm{rhsf}_3{}^{(1)}\mathrm{:rhse}, \mathrm{rhsf}_3{}^{(2)}\mathrm{:rhse}, \mathrm{rhsf}_3{}^{(3)}\mathrm{:rhse}\};$$
$$\mathrm{hsp:hsf} \to \{\mathrm{rhsp:rhsf}, \mathrm{rhsp:rhsf}^{(1)}\}.$$

The above performed data refinements entail various changes in the reaction list of the model. For instance, reactions: $\mathrm{rhsp} + \mathrm{rhsf} \leftrightarrows \mathrm{rhsp:rhsf}$, and $\mathrm{rhsp} + \mathrm{rhsf}^{(1)} \leftrightarrows \mathrm{rhsp:rhsf}^{(1)}$ in the refined model substitute for reaction $\mathrm{hsp} + \mathrm{hsf} \leftrightarrows \mathrm{hsp:hsf}$ in the initial model. As another example, reaction $2\,\mathrm{hsf} \leftrightarrows \mathrm{hsf}_2$ is replaced by three reactions, based on the method described in Section 2:

$$2\,\mathrm{rhsf} \leftrightarrows \mathrm{rhsf}_2; \quad \mathrm{rhsf} + \mathrm{rhsf}^{(1)} \leftrightarrows \mathrm{rhsf}_2{}^{(1)}; \quad 2\mathrm{rhsf}^{(1)} \leftrightarrows \mathrm{rhsf}_2{}^{(2)}.$$

For a complete list of reactions of the refined model, we refer the reader to B. The refined model has 20 species, 39 reactions (some of which are reversible)

and 54 kinetic parameters. The initial model had 10 species, 12 reactions and 16 kinetic parameters.

4.2 The numerical setup of the refined model

In this section we address the numerical setup of the refined model for the heat shock response. We determine these numerical values in such a way that the refinement procedure is quantitatively correct, i.e., the relations (2)-(3) are verified. The refined variables have their initial values set conforming to the following 10 conditions, determined from the data refinement relations:

$$[\mathsf{hsf}](0) = [\mathsf{rhsf}](0) + [\mathsf{rhsf}^{(1)}](0);$$
$$[\mathsf{hsf}_2](0) = [\mathsf{rhsf}_2](0) + [\mathsf{rhsf}_2{}^{(1)}] + [\mathsf{rhsf}_2{}^{(2)}](0);$$
$$[\mathsf{hsf}_3](0) = [\mathsf{rhsf}_3](0) + [\mathsf{rhsf}_3{}^{(1)}](0) + [\mathsf{rhsf}_3{}^{(2)}](0) + [\mathsf{rhsf}_3{}^{(3)}](0);$$
$$[\mathsf{hsp}\colon \mathsf{hsf}](0) = [\mathsf{hsp}\colon \mathsf{rhsf}](0) + [\mathsf{rhsp}\colon \mathsf{rhsf}^{(1)}](0);$$
$$[\mathsf{hsf}_3\colon \mathsf{hse}](0) = [\mathsf{rhsf}_3\colon \mathsf{rhse}](0) + [\mathsf{rhsf}_3{}^{(1)}\colon \mathsf{rhse}](0) + [\mathsf{rhsf}_3{}^{(2)}\colon \mathsf{rhse}](0) + [\mathsf{rhsf}_3{}^{(3)}\colon \mathsf{rhse}](0);$$
$$[\mathsf{hsp}](0) = [\mathsf{rhsp}](0);$$
$$[\mathsf{hsp}\colon \mathsf{mfp}](0) = [\mathsf{rhsp}\colon \mathsf{rmfp}](0);$$
$$[\mathsf{mfp}](0) = [\mathsf{rmfp}](0);$$
$$[\mathsf{prot}](0) = [\mathsf{rprot}](0);$$
$$[\mathsf{hse}](0) = [\mathsf{rhse}](0).$$

The mass-action based system of ODEs for the refined model is illustrated in Table 4. Regarding the numerical setup of the refined model, we aimed to determine it such that it constitutes a quantitative refinement of the model in [17], as defined in Section 2. We start by introducing the following notations.

$$\mathsf{Rhsf} = \mathsf{rhsf} + \mathsf{rhsf}^{(1)};$$
$$\mathsf{Rhsf}_2 = \mathsf{rhsf}_2 + \mathsf{rhsf}_2{}^{(1)} + \mathsf{rhsf}_2{}^{(2)};$$
$$\mathsf{Rhsf}_3 = \mathsf{rhsf}_3 + \mathsf{rhsf}_3{}^{(1)} + \mathsf{rhsf}_3{}^{(2)} + \mathsf{rhsf}_3{}^{(3)};$$
$$\mathsf{Rhsf}_3\colon \mathsf{Rhse} = \mathsf{rhsf}_3\colon \mathsf{rhse} + \mathsf{rhsf}_3{}^{(1)}\colon \mathsf{rhse} + \mathsf{rhsf}_3{}^{(2)}\colon \mathsf{rhse} + \mathsf{rhsf}_3{}^{(3)}\colon \mathsf{rhse};$$
$$\mathsf{Rhsp}\colon \mathsf{Rhsf} = \mathsf{rhsp}\colon \mathsf{rhsf} + \mathsf{rhsp}\colon \mathsf{rhsf}^{(1)}.$$

Our aim is to determine some values for the kinetic parameters of the refined model so that the two systems of ODEs, the one for the refined model and that for the initial model, are indistinguishable, modulo a variable renaming, given that Rhsf, Rhsf_2, Rhsf_3, $\mathsf{Rhsf}_3\colon \mathsf{Rhse}$, and $\mathsf{Rhsp}\colon \mathsf{Rhsf}$ substitute for hsf, hsf_2, hsf_3, $\mathsf{hsf}_3\colon \mathsf{hse}$, and $\mathsf{hsp}\colon \mathsf{hsf}$ respectively. Consequently, we wrote

the ODEs for [Rhsf], [Rhsf_2], [Rhsf_3], [Rhsf_3: Rhse] and for [Rhsp: Rhsf], see Table 5, based on the system of ODEs in Table 4.

We then determined the kinetic parameter values so that the right hand side of each ODE in Table 5 turns into the right hand side of the corresponding ODE in Table 2, modulo the above variable renaming. For instance, we determine the kinetic parameters of the refined model so as to obtain from the ODE corresponding to Rhsf in Table 5 an identical one corresponding to hsf in Table 2, modulo the above variable renaming. To identify the first term of the ODE for Rhsf (in the form written in Table 5) with the first term of the ODE for hsf (in the form written in Table 2), it is sufficient to set $r_1^+ = k_1^+$, $r_2^+ = 2k_1^+$, and $r_3^+ = k_1^+$. Using a similar reasoning for all terms of all ODEs, we can ultimately find a solution. We include in D a table comprising the identified values for the parameters.

The solution is clearly not unique. To find all solutions we should solve the systems of ODEs in Tables 2 and 5. However, these systems cannot be solved analytically. Instead, we chose some values for the kinetic parameters of the refined model as expressions depending on the kinetic parameters of the initial model, so that the ODEs in Table 5 can be translated into the equations in Table 2, modulo the above variable renaming. Our approach in solving this problem is aiming for a minimum number of independent parameters: all (de)acetylation reactions share the same kinetic parameter. On the other hand, when biological knowledge was available that similar reactions might in fact have different kinetic parameters (e.g., the acetylation-induced transcription rate of the hsp-encoding gene, discussed in the next section), we included it in the model as such.

Due to our choice for the initial values of the refined variables, we obtain that the systems of ODE's associated to the initial model and to the refined one have identical initial conditions. Hence, due to Lemma 1, we can conclude that conditions (2) and (3) are satisfied, i.e., the second model is indeed a quantitative refinement of the initial model from Table 1.

4.3 Process refinement of the model

Heat shock factors have a crucial role in regulating protein homeostatic processes such as synthesis, refolding or degradation. They endure a three-step conversion from a latent monomer state to a transcriptionally active DNA binding competent trimer. During the attenuation of the heat shock, DNA binding activity declines. This is also correlated to an increase in the level of acetylation, which enhances the unbinding of the hsf trimers from DNA, regardless of the behaviour of hsp's [22].

Next, we modified the refined model so that it reflects the considerations from [22]. In particular, we replace in the final model the reactions corresponding to the hsp-mediated unbinding of the hsf_3-complex, i.e., q_{30}, ...,

q_{35} from Table 3, with a set of reactions associated to the unbinding of trimers in the absence of hsp:

- $\mathsf{rhsf}_3\colon \mathsf{rhse} \rightarrow \mathsf{rhsf}_3 + \mathsf{rhse}$ with the kinetic rate constant s_1,
- $\mathsf{rhsf}_3{}^{(1)}\colon \mathsf{rhse} \rightarrow \mathsf{rhsf}_3{}^{(1)} + \mathsf{rhse}$ with the kinetic rate constant s_2,
- $\mathsf{rhsf}_3{}^{(2)}\colon \mathsf{rhse} \rightarrow \mathsf{rhsf}_3{}^{(2)} + \mathsf{rhse}$ with the kinetic rate constant s_3,
- $\mathsf{rhsf}_3{}^{(3)}\colon \mathsf{rhse} \rightarrow \mathsf{rhsf}_3{}^{(3)} + \mathsf{rhse}$ with the kinetic rate constant s_4.

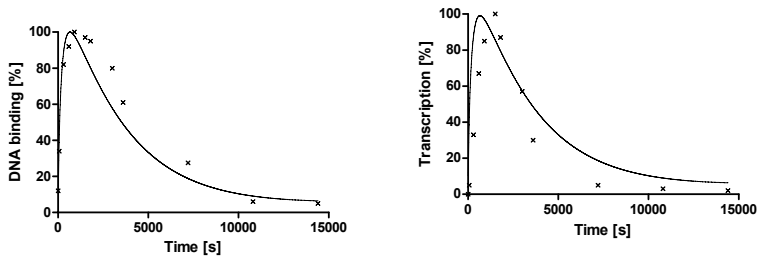
According to [22], acetylation positively regulates DNA unbinding, i.e., $s_1 \leq s_2 \leq s_3 \leq s_4$. Furthermore, we also require the model to obey the following three mass-conservation relations:

- $[\mathsf{rhsf}] + [\mathsf{rhsf}^{(1)}] + 2 \times [\mathsf{rhsf}_2] + 2 \times [\mathsf{rhsf}_2{}^{(1)}] + 2 \times [\mathsf{rhsf}_2{}^{(2)}] + 3 \times [\mathsf{rhsf}_3] + 3 \times [\mathsf{rhsf}_3{}^{(1)}] + 3 \times [\mathsf{rhsf}_3{}^{(2)}] + 3 \times [\mathsf{rhsf}_3{}^{(3)}] + 3 \times [\mathsf{rhsf}_3\colon \mathsf{rhse}] + 3 \times [\mathsf{rhsf}_3{}^{(1)}\colon \mathsf{rhse}] + 3 \times [\mathsf{rhsf}_3{}^{(2)}\colon \mathsf{rhse}] + 3 \times [\mathsf{rhsf}_3{}^{(3)}\colon \mathsf{rhse}] + [\mathsf{hsp}\colon \mathsf{rhsf}] + [\mathsf{rhsp}\colon \mathsf{rhsf}^{(1)}] = C_1$
- $[\mathsf{rprot}] + [\mathsf{rmfp}] + [\mathsf{rhsp}\colon \mathsf{rmfp}] = C_2$
- $[\mathsf{rhse}] + [\mathsf{rhsf}_3\colon \mathsf{rhse}] + [\mathsf{rhsf}_3{}^{(1)}\colon \mathsf{rhse}] + [\mathsf{rhsf}_3{}^{(2)}\colon \mathsf{rhse}] + [\mathsf{rhsf}_3{}^{(3)}\colon \mathsf{rhse}] = C_3$,

where C_1, C_2 and C_3 are the mass constants considered in the original model, introduced in [17]. These relations are the refinements of the mass conservation relations satisfied by the basic model. Finally, we also include a set of reactions to model the acetylation and deacetylation processes. In particular, we set one kinetic rate constant for all the acetylation reactions (independent on the molecule that is acetylated) and one kinetic rate constant for all the deacetylation reactions. The reactions of the final molecular model for the heat shock response are shown in Table 7.

4.4 Numerical data fit of the final model

Most of the parameters of the final model were already fixed in Section 4.2 depending on the parameters of the initial model. The only ones left were the kinetic rate constants for DNA unbinding in the absence of hsp, i.e., s_1, s_2, s_3, and s_4, and the kinetic rate constants for the reactions describing the acetylation and deacetylation of the hsf molecules. In order to estimate these six parameters, we fitted the final model to the experimental data on the DNA binding and gene transcription levels from [11], see Figure 1. The results showed that unbinding rarely happens for $\mathsf{hsf}_3\colon \mathsf{hse}$ and $\mathsf{hsf}_3{}^{(1)}\colon \mathsf{hse}$, while it happens at a high rate for $\mathsf{hsf}_3{}^{(2)}\colon \mathsf{hse}$ and $\mathsf{hsf}_3{}^{(3)}\colon \mathsf{hse}$, thus confirming the findings from [22]. Figure 2 illustrates the acetylation levels of the hse-bound



(a) (b)

FIGURE 1
(a) DNA binding; (b) Dynamics of the transcription level of the hsp-encoding gene.

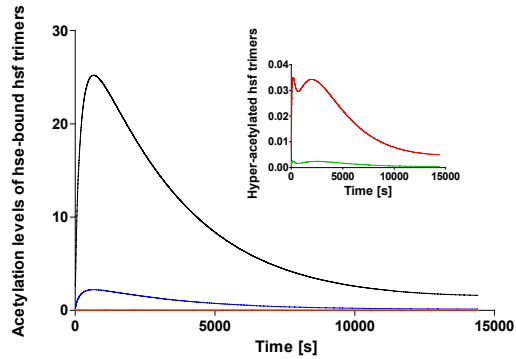


FIGURE 2
The dynamics of the hse-bound hsf trimers, shown separately according to their acetylation level: zero (black), one (blue), two (red) or all three (green) binding sites of the heat shock factor are acetylated.

hsf trimers. Our simulations confirm that acetylation regulates DNA unbinding, thus promoting HSR attenuation. The dynamics of the acetylated hsf shows that the acetylation process is delayed during the heat shock response, but it persists also during the period when the hsf activity and DNA binding have diminished, in accordance with the observations from [22].

5 DISCUSSION

We focused in this paper on quantitative model refinement as an approach for the construction of large reaction models. We propose to start from a core model that is first fit and validated against experimental data. We then

systematically add details to the model through data and process refinement in such a way that the experimental fit of the model is preserved. This approach avoids the repetition of the entire model development procedure, including model fit, which is both time-consuming and computationally-intensive, see [2].

In order to illustrate this method, we chose a recently introduced model for the eukaryotic heat shock response, [17]. Then, we refined this model by including details about the acetylation of the heat shock factors hsf, the promoters of the heat shock proteins. Although each hsf molecule has nine lysine sites that can be acetylated in response to stress, it was suggested in [22] that one has a regulatory role for the DNA binding ability of the hsf molecules and thus implicitly also for the expression level of the heat shock proteins. Thus, we chose to refine the model from [17] by considering only one acetylation site for each hsf molecule. This led to a significant augmentation in the number of the reactions of the refined model and thus also in the number of parameters. While the basic model comprises a set of 12 reactions involving 10 different species and 16 kinetic rate constants, the refined model contains a number of 39 reactions involving 20 species and 54 kinetic rate constants. Fitting a model of this proportion implies a lot of time and computational resources. However, with our approach, we were able to infer the values of all kinetic rate constants from the initial model, avoiding any supplementary model fit. After this step, we did also a second round of refinement in which we included some of the processes reported in [22]. This resulted in the introduction of 6 new kinetic rate constants, which were estimated in such a way that the behaviour predicted by our model would be in agreement with the experimental data in [11]. Moreover, as a validation of our final model, we checked the dynamics of the level of acetylation which proved to be in accordance with the observations in [22]. The final model preserves considerably many values for the kinetic parameters and variables, which resides into a computationally more efficacious parameter estimation.

Among the nine lysine sites in the hsf molecule that can be acetylated in response to stress, $K80$ is located in the DNA binding domain and thus influences its binding ability. Henceforth, in [22] the acetylation of $K80$ is proposed as a regulatory mechanism for the DNA binding ability of the hsf molecules and implicitly for the expression level of the heat shock proteins. Changes in the level of acetylation of the hsf molecules influence its binding ability and thus the expression of the heat shock proteins. In particular, an increase in the level of acetylation caused DNA binding to take place at a lower rate [22]. Moreover, the dynamics of the hsf acetylation is not identical to the dynamics of hsf activation. In particular, it is shown in [22] that the acetylation process is delayed but it continues also when the hsf DNA-binding activity has decreased.

REFERENCES

[1] F. J. Bruggeman and H. V. Westerhoff. (2007). The nature of systems biology. *TRENDS in Microbiology*, 15(1):45–50.

[2] W. W. Chen, B. Schoeberl, P. J. Jasper, M. Niepel, U. B. Nielsen, D. A. Lauffenburger, and P. K. Sorger. (2009). Input–output behavior of erbb signaling pathways as revealed by a mass action model trained against dynamic data. *Molecular systems biology*, 5(1).

[3] C. Choudhary, C. Kumar, F. Gnad, M.L. Nielsen, M. Rehman, T. C. Walther, J. V. Olsen, and M. Mann. (2009). Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science's STKE*, 325(5942):834–840.

[4] El. Czeizler, Eu. Czeizler, B. Iancu, and I. Petre. (2012). Quantitative model refinement as a solution to the combinatorial size explosion of biomodels. *Electronic Notes in Theoretical Computer Science*, 284:35–53.

[5] V. Danos, J. Feret, W. Fontana, R. Harmer, and J. Krivine. (2009). Rule-based modelling and model perturbation. *Transactions on Computational Systems Biology*, 5750:116–137.

[6] I. M. Gessel and R. P. Stanley. (1995). Algebraic enumeration. In R. L. Graham, M. Grötschel, and L. Lovász, editors, *Handbook of combinatorics*, volume 2, pages 1021–1061. North Holland.

[7] M. A. Glozak, N. Sengupta, X. Zhang, and E. Seto. (2005). Acetylation and deacetylation of non-histone proteins. *Gene*, 363:15–23.

[8] R. Harmer. (2009). Rule-based modelling and tunable resolution. In *Proceedings of the Fifth Workshop on Developments in Computational Models–Computational Models From Nature*, volume 9 of *Electronic Proceedings in Theoretical Computer Science*.

[9] M. W. Hirsch, S. Smale, and R. L. Devaney. (2004). *Differential equations, dynamical systems, and an introduction to chaos*, volume 60. Academic Press.

[10] H. Kitano. (2002). Systems biology: a brief overview. *Science*, 295(5560):1662–1664.

[11] M. P. Kline and R. I. Morimoto. (1997). Repression of the heat shock factor 1 transcriptional activation domain is modulated by constitutive phosphorylation. *Molecular and cellular biology*, 17(4):2107–2115.

[12] E. Klipp, R. Herwig, A. Kowald, C. Wierling, and H. Lehrach. (2005). *Systems Biology in Practice: Concepts, Implementation and Application*. WILEY-VCH Verlag GmbH & Co. KGaA.

[13] A. J. Lotka. (1925). *Elements of Physical Biology*. Williams & Wilkins company.

[14] A. Mizera, E. Czeizler, and I. Petre. (2012). Self-assembly models of variable resolution. *LNBI Transactions on Computational Systems Biology*, In Press.

[15] E. Murphy, V. Danos, J. Feret, J. Krivine, and R. Harmer. (2009). Rule-based modeling and model refinement. *Elements of Computational Systems Biology*, pages 83–114.

[16] D. L. Nelson and M. M. Cox. (2000). *Lehninger Principles of Biochemistry*. W. H. Freeman.

[17] I. Petre, A. Mizera, C. L. Hyder, A. Meinander, A. Mikhailov, R. I. Morimoto, L. Sistonen, J. E. Eriksson, and R. J. Back. (2011). A simple mass-action model for the eukaryotic heat shock response and its mathematical validation. *Natural Computing*, 10(1):595–612.

[18] T. R. Rieger, R. I. Morimoto, and V. Hatzimanikatis. (2005). Mathematical modeling of the eukaryotic heat-shock response: dynamics of the hsp70 promoter. *Biophysical journal*, 88(3):1646–1658.

[19] J. Soppa. (2010). Protein acetylation in archaea, bacteria, and eukaryotes. *Archaea*, 2010:1–9.

[20] K. Struhl. (1998). Histone acetylation and transcriptional regulatory mechanisms. *Genes & development*, 12(5):599–606.

[21] V. Volterra. (1926). Animal ecology. In *R. N. Chapman*, pages 409–448. McGraw-Hill, New York.

[22] S. D. Westerheide, J. Anckar, S. M. Stevens Jr, L. Sistonen, and R. I. Morimoto. (2009). Stress-inducible regulation of heat shock factor 1 by the deacetylase sirt1. *Science's STKE*, 323(5917):1063.

A THE ODE-BASED MODEL FOR THE BASIC HEAT SHOCK RESPONSE MODEL

$$
\begin{aligned}
d[\mathsf{hsf}]/dt &= -2k_1^+[\mathsf{hsf}]^2 + 2k_1^-[\mathsf{hsf}_2] - k_2^+[\mathsf{hsf}][\mathsf{hsf}_2] + k_2^-[\mathsf{hsf}_3] \\
&\quad -k_5^+[\mathsf{hsf}][\mathsf{hsp}] + k_5^-[\mathsf{hsp}\colon \mathsf{hsf}] + k_6[\mathsf{hsf}_2][\mathsf{hsp}] \\
&\quad +2k_7[\mathsf{hsf}_3][\mathsf{hsp}] + 2k_8[\mathsf{hsf}_3\colon \mathsf{hse}][\mathsf{hsp}] \\
d[\mathsf{hsf}_2]/dt &= k_1^+[\mathsf{hsf}]^2 - k_1^-[\mathsf{hsf}_2] - k_2^+[\mathsf{hsf}][\mathsf{hsf}_2] + k_2^-[\mathsf{hsf}_3] - k_6[\mathsf{hsf}_2][\mathsf{hsp}] \\
d[\mathsf{hsf}_3]/dt &= k_2^+[\mathsf{hsf}][\mathsf{hsf}_2] - k_2^-[\mathsf{hsf}_3] - k_3^+[\mathsf{hsf}_3][\mathsf{hse}] + k_3^-[\mathsf{hsf}_3\colon \mathsf{hse}] - k_7[\mathsf{hsf}_3][\mathsf{hsp}] \\
d[\mathsf{hse}]/dt &= -k_3^+[\mathsf{hsf}_3][\mathsf{hse}] + k_3^-[\mathsf{hsf}_3\colon \mathsf{hse}] + k_8[\mathsf{hsf}_3\colon \mathsf{hse}][\mathsf{hsp}] \\
d[\mathsf{hsf}_3\colon \mathsf{hse}]/dt &= k_3^+[\mathsf{hsf}_3][\mathsf{hse}] - k_3^-[\mathsf{hsf}_3\colon \mathsf{hse}] - k_8[\mathsf{hsf}_3\colon \mathsf{hse}][\mathsf{hsp}] \\
d[\mathsf{hsp}]/dt &= k_4[\mathsf{hsf}_3\colon \mathsf{hse}] - k_5^+[\mathsf{hsf}][\mathsf{hsp}] + k_5^-[\mathsf{hsp}\colon \mathsf{hsf}] - k_6[\mathsf{hsf}_2][\mathsf{hsp}] \\
&\quad -k_7[\mathsf{hsf}_3][\mathsf{hsp}] - k_8[\mathsf{hsf}_3\colon \mathsf{hse}][\mathsf{hsp}] - k_{11}^+[\mathsf{hsp}][\mathsf{mfp}] \\
&\quad +(k_{11}^- + k_{12})[\mathsf{hsp}\colon \mathsf{mfp}] - k_9[\mathsf{hsp}] \\
d[\mathsf{hsp}\colon \mathsf{hsf}]/dt &= k_5^+[\mathsf{hsf}][\mathsf{hsp}] - k_5^-[\mathsf{hsp}\colon \mathsf{hsf}] + k_6[\mathsf{hsf}_2][\mathsf{hsp}] \\
&\quad +k_7[\mathsf{hsf}_3][\mathsf{hsp}] + k_8[\mathsf{hsf}_3\colon \mathsf{hse}][\mathsf{hsp}] \\
d[\mathsf{mfp}]/dt &= \phi_T[\mathsf{prot}] - k_{11}^+[\mathsf{hsp}][\mathsf{mfp}] + k_{11}^-[\mathsf{hsp}\colon \mathsf{mfp}] \\
d[\mathsf{hsp}\colon \mathsf{mfp}]/dt &= k_{11}^+[\mathsf{hsp}][\mathsf{mfp}] - (k_{11}^- + k_{12})[\mathsf{hsp}\colon \mathsf{mfp}] \\
d[\mathsf{prot}]/dt &= -\phi_T[\mathsf{prot}] + k_{12}[\mathsf{hsp}\colon \mathsf{mfp}]
\end{aligned}
$$

TABLE 2
The system of ODEs associated with the biochemical model proposed in [17].

B THE MOLECULAR HEAT SHOCK RESPONSE MODEL DETAILING THE ACETYLATION STATUS OF hsf

Reaction	Reaction number	Kinetic rate constants
$2\,\mathsf{rhsf} \leftrightarrows \mathsf{rhsf}_2$	$[q_1]$	r_1^+, r_1^-
$\mathsf{rhsf} + \mathsf{rhsf}^{(1)} \leftrightarrow \mathsf{rhsf}_2{}^{(1)}$	$[q_2]$	r_2^+, r_2^-
$2\mathsf{rhsf}^{(1)} \leftrightarrows \mathsf{rhsf}_2{}^{(2)}$	$[q_3]$	r_3^+, r_3^-
$\mathsf{rhsf} + \mathsf{rhsf}_2 \leftrightarrows \mathsf{rhsf}_3$	$[q_4]$	r_4^+, r_4^-
$\mathsf{rhsf}^{(1)} + \mathsf{rhsf}_2 \leftrightarrows \mathsf{rhsf}_3{}^{(1)}$	$[q_5]$	r_5^+, r_5^-
$\mathsf{rhsf} + \mathsf{rhsf}_2{}^{(1)} \leftrightarrows \mathsf{rhsf}_3{}^{(1)}$	$[q_6]$	r_6^+, r_6^-
$\mathsf{rhsf}^{(1)} + \mathsf{rhsf}_2{}^{(1)} \leftrightarrows \mathsf{rhsf}_3{}^{(2)}$	$[q_7]$	r_7^+, r_7^-

TABLE 3
The list of reactions for the refined model that includes the acetylation status of hsf. For an irreversible reaction q_i, r_i denotes its kinetic rate constant. For a reversible reaction q_i, r_i^+ and r_i^- denote the kinetic rate constants of its 'left-to-right' and 'right-to-left' directions, resp.

$\mathrm{rhsf} + \mathrm{rhsf_2}^{(2)} \leftrightarrows \mathrm{rhsf_3}^{(2)}$	$[q_8]$	r_8^+, r_8^-
$\mathrm{rhsf}^{(1)} + \mathrm{rhsf_2}^{(2)} \leftrightarrows \mathrm{rhsf_3}^{(3)}$	$[q_9]$	r_9^+, r_9^-
$\mathrm{rhsf_3} + \mathrm{rhse} \leftrightarrows \mathrm{rhsf_3{:}\ rhse}$	$[q_{10}]$	r_{10}^+, r_{10}^-
$\mathrm{rhsf_3}^{(1)} + \mathrm{rhse} \leftrightarrows \mathrm{rhsf_3}^{(1)}\mathrm{{:}\ rhse}$	$[q_{11}]$	r_{11}^+, r_{11}^-
$\mathrm{rhsf_3}^{(2)} + \mathrm{rhse} \leftrightarrows \mathrm{rhsf_3}^{(2)}\mathrm{{:}\ rhse}$	$[q_{12}]$	r_{12}^+, r_{12}^-
$\mathrm{rhsf_3}^{(3)} + \mathrm{rhse} \leftrightarrows \mathrm{rhsf_3}^{(3)}\mathrm{{:}\ rhse}$	$[q_{13}]$	r_{13}^+, r_{13}^-
$\mathrm{rhsf_3{:}\ rhse} \rightarrow \mathrm{rhsf_3{:}\ rhse} + \mathrm{rhsp}$	$[q_{14}]$	r_{14}
$\mathrm{rhsf_3}^{(1)}\mathrm{{:}\ rhse} \rightarrow \mathrm{rhsf_3}^{(1)}\mathrm{{:}\ rhse} + \mathrm{rhsp}$	$[q_{15}]$	r_{15}
$\mathrm{rhsf_3}^{(2)}\mathrm{{:}\ rhse} \rightarrow \mathrm{rhsf_3}^{(2)}\mathrm{{:}\ rhse} + \mathrm{rhsp}$	$[q_{16}]$	r_{16}
$\mathrm{rhsf_3}^{(3)}\mathrm{{:}\ rhse} \rightarrow \mathrm{rhsf_3}^{(3)}\mathrm{{:}\ rhse} + \mathrm{rhsp}$	$[q_{17}]$	r_{17}
$\mathrm{rhsp} + \mathrm{rhsf} \leftrightarrows \mathrm{rhsp{:}\ rhsf}$	$[q_{18}]$	r_{18}^+, r_{18}^-
$\mathrm{rhsp} + \mathrm{rhsf}^{(1)} \leftrightarrows \mathrm{rhsp{:}\ rhsf}^{(1)}$	$[q_{19}]$	r_{19}^+, r_{19}^-
$\mathrm{rhsp} + \mathrm{rhsf_2} \rightarrow \mathrm{rhsp{:}\ rhsf} + \mathrm{rhsf}$	$[q_{20}]$	r_{20}
$\mathrm{rhsp} + \mathrm{rhsf_2}^{(1)} \rightarrow \mathrm{rhsp{:}\ rhsf} + \mathrm{rhsf}^{(1)}$	$[q_{21}]$	r_{21}
$\mathrm{rhsp} + \mathrm{rhsf_2}^{(1)} \rightarrow \mathrm{rhsp{:}\ rhsf}^{(1)} + \mathrm{rhsf}$	$[q_{22}]$	r_{22}
$\mathrm{rhsp} + \mathrm{rhsf_2}^{(2)} \rightarrow \mathrm{rhsp{:}\ rhsf}^{(1)} + \mathrm{rhsf}^{(1)}$	$[q_{23}]$	r_{23}
$\mathrm{rhsp} + \mathrm{rhsf_3} \rightarrow \mathrm{rhsp{:}\ rhsf} + 2 * \mathrm{rhsf}$	$[q_{24}]$	r_{24}
$\mathrm{rhsp} + \mathrm{rhsf_3}^{(1)} \rightarrow \mathrm{rhsp{:}\ rhsf} + \mathrm{rhsf}^{(1)} + \mathrm{rhsf}$	$[q_{25}]$	r_{25}
$\mathrm{rhsp} + \mathrm{rhsf_3}^{(1)} \rightarrow \mathrm{rhsp{:}\ rhsf}^{(1)} + 2 * \mathrm{rhsf}$	$[q_{26}]$	r_{26}
$\mathrm{rhsp} + \mathrm{rhsf_3}^{(2)} \rightarrow \mathrm{rhsp{:}\ rhsf} + 2\mathrm{rhsf}^{(1)}$	$[q_{27}]$	r_{27}
$\mathrm{rhsp} + \mathrm{rhsf_3}^{(2)} \rightarrow \mathrm{rhsp{:}\ rhsf}^{(1)} + \mathrm{rhsf}^{(1)} + \mathrm{rhsf}$	$[q_{28}]$	r_{28}
$\mathrm{rhsp} + \mathrm{rhsf_3}^{(3)} \rightarrow \mathrm{rhsp{:}\ rhsf}^{(1)} + 2\mathrm{rhsf}^{(1)}$	$[q_{29}]$	r_{29}
$\mathrm{rhsp} + \mathrm{rhsf_3{:}\ rhse} \rightarrow \mathrm{rhsp{:}\ rhsf} + 2\,\mathrm{rhsf} + \mathrm{rhse}$	$[q_{30}]$	r_{30}
$\mathrm{rhsp} + \mathrm{rhsf_3}^{(1)}\mathrm{{:}\ rhse} \rightarrow \mathrm{rhsp{:}\ rhsf}^{(1)} + 2\,\mathrm{rhsf} + \mathrm{rhse}$	$[q_{31}]$	r_{31}
$\mathrm{rhsp} + \mathrm{rhsf_3}^{(1)}\mathrm{{:}\ rhse} \rightarrow \mathrm{rhsp{:}\ rhsf} + \mathrm{rhsf}^{(1)} + \mathrm{rhsf} + \mathrm{rhse}$	$[q_{32}]$	r_{32}
$\mathrm{rhsp} + \mathrm{rhsf_3}^{(2)}\mathrm{{:}\ rhse} \rightarrow \mathrm{rhsp{:}\ rhsf}^{(1)} + \mathrm{rhsf}^{(1)} + \mathrm{rhsf} + \mathrm{rhse}$	$[q_{33}]$	r_{33}
$\mathrm{rhsp} + \mathrm{rhsf_3}^{(2)}\mathrm{{:}\ rhse} \rightarrow \mathrm{rhsp{:}\ rhsf} + 2\mathrm{rhsf}^{(1)} + \mathrm{rhse}$	$[q_{34}]$	r_{34}
$\mathrm{rhsp} + \mathrm{rhsf_3}^{(3)}\mathrm{{:}\ rhse} \rightarrow \mathrm{rhsp{:}\ rhsf}^{(1)} + 2\mathrm{rhsf}^{(1)} + \mathrm{rhse}$	$[q_{35}]$	r_{35}
$\mathrm{rhsp} \rightarrow \emptyset$	$[q_{36}]$	r_{36}
$\mathrm{rprot} \rightarrow \mathrm{rmfp}$	$[q_{37}]$	r_{37}
$\mathrm{rhsp} + \mathrm{rmfp} \leftrightarrows \mathrm{rhsp{:}\ rmfp}$	$[q_{38}]$	r_{38}^+, r_{38}^-
$\mathrm{rhsp{:}\ rmfp} \rightarrow \mathrm{rhsp} + \mathrm{rprot}$	$[q_{39}]$	r_{39}

TABLE 3
The list of reactions for the refined model - Continued

C THE ODE-BASED MODEL OF THE REFINED HEAT SHOCK RESPONSE MODEL

$$
\begin{aligned}
d[\mathrm{rhsf}]/dt = {} & -2r_1^+[\mathrm{rhsf}]^2 + 2r_1^-[\mathrm{rhsf_2}] - r_2^+[\mathrm{rhsf}][\mathrm{rhsf}^{(1)}] + r_2^-[\mathrm{rhsf_2}^{(1)}] \\
& - r_4{}^+[\mathrm{rhsf}][\mathrm{rhsf_2}] + r_4{}^-[\mathrm{rhsf_3}] - r_6{}^+[\mathrm{rhsf}][\mathrm{rhsf_2}^{(1)}] + r_6{}^-[\mathrm{rhsf_3}^{(1)}] \\
& - r_8{}^+[\mathrm{rhsf}][\mathrm{rhsf_2}^{(2)}] + r_8{}^-[\mathrm{rhsf_3}^{(2)}] - r_{18}{}^+[\mathrm{rhsp}][\mathrm{rhsf}] \\
& + r_{18}{}^-[\mathrm{rhsp{:}\ rhsf}] + r_{20}[\mathrm{rhsp}][\mathrm{rhsf_2}] + r_{22}[\mathrm{rhsp}][\mathrm{rhsf_2}^{(1)}] \\
& + 2r_{24}[\mathrm{rhsp}][\mathrm{rhsf_3}] + r_{25}[\mathrm{rhsp}][\mathrm{rhsf_3}^{(1)}] + 2r_{26}[\mathrm{rhsp}][\mathrm{rhsf_3}^{(1)}]
\end{aligned}
$$

