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

Turku Centre *for* Computer Science

TUCS Dissertations
No 138, August 2011

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

Andrzej Mizera

To be presented, with the permission of the Division for Natural Sciences and Technology of Åbo Akademi University, for public criticism in Auditorium Gamma on August 26, 2011, at 12 noon.

Åbo Akademi University
Department of Information Technologies
Joukahaisenkatu 3-5A, FIN-20520 Turku, Finland

2011

Supervisors

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

Dr hab. Barbara Gambin
Department of Ultrasound
Institute of Fundamental Technological Research
Polish Academy of Sciences
Pawińskiego 5B, 02-106 Warsaw
Poland

Reviewers

Prof. Erik de Vink
Department of Mathematics and Computer Science
Technische Universiteit Eindhoven
Den Dolech 2, 5600 MB Eindhoven
The Netherlands

Prof. Corrado Priami
Centre for Computational and Systems Biology
The Microsoft Research - University of Trento
Piazza Manci 17, 38123 Povo (Trento)
Italy

Opponent

Prof. Erik de Vink
Department of Mathematics and Computer Science
Technische Universiteit Eindhoven
Den Dolech 2, 5600 MB Eindhoven
The Netherlands

ISBN 978-952-12-2616-8
ISSN 1239-1883

Abstract

Systems biology is a new, emerging and rapidly developing, multidisciplinary research field that aims to study biochemical and biological systems from a holistic perspective, with the goal of providing a comprehensive, system-level understanding of cellular behaviour. In this way, it addresses one of the greatest challenges faced by contemporary biology, which is to comprehend the function of complex biological systems. Systems biology combines various methods that originate from scientific disciplines such as molecular biology, chemistry, engineering sciences, mathematics, computer science and systems theory. Systems biology, unlike "traditional" biology, focuses on high-level concepts such as: network, component, robustness, efficiency, control, regulation, hierarchical design, synchronization, concurrency, and many others. The very terminology of systems biology is "foreign" to "traditional" biology, marks its drastic shift in the research paradigm and it indicates close linkage of systems biology to computer science.

One of the basic tools utilized in systems biology is the mathematical modelling of life processes tightly linked to experimental practice. The studies contained in this thesis revolve around a number of challenges commonly encountered in the computational modelling in systems biology. The research comprises of the development and application of a broad range of methods originating in the fields of computer science and mathematics for construction and analysis of computational models in systems biology. In particular, the performed research is setup in the context of two biological phenomena chosen as modelling case studies: 1) the eukaryotic heat shock response and 2) the *in vitro* self-assembly of intermediate filaments, one of the main constituents of the cytoskeleton. The range of presented approaches spans from heuristic, through numerical and statistical to analytical methods applied in the effort to formally describe and analyse the two biological processes. We notice however, that although applied to certain case studies, the presented methods are not limited to them and can be utilized in the analysis of other biological mechanisms as well as complex systems in general. The full range of developed and applied modelling techniques as well as model analysis methodologies constitutes a rich modelling framework. Moreover, the presentation of the developed methods,

their application to the two case studies and the discussions concerning their potentials and limitations point to the difficulties and challenges one encounters in computational modelling of biological systems. The problems of model identifiability, model comparison, model refinement, model integration and extension, choice of the proper modelling framework and level of abstraction, or the choice of the proper scope of the model run through this thesis.

Sammanfattning

(abstract in Swedish)

Systembiologi är ett nytt, emergent och snabbt växande, tvärvetenskaplig forskningsområde som fokuserar på systematiskt studium av biokemiska och biologiska system ur ett heltäckande perspektiv, med syftet att uppnå allsidig förståelse av cellulära beteenden på systemnivå. På detta sätt angriper systembiologin en av de största utmaningarna som modern biologi står inför, dvs. att förstå funktionen hos komplexa biologiska system. Systembiologin sammanfogar olika metoder vilka har sitt ursprung i vetenskapliga discipliner, såsom molekylärbiologi, kemi, ingenjörsvetenskap, matematik, datavetenskap och systemteori. Systembiologin, till skillnad från "traditionell" biologi, fokuserar på högnivåbegrepp såsom: nätverk, komponent, robusthet, effektivitet, kontroll, reglering, hierarkisk design, synkronisering, samverkan och många andra. Själva terminologin i systembiologin är "främmande" för "traditionell" biologi; den markerar en drastisk förändring i forskningsparadigmet och tyder på en nära koppling mellan systembiologin och datavetenskap.

Ett av de basala verktygen som används i systembiologin är matematisk modellering av livsprocesser i samband med experimentell forskning. De undersökningar som ingår i denna avhandling kretsar kring ett antal utmaningar som ofta förekommer i beräkningsmodellering inom systembiologin. Den presenterade forskningen består av utveckling och tillämpning av ett brett sortiment av metoder vilka har sitt ursprung i datavetenskap och matematik för konstruktion och analys av datormodeller i systembiologin. I synnerhet är den forskningen utförd i kontexten av två biologiska fenomen utvalda som fallstudier: 1) eukaryotiskt värme-chock respons och 2) *in vitro* självorganisering av intermediära filament, en av huvudbeståndsdelarna av cytoskelettet. Sortimentet av presenterade tillvägagångssätt sträcker sig från heuristiska metoder, via numeriska och statistiska metoder till analytiska metoder, tillämpade i strävan att formellt beskriva de två biologiska processerna. Vi konstaterar att de presenterade metoderna, fastän utnyttjade i de enskilda fallstudierna, inte är begränsade till dessa utan kan tillämpas vid analys av andra biologiska mekanismer samt komplexa system i allmänhet.

Hela sortimentet av utvecklade och tillämpade modelleringstekniker samt metodologier för modellanalys utgör ett rikt modelleringsramverk. Presentationen av de utvecklade metoderna, deras tillämpning i de två fallstudierna och diskussionerna om deras möjligheter och begränsningar, visar därtill på svårigheter och utmaningar man stöter på i beräkningsmodellering av biologiska system. Identifierbarhet av modeller, jämförelse mellan modeller, precisering, integration och utvidgning av modeller, urval av rätt modelleringsramverk och abstraktionsnivå samt urval av ett lämpligt omfång för en modell är problem som diskuteras genom hela denna avhandling.

Acknowledgements

My experience with scientific research started in 2005 when I was in the last stage of my master's degree studies at the Faculty of Mathematics, Informatics and Mechanics at the University of Warsaw. I was then encouraged by Dr hab. Anna Gambin to continue learning and broadening my knowledge as a PhD student. Following her advice, I enrolled in a PhD programme at the Institute of Fundamental Technological Research, Polish Academy of Sciences. The first one and a half years of my studies I spent on studying mechanics and getting to know the field of biomechanics under the supervision of Dr hab. Barbara Gambin. In the meantime, in September 2006 I got in touch with Prof. Ion Petre from Åbo Akademi University, Finland and decided to focus my research on computational modelling in systems biology. In consequence, I joined the *Computational Biomodelling Laboratory* at Åbo Akademi University in January 2007.

First and foremost, I want to express my tremendous and deep gratitude to my both supervisors. I am extremely thankful to my supervisor Prof. Ion Petre for the enormous, comprehensive and wise support I have obtained from him. I wish to thank him for all the time and effort he has dedicated to me, for always being available whenever I had a problem concerning both scientific as well as non-scientific matters, for the patience and understanding with which he treated me, for the great passion with which he has been guiding and advising me ever since we met for the first time, and the confidence that I could always rely on him. I am immensely grateful to my supervisor Dr hab. Barbara Gambin for her extensive, all-embracing and thoughtful assistance. I would especially like to thank her for all the time and energy she has devoted to me, for her constant readiness to support me in all aspects of my studies, for giving me the reassurance that I can turn to her for help with all kinds problems, as well as for the enthusiasm and commitment with which she has been assisting me during all the years of my PhD studies. It is in general difficult to encounter people of such personalities in general, so I feel extremely lucky that I was given the opportunity to work with them and have them as my supervisors. Thank you!

I am very thankful to Prof. Erik de Vink that he kindly agreed to be a reviewer of my dissertation and that he accepted to act as the opponent

at my doctoral defence. I am very grateful to Prof. Corrado Priami for accepting to be a reviewer of my doctoral thesis. I would like to thank both of them for all the time and effort they dedicated to a thorough review of my dissertation, their helpful, valuable and accurate remarks as well as encouraging comments.

I am especially grateful to Elena Czeizler and Eugen Czeizler for our many inspiring and fruitful scientific discussions as well as efficient and productive work which resulted in a number of publications. I am extremely thankful to Anna Gambin for her advice, many discussions, valuable comments and all the support I received from her during my stay in Poland during the academic year 2008/2009.

I would like to thank Prof. Jerzy Tiuryn from the Faculty of Mathematics, Informatics and Mechanics at the University of Warsaw, who gave me the first taste of what scientific research means as my master thesis supervisor. I would like to list and thank all the other members of his *Computational Biology Group* who advised and assisted me during the work on my master thesis as well as are the coauthors of my first scientific article: Norbert Dojer, Anna Gambin, and Bartosz Wilczyński.

I want to list here and express my gratitude to all my coauthors and people with whom I collaborated during my PhD studies. I am thankful to Ralph-Johan Back, Elena Czeizler, Eugen Czeizler, Norbert Dojer, John E. Eriksson, Małgorzata Figurska, Anna Gambin, Barbara Gambin, Claire L. Hyder, Eleonora Kruglenko, Tamara Kujawska, Annika Meinander, Andrey Mikhailov, Richard I. Morimoto, Andrzej Nowicki, Ion Petre, Diana Preoteasa, Lea Sistonen, Maciej Stańczyk, Jerzy Tiuryn, and Bartosz Wilczyński.

These acknowledgements would be incomplete without expressing my gratitude to each of the former and present members of the *Computational Biomodelling Laboratory* at Åbo Akademi University (Combio). I am particularly grateful to my friend Vladimir Rogojin, the first member of Combio whom I met and without whom my stay in Finland would certainly not be this rich and joyful. I would like to thank Artiom Alhazov, Sepinoud Azimi, Ralph-Johan Back, Elena Czeizler, Eugen Czeizler, Vladimir Grigor, Bogdan Iancu, Tseren-Onolt Ishdorj, Miika Langille, Chang Li, and Ion Petre for all the great time I spent with them at work as well as during our less formal meetings and activities.

I would like to thank Åbo Akademi University and the Turku Centre for Computer Science for the excellent conditions they provided me with as well as for the friendly atmosphere that I could enjoy during my doctoral studies. I am truly grateful to the administration of the Department of Information Technologies at the Åbo Akademi University and the administration of Turku Centre for Computer Science for their substantial assistance in all organizational matters.

I gratefully acknowledge the financial support of my doctoral studies and research provided to me by the Turku Centre for Computer Science, Academy of Finland (grant projects 108421, 122426, 129863, and 203667), Department of Information Technologies at Åbo Akademi University, Centre for International Mobility (grant project TM-07-5101), the Institute of Fundamental Technological Research of the Polish Academy of Sciences, and the Polish Ministry of Science and Education (project NN518426936).

I am thankful to Johannes Eriksson for checking and correcting the Swedish version of the abstract.

I would like to express my deep gratefulness to all my friends who motivated, supported and advised me. In particular, I would like to thank Sepinoud Azimi, Elena Czeizler, Eugen Czeizler, Bogdan Iancu, Chang Li, Ion Petre, Vladimir Rogojin, and Robert Stefaniuk. I would especially like to express my gratitude to Mariola, Weronika and Grzegorz Mazerski as well as Marta and Mikołaj Olszewski for all the help, support and understanding I received from them during the years of my stay in Finland.

No words can express my deepest thankfulness to my dear family, without whom this thesis would not have been written. My parents, Monika and Marek Mizera – thank you for always being with me, believing in me, for your absolute support, and for giving me the freedom to choose my own paths in life. My grandparents, Barbara and Henryk Tunia, my uncle, Andrzej Tunia (who was the first to introduce me to the incredible world of physics and mathematics), and my sister, Agnieszka Mizera: I would like to thank you for all the enormous love and support I got from you, for always trusting in me, for all your help and sincere advice.

Above all, I would like to thank my beloved wife, Ilona, for her love, patience and understanding, the extent of which cannot be verbalized. By the tremendous and absolute trust she placed in me, she was the only one who was able to give me the essential further motivation and additional strength to accomplish this work. As an expression of my gratitude, I wish to dedicate this work to Her.

Turku, June 2011 Andrzej Mizera

To my dear wife, Ilona

List of original publications

1. 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.

2. Ion Petre, Andrzej Mizera, Claire L. Hyder, Andrey Mikhailov, John E. Eriksson, Lea Sistonen, and Ralph-Johan Back. A new mathematical model for the heat shock response. In Anne Condon, David Harel, Joost N. Kok, Arto Salomaa, and Erik Winfree, editors, *Algorithmic Bioprocesses*, Natural Computing Series, pages 411-425. Springer, Dordrecht Heidelberg London New York, 2009.

3. Ion Petre, Andrzej Mizera, and Ralph-Johan Back. Computational heuristics for simplifying a biological model. In Klaus Ambos-Spies, Benedikt Löwe, and Wolfgang Merkle, editors, *Mathematical Theory and Computational Practice: 5th Conference on Computability in Europe, CiE 2009, Proceedings*, volume 5635 of *Lecture Notes in Computer Science*, pages 399–408, Berlin Heidelberg New York, 2009. Springer.

4. Andrzej Mizera and Barbara Gambin. Stochastic modelling of the eukaryotic heat shock response. *Journal of Theoretical Biology*, 265(3): 455–466, 2010.

5. Elena Czeizler, Andrzej Mizera, and Ion Petre. A Boolean approach for disentangling the numerical contribution of modules to the system-level behavior of a biomodel. *TUCS Technical Report number 997*, January 2011.

6. Andrzej Mizera, Elena Czeizler, and Ion Petre. Methods for biochemical model decomposition and quantitative submodel comparison. *Israel Journal of Chemistry*, 51(1):151–164, 2011.

7. Eugen Czeizler, Andrzej Mizera, Elena Czeizler, Ralph-Johan Back, John E. Eriksson, and Ion Petre. Quantitative analysis of the self-assembly strategies of intermediate filaments from tetrameric vimentin. *TUCS Technical Report number 963*, December 2009.

8. Andrzej Mizera, Eugen Czeizler, and Ion Petre. Self-assembly models of variable resolution. *TUCS Technical Report number 1014*, June 2011.

9. Andrzej Mizera and Barbara Gambin. Modelling of ultrasound therapeutic heating and numerical study of the dynamics of the induced heat shock response. *Communications in Nonlinear Science and Numerical Simulation*, 16(5):2342–2349, 2011.

Contents

Chapter 1

Introduction

Systems biology is an emerging research field that aims to study biochemical and biological systems from a holistic perspective, with the goal of providing a comprehensive, system-level understanding of cellular behaviour ([138, 137]). Biological processes have long been seen as static systems comprising a vast number of loosely linked, highly detailed, molecular devices ([16]). However, it has already been known for many years that biology is driven by dynamic processes. One of the greatest challenges faced by contemporary biology is to comprehend the function and malfunction of complex biological systems. This is directly related to the problem of profound understanding of what health and disease are. In order to meet this challenge such key issues in systems biology as dynamic processes, interdependent regulatory controls, and the operation of multiple interacting components should be addressed ([16]). Systems biology is a highly multi-disciplinary research area that combines science subjects such as biology, physiology, chemistry, mathematics, computer science, physics, engineering in the effort to investigate interrelationships and interactions of genes, proteins and metabolites. The large interest in topics associated with systems biology by researchers from such a vast range of fields of expertise results in various views on what systems biology is. In consequence, there is no well-established consensus definition of systems biology ([16, 103]). Notions that appear in virtually all definitions are "networks", "computation", "modelling", and "dynamic properties" ([16]). In this thesis, we adopt the definition proposed in [16]:

> "... the objective of systems biology is defined as the understanding of network behaviour, and in particular their dynamic aspects, which requires the utilization of mathematical modelling tightly linked to experiment."

It involves identification, modelling and analysis of biochemical networks, e.g. metabolic pathways, regulatory and signal transduction networks, in close linkage to experiment with the focus on understanding the system's

structure and dynamics. This comprehensive approach enables the capturing of complex properties of a system such as robustness, emergence or adaptation, which are commonly observed in natural systems ([105, 63]). These features are not attributes of certain elements of the system, but rather emerge as a result of the interactions and relationships between the building blocks. In other words, biological systems, e.g. cells, are complex structures of interdependent components whose properties and relationships are determined by their function in the whole ([136]). After [105], systems biology essentially advocates a departure from the reductionist viewpoint, while putting emphasis on the holistic approach towards the analysis of a biological system.

Systems biology, unlike "traditional" biology focuses on high-level concept such as: *network*, *component*, *robustness*, *efficiency*, *control*, *regulation*, *hierarchical design*, *signalling*, *synchronization*, *parallelism*, *competition*, and many others. The very terminology of systems biology is "foreign" to "traditional" biology and it marks its drastic shift in the research paradigm.

Computer science focuses on the study of the scientific foundations for information, computation, and communication, and on the practical techniques for implementing them in computer systems. This is a very broad area of science spanning from the theory of computing, through programming, to cutting-edge development of computing solutions for large distributed systems. The research in computer science typically abstracts from the physical implementation of computing and it rather focuses on high-level concepts such as: *algorithmics*, *computability*, *network*, *component*, *robustness*, *efficiency*, *control*, *regulation*, *hierarchical design*, *signalling*, *synchronization*, *parallelism*, *competition*, and many others.

It is not surprising to see the prominent role that computer science plays in the field of systems biology. A main reason for this is that, as seen also in our definitions above, the key concepts in systems biology have been studied for a long time already in computer science (albeit from different perspectives). A key contribution that computer science brings to systems biology is the ability to manipulate, analyse, and reason about such system-level concepts and structures. For example, mathematical modelling, formal system specifications, control design, and others are by now mainstream techniques in systems biology. Computer science also brings to systems biology research numerical techniques such as modelling and simulation, qualitative and quantitative predictions, sensitivity analysis, model fit and validation, steady state analysis, flux balance analysis, etc.

The doctoral research presented in this thesis concerns a number of challenges of computational modelling in systems biology. It is focused on development and utilization of different methodologies having their origins in the fields of computer science and mathematics. The following list briefly describes the considered problems.

- We address issues related to model construction methodologies such as parameter estimation, model validation, and we discuss the problem of model identifiability.
- We present existing techniques and develop new ones for the problem of model comparison. In particular, we concentrate on the case where the comparison is performed between submodels of a model. In this context we discuss methodologies for model decomposition.
- Further, we address the problem of model modifications. We present various techniques and heuristics useful for applying simplifications or extensions to an already fitted and validated mathematical model in such a way that the desired properties of the model are retained. In particular, we develop the following.
  - A generic model for the process of self-assembly is proposed.
  - The notion of self-assembly model resolution is formally introduced.
  - Both numerical as well as analytical methods for decreasing and increasing the resolution of ordinary differential equations (ODE) models of self-assembly are developed. To the best of our knowledge, this is the first time that formal model refinement is considered in relation to computational ODE-based models.

The presented methodologies are applied in the modelling of two biological processes chosen as modelling case studies:

1. the heat shock response mechanism in eukaryotic cells and
2. the *in vitro* self-assembly of intermediate filaments from tetrameric vimentin.

However, we notice that the developed methodologies are general in nature and can be applied for the modelling of other biological processes as well.

We discuss the above issues in the subsequent chapters. In Chapter 2, we present some generalities concerning modelling in systems biology. The choice of the covered generalities is made so as to provide preliminaries to the theory and techniques used in our original research papers included in this thesis. In Chapter 3, we describe the two case studies considered in this thesis, i.e. the eukaryotic heat shock response and the *in vitro* self-assembly of intermediate filaments. In Chapter 4, we present in a synthetic way a number of issues being subject of the publications included in this thesis and constituting the original contribution of the author. In Chapter 5, we list the original research contribution of each particular publication. Finally, we end with conclusions and perspectives for further research in Chapter 6.

Chapter 2

Computational modelling in systems biology: generalities

2.1 The iterative cycle of systems biology

The great complexity of biological systems enforces the need for representing them in formal models in order to investigate them and to make specific predictions about their behaviour, that can be tested in subsequent experiments. It is difficult, if not impossible, to completely understand such systems based only on intuition and experimental observations. Mathematics is necessary to comprehend and reason about large systems involving many components. Used in an appropriate way, mathematical models can represent such systems in a biologically realistic manner, incorporate a wide range of empirical observations, and provide basis for formulating novel hypotheses. One of the main characteristics of systems biology is the close connection of experiment with theory. As schematically illustrated in Figure 2.1, starting from a model abstracting a biological system, the iterative process of hypothesis generation, experimental design, experimental analysis, and model refinement lies at the core of systems biology ([13, 105, 63, 4]). Even more, this iterative cycle approach is considered as the paradigm of systems biology research ([4]) and proposed as the only logical way for biology to advance ([70]). Development and refinement of a mathematical model of a biochemical process proceeds, in general, in accordance with the following scenario. First, a biochemical model capturing the underlying reaction network of the biological process is constructed. Second, the biochemical model is transformed into an associated mathematical model. This usually involves two steps: obtaining equations describing the dynamics of the system and, next, identifying the model parameter values so that the model fits some experimental data. Finally, the mathematical model is validated against another set of experimental data or qualitative empirical knowledge.

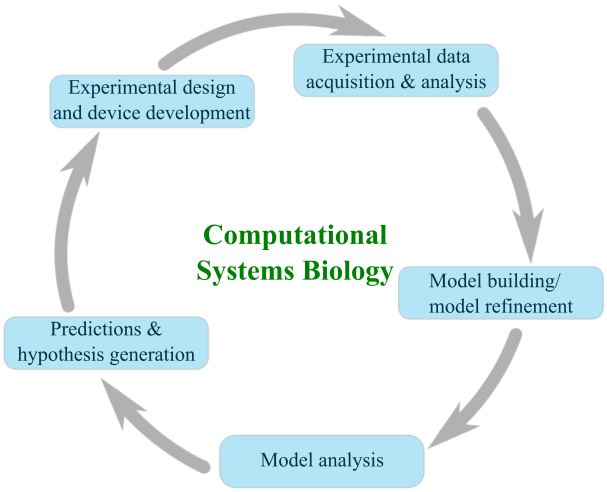


Figure 2.1: The iterative cycle of systems biology: starting from a model abstracting a biological system, the iterative process of hypothesis generation, experimental design, experimental analysis, and model refinement lies at the core of systems biology.

The predictive power of the obtained model is then used to formulate new hypothesis about the considered process. This drives further experimental research, potentially involving novel experimental design. The obtained results are then utilized to refine the model to include more details and better explain the observations.

As stated in [63], a system-level understanding of a biological system can be obtained by insight into four key properties: 1) system structures, 2) system dynamics, 3) the control method, and 4) the design method. System structures include gene interactions network, biochemical pathways, and mechanisms by which such interactions modulate the physical properties of intracellular and multicellular structures, being the large-scale effects of interactions. System dynamics relates to the investigation of the behaviour of a system over time under various conditions. It involves tools such as metabolic control analysis, sensitivity analysis, dynamical systems analysis methods, e.g. phase portrait and bifurcation analysis, and identification of mechanisms giving rise to specific behaviours. The control method investigates the mechanisms that systematically control the state of the system. It involves the identification and adjustment of points that influence the state of the system. This method has potential applications in drug design. Finally, the design method is utilized in the modification of existing biological systems towards "improved" ones having some desired properties and

construction of new such systems. It requires an approach based on design principles and simulations instead of trail-and-error methods, see [63].

2.2 Biochemical reaction networks

As mentioned above, one of possible approaches towards model building is based on distinguishing a certain set of reactions characterizing the considered biological process. We are going to discuss this step in more details in this section. In parallel, we introduce some terminology and notation for describing biological systems that is extensively used in this thesis.

An abstraction of the considered biological process is made by identifying a *biochemical reaction network*, i.e. a finite set of reactions among a finite set of biochemical species, underlying this process. To this aim a relatively small set of biochemical reactions which are capturing the main features of the process' machinery is chosen. The chosen biochemical reactions may be very abstract themselves, i.e. one reaction may in fact encapsulate many real reactions which constitute a whole subprocess in a living organism. We say that the biochemical reaction network constitutes a *biochemical model* of the considered process. In the terminology of computer science this can be expressed by stating that the biochemical reaction network is an *abstraction* of the real biological process. In computer science the term abstraction is used to refer to the process of hiding the details and exposing only the essential features of a particular concept or object, which is exactly what one aims for when constructing a biochemical model. Moreover, computer scientists utilize abstraction as a tool for managing complexity. Coping with biological complexity and understanding it is the ultimate goal of building biochemical models. For example, the simple biochemical model presented in Section 3.1 and discussed in details in [96] is an abstraction of the eukaryotic heat shock response mechanism: in real cells this defence mechanism involves much more than just the 10 biomolecules and 17 biochemical reactions described in the model of [96]. However, despite huge simplifications, the abstraction is sufficient to capture the main regulatory features of the response, as shown in [96, 24, 83].

For this thesis, we adopt and interchangeably use the following terminology: biochemical reaction network, biochemical model, biochemical reaction system, biochemical system. The terms cellular reaction system/network are also commonly seen in the literature.

In a biological cell there are different classes of biochemical reaction networks whose interplay enables the cell to perform its vital functions. The major and commonly recognized classes are: *metabolic networks*, *signal transduction networks* and *gene regulatory networks*, see e.g. [11, 20, 72, 133]. We briefly describe each of them.

Metabolic networks. There are two opposing streams of chemical reactions that take place in the cell: 1) *catabolic reactions* and 2) *anabolic-* or *biosynthetic reactions*. Catabolic reactions break down complex compounds into smaller molecules, thereby generating energy and providing the cell with elementary building blocks. The acquired energy and the basic components are exploited by anabolic reactions to construct complex molecules used in cellular functioning. These two sets of reactions constitute the *metabolism* of the cell ([3, 66]). In other words, metabolism is the sum of all the chemical reactions that take place in every cell of a living organism which provide energy for the processes of life and synthesize new molecules. It is a highly organized process that involves thousands of reactions which are catalysed by enzymes ([66]). Metabolism of the cell is organized in terms of *metabolic pathways*, i.e. sequences of biochemical reactions where the product of one reaction is a substrate for the next one. Metabolic pathways constitute the *metabolic network* underlying the metabolism of the cell.

Signal transduction pathways. In a multicellular organism its cells have to be able to communicate with each other in order to combine into networks that realize higher levels of organization, including, e.g., tissue and organs, and to adjust their own behaviour for the benefit of the organism as a whole ([137, 3]). For example, cells in multicellular organisms must sense the presence of ambient hormones and other neighbouring cells when making decisions such as whether to proliferate, move or die. The study of the mechanisms by which this transfer of biological information comes about is referred to as 'signal transduction', 'cell signalling' or simply 'signalling', see [31].

One can view signal transduction pathways as routes of cellular information. Through them cells monitor their surroundings, as well as their own state. They allow the cell to adjust to environmental changes or hormonal stimuli. Signal transduction pathways orchestrate cellular metabolism, control growth, proliferation and development, establish stress tolerance, and determine morphogenesis ([91]). Typically, the signalling paradigm involves the following sequence of events. First, the 'signal' reaches the proximity of the cell surface. There are then two possible modes by which the cell can import the signal: either 1) the stimulus penetrates the cell membrane and binds to a respective receptor in the cell interior, or 2) the signal is perceived by a transmembrane receptor. In the latter case the signal does not cross the membrane, but instead the state of the receptor is changed from susceptible to active and stimulates an internal signalling cascade, which often involves a series of changes in protein phosphorylation states. The sequence of state changes crosses the nuclear membrane and eventually a transcription factor is either activated or deactivated. This leads to a change in the transcription rates of certain genes and the resulting change in certain pro-

tein concentrations causes the actual response of the cell to the signal. In addition to this downstream program, signalling pathways are regulated by a number of control mechanisms such as feedback and feed-forward modulation ([66]). For a more comprehensive discussion of cell signaling we refer, e.g., to [31, 41, 66, 91, 137].

Gene regulatory networks. A gene regulatory network consists of a set of genes, proteins, small molecules, and their mutual regulatory interactions. A central role in the control of cellular processes of an organism is played by the genome. Proteins synthesized from genes may function as transcription factors binding to regulatory sites of other genes, as enzymes catalysing metabolic reactions, or as components of signal transduction pathways. Gene expression is a complex process involving several stages of regulation. Besides regulation on the level of DNA transcription, the gene expression process may be controlled during RNA processing and transport, RNA translation, or on the stage of posttranslational modifications of proteins. The proteins that are in charge of performing these regulatory functions are synthesized by other genes, which gives rise to *genetic regulatory systems* structured by *regulatory interactions networks* describing the interactions between DNA, RNA, proteins, and small molecules ([29]). Recent development of high-throughput experimental techniques such as cDNA microarrays or oligonucleotide chips, which permit rapid and massively parallel measurement of spatiotemporal expression levels of genes, has contributed to the intensification of the studies on gene regulatory networks. Most of these networks are large and complex. Understanding their dynamics by intuitive approaches alone is very hard and thus, in addition to experimental techniques, formal methods and computer tools for the modelling and simulation of gene regulation processes are indispensable. For a literature review of existing approaches we refer to [29].

2.3 Modelling of biochemical reaction networks

Methods for representing and communicating biological networks in both human- and machine-readable form have become increasingly important ([64]). Biochemical reaction networks can be represented in various ways that can comprise a rich collection of diverse biological and biochemical knowledge. For example, a graphical representation of biological networks leads to *Molecular Interaction Maps* (MIM) ([67, 69]) or *process diagrams* ([64, 68]). In this thesis, biochemical networks are represented in the form of lists of stoichiometric equations that embody the main processes that constitute these systems. Although this representation is less sophisticated than the mentioned graphical forms, it is still capable of capturing the rel-

evant information about the considered network needed for further proper modelling of the network's dynamical behaviour.

The basic components of a biochemical reaction in the considered model are: 1) the substances with their concentrations and 2) the reactions changing the concentrations of the substances. Formally, we consider a set of substances $\mathcal{S} = \{S_1, \ldots, S_N\}$, also referred to as chemical species, and a set of reactions $\mathcal{R} = \{R_1, \ldots, R_M\}$ where N and M are positive integers. Each reaction operates on a subset of substances from $\mathcal{S}$. It models in an abstract and compact way a transformation which takes place in the considered biochemical system. It carries information about what substances in what proportions react and what substances in what proportions are the outcomes of the conversion. There are two types of chemical and biochemical reactions: reversible and irreversible ones. In general, an irreversible reaction R_μ is symbolically written in form of a stoichiometric equation as

$$\begin{aligned} R_\mu : \quad & m_{\mu,1}\, S_1 + m_{\mu,2}\, S_2 + \ldots + m_{\mu,N}\, S_N \rightarrow \\ & m'_{\mu,1}\, S_1 + m'_{\mu,2}\, S_2 + \ldots + m'_{\mu,N}\, S_N, \end{aligned} \tag{2.1}$$

where $m_{\mu,i}$ and $m'_{\mu,j}$ are non-negative integers defining the stoichiometry of the reaction, i.e. the relationship between the amounts of substances that react together in the reaction, and the amounts of substances that are formed ([81]). The reactant substances, i.e. occurring on the left-hand side, are called *substrates* and the resultant substances, i.e. placed on the right-hand side, are referred to as *products*. If a species S_i is not reacting in a reaction R_μ, then $m_{\mu,i}$ is zero and if it is not produced, then $m'_{\mu,i}$ is set zero. Notice that for any $1 \leq i \leq N$ we allow both $m_{\mu,i}$ and $m'_{\mu,i}$ to be non-zero, which means that S_i is both consumed and produced in reaction R_μ. We associate with each reaction R_μ two natural values K_μ and L_μ, $0 \leq K_\mu, L_\mu \leq N$, defined in the following way:

$$K_\mu = |\{S_i | 1 \leq i \leq N \wedge m_{\mu,i} \neq 0\}| \tag{2.2}$$

and

$$L_\mu = |\{S_j | 1 \leq j \leq N \wedge m'_{\mu,j} \neq 0\}|. \tag{2.3}$$

They give the numbers of distinct substrates and products that are involved in reaction R_μ, respectively.

The sum $\sum_{i=1}^{N} m_{\mu,i}$ is the so-called molecularity of the irreversible reaction R_μ, i.e. the number of reactant molecular entities that are involved in the reaction. A reaction with a molecularity of one is called 'unimolecular', one with a molecularity of two 'bimolecular' and of three 'termolecular' ([81]). For example, $A + B \rightarrow C$ is a bimolecular reaction, whereas $A + 2B \rightarrow C$ is a termolecular one. However, due to improbability of three molecular entities colliding at exactly the same time in a suitable orientation

for reaction, termolecular reactions are rarely encountered ([58, 131]). Such reactions occur under unique conditions. We provide, after [58], one example of a termolecular reaction. The reaction of two hydrogen atoms in the gas phase to form a hydrogen molecule cannot take place as a bimolecular reaction. This is due to the fact that the energy released by the formation of the H–H covalent bond can only go into vibrational and rotational energy of the new molecule, and this energy is sufficient to cause the almost immediate reversal of the reaction to split the molecule again into two hydrogen atoms ([58]). If a third molecule is available to absorb the energy, the following reaction can take place:

$$\mathrm{H}^{\bullet} + \mathrm{H}^{\bullet} + \mathrm{M} \rightarrow \mathrm{H\text{–}H} + \mathrm{M}. \tag{2.4}$$

Reactions with molecularity higher than three are virtually impossible.

A reversible reaction R_η is in general in the form of

$$\begin{aligned} R_\eta: \quad & m_{\eta,1}\, S_1 + m_{\eta,2}\, S_2 + \ldots + m_{\eta,N}\, S_N \rightleftarrows \\ & m'_{\eta,1}\, S_1 + m'_{\eta,2}\, S_2 + \ldots + m'_{\eta,N}\, S_N. \end{aligned} \tag{2.5}$$

Any reversible reaction can be split into two irreversible ones. For example, the above reversible reaction R_η can be written as a pair of irreversible reactions $R_{\eta'}$ and $R_{\eta''}$ as follows:

$$\begin{aligned} R'_\eta: \quad & m_{\eta,1}\, S_1 + m_{\eta,2}\, S_2 + \ldots + m_{\eta,N}\, S_N \rightarrow \\ & m'_{\eta,1}\, S_1 + m'_{\eta,2}\, S_2 + \ldots + m'_{\eta,N}\, S_N, \\ R''_\eta: \quad & m_{\eta,1}\, S_1 + m_{\eta,2}\, S_2 + \ldots + m_{\eta,N}\, S_N \leftarrow \\ & m'_{\eta,1}\, S_1 + m'_{\eta,2}\, S_2 + \ldots + m'_{\eta,N}\, S_N. \end{aligned} \tag{2.6}$$

Hence, any reaction network involving reversible reactions can be transformed into a corresponding network based only on irreversible reactions.

There are two special cases of irreversible reactions where either $K_\mu = 1$ and $L_\mu = 0$ or $K_\mu = 0$ and $L_\mu = 1$. Usually, in these variants the respective stoichiometric coefficient is set to 1. Thus, in the former case the reaction is of the form

$$\rightarrow S_j \tag{2.7}$$

and models a constant inflow of product S_j into the system. In the latter case the reaction is

$$S_i \rightarrow \tag{2.8}$$

and describes the outflow of the substrate S_i from the system, commonly referred to as degradation of S_i.

The reactions of a biochemical process are usually interrelated in the following manner: a product of one reaction is a substrate for another one

or one reaction produces an enzyme which catalyses some other reaction of the process. In this sense the reaction set $\mathcal{R}$ forms a network which models the biochemical process under study, in other words forms its biochemical model.

The biochemical model provides us with information on the structure and stoichiometries of the system. However, in order to investigate the dynamics, i.e. the behaviour of the system in time under certain conditions, a mathematical model needs to be introduced.

2.4 Mass-action models for biochemical reaction networks

From the biochemical model an associated *mathematical model* is often derived by deterministic kinetic modelling of individual biochemical reactions. Practical limitations, such as inability to measure interactions in all detail or lack of knowledge on all the properties of the molecules involved, make that the mathematical models of the reaction network cannot be derived from the most basic laws of physical mechanics. Necessarily, the contemporary models aggregate information about mechanistic detail and in this sense can be seen as *macroscopic* or *phenomenological* constructs compared to the *microscopic* approach, where single molecules and their interactions are considered ([66, 138, 137]). The dynamic behaviour of the system is represented in the temporal evolution of its state expressed as the concentrations of all the species considered ([138]). Thus, the deterministic framework based on ordinary differential equations (ODEs) is often chosen as the model of observations made in experiments, i.e. the system of ODEs constitutes a mathematical abstraction of the process under investigation, see, e.g., [138]. The basic quantities are the concentration $[S]$ of a substance S and the rate ν of a reaction ([66]). The concentration is often expressed as the number n of molecules (count of molecules) of substance S per volume V or the number of moles of S per volume V. For the needs of this thesis, we adopt the following notation: the number of molecules of substance S is denoted by $\#S$ and the number of moles of S is referred to simply as S. The number of moles and the number of molecules are related to each other through the Avogadro number $N_A \approx 6.02214179 \cdot 10^{23}$ particles/mol, i.e.

$$\#S = S \cdot N_A. \tag{2.9}$$

The unit of $[S]$ commonly encountered in the literature is $\mathrm{M} = \mathrm{mol} \cdot \mathrm{L}^{-1}$, where L stands for a litre. Biochemical reaction kinetics rely on the assumption that the reaction rate at a certain point in time and space can be expressed as a unique function of the concentrations of all substances at this point in time and space, see [66]. The kinetics is governed by the *mass*

action law (originally introduced in [39] and [40] by Guldberg and Waage in the 19th century), which can be briefly summarized as follows: the rate ν of a reaction is proportional to the product of the reactant masses, with each mass raised to the power equal to the corresponding stoichiometries ([66]). In the classical formulation the rate does not depend directly on time, i.e. $\nu(t) = \nu(\mathbf{S}(t))$, where $\mathbf{S}$ denotes the vector of concentrations of all the substances in a reaction network. For example, for a simple reaction

$$S_1 + S_2 \rightleftarrows 2\,S_3 \tag{2.10}$$

the reaction rate reads

$$\nu = \nu_+ - \nu_- = k_+\,S_1\,S_2 - k_-\,S_3^2, \tag{2.11}$$

where ν is the resultant rate of the reversible reaction, ν_+ is the rate of the forward reaction, ν_- the rate of the backward reaction, while k_+ and k_- are the respective proportionality factors, the so-called *kinetic-* or *rate constants* ([66]). In the general case the mass action rate law for a reversible reaction R_μ reads

$$\nu = \nu_+ - \nu_- = k_+ \prod_{\iota=1}^{K_\mu} [S_{s(\mu,\iota)}]^{m_{\mu,\iota}} - k_- \prod_{\iota=1}^{L_\mu} [S_{p(\mu,\iota)}]^{m'_{\mu,\iota}}. \tag{2.12}$$

If the concentration of substances is measured in M and the time in seconds (s), then the rate is expressed in terms of $\mathrm{M} \cdot s^{-1}$. It follows that the rate constants for bimolecular reactions, i.e. of the form $S_1 + S_2 \to \ldots$, have the unit $\mathrm{M} \cdot s^{-1}$, while monomolecular reactions, i.e. $S \to \ldots$, have the unit s^{-1}. The mass action kinetics model is derived based on the Boltzmann's kinetic theory of gases and is justified under the assumption of constant temperature and fast enough diffusion in the cell, which ensures that the mixture of substances is "well-stirred", i.e. homogenously distributed in a fixed volume V.

The *stoichiometric coefficients* denote the quantitative proportion in which substrate and product molecules are involved in a reaction ([66]). For example, for the reversible reaction $S_1 + S_2 \rightleftarrows 2P$, the stoichiometric coefficients of S_1, S_2 and P are -1, -1, and 2. For a reaction $S_1 + S_2 \rightleftarrows 2P + S_2$ the stoichiometric coefficients are -1, 0, and 2. Thus, if a species S_i is both a reactant and a product in a reaction R_μ, then the resulting stoichiometric coefficient of S_i in R_μ is, in terms of the notation introduced in Section 2.3, $m'_{\mu,i} - m_{\mu,i}$. In general, the stoichiometric numbers are positive for products and negative for reactants ([81]). For the irreversible reaction (2.1) the stoichiometric coefficients read: $m'_{\mu,1} - m_{\mu,1}$, $m'_{\mu,2} - m_{\mu,2}$, ..., $m'_{\mu,N} - m_{\mu,N}$. In the case of a reversible reaction their values depend on the chosen direction. If the direction of the general reversible reaction (2.5) is chosen to

be 'left-to-right', then the stoichiometric coefficients of this reaction are the same as in the irreversible case. The stoichiometric coefficient of substance S_i in reaction R_μ is denoted in this work as $n_{i\mu}$. The stoichiometric coefficients of a reaction network are organized in a matrix of dimensions $N \times M$, a so-called *stoichiometric matrix*, denoted $\mathbf{N}$, i.e.

$$\mathbf{N} = \{n_{i\mu}\}_{i\mu}, \tag{2.13}$$

for $i = 1, \ldots, N$ and $\mu = 1, \ldots, M$.

The dynamics of a biochemical reaction network derived from the law of mass action can be described by a system of first-order, ordinary differential equations (ODEs), also referred to as the rate equations. For a reaction network consisting of N substances and M reactions the dynamics, in particular the change of concentrations in time, is described by the following system of equations:

$$\frac{\mathrm{d}}{\mathrm{d}t}[S_i] = \sum_{\mu=1}^{M} n_{i\mu}\, \nu_\mu, \tag{2.14}$$

for $i = 1, \ldots, N$. In this formulation one assumes that the reactions are the only reason for concentration changes and that no mass flow takes place due to diffusion or to convection ([66]). This can be rewritten in the matrix notation as follows

$$\frac{\mathrm{d}\mathbf{S}}{\mathrm{d}t} = \mathbf{N}\nu, \tag{2.15}$$

where $\mathbf{S}$ is a vector of the concentrations of all the substances in the reaction network, i.e. $\mathbf{S} = ([S_1], \ldots, [S_N])^{\mathrm{T}}$, and $\nu = (\nu_1, \ldots, \nu_M)^{\mathrm{T}}$ is the vector of the reaction rates. Hence, in this framework, the mathematical model of a reaction system consists of the vectors $\mathbf{S}$, ν, the stoichiometric matrix $\mathbf{N}$ and a vector κ consisting of the reaction rate constants, which are constituents of ν.

2.4.1 Steady state and steady state fluxes

One of the basic concepts of dynamical systems theory extensively utilized in systems biology is the notion of a steady state. In steady state it holds for a reaction network that

$$\frac{\mathrm{d}\mathbf{S}}{\mathrm{d}t} = \mathbf{N}\nu = \mathbf{0}. \tag{2.16}$$

The rate vectors satisfying the above steady state condition can be obtained by solving the linear system of algebraic equations denoted by the right equality sign in (2.16). From the theory of linear algebra we know that the equation has nontrivial solutions, i.e. different from the zero vector, only if $\mathrm{Rank}(\mathbf{N}) < M$. This can be expressed in words that a nontrivial

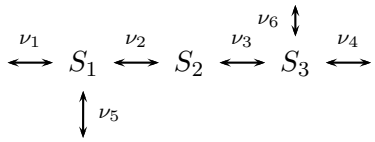


Figure 2.2: An example of a reaction network originally presented in [66].

solution exists provided some of the columns of the stoichiometric matrix $\mathbf{N}$ are linearly dependent. If this is the case, the solutions to this equation form a null space (also referred to as kernel). The vectors forming the basis of the null space arranged into a matrix form the so-called kernel matrix denoted by $\mathbf{K}$. The dimension of the null space, hence the number of columns in $\mathbf{K}$ is $M - \mathrm{Rank}(\mathbf{N})$. Every possible set of steady state fluxes can be expressed as a linear combination of the columns $\mathbf{k}_i$ of matrix $\mathbf{K}$, i.e.

$$\mathbf{J} = \sum_{i=1}^{M-\mathrm{Rank}(\mathbf{N})} \alpha_i \cdot \mathbf{k}_i. \tag{2.17}$$

The choice of the kernel basis vectors is not unique. However, for a network containing irreversible reactions the set of vectors forming $\mathbf{K}$ is restricted by the condition that the rows in $\mathbf{K}$ corresponding to the irreversible reactions cannot contain negative (or positive, depending on the definition of flux direction) entries. Note that a row in $\mathbf{K}$ formed only of zero entries indicates an *equilibrium reaction*, i.e. a reaction that in any steady state must have a zero net rate.

A set of rows in which all basis vectors, i.e. all columns of $\mathbf{K}$, have the same entries indicates an unbranched reaction path. In each steady state, the net rate of all reactions constituting this path is equal ([66]). Let us consider an example. The reaction network originally presented in [66] and reproduced in Figure 2.2 comprises 6 reactions. The stoichiometric matrix for this system reads

$$\mathbf{N} = \begin{pmatrix} 1 & -1 & 0 & 0 & -1 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 1 \end{pmatrix} \tag{2.18}$$

and $\mathrm{Rank}(\mathbf{N}) = 3$. The kernel matrix is, e.g., spanned by the following three basis vectors: $\mathbf{k}_1 = (1\ 1\ 1\ 0\ 0\ -1)^{\mathrm{T}}$, $\mathbf{k}_2 = (1\ 0\ 0\ 0\ 1\ 0)^{\mathrm{T}}$, and

$\mathbf{k}_3 = (-1\ -1\ -1\ -1\ 0\ 0)^{\mathrm{T}}$, i.e.

$$\mathbf{K} = \begin{pmatrix} 1 & 1 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \\ -1 & 0 & 0 \end{pmatrix}. \tag{2.19}$$

The second and the third row of the kernel matrix are the same, i.e. the entries for the second and third reactions are always equal. This indicates that in any steady state the fluxes through reactions 2 and 3 must be equal.

2.4.2 Mass conservation relations

Frequently, the concentrations of several substances involved in biochemical reaction networks are included in so-called conservation sums. A characteristic feature of such substances is that they are neither produced nor degraded, however they can form complexes with other species or be part of other species. For example, in the simple mass-action model for the eukaryotic heat shock response introduced in [96], there are three conservation relations concerning the total amount of heat shock factors, the total amount of proteins other than heat shock proteins and heat shock factors, and the total amount of heat shock elements, see [96] and Section 3.1.1 in Chapter 3 for details.

In general terms, a mass conservation relation can be expressed as: a linear combination of concentrations of species is conserved in time, i.e.

$$\mathbf{g}^{\mathrm{T}}\mathbf{S}(t) = c, \tag{2.20}$$

where $\mathbf{g}$ is a vector of some constant entries and c denotes a constant conservation quantity ([43]). This can be translated as: conservation relations are linear dependencies between some rows of the stoichiometric matrix, i.e.

$$\mathbf{g}^{\mathrm{T}}\mathbf{N} = \mathbf{0}^{\mathrm{T}}. \tag{2.21}$$

The equivalence with (2.20) can be derived by observing that

$$\mathbf{g}^{\mathrm{T}}\dot{\mathbf{S}} = \mathbf{g}^{\mathrm{T}}\mathbf{N}\nu = 0, \tag{2.22}$$

hence by integrating we obtain that $\mathbf{g}^{\mathrm{T}}\mathbf{S} = const.$ ([66]). The number of independent conservation vectors $\mathbf{g}$, thus the number of conservation relations in the considered reaction network, is given by $N - \mathrm{Rank}(\mathbf{N})$. If the stoichiometric matrix is full rank, it follows that the system embraces no conservation relations.

A complete set of linearly independent vectors $\mathbf{g}$ of a reaction network can be arranged into a so-called conservation matrix $\mathbf{G}$ ([43, 66]) fulfilling

$$\mathbf{GN} = \mathbf{0}. \tag{2.23}$$

In other words, $\mathbf{G}^{\mathrm{T}}$ is a kernel matrix of $\mathbf{N}^{\mathrm{T}}$. It can be found, e.g., by using the Gauss algorithm. Notice that a matrix $\mathbf{G}' = \mathbf{PG}$ with $\mathbf{P}$ being any nonsingular matrix of appropriate dimension is a conservation matrix as well. Hence, a conservation matrix is not uniquely defined.

Conservation relations can be used to simplify the system of differential equations $\dot{\mathbf{S}} = \mathbf{N}\nu$ describing the dynamics of a reaction network. Here we explain, after [66], how this can systematically be done. First, the rows in the stoichiometric matrix and the concentration vector are reordered in such a way that the independent rows of $\mathbf{N}$ are at the top and the dependent rows are placed at the bottom, i.e. matrix $\mathbf{N}$ is split into two parts: the independent one denoted $\mathbf{N}^0$ and the dependent one referred to as $\mathbf{N}'$. Also a so-called link matrix $\mathbf{L}$ is introduced in such a way that the following holds:

$$\mathbf{N} = \begin{pmatrix} \mathbf{N}^0 \\ \mathbf{N}' \end{pmatrix} = \mathbf{L}\mathbf{N}^0 = \begin{pmatrix} \mathbf{I}_{\mathrm{Rank}(\mathbf{N})} \\ \mathbf{L}' \end{pmatrix} \mathbf{N}^0, \tag{2.24}$$

where $\mathbf{I}_{\mathrm{Rank}(\mathbf{N})}$ is the identity matrix of size $\mathrm{Rank}(\mathbf{N})$. Now, the system of differential equations for the reaction network may be rewritten in the following form:

$$\frac{\mathrm{d}\mathbf{S}}{\mathrm{d}t} = \begin{pmatrix} \dot{\mathbf{S}}_{\mathrm{indep}} \\ \dot{\mathbf{S}}_{\mathrm{dep}} \end{pmatrix} = \begin{pmatrix} \mathbf{I}_{\mathrm{Rank}(\mathbf{N})} \\ \mathbf{L}' \end{pmatrix} \mathbf{N}^0\nu \tag{2.25}$$

and the change in time of the dependent concentrations satisfy

$$\dot{\mathbf{S}}_{\mathrm{dep}} = \mathbf{L}' \cdot \dot{\mathbf{S}}_{\mathrm{indep}}. \tag{2.26}$$

By integrating we obtain that

$$\mathbf{S}_{\mathrm{dep}} = \mathbf{L}' \cdot \mathbf{S}_{\mathrm{indep}} + \mathrm{constant}. \tag{2.27}$$

In consequence the original system of ODEs can be replaced by a reduced differential equation system and a set of algebraic equations, i.e.

$$\begin{cases} \dot{\mathbf{S}}_{\mathrm{indep}} &= \mathbf{N}^0\nu, \\ \mathbf{S}_{\mathrm{dep}} &= \mathbf{L}' \cdot \mathbf{S}_{\mathrm{indep}} + \mathrm{constant}. \end{cases} \tag{2.28}$$

2.4.3 Elementary flux modes

In the context of metabolic networks the notion of a *pathway* often appears, however it is not easy to define what a pathway in a given metabolic network is. An intuitive definition of a pathway is a sequence of reactions linked by common metabolites ([66]). Examples of metabolic pathways are *glycolysis*, *citric acid cycle* or *oxidative phosphorylation*, see, e.g., [3] for details. As stated in [115], there exist currently two fundamentally different approaches to the definition of metabolic pathways. One is a qualitative identification based on historical groups of reactions in a database setting. The second relies on precise quantitative and systemic definitions based on mathematical studies such as linear algebra and convex analysis. In this work we concentrate on the latter case. An attempt to formalize the notion of pathway has been proposed in [44, 99, 116, 117, 118, 119] in the form of elementary flux modes. The intuitive meaning of an elementary flux mode is a set of reactions whose combined quantitative contribution to the system is zero. In other words, the net loss of substance caused by any reaction in that set is compensated by a net gain in the same substance incurred by some other reactions in the set. From this perspective, the analysis of metabolic pathways uses only the information on the stoichiometric structure and the reversibility or irreversibility of the reactions. First, a flux mode $\mathbf{M}$ is defined as a class of flux vectors that represent direct routes through the metabolic network from one external metabolite to another. Formally,

$$\mathbf{M} = \{\nu \in \mathbb{R}^M \,|\, \nu = \alpha\nu^*, \alpha > 0\}, \tag{2.29}$$

where ν^* is an M-dimensional vector (unequal to the null vector) fulfilling two conditions. First, it satisfies the steady state equation, i.e. $\mathbf{N}\nu^* = \mathbf{0}$. Second, the signs of the entries indicate the corresponding flux directions in agreement with the chosen directions of the irreversible reactions. A flux mode comprising vector ν is called an *elementary flux mode* if ν cannot be represented as a nonnegative linear combination of two vectors that fulfill the two conditions, but contain more zero entries than ν ([66]).

An elegant mathematical solution to the problem of determining metabolic pathways is obtained by applying the theory of convex analysis; for details we refer to, e.g., [116]. For a discussion on the problem of finding elementary flux modes in metabolic networks seen from an algorithmic perspective and covering the complexity issues see, e.g., [1]. For any given metabolic network, the full set of elementary fluxes can be determined using dedicated software such as METATOOL ([99]). The recognition of the elementary flux modes allows the detection of the full set of non-decomposable steady-state flows that the network can support, including cyclic flows. Any steady-state flux pattern can be expressed as a non-negative linear combination of these modes ([116, 117, 118]). The identified elementary flux modes should have

clear biological interpretation: a flux mode is a set of enzymes that operate together at a steady state and a flux mode is elementary if the set of enzymes is minimal, i.e. complete inhibition of any of the enzymes would result in a termination of this flux ([116, 117, 118]). The lack of possibility to interpret the modes in this way is a signal that the model under consideration may not be correct.

2.4.4 Flux balance analysis

The steady state problem (2.16) for a metabolic network may have many mathematical solutions. However, not all of them are biologically sound or interesting. For example, the desired one may be the one that keeps certain metabolites in the correct proportion yet maximizes the growth rate of an organism. Flux balance analysis (FBA) provides means for solving this type of problems. It is a mathematical approach for analysing biochemical networks, in particular the genome-scale metabolic network reconstructions. FBA enables calculation of the flow of metabolites through a metabolic network and, in consequence, makes predictions concerning the growth rate of an organism or the rate of production of certain metabolites possible ([93]). Similarly to the approach of elementary flux modes, FBA is independent of the information concerning concentrations of metabolites or kinetic details of the considered system. It investigates the theoretical capabilities and operative modes of metabolism by introducing further constraints in the stoichiometric analysis ([66]). The first constraint is the steady state requirement. Other constraints may be of thermodynamic nature, regarding the irreversibility of reactions ([66]) or reactions can be given upper and lower bounds, which determine the maximum and minimum permissible fluxes of the reactions ([93]). These constraints are of the form

$$\alpha_i \leq \nu_i \leq \beta_i, \tag{2.30}$$

where ν_i is the flux of i-th reaction, α_i and β_i determine, respectively, the lower and upper bound of the flux. In this way, these constraints impose restrictions on the magnitude of individual fluxes. For example, a thermodynamic constraint forcing only the forward direction of a reaction can be introduced as $0 \leq \nu_i < +\infty$. Some other constraints can also be included, see, e.g. [104].

In FBA the constraints are introduced in two ways. First, as equations that balance reaction inputs and outputs: with the requirement of steady state, the stoichiometric matrix imposes flux balance constraints on the system. Second, as inequalities that restrict the allowable fluxes of the reactions. The constraints confine the steady-state fluxes to a feasible set. Next, a phenotype in the form of a biological objective that is of interest in the problem being studied needs to be defined ([93]). This is often achieved

by adding an artificial reaction to the system, i.e. an additional column of coefficients to the stoichiometric matrix. The determination of a particular metabolic flux distribution is then formulated as a linear programming problem, i.e. the aim is to maximize an objective function Z that quantitatively defines how much each reaction contributes to the phenotype of interest in the problem being studied ([93]). The objective function Z is subject to stoichiometric and capacity constraints and is often of the form

$$Z = \sum_{i=1}^{M} c_i \nu_i = \mathbf{c}^{\mathrm{T}} \nu \rightarrow \max, \tag{2.31}$$

where $\mathbf{c}$ is a vector of weights for the individual rates indicating how much each reaction contributes to the phenotype ([66, 93]). Examples of such objective functions are: maximization of biomass production or ATP production in an organism, maximal growth rate, minimization of nutrient uptake. Through optimization of an objective function, flux balance analysis enables finding an optimal flux distribution placed somewhere at the edge of the restricted solution space ([93]). There exist many computational linear programming algorithms and software packages that are able to coupe with large systems of equations. For example, the COnstraint-Based Reconstruction and Analysis (COBRA) Toolbox ([9]) for Matlab is a free software (distributed under the GNU Library General Public License) that performs such computations.

2.4.5 Local sensitivity analysis

Local sensitivity analysis is a method to estimate the changes brought into the system through small perturbations in the parameters of the model. In this way, one may estimate both the robustness of the model against small changes in the model, as well as identify possibilities for bringing a certain desired change in the system. For example, one question that is often asked of a biochemical model is what changes should be done to the model so that the new steady state satisfies certain properties. We briefly present the theoretical foundations of this analysis, which is extensively applied to verify the dependence of the model results presented in this thesis on the parameter choice. For a review and a more thorough presentation of sensitivity analysis we refer to, e.g., [132]. The robustness of a model with respect to parameter changes spanning the whole admissible range of parameter values can be assessed with the global sensitivity analysis. However, the global sensitivity analysis is out of the scope of this thesis and for more details on its techniques we refer to, e.g., [108].

To start, we rewrite the system of ordinary differential equations (2.14) describing the dynamics of a reaction network in a more general form:

$$\frac{\mathrm{d}}{\mathrm{d}t}[S_i] = f_i([S_1], \ldots, [S_N], \kappa), \qquad \text{for all } 1 \leq i \leq N, \tag{2.32}$$

where $\kappa = (k_1, \ldots, k_M)^{\mathrm{T}}$ is the rate constants vector (we can assume without loss of generality that the biochemical model consists of M irreversible reactions, see Section 2.3, thus there are M rate constants).

The partial derivatives of the solution of the system 2.32 with respect to the parameters of the system are considered. These are called *first-order local concentration sensitivity coefficients*. Let us denote $\mathbf{S}(t, \kappa) = ([S_1](t, \kappa), [S_2](t, \kappa), \ldots, [S_N](t, \kappa))^{\mathrm{T}}$ the solution of the system (2.32) with respect to the parameter vector κ. The concentration sensitivity coefficients are the time functions $\partial[S_i]/\partial k_j(t)$, for all $1 \leq i \leq N$, $1 \leq j \leq M$. Differentiating the system (2.32) with respect to k_j yields the following set of *sensitivity equations*:

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial \mathbf{S}}{\partial k_j} = \mathbf{J}\frac{\partial \mathbf{S}}{\partial k_j} + \frac{\partial \mathbf{f}}{\partial k_j}, \qquad \text{for all } 1 \leq j \leq M, \tag{2.33}$$

where $\partial \mathbf{S}/\partial k_j = (\partial[S_1]/\partial k_j, \ldots, \partial[S_N]/\partial k_j)^{\mathrm{T}}$ is the vector of partial derivatives, $\mathbf{f} = (f_1, \ldots, f_N)^{\mathrm{T}}$ is differentiable with respect to k_j for all $1 \leq j \leq M$, and $\mathbf{J}$ is the Jacobian of the system in (2.32), i.e.

$$\mathbf{J} = \begin{pmatrix} \partial f_1/\partial[S_1] & \partial f_1/\partial[S_2] & \cdots & \partial f_1/\partial[S_N] \\ \partial f_2/\partial[S_1] & \partial f_2/\partial[S_2] & \cdots & \partial f_2/\partial[S_N] \\ \vdots & \vdots & \ddots & \vdots \\ \partial f_N/\partial[S_1] & \partial f_N/\partial[S_2] & \cdots & \partial f_N/\partial[S_N] \end{pmatrix}. \tag{2.34}$$

The initial condition for the system (2.33) is that $\partial \mathbf{S}/\partial k_j(0) = \mathbf{0}$, for all $1 \leq j \leq M$. In practice, the solution of the system (2.33) can be numerically integrated, and in this way a numerical approximation of the time evolution of the sensitivity coefficients can be obtained.

Very often however, the focus is on sensitivity analysis around steady states. If the considered steady state is asymptotically stable, then one may consider the limit $\lim_{t\to\infty}(\partial \mathbf{S}/\partial k_j)(t)$, called *stationary sensitivity coefficients*. They reflect the dependency of the steady state on the parameters of the model. Mathematically, they are given by a set of algebraic equations obtained from (2.33) by setting $d/dt(\partial \mathbf{S}/\partial k_j) = \mathbf{0}$. We then obtain the following algebraic equations:

$$\left(\frac{\partial \mathbf{S}}{\partial k_j}\right) = -\mathbf{J}^{-1}\mathbf{F}_j, \qquad \text{for all } 1 \leq j \leq M, \tag{2.35}$$

where $\mathbf{J}$ is the value of the Jacobian at the steady state and $\mathbf{F}_j$ is the j-th column of the matrix $\mathbf{F} = (\partial f_r/\partial k_s)_{r,s}$ computed at the steady state.

When used for comparing the relative effect of a parameter change in two or more variables, the sensitivity coefficients must have the same physical dimension or be dimensionless, see [132]. Most often, one simply considers the matrix $\mathbf{C}$ of (dimensionless) *normalized* (also called *scaled*) sensitivity coefficients:

$$\mathbf{C}_{ij} = \frac{k_j}{[S_i](t,\kappa)} \cdot \frac{\partial [S_i](t,\kappa)}{\partial k_j} = \frac{\partial \ln\left([S_i](t,\kappa)\right)}{\partial \ln\left(k_j\right)} \tag{2.36}$$

Numerical estimations of the normalized sensitivity coefficients for a steady state may be obtained, e.g., with COPASI ([56]).

A similar sensitivity analysis may also be performed with respect to the initial conditions, see [132]. If we denote by $\mathbf{S}^{(0)} = \mathbf{S}(0,\kappa)$, the initial values of the vector $\mathbf{S}$, for parameters κ, then the *initial concentration sensitivity coefficients* are obtained by differentiating system (2.32) with respect to $\mathbf{S}^{(0)}$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial \mathbf{S}}{\partial \mathbf{S}^{(0)}} = \mathbf{J}\frac{\partial \mathbf{S}}{\partial \mathbf{S}^{(0)}}, \tag{2.37}$$

with the initial condition that $\partial \mathbf{S}/\partial \mathbf{S}^{(0)}(0)$ is the identity matrix.

Similarly as for the parameter-based sensitivity coefficients, it is often useful to consider the normalized, dimensionless coefficients

$$\frac{[S_j](0,\kappa)}{[S_i](t,\kappa)} \cdot \frac{\partial [S_i](t,\kappa)}{\partial [S_j](0,\kappa)} = \frac{\partial \ln\left([S_i](t,\kappa)\right)}{\partial \ln\left([S_j](0,\kappa)\right)}. \tag{2.38}$$

Stationary sensitivity coefficients with respect to the rate constants as well as the initial conditions can be numerically computed in software applications such as COPASI [56] or SBML-SAT [141], a tool for MATLAB®.

2.5 Markov chain models for biochemical reaction networks

Ignoring quantum mechanical effects, biological systems are often viewed as deterministic, with their dynamics entirely specified, given sufficient information on the state of the system (position, orientation and momentum of every single molecule) and a complete understanding of the chemistry and physics of the interactions between biomolecules ([135, 137]). Unfortunately, we are still unable to model biological systems of realistic complexity and size using such a molecular dynamic approach ([135, 137]). Therefore current models admit far-reaching simplifications, which result in a higher level view of the system being modelled.

The mathematical approaches used to model biological processes differ in their underlying assumptions and the level of resolution they can provide. A broad classification of these methods separates the resulting models into two classes: deterministic (*macroscopic* or *phenomenological*) and

stochastic (*mesoscopic*), where each of these two classes embodies various subclasses with their different mathematical formalisms ([109]). In particular, the use of differential equations, a representative of the deterministic class, for describing such processes makes certain assumptions that are not always satisfied. The assumption that variables can attain continuous values and the fact that random fluctuations (stochasticity) are usually not taken into account in the case of the ODE formulation are often brought up for discussion, especially in cases where the populations of involved species are small, see, e.g., [66, 109, 137, 139]. Such approach constitutes a simplification, since the underlying biological objects, molecules, are discrete in nature. As far as molecule numbers are sufficiently large, this is a minor problem. However, where the involved molecule numbers are on the order of dozens or hundreds, the discreteness and the random fluctuations may have a significant impact on the system's dynamics, but the deterministic approach to chemical kinetics may fail to expose these influences ([79, 123]).

An often applied solution to this problem is to consider the stochastic framework in which the investigated system is viewed as a continuous-time Markov chain ([126]). To this aim one considers the so-called *grand probability function* $\Pr(\mathbf{X}; t) \equiv$ probability that at time t there are X_1 molecules of species S_1, ..., X_N molecules of species S_N in the considered volume V, where $\mathbf{X} \equiv (X_1, X_2, \ldots, X_N)$ is a vector of molecular species populations. This function provides information on the probability distribution of all possible states at all times. Next, the probabilities of various reactions to be triggered in the next infinitesimal time interval $(t, t + \mathrm{d}t)$ are considered. A crucial assumption of the stochastic formulation is that the system is well stirred and at thermal equilibrium, see, e.g., [38]. As such, the molecules are at all times randomly and uniformly distributed throughout the volume V. The fundamental hypothesis of the stochastic formulation of chemical kinetics states that the average probability that a particular combination of reactants will react according to a given reaction R_μ in the next infinitesimal time interval $(t, t+\mathrm{d}t)$ is $c_\mu \cdot \mathrm{d}t$, for a certain constant c_μ, see, e.g., [36, 37, 38]. The constant depends on the reaction (the properties of the reactants) and on the temperature of the system. This is in fact a reformulation of the principle of mass action law, confront Section 2.4. Thus, the probability of a reaction R_μ taking place in the next infinitesimal time interval $(t, t+\mathrm{d}t)$ is $N_{R_\mu} \cdot c_\mu \cdot \mathrm{d}t$, where N_{R_μ} is the number of all combinations of R_μ reactants in the current state. Having an infinitesimally small time interval implies that the probability that two or more reactions take place in that interval is at least quadratic in $\mathrm{d}t$, i.e., vanishingly small. Thus, it can be assumed that at most one reaction takes place in that interval. In consequence, there are at most $M+1$ distinct configurations at time t that can lead to the state $\mathbf{X}$ at time $t+\mathrm{d}t$: either there is no state change in the considered time interval or one of the M reactions takes place in the considered time interval. Let

us denote by $a_\mu(\mathbf{X})\mathrm{d}t$ the probability of reaction R_μ occurring in the time interval $(t, t+\mathrm{d}t)$, given the state $\mathbf{X}$ at time t. What is said above, can then be formally written as

$$\begin{aligned}\Pr(\mathbf{X}; t+\mathrm{d}t) =& \Pr(\mathbf{X}; t)(1 - \sum_{\mu=1}^{M} a_\mu(\mathbf{X})\mathrm{d}t) \\ &+ \sum_{\mu=1}^{M} \Pr(\mathbf{X} - \nu_\mu; t) a_\mu(\mathbf{X} - \nu_\mu)\mathrm{d}t, \end{aligned} \tag{2.39}$$

where ν_μ is a stoichiometric vector defining the result of reaction R_μ on state vector $\mathbf{X}$, i.e. reaction R_μ changes system state from $\mathbf{X}$ to $\mathbf{X} + \nu_\mu$. Since

$$\frac{\partial \Pr(\mathbf{X}; t)}{\partial t} = \lim_{\mathrm{d}t \to 0} \frac{\Pr(\mathbf{X}; t+\mathrm{d}t) - \Pr(\mathbf{X}; t)}{\mathrm{d}t}, \tag{2.40}$$

it follows that

$$\frac{\partial \Pr(\mathbf{X}; t)}{\partial t} = [\sum_{\mu=1}^{M} \Pr(\mathbf{X} - \nu_\mu; t) a_\mu(\mathbf{X} - \nu_\mu) - \Pr(\mathbf{X}; t) a_\mu(\mathbf{X})], \tag{2.41}$$

which is a partial differential equation referred to in the literature as the *Chemical Master Equation* (CME). Although for a complex system detailed mathematical analysis based on the "chemical master equation" is intractable ([135]), it is possible to use a stochastic simulation approach that explicitly calculates the change in the number of molecules of the participating molecules. To this aim one can utilize the so-called Gillespie's algorithm ([36]) or its more efficient variant, the next reaction method developed by Gibson & Bruck ([35]). These algorithms are well-established procedures for generating an exact realization of the temporal behaviour of a continuous-time Markov chain being a stochastic model of the considered reaction system. For a detailed discussion on this topic we refer to [36, 37, 35, 135].

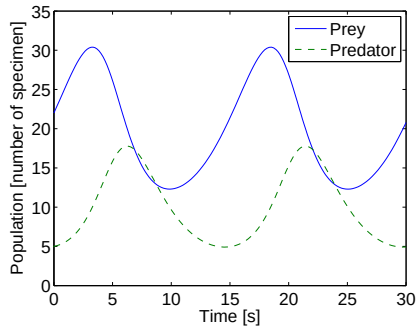
To provide an example of a possible discrepancy between the dynamics demonstrated by the stochastic and the deterministic formalisms, let us consider the famous Lotka-Volterra system of coupled ordinary differential equations describing an ecological predator-prey model. The solutions of this system are known to be periodic (except for the stationary point) independently of the initial size of predator and prey populations, see Figures 2.3a and 2.3b. However, in the stochastic formulation there exist "catastrophic" sequences of events which lead either to depletion of preys by predators and, in consequence, to the extinction of predators as well (Figures 2.3c and 2.3d), or the predators become extinct first and the prey population grows unlimited (Figures 2.3e and 2.3f). When running the model long enough, the

probability of never executing these catastrophic sequences drops to zero. This leads to radical qualitative differences in the trajectories obtained by these two approaches: in the deterministic case the trajectory in the predator versus prey phase space is a closed curve (Figure 2.3b), while in the stochastic case the trajectory either reflects that predators become extinct and the prey population grows without limitations (Figure 2.3f) or it eventually reaches the trivial steady state of no prey and no predator individuals in the system (Figure 2.3d). The expected time it takes for these scenarios to happen depends on the initial number of species. Such discrepancies in the trajectories are especially easily observed when the initial population sizes are small. A number of other examples that illustrate situations where discreteness and random fluctuations have a significant influence can be found in [109]. These examples illustrate the following possible discrepancies in the dynamics of the stochastic and the deterministic formulations: 1) an example of a monostable system depicts how a realization of a stochastic system may significantly vary from the time-course evolution of the deterministic description; 2) an example of a genetic switch exhibiting bistability illustrates how a stochastic realization randomly switches between two equilibrium (steady-state) points of the system, while the deterministic trajectory converges to only one of the equilibria; 3) an example of a genetic oscillator shows that incorporating random fluctuations to a system characterized by a unique deterministic equilibrium can lead to oscillations; finally, 4) a situation where fluctuations in the concentrations of key cellular regulators are of special interest, but they do not induce any change in the dynamical behaviour and are thus not displayed by the deterministic framework. The last example is briefly presented at the end of this section. For more details we refer to [109].