TABLE 4
The system of differential equations of the mathematical model associated with the biochemical model

$$
\begin{aligned}
&+r_{28}[\text{rhsp}][\text{rhsf}_3{}^{(2)}] + 2r_{30}[\text{rhsp}][\text{rhsf}_3\text{: rhse}]\\
&+2r_{31}[\text{rhsp}][\text{rhsf}_3{}^{(1)}\text{: rhse}] + r_{32}[\text{rhsp}][\text{rhsf}_3{}^{(1)}\text{: rhse}]\\
&+r_{33}[\text{rhsp}][\text{rhsf}_3{}^{(2)}\text{: rhse}]\\
d[\text{rhsf}^{(1)}]/dt &= -r_2^+[\text{rhsf}][\text{rhsf}^{(1)}] + r_2^-[\text{rhsf}_2{}^{(1)}] - 2r_3{}^+[\text{rhsf}^{(1)}]^2\\
&+2r_3{}^-[\text{rhsf}_2{}^{(2)}] - r_5{}^+[\text{rhsf}^{(1)}][\text{rhsf}_2] + r_5{}^-[\text{rhsf}_3{}^{(1)}]\\
&-r_7{}^+[\text{rhsf}^{(1)}][\text{rhsf}_2{}^{(1)}] + r_7{}^-[\text{rhsf}_3{}^{(2)}] - r_9{}^+[\text{rhsf}^{(1)}][\text{rhsf}_2{}^{(2)}]\\
&+r_9{}^-[\text{rhsf}_3{}^{(3)}] - r_{19}{}^+[\text{rhsp}][\text{rhsf}^{(1)}] + r_{19}{}^-[\text{rhsp: rhsf}^{(1)}]\\
&+r_{21}[\text{rhsp}][\text{rhsf}_2{}^{(1)}] + r_{23}[\text{rhsp}][\text{rhsf}_2{}^{(2)}] + r_{25}[\text{rhsp}][\text{rhsf}_3{}^{(1)}]\\
&+2r_{27}[\text{rhsp}][\text{rhsf}_3{}^{(2)}] + r_{28}[\text{rhsp}][\text{rhsf}_3{}^{(2)}] + 2r_{29}[\text{rhsp}][\text{rhsf}_3{}^{(3)}]\\
&+r_{32}[\text{rhsp}][\text{rhsf}_3{}^{(1)}\text{: rhse}] + r_{33}[\text{rhsp}][\text{rhsf}_3{}^{(2)}\text{: rhse}]\\
&+2r_{34}[\text{rhsp}][\text{rhsf}_3{}^{(2)}\text{: rhse}] + 2r_{35}[\text{rhsp}][\text{rhsf}_3{}^{(3)}\text{: rhse}]\\
d[\text{rhsf}_2]/dt &= r_1^+[\text{rhsf}]^2 - r_1^-[\text{rhsf}_2] - r_4^+[\text{rhsf}][\text{rhsf}_2] + r_4^-[\text{rhsf}_3]\\
&-r_5^+[\text{rhsf}^{(1)}][\text{rhsf}_2] + r_5^-[\text{rhsf}_3{}^{(1)}] - r_{20}[\text{rhsp}][\text{rhsf}_2]\\
d[\text{rhsf}_2{}^{(1)}]/dt &= r_2^+[\text{rhsf}][\text{rhsf}^{(1)}] - r_2^-[\text{rhsf}_2{}^{(1)}] - r_6^+[\text{rhsf}][\text{rhsf}_2{}^{(1)}]\\
&+r_6^-[\text{rhsf}_3{}^{(1)}] - r_7^+[\text{rhsf}^{(1)}][\text{rhsf}_2{}^{(1)}] + r_7^-[\text{rhsf}_3{}^{(2)}]\\
&-r_{21}[\text{rhsp}][\text{rhsf}_2{}^{(1)}] - r_{22}[\text{rhsp}][\text{rhsf}_2{}^{(1)}]\\
d[\text{rhsf}_2{}^{(2)}]/dt &= r_3^+[\text{rhsf}^{(1)}]^2 - r_3^-[\text{rhsf}_2{}^{(2)}] - r_8^+[\text{rhsf}][\text{rhsf}_2{}^{(2)}]\\
&+r_8^-[\text{rhsf}_3{}^{(2)}] - r_9^+[\text{rhsf}^{(1)}][\text{rhsf}_2{}^{(2)}] + r_9^-[\text{rhsf}_3{}^{(3)}]\\
&-r_{23}[\text{rhsp}][\text{rhsf}_2{}^{(2)}]\\
d[\text{rhsf}_3]/dt &= r_4^+[\text{rhsf}][\text{rhsf}_2] - r_4^-[\text{rhsf}_3] - r_{10}^+[\text{rhsf}_3][\text{rhse}]\\
&+r_{10}^-[\text{rhsf}_3\text{: rhse}] - r_{24}[\text{rhsp}][\text{rhsf}_3]\\
d[\text{rhsf}_3{}^{(1)}]/dt &= r_5^+[\text{rhsf}^{(1)}][\text{rhsf}_2] - r_5^-[\text{rhsf}_3{}^{(1)}] + r_6^+[\text{rhsf}][\text{rhsf}_2{}^{(1)}]\\
&-r_6^-[\text{rhsf}_3{}^{(1)}] - r_{11}^+[\text{rhsf}_3{}^{(1)}][\text{rhse}] + r_{11}^-[\text{rhsf}_3{}^{(1)}\text{: rhse}]\\
&-r_{25}[\text{rhsp}][\text{rhsf}_3{}^{(1)}] - r_{26}[\text{rhsp}][\text{rhsf}_3{}^{(1)}]\\
d[\text{rhsf}_3{}^{(2)}]/dt &= r_7^+[\text{rhsf}^{(1)}][\text{rhsf}_2{}^{(1)}] - r_7^-[\text{rhsf}_3{}^{(2)}] + r_8^+[\text{rhsf}][\text{rhsf}_2{}^{(2)}]\\
&-r_8^-[\text{rhsf}_3{}^{(2)}] - r_{12}^+[\text{rhsf}_3{}^{(2)}][\text{rhse}] + r_{12}^-[\text{rhsf}_3{}^{(2)}\text{: rhse}]\\
&-r_{27}[\text{rhsp}][\text{rhsf}_3{}^{(2)}] - r_{28}[\text{rhsp}][\text{rhsf}_3{}^{(2)}]\\
d[\text{rhsf}_3{}^{(3)}]/dt &= r_9^+[\text{rhsf}^{(1)}][\text{rhsf}_2{}^{(2)}] - r_9^-[\text{rhsf}_3{}^{(3)}] - r_{13}^+[\text{rhsf}_3{}^{(3)}][\text{rhse}]\\
&+r_{13}^-[\text{rhsf}_3{}^{(3)}\text{: rhse}] - r_{29}[\text{rhsp}][\text{rhsf}_3{}^{(3)}]\\
d[\text{rhse}]/dt &= -r_{10}^+[\text{rhsf}_3][\text{rhse}] + r_{10}^-[\text{rhsf}_3\text{: rhse}] - r_{11}^+[\text{rhsf}_3{}^{(1)}][\text{rhse}]\\
&+r_{11}^-[\text{rhsf}_3{}^{(1)}\text{: rhse}] - r_{12}^+[\text{rhsf}_3{}^{(2)}][\text{rhse}] + r_{12}^-[\text{rhsf}_3{}^{(2)}\text{: rhse}]\\
&-r_{13}^+[\text{rhsf}_3{}^{(3)}][\text{rhse}] + r_{13}^-[\text{rhsf}_3{}^{(3)}\text{: rhse}] + r_{30}[\text{rhsp}][\text{rhsf}_3\text{: rhse}]\\
&+r_{31}[\text{rhsp}][\text{rhsf}_3{}^{(1)}\text{: rhse}] + r_{32}[\text{rhsp}][\text{rhsf}_3{}^{(1)}\text{: rhse}]\\
&+r_{33}[\text{rhsp}][\text{rhsf}_3{}^{(2)}\text{: rhse}] + r_{34}[\text{rhsp}][\text{rhsf}_3{}^{(2)}\text{: rhse}]\\
&+r_{35}[\text{rhsp}][\text{rhsf}_3{}^{(3)}\text{: rhse}]\\
d[\text{rhsf}_3\text{: rhse}]/dt &= r_{10}^+[\text{rhsf}_3][\text{rhse}] - r_{10}^-[\text{rhsf}_3\text{: rhse}]\\
&-r_{30}[\text{rhsp}][\text{rhsf}_3\text{: rhse}]\\
d[\text{rhsf}_3{}^{(1)}\text{: rhse}]/dt &= r_{11}^+[\text{rhsf}_3{}^{(1)}][\text{rhse}] - r_{11}^-[\text{rhsf}_3{}^{(1)}\text{: rhse}]\\
&-r_{31}[\text{rhsp}][\text{rhsf}_3{}^{(1)}\text{: rhse}] - r_{32}[\text{rhsp}][\text{rhsf}_3{}^{(1)}\text{: rhse}]\\
d[\text{rhsf}_3{}^{(2)}\text{: rhse}]/dt &= r_{12}^+[\text{rhsf}_3{}^{(2)}][\text{rhse}] - r_{12}^-[\text{rhsf}_3{}^{(2)}\text{: rhse}]\\
&-r_{33}[\text{rhsp}][\text{rhsf}_3{}^{(2)}\text{: rhse}] - r_{34}[\text{rhsp}][\text{rhsf}_3{}^{(2)}\text{: rhse}]\\
d[\text{rhsf}_3{}^{(3)}\text{: rhse}]/dt &= r_{13}^+[\text{rhsf}_3{}^{(3)}][\text{rhse}] - r_{13}^-[\text{rhsf}_3{}^{(3)}\text{: rhse}]\\
&-r_{35}[\text{rhsp}][\text{rhsf}_3{}^{(3)}\text{: rhse}]
\end{aligned}
$$

TABLE 4
The system of differential equations of the mathematical model associated with the biochemical model - Continued

$$
\begin{aligned}
d[\mathrm{rhsp}]/dt &= r_{14}[\mathrm{rhsf_3{:}\,rhse}] + r_{15}[\mathrm{rhsf_3}^{(1)}{:}\,\mathrm{rhse}] + r_{16}[\mathrm{rhsf_3}^{(2)}{:}\,\mathrm{rhse}] \\
&\quad + r_{17}[\mathrm{rhsf_3}^{(3)}{:}\,\mathrm{rhse}] - r_{18}^{+}[\mathrm{rhsp}][\mathrm{rhsf}] + r_{18}^{-}[\mathrm{rhsp{:}\,rhsf}] \\
&\quad - r_{19}^{+}[\mathrm{rhsp}][\mathrm{rhsf}^{(1)}] + r_{19}^{-}[\mathrm{rhsp{:}\,rhsf}^{(1)}] - r_{20}[\mathrm{rhsp}][\mathrm{rhsf_2}] \\
&\quad - r_{21}[\mathrm{rhsp}][\mathrm{rhsf_2}^{(1)}] - r_{22}[\mathrm{rhsp}][\mathrm{rhsf_2}^{(1)}] - r_{23}[\mathrm{rhsp}][\mathrm{rhsf_2}^{(2)}] \\
&\quad - r_{24}[\mathrm{rhsp}][\mathrm{rhsf_3}] - r_{25}[\mathrm{rhsp}][\mathrm{rhsf_3}^{(1)}] - r_{26}[\mathrm{rhsp}][\mathrm{rhsf_3}^{(1)}] \\
&\quad - r_{27}[\mathrm{rhsp}][\mathrm{rhsf_3}^{(2)}] - r_{28}[\mathrm{rhsp}][\mathrm{rhsf_3}^{(2)}] - r_{29}[\mathrm{rhsp}][\mathrm{rhsf_3}^{(3)}] \\
&\quad - r_{30}[\mathrm{rhsp}][\mathrm{rhsf_3{:}\,rhse}] - r_{31}[\mathrm{rhsp}][\mathrm{rhsf_3}^{(1)}{:}\,\mathrm{rhse}] \\
&\quad - r_{32}[\mathrm{rhsp}][\mathrm{rhsf_3}^{(1)}{:}\,\mathrm{rhse}] - r_{33}[\mathrm{rhsp}][\mathrm{rhsf_3}^{(2)}{:}\,\mathrm{rhse}] \\
&\quad - r_{34}[\mathrm{rhsp}][\mathrm{rhsf_3}^{(2)}{:}\,\mathrm{rhse}] - r_{35}[\mathrm{rhsp}][\mathrm{rhsf_3}^{(3)}{:}\,\mathrm{rhse}] - r_{36}[\mathrm{rhsp}] \\
&\quad - r_{38}^{+}[\mathrm{rhsp}][\mathrm{rmfp}] + r_{38}^{-}[\mathrm{rhsp{:}\,rmfp}] + r_{39}[\mathrm{rhsp}][\mathrm{rmfp}] \\
d[\mathrm{rhsp{:}\,rhsf}]/dt &= r_{18}^{+}[\mathrm{rhsp}][\mathrm{rhsf}] - r_{18}^{-}[\mathrm{rhsp{:}\,rhsf}] + r_{20}[\mathrm{rhsp}][\mathrm{rhsf_2}] \\
&\quad + r_{21}[\mathrm{rhsp}][\mathrm{rhsf_2}^{(1)}] + r_{24}[\mathrm{rhsp}][\mathrm{rhsf_3}] \\
&\quad + r_{25}[\mathrm{rhsp}][\mathrm{rhsf_3}^{(1)}] + r_{27}[\mathrm{rhsp}][\mathrm{rhsf_3}^{(2)}] \\
&\quad + r_{30}[\mathrm{rhsp}][\mathrm{rhsf_3{:}\,rhse}] + r_{32}[\mathrm{rhsp}][\mathrm{rhsf_3}^{(1)}{:}\,\mathrm{rhse}] \\
&\quad + r_{34}[\mathrm{rhsp}][\mathrm{rhsf_3}^{(2)}{:}\,\mathrm{rhse}] \\
d[\mathrm{rhsp{:}\,rhsf}^{(1)}]/dt &= r_{19}{}^{+}[\mathrm{rhsp}][\mathrm{rhsf}^{(1)}] - r_{19}{}^{-}[\mathrm{rhsp{:}\,rhsf}^{(1)}] \\
&\quad + r_{22}[\mathrm{rhsp}][\mathrm{rhsf_2}^{(1)}] + r_{23}[\mathrm{rhsp}][\mathrm{rhsf_2}^{(2)}] \\
&\quad + r_{26}[\mathrm{rhsp}][\mathrm{rhsf_3}^{(1)}] + r_{28}[\mathrm{rhsp}][\mathrm{rhsf_3}^{(2)}] \\
&\quad + r_{29}[\mathrm{rhsp}][\mathrm{rhsf_3}^{(3)}] + r_{31}[\mathrm{rhsp}][\mathrm{rhsf_3}^{(1)}{:}\,\mathrm{rhse}] \\
&\quad + r_{33}[\mathrm{rhsp}][\mathrm{rhsf_3}^{(2)}{:}\,\mathrm{rhse}] + r_{35}[\mathrm{rhsp}][\mathrm{rhsf_3}^{(3)}{:}\,\mathrm{rhse}] \\
d[\mathrm{rhsp{:}\,rmfp}]/dt &= r_{38}^{+}[\mathrm{rhsp}][\mathrm{rmfp}] - (r_{38}^{-} + r_{39})[\mathrm{rhsp{:}\,rmfp}] \\
d[\mathrm{rmfp}]/dt &= r_{37}[\mathrm{rprot}] - r_{38}^{+}[\mathrm{rhsp}][\mathrm{rmfp}] + r_{38}^{-}[\mathrm{rhsp{:}\,rmfp}] \\
d[\mathrm{rprot}]/dt &= -r_{37}[\mathrm{rprot}] + r_{39}[\mathrm{rhsp{:}\,rmfp}]
\end{aligned}
$$

TABLE 4
The system of differential equations of the mathematical model associated with the biochemical model - Continued

$$
\begin{aligned}
d[\mathrm{Rhsf}]/dt &= -2(r_1^{+}[\mathrm{rhsf}]^2 + r_2^{+}[\mathrm{rhsf}][\mathrm{rhsf}^{(1)}] + r_3{}^{+}[\mathrm{rhsf}^{(1)}]^2) + 2(r_1^{-}[\mathrm{rhsf_2}] \\
&\quad + r_2^{-}[\mathrm{rhsf_2}^{(1)}] + r_3{}^{-}[\mathrm{rhsf_2}^{(2)}]) - (r_4{}^{+}[\mathrm{rhsf}][\mathrm{rhsf_2}] \\
&\quad + r_6{}^{+}[\mathrm{rhsf}][\mathrm{rhsf_2}^{(1)}] + r_8{}^{+}[\mathrm{rhsf}][\mathrm{rhsf_2}^{(2)}] + r_5{}^{+}[\mathrm{rhsf}^{(1)}][\mathrm{rhsf_2}] \\
&\quad + r_7{}^{+}[\mathrm{rhsf}^{(1)}][\mathrm{rhsf_2}^{(1)}] + r_9{}^{+}[\mathrm{rhsf}^{(1)}][\mathrm{rhsf_2}^{(2)}]) + (r_4{}^{-}[\mathrm{rhsf_3}] + (r_5{}^{-} \\
&\quad + r_6{}^{-})[\mathrm{rhsf_3}^{(1)}] + (r_7{}^{-} + r_8{}^{-})[\mathrm{rhsf_3}^{(2)}] + r_9{}^{-}[\mathrm{rhsf_3}^{(3)}]) \\
&\quad - [\mathrm{rhsp}](r_{18}{}^{+}[\mathrm{rhsf}] + r_{19}{}^{+}[\mathrm{rhsf}^{(1)}]) + (r_{18}{}^{-}[\mathrm{rhsp{:}\,rhsf}] \\
&\quad + r_{19}{}^{-}[\mathrm{rhsp{:}\,rhsf}^{(1)}]) + [\mathrm{rhsp}](r_{20}[\mathrm{rhsf_2}] + (r_{21} + r_{22})[\mathrm{rhsf_2}^{(1)}] \\
&\quad + r_{23}[\mathrm{rhsf_2}^{(2)}]) + 2[\mathrm{rhsp}](r_{24}[\mathrm{rhsf_3}] + (r_{25} + r_{26})[\mathrm{rhsf_3}^{(1)}] + (r_{27} \\
&\quad + r_{28})[\mathrm{rhsf_3}^{(2)}] + r_{29}[\mathrm{rhsf_3}^{(3)}]) + 2[\mathrm{rhsp}](r_{30}[\mathrm{rhsf_3{:}\,rhse}] \\
&\quad + (r_{31} + r_{32})[\mathrm{rhsf_3}^{(1)}{:}\,\mathrm{rhse}] + (r_{33} + r_{34})[\mathrm{rhsf_3}^{(2)}{:}\,\mathrm{rhse}] \\
&\quad + r_{35}[\mathrm{rhsf_3}^{(3)}{:}\,\mathrm{rhse}]) \\
d[\mathrm{Rhsf_2}]/dt &= (r_1^{+}[\mathrm{rhsf}]^2 + r_2^{+}[\mathrm{rhsf}][\mathrm{rhsf}^{(1)}] + r_3^{+}[\mathrm{rhsf}^{(1)}]^2) - (r_1^{-}[\mathrm{rhsf_2}] \\
&\quad + r_2^{-}[\mathrm{rhsf_2}^{(1)}] + r_3^{-}[\mathrm{rhsf_2}^{(2)}]) - (r_4^{+}[\mathrm{rhsf}][\mathrm{rhsf_2}] + r_6^{+}[\mathrm{rhsf}][\mathrm{rhsf_2}^{(1)}] \\
&\quad + r_8^{+}[\mathrm{rhsf}][\mathrm{rhsf_2}^{(2)}] + r_5^{+}[\mathrm{rhsf}^{(1)}][\mathrm{rhsf_2}] + r_7^{+}[\mathrm{rhsf}^{(1)}][\mathrm{rhsf_2}^{(1)}]
\end{aligned}
$$

TABLE 5
The system of ODEs corresponding to Rhsf, $\mathrm{Rhsf_2}$, $\mathrm{Rhsf_3}$, $\mathrm{Rhsf_3}$: Rhse, and Rhsp: Rhsf in the refined model

$$
\begin{aligned}
&+r_9^+[\mathrm{rhsf}^{(1)}][{\mathrm{rhsf}_2}^{(2)}]) + (r_4^-[\mathrm{rhsf}_3] + (r_5^- + r_6^-)[{\mathrm{rhsf}_3}^{(1)}])\\
&+(r_7^- + r_8^-)[{\mathrm{rhsf}_3}^{(2)}] + r_9^-[{\mathrm{rhsf}_3}^{(3)}]) - [\mathrm{rhsp}](r_{20}[\mathrm{rhsf}_2]\\
&+(r_{21} + r_{22})[{\mathrm{rhsf}_2}^{(1)}] + r_{23}[{\mathrm{rhsf}_2}^{(2)}])\\
d[\mathrm{Rhsf}_3]/dt &= (r_4^+[\mathrm{rhsf}][\mathrm{rhsf}_2] + r_6^+[\mathrm{rhsf}][{\mathrm{rhsf}_2}^{(1)}] + r_8^+[\mathrm{rhsf}][{\mathrm{rhsf}_2}^{(2)}]\\
&+r_5^+[\mathrm{rhsf}^{(1)}][\mathrm{rhsf}_2] + r_7^+[\mathrm{rhsf}^{(1)}][{\mathrm{rhsf}_2}^{(1)}] + r_9^+[\mathrm{rhsf}^{(1)}][{\mathrm{rhsf}_2}^{(2)}])\\
&-(r_4^-[\mathrm{rhsf}_3] + (r_5^- + r_6^-)[{\mathrm{rhsf}_3}^{(1)}] + (r_7^- + r_8^-)[{\mathrm{rhsf}_3}^{(2)}]\\
&+r_9^-[{\mathrm{rhsf}_3}^{(3)}]) - [\mathrm{rhse}](r_{10}^+[\mathrm{rhsf}_3] + r_{11}^+[{\mathrm{rhsf}_3}^{(1)}] + r_{12}^+[{\mathrm{rhsf}_3}^{(2)}]\\
&+r_{13}^+[{\mathrm{rhsf}_3}^{(3)}]) + (r_{10}^-[\mathrm{rhsf}_3\colon \mathrm{rhse}] + r_{11}^-[{\mathrm{rhsf}_3}^{(1)}\colon \mathrm{rhse}]\\
&+r_{12}^-[{\mathrm{rhsf}_3}^{(2)}\colon \mathrm{rhse}] + r_{13}^-[{\mathrm{rhsf}_3}^{(3)}\colon \mathrm{rhse}]) - [\mathrm{rhsp}](r_{24}[\mathrm{rhsf}_3]\\
&+(r_{25} + r_{26})[{\mathrm{rhsf}_3}^{(1)}] + (r_{27} + r_{28})[{\mathrm{rhsf}_3}^{(2)}] + r_{29}[{\mathrm{rhsf}_3}^{(3)}])\\
d[\mathrm{Rhsf}_3\colon \mathrm{Rhse}]/dt &= [\mathrm{rhse}](r_{10}^+[\mathrm{rhsf}_3] + r_{11}^+[{\mathrm{rhsf}_3}^{(1)}] + r_{12}^+[{\mathrm{rhsf}_3}^{(2)}]\\
&+r_{13}^+[{\mathrm{rhsf}_3}^{(3)}]) - (r_{10}^-[\mathrm{rhsf}_3\colon \mathrm{rhse}] + r_{11}^-[{\mathrm{rhsf}_3}^{(1)}\colon \mathrm{rhse}]\\
&+r_{12}^-[{\mathrm{rhsf}_3}^{(2)}\colon \mathrm{rhse}] + r_{13}^-[{\mathrm{rhsf}_3}^{(3)}\colon \mathrm{rhse}]) - [\mathrm{rhsp}](r_{30}[\mathrm{rhsf}_3\colon \mathrm{rhse}]\\
&+(r_{31} + r_{32})[{\mathrm{rhsf}_3}^{(1)}\colon \mathrm{rhse}] + (r_{33} + r_{34})[{\mathrm{rhsf}_3}^{(2)}\colon \mathrm{rhse}]\\
&+r_{35}[{\mathrm{rhsf}_3}^{(3)}\colon \mathrm{rhse}])\\
d[\mathrm{Rhsp}\colon \mathrm{Rhsf}]/dt &= [\mathrm{rhsp}](r_{18}^+[\mathrm{rhsf}] + {r_{19}}^+[\mathrm{rhsf}^{(1)}]) - (r_{18}^-[\mathrm{rhsp}\colon \mathrm{rhsf}]\\
&+{r_{19}}^-[\mathrm{rhsp}\colon \mathrm{rhsf}^{(1)}]) + [\mathrm{rhsp}](r_{20}[\mathrm{rhsf}_2] + (r_{21} + r_{22})[{\mathrm{rhsf}_2}^{(1)}]\\
&+r_{23}[{\mathrm{rhsf}_2}^{(2)}]) + [\mathrm{rhsp}](r_{24}[\mathrm{rhsf}_3] + (r_{25} + r_{26})[{\mathrm{rhsf}_3}^{(1)}]\\
&+(r_{27} + r_{28})[{\mathrm{rhsf}_3}^{(2)}] + r_{29}[{\mathrm{rhsf}_3}^{(3)}])\\
&+[\mathrm{rhsp}](r_{30}[\mathrm{rhsf}_3\colon \mathrm{rhse}] + (r_{31} + r_{32})[{\mathrm{rhsf}_3}^{(1)}\colon \mathrm{rhse}]\\
&(r_{33} + r_{34})[{\mathrm{rhsf}_3}^{(2)}\colon \mathrm{rhse}] + r_{35}[{\mathrm{rhsf}_3}^{(3)}\colon \mathrm{rhse}])
\end{aligned}
$$

TABLE 5
The system of ODEs corresponding to Rhsf, Rhsf_2, Rhsf_3, Rhsf_3: Rhse, and Rhsp: Rhsf in the refined model - Continued

D THE NUMERICAL SETUP OF THE REFINED MODEL

$r_1^+ = k_1^+;$	$r_8^+ = k_2^+;$	$r_{16} = k_4;$	$r_{28} = k_7/2;$
$r_1^- = k_1^-;$	$r_8^- = k_2^-/2;$	$r_{17} = k_4;$	$r_{29} = k_7;$
$r_2^+ = 2 \cdot k_1^+;$	$r_9^+ = k_2^+;$	${r_{18}}^+ = k_5^+;$	$r_{30} = k_8;$
$r_2^- = k_1^-;$	$r_9^- = k_2^-;$	${r_{18}}^- = k_5^-;$	$r_{31} = k_8/2;$
$r_3^+ = k_1^+;$	${r_{10}}^+ = k_3^+;$	${r_{19}}^+ = k_5^+;$	$r_{32} = k_8/2;$
$r_3^- = k_1^-;$	${r_{10}}^- = k_3^-;$	${r_{19}}^- = k_5^-;$	$r_{33} = k_8/2;$
$r_4^+ = k_2^+;$	${r_{11}}^+ = k_3^+;$	$r_{20} = k_6;$	$r_{34} = k_8/2;$
$r_4^- = k_2^-;$	${r_{11}}^- = k_3^-;$	$r_{21} = k_6/2;$	$r_{35} = k_8;$
$r_5^+ = k_2^+;$	${r_{12}}^+ = k_3^+;$	$r_{22} = k_6/2;$	$r_{36} = k_9;$
$r_5^- = k_2^-/2;$	${r_{12}}^- = k_3^-;$	$r_{23} = k_6;$	$r_{37} = \Phi_T;$
$r_6^+ = k_2^+;$	${r_{13}}^+ = k_3^+;$	$r_{24} = k_7;$	$r_{38}^- = {k_{11}}^-;$
$r_6^- = k_2^-/2;$	${r_{13}}^- = k_3^-;$	$r_{25} = k_7/2;$	$r_{38}^+ = {k_{11}}^+;$
$r_7^+ = k_2^+;$	$r_{14} = k_4;$	$r_{26} = k_7/2;$	$r_{39} = k_{12}$
$r_7^- = k_2^-/2;$	$r_{15} = k_4;$	$r_{27} = k_7/2;$	

TABLE 6
The numerical values of the parameters of the refined model

E THE FULL MOLECULAR MODEL FOR THE ACETYLATION-INDUCED HEAT SHOCK RESPONSE

Reaction
$2\,\mathrm{rhsf} \leftrightarrows \mathrm{rhsf}_2$
$\mathrm{rhsf} + \mathrm{rhsf}^{(1)} \leftrightarrows \mathrm{rhsf}_2{}^{(1)}$
$2\mathrm{rhsf}^{(1)} \leftrightarrows \mathrm{rhsf}_2{}^{(2)}$
$\mathrm{rhsf} + \mathrm{rhsf}_2 \leftrightarrows \mathrm{rhsf}_3$
$\mathrm{rhsf}^{(1)} + \mathrm{rhsf}_2 \leftrightarrows \mathrm{rhsf}_3{}^{(1)}$
$\mathrm{rhsf} + \mathrm{rhsf}_2{}^{(1)} \leftrightarrows \mathrm{rhsf}_3{}^{(1)}$
$\mathrm{rhsf}^{(1)} + \mathrm{rhsf}_2{}^{(1)} \leftrightarrows \mathrm{rhsf}_3{}^{(2)}$
$\mathrm{rhsf} + \mathrm{rhsf}_2{}^{(2)} \leftrightarrows \mathrm{rhsf}_3{}^{(2)}$
$\mathrm{rhsf}^{(1)} + \mathrm{rhsf}_2{}^{(2)} \leftrightarrows \mathrm{rhsf}_3{}^{(3)}$
$\mathrm{rhsf}_3 + \mathrm{rhse} \leftrightarrows \mathrm{rhsf}_3\colon \mathrm{rhse}$
$\mathrm{rhsf}_3{}^{(1)} + \mathrm{rhse} \leftrightarrows \mathrm{rhsf}_3{}^{(1)}\colon \mathrm{rhse}$
$\mathrm{rhsf}_3{}^{(2)} + \mathrm{rhse} \leftrightarrows \mathrm{rhsf}_3{}^{(2)}\colon \mathrm{rhse}$
$\mathrm{rhsf}_3{}^{(3)} + \mathrm{rhse} \leftrightarrows \mathrm{rhsf}_3{}^{(3)}\colon \mathrm{rhse}$
$\mathrm{rhsf}_3\colon \mathrm{rhse} \rightarrow \mathrm{rhsf}_3\colon \mathrm{rhse} + \mathrm{rhsp}$
$\mathrm{rhsf}_3{}^{(1)}\colon \mathrm{rhse} \rightarrow \mathrm{rhsf}_3{}^{(1)}\colon \mathrm{rhse} + \mathrm{rhsp}$
$\mathrm{rhsf}_3{}^{(2)}\colon \mathrm{rhse} \rightarrow \mathrm{rhsf}_3{}^{(2)}\colon \mathrm{rhse} + \mathrm{rhsp}$
$\mathrm{rhsf}_3{}^{(3)}\colon \mathrm{rhse} \rightarrow \mathrm{rhsf}_3{}^{(3)}\colon \mathrm{rhse} + \mathrm{rhsp}$
$\mathrm{rhsp} + \mathrm{rhsf} \leftrightarrows \mathrm{rhsp}\colon \mathrm{rhsf}$
$\mathrm{rhsp} + \mathrm{rhsf}^{(1)} \leftrightarrows \mathrm{rhsp}\colon \mathrm{rhsf}^{(1)}$
$\mathrm{rhsp} + \mathrm{rhsf}_2 \rightarrow \mathrm{rhsp}\colon \mathrm{rhsf} + \mathrm{rhsf}$
$\mathrm{rhsp} + \mathrm{rhsf}_2{}^{(1)} \rightarrow \mathrm{rhsp}\colon \mathrm{rhsf} + \mathrm{rhsf}^{(1)}$
$\mathrm{rhsp} + \mathrm{rhsf}_2{}^{(1)} \rightarrow \mathrm{rhsp}\colon \mathrm{rhsf}^{(1)} + \mathrm{rhsf}$
$\mathrm{rhsp} + \mathrm{rhsf}_2{}^{(2)} \rightarrow \mathrm{rhsp}\colon \mathrm{rhsf}^{(1)} + \mathrm{rhsf}^{(1)}$
$\mathrm{rhsp} + \mathrm{rhsf}_3 \rightarrow \mathrm{rhsp}\colon \mathrm{rhsf} + 2 * \mathrm{rhsf}$
$\mathrm{rhsp} + \mathrm{rhsf}_3{}^{(1)} \rightarrow \mathrm{rhsp}\colon \mathrm{rhsf} + \mathrm{rhsf}^{(1)} + \mathrm{rhsf}$
$\mathrm{rhsp} + \mathrm{rhsf}_3{}^{(1)} \rightarrow \mathrm{rhsp}\colon \mathrm{rhsf}^{(1)} + 2 * \mathrm{rhsf}$
$\mathrm{rhsp} + \mathrm{rhsf}_3{}^{(2)} \rightarrow \mathrm{rhsp}\colon \mathrm{rhsf} + 2\mathrm{rhsf}^{(1)}$
$\mathrm{rhsp} + \mathrm{rhsf}_3{}^{(2)} \rightarrow \mathrm{rhsp}\colon \mathrm{rhsf}^{(1)} + \mathrm{rhsf}^{(1)} + \mathrm{rhsf}$
$\mathrm{rhsp} + \mathrm{rhsf}_3{}^{(3)} \rightarrow \mathrm{rhsp}\colon \mathrm{rhsf}^{(1)} + 2\mathrm{rhsf}^{(1)}$
$\mathrm{rhsf}_3\colon \mathrm{rhse} \rightarrow \mathrm{rhsf}_3 + \mathrm{rhse}$
$\mathrm{rhsf}_3{}^{(1)}\colon \mathrm{rhse} \rightarrow \mathrm{rhsf}_3{}^{(1)} + \mathrm{rhse}$
$\mathrm{rhsf}_3{}^{(2)}\colon \mathrm{rhse} \rightarrow \mathrm{rhsf}_3{}^{(2)} + \mathrm{rhse}$
$\mathrm{rhsf}_3{}^{(3)}\colon \mathrm{rhse} \rightarrow \mathrm{rhsf}_3{}^{(3)} + \mathrm{rhse}$
$\mathrm{rhsp} + \mathrm{rhsf}_3\colon \mathrm{rhse} \rightarrow \mathrm{rhsp}\colon \mathrm{rhsf} + 2 * \mathrm{rhsf} + \mathrm{rhse}$
$\mathrm{rhsp} + \mathrm{rhsf}_3{}^{(1)}\colon \mathrm{rhse} \rightarrow \mathrm{rhsp}\colon \mathrm{rhsf}^{(1)} + 2 * \mathrm{rhsf} + \mathrm{rhse}$
$\mathrm{rhsp} + \mathrm{rhsf}_3{}^{(1)}\colon \mathrm{rhse} \rightarrow \mathrm{rhsp}\colon \mathrm{rhsf} + \mathrm{rhsf} + \mathrm{rhsf}^{(1)} + \mathrm{rhse}$
$\mathrm{rhsp} + \mathrm{rhsf}_3{}^{(2)}\colon \mathrm{rhse} \rightarrow \mathrm{rhsp}\colon \mathrm{rhsf}^{(1)} + \mathrm{rhsf} + \mathrm{rhsf}^{(1)} + \mathrm{rhse}$
$\mathrm{rhsp} + \mathrm{rhsf}_3{}^{(2)}\colon \mathrm{rhse} \rightarrow \mathrm{rhsp}\colon \mathrm{rhsf} + 2 * \mathrm{rhsf}^{(1)} + \mathrm{rhse}$
$\mathrm{rhsp} + \mathrm{rhsf}_3{}^{(3)}\colon \mathrm{rhse} \rightarrow \mathrm{rhsp}\colon \mathrm{rhsf}^{(1)} + 2 * \mathrm{rhsf}^{(1)} + \mathrm{rhse}$
$\mathrm{rhsp} \rightarrow \emptyset$

TABLE 7
The final acetylation-modulated heat shock response model.