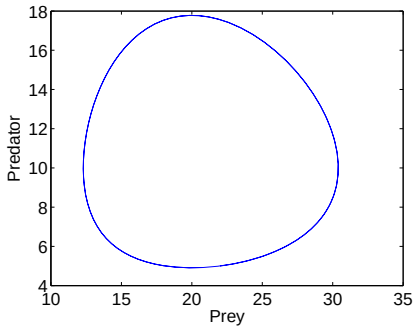
In [139], a careful analysis of the mathematical bases for the stochastic and deterministic framework is performed and the close relationship between them is investigated. Worth noticing is the fact that, as stated in [139], the discrepancies between the simulations in the two frameworks cannot be used as an argument against the use of the deterministic, continuous approach towards modelling of biological systems where the number of molecules of some species is small. In fact, both formalisms are correct and the choice of the framework should be made based on the context and purpose of modelling or whether a biological principle is reflected by the model, see [139] for a thorough discussion on this matter. Just to illustrate this, we consider after [109] the following biochemical reactions:

$$S_1 \underset{k_2}{\overset{k_1}{\rightleftharpoons}} S_2 \xrightarrow{k_3} S_3. \tag{2.42}$$

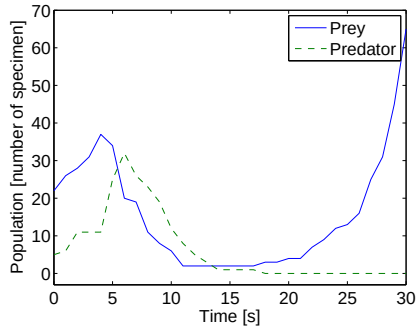
On one hand, when the initial values of S_1, S_2, and S_3 are 100, 0, and 0, respectively, and the parameter values of this system are set as in [109],



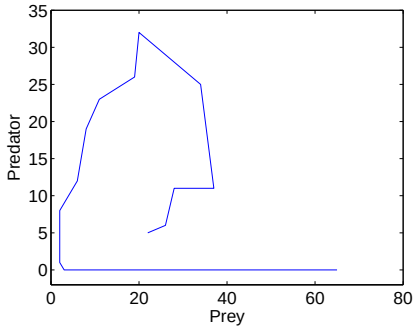
(a) Deterministic time-course simulation



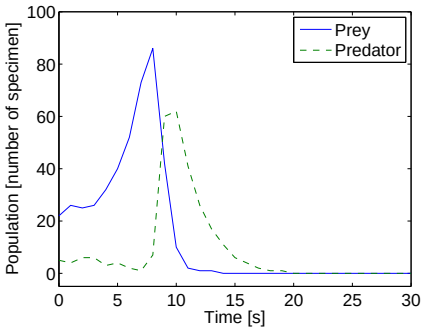
(b) Deterministic phase plane trajectory



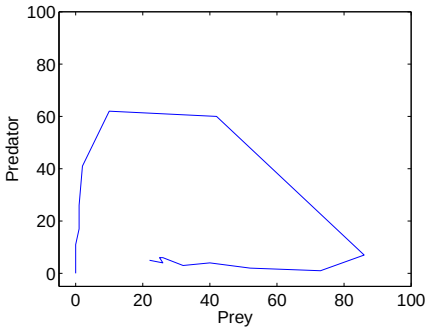
(c) Stochastic time-course simulation: prey population grows unlimited



(d) Stochastic phase plane trajectory: prey population grows unlimited

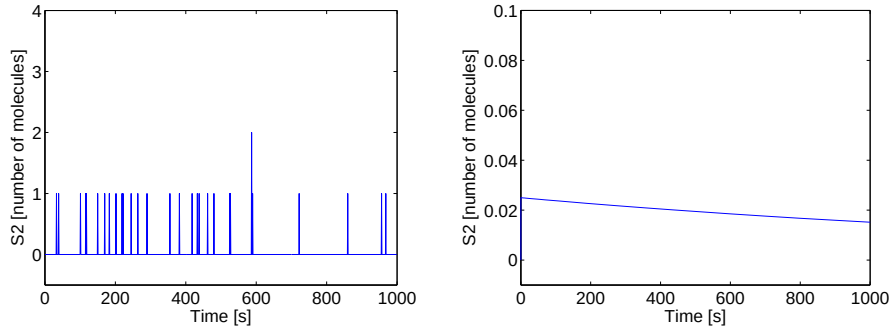


(e) Stochastic time-course simulation: both populations become extinct



(f) Stochastic phase plane trajectory: both populations become extinct

Figure 2.3: Comparison between the deterministic and stochastic modelling of the Lotka-Volterra system. The number of specimen of the prey and predator populations are 22 and 5, respectively. The deterministic trajectory (b) differs significantly from the two stochastic realizations in (d) and (f).



(a) Stochastic time-course simulation

(b) Deterministic time-course simulation

Figure 2.4: Comparison between the stochastic and deterministic modelling of the system in (2.42). (a) A stochastic time-course simulation predicts that the population of the species S_2 is 0 most of the time, sometimes it is 1, occasionally it is 2. (b) A deterministic time-course simulation predicts the level of S_2 to be constantly very close to 0.

i.e. $k_1 = 10$, $k_2 = 4 \cdot 10^4$, and $k_3 = 2$, a stochastic simulation predicts that the population of the species S_2 is 0 most of the time, sometimes it is 1, occasionally it is 2 or 3, and rarely anything more (see Figure 2.4a). On the other hand, the associated deterministic mass-action model predicts the level of S_2 to be constantly very close to 0 (Figure 2.4b). However, if the biochemical reaction system in (2.42) were to be interpreted as in [109], where it is used to model a certain aspect of the heat shock response in the bacterium *Escherichia coli*, then the reaction $S_2 \to S_3$ would correspond to an important event of gene expression. It would be then of interest to track precisely when this event happens, hence motivating a look at the statistics of the S_2 molecule, rather than obtaining an averaged behaviour of this quantity as in the case of a deterministic description of the system.

Chapter 3

Case studies

3.1 The eukaryotic heat shock response

The heat shock response (HSR) is an ancient, highly evolutionary conserved defence mechanism ([75]). It is a global regulatory network found in virtually all living cells[1]. It allows the cell to quickly react to elevated temperatures by the induction of some dedicated proteins called *heat shock proteins* (hsp). Exposure to raised temperature leads to protein misfolding. Misfolded proteins accumulate and tend to form aggregates with disastrous effect for the cell. Stress conditions can be caused not only by increased temperature but also by other forms of environmental, chemical or physical stress, such as addition of ethanol, heavy metals, pollutants, high osmolarity, starvation, etc. The heat shock proteins act as chaperones – they stabilize proteins and refold the denatured ones. They maintain the proper functioning of the cell by preventing the formation of cytotoxic aggregates.

The heat shock response has been the subject of active research, see [101, 18, 134], for at least two reasons. On one hand, as it represents an exceptionally well-conserved regulatory mechanism, it is a good candidate for deciphering the mechanistic and engineering principles underlying gene regulatory networks. On the other hand, heat shock proteins, regardless of the regulatory aspects of the heat shock response, have fundamental importance for many key biological processes such as protein biogenesis, dismantling of damaged proteins, activation of immune responses, and signaling ([60, 100]). Therefore, understanding the details of the heat shock response has broad ramifications for the the biology of the cell. In particular, it would give

[1]Examples of species that lack the classical heat shock response are the Antarctic ciliate *Euplotes focardii* ([129]), an Antarctic sea star *Odontaster validus*, an Antarctic gammarid *Paraceradocus gibber* ([21]), and an Antarctic notothenioid fish *Trematomus bernacchii* ([53]).

better insight into the response to cellular insults and the onset of a number of diseases, including neurodegenerative disorders, cancer, aging, and cardiovascular diseases, see [77, 78, 140, 8, 88].

Although a number of mathematical models describing the heat shock response both in eukaryotes as well as in bacteria have been presented in the literature, see [97, 94, 107, 32, 122, 76, 30, 59, 106, 128], still a comprehensive mechanistic understanding of this process is lacking. In [96], a new simple model was proposed, which captures in mechanistic details all key aspects of the regulation: the heat-induced protein misfolding, the chaperone activity of heat shock proteins, the transactivation of the genes encoding heat shock proteins and the repression of their transcription once the stress is removed. Unlike other previous models, it is based solely on well-documented biochemical reactions and does not include modelling "blackboxes" such as experimentally unsupported components or biochemical reactions. A detailed discussion on the differences between the model in [96] and the previous attempts to model the eukaryotic heat shock response can be found therein.

3.1.1 Biochemical model

In the model of [96], the central role is played by the heat shock proteins (hsp), which act as chaperones for the misfolded proteins (mfp): the heat shock proteins sequester the misfolded proteins ($\mathsf{hsp}\colon\mathsf{mfp}$) and help the misfolded proteins to regain their native conformation (prot). The defence mechanism is controlled through the regulation of the transactivation of the hsp-encoding genes. The transcription is initiated by heat shock factors (hsf), some specific proteins which first form dimers (hsf_2), then trimers (hsf_3) and in this configuration bind to the heat shock elements (hse), i.e. certain DNA sequences in the promotor regions of the hsp-encoding genes. Once the trimers bind to the promoter elements ($\mathsf{hsf}_3\colon\mathsf{hse}$), the transcription and translation of the hsp-encoding genes boosts and, in consequence, new heat shock protein molecules get synthesized at a substantially augmented rate.

When the amount of the heat shock proteins reaches a sufficiently high level that enables coping with the stress conditions, the production of new chaperone molecules is switched off by the excess of the heat shock proteins. To this aim hsp form complexes with the heat shock factors ($\mathsf{hsp}\colon\mathsf{hsf}$) in three independently and concurrently running processes: 1) by binding to the free hsf, 2) by breaking the dimers and trimers, and 3) by breaking the $\mathsf{hsf}_3\colon\mathsf{hse}$, in result of which the trimer gets unbound from the DNA, it is decomposed into three free hsf molecules and one of these hsf molecules forms a complex with hsp. This terminates the enhanced production of new heat shock protein molecules and blocks the formation of new hsf trimers.

As soon as the temperature increases, proteins present in the cell start misfolding. The misfolded proteins titrate hsp away from the $\mathsf{hsp}\colon\mathsf{hsf}$ com-

Reaction	(Reaction number)
$2\,\mathsf{hsf} \leftrightarrows \mathsf{hsf}_2$	$(r1)$
$\mathsf{hsf} + \mathsf{hsf}_2 \leftrightarrows \mathsf{hsf}_3$	$(r2)$
$\mathsf{hsf}_3 + \mathsf{hse} \leftrightarrows \mathsf{hsf}_3\colon \mathsf{hse}$	$(r3)$
$\mathsf{hsf}_3\colon \mathsf{hse} \rightarrow \mathsf{hsf}_3\colon \mathsf{hse} + \mathsf{hsp}$	$(r4)$
$\mathsf{hsp} + \mathsf{hsf} \leftrightarrows \mathsf{hsp}\colon \mathsf{hsf}$	$(r5)$
$\mathsf{hsp} + \mathsf{hsf}_2 \rightarrow \mathsf{hsp}\colon \mathsf{hsf} + \mathsf{hsf}$	$(r6)$
$\mathsf{hsp} + \mathsf{hsf}_3 \rightarrow \mathsf{hsp}\colon \mathsf{hsf} + 2\,\mathsf{hsf}$	$(r7)$
$\mathsf{hsp} + \mathsf{hsf}_3\colon \mathsf{hse} \rightarrow \mathsf{hsp}\colon \mathsf{hsf} + \mathsf{hse} + 2\,\mathsf{hsf}$	$(r8)$
$\mathsf{hsp} \rightarrow$	$(r9)$
$\mathsf{prot} \rightarrow \mathsf{mfp}$	$(r10)$
$\mathsf{hsp} + \mathsf{mfp} \leftrightarrows \mathsf{hsp}\colon \mathsf{mfp}$	$(r11)$
$\mathsf{hsp}\colon \mathsf{mfp} \rightarrow \mathsf{hsp} + \mathsf{prot}$	$(r12)$

Table 3.1: The list of reactions in the molecular model for the eukaryotic heat shock response in [96].

plexes. This enables the accumulation of free hsf molecules, which in turn form trimers and promote the production of new chaperones. In consequence, the response mechanism gets switched on. The full list of biochemical reactions constituting the biochemical model of [96] is presented in Table 3.1.

The model in Table 3.1 includes three mass conservation relations, see [96], for the total amount of hsf, the total amount of proteins (other than hsp and hsf) in the model, as well as for the total amount of hse:

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

for some mass constants C_1, C_2, and C_3.

In [97], an extended model of the eukaryotic heat shock response is presented. With respect to the basic model of [96], the extended version includes the heat-induced misfolding and chaperone-assisted refolding of both hsf and hsp. The justification for this extension is that since both hsf and hsp are proteins, they are exposed to heat-induced misfolding. In this way,

the extended version of the model contains one of the most attractive features of living cells: the repair mechanism is subject to failure, but it has capabilities to repair itself. The extended biochemical model is obtained from the basic one by adding the following 6 new reactions to the list in Table 3.1:

$$\mathsf{hsf} \rightarrow \mathsf{mhsf} \tag{$r13$}$$

$$\mathsf{hsp} \rightarrow \mathsf{mhsp} \tag{$r14$}$$

$$\mathsf{hsp} + \mathsf{mhsf} \leftrightarrows \mathsf{hsp{:}\,mhsf} \tag{$r15$}$$

$$\mathsf{hsp} + \mathsf{mhsp} \leftrightarrows \mathsf{hsp{:}\,mhsp} \tag{$r16$}$$

$$\mathsf{hsp{:}\,mhsf} \rightarrow \mathsf{hsp} + \mathsf{hsf} \tag{$r17$}$$

$$\mathsf{hsp{:}\,mhsp} \rightarrow 2\,\mathsf{hsp} \tag{$r18$}$$

and by substituting reaction $(r4)$ with

$$\mathsf{hsf_3{:}\,hse} \rightarrow \mathsf{hsf_3{:}\,hse} + \mathsf{mhsp}\,. \tag{$r4'$}$$

For more details concerning the extended model, its associated mathematical model and the numerical setup as well as its analysis and validation we refer to [97]. The equivalence between the extended and basic model in terms of numerical behaviour is shown in [95] and is a matter under discussion in Chapter 4.

3.1.2 Mathematical model

By assuming the law of mass action for all reactions $(r1)$-$(r12)$ in Table 3.1 the associated mathematical model consisting of a system of ordinary, nonlinear differential equations is obtained. The rate coefficient of protein misfolding in reaction $(r10)$ is denoted as $\varphi(T)$ and given by the following formula

$$\varphi(T) = (1 - \frac{0.4}{e^{T-37}}) \cdot 1.4^{T-37} \cdot 1.45 \cdot 10^{-5}\ s^{-1}, \tag{3.1}$$

where T is the numerical value of the temperature of the environment in Celsius degrees. The formula is valid for $37 \leq T \leq 45$. It is based on experimental investigations of [74, 73] and was originally proposed in [94]. Expression (3.1) in its current form is obtained by adapting the original formula of [94] to seconds (s), which is the time unit of the mathematical model in [96].

The kinetic rate constants and the initial values of all reactants are obtained by performing extensive parameter estimations. For this purpose a suite of diverse algorithms available in COPASI, a software application for simulation and analysis of biochemical networks and their dynamics ([56]), is used. The numerical setup of the mathematical model is determined so as to satisfy the following three conditions.

1. The model should exhibit no response in the absence of heat shock, i.e. at 37 °C. Hence, for a temperature of 37 °C, the system is at steady state, i.e. the differentials of all model variables are zero.

2. For a temperature of 42 °C, the numerical prediction of the model for $[\mathsf{hsf_3}\colon \mathsf{hse}](t)$ should be in agreement with experimental data of [65] on DNA binding of $\mathsf{hsf_3}$.

3. For a temperature of 42 °C, the numerical prediction of the model for $[\mathsf{hsp}](t)$ should be correlated with experimental data of [96] on a *de-novo* fluorescent reporter-based experiment.

The full list of differential equations constituting the mathematical model together with the numerical values of the rate constants and initial concentrations can be found in [96].

In [85] the basic HSR model of [96] is modelled with the stochastic framework. The dynamics of the HSR is viewed as a continuous-time Markov chain. A proof of the existence and uniqueness of the process' stationary distribution is presented. The outcomes of 1000 stochastic simulations are compared with the results of the deterministic model. For details we refer to Section 4.1.4 and [85].

3.2 *In vitro* self-assembly of intermediate filaments

One of the characteristics of eukaryotic cells is the existence of the cytoskeleton – an intricate network of protein filaments that extends throughout the cytoplasm. It enables the cells to adopt a variety of shapes, interact mechanically with the environment, organize the many components in their interior, carry out coordinated and directed movements. It also provides the machinery for intracellular movements, e.g. transport of organelles in the cytoplasm and the segregation of chromosomes at mitosis ([2, 3]). There are three kinds of protein filaments that form the cytoskeleton: 1) actin filaments, 2) intermediate filaments (IFs), and 3) microtubules. Each kind has different mechanical properties and is assembled from an individual type of proteins. Actin filaments and microtubules are formed from *globular* proteins (*actin* and *tubulin* subunits, respectively), whereas *fibrous proteins* are the building blocks of intermediate filaments ([3, 47]). Thousands of these basic elements assemble into a construction of girders and ropes that spreads throughout the cell.

One of the main functions of intermediate filaments is to provide cells with mechanical strength. Intermediate filaments are especially prominent in the cytoplasm of cells that are exposed to conditions of mechanical tension. For example, IFs are abundantly present along nerve cells axons where they provide crucial internal reinforcement of these long cell extensions.

They can also be observed in great number in muscle cells and epithelial cells. IFs are characterized by great tensile strength. By stretching and distributing the effect of locally applied forces, they protect cells and their membranes against breaking due to mechanical shear. Compared with microtubules and actin filaments, IFs are more stable, tough and durable. For example, they remain intact during exposure of cells to salt solutions and nonionic detergents, while the rest of the cytoskeleton is mostly destroyed ([2]).

Intermediate filaments can be grouped into four classes: (1) *keratin filaments* in epithelial cells; (2) *vimentin filaments* in connective-tissue cells, muscle cells and supporting cells of the nervous system; (3) *neurofilaments* in nerve cells; and (4) *nuclear lamins*, which strengthen the nuclear membrane of all eukaryotic cells, see [2]. Major degenerative diseases of skin, muscle, and neurons are caused by disruptions of the IF cytoskeleton or its connections to other cell structures.

Unlike the other protein filaments which are assembled from globular proteins, see [55, 124, 34], IFs subunits are α-helical rods that assemble into rope-like filaments ([48]). Their assembly proceeds through a series of intermediate structures, which associate by lateral and end-to-end interactions. However, unlike in the case of microtubules and actin filaments where rich literature is available, the assembly principles of IFs, either *in vitro* or *in vivo*, are still poorly understood. In [62] and [25] the quantitative kinetic strategies for the *in vitro* assembly of IFs from human tetrameric vimentin proteins are analysed. In general, the *in vitro* assembly of vimentin IF proteins can be described as a process consisting of three major phases: (i) formation of the unit-length filaments (ULFs); (ii) longitudinal annealing of ULFs and growing filaments; (iii) radial compaction of immature (16 nm diameter) filaments into mature (11 nm diameter) IFs ([50, 51]). In the first phase of their assembly, vimentin proteins rapidly associate parallelly into dimers and then form anti-parallel, half-staggered tetramers, see [49]. Subsequently, tetramers rapidly associate laterally to yield short filaments called unit-length filaments of the same length as the tetramers, see [48]. The first phase of assembly is illustrated in Figure 3.1. In the second phase of the assembly, the ULFs and the emerging longer filaments elongate longitudinally with tetramers, with ULFs, and with other filaments ([48]). The third phase is not considered in the biochemical models introduced in [62], analysed in [62, 25] and briefly presented in the following.

3.2.1 Biochemical models

The *simple model* of [62] treats ULFs as ordinary filaments and describes the assembly process through a sequence of biochemical events as follows.

(i) two tetramers (denoted T) associate laterally into an octamer (denoted

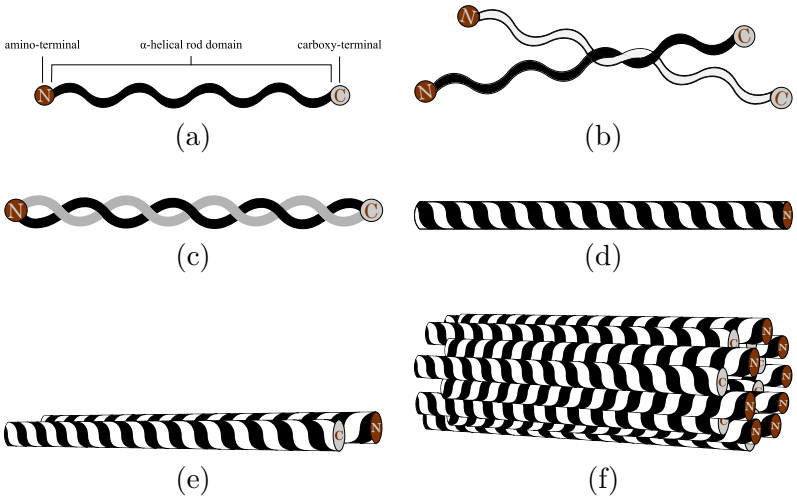


Figure 3.1: The first phase of the assembly of human vimentin proteins. Intermediate filament subunits are α-helical rods, that associate parallelly into coiled-coil dimers, which in turn form anti-parallel, half-staggered tetramers. Tetramers rapidly associate laterally to yield the shortest filaments called *unit-length filaments* (ULFs) of the same length as the tetramers. (a) α-helical rods, (b) dimerization of α-helical rods, (c) coiled-coil dimer, (d) another representation of a coiled-coil dimer, (e) tetramer, (f) ULF. Reprint of the illustration originally presented in [25].

O):

$$2\,\mathrm{T} \rightarrow \mathrm{O}; \tag{3.2}$$

(ii) two octamers associate laterally to yield a hexadecamer (denoted H):

$$2\,\mathrm{O} \rightarrow \mathrm{H}; \tag{3.3}$$

(iii) two hexadecamers associate laterally to form a (unit length) filament (denoted F):

$$2\,\mathrm{H} \rightarrow \mathrm{F}; \tag{3.4}$$

(iv) a tetramer associates longitudinally to a filament to yield an elongated filament:

$$\mathrm{F} + \mathrm{T} \rightarrow \mathrm{F}; \tag{3.5}$$

(v) two filaments associate longitudinally to yield an elongated filament:

$$\mathrm{F} + \mathrm{F} \rightarrow \mathrm{F}\,. \tag{3.6}$$

The *extended model* of [62] adds a distinction between minimal-length filaments (ULFs, denoted U) and longer filaments (consisting of at least two ULFs), treating them as distinct species in the model. In terms of biochemical events, the extended model consists of the following reactions:

(i') two tetramers (denoted T) associate laterally into an octamer (denoted O):

$$2\,\mathrm{T} \to \mathrm{O}; \tag{3.7}$$

(ii') two octamers associate laterally to yield a hexadecamer (denoted H):

$$2\,\mathrm{O} \to \mathrm{H}; \tag{3.8}$$

(iii') two hexadecamers associate laterally to form a unit length filament (denoted U):

$$2\,\mathrm{H} \to \mathrm{U}; \tag{3.9}$$

(iv') two unit length filaments associate longitudinally to form an elongated filament (denoted F):

$$2\,\mathrm{U} \to \mathrm{F}; \tag{3.10}$$

(v') a filament is elongated longitudinally with a tetramer:

$$\mathrm{F} + \mathrm{T} \to \mathrm{F}; \tag{3.11}$$

(vi') a filament is elongated longitudinally with a unit length filament:

$$\mathrm{F} + \mathrm{U} \to \mathrm{F}; \tag{3.12}$$

(vii') two filaments associate longitudinally to yield an elongated filament:

$$\mathrm{F} + \mathrm{F} \to \mathrm{F}\,. \tag{3.13}$$

Models for IFs self-assembly providing distinction between filaments of particular lengths up to n, where n is an arbitrary integer determining the so-called *model resolution*, are introduced in [25]. The problem of increasing and decreasing the resolution of self-assembly models in general is further investigated in [84].

3.2.2 Mathematical models

As for the heat shock response model, also in this case a mathematical formulation based on the mass-action law is considered both for the simple and the extended model of IFs self-assembly. The full lists of differential equations for both models can be found in [25]. The models are fit by performing extensive parameter estimation in COPASI ([56]) with respect to one set of experimental data of [62] and validated against another set. The fit and validation are performed in two cases: with and without taking into account a qualitative property of the IF assembly, reported in [62], that very quickly (within approximately 10 seconds) after the initiation of the assembly, ULF is the most predominant species in the system. Moreover, in the case of the extended model several different knockdown mutant model variants are considered, where various combinations of assembly mechanisms (the so-called strategies) are analysed separately. The performed study provides several conclusions regarding the kinetics of the *in vitro* assembly of human vimentin, see [62, 25] for details.

Relating the models for IF assembly to the quantitative experimental data on the dynamics of the filament length is non-trivial because the considered models do not represent explicitly the information about the length of the emerging filaments. Indeed, the models collect all filaments into a single variable (F), regardless of their length. However, as is shown in [62] and later improved in [25], the dynamics of the mean filament length (MFL) can be deduced based on the variables of the models. For the details and derivations we refer to [25].

As observed in [25], an interesting aspect of the mathematical models is that the mass conservation relation on the total number of tetramers in the models is evident in the molecular models (since there is no synthesis and no degradation in the models), whereas it cannot be deduced as a property of the corresponding mathematical models. This is a consequence of how, for example, the longitudinal association of two filaments is modelled: the information about the lengths of the two input filaments is not explicitly reproduced in a property of the two filaments. However, one can calculate the number of tetramers integrated in the assembled filaments and then use this quantity to reason about the time-dependant dynamics of the mean filament length.

Finally, we notice that the presented models are valid for the early dynamics of the vimentin filament assembly, where the kinetics of the system is fast, with tetramers and ULFs being quickly replaced by emerging filaments of various lengths. During this phase, the presence of a large amount of tetramers and, a little later, of short filaments in the solution make far more likely assembly/elongation events rather than disassembly events. For this reason our models prove to be able to explain the experimental data

during the early phase of the assembly, even though they do not include any disassembly or filament breaking mechanisms. The applicability of the models is, however, tied to the early part of the assembly. Over longer time intervals the lack of a disassembly mechanism in the models makes them limited in their predictive power. For example, a model with no disassembly or filament breaking mechanism would predict that the system will reach (albeit in a huge interval of time) a steady state where all initial tetramers are integrated into one single filament (of huge length).

Chapter 4

Computational modelling challenges

This chapter gathers and presents in a synthetic way a number of issues being subject of the publications included in this thesis and constituting the original contribution of the author. From a general perspective, the contribution consists in perceiving biological systems, mechanisms, processes, etc., as complex arrangements whose properties emerge from much simpler elements interrelated through logical, cause-effect types of relations. As described in Chapter 1, such perception and description of various problems is characteristic for the field of computer science. Hence, seeing biological problems and designs from this perspective makes a natural choice the use of methods, tools and formalisms originally developed in scientific disciplines of computer science and mathematics for the analysis and description of constructs and processes encountered in biology. In this way, the mentioned fields of science serve an auxiliary role for biology. However, also computer science and mathematics extensively benefit from this interdisciplinary cooperation. As mentioned previously, the complexity of designs and arrangements encountered in biological systems, in particular living cells, exceeds virtually any other structure or organization of inanimate matter one could think of. Application of methods and formalisms developed in the fields of computer science and mathematics creates new problems and imposes challenges. On one hand, these issues enforce tuning and advancement of methods and tools of computer science. On the other hand, they display the limitations of these techniques and, in consequence, stimulate the work out of new approaches towards the analysis of complex systems. At the same time this cooperation enables us to better understand what computer science in fact is: what can be explained based on its perspectives of viewing problems and what are the limitations of its approaches. In this sense, although this research is clearly driven by biolog-

ical questions, assumptions, and data, it is highly beneficial for computer science.

In the research constituting the foundations of this thesis a broad range of approaches has been developed and applied to systematically investigate two biological phenomena: the eukaryotic heat shock response and the *in vitro* self-assembly of intermediate filaments, both briefly described in Chapter 3. This range spans from heuristic, through numerical and statistical to analytical methods applied in the effort to formally describe and analyse the two cellular processes. In particular, we present a number of various methodologies and heuristics relevant to the process of generally understood model analysis. The considered techniques pertain to issues such as model construction, the decomposition of a model into certain components and identification of the contribution of each of them to the overall behaviour of the system, simplification or extension of a model, as well as the problem of comparison between various submodels, which is part of an important and difficult question of how models should be compared. A review of existing methodologies with reference to appropriate literature is provided and the original contribution of the author in this subject is indicated. Although the methodologies presented here are mostly discussed and applied in the context of biological systems, we notice that they are, or can be with relatively small amount of effort made useful in different, i.e. other than biological, setups as well. The full range of developed and applied modelling techniques as well as model analysis methodologies described in this chapter constitutes a rich modelling framework.

4.1 Model construction techniques

In developing the mathematical model of the eukaryotic heat shock response we concentrated on constructing a simple model capturing in mechanistic details all key aspects of the regulation: the heat-induced protein misfolding, the chaperone activity of heat shock proteins, the transactivation of the genes encoding heat shock proteins and the repression of their transcription once the stress is removed. In consequence, the resulting model is based solely on well-documented molecular reactions (based on standard molecular biology only) and does not include modelling "blackboxes" such as experimentally unsupported components or biochemical reactions. In this way, the intricate process is viewed from a computer science perspective as a logical arrangement of relatively simple building blocks, i.e. biochemical species, interrelated through cause-effect interactions described in the form of biochemical reactions. We notice however that, although the reactions are simple rules describing the relations between species, they may encapsulate complicated processes from the point of view of molecular biology and

biochemistry. For example, the basic heat shock response model presented in [96] contains a reaction representing the transcription and translation of hsp-encoding genes in the form $\mathsf{hsf}_3 : \mathsf{hse} \rightarrow \mathsf{hsf}_3 : \mathsf{hse} + \mathsf{hsp}$, see Section 3.1.1 and reaction $(r4)$ in Table 3.1. In consequence, we do not model explicitly the transcription machinery binding to the promoter region of the hsp-encoding gene, the mRNA molecules being produced, edited, transported, etc., but only consider that a transcriptionally active hsp-encoding gene will eventually yield the synthesis of new hsp molecules. In this sense, this reaction captures in a simple and compact way the important, from the perspective of the analysis of the heat shock response mechanism, relationship between the transcription machinery bound to the DNA and the enhanced production of heat shock proteins. At the same time it hides the intricate details of the gene transcription and translation process. These considerations are part of a broader context of how to define the scope and choose the abstraction level of a model. It is crucial to carefully consider the purpose of model building: whether it is to obtain a very detailed and accurate replica of the system under study or rather the aim is to construct an abstract model which is relatively simple to analyse yet able to capture the essential characteristic of the system. Depending on which purpose the model is to serve, a suitable choice of the abstraction level can be made and proper modelling frameworks can be applied. We come back to this issue in Section 4.1.4, where we address the problem of proper choice between the deterministic, continuous modelling framework versus the stochastic, discrete one. We also discuss this matter in Section 4.3 in the context of model modifications techniques, as well as Chapter 6.

The full list of biochemical reactions presented in Table 3.1 in Section 3.1.1 constitutes a biochemical model for the eukaryotic heat shock response. The associated mathematical model is obtained based on the mass-action kinetics, see Section 3.1.2. The reason why a simple mass-action formalization rather than more sophisticated approaches such as Michaelis-Menten or Hill equations is chosen is so that we can follow the explicit effect of each individual reaction to the overall response, i.e. to serve our goal of constructing a model from simple and well-specified building-blocks.

Although constructed from plain elements, the behaviour of the resulting mathematical model is far from being simple and is characterized by emergent properties. In response to external stimulus in the form of increased temperature (above 37 °C, which is considered the physiological conditions temperature), the model correctly predicts that the level of hsf trimers is transiently increased, see [54, 96]. It is also able to confirm that the hsf dimers are only a transient state between monomers and trimers and that their level remains low at all times, independent of the temperature. Next, in the case of a heat shock applied in two stages with a recovery period between them, with the second shock applied after the level of hsp has reached a max-

imum, the predicted response of the model to the second heat shock is much milder. This is in complete agreement with the expectation that due to the first heat shock, the level of hsp is already raised, and so the cell may react to the second shock with a lower $\mathsf{hsf}_3 : \mathsf{hse}$ peak. Further, when a heat shock at 43 °C is considered, our model is able to show prolonged transactivation and an experiment where the heat shock at 42 °C is removed at the peak of the response shows a faster attenuation phase, see [96] for a more detailed discussion of these results. All these make the model especially interesting and outstanding: the presented properties of the model are emergent, i.e the model is neither biased nor tweaked towards any of these behaviours by any artificial, undocumented mechanisms. Just to provide an example, in the case of the model in [94], mRNA is not produced as a result of DNA transcription and it is not used directly in a model for protein synthesis, which in our case is the crucial feedback regulatory motif. Instead, mRNA is used in a hypothetical reaction of binding to misfolded proteins. Such a reaction leaves only part of mRNA molecules as "healthy" and their proportion is then used to model the slowing reaction rate of hsp binding to nascent protein chains. First, many of these steps lack experimental support. Second, the same effect can be obtained, as suggested by our model, based on the observation that hsp molecules are competed on, according to the mass-action principle, both by misfolded proteins and by nascent proteins chains. For a detailed comparison of the eukaryotic HSR model originally introduced in [96] with other models described in the literature we refer to [96].

4.1.1 Parameter estimation

The mass-action model of the heat shock response is expressed in terms of ten, ordinary, first-order, non-linear differential equations ([96]). However, based on the three mass-conservation relations concerning the total amount of hsf, proteins other than hsp and hsf, and heat shock elements, only seven equations turned to be independent ([96] and Section 3.1.1). There are 17 independent parameters in the model and 10 initial conditions that must be specified or estimated to fit experimental data, but in fact the three conservation relations leave only seven initial conditions to specify. In addition, the condition that with the same initial values and the same numerical parameters the model is at steady state if the temperature is 37 °C is imposed. This is a natural condition since the model is supposed to reflect the reaction to temperatures raised above 37 °C. This yields 7 independent algebraic relations on the set of parameters and initial values. Thus, we have altogether 17 independent values that we need to estimate. Mathematically, the problem we need to solve is one of global optimization, as formulated below. For each 17-tuple κ of positive numerical values for all kinetic constants, and for each 10-tuple α of positive initial values for all variables in the model, the

function describing the level of DNA binding in time is uniquely defined for a fixed temperature T. We denote the value of this function at time point τ, with parameters κ and α by $x^T(\kappa, \alpha, \tau)$. Note that this property holds for all the other variables in the model and it is valid in general for any mathematical model based on ordinary differential equations (one calls such models *deterministic*). We denote the set of experimental data by

$$E_n = \{(t_i, r_i) \mid t_i, r_i > 0,\ 1 \leq i \leq N\},$$

where $N \geq 1$ is the number of observations, t_i is the time point of each observation and r_i is the value of the reading.

With this setup, we formulate our optimization problem as follows: find $\kappa \in \mathbb{R}_+^{17}$ and $\alpha \in \mathbb{R}_+^{10}$ such that:

(i) $f(\kappa, \alpha) = \frac{1}{N} \sum_{i=1}^{N} (x^{42}(\kappa, \alpha, t_i) - r_i)^2$ is minimal and

(ii) α is a steady state of the model for $T = 37$ and parameter values given by κ.

The function $f(\kappa, \alpha)$ is a cost function (*least mean squares* in the case of the HSR model), indicating numerically how the function $x^T(\kappa, \alpha, t)$, $t \geq 0$, compares with the experimental data. As mentioned above, not all 27 variables (the components of κ and α) are independent: on one hand, we have the three algebraic relations defining the mass conservation relations and, on the other hand, we have seven more independent algebraic relations given by the steady state equations. Consequently, we have 17 independent variables in our optimization problem.

Given the high degree of the system, finding the analytical form of the minimum points of $f(\kappa, \alpha)$ is very challenging. This is typical when the system of equations is non-linear. Adding to the difficulty of the problem is the fact that the seven independent steady state equations cannot be solved analytically, given their high overall degree.

Since an analytical solution to the model fitting problem is often intractable, the practical approach to such problems is to give a numerical simulation of a solution. Several methods exist for this, see [14, 102]. The trade-off with all these methods is that typically they offer an estimate of a *local* optimum, with no guarantee of it being a *global* optimum.

Obtaining a numerical estimation of a local optimum for (i) is not difficult. However, such a solution may not satisfy (ii). To solve this problem, for a given local optimum $(\kappa_0, \alpha_0) \in \mathbb{R}_+^{17} \times \mathbb{R}_+^{10}$ one may numerically estimate a steady state $\alpha_1 \in \mathbb{R}_+^{10}$ for $T = 37$. The pair (κ_0, α_1) then satisfies (ii). Unfortunately, (κ_0, α_1) may not be close to a local optimum of the cost function in (i).

Another approach is to replace the algebraic relations implicitly given by (ii) with an optimization problem similar to that in (i). Formally, we

replace all algebraic relations $R_i = 0$, $1 \leq i \leq 7$, given by (ii) with the condition that

$$g(\kappa, \alpha) = \frac{1}{M} \sum_{j=1}^{M} R_i^2(\kappa, \alpha, \delta_j)$$

is minimal, where $0 < \delta_1 < \cdots < \delta_M$ are some arbitrary (but fixed) time points. Our problem thus becomes one of optimization with cost function (f, g), with respect to the order relation $(a, b) \leq (c, d)$ if and only if $a \leq c$ and $b \leq d$. In the case of our basic ([96]), as well as the extended ([97]) HSR model the solution to the above problem is obtained based on COPASI [56].

In [96] yet another method for this challenging problem of finding parameters that simultaneously implement the stress-induced response of the model (i) and satisfy the steady state condition (ii) is proposed. This method is of the local optimization type and is based on the following two observations. First, the steady state of the model is a function of the parameters and of other variables, such as total mass of various species. Second, the model is continuous in all of the parameters. The main reason why parameter estimation is the most time-consuming part of the work presented in [96] is due to the problem that once a good fit with respect to experimental data is found, the utilized approach is to replace the initial values with the steady state of the obtained model at 37 °C and hope that the model fit at 42 °C is not destroyed. The new idea proposed in the discussion of [96] is that for models for which the mentioned two observations are true, a systematic parameter scan in the space determined by the considered ranges of parameter values can be applied to identify a region in the multidimensional parameter space where a local minimum of the score function is found. Iterating this procedure yields a realization of a local minimum of the score function, while the initial state of the model is a steady state for a temperature of 37 °C. In cases where the direct implementation of this idea is intractable, i.e. for models with more than a few parameters due to the combinatorial explosion of the number of simulations that need to be run, a fast and practical solution is to apply the Latin Hypercube Sampling method (LHS), first introduced in [80]. We describe the sampling scheme briefly in the following, in the case when the parameter values are uniformly distributed in their range interval. One first chooses the desired size Z of the sampling set. The range interval of each parameter is then partitioned into Z non-overlapping intervals of equal length. For each parameter, we randomly select Z numerical values, one from each interval of the partition. We collect the Z sampled values for the i-th parameter of the model on the i-th column of a $Z \times p$ matrix, where p is the number of parameters. One then randomly shuffles the values on each column. The result of the procedure is read from the rows of the matrix: each of the Z rows of the matrix contains numerical values

for each of the p parameters. For a detailed description and applications of this sampling scheme we refer to [80, 45, 46, 92, 96, 83].

4.1.2 Model validation

Model validation is a crucial step in the model development methodology. It enables gaining trust in the predictive power of the model and, in consequence, makes credible potential hypothesis formulated on the basis of model predictions. Such hypotheses are supposed to be subject to experimental verification and hence stimulate new experimental designs. This in turn advances our understanding of the process under consideration and subsequently leads to further model tuning and development. Hence, model validation finds its important place in the iterative circle of systems biology, see Section 2.1.

In the case studies of the eukaryotic heat shock response and *in vivo* self-assembly of intermediate filaments a number of model validation instances are presented, see [96, 97, 25]. These validations are based both on quantitative data as well as qualitative knowledge. One of the more challenging issues is the validation of the basic HSR model described in [96] with respect to newly obtained experimental data. Specifically, the aim is to validate the numerical prediction on the level of hsp of the model over time. Our approach is to use a suitable quantitative reporter system based on yellow fluorescent proteins (yfp). Our experimental setup is designed so that the kinetics of the reporter geneŠs transactivation mimic the results obtained in experimental studies on endogenous hsf target genes. In this way, the dynamics of yfp partially reports on the dynamics of hsp. No assumptions are made on the stability of yfp genes. Rather, this issue is dealt with in the mathematical validation process. Our assumption is that the fluorescence intensity is roughly linear with respect to the level of the yellow fluorescent proteins (yfp). The idea of the validation is to extend the already fit basic model so as to include also yfp. Given that the transactivation of the yfp genes is controlled by their own heat shock elements hse', transcription/translation and degradation kinetics, we obtained that

$$d[\mathsf{yfp}]/dt = k'_4[\mathsf{hsf}_3\colon \mathsf{hse}'] - k'_9[\mathsf{yfp}], \tag{4.1}$$

for some positive constants k'_4, k'_9 standing for the kinetic rate constants of the yfp synthesis and of the yfp degradation, respectively. In the extended model we re-use all the kinetic rate constants of the basic model. We then look for numerical values for parameters k'_4 and k'_9 and for initial values of all variables of the model so that the numerical prediction for yfp fit well with the experimental data. The numerical values of parameters k'_4 and k'_9 are not deduced from the basic model to underline that we make no assumptions on the stability of yfp, or on their gene transcription rates. For more

details, in particular the description of the experimental setup, methods and validation results, we refer to [96]. Importantly, what can be learnt from this approach is the methodology of how to validate against experimental data mathematical models having no variable directly related to the available data. Moreover, it illustrates how to address certain subtle issues related to the use of a reporter system. In particular, how to include the reporter system to an existing model in such a way that the validation procedure is not biased (completely independent rate constants are introduced) and how to deal with the lack of knowledge on the reporters stability or unknown numerical value of their gene transcription rate.

In this context it is worth mentioning that also in the case study of the self-assembly of intermediate filaments the problem of validating the mathematical models with respect to experimental data is not a trivial task. This is due to the fact that the considered models do not represent explicitly the information about the length of the emerging filaments. Thus, relating them to the quantitative data on the dynamics of the filament length is not straightforward and requires some effort to deduce the dynamics of the mean filament length (MFL) based on the variables of the models, as is shown in [25]. Although the idea of relating the mathematical models to the experimental data through the mean filament length was originally presented in [62], the provided mathematical expressions for the MFL introduced an approximation error which is proportional to the length of each filament. The approach in [25] is not influenced by this approximation error and leads to a correct interpretation of the experimental data. For a detailed discussion on this issue we refer to [25].

A methodology for increasing the resolution of the filament self-assembly model is introduced in [25] and further investigated in [84]. We postpone the discussion of this approach to Section 4.3.2, but mention here that this methodology, as is shown in [25], enables the introduction of a high-resolution model for vimentin filament self-assembly, able to capture the detailed dynamics of filaments of arbitrary length. This provides much more predictive power for the model in comparison to previous models, i.e. those in [62, 25], where only the mean length of all filaments in the solution can be analysed. Hence, the resulting model can be directly validated against the raw experimental data of [62], i.e. capturing the lengths of individual filaments, without the need for computing the mean filament length. In this way, this methodology provides means for model validation. For example, if the experimental data contain information on the time-course evolution of the number of objects of certain sizes, than a simpler model under consideration can with use of our methodology be straightforwardly refined to a model of higher resolution, containing variables which numerical evolution in time can be directly related to the experimental data. In this way, the simpler model can be validated and used for further purposes.

4.1.3 Model identifiability problem

Where reaction rate constant values of a mathematical model are obtained by performing parameter estimation with respect to experimental data, there arises the substantial question of the uniqueness of the set of parameters that fulfill the imposed conditions. This is a substantial and commonly encountered problem in systems biology modelling. It is referred to as the *model identifiability problem*. In fact, this issue can be discussed in a broader, more general context: it is related not only to the problem of numerical model fitting but also appears in the problems of model comparison, model simplification, recognition of the contribution of identified modules of a model to its system-level behaviour, and any other problems involving numerical techniques, e.g. sensitivity analysis performed in a specific numerical setup of a mathematical model. Examples of such problems can be found in [96, 95, 83, 24] and we briefly list and summarize them in the following.

In [96], the question of potential alternative fits for the basic HSR model is considered. To this aim a thorough method based on systematic parameter scan in the space determined by the considered ranges of parameter values, e.g. defined through some biological knowledge, is proposed. Based on the Latin Hypercube Sampling method ([80]), the following strategy to look for alternative model fits that are both in agreement with the experimental data of [65] and satisfy the steady-state condition for the initial values is implemented. First, by applying the LHS method, sets of parameter values are sampled. For each set, the steady state of the model under physiological conditions (for a temperature value of 37 °C) is numerically estimated. The initial state of the model as the calculated steady state is then set and the model is simulated in the stress conditions (42 °C). Finally, those parameter samples that lead to low DNA binding level at the peak of the response are classified as non-responsive and excluded from further analysis. For each of the remaining models, for each variable and each parameter a scatter plot is made where characteristic quantitative property values such as the steady state values of the variable under physiological conditions or the model fit scoring function values under stress are plotted against the values of the parameter. The obtained results are compared with the chosen properties values of the basic model and conclusions concerning the identifiability of the model are drawn, see [96] for details. It should be stressed that the discussed methodology does not provide a proof of the uniqueness of parameter values satisfying the imposed conditions. On the contrary, it is likely that a model of this size is in fact not uniquely identifiable. However, the methodology in the case of the basic HSR model shows that finding parameter values satisfying our model constraints is far from being easy.

Also in a series of papers, i.e. [23, 83, 24], focusing on the control-based approach towards model decomposition, analysis and comparison, the issue

of model identifiability, or more generally the dependence of the obtained results on the numerical setups, is recognized and addressed.

In [23] a control driven approach to studying the HSR regulatory network of [96] is taken, the network is decomposed by identifying its main functional modules and three main feedback loops are distinguished. The main question addressed is why such level of complexity is needed for implementing something that, in principle, could also be achieved with an open-loop design. To provide an answer to this question, the numerical behaviour of various knockdown mutants, where one or more feedback loops are missing, is compared. However, as the authors of [23] admit, the results of the analysis remain heavily dependent on the numerical setup of the models, i.e. numerical values of the mass constants and of the kinetic rate constants chosen in [96] for the original (reference) model (for details on how the numerical values of the knockdown mutants were chosen we refer to [23] and to the presentation of local submodels comparison in Section 4.4). To emphasize this problem, the authors of [23] say that their analysis is *local* and state that repeating the analysis in a different numerical setup could result in very different conclusions regarding the role of each of the considered feedbacks. The authors point to the fact that the control in the regulatory network can easily be shifted elsewhere by drastically slowing down some reactions and speeding up others or by changing the mass constants, which is illustrated with an example, see [23] for details. The main conclusions are as follows. First, the local numerical analysis has to be taken in close relationship with the experimental data and available biological knowledge and validated as such. Second, repeating the analysis in different numerical contexts can be very useful for gaining trust in the outcomes of the analysis, however projecting conclusions from one numerical context to another imposes a challenge in itself.

The same issue is encountered in the context of a novel method for identifying the numerical contribution of a component to the system-level behaviour of a larger model that is proposed in [24]. In this case a Boolean logic-based approach for extracting conclusions about the role of each module from the systematic comparison of the numerical behaviour of all knockdown mutants is considered. In this approach a Boolean variable is associated to each module, expressing when the module is included in the architecture (value 'true') and when it is not (value 'false'). For each knockdown mutant a Boolean formula is then written (using the conjunction and negation of the introduced Boolean variables) characterizing the mutant's control architecture, i.e., which of the modules are present in the considered model. The associated Boolean formulas encompass time-independent properties of the models. Moreover, they are parameter independent, i.e. they are not influenced by the parameters used to describe the compared models. Further, the satisfiability of system-level properties of the full model, such as

efficiency, or economical use of resources, is expressed in terms of a Boolean formula indicating in a compact way which model architectures, i.e., which combinations of modules, give rise to the desired property. However, at this stage the parameter independence is lost since, in order to perform numerical simulations of the models, numerical setups for each of the knockdown mutants are needed. This makes the analysis again sensitive to the choice of the numerical values. Again, repeating the analysis for several numerical setups could be a potential, although definitely not completely satisfactory, solution to this challenging problem.

An attempt to handle the problem identified in [23] and [24] of dependence of the submodel comparison results on numerical setup is made in [83]. Therein another original approach for quantitative submodel comparison is proposed, where statistical sampling of the reference model and knockdown mutant behaviours is performed. This method allows to consider more than one numerical context for each submodel. In this way, the conclusions of the analysis are not restricted to some particular values of the kinetic rate constants. Thus, the obtained simulation results provide a basis for comparison between the different potential architectural designs underlying the analysed system. A more detailed presentation and discussion of this method is provided in Section 4.4, where the methodology for submodel comparison is considered. The full presentation of this approach and an example of application in the case study of the basic HSR model of [96] can be found in [83]. Here we just briefly outline this method. First, submodels of the original model are constructed: the considered model is decomposed into modules and a number of knockdown mutants lacking one or more of the modules is considered. Second, the associated mathematical models are obtained. Next, the statistical sampling of the reference model and mutant behaviours is performed by scanning the parameter value space. Further, the initial values of the variables of the reference model and the submodels are determined independently of each other by a systemic property, such as the system being in a steady state in a given setup. Subsequently, numerical simulations are run in order to evaluate the functional effectiveness of the reference model and its knockdown mutants. Finally, the obtained results are summarized and the alternative submodels are compared with the use of some statistical measures. In this way, in this model comparison approach the behaviour of each potential architecture is characterized by some statistical measures summarizing the outcomes of many different numerical setups, as opposed to being considered just in one numerical configuration. Hence, the dependence of the analysis' results on the choice of numerical setup is reduced to a large extent.

Finally, in [95] the problem of dependance on numerical setup appears in the context of model analysis leading to model simplifications. As noticed in [95], all the simplifications that are made on the extended model ([97, 95])

are based on numerical arguments and, in principle, they are dependant on the numerical setup of the model. Thus, in order to address this problem the robustness of the model reductions against changes in the numerical setup is examined. To this aim a set of tests is designed and performed. In each test either the initial values of some variables, or the values of some kinetic rate constants are changed. For each new numerical setup the initial values of all variables are set to their steady state values at 37 °C to underline that the heat shock response is missing under physiological conditions. Finally, comparison of the numerical behaviour of the extended model with that of its simplified version, for temperatures between 37 °C and 45 °C is performed. For the details and outcomes of each particular test we refer to [95]. However, as stated in [95], to evaluate the robustness of the model simplifications in a more comprehensive way, one should compare the two models, i.e. the extended and simplified ones, in several numerical setups, spanning the domain of expected values for the model parameters.

4.1.4 Deterministic versus stochastic modelling framework

As described in Section 2.5, the proper choice of a modelling framework depends on the context and purpose of modelling. Different modelling frameworks provide different levels of abstraction and resolution. Especially in the case of biological systems such as gene expression networks, where the copies of a particular gene are of an order of dozens and transcription factor molecules of a number of few hundreds, the discrete, stochastic framework may provide a more detailed insight into the dynamics of the considered system than just an average tendency captured by the continuous, deterministic description of the system. In [85], a stochastic model corresponding to the basic deterministic mass-action model of [96] for the eukaryotic heat shock response is constructed and the outcomes of these two models are confronted. In particular, the performed analysis shows that in the case of the eukaryotic HSR, the behaviour of the basic model is a very good macroscopic approximation of the mesoscopic dynamics: the stochastic framework does not provide any additional, relevant information with respect to what is already known from the deterministic description. The presented results indicate that the stochastic and deterministic models provide a qualitatively consistent picture of the dynamics of the heat shock response mechanism. This is relevant from the point of view that, in general, performing stochastic simulations is much more expensive in terms of computational resources and time than performing numerical integration of ODEs. The availability of dedicated simulation software equipped with tools for analysis of steadstates, sensitivities, robustness, etc., as well as the expertise in the theory of differential equations makes the deterministic framework more preferable, although the stochastic formulation could seem in many cases more justi-

fied. The analysis of [85] deepens our belief that in the case of the basic model of the eukaryotic HSR the choice of the deterministic formulation does not carry with it any substantial loss in our perception of the heat shock response mechanism.

At the same time, the conducted analysis of a substantial number of performed stochastic simulations, i.e. 1000 independent trajectories, let us gain some more insight into the dynamics of the heat shock response mechanism. In general, the question about the stationarity and stability, i.e. the number of steady-states and whether they are stable or unstable, is important in the examination of the dynamics of biological systems. In [85], by proving the existence and uniqueness of the stationary distribution of the Markov chain underlying the stochastic model, by summarizing the outcomes of the stochastic simulations with some statistics, and by performing clustering and grouping of the stochastic trajectories, the range of behaviour the stochastic model is likely to exhibit is investigated. It is demonstrated that both in the deterministic and stochastic models the conclusions concerning the stability of the system are coherent, i.e. in both case the interpretation of the results indicates that the system is rather monostable. For the formal presentation of the Markov chain underlying the stochastic model of the eukaryotic heat shock response, the proof of the existence and uniqueness of its stationary distribution, the methodology for the analysis and comparison of the dynamics of both the stochastic as well as the deterministic framework, we refer to [85].

4.2 Methods for model decomposition

Much experimental and theoretical effort is invested nowadays in analysing large biochemical systems, e.g. metabolic pathways, regulatory networks, signal transduction networks, aiming to obtain a holistic perspective providing a comprehensive, system-level understanding of cellular behaviour. This often results in the creation and analysis of very large and complex models, often encompassing hundreds of reactions and reactants, see, e.g., [17]. Therefore, obtaining a global picture of the system's architecture, in particular understanding the interactions between various components, or even just distinguishing a high-level functional decomposition of the network, constitutes a significant challenge. Recognizing that similar problems have been encountered, for instance, in engineering sciences ([22]), one strategy towards a system-level understanding of such architectures is to adapt specific methods originating from these disciplines, in particular from control theory, see [42, 63, 71, 120, 121, 125, 136]. Such approach provides a systematic way of identifying the main regulatory components, including feedforward and feedback mechanisms. In consequence, contributes to the

understanding of the reactivity, robustness and efficiency of the considered system design.

4.2.1 Knockdown mutants

To disentangle the individual contribution of the various components of the system design to the overall behaviour, knockdown mutants are often useful to consider. A simple model decomposition consists of isolating a single process or mechanism in the considered system. In this way the model is split into two parts: one comprising the process of interest (e.g. a feedback mechanism) and the second containing all the remaining elements of the system. Although such decomposition might seem unsophisticated, this approach is often very useful in discovering the role of a single mechanism in a larger system. It is widely exploited in reverse engineering, a process aiming at revealing the technological principles of a device, object or system. In Section 4.4 we briefly describe the method of mathematically controlled comparison ([113]), where this simple decomposition approach is at the basis of the method.

This method of decomposition is used in particular in [25], in the case of the extended model of intermediate filaments assembly, where three modes of filament elongation are distinguished: (i) with a tetramer, (ii) with an ULF, or (iii) with another filament. In order to determine the role of each of these modes in the self-assembly of IFs, all possible knockdown mutant models are considered: all eight possible combinations of these three elongation mechanisms are investigated by performing parameter estimation and numerical model validation for each of them. The obtained results allow us to draw some conclusions about the importance of each of these mechanisms in the process of *in vitro* filament self-assembly. For details we refer to [25].

Knockdown mutants are also considered in [24], to disentangle the numerical contribution of modules to the system-level behaviour of the basic model for the eukaryotic heat shock response. Therein modules are systematically included and excluded from the model architecture in all possible ways and the resulting change in the model behaviour is investigated. We discuss this approach more in the context of the problem of submodel comparison in Section 4.4.

4.2.2 Elementary flux modes

Another well-established decomposition method for biochemical models appears in the context of the analysis of metabolic pathways and is concerned with the notion of elementary flux modes, discussed in Section 2.4.3. As stated therein, the recognition of the elementary flux modes allows the detection of the full set of non-decomposable steady-state flows that the network

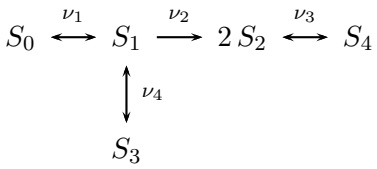


Figure 4.1: A hypothetical metabolic network from [66]. Species S_0, S_3 and S_4 are so-called external metabolites.

can support, including cyclic flows. Any steady-state flux pattern can be expressed as a non-negative linear combination of these modes ([116, 117, 118]). For example, for the network depicted in Figure 4.1, originally presented in [66], the elementary flux modes connect the external metabolites S_0 and S_3, S_0 and S_4, finally S_3 and S_4. They read

$$\begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \\ 0 \\ -1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \\ -1 \end{pmatrix}$$

and can schematically be represented as shown in Figure 4.2.

4.2.3 Control-based decomposition

A control-driven approach to model decomposition enables the recognition of the main functional modules of a system and their individual contribution to the emergent, complex behaviours of the system as a whole. In turn, this can provide great insight about various properties of a given biochemical system, e.g., robustness, efficiency, reactivity, adaptation, regulation, synchronization, etc. In particular, by applying this approach, one usually aims to identify the main regulatory components of a given biochemical system: the process to be regulated, referred to as the *plant*, the *sensors* which monitor the current state of the process and send the collected information to a decision-making module, i.e. the *controller*, and the *actuator* that modifies the state of the process in accordance with the controller's decisions, thus influences the activity of the plant. One of the fundamental concepts in control theory is the *feedback mechanism*, which provides means to cope with uncertainties: the information about the current state of the process is sent back to the controller, which reacts accordingly to facilitate a dynamic compensation for any deviance from the intended behaviour of the system. In the case of a complex system, this decomposition can be performed in different ways depending on what is considered to be the main role of that system, i.e. there may be a few reasonable choices for the plant, and the remaining components are recognized with respect to the choice of the plant.

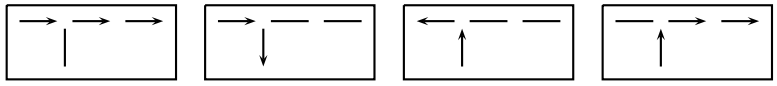


Figure 4.2: Schematic representation of the elementary flux modes of the hypothetical metabolic network in Figure 4.1. The flux modes from left to right read: $(1\ 1\ 1\ 0)^{\mathrm{T}}$, $(1\ 0\ 0\ 1)^{\mathrm{T}}$, $(-1\ 0\ 0\ -1)^{\mathrm{T}}$, and $(0\ 1\ 1\ -1)^{\mathrm{T}}$.

We illustrate these concepts and their interactions on the example of the functioning principles of an air conditioner. Here, the plant is a room which temperature is to be maintained near a desired preset value. The controller module is a thermostat which receives an input from the sensor – a temperature sensing bulb and then determines whether there is any disturbance in the room temperature. The actuator is the whole machinery consisting of coils, coolant, blower, fan, and compressor that blows the cool air into the room. It is regulated by the controller depending on the input sent by the sensor. If the temperature is too high with respect to the preset value, the controller keeps the actuator on to pump the cool air into the room. If the temperature is below the desired value, the controller switches off the cooling system.

How this control-driven approach can be exploited to investigate and understand regulatory networks can be seen in [15, 32, 63, 120, 121, 23, 24, 83]. Here we briefly describe the approach taken in [32]. The authors make a thorough study of the heat shock response mechanism in *Escherichia coli* based on modular decomposition. A model for the system is built and functional modules, i.e. the plant, sensors, controller, and actuator are identified. The decomposition reveals the underlying design of the heat shock response mechanism and its level of complexity, which, as the authors show, is not justified if only the functionality of an operational heat shock system is required. Further, this observation leads to the introduction and analysis of hypothetical design variants (mutants) of the original heat shock response model. In the original model one feedforward (temperature sensing) and two feedback elements (σ^{32} factor sequestration feedback loop and σ^{32} degradation feedback loop) can be isolated. The variants are obtained through the elimination of either the σ^{32} degradation feedback loop or both feedbacks. Moreover, the case without the feedforward element is also considered, see [32] for details. One by one, the variants in order of increasing complexity are considered starting from the simplest architecture containing just the feedforward element (the *open-loop design*). Based on numerical simulations, the authors demonstrate how the addition of subsequent layers of regulation, thereby increase in the complexity of the model, improves the performance of the response in terms of systemic properties such as robustness, noise reduction, speed of response and economical use of cellular

resources. Moreover, this systematic approach enables the identification of the contribution of each of the regulatory layers to the overall behaviour of the system. In consequence the authors succeed to perform an in-depth comparison between different model variants.

We apply the control-based decomposition approach in [24] and [83] to modularize the basic HSR model of [96] and to systematically evaluate the role of each of the distinguished modules in the overall behaviour of the HSR mechanism. This is done by considering the knockdown mutant variants of the basic model where one or more modules are missing and by comparing the mutants between each other. We discuss this more in Section 4.4.

4.3 Techniques for model modifications

In fields such as bioinformatics, genomics, proteomics or molecular biology, the focus is on providing comprehensive and detailed information on cellular components. The extensive knowledge accumulated within these research fields is often summarized in the form of static diagrams of genes and proteins interconnections. With respect to these scientific disciplines, systems biology aims to realize a paradigm shift towards understanding cell function as a well organized interplay of dynamic processes. Systems biology aims at explaining the structural and functional organization of complex biological systems as networks of dynamic interactions ([63, 137]). For analysis of dynamic processes of a cellular system, a model needs to be created ([63]). Constructing a model is in general a complex, multi-phase task, which, as mentioned in Section 2.1, involves an iterative cycle of hypothesis generation, experimental design, experimental analysis, and model refinement. However, for a starting point, one must first define the scope and abstraction level of the model. Accordingly then recognize the relevant processes and components to be incorporated into the model from a tangle of diverse mechanisms and elements that often a real biological system consists of. When designing a new biochemical model for some biological process or network, the choice one has to make on the early stage of the modelling process is whether to strive for a rich model, capturing many details, or on the contrary, to focus on a more abstract model, capturing only a few, main actors of interest. The choice is not obvious and depends heavily on the goals of the modelling project. On one hand, a rich model has the potential of being more realistic, but at the same time it leads to a more complex mathematical model that may be difficult to fit to experimental data, to analyse, and ultimately may be less apt to provide answers to biological queries. On the other hand, a less finely grained molecular model leads to a smaller mathematical model (in terms of the number of variables and equations) that may be easier to work with, but it pays a price in ignor-

ing a number of details. Hence, techniques which would allow, at a later stage of model analysis, seamless modifications of a model in both directions, i.e. either extension or reduction, while preserving certain features of the original model are of utmost importance. For example, when having a model, there may arise a necessity for introducing some modifications into it. However, making the modifications in a straightforward way often leads to the loss of desired properties, e.g. such as the fit to experimental data, of the original model. Some of these properties may have been obtained at a significant cost of computational time, resources, etc. Hence the need for methodologies which would allow to modify models in a clever way such that the desired properties of the original model could be retained without the need for repeating expensive procedures.

In the following subsections some methodologies as well as heuristic approaches towards the problem of model modifications are discussed. The presented material, which constitutes an original contribution of the author in this matter, is presented in the context of the two case studies considered throughout this thesis: the heat shock response in eukaryotic cells and the *in vitro* self-assembly of intermediate filaments.

4.3.1 Computational heuristics for simplifying a biological model

As mentioned above, computational biomodelers adopt either of the following approaches: build rich, as complete as possible models in an effort to obtain very realistic models, or on the contrary, build as simple as possible models focusing only on the core aspects of the process, in an effort to obtain a model that is easier to analyse, fit, and validate. A main difficulty in choosing between a rich and a simplified molecular model is that the potential cost of starting off with a rich model only becomes transparent at a latter stage, in the process of analysing the corresponding mathematical model. Moreover, in the case of choosing a simplified model, the selection of the aspects to be ignored in the model is left up to the subjective choice of the modeler.

In [95], on the example of the extended computational model for the eukaryotic heat shock response discussed in [97] as well as [96], a heuristic approach towards model simplification is presented. The method, starting with a (potentially large, rich) model that has already been fit and validated against experimental data, allows to simplify the model in such a way that its numerical behaviour remains largely unchanged. In consequence, the simplified model is the result of a systematic, numerical analysis of the larger model that preserves the original validation of the extended model. On the other hand, in this context, the extended model can be viewed as being capable of remaining faithful to the biological data and of soundly

identifying those aspects of the biological reality that have insignificant contribution to the overall behaviour. This simplification method is proposed as an intermediate approach between building simple and rich models. Since the models are considered in a certain numerical setup, this approach is susceptible to the problem of model identifiability, discussed in Section 4.1. This issue is discerned in [95] and a discussion on this matter is presented therein.

The simplification of the model is conducted based on a series of numerical observations of the extended computational model of the eukaryotic heat shock response originally proposed in [97]. The extension of this model with respect to the so-called basic model discussed in [96] concerns addition of a number of biochemical reactions which model the misfolding of the heat shock factors and heat shock proteins, i.e. the main actors of the response, see Section 3.1.1 for more details. In this way, the repairing mechanism is subject to failure itself. We notice here that this, by no means, is neither a trivial nor minor modification of the model. On the contrary, it introduces a profound difference on the level of designs of these two models. However, as presented in [95], based on the numerical investigations of the dynamics of these models, the extended one can be reduced to the basic model without altering its numerical behaviour, in particular without loosing its experimental fit and validation. Hence, the basic version can be seen as equivalent, from the point of view of the dynamics, with the extended version and can serve, to the extent of numerical behaviour, as a justified and precise substitute of the latter one. The advantage of being in possession of the simplified version cannot be overestimated, e.g., when combining the model into some larger modelling project. We return to this issue at the end of this section.

We briefly list here the main numerical observations which provide this justification, for the details we refer to [95]. The first observation is that the variables mhsf and hsp: mhsf of the extended model both assume negligible numerical values throughout numerical simulations in the whole range of possible environmental conditions, i.e. for temperatures ranging from 37 °C to 45 °C. Even when their initial values are increased to higher values, their numerical convergence towards their steady state values is very fast. Moreover, if the increase in the initial values of mhsf and hsp: mhsf is so that the total amount of hsf and of hsp remain unchanged, then the experimental fit and validation of the model remain largely unchanged. The reason for this behaviour lies in the negligible flux rates of the reactions having mhsf and hsp: mhsf as a product. On the other hand, the reactions having mhsf and hsp: mhsf as reactants reach much higher flux rates because of larger kinetic constants and high levels of hsp. This provides justification for the elimination of both mhsf and hsp: mhsf from the model, along with the reactions where they take part in, see [95].

Although the situation is somewhat similar for hsf, hsf_2 and hsf_3 in the sense that they all assume small values throughout numerical simulations, there is a crucial difference which points to their significance for the model: increasing the initial level of hsf_3, even in such a way that the total level of hsf is unchanged, drastically changes the fit to the experimental data. Hence, these variables cannot be removed from the model.

Second, the observation that the flux of the hsf misfolding reaction is negligible is the main rationale behind eliminating mhsf and $\mathsf{hsp}\colon\mathsf{mhsf}$ from the model. Also the flux of the hsp misfolding reaction, leading to the formation of mhsp is negligible. The case of mhsp is however different because it is also the end product of reaction $(r4')$ in Section 3.1.1, i.e.

$$\mathsf{hsf}_3\colon\mathsf{hse} \rightarrow \mathsf{hsf}_3\colon\mathsf{hse} + \mathsf{mhsp}\,. \tag{i}$$

Moreover, mhsp plays a central role in the model, being the source of all induced hsp. The numerical values assumed by mhsp throughout simulations for the whole admissible range of temperatures (37 °C-45 °C) are small, but not negligible. They are however negligible relative to the total level of hsp. Moreover, the numerical convergence of mhsp towards its steady state value is very fast, even in the case when the initial level of mhsp is increased several folds. This points to the observation that mhsp plays the role of a transient state towards hsp, having a very high turnover rate. As such, it could be eliminated from the model if only mhsp were replaced in reaction (i) with hsp. The resulting simplified molecular model has only 10 variables and 12 reactions, compared to 14 variables and 18 reactions in the initial model, confront [97] with [96]. The numerical simulations of the simplified model for temperatures between 37 °C and 45 °C are indistinguishable from those of the initial model.

From the biological perspective, the simplified model differs from the initial model in ignoring the misfolded form of hsf and hsp, as well as in ignoring the fact that newly synthesized proteins often need chaperones to gain their native conformation. Excluding the misfolding of hsf and hsp is justified by the numerical levels of misfolded hsf and hsp which are negligible with respect to the level of mfp and thus, their competition for the chaperon resources of the cell is insignificant. Excluding the role of chaperones in assisting the formation of the native conformation of newly synthesized proteins is reasonable because of the high speed of the reaction, relative to the speed of the other reactions in the model. As such, the complex of chaperone and newly synthesized protein is a very fast transient stage in the model and can be ignored.

To summarize, the following aspects contribute to the model simplification succeeding in a given numerical setup. First, variables that have a fast numerical convergence to their steady state values are eliminated (the so-called time-separation principle). An important factor here is the flux rate

of the reactions producing certain variables of the model: if the total flux contributing to the production of a given variable remains very small, then that variable will converge rapidly to its steady state value and it can be eliminated from the model. Second, the condition that the initial values of all variables are a steady state of the model at 37 °C. As stated as a theorem and proved in [95], the model has an interesting property that the steady state values of most of its variables are independent of the temperature. In this way, even at higher temperature, several of the variables of the model start from their steady state values and witness only minor numerical disturbances before returning to the same values.

The basic model is simplified even more in [96]. Based on numerical predictions of the model and on its sensitivity analysis a number of reactions with marginal contribution to the heat shock response are identified. The computed values of scaled steady state sensitivity coefficients of all variables of the model with respect to some of the reactions rate constants suggested that the respective reactions may have negligible effect on the overall quantitative (numerical) behaviour of the model. After eliminating these reactions from the basic model, the reduced model performs equally well as the basic model in all validation tests considered in [96]. In this way, mathematical modelling predicts that hsf dimers and trimers are very stable and do not break spontaneously at a significant rate and that unprompted unbinding of an hsf trimer (without the involvement of hsp) from hse is also taking place at a negligible rate. At the same time, the results of sensitivity analysis help to recognize the most significant reactions regulating the levels of the heat shock proteins and those of the misfolded proteins, for details see [96]. Thus, this heuristic analysis plays a double role: it enables further model reduction, but also deepens our understanding of where the significant control resides in the network. These outcomes indicate the usefulness and still not fully exploited potential of the application of mathematical modelling in biology.

Another example of model simplifications based on sensitivity analysis and heuristic observations can be found in [25], where the quantitative kinetic strategies for the *in vitro* assembly of intermediate filaments from human vimentin proteins are considered. The extended model distinguishes among three modes of filaments elongation: (i) with a tetramer, (ii) with an ULF, or (iii) with another filament, see Section 3.2.1. In the case of this model, a qualitative property of the IFs assembly concerning the very quick tetramers-to-ULF turnover (which was reported in [62]) leads to elimination of the first mode, i.e. the elongation with a tetramer. As argued in [25], in the case of fast tetramers-to-ULF turnover the populations of tetramers, octamers, and hexadecamers are all quickly depleted, leaving only the filaments as the dominant species. Consequently, the longitudinal assembly of tetramers to filaments has a negligible contribution to the overall dynamics

of the model due to the fact that in the beginning the process is strangled by the negligible population of filaments, whereas later on the population of tetramers is depleted. This is in agreement also with the results of the performed sensitivity analysis and, in consequence, confirmes the observation of [62] that this particular elongation plays an insignificant role in the process of filament self-assembly.

Having simple biomodels is very important for being able to analyse their mathematical properties and for their integration into larger models. The presented heuristic method to simplify an already fit model in such a way that the numerical fit to the experimental data is not lost can be utilized to reduce an existing model before extending it with some additional mechanisms not considered in the first approximation. We provide two examples of such situation. First, the model reduction in the case of filament self-assembly with fast tetramers-to-ULF turnover discussed above turns out to be essential for introducing a new, significantly larger model being able to capture the length distribution of filaments in time, i.e. a refined model for the self-assembly that allows capturing the evolution of filaments of length up to n, for any given positive integer n. Based on the kinetic observations that the longitudinal elongation of filaments with tetramers has negligible kinetic influence on the dynamics of the model and that eliminating it leads to a numerically equivalent model, we can ignore a substantial number of variables that otherwise would have to be included in the new model. To show the extent of simplification that is reached in this way, we provide this example: in the case of $n = 10$, without the discussed observations, the refined model would consist of 396 variables instead of 14 as in the version presented in [25]. In other words, the simplifications make it possible to keep the size of the refined model manageable.

Second, in the case of the heat shock response, adding the phosphorylation of hsf in all of its homo- and hetero-polymers, along with its influence on gene transcription leads to a combinatorial explosion in the number of variables of the model. Hence, decreasing the number of variables reduces the difficulty of the problem. However, as discussed in [96], even in the case of the simplified, i.e. basic model, adding phosphorylation of hsf and its role on the hsf activity is very challenging. The difficulty is in distinguishing all phosphorylation states of all known phosphorylation sites of hsf (currently at least 14 of them, see [134, 54]). In [25], an extended model is built where only one phosphorylation site for each hsf molecule is considered and an hsf trimer is only able to promote gene transcription if it has at least two of its three sites phosphorylated. The extended model consists of 61 reactions and 26 reactants. It includes all possible phosphorylation states of hsf, hsf_2, hsf_3 and hsp: hsf, as well as protein kinases and phosphatases which are subject to misfolding/refolding. The model is successfully fit to the data on DNA binding of [65] in such a way that the rate constants of the reactions

of the basic model remain unchanged. However, when considering also the phosphorylation data of [65], the combined fit is very poor. The conclusion is that the rate constants of the basic model should be re-estimated in this case, but this would lead to a very challenging computational task.

The difficulties that are faced both when considering model refinement in [25] as well as when discussing the addition of phoshorylation to the heat shock response model in [96] pointed to an intrinsic problem of modelling with differential equations: they are describing explicitly all variables in the model, even when many of them are essentially just duplicates of each other. A novel mathematical modelling methodology able to describe models in terms of various independent components and the communication between them (such as done in concurrency in Computer Science), may be more suitable in such setups. A potential choice here could be the rule-based modelling approach. For more details on this formalism, see [52, 28, 26, 27].

4.3.2 Model refinement

The generic model for self-assembly, defined notions and introduced methods in [84] came as a spin-off from the research on constructing models being able to capture the evolution of filaments of certain length up to some arbitrarily chosen value, which is presented in [25]. In [84], formal model refinement is considered in the context of mathematical models based on ordinary differential equations for the processes of self-assembly. Model refinement is an important aspect of the model-building process. As stated in Section 2.1, starting from a model abstracting a biological system, the iterative process of hypothesis generation, experimental design, experimental analysis, and model refinement lies at the core of systems biology. Model refinement can be described as a procedure which, starting from an abstract model of a system, performs a number of refinement steps in result of which a more detailed model is obtained. At the same time, in order to be correct, the refinement mechanism has to be capable of preserving already proven systemic quantitative properties of the original model, e.g. model fit, stochastic semantics, etc. One could take all the intended changes into consideration while simply repeating the whole model development procedure from scratch. But such solution would again involve the time-consuming, computationally-intensive model fitting procedure. Another approach, not much investigated in the literature, is to refine the model in such a way that the previously obtained fit is preserved. This basically implies deriving the parameter values of the refined model from the ones of the original model.

In [84], a generic formal model for the process of self-assembly is presented, the notion of model resolution is introduced and the model refinement procedures for a family of ordinary differential equation models describing this process are developed. The refinement procedures concern

increasing and decreasing the model resolution while preserving the fit to experimental data. To the best of our knowledge, this is the first time formal refinement is considered in the context of ODE-based mathematical models and the results of [25] and [84] constitute a significant, novel contribution in this area. For increasing the model resolution an exact, constructive method based on analytical investigations is developed. Decreasing the model resolution is more challenging. This is because obtaining some of the essential symbolic solutions for the derivation of a method which would be fully based on analytical deliberations and that would relate the numerical values of the parameters in the reduced model to the ones in the original model in a straightforward and simple way, yet providing exact equivalence between the two models turns out to be difficult. Hence, the method proposed in [84] for decreasing the model resolution is based both on symbolical computations as well as requires numerical investigations and simulations of the models. For detailed description and formal derivation of these methods we refer to [84]. An example of application of this methodology to the case study of self-assembly of intermediate filaments can be found both in [25] and [84].

The respective methods for model refinement presented in [84] provide means for model modifications, i.e. model extensions or model simplifications. Moreover, it is worth repeating here what is said in Section 4.1.2, that the method for increasing model resolution provides potential means for model validation.

4.4 Methods for submodel comparison

Various experimental investigations of a given biochemical system often lead to generation of a large variety of alternative molecular designs, thus raising questions about comparing their functionality, efficiency, and robustness. Comparing alternative models for a given biochemical system is, in general, a very difficult problem. This is due to the fact that the models may focus on different aspects of the same system and may consist of very different species and reactions. Moreover, the numerical setups of the associated computational models play a crucial role in the quantitative comparison. Hence, model comparison involves a deep analysis of both the underlying network of reactions, the biological assumptions as well as the numerical setup. To decide what are the benefits of one design over another, or to understand what are the selection requirements involved in an evolutionary design, one needs some unbiased methods to objectively compare the alternative designs.

The problem becomes somewhat simpler when the alternative designs are actually submodels of a larger model: the underlying networks are similar, although not identical, and the biological constrains are given by the larger

model. In the following, we concentrate on this particular case: we review several known approaches for quantitative comparison of submodels as well as we present new methods that are developed as part of the research work underlying this thesis.

4.4.1 Mathematically controlled model comparison

One such method is the mathematically controlled comparison ([113]), which provides a structured approach for comparing alternative regulatory designs with respect to some chosen measures of functional effectiveness. Under this approach, mathematical models for both the reference design and the alternatives are first developed in the framework of canonical nonlinear modelling referred to as S-systems, see [110, 111, 112]. This canonical nonlinear representation, developed within the power-law formalism, is a system of non-linear ordinary differential equations with a well-defined structure. Moreover, this framework allows the alternative models to differ from the reference design in only one process, e.g., the existence or not of some feedback mechanisms, which is actually the focus of the comparison. In each of the alternative models one then sets the numerical values of the parameters to be identical with those from the reference model for all processes other than the process of interest. This leads to a so-called internal equivalence between the reference model and the alternatives. Next, various systemic properties are selected and used to impose some constraints for all the other parameters in the alternative designs. In general in this approach, one imposes that some steady state values or logarithmic gains are equal in the reference model and its alternatives. This provides a way to express the parameters of the process of interest in the alternative models as functions of the parameters of the reference model. Thus, one obtains a so-called external equivalence between the reference model and the alternative designs, meaning that to an external observer the considered models are equivalent with respect to the selected systemic properties. Finally, one chooses various measures of functional effectiveness depending on the particularities of the biological context of these models and uses them to compare the alternative designs with the reference model. By doing this, one usually aims to determine analytically the qualitative differences between the compared models. This method was successfully used to compare alternative regulatory designs in, e.g., metabolic pathways, [57], [114], in gene circuits, [43], in immune networks, [10]. Moreover, by introducing specific numerical values for the parameters of the models, one is also able to quantify these differences but, at the same time, the generality of the results is lost. Thus, in [6], the method of mathematically controlled comparison was extended to include some statistical methods, [5], [7], that allow the use of numerical values for the parameters while still preserving the generality of the conclusions.

4.4.2 An extension of the mathematically controlled comparison

The first step of the extension in [6] is to generate a representative ensemble of sets of parameter values. Since usually for biological systems the exact statistical distribution of the parameters values is not known, the most appropriate approach is to sample uniformly a given range of values. There exist different methods for scanning a given interval of values, ranging from (more or less sophisticated) random samplings to some systematic deterministic scanning methods, see, e.g., [108]. Using this ensemble of sets of parameters, we can then construct a large class of numerical models both for the reference and for the alternative designs. In accordance with [7], there are two different methods to construct such a class of systems for which we can then investigate some statistical properties. A *structural class* consists of systems having the same network topology, i.e., generated by the sampling of the parameter space. A *behavioural class* consists of systems that exhibit a particular systemic behaviour, e.g., exhibiting a steady state behaviour under given conditions, or low concentrations of intermediary products, or small values for the parameter sensitivity, see, e.g., [7]. The members of such a class are obtained in two steps: first generate a set of parameters by sampling the parameter space, then test the sample for the desired systemic behaviour and keep only those systems that fulfil the conditions, see [7] for more details.

After constructing this large class of numerical models both for the reference and the alternative architectures, one can start comparing the values of a given systemic property P between the reference model and its alternative designs. One way to do this is by using density plots of the ratio $R = P_{reference}/P_{alternative}$ versus the values $P_{reference}$, where the subscript indicates in which model the property P was measured ([5]). Such density plots can be used for instance to compute rank correlations between the considered property P (measured in the reference model) and the values of the ratio R ([5]). However, this is not easy to do if the density plots are very scattered. One can then construct secondary density plots by using the *moving median technique* ([5]). This technique can be outlined as follows. Basically, the density plot can be interpreted as a list of N pairs of values $(P_{reference}, R)$, which can be arranged in a ordered list L with respect to the first component, $P_{reference}$. We then pick a window size W, usually much smaller than the sample size N and we compute the median $< R >$ of the ratio values and the median $< P >$ of the values $P_{reference}$, for the first W pairs in the list L. We then advance the window by one, we collect the ratios and the values $P_{reference}$ from the second until the $W + 1$st pair and compute the corresponding median values $< R >$ and $< P >$. This process is continued until the last pair of the list L is used for the first time. In

the secondary density plot, we pair the computed values $< R >$ with the corresponding $< P >$ values. This moving median technique is very useful since for a finite ordered sample of size N, the moving median tends to the median of the samples as the value W approaches N ([5]). These secondary density plots can be used to compare the efficiency of two classes of models from the point of view of a given systemic property.

For more details on the extension of the mathematically controlled comparison we refer to [6].

4.4.3 Local submodels comparison

When the alternative designs are actually submodels of the reference architecture, there is also another approach, see [23], for performing the comparison. This is the case when, for instance, one is interested in a functional analysis of various modules of a large system. The underlying reaction networks in the alternative designs are then very similar (although not identical), and both the biological constraints and the kinetics of the reactions are given by those of the reference model. The only remaining question regards the initial distribution of the variables in the alternative models. In the mathematically controlled comparison they are usually taken from the reference model. However, for some biochemical systems this choice might lead to biased comparisons. For instance, in the case of regulatory networks, models should be in a steady state in the absence of the trigger of the response and indeed the initial values of the reference model are usually chosen in such a way to fulfil this condition. However, this will not imply in general that also a submodel will be in its steady state if it uses the same initial values as the reference model. Thus, the dynamic behaviour of the submodel will be the result of two intertwined tendencies: migrating from a possible unstable state and the response to a trigger. If the focus of the comparison is exactly the efficiency of the response of various submodels to a trigger, then the approach proposed in [23] is more appropriate, yielding biologically unbiased results. In this approach, the initial distribution of the reactants is chosen in such a way that the initial setup of each submodel constitutes a steady state of that design in the absence of a trigger.

4.4.4 A discrete approach for comparing continuous submodels

The application of the control-theoretical analysis described in Section 4.2 enables the identification of the main functional modules, their interconnections and control strategies of a biochemical network. In particular, this approach can be very useful for identifying the main regulatory components of a biochemical network, including its feed-forward and feedback

mechanisms. In order to identify and quantify the exact role of each of these regulatory mechanisms, one then usually uses knockdown mutants, see [32] and Section 4.2, lacking one or more of these components. In particular, the knockdown mutant models are submodels of the reference architecture. The approach proposed in [24], associates to each knockdown mutant a Boolean formula describing its control architecture in the following way. First, a Boolean variable is associated to each of the regulating mechanisms. Using the negation and conjunction of Boolean variables, one can then write a Boolean formula for each of the knockdown mutants describing which of the regulating mechanisms are present in their architecture. In particular, these Boolean formulas describe a property of the alternative designs which is independent of time, i.e., their regulatory network. Moreover, one can go one step further and write a Boolean formula describing all those mutant architectures that show a given behavioural property, e.g., a high level of a given reactant or a given correlation between two reactants. This formula is actually the conjunction of all Boolean formulas characterizing the architectures of the mutants exhibiting the required property. The numerical comparison of the mutants is then performed by analysing the Boolean formulas associated to various behavioural properties.

In [24], this method is applied to a computational, *in-silico* model of the eukaryotic heat shock response and for that, the numerical properties of this model and of all its knockdown mutants are analysed. However, as concluded in [24], the applicability of this approach is more general: the same method could be applied to describe how properties of a *wet-lab* biomodel emerge from the combination of its modules. To this aim, one would simply replace the numerical simulation of the computational models with the experimental measurement of the behaviour of the wet-lab biomodel and that of its knockdown mutant variants.

4.4.5 A new statistical method for quantitative submodel comparison

The new statistical method for quantitative submodel comparison originally introduced in [83] can be outlined as follows. First, starting with biochemical model of some biological mechanism, referred to as the reference model (or reference architecture) of this system, we construct a submodel (or alternative architecture) by eliminating certain reactions from the list of biochemical reactions of the reference model. At this stage, we can for example apply control-based decomposition techniques, see Section 4.2, to identify a number of modules, and then study them separately by considering a number of knockdown mutants lacking one or more of the modules. Second, the associated mathematical models are formulated, both for the reference and the alternative architecture. Notice that this procedure assures that all the pa-

rameters of the alternative architecture match a subset of parameters of the reference model. Next, we perform the statistical sampling of the reference model and mutant behaviour. To accomplish this, we scan the parameter value space of the reference model (for this, the Latin Hypercube Sampling method ([80]) can be used, see [83]). This provides us with a set of parameter value vectors. Each coordinate of these vectors is associated with one of the parameters in the reference model, and determines the value of the corresponding parameter. We consider each of the vectors one by one. We set the parameters of the reference model and the submodel in accordance with the considered vector. Since, as mentioned above, the alternative architecture contains only a subset of the reference model parameters, only the values of certain coordinates are used when setting the parameters of the submodel. Further, the initial values of the variables of the reference model and the submodel are determined independently of each other by a systemic property, such as the system being in a steady state in a given setup. For example, in the general case of stress response, we expect in accordance with biological observations that a feasible mathematical model is in a steady state under the unstressed, physiological conditions. Assuring that both mathematical submodels satisfy such systemic properties makes them suitable to be considered as viable alternative formal descriptions of the biological mechanism being analysed. As a result, we obtain the numerical instantiations of the reference model and the submodel and we run numerical simulations for both of them in order to evaluate their functional effectiveness. Finally, having done this for all sampled vectors, we summarize the obtained results for the variants and compare the models by use of some statistical measures (in the case study presented in [83], the moving median technique, briefly described before while presenting the extension of the mathematically controlled comparison, is applied). As already mentioned in 4.1.3, where the model identifiability problem is considered, in this method the dependence of the analysis' results on numerical setup of the models is highly reduced. Moreover, as noticed in [24], although this new approach for model comparison is designed and presented for the deterministic framework, it can be also adapted for the stochastic framework.

4.5 Exploitation of a computational model – an example

There are many ways in which a computational model can be exploited. We list a few of them. First, the mere fact that a model is capable of recapitulating the phenomenon under study makes mathematical modelling very appealing: the ability to construct such a model indicates that one has recognized all the relevant components and interaction. On the other hand,

a negative result in this matter may suggest that the qualitative mechanisms underlying the process of interest are not yet understood or recognized. Next, having a model provides us with possibilities to identify the critical parts of the system and these which play a less significant role. Finally, a computational model can be utilized to make predictions about the behaviour of the system in various conditions without the need for tedious, expensive or infeasible in practice lab experiments. These predictions in turn can form basis for proposing new hypotheses concerning the process under study or provide essential expertise for the design of new methods or tools for various practical applications. In this section we describe one example of the latter case, where the basic heat shock response model is used to verify the possibilities of exploiting hyperthermia, a procedure of raising the temperature above 37 °C, in clinical treatment.

Theoretically, a properly tuned tempo-spatial temperature distribution in a tissue would lead to a desired heat shock response in the tissue forming cells and, in consequence, enhanced expression of heat shock proteins which are important from the therapeutic point of view. One of the most relevant problems which arise in this context is related to the question whether in the considered type of tissue a controlled and effective application of hyperthermia is practically feasible. The application has to be strictly controlled since it is important to assure that the temperature itself is kept within the therapeutic range, i.e. up to 43 °C. Furthermore, the tissue area and exposure time to heating must be precisely defined in order to activate the finely tuned heat shock response, on which the effectiveness of the treatment depends. As explained in [86], utilization of ultrasonic technique for inducing hyperthermia in tissue seems a promising approach. Technical improvements of the focused ultrasound ensure the non-invasive and strictly controlled heating of the target tissue volumes. However, the control over the spatial temperature distribution in a tissue is of essential importance for the appropriate induction of gene expression on the cellular level. It is hoped that a proper ultrasonic regime can be tuned by adjusting the ultrasound beamŠs parameters (such as intensity, frequency, pulse duration, duty-cycle) or the exposure time. The aim is to establish safe protocols for inducing heat shock response by ultrasound irradiation, which could be applied in clinical treatment.

In [86], a simple soft tissue heating model based on the Pennes' bioheat equation (see, e.g., [12, 19]) is introduced. It is utilized to establish an ultrasound heating scheme that meets the requirement of not exceeding the temperature of 43 °C at the transducerŠs focal point. Next, the resulting temperature time-course profile is combined with the heat-induced protein denaturation formula of the basic HSR mathematical model: the temperature profile is incorporated into the HSR model through the rate for protein misfolding, i.e. Equation (3.1). Finally, the obtained numerical simulation

results concerning the response at the focal point in the tissue form the basis for a discussion on the potential application of ultrasound induced soft tissue heating for therapeutic purposes.

Chapter 5

Original research contributions

In this chapter we list the original contribution of each paper contained in this thesis.

1. 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.

    - A simple model of the heat shock response is proposed. The model captures in mechanistic details all key aspects of the regulation: the heat-induced protein misfolding, the chaperone activity of heat shock proteins, the transactivation of the genes encoding heat shock proteins and the repression of their transcription once the stress is removed.
    - In contrast with previous attempts to model the eukaryotic heat shock response, our model is based solely on well-documented molecular reactions (based on standard molecular biology only) and does not include modelling "blackboxes" such as experimentally unsupported components and biochemical reactions.
    - An associated mathematical mass-action model is presented.
    - Extensive parameter estimation is performed to fit the model to the experimental data of [65] on DNA binding under stress conditions (constant heat shock of 42 °C). Moreover, the requirement of the model to be at steady state in the absence of stress is imposed.

- The model is validated with respect to new quantitative experimental data on the fluorescence level of reporter genes, i.e. yellow fluorescence protein genes (yfp). We propose an approach that enables performing the validation of the model although the original model does not contain any variable directly related to yfp. Moreover, when a heat shock applied in two stages, with a recovery period between them, with the second shock applied after the level of hsp has reached a maximum is considered, the model correctly predicts that the response to the second heat shock is much milder. The model is scalable: in the case where a constant heat shock at 43 °C is considered, our model shows a prolonged transactivation and in an experiment where the heat shock at 42 °C is removed at the peak of the response, the model shows a faster attenuation phase.
- Based on numerical predictions of the model and on its sensitivity analysis, we minimize the model by identifying the reactions with marginal contribution to the heat shock response: we use the model, in particular its sensitivity coefficients, to identify a number of reactions that have a negligible effect on the model and could be eliminated from the model without affecting its quantitative (numerical) behaviour.
- We identify the most significant reactions regulating the levels of the heat shock proteins and those of the misfolded proteins. This analysis deepens our understanding of where the significant control resides in the network.
- We address the model identifiability problem, i.e. the question of the uniqueness of the set of parameters that fulfill the imposed conditions.
- We propose a local optimization method for model fitting based on parameter scanning.
- We show how mathematical modelling of biological processes may allow reasoning about uncertain or incomplete subparts of the HSR process.
- We notice that the numerical techniques that were used in this paper for identifying the essential components of the regulatory network may also be applicable in other mathematical modelling projects.

2. Ion Petre, Andrzej Mizera, Claire L. Hyder, Andrey Mikhailov, John E. Eriksson, Lea Sistonen, and Ralph-Johan Back. A new mathematical model for the heat shock response. In Anne Condon, David Harel,

Joost N. Kok, Arto Salomaa, and Erik Winfree, editors, *Algorithmic Bioprocesses*, Natural Computing Series, pages 411–425. Springer, Dordrecht Heidelberg London New York, 2009.

- An extended model of the eukaryotic heat shock response, where the repairing mechanism is subject to failure itself is proposed.
- An associated continuous mathematical model based on the law of mass-action is fitted to the data of [65] with the condition of being in steady state under physiological conditions (constant temperature of 37 °C).

3. Ion Petre, Andrzej Mizera, and Ralph-Johan Back. Computational heuristics for simplifying a biological model. In Klaus Ambos-Spies, Benedikt Löwe, and Wolfgang Merkle, editors, *Mathematical Theory and Computational Practice: 5th Conference on Computability in Europe, CiE 2009, Proceedings*, volume 5635 of *Lecture Notes in Computer Science*, pages 399–408, Berlin Heidelberg New York, 2009. Springer.

   - We present heuristic methods to simplify an already fit model in such a way that the numerical fit to the experimental data is not lost. We focus in particular on eliminating some of the variables of the model and the reactions they take part in, while also modifying some of the remaining reactions.
   - We illustrate the methods by simplify the extended model of the heat shock response to the basic one without loosing the fit and validation. In this way the basic version can be seen as equivalent, from the point of view of the dynamics, with the extended version and can serve, to the extent of numerical behaviour, as a justified and precise substitute of the latter one.
   - We discuss the limitations of the proposed methodology.

4. Andrzej Mizera and Barbara Gambin. Stochastic modelling of the eukaryotic heat shock response. *Journal of Theoretical Biology*, 265(3): 455–466, 2010.

   - In this paper a stochastic model of the heat shock response corresponding to the deterministic one is constructed and the out comes of these two models are confronted. The aim with this comparison is two show that, in the case of the heat shock response, the approximation of a discrete system with a continuous model is a reasonable approach.

- A proof of the existence and uniqueness of the stationary distribution of the Markov chain underlying the stochastic model is given.
- We perform 1000 stochastic simulations and we analyse them. By summarizing the outcomes of the stochastic simulations with some statistics and by performing clustering and grouping of the stochastic trajectories, we investigate the range of behaviour the stochastic model is likely to exhibit.
- We demonstrate that the obtained results agree well with the dynamics displayed by the continuous model, which strengthens the trust in the deterministic description.
- Moreover, we show that both in the deterministic and stochastic models the conclusions concerning the stability of the system are coherent, i.e. in both cases the interpretation of the results leads to the conclusion that the system is monostable.

5. Elena Czeizler, Andrzej Mizera, and Ion Petre. A Boolean approach for disentangling the numerical contribution of modules to the system-level behavior of a biomodel. *TUCS Technical Report number 997*, January 2011.

    - To disentangle the contribution of modules to the system-level behaviour of a given biomodel, one often considers knockdown mutant models investigating the change in the model behaviour when modules are systematically included and excluded from the model architecture in all possible ways. We propose in this paper a Boolean logic-based approach for extracting conclusions about the role of each module from the systematic comparison of the numerical behaviour of all knockdown mutants. We associate a Boolean variable to each module, expressing when the module is included in the architecture and when it is not. We express the satisfiability of system-level properties of the full model, such as efficiency, or economical use of resources, in terms of a Boolean formula expressing in a compact way which model architectures, i.e., which combinations of modules, give rise to the desired property.
    - In our comparison of the numerical knockdown mutant models we aim to focus on the differences stemming from the intrinsic dissimilarities in their architectures and eliminate as much as possible differences coming from unfavourable numerical setups chosen for the various models.

- We demonstrate this methodology on the basic model for the heat shock response in eukaryotes. We consider all models to be viable alternatives for the biological system and, as such, we take for each of them the most favourable numerical setup.
- We describe the contribution of each of the model's three feedback loops towards achieving an economical and effective heat shock response.
- We point to the generality of our methodology: 1) our approach is independent of the ODE formulation and it would work equally well with other formulations; 2) we argue that it can be applied not only to a computational, *in-silico* model as in the case of the paper, but also in *wet-lab* research.

6. Andrzej Mizera, Elena Czeizler, and Ion Petre. Methods for biochemical model decomposition and quantitative submodel comparison. *Israel Journal of Chemistry*, 51(1):151–164, 2011.

    - We address the problem of objective quantitative comparison of several alternative submodels for the same biological process Ű- a special case of the general problem of alternative model comparison.
    - In the first part of our study we review several known methods for model decomposition and for quantitative comparison of submodels. We describe the knockdown mutants, elementary flux modes, control-based decomposition, mathematically controlled comparison and its extension, local submodels comparison and a discrete approach for comparing continuous submodels.
    - In the second part of the paper we present a new statistical method for comparing submodels that complements the methods presented in the review.
    - Similarly as in the case of our Boolean approach, each alternative model is assumed to start from its own steady state under basal conditions. This is the main difference between our approaches and the other reviewed methods, where the comparison is made in the numerical context of the reference model.
    - We address the problem of sensitivity of the comparison results to numerical setup. In our approach for quantitative comparison of alternative submodels we adopt some statistical, parameter-independent methods (sampling of the whole parameter space, moving median technique). In this way, our method enables a global, parameter-independent analysis of the numerical role of each module.

- We demonstrate our approach on a case study focusing on the heat shock response in eukaryotes.

7. Eugen Czeizler, Andrzej Mizera, Elena Czeizler, Ralph-Johan Back, John E. Eriksson, and Ion Petre. Quantitative analysis of the self-assembly strategies of intermediate filaments from tetrameric vimentin. *TUCS Technical Report number 963*, December 2009.

    - We focus on a systematic quantitative investigation of two molecular models for filament assembly, recently proposed in [62]), through mathematical modelling, model fitting, and model validation. We focus on the quantitative kinetic strategies for the *in vitro* assembly of IFs from human vimentin proteins.
    - We perform a quantitative analysis of the predictive capabilities of these models. We construct two mass action-based mathematical models corresponding to the two molecular models. For each of them we consider several different knockdown mutant model variants where various combinations of assembly mechanisms are analysed separately.
    - We consider a qualitative property of the IF assembly, reported in [62] that very quickly (within approximately 10 seconds) after the initiation of the assembly, ULF is the most predominant species in the system. However, this observation only applies for the *ab initio in vitro* assembly of intermediate filaments. *In vivo* there exists also a mechanism of tetramer synthesis that contributes an influx of tetramers to the model. For this reason we consider two different strategies for the fitting of our models: one where tetramers are quickly depleted in the model, and one where no such condition is imposed.
    - We perform parameter estimation and validation of our models with respect to separate data sets of [62].
    - We demonstrate how to enhance the existing filament assembly models with the dynamics of the filament length distribution. The size of this detailed model is considerably higher than that of the basic model, both in terms of molecular species, as well as in terms of molecular reactions. Based on kinetic observations on the basic model, we show however how the size of the high-resolution model can be drastically reduced. In this way, we introduce a high-resolution model for vimentin filament self-assembly, able to capture the detailed dynamics of filaments of arbitrary length yet of manageable size. Our approach towards high-resolution models for protein self-assembly is independent of

the particulars of vimentin filaments and can be applied to other instances of protein-protein interactions and protein assemblies.

8. Andrzej Mizera, Eugen Czeizler, and Ion Petre. Self-assembly models of variable resolution. *TUCS Technical Report number 1014*, June 2011.

    - We concentrate on quantitative model refinement in the case of self-assembly ODE-based models.
    - We develop a generic formal model for the self-assembly process and introduce a notion of model resolution capturing the maximum size up to which objects can be distinguished individually in the model. All bigger objects are treated homogenously in the model.
    - We show how this self-assembly model can be systematically refined in such a way that its resolution can be increased and decreased while preserving the original model fit to experimental data, without the need for tedious, computationally expensive process of parameter refitting.
    - We demonstrate how the introduced methodology can be applied to a previously published model: we consider the case-study of *in vitro* self-assembly of intermediate filaments.

9. Andrzej Mizera and Barbara Gambin. Modelling of ultrasound therapeutic heating and numerical study of the dynamics of the induced heat shock response. *Communications in Nonlinear Science and Numerical Simulation*, 16(5):2342–2349, 2011.

    - The basic heat shock response model is used to verify the possibilities of exploiting hyperthermia, a procedure of raising the temperature above 37 °C, in clinical treatment.
    - A simple soft tissue heating model based on the Pennes' bioheat equation is introduced. Ultrasonic irradiation technique for inducing hyperthermia in tissue is considered. The model is utilized to establish an ultrasound heating scheme that meets the requirement of not exceeding the critical temperature of 43 °C at the transducerŠs focal point. The resulting temperature time-course profile is combined with the heat-induced protein denaturation formula of the basic HSR mathematical model.
    - The obtained results of numerical simulations concerning the response at the focal point in the tissue are discussed in the context of potential application of ultrasound induced soft tissue heating for therapeutic purposes.

Chapter 6

Conclusions and perspectives

The doctoral research constituting the foundations of this thesis revolves around a number of challenges commonly encountered in the computational modelling in systems biology. The research comprises of the development and application of a broad range of methods originating in the fields of computer science and mathematics for construction and analysis of computational models in systems biology. In particular, the performed research is setup in the context of two biological phenomena chosen as modelling case studies: 1) the eukaryotic heat shock response and 2) *in vitro* self-assembly of intermediate filaments from tetrameric vimentin. The range of presented approaches spans from heuristic, through numerical and statistical to analytical methods applied in the effort to formally describe and analyse the two biological processes. Although applied to certain case studies, these methods are not limited to them and can be utilized in the analysis of other biological mechanisms as well as complex systems in general. The full range of developed and applied modelling techniques, as well as model analysis methodologies, constitutes a rich modelling framework.

In this thesis, we address issues related to model construction methodologies such as parameter estimation and model validation with respect to separate sets of experimental data, both quantitative as well as qualitative. We discuss the problem of model identifiability in various contexts (e.g. uniqueness of parameters satisfying imposed conditions, choice of the numerical setup for model comparison, generality of the drawn conclusions with respect to numerical setup of the model). We review existing techniques and develop new ones for performing comparison between submodels of a larger model. We address the problem of model modifications: we develop various techniques and show a number of heuristics useful for applying simplifications or extensions to an already fitted and validated mathematical model in such a way that the desired properties of the original model are retained. In particular, in the context of self-assembly, we provide both numerical

as well as analytical methods for decreasing and increasing the resolution of models based on the ODE formulation. These methods can be viewed as examples of adaptations of formal model refinement techniques from the field of computer science to systems biology. Although such attempts have been made previously in the case of the rule-based modelling ([26, 89]), to the best of our knowledge this is the first time that formal model refinement is considered in relation to computational models based on ODEs. The techniques are developed in the case of a generic self-assembly model, however we notice that the formulation of the refinement problem presented in [84] is valid for ODE-based models in general and the discussed approaches of obtaining the refinement method can be used for other ODE-based models as well.

In addition to introducing new methodologies, the performed research provides some insight into the general question concerning the role of mathematical modelling in biology. On the example of the model for the eukaryotic heat shock response we show how mathematical modelling of biological processes may allow reasoning about uncertain or incomplete subparts of the process such as reversibility or irreversibility of certain reactions. Moreover, constructing and analysing mathematical models provides means for identifying the essential components of the HSR regulatory network. Similarly, in the case of *in vitro* self-assembly of intermediate filaments mathematical modelling and the analysis of various potential scenarios of self-assembly allows us to draw conclusions and formulate some hypothesis regarding the still poorly understood process of intermediate filaments formation. In this way, our research helps recognizing the potential of mathematical modelling in biology. We notice two things here. First, with respect to clarifying the intricacies of the two considered biological processes, the performed research is by no means completed and the predictions of our mathematical models require, as the next step, further experimental validation. For example, as stated in [25], an *in vitro* experiment where tetramers are added either continuously or at well-chosen time points could offer more insight into the role of tetramer longitudinal aggregation for the process of filament elongation. Second, in the context of the two case studies, the presented work finds its well-defined place in the iterative circle of systems biology.

In this thesis, we face the problem of choosing between the deterministic or stochastic framework. As mentioned in Section 2.5, this issue is often brought up for discussion and, as argued therein, the choice between obtaining an averaged characterization of the dynamics or more detailed view where stochastic effects are taken into account depends on the scope and purpose of modelling. Here we draw this discussion further and consider the problem of the choice of the proper modelling framework in a broader context than in the case of selecting between the averaged or more detailed view of the system dynamics. There exists a number of other modelling frame-

works such as process algebras, with π-calculus being one of its representatives ([82]), Boolean networks ([61]), generalized logical networks ([130]), Petri Nets ([98]), stochastic Petri Nets ([127, 90, 87]), rule-based formalisms ([52, 28, 26, 27]), etc. One could view these frameworks as providing a higher level, system-view of a modelled biochemical reaction network than the formulation based on ordinary differential equations. In the latter case, the equations describe the changes in time of the state of the system (often expressed in terms of concentrations), i.e. they provide information on the evolution of populations of molecules in time. They hide the 'cause-effect' relations and interactions. Often the equations are obtained from a biochemical model consisting of a set of biochemical reactions, which contains the information on the network structure and interactions between considered types of molecules. However, this is not the only possible scenario and one could imagine a situation where the system of ODEs is not accompanied by any explicit knowledge about the underlying reaction network. The ODEs specify a transfer function, which relates different numerical quantities to each other ([33]). In other words, the equations describe the changes in the model variables' values when the considered system moves from one state to another. They do not highlight why and how that system transition occurs ([103]). In this sense they provide a view of the system on a "molecular" level, i.e. they describe the evolution of the molecular populations. On the other hand, in the case of frameworks such as Petri Nets, Boolean networks or rule-based formalisms, the information on the network structure, pathways or interactions between elements of the system is inherent in the framework. The paradigm in this type of formalisms is the thinking in terms of 'cause-effect' rather than rates of change as in the case of ODEs ([33]). We can view this formalisms as providing a higher level, system-view of the process under study.

When deciding on a particular framework, it is important to realize that expressing something on a lower level does not eliminate the possibilities to capture and analyse global, emergent, or structural properties of the modelled system. This is to some extent reflected in the fact that, as in the case of the deterministic and stochastic frameworks, there often exist ways to make transitions between different formalisms. Moreover, it is important to remember that each framework has its advantages and limitations. For example, the ODEs are well-suited where modelling involves dealing with large populations. However, they may become a bottleneck where further model refinement is required due to the combinatorial explosion in the number of variables of the considered models (see, e.g., the cases discussed in this thesis: extending the HSR model with the process of phosphorylation or increasing the resolution of the IFs self-assembly model by straightforward refinement) and in consequence the manageability of the model may become impractical. In the case of rule-based formalism, the situation is the oppo-

site: the refinement procedure does not present any essential problem, but executing a model containing big number of molecules is rather inefficient. Finally, the decision may be influenced by other conditions. For example, the choice of ODEs for modelling biochemical networks is often based on the fact that this formalism is very well-established in many disciplines of science, the underlying theory is very profound and rich, there exist a lot of expertise with respect to how to analyse analytically as well as numerically systems of differential equations, and how to efficiently simulate them. On the other hand, the frameworks originating from the field of computer science, although very promising and suitable, are still young, under intensive development and known to a relatively small community. Since in the case of systems biology the communication between researchers having their background in very diverse disciplines is highly important and, as argued in Section 2.4, ODEs are suitable as the model of observations made in experiments, they are commonly used to model biochemical reaction networks. This may however change with further development of other formalisms as well as increasing expertise in them.

We conclude the discussion on this interesting problem by giving the analogy with the problem of choosing the proper programming language. Again, different programming languages are characterized by different programming paradigms. There exist for example imperative languages (e.g. Fortran, Basic, C), object-oriented languages (e.g. Smalltalk, Java) functional languages (e.g. Haskell, Lisp) logic languages (e.g. Prolog), etc. Moreover, programming languages are characterized by the level of abstraction and efficiency (usually determined by the compiler) they provide. However, anything that is programmed in one of them can be expressed in the assembler language or machine code consisting of very simple, low-level processor instructions. There are many reasons for choosing a particular language depending on the purposes, expertise, or preferences of the programmer and often it is a subjective choice. Although sometimes the choice may seem awkward or unnatural, there is no wrong decision as far as the preset goals can be achieved. The same remains true with respect to the choice of the proper modelling framework. For more discussion on this exciting issue we refer to [33] and [103].

In the course of the research work we left unanswered a few open technical problems. We list three of them here. First, in [95], on the basis of some numerical observations the extended HSR model is reduced to the basic one. In order to analyse the robustness of our model reductions with respect to different numerical setups, a number of tests are performed. In each test, a perturbation with respect to the original numerical setup either in the initial values of some variables, or in the numerical values of some of the rate constants is introduced. For each new numerical setup the initial values of all variables are set to their steady state values at 37 °C. Finally,

the numerical behaviour of the model with that of its simplified version is compared, for temperatures between 37 °C and 45 °C. In one of the tests, in which the total amount of hsp is increased, we observed that the steady state values of the model at 37 °C were identical with those of the initial model. This observation raised the following intriguing question: is the steady state of the extended HSR model independent of the initial total level of hsp? In fact, the same question is valid also for the basic HSR model. We state this as a conjecture.

Conjecture. *Let $C_1, C_2, C_3 \in \mathbb{R}_+$ be positive real constants in the mass conservation relations of the basic mass-action HSR model presented in [96], i.e.*

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

For any fixed set of parameter values the steady state of the basic HSR model is then independent of the choice of the set of initial concentrations satisfying the conservation relations with mass constants C_1, C_2, and C_3.

Second, we derive in [84] quantitative refinement methods for the generic self-assembly mathematical model expressed in terms of ODEs. These methods determine how to set the rate constants of the new model with respect to the numerical setup of the original one in such way that the numerical fit of the original model is preserved without the need of performing parameter estimation. These methods are derived from a refinement condition which is formulated in terms of equalities between certain variables or sums of variables of the two models. These equalities have to be satisfied at any time point (for more details we refer to [84]). One could think of similar methods in the case of the stochastic formulation. However, the time evolution of a stochastic system is characterized by probability distributions, i.e. the grand probability functions, instead of deterministic concentration functions of time as in the case of ODEs. Thus, one would have to start derivations of the methods with expressing the refinement condition in terms of equalities between respective probability distributions. Subsequent steps would most probably require manipulations on the probability distributions and their derivatives with respect to time, i.e. the corresponding chemical master equations. The mathematical considerations in this case seem more involved than in the case of ODE-based models.

Finally, as proposed in [84], one could think of deriving a refinement method for a generic self-assembly model of resolution n presented in [84] to the model of infinite resolution. Although we believe that our methodology in [84] would also work in this case, formal theoretical considerations of

this issue are much more intricate than in the finite case. Already at the stage of writing the differential equations of the model one needs to make sure that the appearing infinite function series are (uniformly) convergent. For example, let us consider as in [84] a generic self-assembly model of resolution 0, i.e. $F+F \xrightarrow{k} F$, and let us consider its refinement to the infinite resolution. This amounts to considering an infinite set of reactions, i.e. $F_i + F_j \xrightarrow{k_{i,j}} F_{i+j}$ for all $i, j \geq 1$, where F_i and F_j describe the concentration in time of elements of size exactly i and j, respectively (for the formal definition of the notion of self-assembly model resolution we refer to [84]). In the case of the infinite resolution model one already faces a problem of function series convergence while writing the differential equations for the model variables F_is. For each fixed i, the expression for the derivative $\mathrm{d}F_i/\mathrm{d}t$ contains a finite number of terms $k_{l,j}F_iF_j$ where $l + j = i$ with $1 \leq l \leq j < i$, and an infinite number of terms $-k_{i,j}F_iF_j$ where $j \geq 1$. The trouble is whether the infinite series $\sum_{j=1}^{\infty} k_{i,j}F_iF_j$ is convergent for all $t \geq 0$ or whether the terms can be reordered in such a way that the requirement of convergence is satisfied. More details on this can be found in the discussion section of [84].

As stated above, the wide range of modelling techniques as well as model analysis methodologies discussed in this thesis provides a rich modelling framework. Moreover, the presentation of the developed methods, their application to the two case studies and the discussions concerning their potentials and limitations point to the difficulties and challenges one encounters in computational modelling of biological systems. The problems of model identifiability, model comparison, model refinement, model integration and extension, choice of the proper modelling framework and level of abstraction, or the choice of the proper scope of the model run through this thesis. The aim with the presented research work underlying this thesis is to contribute more understanding of these important issues and, hopefully, to make a step forward on the long way towards gaining methodologies that would provide means to cope with the current computational challenges of the emerging field of systems biology.

Bibliography

[1] Vicente Acuña, Flavio Chierichetti, Vincent Lacroix, Alberto Marchetti-Spaccamela, Marie-France Sagot, and Leen Stougie. Modes and cuts in metabolic networks: Complexity and algorithms. *BioSystems*, 95(1):51–60, 2009.

[2] Bruce Alberts, Dennis Bray, Karen Hopkin, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Essential Cell Biology*. Garland Science, New York, 2nd edition, 2004.

[3] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell*. Garland Science, New York, 4th edition, 2002.

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

[5] Rui Alves and Michael A. Savageau. Comparing systemic properties of ensembles of biological networks by graphical and statistical methods. *Bioinformatics*, 16(6):527–533, 2000.

[6] Rui Alves and Michael A. Savageau. Extending the method of mathematically controlled comparison to include numerical comparisons. *Bioinformatics*, 16(9):786–798, 2000.

[7] Rui Alves and Michael A. Savageau. Systemic properties of ensembles of metabolic networks: application of graphical and statistical methods to simple unbranched pathways. *Bioinformatics*, 16(6):534–547, 2000.

[8] William E. Balch, Richard I. Morimoto, Andrew Dillin, and Jeffery W. Kelly. Adapting proteostasis for disease intervention. *Science*, 319(5865):916–919, 2008.

[9] Scott A. Becker, Adam M. Feist, Monica L. Mo, Gregory Hannum, Bernhard Ø. Palsson, and Markus J. Herrgard. Quantitative predic-

tion of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nature Protocols*, 2(3):727–738, 2007.

[10] Rob J. De Boer and Pauline Hogeweg. Stability of symmetric idiotypic networks – a critique of Hoffmann's analysis. *Bulletin of Mathematical Biology*, 51:217–222, 1989.

[11] Rainer Breitling, David Gilbert, Monika Heiner, and Richard Orton. A structured approach for the engineering of biochemical network models, illustrated for signalling pathways. *Briefings in Bioinformatics*, 9(5):404–421, 2008.

[12] Joseph D. Bronzino, editor. *The Biomedical Engineering*, volume 2. CRC Press LLC & Springer-Verlag, Heidelberg, 2nd edition, 2000.

[13] Frank J. Bruggeman and Hans V. Westerhoff. The nature of systems biology. *Trends in Microbiology*, 15(1):45–50, 2007.

[14] Richard L. Burden and J. Douglas Faires. *Numerical Analysis*. Thomson Brooks/Cole, Belmont, USA, 8th edition, 2005.

[15] Jean M. Carlson and John C. Doyle. Complexity and robustness. *Proc. Natl. Acad. Sci. USA*, 99(suppl. 1):2538–2545, 2002.

[16] Marvin Cassman, Adam Arkin, Frank Doyle, Fumiaki Katagiri, Douglas Lauffenburger, and Cynthia Stokes. WTEC Panel Report on International Research and Development in Systems Biology. Technical report, World Technology Evaluation Center, Inc., 2005.

[17] William W. Chen, Birgit Schoeberl, Paul J. Jasper, Mario Niepel, Ulrik B. Nielsen, Douglas A. Lauffenburger, and Peter K. Sorger. InputŰoutput behavior of ErbB signaling pathways as revealed by a mass action model trained against dynamic data. *Molecular Systems Biology*, 5:239, 2009.

[18] Yu Chen, Tracy S. Voegeli, Peter P. Liu, Earl G. Noble, and R. William Currie. Heat shock paradox and a new role of heat shock proteins and their receptors as anti-inflammation targets. *Inflammation & Allergy-Drug Targets*, 6(2):91–100, 2007.

[19] Young I. Cho, editor. *Bioengineering Heat Transfer*. Academic Press Limited, London, UK, 1992.

[20] Federica Ciocchetta and Jane Hillston. Bio-PEPA: an extension of the process algebra PEPA for biochemical networks. *Electronic Notes in Theoretical Computer Science*, 194(3):103–117, 2008.

[21] Melody S. Clark, Keiron P. P. Fraser, and Lloyd S. Peck. Lack of an HSP70 heat shock response in two Antarctic marine invertebrates. *Polar Biology*, 31(9):1059–1065, 2008.

[22] Marie E. Csete and John C. Doyle. Reverse engineering of biological complexity. *Science*, 295(5560):1664–1669, 2002.

[23] Elena Czeizler, Eugen Czeizler, Ralph-Johan Back, and Ion Petre. Control strategies for the regulation of the eukaryotic heat shock response. In Pierpaolo Degano and Roberto Gorrieri, editors, *Computational Methods in Systems Biology*, volume 5688 of *Lecture Notes in Computer Science*, pages 111–125, Heidelberg, 2009. Springer-Verlag.

[24] Elena Czeizler, Andrzej Mizera, and Ion Petre. A Boolean approach for disentangling the numerical contribution of modules to the system-level behavior of a biomodel. *Submitted to BMC Systems Biology*, 2011.

[25] Eugen Czeizler, Andrzej Mizera, Elena Czeizler, Ralph-Johan Back, John E. Eriksson, and Ion Petre. Quantitative analysis of the self-assembly strategies of intermediate filaments from tetrameric vimentin. *Submitted to Transactions on Computational Biology and Bioinformatics*, 2011.

[26] Vincent Danos, Jérôme Feret, Walter Fontana, Russ Harmer, and Jean Krivine. Rule-based modelling, symmetries, refinements. In Jasmin Fisher, editor, *Formal Methods in Systems Biology. First International Workshop, FMSB 2008, Proceedings*, volume 5054 of *Lecture Notes in Bioinformatics*, pages 103–122, Berlin Heidelberg, 2008. Springer-Verlag.

[27] Vincent Danos, Jérôme Feret, Walter Fontana, Russ Harmer, and Jean Krivine. Rule-based modelling and model perturbation. In Corrado Priami, Ralph-Johan Back, and Ion Petre, editors, *Transactions on Computational Systems Biology XI*, volume 5750 of *Lecture Notes in Bioinformatics*, pages 116–137. Springer-Verlag, Berlin Heidelberg, 2009.

[28] Vincent Danos, Jérôme Feret, Walter Fontana, and Jean Krivine. Scalable simulation of cellular signaling networks. In Zhong Shao, editor, *Programming Languages and Systems. 5th Asian Symposium, APLAS 2007, Proceedings*, volume 4807 of *Lecture Notes in Computer Science*, pages 139–157, Berlin Heidelberg, 2007. Springer-Verlag.

[29] Hidde de Jong. Modeling and simulation of genetic regulatory systems: A literature review. *Journal of Computational Biology*, 9(1):97–103, 2002.

[30] Yves R. A. Donati, Daniel O. Slosman, and Barbara S. Polla. Oxidative injury and the heat shock response. *Biochemical Pharmacology*, 40(12):2571–2577, 1990.

[31] Julian Downward. The ins and outs of signalling. *Nature*, 411:759–762, 2001.

[32] Hana El-Samad, Hiroyuki Kurata, John C. Doyle, Carol A. Gross, and Mustafa Khammash. Surviving heat shock: Control strategies for robustness and performance. *Proc. Natl. Acad. Sci. USA*, 102(8):2736–2741, 2005.

[33] Jasmin Fisher and Thomas A. Henzinger. Executable cell biology. *Nature Biotechnology*, 25(11):1239–1249, 2007.

[34] Tito Fojo, editor. *The Role of Microtubules in Cell Biology, Neurobiology, and Oncology.* Cancer Drug Discovery and Development. Humana Press, 2008.

[35] Michael Gibson and Jehoshua Bruck. An efficient algorithm for generating trajectories of stochastic gene regulation reactions. Technical report, California Institute of Technology, 1998.

[36] Daniel T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22:403–434, 1976.

[37] Daniel T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry*, 81(25):2340–2361, 1977.

[38] Daniel T. Gillespie. A rigorous derivation of the chemical master equation. *Physica A*, 188(1-3):404–425, 1992.

[39] Cato M. Guldberg and Peter Waage. Studies concerning affinity. *C. M. Forhandlinger: Videnskabs-Selskabet i Christiana*, 35, 1864.

[40] Cato M. Guldberg and Peter Waage. Concerning chemical affinity. *Erdmann's Journal für Practische Chemie*, 127:69–114, 1879.

[41] John T. Hancock. *Cell Signalling.* Oxford University Press, New York, 3rd edition, 2010.

[42] Bradford A. Hawkins and Howard Vernon Cornell, editors. *Theoretical Approaches to Biological Control*. Cambridge University Press, Cambridge, UK, 1999.

[43] Reinhart Heinrich and Stefan Schuster. *The regulation of cellular systems*. Chapman & Hall, New York, 1996.

[44] Reinhart Heinrich and Stefan Schuster. *The regulation of cellular systems*. Chapman & Hall, New York, 1996.

[45] Jon C. Helton and Freddie J. Davis. Illustration of sampling-based methods for uncertainty and sensitivity analysis. *Risk Analysis*, 22(3):591–622, 2002.

[46] Jon C. Helton and Freddie J. Davis. Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. *Reliability Engineering & System Safety*, 81(1):23–69, 2003.

[47] Ray C. Henrikson, Gordon I. Kaye, and Joseph E. Mazurkiewicz. *NMS Histology*. National Medical Series for Independent Study. Lippincott Williams & Wilkins, Maryland Pennsylvania, 1997.

[48] Harald Herrmann and Ueli Aebi. Intermediate filaments: molecular structure, assembly mechanism, and integration into functionally distinct intracellular scaffolds. *Annual Review of Biochemistry*, 73:749–789, 2004.

[49] Harald Herrmann, Harald Bär, Laurent Kreplak, Sergei V. Strelkov, and Ueli Aebi. Intermediate filaments: from cell architecture to nanomechanics. *Nature Reviews Molecular Cell Biology*, 8:562–573, 2007.

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

[51] Harald Herrmann, Markus Häner, Monika Brettel, Shirley A. Müller, Kenneth N. Goldie, Bettina Fedtke, Ariel Lustig, Werner W. Franke, and Ueli Aebi. Structure and assembly properties of the intermediate filament protein vimentin: the role of its head, rod and tail domains. *Journal of Molecular Biology*, 264(5):933–953, 1996.

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

[53] Gretchen E. Hofmann, Bradley A. Buckley, Susanna Airaksinen, John E. Keen, and George N. Somero. Heat-shock protein expression is absent in the Antarctic fish *Trematomus Bernacchii* (family nototheniidae). *Journal of Experimental Biology*, 203(15):2331–2339, 2000.

[54] Carina I. Holmberg, Stefanie E. F. Tran, John E. Eriksson, and Lea Sistonen. Multisite phosphorylation provides sophisticated regulation of transcription factors. *Trends in Biochemical Sciences*, 27(12):619–627, 2002.

[55] Kenneth C. Holmes, David Popp, Werner Gebhard, and Wolfgang Kabsch. Atomic model of the actin filament. *Nature*, 347:44–49, 1990.

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

[57] Axel Hunding. Limit-cycles in enzyme-systems with nonlinear negative feedback. *Biophysics of Structure & Mechanism*, 1:47–54, 1974.

[58] Richard A. Jackson. *Mechanisms in organic reactions*. Royal Society of Chemistry, 2004.

[59] Colleen M. Jones, Eric R. Henry, Yi Hu, Chi-Kin Chan, Stan D. Luck, Abani Bhuyan, Heinrich Roder, James Hofrichter, and William A. Eaton. Fast events in protein folding initiated by nanosecond laser photolysis. *Proc. Natl. Acad. Sci. USA*, 90(24):11860Ű11864, 1993.

[60] Harm K. Kampinga. Thermotolerance in mammalian cells: protein denaturation and aggregation, and stress proteins. *Journal of Cell Science*, 104:11–17, 1993.

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

[62] Robert Kirmse, Stephanie Portet, Norbert Mücke, Ueli Aebi, Harald Herrmann, and Jörg Langowski. A quantitative kinetic model for the *in Vitro* assembly of intermediate filaments from tetrameric vimentin. *Journal of Biological Chemistry*, 282(25):18563–18572, 2007.

[63] Hiroaki Kitano. Systems biology: A brief overview. *Science*, 295(5560):1662–1664, 2002.

[64] Hiroaki Kitano, Akira Funahashi, Yukiko Matsuoka, and Kanae Oda. Using process diagrams for the graphical representation of biological networks. *Nature Biotechnology*, 23(8):961–966, 2005.

[65] 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.

[66] Edda Klipp, Ralf Herwig, Axel Kowald, Christoph Wierling, and Hans Lehrach. *Systems Biology in Practice. Concepts, Implementation and Application*. Wiley-VCH, 2005.

[67] Kurt W. Kohn. Molecular interaction maps as information organizers and simulation guides. *Chaos*, 11(1):84–97, 2001.

[68] Kurt W. Kohn and Mirit I. Aladjem. Circuit diagrams for biological networks. *Molecular Systems Biology*, 2006.

[69] Kurt W. Kohn, Mirit I. Aladjem, John N. Weinstein, and Yves Pommier. Molecular interaction maps of bioregulatory networks: A general rubric for systems biology. *Molecular Biology of the Cell*, 17(1):1–13, 2006.

[70] Arthur D. Lander. The edges of understanding. *BMC Biology*, 8:40, 2010.

[71] Yuri Lazebnik. Can a biologist fix a radio? – Or, what I learned while studying apoptosis. *Cancer Cell*, 2(3):179–182, 2002.

[72] Jong Min Lee, Erwin P. Gianchandani, James A. Eddy, and Jason A. Papin. Dynamic analysis of integrated signaling, metabolic, and regulatory networks. *PLoS Computational Biology*, 4(5):e1000086, 2008.

[73] James R. Lepock, Harold E. Frey, and Kenneth P. Ritchie. Protein denaturation in intact hepatocytes and isolated cellular organelles during heat shock. *Journal of Cell Biology*, 122(6):1267–1276, 1993.

[74] James R. Lepock, Harold E. Frey, A. Michael Rodahl, and Jack Kruuv. Thermal analysis of chl v79 cells using differential scanning calorimetry: Implications for hyperthermic cell killing and the heat shock response. *Journal of Cellular Physiology*, 137(1):14–24, 1988.

[75] Susan Lindquist and Elizabeth A. Craig. The heat-shock proteins. *Annual Review of Genetics*, 22:631–677, 1988.

[76] Ovidiu Lipan, Jean-Marc Navenot, Zixuan Wang, Lei Huang, and Stephen C. Peiper. Heat shock response in CHO mammalian cells is controlled by a nonlinear stochastic process. *PLoS Computational Biology*, 3(10):1859–1870, 2007.

[77] Bei Liu, Anna M. DeFilippo, and Zihai Li. Overcomming immune toerance to cancer by heat shock protein vaccines. *Molecular Cancer Therapeutics*, 1(12):1147–1151, 2002.

[78] Katalin Lukacs, Olivier Pardo, M. Jo Colston, Duncan Geddes, and Eric Alton. Heat shock proteins in cancer therapy. In Nagy A. Habib, editor, *Cancer Gene Therapy: Past Achievements and Future Challenges*, volume 465 of *Advances in Experimental Medicine and Biology*, pages 363–368. Kluwer Academic Publishers, New York, Boston, Dordrecht, London, Moscow, 2002.

[79] Harley H. McAdams and Adam Arkin. It's a noisy business! Genetic regulation at the nanomolar scale. *Trends in Genetics*, 15(2):65–69, 1999.

[80] Michael D. McKay, Richard J. Beckman, and William J. Conover. A 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.

[81] Alan D. McNaught and Andrew Wilkinson. *IUPAC. Compendium of Chemical Terminology. The Gold Book*. Blackwell Scientific Publications, Oxford, 2nd edition, 1997.

[82] Robin Milner, Joachim Parrow, and David Walker. A calculus of mobile processes, I. *Information and Computation*, 100(1):1–40, 1992.

[83] Andrzej Mizera, Elena Czeizler, and Ion Petre. Methods for biochemical model decomposition and quantitative submodel comparison. *Israel Journal of Chemistry*, 51(1):151–164, 2011.

[84] Andrzej Mizera, Eugen Czeizler, and Ion Petre. Self-assembly models of variable resolution. *Submitted to Transactions on Computational Systems Biology*, 2011.

[85] Andrzej Mizera and Barbara Gambin. Stochastic modelling of the eukaryotic heat shock response. *Journal of Theoretical Biology*, 265(3):455–466, 2010.

[86] Andrzej Mizera and Barbara Gambin. Modelling of ultrasound therapeutic heating and numerical study of the dynamics of the induced

heat shock response. *Communications in Nonlinear Science and Numerical Simulation*, 16(5):2342–2349, 2011.

[87] Michael K. Molloy. *On the integration of delay and throughput measures in distributed processing models*. PhD thesis, Department of Computer Science, University of California, Los Angeles, CA, 1981.

[88] Richard I. Morimoto. Proteotoxic stress and inducible chaperone networks in neurodegenerative disease and aging. *Genes & Development*, 22(11):1427–1438, 2008.

[89] Elaine Murphy, Vincent Danos, Jérôme Feret, Jean Krivine, and Russell Harmer. Rule-based modeling and model refinement. In Huma M. Lodhi and Stephen H. Muggleton, editors, *Elements of Computational Systems Biology*. John Wiley & Sons, Inc., Hoboken, NJ, USA, 2010.

[90] Stéphane Natkin. *Les réseaux de Petri stochastiques et leur application a l'evaluation des systemes informatiques*. PhD thesis, Conservatoire National des Arts et Metier, Paris, 1980.

[91] Bodil Nordlander, Edda Klipp, Bente Kofahl, and Stefan Hohmann. Modelling signalling pathways Ű- a yeast approach. In Lilia Alberghina and Hans V. Westerhoff, editors, *Systems Biology: Definitions and Perspectives*, volume 13 of *Topics in Current Genetics*, pages 277–302. Springer-Verlag, Berlin Heidelberg, 2005.

[92] Michael Oberguggenberger, Julian King, and Bernhard Schmelzer. Classical and imprecise probability methods for sensitivity analysis in engineering: A case study. *International Journal of Approximate Reasoning*, 50(4):680–693, 2009.

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

[94] Abraham Peper, Cornelis A. Grimbergen, Jos A. E. Spaan, Jan E. M. Souren, and Roeland van Wijk. A mathematical model of the hsp70 regulation in the cell. *International Journal of Hyperthermia*, 14(1):97–124, 1998.

[95] Ion Petre, Andrzej Mizera, and Ralph-Johan Back. Computational heuristics for simplifying a biological model. In Klaus Ambos-Spies, Benedikt Löwe, and Wolfgang Merkle, editors, *Mathematical Theory and Computational Practice: 5th Conference on Computability in Europe, CiE 2009, Proceedings*, volume 5635 of *Lecture Notes in Computer Science*, pages 399–408, Berlin Heidelberg New York, 2009. Springer.

[96] 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*, 2011. (to appear).

[97] Ion Petre, Andrzej Mizera, Claire L. Hyder, Andrey Mikhailov, John E. Eriksson, Lea Sistonen, and Ralph-Johan Back. A new mathematical model for the heat shock response. In Anne Condon, David Harel, Joost N. Kok, Arto Salomaa, and Erik Winfree, editors, *Algorithmic Bioprocesses*, Natural Computing Series, pages 411–425. Springer, Dordrecht Heidelberg London New York, 2009.

[98] Carl A. Petri. Communication with automata. Technical Report AD0630125, DTIC, 1966.

[99] Thomas Pfeiffer, Ignacio Sánchez-Valdenebro, Juan C. Nuño, Francisco Montero, and Stefan Schuster. METATOOL: for studying metabolic networks. *Bioinformatics*, 15:251–257, 1999.

[100] Alan G. Pockley. Heat shock proteins as regulators of the immune response. *The Lancet*, 362(9382):469–476, 2003.

[101] Marissa V. Powers and Paul Workman. Inhibitors of the heat shock response: Biology and pharmacology. *FEBS Letters*, 581(19):3758–3769, 2007.

[102] William H. Press, Saul A. Teukolsky, Wiliam T. Vetterling, and Brian P. Flannery. *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, New York, USA, 3rd edition, 2007.

[103] Corrado Priami. Algorithmic systems biology. *Communications of the ACM*, 52(5):80–88, 2009.

[104] Nathan D. Price, Jennifer L. Reed, and Bernhard Ø. Palsson. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nature Reviews Microbiology*, 2(11):886–897, 2004.

[105] Karthik Raman and Nagasuma Chandra. Systems biology. *Resonance*, 15(2):131–153, 2010.

[106] Daniel Remondini, Chiara Bernardini, Monica Forni, Ferdinando Bersani, Gastone C. Castellani, and Maria L. Bacci. Induced metastable memory in heat shock response. *Journal of Biological Physics*, 32(1):49–59, 2006.

[107] Theodore R. Rieger, Richard I. Morimoto, and Vassily Hatzimanikatis. Mathematical modeling of the eukaryotic heat shock response: Dynamics of the hsp70 promoter. *Biophysical Journal*, 88(3):1646–1658, 2005.

[108] Andrea Saltelli, Stefano Tarantola, Francesca Campolongo, and Marco Ratto. *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models*. John Wiley & Sons Ltd, Chichester, England, 2004.

[109] Hana El Samad, Mustafa Khammash, Linda Petzold, and Dan Gillespie. Stochastic modelling of gene regulatory networks. *International Journal of Robust and Nonlinear Control*, 15:691–711, 2005.

[110] Michael A. Savageau. Biochemical systems analysis: I. Some mathematical properties of the rate law for the component enzymatic reactions. *Journal of Theoretical Biology*, 25(3):365–369, 1969.

[111] Michael A. Savageau. Biochemical systems analysis: II. The steady state solution for an n-pool system using a power law approximation. *Journal of Theoretical Biology*, 25(3):370–379, 1969.

[112] Michael A. Savageau. Biochemical systems analysis: III. Dynamic solutions using a power-law approximation. *Journal of Theoretical Biology*, 26(2):215–226, 1970.

[113] Michael A. Savageau. The behavior of intact biochemical control systems. *Current Topics in Cellular Regulation*, 6:63–130, 1972.

[114] Michael A. Savageau. Optimal design of feedback control by inhibition: steady state considerations. *Journal of Molecular Evolution*, 4:139–156, 1974.

[115] Christophe H. Schilling, David Letscher, and Bernhard Ø. Palsson. Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *Journal of Theoretical Biology*, 203(3):229–248, 2000.

[116] Christophe H. Schilling, Stefan Schuster, Bernhard O. Palsson, and Reinhart Heinrich. Metabolic Pathway Analysis: Basic Concepts and Scientific Applications in the Post-genomic Era. *Biotechnological Progress*, 15(3):296–303, 1999.

[117] Stefan Schuster, Thomas Dandekar, and David A. Fell. Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends in biotechnology*, 17(2):53–60, 1999.

[118] Stefan Schuster, David A. Fell, and Thomas Dandekar. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nature Biotechnology*, 18:326–332, 2000.

[119] Stefan Schuster, Claus Hilgetag, John H. Woods, and David A. Fell. Reaction routes in biochemical reaction systems: algebraic properties, validated calculation procedure and example from nucleotide metabolism. *Journal of Mathematical Biology*, 45(2):153–181, 2002.

[120] Eduardo D. Sontag. Some new directions in control theory inspired by systems biology. *IEE Systems Biology*, 1(1):9–18, 2004.

[121] Eduardo D. Sontag. Molecular systems biology and control. *European Journal of Control*, 11(4):396–435, 2005.

[122] Ranjan Srivastava, Marvin S. Peterson, and William E. Bentley. Stochastic kinetic analysis of the Escherichia coli stress circuit using σ^{32}-targeted antisense. *Biotechnology and Bioengineering*, 75(1):120–129, 2001.

[123] Ranjan Srivastava, Lingchong You, Jesse Summers, and John Yin. Stochastic vs. deterministic modeling of intracellular viral kinetics. *Journal of Theoretical Biology*, 218(3):309–321, 2002.

[124] Michel O. Steinmetz, Daniel Stoffler, Andreas Hoenger, Andreas Bremer, and Ueli Aebi. Actin: From cell biology to atomic detail. *Journal of Structural Biology*, 119(3):295–320, 1997.

[125] Jörg Stelling, Uwe Sauer, Zoltan Szallasi, Francis J. Doyle III, and John Doyle. Robustness of cellular functions. *Cell*, 118(6):675–685, 2004.

[126] William J. Stewart. *Probability, Markov Chains, Queues, and Simulation. The Mathematical Basis of Performance Modeling.* Princeton University Press, Princeton, New Jersey, 2009.

[127] Frederick J. W. Symons. *Modelling and analysis of communication protocols using numerical Petri nets.* PhD thesis, Department of Electrical Engineering Science, University of Essex, Essex, England, 1978.

[128] Zuzanna Szymańska and Maciej Żylicz. Mathematical modeling of heat shock protein synthesis in response to temperature change. *Journal of Theoretical Biology*, 259(3):562–569, 2009.

[129] Antonella La Terza, Giampaolo Papa, Cristina Miceli, and Pierangelo Luporini. Divergence between two Antarctic species of the ciliate *Euplotes*, *E. focardii* and *E. nobilii*, in the expression of heat-shock protein 70 genes. *Molecular Ecology*, 10(4):1061–1067, 2001.

[130] René Thomas. Regulatory networks seen as asynchronous automata: A logical description. *Journal of Theoretical Biology*, 153(1):1–23, 1991.

[131] Nivaldo J. Tro. *Chemistry: A Molecular Approach*. Prentice Hall, 2nd edition, 2010.

[132] Tamás Turányi. Sensitivity analysis of complex kinetic systems. Tools and applications. *Journal of Mathematical Chemistry*, 5(3):203–248, 1990.

[133] Natal A. W. van Riel. Dynamic modelling and analysis of biochemical networks: mechanism-basedmodels and model-based experiments. *Briefings in Bioinformatics*, 7(4):364–374, 2006.

[134] Richard Voellmy and Frank Boellmann. Chaperone regulation of the heat shock protein response. In Peter Csermely and László Vígh, editors, *Molecular Aspects of the Stress Response: Chaperones, Membranes and Networks*, volume 594 of *Advances in Experimental Medicine and Biology*, chapter 9, pages 89–99. Springer Science+Business Media, LLC/Landes Bioscience/Eurekah.com, New York, 2007.