$\mathrm{rprot} \rightarrow \mathrm{rmfp}$
$\mathrm{rhsp} + \mathrm{rmfp} \leftrightarrows \mathrm{rhsp{:}\ rmfp}$
$\mathrm{rhsp{:}\ rmfp} \rightarrow \mathrm{rhsp} + \mathrm{rprot}$
$\mathrm{rhsf} \rightarrow \mathrm{rhsf}^{(1)}$
$\mathrm{rhsf}_2 \rightarrow {\mathrm{rhsf}_2}^{(1)}$
${\mathrm{rhsf}_2}^{(1)} \rightarrow {\mathrm{rhsf}_2}^{(2)}$
$\mathrm{rhsf}_3 \rightarrow {\mathrm{rhsf}_3}^{(1)}$
${\mathrm{rhsf}_3}^{(1)} \rightarrow {\mathrm{rhsf}_3}^{(2)}$
${\mathrm{rhsf}_3}^{(2)} \rightarrow {\mathrm{rhsf}_3}^{(3)}$
$\mathrm{rhsp{:}\ rhsf} \rightarrow \mathrm{rhsp{:}\ rhsf}^{(1)}$
$\mathrm{rhsf}_3\mathrm{{:}\ rhse} \rightarrow {\mathrm{rhsf}_3}^{(1)}\mathrm{{:}\ rhse}$
${\mathrm{rhsf}_3}^{(1)}\mathrm{{:}\ rhse} \rightarrow {\mathrm{rhsf}_3}^{(2)}\mathrm{{:}\ rhse}$
${\mathrm{rhsf}_3}^{(2)}\mathrm{{:}\ rhse} \rightarrow {\mathrm{rhsf}_3}^{(3)}\mathrm{{:}\ rhse}$
$\mathrm{rhsf}^{(1)} \rightarrow \mathrm{rhsf}$
${\mathrm{rhsf}_2}^{(1)} \rightarrow \mathrm{rhsf}_2$
${\mathrm{rhsf}_2}^{(2)} \rightarrow {\mathrm{rhsf}_2}^{(1)}$
${\mathrm{rhsf}_3}^{(1)} \rightarrow \mathrm{rhsf}_3$
${\mathrm{rhsf}_3}^{(2)} \rightarrow {\mathrm{rhsf}_3}^{(1)}$
${\mathrm{rhsf}_3}^{(3)} \rightarrow {\mathrm{rhsf}_3}^{(2)}$
$\mathrm{rhsp{:}\ rhsf}^{(1)} \rightarrow \mathrm{rhsp{:}\ rhsf}$
${\mathrm{rhsf}_3}^{(1)}\mathrm{{:}\ rhse} \rightarrow \mathrm{rhsf}_3\mathrm{{:}\ rhse}$
${\mathrm{rhsf}_3}^{(2)}\mathrm{{:}\ rhse} \rightarrow {\mathrm{rhsf}_3}^{(1)}\mathrm{{:}\ rhse}$
${\mathrm{rhsf}_3}^{(3)}\mathrm{{:}\ rhse} \rightarrow {\mathrm{rhsf}_3}^{(2)}\mathrm{{:}\ rhse}$

TABLE 7
The final acetylation-modulated heat shock response model - Continued

Publication II

ODE analysis of biological systems

Diana-Elena Gratie, Bogdan Iancu and Ion Petre

Originally published in: *Formal Methods for Dynamical Systems*, Editors: Marco Bernardo, Erik de Vink, Alessandra di Pierro and Herbert Wiklicky, LNCS 7938, pages 29–62, Springer, 2013.

ODE Analysis of Biological Systems

Diana-Elena Gratie, Bogdan Iancu, and Ion Petre

Computational Biomodeling Laboratory
Turku Centre for Computer Science and Åbo Akademi University
Turku, Finland
{dgratie,biancu,ipetre}@abo.fi

Abstract. This chapter aims to introduce some of the basics of modeling with ODEs in biology. We focus on computational, numerical techniques, rather than on symbolic ones. We restrict our attention to reaction-based models, where the biological interactions are mechanistically described in terms of reactions, reactants and products. We discuss how to build the ODE model associated to a reaction-based model; how to fit it to experimental data and estimate the quality of its fit; how to calculate its steady state(s), mass conservation relations, and its sensitivity coefficients. We apply some of these techniques to a model for the heat shock response in eukaryotes.

Keywords: Biomodeling, reaction-based models, ODE-based models, ODE analysis, parameter estimation, model identifiability, model refinement, heat shock response.

1 Introduction

Mathematical modeling with ordinary differential equations (ODEs) has a very long tradition in biology and ecology. Efforts to apply ODEs to understand population dynamics started already in the 18th century (see, e.g., Malthus's growth model [40]) as an effort to apply the principles of physical sciences to biological sciences as well. This research area led to major developments both in biology and ecology, as well as in mathematics. The field has long been called *biomathematics*, *mathematical biology* or *theoretical biology* and it typically involved researchers from life sciences (biology, biochemistry and ecology in particular), mathematics and more recently, from computer science and engineering. It has recently witnessed an explosion of interest in the computer science community due to the fast-paced developments in quantitative laboratory technologies. The developments on the computational side have also been influential, allowing for analyzing ever larger models and opening the door to new fields of research such as computational drug design or personalized medicine.

This chapter is primarily targeting the computer science community. Many computer scientists working in biomodeling seem to prefer a discrete stochastic approach rather than one based on ODEs. Such a choice is in some ways natural for computer scientists as it leads to new types of applications of formalisms that are well-studied in computer science, such as Petri nets, process algebra, finite automata, etc. On the other hand, such

M. Bernardo et al. (Eds.): SFM 2013, LNCS 7938, pp. 29–62, 2013.

methods come with their own limitations, especially in terms of numerical simulations of large models and in parameter estimation. Moreover, ODE-based modeling offers a huge array of analysis methods, some of which do not have a correspondent on the discrete side. Our chapter aims to introduce some of the basics of modeling with ODEs in biology, especially in terms of building such a model, analyzing some of its properties, and estimating its parameters. We focus on computational, numerical techniques, rather than on symbolic ones. We restrict our attention to reaction-based models, where the biological interactions are mechanistically described in terms of reactions, reactants and products.

The chapter is structured as follows. We discuss in Section 2 the notion of reaction-based models and introduce briefly the stochastic modeling approach in terms of continuous time Markov chains. We then discuss in more details the modeling with ODEs in Section 3. The parameter estimation problem is discussed in Section 4. We then introduce in Section 5 several analysis techniques, including steady state analysis, sensitivity analysis, and identification of mass conservation relations. As a case-study we discuss the modeling of the heat shock response in Section 6. We conclude with discussions in Section 7.

2 Reaction-Based Models

Reaction-based models are formalized as sets of reactions that describe the given system in terms of mechanistic interactions between the species of interest. We discuss separately two types of reactions: reversible and irreversible. In the following we consider a model M consisting of a set of m species $\Sigma = \{S_1, S_2, ..., S_m\}$ and n (reversible or irreversible) reactions R_j, $1 \leq j \leq n$.

Generalities. If reaction R_j, $1 \leq j \leq n$ is irreversible, then it has the following form:

$$R_j \ : \ c_{1,j}S_1 + c_{2,j}S_2 + \ldots + c_{m,j}S_m \xrightarrow{k_j} c'_{1,j}S_1 + c'_{2,j}S_2 + \ldots + c'_{m,j}S_m. \tag{1}$$

On the other hand, if it is reversible, then it is of the following form:

$$R_j \ : \ c_{1,j}S_1 + c_{2,j}S_2 + \ldots + c_{m,j}S_m \underset{k_j^-}{\overset{k_j^+}{\rightleftarrows}} c'_{1,j}S_1 + c'_{2,j}S_2 + \ldots + c'_{m,j}S_m. \tag{2}$$

In both cases, $k_j \geq 0$ ($k_j^+, k_j^- \geq 0$, resp.) is the *kinetic rate constant* of the irreversible (reversible, resp.) reaction R_j and $c_{1,j}, \ldots, c_{m,j}, c'_{1,j}, \ldots, c'_{m,j} \geq 0$ are non-negative integers that represent the quantitative proportion in which species participate in a reaction. The *stoichiometric coefficient* of molecular species S_i in reaction R_j is $n_{i,j} = c'_{i,j} - c_{i,j}$. The stoichiometric coefficients can be represented in a *stoichiometric matrix* $\mathbf{N} = (n_{i,j})_{m \times n}$. The (i, j) entry of the matrix is the stoichiometric coefficient of species S_i in reaction R_j. If $n_{i,j} > 0$ ($n_{i,j} < 0$, resp.), then we say that S_i is *produced* (*consumed*, resp.) in reaction R_j.

The reactants (the species indicated on the left-hand side of the reaction) are referred to as *substrates*, while the species produced as a result of the reaction being triggered

(indicated on the right-hand side of the reaction) are called *products*. A species S_i with $c_{i,j} = 0$ ($c'_{i,j} = 0$, resp.) is usually omitted from the left- (right-, resp.) hand side of reaction R_j.

Note that a reversible reaction can be divided into two different irreversible reactions, as follows:

$$R_j{}^{(1)} \ : \ c_{1,j}S_1 + c_{2,j}S_2 + \ldots + c_{m,j}S_m \xrightarrow{k_j{}^+} c'_{1,j}S_1 + c'_{2,j}S_2 + \ldots + c'_{m,j}S_m$$

$$R_j{}^{(2)} \ : \ c'_{1,j}S_1 + c'_{2,j}S_2 + \ldots + c'_{m,j}S_m \xrightarrow{k_j{}^-} c_{1,j}S_1 + c_{2,j}S_2 + \ldots + c_{m,j}S_m$$

The sum $\sum_{i=1}^{m} c_{i,j}$ for an irreversible reaction R_j is called the *molecularity* of reaction R_j. We consider here only reactions with molecularity at most two. Reactions with molecularity three are very rare, due to the high improbability of having three molecular entities simultaneously colliding and forming a correct configuration that leads to the constitution of a molecular complex; a molecularity greater than three for an elementary reaction is unattainable, since a number of molecules greater than three cannot concomitantly collide [49].

Example 1. We consider here the representation of a simple ecological prey-predator model through coupled chemical reactions: the Lotka-Volterra system, [39, 60]. The model consists of species *Prey* and *Predator* and its reactions are shown in Table 1.

Table 1. The Lotka-Volterra model [39, 60]

$Prey \xrightarrow{k_1} 2 \times Prey,$	growth of prey population
$Prey + Predator \xrightarrow{k_2} 2 \times Predator,$	consumption of preys
$Predator \xrightarrow{k_3} \emptyset$	death of predators

The dynamics of the Lotka-Volterra system is periodical: the population of preys grows at a rate proportional to the current population, the presence of predators in the system induces a decrease in the population of preys at a rate proportional to the number of prey-predator encounters, the population of predators declines at a rate proportional to the current population of predators. We return to this example in the next section where we associate to it an ODE-based model. A plot of its numerical simulation is shown in Figure 1.

Associating a mathematical model. After building a reaction-based model, one then associates to it a mathematical model to facilitate quantitative analysis and simulation of the model. There are many approaches available for this, see e.g. [9, 12]. The two approaches that are most used (either in a direct way, or an indirect way, as the underlying semantic of a higher-level model) are the ODE-based approach and the one based

on continuous time Markov chains (CTMCs). The modeling with CTMCs is described in more details elsewhere in this book; we only give it here a very brief presentation so that we can draw some comparison between the two in Section 7. We introduce the modeling with ODEs in more details in Section 3.

The stochastic approach is typically argued for on the basis of physical difficulties of ODE-based models with small populations [14, 15], or in terms of the network being too complex to describe in a deterministic way [61]. The stochastic formulation of a biochemical reaction network assumes homogeneity of substances and thermal equilibrium, see [16]. In this case, the model is usually described mathematically as a *continuous time Markov chain*, see [57]. Each species of the model becomes a time-dependent discrete stochastic variable indicating the number of individuals in that species, where time is modeled as a continuous variable. Formally, a stochastic process, $\{X(t), t \geq 0\}$, is a continuous-time Markov chain if for all $s,t \geq 0$, the following property is satisfied:

$$Pr\{X(s+t) = x_{s+t} | X(s) = x_s, X(u) = x_u, 0 \leq u \leq s\} = Pr\{X(s+t) = x_{s+t} | X(s) = x_s\}.$$

Intuitively, we say that the Markov chain is *memoryless*: its future dynamics depends only on the current state and not on the past states.

A continuous-time Markov chain is *time-homogeneous* if the following relation is satisfied:

$$Pr\{X(s+t) = j | X(s) = i\} = Pr\{X(t) = j | X(0) = i\}.$$

We discuss here only time-homogeneous systems.

Given a vector of non-negative integers $\mathbf{X} = (X_1, X_2, ..., X_m)$ and species $S_1, S_2, ..., S_m$ the *grand probability function* of the model, $Pr(\mathbf{X}, t)$, is the probability that there are X_1 species S_1, X_2 species S_2, ..., X_m species S_m at time t. We consider all species to be distributed randomly and homogeneously in the volume V. The central hypothesis for the stochastic formulation of chemical kinetics is that the probability of a particular combination of reactants to react according to a given reaction R in the next infinitesimal time interval $(t, t+dt)$ is $c_R dt$, for a certain constant c_R, called the *stoichiometric constant* of the reaction. The probability of a reaction occurring in the interval $(t, t+dt)$ is given by the formula $N_R \cdot c_R \cdot dt$, where by N_R we denote the number of combinations of reactants in the current state. For instance, for reaction $R^{(1)} : S_1 + S_2 \rightarrow S_3$, we have $N_{R^{(1)}} = X_1 \cdot X_2$. For reaction $R^{(2)} : 2S_1 \rightarrow S_3$, $N_{R^{(2)}} = X_1 \cdot (X_1 - 1)/2$.

For an infinitesimally small dt, the probability of the system being in a certain state at time $t+dt$ may be given by the following two scenarios: the system was in the current state at time t and no reaction occurred, or the system reached the current state as a result of a single reaction being triggered (the probability of having had two or more reactions is negligible). Denote by $a_k dt$ the probability of a reaction R_k occurring in the interval $(t, t+dt)$, given the state characterized by $\mathbf{X}$ at time t, and by $B_k dt$ the probability that reaction R_k occurs in the time interval $(t, t+dt)$ resulting in a state characterized by $\mathbf{X}$. The reasoning above can be formally written as follows:

$$Pr(\mathbf{X},t+dt) = Pr(\mathbf{X},t)(1-\sum_{k=1}^{n} a_k dt) + \sum_{k=1}^{n} B_k dt, \text{ i.e.,}$$

$$(Pr(\mathbf{X},t+dt) - Pr(\mathbf{X},t))/dt = -\sum_{k=1}^{n} a_k Pr(\mathbf{X},t) + \sum_{k=1}^{n} B_k, \text{ and so,}$$

$$\frac{\partial Pr(\mathbf{X},t)}{\partial t} = \sum_{k=1}^{n} (B_k - a_k Pr(\mathbf{X},t)). \tag{3}$$

Equation (3) is known in the literature as the *Chemical Master Equation*. A detailed mathematical analysis of a complex system using the chemical master equation has been proven to be intractable, see [61]. However, an alternative to the aforementioned approach is Gillespie's algorithm, introduced in [14, 15], that generates a random walk through the state space of the model, avoiding the solving of the master equation.

3 ODE-Based Models

We discuss in this section how to associate an ODE-based model to a reaction model. In this case, the dynamic behavior of the system is expressed in terms of the time-dependent evolution of each species' concentration. The deterministic framework of ordinary differential equations (ODEs) is often chosen as the default mathematical counterpart of a reaction-based system, sometimes followed-up by other modeling approaches. The basic quantities describing the ODE model are the concentrations $[S_1]$, $[S_2]$, ..., $[S_m]$ of the m species in the model, and the fluxes $v_1, v_2, ..., v_n$ of the n reactions in the model. The concentration is generally expressed either in terms of particle numbers (i.e. the number of molecules of species S, denoted $\#S$, in a solution with volume V), or in terms of moles of species S per volume V. The correspondence between the number of molecules and the number of moles is given by the relation:

$$\#S = [S] \cdot N_A,$$

where $N_A \approx 6.02214179 \cdot 10^{23}$particles/mol. The unit of $[S]$ is commonly denoted by M = mol $\cdot$L^{-1}, where L is litre.

Without loss of generality, we will assume that all reactions are reversible and have the form in (2); an irreversible reaction is then a particular case, where one of the two kinetic constants is zero.

Each species S_i of the reaction model can be modeled as a function $[S_i] : \mathbb{R}_+ \to \mathbb{R}_+$ representing the time evolution of its concentration. The dependencies between the species can then be expressed in terms of a systems of ODEs in the variables $[S_i]$ modeling the change in $[S_i]$ as a function of all other variables:

$$d[S_i]/dt = \sum_{j=1}^{n} n_{i,j} v_j,$$

where v_j is the flux of reaction r_j and $n_{i,j}$ is the (i,j) stoichiometric coefficient. Here, we make the assumption that the only factor affecting the concentrations of the species are the reactions. Considering the vector of all reaction fluxes $v = (v_1, v_2, ..., v_n)^T$, the ODE representation of the entire reaction model can be written in a compact way as follows:

$$d[S]/dt = Nv, \tag{4}$$

where $[S] = ([S_1], [S_2], ..., [S_m])^T$ is a vector of the concentrations of all species in the reaction-based model, see [25].

Example 2. Consider the following reaction model:

$$2A \to B,$$
$$A + B \rightleftarrows C.$$

Denote by v_1 and v_2 the fluxes of the two reactions in the model, respectively. (We discuss in the next section how the flux of a reaction is defined, depending on the kinetic law the modeler chooses.) Then the corresponding ODE model is:

$$\begin{bmatrix} \frac{d[A]}{dt} \\ \frac{d[B]}{dt} \\ \frac{d[C]}{dt} \end{bmatrix} = \begin{bmatrix} -2 & -1 \\ 1 & -1 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}.$$

While the stoichiometries in a reaction-based model are constant, the concentrations of all species will vary in time as a function of the reaction fluxes, which are in turn dependent on the kinetics of each reaction and on the concentrations of all reactants. In the following, we describe in details two of the most common reaction kinetics: the mass-action principle, and Michaelis-Menten kinetics.

3.1 Law of Mass-Action

The most common biochemical kinetics follow the *mass action law*. It was introduced in [20,21], and it states that the *flux* (also called sometimes *rate*) of a reaction is proportional to the probability of the reactants colliding. Assuming a well-stirred environment, the probability of the substrates of a reaction colliding is proportional to their concentration to the power of their molecularity.

Example 3. For a simple reaction of the form

$$S_1 + S_2 \underset{k^-}{\overset{k^+}{\rightleftarrows}} 2P,$$

the reaction flux is

$$v = v_+ - v_- = k^+[S_1][S_2] - k^-[P]^2,$$

where v_+ represents the left-to-right (forward) reaction rate, v_- represents the right-to-left (backward) flux, k^+ is the left-to-right kinetic rate constant, and k^- represents the right-to-left kinetic rate constant. For the forward reaction, the molecularity of each of

the two substrates S_1, S_2 is 1, and for the backward reaction the molecularity is 2. If the time is measured in seconds (s), and the concentration in M, then the unit for the reaction rates is $\mathrm{M} \cdot \mathrm{s}^{-1}$. It follows that for monomolecular reactions (e.g. $S \to \emptyset$), the rate constant has unit s^{-1}, while for bimolecular reactions, the rate constant is measured in $(\mathrm{M} \cdot \mathrm{s})^{-1}$.

Considering a general reversible reaction of the form (2), the reaction rate reads

$$v = v_+ - v_- = k_j^+ \prod_{i=1}^{m} [S_i]^{c_{i,j}} - k_j^- \prod_{i=1}^{m} [S_i]^{c'_{i,j}}.$$

The corresponding system of ODEs, following (4), is

$$\frac{d[S_i]}{dt} = n_{i,j} v = (c'_{i,j} - c_{i,j}) \left(k_j^+ \prod_{l=1}^{m} [S_l]^{c_{l,j}} - k_j^- \prod_{l=1}^{m} [S_l]^{c'_{l,j}} \right), 1 \le l \le m.$$

For reversible reactions, the ratio of substrate and product at steady state (i.e., when the forward and backward reaction rates are equal, $v_+ = v_-$) is a constant, K_{eq}, called the equilibrium constant:

$$K_{eq} = \frac{k_j^+}{k_j^-} = \frac{\prod_{i=1}^{m} [S_i]_{eq}^{c'_{i,j}}}{\prod_{i=1}^{m} [S_i]_{eq}^{c_{i,j}}},$$

where $[S_i]_{eq}$ represents the equilibrium concentration of species S_i.

The time course for a species S is obtained by integrating the corresponding ODE. For a simple decomposition reaction $S \xrightarrow{k} P_1 + P_2$, the time dynamics is described by the ODE $d[S]/dt = -k[S]$. Integrating over the interval [0, t) yields the analytical solution

$$\int_{S_0}^{S} d[S]/dt = -\int_{t=0}^{t} k dt \Rightarrow [S](t) = S_0 e^{-kt}.$$

Calculating the analytical solution for more complex models is however rarely possible.

Example 4. For the Lotka-Volterra model introduced in Example 1, the mass-action reaction fluxes for the three reactions in the system are the following:

$$v_1 = k_1[Prey], \quad v_2 = k_2[Prey][Predator], \quad v_3 = k_3[Predator].$$

The system of ODEs describing the dynamics of the Lotka-Volterra model is:

$$\begin{aligned} d[Prey]/dt &= v_1 - v_2 = k_1[Prey] - k_2[Prey][Predator] \\ d[Predator]/dt &= v_2 - v_3 = k_2[Prey][Predator] - k_3[Predator]. \end{aligned} \tag{5}$$

The periodic dynamics of the Lotka-Volterra model is depicted in Figure 1.

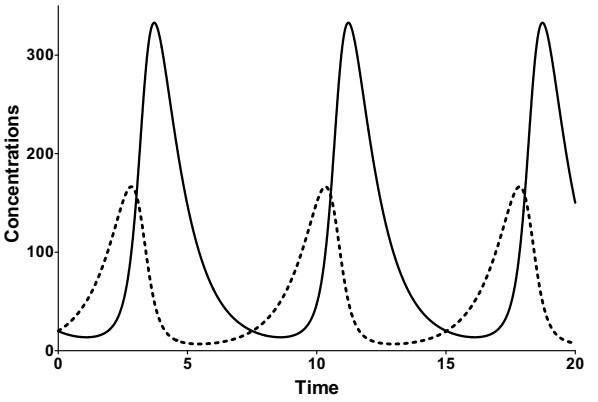


Fig. 1. The periodic time dynamics of the Lotka-Volterra model. The solid line represents the concentration of *Predator*, while the dotted line represents that of *Prey*. As the *Prey* population grows, the *Predator* population also grows; then there are more encounters *Predator-Prey* that reduce the *Prey* population; this reflects on the *Predator*, as they only multiply as long as they find food. When the size of *Predator* drops, the *Prey* population starts to grow, and the cycle repeats.

3.2 Kinetics of Enzymatic Reactions

Enzymatic reactions are a special class of biochemical reactions, where an enzyme is required for a reaction to take place, but the enzyme itself is not consumed during the reaction. The general form of an enzymatic reaction, as proposed in [7] based on previous experimental results of [28, 43], is:

$$S + E \underset{k_2}{\overset{k_1}{\rightleftarrows}} S:E \overset{k_3}{\rightarrow} P + E,$$

where E is an enzyme, S is the substrate of the reaction, $S:E$ is a substrate-enzyme complex, and P is the product. This system of reactions represents in fact the reaction $S \rightarrow P$, catalyzed by enzyme E. The system can be represented using mass-action kinetics, considering the following irreversible reactions:

$$S + E \overset{k_1}{\rightarrow} S:E, \quad S:E \overset{k_2}{\rightarrow} S + E, \quad S:E \overset{k_3}{\rightarrow} P + E. \tag{6}$$

The system of ODEs describing the mass-action dynamics of the reaction-based model (6) is the following:

$$\frac{d[S]}{dt} = -k_1[S][E] + k_2[S:E]; \tag{7}$$

$$\frac{d[E]}{dt} = -k_1[S][E] + k_2[S:E] + k_3[S:E]; \tag{8}$$

$$\frac{d[S:E]}{dt} = k_1[S][E] - k_2[S:E] - k_3[S:E]; \tag{9}$$

$$\frac{d[P]}{dt} = k_3[S:E]. \tag{10}$$

Michaelis-Menten kinetics. Because the system of ODEs (7) - (10) cannot be solved analytically, simplifying assumptions have been proposed. For example, the kinetic constants k_1, k_2 could be assumed to be much greater than k_3 ($k_1, k_2 \gg k_3$, see [43]), i.e. $[S:E]$ is negligible compared to $[S]$ and $[P]$, because the substrate-enzyme complex concentration is very low. This is called the *quasi-equilibrium* between the free enzyme E and the compound $S:E$.

This assumption has been further extended (see [7]) to considering that the system will eventually reach a state where the concentration of substrate-enzyme complex remains unchanged (*quasi-steady state* of $S:E$); the assumption only holds when $S_0 \gg E_0$. In this case we obtain:

$$d[S:E]/dt = 0, \text{ i.e., } k_1[S][E] - k_2[S:E] - k_3[S:E] = 0. \tag{11}$$

Note that the right hand side of (8) is the complement of the right hand side of (9). Adding them we get that $d[E]/dt + d[S:E]/dt = 0$. Equivalently,

$$[E] + [S:E] = E_{tot}, \text{or equivalently } [E] = E_{tot} - [S:E], \tag{12}$$

where E_{tot} is constant, standing for the total amount of enzyme in the system, either free or as part of the substrate-enzyme complex.

Considering the *quasi-steady state* assumption and (12), equation (11) can be rewritten as follows:

$$\begin{aligned} &k_1[S]E_{tot} = k_1[S][S:E] + k_2[S:E] + k_3[S:E], \text{ i.e.,} \\ &[S:E] = \frac{k_1[S]E_{tot}}{k_1[S] + k_2 + k_3}, \text{ i.e.,} \\ &[S:E] = \frac{[S]E_{tot}}{[S] + \frac{k_2+k_3}{k_1}} \end{aligned} \tag{13}$$

Introducing (13) into (10) yields the result

$$\frac{d[P]}{dt} = \frac{k_3[S]E_{tot}}{[S] + \frac{k_2+k_3}{k_1}}. \tag{14}$$

The Michaelis-Menten equation relates the reaction rate v of synthesizing the product P to the concentration of the substrate, $[S]$, by the relation:

$$v = \frac{d[P]}{dt} = \frac{V_{max}[S]}{[S] + K_m}, \tag{15}$$

where V_{max} represents the maximum rate achieved by the system, for saturated values of $[S]$. The Michaelis constant K_m is the concentration of substrate for which the reaction rate is half-maximal. Identifying the parameters of (15) into (14) yields the connection between the Michaelis-Menten kinetics and the mass-action deduced kinetics of an enzymatic reaction:

$$V_{max} = k_3 E_{tot}, \quad K_m = \frac{k_2 + k_3}{k_1}.$$

Assuming the *quasi-equilibrium*, the quantity k_3/k_1 is negligible, thus $K_m \cong k_2/k_1$. Figure 2 shows the dependency of the reaction rate v with $[S]$. For more details on Michaelis-Menten kinetics, we refer the reader to [37].

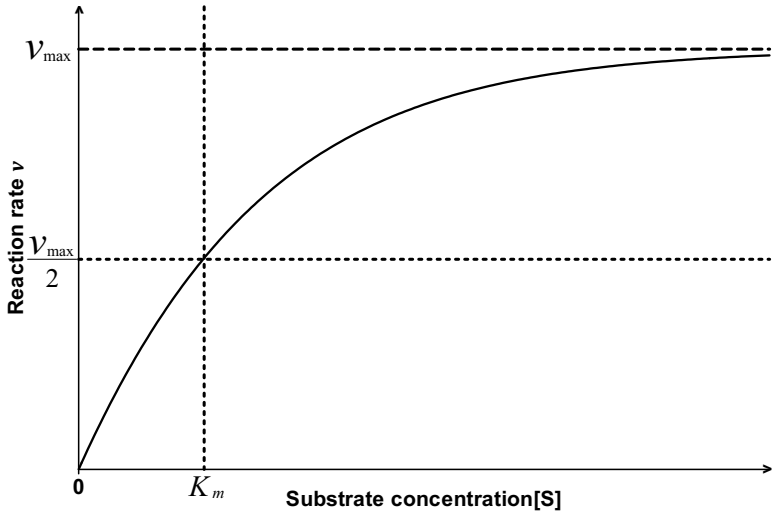


Fig. 2. Dependency of the reaction rate v with $[S]$ for Michaelis-Menten kinetics. v_{max} represents the maximum velocity, and K_m is the concentration of substrate for which the reaction rate is half-maximal.

Reversible Michaelis-Menten kinetics. Enzyme kinetics can often be reversible, and the Michaelis-Menten equation can be extended to a reversible reaction of the form

$$S + E \underset{k_2}{\overset{k_1}{\rightleftarrows}} X \underset{k_4}{\overset{k_3}{\rightleftarrows}} P + E, \tag{16}$$

where S and P are substrates, E is the enzyme, and X represents the intermediary enzyme-substrate compound. The mass-action irreversible reactions describing this system are:

$$\begin{aligned} S+E &\xrightarrow{k_1} X;\\ X &\xrightarrow{k_2} S+E;\\ X &\xrightarrow{k_3} P+E;\\ P+E &\xrightarrow{k_4} X. \end{aligned} \tag{17}$$

The system of ODEs describing the dynamics of the reaction-based model (17) is:

$$\frac{d[S]}{dt} = -k_1[S][E] + k_2[X]; \tag{18}$$

$$\frac{d[E]}{dt} = -k_1[S][E] + k_2[X] + k_3[X] - k_4[P][E]; \tag{19}$$

$$\frac{d[X]}{dt} = k_1[S][E] - k_2[X] - k_3[X] + k_4[P][E]; \tag{20}$$

$$\frac{d[P]}{dt} = k_3[X] - k_4[P][E]. \tag{21}$$

Following the reasoning for simple Michaelis-Menten equations, adding (19) and (20) yields

$$\frac{d[E]}{dt} + \frac{d[X]}{dt} = 0 \Rightarrow [E] + [X] = E_{tot}.$$

For the *quasi-steady state*, $d[X]/dt = 0$, i.e., $k_1[S](E_{tot} - [X]) - [X](k_2 + k_3) + k_4[P](E_{tot} - [X]) = 0$, which leads to

$$[X] = \frac{k_1[S]E_{tot} + k_4[P]E_{tot}}{k_1[S] + k_4[P] + k_2 + k_3}. \tag{22}$$

Introducing (22) into equation (21), after a few computations the formula reads

$$v = \frac{k_1k_3[S]E_{tot} - k_2k_4[P]E_{tot}}{k_1[S] + k_4[P] + k_2 + k_3} = \frac{k_3E_{tot}\frac{k_1[S]}{k_2+k_3} - k_2E_{tot}\frac{k_4[P]}{k_2+k_3}}{1 + \frac{k_1[S]}{k_2+k_3} + \frac{k_4[P]}{k_2+k_3}} = \frac{\frac{V_{fw}}{K_{mS}}[S] - \frac{V_{bw}}{K_{mP}}[P]}{1 + \frac{[S]}{K_{mS}} + \frac{[P]}{K_{mP}}},$$

where $K_{mS} = (k_2 + k_3)/k_1$ and $K_{mP} = (k_2 + k_3)/k_4$ are the Michaelis-Menten constants (i.e. for half-maximal forward and backward rate) for the substrate and product, respectively, and $V_{fw}(V_{bw})$ denotes the maximal rate in forward (backward) direction. An exact solution to this equation can be found in [44]. For details on the reversible Michaelis-Menten kinetics, we refer the reader to [22].

Other kinetic laws. Mass action and Michaelis-Menten are not the only existing kinetics. Some enzymatic reaction can follow Hill kinetics, Goldbeter-Koshland kinetics, or be subject to inhibition. We only introduce them briefly, discussing the types of reactions that are typically modeled in this way, and skipping the derivation of their mathematical formulations.

Goldbeter-Koshland kinetics, introduced in [18], applies to reversible reactions from substrate to product and back, catalyzed by different enzymes (e.g. phosphorylation and

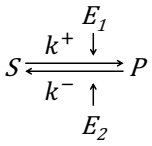


Fig. 3. Goldbeter-Koshland kinetics. P is produced from S in presence of enzyme E_1 and S is produced from P in presence of enzyme E_2.

dephosphorylation of proteins). The forward and backward reactions have Michaelis-Menten kinetics. The general form of such reactions is shown in Figure 3.

Hill kinetics, introduced in [29], are suitable for reactions where the enzyme can bind more molecules from the substrate S. Usually, the binding of the first S molecule changes the binding rate of the second molecule. The rate can either increase (called *positive cooperativity*), or decrease (called *negative cooperativity*). A general form of such reactions is the following:

$$E + S_1 \underset{k_1^-}{\overset{k_1^+}{\rightleftarrows}} ES_1;$$

$$ES_1 \overset{k_2}{\to} P + E;$$

$$ES_1 + S_2 \underset{k_3^-}{\overset{k_3^+}{\rightleftarrows}} ES_1S_2;$$

$$ES_1S_2 \overset{k_4}{\to} P + E.$$

Inhibition in a system with Michaelis-Menten kinetics (see (16)) can occur at different levels. An inhibitor I can bind to an enzyme in different states of the enzyme. When it binds (in a reversible reaction) to the free enzyme, the inhibition is called *competitive*, as both the substrate and the inhibitor are competing for binding the enzyme. When I binds reversibly to the enzyme-substrate complex, the reaction is called *uncompetitive inhibition*, as the enzyme is already bound to the substrate. When the inhibitor binds both to the free enzyme and to the enzyme-substrate complex, the inhibition is called *noncompetitive*. For a more detailed description of enzyme inhibition reactions, we refer the reader to [37].

4 Parameter Estimation

We discuss in this section the parameter estimation problem, including aspects of model identifiability, quantitative measures for model fit quality, model validation, and methods for model fitting.

4.1 Generalities: Relating the Mathematical Model to the Experimental Data

Relating the mathematical model to the experimental data is an essential step in the process of model building. This includes the validation of the model in terms of how

well it can explain existing (quantitative or qualitative) experimental data and how well its predictions correspond to existing (non-quantitative) knowledge. There are several ways to approach this problem.

1. The modeler might have no a-priori hypothesis regarding the mathematical form and is instead strongly guided by data. The focus here is to capture the trend of the data and to predict the behavior in-between the data points and the emphasis is on the data. This approach is called *interpolation*.
2. The modeler has a clear hypothesis regarding the mathematical form she is building. For example, she might start from a reaction-based model and then associate to it an ODE-based model with mass-action kinetics as discussed in Section 3. The focus here is on finding values for all model parameters and the emphasis is on the model. This approach is called *model fitting*.
3. The modeler might replace a fitted model with an interpolating curve because of a need for better mathematical properties in further manipulations/analsysis of the model. This approach is called sometimes *model approximation*.

We only focus in this section on aspects related to model fitting. For a basic introduction to other approaches we refer to [17].

The main focus in model fitting is on the estimation of the unknown kinetic parameters of the model so that its predictions are consistent with a given set of data, usually presented in terms of time series. This step is often followed by a model validation step, where the model is compared with another set of data, that was not used in the fitting stage. In both cases, the task can be formulated as a mathematical optimization problem to minimize a cost function that quantifies the differences between the model predictions and the experimental measurements. The cost function can be seen as a distance measure between two vectors with non-negative real numbers as entries, one holding the experimental data, the other the model prediction for the time points where the data was collected. Some of the most widely used cost measures in this context are based on the *Chebyshev criterion*, *sums of absolute deviations*, and *least-squares*. We introduce briefly each of them in the following. In all cases, we are given a data set (x_i, y_i), $1 \leq i \leq m$ and a model $y = f(k,x)$, where $f : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ and $k \in \mathbb{R}^n$ is the vector of parameters, often with non-negative values.

In the Chebyshev criterion, the goal is to find $k \in \mathbb{R}^n$ that minimizes $\max\{|y_i - f(k,x_i)|, 1 \leq i \leq m\}$. In other words, the goal is to *minimize the largest absolute deviation* of a model value from the corresponding experimental value. The effect is that more weight is given to the worst outlier.

Another approach is to find $k \in \mathbb{R}^n$ that minimizes $\sum_{1 \leq i \leq m} |y_i - f(k,x_i)|$. In other words, the goal is to *minimize the sum of absolute deviations*. The effect is to treat each data point equally and to average the deviations over all experimental points.

In the third approach we mention here, the goal is to find $k \in \mathbb{R}^n$ that minimizes $\sum_{1 \leq i \leq m} |y_i - f(k,x_i)|^2$. This is the most widely used criterion in model fitting because the resulting optimization problem can be approached using calculus if f is a differentiable function (such as those obtained through the methods in Section 3).

The problem of estimating the parameters of kinetic models in systems biology is computationally difficult, see e.g., [4,42,45]. Regardless of which fitting criterion (score

function) is used, the high number of variables in a typical biomodel makes an exact solution to the problem unfeasible in practice. There are however many approximation methods. Some of them are based on local approximation algorithms; they are faster in practice, but tend to converge to local optima. Others are based on global optimization algorithms; they are in general slower, but tend to converge to a global optimum. The global optimization methods can be based on deterministic searches [19, 33] or on stochastic ones [2,6]. Even though the deterministic methods guaranty the convergence to a global optimum, the speed of the convergence is typically a major concern and in general, these methods cannot ensure the termination of the algorithm within a given finite time interval [45]. On the other hand, the intrinsic randomness of the stochastic approaches does not guarantee their convergence to an optimum [45]. However, many stochastic methods exhibit a good performance in practice – they are often capable of efficiently identifying a point in the vicinity of global solutions, see [45].