[135] Darren J. Wilkinson. *Stochastic Modelling for Systems Biology*. Mathematical and Computational Biology Series. Chapman & Hall/CRC, Boca Raton, FL, USA, 2006.

[136] Olaf Wolkenhauer. Systems biology: The reincarnation of systems theory applied in biology? *Briefings in Bioinformatics*, 2(3):258–270, 2001.

[137] Olaf Wolkenhauer. *Systems Biology. Dynamic Pathway Modelling*. In preparation, 2010. `http://www.sbi.uni-rostock.de/dokumente/t_sb.pdf`.

[138] Olaf Wolkenhauer and Mihajlo Mesarović. Feedback dynamics and cell function: Why systems biology is called Systems Biology. *Molecular BioSystems*, 1:14–16, 2005.

[139] Olaf Wolkenhauer, Mukhtar Ullah, Walter Kolch, and Kwang-Hyun Cho. Modelling and simulation of intracellular dynamics: Choosing

an appropriate framework. *IEEE Transactions on NanoBioscience*, 3:200–207, 2004.

[140] Paul Workman and Emmanuel de Billy. Putting the heat on cancer. *Nature Medicine*, 13(12):1415–1417, 2007.

[141] Zhike Zi, Yanan Zheng, Ann E. Rundell, and Edda Klipp. SBML-SAT: a systems biology markup language (SBML) based sensitivity analysis tool. *BMC Bioinformatics*, 9:342, 2008.

Paper I

A simple mass-action model for the eukaryotic heat shock response and its mathematical validation

Ion Petre, Andrzej Mizera, Claire L. Hyder, Annika Meinander, Andrey Mikhailov, Richard I. Morimoto, Lea Sistonen, John E. Eriksson, and Ralph-Johan Back

Originally published in *Natural Computing*, 10(1):595-612, 2011.

Nat Comput
DOI 10.1007/s11047-010-9216-y

A simple mass-action model for the eukaryotic heat shock response and its mathematical validation

Ion Petre · Andrzej Mizera · Claire L. Hyder · Annika Meinander · Andrey Mikhailov · Richard I. Morimoto · Lea Sistonen · John E. Eriksson · Ralph-Johan Back

Abstract The heat shock response is a primordial defense mechanism against cell stress and protein misfolding. It proceeds with the minimum number of mechanisms that any regulatory network must include, a stress-induced activation and a feedback regulation, and can thus be regarded as the archetype for a cellular regulatory process. We propose here a simple mechanistic model for the eukaryotic heat shock response, including its mathematical validation. Based on numerical predictions of the model and on its sensitivity analysis, we minimize the model by identifying the reactions with marginal contribution to the heat shock response. As the heat shock response is a very basic and conserved regulatory network, our analysis of the network provides a useful foundation for modeling strategies of more complex cellular processes.

Keywords Heat shock response · Heat shock protein · Heat shock factor · Heat shock element · Mathematical model · Validation · Regulatory network

1 Introduction

The heat shock response is an ancient, evolutionary conserved regulatory mechanism that allows the cell to quickly react to elevated temperatures and other forms of physiological and environmental stress. The heat shock response has been subject of active research (see

I. Petre (✉) · A. Mizera · R.-J. Back
Department of Information Technologies, Åbo Akademi University, Turku 20520, Finland
e-mail: ipetre@abo.fi

C. L. Hyder · A. Meinander · A. Mikhailov · L. Sistonen · J. E. Eriksson
Turku Centre for Biotechnology, Turku, Finland

C. L. Hyder · A. Meinander · A. Mikhailov · L. Sistonen · J. E. Eriksson
Department of Biosciences, Åbo Akademi University, Turku 20520, Finland

R. I. Morimoto
Department of Biochemistry, Molecular Biology and Cell Biology, Rice Institute for Biomedical Research, Northwestern University, Evanston, IL 60208, USA

Published online: 02 September 2010

Powers and Workman 2007; Chen et al. 2007; Voellmy and Boellmann 2007) for at least two reasons. On one hand, as it represents an exceptionally well-conserved signaling mechanism, it is a good candidate for deciphering the mechanistic principles of gene regulatory networks. On the other hand, heat shock proteins have essential roles in all aspects of protein biogenesis, regardless of the regulatory aspects of the heat shock response, and have fundamental importance for many key biological processes. Therefore, understanding the details of the heat shock response has broad ramifications for the biology of the cell and response to cellular insults and for the onset and treatment of a number of diseases, including neurodegenerative disorders, cancer, aging, and cardiovascular diseases (see Balch et al. 2008; Morimoto 2008).

Despite intense research and a number of models that have been presented to cover the heat shock response, a comprehensive mechanistic understanding of this process is lacking. Here, we propose a simple model capturing in mechanistic details all key aspects of the regulation: the heat-induced protein misfolding, the chaperone activity of heat shock proteins, the transactivation of the genes encoding heat shock proteins and the repression of their transcription once the stress is removed. In contrast with previous attempts to model the eukaryotic heat shock response, our model is based solely on well-documented molecular reactions and does not include modeling "blackboxes" such as experimentally unsupported components and biochemical reactions.

We also present a mathematical model associated with the model and its experimental validation. For specific parameter estimation and model validation, we use already published data (Kline and Morimoto 1997), as well as new experimental data. The model predictions correlate well with experimental data on the heat-induced transactivation of the genes encoding heat shock proteins at various temperatures, its return to the original level once the stress is removed, and a lower response to a second consecutive heat shock. We use the model to identify a number of reactions that could be eliminated from the model without affecting its quantitative behavior. We also identify the most significant reactions regulating the levels of the heat shock proteins and those of the misfolded proteins. This analysis deepens our understanding of where the significant control resides in the network.

2 Results

2.1 Molecular model

The heat shock protein (hsp) plays the central role as a chaperone to prevent misfolding, to capture intermediates, and to facilitate protein folding. Even though there are multiple classes of hsps, with various molecular masses and different regulatory mechanisms, we treated them all uniformly in our model, with hsp 70 as the base denominator. The hsp-encoding genes are transactivated through the binding of heat shock factors (hsf) to the heat shock element (hse) found on the DNA upstream of the gene. Even though several types of heat shock factors exist (HSFs1-4) (see Holmberg et al. 2002), we focused on HSF1 in our model. The binding of a heat shock factor trimer (hsf_3) to a heat shock element was denoted as $hsf_3 : hse$. Heat shock proteins may bind to heat shock factors; we denoted such a bond as hsp:hsf. The drivers of the whole heat shock response are the heat-induced misfolded proteins, denoted mfp. Binding of a heat shock protein to a misfolded protein was denoted as hsp:mfp. We made no distinction among the many types of protein substrates that exist in the cell. From the point of view of the heat shock response, we were only interested in whether they are correctly folded (collected globally under the name

prot), or misfolded (collected globally under the name mfp). What drives the heat shock response is the race to keep the level of misfolded proteins under control, in such a way that they are not able to accumulate, form aggregates, and eventually lead to cell death.

Our molecular model for the heat shock response consists of three parts: the dynamic transactivation of the hsp-encoding genes, their backregulation, and the chaperone activity of the hsp. In the absence of the heat stress, the heat shock factors are present as monomers, mainly bounded to heat shock proteins. There is insignificant variation in their concentration with stress. Upon heat stress, however, the heat shock factors form trimers, which are the active components, able to bind to heat shock elements (see Voellmy 1994; Morimoto et al. 1994). Once hsf_3 is bound to the heat shock element, we assumed that the hsp-encoding gene is transcriptionally active. We did not model explicitly the transcription machinery binding to the promoter region of the hsp-encoding gene, the mRNA molecules being produced, edited, transported, etc., but only represented that a transcriptionally active hsp-encoding gene will eventually yield the synthesis of new hsp molecules, see reaction (4) in Table 1. Heat shock proteins have an affinity for heat shock factors and so, if present in sufficient amounts, are able to shut down their own synthesis: a heat shock protein hsp contributes to unbinding a trimer hsf_3 from the heat shock element, see reaction (8) in Table 1 and (Abravaya et al. 1992; Shi et al. 1998).

The heat-induced misfolding of proteins was represented in our model as a reaction switching an unfolded or native protein (prot) to misfolded (mfp). The reaction rate depends exponentially on the temperature of the environment (see Peper et al. 1997; Lepock et al. 1993). A heat shock protein may chaperone a misfolded protein and facilitate its refolding. The list of all reactions in our molecular model is given in Table 1.

There are three conservation relations in our model. One concerns the total amount of hsf:

$$[\mathsf{hsf}] + 2 \times [\mathsf{hsf}_2] + 3 \times [\mathsf{hsf}_3] + 3 \times [\mathsf{hsf}_3 : \mathsf{hse}] + [\mathsf{hsp} : \mathsf{hsf}] = \text{constant}. \quad \text{(C1)}$$

The second concerns the total amount of proteins, other than hsp and hsf:

$$[\mathsf{prot}] + [\mathsf{mfp}] + [\mathsf{hsp} : \mathsf{mfp}] = \text{constant}. \quad \text{(C2)}$$

The third concerns the total amount of heat shock elements:

Table 1 The list of reactions in the molecular model for the heat shock response

Reaction	(Reaction number)
$2\mathsf{hsf} \leftrightarrow \mathsf{hsf}_2$	(1)
$\mathsf{hsf} + \mathsf{hsf}_2 \leftrightarrow \mathsf{hsf}_3$	(2)
$\mathsf{hsf}_3 + \mathsf{hse} \leftrightarrow \mathsf{hsf}_3 : \mathsf{hse}$	(3)
$\mathsf{hsf}_3 : \mathsf{hse} \rightarrow \mathsf{hsf}_3 : \mathsf{hse} + \mathsf{hsp}$	(4)
$\mathsf{hsp} + \mathsf{hsf} \leftrightarrow \mathsf{hsp} : \mathsf{hsf}$	(5)
$\mathsf{hsp} + \mathsf{hsf}_2 \rightarrow \mathsf{hsp} : \mathsf{hsf} + \mathsf{hsf}$	(6)
$\mathsf{hsp} + \mathsf{hsf}_3 \rightarrow \mathsf{hsp} : \mathsf{hsf} + 2\mathsf{hsf}$	(7)
$\mathsf{hsp} + \mathsf{hsf}_3 : \mathsf{hse} \rightarrow \mathsf{hsp} : \mathsf{hsf} + \mathsf{hse} + 2\mathsf{hsf}$	(8)
$\mathsf{hsp} \rightarrow$	(9)
$\mathsf{prot} \rightarrow \mathsf{mfp}$	(10)
$\mathsf{hsp} + \mathsf{mfp} \leftrightarrow \mathsf{hsp} : \mathsf{mfp}$	(11)
$\mathsf{hsp} : \mathsf{mfp} \rightarrow \mathsf{hsp} + \mathsf{prot}$	(12)

$$[\mathsf{hse}] + [\mathsf{hsf_3} : \mathsf{hse}] = \text{constant}. \tag{C3}$$

The only variable of the model not covered by the conservation relations is hsp, which is the regulatory target of the heat shock response.

2.2 Mathematical model and parameter estimation

In developing the mathematical model, we assumed for all reactions the principle of *mass-action*, that can be briefly summarized as follows: the flux of each reaction is proportional to the amount of input to the reaction (see Guldberg and Waage 1864, 1879). The reason why we preferred a simple mass-action formalization rather than more sophisticated approaches such as Michaelis-Menten or Hill equations was so that we could follow the explicit effect of each individual reaction to the overall response. We expressed our model in terms of differential equations, with one function associated to each component in the model. The resulting mathematical model consists of 10 differential equations and is shown in Table 2. Of these 10 equations, based on the conservation relations (C1–C3), only seven equations are independent.

In Table 2, we denoted by k_i the reaction rate constant of the irreversible reaction (i) in Table 1, by k_i^+, the reaction rate constant corresponding to the 'left-to-right' direction of the reversible reaction (i) in the same table, while k_i^- denotes the rate constant corresponding to its 'right-to-left' direction, for all $1 \leq i \leq 12$. We denoted by T the temperature of the environment.

The extent of heat-induced protein denaturation in CHL V79 cells has been investigated in Lepock et al. (1993). Based on that study, the fractional protein denaturation per hour was deduced in Peper et al. (1997). Since our model uses the second as time unit, we

Table 2 The differential equations of the associated mathematical model

Equation	(Equation number)
$d[\mathsf{hsf}]/dt = -2k_1^+[\mathsf{hsf}]^2 + 2k_1^-[\mathsf{hsf_2}] - k_2^+[\mathsf{hsf}][\mathsf{hsf_2}] + k_2^-[\mathsf{hsf_3}] - k_5^+[\mathsf{hsf}][\mathsf{hsp}] + k_5^-[\mathsf{hsp} : \mathsf{hsf}] + k_6[\mathsf{hsf_2}][\mathsf{hsp}] + 2k_7[\mathsf{hsf_3}][\mathsf{hsp}] + 2k_8(\mathsf{hsf_3} : hse)\mathsf{hsp}$	(13)
$d[\mathsf{hsf_2}]/dt = k_1^+[\mathsf{hsf}]^2 - k_1^-[\mathsf{hsf_2}] - k_2^+[\mathsf{hsf}][\mathsf{hsf_2}] + k_2^-[\mathsf{hsf_3}] - k_6[\mathsf{hsf_2}][\mathsf{hsp}]$	(14)
$d[\mathsf{hsf_3}]/dt = k_2^+[\mathsf{hsf}][\mathsf{hsf_2}] - k_2^-[\mathsf{hsf_3}] - k_3^+[\mathsf{hsf_3}][\mathsf{hse}] + k_3^-[\mathsf{hsf_3} : \mathsf{hse}] - k_7[\mathsf{hsf_3}][\mathsf{hsp}]$	(15)
	(16)
$d[\mathsf{hse}]/dt = -k_3^+[\mathsf{hsf_3}][\mathsf{hse}] + k_3^-[\mathsf{hsf_3} : \mathsf{hse}] + k_8[\mathsf{hsf_3} : \mathsf{hse}][\mathsf{hsp}]$	(17)
$d[\mathsf{hsf_3} : \mathsf{hse}]/dt = k_3^+[\mathsf{hsf_3}][\mathsf{hse}] - k_3^-[\mathsf{hsf_3} : \mathsf{hse}] - k_8[\mathsf{hsf_3} : \mathsf{hse}][\mathsf{hsp}]$	(18)
$d[\mathsf{hsp}]/dt = k_4[\mathsf{hsf_3} : \mathsf{hse}] - k_5^+[\mathsf{hsf}][\mathsf{hsp}] + k_5^-[\mathsf{hsp} : \mathsf{hsf}] - k_6[\mathsf{hsf_2}][\mathsf{hsp}] - k_7[\mathsf{hsf_3}][\mathsf{hsp}] - k_8[\mathsf{hsf_3} : \mathsf{hse}][\mathsf{hsp}] - k_{11}^+[\mathsf{hsp}][\mathsf{mfp}] + (k_{11}^- + k_{12})[\mathsf{hsp} : \mathsf{mfp}] - k_9[\mathsf{hsp}]$	(19)
$d[\mathsf{hsp} : \mathsf{hsf}]/dt = k_5^+[\mathsf{hsf}][\mathsf{hsp}] - k_5^-[\mathsf{hsp} : \mathsf{hsf}] + k_6[\mathsf{hsf_2}][\mathsf{hsp}] + k_7[\mathsf{hsf_3}][\mathsf{hsp}] + k_8[\mathsf{hsf_3} : \mathsf{hse}][\mathsf{hsp}]$	(20)
$d[\mathsf{mfp}]/dt = \phi_T[\mathsf{prot}] - k_{11}^+[\mathsf{hsp}][\mathsf{mfp}] + k_{11}^-[\mathsf{hsp} : \mathsf{mfp}]$	(21)
$d[\mathsf{hsp} : \mathsf{mfp}]/dt = k_{11}^+[\mathsf{hsp}][\mathsf{mfp}] - (k_{11}^- + k_{12})[\mathsf{hsp} : \mathsf{mfp}]$	(22)
$d[\mathsf{prot}]/dt = -\phi_T[\mathsf{prot}] + k_{12}[\mathsf{hsp} : \mathsf{mfp}]$	(23)

adapted the fractional protein denaturation per second ϕ_T from Peper et al. (1997) to obtain the temperature-dependant formula

$$\phi_T = \left(1 - \frac{0.4}{e^{T-37}}\right) \times 1.4^{T-37} \times 1.45 \times 10^{-5} \quad \mathrm{s}^{-1},$$

where T is the temperature of the environment in Celsius degrees. According to Lepock et al. (1993), this formula is valid for temperatures between 37 and 45°C.

There are 17 independent parameters in our model and 10 initial conditions that must be specified or estimated. We had on the other hand the three conservation relations (C1–C3) that leave only seven initial conditions to specify. In estimating our parameters we used experimental data of Kline and Morimoto (1997) on the rate of $\mathsf{hsf}_3 : \mathsf{hse}$ during a heat shock of HeLa cells at 42°C. In addition, we also imposed the condition that with the same initial values and the same numerical parameters, the model is at steady state if the temperature is 37°C (by definition, the heat shock response is triggered for temperatures upwards of 37°C). This yields 7 independent algebraic relations on the set of parameters and initial values. Thus, we have altogether 17 independent values that we need to estimate.

By performing parameter estimation in COPASI (Hoops et al. 2006), we obtained the values shown in Table 3 that satisfy the conditions above. The model fit with respect to the data in Kline and Morimoto (1997) is shown in Fig. 1a.

2.3 Model validation

In the final model, we obtained that protein misfolding occurs at 37°C at very low rate, that hsp are long-lived molecules, and that the protein folding is a fast reaction, which is in accordance with Jones et al. (1993), Ballew et al. (1996) (We are disregarding in the model folding intermediates.). Moreover, the model correctly predicted (see Holmberg et al. 2002), that under heat shock, the level of hsf trimers is transiently increased. The model was also able to confirm that the hsf dimers are only a transient state between monomers and trimers and that their level remains low at all times, independent of the temperature.

In another validation test, we considered a heat shock applied in two stages, with a recovery period between them, with the second shock applied after the level of hsp has reached a maximum. We observed, similarly as in Peper et al. (1997), that the predicted response of the model to the second heat shock is much milder, see Fig. 2a. This is consistent with the expectation that due to the first heat shock, the level of hsp is already raised, and so the cell may react to the second shock, with a lower $[\mathsf{hsf}_3 : \mathsf{hse}]$ peak.

We also considered a heat shock at 43°C and compared our prediction to that of Rieger et al. (2005). Similarly as shown by the experimental data in Abravaya et al. (1991), our model was able to show prolonged transactivation, see Fig. 2b, unlike the model in Rieger et al. (2005). An experiment where the heat shock at 42°C is removed at the peak of the response showed a faster attenuation phase, similarly as reported in Rieger et al. (2005), see Fig. 2b. Several sensitivity analysis experiments, where some parameters are set to lower or higher values agreed with the predictions made in similar experiments by Rieger et al. (2005).

For further verification of our model and its prediction abilities, we performed a set of experiments. Specifically, we aimed to validate the numerical prediction on the level of hsp over time. Our approach was to use a suitable quantitative reporter system based on yellow fluorescent proteins (yfp). Our setup was designed so that the kinetics of the reporter gene's transactivation mimics the results obtained in experimental studies on endogenous hsf target

Table 3 The numerical values of the parameters and the initial valus of the variables of the heat shock response model

Param.	Value	Units
The numerical values of the parameters		
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^-	1.21×10^{-6}	s^{-1}
k_4	8.3×10^{-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	4.31×10^{-5}	$\frac{ml}{\#\cdot s}$
k_8	2.73×10^{-7}	$\frac{ml}{\#\cdot s}$
k_9	3.2×10^{-5}	s^{-1}
k_{11}^+	3.32×10^{-3}	$\frac{ml}{\#\cdot s}$
k_{11}^-	4.44	s^{-1}
k_{12}	13.94	s^{-1}
Variable		Initial conc.
The initial values of all variables		
[hsf]		0.67
[hsf_2]		8.7×10^{-4}
[hsf_3]		1.2×10^{-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

genes. In this way, the dynamics of yfp partially reports on the dynamics of hsp. We did not make any assumptions on the stability of the yfp proteins. Rather, this issue was dealt with in the mathematical validation process. To this aim, we employed K562 cells, expressing a 712 bp fragment of the hsp70 promoter fused to a yellow fluorescent protein (yfp) reporter gene. The cells were subjected to a continuous heat shock at 42°C and samples were taken at indicated time points (for details, see "Materials and methods" section). hsp70 promoter activity as a result of expression of yfp was analyzed by flow cytometry to give a measure of the heat shock response in individual cells.

In three independent biological repeats, we measured the fluorescence intensity of 10000 cells for each time point (15 of them up to 36 h). Our assumption was that the fluorescence intensity is roughly linear with respect to the level of the yellow fluorescent proteins (yfp) in our sample. Given that the transactivation of the yfp genes is controlled by

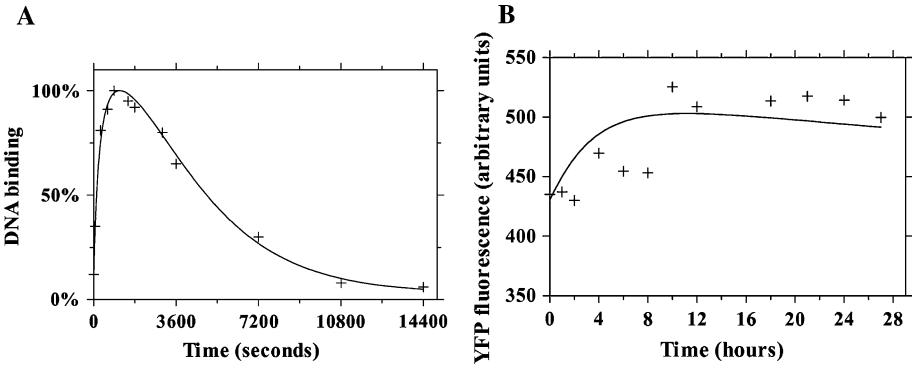


Fig. 1 Comparison of the numerical predictions of the model with two sets of experimental data. **a** The model fit with respect to the experimental data in Kline and Morimoto (1997). The *thick line* is the model prediction regarding $[\mathsf{hsf}_3 : \mathsf{hse}]$, that is compared with the experimental data showed with crossed points. Both plots are relative to their maximum value. **b** Model validation based on fluorescence intensity of cells transfected by hse-controlled genes coding for yellow fluorescent proteins. The *crossed dots* are the mean values of the experimental data, while the *continuous line* is the numerical integration of the benchmark variable

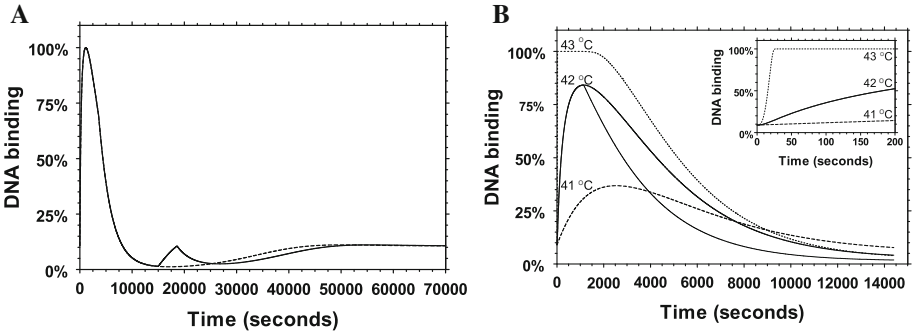


Fig. 2 Numerical predictions of the model. **a** The model correctly predicts that DNA binding peaks at a much lower level in a second consecutive heat shock. The experiment with a single heat shock is shown with a dashed line. **b** The model correctly predicts longer transactivation with higher heat shock: the behaviors at 41, 42, and 43°C are shown. We also plot on the same graph the correct prediction that the DNA binding attenuates more rapidly in an experiment where the heat shock at 42°C is removed at the peak of the response

their own heat shock elements hse', transcription/translation and degradation kinetics k_4' and k_9', resp., we obtained that

$$d[\mathsf{yfp}]/dt = k_4'[\mathsf{hsf}_3 : \mathsf{hse}'] - k_9'[\mathsf{yfp}],$$

for some positive constants k_4', k_9' standing for the kinetic rate constants of the yfp synthesis and of the yfp degradation, respectively. The numerical values of parameters were not deduced from the basic model to underline that we made no assumptions on the stability of yfp, or on their gene transcription rates. The idea of the validation was to extend the already fit basic model so as to include also yfp. In the extended model we re-used all the kinetic rate constants of the basic model. We then looked for numerical values for parameters k_4' and k_9' and for initial values of all variables of the model so that the

numerical prediction for yfp fit well with the experimental data. The result of the validation is shown in Fig. 1b, where the crossed points represent the mean values of the experimental data at each time point and the continuous line is the numerical integration of yfp.

2.4 Model analysis

We estimated the scaled steady state sensitivity coefficients (see Turanyi 1990), of all variables of the model with respect to reaction rate constants and with respects to initial concentrations. For a variable X of the model and a parameter p, the scaled steady state sensitivity coefficient of X with respect to p is $\lim_{t\to\infty} \partial \ln(X)/\partial \ln(p)(t)$. These coefficients measure the relative change in steady state when some parameter is changed with an infinitesimally small amount. They help identify the most important steps in the heat shock response network. A first observation was that the sensitivity coefficients of all variables of the model with respect to reaction rate constants k_1^-, k_2^-, k_3^-, and k_7 are negligible. This suggested that the respective reactions may have negligible effect on the overall behavior of the model. To test this prediction, we removed reaction (7) and the right-to-left directions of reactions (1), (2) and (3). The reactions of the reduced model are in Table 4 and their kinetic constants are unchanged with respect to the basic model. It turned out that the reduced model performs equally well as the basic model in all validation tests described above. Our model thus predicted that hsf dimers and trimers are very stable and do not break spontaneously at a significant rate. The spontaneous unbinding of an hsf trimer from hse (without the involvement of hsp) was also insignificant. Interestingly, while reaction (7) (hsp breaking hsf trimers) did not have a significant role and could be eliminated from the model, reaction (6) (hsp breaking hsf dimers) did have a significant influence on several variables of the model, including hsp and mfp.

We focused on the sensitivity coefficients of hsp and mfp, the main drivers of the response. They showed a direct correlation between variations in the steady state levels of hsp and mfp, not surprising given the chaperoning role of hsp. Their largest sensitivity coefficients are in Table 5 and can be interpreted as follows. The coefficients with respect to k_5^+ and k_5^- being the largest identified reaction (5) in Table 1 as the most important feedback loop in our model. In one direction of reaction (5), hsf is sequestrated, leading eventually to a suppression of the transcription, in concert with reaction (8), and consequently, to a reduction in hsp and an increase in mfp. In the other direction of reaction (5),

Table 4 The list of reactions in the reduced molecular model

Reaction
$2\mathsf{hsf} \rightarrow \mathsf{hsf}_2$
$\mathsf{hsf} + \mathsf{hsf}_2 \rightarrow \mathsf{hsf}_3$
$\mathsf{hsf}_3 + \mathsf{hse} \rightarrow \mathsf{hsf}_3 : \mathsf{hse}$
$\mathsf{hsf}_3 : \mathsf{hse} \rightarrow \mathsf{hsf}_3 : \mathsf{hse} + \mathsf{hsp}$
$\mathsf{hsp} + \mathsf{hsf} \leftrightarrow \mathsf{hsp} : \mathsf{hsf}$
$\mathsf{hsp} + \mathsf{hsf}_2 \rightarrow \mathsf{hsp} : \mathsf{hsf} + \mathsf{hsf}$
$\mathsf{hsp} + \mathsf{hsf}_3 : \mathsf{hse} \rightarrow \mathsf{hsp} : \mathsf{hsf} + \mathsf{hse} + 2\mathsf{hsf}$
$\mathsf{hsp} \rightarrow$
$\mathsf{prot} \rightarrow \mathsf{mfp}$
$\mathsf{hsp} + \mathsf{mfp} \leftrightarrow \mathsf{hsp} : \mathsf{mfp}$
$\mathsf{hsp} : \mathsf{mfp} \rightarrow \mathsf{hsp} + \mathsf{prot}$

Reactions (1) (right-to-left), (2) (right-to-left), (3) (right-to-left), and (7) were eliminated from the basic model in Table 1 without affecting its numerical behavior

Table 5 The largest scaled steady state sensitivity coefficients of hsp and mfp

Description	p	Sensitivity $\frac{\partial \ln(\mathsf{hsp})}{\partial \ln(p)}\big\vert_{t\to\infty}$	Sensitivity $\frac{\partial \ln(\mathsf{mfp})}{\partial \ln(p)}\big\vert_{t\to\infty}$
Sequestration of hsf by hsp	k_5^+	−0.50	0.50
Dissipation of hsp:hsf	k_5^-	0.50	−0.50
Formation of hsf dimers	k_1^+	0.17	−0.17
Formation of hsf 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
Refolding	k_{12}	0.00	−0.24
Initial level of hsp:hsf	hsp:hsf(0)	0.50	−0.50

The coefficients are identical at 37 and 42°C

hsp and hsf levels are increased, both leading to increasing hsp and decreasing mfp. The next largest coefficients are with respect to k_1^+, k_2^+ and k_4: reactions (1), (2), and (4) all contribute to increasing the level of transcription and by consequence, the level of hsp as follows: hsf dimers or trimers form at a higher rate, or hsf_3 binds to hse at a higher rate. Reactions (6), (8) and (9) (see the sensitivity coefficients with respect to k_6, k_8, k_9) have a countering effect on the level of transcription or directly on that of hsp:hsf: dimers are dissipated at a higher rate and are less able to form trimers, hsf_3 unbinds from hse at a higher rate, or hsp degrades at a higher rate. The only reactions that influenced the level of mfp but not that of hsp are (11) and (12), see the sensitivity coefficients with respect to k_{11}^+, k_{11}^-, and k_{12} in Table 5. These reactions control the chaperoning and the refolding of mfp, while not consuming hsp.

The most significant sensitivity coefficient of hsp and of mfp with respect to initial concentrations was that depending on hsp:hsf(0), where hsp:hsf(0) denotes the initial level of hsp:hsf, with similar notations for the other variables of the model. On the other hand, the sensitivity coefficients of both hsp and of mfp on the other forms of hsf (monomer, dimer, trimer) were negligible. This is a direct consequence of the fact that almost all initial amount of hsf is sequestered by hsp, while the initial levels of dimers and trimers are very low (in line with experimental observations of Holmberg et al. (2002)). As such the dependency on hsp:hsf(0) should rather be interpreted as a dependency on the total initial amount of hsf. This interpretation was supported by the following numerical experiment. We set hsp:hsf(0) to 0 and increase correspondingly hsp(0) and hsf(0) (or, alternatively, $\mathsf{hsf}_2(0)$, or $\mathsf{hsf}_3(0)$) in such a way that the initial total amount of hsp and of hsf is unchanged. Then hsp and mfp got significant sensitivity coefficients with respect to hsf(0)($\mathsf{hsf}_2(0)$, or $\mathsf{hsf}_3(0)$, respectively) and negligible with respect to hsp : hsf(0). Distributing the initial amount of hsf among its various forms had, however, a crucial effect on the speed and on the peak of the response.

The scaled steady state sensitivity coefficients of both hsp and mfp with respect to hse(0) were negligible. This result is explained by the fact that we considered the sensitivities around the steady state. For example, with fewer hse(0), the response will eventually be able to approach the same steady state, albeit the transcription stays at the 100% level for a longer time (because a lower [hse] becomes a bottleneck of the

response). Interestingly, with a higher hse(0), the time evolution of the response remained unchanged, indicating that as long as hse(0) was higher than a certain threshold, its numerical value was irrelevant for the model prediction. This was indeed confirmed by numerical simulations.

The sensitivities of both hsp and mfp (and in fact those of all variables) with respect to hsp(0) were also negligible. The reason for this is that the system was able to self-regulate a lower/higher hsp(0) and eventually approach the same steady state. On the other hand, a lower/higher hsp(0) did have an impact on the time evolution of the response.

2.5 Alternative numerical model fits

The reaction rate constant values of our mathematical model were obtained by performing parameter estimation with respect to the experimental data of Kline and Morimoto (1997). We address in this section the question of the uniqueness of the set of parameters that fulfill the imposed conditions, a problem that is also known as the *model identifiability*. By repeating from scratch the whole parameter estimation procedure, we obtained several different sets of parameter values that result both in a good fit of the model to the experimental data, as well as in initial values that are steady-states of the model at 37°C. It turned out, however, that all these parameter sets failed the model validation tests discussed above with respect to the qualitative observations concerning the behavior of cells under stress. This does not prove that our heat shock response model is uniquely identifiable. However, it does suggest that fitting the model to the experimental data in Kline and Morimoto (1997) and to the steady-state condition for the initial values is a difficult numerical problem.

A thorough method of searching for alternative numerical model fits is to perform a systematic parameter scan in the space determined by the considered ranges of parameter values. This means that for each parameter, one partitions its value range into a large number of subintervals (say, tens of thousands of them) and samples values for the parameter from all of them. One then tests the quality of the model fit for all possible combinations of parameter samples to yield a thorough sampling of the model behavior throughout the multi-dimensional parameter space. Unfortunately, the direct implementation of this idea is intractable for models with more than a few parameters due to the combinatorial explosion of the number of simulations that need to be run. A fast, practical solution to this problem is to apply the *Latin Hypercube Sampling* method (LHS), first introduced in McKay et al. (1979). This method provides samples which are uniformly distributed over each parameter while the number of samples is independent of the number of parameters (see also Helton and Davis 2002, 2003; Oberguggenberger et al. 2009) for applications of this method. We describe the sampling scheme briefly in the following, in the simpler case when the parameter values are uniformly distributed in their range interval. One first chooses the desired size N of the sampling set. The range interval of each parameter is then partitioned into N non-overlapping intervals of equal length. For each parameter, we randomly select N numerical values, one from from each interval of the partition. We collect the N sampled values for the i-th parameter of the model on the i-th column of a $N \times p$ matrix, where p is the number of parameters. One then randomly shuffles the values on each column. The result of the procedure is read from the rows of the matrix: each of the N rows of the matrix contains numerical values for each of the p parameters. For a detailed description of this sampling scheme we refer to McKay et al. (1979).

Based on the LHS method we have implemented the following strategy to look for alternative models fits that are in agreement with the experimental data of Kline and Morimoto (1997), and satisfy the steady-state condition for the initial values. First, by applying the LHS method, we sampled $N = 100.000$ sets of parameter values. For each set, we estimated numerically the steady state of the model for a temperature value of 37°C. We then set the initial state of the model as the calculated steady state. We simulated the model for 14400 s at a temperature of 42°C. Finally, we classified as *non-responsive* those parameter samples that led to low DNA binding level at the peak of the response, and excluded them from further analysis. We obtained that only 31.506 out of the 100.000 samples were responsive, already a result pointing to difficulties in finding satisfactory alternative numerical fits. We analyzed each of these models as follows. For each model, we made a scatter plot for each variable and each parameter where we plotted the steady state values of each variable at 37°C, against the values of the parameter. We discuss here only a few of the plots. All plots are available as supplementary materials at http://combio.abo./hsr/plots.zip.

We compared the obtained results with the steady state values of our basic model (called also reference model in the following) at 37°C. As can be seen in Fig. 3a, only very few of the sampled models were capable of reaching low levels of DNA binding at the steady state. This showed that most of the alternative fits predicted high levels of gene

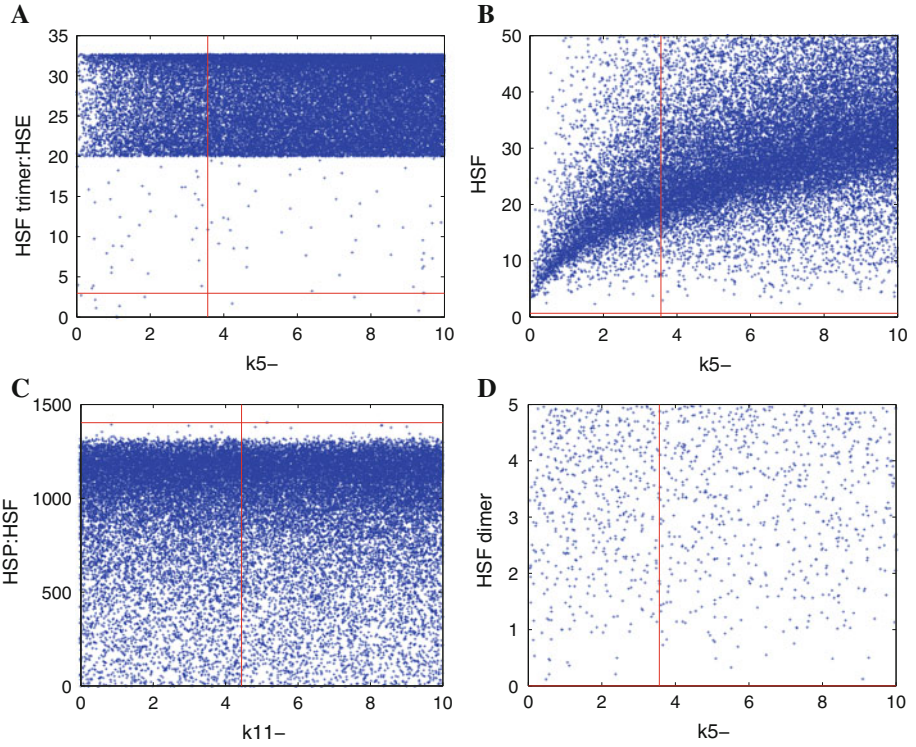


Fig. 3 Scatter plots of the steady state values at 37°C of the sampled models (*blue crosses*) and the basic model (*red horizontal line*). The *red vertical line* indicates the parameter value of the basic model. The plots of hsf (**b**) and hsf_2 (**d**) are zoomed in, hence not all points are present, i.e., the values of the remaining steady states were higher than the maximum value on the y-axes

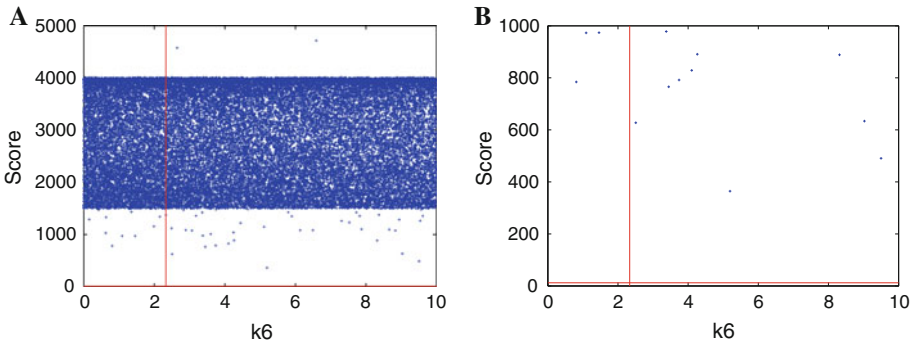


Fig. 4 Scatter plot (**a**) and its zoomed in version (**b**) of the score measuring the fit of the sampled models (*blue crosses*) and the basic model (*red horizontal line*) with respect to the experimental data. The *red vertical line* indicates the parameter value of the basic model

transcription in the absence of the heat shock, a contradiction of the available biological evidence (see Holmberg et al. 2002). In the case of hsf, none of the sampled models reached such a low level as the reference model, see Fig. 3b. Moreover, the reference model is one of the very few models in which most of the hsf molecules are sequestered by hsp, see Fig. 3c. This indicates that at the temperature of 37°C the response mechanism is turned off, which is in excellent agreement with biological observations (see Holmberg et al. 2002). These outcomes are also supported by the results obtained for hsf dimers presented in Fig. 3d, where the basic model reaches the lowest values. This is also in agreement with the observation that hsf dimers are unnoticeable in biological experiments.

We also compared the predictions of all the sampled models at 42°C with respect to the experimental data of Kline and Morimoto (1997). The same score function which was used in the case of parameter estimation, i.e., the sum of of squares of the residues, was computed for all considered models. The results are depicted in the form of a scatter plot and its zoomed in version in Fig. 4a and b, respectively. Our reference model obtained the lowest score of around 12, while the 13 best fits of the sampled models were in the range between 300 and 1000. All the other models had much worse scores, of more than 1000.

While it is likely that a model of this size is not uniquely identifiable, our parameter scan showed that finding parameter values satisfying our model constraints is far from being easy. This is evident both from the plots of the model deviation from the experimental data under stress (as measured by the score function, Fig. 4a, b) as well as from the plots of the model behavior in the absence of stress (Fig. 3a–d). Even more, about two thirds of the parameter samples led to none-responsive models, i.e., models that yield an insufficient response under stress.

3 Discussion

We presented a simple molecular model for the heat shock response, based on standard molecular biology only. The mathematical model was validated based both on existing data from the literature, as well as on our novel experimental evidence. The numerical simulations of the model correlate well with predictions reported elsewhere in the literature.

Using sensitivity coefficients we predict that a number of reactions have a negligible effect on the model and could be removed without affecting its numerical behavior. We also identify the reactions with the most significant effect on hsp and on mfp. This is a useful, still not fully exploited potential of mathematical modeling in biology. We have started from a molecular model that incorporated a number of reactions that could in principle take place even though no direct experimental evidence in their support exists: the dissipation of dimers and of trimers, or the spontaneous unbinding of hsf_3 from hse. The mathematical analysis of the model points out to the fact that these reactions have a negligible effect on the overall behavior of the model and it suggests that they could be eliminated from the model. These results help simplify the molecular model which in turn is important for further, more complex analysis of the associated mathematical model and for their integration into larger models. They can also be regarded as predictions that could be used in further validation experiments. It is important to recognize, however, that these results are dependent on the numerical values of the reaction rate constants and those of the initial concentrations. Different numerical values for these parameters may lead to different results. This is a general problem of any mathematical modeling project (see Chen et al. 2009) for a discussion on the computational difficulties of this task. Clearly, having the model validated in a number of experimental setups helps increase the confidence in the numerical values we report.

3.1 Related models

Several mathematical models for the heat shock response, both for prokaryotes and for eukaryotes have been proposed in Peper et al. (1997), Rieger et al. (2005), El Samad et al. (2005), Srivastava et al. (2001), Lipan et al. (2007), Remondini et al. (2006). We compare in here our model with the ones in Peper et al. (1997) and Rieger et al. (2005) that seem most related to ours.

The model in Peper et al. (1997) considers, as we do, hsf and its dimerization and trimerization, heat-induced misfolding, but it also considers other components: mRNA molecules and nascent proteins chains, including their interactions with HSP. The model was tested against experimental data obtained from Reuber H35 rat hepatoma cells on the synthesis of the hsp70 family members. One of the shortcomings of the model in Peper et al. (1997) is that it does not consider the details of the hsp-regulated transcription. Instead the control is realized in the model through hsp-blocking of mRNA and through hsp:hsf bindings. Another concern has to do with the treatment of mRNA: it is not produced as a result of DNA transcription and it is not used directly in a model for protein synthesis, the crucial feedback regulatory motif in our model. Instead, mRNA is used in a hypothetical reaction of binding to misfolded proteins. Such a reaction leaves only part of mRNA molecules as "healthy" and their proportion is then used to model the slowing reaction rate of hsp binding to nascent protein chains (Many of these steps lack experimental support.). The same effect can, however, be obtained, as suggested in our model, based on the observation that hsp molecules are competed on, according to the mass-action principle, both by misfolded proteins (present on a massive scale under stress), and by nascent proteins chains.

The model in Rieger et al. (2005) examines the eukaryotic heat shock response based on hsp, hsf and hse, as we have, but also includes hsp mRNA molecules, a stimulus signal, and a stress kinase. The hsp synthesis is controlled through hsp-regulated DNA transcription, through hsp:hsf binding, but also through the fact that the stability of hsp mRNA molecules is increased due to stress. Moreover, the model considers the activation

of hsf molecules when bounded to hse, mediated by the stress kinase. In turn, the stress kinase is activated by the stimulus signal. On the other hand, dimerization and trimerization of hsf molecules is not considered, and neither is the degradation of hsp.

The model is tested against experimental data from HeLa cells (Kline and Morimoto 1997). The main difference with respect to our model is the fact that the heat shock is modeled in an abstract way through the stimulus signal and the stress kinase, rather than mechanistically through mfp as the initiating signal, as we do.

A recent paper (Lipan et al. 2007), takes a completely different modeling approach. Starting from available experimental data on the response of Chinese hamster ovary cells to heat shock, rather than a set of reactions, they develop a stochastic theoretical model accounting for the observed mean response. Interestingly, they rediscover in this way the hsf-regulated transactivation of hsp-encoding genes.

In another recent paper (Remondini et al. 2006), the molecular model (summarized from Morimoto 1998) includes several of the reactions in our model. Importantly, they do not consider the heat-induced protein misfolding. Also, in the associated mathematical model, only a part of the molecular model is analyzed.

A molecular model that is similar to the one we consider in this paper has been recently presented in Szymańska and Zylicz (2009). Some of the molecular details of the model in Szymańska and Zylicz (2009) are, however, different and in fact, their model includes reactions (such as the concomitant binding of three different molecules) whose kinetics are highly unfavorable. The major differences, however, are in the numerical evaluations of the model. While the authors of Szymańska and Zylicz (2009) have an ad-hoc choice of parameter values, the bulk of our work is in extensive parameter estimation and numerical validation of the model, based both on literature data, as well as on novel experiments.

3.2 Extensions

The current model can be extended to include several other aspects of the heat shock response. For example, one may include in the model the heat-induced misfolding and chaperon-assisted refolding of both hsp and hsf. Indeed, since both hsp and hsf are proteins, they are exposed to heat-induced misfolding. This extension includes in the model a most attractive feature of living cells: the repair mechanism is subject to failure, but it has capabilities to repair itself. In terms of the molecular model, the model extension consists of adding 6 reactions:

Misfolding: $\mathsf{hsp} \rightarrow \mathsf{mhsp}$

$\mathsf{hsf} \rightarrow \mathsf{mhsf}$

Sequestration: $\mathsf{hsp} + \mathsf{mhsp} \leftrightarrow \mathsf{hsp{:}\ mhsp}$

$\mathsf{hsp} + \mathsf{mhsf} \leftrightarrow \mathsf{hsp{:}\ mhsf}$

Refolding: $\mathsf{hsp{:}\ mhsp} \rightarrow 2\mathsf{hsp}$

$\mathsf{hsp{:}\ mhsf} \rightarrow \mathsf{hsp} + \mathsf{hsf}$

One way to include this model extension in the mathematical model is to assign each reaction a new kinetic parameter and measure or estimate their numerical values in such a way that the fit and the model validation with respect to experimental data remain excellent. Another way, that we adopted, is to assume a principle of uniform biochemistry: every two similar reactions in the model should be driven by the same kinetic constants. We observe that each of the reactions in the model extension above has a correspondent in

the basic model: the misfolding reactions are similar to reaction (10), the sequestration reactions are similar to (11) and the refolding reactions are similar to (12). Therefore, we can use for the model extension similar kinetics as in the basic model: ϕ_T as reaction rate coefficient for the misfolding reactions, k_{11}^+, k_{11}^- for the sequestration reactions, and k_{12} for the refolding reactions (with the same numerical values as for the basic model). Remarkably, the fit and the validation of the extended model remains essentially unchanged. For details we refer to Petre et al. (2009).

Including the phosphorylation of hsf and its role on the hsf activity is attractive, but it appears to be very challenging. The difficulty is in distinguishing all phosphorylation states of all known phosphorylation sites (currently at least 14 of them, see Voellmy and Boellmann 2007; Holmberg et al. 2002) of hsf. This leads to an exponential increase in the number of variables in our model. To start with, we have considered only one phosphorylation site for each hsf. We also asked in the extended model that an hsf trimer is only able to promote gene transcription if it has at least two of its three sites phosphorylated. The extended model includes all possible phosphorylation states of hsf, hsf_2, hsf_3, hsp:hsf, as well as protein kinases and phosphatases (which may be misfolded/refolded). The new model consists of 61 reactions and 26 reactants (I.Petre et al, unpublished data). We succeeded fitting the model to the data on DNA binding from Kline and Morimoto (1997) in such a way that the rate constants of the reactions of the basic model remain unchanged. When considering also the phosphorylation data of Kline and Morimoto (1997), the combined fit was very poor. This may indicate that the rate constants of the basic model should be re-estimated in this case, leading to a very challenging computational task. This difficulty points also to an intrinsic problem of modeling with differential equations: they are describing explicitly all variables in the model, even when many of them are essentially just duplicates of each other. A novel mathematical modeling methodology able to describe models in terms of various independent components and the communication between them (such as done in concurrency in Computer Science), may be more suitable in such setups.

3.3 Parameter scanning as a local optimization method

A major difficulty we have encountered when performing parameter estimation was to fit the time-dependent behavior of the model with respect to experimental data, while making sure that the initial values are an approximation of a steady state of the model. Indeed, the steady state of the model is a function of the parameters (and of other variables, such as total mass of various species). Once a good fit with respect to experimental data was found, our approach was to replace the initial values with the steady state of the obtained model at 37°C and hope that the model fit at 42°C is not destroyed. This is the main reason why parameter estimation was the most time-consuming part of the work.

The parameter scanning method that we have used when analyzing our model could in fact be used as a local optimization method that takes into consideration simultaneously the steady state condition and the stress-induced response of the model. The idea is that for a model that is continuous in all of the parameters (as ours is), the procedure identifies a region in the multi-dimensional parameter space where a local minimum of the score function is found. Iterating this procedure yields a realization of a local minimum of the score function, while the initial state of the model is a steady state for a temperature of 37°C.

3.4 Applicability

Mathematical modeling of biological processes may allow reasoning about uncertain or incomplete subparts of the process. For example, when constructing our molecular model, see Table 1, all reactions were considered reversible, unless they were definitely known to be unidirectional. E.g., we decided to include also reaction $\mathsf{hsf}_3 \rightarrow \mathsf{hsf} + \mathsf{hsf}_2$, although arguments based on the stability of trimers and the transient nature of dimers could be used against it. The corresponding mathematical model and its fitting help handle such incomplete information. It turns out that our model fit gives a very low rate constant for that reaction, suggesting that the reaction could be omitted altogether from the model. Arguments based on sensitivity analysis help identify more reactions that can be eliminated from the model without affecting its time evolution.

The heat shock response was amongst the primordial gene networks given the fluctuating environment and the necessity to establish proteostasis networks. The minimal mathematical model we proposed in this paper, based on stress-induced activation and feedback regulation only, may be useful also for the understanding of other forms of stress signalling or gene expression. The numerical techniques that we have used in this paper for identifying the essential components of the regulatory network may also be applicable in other mathematical modeling projects.

4 Materials and methods

4.1 Construct information

To make the hsp70promoter700-yfp construct, the CMV promoter was removed from pEYFP-N1 (Clontech) by inserting an XhoI site before the start of the CMV promoter by site directed mutagenesis using the primer 5′-TCTGTGGATAAGATCTCGAGCG CCATGCAT-3′ and its complement. The CMV promoter was deleted by digesting with XhoI, which cleave the new plasmid both in front of the CMV sequence and after this sequence in the MCS. The cleaved fragments were separated by electrophoresis, and the 4.1 kb fragment lacking the CMV promoter sequence was isolated and ligated to form pEYFPΔCMV. To add the hsp70 promoter in front of yfp, the 712 bp fragment of the hsp70 promoter was digested from pGL-712-hsp70 (a kind gift from A. Stanhill and D. Engelberg, Jerusalem, Israel) using XhoI and HindIII, and subcloned into the pEYFPΔCMV plasmid.

4.2 Cell culture and heat shock experiments

K562 cells were maintained in RPMI-1640 medium supplemented with 10% fetal calf serum, 2 mM L-glutamine, penicillin and streptomycin at 37°C in a 5% CO2 humidified atmosphere. 5.0×10^6 K562 cells were transfected with hsp70promoter700-yfp plasmid by electroporation (250 V per 975 μF; GenePulser II electroporator, BioRad laboratories). hsp70promoter700-yfp stable cell pools were selected with geneticin. For heat shock treatments, 0.5×10^6 ml^{-1} hsp70promoter700-yfp stably expressing K562 cells were transferred to RPMI-1640 medium with supplements pre-warmed to 42°C. Heat shock was induced at 42°C in a 5% CO2 humidified atmosphere for the following time points prior to sampling: 36, 33, 30, 27, 24, 21, 18, 12, 10, 8, 6, 4, 2, 1, and 0 h (control). Cells were allowed to recover post-heat shock for 2 h at 37°C. Fluorescence intensity of yfp was

measured by flow cytometry with FACScan (Becton Dickinson). Samples from heat-shocked cells were lysed and separated by SDS-PAGE and analyzed by western bloting.

Acknowledgments This work has been partially supported by the following grants from Academy of Finland: project 108421 (IP), project 203667 (A.Mizera), the Center of Excellence on Formal Methods in Programming (R-J.B.).

References

Abravaya K, Philips B, Morimoto RI (1991) 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 Dev 5(11):2117–2127

Abravaya K, Myers M, Murphy S, Morimoto RI (1992) Human heat shock protein HSP70 interacts with HSF, the transcription factor that regulates heat shock gene expression. Genes Dev 6:1153–1164

Balch WE, Morimoto RI, Dillin A, Kelly JW (2008) Adapting proteostasis for disease intervention. Science 319:916–919

Ballew RM, Sabelko J, Gruebele M (1996) Direct observation of fast protein folding: the initial collapse of apomyoglobin. Proc Natl Acad Sci USA 93:5759-64

Chen Y, Voegli TS, Liu PP, Noble EG, Currie RW (2007) Heat shock paradox and a new role of heat shock proteins and their receptors as anti-inflammation targets. Inflamm Allergy Drug Targets 6(2):91–100

Chen WW, Schorberl B, Jasper PJ, Niepel M, Nielsen UB, Lauffenburger DA, Sorger PK (2009) Input–output behavior of ErbB signaling pathways as revealed by a mass action model trained against dynamic data. Mol Syst Biol 5:1–19

Ciocca DR, Calderwood SK (2005) Heat shock proteins in cancer: diagnostic, prognostic, predictive, and treatment implications. Cell Stress Chaperones 10(2):86–103

Donati YRA, Slosman DO, Polla BS (1990) Oxidative injury and the heat shock response. Biochem Pharmacol 40:2571–2577

El Samad H, Kurata H, Doyle JC, Gross CA, Khammash M (2005) Surviving heat shock: control strategies for robustness and performance. Proc Natl Acad Sci USA 102(8):2736–2741

Guldberg CM, Waage P (1864) Studies concerning affinity. C. M. Forhandlinger: Videnskabs-Selskabet i Christiana 35

Guldberg CM, Waage P (1879) Concerning chemical affinity. Erdmann's Journal fr Practische Chemie 127:69–114

Helton JC, Davis FJ (2002) Illustration of sampling-based methods for uncertainty and sensitivity analysis. Risk Anal 22(3):591–622

Helton JC, Davis FJ (2003) Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. Reliab Eng Syst Saf 81:23–69

Holmberg CI, Tran SE, Eriksson JE, Sistonen L (2002) Multisite phosphorylation provides sophisticated regulation of transcription factors. Trends Biochem Sci 27(12):619–627

Hoops S, Sahle S, Gauges R, Lee C, Pahle J, Simus N, Singhal M, Xu L, Mendes P, Kummer U (2006) COPASI—a COmplex PAthway SImulator. Bioinformatics 22:3067–3074

Jones CM, Henry ER, Hu Y, Chan C, Luck SD, Bhuyan A, Roder H, Hofrichter J, Eaton WA, et al. (1993) Fast events in protein folding initiated by nanosecond laser photolysis. Proc Natl Acad Sci USA 90:11860–64

Kline MP, Morimoto RI (1997) Repression of the heat shock factor 1 transcriptional activation domain is modulated by constitutive phosphorylation. Mol Cell Biol 17(4):2107–2115

Lepock JR, Frey HE, Ritchie KP (1993) Protein denaturation in intact hepatocytes and isolated cellular organelles during heat shock. J Cell Biol 122(6):1267–1276

Lipan O, Navenot J-M, Wang Z, Huang L, Peiper SC (2007) Heat shock response in CHO mammalian cells is controlled by a nonlinear stochastic process. PLoS Comput Biol 3(10):1859–1870

McKay MD, Beckman RJ, Conover WJ (1979) A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21(2):239-245

Morimoto RI (1998) Regulation of the heat shock transcriptional response: cross talk between a family of heat shock factors, molecular chaperones, and negative regulators. Genes Dev 12:3788–3796

Morimoto RI (2008) Proteotoxic stress and inducible chaperone networks in neurodegenerative disease and aging. Genes Dev 22:1427–1438

Morimoto RI, Jurivich DA, Kroger PE, Mathur SK, Murphy SP, et al (1994) Regulation of heat shock gene transcription by a family of heat shock factors. In: Morimoto RI, Tissières A, Georgopoulos C (eds) The biology of the heat shock proteins and molecular chaperones. Cold Spring Harbor Laboratory, New York, pp. 417–455

Oberguggenberger M, King J, Schmelzer B (2009) Classical and imprecise probability methods for sensitivity analysis in engineering: a case study. Int J Approx Reason 50:680-693

Parsell DA, Lindquist S (1993) The function of heat-shock proteins in stress tolerance: degradation and reactivation of damaged proteins. Ann Rev Genetics 27:437–496

Peper A, Grimbergen CA, Spaan JAE, Souren JEM, van Wijk R (1997) A mathematical model of the hsp70 regulation in the cell. Int J Hyperth 14(1):97–124

Petre I, Mizera A, Hyder CL, Mikhailov A, Eriksson JE, Sistonen L, Back R-J (2009) A new mathematical model for the heat shock response. In: Condon A, Harel D, Kok J, Salomaa A (eds) Algorithmic bioprocesses. Springer, New York, pp. 411–428

Pockley AG (2003) Heat shock proteins as regulators of the immune response. The Lancet 362 (9382):469–476

Powers MV, Workman P (2007) Inhibitors of the heat shock response: biology and pharmacology. FEBS Lett 581(19):3758–3769

Remondini D, Bernardini C, Forni M, Bersani F, Castellani GC, Bacci ML (2006) Induced metastable memory in heat shock response. J Biol Phys 32:49–59

Rieger TR, Morimoto RI, Hatzimanikatis V (2005) Mathematical modeling of the eukaryotic heat shock response: dynamics of the Hsp70 promoter. Biophys J 88:1646–1658

Shi Y, Mosser D, Morimoto RI (1998) Molecular chaperones as HSF1 specific transcriptional repressors. Genes Dev 12:654–666

Srivastava R, Peterson MS, Bentley WE (2001) Stochastic kinetic analysis of the *Escherichia coli* stress circuit using $\sigma^3$2-trageted antisense. Biotechnol Bioeng 75(1):120–129

Szymańska Z, Zylicz M (2009) Mathematical modeling of heat shock protein synthesis in response to temperature change. J Theoret Biol 259:562–569

Turanyi T (1990) Sensitivity analysis of complex kinetic systems—tools and applications. J Math Chem 5(3):203–248

Vastag B (2006) HSP-90 inhibitors promise to complement cancer therapies. Nat Biotechnol 24(11):1307

Voellmy R (1994) Transduction of the stress signal and mechanisms of transcriptional regulation of heat shock/stress protein gene expression in higher eukaryotes. Crit Rev Eukaryot Gene Expr 4:357–401

Voellmy R, Boellmann F (2007) Chaperone regulation of the heat shock protein response. Adv Exp Med Biol 594:89–99

Paper II

A new mathematical model for the heat shock response

Ion Petre, Andrzej Mizera, Claire L. Hyder, Andrey Mikhailov, John E. Eriksson, Lea Sistonen, and Ralph-Johan Back

Originally published in Anne Condon, David Harel, Joost N. Kok, Arto Salomaa, and Erik Winfree (Eds.), *Algorithmic Bioprocesses*, Natural Computing Series, pages 411-425. Springer, Dordrecht Heidelberg London New York, 2009.

A new mathematical model for the heat shock response

Ion Petre[1,2], Andrzej Mizera[2], Claire L. Hyder[3,4], Andrey Mikhailov[3,4], John E. Eriksson[3,4], Lea Sistonen[1,3,4], and Ralph-Johan Back[1,2]

[1] Academy of Finland
[2] Department of Information Technologies, Åbo Akademi University, Turku 20520, Finland ipetre, amizera, backrj@abo.fi
[3] Turku Centre for Biotechnology chyder, andrey.mikhailov, john.eriksson, lea.sistonen@btk.fi
[4] Department of Biochemistry, Åbo Akademi University, Turku 20520, Finland

Summary. We present in this paper a new molecular model for the gene regulatory network responsible for the eukaryotic heat shock response. Our model includes the temperature-induced protein misfolding, the chaperone activity of the heat shock proteins and the backregulation of their gene transcription. We then build a mathematical model for it, based on ordinary differential equations. Finally, we discuss the parameter fit and the implications of the sensitivity analysis for our model.

Key words: Heat shock response, heat shock protein, heat shock factor, mathematical model, differential equations, model fit, sensitivity analysis

1 Introduction

One of the most impressive algorithmic-like bioprocesses in living cells, crucial for the very survival of cells is the *heat shock response*: the reaction of the cell to elevated temperatures. One of the effects of raised temperature in the environment is that proteins get misfolded, with a rate that is exponentially dependent on the temperature. In turn, as an effect of their hydrophobic core being exposed, misfolded proteins tend to form bigger and bigger aggregates, with disastrous consequences for the cell, see [1]. To survive, the cell needs to increase quickly the level of chaperons (proteins that are assisting in the folding or refolding of other proteins). Once the heat shock is removed, the cell eventually re-establishes the original level of chaperons, see [10, 18, 22].

The heat shock response has been subject of intense research in the last few years, for at least three reasons. First, it is a well-conserved mechanism across all eukaryotes, while bacteria exhibit only a slightly different response, see [5, 12, 23]. As such, it is a good candidate for studying the engineering principle of gene regulatory networks, see [4, 5, 12, 25]. Second, it is a tempting mechanism

to model mathematically, since it involves only very few reactants, at least in a simplified presentation, see [18, 19, 22]. Third, the heat shock proteins (the main chaperons involved in the eukaryotic heat shock response) play a central role in a large number of regulatory and of inflammatory processes, as well as in signaling, see [9, 20]. Moreover, they contribute to the resilience of cancer cells, which makes them attractive as targets for cancer treatment, see [3, 15, 16, 27].

We focus in this paper on a new molecular model for the heat shock response, proposed in [19]. We consider here a slight extension of the model in [19] where, among others, the chaperons are also subject to misfolding. After introducing the molecular model in Section 2, we build a mathematical model in Section 3, including the fitting of the model with respect to experimental data. We discuss in Section 4 the results of the sensitivity analysis of the model, including its biological implications.

2 A new molecular model for the eukaryotic heat shock response

The heat shock proteins (hsp) play the key role in the heat shock response. They act as chaperons, helping misfolded proteins (mfp) to refold. The response is controlled in our model through the regulation of the transactivation of the hsp-encoding genes. The transcription of the gene is promoted by some proteins called heat shock factors (hsf) that trimerize and then bind to a specific DNA sequence called heat shock element (hse), upstream of the hsp-encoding gene. Once the hsf trimer is bound to the heat shock element, the gene is transactivated and the synthesis of hsp is thus switched on (for the sake of simplicity, the role of RNA is ignored in our model). Once the level of hsp is high enough, the cell has an ingenious mechanism to switch off the hsp synthesis. For this, hsp bind to free hsf, as well as break the hsf trimers (including those bound to hse, promoting the gene activation), thus effectively halting the hsp synthesis.

Under elevated temperatures, some of the proteins (prot) in the cell get misfolded. The heat shock response is then quickly switched on simply because the heat shock proteins become more and more active in the refolding process, thus leaving the heat shock factors free and able to promote the synthesis of more heat shock proteins. Note that several types of heat shock proteins exist in an eukaryotic cell. We treat them all uniformly in our model, with hsp90 as common denominator. The same comment applies also to the heat shock factors.

Our molecular model for the eukaryotic heat shock response consists of the following molecular reactions:

1. $2\ \mathsf{hsf} \leftrightarrows \mathsf{hsf}_2$
2. $\mathsf{hsf} + \mathsf{hsf}_2 \leftrightarrows \mathsf{hsf}_3$

3. $\mathsf{hsf}_3 + \mathsf{hse} \leftrightarrows \mathsf{hsf}_3\colon \mathsf{hse}$
4. $\mathsf{hsf}_3\colon \mathsf{hse} \to \mathsf{hsf}_3\colon \mathsf{hse} + \mathsf{mhsp}$
5. $\mathsf{hsp} + \mathsf{hsf} \leftrightarrows \mathsf{hsp}\colon \mathsf{hsf}$
6. $\mathsf{hsp} + \mathsf{hsf}_2 \to \mathsf{hsp}\colon \mathsf{hsf} + \mathsf{hsf}$
7. $\mathsf{hsp} + \mathsf{hsf}_3 \to \mathsf{hsp}\colon \mathsf{hsf} + 2\ \mathsf{hsf}$
8. $\mathsf{hsp} + \mathsf{hsf}_3\colon \mathsf{hse} \to \mathsf{hsp}\colon \mathsf{hsf} + 2\ \mathsf{hsf} + \mathsf{hse}$
9. $\mathsf{hsp} \to \emptyset$
10. $\mathsf{prot} \to \mathsf{mfp}$
11. $\mathsf{hsp} + \mathsf{mfp} \leftrightarrows \mathsf{hsp}\colon \mathsf{mfp}$
12. $\mathsf{hsp}\colon \mathsf{mfp} \to \mathsf{hsp} + \mathsf{prot}$
13. $\mathsf{hsf} \to \mathsf{mhsf}$
14. $\mathsf{hsp} \to \mathsf{mhsp}$
15. $\mathsf{hsp} + \mathsf{mhsf} \leftrightarrows \mathsf{hsp}\colon \mathsf{mhsf}$
16. $\mathsf{hsp}\colon \mathsf{mhsf} \to \mathsf{hsp} + \mathsf{hsf}$
17. $\mathsf{hsp} + \mathsf{mhsp} \leftrightarrows \mathsf{hsp}\colon \mathsf{mhsp}$
18. $\mathsf{hsp}\colon \mathsf{mhsp} \to 2\ \mathsf{hsp}$

It is important to note that the main addition we consider here with respect to the model in [19] is to include the misfolding of hsp and hsf. This is, in principle, no minor extension since in the current model the repairing mechanism is subject to failure, but it is capable to fix itself.

Several criteria were followed when introducing this molecular model:

(i) as few reactions and reactants as possible;
(ii) include the temperature-induced protein misfolding;
(iii) include hsf in all its three forms: monomers, dimers, and trimers;
(iv) include the hsp-backregulation of the transactivation of the hsp-encoding gene;
(v) include the chaperon activity of hsp;
(vi) include only well-documented, textbook-like reactions and reactants.

For the sake of keeping the model as simple as possible, we are ignoring a number of details. E.g., note that there is no notion of locality in our model: we make no distinction between the place where gene transcription takes place (inside nucleus) and the place where protein synthesis takes place (outside nucleus). Note also that protein synthesis and gene transcription are greatly simplified in reaction 4: we only indicate that once the gene is transactivated, protein synthesis is also switched on. On the other hand, reaction 4 is faithful to the biological reality, see [1] in indicating that newly synthesized proteins often need chaperons to form their native fold.

As far as protein degradation is concerned, we only consider it in the model for hsp. If we considered it also for hsf and prot, then we should also consider the compensating mechanism of protein synthesis, including its control. For the sake of simplicity and also based on experimental evidence that the total amount of hsf and of prot is somewhat constant, we ignore the details of synthesis and degradation for hsf and prot.

3 The mathematical model

We build in this section a mathematical model associated to the molecular model 1–18. Our mathematical model is in terms of coupled ordinary differential equations and its formulation is based on the principle of mass-action.

3.1 The principle of mass-action

The mass-action law is widely used in formulating mathematical models in physics, chemistry, and engineering. Introduced in [6, 7], it can be briefly summarized as follows: *the rate of each reaction is proportional to the concentration of reactants*. In turn, the rate of each reaction gives the rate of consuming the reactants and the rate of producing the products. E.g., for a reaction

$$R_1 : \ A + B \to C,$$

the rate according to the principle of mass action is $f_1(t) = kA(t)B(t)$, where $k \geq 0$ is a constant and $A(t)$, $B(t)$ are functions of time giving the level of the reactants A and B, respectively. Consequently, the rate of consuming A and B, and the rate of producing C is expressed by the following differential equations:

$$\frac{dA}{dt} = \frac{dB}{dt} = -k\,A(t)\,B(t), \qquad \frac{dC}{dt} = k\,A(t)\,B(t).$$

For a reversible reaction

$$R_2 : \ A + B \leftrightarrows C,$$

the rate is $f_2(t) = k_1\,A(t)\,B(t) \ - \ k_2\,C(t)$, for some constants k_1, $k_2 \geq 0$. The differential equations are written in a similar way:

$$\frac{dA}{dt} = \frac{dB}{dt} = -f_2(t), \qquad \frac{dC}{dt} = f_2(t). \tag{*}$$

For a set of coupled reactions, the differential equations capture the combined rate of consuming and producing each reactant as an effect of all reactions taking place simultaneously. E.g., for reactions

$$R_3 : \ A + B \leftrightarrows C, \qquad R_4 : \ B + C \leftrightarrows A, \qquad R_5 : \ A + C \leftrightarrows B,$$

the associated system of differential equations is

$$\begin{aligned} dA/dt &= -f_3(t) + f_4(t) - f_5(t),\\ dB/dt &= -f_3(t) - f_4(t) + f_5(t),\\ dC/dt &= f_3(t) - f_4(t) - f_5(t), \end{aligned}$$

where $f_i(t)$ is the rate of reaction R_i, for all $3 \leq i \leq 5$, formulated according to the principle of mass action.