There are many modeling software environments, some commercial, others offering free access, that are used for model fitting. In most of our projects we chose COPASI [31] as a computational environment for parameter estimation. This software is a widely used tool in computational systems biology, having a documented good performance, see [4, 42, 45]. It includes a suite of various local and global, deterministic and stochastic parameter estimation algorithms, such as simulated annealing, genetic algorithms, evolution strategy using stochastic ranking, and particle swarm.

4.2 Alternative Model Fits and Model Identifiability

The problem of *model identifiability* adds to the difficulty of model fitting; it has to do with a model having several (sometimes very) different sets of parameter values, all yielding good model fits. The problem is that some numerical properties of the model, such as sensitivity coefficients, might be drastically different in different numerical setups, even if they all fit well the available data. This implies that there exist several models (or model setups) offering equally good, but different explanations for the available data. In such a situation, additional data is needed, focusing on the domains where the candidate models exhibit different behavior.

Even when only one model fit has been achieved, the modeler should evaluate the uniqueness of the parameter set. One way of doing this is to repeat the parameter estimation procedure, using some other available algorithms but the same data set. Such a procedure can in principle yield several different results, as demonstrated e.g. in [8,53].

When searching for alternative numerical model fits, one can sample the distribution of the score functions measuring the distance between the model predictions through a simultaneous sampling on the range of all parameter values. For each parameter, one can generate a large sample, e.g. through partitioning its value range into a large number of equal sized subintervals (say, on the scale of tens of thousands) and randomly select a value from each of them. For all combinations of values for all parameters, one can then calculate the score of the model fit and thus sample the distribution of the score function. However, the direct implementation of this idea is clearly intractable for models with more than a few parameters due to the combinatorial explosion of the number of model simulations that need to be run. A fast, practical solution to this problem is the *Latin Hypercube Sampling* method (LHS) of [41]. This is a method to generate samples

which are uniformly distributed over each parameter space, with the number of samples being independent of the number of parameters, see [26,27,50] for several applications. Let p be the number of parameters. The first step is to choose the size of the sample, N; this will also serve as the number of samples for each parameter. The range of each parameter is then partitioned into N intervals, with the length of each interval proportional to the probability of the parameter's value to fall in that interval; in particular, if the parameter is uniformly distributed in its range, then all subintervals are equal-sized. We then randomly select a value from each subinterval to generate a sample of N values for each parameter. The N values for parameter i are then stored on the i-th column of an $N \times p$ matrix. Finally, we randomly shuffle the values on all the columns of the matrix. The result is read from the matrix row-by-row, giving a sample of N combinations of parameter values. For a detailed description of this sampling scheme we refer to [41]. We discuss this method in the case of the heat shock response model in Section 6.

4.3 Fit-Preserving Model Refinement

Altering an already-fitted model, for example by adding a new component to it, replacing a module with another one, or adding new variables and reactions to it, will lead to losing its numerical fit. The problem is especially difficult in cases where the number of parameters in the new model is much larger than in the starting model. Rather then attempting to re-estimate all parameters, including those that were already fitted in the starting model, a computationally more efficient way is to build the larger model in an iterative way, ensuring in each step that its quantitative model fit is preserved. This method is called *quantitative model refinement* and has already been investigated in several different setups in [3, 34, 46]. We follow here the presentation of [34].

A given reaction-based model can undergo several types of refinement, for instance depending whether the focus lies on the reactants or reactions of the model. If one's focus lies on model's data, then the model could be refined so as to include more details regarding a species by having it substituted for several of its subspecies. The main interest in this type of refinement originates in the analysis of the possible behavioral intricacies the model refined as such would depict. This type of refinement is called *data refinement* and it consists in refining a set of variables so as to include more details about their internal states, attributes, etc. If the interest lies on the reactions of the model, one could refine the model by replacing for instance a reaction in the model describing a certain process by a set of reactions detailing on some intermediate steps of the process. This type of refinement is called *process refinement*.

Formal refinement arose from the field of software engineering as a necessity to embed an elementary set of specifications in a system's final implementation. The problem of quantitative model refinement has been addressed before in systems biology in particular related to rule-based modeling, which integrates *data refinement* through the notion of agent resolution ([23]). The focus lies here on rule refinement, a method designed to refine rules ensuring model fit preservation. Nevertheless, the refinement technique must preserve the quantitative systemic properties of the model, such as numerical fit and validation, see [46].

A model M consisting of a set of reactions of the form (1) could be formalized through a discrete or continuous approach, a deterministic or non-deterministic

evolution, etc. This discussion focuses on a continuous mass-action formulation. To each variable $S_i, 1 \leq i \leq m$ we associate a time-dependent function $[S_i] : \mathbb{R}_+ \to \mathbb{R}_+$, which denotes the concentration of the species over time. According to the principle of mass action, see [37], the time evolution of the system may be specified by a system of ODEs as follows:

$$\frac{d[S_i]}{dt} = -\sum_{j=1}^{n}\left(k_j c_{i,j}\prod_{l=1}^{m}[S_l]^{c_{l,j}}\right) + \sum_{j=1}^{n}\left(k_j c'_{i,j}\prod_{l=1}^{m}[S_l]^{c'_{l,j}}\right), \quad 1 \leq i \leq m. \tag{23}$$

Assume now model M is refined discerning among various subspecies of S_1. The distinction among the subspecies of S_1 can be made by either different classes of S_1 or several biochemical configurations of S_1, as a result of various post-translational modifications such as acetylation, phosphorylation, etc. All subspecies characterized as such participate in all reactions S_1 took part in (in model M), with possible variations in the kinetics. Replacing species S_1 in model M by subspecies $V_1, \ldots, V_l$ brings about a new model M_R, whose set of species consists of the new variables $\{S'_2, S'_3, ..., S'_m\} \cup \{V_1, ..., V_l\}$, for some $l \geq 2$, where variables S'_j, $2 \leq j \leq m$ of M_R, match variables S_j of model M and $V_1, ..., V_l$ replace species S_1 in M_R. Moreover, each reaction R_j of model M is substituted for in model M_R by a reaction R'_j as follows:

$$\begin{array}{rl} R'_j \ : & (T_{1,j}V_1 + ... + T_{l,j}V_l) + c_{2,j}S'_2 + ... + c_{m,j}S'_m \xrightarrow{k'_j} \\ & (T'_{1,j}V_1 + ... + T'_{l,j}V_l) + c'_{2,j}S'_2 + ... + c'_{m,j}S'_m, \end{array}$$

where k'_j is the kinetic rate constant, and $T_{1,j}, ..., T_{l,j}, T'_{1,j}, ..., T'_{l,j}$ are nonnegative integers so that $T_{1,j} + ... + T_{l,j} = c_{1,j}$ and $T'_{1,j} + ... + T'_{l,j} = c'_{1,j}$.

Model M_R is a *data refinement of model M on variable* S_1 if and only if the subsequent conditions hold, see [34]:

$$[S_j](t) = [S'_j](t), \text{ for all } 2 \leq j \leq m, \tag{24}$$

$$[S_1](t) = [V_1](t) + \ldots + [V_l](t), \text{ for all } t \geq 0. \tag{25}$$

The refined model, M_R, involves a number of $m + l - 1$ species, while model M comprises only m species, M_R evolving linearly in the size of its data set. The number of reactions in M_R substituting for reaction R_j of M is the number of non-negative integer solutions of the subsequent system of equations:

$$T_{1,j} + T_{2,j} + ... + T_{l,j} = c_{1,j};$$
$$T'_{1,j} + T'_{2,j} + ... + T'_{l,j} = c'_{1,j};$$

over the independent unknowns $T_{k,j}, T'_{k,j}, 1 \leq k \leq l$. The number of solutions of the first equation is given by the *multinomial coefficient* "l multichooses $c_{1,j}$", see [13]:

$$\left(\!\!\binom{l}{c_{1,j}}\!\!\right) = \binom{l + c_{1,j} - 1}{c_{1,j}} = \frac{(l + c_{1,j} - 1)!}{c_{1,j}!\,(l-1)!}.$$

Some values for the new kinetic parameters of M_R may be attained from the literature or they can be estimated experimentally. The parameters not attained as such require

calculation through computational methods so that conditions (24) and (25) are fulfilled. The reiteration of the parameter estimation process is however computationally expensive. As an alternative, the method proposed in [34] describes an approach for setting the values of the unknown parameters in the refined model so that relations (24) and (25) hold. The approach promotes a choice of parameters symmetrical in $V_1, ..., V_l$.

4.4 Quantitative Measures for the Model Fit Quality

Given parameter estimation may yield several different outputs, depending on the methods that were used in the fitting, it is important to quantify the goodness of a model fit. In this way, the results of different parameter estimation rounds can be compared. Moreover, through a suitable normalization, even the fitting of different models, using different sets of data, may also be compared. Part of the challenge here is to avoid to discriminate against models deviations that may be large in absolute values, but relatively small compared to the experimental data.

We discuss here briefly a notion of model fit quality introduced in [38]. Their fit quality only takes into account one set of experimental data at a time and aims to give a measure of the average deviation of the model from the data, normalized on the scale of the numerical values of the model predictions. For a given experimental data set $\mathcal{E} = \{(x_i, y_i) \mid 1 \leq i \leq n\}$ and a model $M = f(k,x)$, the quality of M's fit with respect to $\mathcal{E}$ is denoted as $q(M, \mathcal{E})$ and is defined as follows:

$$q(M, \mathcal{E}) = \frac{\sqrt{\sum_{i=1}^{n} (f(k,x_i) - y_i)^2 / n}}{\sum_{i=1}^{n} f(k,x_i)/n} \cdot 100\%.$$

It was argued in [38] that a low (say, lower than $15-20\%$) value of $q(M, \mathcal{E})$ could be considered as an indicator of a successful fit. We discuss the quality of the best fit for the heat shock response model in Section 6 and refer to [10] for more details on applying this measure.

5 Analysis of ODE-Based Models

We discuss in this section several computational analysis techniques for ODE-based models. We apply some of these techniques in the next section, on the heat shock response model.

5.1 Steady State Analysis

Steady states (also called *stationary states*, *fixed points*, *equilibrium points*) have the property that when taken as initial values for the model, they yield a constant dynamics; in other words, there is no change in the concentration of any of the species when starting from steady state values. This is one of the basic concepts in dynamical systems theory, extensively employed in modeling biological systems. There are several types of steady states: *stable*, *asymptotically stable*, *unstable* etc.

Consider a dynamical system $dx/dt = f(x(t))$, $x(0) = x_0$, where $f : \mathbb{R}^n \to \mathbb{R}^n$ is a continuous function with equilibrium point x_e. The equilibrium is *stable* if for every $\varepsilon > 0$ there exists δ_ε such that, if $||x_0 - x_e|| < \delta_\varepsilon$, then $||x(t) - x_e|| < \varepsilon, \forall t \geq 0$. A steady state is called *asymptotically stable* if there exists $\delta > 0$ such that if $||x_0 - x_e|| < \delta$, then $\lim_{t\to\infty} ||x(t) - x_e|| = 0$. A steady state is *unstable* if the conditions for stability are not met.

For a reaction-based model, the steady state behavior is characterized by the equation

$$\frac{d[S]}{dt} = 0,$$

or equivalently, considering Equation (4),

$$Nv = 0. \tag{26}$$

The rate vector v that satisfies the steady state condition (26) can be obtained by solving the corresponding system of algebraic equations with the variables $[S_1], [S_2], \ldots, [S_m]$. The equation has nontrivial solutions (not all variables are zero) only if $\text{rank}(N) < n$, where n is the number of reactions in the system, i.e. matrix N contains at least one pair of linearly dependent columns. The dependencies can be expressed by a so-called kernel matrix K, such that

$$NK = 0, \tag{27}$$

where K has $c = n - \text{rank}(N)$ columns. The columns k_i of matrix K are the vectors that span the null space (also termed kernel) of N, i.e. the subspace of the reaction rates space that contains all solutions to Equation (26), see [24]. Consequently, any vector J of steady-state fluxes can be expressed as a linear combination of K's columns,

$$J = \sum_{i=1}^{c} \alpha_i k_i.$$

The kernel matrix K is not uniquely determined. Another kernel matrix K' could be obtained for example by a multiplication $K' = KQ$, where Q has dimensions $[n - \text{rank}(N)] \times [n - \text{rank}(N)]$. Since K is a solution to Equation (27), so is K'. For details on how to determine the kernel matrix using Gauss's algorithm, we refer the reader to [37].

Example 5. Consider the following system of reactions:

$$2A \xrightarrow{k_1} B;$$
$$A + B \underset{k_2^-}{\overset{k_2^+}{\rightleftarrows}} C.$$

To compute the steady state, one needs to solve Equation (26), which reads as the following system of algebraic equations:

$$\underbrace{\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}}_{\mathbf{0}} = \underbrace{\begin{bmatrix} -2 & -1 \\ 1 & -1 \\ 0 & 1 \end{bmatrix}}_{\mathbf{N}} \cdot \underbrace{\begin{bmatrix} v_1 \\ v_2 \end{bmatrix}}_{\mathbf{v}}.$$

Considering mass action kinetics and denoting by $[A]_0, [B]_0, [C]_0$ the steady-state concentrations for the species in the model, the system reads:

$$\begin{aligned} &-2k_1[A]_0^2 - k_2^+[A]_0[B]_0 + k_2^-[C]_0 = 0,\\ &k_1[A]_0^2 - k_2^+[A]_0[B]_0 + k_2^-[C]_0 = 0,\\ &k_2^+[A]_0[B]_0 - k_2^-[C]_0 = 0. \end{aligned}$$

Solving the steady-state system of equations gives the solution $[A]_0 = 0, [B]_0 = \alpha, [C]_0 = 0$, where $\alpha > 0$ is arbitrary.

Example 6. Let us consider the Lotka-Volterra model expressed in Table 1. The ODEs characterising the system's dynamics are expressed in Equation (5). The steady state analysis leads to the system

$$\begin{aligned} &k_1[Prey] - k_2[Prey][Predator] = 0,\\ &k_2[Prey][Predator] - k_3[Predator] = 0. \end{aligned}$$

Solving this system of two equations gives the steady state points

$$([Prey]_s, [Predator]_s) \in \{(0,0), (k_3/k_2, k_1/k_2)\}.$$

To study the behavior of the Lotka-Volterra model around the steady states, one needs to examine the behavior of the concentrations around each equilibrium point, i.e. their tendency to increase or decrease. To do that, one studies the sign of the derivatives:

$$\begin{aligned} &\tfrac{d[Prey]}{dt} \geq 0 \Rightarrow k_1 - k_2[Predator] \geq 0 \Rightarrow [Predator] \leq \tfrac{k_1}{k_2};\\ &\tfrac{d[Predator]}{dt} \geq 0 \Rightarrow k_2[Prey] - k_3 \geq 0 \Rightarrow [Prey] \geq \tfrac{k_3}{k_2}. \end{aligned} \tag{28}$$

The behavior around the steady states is depicted in Figure 4.

5.2 Mass Conservation Relations

In this section we introduce mass conservation relations and their importance in modeling reaction-based systems. For a more detailed presentation and additional examples we refer to [24].

Identifying the mass conservation relations in a given model is one of the first analyzes that a modeler typically performs. It gives an insight into the dynamics of the model, but at the same time it reduces the number of free variables in the model. Mathematically, a mass conservation relation is a linear combination of concentrations of species that is constant in time:

$$g^T S = C, \tag{29}$$

where g is a vector with some constant entries, S is the species concentrations vector, and C is some constant. An implication of mass conservation relations is that some of the stoichiometric matrix rows are linearly dependent, i.e.

$$g^T N = 0^T. \tag{30}$$

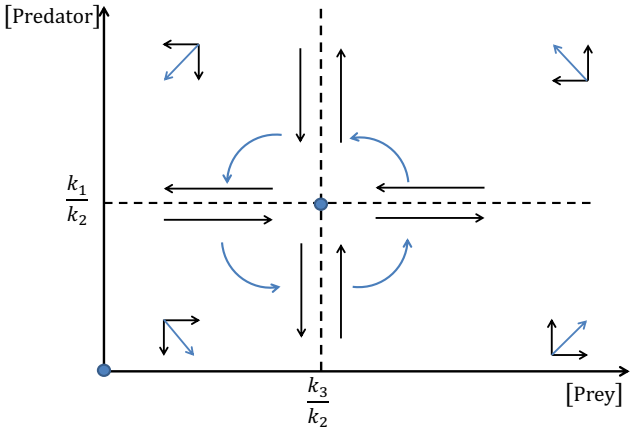


Fig. 4. Steady-state analysis of the Lotka-Volterra system. The blue dots are the two steady states of the system. The black arrows indicate how the concentration of each species increases or decreases (as established in (28)). The blue arrows are the combination of *Predator* and *Prey* concentrations tendencies, and they show how the dynamics of the system changes between the four areas delimited by the dotted lines. The behavior around the $(k_3/k_2, k_1/k_2)$ point suggests periodicity; this is confirmed by Figure 1. Both equilibrium points are unstable, as indicated by the blue arrows.

Equations (29) and (30) are equivalent. Derivating the former equation and taking into account Equation (4) yields

$$(g^T S)' = g^T \dot{S} = g^T N v = 0.$$

There may be more linearly independent vectors g that satisfy Equation (30), each denoting a different mass conservation relation. The number of mass conservation relations is given by $m-$ rank(N), where m is the number of species in the system. The full set of vectors g describing these mass conservation relations form a so-called *conservation matrix* G , see [24], with the property

$$GN = 0.$$

Consequently, G^T is a kernel matrix for N^T. A conservation matrix G can be determined using the Gauss algorithm, and it is not unique (any other matrix $G' = PG$, where P is any nonsingular matrix of appropriate dimensions, is a valid conservation matrix).

Example 7. Consider the following system of biochemical reactions:

$$\begin{aligned}
&2A \rightleftarrows A_2;\\
&A_2 + B \rightleftarrows A_2 : B;\\
&A_2 : B \to C + A_2 : B;\\
&C \to \emptyset.
\end{aligned}$$

The species vector, stoichiometric coefficients matrix and the conservation matrix read:

$$S = \begin{bmatrix} A \\ A_2 \\ B \\ A_2 : B \\ C \end{bmatrix}, \quad N = \begin{bmatrix} -2 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}, \quad G = \begin{bmatrix} 1 & 2 & 0 & 2 & 0 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix}.$$

The mass conservation relations induced by G are:

$$\begin{aligned} &[A] + 2[A_2] + 2[A_2 : B] = C_1, \\ &[B] + [A_2 : B] = C_2, \end{aligned} \tag{31}$$

for some constants C_1, C_2.

The mass conservation relations are important for reducing the system of differential equations $\dot{S} = Nv$ that describe the dynamics of the model. Each mass conservation relation introduces one dependent variable, which can be expressed in terms of the independent variables, and thus eliminated from the system of ODEs. The two mass conservation relations in Equation (31) could be used to express the dependency between $[A], [B]$ and the rest of the species concentrations:

$$\begin{aligned} [A] &= C_1 - 2[A_2] - 2[A_2 : B], \\ [B] &= C_2 - [A_2 : B]. \end{aligned}$$

This reduces the initial system of ODEs from 5 to 3 equations.

5.3 Sensitivity Analysis

Sensitivity analysis is a method of estimating the changes that small perturbations in the parameters of a model induce in the system. With this type of analysis, one can estimate the robustness of a model against small changes, and also identify ways of inducing a desired change into the model. There exist many methods for sensitivity analysis, some suitable for spatially homogeneous constant-parameter reaction-based models, others suitable for systems with space- and time-dependent parameters, or stochastic models. For a review of multiple methods, we refer the reader to [59, 62]. One of the questions often encountered in biochemical systems is what changes should the system undergo such that the new steady state satisfies certain properties.

There are two types of sensitivity analysis: *local sensitivity analysis*, and *global sensitivity analysis*. In the global approach, all parameters are varied at once, and the sensitivity is measured over the entire range of each parameter. In the local analysis, only one parameter is varied at a time, within a small interval around some nominal value. Generally, it is assumed that input-output relationships are linear. We only focus here on local sensitivity analysis.

We consider the system of ODEs describing a system to be expressed as a function of the concentrations of all species and all the parameter values:

$$\frac{d[S_i]}{dt} = f_i([S_1], [S_2], ..., [S_m], \kappa), \tag{32}$$

where $\kappa = (k_1, k_2, ..., k_n)^T$ is the rate constants vector (assuming without loss of generality that the system comprises n irreversible reactions). Let $\mathcal{S}(t,\kappa) = ([S_1](t,\kappa), [S_2](t,\kappa), ..., [S_m](t,\kappa))^T$ be the solution of Equation (32) with respect to κ, also called *sensitivity matrix*. The elements of the matrix are the partial derivatives $\partial[S_i]/\partial k_j$, also called *first-order local sensitivity coefficients*.

There are many ways of determining the local sensitivity of the concentrations. The simplest method is the *brute force method* (also called *indirect method*, or *finite-difference method*), that uses the finite difference approximation. The j-th parameter, k_j, changes with the amount δk_j at time point t_1, and all other parameters remain unchanged. One can compute the new matrix $[S]$ using the change between the initial and the perturbed solution, see Equation (33). The method requires $n+1$ runs, one for the initial values of the parameters and n modifying each of the parameters at a time.

$$\frac{\partial[S](t_2)}{\partial k_j(t_1)} = \frac{[S](t2, k_j + \delta k_j) - [S](t_2, k_j)}{\delta k_j}, 1 \leq j \leq n. \tag{33}$$

This method is widely used because of its simplicity, but other more efficient methods exist, e.g. the *direct method*. This method solves the differential equations for the sensitivity coefficients $\partial[S_i]/\partial k_j$, by differentiating Equation (32). This results in the following set of sensitivity equations:

$$\frac{d}{dt}\frac{\partial[S]}{\partial k_j} = \mathcal{J}\frac{\partial[S]}{\delta k_j} + \frac{\partial f}{\partial k_j}, 1 \leq j \leq n,$$

where $\mathcal{J}$ is the Jacobian for Equation (32). For a complete mathematical derivation of this result, see [62].

Perturbations should be small enough to yield small errors in the indirect method, and large enough to surpass the simulation inaccuracies of ODE solvers, for the direct method, see [62]. Other methods of computing the sensitivity of a model to parameter changes exist, e.g. the Green function method, polynomial approximation method, AIM method, detailed in [54, 59].

Very often, sensitivity analysis is focused on the steady states, when concentrations are constant. In this case, the sensitivity coefficients are computed as solutions to the system

$$\frac{d}{dt}\frac{\partial[S]}{\partial k_j} = 0,$$

and reflect the dependency of the steady state on the parameters. If the steady state is asymptotically stable, then one can consider the limit $\lim_{t\to\infty}(\partial[S]/\partial k_j)(t), 1 \leq j \leq n$, called *stationary sensitivity coefficients*. The system can be written as

$$\frac{\partial[S]}{\partial k_j} = -\mathcal{J}F_j, 1 \leq j \leq n,$$

where $\mathcal{J}$ is the value of the jacobian at steady state, and F_j is the j-th column in the matrix $F = (\partial f_r/\partial k_s)_{m\times n}$ computed at steady state. Sensitivity coefficients can be computed in many software applications, e.g. in COPASI [31].

6 The Heat Shock Response Model

We consider in this section a larger modeling case-study to which we apply some of the techniques discussed in the chapter. The eukaryotic heat shock response is an evolutionarily conserved bio-regulatory network, crucial to cell survival. It acts as a defence mechanism that regulates the cellular response to proteotoxicity induced by diverse physiological and environmental stressors, such as elevated temperatures. Exposure of proteins to elevated temperatures causes protein misfolding, which results in the constitution of large aggregates that eventually induce apoptosis (controlled cell death). Protein homeostasis is promoted by augmenting the level of molecular chaperons.

6.1 The Reaction-Based Model

We consider here the basic molecular model for the heat shock response introduced in [53]. Elevated temperatures cause protein misfolding and accumulation of misfolded proteins in large conglomerates that induce cell death. The key role in homeostasis restoration is played by *heat shock proteins* (hsp), which chaperone the misfolded proteins, promoting the folding of proteins. The transactivation of hsp-encoding genes regulates the heat shock response. Heat shock factors (hsf) activate gene transcription. In the absence of stress, heat shock factors are present in a monomeric conformation and they are bound to a great extent to heat shock proteins. However, heat stress actuates the dimerization (hsf_2) and consequently trimerization (hsf_3) of heat shock factors, a DNA binding-competent conformation. Due to their high affinity toward the heat shock element (hse), hsf trimers bind to the heat shock elements, promoting the transcription and translation of the gene. Consequently, DNA binding activates hsp synthesis, see [53, 55].

Once the heat stress is removed, hsp synthesis is turned off as follows: hsp's sequestrate free hsf's (residing in the constitution of hsp:hsf complexes), break hsf_2 and hsf_3 and induce DNA unbinding, see [53, 55]. Subsequently, DNA transcription is turned off and the formation of new hsf trimers repressed. The heat shock response mechanism is switched back on when the temperature is again elevated, impelling the proteins in the cell (prot) to misfold and hsp:hsf complexes to break down. The reactions of the molecular model in [53] are shown in Table 2.

Table 2. The molecular model for the eukaryotic heat shock response proposed in [53]

$2\,\mathsf{hsf} \rightleftarrows \mathsf{hsf}_2$	$\mathsf{hsp} + \mathsf{hsf}_3 \to \mathsf{hsp}:\mathsf{hsf} + 2\,\mathsf{hsf}$
$\mathsf{hsf} + \mathsf{hsf}_2 \rightleftarrows \mathsf{hsf}_3$	$\mathsf{hsp} + \mathsf{hsf}_3:\mathsf{hse} \to \mathsf{hsp}:\mathsf{hsf} + 2\,\mathsf{hsf} + \mathsf{hse}$
$\mathsf{hsf}_3 + \mathsf{hse} \rightleftarrows \mathsf{hsf}_3:\mathsf{hse}$	$\mathsf{hsp} \to \emptyset$
$\mathsf{hsf}_3:\mathsf{hse} \to \mathsf{hsf}_3:\mathsf{hse} + \mathsf{hsp}$	$\mathsf{prot} \to \mathsf{mfp}$
$\mathsf{hsp} + \mathsf{hsf} \rightleftarrows \mathsf{hsp}:\mathsf{hsf}$	$\mathsf{hsp} + \mathsf{mfp} \rightleftarrows \mathsf{hsp}:\mathsf{mfp}$
$\mathsf{hsp} + \mathsf{hsf}_2 \to \mathsf{hsp}:\mathsf{hsf} + \mathsf{hsf}$	$\mathsf{hsp}:\mathsf{mfp} \to \mathsf{hsp} + \mathsf{prot}$

This molecular model is clearly on a high level of abstraction, for the sake of easing its analysis. For example, note that the eukaryotic cell presents various classes of heat shock proteins, denominated according to their molecular weight, e.g., Hsp60, Hsp70, Hsp90. However, in this molecular model, they are all referred to as belonging to the same class, with Hsp70 as common denominator. The same assumptions are made for hsf and hse. Furthermore, the model considers all proteins uniformly, distinguishing only between the ones that are correctly folded (prot) and the misfolded ones (mfp). The model contains also simplified representations of some cellular mechanisms, e.g., protein synthesis and degradation, see [53] for more details.

The molecular model in [53] satisfies the following three mass-conservation relations, for the total amount of hsf, the total amount of proteins (excluding hsp and hsf) and for the total amount of hse:

- $[\mathsf{hsf}] + 2[\mathsf{hsf}_2] + 3[\mathsf{hsf}_3] + 3[\mathsf{hsf}_3\colon\mathsf{hse}] + [\mathsf{hsp}\colon\mathsf{hsf}] = C_1$,
- $[\mathsf{prot}] + [\mathsf{mfp}] + [\mathsf{hsp}\colon\mathsf{mfp}] = C_2$,
- $[\mathsf{hse}] + [\mathsf{hsf}_3\colon\mathsf{hse}] = C_3$,

where C_1, C_2 and C_3 are constants.

6.2 The Mathematical Model

Given the molecular model in Table 2, we consider a mathematical model derived through the principle of mass action, formulated as a system of ordinary differential equations ([37]). The rate coefficient for protein misfolding ($\mathsf{prot} \rightarrow \mathsf{mfp}$) is described by the following formula:

$$\varphi(T) = (1 - \frac{0.4}{e^{T-37}}) \cdot 1.4^{T-37} \cdot 1.45 \cdot 10^{-5} s^{-1},$$

where T is the temperature of the environment, expressed in $^\circ C$, in accordance to [52]. Each species X in the molecular model is associated to a continuous, time-dependent function $[X](t)$, expressing the concentration of the respective reactant. The dynamics of the system is described through the system of differential equations in Table 3.

The initial values of all species and the kinetic rate constants were estimated in [53], by imposing the following three conditions:

(i) At $37^\circ C$ the system is in a steady state, since the model should not reveal any response in the absence of the heat stress;
(ii) At $42^\circ C$, the numerical predictions for DNA binding ($[\mathsf{hsf}_3\colon\mathsf{hse}](t)$) should be in accordance with the experimental data reported in [36];
(iii) At $42^\circ C$, the numerical prediction of the model for $[\mathsf{hsp}](t)$ should confirm the data obtained in [53] through a de-novo fluorescent reporter-based experiment.

The numerical setup obtained in [53] for the heat shock response model is shown in Table 4.

The estimation of parameters was based on the experimental data in [36] on DNA binding in HeLa cells for a temperature of $42^\circ C$. Moreover, the model should also be in a steady state at $37^\circ C$. Hence, seven more independent algebraic relations on the set of

Table 3. The system of ODE's associated with the biochemical model proposed in [53]

$$
\begin{aligned}
d[\mathsf{hsf}]/dt &= -2k_1^+[\mathsf{hsf}]^2 + 2k_1^-[\mathsf{hsf}_2] - k_2^+[\mathsf{hsf}][\mathsf{hsf}_2] + k_2^-[\mathsf{hsf}_3] \\
&\quad - k_5^+[\mathsf{hsf}][\mathsf{hsp}] + k_5^-[\mathsf{hsp{:}\,hsf}] + k_6[\mathsf{hsf}_2][\mathsf{hsp}] \\
&\quad + 2k_7[\mathsf{hsf}_3][\mathsf{hsp}] + 2k_8[\mathsf{hsf}_3{:}\,\mathsf{hse}][\mathsf{hsp}]; \\
d[\mathsf{hsf}_2]/dt &= k_1^+[\mathsf{hsf}]^2 - k_1^-[\mathsf{hsf}_2] - k_2^+[\mathsf{hsf}][\mathsf{hsf}_2] + k_2^-[\mathsf{hsf}_3] \\
&\quad - k_6[\mathsf{hsf}_2][\mathsf{hsp}]; \\
d[\mathsf{hsf}_3]/dt &= k_2^+[\mathsf{hsf}][\mathsf{hsf}_2] - k_2^-[\mathsf{hsf}_3] - k_3^+[\mathsf{hsf}_3][\mathsf{hse}] + k_3^-[\mathsf{hsf}_3{:}\,\mathsf{hse}] \\
&\quad - k_7[\mathsf{hsf}_3][\mathsf{hsp}]; \\
d[\mathsf{hse}]/dt &= -k_3^+[\mathsf{hsf}_3][\mathsf{hse}] + k_3^-[\mathsf{hsf}_3{:}\,\mathsf{hse}] + k_8[\mathsf{hsf}_3{:}\,\mathsf{hse}][\mathsf{hsp}]; \\
d[\mathsf{hsf}_3{:}\,\mathsf{hse}]/dt &= k_3^+[\mathsf{hsf}_3][\mathsf{hse}] - k_3^-[\mathsf{hsf}_3{:}\,\mathsf{hse}] - k_8[\mathsf{hsf}_3{:}\,\mathsf{hse}][\mathsf{hsp}]; \\
d[\mathsf{hsp}]/dt &= k_4[\mathsf{hsf}_3{:}\,\mathsf{hse}] - k_5^+[\mathsf{hsf}][\mathsf{hsp}] + k_5^-[\mathsf{hsp{:}\,hsf}] - k_6[\mathsf{hsf}_2][\mathsf{hsp}] \\
&\quad - k_7[\mathsf{hsf}_3][\mathsf{hsp}] - k_8[\mathsf{hsf}_3{:}\,\mathsf{hse}][\mathsf{hsp}] - k_{11}^+[\mathsf{hsp}][\mathsf{mfp}] \\
&\quad + (k_{11}^- + k_{12})[\mathsf{hsp{:}\,mfp}] - k_9[\mathsf{hsp}]; \\
d[\mathsf{hsp{:}\,hsf}]/dt &= k_5^+[\mathsf{hsf}][\mathsf{hsp}] - k_5^-[\mathsf{hsp{:}\,hsf}] + k_6[\mathsf{hsf}_2][\mathsf{hsp}] \\
&\quad + k_7[\mathsf{hsf}_3][\mathsf{hsp}] + k_8[\mathsf{hsf}_3{:}\,\mathsf{hse}][\mathsf{hsp}]; \\
d[\mathsf{mfp}]/dt &= \varphi(T)[\mathsf{prot}] - k_{11}^+[\mathsf{hsp}][\mathsf{mfp}] + k_{11}^-[\mathsf{hsp{:}\,mfp}]; \\
d[\mathsf{hsp{:}\,mfp}]/dt &= k_{11}^+[\mathsf{hsp}][\mathsf{mfp}] - (k_{11}^- + k_{12})[\mathsf{hsp{:}\,mfp}]; \\
d[\mathsf{prot}]/dt &= -\varphi(T)[\mathsf{prot}] + k_{12}[\mathsf{hsp{:}\,mfp}].
\end{aligned}
$$

parameters and initial values are derived. Therefore, the model comprises 17 independent values that require estimation. The above-mentioned conditions are satisfied by the values in Table 4. These values have been attained by means of parameter estimation in COPASI [31]. The model is fit with regard to the DNA binding experimental data in [36]. The model predictions regarding $\mathsf{hsf}_3{:}\,\mathsf{hse}$ compared with the experimental data of [36] are shown in Figure 5.

6.3 Model Validation

The model exhibits a very low rate for protein misfolding for a temperature of $37°C$ and a high rate for protein folding, in compliance with [5] and [35]. The model also predicts a transient increase in the level of hsf trimers, in accordance with [30]. The model confirms that dimers are only a transient form between monomers and trimers, and that the level of dimers is low throughout the simulation, regardless of the temperature.

Another validation test consisted in applying the heat shock response twice subsequently. The second heat shock was applied after the heat shock proteins had attained a maximal level. The model in [53] predicted the response to the second heat shock to be

Table 4. The numerical values of the parameters (A) and the initial values of the variables (B) of the heat shock response model proposed in [53]

A

Param.	Value	Units
k_1^+	3.49	$\frac{ml}{\#\cdot s}$
k_1^-	0.19	s^{-1}
k_2^+	1.07	$\frac{ml}{\#\cdot s}$
k_2^-	10^{-9}	s^{-1}
k_3^+	0.17	$\frac{ml}{\#\cdot s}$
k_3^-	$1.21 \cdot 10^{-6}$	s^{-1}
k_4	$8.3 \cdot 10^{-3}$	s^{-1}
k_5^+	9.74	$\frac{ml}{\#\cdot s}$
k_5^-	3.56	s^{-1}
k_6	2.33	$\frac{ml}{\#\cdot s}$
k_7	$4.31 \cdot 10^{-5}$	$\frac{ml}{\#\cdot s}$
k_8	$2.73 \cdot 10^{-7}$	$\frac{ml}{\#\cdot s}$
k_9	$3.2 \cdot 10^{-5}$	s^{-1}
k_{11}^+	$3.32 \cdot 10^{-3}$	$\frac{ml}{\#\cdot s}$
k_{11}^-	4.44	s^{-1}
k_{12}	13.94	s^{-1}

B

Variable	Initial conc.
[hsf]	0.67
[hsf_2]	$8.7 \cdot 10^{-4}$
[hsf_3]	$1.2 \cdot 10^{-4}$
[hse]	29.73
[hsf_3: hse]	2.96
[hsp]	766.88
[hsp: hsf]	1403.13
[mfp]	517.352
[hsp: mfp]	71.65
[prot]	1.15×10^8

greatly diminished in intensity. Indeed, a diminished response for the second heat shock could be anticipated since the level of heat shock proteins (hsp's) is already elevated as a consequence of the first heat shock. A similar result was reported in [52].