We recall that for a system of differential equations

$$\begin{aligned} \frac{dX_1}{dt} &= f_1(X_1, \ldots, X_n), \\ &\cdots \\ \frac{dX_n}{dt} &= f_n(X_1, \ldots, X_n), \end{aligned}$$

we say that $(x_1, x_2, \ldots, x_n)$ is a *steady states* (also called *equilibrium points*) if it is a solution of the algebraic system of equations $f_i(X_1, \ldots, X_n) = 0$, for all $1 \leq i \leq n$, see [24, 28]. Steady states are particularly interesting because they characterize situations where although reactions may have non-zero rates, their combined effect is zero. In other words, the concentration of all reactants and of all products are constant.

We refer to [11, 17, 29] for more details on the principle of mass action and its formulation based on ordinary differential equations.

3.2 Our mathematical model

Let $\mathbb{R}_+$ be the set of all positive real numbers and $\mathbb{R}_+^n$ the set of all n-tuples of positive real numbers, for $n \geq 2$. We denote each reactant and bond between them in the molecular model 1–18 according to the convention in Table 1. We also denote by $\kappa \in \mathbb{R}_+^{17}$ the vector with all reaction rate constants as its components, see Table 2: $\kappa = (k_1^+, k_1^-, k_2^+, k_2^-, k_3^+, k_3^-, k_4, k_5^+, k_5^-, k_6, k_7, k_8, k_9, k_{11}^+, k_{11}^-, k_{12}, k_{13}^+, k_{13}^-, k_{14}, k_{15}^+, k_{15}^-, k_{16})$.

Table 1. The list of variables in the mathematical model, their initial values, and their values in one of the steady states of the system, for $T = 42$. Note that the initial values give one of the steady states of the system for $T = 37$.

Metabolite	Variable	Initial value	A steady state (T=42)
hsf	X_1	0.669	0.669
hsf_2	X_2	$8.73 \cdot 10^{-4}$	$8.73 \cdot 10^{-4}$
hsf_3	X_3	$1.23 \cdot 10^{-4}$	$1.23 \cdot 10^{-4}$
hsf_3: hse	X_4	2.956	2.956
mhsf	X_5	$3.01 \cdot 10^{-6}$	$2.69 \cdot 10^{-5}$
hse	X_6	29.733	29.733
hsp	X_7	766.875	766.875
mhsp	X_8	$3.45 \cdot 10^{-3}$	$4.35 \cdot 10^{-2}$
hsp: hsf	X_9	1403.13	1403.13
hsp: mhsf	X_{10}	$4.17 \cdot 10^{-7}$	$3.72 \cdot 10^{-6}$
hsp: mhsp	X_{11}	$4.78 \cdot 10^{-4}$	$6.03 \cdot 10^{-3}$
hsp: mfp	X_{12}	71.647	640.471
prot	X_{13}	$1.14 \cdot 10^8$	$1.14 \cdot 10^8$
mfp	X_{14}	517.352	4624.72

Table 2. The numerical values for the fitted model.

Kinetic constant	Reaction	Numerical value
k_1^+	(1), forward	3.49091
k_1^-	(1), backward	0.189539
k_2^+	(2), forward	1.06518
k_2^-	(2), backward	$1 \cdot 10^{-9}$
k_3^+	(3), forward	0.169044
k_3^-	(3), backward	$1.21209 \cdot 10^{-6}$
k_4	(4)	0.00830045
k_5^+	(5), forward	9.73665
k_5^-	(5), backward	3.56223
k_6	(6)	2.33366
k_7	(7)	$4.30924 \cdot 10^{-5}$
k_8	(8)	$2.72689 \cdot 10^{-7}$
k_9	(9)	$3.2 \cdot 10^{-5}$
k_{11}^+	(11), forward	0.00331898
k_{11}^-	(11), backward	4.43952
k_{12}	(12)	13.9392
k_{13}^+	(15), forward	0.00331898
k_{13}^-	(15), backward	4.43952
k_{14}	(16)	13.9392
k_{15}^+	(17), forward	0.00331898
k_{15}^-	(17), backward	4.43952
k_{16}	(18)	13.9392

The mass action-based formulation of the associated mathematical model in terms of differential equations is straightforward, leading to the following system of equations:

$$dX_1/dt = f_1(X_1, X_2, \ldots, X_{14}, \kappa) \quad (1)$$
$$dX_2/dt = f_2(X_1, X_2, \ldots, X_{14}, \kappa) \quad (2)$$
$$dX_3/dt = f_3(X_1, X_2, \ldots, X_{14}, \kappa) \quad (3)$$
$$dX_4/dt = f_4(X_1, X_2, \ldots, X_{14}, \kappa) \quad (4)$$
$$dX_5/dt = f_5(X_1, X_2, \ldots, X_{14}, \kappa) \quad (5)$$
$$dX_6/dt = f_6(X_1, X_2, \ldots, X_{14}, \kappa) \quad (6)$$
$$dX_7/dt = f_7(X_1, X_2, \ldots, X_{14}, \kappa) \quad (7)$$
$$dX_8/dt = f_8(X_1, X_2, \ldots, X_{14}, \kappa) \quad (8)$$
$$dX_9/dt = f_9(X_1, X_2, \ldots, X_{14}, \kappa) \quad (9)$$
$$dX_{10}/dt = f_{10}(X_1, X_2, \ldots, X_{14}, \kappa) \quad (10)$$
$$dX_{11}/dt = f_{11}(X_1, X_2, \ldots, X_{14}, \kappa) \quad (11)$$
$$dX_{12}/dt = f_{12}(X_1, X_2, \ldots, X_{14}, \kappa) \quad (12)$$

$$dX_{13}/dt = f_{13}(X_1, X_2, \ldots, X_{14}, \kappa) \tag{13}$$
$$dX_{14}/dt = f_{14}(X_1, X_2, \ldots, X_{14}, \kappa) \tag{14}$$

where

$$\begin{aligned}
f_1 &= -k_2^+ X_1 X_2 + k_2^- X_3 - k_5^+ X_1 X_7 + k_5^- X_9 + 2\,k_8 X_4 X_7 + k_6 X_2 X_7 \\
&\quad -\varphi(T) X_1 + k_{14} X_{10} + 2\,k_7 X_3 X_7 - 2\,k_1^+ X_1^2 + 2\,k_1^- X_2 \\
f_2 &= -k_2^+ X_1 X_2 + k_2^+ X_3 - k_6 X_2 X_7 + k_1^+ X_1^2 - k_1^- X_2 \\
f_3 &= -k_3^+ X_3 X_6 + k_2^+ X_1 X_2 - k_2^- X_3 + k_3^- X_4 - k_7 X_3 X_7 \\
f_4 &= k_3^+ X_3 X_6 - k_3^- X_4 - k_8 X_4 X_7 \\
f_5 &= \varphi(T) X_1 - k_{13}^+ X_5 X_7 + k_{13}^- X_{10} \\
f_6 &= -k_3^+ X_3 X_6 + k_3^- X_4 + k_8 X_4 X_7 \\
f_7 &= -k_5^+ X_1 X_7 + k_5^- X_9 - k_{11}^+ X_7 X_{14} + k_{11}^- X_{12} - k_8 X_4 X_7 - k_6 X_2 X_7 \\
&\quad -k_{13}^+ X_5 X_7 + (k_{13}^- + k_{14}) X_{10} - (\varphi(T) + k_9) X_7 - k_{15}^+ X_7 X_8 \\
&\quad -k_7 X_3 X_7 + (k_{15}^- + 2\,k_{16}) X_{11} + k_{12} X_{12} \\
f_8 &= k_4 X_4 + \varphi(T) X_7 - k_{15}^+ X_7 X_8 + k_{15}^- X_{11} \\
f_9 &= k_5^+ X_1 X_7 - k_5^- X_9 + k_8 X_4 X_7 + k_6 X_2 X_7 + k_7 X_3 X_7 \\
f_{10} &= k_{13}^+ X_5 X_7 - (k_{13}^- + k_{14}) X_{10} \\
f_{11} &= k_{15}^+ X_7 X_8 - (k_{15}^- + k_{16}) X_{11} \\
f_{12} &= k_{11}^+ X_7 X_{14} - (k_{11}^- + k_{12}) X_{12} \\
f_{13} &= k_{12} X_{12} - \varphi(T) X_{13} \\
f_{14} &= -k_{11}^+ X_7 X_{14} + k_{11}^- X_{12} + \varphi(T) X_{13}
\end{aligned}$$

The rate of protein misfolding $\varphi(T)$ with respect to temperature T has been investigated experimentally in [13, 14], and a mathematical expression for it has been proposed in [18]. We have adapted the formula in [18] to obtain the following misfolding rate per second:

$$\varphi(T) = (1 - \frac{0.4}{e^{T-37}}) \cdot 0.8401033733 \cdot 10^{-6} \cdot 1.4^{T-37}\ s^{-1},$$

where T is the temperature of the environment in Celsius degrees, with the formula being valid for $37 \leq T \leq 45$.

The following result gives three mass-conservation relations for our model.

Theorem 1. *There exists K_1, K_2, $K_3 \geq 0$ such that:*

(i) $X_1(t) + 2\,X_2(t) + 3\,X_3(t) + 3\,X_4(t) + X_5(t) + X_9(t) = K_1$,
(ii) $X_4(t) + X_6(t) = K_2$,
(iii) $X_{13}(t) + X_{14}(t) + X_{12}(t) = K_3$,

for all $t \geq 0$.

Proof. We only prove here part (ii), as the others may be proved analogously. For this, note that from equations (4) and (6), it follows that

$$\frac{d(X_4 + X_6)}{dt} = (f_4 + f_6)(X_1, \ldots, X_{14}, \kappa, t) = 0,$$

i.e., $(X_4 + X_6)(t)$ is a constant function. ■

The steady states of the model (1)-(14) satisfy the following algebraic relations, where x_i is the numerical value of X_i in the steady state, for all $1 \leq i \leq 14$.

$$\begin{aligned}
0 &= -k_2^+ x_1 x_2 + k_2^- x_3 - k_5^+ x_1 x_7 + k_5^- x_9 + 2 k_8 x_4 x_7 + k_6 x_2 x_7 \\
&\quad -\varphi(T) x_1 + k_{14} x_{10} + 2 k_7 x_3 x_7 - 2 k_1^+ x_1^2 + 2 k_1^- x_2 && (15)\\
0 &= -k_2^+ x_1 x_2 + k_2^+ x_3 - k_6 x_2 x_7 + k_1^+ x_1^2 - k_1^- x_2 && (16)\\
0 &= -k_3^+ x_3 x_6 + k_2^+ x_1 x_2 - k_2^- x_3 + k_3^- x_4 - k_7 x_3 x_7 && (17)\\
0 &= k_3^+ x_3 x_6 - k_3^- x_4 - k_8 x_4 x_7 && (18)\\
0 &= \varphi(T) x_1 - k_{13}^+ x_5 x_7 + k_{13}^- x_{10} && (19)\\
0 &= -k_3^+ x_3 x_6 + k_3^- x_4 + k_8 x_4 x_7 && (20)\\
0 &= -k_5^+ x_1 x_7 + k_5^- x_9 - k_{11}^+ x_7 x_{14} + k_{11}^- x_{12} - k_8 x_4 x_7 - k_6 x_2 x_7 \\
&\quad -k_{13}^+ x_5 x_7 + (k_{13}^- + k_{14}) x_{10} - (\varphi(T) + k_9) x_7 - k_{15}^+ x_7 x_8 - k_7 x_3 x_7 \\
&\quad +(k_{15}^- + 2 k_{16}) x_{11} + k_{12} x_{12} && (21)\\
0 &= k_4 x_4 + \varphi(T) x_7 - k_{15}^+ x_7 x_8 + k_{15}^- x_{11} && (22)\\
0 &= k_5^+ x_1 x_7 - k_5^- x_9 + k_8 x_4 x_7 + k_6 x_2 x_7 + k_7 x_3 x_7 && (23)\\
0 &= k_{13}^+ x_5 x_7 - (k_{13}^- + k_{14}) x_{10} && (24)\\
0 &= k_{15}^+ x_7 x_8 - (k_{15}^- + k_{16}) x_{11} && (25)\\
0 &= k_{11}^+ x_7 x_{14} - (k_{11}^- + k_{12}) x_{12} && (26)\\
0 &= k_{12} x_{12} - \varphi(T) x_{13} && (27)\\
0 &= -k_{11}^+ x_7 x_{14} + k_{11}^- x_{12} + \varphi(T) x_{13} && (28)
\end{aligned}$$

It follows from Theorem 1 that only eleven of the relations above are independent. E.g., relations (15)-(17), (19), (21)-(27) are independent. The system consisting of the corresponding differential equations is called the *reduced system* of (1)-(14).

3.3 Fitting the model to experimental data

The experimental data available for the parameter fit is from [10] and reflects the level of DNA binding, i.e., variable X_4 in our model, for various time points up to 4 hours, with continuous heat shock at 42 °C. Additionally, we require that the initial value of the variables of the model is a steady state for

temperature set to 37. This is a natural condition since the model is supposed to reflect the reaction to temperatures raised above 37 °C.

Mathematically, the problem we need to solve is one of global optimization, as formulated below. For each 17-tuple κ of positive numerical values for all kinetic constants, and for each 14-tuple α of positive initial values for all variables in the model, the function $X_4(t)$ is uniquely defined for a fixed temperature T. We denote the value of this function at time point τ, with parameters κ and α by $x_4^T(\kappa, \alpha, \tau)$. Note that this property holds for all the other variables in the model and it is valid in general for any mathematical model based on ordinary differential equations (one calls such models *deterministic*). We denote the set of experimental data in [10] by

$$E_n = \{(t_i, r_i) \mid t_i, r_i > 0,\, 1 \leq i \leq N\},$$

where $N \geq 1$ is the number of observations, t_i is the time point of each observation and r_i is the value of the reading.

With this setup, we can now formulate our optimization problem as follows: find $\kappa \in \mathbb{R}_+^{17}$ and $\alpha \in \mathbb{R}_+^{14}$ such that:

(i) $f(\kappa, \alpha) = \frac{1}{N} \sum_{i=1}^{N} (x_4^{42}(\kappa, \alpha, t_i) - r_i)^2$ is minimal and
(ii) α is a steady state of the model for $T = 37$ and parameter values given by κ.

The function $f(\kappa, \alpha)$ is a cost function (in this case *least mean squares*), indicating numerically how the function $x_4^T(\kappa, \alpha, t)$, $t \geq 0$, compares with the experimental data.

Note that in our optimization problem, not all 31 variables (the components of κ and α) are independent. On one hand, we have the three algebraic relations given by Theorem 1. On the other hand, we have eleven more independent algebraic relations given by the steady state equations (15)-(17), (19), (21)-(27). Consequently, we have 17 independent variables in our optimization problem.

Given the high degree of the system (1)-(14), finding the analytical form of the minimum points of $f(\kappa, \alpha)$ is very challenging. This is a typical problem whenever the system of equations is non-linear. Adding to the difficulty of the problem is the fact that the eleven independent steady state equations cannot be solved analytically, given their high overall degree.

Since an analytical solution to the model fitting problem is often intractable, the practical approach to such problems is to give a numerical simulation of a solution. Several methods exist for this, see [2, 21]. The trade-off with all these methods is that typically they offer an estimate of a *local* optimum, with no guarantee of it being a *global* optimum.

Obtaining a numerical estimation of a local optimum for (i) is not difficult. However, such a solution may not satisfy (ii). To solve this problem, for a given local optimum $(\kappa_0, \alpha_0) \in \mathbb{R}_+^{17} \times \mathbb{R}_+^{14}$ one may numerically estimate a steady state $\alpha_1 \in \mathbb{R}_+^{14}$ for $T = 37$. Then the pair (κ_0, α_1) satisfies (ii). Unfortunately, (κ_0, α_1) may not be close to a local optimum of the cost function in (i).

Another approach is to replace the algebraic relations implicitly given by (ii) with an optimization problem similar to that in (i). Formally, we replace all algebraic relations $R_i = 0$, $1 \leq i \leq 11$, given by (ii) with the condition that

$$g(\kappa, \alpha) = \frac{1}{M} \sum_{j=1}^{M} R_i^2(\kappa, \alpha, \delta_j)$$

is minimal, where $0 < \delta_1 < \cdots < \delta_M$ are some arbitrary (but fixed) time points. Our problem thus becomes one of optimization with cost function (f, g), with respect to the order relation $(a, b) \leq (c, d)$ if and only if $a \leq c$ and $b \leq d$. The numerical values in Table 2 give one solution to this problem obtained based on Copasi [8]. The plot in Figure 1 shows the time evolution of function $X_4(t)$ up to $t = 4$ hours, with the experimental data of [10] indicated with crosses.

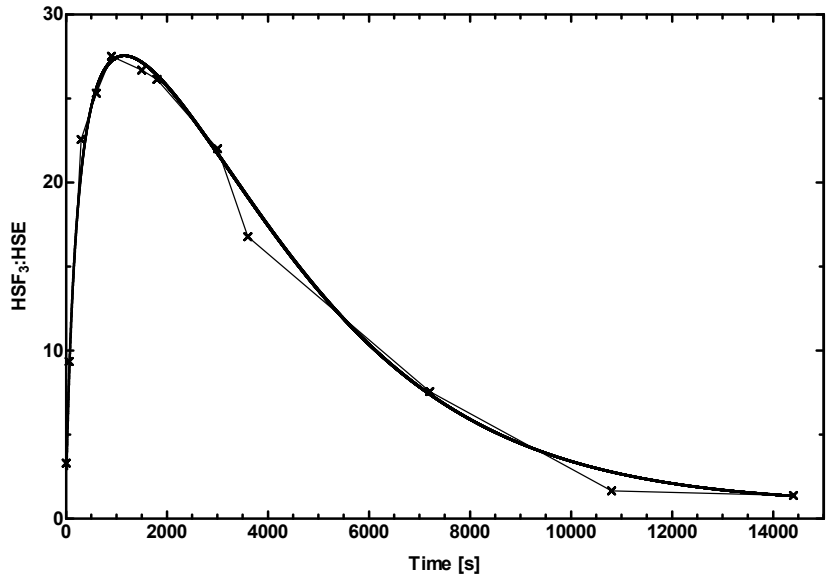


Fig. 1. The continuous line shows a numerical estimation of function $X_4(t)$, standing for DNA binding, for the initial data in Table 1 and the parameter values in Table 2. With crossed points we indicated the experimental data of [10].

The solution in Table 2 has been compared with a number of other available experimental data (such as behavior at 41 °C and at 43 °C), as well as against qualitative, non-numerical data. The results were satisfactory and better than those of previous models reported in the literature, such as [18, 22]. For details on the model validation analysis we refer to [19].

Note that the steady state of the system of differential equations (1)-(14), for the initial values in Table 1 and the parameter values in Table 2 is *asymptotically stable*. To prove it, it is enough to consider its associated *Jacobian*:

$$J(t) = \begin{pmatrix} \partial f_1/\partial X_1 & \partial f_1/\partial X_2 & \dots & \partial f_1/\partial X_{14} \\ \partial f_2/\partial X_1 & \partial f_2/\partial X_2 & \dots & \partial f_2/\partial X_{14} \\ \vdots & \vdots & & \vdots \\ \partial f_{14}/\partial X_1 & \partial f_{14}/\partial X_2 & \dots & \partial f_{14}/\partial X_{14} \end{pmatrix}$$

As it is well-known, see [28, 24], a steady state is asymptotically stable if and only if all eigenvalues of the Jacobian at the steady state have negative real parts. A numerical estimation done with *Copasi* [8] shows that the steady state for $T = 42$, see Table 1, is indeed asymptotically stable.

4 Sensitivity analysis

Sensitivity analysis is a method to estimate the changes brought into the system through small changes in the parameters of the model. In this way one may estimate both the robustness of the model against small changes in the model, as well as identify possibilities for bringing a certain desired changed in the system. E.g., one question that is often asked of a biochemical model is what changes should be done to the model so that the new steady state satisfies certain properties. In our case we are interested in changing some of the parameters of the model so that the level of mfp in the new steady state of the system is smaller than in the standard model, thus presumably making it easier for the cell to cope with the heat shock. We also analyze a scenario in which we are interested in increasing the level of mfp in the new steady state, thus increasing the chances of the cell not being able to cope with the heat shock. Such a scenario is especially meaningful in relation with cancer cells that exhibit the properties of an excited cell, with increased levels of hsp, see [3, 15, 16, 27]. In this section we follow in part a presentation of sensitivity analysis due to [26].

We consider the partial derivatives of the solution of the system with respect to the parameters of the system. These are called *first-order local concentration sensitivity coefficients*. Second- or higher-order sensitivity analysis considering the simultaneous change of two or more parameters is also possible. If we denote $X(t,\kappa) = (X_1(t,\kappa), X_2(t,\kappa), \dots, X_{14}(t,\kappa))$ the solution of the system (1)-(14) with respect to the parameter vector κ, then the concentration sensitivity coefficients are the time functions $\partial X_i/\partial \kappa_j(t)$, for all $1 \le i \le 14$, $1 \le j \le 17$. Differentiating the system (1)-(14) with respect to κ_j yields the following set of *sensitivity equations*:

$$\frac{d}{dt}\frac{\partial X}{\kappa_j} = J(t)\frac{\partial X}{\partial \kappa_j} + \frac{\partial f(t)}{\partial \kappa_j}, \qquad \text{for all } 1 \le j \le 17, \tag{29}$$

where $\partial X/\partial\kappa_j = (\partial X_1/\partial\kappa_j, \ldots, \partial X_{14}/\partial\kappa_j)$ is the component-wise vector of partial derivatives, $f = (f_1, \ldots, f_{14})$ is the model function in (1)-(14), and $J(t)$ is the corresponding Jacobian. The initial condition for the system (29) is that $\partial X/\partial\kappa_j(0) = 0$, for all $1 \leq j \leq 17$.

The solution of the system (29) can be numerically integrated, thus obtaining a numerical approximation of the time evolution of the sensitivity coefficients. Very often however, the focus is on sensitivity analysis around steady states. If the considered steady state is asymptotically stable, then one may consider the limit $\lim_{t\to\infty}(\partial X/\partial\kappa_j)(t)$, called *stationary sensitivity coefficients*. They reflect the dependency of the steady state on the parameters of the model. Mathematically, they are given by a set of algebraic equations obtained from (29) by setting $d/dt(\partial X/\kappa_j) = 0$. We then obtain the following algebraic equations:

$$\left(\frac{\partial X}{\partial\kappa_j}\right) = -J^{-1}F_j, \qquad \text{for all } 1 \leq j \leq 17, \tag{30}$$

where J is the value of the Jacobian at the steady state and F_j is the j-th column of the matrix $F = (\partial f_r/\partial\kappa_s)_{r,s}$ computed at the steady state.

When used for comparing the relative effect of a parameter change in two or more variables, the sensitivity coefficients must have the same physical dimension or be dimensionless, see [26]. Most often, one simply considers the matrix S' of (dimensionless) *normalized* (also called *scaled*) sensitivity coefficients:

$$S'_{ij} = \frac{\kappa_j}{X_i(t,\kappa)} \cdot \frac{\partial X_i(t,\kappa)}{\partial\kappa_j} = \frac{\partial \ln X_i(t,\kappa)}{\partial \ln \kappa_j}$$

Numerical estimations of the normalized sensitivity coefficients for a steady state may be obtained, e.g. with Copasi. For X_{14} (standing for the level of mfp in the model), the most significant (with the largest module) sensitivity coefficients are the following:

- $\partial\ln(X_{14})/\partial\ln(T) = 14.24$,
- $\partial\ln(X_{14})/\partial\ln(k_1^+) = -0.16$,
- $\partial\ln(X_{14})/\partial\ln(k_2^+) = -0.16$,
- $\partial\ln(X_{14})/\partial\ln(k_5^+) = 0.49$,
- $\partial\ln(X_{14})/\partial\ln(k_5^-) = -0.49$,
- $\partial\ln(X_{14})/\partial\ln(k_6) = 0.16$,
- $\partial\ln(X_{14})/\partial\ln(k_9) = 0.15$,
- $\partial\ln(X_{14})/\partial\ln(k_{11}^+) = -0.99$,
- $\partial\ln(X_{14})/\partial\ln(k_{11}^-) = 0.24$,
- $\partial\ln(X_{14})/\partial\ln(k_{12}) = -0.24$.

These coefficients being most significant is consistent with the biological intuition that the level of mfp in the model is most dependant on the temperature (parameter T), on the rate of mfp being sequestered by hsp (parameters k_{11}^+ and k_{11}^-) and the rate of protein refolding (parameter k_{12}). However, the sensitivity coefficients also reveal less intuitive, but significant dependencies such as the one on the reaction rate of hsf being sequestered by hsp (parameters k_5^+ and k_5^-), on the rate of dissipation of hsf dimers (parameter k_6), or on the rate of dimer- and trimer-formation (parameters k_1^+ and k_2^+).

Note that the sensitivity coefficients reflect the changes in the steady state for *small* changes in the parameter. E.g., increasing the temperature from

42 with 0.1% yields an increase in the level of mfp with 1.43%, roughly as predicted by $\partial \ln(X_{14})/\partial \ln(T) = 14.24$. An increase of the temperature from 42 with 10% yields however an increase in the level of mfp of 311.93%.

A similar sensitivity analysis may also be performed with respect to the initial conditions, see [26]. If we denote by $X^{(0)} = X(0, \kappa)$, the initial values of the vector X, for parameters κ, then the *initial concentration sensitivity coefficients* are obtained by differentiating system (1)-(14) with respect to $X^{(0)}$:

$$\frac{d}{dt}\frac{\partial X}{\partial X^{(0)}} = J(t)\frac{\partial X}{\partial X^{(0)}}(t), \tag{31}$$

with the initial condition that $\partial X/\partial X^{(0)}(0)$ is the identity matrix. It follows then that the initial concentration sensitivity matrix is given by the following matrix exponential:

$$\frac{\partial X}{\partial X^{(0)}}(t) = e^{J(t)} = \sum_{k=0}^{\infty} \frac{J(t)^k}{k!}.$$

Similarly as for the parameter-based sensitivity coefficients, it is often useful to consider the normalized, dimensionless coefficients

$$\frac{\partial X_i}{\partial {X^{(0)}}_j}(t) \cdot \frac{{X^{(0)}}_j(t)}{X_i(t)} = \frac{\partial \ln(X_i)}{\partial \ln({X^{(0)}}_j)}.$$

A numerical estimation of the initial concentration sensitivity coefficient of mfp around the steady state given in Table 2 for $T = 42$, shows that all are negligible except for the following two coefficients: $\partial \ln(X_{14})/\partial \ln(X_9^{(0)}) = -0.497748$ and $\partial \ln(X_{14})/\partial \ln(X_{13}^{(0)}) = 0.99$. While the biological significance of the dependency of mfp on the initial level of prot is obvious, its dependency on the initial level of hsp: hsf is perhaps not. Moreover, it turns out that several other variables have a significant dependency on the initial level of hsp: hsf:

- $\partial \ln(X_1)/\partial \ln(X_9(0)) = 0.49$,
- $\partial \ln(X_2)/\partial \ln(X_9(0)) = 0.49$,
- $\partial \ln(X_3)/\partial \ln(X_9(0)) = 1.04$,
- $\partial \ln(X_4)/\partial \ln(X_9(0)) = 0.49$,
- $\partial \ln(X_{10})/\partial \ln(X_9(0)) = 0.49$,
- $\partial \ln(X_6)/\partial \ln(X_9(0)) = -0.04$,
- $\partial \ln(X_7)/\partial \ln(X_9(0)) = 0.49$,
- $\partial \ln(X_9)/\partial \ln(X_9(0)) = 0.99$,
- $\partial \ln(X_{14})/\partial \ln(X_9(0)) = -0.49$,
- $\partial \ln(X_{11})/\partial \ln(X_9(0)) = 0.49$,

E.g., increasing $X_9^{(0)}$ by 1% increases the steady state values of X_7 by 0.49% and decreases the level of X_{14} by 0.49%. Increasing $X_9^{(0)}$ by 10% increases the steady state values of X_7 by 4.85% and decreases the level of X_{14} by 4.63%.

The biological interpretation of this significant dependency of the model on the initial level of hsp: hsf is based on two arguments. On one hand, the most significant part (about two thirds) of the initial available molecules of hsp in our model are present in bonds with hsf. On the other hand, the vast majority of hsf molecules are initially bound to hsp. Thus, changes in the

initial level of hsp: hsf have an immediate influence on the two main drivers of the heat shock response: hsp and hsf. Interestingly, the dependency of the model on the initial levels of either hsp or hsf is negligible.

Acknowledgments

This work has been partially supported by the following grants from Academy of Finland: project 108421 and 203667 (to I.P.), the Center of Excellence on Formal Methods in Programming (to R-J.B.).

References

1. Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Essential Cell Biology, 2nd edition*. Garland Science, 2004.
2. R.L. Burden and J. Douglas Faires. *Numerical Analysis*. Thomson Brooks/Cole, 1996.
3. Daniel R. Ciocca and Stuart K. Calderwood. Heat shock proteins in cancer: diagnostic, prognostic, predictive, and treatment implications. *Cell Stress and Chaperones*, 10(2):86–103, 2005.
4. H. El-Samad, H. Kurata, J. Doyle, C.A. Gross, and M. Khamash. Surviving heat shock: control strategies for robustness and performance. *PNAS*, 102(8):2736–2741, 2005.
5. H. El-Samad, S. Prajna, A. Papachristodoulu, M. Khamash, and J. Doyle. Model validation and robust stability analysis of the bacterial heat shock response using sostools. In *Proceedings of the 42nd IEEE Conference on Decision and Control*, pages 3766–3741, 2003.
6. C.M. Guldberg and P. Waage. Studies concerning affinity. *C. M. Forhandlinger: Videnskabs-Selskabet i Christiana*, 35, 1864.
7. C.M. Guldberg and P. Waage. Concerning chemical affinity. *Erdmann's Journal fr Practische Chemie*, 127:69–114, 1879.
8. Stefan Hoops, Sven Sahle, Ralph Gauges, Christine Lee, Jrgen Pahle, Natalia Simus, Mudita Singhal, Liang Xu, Pedro Mendes, and Ursula Kummer. Copasi – a COmplex PAthway SImulator. *Bioinformatics*, 22(24):3067–3074, 2006.
9. Harm K. Kampinga. Thermotolerance in mammalian cells: protein denaturation and aggregation, and stress proteins. *J. Cell Science*, 104:11–17, 1993.
10. 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.
11. E. Klipp, R. Herwig, A. Kowald, C. Wierling, and H. Lehrach. *Systems Biology in Practice*. Wiley-VCH, 2006.
12. H. Kurata, H. El-Samad, T.M. Yi, M. Khamash, and J. Doyle. Feedback regulation of the heat shock response in e.coli. In *Proceedings of the 40th IEEE Conference on Decision and Control*, pages 837–842, 2001.
13. James R. Lepock, Harold E. Frey, and Kenneth P. Ritchie. Protein denaturation in intact hepatocytes and isolated cellular organelles during heat shock. *The Journal of Cell Biology*, 122(6):1267–1276, 1993.

14. James R. Lepock, Harold E. Frey, A. Michael Rodahl, and Jack Kruuv. Thermal analysis of chl v79 cells using differential scanning calorimetry: Implications for hyperthermic cell killing and the heat shock response. *Journal of Cellular Physiology*, 137(1):14–24, 1988.
15. Bei Liu, Anna M. DeFilippo, and Zihai Li. Overcomming immune toerance to cancer by heat shock protein vaccines. *Molecular cancer therapeutics*, 1:1147–1151, 2002.
16. Katalin V. Lukacs, Olivier E. Pardo, M.Jo Colston, Duncan M. Geddes, and Eric WFW Alton. Heat shock proteins in cancer therapy. In Habib, editor, *Cancer Gene Therapy: Past Achievements and Future Challenges*, pages 363–368. 2000.
17. David L. Nelson and Michael M. Cox. *Principles of Biochemistry, 3rd edition*. Worth Publishers, 2000.
18. A. Peper, C.A. Grimbergent, J.A.E. Spaan, J.E.M. Souren, and R. van Wijk. A mathematical model of the hsp70 regulation in the cell. *Int. J. Hyperthermia*, 14:97–124, 1997.
19. Ion Petre, Claire L. Hyder, Andrzej Mizera, Andrey Mikhailov, John E. Eriksson, Lea Sistonen, and Ralph-Johan Back. Two metabolites are enough to drive the eukaryotic heat shock response.
20. A. Graham Pockley. Heat shock proteins as regulators of the immune response. *The Lancet*, 362(9382):469–476, 2003.
21. W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flammery. *Numerical Recipes: The Art of Scientific Computing*. Cambridge Universoty Press, 2007.
22. Theodore R. Rieger, Richard I. Morimoto, and Vassily Hatzimanikatis. Mathematical modeling of the eukaryotic heat shock response: Dynamics of the hsp70 promoter. *Biophysical Journal*, 88(3):1646–58, 2005.
23. R. Srivastava, M.S. Peterson, and W.E. Bentley. Stochastic kinetic analysis of the escherichia coli stres circuit using σ^{32}-targeted antisense. *Biotechnology and Bioengineering*, 75(1):120–129, 2001.
24. Clifford Henry Taubes. *Modeling Differential Equations in Biology*. Cambridge University Press, 2001.
25. Claire J. Tomlin and Jeffrey D. Axelrod. Understanding biology by reverse engineering the control. *PNAS*, 102(12):4219–4220, 2005.
26. Tamás Turányi. Sensitivity analysis of complex kinetic systems. tools and applications. *Journal of Mathematical Chemistry*, 5:203–248, 1990.
27. Paul Workman and Emmanuel de Billy. Putting the heat on cancer. *Nature Medicine*, 13(12):1415–1417, 2007.
28. Dennis G. Zill. *A First Course in Differential Equations*. Thomson, 2001.
29. Dennis G. Zill. *A First Course in Differential Equations with Modeling Applications*. Thomson, 2005.

Paper III

Computational heuristics for simplifying a biological model

Ion Petre, Andrzej Mizera, and Ralph-Johan Back

Originally published in Klaus Ambos-Spies, Benedikt Löwe, and Wolfgang Merkle (Eds.), *Mathematical Theory and Computational Practice: 5th Conference on Computability in Europe, CiE 2009, Proceedings*, volume 5635 of *Lecture Notes in Computer Science*, pages 399-408. Springer, Berlin Heidelberg New York, 2009.

Computational Heuristics for Simplifying a Biological Model

Ion Petre, Andrzej Mizera, and Ralph-Johan Back

Department of Information Technologies, Åbo Akademi University
Computational Biomodeling Laboratory, Turku Centre for Computer Science
FIN-20520 Turku, Finland
{ipetre,amizera,backrj}@abo.fi

Abstract. Computational biomodelers adopt either of the following approaches: build rich, as complete as possible models in an effort to obtain very realistic models, or on the contrary, build as simple as possible models focusing only on the core aspects of the process, in an effort to obtain a model that is easier to analyze, fit, and validate. When the latter strategy is adopted, the aspects that are left outside the models are very often up to the subjective options of the modeler. We discuss in this paper a heuristic method to simplify an already fit model in such a way that the numerical fit to the experimental data is not lost. We focus in particular on eliminating some of the variables of the model and the reactions they take part in, while also modifying some of the remaining reactions. We illustrate the method on a computational model for the eukaryotic heat shock response. We also discuss the limitations of this method.

Keywords: Model reduction, heat shock response, mathematical model.

1 Introduction

When designing a new molecular model for some biological process or network, the choice one has to make early on in the modeling process is whether to strive for a rich model, capturing many details, or on the contrary, to focus on a more abstract model, capturing only a few, main actors of interest. The choice is not obvious and depends heavily on the goals of the modeling project. On one hand, a rich model has the potential of being more realistic but it leads to a more complex mathematical model that may be difficult to fit to experimental data, to analyze, and ultimately may be less apt to answer to biological queries. On the other hand, a less finely grained molecular model leads to a smaller mathematical model (in terms of the number of variables and equations) that may be easier to work with, but it pays a price in ignoring a number of details. A main difficulty in choosing between a rich and a simplified molecular model is that the potential cost of starting off with a rich model only becomes transparent at a latter stage, in the process of analyzing the corresponding mathematical model. Moreover, in the case of choosing a simplified model, the selection of the aspects to be ignored in the model is left up to the subjective choice of the modeler. We

K. Ambos-Spies, B. Löwe, and W. Merkle (Eds.): CiE 2009, LNCS 5635, pp. 399–408, 2009.

discuss in this paper an intermediate approach where we start with a (potentially large, rich) model that has already been fit and validated against experimental data and we aim to simplify it in such a way that its numerical behavior remains largely unchanged. In this way, the simplified model is the result of a systematic, numerical analysis of the larger model that preserves its validation. We illustrate the approach on a computational model for the eukaryotic heat shock response and discuss the biological relevance of the simplifications we operate on the model. We also discuss the strong dependency of this approach on the numerical setup of the model; we show that our approach in the case of the heat shock response model is robust to some changes in the numerical values of the parameters, but it is sensitive to others.

2 The Heat Shock Response Model

The heat shock response is a well-conserved defence mechanism across all eukaryotic cells that enables them to survive under conditions of elevated temperatures. When exposed to heat shock, proteins inside cells tend to misfold. In turn, as an effect of their hydrophobic core being exposed, misfolded proteins form bigger and bigger aggregates with disastrous consequences for the cell, see [1]. In order to survive, the cell has to immediately react by increasing the level of chaperons (proteins that assist other proteins in the process of folding or refolding). Once the heat shock is removed, the defence mechanism is turned off and the cell eventually re-establishes the original level of chaperons, see [7,11,17].

The heat shock response has been intensively investigated in recent years for at least three main reasons. First, as a well-conserved mechanism in all eukaryotes, it is considered a promising candidate for investigating the engineering principles of gene regulatory networks, see [3,4,8,18]. Second, heat shock proteins (hsp) act as main components in a large number of cellular processes such as signaling, regulation and inflammation, see [6,16]. Moreover, their contribution to the resilience of cancer cells makes them an attractive target for cancer treatment, see [2,9,10,19].

We consider in this paper the molecular model proposed in [14] for the eukaryotic heat shock response. This model consists of only the minimum number of components that any regulatory network must contain: an activation mechanism and a feedback mechanism. Moreover, the model consists of only well-documented reactions, without using any hypothetical, unknown cellular mechanism. The control over the cellular defence mechanism against protein misfolding is implemented through the regulation of the transactivation of the hsp-encoding gene. The transcription of the gene is activated by heat shock factors (hsf) which trimerize (the trimerization includes a transient dimerization phase) and in this form bind to the heat shock element (hse), which is the promoter of the hsp-encoding gene. Once the hsf trimer is bound to the specific DNA sequence, the gene is transactivated and the transcription and translation take place. As a result, new hsp molecules are eventually synthesized. When the level of hsp is high enough, the synthesis is switched off by the following

Table 1. The reactions of the heat shock response model of [14]

(i)	$2\,\mathsf{hsf} \leftrightarrows \mathsf{hsf}_2$	(x)	$\mathsf{prot} \to \mathsf{mfp}$
(ii)	$\mathsf{hsf} + \mathsf{hsf}_2 \leftrightarrows \mathsf{hsf}_3$	(xi)	$\mathsf{hsp} + \mathsf{mfp} \leftrightarrows \mathsf{hsp{:}\,mfp}$
(iii)	$\mathsf{hsf}_3 + \mathsf{hse} \leftrightarrows \mathsf{hsf}_3\mathsf{{:}\,hse}$	(xii)	$\mathsf{hsp{:}\,mfp} \to \mathsf{hsp} + \mathsf{prot}$
(iv)	$\mathsf{hsf}_3\mathsf{{:}\,hse} \to \mathsf{hsf}_3\mathsf{{:}\,hse} + \mathsf{mhsp}$	(xiii)	$\mathsf{hsf} \to \mathsf{mhsf}$
(v)	$\mathsf{hsp} + \mathsf{hsf} \leftrightarrows \mathsf{hsp{:}\,hsf}$	(xiv)	$\mathsf{hsp} \to \mathsf{mhsp}$
(vi)	$\mathsf{hsp} + \mathsf{hsf}_2 \to \mathsf{hsp{:}\,hsf} + \mathsf{hsf}$	(xv)	$\mathsf{hsp} + \mathsf{mhsf} \leftrightarrows \mathsf{hsp{:}\,mhsf}$
(vii)	$\mathsf{hsp} + \mathsf{hsf}_3 \to \mathsf{hsp{:}\,hsf} + 2\ \mathsf{hsf}$	(xvi)	$\mathsf{hsp{:}\,mhsf} \to \mathsf{hsp} + \mathsf{hsf}$
(viii)	$\mathsf{hsp} + \mathsf{hsf}_3\mathsf{{:}\,hse} \to \mathsf{hsp{:}\,hsf} + 2\ \mathsf{hsf} + \mathsf{hse}$	(xvii)	$\mathsf{hsp} + \mathsf{mhsp} \leftrightarrows \mathsf{hsp{:}\,mhsp}$
(ix)	$\mathsf{hsp} \to \emptyset$	(xviii)	$\mathsf{hsp{:}\,mhsp} \to 2\ \mathsf{hsp}$

mechanism: hsp bind to free hsf as well as break the hsf trimers (both free and those bound to DNA). This turns off DNA transcription and blocks the forming of new hsf trimers. The whole defense mechanism is turned on again when, as a result of raised temperature, the proteins (prot) in the cell begin misfolding again. To counteract, the heat shock proteins become involved in refolding them and they free the hsf, which in turn trimerize and activate the synthesis of hsp, etc. What drives the heat shock response is the race to keep under control the level of misfolded proteins, in such a way that they are not able to accumulate, form aggregates, and eventually lead to cell death. The model consists of the molecular reactions in Table 1.

When designing this molecular model, several criteria were followed, see [14], including that only well-documented reactions should be included and that the model should explicitly consider the temperature-induced protein misfolding as the trigger of the response. The model was also designed in such a way that is consistent with itself and with the kinetic principles of biochemistry. E.g., although hsf dimers are not experimentally detectable, they should be included in the model to account as a transient step in the formation of hsf trimers. Also, since hsp and hsf are themselves proteins, they should be subject to temperature-induced misfolding just like the regular proteins prot. Moreover, the refolding of mhsf and mhsp is controlled by the same kinetic constants as the refolding of mfp. The proper folding of newly synthesized hsp is assisted by chaperons as in the case of most proteins, see [1]. The degradation of hsf, prot, and mfp was on the other hand not included in the model so that intricate compensating mechanisms of protein synthesis could be ignored, see [14].

The mathematical model associated with the molecular model in Table 1 is in terms of ordinary differential equations and it is obtained by assuming for all reactions the law of mass-action. The reasons for this choice is so that the explicit contribution of each reaction to the overall behavior could be followed. Let us denote the reactants occurring in the model according to the convention in Table 2(a). We use $\kappa \in \mathbb{R}_+^{25}$ to denote the vector with all reaction rate constants as its components, see Table 2(b): $\kappa = (k_1^+, k_1^-, k_2^+, k_2^-, k_3^+, k_3^-, k_4, k_5^+, k_5^-, k_6, k_7, k_8, k_9, \phi(T), k_{11}^+, k_{11}^-, k_{12}, \phi(T), \phi(T), k_{11}^+, k_{11}^-, k_{12}, k_{11}^+, k_{11}^-, k_{12})$.

The corresponding mathematical model consists of the following differential equations:

$$\begin{aligned}
dX_1/dt &= -k_2^+ X_1 X_2 + k_2^- X_3 - k_5^+ X_1 X_7 + k_5^- X_9 + 2\,k_8 X_4 X_7 + k_6 X_2 X_7 \\
&\quad -\varphi(T) X_1 + k_{14} X_{10} + 2\,k_7 X_3 X_7 - 2\,k_1^+ X_1^2 + 2\,k_1^- X_2 && (1)\\
dX_2/dt &= -k_2^+ X_1 X_2 + k_2^+ X_3 - k_6 X_2 X_7 + k_1^+ X_1^2 - k_1^- X_2 && (2)\\
dX_3/dt &= -k_3^+ X_3 X_6 + k_2^+ X_1 X_2 - k_2^- X_3 + k_3^- X_4 - k_7 X_3 X_7 && (3)\\
dX_4/dt &= k_3^+ X_3 X_6 - k_3^- X_4 - k_8 X_4 X_7 && (4)\\
dX_5/dt &= \varphi(T) X_1 - k_{13}^+ X_5 X_7 + k_{13}^- X_{10} && (5)\\
dX_6/dt &= -k_3^+ X_3 X_6 + k_3^- X_4 + k_8 X_4 X_7 && (6)\\
dX_7/dt &= -k_5^+ X_1 X_7 + k_5^- X_9 - k_{11}^+ X_7 X_{14} + k_{11}^- X_{12} - k_8 X_4 X_7 - k_6 X_2 X_7 \\
&\quad -k_{13}^+ X_5 X_7 + (k_{13}^- + k_{14}) X_{10} - (\varphi(T) + k_9) X_7 - k_{15}^+ X_7 X_8 \\
&\quad -k_7 X_3 X_7 + (k_{15}^- + 2\,k_{16}) X_{11} + k_{12} X_{12} && (7)\\
dX_8/dt &= k_4 X_4 + \varphi(T) X_7 - k_{15}^+ X_7 X_8 + k_{15}^- X_{11} && (8)\\
dX_9/dt &= k_5^+ X_1 X_7 - k_5^- X_9 + k_8 X_4 X_7 + k_6 X_2 X_7 + k_7 X_3 X_7 && (9)\\
dX_{10}/dt &= k_{13}^+ X_5 X_7 - (k_{13}^- + k_{14}) X_{10} && (10)\\
dX_{11}/dt &= k_{15}^+ X_7 X_8 - (k_{15}^- + k_{16}) X_{11} && (11)\\
dX_{12}/dt &= k_{11}^+ X_7 X_{14} - (k_{11}^- + k_{12}) X_{12} && (12)\\
dX_{13}/dt &= k_{12} X_{12} - \varphi(T) X_{13} && (13)\\
dX_{14}/dt &= -k_{11}^+ X_7 X_{14} + k_{11}^- X_{12} + \varphi(T) X_{13} && (14)
\end{aligned}$$

The rate coefficient of protein misfolding $\varphi(T)$ with respect to temperature T has been investigated experimentally in [12,13], and a mathematical expression describing the relation has been proposed in [11]. After adapting this formula in [11] to the time unit of our mathematical model (second), we obtain the following misfolding rate coefficient:

$$\varphi(T) = (1 - \frac{0.4}{e^{T-37}}) \cdot 1.4^{T-37} \cdot 1.45 \cdot 10^{-5}\ s^{-1}, \qquad (15)$$

where T is the numerical value of the temperature of the environment in Celsius degrees. The formula is valid for $37 \leq T \leq 45$.

For the numerical fit of the model, data of [7] on DNA binding at $42°C$ was used to relate it to $\mathsf{hsf_3}$: hse. Moreover, the initial values of the model were sought so that they give a steady state of the model at $37°C$. This latter restriction was imposed since the heat shock response is absent at $37°C$. Once suitable numerical values for the parameters were found, the model was subjected to a number of other validation tests. For a detailed discussion on the fit and the validation of the model we refer to [14] and [15]. The final numerical setup of the model is shown in Tables 2(a) and 2(b).

3 Simplifying the Model

We discuss in this section a series of numerical observations leading to several simplifications we can operate on our model, without changing its numerical

Table 2. (a) The list of variables in the mathematical model, their initial concentration values and their concentration values in one of the steady states of the system, for $T = 42$. Note that the initial state of the model is a steady state for $T = 37$. All concentrations are in $\frac{\#}{\text{cell}}$, where # denotes the number of molecules. The values should be interpreted as an average of a population of cells. [14,15]; (b) The numerical values of parameters for the fitted model [14,15].

Metabolite	Variable	Initial conc.
hsf	X_1	0.67
hsf_2	X_2	$8.73 \cdot 10^{-4}$
hsf_3	X_3	$1.22 \cdot 10^{-4}$
hsf_3: hse	X_4	3
hse	X_5	30
hsp	X_6	766.92
hsp: hsf	X_7	1403.26
hsp: mfp	X_8	71.65
prot	X_9	$1.14915 \cdot 10^8$
mfp	X_{10}	517.32
mhsf	X_{11}	$3.01 \cdot 10^{-6}$
mhsp	X_{12}	0.02
hsp: mhsf	X_{13}	$4.17 \cdot 10^{-7}$
hsp: mhsp	X_{14}	$2.24 \cdot 10^{-3}$

(a)

Constant	Reaction	Nr. value	Unit
k_1^+	(i), forward	3.49	$\frac{\text{cell}}{\#\cdot s}$
k_1^-	(i), backward	0.19	s^{-1}
k_2^+	(ii), forward	1.07	$\frac{\text{cell}}{\#\cdot s}$
k_2^-	(ii), backward	10^{-9}	s^{-1}
k_3^+	(iii), forward	0.17	$\frac{\text{cell}}{\#\cdot s}$
k_3^-	(iii), backward	$1.21 \cdot 10^{-6}$	s^{-1}
k_4	(iv)	$8.3 \cdot 10^{-3}$	s^{-1}
k_5^+	(v), forward	9.74	$\frac{\text{cell}}{\#\cdot s}$
k_5^-	(v), backward	3.56	s^{-1}
k_6	(vi)	2.33	$\frac{\text{cell}}{\#\cdot s}$
k_7	(vii)	$4.31 \cdot 10^{-5}$	$\frac{\text{cell}}{\#\cdot s}$
k_8	(viii)	$2.73 \cdot 10^{-7}$	$\frac{\text{cell}}{\#\cdot s}$
k_9	(ix)	$3.2 \cdot 10^{-5}$	s^{-1}
k_{11}^+	(xi), forward	$3.32 \cdot 10^{-3}$	$\frac{\text{cell}}{\#\cdot s}$
k_{11}^-	(xi), backward	4.44	s^{-1}
k_{12}	(xii)	13.94	s^{-1}

(b)

behavior, in particular without losing its experimental fit and validation. We then discuss the extent to which these simplifications are dependent on the numerical values of our parameters.

The first observation is that the variables mhsf and hsp: mhsf both assume negligible numerical values throughout numerical simulations for temperatures from $37°C$ to $45°C$. Even when their initial values are increased to higher values, e.g. to 100 each, their numerical convergence towards their steady state values is very fast. Moreover, if the increase in the initial values of mhsf and hsp: mhsf is so that the total amount of hsf and of hsp remain unchanged, then the experimental fit and validation of the model remain largely unchanged. The reason for this behavior is that the reactions having mhsf as a product, i.e. reactions (xiii) and the reverse reaction (xv) have a negligible flux rate, primarily due to the small kinetic rate constant of the protein misfolding law, see (15). Consequently, the reaction producing hsp: mhsf, i.e. reaction (xv), also has negligible flux rate. On the other hand, the reactions having mhsf and hsp: mhsf as reactants reach much higher flux rates because of larger kinetic constants and high levels of hsp, a co-reactant in reaction (xv). We decide then to eliminate both mhsf and hsp: mhsf from the model, along with the reactions where they take part in, i.e., reactions (xiii), (xv), and (xvi).

Note now that the situation is somewhat similar for hsf, hsf_2 and hsf_3: they all assume small (albeit not negligible) values throughout numerical simulations. There is however a crucial difference which points to their significance for the model: when increasing the initial level of hsf_3, even in such a way that the total level of hsf is unchanged, the fit to the DNA binding experimental data of [7] is drastically changed.

The observation that the flux of the hsf misfolding reaction is negligible was the main rationale behind eliminating mhsf and hsp: mhsf from the model. This leads to the observation that the flux of the hsp misfolding reaction, leading to the formation of mhsp is also negligible. The case of mhsp is however different because it is also the end product of reaction (iv). Moreover, mhsp plays a central role in our model, being the source of all induced hsp through reactions (iv), (xvii) and (xviii). The numerical values assumed by mhsp throughout simulations for temperatures between $37°C$ and $45°C$ are small, but not negligible. They are however negligible relative to the total level of hsp. Moreover, the numerical convergence of mhsp towards its steady state value is very fast, even in the case when the initial level of mhsp is increased several folds. This points to the observation that mhsp plays the role of a transient state towards hsp, having a very high turnover rate. As such, it could be eliminated from the model if only mhsp were replaced in reaction (iv) with hsp. Consequently, we eliminate mhsp from the model, along with reactions (xiv), (xvii) and (xviii). At the same time, we replace reaction (iv) with

(iv') $\mathsf{hsf_3}\colon \mathsf{hse} \rightarrow \mathsf{hsf_3}\colon \mathsf{hse} + \mathsf{hsp}$

The simplified molecular model has only 10 variables and 12 reactions, compared to 14 variables and 18 reactions in the initial model. The numerical simulations of the simplified model for temperatures between $37°C$ and $45°C$ are indistinguishable from those of the initial model.

Regarding the biological relevance, the simplified model differs from the initial model in ignoring the misfolded form of hsf and hsp, as well as ignoring that newly synthesized proteins often need chaperons to form their native fold. Excluding the misfolding of hsf and hsp is reasonable because the numerical levels of misfolded hsf and hsp are negligible with respect to the level of mfp and thus, their competition for the chaperon resources of the cell is insignificant. Excluding the role of chaperons in assisting the formation of the native fold of newly synthesized proteins is justified by the high speed of the reaction, relative to the speed of the other reactions in our model. As such, the complex chaperon - newly synthesized protein is a very fast transient stage in the model and can be ignored.

It should be noted that the simplifications we have made on the model are based on numerical arguments and so, in principle, they are dependant on the numerical values of the parameters of the model. To test the robustness of the model reductions against changes in the numerical setup of the model, we perform several tests. In each test, we either change the initial values of some variables, or we change the values of some kinetic rate constants. For each new numerical setup we set the initial values of all variables to their steady state

values at $37°C$, similarly as done in [15] (to underline that the heat shock response is missing at $37°C$). Finally, we compare the numerical behavior of the model with that of its simplified version obtained as above, for temperatures between $37°C$ and $45°C$.

We first consider a numerical setup where the total level of hsf is increased by 1000 to a value of around 2400. In a second test, we increase both the total level of hsf by 1000 and the total level of hse by 100. In both tests, the numerical behaviors of the models and those of their simplified versions are undistinguishable. In a third test, we increase the total level of hsp by 1000. When estimating the steady state values of the model at $37°C$, we note that they are identical with those of the initial model, summarized in Table 2(a). This raises an intriguing problem of independent interest: is the steady state of the model independent of the initial total level of hsp?

A test where the complex chaperon–misfolded protein is made more unstable by increasing the kinetic rate constant k_{11}^- to 25 yields a numerically equivalent simplified model. In a final test, we decrease the value of the kinetic rate constant k_{12} of the refolding reaction (xii) from almost 14 to 1. In this way, we induce a great increase in the values of misfolded proteins of all types to test whether eliminating mhsf and mhsp is still possible in this context. It turns out that eliminating mhsf and hsp: mhsf is possible and yields a numerically equivalent simplified model. On the other hand, eliminating mhsp and hsp: mhsp changes the behavior of the model pronouncedly. E.g., mfp peaks at a lower value showing that the simplified model, where hsp is not subject to misfolding, is more efficient in fighting off the accumulation of mfp. A main reason why the elimination of misfolded hsp fails is because, unlike in the previous tests, the change in the refolding rate is not accounted for when setting the initial values of the variables to the steady state values at $37°C$, since the refolding reaction has a negligible flux at that temperature. At $42°C$ however, protein refolding, in particular that of mhsp, becomes very important and removing it from the model makes a big difference.

4 Discussion

Having simple biomodels is very important for being able to analyze their mathematical properties and for their integration into larger models. In the case of the heat shock response, adding the phosphorylation of hsf in all of its homo- and hetero-polymers, along with its influence on gene transcription leads to a combinatorial explosion in the number of variables of the model. As such, decreasing the number of variables, in particular the elimination of mhsf and hsp: mhsf reduces the difficulty of the problem.

Several aspects contribute to the model simplification succeeding in a given numerical setup. The most important is that we eliminate variables that have a fast numerical convergence to their steady state values. This procedure is often referred to as a time-separation principle. A factor here is the flux rate of the reactions producing certain variables of the model. If the total flux contributing

to producing a given variable remains very small, then that variable will converge fast to its steady state value and it can be eliminated from the model. There are at least two reasons why a flux rate can be small: a small kinetic constant, or much higher kinetic constant in reactions using some of the same reactants. In the context of the heat shock response model, one more factor plays an important role: the condition that the initial values of all variables are a steady state of the model at $37°C$. It turns out that the model has an interesting property, formulated as a theorem in the appendix: the steady state values of most of its variables are independent of the temperature. In this way, even at higher temperature, several of the variables of the model start from their steady state values and witness only minor numerical disturbances before returning to the same values.

The model simplification discussed in this paper is dependant on the numerical setup of the model: on the initial values of the variables and on the numerical values of the kinetic constants. Even if the initial and the simplified models appear to be numerically equivalent in one particular setup, they may be very different in other setups. To evaluate the robustness of the model simplifications, one should compare the two models in several numerical setups, spanning the domain of expected values for the model parameters. Some of the simplifications may turn out to be robust against numerical variations, as it is the case with eliminating hsf and mhsf in the heat shock model, while others may be valid only in special numerical setups.

The main difficulty in designing a simple biomodel is that the decision to exclude variables and reactions from the model is most often done at the early stage of considering the molecular basis of the model. At that stage however it is crucial to ensure that all aspects of potential interest are included in the model. Appreciating the potentially insignificant contribution of some of the aspects is very difficult at that stage, without having first a well-validated numerical setup for the model. The approach we have discussed in this paper takes an intermediate view: one may start with a rich model that is first numerically fit and validated against experimental data and then it is subjected to a numerical analysis to identify the components that can be eliminated without changing the numerical behavior of the model. In this way, the result is a model that remains faithful to the biological data and soundly identifies those aspects of the biological reality that have insignificant contribution to the overall behavior.

Acknowledgments. This work has been partially supported by projects 108421 and 203667 (to I.P.) from Academy of Finland. The numerical simulations and estimations discussed in this paper were performed with Copasi, see [5].

References

1. Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P.: Essential Cell Biology, 2nd edn. Garland Science (2004)
2. Ciocca, D.R., Calderwood, S.K.: Heat shock proteins in cancer: diagnostic, prognostic, predictive, and treatment implications. Cell Stress and Chaperones 10(2), 86–103 (2005)

3. El-Samad, H., Kurata, H., Doyle, J., Gross, C.A., Khamash, M.: Surviving heat shock: control strategies for robustness and performance. PNAS 102(8), 2736–2741 (2005)
4. El-Samad, H., Prajna, S., Papachristodoulu, A., Khamash, M., Doyle, J.: Model validation and robust stability analysis of the bacterial heat shock response using sostools. In: Proceedings of the 42nd IEEE Conference on Decision and Control, pp. 3766–3741 (2003)
5. 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)
6. Kampinga, H.K.: Thermotolerance in mammalian cells: protein denaturation and aggregation, and stress proteins. J. Cell Science 104, 11–17 (1993)
7. 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)
8. Kurata, H., El-Samad, H., Yi, T.M., Khamash, M., Doyle, J.: Feedback regulation of the heat shock response in e.coli. In: Proceedings of the 40th IEEE Conference on Decision and Control, pp. 837–842 (2001)
9. Liu, B., DeFilippo, A.M., Li, Z.: Overcomming immune toerance to cancer by heat shock protein vaccines. Molecular cancer therapeutics 1, 1147–1151 (2002)
10. Lukacs, K.V., Pardo, O.E., Colston, M.J., Geddes, D.M., Eric WFW Alton: Heat shock proteins in cancer therapy. In: Habib (ed.) Cancer Gene Therapy: Past Achievements and Future Challenges, pp. 363–368 (2000)
11. Peper, A., Grimbergent, C.A., Spaan, J.A.E., Souren, J.E.M., van Wijk, R.: A mathematical model of the hsp70 regulation in the cell. Int. J. Hyperthermia 14, 97–124 (1997)
12. Lepock, J.R., Frey, H.E., Ritchie, K.P.: Protein denaturation in intact hepatocytes and isolated cellular organelles during heat shock. The Journal of Cell Biology 122(6), 1267–1276 (1993)
13. Lepock, J.R., Frey, H.E., Rodahl, A.M., Kruuv, J.: Thermal analysis of chl v79 cells using differential scanning calorimetry: Implications for hyperthermic cell killing and the heat shock response. Journal of Cellular Physiology 137(1), 14–24 (1988)
14. Petre, I., Mizera, A., Hyder, C.L., Mikhailov, A., Eriksson, J.E., Sistonen, L., Back, R.-J.: A new mathematical model for the heat shock response. In: Kok, J. (ed.) Algorithmic bioprocesses, Natural Computing. Springer, Heidelberg (2008)
15. Petre, I., Hyder, C.L., Mizera, A., Mikhailov, A., Eriksson, J.E., Sistonen, L., Back, R.-J.: A simple mathematical model for the eukaryotic heat shock response (manuscript, 2009)
16. Graham Pockley, A.: Heat shock proteins as regulators of the immune response. The Lancet 362(9382), 469–476 (2003)
17. 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), 1646–1658 (2005)
18. Tomlin, C.J., Axelrod, J.D.: Understanding biology by reverse engineering the control. PNAS 102(12), 4219–4220 (2005)
19. Workman, P., de Billy, E.: Putting the heat on cancer. Nature Medicine 13(12), 1415–1417 (2007)

Appendix

The next theorem formulates an interesting property of the heat shock response model. We formulate the property for the simplified model of the heat shock response.

Theorem 1. *Let* $c^1 = (c^1_1, c^1_2, c^1_3, \ldots, c^1_{10})$ *be a steady state of the system at temperature* T_1 *and* $c^2 = (c^2_1, c^2_2, c^2_3, \ldots, c^2_{10})$ *a steady state at temperature* T_2*, where* c^1_i *and* c^2_i *for* $i = 1, \ldots, 10$ *are steady state concentrations of metabolite* X_i *at temperatures* T_1 *and* T_2 *respectively. Then* $c = (c^1_1, \ldots, c^1_7, c^2_8, c^2_9, c^2_{10})$ *is a steady state of the system at temperature* T_2*.*

Proof. Let c^1 and c^2 be steady states at temperatures T_1 and T_2, respectively. Further, let us split the system of differential equations (1)-(10) into two subsystems: one containing equations (1)-(7) and the other consisting of equations (8)-(10). Equation (6) is the only one in the first subsystem with right-hand side containing functions defined by the second subsystem, i.e. $X_8(t)$, $X_9(t)$ and $X_{10}(t)$, and can be by (9) rewritten in the following form:

$$\begin{aligned} dX_6/dt = {} & k_4\, X_4 - k_5^+\, X_1\, X_6 + k_5^-\, X_7 - k_8\, X_4\, X_6 - k_6\, X_2\, X_6 \\ & - k_7\, X_3\, X_6 - k_9\, X_6 - dX_8/dt. \end{aligned} \tag{16}$$

When considering the steady states, the left-hand sides of (1)-(10) are set to 0 and in consequence equation (16) can be written as

$$0 = k_4\, X_4 - k_5^+\, X_1\, X_6 + k_5^-\, X_7 - k_8\, X_4\, X_6 - k_6\, X_2\, X_6 - k_7\, X_3\, X_6 - k_9\, X_6.$$

This algebraic relation does not contain any of functions $X_8(t)$, $X_9(t)$ or $X_{10}(t)$ and hence the steady state algebraic relations of subsystem (1)-(7) become independent of them. As a consequence, the relations do not contain temperature as a parameter and are the same both for T_1 and T_2. Since the same equations have the same solutions, it follows that $c = (c^1_1, \ldots, c^1_7, c^2_8, c^2_9, c^2_{10})$ is a steady state of the whole system at temperature T_2.

The biological significance of Theorem 1 deserves some comments. Even though the cell approaches similar steady state levels regardless of the temperature values, the *time* it takes to arrive in a certain neighborhood of the steady state is longer for higher temperature values. Even if one starts in the steady state, the *effort* required of the cell is higher for higher temperatures: the fluxes of all reactions are higher for higher temperatures. The intuitive reason for this is that the misfolding rate is vastly accelerated for higher temperatures, eventually accelerating all other reactions.

Paper IV

Stochastic modelling of the eukaryotic heat shock response

Andrzej Mizera and Barbara Gambin

Originally published in *Journal of Theoretical Biology*, 265(3): 455–466, 2010.

Journal of Theoretical Biology 265 (2010) 455–466

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Stochastic modelling of the eukaryotic heat shock response

Andrzej Mizera [a,b,*], Barbara Gambin [a]

[a] Institute of Fundamental Technological Research, Polish Academy of Sciences, Pawińskiego 5B, 02-106 Warsaw, Poland
[b] Department of Information Technologies, Åbo Akademi University & Turku Centre for Computer Science, Joukahaisenkatu 3-5A, FIN-20520 Turku, Finland

ARTICLE INFO

Article history:
Received 9 July 2009
Received in revised form
27 April 2010
Accepted 27 April 2010
Available online 6 May 2010

Keywords:
Stochastic model
Computer simulations
Markov chain
Gillespie algorithm
Stationary distribution

ABSTRACT

The heat shock response (HSR) is a highly evolutionarily conserved defence mechanism allowing the cell to promptly react to elevated temperature conditions and other forms of stress. It has been subject to intense research for at least two main reasons. First, it is considered a promising candidate for deciphering the engineering principles underlying regulatory networks. Second, heat shock proteins (main actors of the HSR) play crucial role in many fundamental cellular processes. Therefore, profound understanding of the heat shock response would have far-reaching ramifications for the cell biology.

Recently, a new deterministic model of the eukaryotic heat shock response has been proposed in the literature. It is very attractive since it consists of only the minimum number of components required by any functional regulatory network, while yet being capable of biological validation. However, it admits small molecule populations of some of the considered metabolites. In this paper a stochastic model corresponding to the deterministic one is constructed and the outcomes of these two models are confronted. The aim with this comparison is to show that, in the case of the heat shock response, the approximation of a discrete system with a continuous model is a reasonable approach. This is not always the truth, especially when the numbers of molecules of the considered species are small. By making the effort of performing and analysing 1000 stochastic simulations, we investigate the range of behaviour the stochastic model is likely to exhibit. We demonstrate that the obtained results agree well with the dynamics displayed by the continuous model, which strengthens the trust in the deterministic description. A proof of the existence and uniqueness of the stationary distribution of the Markov chain underlying the stochastic model is given. Moreover, the obtained view of the stochastic dynamics and the performed comparison to the outcome of the continuous formulation provide more insight into the dynamics of the heat shock response mechanism.

1. Introduction

The heat shock response is the most highly evolutionarily conserved defence mechanism (Lindquist and Craig, 1988). It exists in all eukaryotic cells, protects them from the damaging influence of elevated temperature and allows them to promptly react to other forms of environmental stress. The heat shock response has been subject to intense research recently, see Chen et al. (2007), Powers and Workman (2007) and Voellmy and Boellmann (2007), for at least two reasons. On one hand, as a well-conserved mechanism, it is considered a promising candidate for deciphering the engineering principles underlying regulatory networks. On the other hand, heat shock proteins play crucial roles in many fundamental cellular processes such as protein biogenesis, dismantling of damaged proteins, activation of immune responses and signalling, see Kampinga (1993) and Pockley (2003). Therefore, profound understanding of the heat shock response would have far-reaching ramifications for the cell biology and could potentially allow for treatment of a number of diseases, such as neurodegenerative and cardiovascular disorders, cancer, ageing, see Balch et al. (2008), Liu et al. (2002), Lukacs et al. (2000), Morimoto (2008) and Workman and de Billy (2007).

Although a number of mathematical models describing the heat shock response both in eukaryotes and in bacteria have been presented in the literature, see Donati et al. (1990), Jones et al. (1993), Parsell and Lindquist (1993), Peper et al. (1997), Petre et al. (2009b), Remondini et al. (2006), Rieger et al. (2005) and Szymańska and Żylicz (2009), still a comprehensive mechanistic understanding of this process is lacking. In Petre et al. (2009b) a new model of the eukaryotic heat shock response together with an associated continuous mathematical model based on ordinary differential equations have been discussed. The novelty of the model in Petre et al. (2009b) is due to the fact that, unlike other previous models, it is based solely on well-documented reactions

* Corresponding author at: Department of Information Technologies, Åbo Akademi University, Joukahaisenkatu 3-5A, FIN-20520 Turku, Finland.
Tel.: +358 2 2154045.
E-mail addresses: amizera@abo.fi, amizera@ippt.gov.pl (A. Mizera), bgambin@ippt.gov.pl (B. Gambin).

0022-5193/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2010.04.029

and does not incorporate modelling "blackboxes" such as hypothetical, experimentally unsupported cellular mechanisms whose only purpose is to enforce appropriate behaviour. The simplified version of the model (see Petre et al., 2009a, for details) includes the temperature-induced protein misfolding, all three forms of heat shock factors: monomers, dimers and trimers, the backregulation of the transactivation of the heat shock protein encoding gene and the chaperone activity of heat shock proteins. At the same time, it contains as few reactions and reactants as possible. It is worth noticing that the model consists of only the minimum number of components required by any functional regulatory network: an activation mechanism and a feedback mechanism. Nonetheless, the associated continuous model predictions correlate well with experimental observations on the heat-induced transactivation of the hsp-encoding genes at different temperatures from the range 37 to 43 °C (in particular, the prolonged transcription at 43 °C is confirmed) and the return to the original level of hsp production once the stress is removed (publication in preparation). Moreover, the model perfectly illustrates the experimentally observed process of "self-learning" of the HSR system: the response to a second consecutive heat shock is significantly lower. This is due to a transient increase in the free hsp level caused by the preliminary heat shock. In other words, the increase is a form of temporary memory of the fact that the cell was recently exposed to heat shock conditions.