Another validation method consisted in simulating the model for a temperature of $43°C$ and comparing the results with those of [55]. The model in [53] predicts a prolonged transactivation for DNA binding, as opposed to the model in [55], but it is consistent with the experimental data in [1]. An experiment consisting in the removal of the heat shock at $42°C$ at the peak of the response exhibited an accelerated attenuation phase, complying with the results reported by [55].

An alternative verification scenario focused on the prediction of the evolution of heat shock proteins (hsp's) over time. This method required the use of a quantitative reporter system founded on *yellow fluorescent proteins* (yfp's). This method was based on the assumption that fluorescence intensity is virtually linear reported to the level of yfp's. As yfp's transactivation is regulated by their own heat shock elements, denoted in [53] by hse′, transcription and degradation kinetics ($k_4{}'$ and $k_9{}'$ respectively), their evolution in time may be described by the following differential equation:

$$d[\mathsf{yfp}]/dt = k_4'[\mathsf{hsf}_3\mathsf{:hse}'] - k_9'[\mathsf{yfp}], \tag{34}$$

for some positive constants k_4', k_9' accounting for the kinetic rate constants of yfp synthesis and of yfp degradation. The extended model, including equation (34), takes into

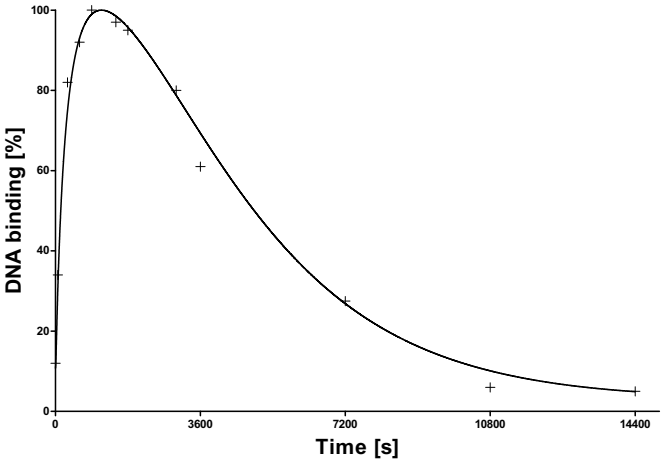


Fig. 5. The dynamic behavior of hsf_3:hse in the best fitted model. The continuous line is the model prediction and the crossed points indicate the experimental data of [36].

account all numerical values from the basic model, described in Table 2, and the numerical values for the new rate constants $k_4{}'$ and $k_9{}'$ were estimated so that the fit for yfp's complies with the experimental data.

6.4 Model Analysis

Sensitivity analysis. The first analysis approach consisted in estimating the scaled steady state sensitivity coefficients, see [59], of all variables against reaction rate constants and initial concentrations. Given a variable X and a parameter p, the *scaled steady state sensitivity coefficient* of variable X against parameter p is defined by:

$$\lim_{t\to\infty} \frac{\partial ln(X)}{\partial \ln(p)(t)}.$$

The coefficients described above represent the relative variance of the steady state when the model undergoes infinitesimal changes in parameter p. The sensitivity coefficients of all variables against reaction rate constants $k_1{}^-, k_2{}^-, k_3{}^-, k_7$ proved to be all insignificant, suggesting that the reactions corresponding to those rate constants may not be crucial to the global behavior of the model. For this aim, the model was altered so as to exclude the reactions corresponding to the aforementioned kinetic rate constants, namely the backward reactions for dimerization, trimerization, DNA binding and DNA unbinding. The new model attained as such satisfies the validation tests described in Section 6.3. This suggests that hsf dimers and trimers are steady configurations and that non-hsp-mediated DNA unbinding is negligible. While the breaking of trimers

(by hsp) does not affect greatly the overall behavior of the model, the reaction describing the breaking of trimers (by hsp) proved to have a substantial impact on the evolution of hsp and mfp.

The variation between the steady state levels of hsp and mfp are correlated, see [53], which is consistent with the biological knowledge that hsp's have a major role in chaperoning mfp's. Table 5 shows the largest sensitivity coefficients for hsp and mfp. The coefficients with respect to ${k_5}^+$ and ${k_5}^-$ are the highest, suggesting that the reaction describing hsf sequestration (forward)/dissipation of hsp:hsf ($\mathsf{hsp} + \mathsf{hsf} \leftrightarrows \mathsf{hsp{:}hsf}$) is the main feedback loop. The forward direction, hsf sequestration, hereafter compels the ceasing of transcription, inducing an augmentation in the level of mfp and a decrease in that of hsp. The backward direction (dissipation of hsp:hsf), however, actuates an increase in the level of hsp and hsf and a reduction of mfp. Considering the coefficients in Table 5 in descending order, the next set of coefficients to discuss consists of ${k_1}^+$, ${k_2}^+$ and k_4, corresponding to the forward directions of dimer(trimer) formation and DNA binding respectively, suggesting the augmentation of the transcription level and thereupon the level of hsp. On the contrary, the reactions describing the breaking of dimers, hsp degradation and protein misfolding, diminish the transcription level. The reactions influencing the level of mfp alone are the reactions corresponding to the sequestration of mfp's/dissipation of hsp:mfp (see coefficients corresponding to ${k_{11}}^+$ and ${k_{11}}^-$ in Table 5) and protein refolding (same for k_{12}).

Among the sensitivity coefficients of hsp and mfp with respect to the initial concentrations, the one dependent on the initial level of hsp:hsf ($\mathsf{hsp{:}hsf}(0)$) was the most relevant. On the other hand, the sensitivity coefficients of hsp and mfp with respect to the level of any of the hsf species (monomers, dimers or trimers) were insignificant. This is to be expected since initially the majority of hsf's is sequestrated by hsp's and the initial levels of dimers and trimers are reduced, which is consistent with [30]. Consequently, the sensitivity coefficient with respect to $\mathsf{hsp{:}hsf}(0)$ should be conceived as describing a dependency over the total initial amount of hsf.

The sensitivity coefficients with respect to the initial amount of hse were insignificant, which is justified by the consideration of the sensitivity coefficients around the steady state. For instance, for a lower initial amount of hse, the response reaches hereafter the same steady state. A higher level of $\mathsf{hsf}(0)$ brings no change in the evolution of the response. The sensitivity coefficients of hsp and mfp with respect to $\mathsf{hsp}(0)$ were also insignificant.

Model identifiability. Looking into the model identifiability problem, alternative good numerical fits were searched for, using the same fitting data as in the model fitting procedure described above. Several were found, but none of them passed the additional validation tests described in the previous section. Then the *Latin Hypercube Sampling* method was applied to sample the distribution of the fitting score function. The first step was to generate a sample of $N = 100000$ combinations of parameter values, as described in Section 4. For each of them, the initial values were chosen so that they are a steady state of the model at 37°C. Out of these, the analysis was continued only for those combinations that were "responsive", where a model was declared responsive if $\mathsf{hsf_3{:}hse}(900) \geq 20$ (note that the experimental data indicated that the peek of the

Table 5. The largest scaled steady state sensitivity coefficients of hsp and mfp. The coefficients are identical for both $37°C$ and $42°C$ [53]

Parameter description	p	$\frac{\partial ln(\mathsf{hsp})}{\partial ln(p)}\|_{t\to\infty}$	$\frac{\partial ln(\mathsf{mfp})}{\partial ln(p)}\|_{t\to\infty}$
Sequestration of hsf	${k_5}^+$	−0.50	0.50
Dissipation of hsp:hsf	${k_5}^-$	0.50	−0.50
Formation of dimers	${k_1}^+$	0.17	−0.17
Formation of trimers	${k_2}^+$	0.17	−0.17
Transcription, translation	k_4	0.17	−0.17
Affinity of hsp for hsf_2	k_6	−0.17	0.17
Affinity of hsp for hsf_3:hse	k_8	−0.17	0.17
Degradation of hsp	k_9	−0.17	0.17
Affinity of hsp for mfp	${k_{11}}^+$	0.00	−1.00
Dissipation of hsp:mfp	${k_{11}}^-$	0.00	0.24
Protein refolding	k_{12}	0.00	−0.24
Initial level of hsp:hsf	$\mathsf{hsp:hsf}(0)$	0.50	−0.50

response is reached after 900 time units). The result was interesting: there were only 31506 models satisfying the constraint, already suggesting that finding suitable alternative model fits is a difficult problem. For each of these models we calculated the fit quality as discussed in Section 4; the result is plotted in Figure 6, showing clearly our best fit as an outlier in the fit quality distribution. More details on the identifiability of the heat shock response model can be found in [53]. This suggests that fitting the simple heat shock response model in Table 2 to the experimental data in [36] and to the steady-state condition for the initial values is indeed a difficult numerical problem.

7 Discussion

The focus of our chapter has been on the practical use of modeling with ordinary differential equations in biology. Our choice of topics to discuss has been driven by targeting primarily the computer science community and by the space limitations. This chapter should only be seen as a "teaser" for modeling with ODEs in biology; for a more comprehensive reading on this topic, many excellent textbooks exist, such as [11,32,47,48, 56, 58]. We only considered in this chapter reaction-based models and started by discussing how to associate to them an ODE-based model; we presented briefly several laws for biochemical kinetics: mass-action, Michaelis-Menten, Goldbeter-Koshland, Hill, and inhibition. One should note that many other types of models exist, see, e.g., [9]. We then discussed the parameter estimation problem, including model identifiability, measures for fit quality, and fit-preserving model refinement. We then introduced several analysis methods for ODE-based models: steady state analysis, mass conservation, and sensitivity analysis. In addition to some smaller examples discussed throughout the chapter, we dedicated a separate section to a larger case-study on the eukaryotic heat shock response.

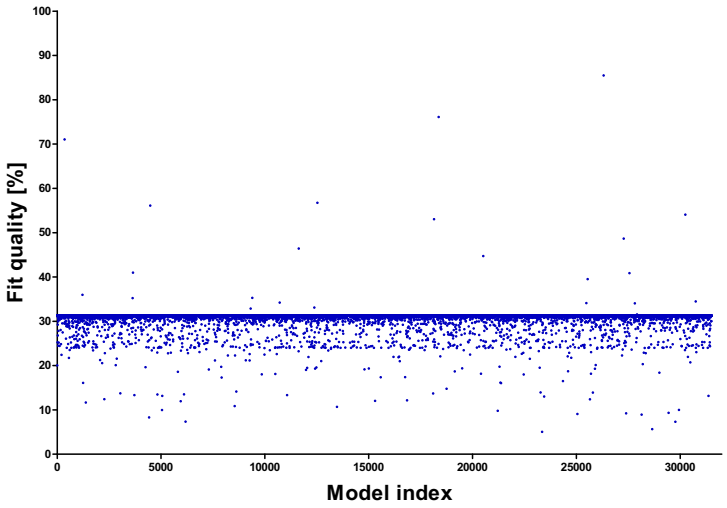


Fig. 6. The distribution of the model fit quality among 31506 model variants obtained through the Latin hypercube sampling method. Most models exhibit a constant level of $\mathsf{hsf_3{:}hse}$, very different from the dynamic behavior in Figure 5; these models yield a numerical value for the fit quality around 30%. The quality of our best fit is around 10^{-30}.

There are many computational benefits that modeling with ODEs brings, including fast numerical simulations, many methods for parameter estimation, several highly useful static and dynamic analysis techniques, such as mass conservation, steady state analysis, flux-balance analysis, metabolic control analysis, sensitivity analysis, etc. At the same time, the ODE-based approach also suffers from several difficulties. The one that is most discussed is the inability to account for stochastic noise in a system, which might be problematic especially in cases where there are relatively small species; a detailed discussion about the physical limitations of the ODE-based approach is in [14, 15]. Another difficulty is in the need for knowing a potentially large number of kinetic parameters; measuring them experimentally is sometimes impossible, while estimating them computationally suffers from model identifiability issues. A partial solution here is the approach based on quantitative model refinement, see [34]. Another partial solution is in terms of static, rather than dynamic analysis, often performed around the steady states; such an approach is modeling based on flux balance analysis, see [51].

The stochastic approach, either in terms of continuous time Markov chains (CTMC) and the chemical master equation, or in terms of higher-level formalisms (such as Petri nets or process algebra) based on a CTMC semantic, is often offered as a solution to the physical limitations of the ODE-based approach. It is important however to understand the limitations of both approaches so that we can take advantage of the benefits of either one, whenever they are applicable. In Table 6 we summarized several aspects about modeling with ODEs and with CTMCs, and placed them in mirror for an easy comparison. It is also important to point out that in the case of very large models, both approaches are insufficient, see Figure 7.

Table 6. Some of the differences between the deterministic and the stochastic modeling approaches

	Deterministic approach	*Stochastic approach*
Fundamental assumptions	the system is well-stirred and at thermodynamical equilibrium	the system is well-stirred and at thermodynamical equilibrium
Modeling goal	it models the average behavior of the system	it models individual runs of the system
Concept	based on the concept of diffusion-like reactions	based on the concept of reactive molecular collisions
Type of model	the time evolution of the model is a continuous process	the time evolution of the model is a random-walk process through the possible states
Math model	governed by a set of ODEs	governed by a single ODE: the chemical master equation
Analytic solution	the system of ODEs is often impossible to solve analytically	the chemical master equation is often impossible to solve
Small populations	conceptual difficulties when small populations are involved	no difficulties with small populations
Numerical simulations	fast	Gillespies algorithm is slow; many runs are needed

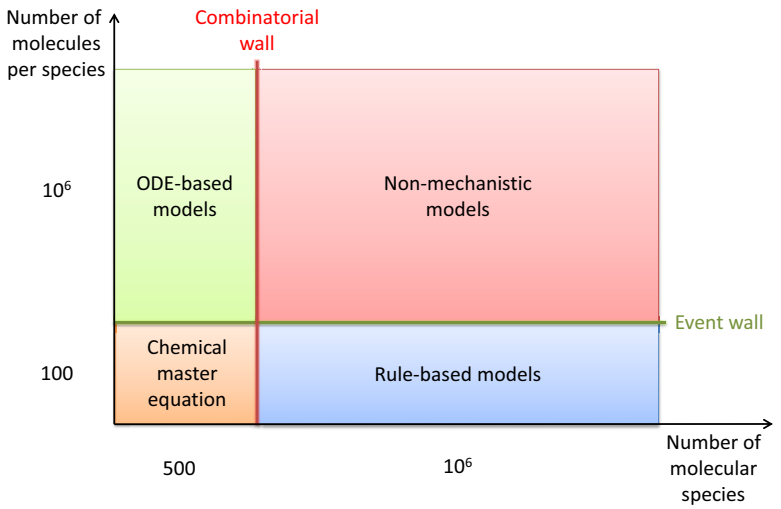


Fig. 7. Modeling limitations depending on the size of the model. Adapted from Walter Fontana `http://fontana.med.harvard.edu/`

The ODE-based approach to computational modeling is (still) arguably the standard choice for biomodelers, especially on the biological side of the community. There are many advantages that it brings, as there are clear limitations. Even in cases where another modeling approach is taken, the corresponding ODE-based model is often also built to serve as comparison to related (ODE-based) models and to make available tools such as parameter estimation or steady state analysis. Moreover, on top of the ODE-based semantic there are many other discrete techniques that can be added to give further insight into the model: Petri net tools, control analysis, network motif identification, etc. In the continuing debate of 'discrete vs. continuous biomodeling' we argue that it is good to retain the advantages of both worlds and use them to their full potential whenever applicable.

References

1. Abravaya, K., Phillips, B., Morimoto, R.I.: Attenuation of the heat shock response in hela cells is mediated by the release of bound heat shock transcription factor and is modulated by changes in growth and in heat shock temperatures. Genes & Development 5(11), 2117–2127 (1991)
2. Ali, M.M., Storey, C., Törn, A.: Application of stochastic global optimization algorithms to practical problems. Journal of Optimization Theory and Applications 95(3), 545–563 (1997)
3. Azimi, S., Gratie, D.-E., Iancu, B., Petre, I.: Three approaches to quantitative model refinement with applications to the heat shock response. Technical report, Turku Centre for Computer Science (2013)
4. Baker, S.M., Schallau, K., Junker, B.H.: Comparison of different algorithms for simultaneous estimation of multiple parameters in kinetic metabolic models. Journal of Integrative Bioinformatics 7(3), 1–9 (2010)
5. Ballew, R.M., Sabelko, J., Gruebele, M.: Direct observation of fast protein folding: the initial collapse of apomyoglobin. Proceedings of the National Academy of Sciences 93(12), 5759–5764 (1996)
6. Boender, G., Romeijn, E.: Stochastic methods. In: Handbook of Global Optimization: nonconvex optimization and its applications, pp. 829–869. Kluwer Academic Publishers (1995)
7. Briggs, G.E., Haldane, J.B.S.: A note on the kinetics of enzyme action. Biochemical Journal 19(2), 338–339 (1925)
8. Chen, W.W., Schoeberl, B., Jasper, P.J., Niepel, M., Nielsen, U.B., Lauffenburger, D.A., Sorger, P.K.: Input-output behavior of erbb signaling pathways as revealed by a mass action model trained against dynamic data. Molecular Systems Biology 5, 239 (2009)
9. Ciobanu, G., Rozenberg, G. (eds.): Modeling in Molecular Biology. Springer (2004)
10. Czeizler, E., Rogojin, V., Petre, I.: The phosphorylation of the heat shock factor as a modulator for the heat shock response. IEEE-ACM Trans. Comp. Biol. Bioinf. 9(5), 1326–1337 (2012)
11. Edelstein-Keshet, L.: Mathematical Models in Biology. McGraw-Hill, New York (1988)
12. Fisher, J., Henzinger, T.A.: Executable cell biology. Nature Biotechnology 25, 1239–1249 (2007)
13. Gessel, I.M., Stanley, R.P.: Algebraic enumeration. Handbook of Combinatorics 2, 1021–1061 (1995)
14. Gillespie, D.T.: A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. Journal of Computational Physics 22(4), 403–434 (1976)
15. Gillespie, D.T.: Exact stochastic simulation of coupled chemical reactions. The Journal of Physical Chemistry 81(25), 2340–2361 (1977)

16. Gillespie, D.T.: A rigorous derivation of the chemical master equation. Physica A: Statistical Mechanics and its Applications 188(1), 404–425 (1992)
17. Giordano, F., Weir, M., Fox, W.: A first course in mathematical modeling, 3rd edn. Thomson (2003)
18. Goldbeter, A., Koshland, D.E.: An amplified sensitivity arising from covalent modification in biological systems. Proceedings of the National Academy of Sciences 78(11), 6840–6844 (1981)
19. Grossmann, I.E.: Global optimization in engineering design. Kluwer Academic Publishers (1996)
20. Guldberg, C.M., Waage, P.: Studies concerning affinity. CM Forhandlinger: Videnskabs-Selskabet i Christiana 35, 92–111 (1864)
21. Guldberg, C.M., Waage, P.: Etudes sur les affinités chimiques. Brøgger & Christie (1867)
22. Haldane, J.B.S.: Enzymes (1930, 1965)
23. Harmer, R.: Rule-based modelling and tunable resolution. EPTCS 9, 65–72 (2009)
24. Heinrich, R., Schuster, S.: The regulation of cellular systems, vol. 416. Chapman & Hall, New York (1996)
25. Heinrich, R., Schuster, S.: The modelling of metabolic systems. structure, control and optimality. Biosystems 47(1), 61–77 (1998)
26. Helton, J.C., Davis, F.J.: Illustration of sampling-based methods for uncertainty and sensitivity analysis. Risk Analysis 22(3), 591–622 (2002)
27. Helton, J.C., Davis, F.J.: Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. Reliability Engineering and System Safety 81, 23–69 (2003)
28. Henri, V.: Lois générales de l'action des diastases. Librairie Scientifique A. Hermann (1903)
29. Hill, A.V.: A new mathematical treatment of changes of ionic concentration in muscle and nerve under the action of electric currents, with a theory as to their mode of excitation. The Journal of Physiology 40(3), 190–224 (1910)
30. Holmberg, C.I., Tran, S.E.F., Eriksson, J.E., Sistonen, L.: Multisite phosphorylation provides sophisticated regulation of transcription factors. Trends in Biochemical Sciences 27(12), 619–627 (2002)
31. 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)
32. Hoppenstaedt, F.C., Peskin, C.S.: Modeling and Simulation in Medicine and the Life Sciences. Springer, New York (2002)
33. Horst, R., Tuy, H.: Global optimization: Deterministic approaches. Springer, Berlin (1990)
34. Iancu, B., Czeizler, E., Czeizler, E., Petre, I.: Quantitative refinement of reaction models. International Journal of Unconventional Computing (page to appear, 2013)
35. Jones, C.M., Henry, E.R., Hu, Y., Chan, C.-K., Luck, S.D., Bhuyan, A., Roder, H., Hofrichter, J., Eaton, W.A.: Fast events in protein folding initiated by nanosecond laser photolysis. Proceedings of the National Academy of Sciences 90(24), 11860–11864 (1993)
36. Kline, M.P., Morimoto, R.I.: Repression of the heat shock factor 1 transcriptional activation domain is modulated by constitutive phosphorylation. Molecular and Cellular Biology 17(4), 2107–2115 (1997)
37. Klipp, E., Herwig, R., Kowald, A., Wierling, C., Lehrach, H.: Systems biology in practice: concepts, implementation and application. Wiley-Vch (2005)
38. Kühnel, M., Mayorga, L.S., Dandekar, T., Thakar, J., Schwarz, R., Anes, E., Griffiths, G., Reich, J.: Modelling phagosomal lipid networks that regulate actin assembly. BMC Systems Biology 2, 107–121 (2008)
39. Lotka, A.J.: Elements of Physical Biology. Williams & Wilkins Company (1925)
40. Malthus, T.R.: An Essay on the Principle of Population. 1798

41. McKay, M.D., Beckman, R.J., Conover, W.J.: Comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21(2), 239–245 (1979)
42. Mendes, P., Kell, D.: Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. Bioinformatics 14(10), 869–883 (1998)
43. Menten, L., Michaelis, M.: Die kinetik der invertinwirkung. Biochem. Z. 49, 333–369 (1913)
44. Miller, W.G., Alberty, R.A.: Kinetics of the reversible michaelis-menten mechanism and the applicability of the steady-state approximation. Journal of the American Chemical Society 80(19), 5146–5151 (1958)
45. Moles, C.G., Mendes, P., Banga, J.R.: Parameter estimation in biochemical pathways: a comparison of global optimization methods. Genome Research 13(11), 2467–2474 (2003)
46. Murphy, E., Danos, V., Féret, J., Krivine, J., Harmer, R.: Rule-based modeling and model refinement. Elements of Computational Systems Biology, 83–114 (2009)
47. Murray, J.D.: Mathematical Biology I: An Introduction. Springer, New York (2002)
48. Murray, J.D.: Mathematical Biology II: Spatial Models and Biomedical Applications. Springer, New York (2002)
49. Nelson, D.L., Cox, M.M.: Lehninger principles of biochemistry. Worth Publishers (2000)
50. Oberguggenberger, M., King, J., Schmelzer, B.: Classical and imprecise probability methods for sensitivity analysis in engineering: A case study. International Journal of Approximate Reasoning 50, 680–693 (2009)
51. Orth, J.D., Thiele, I., Palsson, B.: What is flux balance analysis? Nature Biotechnology 28, 245–248 (2010)
52. Peper, A., Grimbergen, C.A., Spaan, J.A.E., Souren, J.E.M., Van Wijk, R.: A mathematical model of the hsp70 regulation in the cell. International Journal of Hyperthermia 14(1), 97–124 (1998)
53. Petre, I., Mizera, A., Hyder, C.L., Meinander, A., Mikhailov, A., Morimoto, R.I., Sistonen, L., Eriksson, J.E., Back, R.-J.: A simple mass-action model for the eukaryotic heat shock response and its mathematical validation. Natural Computing 10(1), 595–612 (2011)
54. Rabitz, H., Kramer, M., Dacol, D.: Sensitivity analysis in chemical kinetics. Annual Review of Physical Chemistry 34(1), 419–461 (1983)
55. Rieger, T.R., Morimoto, R.I., Hatzimanikatis, V.: Mathematical modeling of the eukaryotic heat-shock response: Dynamics of the hsp70 promoter. Biophysical Journal 88(3), 1646 (2005)
56. Rubinow, S.I.: Introduction to Mathematical Biology. John Wiley, New York (1975)
57. Stewart, W.J.: Probability, Markov chains, queues, and simulation. The mathematical basis of performance modeling. Princeton University Press, Princeton (2009)
58. Taubes, C.: Modeilng Differential Equations in Biology. Prentice Hall, Upper Saddle River (2001)
59. Turányi, T.: Sensitivity analysis of complex kinetic systems. tools and applications. Journal of Mathematical Chemistry 5(3), 203–248 (1990)
60. Volterra, V.: Animal ecology. In: Chapman, R.N. (ed.), pp. 409–448. McGraw-Hill, New York (1926)
61. Wilkinson, D.J.: Stochastic modelling for systems biology. Chapman & Hall/CRC Mathematical Biology and Medicine Series (2006)
62. Zi, Z.: Sensitivity analysis approaches applied to systems biology models. Systems Biology, IET 5(6), 336–346 (2011)

Publication III

Quantitative model refinement in four different frameworks, with applications to the heat shock response

Diana-Elena Gratie, Bogdan Iancu, Sepinoud Azimi and Ion Petre

Accepted, In: *From Action System to Distributed Systems*, Taylor & Francis Group, Editors: Luigia Petre, Emil Sekerinski, To appear 2015.

A preliminary shorter version of this article was published in:

Bogdan Iancu, Diana-Elena Gratie, Sepinoud Azimi and Ion Petre, *On the implementation of quantitative model refinement.* In Proceedings of the *First International Conference on Algorithms for Computational Biology*, Adrian-Horia Dediu, Carlos Martín Vide, Bianca Truthe Eds., LNBI 8542, 95–106, Springer, 2014.

Chapter 1

Quantitative model refinement in four different frameworks, with applications to the heat shock response

Diana-Elena Gratie
Computational Biomodeling Laboratory, Turku Centre for Computer Science and Åbo Akademi University

Bogdan Iancu
Computational Biomodeling Laboratory, Turku Centre for Computer Science and Åbo Akademi University

Sepinoud Azimi
Computational Biomodeling Laboratory, Turku Centre for Computer Science and Åbo Akademi University

Ion Petre
Computational Biomodeling Laboratory, Turku Centre for Computer Science and Åbo Akademi University

1.1 Introduction

Research in molecular biology has been traditionally conducted targeting individual molecular entities of a biological system. Discovery of the functional specificity of these molecular entities has brought about tremendous progress in the understanding of biological systems. However, the understanding of individual molecules alone does not suffice to tackle complex biological phenomena, such as those involved in diseases. Systems biology promotes precisely an integrative bottom-up approach for the analysis of such systems, which takes advantage of previous knowledge about individual proteins, but which is essentially concerned with a holistic analysis of the system: understanding the structure and the dynamics of the system, see [28, 16]. Moreover, in the past decades, we have witnessed a convergence of the fields of computer science and biology towards a new field that is expanding evermore, bioinformatics, see [27].

At the core of systems biology lies the concept of computational modeling, driven by the production of massive sets of experimental data which necessitate computer analysis, see [23]. In this context, formal frameworks prove to be essential in the synthesis and analysis of large biological models as an effort to predict the system-level behavior of such systems. This approach commences with an abstraction of the biological phenomena, which is ultimately converted into a model derived through an iterative process of model building involving system design, model analysis, hypothesis generation, hypothesis testing, experimental verification and model refinement, see [28]. The model obtained as such very often needs to be refined to include more details regarding some of its biological processes, encapsulating presumably new experimental data over a number of new parameters. A reiteration of the whole process of quantitative model fitting and validation is, however, unfeasible for a large model. The alternative we discuss in this paper is that of quantitative model refinement.

Quantitative model refinement focuses on the step-wise construction of models, from small abstract models to large, detailed ones. Each refinement step consists of two parts. The structural part of the refinement consists of fixing the details to add to the model (e.g., new attributes in existing species, new species, new interactions, new modules, etc.) and identifying the new set of species and interactions yielded by adding these details. The quantitative part of the refinement consists in fixing the numerical setup of the refined model (kinetic parameters and initial values) in such a way that the quantitative behavior of the model, in particular its experimental fit, remains unchanged.

The goal of this paper is to give an overview of our approach to quantitative model refinement. We introduce first the main mathematical concepts we use in our framework: reaction-based model, refinement relation, structural refinement, and quantitative refinement. We then discuss the implementation

of model refinement in four of the most widely used approaches in biomodeling: ODE-based models, rule-based models, Petri net models, and guarded command models. The structural part of the refinement has a different solution in each approach, in some cases leading to a compact representation of the refined models. The quantitative part of the refinement aims to avoid the computationally expensive procedure of parameter estimation (especially since the models get larger in each refinement step); instead, we apply in each approach the sufficient condition recently proposed in [8]. As a case study, we consider the heat shock response in eukaryotes. We refer to the model in [26] as the *basic heat shock response model* and to the model in [13] as its *refined model*. A short version of this paper was presented in [15].

The paper is organised as follows: we discuss the concept of fit-preserving refinement in Section 1.2. We succinctly describe in Section 1.3 the heat shock response and its underlying reaction-based model. We focus in Section 1.4 on the refinement of ODE-based models. In Section 1.5, we present rule-based modelling and the rule-based implementation of the heat shock response and of its corresponding refinement. In Section 1.6, we discuss Petri nets and their capabilities with regards to the implementation of the heat shock response and its refinement to include acetylation using colored Petri nets. Section 1.7 comprises a brief description of guarded command languages, focusing on PRISM, and a discussion regarding the implementation of the basic and the refined model for the heat shock response. We conclude the paper with a discussion in Section 1.8. All models developed in this paper can be downloaded at [14].

1.2 Quantitative model refinement

Model refinement has been extensively investigated in the field of software engineering, especially in connection to parallel computing. For example, among other approaches, such studies brought about a logical framework for the construction of computer programs, called *refinement calculus*. This framework tackles the derivation of computer programs correct by construction and refinement of computer programs ensuring correctness preservation, see [2].

Quantitative model refinement is a concept introduced in systems biology as an approach to step-wise construction of biomodels. It focuses on preserving the quantitative behavior of such models, especially their model fit, while avoiding parameter estimation in the context of the combinatorial explosion in the size of the models. Quantitative model refinement was introduced for rule-based models in [25, 5] and for reaction-based models in [24, 13]. We introduce in the following the quantitative refinement of reaction-based models following the approach of [8] based on refinement relations.

A model M comprises *species* $\Sigma = \{A_1, \ldots, A_m\}$ and *reactions* $R = \{r_1, \ldots,$

$r_n\}$, where reaction $r_j \in R$ can be expressed as a rewriting rule of the form:

$$r_j \ : \ s_{1,j}A_1 + \ldots + s_{m,j}A_m \xrightarrow{k_{r_j}} s'_{1,j}A_1 + \ldots + s'_{m,j}A_m, \tag{1.1}$$

where $s_{1,j}, \ldots, s_{m,j}, s'_{1,j}, \ldots, s'_{m,j} \in \mathbb{N}$ are the *stoichiometric coefficients* of r_j and $k_{r_j} \geq 0$ is the *kinetic rate constant* of reaction r_j. We denote by $r_j^{(1)} = [s_{1,j}, \ldots, s_{m,j}]$ the vector of stoichiometric coefficients on the left hand side of reaction r_j and by $r_j^{(2)} = [s'_{1,j}, \ldots, s'_{m,j}]$ the vector of stoichiometric coefficients on its right hand side. We also denote reaction r_j as $r_j^{(1)} \xrightarrow{k_{r_j}} r_j^{(2)}$.

The goal of the refinement is to introduce details into the model, in the form of distinguishing several subspecies of a given species. The distinction between subspecies may represent post-translational modifications such as phosphorylation, acetylation, etc., but it could also account for different possible types of a particular trait (e.g. fur color of animals in a breeding experiment).

We consider that all species are refined at once. Thus, each species in some initial model M will be replaced by a non-empty set of species in its refined model M_R, according to a *species refinement relation* ρ. The refinement of a set of species to a new set of [sub]species is formalized in Definition 1.

Definition 1 ([8]) *Given two sets of species Σ and Σ', and a relation $\rho \subseteq \Sigma \times \Sigma'$, we say that ρ is a* species refinement relation *iff it satisfies the following conditions:*

1. *for each $A \in \Sigma$ there exists $A' \in \Sigma'$ such that $(A, A') \in \rho$;*
2. *for each $A' \in \Sigma'$ there exists* exactly one *$A \in \Sigma$ such that $(A, A') \in \rho$;*

We denote $\rho(A) = \{A' \in \Sigma' \mid (A, A') \in \rho\}$. We say that all species $A' \in \rho(A)$ are siblings.

Intuitively, each species $A \in \Sigma$ is refined to the set of species $\rho(A)$, and replaced in the refined model with its refinements. Each species must be refined to at least a singleton set (and in the singleton case one may say that the refinement is trivial and the species does not change, although it may be denoted by a different symbol in Σ'), and no two species in Σ can be refined to the same species $A' \in \Sigma'$.

We introduce in the following definition the refinement of a vector (of stoichiometric coefficients), of a reaction, and of a reaction-based model.

Definition 2 ([8]) *Let $\Sigma = \{A_1, \ldots, A_m\}$ and $\Sigma' = \{A'_1, \ldots, A'_p\}$ be two sets of species, and $\rho \subseteq \Sigma \times \Sigma'$ a species refinement relation.*

1. *Let $\alpha = (\alpha_1, \ldots, \alpha_m) \in \mathbb{N}^{\Sigma}$ and $\alpha' = (\alpha'_1, \ldots, \alpha'_p) \in \mathbb{N}^{\Sigma'}$. We say that α' is a* ρ-refinement *of α, denoted $\alpha' \in \rho(\alpha)$, if*

$$\sum_{\substack{1 \leq j \leq p \\ A'_j \in \rho(A_i)}} \alpha'_j = \alpha_i, \text{ for all } 1 \leq i \leq m.$$

2. *Let r and r' be two reactions over Σ and Σ', resp.:*

$$r \ : \ s_1A_1 + \ldots + s_mA_m \xrightarrow{k_r} s'_1A_1 + \ldots + s'_mA_m;$$

$$r' : t_1A'_1 + t_2A'_2 + \ldots + t_pA'_p \xrightarrow{k'_r} t'_1A'_1 + t'_2A'_2 + \ldots + t'_pA'_p.$$

We say that r' is a ρ-refinement *of r, denoted $r' \in \rho(r)$, if*

$$r'^{(1)}_j \in \rho(r^{(1)}_j) \text{ and } r'^{(2)}_j \in \rho(r^{(2)}_j).$$

3. *Let $M = (\Sigma, R)$ and $M' = (\Sigma', R')$ be two reaction-based models, and $\rho \subseteq \Sigma \times \Sigma'$ a species refinement relation. We say that M' is a* ρ-structural refinement *of M, denoted $M' \in \rho(M)$, if*

$$R' \subseteq \bigcup_{r \in R} \rho(r) \text{ and } \rho(r) \cap R' \neq \emptyset \ \forall r \in R.$$

In case $R' = \bigcup_{r \in R} \rho(r)$, we say M' is the full structural ρ-refinement *of M.*

Let A be a species with stoichiometry coefficient s in a reaction r, that is refined to a set of species $\rho(A)$. There are $\left(\!\!\binom{|\rho(A)|}{s}\!\!\right)$ ways of choosing s species (not necessarily distinct) from the refined set $\rho(A)$, where $\left(\!\!\binom{n}{k}\!\!\right) = \binom{n+k-1}{k}$ is the *multiset coefficient*, denoting the number of multisets of cardinality k taken from a set of cardinality n. It follows that the number of all possible ρ-refinements of a reaction r of the form (1.1) is

$$\prod_{i=1}^{m} \left(\!\!\binom{|\ \rho(A_i)\ |}{s_i}\!\!\right) \cdot \left(\!\!\binom{|\ \rho(A_i)\ |}{s'_i}\!\!\right). \tag{1.2}$$

We introduce in the following definition the notion of *quantitative refinement* of a model. We denote by $[A](t)$ the concentration of species A at time t. Associating time-dependent concentration functions to the variables of a model can be done either directly from the reaction model by choosing a kinetic law for each reaction, see, eg., [18], or by translating the reaction model to another modeling framework, such as Petri nets, rule-based model, or guarded command language, and using a suitable semantic for that translation.

Definition 3 ([13, 8]) *Given two reaction-based models $M = (\Sigma, R)$ and $M' = (\Sigma', R')$ and a species refinement relation $\rho \subseteq \Sigma \times \Sigma'$ such that M' is a ρ-refinement of M, we say that M' is a* quantitative ρ-refinement of M *if the following condition holds:*

$$[A](t) = \sum_{B \in \rho(A)} [B](t), \text{ for all } A \in \Sigma, t \geq 0. \tag{1.3}$$

A simple sufficient condition for a model M' to be a quantitative refinement of a model M is given in [8].