However, the undertaken modelling approach that utilises ordinary differential equations is just one of many other modelling paradigms (e.g. stochastic formulation, process calculi, Petri nets, etc.), which could be exploited in the context of the heat shock response. In this paper we follow one of the other formalisms: we develop a stochastic model associated with the simplified version of the model from Petre et al. (2009b) which has been described in Petre et al. (2009a). According to current scientific knowledge, ignoring quantum mechanical effects, biological systems can be viewed as deterministic of their very nature, with their dynamics entirely specified, given sufficient information on the state of the system (position, orientation and momentum of every single molecule) and a complete understanding of the chemistry and physics of the interactions between biomolecules. Unfortunately, we are still unable to model biological systems of realistic complexity and size using such a molecular dynamic approach (Wilkinson, 2006). Therefore the current models admit far-reaching simplifications, which result in a higher level view of the system being modelled. However, these abstractions change the character of the dynamics, which becomes intrinsically stochastic and requires consideration of statistical physics to describe the stochastic process governing it. Especially at low concentrations of the involved reactants, random fluctuations may have a significant impact on the reaction dynamics, but the deterministic approach to chemical kinetics fails to capture such phenomena, see McAdams and Arkin (1999) and Srivastava et al. (2002). For example, let us consider the famous Lotka–Volterra system of coupled ordinary differential equations describing an ecological predator–prey model. The solutions of this system are known to be periodic (except for the stationary point) independently of the initial size of predator and prey populations. However, in the stochastic formulation there exists a "catastrophic" sequence of events which leads to depletion of preys by predators and, in consequence, to the extinction of predators as well. When running the model long enough, the probability of not executing this catastrophic sequence drops to zero. This leads to radical qualitative differences in the trajectories obtained by these two approaches: in the deterministic case the trajectory in the predator versus prey phase space is an ellipse, while in the stochastic case the trajectory eventually reaches the trivial steady state of no predators and no prey individuals in the system. The expected time it takes to reach this state depends on the initial number of species. Such discrepancy in the trajectories is especially easily observed when the initial population sizes are small.

Another significant impact of random fluctuations can be observed in the model of T cell receptor signalling presented in Lipniacki et al. (2008), where it is shown that, because of bistability of the system and the fact that the T cell activation is due to a small number of foreign peptides, the responses are highly stochastic. This results in stochastic trajectories not following the deterministic trajectory, which converges to a steady state. Instead, the stochastic realisations may occasionally jump between the basins of attraction of two possible states. In particular, as was shown in Lipniacki et al. (2008), stochastic noise can cause a transition from the higher stable state to the lower one and most of the stochastic trajectories are trapped in the basin of attraction of the latter steady state in contrast to the deterministic case. As a result, the qualitative behaviour revealed by the stochastic approach differs significantly from the behaviour obtained from the deterministic description. For details we refer the reader to Lipniacki et al. (2008).

Although for a complex system detailed mathematical analysis based on the "chemical master equation" is intractable (Wilkinson, 2006), it is possible to gain insight into the system's dynamics by performing a series of stochastic simulations of the time-evolution of such system by so-called Gillespie (1976) algorithm. The algorithm is a well-established procedure for generating a stochastic realisation of the system's temporal behaviour. However, due to reasons such as computational efficiency, availability of dedicated simulation software with analysis tools (steady-state, sensitivity analysis, etc.), and expertise in the theory of differential equations, the deterministic modelling approach is commonly used in examination of biological systems, although the stochastic formulation in many cases would be more justified.

Bearing in mind the above mentioned merits of the new simplified heat shock response model described in Petre et al. (2009a), the aim of this paper is to show that in this particular case approximating a discrete system with a continuous model is a valid approach. A stochastic model complementary to the deterministic one is developed. An effort to perform 1000 stochastic simulations is made in order to investigate whether the qualitative results of the stochastic model agree with the deterministic outcome. Having the problem of small number of molecules of some of the reactants in mind (initial number concentrations of hsf, hsf_2, hsf_3, hsf_3:hse, hse, hsp:mfp, see Petre et al., 2009a, for details), as explained above, one could expect the time-course trajectories obtained with the stochastic model to be substantially different from the trajectories computed in the deterministic formalism. However, we show that the influence of the random fluctuations does not invalidate the continuous approach and the obtained results support the use of the deterministic formulation in this case. In particular, we investigate the number of steady states of the deterministic model and compare the obtained results with the dynamics demonstrated by the stochastic model. We show that the underlying stochastic process of our model has a unique stationary distribution and that the performed stochastic simulation results reveal no evidence of multistationarity, which is consistent with the deterministic description. Additionally, this analysis let us gain some more insight into the dynamics of the heat shock response mechanism. The question about the stationarity and stability, i.e. the number of steady-states and whether they are stable or unstable, is important in the examination of the dynamics of biological systems. For example, bistability in biological systems is, in

general, accompanied by hysteresis, which in turn promotes robustness (Karmakar and Bose, 2007; Lipniacki et al., 2008).

The paper is organised as follows. In Section 2 we briefly describe the simplified deterministic model (named *deterministic model* for compactness in the continuation) of the heat shock response in eukaryotic cells which was proposed in Petre et al. (2009a). Next, in Section 3, we discuss the Markov jump process which constitutes the corresponding stochastic model and show that it has a unique stationary distribution. Further, in Section 4, the stochastic simulation results are discussed and a comparison between the deterministic and stochastic model is presented. Finally, we end with conclusions in Section 5.

2. Deterministic model

The model of the eukaryotic heat shock response consists of four main modules: the heat-induced protein misfolding, the dynamic transactivation of the genes encoding heat shock proteins, their backregulation and the chaperone activity of the heat shock proteins.

At elevated temperatures proteins tend to misfold and create aggregates, which has disastrous effects on the cell. In order to survive, the cell has to promptly increase the level of *heat shock proteins* (hsp), which is the main task of the heat shock response mechanism. Heat shock proteins act as chaperones: they interact with the *misfolded proteins* (mfp) and assist them in refolding to their native conformation (prot). The control over the defence mechanism against the temperature-induced harmful phenomena is implemented through the regulation of the transactivation of the hsp-encoding gene. Activation of the transcription proceeds along the following scheme: *heat shock factors* (hsf) trimerize (through a transient dimerisation) and in this form bind to the *heat shock element* (hse), i.e. the promoter of the hsp-encoding gene. Once the hsf-trimer (hsf_3) is bound to the specific DNA sequence (hsf_3:hse), the gene is transactivated and new hsp molecules are eventually synthesised. Finally, when the level of hsps is high enough to cope with the thermal stress, the production is switched off: hsps bind both to free hsfs and hsfs that occur in compound forms (hsf_2, hsf_3, hsf_3:hse), which, in consequence, get disassembled. As a result, DNA transcription of hsp-encoding gene is turned off and the formation of new hsf trimers is blocked. The full list of molecular reactions constituting the model is presented in Table 1. By assuming the law of mass-action for all reactions (R_1)–(R_{17}), the associated mathematical model based on ordinary differential equations is obtained. The rate coefficient of protein misfolding with respect to the temperature ($\varphi(T)$) in reaction (R_{14}) is given by the following formula:

$$\varphi(T)=\left(1-\frac{0.4}{e^{T-37}}\right)\times 1.4^{T-37}\times 1.45\times 10^{-5}\,\mathrm{s}^{-1}, \tag{1}$$

where T is the numerical value of the temperature of the environment in Celsius degrees. The formula is valid for $37\le T\le 45$. It is based on experimental investigations in Lepock et al. (1993, 1988) and was originally proposed in Peper et al. (1997). Expression (1) in its current form was obtained by adapting the original formula to the time unit of the discussed mathematical model (see Petre et al., 2009a). In our survey the temperature is set to 42 °C, i.e. the cells are exposed to heat shock conditions.

As shown in Petre et al. (2009a), there are three mass-conservation relations in the model: the total number of heat shock factor molecules, heat shock elements and protein molecules (either misfolded or in native conformation) is conserved in time. This can be written formally as

$$C_1=\mathrm{hse}(t)+3\mathrm{hsf}_3:\mathrm{hse}(t), \tag{2}$$

$$C_2=\mathrm{hsf}(t)+2\mathrm{hsf}_2(t)+3\mathrm{hsf}_3(t)+3\mathrm{hsf}_3:\mathrm{hse}(t)+\mathrm{hsp}:\mathrm{hsf}, \tag{3}$$

$$C_3=\mathrm{prot}(t)+\mathrm{mfp}(t)+\mathrm{hsp}:\mathrm{mfp}(t) \tag{4}$$

for all $t\ge 0$, where $C_1, C_2, C_3\ge 0$ are some constants determined by initial conditions, i.e. right-hand side expressions at $t=0$ in the above Eqs. (2)–(4).

The described model of eukaryotic heat shock response is based solely on well-documented reactions and does not include any "artificial" elements such as experimentally unsupported components or biochemical reactions. For a detailed discussion of the model, we refer the reader to Petre et al. (2009a).

Table 1
The simplified model for the eukaryotic heat shock response.

$2\mathrm{hsf}\to\mathrm{hsf}_2$	(R1)
$\mathrm{hsf}_2\to 2\mathrm{hsf}$	(R2)
$\mathrm{hsf}+\mathrm{hsf}_2\to\mathrm{hsf}_3$	(R3)
$\mathrm{hsf}_3\to\mathrm{hsf}+\mathrm{hsf}_2$	(R4)
$\mathrm{hsf}_3+\mathrm{hse}\to\mathrm{hsf}_3:\mathrm{hse}$	(R5)
$\mathrm{hsf}_3:\mathrm{hse}\to\mathrm{hsf}_3+\mathrm{hse}$	(R6)
$\mathrm{hsf}_3:\mathrm{hse}\to\mathrm{hsf}_3:\mathrm{hse}+\mathrm{hsp}$	(R7)
$\mathrm{hsp}+\mathrm{hsf}\to\mathrm{hsp}:\mathrm{hsf}$	(R8)
$\mathrm{hsp}:\mathrm{hsf}\to\mathrm{hsp}+\mathrm{hsf}$	(R9)
$\mathrm{hsp}+\mathrm{hsf}_2\to\mathrm{hsp}:\mathrm{hsf}+\mathrm{hsf}$	(R10)
$\mathrm{hsp}+\mathrm{hsf}_3\to\mathrm{hsp}:\mathrm{hsf}+2\mathrm{hsf}$	(R11)
$\mathrm{hsp}+\mathrm{hsf}_3:\mathrm{hse}\to\mathrm{hsp}:\mathrm{hsf}+\mathrm{hse}+2\mathrm{hsf}$	(R12)
$\mathrm{hsp}\to$	(R13)
$\mathrm{prot}\to\mathrm{mfp}$	(R14)
$\mathrm{hsp}+\mathrm{mfp}\to\mathrm{hsp}:\mathrm{mfp}$	(R15)
$\mathrm{hsp}:\mathrm{mfp}\to\mathrm{hsp}+\mathrm{mfp}$	(R16)
$\mathrm{hsp}:\mathrm{mfp}\to\mathrm{hsp}+\mathrm{prot}$	(R17)

3. Stochastic model

Stochastic modelling of biochemical networks is today well-established. The time-evolution of a reaction system can be regarded as a stochastic process (cf. Wilkinson, 2006). In particular, the dynamics of a biochemical network can be viewed as a continuous-time Markov process. A continuous-time stochastic process $\{X(t), t\ge 0\}$ with discrete state space $\mathcal{S}$ is said to be a continuous-time Markov chain (CTMC for short) if

$$P\{X(t_n)=i_n|X(t_0)=i_0,\ldots,X(t_{n-1})=i_{n-1}\}=P\{X(t_n)=i_n|X(t_{n-1})=i_{n-1}\}$$

for all $0\le t_0<\cdots<t_{n-1}<t_n$ and $i_0,\ldots,i_{n-1},i_n\in\mathcal{S}$. The Markov property expresses that the conditional distribution of a future state given the present and past states depends only on the present state and is independent of the past.

We consider a time-homogeneous Markov chain for which the transition probability $P\{X(t+u)=j|X(u)=i\}$ is independent of u.

Let $\mathbf{Q}=(q_{ij})_{i,j\in\mathcal{S}}$ be the infinitesimal transition rate matrix of the continuous-time Markov chain $\{X(t)\}$ such that the following assumption is satisfied.

Assumption 1. $v_i=\sum_{j\neq i}q_{ij}$ are positive and bounded in $i\in\mathcal{S}$.

As stated in Tijms (2003), if Assumption 1 is fulfilled, it can be shown that the infinitesimal transition rates determine a unique continuous-time Markov chain which is precisely a Markov jump process constructed in the following way:

(a) if the system jumps to state i, it then stays in state i for an exponentially distributed time with mean $1/v_i$ independently of how the system reached state i and how long it took to get there (this explains the name *sojourn-time rates* used for $\{v_i\}$);
(b) if the system leaves state i, it jumps to state j $(j\neq i)$ with probability p_{ij} independently of the duration of the stay in state i.

Further, let $\{X_n, n=0,1,\dots\}$ be the *embedded Markov chain*, i.e. X_n is defined as the state of $\{X(t)\}$ just after the n-th transition with the convention that $X_0=X(0)$. The one-step transition probabilities of the discrete-time Markov chain $\{X_n\}$ are given by

$$p_{ij}=\begin{cases} q_{ij}/v_i, & j\neq i,\\ 0, & j=i\end{cases}$$

for all $i,j\in\mathcal{S}$ (e.g. see Tijms, 2003).

The corresponding stochastic formulation of the HSR molecular model presented in Section 2 is a system of 10 chemically active species S_i $(i=1,\dots,10)$, ($S_1\equiv$ hse, $S_2\equiv$ hsf, $S_3\equiv \mathrm{hsf}_2$, $S_4\equiv \mathrm{hsf}_3$, $S_5\equiv \mathrm{hsf}_3$: hse, $S_6\equiv$ hsp, $S_7\equiv$ hsp : hsf, $S_8\equiv$ hsp : mfp, $S_9\equiv$ mfp, $S_{10}\equiv$ prot) that participate in 17 chemical reactions (R_1)–(R_{17}) in some volume V. More specifically, in our case the abstract volume V is simply a eukaryotic cell. The system's state space $\mathcal{S}\subset\mathbb{N}^{10}$ is defined as

$$\begin{aligned}\mathcal{S}=\{&(s_1,\dots,s_{10})^T:\forall_{i\in\{1,\dots,10\}}s_i\in\mathbb{N},\ s_1+s_5=C_1,\\ &s_2+2s_3+3s_4+3s_5+s_7=C_2,\ s_8+s_9+s_{10}=C_3,\ s_6+s_7+s_8\leq K\},\end{aligned}\tag{5}$$

where $C_1\geq 1$, $C_2\geq 3$ and $C_3\geq 1$ are constants describing the fixed in time total number of hse, hsf and protein (either misfolded or in native conformation) molecules present in the system, respectively. The last inequality, i.e. $s_6+s_7+s_8\leq K$, requires some comment. In the absence of it the model would make allowance for any unbounded number of free hsp molecules to co-exist. However, this is certainly contrary to the fact that any living cell has a limited volume and, in consequence, can only contain a finite number of hsp molecules. Thus, in order to make the model more realistic, a big enough (in the sense that it allows for the appropriate number of free hsp molecules to be present in the system) constant K is introduced and an upper bound on the value of the S_6 variable is imposed by the last inequality. The direct consequence of adding it is that the state space $\mathcal{S}$ becomes finite.

The system is in state $s=(s_1,\dots,s_{10})^T$ at time t if and only if the number of molecules of species S_i at time t is s_i for all $i=1,\dots,10$. The conditions posed on the constants C_1, C_2 and C_3 ensure that the HSR mechanism is operational, i.e. that at least one hse molecule is present in the system, that the system is able to produce at least one hsf_3 molecule that can bind to the DNA and, in consequence, initiate the transcription and translation of hsp. Finally, that at least one generic protein prone to misfolding exists in the system.

Each reaction R_μ is characterised by a stochastic rate constant c_μ $(\mu=1,\dots,17)$, see Table 3. The values were obtained from the deterministic model (Petre et al., 2009a). In the case of reaction (R_1), the deterministic rate constant value was multiplied by 2 in order to obtain the value for the corresponding stochastic rate constant.

Let v_μ be the μ-th column of the stoichiometry matrix of the HSR system presented in Table 2. Reaction R_μ causes the system to make a transition from some state $i\in\mathcal{S}$ to state $j=i+v_\mu$. The *fundamental hypothesis* of the stochastic formulation of chemical kinetics (Gillespie, 1976) is that the reaction parameter c_μ can be defined as follows:

$c_\mu\,\mathrm{d}t\equiv$ average probability, to first order in $\mathrm{d}t$, that a particular combination of R_μ reactant molecules will react accordingly in the next time interval $\mathrm{d}t$.

As shown by Gillespie (1976), c_μ is dependent on the radii of the molecules involved in the reaction and their average relative velocities, where the average relative velocity is a function of the temperature of the system and the individual molecular masses. Further, it is shown that the probability of reaction R_μ occurring in V in the time interval $(t,t+\mathrm{d}t)$, given that the system is in state i at time t, has the form $h_\mu^i c_\mu\,\mathrm{d}t$. h_μ^i denotes the number of possible combinations of reactant molecules involved in reaction R_μ when the system is in state i. However, since the total number of hsp molecules that might co-exist in a cell is limited, no further hsp production (reaction (R_7)) should take place when the system is in any of the states in which the limit is reached. Let us denote all these states by $\mathcal{S}_K$, i.e.

$$\mathcal{S}_K=\{s|s_6+s_7+s_8=K\}.\tag{6}$$

Thus, the probability of reaction (R_7) occurring in V in the time interval $(t,t+\mathrm{d}t)$ when the system is in $i\in\mathcal{S}_K$ should be 0. Hence

$$h_7^i=\begin{cases}\text{No. of combinations of reactant molecules of } R_7, & i\in\mathcal{S}\backslash\mathcal{S}_K,\\ 0, & i\in\mathcal{S}_K.\end{cases}\tag{7}$$

Due to the fact that the reaction hazards depend only on the current state of the system, the time-evolution of the state of the reaction system can be regarded as a CTMC. Since the state space $\mathcal{S}$ is finite, Assumption 1 is fulfilled and hence the chain is a Markov jump process constructed as described above. The infinitesimal transition rates of the Markov jump process are

$$q_{ij}=h_\mu^i c_\mu,\tag{8}$$

where $j=i+v_\mu$.

Now, let us consider the *reaction probability density function* $P^i(\tau,\mu)\,\mathrm{d}\tau$ of the HSR system, which forms the basis for the Gillespie's simulation algorithm. It is defined by

$P^i(\tau,\mu)\,\mathrm{d}\tau\equiv$ probability at time t (when the system is in state $i\in\mathcal{S}$) that the next reaction in V will occur in the differential time interval $(t+\tau,t+\tau+\mathrm{d}\tau)$, and will be an R_μ reaction.

Table 2
The stoichiometry matrix of the heat shock response model.

$$\begin{bmatrix}
0&0&0&0&-1&1&0&0&0&0&0&1&0&0&0&0&0\\
-2&2&-1&1&0&0&0&-1&1&1&2&2&0&0&0&0&0\\
1&-1&-1&1&0&0&0&0&0&-1&0&0&0&0&0&0&0\\
0&0&1&-1&-1&1&0&0&0&0&-1&0&0&0&0&0&0\\
0&0&0&0&1&-1&0&0&0&0&0&-1&0&0&0&0&0\\
0&0&0&0&0&0&1&-1&1&-1&-1&-1&-1&0&-1&1&1\\
0&0&0&0&0&0&0&1&-1&1&1&1&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&-1&-1\\
0&0&0&0&0&0&0&0&0&0&0&0&0&1&-1&1&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&-1&0&0&1
\end{bmatrix}$$

The columns correspond to reactions (R_1)–(R_{17}) and the rows to metabolites in the order: hse, hsf, hsf_2, hsf_3, hsf_3: hse, hsp, hsp:hsf, hsp:mfp, mfp, prot.

As shown in Gillespie (1976), the density function can be expressed as

$$P^i(\tau,\mu)\,\mathrm{d}\tau = h^i_\mu c_\mu \cdot \exp\left[-\sum_{v=1}^{17} h^i_v c_v \tau\right]\mathrm{d}\tau. \tag{9}$$

The superscript $i \in \mathcal{S}$ indicates that in fact we deal with a whole family of such functions. Which of them is considered at time t depends on the state of the system at time t. In the continuation, in order to lighten the language, "probability at time t" will be a shorthand for "probability at time t when the system is in state i". Let $P^i_1(\tau)\,\mathrm{d}\tau$ denote the probability at time t that the next reaction will occur between times $t+\tau$ and $t+\tau+\mathrm{d}\tau$, irrespective of which reaction it might be. By the definition of $P^i(\tau,\mu)$ we have that

$$P^i_1(\tau) = \sum_{\mu=1}^{17} P^i(\tau,\mu) = \left(\sum_{k=1}^{17} h^i_k c_k\right) \cdot \exp\left[-\left(\sum_{k=1}^{17} h^i_k c_k\right) \cdot \tau\right]. \tag{10}$$

Hence, the sojourn-time rates v_i of the Markov jump process are given by

$$v_i = \sum_{k=1}^{17} h^i_k c_k. \tag{11}$$

The probability of the transition from state i to state $j = i + v_\mu$ of the Markov jump process (and, in consequence, of the embedded Markov chain $\{X_n\}$) is the probability at time t that the next reaction in V will be an R_μ reaction. Using Eq. (9), it can be expressed as

$$p_{ij} = \int_0^\infty P^i(\tau,\mu)\,\mathrm{d}\tau = \frac{h^i_\mu c_\mu}{\sum_{k=1}^{17} h^i_k c_k} \tag{12}$$

if $j \neq i$ and $p_{ii} = 0$ for all $i \in \mathcal{S}$.

Lemma 1. *The embedded Markov chain $\{X_n\}$ is irreducible.*

Before presenting the proof let us divide the species into two groups. The first one, called *elementary species group* (denoted by $G_{elementary}$), contains hsf, hse, hsp, mfp and prot species. The other one, denoted by $G_{compound}$ and named *compound species group*, is made of all the remaining species.

Proof of Lemma 1. Let i, j be any two states from the state space $\mathcal{S}$. By $i \rightsquigarrow j$ we denote that the state j is reachable from the state i, i.e. that there exists a sequence of reactions (R_1)–(R_{17}) which leads the system from the state i to the state j.

In order to prove that $\{X_n\}$ is irreducible it is enough to show that $i \rightsquigarrow j$, since i and j are two arbitrarily chosen states. Let hsp(k), mfp(k) and prot(k) be the total number of hsp, mfp and prot molecules present in the system when in the state k, respectively. Further, let

$$z = (C_1, C_2, 0, 0, 0, \mathrm{hsp}(i), 0, 0, \mathrm{mfp}(i), \mathrm{prot}(i))^T.$$

z is obtained from i by disassembling all compound species from $G_{compound}$. Thus, in the state z the number of molecules of any species from $G_{compound}$ is 0 and the number of molecules of any $s \in G_{elementary}$ is equal to the total number of s molecules in the system in the state i. Clearly $z \in \mathcal{S}$ and $i \rightsquigarrow z$, since for any $s \in G_{compound}$ there exists a sequence of reactions (R_1)–(R_{17}) which disassembles s into elements from $G_{elementary}$.

Let

$$z' = (C_1, C_2, 0, 0, 0, \mathrm{hsp}(j), 0, 0, \mathrm{mfp}(i), \mathrm{prot}(i))^T$$

and

$$z'' = (C_1, C_2, 0, 0, 0, \mathrm{hsp}(j), 0, 0, \mathrm{mfp}(j), \mathrm{prot}(j))^T.$$

$z' \in \mathcal{S}$ and we continue to show that $z \rightsquigarrow z'$. There are three cases: hsp(i) = hsp(j), hsp(j) < hsp(i) or hsp(j) > hsp(i). In the first case $z = z'$ and trivially $z \rightsquigarrow z'$. If hsp(j) < hsp(i), z' can be reached from z by applying reaction (R_{13}) hsp(i) − hsp(j) times. If finally hsp(j) > hsp(i), first hsf_3:hse is produced (this is doable since $C_1 \geq 1$ and $C_2 \geq 3$). Next, by applying reaction (R_7) hsp(j) − hsp(i) times the required number of additional hsp molecules is produced. Finally, hsf_3:hse is disassembled by applying a sequence of reactions $\langle (R_6), (R_3), (R_2) \rangle$. Hence $z \rightsquigarrow z'$.

We continue to show that $z' \rightsquigarrow z''$. There are two cases. Either mfp(i) > mfp(j) or prot(i) ≥ prot(j) since mfp(k) + prot(k) = C_3 for any state $k \in \mathcal{S}$. In the first case, if hsp(j) = 0, first one hsp molecule is produced by applying reaction sequence $\langle (R_1), (R_3), (R_5), (R_7), (R_6), (R_4), (R_2) \rangle$, which leads to state $(C_1, C_2, 0, 0, 0, 1, 0, 0, \mathrm{mfp}(i), \mathrm{prot}(i)^T)$. Then, by applying reaction sequence $\langle (R_{15}), (R_{17}) \rangle$ mfp(i) − mfp(j) times the system reaches state:

$$(C_1, C_2, 0, 0, 0, 1, 0, 0, \mathrm{mfp}(j), \mathrm{prot}(j))^T$$

since, as mentioned before, mfp(k) + prot(k) = C_3 for any state $k \in \mathcal{S}$. Finally, the only hsp molecule is degraded by applying reaction (R_{13}) and the system arrives in state z''. If hsp(j) is greater than 0, state z'' can be reached by less steps since neither production nor degradation of the one additional hsp molecule is required.

In the second case, when prot(i) ≥ prot(j), state z'' can be reached by applying misfolding reaction (R_{14}) prot(i) − prot(j) times. Hence $z' \rightsquigarrow z''$.

At last $z'' \rightsquigarrow j$. State j is reached by producing the appropriate numbers of molecules of all compound species. Since in state z'' the required number of molecules of all elementary species is already present, by applying appropriate reactions all compound species molecules can be produced. The mass-conservation law ensures that the numbers of elementary species molecules will be decreased appropriately and that the system will reach state j. Hence $\{X_n\}$ is irreducible. □

The irreducibility of the embedded chain $\{X_n\}$ implies the irreducibility of the continuous-time Markov chain $\{X(t)\}$. Since the state space $\mathcal{S}$ is finite, it follows that the CTMC $\{X(t)\}$ is positive recurrent. In consequence, it has an invariant measure η which is unique up to multiplicative factors and can be found as the solution of the equation $\eta^T \mathbf{Q} = 0$. Moreover, $\sum_{i \in \mathcal{S}} \eta_i < \infty$ since $\mathcal{S}$ is finite and there exists a unique stationary distribution π of $\{X(t)\}$ given by

$$\pi = \left(\frac{\eta_i}{\sum_{k \in \mathcal{S}} \eta_k}\right)_{i \in \mathcal{S}}. \tag{13}$$

For the theoretical details we refer the reader to, e.g., Norris (1998) and Resnick (1992).

4. Results and discussion

The deterministic approach, based on the law of mass action, yields a system of ordinary differential equations for molecular concentrations. In consequence, the biochemical system is modelled as being continuous. But such description does not capture effects that occur due to either the discreteness of molecular quantities or the stochastic nature of chemical reactions (McAdams and Arkin, 1999; Pahle, 2009; Sandmann, 2008; Wilkinson, 2006). As discussed in Section 1, random fluctuations may have a significant impact on the reaction dynamics, especially as the numbers of molecules of some reactants become

Table 3
The numerical values of the parameters and the initial numbers of molecules in the stochastic model.

Param.	Reaction	Value	Unit	Metabolite	Init. no.
k_1^+	(R_1)	6.98	$V/(\#\,s)$	hsf	0
k_1^-	(R_2)	0.19	s^{-1}	hsf_2	0
k_2^+	(R_3)	1.07	$V/(\#\,s)$	hsf_3	0
k_2^-	(R_4)	10^{-9}	s^{-1}	hse	29
k_3^+	(R_5)	0.17	$V/(\#\,s)$	hsf_3:hse	2
k_3^-	(R_6)	1.21×10^{-6}	s^{-1}	hsp	766
k_4	(R_7)	8.3×10^{-3}	s^{-1}	hsp:hsf	1403
k_5^+	(R_8)	9.74	$V/(\#\,s)$	mfp	517
k_5^-	(R_9)	3.56	s^{-1}	hsp:mfp	71
k_6	(R_{10})	2.33	$V/(\#\,s)$	prot	1.15×10^{8}
k_7	(R_{11})	4.31×10^{-5}	$V/(\#\,s)$		
k_8	(R_{12})	2.73×10^{-7}	$V/(\#\,s)$		
k_9	(R_{13})	3.2×10^{-5}	s^{-1}		
k_{10}	(R_{14})	$\varphi(42)=7.77\times10^{-5}$	s^{-1}		
k_{11}^+	(R_{15})	3.32×10^{-3}	$V/(\#\,s)$		
k_{11}^-	(R_{16})	4.44	s^{-1}		
k_{12}	(R_{17})	13.94	s^{-1}		

The numerical quantities are obtained by adopting the corresponding values in Petre et al. (2009a): the initial numbers of molecules are truncated to natural numbers, the value of the rate constant k_1^+ is twice the value of the corresponding deterministic rate constant. # denotes the number of molecules, V is the cell volume and s is the second.

smaller (McAdams and Arkin, 1999; Srivastava et al., 2002). This is the case of the deterministic heat shock response model being discussed: except for prot, hsp, hsp:hsf and mfp, all the other species have very small initial number of molecules (Table 3) and, as can be seen from the continuous simulation results, stay at the low level throughout the time of simulation. This might be the main objection to the continuous approach applied in Petre et al. (2009a, b). Since the stochastic modelling seems more reasonable in this case, we made the effort to run 1000 stochastic simulations in order to check whether the dynamics of the continuous description agrees qualitatively with the behaviour demonstrated by the discrete system. The results of 1000 independent stochastic simulation runs (blue and green points) for five species: hsf_3:hse, hsp, mfp, hsp:hsf and hsp:mfp, overlaid with the deterministic outcome (yellow line) are shown in Fig. 1. The mean together with the mean $\pm$ standard deviation are shown in Fig. 2. The ratios of the sample standard deviation to the sample mean were computed for the five considered species and are depicted in Fig. 3. According to Gillespie (1976), since the ratios are small (less than 0.12 in the case of mfp and hsp, see Fig. 3c and b) and very small (less than 0.035 for hsp:mfp and less than 0.007 for hsp:hsf, see Fig. 3e and d, respectively), the results of independent runs of the system are expected not to vary much and the presented outcomes of 1000 stochastic simulations together with the estimated mean should provide a statistically adequate picture of the evolution of the chemical system in time. One might argue that the ratio for hsf_3:hse is, however, quite big: it peaks at about 1.1 and stabilises below 0.6 (see Fig. 3a). In this particular case the mean converges to approximately 3 molecules of hsf_3:hse and the standard deviation is around 1.6, which all in all gives a narrow range of possible values of molecule number and hence this result can be accepted.

We first investigated the number of steady-states of the deterministic model. Since our attempts to analytically solve the algebraic system of steady state equations obtained from the differential ones did not bring any results, we performed some numerical investigations. We randomly chose 10 000 sets of initial particle numbers for the continuous model from a wide range of values, but in such a way that the total amounts of hse, hsf and proteins in the resulting system would always be the same as in the case of the original deterministic model presented in Petre et al. (2009a). For each of these sets we run numerical time-course simulations and waited for the considered system to stabilise. In all these cases the systems converged to exactly the same state as the original model, i.e. no other steady states were found by this method. Additionally, bifurcation analysis performed with the AUTO software (XPPAUT was used as the front-end, Doedel et al., 1997; Ermentrout, 2002) with respect to parameter values did not reveal multistationarity (data not shown). These results suggest that the heat shock response mechanism is rather monostable.

Next, we were interested in investigating the range of behaviour the stochastic model was likely to exhibit. As shown in Section 3, there exists only one stationary limit distribution π given by Eq. (13), which governs the transitions of the Markov jump process when the number of iterations goes to infinity. In particular, we analysed the unimodality of the hsp level by computing some appropriate statistics from the performed 1000 stochastic realisations.

First, we computed the median $m(t)$ of the 1000 stochastic realisations on the time interval $T=\{130\,000\,\text{s}, \ldots, 150\,000\,\text{s}\}$. It is depicted in Fig. 4 as the middle black line. The upper and lower black lines are $m(t) \pm \frac{1}{4}\cdot s$, respectively, where s is the range of dynamics the model exhibits in the 1000 realisations on the considered time interval, i.e.

$$s = \max_{t\in T, i\in I}\{r_i(t)\} - \min_{t\in T, i\in I}\{r_i(t)\},$$

where $I=\{1,\ldots,1000\}$ and r_i is the i-th realisation. The mean (brown line) basically coincides with the median on the whole time interval.

Next, in order to check whether the realisations r_i, $i = 1,\ldots,1000$, can be divided into subgroups such that the means of the subgroups would differ significantly from each other, we applied the following procedure. We defined two subsets:

$$S_U = \left\{r_i : \forall_{t\in T} r_i(t) > m(t) - \frac{s}{4} \wedge \exists_{t\in T} r_i(t) > m(t) + \frac{s}{4}\right\}$$

and

$$S_L = \left\{r_i : \forall_{t\in T} r_i(t) < m(t) + \frac{s}{4} \wedge \exists_{t\in T} r_i(t) < m(t) - \frac{s}{4}\right\}.$$

In our case, there are 253 realisations in S_U and 189 in S_L. The means computed from the realisations of each of these subsets are depicted by red lines in Fig. 4. The means are close to the global mean on the whole considered time interval and since the numbers of elements in the S_U and S_L subsets are rather small, i.e. approximately $\frac{1}{4}$ and $\frac{1}{5}$ of all the 1000 considered realisations, this result does not indicate any significant split.

Further, a clustering algorithm was applied in order to determine whether some subsets of realisations could be isolated and the computed means would point to potential multimodality. To this aim, we utilised the *Agnes* algorithm (implementation of an agglomerative hierarchical clustering method, Kaufman and Rousseeuw, 1990) with the *manhattan* metric, i.e. the distance between two realisations r_i and r_j is defined as $d(r_i,r_j)=\sum_{t\in T}|r_i(t)-r_j(t)|$, thus the realisations are treated as points in a $|T|$-dimensional space. By applying this metric the characteristics of the realisations on the whole considered time interval are taken into account, hence they are compared in a "global" sense. The obtained dendrogram is presented in Fig. 5. The *agglomerative coefficient* (AC), which measures the clustering structure of the dataset, is 0.82. This indicates that the clustering algorithm did find some rather clear structuring.[1] We isolated two

[1] AC is a dimensionless quantity, varying between 0 and 1 – AC close to 1 shows that a very clear structure has been found, while value 0 implies that the data consist of only one big cluster, see e.g. Kaufman and Rousseeuw (1990) for details.

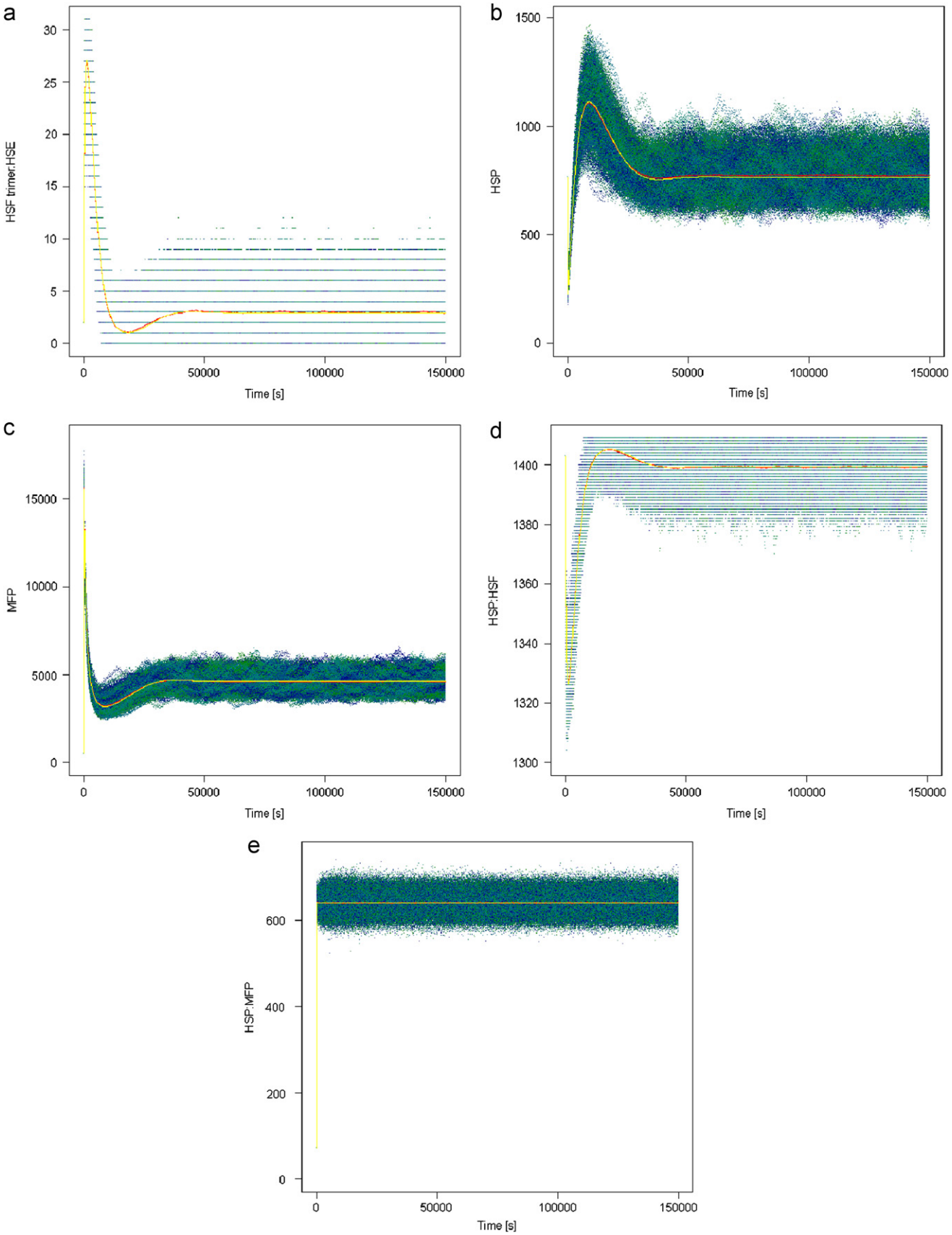


Fig. 1. Results of 1000 independent discrete stochastic simulation runs. The trajectories of individual realisations are plotted with blue and green points (each run with separate shade). The red points show the average taken over all runs and the yellow line is the outcome of the continuous deterministic simulation: (a) hsf_3:hse, (b) hsp, (c) mfp, (d) hsp:hsf, (e) hsp:mfp. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

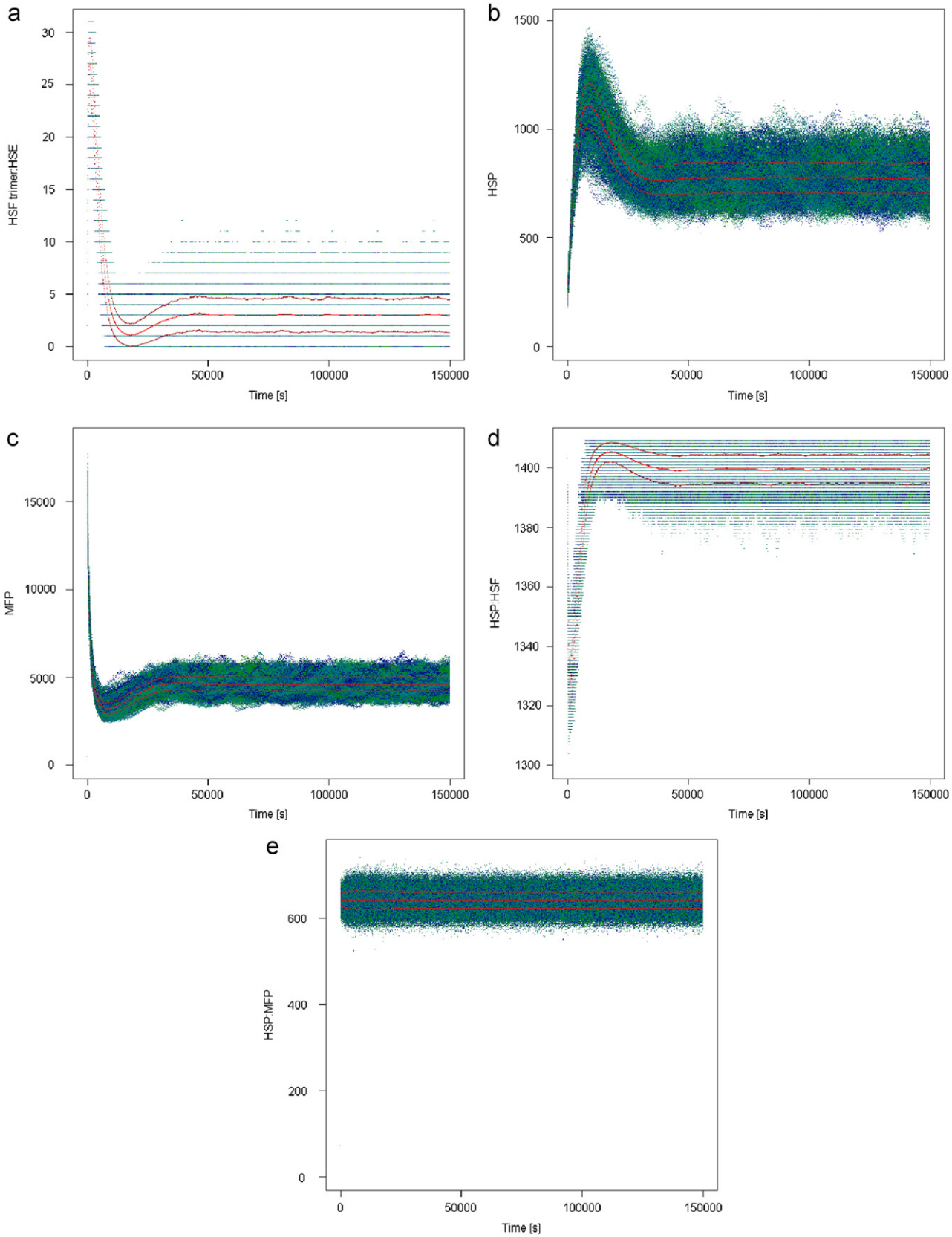


Fig. 2. The mean taken over the outcome of 1000 independent stochastic simulations of the system (red points) and the mean $\pm$ standard deviation (upper/lower brown points): (a) hsf_3:hse, (b) hsp, (c) mfp, (d) hsp:hsf, (e) hsp:mfp. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

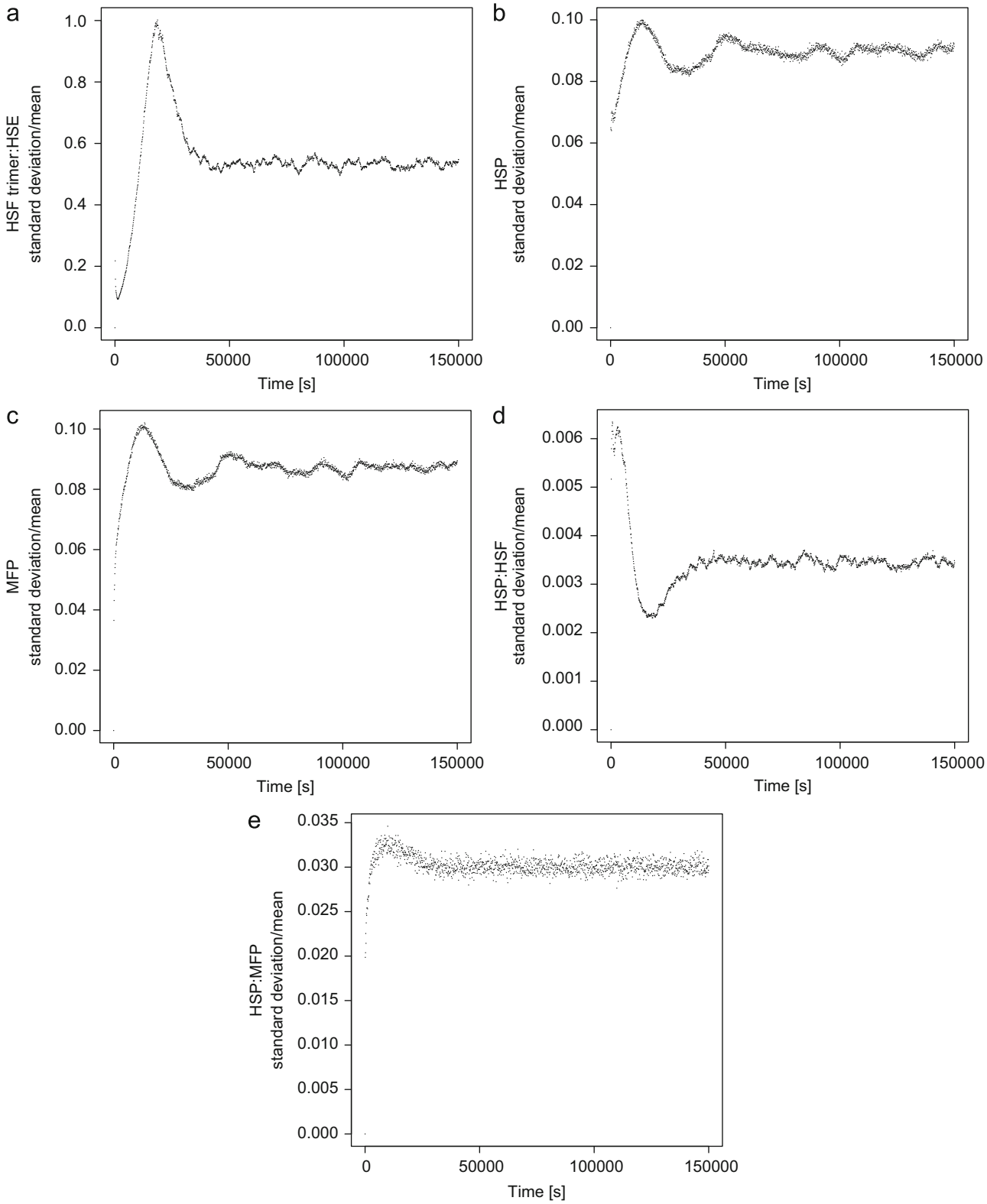


Fig. 3. The ratios of the standard deviation to the sample mean at each considered time point: (a) hsf_3:hse, (b) hsp, (c) mfp, (d) hsp:hsf, (e) hsp:mfp.

groups of realisations that stand out on the obtained dendrogram. They are marked in Fig. 5 by two rectangles which enclose the dendrogram branches constituting these groups. The two resulting subclusters are at almost the same height in the clustering tree. The means of the stochastic realisations belonging to these two groups at time point $t=150\,000$ s are 757 (left subcluster) and 794 (right subcluster). Although the agglomerative coefficient indicates some clustering structure of the realisations, the mean values are very close to each other and agree well with the steady state value of the deterministic model (767).

Finally, as suggested in Wilkinson (2006), we investigated the empirical probability mass function by drawing histograms of the realisations at some time point in the considered time interval T.

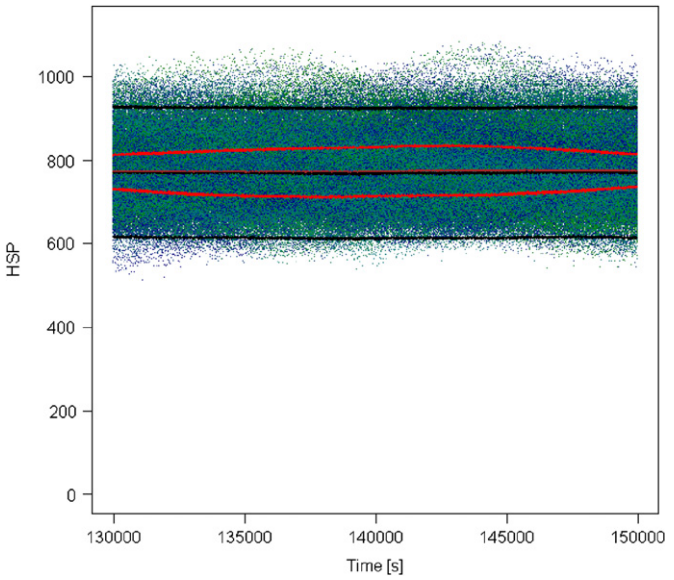


Fig. 4. The median of the 1000 realisations on the time interval $T=\{130\,000$ s,…,150 000 s$\}$ (middle black line). The upper and lower black lines are the median $\pm\frac{1}{4}$ of the range of dynamics the model exhibits in the 1000 realisations on the considered time interval. The mean of all the realisations, of the subset S_U and S_L plotted with brown, upper red and lower red lines, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

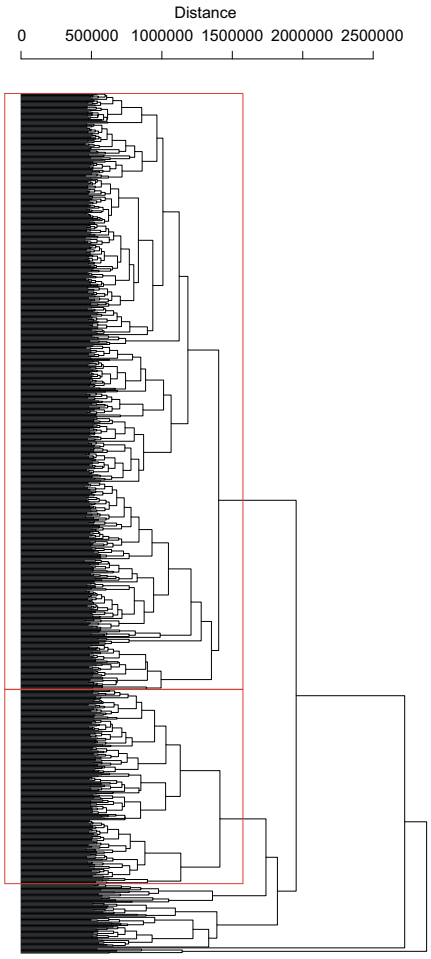


Fig. 5. The clustering tree (dendrogram) obtained with the *Agnes* clustering algorithm with the *average* method and the *manhattan* metric applied to the 1000 stochastic realisations considered on the time interval 130 000–150 000 s. The leaves of the clustering tree are the original realisations. Two branches come together at the distance between the two clusters being merged. The agglomerative coefficient equals 0.82. The rectangles distinguish two subclusters discussed in Section 4.

Fig. 6 shows the histograms overlaid with the normal distribution curve with mean and standard deviation computed from all 1000 realisations at time point $t=150\,000$ s. In the case of Fig. 6a, where the bin width is set to 20, the obtained results indicate that the distribution is unimodal. Changing the bin size to 10 (Fig. 6b) does not change the picture significantly.

Although due to the small particle numbers of some of the reagents the stochastic modelling is more reasonable, the presented results do not reveal any qualitative discrepancy in the dynamics of the two considered models of the heat shock response. The range of behaviour the stochastic model is likely to exhibit, which can be observed based on the performed 1000 simulations, confirms the dynamics of the continuous model. The performed analysis of the stochastic realisations does not reveal any clear signs of multistationarity of the HSR mechanism. Although unimodality of a stationary probability density function does not necessarily imply the uniqueness of the stable steady state of the deterministic approximation (as well as bimodality does not determine the existence of bistability, etc.), usually this is the case and to this extent the stochastic results agree with the deterministic outcomes indicating that there exists only one stable steady state. This shows that the approximation of a discrete system with a continuous model is valid and strengthens the trust in the deterministic description. Additionally, the presented stochastic formulation, together with the performed analysis of its behaviour and comparison to the continuous description, let us gain more insight into the dynamics of the HSR mechanism, especially in respect of the number of steady states, which, as discussed previously, is important from a biological point of view.

5. Conclusions and further research

In this paper we presented a stochastic model associated with a previously described (Petre et al., 2009a) model of the heat shock response in eukaryotic cells. The stochastic model was viewed as a Markov jump process and the existence and uniqueness of the stationary distribution was shown. Further, the model was compared to the deterministic description of heat shock response (Petre et al., 2009a). The aim with the comparison was to show that in this particular case the approximation of a discrete system with a continuous model is reasonable. This is not true in general, especially when the numbers of metabolites in the considered biochemical system are small. The presented results indicate that the stochastic and deterministic models provide a qualitatively consistent picture of the dynamics of the heat shock response mechanism. Additionally, the development of the stochastic model and the effort of performing 1000 stochastic simulations enabled gaining some more information about the dynamics of the heat shock response. The outcomes of the analysis of the stochastic realisations lead towards the conclusion that the heat shock response mechanism is a rather monostable system. Moreover, this is in agreement with the results of the analysis performed on the deterministic model. All in all, the presented results strengthen the trust in the deterministic description of the HSR mechanism in eukaryotic cells proposed in Petre et al. (2009a).

Although it was shown in Section 3 that the Markov jump process has a unique stationary distribution, there is no certainty

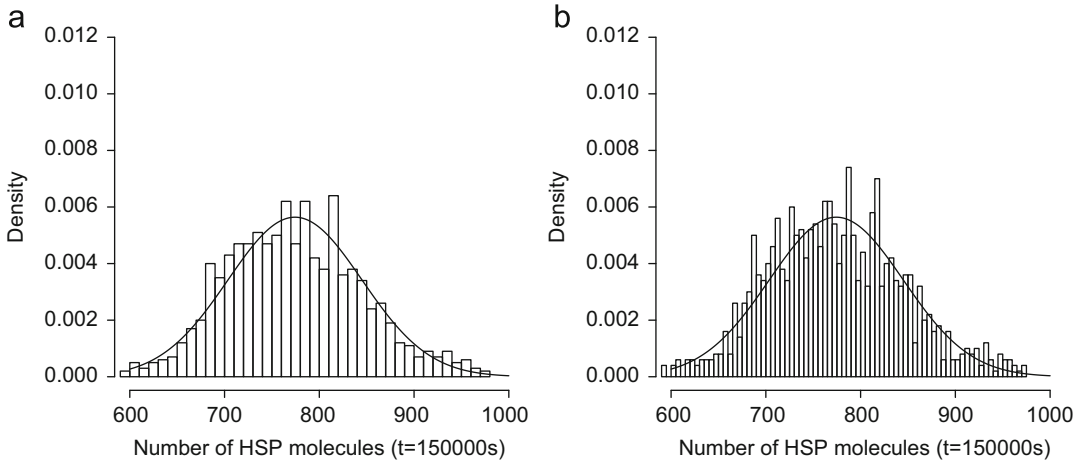


Fig. 6. Histograms overlaid with the normal distribution curve with mean and standard deviation computed at time point $t=150\,000$ s from all 1000 realisations: (a) bin width set to 20, (b) bin width set to 10.

that it was reached already in the considered time interval $T=\{130\,000\text{ s}, \ldots, 150\,000\text{ s}\}$. It was chosen based on the results of many stochastic simulations, which suggest that the process stabilises relatively long before the time point $t=130\,000$ s. Nevertheless, some assessment of the convergence to the stationary distribution in this case would be desired. One of possible approaches is to measure the rate of convergence by the *mixing time* (Sinclair, 1992). For ergodic Markov chains the rate is governed by the second largest eigenvalue in absolute value λ_2, in particular the *spectral gap* $1-\lambda_2$ is both a necessary and sufficient condition for rapid mixing, see Sinclair (1992) for details. The problem of determining λ_2 of the presented Markov chain underlying the stochastic model of heat shock response is subject of further research.

The rate constant values for the presented stochastic model were obtained from the corresponding values of the deterministic model presented in Petre et al. (2009b), which in turn were fitted to available experimental data. As suggested in Wilkinson (2006), another way of deducing the rate constant values for the stochastic model could utilise methods that are based on Bayesian inference and take advantage of Markov chain Monte Carlo (MCMC) algorithms such as the Metropolis–Hastings algorithm or the Gibbs Sampler. However, such methods demand high-quality, calibrated, high-resolution time-course measurements for a reasonably large subset of model metabolites (Wilkinson, 2006). Unfortunately, experimental data of such quality are still seldom if ever available and make a challenge for experimental biology.

Acknowledgements

Both deterministic and stochastic models were implemented and run in Copasi, a software application for simulation and analysis of biochemical networks (Hoops et al., 2006). The stochastic simulations were performed using the Gibson and Bruck (1998) algorithm. The obtained time-course data were analysed and plotted in R, a software environment for statistical computing and graphics (R Development Core Team, 2008).

The authors give special thanks to Anna Gambin from the Faculty of Mathematics, Informatics and Mechanics, University of Warsaw for helpful advice as well as thorough and detailed comments.

Andrzej Mizera would like to express his gratefulness to Jan Westerholm, Mats Aspnäs and Evren Yurtesen from the Department of Information Technologies, Åbo Akademi University for making available their computational resources and providing essential help for performing the stochastic simulations. The author would like to thank Ion Petre from the Computational Biomodelling Laboratory, Åbo Akademi University for invaluable discussions.

The authors were partially supported by Ministry of Science and Higher Education (Grant number N N518 426936).

References

Balch, W.E., Morimoto, R.I., Dillin, A., Kelly, J.W., 2008. Adapting proteostasis for disease intervention. Science 319, 916–919.

Chen, Y., Voegli, T., Liu, P., Noble, E., Currie, R., 2007. Heat shock paradox and a new role of heat shock proteins and their receptors as anti-inflammation targets. Inflammation and Allergy Drug Targets 6 (2), 91–100.

Doedel, E.J., Champneys, A.R., Fairgrieve, T.F., Kuznetsov, Y.A., Sandstede, B., Wang, X., 1997. AUTO 97, software for continuation and bifurcation problems in ordinary differential equations. Technical Report, Concordia University, Montreal, Canada.

Donati, Y., Slosman, D., Polla, B., 1990. Oxidative injury and the heat shock response. Biochemical Pharmacology 40, 2571–2577.

Ermentrout, B., 2002. Simulating, Analyzing, and Animating Dynamical Systems: A Guide to XPPAUT for Researchers and Students. Software, Environment and Tools, vol. 14. SIAM, Philadelphia, USA.

Gibson, M., Bruck, J., 1998. An efficient algorithm for generating trajectories of stochastic gene regulation reactions. Technical Report, California Institute of Technology.

Gillespie, D.T., 1976. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. Journal of Computational Physics 22, 403–434.

Hoops, S., Sahle, S., Gauges, R., Lee, C., Pahle, J., Simus, N., Singhal, M., Xu, L., Mendes, P., Kummer, U., 2006. Copasi—a COmplex PAthway SImulator. Bioinformatics 22 (24), 3067–3074.

Jones, C.M., Henry, E.R., Hu, Y., Chan, C.K., Luck, S.D., Bhuyan, A., Roder, H., Hofrichter, J., Eaton, W.A., 1993. Fast events in protein folding initiated by nanosecond laser photolysis. Proceedings of the National Academy of Sciences of the United States of America 90, 11860–11864.

Kampinga, H.K., 1993. Thermotolerance in mammalian cells: protein denaturation and aggregation, and stress proteins. Journal of Cell Science 104, 11–17.

Karmakar, R., Bose, I., 2007. Positive feedback, stochasticity and genetic competence. Physical Biology 4, 29–37.

Kaufman, L., Rousseeuw, P.J., 1990. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons, New York, USA.

Lepock, J.R., Frey, H.E., Ritchie, K.P., 1993. Protein denaturation in intact hepatocytes and isolated cellular organelles during heat shock. The Journal of Cell Biology 122 (6), 1267–1276.

Lepock, J.R., Frey, H.E., Rodahl, A.M., Kruuv, J., 1988. Thermal analysis of chl v79 cells using differential scanning calorimetry: implications for hyperthermic cell killing and the heat shock response. Journal of Cellular Physiology 137 (1), 14–24.

Lindquist, S., Craig, E.A., 1988. The heat-shock proteins. Annual Review of Genetics 22, 631–677.

Lipniacki, T., Hat, B., Faeder, J.R., Hlavacek, W.S., 2008. Stochastic effects and bistability in T cell receptor signaling. Journal of Theoretical Biology 254 (1), 110–122.

Liu, B., DeFilippo, A.M., Li, Z., 2002. Overcoming immune tolerance to cancer by heat shock protein vaccines. Molecular Cancer Therapeutics 1, 1147–1151.

Lukacs, K.V., Pardo, O.E., Colston, M., Geddes, D.M., Alton, E.W., 2000. Heat shock proteins in cancer therapy. In: Habib, N.A. (Ed.), Cancer Gene Therapy: Past Achievements and Future Challenges. Kluwer, New York, USA, pp. 363–368.

McAdams, H.H., Arkin, A., 1999. It's a noisy business! Genetic regulation at the nanomolar scale. Trends in Genetics 15 (2), 65–69.
Morimoto, R., 2008. Proteotoxic stress and inducible chaperone networks in neurodegenerative disease and aging. Genes and Development 22, 1427–1438.
Norris, J.R., 1998. Markov chains. In: Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, New York, USA.
Pahle, J., 2009. Biochemical simulations: stochastic, approximate stochastic and hybrid approaches. Briefings in Bioinformatics 10 (1), 53–64.
Parsell, D., Lindquist, S., 1993. The function of heat-shock proteins in stress tolerance: degradation and reactivation of damaged proteins. Annual Review of Genetics 27, 437–496.
Peper, A., Grimbergent, C., Spaan, J., Souren, J., van Wijk, R., 1997. A mathematical model of the hsp70 regulation in the cell. International Journal of Hyperthermia 14, 97–124.
Petre, I., Mizera, A., Back, R.-J., 2009a. Computational heuristics for simplifying a biological model. In: Ambos-Spies, K., Löwe, B., Merkle, W. (Eds.), Mathematical Theory and Computational Practice: 5th Conference on Computability in Europe, CiE 2009, Heidelberg, Germany, July 19–24, 2009, Proceedings. Lecture Notes in Computer Science, vol. 5635. Springer, Heidelberg, Germany, pp. 399–408.
Petre, I., Mizera, A., Hyder, C.L., Mikhailov, A., Eriksson, J.E., Sistonen, L., Back, R.-J., 2009b. A new mathematical model for the heat shock response. In: Condon, A., Harel, D., Kok, J.N., Salomaa, A., Winfree, E. (Eds.), Algorithmic Bioprocesses. Natural Computing Series. Springer, Heidelberg, Germany, pp. 411–425.
Pockley, A., 2003. Heat shock proteins as regulators of the immune response. The Lancet 362 (9382), 469–476.
Powers, M., Workman, P., 2007. Inhibitors of the heat shock response: biology and pharmacology. FEBS Letters 581 (19), 3758–3769.
R Development Core Team, 2008. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0. 〈http://www.R-project.org〉.
Remondini, D., Bernardini, C., Forni, M., Bersani, F., Castellani, G., Bacci, M., 2006. Induced metastable memory in heat shock response. Journal of Biological Physics 32, 49–59.
Resnick, S., 1992. Adventures in Stochastic Processes. Birkhäuser, Boston, USA.
Rieger, T.R., Morimoto, R.I., Hatzimanikatis, V., 2005. Mathematical modeling of the eukaryotic heat shock response: dynamics of the hsp70 promoter. Biophysical Journal 88 (3), 1646–1658.
Sandmann, W., 2008. Discrete-time stochastic modeling and simulation of biochemical networks. Computational Biology and Chemistry 32, 292–297.
Sinclair, A., 1992. Improved bounds for mixing rates of Markov chains and multicommodity flow. Combinatorics, Probability and Computing 1, 351–370.
Srivastava, R., You, L., Summers, J., Yin, J., 2002. Stochastic vs. deterministic modeling of intracellular viral kinetics. Journal of Theoretical Biology 218, 309–321.
Szymańska, Z., Żylicz, M., 2009. Mathematical modeling of heat shock protein synthesis in response to temperature change. Journal of Theoretical Biology 259, 562–569.
Tijms, H.C., 2003. A First Course in Stochastic Models. John Wiley & Sons, West Sussex, UK.
Voellmy, R., Boellmann, F., 2007. Chaperone regulation of the heat shock protein response. Advances in Experimental Medicine and Biology 594, 89–99.
Wilkinson, D.J., 2006. Stochastic modelling for systems biology. In: Mathematical and Computational Biology. Chapman & Hall, CRC, London, UK.
Workman, P., de Billy, E., 2007. Putting the heat on cancer. Nature Medicine 13 (12), 1415–1417.

Paper V

A Boolean approach for disentangling the numerical contribution of modules to the system-level behavior of a biomodel

Elena Czeizler, Andrzej Mizera, and Ion Petre

TUCS Technical Report number 997, January 2011.

A Boolean approach for disentangling the numerical contribution of modules to the system-level behavior of a biomodel

Elena Czeizler
Department of Information Technologies, Åbo Akademi University,
FI-20520 Turku, Finland
University of Helsinki, Computational Systems Biology group, Biomedicum,
FI-00014 Helsinki, Finland
elena.czeizler@helsinki.fi

Andrzej Mizera
Department of Information Technologies, Åbo Akademi University,
FI-20520 Turku, Finland
amizera@abo.fi

Ion Petre
Department of Information Technologies, Åbo Akademi University,
FI-20520 Turku, Finland
ipetre@abo.fi

TUCS Technical Report
No 997, January 2011

Abstract

To disentangle the numerical contribution of modules to the system-level behavior of a given biomodel, one often considers knockdown mutant models, investigating the change in the model behavior when modules are systematically included and excluded from the model architecture in all possible ways. We propose in this paper a Boolean logic-based approach for extracting conclusions about the role of each module from the systematic comparison of the numerical behavior of all knockdown mutants. We associate a Boolean variable to each module, expressing when the module is included in the architecture (value 'true') and when it is not (value 'false'). We can then express the satisfiability of system-level properties of the full model, such as efficiency, or economical use of resources, in terms of a Boolean formula expressing in a compact way which model architectures, i.e., which combinations of modules, give rise to the desired property. We apply this methodology on a recently proposed model for the heat shock response in eukaryotes. We describe the contribution of each of its three feedback loops towards achieving an economical and effective heat shock response.

Keywords: mathematical model — modularization — Boolean logic — heat shock response

TUCS Laboratory
Computational Biomodelling

1 Introduction

Modularization of biomodels. There is a sustained experimental and computational effort nowadays towards building large, system-level models for biochemical processes, including regulatory networks, signaling pathways, metabolic pathways, etc. Models can encompass thousands of reactants and reactions, see [1]. On this scale, understanding the details of the network, especially its regulatory mechanisms, becomes a considerable challenge. Similar problems have also been encountered in engineering (and elsewhere), see [2]. Thus, one strategy for elucidating the structure of a biological system, is to adapt to systems biology methods coming from engineering sciences, in particular from control theory, [3], [4], [5], [6]. Applying a control-theoretical analysis to a biological system can provide a systematic way to identify the main regulatory components of a biological system, including its feedforward and feedback mechanisms, see [7]. This, in turn, contributes to the understanding of the reactivity, robustness and efficiency of the biological system. To disentangle the individual contribution of the various components to the network, knockdown mutants are often useful to consider, see [7]. The mutants are numerically compared to each other and to the reference model in an effort to extract the individual contribution of each mechanism to the overall behavior of the system.

Our approach for comparing knockdown mutant models. We propose in this paper a novel approach for identifying the numerical contribution of a component to the system-level behavior of a larger model. The core technique we use throughout the paper is to associate a Boolean variable to each of the components of the model. For each knockdown mutant we write a Boolean formula describing the presence or the absence of each component (using the conjunction and the negation of Boolean variables). The obtained Boolean formulas encompass properties of the architecture of the knockdown mutant models. Going one step further, we can also write a Boolean formula characterizing all mutant architectures that exhibit a given property: we select all knockdown mutants that exhibit that property and construct the disjunction of their Boolean formulas. The formula thus obtained describes which components must be present/absent and in which configurations in order for the system to exhibit the desired property. Iterating this technique for several well chosen systemic properties may help to identify (at least qualitatively) the roles of each model component.

Our approach is essentially different from the Boolean network framework often used for qualitative modeling and analysis of biological systems, see, e.g., [8], [9], [10], [11], [12], [13]. The Boolean network model was first introduced by Kauffman, see [14] and [10], as a way to investigate the qualitative properties of a continuous biochemical regulatory network which depend on the logical structure of that particular network and not on the parameters used to describe it. In this framework, one usually associates to each species a Boolean variable, which as-

sumes the value 1 if that particular species is active, i.e., its activity is biologically detectable, or 0 otherwise. Moreover, using Boolean functions one can update the values of the variables at any time point, depending on the values from the previous time step. Depending on how time is handled in the model, there are two main paradigms for such Boolean simulations: synchronous, where at each time step all the variables are updated simultaneously, and asynchronous, where at each time step we update only one variable.

Case study: The eukaryotic heat shock response. The heat shock response (HSR) is an evolutionary-conserved global regulatory network found in virtually all living cells. It allows the cell to quickly react to elevated temperatures by the induction of some proteins called *heat shock proteins* (hsp). Exposure to raised temperature leads to protein misfolding, which then accumulate and form aggregates with disastrous effect for the cell. Stress conditions can be caused not only by increased temperature but also by other forms of environmental, chemical or physical stress, such as addition of ethanol, heavy metals, pollutants, high osmolarity, starvation, etc. The heat shock proteins act as chaperones – they stabilize proteins and help to refold the denatured ones. They maintain the proper functioning of the cell by preventing the formation of cytotoxic aggregates.

The heat shock response has been subject to intense scrutiny, see e.g., [15], [16], [17], for at least two main reasons. First, as a primordial, very well-conserved mechanism it is considered a promising candidate for providing insight into the design principles of regulatory networks in general, see e.g., [18], [7]. Second, the heat shock proteins, which are the main actors of the HSR, play crucial role also in many other fundamental cellular processes, see e.g., [19], [20]. We use as a case study in this paper a model for the heat shock response introduced in [21]. We take a control-based approach to identify three feedback mechanisms in this model. We then apply our Boolean approach for knockdown mutant comparison to identify the contribution of each of the three feedbacks to having a response where the level of misfolded proteins remains low, with a relatively low cost in terms of transactivating the heat shock protein genes.