1.3 Case study: the heat shock response (HSR)

The heat shock response is a cellular defence mechanism against stress (high temperatures, toxins, bacterial infection, etc.) that is highly conserved among eukaryotes. We consider here the heat shock response model proposed in [26], consisting of the set of reactions listed in Table 1.1.

Upon exposure to stress, proteins misfold (reaction (10) in Table 1.1) or aggregate into multi-protein complexes that impair cellular functions up to cell death; the flux of the misfolding reaction depends exponentially on the temperature. To counter these proteotoxic effects of thermal stress the expression of a special family of molecular chaperones, called heat shock proteins (hsp's), increases. The chaperone role of hsp's is to bind to misfolded proteins and assist them in their correct refolding (reactions (11),(12) in Table 1.1) thus preventing multi-protein aggregation and cell death.

TABLE 1.1: The molecular model of the eukaryotic heat shock response proposed in [26].

No.	Reaction	No.	Reaction
(1)	$2\,\mathsf{hsf} \rightleftarrows \mathsf{hsf}_2$	(7)	$\mathsf{hsp} + \mathsf{hsf}_3 \rightarrow \mathsf{hsp{:}hsf} + 2\,\mathsf{hsf}$
(2)	$\mathsf{hsf} + \mathsf{hsf}_2 \rightleftarrows \mathsf{hsf}_3$	(8)	$\mathsf{hsp} + \mathsf{hsf}_3{:}\mathsf{hse} \rightarrow \mathsf{hsp{:}hsf} + 2\,\mathsf{hsf} + \mathsf{hse}$
(3)	$\mathsf{hsf}_3 + \mathsf{hse} \rightleftarrows \mathsf{hsf}_3{:}\mathsf{hse}$	(9)	$\mathsf{hsp} \rightarrow \emptyset$
(4)	$\mathsf{hsf}_3{:}\mathsf{hse} \rightarrow \mathsf{hsf}_3{:}\mathsf{hse} + \mathsf{hsp}$	(10)	$\mathsf{prot} \rightarrow \mathsf{mfp}$
(5)	$\mathsf{hsp} + \mathsf{hsf} \rightleftarrows \mathsf{hsp{:}hsf}$	(11)	$\mathsf{hsp} + \mathsf{mfp} \rightleftarrows \mathsf{hsp{:}mfp}$
(6)	$\mathsf{hsp} + \mathsf{hsf}_2 \rightarrow \mathsf{hsp{:}hsf} + \mathsf{hsf}$	(12)	$\mathsf{hsp{:}mfp} \rightarrow \mathsf{hsp} + \mathsf{prot}$

The expression of hsp's is regulated by a family of proteins called heat shock transcription factors (hsf's). In a trimeric state (hsf_3) they bind to heat shock elements (hse's - the hsp-encoding gene promoter regions), forming $\mathsf{hsf}_3{:}\mathsf{hse}$ complexes, and activate the transcription of hsp's, process modeled through reactions (1)-(4) in Table 1.1. The concentration levels of $\mathsf{hsf}_3{:}\mathsf{hse}$ measure the DNA binding activity. Reation (4) implies that with a higher level of DNA binding we get a faster transcription/synthesis of hsp. The hsp's downregulate their expression levels by binding to $\mathsf{hsf}_3{:}\mathsf{hse}$'s, hsf_3's, hsf_2's and hsf's and breaking down the complexes, thus stopping the expression activity (reactions (5)-(8)). The degradation of hsp molecules is modeled in reaction (9).

The hsf protein can undergo post-translational modifications (phosphorylation, acetylation, sumoylation), some of which influence hsf binding activity, see [1]. In particular, the acetylation of hsf's plays a role in the attenuation of the heat shock response. We consider in this paper the refinement of hsf molecules as described in [13]. The refinement considers the acetylation status (ON/OFF) of hsf proteins. The order of acetylated sites is not important in a compound with two or more hsf molecules, only their count. Thus,

- hsf is refined to $\{\mathsf{rhsf}^{(0)}, \mathsf{rhsf}^{(1)}\}$,
- a dimer molecule $\mathsf{hsf_2}$ is refined to $\{\mathsf{rhsf_2}^{(0)}, \mathsf{rhsf_2}^{(1)}, \mathsf{rhsf_2}^{(2)}\}$,
- and a trimer molecule $\mathsf{hsf_3}$ is refined to $\{\mathsf{rhsf_3}^{(0)}, \mathsf{rhsf_3}^{(1)}, \mathsf{rhsf_3}^{(2)}, \mathsf{rhsf_3}^{(3)}\}$,

where the superscript denotes the number of acetylated sites. This leads to an expansion of the model from 10 species and 17 irreversible reactions to 20 species and 55 irreversible reactions.

This refinement can be described via the following species refinement relation:

$$\begin{aligned}\rho = \{&(\mathsf{hse}, \mathsf{rhse}), (\mathsf{hsp}, \mathsf{rhsp}), (\mathsf{prot}, \mathsf{rprot}), (\mathsf{mfp}, \mathsf{rmfp}), (\mathsf{hsp{:}mfp}, \mathsf{rhsp{:}rmfp}),\\ &(\mathsf{hsf}, \mathsf{rhsf}^{(0)}), (\mathsf{hsf}, \mathsf{rhsf}^{(1)}),\\ &(\mathsf{hsf_2}, \mathsf{rhsf_2}^{(0)}), (\mathsf{hsf_2}, \mathsf{rhsf_2}^{(1)}), (\mathsf{hsf_2}, \mathsf{rhsf_2}^{(2)}),\\ &(\mathsf{hsf_3}, \mathsf{rhsf_3}^{(0)}), (\mathsf{hsf_3}, \mathsf{rhsf_3}^{(1)}), (\mathsf{hsf_3}, \mathsf{rhsf_3}^{(2)}), (\mathsf{hsf_3}, \mathsf{rhsf_3}^{(3)}),\\ &(\mathsf{hsp{:}hsf}, \mathsf{rhsp{:}rhsf}^{(0)}), (\mathsf{hsp{:}hsf}, \mathsf{rhsp{:}rhsf}^{(1)}),\\ &(\mathsf{hsf_3{:}hse}, \mathsf{rhsf_3^{(0)}{:}rhse}), (\mathsf{hsf_3{:}hse}, \mathsf{rhsf_3}^{(1)}\mathsf{{:}rhse}), (\mathsf{hsf_3{:}hse}, \mathsf{rhsf_3}^{(2)}\mathsf{{:}rhse}),\\ &(\mathsf{hsf_3{:}hse}, \mathsf{rhsf_3}^{(3)}\mathsf{{:}rhse})\}.\end{aligned}$$

The full set of the refined reactions is given in Table 1.2.

TABLE 1.2: The list of reactions for the refined model that includes the acetylation status of hsf. A reaction (i.j) is a refinement of reaction (i) of the basic model, see Table 1.1

No.	**Reaction**
(1.1)	$2\,\mathsf{rhsf}^{(0)} \rightleftarrows \mathsf{rhsf_2}^{(0)}$
(1.2)	$\mathsf{rhsf}^{(0)} + \mathsf{rhsf}^{(1)} \rightleftarrows \mathsf{rhsf_2}^{(1)}$
(1.3)	$2\,\mathsf{rhsf}^{(1)} \rightleftarrows \mathsf{rhsf_2}^{(2)}$
(2.1)	$\mathsf{rhsf}^{(0)} + \mathsf{rhsf_2}^{(0)} \rightleftarrows \mathsf{rhsf_3}^{(0)}$
(2.2)	$\mathsf{rhsf}^{(1)} + \mathsf{rhsf_2}^{(0)} \rightleftarrows \mathsf{rhsf_3}^{(1)}$
(2.3)	$\mathsf{rhsf}^{(0)} + \mathsf{rhsf_2}^{(1)} \rightleftarrows \mathsf{rhsf_3}^{(1)}$
(2.4)	$\mathsf{rhsf}^{(1)} + \mathsf{rhsf_2}^{(1)} \rightleftarrows \mathsf{rhsf_3}^{(2)}$
(2.5)	$\mathsf{rhsf}^{(0)} + \mathsf{rhsf_2}^{(2)} \rightleftarrows \mathsf{rhsf_3}^{(2)}$
(2.6)	$\mathsf{rhsf}^{(1)} + \mathsf{rhsf_2}^{(2)} \rightleftarrows \mathsf{rhsf_3}^{(3)}$
(3.1)	$\mathsf{rhsf_3}^{(0)} + \mathsf{rhse} \rightleftarrows \mathsf{rhsf_3^{(0)}{:}rhse}$
(3.2)	$\mathsf{rhsf_3}^{(1)} + \mathsf{rhse} \rightleftarrows \mathsf{rhsf_3}^{(1)}\mathsf{{:}rhse}$
(3.3)	$\mathsf{rhsf_3}^{(2)} + \mathsf{rhse} \rightleftarrows \mathsf{rhsf_3}^{(2)}\mathsf{{:}rhse}$
(3.4)	$\mathsf{rhsf_3}^{(3)} + \mathsf{rhse} \rightleftarrows \mathsf{rhsf_3}^{(3)}\mathsf{{:}rhse}$
(4.1)	$\mathsf{rhsf_3^{(0)}{:}rhse} \to \mathsf{rhsf_3^{(0)}{:}rhse} + \mathsf{rhsp}$
(4.2)	$\mathsf{rhsf_3}^{(1)}\mathsf{{:}rhse} \to \mathsf{rhsf_3}^{(1)}\mathsf{{:}rhse} + \mathsf{rhsp}$

TABLE 1.2: The list of reactions for the refined model - Continued

(4.3)	$\mathsf{rhsf_3}^{(2)}\!:\!\mathsf{rhse} \to \mathsf{rhsf_3}^{(2)}\!:\!\mathsf{rhse} + \mathsf{rhsp}$
(4.4)	$\mathsf{rhsf_3}^{(3)}\!:\!\mathsf{rhse} \to \mathsf{rhsf_3}^{(3)}\!:\!\mathsf{rhse} + \mathsf{rhsp}$
(5.1)	$\mathsf{rhsp} + \mathsf{rhsf}^{(0)} \rightleftarrows \mathsf{rhsp}\!:\mathsf{rhsf}^{(0)}$
(5.2)	$\mathsf{rhsp} + \mathsf{rhsf}^{(1)} \rightleftarrows \mathsf{rhsp}\!:\mathsf{rhsf}^{(1)}$
(6.1)	$\mathsf{rhsp} + \mathsf{rhsf_2}^{(0)} \to \mathsf{rhsp}\!:\mathsf{rhsf}^{(0)} + \mathsf{rhsf}^{(0)}$
(6.2)	$\mathsf{rhsp} + \mathsf{rhsf_2}^{(1)} \to \mathsf{rhsp}\!:\mathsf{rhsf}^{(0)} + \mathsf{rhsf}^{(1)}$
(6.3)	$\mathsf{rhsp} + \mathsf{rhsf_2}^{(1)} \to \mathsf{rhsp}\!:\mathsf{rhsf}^{(1)} + \mathsf{rhsf}^{(0)}$
(6.4)	$\mathsf{rhsp} + \mathsf{rhsf_2}^{(2)} \to \mathsf{rhsp}\!:\mathsf{rhsf}^{(1)} + \mathsf{rhsf}^{(1)}$
(7.1)	$\mathsf{rhsp} + \mathsf{rhsf_3}^{(0)} \to \mathsf{rhsp}\!:\mathsf{rhsf}^{(0)} + 2 * \mathsf{rhsf}^{(0)}$
(7.2)	$\mathsf{rhsp} + \mathsf{rhsf_3}^{(1)} \to \mathsf{rhsp}\!:\mathsf{rhsf}^{(0)} + \mathsf{rhsf}^{(1)} + \mathsf{rhsf}^{(0)}$
(7.3)	$\mathsf{rhsp} + \mathsf{rhsf_3}^{(1)} \to \mathsf{rhsp}\!:\mathsf{rhsf}^{(1)} + 2 * \mathsf{rhsf}^{(0)}$
(7.4)	$\mathsf{rhsp} + \mathsf{rhsf_3}^{(2)} \to \mathsf{rhsp}\!:\mathsf{rhsf}^{(0)} + 2\,\mathsf{rhsf}^{(1)}$
(7.5)	$\mathsf{rhsp} + \mathsf{rhsf_3}^{(2)} \to \mathsf{rhsp}\!:\mathsf{rhsf}^{(1)} + \mathsf{rhsf}^{(1)} + \mathsf{rhsf}^{(0)}$
(7.6)	$\mathsf{rhsp} + \mathsf{rhsf_3}^{(3)} \to \mathsf{rhsp}\!:\mathsf{rhsf}^{(1)} + 2\,\mathsf{rhsf}^{(1)}$
(8.1)	$\mathsf{rhsp} + \mathsf{rhsf_3}^{(0)}\!:\mathsf{rhse} \to \mathsf{rhsp}\!:\mathsf{rhsf}^{(0)} + 2\,\mathsf{rhsf}^{(0)} + \mathsf{rhse}$
(8.2)	$\mathsf{rhsp} + \mathsf{rhsf_3}^{(1)}\!:\mathsf{rhse} \to \mathsf{rhsp}\!:\mathsf{rhsf}^{(1)} + 2\,\mathsf{rhsf}^{(0)} + \mathsf{rhse}$
(8.3)	$\mathsf{rhsp} + \mathsf{rhsf_3}^{(1)}\!:\mathsf{rhse} \to \mathsf{rhsp}\!:\mathsf{rhsf}^{(0)} + \mathsf{rhsf}^{(1)} + \mathsf{rhsf}^{(0)} + \mathsf{rhse}$
(8.4)	$\mathsf{rhsp} + \mathsf{rhsf_3}^{(2)}\!:\mathsf{rhse} \to \mathsf{rhsp}\!:\mathsf{rhsf}^{(1)} + \mathsf{rhsf}^{(1)} + \mathsf{rhsf}^{(0)} + \mathsf{rhse}$
(8.5)	$\mathsf{rhsp} + \mathsf{rhsf_3}^{(2)}\!:\mathsf{rhse} \to \mathsf{rhsp}\!:\mathsf{rhsf}^{(0)} + 2\,\mathsf{rhsf}^{(1)} + \mathsf{rhse}$
(8.6)	$\mathsf{rhsp} + \mathsf{rhsf_3}^{(3)}\!:\mathsf{rhse} \to \mathsf{rhsp}\!:\mathsf{rhsf}^{(1)} + 2\,\mathsf{rhsf}^{(1)} + \mathsf{rhse}$
(9.1)	$\mathsf{rhsp} \to \emptyset$
(10.1)	$\mathsf{rprot} \to \mathsf{rmfp}$
(11.1)	$\mathsf{rhsp} + \mathsf{rmfp} \rightleftarrows \mathsf{rhsp}\!:\mathsf{rmfp}$
(12.1)	$\mathsf{rhsp}\!:\mathsf{rmfp} \to \mathsf{rhsp} + \mathsf{rprot}$

An ODE-based model for the basic HSR model was introduced and analyzed in [26]. A similar model for the refined HSR model was introduced in [13].

1.4 Quantitative refinement for ODE models

The main problem to solve for quantitative model refinement of ODE-based models is to identify the kinetic rate constants of the refined model that lead to a solution of the refinement condition (1.3). An attempt to obtain all solutions of (1.3) would require solving the system of ODEs corresponding to the mass-action model for the basic and the refined models; in general, this cannot be done analytically because the ODEs can be non-linear.

An alternative was proposed in the form of a sufficient condition in [8]. We recall that result in the following. For two vectors of nonnegative integers $\alpha = (\alpha_1, \ldots, \alpha_m)$, $\alpha' = (\alpha'_1, \ldots, \alpha'_{m'})$, we denote

$$\binom{\alpha}{\alpha'} = \frac{\prod_{i=1}^{m} \alpha_i!}{\prod_{j=1}^{m'} \alpha'_j!}.$$

Theorem 1 ([8]) *Let Σ and Σ' be two sets of species and $\rho \subseteq \Sigma \times \Sigma'$ a species refinement relation. Let $M = (\Sigma, R)$ be a reaction-based model and $M' = (\Sigma', R')$ be the full structural ρ-refinement of M; we use letters k (indexed with the reaction name) to indicate the kinetic rate constants of M and letters k' those of M'. If for every $\alpha \to \beta \in R$ and for any $\alpha' \in \rho(\alpha)$ we have that*

$$\sum_{\beta' \in \rho(\beta)} k'_{\alpha' \to \beta'} = \binom{\alpha}{\alpha'} k_{\alpha \to \beta}, \tag{1.4}$$

then M' is a fit-preserving data refinement of M.

The solution thus obtained is evidently not unique. For example, in the heat shock response refined model the kinetic rate constants of all reactions involving at least one form of acetylated hsf could be set to zero; such a choice would cancel the refinement since the influence of all acetylated variables would be ignored in the model. Theorem 1 allows one to choose a multitude of different solutions. One can, for example, include in the solution some parameter values that were obtained from experiments or literature, while using condition (1.4) to choose suitable values for the remaining parameters. In fact, condition (1.4) can also be used to check if a set of given parameter values (for all parameters) leads to a fit-preserving refined model.

The main disadvantage of ODE-based models is that each species gets its own variable and then its own ODE. The framework does not allow for the implicit specification of some of its variables, even when the semantic difference between them is minor (as it could be between sibling subspecies). This leads to an explosion in the size of the refined ODE-based model with respect to the size of the basic ODE-based model. We discuss in the following sections the model refinement approach in three other widely used modelling frameworks. In each case, we focus on whether a more compact specification of the refined model is possible.

1.5 Quantitative refinement for rule-based models

Rule-based modelling is an approach for tackling the combinatorial explosion induced by expanding reaction-based models. The key feature is that rules

only specify those aspects of the input species that are critical for that interaction, while omitting all their other attributes. The rules can be translated either into a set of ODEs, following a continuous, deterministic interpretation, or into a stochastic process, following a stochastic, discrete interpretation of the biological phenomena. Two description languages for the implementation of such models are *BioNetGen*, see [3], and *Kappa*, see [4]. We use BioNetGen in the following. We refer to [6, 7] for details on how models are represented in BioNetGen. A BioNetGen input file, for instance, is essentially a description of the molecular species and their components, reaction rules, kinetic rate constants, initial concentrations and simulation commands. The reaction network generated by BioNetGen can be used to emulate system's dynamics deterministically or stochastically, see [29]. RuleBender is an open source editor for rule-based models which allows for the construction of large models. The simulation, based on a BioNetGen simulator, see [30], generates the reaction network, in *SBML* and *NET* format, corresponding to the given rule-based model. Simulations can run either deterministically, using ODEs, or stochastically, using SSA algorithms, see [29, 30].

A BioNetGen implementation of the HSR model

We discussed in [15] the implementation of the basic heat shock response model of [26] with BioNetGen and RuleBender. The model can be found in [14]. All reactions in our implementation follow the principle of mass action. The BioNetGen model consists of 12 rules, which produce 17 irreversible reactions; kinetic rate constants and initial values are set according to [26]. For example, the RuleBender implementation of the dimerization of hsf is illustrated in Figure 1.1. A deterministic simulation for the BioNetGen model revealed identical simulation results for DNA binding for a temperature of 42°C as the ODE-based model in [26].

To implement in BioNetGen the refinement described in Section 1.3 required only one change: the addition of a site to hsf, having two possible states: acetylated and non-acetylated. The initial concentrations were set conforming to [13]. For more details regarding the implementation, we refer the reader to [15].

1.6 Quantitative refinement for Petri net models

In this section we model the heat shock response and its refinement using the framework of Petri nets. We implemented our models using Snoopy, a visualization, modeling and simulation tool with support for many types of Petri nets, see [11].

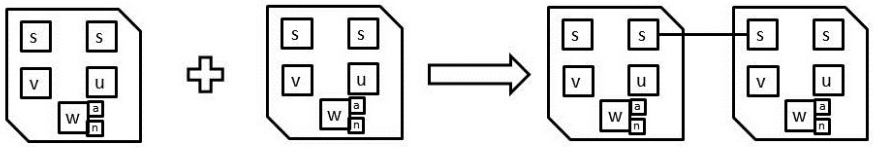


FIGURE 1.1: A graphical representation of the species hsf (containing sites 's','u','v','w') and of the rule showing the dimerization of hsf, illustrated by binding one of the 's' sites of the hsf species with one of the 's' site of the other hsf species. Note the two possible states of the site 'w', namely 'a' and 'n', which depict two possible states of the species, acetylated or non-acetylated respectively.

The basic HSR model as a Petri net

For the general method of building a (standard) Petri net model from a given set of biochemical reactions we refer to [19]. We built our implementation of the heat shock response following the standard procedure: each species is represented as a place, and each irreversible reaction is represented as a transition having as pre-places the places corresponding to species on the left hand side of the reaction, and as post-places the places representing species on the right hand side of the reaction; arc multiplicities denote the stoichiometric coefficients of the species involved in the reaction.

We checked several properties of the model to ensure that our implementation is correct. For example, the P-invariants of the Petri net encode the three mass conservation relations of the biological model, as described in [26]. The net is covered by T-invariants, and all places except for the place representing species hsp are covered by P-invariants, which means they are bounded. Our PRISM implementation uses as bounds for the species (except hsp) the constants from the three mass conservation relations, namely the total amount of hsf, hse and prot. We also simulated the model with the numerical setup of [26] and obtained the same DNA binding curve for a heat shock of 42°C as that shown in [26] for the corresponding ODE-based model. We refer to [15] for details.

The refined HSR model as a colored Petri net

Implementing a model as a (standard) Petri net means that each reaction is represented as a transition. In doing so, the refined model is inevitably larger than the initial model, similarly as in the case of ODE-based models. However, the framework of *colored Petri nets* allows for stacking more than one species in a colored place, and for identifying each species via a color, see Figure 1.2 for an example.

We implemented the refinement of the heat shock response model as a colored continuous Petri net, in order to maintain a compact representation. Multiple coloring strategies are possible; we considered two. One of them aimed at using as few colors as possible. In this approach, trimers are represented with four colors $0, 1, 2, 3$, denoting the number of acetylated sites.

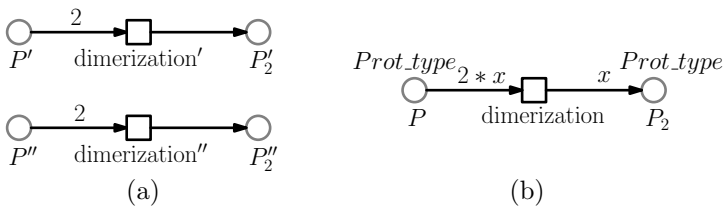


FIGURE 1.2: Representing the dimerization of two different proteins, P' and P'' with (a) a transition for each of them, and (b) a single colored transition for both. In (b) we use a color set with two colors, $Prot\_type = \{1, 2\}$. The choice between colors 1 and 2 is done by the variable x; when $x = 1$ the reaction will consume two proteins with color 1 and produce one dimer with color 1, and when $x = 2$ the reaction will consume two proteins with color 2 and produce one dimer with color 2. In the figure, all places and transitions have identifiers, and in (b) we also list the color set for each place (italic text).

A possible refinement of reaction (7) with a single-acetylated trimer is

$$\mathsf{rhsp} + \mathsf{rhsf_3}^{(1)} \rightarrow \mathsf{rhsp{:}\,rhsf}^{(0)} + \mathsf{rhsf}^{(0)} + \mathsf{rhsf}^{(1)},$$

and another one is

$$\mathsf{rhsp} + \mathsf{rhsf_3}^{(1)} \rightarrow \mathsf{rhsp{:}\,rhsf}^{(1)} + 2\,\mathsf{rhsf}^{(0)}\,.$$

In order to differentiate between the two refined reactions, we had to use two transitions, thus increasing the number of transitions compared to that of the basic model. The second strategy was to preserve the structure of the network in terms of number of places, transitions and the connections between them. We were able to do so by considering the color sets of places denoting some of the species (e.g. $\mathsf{hsf_2}$, $\mathsf{hsf_3}, \mathsf{hsf_3{:}\,hse}, \mathsf{hsp{:}\,hsf}$) as Cartesian products of the color sets of the places corresponding to the species they consist of. This corresponds to a refinement where the order of the acetylated monomers in a trimer is explicitly described. An example of the reversible dimerization reaction using the second coloring approach is presented in Figure 1.3. We refer to [9] for more details about the two modeling strategies based on colored Petri nets.

1.7 Quantitative refinement for PRISM models

PRISM is a free and open source guarded command language and probabilistic model checker. It can be used to model and analyze a wide range of

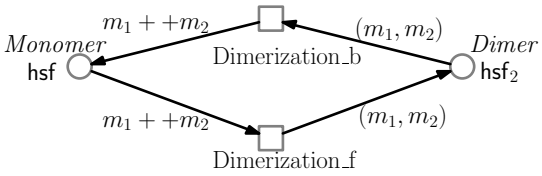


FIGURE 1.3: Modeling the hsf dimers using a compound color set *Dimer* = *Monomer*× *Monomer*. The regular text next to places and transitions denotes their respective identifier, while the color sets are written in italic font. The hsf monomers are represented using the color set *Monomer* = $\{0, 1\}$. The preplaces of the forward reaction are two monomers, with colors m_1 and m_2. The result will be the production of one dimer with color (m_1,m_2). In the reverse reaction, one dimer with color (m_1, m_2) is split into the two monomers m_1 and m_2.

probabilistic systems. PRISM supports various types of probabilistic models: probabilistic automata (PAs), probabilistic timed automata (PTAs), discrete-time Markov chains (DTMCs), continuous-time Markov chains (CTMCs), Markov decision processes (MDPs). A PRISM model consists of a keyword which describes the model type (e.g., CTMC) and a set of modules whose states are defined by the state of their finite range variables (e.g., hsf). The state of the variables in each module is specified by some commands including a guard and one or more updates, see [21].

The HSR models as PRISM implementations

We implemented the basic heat shock response as a CTMC model within a single module. The PRISM model consists of 10 variables, each of them corresponding to one of the reactants in the model, and 17 guards representing the 17 irreversible reactions of the system. For example, the guard corresponding to *dimerization*, reaction (1) in Table 1.1, is expressed in Table 1.4(a).

For the refined heat shock response model, the corresponding PRISM model was built in a similar way, following its reactions in Table 1.2. For example, the guards corresponding to *dimerization* are presented in Table 1.4(b). The complete models can be found at [14] and more details on how we built them can be found in [15].

TABLE 1.4: PRISM code for the dimerization in (a) the basic and (b) the refined HSR models (N_{hsf} is the upper bound for hsf and N_{hsf_2} is the upper bound for hsf_2).

(a) Dimerization in the basic HSR model

$[]2 \leq \mathsf{hsf} \leq N_{\mathsf{hsf}} \wedge 0 \leq \mathsf{hsf}_2 \leq N_{\mathsf{hsf}_2} - 1 \rightarrow \mathsf{hsf} * \mathsf{hsf} * 0.5 * k_1 :$
$(\mathsf{hsf}' = \mathsf{hsf} - 2) \wedge (\mathsf{hsf_2}' = \mathsf{hsf}_2 + 1);$

TABLE 1.4: PRISM code for the dimerization in (a) the basic and (b) the refined HSR models (N_{hsf} is the upper bound for hsf and N_{hsf_2} is the upper bound for hsf_2).

(b) Dimerization in the refined HSR model

$[]2 \leq \mathsf{rhsf}^{(0)} \leq N_{\mathsf{hsf}} \wedge 0 \leq {\mathsf{rhsf}_2}^{(0)} \leq N_{\mathsf{hsf}_2} - 1 \rightarrow \mathsf{rhsf}^{(0)} * \mathsf{rhsf}^{(0)} * 0.5 * k_1 :$
$({\mathsf{rhsf}^{(0)}}' = \mathsf{rhsf}^{(0)} - 2) \wedge ({\mathsf{rhsf}_2^{(0)}}' = {\mathsf{rhsf}_2}^{(0)} + 1);$
$[]1 \leq \mathsf{rhsf}^{(0)} \leq N_{\mathsf{hsf}} \wedge 1 \leq \mathsf{rhsf}^{(1)} \leq N_{\mathsf{hsf}} \wedge 0 \leq {\mathsf{rhsf}_2}^{(1)} \leq N_{\mathsf{hsf}_2} - 1 \rightarrow$
$\mathsf{rhsf}^{(0)} * \mathsf{rhsf}^{(1)} * k_1 : ({\mathsf{rhsf}^{(0)}}' = \mathsf{rhsf}^{(0)} - 1) \wedge ({\mathsf{rhsf}^{(1)}}' = \mathsf{rhsf}^{(1)} - 1)$
$\wedge({\mathsf{rhsf}_2^{(1)}}' = \mathsf{rhsf}^{(1)} + 1);$
$[]2 \leq \mathsf{rhsf}^{(1)} \leq N_{\mathsf{hsf}} \wedge 0 \leq {\mathsf{rhsf}_2}^{(2)} \leq N_{\mathsf{hsf}_2} - 1 \rightarrow \mathsf{rhsf}^{(1)} * \mathsf{rhsf}^{(1)} * 0.5 * k_1 :$
$({\mathsf{rhsf}^{(1)}}' = \mathsf{rhsf}^{(1)} - 2) \wedge ({\mathsf{rhsf}_2^{(2)}}' = \mathsf{rhsf}_2^{(2)} + 1);$

Stochastic model checking of the PRISM HSR models

The maximum number of states that PRISM can handle for CTMCs does not exceed 10^{10}, see [20], which leads to difficulties in handling the *state space explosion* problem, see [10]. To avoid this problem, we used *approximate verification*, see [22, 12, 10], to verify our two PRISM models.

We are interested in verifying two properties discussed in [26]: (i) the existence of three mass-conservation relations and (ii) the level of DNA binding eventually returns to the basal values, both at $37°C$ and at $42°C$.

The following three properties are used to check whether the mass-conservation relations, corresponding to the level of hsf, hse and prot, are valid in all states along the path:

- $p =?$ [G $\mathsf{hsf} + 2\,\mathsf{hsf}_2 + 3\,\mathsf{hsf}_3 + 3\,\mathsf{hsf}_3{:}\mathsf{hse} + \mathsf{hsp}{:}\mathsf{hsf} = \mathsf{hsf}_{\text{const}}$],
- $p =?$ [G $\mathsf{hse} + \mathsf{hsf}_3{:}\mathsf{hse} = \mathsf{hse}_{\text{const}}$],
- $p =?$ [G $\mathsf{prot} + \mathsf{mfp} + \mathsf{hsp}{:}\mathsf{mfp} = \mathsf{prot}_{\text{const}}$].

As expected, the value of p was confirmed to be 1 in all cases, with confidence level 95%, i.e. the three mass conservation relations are respected in the model.

We verified in PRISM that for time points larger than 14400, the value of $\mathsf{hsf}_3{:}\mathsf{hse}$ complex returns to the initial value by formulating the property, $p =? \quad [F >= 14400 \quad \mathsf{hsf}_3{:}\mathsf{hse} = 3]$. We chose 14400 as a time point reference to correspond to the upper limit of the simulation time for the model in [26]. The probability value calculated by PRISM was 1 for this property as well, with confidence level 95%.

Finally, we also checked if the model confirms the experimental data of [17] on DNA binding. Due to the memory issues of PRISM, it was not possible to run the simulation many times and use the average run plot to verify the experimental data. Therefore, as an alternative approach, we checked the

probability of having a data point within the interval $[0.9 \cdot d, 1.1 \cdot d]$ in the time period $[0.9 \cdot t, 1.1 \cdot t]$, where d is the experimental data point at time t. The results in both cases show high value probabilities, which confirms that our two PRISM models are in accordance with the experimental data of [17]; we refer to [9] for the numerical results.

1.8 Discussion

We discussed in this paper quantitative model refinement, an approach to step-wise construction of biological models. Preserving the quantitative behavior of models throughout the model construction process is at the forefront of this approach. This allows the modeler to avoid repeating the computationally expensive process of parameter estimation at every step of the process, thus avoiding the need of collecting larger and larger sets of high-quality time data. Quantitative refinement also allows the modeler to deal with partial and incomplete information about some of the parameters of the model-to-build, including such information when available, checking its consistency with the other parameters and with the data, and compensating for lack of information about parameters with an algorithmic solution.

We investigated in this paper the versatility of the fit-preserving refinement method with respect to four broadly used frameworks: reaction models (with ODEs), rule-based models (with BioNetGen), Petri net models (with Snoopy), and guarded command-based models (with PRISM). Dealing with the combinatorial explosion to account for post-translational modifications was considerably different from one framework to another. We conclude that the method is cross-platform: it is implementable in all chosen frameworks, despite their distinct underlying modeling paradigms.

In our case study based on the heat shock response, the data structure provided by BioNetGen proves to be suitable for modeling the refinement of biological systems. One could effectively employ species, sites, links, etc., to produce a compact representation of the refined model. On the other hand, using the colored version of Petri nets provides the modeler with appropriate tools to introduce data types into the places of the network. The modeling choices in the definition of the new data types and associating biological meanings to them directly affect the compactness of the representation as well as the corresponding model's complexity. In contrast, PRISM supports only elementary data types for the variables in the model, which leads to an explicit detailing of all elements of the refined model, similarly as in the case of ODE-based models.

We show that our approach toward quantitative model refinement is a potentially suitable one to build a large biomodel which can be implemented within a wide spectrum of modeling frameworks. In this method the modeler

is able to easily modify the level of details in the model in an algorithmic fashion, while ensuring that the model fit is preserved from one level of detail to another. It should also be noted that one can switch from one modeling framework to another in order to use the advantages that each model formulation offers in terms of fast simulations, model checking or compact model representation.

Acknowledgments

The authors thank Monika Heiner for help on Snoopy and Charlie, James Faeder and Leonard Harris for advice on the BioNetGen implementation of the heat shock response, and Adam Smith for technical support regarding RuleBender. We gratefully acknowledge support from Academy of Finland through project 267915.

Bibliography

[1] Malin Åkerfelt, Richard I. Morimoto, and Lea Sistonen. Heat shock factors: Integrators of cell stress, development and lifespan. *Nature Reviews Molecular Cell Biology*, 11(8):545–555, 2010.

[2] Ralph-Johan Back and Joakim Wright. *Refinement calculus: a systematic introduction*. Springer Heidelberg, 1998.

[3] Michael L. Blinov, James R. Faeder, Byron Goldstein, and William S. Hlavacek. BioNetGen: Software for rule-based modeling of signal transduction based on the interactions of molecular domains. *Bioinformatics*, 20(17):3289–3291, 2004.

[4] Vincent Danos, Jérôme Feret, Walter Fontana, Russell Harmer, and Jean Krivine. Rule-based modelling, symmetries, refinements. In *Formal Methods in Systems Biology*, pages 103–122. Springer, 2008.

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

[6] James R. Faeder, Michael L. Blinov, Byron Goldstein, and William S. Hlavacek. Rule-based modeling of biochemical networks. *Complexity*, 10(4):22–41, 2005.

[7] James R. Faeder, Michael L. Blinov, and William S. Hlavacek. Graphical rule-based representation of signal-transduction networks. In *Proceedings of the 2005 ACM Symposium on Applied Computing*, pages 133–140. ACM, 2005.

[8] Cristian Gratie and Ion Petre. Fit-preserving data refinement of mass-action reaction networks. In Arnold Beckmann, Erzsébet Csuhaj-Varjú, and Klaus Meer, editors, *Language, Life, Limits*, volume 8493 of *Lecture Notes in Computer Science*, pages 204–213. Springer, 2014.

[9] Diana-Elena Gratie, Bogdan Iancu, Sepinoud Azimi, and Ion Petre. Quantitative model refinement in four different frameworks, with applications to the heat shock response. Technical Report 1067, TUCS, 2013.

[10] John Heath, Marta Kwiatkowska, Gethin Norman, David Parker, and Oksana Tymchyshyn. Probabilistic model checking of complex biological

pathways. In *Computational Methods in Systems Biology*, pages 32–47. Springer, 2006.

[11] Monika Heiner, Mostafa Herajy, Fei Liu, Christian Rohr, and Martin Schwarick. Snoopy – a unifying Petri net tool. In Serge Haddad and Lucia Pomello, editors, *Application and Theory of Petri Nets*, volume 7347 of *Lecture Notes in Computer Science*, pages 398–407. Springer Berlin Heidelberg, 2012.

[12] Andrew Hinton, Marta Kwiatkowska, Gethin Norman, and David Parker. PRISM: A tool for automatic verification of probabilistic systems. *Tools and Algorithms for the Construction and Analysis of Systems*, pages 441–444, 2006.

[13] 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.

[14] Bogdan Iancu, Diana-Elena Gratie, Sepinoud Azimi, and Ion Petre. Computational modeling of the eukaryotic heat shock response: the BioNetGen implementation, the Petri net implementation and the PRISM implementation, 2013. Available at: http://combio.abo.fi/research/computational-modeling-of-the-eukaryotic-heat-shock-response/.

[15] Bogdan Iancu, Diana-Elena Gratie, Sepinoud Azimi, and Ion Petre. On the implementation of quantitative model refinement. In Adrian-Horia Dediu, Carlos Martín-Vide, and Bianca Truthe, editors, *Algorithms for Computational Biology*, volume 8542 of *Lecture Notes in Computer Science*, pages 95–106. Springer International Publishing, 2014.

[16] Hiroaki Kitano. Systems biology: a brief overview. *Science*, 295(5560):1662–1664, 2002.

[17] Michael P. Kline and Richard I. Morimoto. Repression of the heat shock factor 1 transcriptional activation domain is modulated by constitutive phosphorylation. *Molecular and Cellular Biology*, 17(4):2107–2115, 1997.

[18] Edda Klipp, Ralf Herwig, Axel Kowald, Christoph Wierling, and Hans Lehrach. *Systems Biology in Practice: Concepts, Implementation and Application*. Wiley-Vch, 2005.

[19] Ina Koch, Wolfgang Reisig, and Falk Schreiber. *Modeling in Systems Biology: the Petri Net Approach*. Springer, 2010.

[20] Marta Kwiatkowska, Gethin Norman, and David Parker. Quantitative analysis with the probabilistic model checker PRISM. *Electronic Notes in Theoretical Computer Science*, 153(2):5–31, 2006.

[21] Marta Kwiatkowska, Gethin Norman, and David Parker. PRISM 4.0: Verification of probabilistic real-time systems. In Ganesh Gopalakrishnan and Shaz Qadeer, editors, *Proceedings of the 23rd International Conference on Computer Aided Verification (CAV'11)*, volume 6806 of *Lecture Notes in Computer Science*, pages 585–591. Springer, 2011.

[22] Richard Lassaigne and Sylvain Peyronnet. Approximate verification of probabilistic systems. *Process Algebra and Probabilistic Methods: Performance Modeling and Verification*, 2399:277–295, 2002.

[23] Wayne Materi and David S. Wishart. Computational systems biology in drug discovery and development: methods and applications. *Drug Discovery Today*, 12(7):295–303, 2007.

[24] Andrzej Mizera, Eugen Czeizler, and Ion Petre. Self-assembly models of variable resolution. *LNBI Transactions on Computational Systems Biology*, 7625:181–203, 2011.

[25] Elaine Murphy, Vincent Danos, Jérôme Feret, Jean Krivine, and Russell Harmer. *Elements of Computational Systems Biology*, chapter Rule Based Modelling and Model Refinement, pages 83–114. Wiley Book Series on Bioinformatics. John Wiley & Sons, Inc., 2010.

[26] Ion Petre, Andrzej Mizera, Claire L. Hyder, Annika Meinander, Andrey Mikhailov, Richard I. Morimoto, Lea Sistonen, John E. Eriksson, and Ralph-Johan Back. A simple mass-action model for the eukaryotic heat shock response and its mathematical validation. *Natural Computing*, 10(1):595–612, 2011.

[27] Corrado Priami. Algorithmic systems biology. *Communications of the ACM*, 52(5):80–88, 2009.

[28] Karthik Raman and Nagasuma Chandra. Systems biology. *Resonance*, 15(2):131–153, 2010.

[29] Adam M. Smith, Wen Xu, Y. Sun, James R. Faeder, and G. Elisabeta Marai. RuleBender: integrated modeling, simulation and visualization for rule-based intracellular biochemistry. *BMC Bioinformatics*, 13(Suppl 8):S3, 2012.

[30] Wen Xu, Adam M. Smith, James R. Faeder, and G. Elisabeta Marai. RuleBender: a visual interface for rule-based modeling. *Bioinformatics*, 27(12):1721–1722, 2011.

Publication IV

Refinement-based modelling of the ErbB signalling pathway – Extended abstract

Bogdan Iancu, Cristian Gratie and Ion Petre

Originally published in: Annals of University of Bucharest LXI, pages 7-14, The Bucharest University Press, 2014.

REFINEMENT-BASED MODELLING OF THE ERBB SIGNALLING PATHWAY

BOGDAN IANCU, CRISTIAN GRATIE and ION PETRE

Computational Biomodeling Laboratory,
Turku Centre for Computer Science and
Department of IT, Åbo Akademi University
Joukahainengatan 3-5, FIN-20520 Åbo
{biancu,cgratie,ipetre}@abo.fi

Abstract

Building large biological models is a difficult task, often attained by iteratively adding details to an initial abstraction of the modeled process. Refitting the model at every step of the development is computationally intensive. *Fit-preserving data refinement* offers an efficient alternative by providing adequate parameter values that preserve the fit from the previous step. We focus here on the implementation of fit-preserving data refinement of a model of the ErbB signalling pathway, which is extended to include details regarding the types of ligands and receptors involved. We obtained an extensive model ensuring a good fit by construction, with significantly less effort than any parameter estimation routine would require.

1 Introduction

Research in cell biology has been governed by a reductionist view for the past century and it has brought about tremendous insights regarding the functionality of singular molecular components, see [2]. However, most cellular functions can hardly ever be ascribed to an individual molecular component. The control mechanisms that regulate cell's adaptation to the environment are highly complex and the predictions regarding their system-level behaviour are exceedingly difficult. Such predictions can be crucial in understanding disease and manipulating cellular dynamics to reverse cellular impairment. A rigorous prediction of the behaviour of a large model reflects the ability of understanding the system of interest in its entirety. Such predictions are very well captured through mathematical modelling, but the field of practical applications exhibits a great deficit in accommodating mathematical modelling as a standard procedure. One reason is the massive number of parameters that need to be either fixed or estimated in a model of large proportions, see [11]. With many parameters to estimate, one needs huge amounts of detailed data that on one hand may be hard to get, and on the other hand makes the model fitting process computationally difficult.

As an alternative, we focus on *refinement-based model construction*. Model refinement originates from the field of software engineering. Stepwise refinement has been introduced in the context of parallel computing and it gave rise to the framework of refinement calculus, used in program refinement to ensure correctness preservation, see [1]. In systems biology, model refinement is essential in the model development cycle. Adding species or reactions to a model, as well as adding or removing certain modules, generates changes in the model fit. The process of model fitting is computationally intensive due to parameter estimation routines, given that most often the refined model has a considerably large number of parameters. An iterative approach based on a traditional reiteration of the whole model fitting procedure is dissipative resource-wise. Alternatively, we can consider an approach where we can build the model iteratively, ensuring model fit preservation at every step of the refinement; the method is called *quantitative model refinement* and it has been addressed formerly in [6, 7, 9, 10, 14, 16].

We examine in this paper the implementation of *fit-preserving quantitative model refinement* of a model describing the ErbB signalling pathway. Our approach is based on *data refinement*, where several subspecies of a given species in the initial model are substituted for their parent species. Our starting point is the ErbB signalling pathway proposed in [13] and [18]. Throughout the paper, the model in [13] will be referred to as the *basic model*. We refined it to include four different types of growth factors and two types of ligands and compared the computational effort with the one of [5].

The paper is organized as follows: we start with a description of the fit-preserving quantitative model refinement approach in Section 2. We introduce in Section 3 the main biological processes involved in the ErbB signalling pathway, as described in [13] and [18] and then we discuss in Section 4 the implementation of fit preserving refinement on the model in [13]. We conclude the paper with an analysis of our results.

2 Quantitative model refinement

In this section we discuss model refinement as proposed in [14] and later extended in [10] to address the assignment of kinetic rate constants of the refined model. We have adjusted the formal notation so as to obtain a uniform presentation of the results and also to avoid abstractions that are unnecessary for the work presented here.

2.1 Data refinement

As addressed in [14], a reaction-based model M consisting of m species, $\Sigma = \{A_1, \ldots, A_m\}$ and n reactions $R = \{r_1, r_2, \ldots, r_n\}$, where $m,n \in \mathbb{N}^*$, can be refined to include more details regarding its reactants, for instance to draw distinction between the subspecies of a given species. Even when the triggering distinction refers to a single atomic species, one needs to propagate this differentiation to all the complex species that contain it. Without loss of generality, we can assume that all species A_i are refined into $\{B_{i1}, \ldots, B_{ip_i}\}$, where $p_i \geqslant 1$, with $p_i = 1$ for species A_i that are not actually refined.

The new model obtained in this way, M_R, consists of species $\{B_{iq} \mid 1 \leqslant i \leqslant m, 1 \leqslant q \leqslant p_i\}$. It is assumed that all subspecies of A_i will participate in all the reactions A_i took part in, but presumably following a different kinetic law than the one of the parent species. Thus, each reaction from M of the form

$$r_j \ : \ c_{j1}A_1 + c_{j2}A_2 + \ldots + c_{jm}A_m \xrightarrow{k_j} d_{j1}A_1 + d_{j2}A_2 + \ldots + d_{jm}A_m$$

is replaced in the refined model M_R by all possible reactions $r_j^{(l)}$ of the following form:

$$r_j^{(l)} \ : \ \sum_{q=1}^{p_1} c_{j1q}^{(l)} B_{1q} + \ldots + \sum_{q=1}^{p_m} c_{jmq}^{(l)} B_{mq} \xrightarrow{k_j^{(l)}} \sum_{q=1}^{p_1} d_{j1q}^{(l)} B_{1q} + \ldots + \sum_{q=1}^{p_m} d_{jmq}^{(l)} B_{mq},$$

where $k_j^{(l)}$ is the kinetic rate constant of reaction $r_j^{(l)}$ and the coefficients $c_{jiq}^{(l)}$ and $d_{jiq}^{(l)}$ are non-negative integers such that $c_{ji1}^{(l)} + \ldots + c_{jip_i}^{(l)} = c_{ji}$ and $d_{ji1}^{(l)} + \ldots + d_{jip_i}^{(l)} = d_{ji}$.

Thus, from a structural point of view, the refined model M_R captures the same kind of interactions as the original model M, only written in terms of the refined species B_{iq} and allowing for different values of the kinetic rate constants.

2.2 Fit-preserving refinement

Given a refined model M_R constructed as explained in the previous subsection, we aim to assign values to its rate constants in such a way as to characterize the same dynamics with respect to parent species, i.e.:

$$[A_i](t) = [B_{i1}](t) + \ldots + [B_{ip_i}](t) \tag{1}$$

for all $1 \leqslant i \leqslant m$ and $t \geqslant 0$. These constraints define what is referred to as *fit-preserving refinement* in [10]. In the paper, the authors provide a set of linear constraints on the kinetic rate constants that is sufficient to ensure the satisfaction of (1). In order to write these conditions formally, we need to introduce some additional notations. We are going to pack the stoichiometric coefficients of the original reactions r_j into vectors $\boldsymbol{c_j}$ and $\boldsymbol{d_j}$, respectively. Similarly, the left and right hand sides of refined reactions $r_j^{(l)}$ are collected in vectors $\boldsymbol{c_j^{(l)}}$ and $\boldsymbol{d_j^{(l)}}$. With this, the sufficient conditions for M_R to be a fit-preserving refinement of M are:

$$\sum_{l \text{ s.t. } \boldsymbol{c_j^{(l)}} = \boldsymbol{c_j^{(s)}}} k_j^{(l)} = \binom{\boldsymbol{c_j}}{\boldsymbol{c_j^{(s)}}} k_j \ , \tag{2}$$

for any reaction r_j and any selected left hand side $\boldsymbol{c_j^{(s)}}$ of corresponding refined reactions. The sum is taken over all refined reactions that share the selected left hand side. The formula uses the generalized binomial coefficient for vectors $\boldsymbol{x}, \boldsymbol{y}$:

$$\binom{\boldsymbol{x}}{\boldsymbol{y}} = \frac{\prod_i x_i!}{\prod_j y_j!}, \ .$$

In the model we built, we chose equal values for all the rate constants that appear in the same sum of (2), by splitting the value prescribed by the constraint to the actual number of reactions that share the selected left hand side.

3 Case-study: the EGFR signalling pathway

The EGFR signalling network is an evolutionary developmental pathway, which regulates various physiological responses of the mammalian cell, such as growth, survival, proliferation, differentiation and motility, and plays a major role in oncogenesis, see [3, 5, 17]. Anomalies found in the EGFR signalling pathway have been associated with various cancer types and the pharmaceutical intervention developed to tackle these abnormalities proved to be successful in the treatment of some cancer types, see [3, 19].

Epidermal growth factors (EGFRs) are receptor tyrosine kinases (RTKs), which regulate all the aforementioned physiological responses. Signalling is activated by binding to the extracellular domain of the epidermal growth factor (e.g. EGF) or another factor which belongs to the EGF family, see [15]. This binding induces EGFR

dimerizatin and, consequently, an accelerated auto-phosphorylation of its intracellular domain. Signal propagation is promoted by two distinct pathways: Shc-dependent and Shc-independent, both compelling the activation of the Ras-GTP protein. The Shc -dependent pathway is initiated by binding of Shc to the autophosphorylated, ligand-bound, dimerized receptor and is promoted by binding to the growth factor receptor-binding protein 2, Grb2 . The Shc-independent pathway is initiated by a direct binding of the autophosphorylated, ligand-bound, dimerized receptor with Grb2 . Both pathways induce the recruitment of Sos to the membrane; protein Ras is docked onto the membrane and its association with Sos activates it by the formation of Ras-GTP. This process induces consecutively the activation of the MAPK cascade through the Raf, MEK and ERK kinases, see [13, 18].

The initial model, introduced in [13], is a reaction-based model of the EGF-induced signal transduction through the MAPK cascade and it consists of 148 reactions, 103 reactants and 90 kinetic rate constants. It is an updated version of the models in [15] and [18]. The model comprises a negative feedback loop from the dual phosphorylated ERK (ERK-PP) to the Sos protein, which brings about the unbinding of Grb2-Sos from the receptor complex; the process has been previously documented in the literature, see [4, 8]. The model does not take into consideration protein isoform specificity (several forms of the same protein), see [13]. In the absence of the stimulus EGF, the system depicted by the model in [13] exhibits a stable steady state, which corresponds to a state of unphosphorylated ERK. The model draws a distinction between two pools of double-phosphorylated ERK, one of which is situated in the cytoplasm and one correlated with the internalization process, see [13]. The model consists of a set of 13 biochemical processes: EGFR activation (e.g.: the binding of the receptor to the ligand: $\mathsf{EGF} + \mathsf{EGFR} \rightarrow \mathsf{EGF\text{-}EGFR}$), Shc, Grb2, Sos recruitment, activation and inactivation of Ras, activation of Raf, dephosphorylation of Raf, phosphorylation and dephosphorylation of MEK, ERK dephosphorylation, negative feedback from ERK to Sos, internalization of complexes involving EGFR and degradation reactions. For more details, we refer the reader to [13]. We have imported the model in COPASI ([12]), it is available at: `http://combio.abo.fi/www_includes/qr_pro/HornbergErbBinitial.zip`.

4 The implementation of fit-preserving model refinement

In this section, we focus on expanding the EGFR signalling pathway model from [13] by considering all receptors of the ErbB family, ErbB1-4 receptors: ErbB1 (EGFR),

ErbB2 (HER2), ErbB3, ErbB4. Furthermore, all complex species which involve the aforementioned species are to be refined as well. We take into account all combinations of hetero- and homodimers in our model, as well as two types of ligands: EGF and HRG. We also consider that all ligands bind to all dimers represented in the model.

With respect to their data refinement, the species of the model can be classified in the following categories: species that are not affected by refinement (atomic or complex species which don't involve any of the species to be refined), atomic species that are to be refined, and complexes that involve species to be refined.

The basic model in [13] consists of a number of 103 species and 148 reactions, while the refined model comprises a number of 421 species involved in 928 reactions. We refined the reactions of the basic model in [13] following the approach of [10, 14], briefly presented in Section 2. The refined model can be found at: `http://combio.abo.fi/www_includes/qr_pro/ErbB1-4Refinement.zip`.

5 Discussion

Model refinement is a fundamental step in the model development cycle. Developing an extensive system-level biological model requires in the initial stages an abstraction of the biological processes to be considered. However, these abstractions have to be very often refined to include more accurate details regarding the aforementioned processes. A traditional approach would be to reiterate the whole model development cycle; however, this method is overly ineffective as it requires refitting the model, a process that is time-consuming and computationally-expensive. For instance, for a model consisting of 828 reactions and 499 reactants, a good fit was obtained by running about 100 times annealing methods, over 24 hours on a cluster consisting of 100 nodes, see [5]. Fit-preserving quantitative model refinement ensures a good fit, starting from an already fit original model; further refinement steps can be applied after this original refinement so as to include more details regarding biological knowledge of the model.

We refined the model from [13], considering two types of ligands: EGF and HRG and four types of receptors: ErbB1 (EGFR), ErbB2 (HER2), ErbB3, ErbB4. This gave rise to a significant augmentation in the number of reactants and the number of reactions. While the initial model consists of a number of 103 reactants and 148 reactions, the refined model comprises a number of 421 reactants involved in 928 reactions. Fitting a model of this proportion would be comparable effort-wise to that of [5].

However, our approach has proved to be efficient in building a refined model, with a good numerical behaviour, avoiding any supplementary model fit.

References

1. R. Back and J. von Wright. *Refinement Calculus*. Springer (New York), 1998.
2. A.-L. Barabasi and Z. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004.
3. M. Birtwistle, M. Hatakeyama, N. Yumoto, B. Ogunnaike, J. Hoek, and B. Kholodenko. Ligand-dependent responses of the ErbB signaling network: experimental and modeling analyses. *Molecular systems biology*, 3(1), 2007.
4. L. Buday, P. Warne, and J. Downward. Downregulation of the Ras activation pathway by MAP kinase phosphorylation of Sos. *Oncogene*, 11(7):1327–1331, 1995.
5. W. Chen, B. Schoeberl, P. Jasper, M. Niepel, U. Nielsen, D. Lauffenburger, and P. Sorger. Input–output behavior of ErbB signaling pathways as revealed by a mass action model trained against dynamic data. *Molecular systems biology*, 5(1), 2009.
6. V. Danos, J. Feret, W. Fontana, R. Harmer, and J. Krivine. Rule-based modelling, symmetries, refinements. In *Formal Methods in Systems Biology*, pages 103–122. Springer, 2008.
7. V. Danos, J. Feret, W. Fontana, R. Harmer, and J. Krivine. Rule-based modelling and model perturbation. In *Transactions on Computational Systems Biology XI*, pages 116–137. Springer, 2009.
8. C. Dong, S. Waters, K. Holt, and J. Pessin. Sos phosphorylation and disassociation of the Grb2-SOS complex by the ERK and JNK signaling pathways. *Journal of Biological Chemistry*, 271(11):6328–6332, 1996.
9. J. Faeder, M. Blinov, B. Goldstein, and W. Hlavacek. Rule-based modeling of biochemical networks. *Complexity*, 10(4):22–41, 2005.
10. C. Gratie and I. Petre. Fit-preserving data refinement of mass-action reaction networks. In *Proceedings CiE 2014, Language, Life, Limits*, volume 8493 of *Lecture Notes in Computer Science*, pages 204–213. Springer, 2014.
11. W. Hlavacek. How to deal with large models? *Molecular Systems Biology*, 5(1), 2009.
12. S. Hoops, S. Sahle, R. Gauges, C. Lee, J. Pahle, N. Simus, M. Singhal, L. Xu, P. Mendes, and U. Kummer. Copasi–a complex pathway simulator. *Bioinformatics*, 22(24):3067–3074, 2006.
13. J. Hornberg, B. Binder, F. Bruggeman, B. Schoeberl, R. Heinrich, and H. Westerhoff. Control of MAPK signalling: from complexity to what really matters. *Oncogene*, 24(36):5533–5542, 2005.
14. B. Iancu, E. Czeizler, E. Czeizler, and I. Petre. Quantitative refinement of reaction models. *International Journal of Unconventional Computing*, 8(5-6):529–550, 2012.
15. B. Kholodenko, O. Demin, G. Moehren, and J. Hoek. Quantification of short term signaling by the epidermal growth factor receptor. *Journal of Biological Chemistry*, 274(42):30169–30181, 1999.
16. E. Murphy, V. Danos, J. Feret, J. Krivine, and R. Harmer. Rule based modeling and model refinement. *Elements of Computational Systems Biology*, pages 83–114, 2009.
17. K. Oda, Y. Matsuoka, A. Funahashi, and H. Kitano. A comprehensive pathway map of epidermal growth factor receptor signaling. *Molecular systems biology*, 1(1), 2005.

18. B. Schoeberl, C. Eichler-Jonsson, E. Gilles, and G. Müller. Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. *Nature biotechnology*, 20(4):370–375, 2002.
19. Y. Yarden and M. Sliwkowski. Untangling the ErbB signalling network. *Nature reviews Molecular cell biology*, 2(2):127–137, 2001.

Publication V

Reaction system models for the heat shock response

Sepinoud Azimi, Bogdan Iancu and Ion Petre

Originally published in: *Fundamenta Informaticae*, 131(3), 299–312, IOS Press, 2014.

Fundamenta Informaticae 131 (2014) 299–312
DOI 10.3233/FI-2014-1016
IOS Press

Reaction System Models for the Heat Shock Response

Sepinoud Azimi*, Bogdan Iancu, Ion Petre
Computational Biomodeling Laboratory
Turku Centre for Computer Science
Åbo Akademi University
20520 Turku, Finland
{*Sepinoud.Azimi, Bogdan.Iancu, Ion.Petre*}*@abo.fi*

Abstract. Reaction systems are a formal framework for modeling processes driven by biochemical reactions. They are based on the mechanisms of facilitation and inhibition. A main assumption is that if a resource is available, then it is present in sufficient amounts and as such, several reactions using the same resource will not compete concurrently against each other; this makes reaction systems very different as a modeling framework than traditional frameworks such as ODEs or continuous time Markov chains. We demonstrate in this paper that reaction systems are rich enough to capture the essential characteristics of ODE-based models. We construct a reaction system model for the heat shock response in such a way that its qualitative behavior correlates well with the quantitative behavior of the corresponding ODE model. We construct our reaction system model based on a novel concept of dominance graph that captures the competition on resources in the ODE model. We conclude with a discussion on the expressivity of reaction systems as compared to that of ODE-based models.

Keywords: Reaction systems; heat shock response; quantitative model; qualitative model; model comparison.

1. Introduction

Reaction systems (RS in short) are a formal framework for modeling processes driven by biochemical reactions. They were introduced in [2], see also [1] and references therein. The fundamental idea in this

*Address for correspondence: Department of Information Technologies, Åbo Akademi University, Joukahaisenkatu 3-5, 20520 Turku, Finland

framework is that biochemical reactions are based on the mechanisms of *facilitation* and *inhibition*. A reaction is modeled as a triplet: a set of reactants, a set of inhibitors, and a set of products. A reaction can take place in a given state if all its reactants are present in that state and none of its inhibitors; when triggered, the reaction creates its products. Two major assumptions in reaction systems set them apart from standard methods for biomodelling (such as ordinary differential equations, stochastic processes, Petri nets, and process algebras):

- *The threshold assumption:* if a resource is present, then it is present in a "sufficient amount" and it will not cause any conflict between several reactions needing that resource. In other words, several reactions needing the same reactant will not be in conflict.
- *No permanency assumption:* an entity will vanish from the current state unless it is produced by one of the reactions enabled in that state.

The goal of this paper is to demonstrate that reaction systems are capable of capturing the essential characteristics of complex ODE models. We construct an RS model for the molecular heat shock response introduced in [6]. Our focus is on building the model in such a way that a number of properties of the ODE-based model of [6] for the heat shock response are preserved: mass-conservation, steady state configuration with and without stress, behavior under continuous stress. The challenge here is that these properties are essentially numerical, correlating well to numerical experimental data and knowledge, whereas the RS framework is qualitative, as shown by the threshold assumption. Moreover, special attention has to be given to overcoming the no permanency assumption to make sure that, e.g., a gene is not removed from the system, even when no gene activity has occurred. We first take the straightforward approach of building the RS model through translating the reactants/products from the molecular model to an RS with the same reactants/products and no inhibitors. It turns out however that the resulting RS model leads to a behavior that is very different than that of the ODE-based model. We show however that an RS model can be built in a different way, qualitatively replicating the numerical behavior of the ODE model.

The paper is structured as follows. In Section 2 we introduce some basics of modeling with ODEs and some basic notions of reaction systems. In Section 3 we introduce our molecular model for the heat shock response and discuss some of the numerical properties of its corresponding mass-action-based ODE model. In Section 4 we build a direct translation of the molecular model to an RS model with no inhibitors and look at some of its interactive processes. In Section 5 we build a different RS model whose interactive processes correlate well with the numerical behavior of the ODE model. We conclude with some discussion in Section 6.

2. Preliminaries

2.1. Reaction-based molecular models and their ODE-based representation

Biochemical networks can be represented as reaction-based molecular models. Such models consist of sets of coupled chemical reactions describing the system of interest; the reactions can be reversible or irreversible. Formally, the fundamental constituents of a model M are represented by a set of species $\Sigma = \{S_i \mid 1 \leq i \leq m\}$ and a set of reactions $\{R_j \mid 1 \leq j \leq n\}$, where n and m are nonnegative integers.

An irreversible reaction R_j, $1 \leq j \leq n$, is formalized as a rewriting rule as follows:

$$R_j \ : \ c_{1,j}S_1 + c_{2,j}S_2 + \ldots + c_{m,j}S_m \xrightarrow{k_j} c'_{1,j}S_1 + c'_{2,j}S_2 + \ldots + c'_{m,j}S_m.$$

A reversible reaction can be written in the following form:

$$R_j \ : \ c_{1,j}S_1 + c_{2,j}S_2 + \ldots + c_{m,j}S_m \underset{k_j^-}{\overset{k_j^+}{\rightleftarrows}} c'_{1,j}S_1 + c'_{2,j}S_2 + \ldots + c'_{m,j}S_m.$$

The nonnegative integers k_j and k_j^-, k_j^+ represent the *kinetic rate constants* of an irreversible, respectively reversible reaction R_j. The coefficients $c_{1,j}, \ldots, c_{m,j}, c'_{1,j}, \ldots, c'_{m,j}$ are positive integers characterizing the stoichiometry of the reaction. The stoichiometric coefficient of species S_i in reaction R_j is defined by: $n_{i,j} = c'_{i,j} - c_{i,j}$.

The reactant species, found on the left-hand side of the reaction, are called *substrates*, while the species produced, occurring on the right-hand side, are referred to as *products*. A species S_i with $c_{i,j} = 0$ ($c'_{i,j} = 0$, resp.) can be omitted from the left-hand (right-hand, resp.) side of reaction R_j corresponding to the null coefficient. A reversible reaction can always be regarded as a pair of two irreversible reactions, see [3].

The *molecularity* of a reaction R_j is defined by the following sum: $\sum_{i=1}^{m} c_{i,j}$. One generally considers systems comprising reactions of a molecularity of at most two. Reactions exhibiting a molecularity of three are very infrequent due to the high improbability of simultaneous collisions between three molecules leading to the formation of a complex. Reactions with a molecularity greater than three are completely disregarded due to the impossibility of collision between more than three molecules synchronously, see [5].

A molecular model M can be represented as a mathematical model in various manners, following a continuous or discrete time evolution, based on continuous or discrete species variables. A common representation is based on ordinary differential equations (ODEs) and the principle of mass-action, see [4]. To each species S_i, $1 \leq i \leq m$, one associates a time-dependent function: $[S_i] : \mathbb{R}_+ \to \mathbb{R}_+$, representing the concentration of the species over time. Therefore, the evolution of the system is described by a system of differential equations of the following form:

$$\frac{d[S_i]}{dt} = -\sum_{j=1}^{n} \left(k_j^+ c_{i,j} \prod_{l=1}^{m} [S_l]^{c_{l,j}}(t) \right) + \sum_{j=1}^{n} \left(k_j^- c'_{i,j} \prod_{l=1}^{m} [S_l]^{c'_{l,j}}(t) \right), \ \ 1 \leq i \leq m.$$

Such systems of ODEs can be rarely solved analytically, but many numerical methods for analyzing them exist. In particular, a numerical integration of the ODE system is interpreted as a numerical simulation of the corresponding molecular model.

2.2. Reaction systems

Reaction systems have been introduced in [2] as a formal framework for the analysis of biochemical networks. A biochemical reaction in this framework is based on a finite set of reactants and it is triggered provided that all the reactants involved in that particular reaction are *present* in a given state and all of its inhibitors are *absent*, see [8].

We recall in this section the basic definitions we need throughout the paper. For more details we refer to [2, 8].

Definition 2.1. [2] 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 *rac*(S).

Definition 2.2. [2] Let A be a 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

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

Definition 2.3. [2] A *reaction system* (RS in short) is defined as an ordered pair $\mathcal{A} = (S, A)$, where S is a finite set and $A \subseteq$ *rac*(S). The set S is called the *background* (set) of A.

Definition 2.4. [2] Let $\mathcal{A}$ be a reaction system. 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 $D_1 =$ *res*$_{\mathcal{A}}(C_0)$ and, for each $1 < i \leq n$, $D_i =$ *res*$_{\mathcal{A}}(C_{i-1} \cup D_{i-1})$.

The sequences γ and δ are the *context sequence* of π, *con*(π), and the *result sequence* of π, *res*(π), resp. The *state sequence* of π is $\tau = W_0, W_1, ..., W_n$, where $W_i = C_i \cup D_i$, for all $i \in \{0, ..., n\}$ and $W_0 = C_0$. W_0 is the *initial state* of π, init(π), and W_n is the *final state* of π, fst(π).

Definition 2.5. 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 *res*$_{\mathcal{A}}(C \cup D) = D$.

3. A molecular model for the heat shock response

The heat shock response in eukaryotes is a fundamental, well conserved defense mechanism, which allows the cell to react to environmental stressors such as elevated temperatures. The increase of temperature in the environment causes proteins in the cell to misfold and build up large conglomerates that ultimately result in cell death. The key elements for the heat shock response mechanism are the heat shock proteins (hsp), which operate as molecular chaperones for misfolded proteins (mfp), facilitating their refolding process. The response is regulated by the transactivation of the hsp-encoding genes. Gene transcription is mediated by heat shock factors (hsf), which, under stress, dimerize (hsf_2), subsequently trimerize (hsf_3) and then bind to a promoter-site of the hsp-encoding gene, the heat shock-element (hse). As soon as trimers are bound, protein synthesis is activated and new hsp molecules are produced. When the level of hsp is sufficiently uplifted, hsp synthesis is turned off, see [6, 7]. Heat shock proteins sequestrate hsf molecules, hence promoting DNA binding. When the temperature is elevated, proteins in the

Table 1. The molecular model for the eukaryotic heat shock response proposed in [6].

Reaction	Reaction
$2\,\mathsf{hsf} \leftrightarrows \mathsf{hsf}_2$	$\mathsf{hsp} + \mathsf{hsf}_3 \to \mathsf{hsp{:}\,hsf} + 2\,\mathsf{hsf}$
$\mathsf{hsf} + \mathsf{hsf}_2 \leftrightarrows \mathsf{hsf}_3$	$\mathsf{hsp} + \mathsf{hsf}_3{:}\,\mathsf{hse} \to \mathsf{hsp{:}\,hsf} + 2\,\mathsf{hsf} + \mathsf{hse}$
$\mathsf{hsf}_3 + \mathsf{hse} \leftrightarrows \mathsf{hsf}_3{:}\,\mathsf{hse}$	$\mathsf{hsp} \to \emptyset$
$\mathsf{hsf}_3{:}\,\mathsf{hse} \to \mathsf{hsf}_3{:}\,\mathsf{hse} + \mathsf{hsp}$	$\mathsf{prot} \to \mathsf{mfp}$
$\mathsf{hsp} + \mathsf{hsf} \leftrightarrows \mathsf{hsp{:}\,hsf}$	$\mathsf{hsp} + \mathsf{mfp} \leftrightarrows \mathsf{hsp{:}\,mfp}$
$\mathsf{hsp} + \mathsf{hsf}_2 \to \mathsf{hsp{:}\,hsf} + \mathsf{hsf}$	$\mathsf{hsp{:}\,mfp} \to \mathsf{hsp} + \mathsf{prot}$

cell (prot) tend to misfold, causing hsp: hsf complexes to break up. The heat shock response is switched on again, facilitating hsp synthesis, see [6]. We list in Table 1 the reactions of the molecular model in [6].

A mathematical model is associated with the molecular model in Table 1. The mathematical model consists in a mass-action-based system of ODEs, see [4] for a brief discussion on the principle of mass-action. We refer to [6] for the system of ODEs and the numerical setup of the ODE model.

For a constant temperature of $37°C$ the model reaches a steady state where most hsf's are bound in a complex hsp: hsf, most hse's are available for binding to trimers and there are very few misfolded proteins. For a constant temperature of $42°C$ the model reaches a steady state different from the one at $37°C$ in that the level of misfolded proteins, in both forms mfp and hsp: mfp are relatively high. Upon removal of the stress and return to $37°C$, the model returns to the basal values attained under a constant temperature of $37°C$. For a more detailed discussion about the steady states of the model and numerical simulations we refer to [6].

4. From the molecular model to a reaction system model: a direct translation

In this section, we formulate a reaction system model for the heat shock response, based on the mechanisms of facilitation and inhibition. We use a direct translation approach, i.e. we translate each reaction in the molecular model to a reaction in the corresponding reaction system. We disregard in the current model the reaction $\mathsf{hsp} \to \emptyset$ in Table 1 due to its very low reaction rate in the ODE model; similarly, we ignore reactions $\mathsf{hsf}_3{:}\,\mathsf{hse} \to \mathsf{hsf}_3 + \mathsf{hse}$ and $\mathsf{hsp{:}\,mfp} \to \mathsf{hsp} + \mathsf{mfp}$ (both the reverse directions of some irreversible reactions in Table 1). We also disregard the dimer form hsf_2 of hsf; indeed, the dimer is only a transient state from hsf to hsf_3 and their levels remain insignificant regardless of the stress. The simplified molecular model for the heat shock response is in Table 2.

Table 2. The simplified molecular model for the eukaryotic heat shock response.

Reaction		Reaction	
$3\,\mathsf{hsf} \leftrightarrows \mathsf{hsf}_3$	(1)	$\mathsf{hsp} + \mathsf{hsf}_3{:}\,\mathsf{hse} \to \mathsf{hsp{:}\,hsf} + 2\,\mathsf{hsf} + \mathsf{hse}$	(6)
$\mathsf{hsf}_3 + \mathsf{hse} \to \mathsf{hsf}_3{:}\,\mathsf{hse}$	(2)	$\mathsf{prot} \to \mathsf{mfp}$	(7)
$\mathsf{hsf}_3{:}\,\mathsf{hse} \to \mathsf{hsf}_3{:}\,\mathsf{hse} + \mathsf{hsp}$	(3)	$\mathsf{hsp} + \mathsf{mfp} \to \mathsf{hsp{:}\,mfp}$	(8)
$\mathsf{hsp} + \mathsf{hsf} \leftrightarrows \mathsf{hsp{:}\,hsf}$	(4)	$\mathsf{hsp{:}\,mfp} \to \mathsf{hsp} + \mathsf{prot}$	(9)
$\mathsf{hsp} + \mathsf{hsf}_3 \to \mathsf{hsp{:}\,hsf} + 2\,\mathsf{hsf}$	(5)		

4.1. The first reaction system model

We describe first a simple method for translating a set of molecular reactions to a reaction system. A unary molecular reaction $A \rightarrow C$ is translated into the RS reaction $\{\{A\}, \{\mathsf{d_I}\}, \{C\}\}$, where $\mathsf{d_I}$ is a "dummy" variable. A binary molecular reaction $A + B \rightarrow C$ is translated into the RS reaction $\{\{A, B\}, \{\mathsf{d_I}\}, \{C\}\}$. The dummy variable is only used here to comply with the constraint that the set of inhibitors of all RS reactions should be non-empty; none of the molecular reactions specify any explicit inhibitor. The ternary reaction (1) is treated as a unary reaction $\mathsf{hsf} \leftrightarrow \mathsf{hsf}_3$. The case of reversible molecular reactions is handled analogously, provided that we first replace it with two irreversible molecular reactions, standing for the two directions of the original reaction, and then define their correspondents in a reaction system as above. The full RS model obtained as a result is given in Table 3.