2 Models

The eukaryotic heat shock response: a molecular model. The central role in the heat shock response is played by the *heat shock proteins* (hsp), which act as chaperones for the *misfolded proteins* (mfp) by forming hsp: mfp complexes and helping them to refold. In the model presented in [21], the regulation of the heat shock response is done by controlling the transactivation of the hsp-encoding genes. The transcription of these genes is initiated by some specific proteins called *heat shock factors* (hsf) that first dimerize (hsf_2), then trimerize (hsf_3) to finally bind to the promoters of the hsp-encoding genes, called *heat shock elements* (hse).

After the trimers bind to the promoter sites (hsf_3: hse) the transcription and translation of the hsp-encoding genes starts, ultimately producing new hsp molecules.

Once the level of hsp molecules is high enough, the transcription process is turned off through a self-regulating mechanism. The hsp molecules sequestrate the heat shock factors (hsp: hsf), thus preventing them to trimerize and bind to the heat shock elements. The sequestration of the heat shock factors by the heat shock proteins can be done in three different ways: by binding to free hsf, by breaking dimers and trimers, and by unbinding hsf_3 from the DNA promoter sites with simultaneous breaking of the trimer. Once the temperature increases, some of the proteins (prot) start to misfold, driving hsp away from hsf. Thus, the heat shock response is quickly switched on since the heat shock factors are again free and able to promote the synthesis of more heat shock proteins. The reaction rules of the molecular model introduced in [21] are presented in Table 1.

Clearly, the model in Table 1 is very generic in nature. For instance, the protein synthesis and degradation (i.e., reactions 4 and 9) are greatly simplified. Also, although there exist several types of slightly different heat shock proteins, see [22], here they are all treated uniformly, with hsp 70 as base denominator. This is also the case for the heat shock factors and the heat shock elements. Furthermore, in this model all proteins are treated generically, through the prism of whether they are properly folded (prot), or misfolded (mfp). Nevertheless, the model is well suited for the purpose of demonstrating our method for knockdown mutant analysis: the formal results of the analysis can be easily related to an intuitive understanding of the model.

The model in Table 1 includes three mass conservation relations, see [21], for the total amount of hsf, the total amount of proteins (other than hsp and hsf) in the model, as well as for the total amount of hse:

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

for some mass constants C_1, C_2, C_3.

The mathematical model. We associate with the molecular model in Table 1 a mathematical model in terms of differential equations, where for each reaction we assume the principle of mass action, see, e.g., [23]. We associate with each reactant a continuous, time-dependant variable that gives its concentration level. For each variable, its differential equation gives the cumulated consumption and production rates of the reactant corresponding to it in the molecular model. Thus, the dynamic behavior of the molecular model is described through the set of all resulting differential equations. For the full set of differential equations we refer to [21] and [24]. For details on the parameter estimation and the experimental

validation of the model, we refer to [21]. The resulting model exhibits four major numerical achievements, see [21]:

(A) *It uses economically the cellular resources:* In the absence of heat shock, the transcription of the hsp-encoding gene is almost non-existent. This gene is transactivated only for a short period of time after the temperature increases.

(B) *It is fast to respond to a heat shock:* Upon temperature upshift, the hsp-encoding gene is quickly activated.

(C) *The response is effective:* The level of mfp is kept low when the heat shock is mild.

(D) *The response is scalable:* The cell exhibits a higher response when exposed to higher temperature.

A control-based modularization of the heat shock response model. A control-driven analysis of the heat shock response model of [21] was introduced in [25] to decompose the heat shock response model. The model was divided into the following submodules: *the plant*, i.e., the process to be regulated, *the controller*, i.e., the decision-making module, and *the actuator*, i.e., the module which modifies the current state of the system, thus influencing the activity of the plant. A *sensor* which measures the current state of the system and sends this information to the controller and three *feedback mechanisms* regulating this process were also identified. This decomposition of the heat shock model is presented in Table 2, where the reaction numbers refer to the reactions in Table 1.

For a more intuitive understanding of this modularization, we also include a graphical illustration in Figure 1. The three identified feedback loops and their points of interaction with the mainstream process are depicted in Figure 2.

Knockdown mutant models. In order to disentangle the role of the feedback mechanisms within the full model, we consider eight knockdown mutants obtained by eliminating from the basic model all combinations of the feedbacks FB_1, FB_2, and FB_3. We will denote each of these mutants as M_X, where $X \subseteq \{1, 2, 3\}$ represents the set of indexes of the feedbacks included in the model M_X:

- M_0 includes no feedback, i.e., it consists of reactions $[\mathbf{r1}]$-$[\mathbf{r4}]$, $[\mathbf{r9}]$-$[\mathbf{r12}]$ and the backward direction of reaction $[\mathbf{r5}]$. In the control-theory terminology, this model is called the *open-loop design*.
- M_1 includes feedback FB_1, i.e., it consists of reactions $[\mathbf{r1}]$-$[\mathbf{r5}]$, $[\mathbf{r9}]$-$[\mathbf{r12}]$.
- M_2 includes feedback FB_2, i.e., it consists of reactions $[\mathbf{r1}]$-$[\mathbf{r4}]$, $[\mathbf{r6}]$-$[\mathbf{r7}]$, $[\mathbf{r9}]$-$[\mathbf{r12}]$, and the backward direction of reaction $[\mathbf{r5}]$.

- M_3 includes feedback FB_3, i.e., it consists of reactions [**r1**]-[**r4**], [**r8**]-[**r12**], and the backward direction of reaction [**r5**].

- $M_{1,2}$ includes feedbacks FB_1, FB_2, i.e., it consists of reactions [**r1**]-[**r7**], [**r9**]-[**r12**].

- $M_{1,3}$ includes feedbacks FB_1, FB_3, i.e., it consists of reactions [**r1**]-[**r5**], [**r8**]-[**r12**].

- $M_{2,3}$ includes feedbacks FB_2, FB_3, i.e., it consists of reactions [**r1**]-[**r4**], [**r6**]-[**r12**], and the backward direction of reaction [**r5**].

- $M_{1,2,3}$ is the full, reference model, consisting of reactions [**r1**]-[**r12**].

To identify the individual contributions of the three feedback mechanisms, we compare the dynamics of these eight models at $42°C$. We choose this temperature since at $42°C$ the experimental data shows a heat shock response both in terms of increased level of misfolded proteins and in terms of transcription activity of the hsp-encoding genes, see [21].

Numerical setup of the knockdown mutant models. In our comparison of the numerical knockdown mutant models we aim to focus on the differences stemming from the intrinsic dissimilarities in their architectures and eliminate as much as possible differences coming from unfavorable numerical setups chosen for the various models. For example, we consider all knockdown mutants as viable alternatives for the heat shock response model. We impose the following three constraints:

(1) The kinetic rate constants for the reactions of each of the eight knockdown mutants should be chosen in such a way that the numerical prediction for the time evolution of the level of hsf_3: hse should fit in with the experimental data given in [26] on DNA binding of hsf_3.

(2) The initial distribution of the reactants of each mutant should be chosen in such a way that they form a steady state at $37°C$ for that particular model.

(3) For all knockdown mutants, the values of the mass constants C_1, C_2, C_3 are chosen to be identical to those of the reference model $M_{1,2,3}$.

Note that our constraint (1) is fundamentally different from the one used in [25], where the mutants were regarded as submodels of the reference model. As such, in [25], all mutants assumed the same kinetic rate constants as the reference model. Instead, we perform here parameter estimation to determine the kinetic rates for each of the alternative models. Our aim is to find for each alternative architecture a favorable numerical setup that provides numerical predictions which

verify the existent experimental data. This way, each mutant establishes itself as a possible alternative model for the heat shock response.

Condition (2) was used in the same way both in [21] when choosing the initial setup of the reference model and in [25] for the knockdown mutants. However, it is essentially different from the condition used in [7], where the authors take an approach based on mathematically controlled comparison, see [27]. As such, in [7] the authors set all submodels to start from the same initial setup as the reference model. Instead, our approach is based on a more biologically meaningful constraint, i.e., all models should exhibit a steady state behavior in the absence of a heat shock. In particular, this means that each mutant will present a different initial distribution of the reactants, depending on the kinetics of the underlying reaction network and the mass constants.

3 Results

When comparing the performance of the eight alternative models we focused on two aspects: the total amount of hsp and the total amount of mfp both at $37°C$ and at $42°C$. We were interested mainly in these two aspects since a very high level of mfp indicates a non-effective response while a very high level of hsp indicates a non-economical response. We first chose empirically four numerical thresholds, denoted by l^{37}_{mfp}, l^{42}_{mfp}, l^{37}_{hsp}, and l^{42}_{hsp}, which differentiated between 'low' and 'high' levels for the total amount of mfp and hsp at $37°C$ and at $42°C$, respectively.

We associated to each of the three feedback mechanisms a Boolean variable, denoted by F_1, F_2 and F_3, respectively. Then, for each knockdown mutant we wrote a Boolean formula expressing which of the feedback mechanisms are present in the model, see Table 3 where we denoted by $\wedge$ the conjunction operator and by $\overline{F_i}$ the negation of the variable F_i. For example, to knockdown mutant $M_{1,2}$ we associated the Boolean formula $F_1 \wedge F_2 \wedge \overline{F_3}$ to express that feedbacks FB_1, FB_2 are included in the model, while feedback FB_3 is not.

Note that this approach is different from the Boolean modeling framework used in the literature, see e.g., [8], [9], [10], [12], for qualitative modeling and simulation of biological systems. When working with Boolean networks, one usually associates to each species a Boolean variable, which is either 0 if that particular species is inactive, i.e., its activity is biologically undetectable, or 1 otherwise. At the same time, knowing the values of the Boolean variables associated to all species at some time point t, Boolean functions are used to compute the values for the next time point. In our approach a Boolean formula describes the control architecture of the model, i.e., which of the three feedbacks are present in that particular model.

Going one step further in our approach, we considered all knockdown mutant models having 'low' total amount of hsp at $37°C$ and at $42°C$, respectively. By writing the disjunction, denoted by $\vee$, of the formulas corresponding to these mu-

tants we obtained a Boolean formula describing the contribution of each feedback to achieving the property: which feedbacks must be present in the model in order for it to exhibit the desired property. We applied the same technique to describe the architectures which exhibit low levels for the total amount of mfp at $37°C$ or at $42°C$.

The open-loop design. We started our analysis with the mutant M_0, which does not include any of the three feedback mechanisms. Using the notations from Table 4, the system of differential equations corresponding to M_0 is in Table 5. The steady state equations (obtained by equating all differential equations to 0) are in Table 6. These equations showed that if the mutant M_0 starts from its steady state at $37°C$, then at any temperature the differentials for $X_1, \ldots, X_5$, and X_7 are zero. That is, those functions remain constant at their steady state values independent of temperature. In particular, the DNA binding level, i.e., hsf_3: hse, remains constant even when we increase the temperature. So, for no numerical setup, this mutant can provide numerical predictions in agreement with the data from [26] if it starts from its steady state at $37°C$. Thus, we discarded this knockdown mutant from our considerations.

Numerical analysis of the remaining knockdown mutant models. For each of the mutants M_1, M_2, M_3, $M_{1,2}$, $M_{1,3}$, and $M_{2,3}$ we performed parameter estimation to identify a numerical setup, i.e., a set of values for the kinetic rate constants, that provides numerical predictions in accordance with the experimental data of [26]. The results are shown in Table 7 A. We then numerically estimated the steady state of each model at $37°C$; the results are given in Table 7 B. Finally, we numerically integrated the mathematical model corresponding to each knockdown mutant starting from its own steady state values in Table 7 B. We integrated the ODEs up to 14400 seconds (in model time), for a temperature value of $42°C$. We collected in Table 8 the maximal values for the total amount of hsp and mfp in each of these models, both at $37°C$ and at $42°C$. For the numerical integration we used the software COPASI [28].

We chose empirically four numerical thresholds separating the 'low' and 'high' values for the total amount of: (i) hsp proteins at $37°C$; (ii) mfp proteins at $37°C$; (iii) hsp proteins at $42°C$; and (iv) mfp proteins at $42°C$. The thresholds we selected were the following: $l^{37}_{\mathsf{hsp}} = 8000$, $l^{37}_{\mathsf{mfp}} = 3000$, $l^{42}_{\mathsf{hsp}} = 8 \times 10^4$, and $l^{42}_{\mathsf{mfp}} = 2.5 \times 10^6$, respectively, all in terms of number of molecules. We plotted the behavior of each knockdown mutant models with respect to these thresholds in Figures 3 and 4.

We considered the following four properties:

- *Property P_1: Low level for the total amount of* hsp *at* $37°C$. This property is exhibited only by the mutants M_1, M_3, $M_{1,2}$, $M_{1,3}$, and $M_{1,2,3}$. Using the Boolean formulas in Table 3 expressing each mutant in terms of their

feedback structure, we constructed a Boolean formula for property P_1. This is easily obtained as a disjunctive formula (logical OR) among the Boolean formulas for M_1, M_3, $M_{1,2}$, $M_{1,3}$, and $M_{1,2,3}$:

$$(F_1 \wedge \overline{F_2} \wedge \overline{F_3}) \vee (\overline{F_1} \wedge \overline{F_2} \wedge F_3) \vee (F_1 \wedge F_2 \wedge \overline{F_3}) \vee \\ \vee (F_1 \wedge \overline{F_2} \wedge F_3) \vee (F_1 \wedge F_2 \wedge F_3),$$

which could be rewritten in a compact form as:

$$F_1 \vee (\overline{F_1} \wedge \overline{F_2} \wedge F_3). \tag{1}$$

Thus, property P_1 can be satisfied if and only if either feedback F_1 is present (regardless of whether F_2 and F_3 are included or not) or feedback F_3 is present while feedbacks F_1 and F_2 are absent.

- *Property P_2: Low level for the maximal value of the total amount of* hsp *at* $42°C$. This property is exhibited again only by the mutants M_1, M_3, $M_{1,2}$, $M_{1,3}$, and $M_{1,2,3}$. So, we obtained for property P_2 the Boolean formula

$$F_1 \vee (\overline{F_1} \wedge \overline{F_2} \wedge F_3). \tag{2}$$

- *Property P_3: Low level for the total amount of* mfp *at* $37°C$. This property is exhibited only by the mutants M_1, M_3, and $M_{1,2,3}$. So, in this case we obtained the Boolean formula

$$(F_1 \wedge \overline{F_2} \wedge \overline{F_3}) \vee (\overline{F_1} \wedge \overline{F_2} \wedge F_3) \vee (F_1 \wedge F_2 \wedge F_3). \tag{3}$$

- *Property P_4: Low level for the maximal value of the total amount of* mfp *at* $42°C$. This property is exhibited by the mutants M_1, M_2, $M_{1,2}$, and $M_{1,2,3}$. In this case, we obtained the Boolean formula

$$(F_1 \wedge \overline{F_3}) \vee (F_1 \wedge F_2 \wedge F_3) \vee (\overline{F_1} \wedge F_2 \wedge \overline{F_3}). \tag{4}$$

To investigate which knockdown mutants can be both effective and economic, we looked at the models that exhibit low levels for both hsp and mfp. For a temperature of $37°C$, we considered the models that verify simultaneously properties P_1 and P_3. The Boolean formula describing these architectures was easily obtained as a conjunctive formula (logical AND) among the formulas for properties P_1 and P_3, which could then be rewritten in a compact form as

$$(F_1 \wedge \overline{F_2} \wedge \overline{F_3}) \vee (\overline{F_1} \wedge \overline{F_2} \wedge F_3) \vee (F_1 \wedge F_2 \wedge F_3).$$

Since this was identical with (3), we concluded that at $37°C$, once a mutant achieved a low level for the total amount of mfp, it would also exhibit a low level for the total amount of hsp. For the similar analysis at $42°C$ we were interested

in the models that verify simultaneously properties P_2 and P_4. In this case, the Boolean formula describing these architectures is

$$F_1 \wedge (F_2 \vee (\overline{F_2} \wedge \overline{F_3})).$$

This shows that to obtain low values for both hsp and mfp at $42°C$ the first feedback is essential. Moreover, only two types of mutant architectures predicted this outcome: if both F_1 and F_2 were present in the model (regardless of whether F_3 is included or not), or if F_1 was included while F_2 and F_3 are not. Furthermore, it showed that the second feedback, in addition to the first one, has a role in decreasing the levels of both hsp and mfp at $42°C$. The second type of architecture, i.e., when F_1 was present in the model while F_2 and F_3 were absent, showed that the first feedback alone is sufficient to ensure a low enough level of both hsp and mfp at $42°C$. However, when we compared the values predicted by M_1 and $M_{1,2,3}$, see Figure 4, we noticed that the cumulative effect of the second and the third feedbacks added to the first one is to further reduce the level of total mfp.

We noticed that Boolean formulas corresponding to properties P_1 and P_2 were identical. This means that once a knockdown mutant is able to keep a low level of hsp at $37°C$, it will also be able to respond to heat shock with a relatively low level of hsp. Moreover, this was the case only for two types of mutant architectures: either when the feedback F_1 was present (regardless of whether F_2 and F_3 were included or not) or when feedback F_3 was present while feedbacks F_1 and F_2 were absent. This showed that the first and the third feedback have roles in lowering the level of hsp both at $37°C$ and at $42°C$. The first type of mutant architecture, having the feedback F_1 present, was insensitive to the second and the third feedbacks: whether they were included in the model or not did not change the behavior of the model with respect to P_2 and P_4. The second type of mutant architecture that satisfies the Boolean formula (1) showed that in the absence of the first feedback, the third one is necessary to obtain low levels of hsp both at $37°C$ and at $42°C$. Moreover, if we required also properties P_3 and P_4 to be satisfied, i.e., if we asked for low levels of mfp both at $37°C$ and at $42°C$, then we saw that the first feedback had to be present in the model. Otherwise, i.e., if $F_1 = 0$, the two Boolean formulas (3) and (4) become $\overline{F_2} \wedge F_3$ and $F_2 \wedge \overline{F_3}$, respectively, which obviously cannot be simultaneously satisfied. This confirmed again our conclusion that the first feedback is essential for the model to satisfy all four properties P_1, P_2, P_3, and P_4, i.e., for the model to exhibit low levels for both hsp and mfp, both at $37°C$ and at $42°C$.

4 Discussion

Carrying out a numerical comparison between two alternative computational models for a biological system is, in general, a difficult problem. It involves a detailed analysis of various aspects of the models: the underlying networks, the biological

constrains, the initial distribution of the reactants and the kinetics of the models. The problem becomes somewhat simpler if the considered alternative models are submodels of a larger model: the underlying networks are similar, although not identical, and the biological constrains are the same.

Previous approaches for model comparison. The technique of mathematically controlled comparison, [27], provides a structured approach for comparing several alternative designs with respect to some chosen measures of functional effectiveness. However, this framework imposes one important constraint on the alternative designs: they are allowed to differ from the reference design in only one component. Moreover, the mathematical models both for the reference design and for the alternative architectures are developed in the framework of canonical nonlinear modeling referred to as S-systems, [29], [30], and [31]. Then, using various systemic properties, one imposes some constraints on all parameters of the alternative designs, setting their values depending on the parameters of the reference model. Finally, one chooses some numerical measures of functional effectiveness and uses them to compare the alternative designs with the reference model. This way, one can determine analytically the qualitative differences between the compared models. If one is also interested in quantifying these differences, then numerical values can be introduced for the parameters of the models. However, by doing this the generality of the results is lost. An extension of the method of mathematically controlled comparison was proposed in [32] to include some statistical methods, see [33] and [34], which allow the use of numerical values for the parameters while still preserving the generality of the conclusions.

Another approach for model comparison was proposed in [25]. Since the alternative designs are submodels of the reference model, the underlying reaction networks of these models are very similar (although not identical), and both the biological constraints and the kinetics of the reactions are taken from the reference model. The only remaining question is how to chose the initial distribution of the variables in the alternative designs. In the mathematically controlled comparison they are usually taken from the reference model, see [7] for a case study using this method. However, this might lead to biased comparisons for some biochemical systems. For instance, for regulatory networks, models should be in a steady state in the absence of the trigger of the response. In particular, the initial values of the reference model are usually chosen in such a way to fulfil this condition. However, this does not imply in general that also a submodel will be in its steady state if it starts from the same initial values as the reference model. As a consequence, the dynamic behavior of the submodel will exhibit two intertwined tendencies: the migration from a possible unstable state and the response to a particular stimulus. Thus, if the purpose of the comparison is to determine the efficiency of the response of various submodels to a particular trigger, then the approach proposed in [25] is more appropriate, leading to biologically unbiased results. In this approach, the initial values of the reactants are chosen in such a

way that they constitute a steady state of that design in the absence of a trigger. However, also in this approach, the comparison is done locally, for a particular set of parameters. In [35], this method was combined with some statistical methods of [33] and [34], leading to general comparison results independent of the values of the parameters.

Our approach for knockdown mutant model comparison: advantages and limitations. In this paper, we proposed a novel approach to the knockdown mutant model comparison problem. First, we associated a Boolean variable to each of the three feedback mechanisms identified in [25] for the reference model of the eukaryotic heat shock response. Then, for each knockdown mutant we wrote a Boolean formula (using the conjunction and negation of the three introduced Boolean variables) characterizing its control architecture, i.e., which of the three feedback mechanisms are present in the model. As such, each of these formulas encompass time-independent properties of the models. This makes our approach very different from the Boolean network framework for modeling biological systems, see [8], [11], [12], [13], where one usually associates a Boolean variable to each species present in the system. Boolean formulas are then used to simulate the time evolution of the species. However, in our approach the associated Boolean formulas are parameter independent, i.e., they are not influenced by the parameters used to describe the compared models. Going one step further, we could introduce a Boolean formula characterizing all those mutant architectures that exhibit a given behavioral property, e.g., low levels of hsp or mfp. This can be easily obtained as a disjunctive formula (logical OR) of the Boolean formulas describing the architectures of the mutants exhibiting the required property. However, in order to perform numerical simulations of the models we needed numerical setups for each of the knockdown mutants, i.e., specific values both for the initial distribution of the reactants and for the kinetic rate constants of the models. For the initial values of the variables, we chose the approach proposed in [25], i.e., we set them separately for each knockdown mutant in such a way that they form a steady state for that particular model. Regarding the kinetic rate constants in each of the knockdown mutants, one approach is to take them from the reference model, see [25]. The idea in this case is to make the whole comparison in the numerical setup of the reference model. Alternatively, we proposed here to separately estimate the kinetic constants of each alternative model with respect to available experimental data. In other words, we considered all models to be viable alternatives for the biological system and, as such, we took for each of them a most favorable numerical setup.

Since the numerical setup giving a good model fit is in general not unique, it means that our analysis is sensitive with respect to the choice of the values for the kinetic constants. This is often the case when model fitting is involved, see [1]. Repeating the analysis for several numerical setups (all of them as good in terms of fitting the model to the experimental data) would enrich the conclusions, by po-

tentially showing that the same model architecture can exhibit different properties depending on the numerical setup. The conclusions of the analysis also depend on the numerical values chosen for the thresholds l^{37}_{hsp}, l^{37}_{mfp}, l^{42}_{hsp}, and l^{42}_{mfp}.

It is crucial for our approach that all knockdown mutant models are considered in the analysis, i.e., all possible combinations ON/OFF of the model components are included in the comparison. In this way, we obtain a complete characterization of the properties being analyzed in terms of *all* model architectures that can exhibit those properties. For a large number of components, this approach becomes quickly computationally challenging: for n components to be analyzed, there are 2^n knockdown mutant models to be compared. Including in the comparison only a part of those mutants is also possible but then the output of the method is partial: one only discovers *some*, potentially not all, model architectures exhibiting the property of interest.

When we compared the numerical behavior of the knockdown mutants, we chose a mathematical model formulation in terms of ordinary differential equations. However, our approach is independent of this formulation and it would work equally well with other formulations, such as continuous-time Markov chains and their numerical simulations based on Gillespie's algorithm, see [36, 37].

Our approach can be easily extended to a more refined analysis, where the range of the properties to be analyzed is divided into more domains than just 'low' and 'high'. The range could in fact be divided into an arbitrarily high number of intermediate domains, depending on the details of the case study. A Boolean formula could be associated to characterize each of those domains in a manner similar to that demonstrated in this paper.

Acknowledgments. This work was supported by Academy of Finland, grants 129863, 108421, and 122426. Andrzej Mizera is on leave of absence from the Institute of Fundamental Technological Research, Polish Academy of Sciences, Warsaw, Poland.

References

[1] Chen WW, Schoeberl B, Jasper PJ, Niepel M, Nielsen UB, Lauffenburger DA, Sorger PK (2009) Inputoutput behavior of ErbB signaling pathways as revealed by a mass action model trained against dynamic data. *Molecular Systems Biology* 5: 239.

[2] Csete ME, Doyle JC (2002) Reverse Engineering of Biological Complexity. *Science* 295: 1664-1669.

[3] Hawkins BA, Cornell HV (1999) *Theoretical Approaches to Biological Control*. Cambridge University Press.

[4] Kitano H (2002) Systems biology: A brief overview. *Science* 295: 16621664.

[5] Sontag E (2005) Molecular systems biology and control. *Eur J Control* 11: 396435.

[6] Wolkenhauer O (2001) Systems biology: the reincarnation of systems theory applied in biology? *Brief Bioinform* 2(3): 258270.

[7] El-Samad H, Kurata H, Doyle JC, Gross CA, Khammash M (2005) Surviving heat shock: Control strategies for robustness and performance. *Proc Natl Acad Sci USA* 102(8): 2736-2741.

[8] Chaves M, Albert R, Sontag ED (2005) Robustness and fragility of Boolean models for genetic regulatory networks. *J Theor Biol* 235(3):431-49.

[9] Faure A, Naldi A, Chaouiya C, Thieffry D (2006) Dynamical analysis of a generic Boolean model for the control of the mammalian cell cycle. *Bioinformatics* 22(14): 124-131.

[10] Glass L, Kauffman SA (1973) The logical analysis of continuous, non-linear biochemical control networks. *J Theoret Biol* 39(1): 103-129.

[11] Kauffman S, Peterson C, Samuelsson B, Troein C (2003) Random Boolean network models and the yeast transcriptional network. *Proc Natl Acad Sci USA* 100: 1479614799.

[12] Kervizic G, Corcos L (2008) Dynamical modeling of the cholesterol regulatory pathway with Boolean networks. *BMC Systems Biology* 2: 99.

[13] Shmulevich I, Dougherty R, Zhang W (2002) From Boolean to Probabilistic Boolean Networks as Models of Genetic Regulatory Networks. *Proc IEEE* 90(11): 1778-1792.

[14] Kauffman S (1969) Metabolic stability and epigenesis in randomly constructed genetic nets. *J Theoret Biol* 22: 437-467.

[15] Chen Y, Voegeli TS, Liu PP, Noble EG, Currie RW (2007) Heat shock paradox and a new role of heat shock proteins and their receptors as anti-inflammation targets. *Inflamm Allergy Drug Targets* 6(2):91-100.

[16] Powers MV, Workman P (2007) Inhibitors of the heat shock response: Biology and pharmacology. *FEBS Lett* 581(19): 3758-3769.

[17] Voellmy R, Boellmann F (2007) Chaperone regulation of the heat shock protein response. *Adv Exp Med Biol* 594: 89-99.

[18] Kurata H, El-Samad H, Yi TM, Khamash M, Doyle J (2001) Feedback regulation of the heat shock response in E.coli. In *Proc 40th IEEE Conference on Decision and Control* 837842.

[19] Kampinga HK (1993) Thermotolerance in mammalian cells: protein denaturation and aggregation, and stress proteins. *J. Cell Science* 104: 1117.

[20] Pockley AG (2003) Heat shock proteins as regulators of the immune response. *The Lancet* 362(9382): 469476.

[21] Petre I, Mizera A, Hyder CL, Mikhailov A, Eriksson JE, Sistonen L, Back R-J (2011) A simple mathematical model for the eukaryotic heat shock response and its mathematical validation. *J Nat Comp*, to appear.

[22] Holmberg CI, Tran SE, Eriksson JE, Sistonen L (2002) Multisite phosphorylation provides sophisticated regulation of transcription factors. *Trends Biochem Sci* 27(12): 619-627.

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

[24] Petre I, Mizera A, Hyder CL, Mikhailov A, Eriksson JE, Sistonen L, Back R-J (2008) in *Algorithmic bioprocesses*, eds Condon A, Harel D, Kok JN, Salomaa A, Winfree E (Springer, Dordrecht Heidelberg London New York), pp 411-425.

[25] Czeizler El, Czeizler Eu, Back R-J, Petre I (2009) in *Lecture Notes in Bioinformatics* 5688, eds Degano P, Gorrieri R (Springer, Berlin, Heidelberg), pp 111-125.

[26] Kline MP, Morimoto RI (1997) Repression of the heat shock factor 1 transcriptional activation domain is modulated by constitutive phosphorylation. *Mol Cel Biol* 17(4): 2107-2115.

[27] Savageau MA (1972) The behavior of intact biochemical control systems. *Curr Top Reg* 6: 63-130.

[28] Hoops S, Sahle S, Gauges R, Lee C, Pahle J, Simus N, Singhal M, Xu L, Mendes P, Kummer U (2006) Copasi a COmplex PAthway SImulator. *Bioinformatics* 22(24): 30673074.

[29] Savageau MA (1969) Biochemical systems analysis: I. Some mathematical properties of the rate law for the component enzymatic reactions. *J Theoret Biol* 25(3): 365-369.

[30] Savageau MA (1969) Biochemical systems analysis: II. The steady state solution for an n-pool system using a power law approximation. *J Theoret Biol* 25(3): 370-379.

[31] Savageau MA (1970) Biochemical systems analysis: III. Dynamic solutions using a power-law approximation. *J Theoret Biol* 26(2), 215-226.

[32] Alves R, Savageau MA (2000) Extending the method of mathematically controlled comparison to include numerical comparisons. *Bioinformatics* 16: 786798.

[33] Alves R, Savageau MA (2000) Comparing systemic properties of ensembles of biological networks by graphical and statistical methods. *Bioinformatics* 16 (6): 527-533.

[34] Alves R, Savageau MA (2000) Systemic properties of ensembles of metabolic networks: application of graphical and statistical methods to simple unbranched pathways. *Bioinformatics* 16(6): 534-547.

[35] Mizera A, Czeizler E, Petre I (2011) Methods for biochemical model decomposition and quantitative submodel comparison. *Israel J Chem*, to appear.

[36] Gillespie DT (1976) A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J Comput Phys* 22: 403-434.

[37] Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. *J Phys Chem* 81(25): 2340-2361.

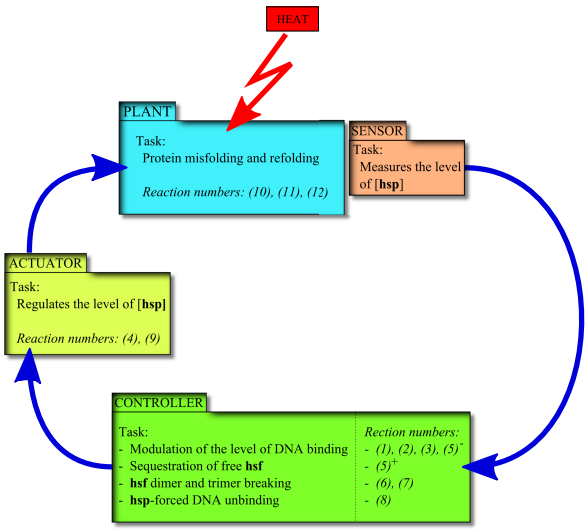


Figure 1: The control structure of the heat shock response network.

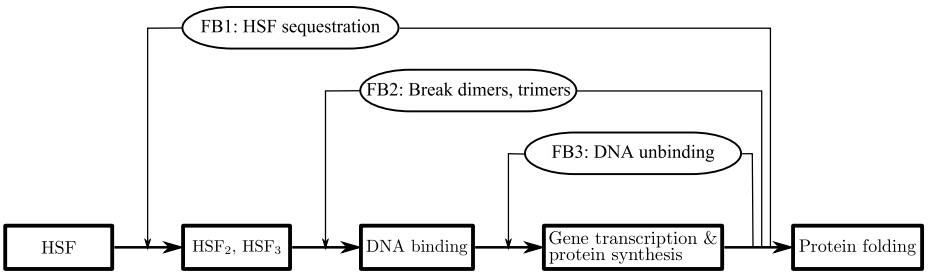


Figure 2: The control structure of the heat shock response network.

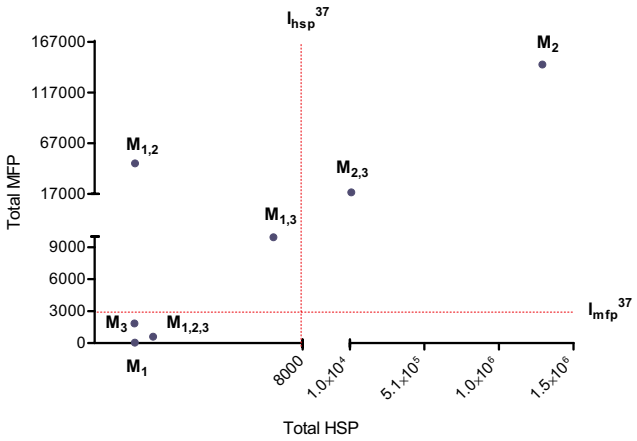


Figure 3: The total amount of hsp and mfp for each of the seven models at $37°C$. Values on the axes are in terms of number of molecules and should be interpreted as an average of a population of cells.

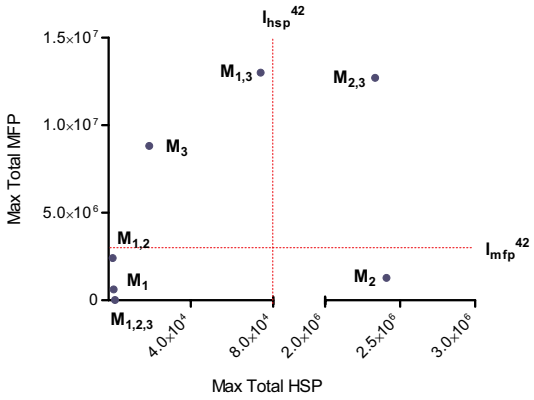


Figure 4: The maximal value for the total amount of hsp and mfp for each of the seven models at $42°C$. Values on the axes are in terms of number of molecules and should be interpreted as an average of a population of cells.

Table 1: The molecular model for the eukaryotic heat shock response proposed in [21].

Reaction	Reaction name	
$2\,\mathsf{hsf} \leftrightarrows \mathsf{hsf}_2$	(hsf dimerization)	$[\mathbf{r1}]$
$\mathsf{hsf} + \mathsf{hsf}_2 \leftrightarrows \mathsf{hsf}_3$	(hsf trimerization)	$[\mathbf{r2}]$
$\mathsf{hsf}_3 + \mathsf{hse} \leftrightarrows \mathsf{hsf}_3\colon \mathsf{hse}$	(DNA binding)	$[\mathbf{r3}]$
$\mathsf{hsf}_3\colon \mathsf{hse} \rightarrow \mathsf{hsf}_3\colon \mathsf{hse} + \mathsf{hsp}$	(hsp synthesis)	$[\mathbf{r4}]$
$\mathsf{hsp} + \mathsf{hsf} \leftrightarrows \mathsf{hsp}\colon \mathsf{hsf}$	(hsf sequestration)	$[\mathbf{r5}]$
$\mathsf{hsp} + \mathsf{hsf}_2 \rightarrow \mathsf{hsp}\colon \mathsf{hsf} + \mathsf{hsf}$	(hsf_2 breaking)	$[\mathbf{r6}]$
$\mathsf{hsp} + \mathsf{hsf}_3 \rightarrow \mathsf{hsp}\colon \mathsf{hsf} + 2\,\mathsf{hsf}$	(hsf_3 breaking)	$[\mathbf{r7}]$
$\mathsf{hsp} + \mathsf{hsf}_3\colon \mathsf{hse} \rightarrow \mathsf{hsp}\colon \mathsf{hsf} + 2\,\mathsf{hsf} + \mathsf{hse}$	(hsp-forced hsf_3 unbinding)	$[\mathbf{r8}]$
$\mathsf{hsp} \rightarrow \emptyset$	(hsp degradation)	$[\mathbf{r9}]$
$\mathsf{prot} \rightarrow \mathsf{mfp}$	(protein misfolding)	$[\mathbf{r10}]$
$\mathsf{hsp} + \mathsf{mfp} \leftrightarrows \mathsf{hsp}\colon \mathsf{mfp}$	(protein chaperoning)	$[\mathbf{r11}]$
$\mathsf{hsp}\colon \mathsf{mfp} \rightarrow \mathsf{hsp} + \mathsf{prot}$	(protein refolding)	$[\mathbf{r12}]$

Table 2: The control-based decomposition of the model in Table 1. We denote the 'left-to-right' direction of reaction $[\mathbf{r5}]$ by $[\mathbf{r5}]^+$ and by $[\mathbf{r5}]^-$ its 'right-to-left' direction.

	Main Task	Reactions
Plant	Protein misfolding and refolding	$[\mathbf{r10}]$, $[\mathbf{r11}]$, $[\mathbf{r12}]$
Actuator	Regulate the level of hsp	$[\mathbf{r4}]$, $[\mathbf{r9}]$
Sensor	Measure the level of hsp	
Controller	Modulate the level of DNA binding	$[\mathbf{r1}]$, $[\mathbf{r2}]$, $[\mathbf{r3}]$, $[\mathbf{r5}]^-$
Feedback FB_1	Sequestration of hsf	$[\mathbf{r5}]^+$
Feedback FB_2	Dimer and trimer breaking	$[\mathbf{r6}]$, $[\mathbf{r7}]$
Feedback FB_3	hsp-forced DNA unbinding	$[\mathbf{r8}]$

Table 3: Boolean formulas encoding the presence or absence of the three feedback mechanisms in each of the eight models.

Mutant	Boolean formula
M_0	$\overline{F_1} \wedge \overline{F_2} \wedge \overline{F_3}$
M_1	$F_1 \wedge \overline{F_2} \wedge \overline{F_3}$
M_2	$\overline{F_1} \wedge F_2 \wedge \overline{F_3}$
M_3	$\overline{F_1} \wedge \overline{F_2} \wedge F_3$
$M_{1,2}$	$F_1 \wedge F_2 \wedge \overline{F_3}$
$M_{1,3}$	$F_1 \wedge \overline{F_2} \wedge F_3$
$M_{2,3}$	$\overline{F_1} \wedge F_2 \wedge F_3$
$M_{1,2,3}$	$F_1 \wedge F_2 \wedge F_3$

Table 4: The molecular model for knockdown mutant M_0, the open-loop design.

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

Variable	Molecule or complex
X_1	hsf
X_2	hsf_2
X_3	hsf_3
X_4	$\mathsf{hsf}_3\colon \mathsf{hse}$
X_5	hse
X_6	hsp
X_7	$\mathsf{hsp}\colon \mathsf{hsf}$
X_8	$\mathsf{hsp}\colon \mathsf{mfp}$
X_9	prot
X_{10}	mfp

Table 5: The ODE model corresponding to knockdown mutant model M_0, the open-loop design. For the expression of the temperature-dependant parameter ϕ_T we refer to [21] and [24].

$$
\begin{aligned}
dX_1/dt = &\quad k_2^+ X_1 X_2 + k_2^- X_3 + k_5^- X_7 \quad 2 k_1^+ X_1^2 + 2 k_1^- X_2 \\
dX_2/dt = &\quad k_2^+ X_1 X_2 + k_2^- X_3 + k_1^+ X_1^2 \quad k_1^- X_2 \\
dX_3/dt = &\quad k_3^+ X_3 X_5 + k_2^+ X_1 X_2 \quad k_2^- X_3 + k_3^- X_4 \\
dX_4/dt = &\; k_3^+ X_3 X_5 \quad k_3^- X_4 \\
dX_5/dt = &\quad k_3^+ X_3 X_5 + k_3^- X_4 \\
dX_6/dt = &\; k_4 X_4 + k_5^- X_7 \quad k_{11}^+ X_6 X_{10} + (k_{11}^- + k_{12}) X_8 \quad k_9 X_6 \\
dX_7/dt = &\quad k_5^- X_7 \\
dX_8/dt = &\; k_{11}^+ X_6 X_{10} \quad (k_{11}^- + k_{12}) X_8 \\
dX_9/dt = &\; k_{12} X_8 \quad \varphi_T X_9 \\
dX_{10}/dt = &\quad k_{11}^+ X_6 X_{10} + k_{11}^- X_8 + \varphi_T X_9.
\end{aligned}
$$

Table 6: The steady state equations of knockdown mutant model M_0, the open-loop design.

$$
\begin{aligned}
k_1^- X_2 &= k_1^+ X_1^2 \\
k_2^- X_3 &= k_2^+ X_1 X_2 \\
k_3^- X_4 &= k_3^+ X_3 X_5 \\
k_9 X_6 &= k_4 X_4 \\
0 &= \quad k_5^- X_7 \\
(k_{11}^- + k_{12}) X_8 &= k_{11}^+ X_6 X_{10} \\
\varphi(T) X_9 &= k_{12} X_8
\end{aligned}
$$

Table 7: The numerical values of the parameters and the initial values of the variables of the knockdown mutants. A. The numerical values of the parameters in each of the six knockdown mutants. k_i denotes the kinetic rate constant of the irreversible reaction (i). k_i^+ denotes the 'left-to-right' direction of reaction (i), while k_i^- denotes its 'right-to-left' direction. Notice that there is no parameter k_{10} in the table. It is assumed to be the temperature-dependant parameter ϕ_T whose value is determined from the expression presented and discussed in [21] and [24]. B. The initial values of all variables in each of the six knockdown mutants.

A

	M_1	M_2	M_3	$M_{1,2}$	$M_{1,3}$	$M_{2,3}$
k_1^+	0 02	10 00	4 36 10^{-7}	7 24	0 04	10 00
k_1^-	0 01	9 90	1 36 10^{-7}	1 84 10^{-5}	0 26	0 01
k_2^+	9 90	6 02	0 23	0 34	0 00	0 01
k_2^-	0 01	0 01	1 22 10^{-6}	1 05 10^{-5}	0 03	8 04 10^{-5}
k_3^+	0 08	3 04	0 01	0 70	0 13	10 00
k_3^-	0 66	0 00	0 17	0 15	0 00	0 00
k_4	0 01	10 00	0 19	0 00	0 51	1 59
k_5^-	0 00	10 00	9 98	1 23	3 41	10 00
k_5^+	0 15	-	-	10 00	1 00 10^{-9}	-
k_6	-	0 60	-	1 00 10^{-9}	-	0 13
k_7	-	0 24	-	10 00	-	2 08 10^{-7}
k_8	-	-	0 51	-	0 23	3 20
k_9	3 20 10^{-5}	3 20 10^{-5}	3 20 10^{-5}	3 20 10^{-5}	3 20 10^{-5}	3 20 10^{-5}
k_{11}^+	9 75	10 00	0 38	0 00	0 00	0 57
k_{11}^-	6 52	1 00 10^{-9}	10 00	1 30 10^{-8}	0 32	5 01
k_{12}	32 08	0 01	0 70	16 47	0 17	0 05

B

	M_1	M_2	M_3	$M_{1,2}$	$M_{1,3}$	$M_{2,3}$
[hsf]	0 03	36 95	1399 68	1 27	67 96	33 45
[hsf_2]	0 00	0 02	0 00	27 33	667 19	130 23
[hsf_3]	0 11	1 52 10^{-5}	4 28	0 01	2 39	0 09
[hse]	32 28	28 97	32 67	31 41	32 64	32 68
[hsf_3: hse]	0 41	3 72	0 02	1 28	0 05	0 01
[hsp]	99 31	1 16262 10^6	100 05	130 39	839 82	662 90
[hsp: hsf]	1411 09	1364 52	0 09	1352 88	3 00	1118 47
[mfp]	1 24	8 58 10^{-5}	405 56	47164 90	3915 46	244 61
[hsp: mfp]	31 14	144533	1426 09	60 62	6024 21	18259 40
[prot]	1 14916 10^8	1 14771 10^8	1 14914 10^8	1 14868 10^8	1 14906 10^8	1 14897 10^8

Table 8: a) The numerical values for the total amount of hsp and mfp at $37°C$, b) The maximal numerical values for the total amount of hsp and mfp at $42°C$. All values are in terms of number of molecules and should be interpreted as an average of a population of cells.

	Total hsp	Total mfp
M_1	1541	32,3
M_2	$1,3 \times 10^6$	144533
M_3	1526	1832
$M_{1,2}$	1544	47225
$M_{1,3}$	6867	9939,6
$M_{2,3}$	20040,8	18504
$M_{1,2,3}$	2241	589

(a)

	Max Total hsp	Max Total mfp
M_1	2458	623997
M_2	$2,41 \times 10^6$	$1,28 \times 10^6$
M_3	19782,5	$8,8 \times 10^6$
$M_{1,2}$	1978,6	$2,41 \times 10^6$
$M_{1,3}$	73931,4	$1,3 \times 10^7$
$M_{2,3}$	233778	$1,27 \times 10^7$
$M_{1,2,3}$	3157	16116

(b)

Paper VI

Methods for biochemical model decomposition and quantitative submodel comparison

Andrzej Mizera, Elena Czeizler, and Ion Petre

Originally published in *Israel Journal of Chemistry*, 51(1):151–164, 2011.

Review

Israel Journal of Chemistry

DOI: 10.1002/ijch.201000067

Methods for Biochemical Model Decomposition and Quantitative Submodel Comparison

Andrzej Mizera,[a] Elena Czeizler,[a] and Ion Petre*[a]

Abstract: Comparing alternative models for a given biochemical system is in general a very difficult problem: the models may focus on different aspects of the same system and may consist of very different species and reactions. The numerical setups of the models also play a crucial role in the quantitative comparison. When the alternative designs are submodels of a reference model, for example, knockdown mutants of a model, the problem of comparing them becomes simpler: they all have very similar, although not identical, underlying reaction networks, and the biological constraints are given by those in the reference model. In the first part of our study, we review several known methods for model decomposition and for quantitative comparison of submodels. We describe knockdown mutants, elementary flux modes, control-based decomposition, mathematically controlled comparison and its extension, local submodel comparison and a discrete approach for comparing continuous submodels. In the second part of the paper we present a new statistical method for comparing submodels, which complements the methods presented in the review. The main difference between our approach and the known methods is related to the important question of how to chose the numerical setup in which to perform the comparison. In the case of the reviewed methods, the comparison is made in the numerical context of the reference model, i.e., in each of the alternative models both the kinetics of the reactions and the initial values of all variables are chosen to be identical to those from the reference model. We propose in this paper a different approach, better suited for response networks, where each alternative model is assumed to start from its own steady state under basal conditions. We demonstrate our approach with a case study focusing on the heat shock response in eukaryotes.

Keywords: chaperone proteins · control-based model decomposition · protein–protein interactions · quantitative model comparison · statistical methods

1. Introduction

Much experimental and theoretical effort is invested nowadays in analyzing large biochemical systems, e.g., metabolic pathways, regulatory networks, signal transduction networks, aiming to obtain a holistic perspective that can provide a comprehensive, system-level understanding of cellular behavior. This often results in the creation and analysis of very large and complex models, often encompassing hundreds of reactions and reactants (see, for example, ref. [1]). Therefore, obtaining a global picture of the system's architecture, in particular understanding the interactions between various components, or even just distinguishing a high-level functional decomposition of the network, constitutes a significant challenge. An important insight here is that the architecture of some biological systems, for example, some regulatory networks, is a consequence of functional requirements of the entire system. Even though evolution is driven by random events, some designs, such as having an extra feedback loop helping the system to correlate better the response of the system with its trigger, may offer a selective advantage and in time, may get to dominate the population.[2] Thus, when comparing the performance of different alternative designs in terms of sub-components being on or off, one aims to formulate general principles for how functional requirements correlate biologically with various designs.

Similar problems have been encountered, for instance, in engineering sciences,[3] and a variety of strategies and approaches for solving such problems have been already developed in this framework. Thus, when aiming to obtain a system-level understanding of such large biochemical networks, one possible approach is to adapt to systems biology some of the methods originating from engineering sciences, especially from control theory.[4–10]

[a] A. Mizera, E. Czeizler, I. Petre
Department of Information Technologies
Åbo Akademi University
Joukahaisenkatu 3-5 A
FIN-20520 Turku, Finland
phone: +358 (0)2 215 3361
fax: +358 (0)2 215 4732
e-mail: ipetre@abo.fi

Such methods have been used, as we also do in this paper, to identify various functional modules of a model, including feedback and feedforward mechanisms. To identify the quantitative contribution of each of the modules to the global behavior of the model, the general approach is to consider knockdown mutants of the initial model, missing one or several of the modules. The main problem then becomes an objective quantitative comparison of several alternative submodels for the same biological process. We focus on this problem in our paper, that is, we concentrate on the comparison of submodels of a given reference model. This issue is a special case of the general problem of alternative model comparison. In the general case it is a very difficult issue and is not in the scope of this study.

The first part of our paper contains a review of existing techniques for model decomposition, and for quantitative comparison of submodels. We describe knockdown mutants, elementary flux modes, control-based decomposition, mathematically controlled comparison and its extension, local submodels comparison and a discrete approach for comparing continuous submodels. In the second part of the paper we introduce a new approach to quantitative submodel comparison. A main difference in our approach compared to previous methods is that we allow the alternative models to start from different initial states, rather than assuming the initial state of the reference model. We argue that this is a better approach, at least in the case of response networks, where the system is assumed to be in a steady state under basal conditions and to exhibit a response only as an effect of an external trigger. In order to treat each model as a genuine alternative for the biological process under study, we allow it to start from its own steady state under basal conditions. Finally, we illustrate our approach with a case study focusing on the heat shock response in eukaryotes.

The numerical behavior of any model is clearly sensitive to the numerical setup, i.e., the numerical values of the kinetic constants and of the initial values of the model variables. In our approach for quantitative comparison of alternative submodels we adopt some statistical, parameter-independent methods introduced in refs. [11,12]. These methods aim to sample the numerical behavior of the model through a sampling of the parameter space. We adopt in this paper the Latin hypercube sampling method of ref. [13], which gives uniformly distributed samples over each parameter, of size independent of the number of parameters. We briefly survey this method and apply it to the heat shock response in eukaryotes.

The heat shock response is an evolutionary conserved mechanism protecting the cell against protein misfolding. In the case study for our new approach to quantitative submodel comparison we consider a model recently introduced in ref. [14]. The model was analyzed in ref. [15] using control-driven methods where it was decomposed into several modules, including three feedback loops. We focus in our case study on identifying the numerical contribution of each of these feedback loops to the global behavior of the model. A local, point-wise comparison of the three feedbacks was already done in ref. [15], in the kinetic setup of the reference model. In this paper we do a global, parameter-independent analysis of the numerical role of each feedback, through a sampling of the whole parameter space.

Andrzej Mizera is a Ph.D. candidate pursuing a double doctoral degree at the Computational Biomodeling Laboratory at the Department of Information Technologies, Åbo Akademi University, Turku, Finland, and at the Institute of Fundamental Technological Research, Polish Academy of Sciences, Warsaw, Poland. He received his M.Sc. from the Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, in 2005. His research interests revolve around mathematical modeling in systems biology.

Elena Czeizler received her Ph.D. in Mathematics with honors in 2007 from the University of Turku, Finland. Since then she was a member of the Computer Science Department in University of Western Ontario, Canada, and the Department of Information Technologies, Åbo Akademi, Finland. Her research interests include computational modeling of complex biochemical systems, combinatorics on strings, combinatorics on DNA. Currently, she works in the Computational Systems Biology group, University of Helsinki, where she focuses on the computational analysis of complex biological networks, such as finding biologically-relevant motifs within these networks.

Ion Petre is a Professor of Computer Science at Åbo Akademi University, Turku, Finland, and the leader of the Computational Biomodeling Laboratory at Turku Centre for Computer Science. He holds a Ph.D. in Discrete Mathematics from the University of Turku, Finland. He is a former Research Fellow of the Academy of Finland. His research interests are in the computational biomodeling of biochemical systems. Among his recent research projects are studies on gene assembly in ciliates, the eukaryotic heat shock response, and the self-assembly of intermediate filaments.

Israel Journal of Chemistry

2. Methods for Model Decomposition

2.1 Knockdown Mutants

A simple model decomposition consists of isolating a single process or mechanism in the considered system. In this way the model is split into two parts, the first one comprising the process of interest and the second containing all the remaining elements of the system. Although such decomposition might seem unsophisticated, this approach is often very useful in discovering the role of a single mechanism in a larger system. It is widely exploited in reverse engineering, a process aiming at revealing the technological principles of a device, object, or system. In Section 3 we briefly describe the method of mathematically controlled comparison,[2] where this simple decomposition approach is at the basis of the method.

2.2. Elementary Flux Modes

Another well-established decomposition method for biochemical models appears in the context of the analysis of metabolic pathways. It is not easy to define a pathway in a given metabolic network. An intuitive definition of a pathway is a sequence of reactions linked by common metabolites.[16] Examples of metabolic pathways are glycolysis or amino acid synthesis. Discovering new pathways in a large model driven only by biological intuition is even more difficult. An attempt to formalize the notion of pathway has been proposed in refs. [17–22] in the form of elementary flux modes. The intuitive meaning of an elementary flux mode is a set of reactions whose combined quantitative contribution to the system is zero. In other words, the net loss of substance caused by any reaction in that set is compensated by a net gain in the same substance incurred by some other reactions in the set. A formal definition of elementary flux modes is beyond the scope of this paper; instead we refer the reader to refs. [16,17,19–22] for details. For any given metabolic network, the full set of elementary fluxes can be determined using methods of linear algebra or dedicated software such as METATOOL.[18] Recognition of the elementary flux modes allows the detection of the full set of non-decomposable steady-state flows that the network can support, including cyclic flows. Any steady-state flux pattern can be expressed as a non-negative linear combination of these modes.[19–21] The identified elementary flux modes should have a clear biological interpretation: a flux mode is a set of enzymes that operate together at a steady state and a flux mode is elementary if the set of enzymes is minimal, that is, complete inhibition of any of the enzymes would result in a termination of this flux.[19,20,21] The lack of possibility to interpret the modes in this way is a signal that the model under consideration may not be correct.

2.3. Control-Based Decomposition

A control-driven approach to model decomposition enables the recognition of the main functional modules of a system and their individual contribution to the emergent, complex behaviors of the system as a whole. In turn, this can provide great insight about various properties of a given biochemical system, e.g., robustness, efficiency, reactivity, adaptation, regulation, synchronization, etc. In particular, by applying this approach, one usually aims to identify the main regulatory components of a given biochemical system: the process to be regulated, referred to as the *plant*; the *sensors* which monitor the current state of the process and send the collected information to a decision-making module, that is, the *controller*; and the *actuator* that modifies the state of the process in accordance with the controller's decisions, thus influencing the activity of the plant. One of the fundamental concepts in control theory is the *feedback mechanism*, which provides the means to cope with the uncertainties: the information about the current state of the process is sent back to the controller, which reacts accordingly to facilitate a dynamic compensation for any deviance from the intended behavior of the system. In the case of a complex system this decomposition can be performed in different ways depending on what is considered to be the main role of the system in question; that is, there may be a few reasonable choices for the plant, and the remaining components are recognized with respect to the choice of the plant.

An easy example illustrating these concepts and their interactions is given by the functioning principles of a motion-activated spotlight. Here, the controller module is an electronic unit which receives an input from the motion sensor and then determines whether there are any changes in the environment. The actuator is a relay switch that operates the lighting system. This actuator is activated by the controller depending on the input sent by the sensor. Then, the switch is kept on by the controller as long as movement is detected by the sensor.

How this control-driven approach can be exploited to investigate and understand regulatory networks can be seen in refs. [3,5,7,8,23]. Here we briefly describe the approach taken in ref. [23]. The authors make a thorough study of the heat shock response mechanism in *Escherichia coli* based on modular decomposition. A model for the system is built and functional modules — the plant, sensors, controller, and actuator — are identified. The decomposition reveals the underlying design of the heat shock response mechanism and its level of complexity, which, as the authors show, is not justified if only the functionality of an operational heat shock system is required. Further, this observation leads to the introduction and analysis of hypothetical design variants (mutants) of the original heat shock response model. In the original model one feedforward (temperature sensing) and two feedback elements (σ^{32} factor sequestration feedback

loop and σ^{32} degradation feedback loop) can be isolated. The variants are obtained through the elimination of either the σ^{32} degradation feedback loop or both feedbacks. Moreover, the case without the feedforward element is also considered, see ref. [23] for details. One by one the variants are considered in order of increasing complexity, starting from the simplest architecture containing just the feedforward element (the open-loop design). Based on numerical simulations, the authors demonstrate how the addition of subsequent layers of regulation, thereby increasing the complexity of the model, improves the performance of the response in terms of systemic properties such as robustness, noise reduction, speed of response, and economical use of cellular resources. Moreover, this systematic approach enables the identification of the contribution of each of the regulatory layers to the overall behavior of the system. In consequence the authors succeed in performing an in-depth comparison between different model variants.

3. Known Methods for Submodel Comparison

Comparing alternative models for a given biochemical system is, in general, a very difficult problem, involving a deep analysis of both the underlying network of reactions, the biological assumptions as well as the numerical setup. To decide the benefits of one design over another, or to understand the selection requirements involved in an evolutionary design, one needs some unbiased methods to objectively compare the alternative designs.

3.1. Mathematically Controlled Model Comparison

One such method is the mathematically controlled comparison,[2] which provides a structured approach for comparing alternative regulatory designs with respect to some chosen measures of functional effectiveness. Under this approach, mathematical models for both the reference design and the alternatives are first developed in the framework of canonical nonlinear modeling, referred to as S-systems.[24–26] This canonical nonlinear representation, developed within the power-law formalism, is a system of non-linear ordinary differential equations with a well-defined structure. Moreover, this framework allows the alternative models to differ from the reference design in only one process, e.g., the existence or absence of some feedback mechanisms, which is actually the focus of the comparison. Then, in each of the alternative models one sets the numerical values of the parameters to be identical with those from the reference model for all processes other than the process of interest. This leads to a so-called internal equivalence between the reference model and the alternatives. Next, various systemic properties are selected and used to impose some constraints for all the other parameters in the alternative designs. In general in this approach, one imposes that some steady state values or logarithmic gains are equal in the reference model and its alternatives. This provides a way to express the parameters of the process of interest in the alternative models as functions of the parameters of the reference model. Thus, one obtains a so-called external equivalence between the reference model and the alternative designs, meaning that to an external observer the considered models are equivalent with respect to the selected systemic properties. Finally, one chooses various measures of functional effectiveness depending on the particularities of the biological context of these models, and uses them to compare the alternative designs with the reference model. By doing this, one usually aims to determine analytically the qualitative differences between the compared models. This method was successfully used to compare alternative regulatory designs in, e.g., metabolic pathways,[27,28] gene circuits,[29] and immune networks.[30] Moreover, by introducing specific numerical values for the parameters of the models, one is also able to quantify these differences but, at the same time, the generality of the results is lost. Thus, in ref. [12], the method of mathematically controlled comparison was extended to include some statistical methods[11,31] that allow the use of numerical values for the parameters, while still preserving the generality of the conclusions.

3.2. An Extension of the Mathematically Controlled Comparison

The first step of this extension is to generate a representative ensemble of sets of parameter values. Since usually for biological systems the exact statistical distribution of the parameters values is not known, the most appropriate approach is to uniformly sample a given range of values. There exist different methods for scanning a given interval of values, ranging from (more or less sophisticated) random samplings to some systematic deterministic scanning methods; see, for example, ref. [32]. Using this ensemble of sets of parameters, we can then construct a large class of numerical models both for the reference and for the alternative designs. There are two different methods to construct such a class of systems for which we can then investigate some statistical properties. A structural class consists of systems having the same network topology, i.e., generated by the sampling of the parameter space. A behavioral class consists of systems that exhibit a particular systemic behavior, e.g., exhibiting a steady state behavior under given conditions, or low concentrations of intermediary products, or small values for the parameter sensitivity; see, for example, ref. [31]. The members of such a class are obtained in two steps: first generate a set of parameters by sampling the parameter space, then test the sample for the desired systemic behavior and keep only those systems that fulfill the conditions.

After constructing this large class of numerical models both for the reference and the alternative architectures,

Israel Journal of Chemistry

one can start comparing the values of a given systemic property P between the reference model and its alternative designs. One way to do this is by using density plots of the ratio $R = P_{reference}/P_{alternative}$ versus the values $P_{reference}$, where the subscript indicates in which model the property P was measured. Such density plots can be used, for instance, to compute rank correlations between the considered property P (measured in the reference model) and the values of the ratio R. However, this is not easy to do if the density plots are very scattered. Then, one can construct secondary density plots by using the moving median technique as follows. Basically, the density plot can be interpreted as a list of N pairs of values $(P_{reference}, R)$, which can be arranged in a ordered list L with respect to the first component, $P_{reference}$. Then, we pick a window size W, usually much smaller than the sample size N, and we compute the median $<R>$ of the ratio values and the median $<P>$ of the values $P_{reference}$, for the first W pairs in the list L. Then, we advance the window by one, we collect the ratios and the values $P_{reference}$ from the second until the $W+1$st pair, and compute the corresponding median values $<R>$ and $<P>$. This process is continued until the last pair in list L is used for the first time. In the secondary density plot, we pair the computed values $<R>$ with the corresponding $<P>$ values. This moving median technique is very useful since, for a finite ordered sample of size N, the moving median tends to the median of the samples as the value W approaches N. These secondary density plots can be used to compare the efficiency of two classes of models from the point of view of a given systemic property.

3.3 Local Submodel Comparison

When the alternative designs are actually submodels of the reference architecture, there is also another approach for performing the comparison; see ref. [15]. This is the case when, for instance, one is interested in a functional analysis of various modules of a large system. Then, the underlying reaction networks in the alternative designs are very similar (although not identical), and both the biological constraints and the kinetics of the reactions are given by those of the reference model. The only remaining question regards the initial distribution of the variables in the alternative models. In a mathematically controlled comparison they are usually taken from the reference model. However, for some biochemical systems this choice might lead to biased comparisons. For instance, in the case of regulatory networks, models should be in a steady state in the absence of the trigger of the response and, indeed, the initial values of the reference model are usually chosen in such a way as to fulfill this condition. However, this will not imply in general that also a submodel is in its steady state if it uses the same initial values as the reference model. Thus, the dynamic behavior of the submodel will be the result of two intertwined tendencies: migrating from a possible unstable state, and the response to a trigger. If the focus of the comparison is specifically the efficiency of the response of various submodels to a trigger, then the approached proposed in ref. [15] is more appropriate, yielding biologically unbiased results. In this approach, the initial distribution of the reactants is chosen in such a way that the initial setup of each submodel constitutes a steady state of that design in the absence of a trigger.

3.4. A Discrete Approach for Comparing Continuous Submodels

The application of the control–theoretical analysis described in Section 2 enables the identification of the main functional modules, their interconnections, and the control strategies of a biochemical network. In particular, this approach can be very useful for identifying the main regulatory components of a biochemical network, including its feed-forward and feedback mechanisms. Then, in order to identify and quantify the exact role of each of these regulatory mechanisms, one usually uses knockdown mutants[23] lacking one or more of these components. In particular, the knockdown mutant models are submodels of the reference architecture. The approached proposed in ref. [33] associates to each knockdown mutant a Boolean formula describing its control architecture in the following way. First, a Boolean variable is associated to each of the regulating mechanisms. Then, using the negation and conjunction of Boolean variables, one can write a Boolean formula for each of the knockdown mutants describing which of the regulating mechanisms are present in their architecture. In particular, these Boolean formulas describe a property of the alternative designs which is independent of time, i.e., their regulatory network. Moreover, one can go one step further and write a Boolean formula describing all those mutant architectures that show a given behavioral property, e.g., a high level of a given reactant or a given correlation between two reactants. This formula is actually the conjunction of all Boolean formulas characterizing the architectures of the mutants exhibiting the required property. The numerical comparison of the mutants is then performed by analyzing the Boolean formulas associated with various behavioral properties.

4. A New Approach for Quantitative Submodel Comparison

Here we propose a new approach for quantitative comparison of biological models. Before presenting the method itself, we clarify the adopted terminology which is used in the description of our new approach. Usually biological models are expressed in terms of biochemical reactions. We will refer to a list of such reactions describ-

ing a biological mechanism as its biochemical model. From the biochemical model, an associated mathematical model is derived by choosing one of the two commonly used frameworks: either a deterministic or a stochastic formulation. In the first case, the biochemical reaction kinetics rely on the assumption that the reaction rate at a certain point in time and space can be expressed as a unique function of the concentrations of all substances at this point in time and space.[16] It is governed by the *mass action law*, which can be briefly summarized as follows: the rate of each reaction is proportional to the product of the reactant masses, with each mass raised to the power equal to the corresponding stoichiometric coefficient.[16] With this assumption, the mathematical formulation of a biochemical model results in a system of ordinary differential rate equations constituting the associated deterministic mathematical model. In the second case, single molecules and their interactions are considered, and the changes in molecular populations are described in terms of stochastic processes. In the stochastic framework the associated mathematical model is a continuous-time Markov chain, defined by a chemical master equation describing the time evolution of the probability of the biochemical system to be in a certain state. Our new approach for model comparison is designed and presented for the deterministic framework; however, it can be easily adapted for the stochastic formulation.

As mentioned above, the assumption of the mass action kinetics leads to a system of ordinary differential equations (ODE) constituting the mathematical model. The ODE system contains a certain number of parameters representing the kinetic rate constants of the biochemical reactions. By assigning numerical values to the parameters and setting the initial conditions for the equations, we obtain an *instantiation* of the mathematical model.

Our model comparison method can be outlined as follows. First, starting with a biochemical model of some biological mechanism, referred to as the *reference model* (or *reference architecture*) of this system, we construct a *submodel* (or *alternative architecture*) by eliminating certain reactions from the list of biochemical reactions of the reference model. At this stage, we can for example apply control-based decomposition techniques to identify a number of modules, and then study them separately by considering a number of knockdown mutants lacking one or more of the modules. Second, the associated mathematical models are formulated, both for the reference and the alternative architecture. Notice that this procedure assures that all the parameters of the alternative architecture match a subset of parameters of the reference model. Next, we perform the statistical sampling of the reference model and mutant behavior. To accomplish this, we scan the parameter value space of the reference model. This provides us with a set of parameter value vectors. Each coordinate of these vectors is associated with one of the parameters in the reference model, and determines the value of the corresponding parameter. We consider each of the vectors one by one. We set the parameters of the reference model and the submodel in accordance with the considered vector. Since, as mentioned above, the alternative architecture contains only a subset of the reference model parameters, only the values of certain coordinates are used when setting the parameters of the submodel. Further, the initial values of the variables of the reference model and the submodel are determined independently of each other by a systemic property, such as the system being in a steady state in a given setup. For example, in the general case of stress response, we expect in accordance with biological observations that a feasible mathematical model is in a steady state under the unstressed, physiological conditions. We call steady state a numerical configuration of the model (given by numerical values for all variables and parameters of the model) such that starting from that configuration, the model shows no change in the level of any of the variables. In other words, the net loss per unit of time in every variable is exactly compensated by the net gain per unit of time in that variable. The steady states of a model are defined by the values of its parameters and by the initial values of its variables. Assuring that both mathematical submodels satisfy such systemic properties makes them suitable to be considered as viable alternative formal descriptions of the biological mechanism being analyzed. As a result, we obtain the instantiations of the reference model and the submodel and we run numerical simulations for both of them in order to evaluate their functional effectiveness. Finally, having done this for all sampled vectors, we summarize the obtained results for the variants and compare the models by use of some statistical measures. Moreover, the methodology allows us to consider more than one submodel, and thus the obtained results provide a basis for comparison between the different potential architectural designs underlying the analyzed biological mechanism.

For the parameter scanning, in the above procedure we use the Latin hypercube sampling method (LHS) originally introduced in ref. [34]. It provides samples which are uniformly distributed over each parameter while the number of samples is independent of the number of parameters. The sampling scheme can be briefly described as follows: First, the desired size N of the sampling set is chosen. Next, the range interval of each parameter is partitioned into N non-overlapping intervals of equal length. For each parameter, N numerical values are randomly selected, one from each interval of the partition according to a uniform distribution on that interval. Finally, the N sampled values for the i-th parameter of the model are collected on the i-th column of a $N \times p$ matrix, where p is the number of model parameters and the values in each column are shuffled randomly. As a result, each of the N rows of the matrix contains numerical values for each of

the p parameters. For a detailed description of this sampling scheme we refer the reader to ref. [13,34]; see also ref. [14] for an example of the application of this sampling method in the context of model identifiability problem.

In the next sections we show how the described method, where the sampling is performed with the LHS approach, can be utilized in the case of a recently introduced mathematical model for the eukaryotic heat shock response. In particular, we present how this method makes it possible to discriminate between different variants of the model and to determine the roles of certain control mechanisms of the response system.

5. Case Study

5.1 A Biochemical Model for the Heat Shock Response

The heat shock response (HSR) is a highly evolutionarily conserved defense mechanism among organisms.[35] It serves to prevent and repair protein damage induced by elevated temperature and other forms of environmental, chemical, or physical stress. Such conditions induce the misfolding of proteins, which in turn accumulate and form aggregates with disastrous effect for the cell. In order to survive, the cell has to abruptly increase the expression of heat shock proteins. These proteins operate as intra-cellular chaperones, that is, play a crucial role in folding of proteins and re-establishment of proper protein conformation. They prevent the destructive protein aggregation. We discern two main reasons that account for the strong interest in the heat shock response mechanism observed in recent years.[36–38] First, as a well-conserved mechanism among organisms, it is considered a promising candidate for disentangling the engineering principles fundamental for any regulatory network.[23,39–41] Second, besides their functions in the HSR, heat shock proteins have fundamental importance to many key biological processes such as protein biogenesis, dismantling of damaged proteins, activation of immune responses, and signaling.[42,43] In consequence, a thorough insight into the HSR mechanism would have significant implications for the advancement in understanding the cell biology.