Table 3. The direct translation of the biochemical reactions of the simplified model of the heat shock response to a reaction system.

Reaction in the chemical network	Reaction in the reaction system	
$3\,\mathsf{hsf} \leftrightarrow \mathsf{hsf}_3$	$(\{\mathsf{hsf}\}, \{\mathsf{d_I}\}, \{\mathsf{hsf}_3\})$	(i)
	$(\{\mathsf{hsf}_3\}, \{\mathsf{d_I}\}, \{\mathsf{hsf}\})$	(ii)
$\mathsf{hsf}_3 + \mathsf{hse} \rightarrow \mathsf{hsf}_3\colon \mathsf{hse}$	$(\{\mathsf{hsf}_3, \mathsf{hse}\}, \{\mathsf{d_I}\}, \{\mathsf{hsf}_3\colon \mathsf{hse}\})$	(iii)
$\mathsf{hsf}_3\colon \mathsf{hse} \rightarrow \mathsf{hsf}_3\colon \mathsf{hse} + \mathsf{hsp}$	$(\{\mathsf{hsf}_3\colon \mathsf{hse}\}, \{\mathsf{d_I}\}, \{\mathsf{hsf}_3\colon \mathsf{hse}, \mathsf{hsp}\})$	(iv)
$\mathsf{hsp} + \mathsf{hsf} \leftrightarrow \mathsf{hsp}\colon \mathsf{hsf}$	$(\{\mathsf{hsp}, \mathsf{hsf}\}, \{\mathsf{d_I}\}, \{\mathsf{hsp}\colon \mathsf{hsf}\})$	(v)
	$(\{\mathsf{hsp}\colon \mathsf{hsf}\}, \{\mathsf{d_I}\}, \{\mathsf{hsp}, \mathsf{hsf}\})$	(vi)
$\mathsf{hsp} + \mathsf{hsf}_3 \rightarrow \mathsf{hsp}\colon \mathsf{hsf} + 2\,\mathsf{hsf}$	$(\{\mathsf{hsp}, \mathsf{hsf}_3\}, \{\mathsf{d_I}\}, \{\mathsf{hsp}\colon \mathsf{hsf}, \mathsf{hsf}\})$	(vii)
$\mathsf{hsp} + \mathsf{hsf}_3\colon \mathsf{hse} \rightarrow \mathsf{hsp}\colon \mathsf{hsf} + \mathsf{hse} + 2\,\mathsf{hsf}$	$(\{\mathsf{hsp}, \mathsf{hsf}_3\colon \mathsf{hse}\}, \{\mathsf{d_I}\}, \{\mathsf{hsp}\colon \mathsf{hsf}, \mathsf{hsf}, \mathsf{hse}\})$	(viii)
$\mathsf{prot} \rightarrow \mathsf{mfp}$	$(\{\mathsf{prot}\}, \{\mathsf{d_I}\}, \{\mathsf{mfp}\})$	(ix)
$\mathsf{hsp} + \mathsf{mfp} \rightarrow \mathsf{hsp}\colon \mathsf{mfp}$	$(\{\mathsf{hsp}, \mathsf{mfp}\}, \{\mathsf{d_I}\}, \{\mathsf{hsp}\colon \mathsf{mfp}\})$	(x)
$\mathsf{hsp}\colon \mathsf{mfp} \rightarrow \mathsf{hsp} + \mathsf{prot}$	$(\{\mathsf{hsp}\colon \mathsf{mfp}\}, \{\mathsf{d_I}\}, \{\mathsf{hsp}, \mathsf{prot}\})$	(xi)

4.2. Interactive processes in the first model

We analyze in this subsection the dynamics of the reaction system in terms of interactive processes and compare it with that of the corresponding ODE model.

Note that in the ODE model the temperature was taken into account through the temperature dependent protein misfolding rate constant. Since our construction only translates the reactions and not their quantitative values, the effect of the temperature on the model is lost in the translation. Consequently, the RS model can only at best capture some of the behaviour of the model in the absence of stress; we show below that the model in fact even fails to do that.

In our first interactive process we start from an initial state consisting of a minimal set of species needed in the heat shock response model: hsf, hse, and prot; all other species and complexes can be

obtained in the molecular model (as well as in the ODE model) starting from these main ingredients. Thus, let the initial context of our reaction system be $C_0 = \{\mathsf{hsf}, \mathsf{prot}, \mathsf{hse}\}$. Throughout this interactive process all subsequent contexts are empty: $C_i = \emptyset$, for all $i \geq 1$. This interactive process is represented in Table 4 and it shows that for every $k \geq 1$, $D_{2k} = \{\mathsf{hsf}\}$ and $D_{2k+1} = \{\mathsf{hsf}_3\}$. The prediction of the RS model is thus that the model will enter into a loop of length two when starting from $\{\mathsf{hsf}, \mathsf{prot}, \mathsf{hse}\}$, for an empty context. This is in contradiction with the prediction of the ODE model, which shows the system converging to a steady state at $37°C$.

Table 4. An interactive process for the direct translation of the simplified model of the heat shock reponse for the first setting.

State	C_i	D_i	W_i	r_i
0	$\{\mathsf{hsf}, \mathsf{prot}, \mathsf{hse}\}$	$\emptyset$	$\{\mathsf{hsf}, \mathsf{prot}, \mathsf{hse}\}$	{(i), (ix)}
1	$\emptyset$	$\{\mathsf{hsf}_3, \mathsf{mfp}\}$	$\{\mathsf{hsf}_3, \mathsf{mfp}\}$	{ (ii) }
2	$\emptyset$	$\{\mathsf{hsf}\}$	$\{\mathsf{hsf}\}$	{ (i) }
3	$\emptyset$	$\{\mathsf{hsf}_3\}$	$\{\mathsf{hsf}_3\}$	{ (ii) }
4	$\emptyset$	$\{\mathsf{hsf}\}$	$\{\mathsf{hsf}\}$	{ (i) }

For our second interactive process, the initial state consists of all species included in the $37°C$ steady-state of the ODE model: $C_0 = \{\mathsf{hse}, \mathsf{hsp}\colon \mathsf{hsf}, \mathsf{prot}\}$. The interactive process starting from this state and using an empty context is represented in Table 5. Thus, the prediction of the RS model is that $D_{2k-1} = D_5$ and $D_{2k} = D_6$, for all $k \geq 3$. This again shows a contradiction with the ODE model. Indeed, starting from an initial state corresponding to the $37°C$ steady state of the ODE model, the RS model eventually enters into a loop of length 2.

Table 5. An interactive process for the direct translation of the simplified model of the heat shock response for the second setting.

State	C_i	D_i	W_i	r_i
0	$\{\mathsf{hse}, \mathsf{hsp}\colon \mathsf{hsf}, \mathsf{prot}\}$	$\emptyset$	$\{\mathsf{hse}, \mathsf{hsp}\colon \mathsf{hsf}, \mathsf{prot}\}$	(vi) , (ix)
1	$\emptyset$	$\{\mathsf{hsp}, \mathsf{hsf}, \mathsf{mfp}\}$	$\{\mathsf{hsp}, \mathsf{hsf}, \mathsf{mfp}\}$	(i), (v), (x)
2	$\emptyset$	$\{\mathsf{hsf}_3, \mathsf{hsp}\colon \mathsf{hsf}, \mathsf{hsp}\colon \mathsf{mfp}\}$	$\{\mathsf{hsf}_3, \mathsf{hsp}\colon \mathsf{hsf}, \mathsf{hsp}\colon \mathsf{mfp}\}$	(ii), (vi), (xi)
3	$\emptyset$	$\{\mathsf{hsp}, \mathsf{hsf}, \mathsf{prot}\}$	$\{\mathsf{hsf}, \mathsf{hsp}\}$	(i), (v), (ix)
4	$\emptyset$	$\{\mathsf{hsf}_3, \mathsf{hsp}\colon \mathsf{hsf}, \mathsf{mfp}\}$	$\{\mathsf{hsf}_3, \mathsf{hsp}\colon \mathsf{hsf}, \mathsf{mfp}\}$	(ii), (vi)
5	$\emptyset$	$\{\mathsf{hsf}, \mathsf{hsp}\}$	$\{\mathsf{hsf}, \mathsf{hsp}\}$	(i), (v)
6	$\emptyset$	$\{\mathsf{hsf}_3, \mathsf{hsp}\colon \mathsf{hsf}\}$	$\{\mathsf{hsf}_3, \mathsf{hsp}\colon \mathsf{hsf}\}$	(ii), (vi)
7	$\emptyset$	$\{\mathsf{hsp}, \mathsf{hsf}\}$	$\{\mathsf{hsp}, \mathsf{hsf}\}$	(i), (v)

5. A second reaction system model for heat shock response

We introduce in this section a second RS model for the heat shock response. Our strategy this time is completely different than in the last section: we will formulate a number of essential properties of the heat shock response model and build the RS model to satisfy them. These properties will be achieved in the RS model *only through the mechanism of inhibition*, whereas they are emerging in the ODE model through a 'game of numbers', i.e., through the numerical values of the kinetic rate constants.

We introduce two new resources, nostress and stress, to model the system in the absence and the presence of the heat shock, resp., mirroring the behavior of the ODE model for temperature values of $37°C$ and $42°C$, resp.

We build the model so that the following properties hold in any state W of the reaction system $\mathcal{A}$, where either $\mathsf{stress} \in W$, or $\mathsf{nostress} \in W$, but not both:

P1. *mass-conservation of* hse: if $\{\mathsf{hse}, \mathsf{hsf_3}\colon \mathsf{hse}\} \cap W \neq \emptyset$, then $\{\mathsf{hse}, \mathsf{hsf_3}\colon \mathsf{hse}\} \cap \mathit{res}_{\mathcal{A}}(W) \neq \emptyset$;

P2. *a single form of* hse: if $\{\mathsf{hse}, \mathsf{hsf_3}\colon \mathsf{hse}\} \not\subseteq W$, then $\{\mathsf{hse}, \mathsf{hsf_3}\colon \mathsf{hse}\} \not\subseteq \mathit{res}_{\mathcal{A}}(W)$;

P3. *mass-conservation of* prot: if $\mathsf{prot} \in W$, then $\mathsf{prot} \in \mathit{res}_{\mathcal{A}}(W)$;

P4. *misfolded proteins must be addressed*: if $\mathsf{mfp} \in W$, then $\{\mathsf{mfp}, \mathsf{hsp}\colon \mathsf{mfp}\} \cap \mathit{res}_{\mathcal{A}}(W) \neq \emptyset$;

P5. *a single form of* hsf: let $\mathsf{HSF} = \{\mathsf{hsf}, \mathsf{hsf_3}, \mathsf{hsf_3}\colon \mathsf{hse}, \mathsf{hsp}\colon \mathsf{hsf}\}$; if $|W \cap \mathsf{HSF}| \leq 1$, then $|\mathit{res}_{\mathcal{A}}(W) \cap \mathsf{HSF}| \leq 1$;

P6. *stability of* hsp: hsf *in the absence of stress*: if $\{\mathsf{nostress}, \mathsf{hsp}\colon \mathsf{hsf}\} \subseteq W$, then $\mathsf{hsp}\colon \mathsf{hsf} \in \mathit{res}_{\mathcal{A}}(W)$.

Our main challenge is in building an RS model that captures qualitatively a behavior driven by numerical competition on resources, *in the absence* of an explicit mechanism for concurrency. Since our model consists of only unary and binary reactions, our main observation is that we can capture the competition between two binary molecular reactions using resources $\{A, B_1\}$ and $\{A, B_2\}$, resp., in terms of a preference (*binding affinity*) of A over, say, B_1, rather than B_2. We can formulate these relationships in the form of a *dominance graph*, where the graph nodes represent the molecular reactions (we indicate by '+' the left-to-right direction of a reversible reaction and by '-' the reverse direction) and a directed edge $u \to v$ indicates that u, v compete on a common resource and u is favoured over v.

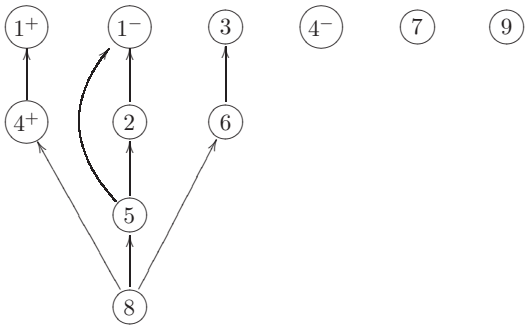


Figure 1. The reaction dominance graph of the simplified heat shock response model. The node labels refer to the molecular reactions in Table 2. We indicate with a directed edge $u \to v$ the property that u is favoured over v.

We build the dominance graph in Figure 1 for the molecular model in Table 2 based on the following assumptions:

- hsf has a higher affinity for hsp than for another hsf (edge $4^+ \to 1^+$);
- hsf_3 has a higher affinity to bind to hsp, than to break into hsf monomers or to interact with hse (edges $5 \to 1^-$ and $5 \to 2$);
- hsf_3 has a higher affinity to bind to hse, than to break into hsf monomers (edge $2 \to 1^-$);
- hsf_3: hse has a higher affinity to interact with hsp than to promote gene transcription (edge $6 \to 3$);
- hsp has a higher affinity for mfp than for hsf, hsf_3, hsf_3: hse (edges $8 \to 4^+$, $8 \to 5$, $8 \to 6$).

All these assumptions correspond to numerical observations regarding the reaction rates of the ODE model in [6].

5.1. Building the second model

We discuss now the construction of the RS model going through the reactions of the simplified molecular model in Table 2 one by one. The corresponding RS reactions are in Table 6. In the following we extend the terminology of 'enabled reactions' to the molecular model in Table 2; we will say that a molecular reaction is enabled in the current state W of our RS system, if all its reactants are in W.

Molecular reaction (1^+) is first modelled through RS reaction (10), where hsp is indicated as an inhibitor, as suggested by the edge $4^+ \to 1^+$. Note however based on Figure 1 that reaction (1^+) is enabled also when (8) is enabled, since in this case (4^+) is disabled. In other words, (1^+) is enabled in the presence of hsp, mfp, with no inhibitor. This leads to formulating RS reaction (11).

Molecular reaction (1^-) is first modelled through RS reaction (12), where hse, hsp are indicated as inhibitors, as suggested by edges $5 \to 1^-$ and $2 \to 1^-$. Similarly, as above, we note based on Figure 1 that if (8) is enabled, i.e., if hsp and mfp are in the current state, then (5) is disabled and so, in this case only reaction (2) supersedes (1^-). This leads to formulating RS reaction (13).

Molecular reaction (2) is modelled through RS reactions (14) and (15). The reasoning in this case is similar to that corresponding to (1^+). Additionally however, we need to introduce several other RS reactions to make sure that, in case (14) and (15) are disabled in the current state, hse is not lost, as required by property **P1**. In other words, this is analogous to building the RS correspondent of a molecular reaction hse $\to$ hse that in the dominance graph would have an incoming edge from node 2. With a similar reasoning as above, this leads to adding RS reaction (16): hse is preserved unless hsf_3 is present, when (2) is enabled. Moreover, based on the graph in Figure 1, we note that if (5) is enabled, i.e., if hsf_3 and hsp were in the current state, but not mfp, then (2) is disabled, i.e., the molecular reaction hse $\to$ hse would be enabled. This leads to introducing RS reaction (17).

Molecular reaction (3) is modelled through RS reactions (18) and (19). The reasoning in this case is similar to that corresponding to (1^+).

Modeling molecular reactions (4^+), (5) and (6) are all similar. They are modelled through RS reactions (20), (23), and (24), resp. We only note that when writing (23) we took into account **P5** and excluded hsf from the result of the RS reaction; this is in slight disagreement with the molecular reaction (5) where hsf is a product of the reaction. Applying **P5** we decided to keep in the set of products only one form of hsf and we favoured hsp: hsf over hsf, since it is the more common form of hsf in the ODE model.

To model the molecular reaction (4^-) we took into account **P6** and formulated it through RS reactions (21) and (22), depending on whether stress or nostress are in the current state.

Molecular reaction (7) is modelled through RS reactions (25) and (26), where we also took into account **P3**.

Molecular reaction (8) is modelled through RS reaction (27) and (28), where we also took into account **P4**.

Finally, molecular reaction (9) is modelled in a straightforward way through RS reaction (29). The complete list of reactions of the RS model for heat shock response is given in Table 6.

Table 6. The list of reactions of the second reaction system model for heat shock response.

Reaction		Reaction	
$(\{\mathsf{hsf}\}, \{\mathsf{hsp}\}, \{\mathsf{hsf}_3\})$	(10)	$(\{\mathsf{hsp}, \mathsf{hsf}\}, \{\mathsf{mfp}\}, \{\mathsf{hsp}\colon \mathsf{hsf}\})$	(20)
$(\{\mathsf{hsf}, \mathsf{hsp}, \mathsf{mfp}\}, \{\mathsf{d_I}\}, \{\mathsf{hsf}_3\})$	(11)	$(\{\mathsf{hsp}\colon \mathsf{hsf}, \mathsf{stress}\}, \{\mathsf{nostress}\}, \{\mathsf{hsp}, \mathsf{hsf}\})$	(21)
$(\{\mathsf{hsf}_3\}, \{\mathsf{hse}, \mathsf{hsp}\}, \{\mathsf{hsf}\})$	(12)	$(\{\mathsf{hsp}\colon \mathsf{hsf}, \mathsf{nostress}\}, \{\mathsf{stress}\}, \{\mathsf{hsp}\colon \mathsf{hsf}\})$	(22)
$(\{\mathsf{hsf}_3, \mathsf{hsp}, \mathsf{mfp}\}, \{\mathsf{hse}\}, \{\mathsf{hsf}\})$	(13)	$(\{\mathsf{hsp}, \mathsf{hsf}_3\}, \{\mathsf{mfp}\}, \{\mathsf{hsp}\colon \mathsf{hsf}\})$	(23)
$(\{\mathsf{hsf}_3, \mathsf{hse}\}, \{\mathsf{hsp}\}, \{\mathsf{hsf}_3\colon \mathsf{hse}\})$	(14)	$(\{\mathsf{hsp}, \mathsf{hsf}_3\colon \mathsf{hse}\}, \{\mathsf{mfp}\}, \{\mathsf{hsp}\colon \mathsf{hsf}, \mathsf{hse}\})$	(24)
$(\{\mathsf{hsf}_3, \mathsf{hse}, \mathsf{hsp}, \mathsf{mfp}\}, \{\mathsf{d_I}\}, \{\mathsf{hsf}_3\colon \mathsf{hse}\})$	(15)	$(\{\mathsf{prot}, \mathsf{stress}\}, \{\mathsf{nostress}\}, \{\mathsf{prot}, \mathsf{mfp}\})$	(25)
$(\{\mathsf{hse}\}, \{\mathsf{hsf}_3\}, \{\mathsf{hse}\})$	(16)	$(\{\mathsf{prot}, \mathsf{nostress}\}, \{\mathsf{stress}\}, \{\mathsf{prot}\})$	(26)
$(\{\mathsf{hse}, \mathsf{hsf}_3, \mathsf{hsp}\}, \{\mathsf{mfp}\}, \{\mathsf{hse}\})$	(17)	$(\{\mathsf{hsp}, \mathsf{mfp}\}, \{\mathsf{d_I}\}, \{\mathsf{hsp}\colon \mathsf{mfp}\})$	(27)
$(\{\mathsf{hsf}_3\colon \mathsf{hse}\}, \{\mathsf{hsp}\}, \{\mathsf{hsf}_3\colon \mathsf{hse}, \mathsf{hsp}\})$	(18)	$(\{\mathsf{mfp}\}, \{\mathsf{hsp}\}, \{\mathsf{mfp}\})$	(28)
$(\{\mathsf{hsf}_3\colon \mathsf{hse}, \mathsf{hsp}, \mathsf{mfp}\}, \{\mathsf{d_I}\}, \{\mathsf{hsf}_3\colon \mathsf{hse}, \mathsf{hsp}\})$	(19)	$(\{\mathsf{hsp}\colon \mathsf{mfp}\}, \{\mathsf{d_I}\}, \{\mathsf{hsp}, \mathsf{prot}\})$	(29)

Our following result shows that the resulting model satisfies properties **P1-P6**.

Theorem 5.1. The reaction system in Table 6 satisfies properties **P1-P6**.

Proof:
To prove **P1**, note that if $\mathsf{hse} \in W$, then either (14), (15), (16), or (17) are enabled, all leading to a state containing hse or $\mathsf{hsf}_3\colon \mathsf{hse}$. If $\mathsf{hsf}_3\colon \mathsf{hse} \in W$, then the same argument holds, noting that either (18), or (19), or (24) are enabled.

To prove **P2** we first observe that there is no reaction with both hse and $\mathsf{hsf}_3\colon \mathsf{hse}$ in its list of products. To prove that the model satisfies the property it is enough to show that no two reactions, one having hse, the other having $\mathsf{hsf}_3\colon \mathsf{hse}$ in its list of products, are enabled simultaneously. The reactions having hse in their list of products are (16), (17) and (24); those having $\mathsf{hsf}_3\colon \mathsf{hse}$ in their list of products are (14), (15), (18), and (19). Reaction (16) cannot be enabled simultaneously with either (14) or (15) because hsf_3 is an inhibitor for (16) and it is a reactant for the others. Also, if (16) were enabled simultaneously with (18) or (19), then $\mathsf{hse}, \mathsf{hsf}_3\colon \mathsf{hse} \in W$, a contradiction with the hypothesis of **P2**. Similar arguments show that (17) and (24) cannot be enabled simultaneously with (14), (15), (18), and (19).

To prove **P3** note that the only reactions involving prot are (25) and (26), that exactly one of them is triggered if stress $\in W$ or nostress $\in W$ but not both, and that both have prot in their list of products.

To prove **P4**, it is enough to observe that if $\mathsf{mfp}, \mathsf{hsp} \in W$, then (27) is enabled, while if $\mathsf{mfp} \in W$ and $\mathsf{hsp} \notin W$, then (28) is enabled.

We prove now **P5**. First, it is easy to see that if $W \cap \mathsf{HSF} = \emptyset$, then $res_{\mathcal{A}}(W) \cap \mathsf{HSF} = \emptyset$. Second, note that there is no reaction with more than one element of HSF in its set of products. If $\mathsf{hsf} \in W$, then the reactions that may be enabled are (10), (11), (20); no two of them can be enabled simultaneously. If $\mathsf{hsf}_3 \in W$, then the reactions that may be enabled are (12), (13), (14), (15), (17), and (23); of these, only (17) and (23) can be enabled simultaneously, but the products of (17) do not include any element from HSF. If $\mathsf{hsf}_3\colon \mathsf{hse} \in W$, then the reactions that may be enabled are (18), (19), and (24); no two of them can be enabled simultaneously. If $\mathsf{hsp}\colon \mathsf{hsf} \in W$, then the reactions that may be enabled are (21) and (22), which cannot be enabled simultaneously.

Property **P6** follows from reaction (22). □

5.2. Interactive processes in the second model

We analyze here the interactive processes of the second RS model of the heat shock response and compare it with the results previously attained in the ODE model of [6]. Taking into account the qualitative nature of our model, by the presence of a resource in the environment, except for stress and nostress, conforming to the *threshold assumption*, we assume that there is a sufficient amount of the respective resource in the environment.

Similarly as in the case of our first reaction system model, we start our first interactive process from an initial state consisting of a minimal set of species needed in the heat shock response model: hsf, hse, and prot. To draw a parallel with the numerical simulations of the ODE model at $37°C$, the subsequent contexts of our reaction system consists of the resource nostress. The result is shown in Table 7.

Table 7. An interactive process of the second RS model for heat shock response, in the absence of stress. The first column of the table represents the state of the system, C_i is the context given to the system in state i, $D_i = res_{\mathcal{A}}(C_{i-1} \cup D_{i-1})$ and $W_i = C_i \cup D_i$. The last column provides the list of the reactions triggered in each state.

State	C_i	D_i	W_i	r_i
0	{hsf, prot, hse, nostress}	$\emptyset$	{hsf, prot, hse, nostress}	(10), (16), (26)
1	{nostress}	{hsf_3, prot, hse}	{hsf_3, prot, hse, nostress}	(14), (26)
2	{nostress}	{hsf_3: hse, prot}	{hsf_3: hse, prot, nostress}	(18), (26)
3	{nostress}	{hsp, hsf_3: hse, prot}	{hsp, hsf_3: hse, prot, nostress}	(24), (26)
4	{nostress}	{hsp: hsf, hse, prot}	{hsp: hsf, hse, prot, nostress}	(16), (22), (26)
5	{nostress}	{hsp: hsf, hse, prot}	{hsp: hsf, hse, prot, nostress}	(16), (22), (26)

The result shows that the reaction system in this case enters into a steady state, that is similar to the steady state of the ODE model in the absence of stress.

Our next interactive process follows the behavior of the RS model when the context introduces stress in every state, corresponding to the situation when the temperature is set to $42°C$ in the ODE model. As

the initial state we take the steady state achieved in the previous interactive process: $\{\mathsf{hse}, \mathsf{prot}, \mathsf{hsp}\colon \mathsf{hsf}\}$. The result is shown in Table 8. We note that also in this case the system is reaching a steady state, similar to the steady state of the ODE model for a temperature of $42°C$.

Table 8. An interactive process of the second reaction system model for heat shock response at $42°C$.

State	C_i	D_i	W_i	r_i
0	$\{\mathsf{hse}, \mathsf{prot}, \mathsf{hsp}\colon \mathsf{hsf}, \mathsf{stress}\}$	$\emptyset$	$\{\mathsf{hse}, \mathsf{prot}, \mathsf{hsp}\colon \mathsf{hsf}, \mathsf{stress}\}$	(16),(21),(25)
1	$\{\mathsf{stress}\}$	$\{\mathsf{hse}, \mathsf{hsp}, \mathsf{hsf}, \mathsf{prot}, \mathsf{mfp}\}$	$\{\mathsf{hse}, \mathsf{hsp}, \mathsf{hsf}, \mathsf{prot}, \mathsf{mfp}, \mathsf{stress}\}$	(11),(16),(25),(27)
2	$\{\mathsf{stress}\}$	$\{\mathsf{prot}, \mathsf{mfp}, \mathsf{hsp}\colon \mathsf{mfp}, \mathsf{hsf}_3, \mathsf{hse}\}$	$\{\mathsf{prot}, \mathsf{mfp}, \mathsf{hsp}\colon \mathsf{mfp}, \mathsf{hsf}_3, \mathsf{hse}, \mathsf{stress}\}$	(14),(25),(28),(29)
3	$\{\mathsf{stress}\}$	$\{\mathsf{hsp}, \mathsf{prot}, \mathsf{mfp}, \mathsf{hsf}_3\colon \mathsf{hse}\}$	$\{\mathsf{hsp}, \mathsf{prot}, \mathsf{mfp}, \mathsf{hsf}_3\colon \mathsf{hse}, \mathsf{stress}\}$	(19),(25),(27)
4	$\{\mathsf{stress}\}$	$\{\mathsf{hsp}, \mathsf{prot}, \mathsf{mfp}, \mathsf{hsf}_3\colon \mathsf{hse}, \mathsf{hsp}\colon \mathsf{mfp}\}$	$\{\mathsf{hsp}, \mathsf{prot}, \mathsf{mfp}, \mathsf{hsf}_3\colon \mathsf{hse}, \mathsf{hsp}\colon \mathsf{mfp}, \mathsf{stress}\}$	(19),(25),(27),(29)
5	$\{\mathsf{stress}\}$	$\{\mathsf{hsp}, \mathsf{prot}, \mathsf{mfp}, \mathsf{hsf}_3\colon \mathsf{hse}, \mathsf{hsp}\colon \mathsf{mfp}\}$	$\{\mathsf{hsp}, \mathsf{prot}, \mathsf{mfp}, \mathsf{hsf}_3\colon \mathsf{hse}, \mathsf{hsp}\colon \mathsf{mfp}, \mathsf{stress}\}$	(19),(25),(27),(29)

In our final interactive process we start from the steady state achieved in the previous one and consider the case when the context consists in all subsequent state of only nostress. This corresponds to a case when the ODE model is stabilized at $42°C$, followed then by a temperature of $37°C$. The result is shown in Table 9. The model reaches again a steady state, the same as that reached in the first interactive process. The situation is similar to that of the ODE model, where upon removal of the stress, the model eventually returns to its basal physiological values.

Table 9. Interactive process for the recovery (at $37°C$) of the second reaction system model after several steps of heat shock (at $42°C$). The process starts from the steady state obtained in Table 8.

State	C_i	D_i	W_i	r_i
0	$\{\mathsf{hsp}, \mathsf{prot}, \mathsf{hsf}_3\colon \mathsf{hse}, \mathsf{mfp}, \mathsf{hsp}\colon \mathsf{mfp}, \mathsf{nostress}\}$	$\emptyset$	$\{\mathsf{hsp}, \mathsf{prot}, \mathsf{hsf}_3\colon \mathsf{hse}, \mathsf{mfp}, \mathsf{hsp}\colon \mathsf{mfp}, \mathsf{nostress}\}$	(19),(26),(27), (29)
1	$\{\mathsf{nostress}\}$	$\{\mathsf{hsp}, \mathsf{prot}, \mathsf{hsp}\colon \mathsf{mfp}, \mathsf{hsf}_3\colon \mathsf{hse}\}$	$\{\mathsf{hsp}, \mathsf{prot}, \mathsf{hsp}\colon \mathsf{mfp}, \mathsf{hsf}_3\colon \mathsf{hse}, \mathsf{nostress}\}$	(24),(26),(29)
2	$\{\mathsf{nostress}\}$	$\{\mathsf{hse}, \mathsf{hsp}, \mathsf{hsp}\colon \mathsf{hsf}, \mathsf{prot}\}$	$\{\mathsf{hse}, \mathsf{hsp}, \mathsf{hsp}\colon \mathsf{hsf}, \mathsf{prot}, \mathsf{nostress}\}$	(16),(22),(26)
3	$\{\mathsf{nostress}\}$	$\{\mathsf{hse}, \mathsf{prot}, \mathsf{hsp}\colon \mathsf{hsf}\}$	$\{\mathsf{hse}, \mathsf{prot}, \mathsf{hsp}\colon \mathsf{hsf}, \mathsf{nostress}\}$	(16),(22),(26)
4	$\{\mathsf{nostress}\}$	$\{\mathsf{hse}, \mathsf{prot}, \mathsf{hsp}\colon \mathsf{hsf}\}$	$\{\mathsf{hse}, \mathsf{prot}, \mathsf{hsp}\colon \mathsf{hsf}, \mathsf{nostress}\}$	(16),(22),(26)

6. Discussion

We demonstrated in this paper the ability of the reaction system framework to capture the intricate behaviour of ODE-based models. Given the qualitative nature of reaction systems and their unusual two major assumptions, the non-permanency and the threshold assumptions, the result bridges two fundamentally different modelling paradigms. As a case-study we chose the heat shock response. We focused on emulating the behavior of the corresponding mass-action ODE model for the heat shock response. Whereas the ODE model is driven by the numerical values of its kinetic constants, the reaction system framework only allows the specification of logical dependencies in terms of facilitators and inhibitors. Moreover, its two fundamental principles, the non-permanency of resources and the uncompetitive triggering of reactions, make the reaction system framework fundamentally different than that of the ODE-based modelling.

Our first approach, where each molecular reaction would be directly translated to a reaction system with no inhibitors failed to reproduce the biological knowledge about the heat shock response and the behavior of the ODE model. The *non-permanency principle* of the reaction systems has a major role in the behavior of the reaction system being different than that of the ODE model. For example, the heat shock element hse is lost in an interactive process in any state where reaction (iii) is not triggered; this is in disagreement with the common intuition that a gene promoter is not lost when there is no gene binding activity; it is also in disagreement with the ODE model, where the total amount of heat shock element, either in the form of hse , or as $\mathsf{hsf}_3\colon\mathsf{hse}$, is conserved. The solution here, that we implement in our second reaction system model, is to have several rules making sure that hse is involved in at least one reaction regardless of the context, and thus preserved throughout the interactive process. The *threshold assumption* of the reaction systems also plays a major role in the disagreement between the models. This is seen, e.g., in the treatment of proteins in the two models. While the ODE model exhibits a *gradual* misfolding (and refolding) of prot, in the RS model *all* proteins (prot) are converted into mfp's in a single step. Another weakness of the RS model obtained through the direct translation is that it does not take into consideration the main driving factor of the heat shock response: the temperature.

Our second RS model was successful in emulating the behavior of the ODE model. The key observation here was that in building our first RS model we lost the information about all kinetic constants of the model and did not compensate for it in the RS model. In building our second model we developed an approach where we capture the affinity of a species for another species through carefully selected inhibitors rather than kinetic constants as in the ODE model. We also focused on a number of properties, including mass conservation, that the RS model should satisfy. Instrumental here was the notion of dominance graph that aims to translate numerical observations about the reaction rates of the ODE model into a binary relation on the set of RS reactions. The competition between two reactions on the same resource is captured by an ODE-based model through the reaction rates: the higher the rate of a reaction, the more it will consume the competed resource; note however that all reactions, even those with the lowest rate, will get some part of the resource and will be able to run. Going towards the qualitative realm of reaction systems, we translated the competition between two reactions in terms of a directed graph where an edge $i \to j$ represents that reactions i, j compete on a common resource and that in the ODE model, reaction i has a higher rate than reaction j. This is then translated into a reaction system where, through the use of suitable inhibitors, the RS reaction corresponding to j cannot be triggered if the RS reaction corresponding to i is enabled. This then helps to capture cascading effects: if reaction

i dominates reaction j and j dominates k, then having i enabled effectively enables k; this is similar to the behaviour of ODE-based models.

Our approach based on dominance graphs might be possible to generalize to arbitrary molecular models consisting of unary and binary molecular reactions. The main problem here is the algorithmic construction of the dominance graph starting from a molecular model, its kinetic rate constants and the initial values of all variables. How much a-priori knowledge about the behaviour of the model is needed, as well as how well the procedure scales up with the size of the model appear as interesting questions, worthy of further attention. A particularly interesting question is how, if at all possible, can one capture a dynamic situation where the dominance relations between reactions changes in time, as it may well be possible in ODE models.

The reaction system framework forces the modeler to make explicit a number of assumptions about the model, that in other framework are typically hidden in some numerical values. Moreover, the explicit list of inhibitors is shedding light into the causality relations between the various reactions in the system. This kind of insight is highly valuable and may be very difficult to obtain through other frameworks, either from the presentation of the model, or from numerical simulations.

Acknowledgments

We thank Daniela Besozzi for reading through and commenting on the manuscript. We acknowledge the detailed comments from the two referees, that have helped to improve the presentation of the paper.

References

[1] R. Brijder, A. Ehrenfeucht, M. Main, and G. Rozenberg. A tour of reaction systems. *International Journal of Foundations of Computer Science*, 22(07):1499–1517, 2011.

[2] A. Ehrenfeucht and G. Rozenberg. Reaction systems. *Fundamenta Informaticae*, 75(1):263–280, 2007.

[3] F.G. Helfferich. *Kinetics of multistep reactions*, volume 40. Elsevier Science, 2004.

[4] E. Klipp, R. Herwig, A. Kowald, C. Wierling, and H. Lehrach. *Systems biology in practice: concepts, implementation and application*. Wiley-Vch, 2005.

[5] D. L. Nelson and M. M. Cox. Lehninger principles of biochemistry. *Worth Publishers*, 2000.

[6] I. Petre, A. Mizera, C.L. Hyder, A. Meinander, A. Mikhailov, R.I. Morimoto, L. Sistonen, J.E. Eriksson, and R. Back. A simple mass-action model for the eukaryotic heat shock response and its mathematical validation. *Natural Computing*, 10(1):595–612, 2011.

[7] T.R. Rieger, R.I. Morimoto, and V. Hatzimanikatis. Mathematical modeling of the eukaryotic heat-shock response: dynamics of the hsp70 promoter. *Biophysical journal*, 88(3):1646–1658, 2005.

[8] G. Rozenberg, A. Ehrenfeucht, and M. Main. Combinatorics of life and death for reaction systems. *International Journal of Foundations of Computer Science*, 21(3):345–356, 2010.

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

Turku Centre *for* Computer Science

Joukahaisenkatu 3-5 B, 20520 Turku, Finland | www. tucs.fi

University of Turku
Faculty of Mathematics and Natural Sciences
- Department of Information Technology
- 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 978-952-12-3226-8
ISSN 1239-1883

Bogdan Iancu

Quantitative Refinement of Reaction-Based Biomodels

Bogdan Iancu

Quantitative Refinement of Reaction-Based Biomodels