In order to coherently investigate the HSR a number of mathematical models has been proposed in the literature.[23,44–47] In this study we consider a recently introduced model of the eukaryotic heat shock response.[14,48] In this model the central role is played by the heat shock proteins (**hsp**), which act as chaperones for the misfolded proteins (**mfp**): the heat shock proteins sequester the misfolded proteins (**hsp : mfp**) and help the misfolded proteins to regain their native conformation (**prot**). The defense mechanism is controlled through the regulation of the transactivation of the **hsp**-encoding genes. The transcription is initiated by heat shock factors (**hsf**), some specific proteins which first form dimers ($\mathbf{hsf_2}$), then trimers ($\mathbf{hsf_3}$), and in this configuration bind to the heat shock elements (**hse**), that is, certain DNA sequences in the promotor regions of the **hsp**-encoding genes. Once the trimers bind to the promoter elements ($\mathbf{hsf_3 : hse}$), the transcription and translation of the **hsp**-encoding genes boosts and, in consequence, new heat shock protein molecules get synthesized at a substantially augmented rate.

When the amount of the heat shock proteins reaches a high enough level to enable coping with the stress conditions, the production of new chaperone molecules is switched off by the excess of the heat shock proteins. To this aim **hsp** form complexes with the heat shock factors (**hsp : hsf**) in three independently and concurrent processes: 1) by binding to the free **hsf**, 2) by breaking the dimers and trimers, and 3) by breaking the $\mathbf{hsf_3 : hse}$, as a result of which the trimer gets unbound from the DNA, it is decomposed into three free **hsf** molecules and one of these **hsf** molecules forms a complex with **hsp**. This terminates the enhanced production of new heat shock protein molecules and blocks the formation of new **hsf** trimers. As soon as the temperature increases, proteins present in the cell start misfolding. The misfolded proteins titrate **hsp** away from the **hsp : hsf** complexes. This enables the accumulation of free **hsf** molecules, which in turn form trimers and promote the production of new chaperones. In consequence the response mechanism gets switched on. The full list of biochemical reactions constituting the biochemical model from ref. [14] is presented in Table 1. The model is based only on well-documented reactions, without introducing any hypothetical mechanisms or experimentally unsupported biochemical reactions. For a full presentation and discussion of this model we refer the reader to ref. [14].

Based on the assumption of mass-action law for all the Reactions (1)–(12), an associated mathematical model of the eukaryotic heat shock response is obtained. The resulting mathematical model is expressed in terms of ten first-order, ordinary differential equations. The full ODE system is shown in Table 2, where by k_i we denote the reaction rate constant of the irreversible reaction (i) in Table 1, by k_i^+ the rate constant associated with the "left-

Table 1. The list of reactions of the biochemical model for the heat shock response originally introduced in ref. [14].

Reaction	(Reaction number)
$2\,hsf \leftrightarrow hsf_2$	(1)
$hsf + hsf_2 \leftrightarrow hsf_3$	(2)
$hsf_3 + hse \leftrightarrow hsf_3 : hse$	(3)
$hsf_3 : hse \rightarrow hsf_3 : hse + hsp$	(4)
$hsp + hsf \leftrightarrow hsp : hsf$	(5)
$hsp + hsf_2 \rightarrow hsp : hsf + hsf$	(6)
$hsp + hsf_3 \rightarrow hsp : hsf + 2\,hsf$	(7)
$hsp + hsf_3 : hse \rightarrow hsp : hsf + hse + 2\,hsf$	(8)
$hsp \rightarrow$	(9)
$prot \rightarrow mfp$	(10)
$hsp + mfp \leftrightarrow hsp : mfp$	(11)
$hsp : mfp \rightarrow hsp + prot$	(12)

Table 2. The system of differential equations of the mathematical model associated with the biochemical model in Table 1.

Equation		(Equation number)
$d[hsf]/dt =$	$-2k_1^+[hsf]^2+2k_1^-[hsf_2]-k_2^+[hsf][hsf_2]+k_2^-[hsf_3]$ $-k_5^+[hsf][hsp]+k_5^-[hsp:hsf]+k_6[hsf_2][hsp]$ $+2k_7[hsf_3][hsp]+2k_8[hsf_3:hse][hsp]$	(13)
$d[hsf_2]/dt =$	$k_1^+[hsf]^2-k_1^-[hsf_2]-k_2^+[hsf][hsf_2]+k_2^-[hsf_3]$ $-k_6[hsf_2][hsp]$	(14)
$d[hsf_3]/dt =$	$k_2^+[hsf][hsf_2]-k_2^-[hsf_3]-k_3^+[hsf_3][hse]+k_3^-[hsf_3:hse]$ $-k_7[hsf_3][hsp]$	(15)
$d[hse]/dt =$	$-k_3^+[hsf_3][hse]+k_3^-[hsf_3:hse]+k_8[hsf_3:hse][hsp]$	(16)
$d[hsf_3:hse]/dt =$	$k_3^+[hsf_3][hse]-k_3^-[hsf_3:hse]-k_8[hsf_3:hse][hsp]$	(17)
$d[hsp]/dt =$	$k_4[hsf_3:hse]-k_5^+[hsf][hsp]+k_5^-[hsp:hsf]-k_6[hsf_2][hsp]$ $-k_7[hsf_3][hsp]-k_8[hsf_3:hse][hsp]-k_{11}^+[hsp][mfp]$ $+(k_{11}^-+k_{12})[hsp:mfp]-k_9[hsp]$	(18)
$d[hsp:hsf]/dt =$	$k_5^+[hsf][hsp]-k_5^-[hsp:hsf]+k_6[hsf_2][hsp]$ $+k_7[hsf_3][hsp]+k_8[hsf_3:hse][hsp]$	(19)
$d[mfp]/dt =$	$\phi_T[prot]-k_{11}^+[hsp][mfp]+k_{11}^-[hsp:mfp]$	(20)
$d[hsp:mfp]/dt =$	$k_{11}^+[hsp][mfp]-(k_{11}^-+k_{12})[hsp:mfp]$	(21)
$d[prot]/dt =$	$-\phi_T[prot]+k_{12}[hsp:mfp]$	(22)

to-right" direction of the reversible reaction (i), while k_i^- denotes the rate constant corresponding to its "right-to-left" direction. By T we denote the numerical value of the temperature of the environment in degrees Celsius. The rate coefficient of protein misfolding with respect to the temperature ($\varphi(T)$) in Reaction (10) is given by the following formula:

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

which is valid for T in the range from 37 to 45. The formula was obtained based on experimental investigations described in refs. [49,50], originally proposed in ref. [45], and adapted for use in the mathematical model of HSR in ref. [14]. The mathematical model comprises 16 independent kinetic parameters and 10 initial conditions. In the case of our method, we do not fix the parameter values as was done in ref. [14]: we neither fit nor validate the model with respect to experimental data. Instead, we sample the HSR model behavior by randomly choosing different sets of parameter values. This results in not one, but a collection of instances of the HSR model. Notice that in the process of obtaining these instances, no experimental data are considered. Thus, the instances are not required to confirm any experimental results. We discuss in details how the parameter values for the HSR model are obtained in Subsection 5.4.

5.2 Control-Based Decomposition

In ref. [15], a control-driven modular decomposition of the heat shock response model was performed. As a result, the model has been divided into four main functional submodules usually distinguished in control engineering: the plant, the sensor, the controller, and the actuator. In the case of the HSR model the plant is the misfolding and refolding of proteins, the actuator consists of the synthesis and degradation of the chaperones, the sensor measures the level of **hsp** in the system, and the controller regulates the level of DNA binding. Moreover, within the controller we distinguish three feedback mechanisms. The feedback loops are responsible for sequestering the heat shock factors in different forms by the chaperones. In this way, the feedback loops decrease the level of DNA binding. The three identified feedback mechanisms are the following:

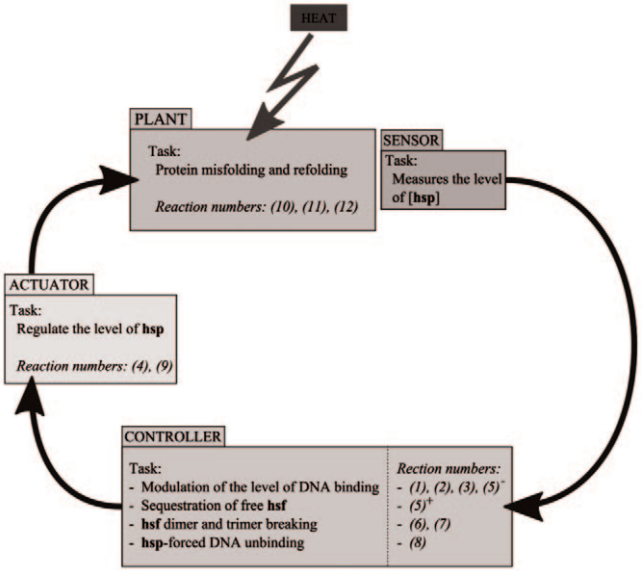


Figure 1. The control-based decomposition of the heat shock response network. The reaction numbers refer to the reactions in Table 1. We denote the "left-to-right" direction of reaction (5) by $(5)^+$ and by $(5)^-$ its "right-to-left" direction.

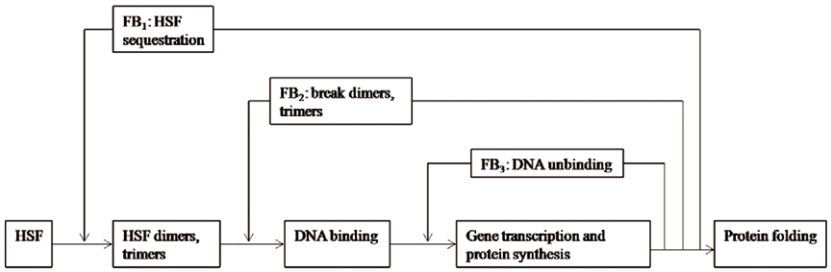


Figure 2. The control structure of the heat shock response network. The three identified feedback loops and their points of interaction with the mainstream process are depicted.

- FB1: Sequestration of free **hsf**, that is, Reaction $(5)^+$ (the "left-to-right" direction of Reaction (5));
- FB2: Breaking of **hsf** dimers and trimers, that is, Reactions (6) and (7);
- FB3: Unbinding of $\mathbf{hsf_3}$ from **hse** and breaking the trimers, that is, Reaction (8).

The control-driven functional decomposition of the eukaryotic heat shock response model is shown in Figure 1, where the reaction numbers refer to the reactions in Table 1. In Figure 2 a graphical illustration of the control structure, that is, the three feedback loops and their points of interactions with the mainstream process, is presented.

5.3 The Knockdown Mutants

In refs. [15] and [33], the reference architecture and seven knockdown mutants (alternative architectures) were considered. The mutants were obtained by eliminating from the reference architecture all possible combinations of the three feedback loops FB1, FB2, and FB3. The mutants were denoted as M_X, where $X \subset \{1,2,3\}$ is the set of numbers of the feedback mechanisms present in M_X:

- M_0 is determined by Reactions (1)–(4), (9)–(12) and, in the terminology of control theory, is characterized by the *open-loop design*;
- M_1 is determined by Reactions (1)–(5), (9)–(12);
- M_2 is determined by Reactions (1)–(4), (6)–(7), (9)–(12), and the "right-to-left" direction of reaction (5);
- M_3 is determined by Reactions (1)–(4), (8)–(12), and the "right-to-left" direction of reaction (5);
- $M_{1,2}$ is determined by Reactions (1)–(7), (9)–(12);
- $M_{1,3}$ is determined by Reactions (1)–(5), (8)–(12);
- $M_{2,3}$ is determined by Reactions (1)–(4), (6)–(12), and the "right-to-left" direction of reaction (5);
- $M_{1,2,3}$ is the reference architecture consisting of all Reactions (1)–(12).

5.4 Statistical Sampling of the Mutant Behavior

We apply our model comparison method described in Section 3 to the presented model of eukaryotic heat shock response in order to investigate the functional role of the feedback mechanisms. It is easy to see that M_0 is non-responsive: starting from a steady state at physiological conditions, that is, 37 °C, M_0 shows no increase in DNA binding for any arbitrarily high temperature; see ref. [33]. We remove M_0 from further considerations. In our study we analyze the six knockdown mutants M_1, M_2, M_3, $M_{1,2}$, $M_{1,3}$ and $M_{2,3}$ as the variants of the reference architecture $M_{1,2,3}$. Our comparison method is applied in the following way. First, a sample of 10,000 vectors of parameter values for the reference architecture is obtained by the Latin hypercube sampling described above. In our case, the sampled vectors are of length 15, the number of unknown reference architecture parameters. The value of the 16th remaining parameter, that is, the degradation rate constant, is assumed to be known and is obtained based on the fact that heat shock proteins are generally long-lived proteins.[51] Here we choose their half-life to be 6 h. Then, the procedure described next is repeated separately for each of the six mutants. To begin with, each sampled vector of parameter values is used to set up the parameters in the mathematical models of the considered mutant and the reference architecture ($M_{1,2,3}$). It follows from the construct of the mutant that the corresponding mathematical model contains only a subset of the parameters of the reference model, so this step can be performed. Next, the steady state concentrations at 37 °C both for the mutant and the reference model are numerically computed and set as their respective initial states. In this way we obtain two instances of the mathematical models, that is, one for the mutant and the second for the reference model. Further, the temperature is increased to 42 °C and the quantities

$$\Theta_1 = \max_{t\in[0s,1800s]} (\text{total mfp}(t))$$

$$\Theta_2 = \max_{t\in[0s,1800s]} (\text{hsf}_3 : \text{hse}(t) - \text{hsf}_3 : \text{hse}(0))$$

$$\Theta_3 = \frac{1}{T}\int_0^T (\text{total hsp}(t))\mathrm{d}t$$

$$\Theta_4 = \frac{1}{T}\int_0^T (\text{total mfp}(t))\mathrm{d}t$$

are computed both for the mutant and the reference instance. The initial 30 min of the response are considered for the computation of Θ_1 and Θ_2. In the case of Θ_3 and Θ_4 the time range of 4 h ($T = 14400\,s$) is taken into account. These quantities are used to evaluate the functional effectiveness of the mutant. Having these quantities computed for all the 10,000 sampled parameter values, the scatter plot of the $R_1 = \Theta_1^m/\Theta_1^r$ against Θ_1^r values is made, where the superscripts m and r indicate the instance for which Θ_1 was computed, that is, the instance of the mutant or the reference model, respectively. Finally, the moving median technique is applied to the scatter plot, with the window size set to 500. These result in a trend curve summarizing the data of the scatter plot and revealing the overall dependency between the considered quantities. Analogical plots are computed for $R_2 = \Theta_2^m/\Theta_2^r$. Moreover, scatter plots of Θ_3 versus Θ_4 are made both for the mutant and the reference architecture, and the moving median technique is applied to each of these plots.

The mutants represent six different potential architectures of the heat shock response mechanism and the sampling procedure, as explained above, provides us with 10,000 different instantiations of each of the mutants and the reference architecture.

5.5 Results

In our analysis of the obtained results we assume that the heat shock response at raised temperatures is accompanied, and hence characterized, by the following three phenomena:

1. Increase in DNA-binding with respect to the steady-state level at 37 °C,
2. Increase in the level of **mfp**, and
3. Increase in the level of **hsp** as the response to the higher level of **mfp** in the cell.

We base our analysis of the architecture properties of the six mutants with respect to the reference architecture on the following plots: R_1 vs Θ_1^r, R_2 vs Θ_2^r, Θ_3 vs Θ_4 made for each of the mutants. We refer to the Θ_3 vs Θ_4 plot as the cost plot (or simply the cost) of the corresponding architecture. This is motivated by the fact that the efficiency of the heat shock response mechanism could be measured by the amount of chaperones needed to cope with the intensified misfolding of proteins. Hypothetically, a cell which produces smaller amounts of **hsp** than some other cell to cope with the heat shock would be considered the one which manages stress conditions at a lower cost in terms of its resources than the latter one. Notice, however, that in our case we are not assessing the ability of particular models to cope with heat shock. That is, the sampled models are neither validated against experimental data nor classified by any other means as to whether they enable the cell to survive or not under the stress conditions. Hence the cost plots reflect just the general tendency of the models instantiating a particular architecture to keep certain average in time amounts of **hsp** in response to different average levels of **mfp** present in the system. The reference trend line indicates a clear linear dependency between the average levels of **hsp** and **mfp**; see Figure 3. The trend lines of all mutants, despite some more or less pronounced fluctuations in the region of small Θ_4 values, can be seen as increasing (Figure 4), which is in agreement with our characterization of the heat shock response.

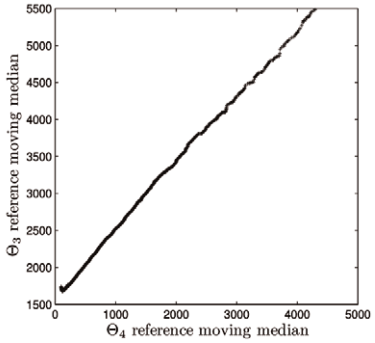


Figure 3. The plot shows the result of applying the moving median technique to the scatter plots of the cost, that is, Θ_3 versus Θ_4, obtained for the reference architecture. For each sampled vector of parameters, the values of Θ_3 and Θ_4 were computed and plotted against each other. Then, the moving median technique was applied to discern the overall trend in the data depicted in the obtained scatter plot. The window size of the moving median was set to 500 and the sample size of the vectors of parameter values was 10,000.

Considering the three mutants with only one feedback — M_1, M_2, and M_3 — we observe that the **mfp** level peak value in the first 30 min. of heat shock is smaller than in the reference architecture: the ratio R_1 in Figure 5 a, b, and c is always smaller than 1. This is especially pronounced in mutant instances obtained with samples characterized by high **mfp** peak values in the case of the reference architecture. However, for all these mutants the cost is definitely higher than in the reference architecture; compare Figure 4 a, b, and c with Figure 3. Notice also that the M_2 mutant is more economic in terms of cost than the two other mutants with only one feedback.

In the mutants $M_{1,2}$, $M_{1,3}$ and $M_{2,3}$ the **mfp** level also peaks at a lower value than in the reference case, al-

Israel Journal of Chemistry

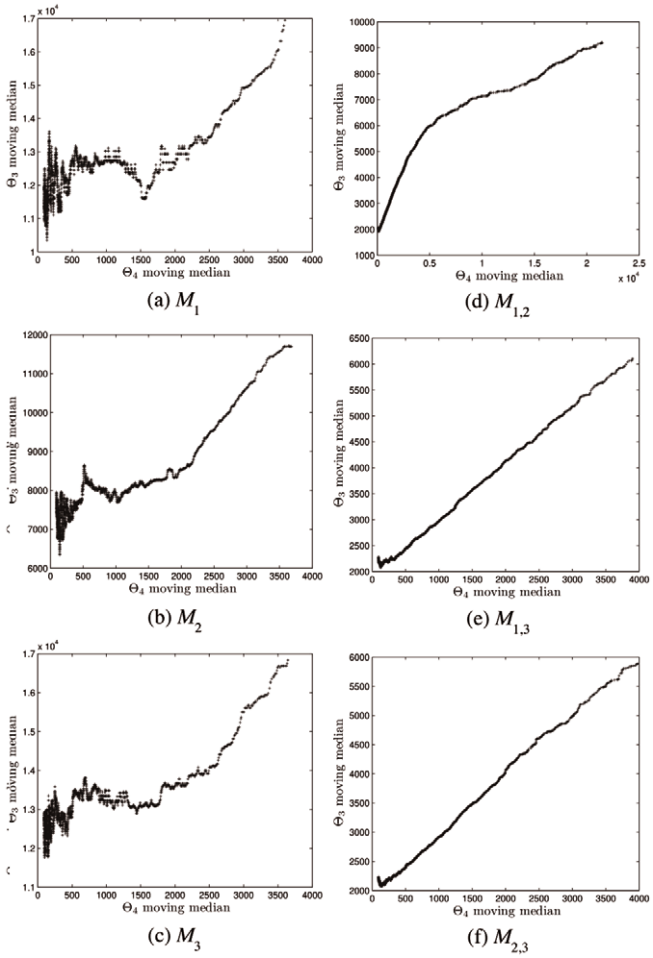


Figure 4. The plots show the result of applying the moving median technique to the scatter plots of the cost, that is, Θ_3 versus Θ_4, obtained individually for each of the six considered mutants. For each mutant and each sampled vector of parameters, the values of Θ_3 and Θ_4 were computed and plotted against each other. Then, the moving median technique was applied to discern the overall trend in the data depicted in the obtained scatter plots. The window size of the moving median was set to 500 and the sample size of the vectors of parameter values was 10,000.

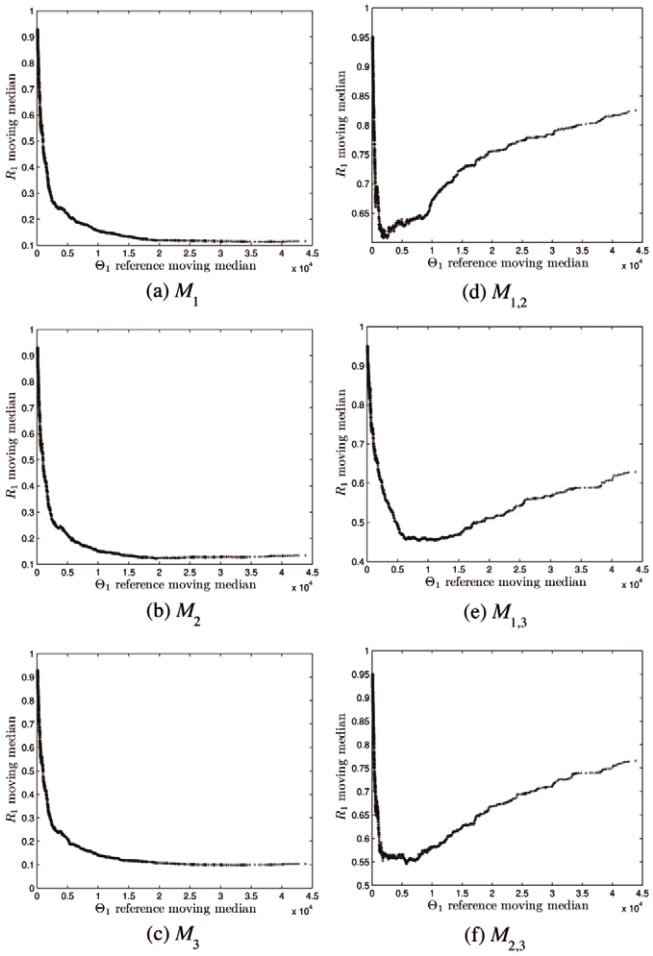


Figure 5. The plots show the result of applying the moving median technique to the scatter plots of R_1 vs Θ_1^r obtained individually for each of the six considered mutants. For each mutant and each sampled vector of parameters, the value of R_1 was computed and plotted against the value of Θ_1 obtained for the reference architecture with the same parameter vector. Then, the moving median technique was applied to discern the overall trend in the data depicted in the obtained scatter plots. The window size of the moving median was set to 500 and the sample size of the vectors of parameter values was 10,000.

though this time the $M_{1,3}$ and $M_{2,3}$ mutants have the cost comparable with the one of the reference architecture. Both $M_{1,3}$ and $M_{2,3}$ reveal the same linear relationship between the average amounts of **hsp** and **mfp** as is observable in the reference case; however, in both cases the trend line is slightly shifted upwards with respect to the reference. This indicates that the mutants have a tendency to keep a bit higher amount of **hsp** than the reference with a certain amount of misfolded proteins (Figure 4e,f and Figure 3). The same is true also for the $M_{1,2}$ mutant. Although it admits an order of magnitude larger range of observable average amounts of misfolded proteins (Figure 4d), the cost plot restricted to Θ_4 moving median values ≤ 4000 is basically identical with the cost plots of the two other mutants, see Figure 6.

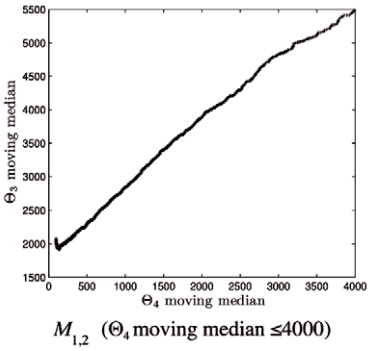


Figure 6. The zoomed in version of Figure 4d where Θ_4 moving median is not greater than 4000.

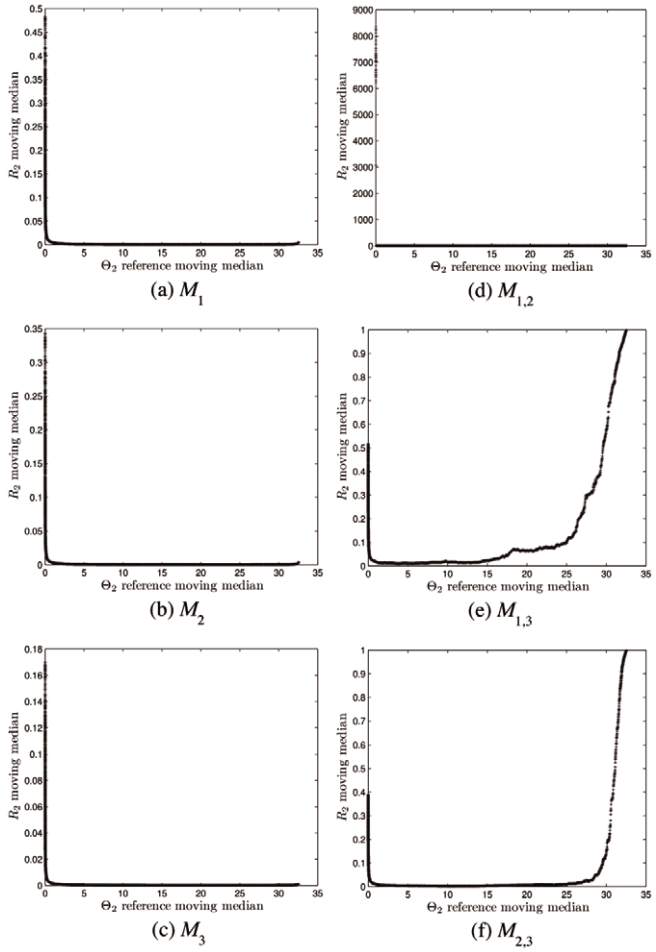


Figure 7. The plots show the result of applying the moving median technique to the scatter plots of R_2 vs $\Theta_2{}^r$ obtained individually for each of the six considered mutants. For each mutant and each sampled vector of parameters, the value of R_2 was computed and plotted against the value of Θ_2 obtained for the reference architecture with the same parameter vector. Then, the moving median technique was applied to discern the overall trend in the data depicted in the obtained scatter plots. The window size of the moving median was set to 500 and the sample size of the vectors of parameter values was 10,000.

Another thing which we observe for the three mutants missing one feedback is that the samples characterized by significant increase in DNA-binding in the reference architecture, that is, by 15 and more, span a wide range of possible behaviors in the mutants: from almost no DNA-binding increase to an increase comparable with the one observed for the reference architecture. This feature is clearly visible in Figure 7e and f for the mutants $M_{1,3}$ and $M_{2,3}$, respectively. In the case of the $M_{1,2}$ mutant we need to zoom in Figure 7d. To this aim we observe in the scatter plot R_2 vs $\Theta_2{}^r$ for the $M_{1,2}$ mutant that all points with $R_2 > 1000$ are concentrated in the range [0,0.0307] of $\Theta_2{}^r$ values (not shown). We exclude all samples with $\Theta_2{}^r$ in

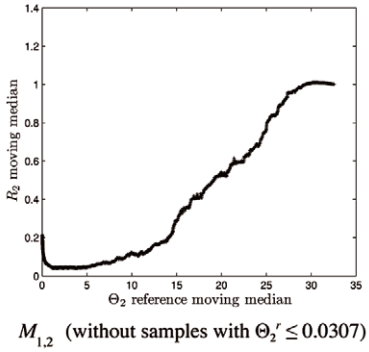


Figure 8. A version of Figure 7d where samples with $\Theta_2{}^r \leq 0.0307$ were not considered. It shows that the samples characterized by significant increase in DNA-binding in the reference architecture (by 15 and more) span a wide range of possible behaviors in the $M_{1,2}$ mutant: from almost no DNA-binding increase (the moving median of $R_2 = 0.2$) to an increase comparable with the one observed for the reference architecture (the moving median of $R_2 \geq 1$).

this range, irrespective of the R_2 value they admit in the mutant. All in all, 2247 samples are filtered out and we apply the moving median technique to the remaining ones. The resulting plot is shown in Figure 8. It clearly illustrates that the discussed feature is also a characteristic of the $M_{1,2}$ mutant. This is not true for the three mutants with only one feedback. In these cases we do not observe any substantial increase in the DNA-binding with respect to the steady-state levels at 37 °C for samples which generate such increase in the reference case (Figure 7a, b, and c).

On the basis of the presented results, we notice that all the mutants lacking two feedbacks exhibit no heat shock response in the sense of the above definition: as observed previously, there is no increase in the DNA-binding. This is in agreement with the results presented in ref. [15], where the models with only one feedback kept the DNA-binding at the maximum possible level both at 37 °C and 42 °C throughout the simulation time of 50,000 s. The HSR can be observed, however, in the mutants $M_{1,3}$ and $M_{1,2}$. In the case of the $M_{2,3}$ mutant the HSR is still observed, but only for a fraction of the 10,000 sampled models, that is, only those parameter values for which the reference architecture displays the maximal possible increase in the peak of DNA-binding with respect to the steady-state level at 37 °C. This is in complete agreement with previous observations that FB1 is the most powerful feedback.[15] Since FB2 and FB3 include sequestration as one of their features, they compensate partially for the lack of FB1. However, only FB2 or only FB3 is not enough to enforce the system's behavior to have the HSR characteristics. Despite its power, FB1 alone is also not enough and one of the other feedbacks is also needed in order to implement a response mechanism with the features describing the heat shock response.

6. Discussion

Very often, various experimental investigations of a given biochemical system generate a large variety of alternative molecular designs, thus raising questions about comparing their functionality, efficiency, and robustness. Comparing alternative models for a given biochemical system is, in general, a very difficult problem which involves a deep analysis of various aspects of the models: the underlying networks, the biological constraints, and the numerical setup. The problem becomes somewhat simpler when the alternative designs are actually submodels of a larger model: the underlying networks are similar, although not identical, and the biological constrains are given by the larger model. It only remains to decide how to chose the numerical setup for each of the alternative submodels, i.e., the initial conditions and the kinetics.

In the first part of our study we review several known methods for model decomposition and for quantitative comparison of submodels. We describe knockdown mutants, elementary flux modes, control-based decomposition, mathematically controlled comparison and its extension, local submodels comparison and a discrete approach for comparing continuous submodels. In the second part of the paper we present a new statistical method for comparing submodels that complements the methods presented in the review. When choosing the initial setup for the alternative submodels, i.e., the initial values of all variables, one approach is to take them from the reference model. This approach is based on the technique of mathematically controlled comparison;[2] see also refs. [23] and [52] for some case studies using this method. However, in the case of biological systems this approach may lead to biased conclusions. For instance, regulatory networks exhibit a steady-state behavior in the absence of stimulus. In general for the reference model, the initial values of the variables are chosen such that it exhibits a steady state behavior in the absence of a trigger. However, the submodels of the reference model may not exhibit the same property when starting from the same initial values. Thus, the dynamic behaviors of the considered submodels will exhibit the intertwined influences of two tendencies: the migration from a (possibly) unstable state and the response to the stimulus. In this context, an analysis of the efficiency of the response and the robustness of the alternative models may lead to erroneous conclusions. As an alternative, we propose in this paper to chose the initial values in such a way that each alternative design starts from its own steady state. Our main motivation for this is that we considered all submodels to be viable alternatives for the biological system and, as such, they should exhibit (some of) its main properties. Regarding the values of the kinetic parameters in each of the alternative submodes, there are several approaches in the literature. In the mathematically controlled comparison approach, the values of the kinetic parameters in each of the alternative designs are uniquely determined from the parameters of the reference model; see, for example, refs. [2] and [23]. Another approach is to chose in each alternative submodel independent values for the kinetic parameters, e.g., through parameter estimation and validation against experimental data; see, for example, refs. [15]. However, restricting to some particular values for the kinetic rate constants will also confine the conclusions of our analysis to that particular system. Instead, we take the approach proposed in refs. [11] and [12] and we sample a large set of parameter values from a given range of values. Then we use some statistical techniques to analyze various properties of a general class of systems which includes the considered system. In particular, for each sampled parameter vector, various functional effectiveness measures are computed both in the reference and in the alternative models. Then by analyzing both the density of ratios plots and the moving median plots one can identify and quantify the differences in the dynamic behaviors of the considered models. See e.g., for example, refs. [31] and [53] for some case studies where these methods were applied.

Acknowledgments

The work of Elena Czeizler, Andrzej Mizera and Ion Petre was supported by Academy of Finland, grants 129863, 108421, and 122426. Andrzej Mizera is on leave of absence from the Institute of Fundamental Technological Research, Polish Academy of Sciences, Warsaw, Poland.

References

[1] W. W. Chen, B. Schoeberl, P. J. Jasper, M. Niepel, U. B. Nielsen, D. A. Lauffenburger, P. K. Sorger, *Mol. Syst. Biol.* **2009**, *5*, 239.
[2] M. A. Savageau, *Curr. Top. Cell. Regul.* **1972**, *6*, 63–130.
[3] M. E. Csete, J. C. Doyle, *Science* **2002**, *295*, 1664–1669.
[4] B. A. Hawkins, H. V. Cornell, *Theoretical Approaches to Biological Control*, Cambridge Univ. Press, Cambridge, UK, **1999**.
[5] H. Kitano, *Science* **2002**, *295*, 1662–1664.
[6] Y. Lazebnik, *Cancer Cell* **2002**, *2*, 179–182.
[7] E. D. Sontag, *IEE Proc. Syst. Biol.* **2004**, *1*, 9–18.
[8] E. D. Sontag, *European J. Control* **2005**, *11*, 396–435.
[9] J. Stelling, U. Sauer, Z. Szallasi, F. J. Doyle, J. Doyle, *Cell* **2004**, *118*, 675–685.
[10] O. Wolkenhauer, *Briefings Bioinf.* **2001**, *2*, 258–270.
[11] R. Alves, M. A. Savageau, *Bioinformatics* **2000**, *16*, 527–533.
[12] R. Alves, M. A. Savageau, *Bioinformatics* **2000**, *16*, 786–798.
[13] J. C. Helton, F. J. Davis, *Rel. Eng. and Systems Safety* **2003**, *81*, 23–69.
[14] I. Petre, A. Mizera, C. L. Hyder, A. Meinander, A. Mikhailov, R. I. Morimoto, L. Sistonen, J. E. Eriksson, R.-J. Back, *Natural Computing* **2011**, DOI: 10.1007/s11047-010-9216-y.

[15] El. Czeizler, E. Czeizler, R.-J. Back, I. Petre in *Computational Methods in Systems Biology, 7th International Conference, CMSB 2009* (Eds.: P. Degano, R. Gorrieri), Springer-Verlag, Heidelberg, **2009**, pp. 111–125.
[16] E. Klipp, R. Herwig, A. Kowald, C. Wierling, H. Lehrach, *Systems Biology in Practice. Concepts, Implementation and Application*, Wiley-VCH, Weinheim, **2005**.
[17] R. Heinrich, S. Schuster, *The Regulation of Cellular Systems*, Chapman & Hall, New York, **1996**.
[18] T. Pfeiffer, I. Sanchez-Valdenebro, J. C. Nuno, F. Montero, S. Schuster. *Bioinformatics* **1999**, *15*, 251–257.
[19] C. H. Schilling, S. Schuster, B. O. Palsson, R. Heinrich, *Biotechnol. Prog.* **1999**, *15*, 296–303.
[20] S. Schuster, T. Dandekar, D. A. Fell, *Trends Biotechnol.* **1999**, *17*, 53–60.
[21] S. Schuster, D. A. Fell, T. Dandekar, *Nat. Biotechnol.* **2000**, *18*, 326–332.
[22] S. Schuster, C. Hilgetag, J. H. Woods, D. A. Fell, *J. Math. Biol.* **2002**, *45*, 153–181.
[23] H. El-Samad, H. Kurata, J. C. Doyle, C. A. Gross, M. Khammash, *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 2736–2741.
[24] M. A. Savageau, *J. Theor. Biol.* **1969**, *25*, 365–369.
[25] M. A. Savageau, *J. Theor. Biol.* **1969**, *25*, 370–379.
[26] M. A. Savageau, *J. Theor. Biol.* **1970**, *26*, 215–226.
[27] A. Hunding, *Biophys. Struct. Mech.* **1974**, *1*, 47–54.
[28] M. A. Savageau, *J. Mol. Evol.* **1974**, *4*, 139–156.
[29] W. S. Hlavacek, M. A. Savageau, *J. Mol. Biol.* **1996**, *255*, 121–139.
[30] R. J. De Boer, P. Hogeweg, *Bull. Math. Biol.* **1989**, *51*, 217–222.
[31] R. Alves, M. A. Savageau, *Bioinformatics* **2000**, *16*, 534–547.
[32] A. Saltelli, S. Tarantola, F. Campolongo, M. Ratto, *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models*, John Wiley & Sons Ltd, Chichester, UK, **2004**.
[33] El. Czeizler, A. Mizera, I. Petre, A Boolean logic-based approach for comparing biomodels, *unpublished data*, **2011**.
[34] M. D. McKay, R. J. Beckman, W. J. Conover, *Technometrics* **1979**, *21*, 239–245.
[35] S. Lindquist, E. A. Craig, *Annu. Rev. Genet.* **1988**, *22*, 631–677.
[36] Y. Chen, T. S. Voegeli, P. P. Liu, E. G. Noble, R. W. Currie, *Inflammation Allergy: Drug Targets* **2007**, *6*, 91–100.
[37] M. V. Powers, P. Workman, *FEBS Lett.* **2007**, *581*, 3758–3769.
[38] R. Voellmy, F. Boellmann, *Adv. Exp. Med. Biol.* **2007**, *594*, 89–99.
[39] H. El-Samad, S. Prajna, A. Papachristodoulou, M. Khammash, J. C. Doyle, *Proceedings of the 42th IEEE Conference on Decision and Control*, **2003**, pp. 3766–3771.
[40] H. Kurata, H. El-Samad, T.-M. Yi, M. Khammash, J. C. Doyle, *Proceedings of the 40th IEEE Conference on Decision and Control*, **2001**, pp. 837–842.
[41] C. J. Tomlin, J. D. Axelrod, *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 4219–4220.
[42] H. K. Kampinga, *J. Cell Sci.* **1993**, *104*, 11–17.
[43] A. G. Pockley, *Lancet* **2003**, *362*, 469–476.
[44] O. Lipan, J.-M. Navenot, Z. Wang, L. Huang, S. C. Peiper, *PLoS Comput. Biol.* **2007**, *3*, 1859–1870.
[45] A. Peper, C. A. Grimbergent, J. A. E. Spaan, J. E. M. Souren, R. van Wijk, *Int. J. Hyperthermia* **1997**, *14*, 97–124.
[46] T. R. Rieger, R. I. Morimoto, V. Hatzimanikatis, *Biophys. J.* **2005**, *88*, 1646–1658.
[47] R. Srivastava, M. S. Peterson, W. E. Bentley, *Biotechnol. Bioeng.* **2001**, *75*, 120–129.
[48] I. Petre, A. Mizera, C. L. Hyder, A. Mikhailov, J. E. Eriksson, L. Sistonen, R.-J. Back in *Algorithmic Bioprocesses* (Eds.: A. Condon, D. Harel, J. N. Kok, A. Salomaa, E. Winfree), Springer, Heidelberg, **2009**, pp. 411–425.
[49] J. R. Lepock, H. E. Frey, K. P. Ritchie, *J. Cell Biol.* **1993**, *122*, 1267–1276.
[50] J. R. Lepock, H. E. Frey, A. M. Rodahl, J. Kruuv, *J. Cell. Physiol.* **1988**, *137*, 14–24.
[51] A. M. Sapozhnikov, G. A. Gusarova, E. D. Ponomarev, W. G. Telford, **2002**, *35*, 193–206.
[52] M. E. Wall, W. S. Hlavacek, M. A. Savageau, *J. Mol. Biol.* **2003**, *332*, 861–876.
[53] J. H. Schwacke, E. O. Voit, *Theor. Biol. Med. Modell.* **2004**, 1.

Received: October 1, 2010
Accepted: November 7, 2010
Published online: January 27, 2011

Paper VII

Quantitative analysis of the self-assembly strategies of intermediate filaments from tetrameric vimentin

Eugen Czeizler, Andrzej Mizera, Elena Czeizler, Ralph-Johan Back, John E. Eriksson, and Ion Petre

TUCS Technical Report number 963, December 2009.

Quantitative analysis of the self-assembly strategies of intermediate filaments from tetrameric vimentin

Eugen Czeizler
Department of Information Technologies, Åbo Akademi University,
(eczeizle@abo.fi)
Andrzej Mizera
Department of Information Technologies, Åbo Akademi University,
(amizera@abo.fi)
Elena Czeizler
Department of Information Technologies, Åbo Akademi University,
(elena.czeizler@abo.fi)
Ralph-Johan Back
Department of Information Technologies, Åbo Akademi University,
(backrj@abo.fi)
John E. Eriksson
Turku Centre for Biotechnology and Department of Biochemistry,
Åbo Akademi University,
(john.eriksson@btk.fi)
Ion Petre
Department of Information Technologies, Åbo Akademi University,
(ipetre@abo.fi)

TUCS Technical Report

No 963, December 2009

Abstract

In vitro assembly of intermediate filaments from tetrameric vimentin consists of a very rapid phase of tetramers laterally associating into unit-length filaments and a slow phase of filament elongation. We focus in this paper on a systematic quantitative investigation of two molecular models for filament assembly, recently proposed in (Kirmse et al *J. Biol. Chem. 282*, 52 (2007), 18563–18572), through mathematical modeling, model fitting, and model validation. We analyze the quantitative contribution of each filament elongation strategy: with tetramers, with unit-length filaments, with longer filaments, or combinations thereof. In each case, we discuss the numerical fitting of the model with respect to one set of data, and its separate validation with respect to a second, different set of data. We introduce a high-resolution model for vimentin filament self-assembly, able to capture the detailed dynamics of filaments of arbitrary length. This provides much more predictive power for the model, in comparison to previous models where only the mean length of all filaments in the solution could be analyzed. We show how kinetic observations on low-resolution models can be extrapolated to the high-resolution model and used for lowering its complexity.

Keywords: Mathematical modeling — Protein self-assembly — Quantitative self-assembly strategies — Model resolution — Sensitivity analysis — Filament length distribution.

TUCS Laboratory
Computational Biomodelling

1 Introduction

The cytoskeleton of eukaryotic cells is an intricate network of protein filaments that extends throughout the cytoplasm. There are three types of protein filaments: *intermediate filaments* (IFs), *microtubules*, and *actin filaments*, [24]. Together with other proteins that attach to them, they form a system of girders, ropes, and motors that gives the cell its mechanical strength, controls its shape, and drives and guides its movements, see [17]. Compared with microtubules and actin filaments, IFs are more stable, tough and durable; in particular, IFs are the most insoluble part of the cell, see [8]. IFs have an important structural function in reinforcing the cells, organize cells into tissues, and most importantly, distribute the tensile forces across the cells in a tissue, see [17]. Major degenerative diseases of skin, muscle, and neurons are caused by disruptions of the IF cytoskeleton or its connections to other cell structures. Currently, around 80 diseases have been associated with the IF gene family, including various skin fragility disorders, as well as *laminopathies*, a family of afflictions caused by point mutations in the lamin A genes, [4, 5, 26]. A thorough understanding of the assembling principles of IFs can provide new insights on comprehending these abnormal conditions, as well as a better basis for diagnostic and possible treatment.

Contrary to the other protein filaments which are assembled from globular proteins, see [11, 25, 22], IFs subunits are α-helical rods that assemble into rope-like filaments [8]. Their assembly proceeds through a series of intermediate structures, which associate by lateral and end-to-end interactions. However, unlike in the case of microtubules and actin filaments where rich literature is available, the assembly principles of IFs are still poorly understood. We focus in this paper on the quantitative kinetic strategies for the *in vitro* assembly of IFs from human vimentin proteins (several other IF proteins exist, see [10]). On a first level of their assembly, vimentin proteins rapidly associate parallelly into dimers and then form anti-parallel, half-staggered tetramers, see [9] and Figure 1 (a)-(e). Tetramers then rapidly associate laterally to yield short filaments called *unit-length filaments* (ULFs) of the same length as the tetramers, see [8] and Figure 1 (f). On a second level of the assembly, the ULFs and the emerging longer filaments elongate longitudinally with tetramers, with ULFs, and with other filaments, [8] and Figure 2. On a third level, filaments undergo a radial compaction from an ULF diameter of about 15 nm to a filament diameter of about 11 nm, see [8] for details.

We investigate in this paper two molecular models (the so-called *simple* and *extended* models) introduced in [15] for the *in vitro* assembly of intermediate filaments from tetrameric vimentin. We perform a quantitative analysis of the predictive capabilities of these models. We construct two mass action-based mathematical models corresponding to the two molecular models. For each of them we consider several different knockdown mutant model variants where various combinations of assembly mechanisms are analyzed separately. We use COPASI [12] as a computational environment for the experimental data fitting (based on data

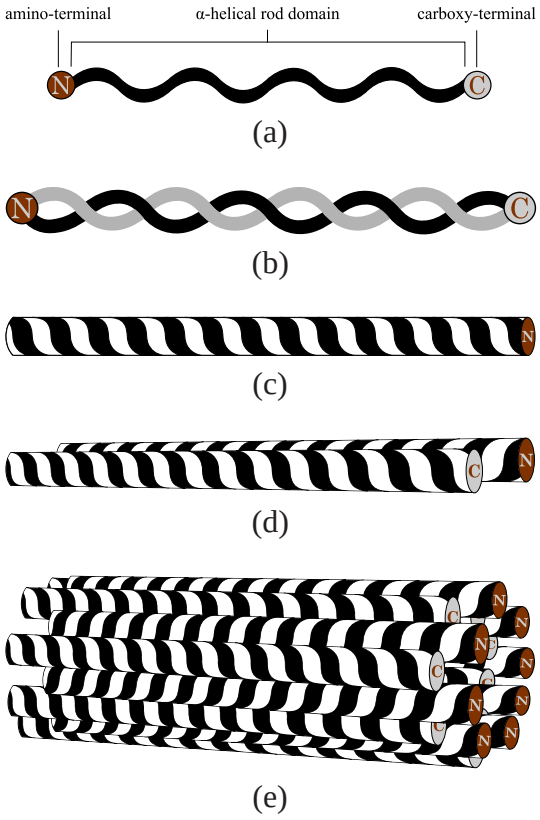


Figure 1: The first stage in the assembly of human vimentin proteins. Intermediate filament subunits are α-helical rods, that associate parallelly into coiled-coil dimers, which in turn form anti-parallel, half-staggered tetramers. Tetramers rapidly associate laterally to yield the shortest filaments called *unit-length filaments* (ULFs) of the same length as the tetramers. (a) α-helical rods, (b) coiled-coil dimer, (c) another representation of a coiled-coil dimer, (d) tetramer, (e) ULF.

of [15] and [14]), the model validation, and the sensitivity analysis. Our approach for the numerical analysis of the models differs markedly from that of [15], see Section 4 for a discussion.

Our study provides several conclusions regarding the kinetics of the *in vitro* assembly of human vimentin. On one hand, we show that the filament elongation process requires the end-to-end annealing of filaments as one of its features, which is in agreement with the results of [15]. Indeed, in all of our models where this reaction was missing, either the model did not fit the experimental data or the model was rejected in the validation round. Moreover, in almost all cases where the reaction modeling the end-to-end annealing of filaments is present, its rate constant is estimated to roughly the same value, although the other kinetic constants differ from model to model. On the other hand, the quantitative contribution of the filament elongation with tetramers depends on the turnover rate of tetramers into unit length filaments. If tetramers are quickly depleted from the system, e.g., through

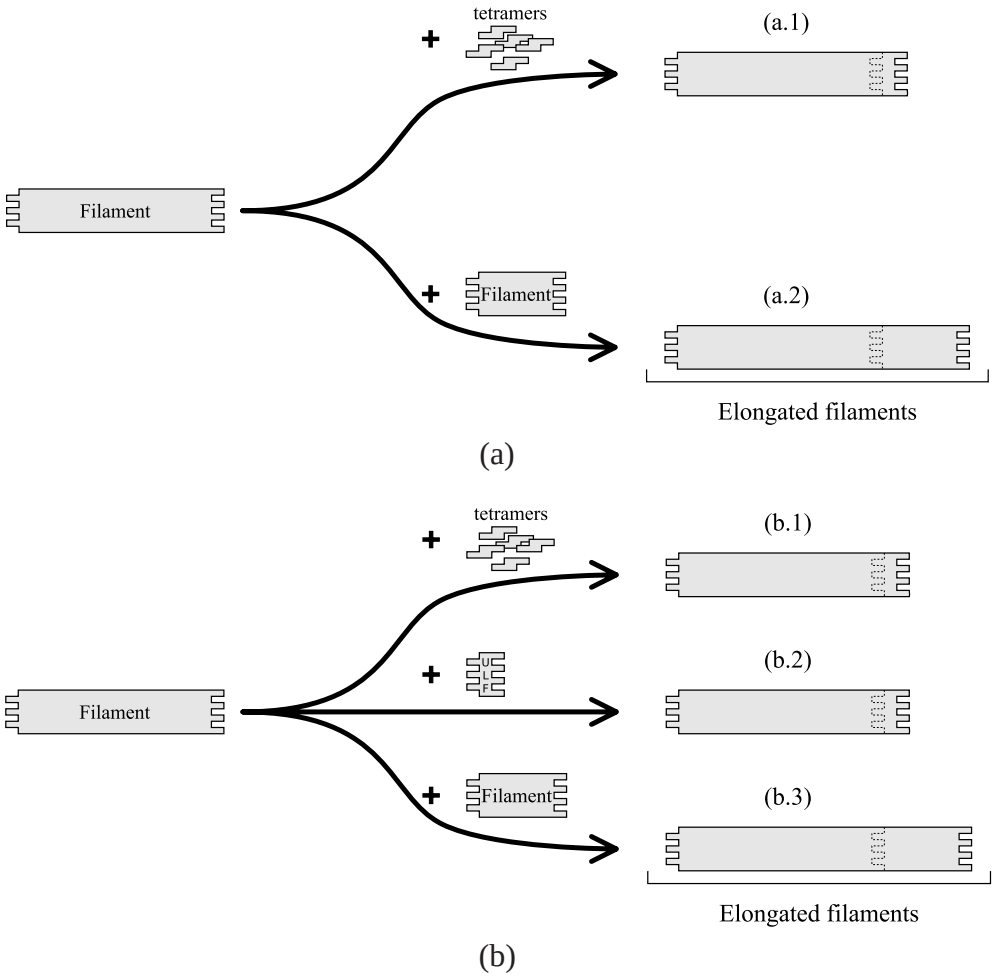


Figure 2: The two molecular models of the *in vitro* assembly of vimentin IF tetramers. (a) In the *simple model* filaments undergo elongation either by (a.1) longitudinal association of tetramers or (a.2) by end-to-end annealing of another filament. (b) The *extended model* adds a distinction between minimal-length filaments (ULFs) and longer filaments (consisting of at least 2 ULFs). In this case, there is one extra possibility for filament elongation: (b.1) by tetramer, (b.2) by the longitudinal association of a ULF, and (b.3) by another filament.

a high tetramer-to-ULF turnover rate as documented in *in vitro* experiments of [15], then only one of eight possible assembly strategies correlates well with the available experimental data, in agreement with conclusions of [15]. If free tetramers are however available throughout the assembly, then we show that several different assembly strategies correlate similarly well with the experimental data.

One of the modeling challenges identified in [15] was to increase the resolution of the model: instead of collecting all filaments into a single variable, regardless of their length, one should describe separately the dynamics of filaments of various lengths, at least up to a certain fixed, but arbitrarily high length, that we

call the resolution of the model. Indeed, the quantitative experimental data of [15] captures the levels of filaments of various lengths, but the data is only used in [15] to calculate the mean length of all filaments in the solution. We provide in this paper a generic solution to this problem, demonstrating how to enhance the existing filament assembly models with the dynamics of the filament length distribution. Our enhanced model can have arbitrarily high resolution, being able to capture the dynamics of filaments of arbitrarily high length. The size of this detailed model is considerably higher than that of the basic model, both in terms of molecular species, as well as in terms of molecular reactions. Based on kinetic observations on the basic model, we show however how the size of the high-resolution model can be drastically reduced. Our approach towards high-resolution models for protein self-assembly is independent of the particulars of vimentin filaments and can be applied to other instances of protein-protein interactions and protein assemblies.

2 Models and methodology

2.1 Two molecular models for the assembly of vimentin IFs

The *in vitro* assembly of vimentin IF proteins consists of three major phases, see [10]: (i) formation of the unit-length filaments (ULF) structures; (ii) longitudinal annealing of ULFs and growing filaments; (iii) radial compaction of immature filaments into mature IFs. We consider here two molecular models for this process, originally introduced in [15]. Both of them focus on the first two phases of the assembly, ignoring the third.

The *simple model* of [15] treats ULFs as ordinary filaments and describes the assembly process through a sequence of molecular events as follows, see also Figure 2 (a):

(i) two tetramers (denoted T) associate laterally into an octamer (denoted O):

$$2\,\mathrm{T} \to \mathrm{O}; \tag{1}$$

(ii) two octamers associate laterally to yield a hexadecamer (denoted H):

$$2\,\mathrm{O} \to \mathrm{H} \tag{2}$$

(iii) two hexadecamers associate laterally to form a (unit length) filament (denoted F):

$$2\,\mathrm{H} \to \mathrm{F} \tag{3}$$

(iv) a tetramer associates longitudinally to a filament to yield an elongated filament:

$$\mathrm{F} + \mathrm{T} \to \mathrm{F}; \tag{4}$$

(v) two filaments associate longitudinally to yield an elongated filament:

$$\mathrm{F} + \mathrm{F} \rightarrow \mathrm{F}\,. \tag{5}$$

The *extended model* of [15] adds a distinction between minimal-length filaments (ULFs, denoted U) and longer filaments (consisting of at least two ULFs), treating them as distinct species in the model, see Figure 2 (b). In terms of molecular events, the extended model consists of the following reactions:

(i') two tetramers (denoted T) associate laterally into an octamer (denoted O):

$$2\,\mathrm{T} \rightarrow \mathrm{O}; \tag{6}$$

(ii') two octamers associate laterally to yield a hexadecamer (denoted H):

$$2\,\mathrm{O} \rightarrow \mathrm{H} \tag{7}$$

(iii') two hexadecamers associate laterally to form a unit length filament (denoted U):

$$2\,\mathrm{H} \rightarrow \mathrm{U} \tag{8}$$

(iv') two unit length filaments associate longitudinally to form an elongated filament (denoted F):

$$2\,\mathrm{U} \rightarrow \mathrm{F} \tag{9}$$

(v') a filament is elongated longitudinally with a tetramer:

$$\mathrm{F} + \mathrm{T} \rightarrow \mathrm{F} \tag{10}$$

(vi') a filament is elongated longitudinally with a unit length filament:

$$\mathrm{F} + \mathrm{U} \rightarrow \mathrm{F} \tag{11}$$

(vii') two filaments associate longitudinally to yield an elongated filament:

$$\mathrm{F} + \mathrm{F} \rightarrow \mathrm{F} \tag{12}$$

2.2 Mathematical models

We consider a mathematical formulation of the simple and the extended models for IF assembly based on the mass-action law, where each molecular species is represented by a continuous non-negative real function denoting its concentration

in time. The system of differential equations corresponding to the simple model is the following:

$$d[\mathrm{T}]/dt = -2k_1^s[\mathrm{T}]^2 - k_t^s[\mathrm{T}][\mathrm{F}] \tag{13}$$
$$d[\mathrm{O}]/dt = k_1^s[\mathrm{T}]^2 - 2k_2^s[\mathrm{O}]^2 \tag{14}$$
$$d[\mathrm{H}]/dt = k_2^s[\mathrm{O}]^2 - 2k_3^s[\mathrm{H}]^2 \tag{15}$$
$$d[\mathrm{F}]/dt = k_3^s[\mathrm{H}]^2 - k_f^s[\mathrm{F}]^2 \tag{16}$$

where $k_1^s, k_2^s, k_3^s, k_t^s, k_f^s$ are the kinetic rate constants of reactions (1)-(5), respectively.

The mathematical model corresponding to the extended model consists of the following system of differential equations:

$$d[\mathrm{T}]/dt = -2k_1^e[\mathrm{T}]^2 - k_t^e[\mathrm{T}][\mathrm{F}] \tag{17}$$
$$d[\mathrm{O}]/dt = k_1^e[\mathrm{T}]^2 - 2k_2^e[\mathrm{O}]^2 \tag{18}$$
$$d[\mathrm{H}]/dt = k_2^e[\mathrm{O}]^2 - 2k_3^e[\mathrm{H}]^2 \tag{19}$$
$$d[\mathrm{U}]/dt = k_3^e[\mathrm{H}]^2 - 2k_4^e[\mathrm{U}]^2 - k_u^e[\mathrm{U}][\mathrm{F}] \tag{20}$$
$$d[\mathrm{F}]/dt = k_4^e[\mathrm{U}]^2 - k_f^e[\mathrm{F}]^2 \tag{21}$$

where $k_1^e, k_2^e, k_3^e, k_4^e, k_t^e, k_u^e, k_f^e$ are the kinetic rate constants of reactions (6)–(12), respectively.

An interesting aspect here is that the mass conservation relation on the total number of tetramers in the model is evident in the molecular models (since there is no synthesis and no degradation in the model), whereas it cannot be deduced as a property of either of the two corresponding mathematical models. This is a consequence of how, for example, the longitudinal association of two filaments is modeled: the information about the lengths of the two input filaments is not explicitly reproduced in a property of the two filaments. One can however calculate the number of tetramers integrated in the assembled filaments, as we do in Section 2.3, and then use this quantity to reason about the time-dependant dynamics of the mean filament length (MFL). We relate MFL to the experimental data of [15] and discuss the numerical fit of the models in Section 3.

2.3 Calculating the mean filament length

Relating the models proposed in the previous section for IF assembly to the quantitative data on the dynamics of the filament length is non-trivial because the two models do not represent explicitly the information about the length of the emerging filaments. Indeed, both models collect all filaments into a single variable (F), regardless of their length. We show however in this section that the dynamics of the mean filament length can in fact be deduced based on the variables of the two models.

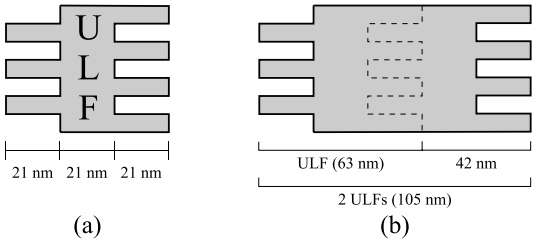


Figure 3: (a) The unit-length filament is approximately 63 nm long ([2]). (b) However, each ULF associated longitudinally at the end of an existing filament (or ULF) elongates it by approximately 42 nm ([2]). This is due to the interdigitation by which two ULFs anneal longitudinally.

During the process of ULFs aggregation atomic force microscopy (AFM) shows that each ULF associated longitudinally at the end of an existing filament adds to the length of that filament less than the stand-alone length of a ULF, see [2]. In the model for vimentin assembly of [2] this is due to interdigitation of the ULF and the filament to each other, see Figure 3. The stand-alone unit-length filament is approximately 63 nm long ([2]), while each additional ULF elongates a filament by approximately 42 nm ([2]).

We denote by $L_m(t)$ the time-dependent expression for the mean filament length (MFL) at time t. We also denote by $\#T_F(t)$ the total number of all tetramers integrated in the assembled filaments at time t. Since we consider two categories of filaments, U and F, we obtain that

$$L_m(t) = \frac{l_F(t) + l_U(t)}{\#F(t) + \#U(t)}, \tag{22}$$

where $l_F(t)$ and $l_U(t)$ denote the total length of all filaments and the total-length of all ULFs at time t, while $\#F(t)$ and $\#U(t)$ denote the total number of all filaments and that of all ULFs, respectively. Since in each filament the first ULF accounts for $l_{ULF} \simeq 63$ nm of the total length of that filament and all the additional ULFs elongate the filament by $l_{addULF} \simeq 42$ nm, we have that

$$\begin{aligned} l_F(t) &= (\#U_F(t) - \#F(t)) \cdot l_{addULF} + \#F(t) \cdot l_{ULF} \\ &= \#U_F(t) \cdot l_{addULF} + \#F(t) \cdot (l_{ULF} - l_{addULF}), \end{aligned}$$

where $\#U_F(t)$ denotes the the total number of all ULFs in all filaments, in time. Since ULFs consist on average of eight tetramers, we have that

$$\#U_F(t) = \frac{\#T_F(t)}{8},$$

where $\#T_F(t)$ is the number of tetramers already assembled into filaments.

We denote by c_0 the initial molar concentration of all tetramers in the system (occurring in any of the molecular species of the model: tetramers, octamers,

hexadecamers, ULFs, or filaments). Then, in the case of the extended model we obtain

$$\begin{aligned}\#T_F(t) = (c_0 - [T](t) - 2\,[O](t) - 4\,[H](t)\\ - 8\,[U](t)) \cdot N_A \cdot V,\end{aligned}$$

where N_A is the Avogadro constant and V is the volume of the system. Thus, (22) becomes

$$\begin{aligned}L_m(t) =& \frac{\frac{c_0 - [T](t) - 2\,[O](t) - 4\,[H](t) - 8\,[U](t)}{8} \cdot l_{addULF}}{([F](t) + [U](t))}\\ &+ \frac{[F](t) \cdot (l_{ULF} - l_{addULF}) + l_{ULF} \cdot [U](t)}{([F](t) + [U](t))}.\end{aligned}$$

In the case of the simple model, we obtain that

$$\#T_F(t) = (c_0 - [T](t) - 2\,[O](t) - 4\,[H](t)) \cdot N_A \cdot V.$$

Thus, (22) becomes

$$\begin{aligned}L_m(t) =& \frac{\frac{c_0 - [T](t) - 2\,[O](t) - 4\,[H](t)}{8} \cdot l_{addULF}}{[F](t)}\\ &+ (l_{ULF} - l_{addULF}).\end{aligned}$$

Since the volume V of the considered system does not change, the molar concentrations are expressed simply in terms of micromoles (without reciprocal of the volume unit) in the continuation.

2.3.1 Experimental data and model fitting

For the parameter estimations and model validations we used the experimental data from [14] on the *in vitro* assembly process of recombinant vimentin at $37\,^{\circ}\mathrm{C}$. The data consists of two sets, each containing the length distributions of growing filaments at distinct time points up to 20 min. The data sets were obtained by adsorption of the filaments onto carbon-coated copper grids and measurements of the filament lengths from images recorded with electron microscopy (EM) in two cases: when the initial amount of tetramers was $0.45\ \mu$M and $0.9\ \mu$M. For each set the time-dependent mean filament length (MFL) was calculated. The MFL values together with the 0.95 confidence intervals are presented in Table 1. For detailed description of experimental procedures and discussion on the independence of the measured MFLs from the support medium we refer to [15].

For fitting our mathematical models, we used the MFL data obtained for an initial tetramer concentration of $0.45\ \mu$M. For model validation, we then compared the numerical prediction for the mean filament length with the experimental data in Table 1 for an initial tetramer concentration of $0.9\ \mu$M.

Table 1: Measurements on the mean filament length of vimentin protein IFs, based on EM data of [14] (data in [nm]); a preliminary version of the data (containing a few minor errors) is in [15].

Time [s]	Initial molar concentration of all tetramers (c_0)	
	0.45 μM	0.9 μM
10	65.1±1.4	62.8±2.1
20	68.2±2.0	
30	76.5±2.1	84.1±2.0
60	112.9±4.0	131.4±5.2
180	172.6±8.4	
300	233.0±10.0	289.1±15.8
600	320.7±18.5	418.6±24.7
900		544.1±34.8
1200	474.9±37.2	821.3±41.5

We set the initial molar concentrations of all molecular species other than tetramers to 0, based on the setup of the experimental assays. Thus, there remained to be estimated five independent parameters (rate constants k_1^s, k_2^s, k_3^s, k_t^s and k_f^s) for the simple model and seven of them (rate constants k_1^e, k_2^e, k_3^e, k_4^e, k_t^s, k_u^e and k_f^e) for the extended model. Parameter estimations were performed in COPASI [12].

We also considered a qualitative property of the IF assembly, reported in [15]: very quickly (within approximately 10 seconds) after the initiation of the assembly, ULF is the most predominant species in the system, while tetramers are depleted. This observation only applied for the *ab initio in vitro* assembly of intermediate filaments. The dynamics could however be very different if more free tetramers were available for longer throughout the assembly (e.g., through an additional tetramer synthesis mechanism). To test it, we considered two different strategies for fitting our models: one where the tetramer-to-ULF turnover is fast, and another where it is slow. While the latter setup does not mimic the presence of a tetramer synthesis mechanism (introducing one would make it difficult to compare the models), it does allow us to analyze the system in the case where tetramers are available for a longer period for the assembly. We demonstrate in the next section that the two situations are indeed very different, in terms of which filament elongation mechanisms (with tetramers, with ULFs, or with other filaments) can explain the available experimental data.

The problem of estimating the parameters of computational models in systems biology is difficult, see e.g., [3, 20, 21]. This problem can be formulated as a minimization of a cost function which quantifies the differences between the values predicted by the model and the experimental measurements. There are numerous methods, both local and global, which can be used to tackle this problem, each with its own advantages and disadvantages. For instance, while local methods work faster to find a solution, they tend to converge to local optima. On the other

hand, global optimization methods are typically slower, but they tend to converge to a global optimum. The global optimization methods can be further divided into deterministic [6, 13] and stochastic approaches [1, 7]. Although the deterministic methods guaranty the convergence to a global optimum, they cannot ensure the termination of this process within a finite time interval [21]. On the other hand, the inherent randomness of the stochastic approaches makes it very hard to guaranty that these methods actually converge to the global optimum [21]. However, many stochastic methods are capable of locating the vicinity of global solutions with relative efficiency, i.e. they provide a very good approximation of the solution in acceptable computation time [21]. This makes the stochastic global optimization methods to be usually preferred for parameter estimation problems. We chose COPASI, [12], as a computational environment for parameter fitting since it includes a number of various optimization algorithms, searching for either local or global optimum values, see e.g., [19, 23]. This software is a widely used tool in the computational systems biology modeling community, having a documented good performance, see e.g. [3, 20, 21]. In particular, for determining the best numerical fits of our models, a suite of various global, stochastic parameter estimation procedures was used, comprising of methods such as Simulated Annealing, Genetic Algorithm, Evolution Strategy using Stochastic Ranking, and Particle Swarm. All these methods use specific strategies for sampling the parameter space looking for combinations of parameter numerical values that give better and better fits of the model predictions to the experimental data.

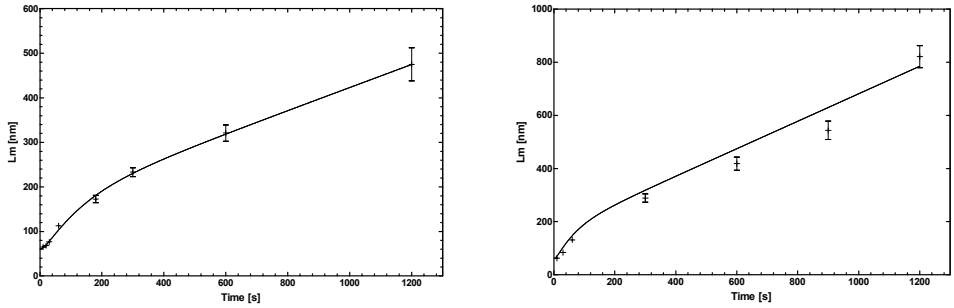
The fit of a model was performed by searching for a set of parameter values that minimizes the sum of squared deviations SS_f of the values predicted by the model from the 0.45 μM experimental data. The validation of a fitted model was performed by numerically simulating the model and by computing the sum of squared deviations SS_v of the values predicted by the model from the 0.9 μM experimental data. Moreover, the quality of the fit/validation for each model was estimated by a dimensionless number expressing the deviation of the model from the experimental data, normalized by the mean of the predicted values. This method for estimating the quality of model fit/validation was originally proposed in [16] and it allows for comparison of different models and different data sets. The formula for the quality of the fit (fq) is:

$$fq = \frac{\sqrt{SS_f/N_f}}{\text{mean of predicted values}} \cdot 100\%, \tag{23}$$

where N_f is the number of 0.45 μM experimental data points (in our case $N_f = 8$). Similarly, the formula for the quality of the validation (vq) is:

$$vq = \frac{\sqrt{SS_v/N_v}}{\text{mean of predicted values}} \cdot 100\%, \tag{24}$$

where N_v is the number of 0.9 μM experimental data points (in our case $N_v = 7$). It was argued in [16] that a low (say, lower than 15%) value of fq (vq) was con-



(a) (b)

Figure 4: Time-dependent MFL growth corresponding to the simple model without the quick filament formation requirement. (a) The model fit with respect to the EM 0.45 μM experimental data set. (b) Model validation based on the EM 0.9 μM experimental data set. The continuous line is the model prediction regarding $L_m(t)$, that is compared with the experimental data showed with crossed points. The short vertical lines represent the 0.95 confidence intervals for the experimental data.

Table 2: Kinetic rate constant values in $\mu\text{M}^{-1}s^{-1}$ for the simple model.

k_1^s	k_2^s	k_3^s	k_t^s	k_f^s
$3.39 \cdot 10^{-3}$	30	30	0.83	0.11

sidered as an indicator of a successful fit (validation). We discuss the numerical values of fq and vq for all our models in Section 3.

3 Results

3.1 Data fitting the simple model

The kinetic rate constants in Table 2 yield an excellent fit ($fq = 2.52\%$) of the simple model for the experimental data from the assay with 0.45 μM tetramers and a good validation ($vq = 12.07\%$) of the model when compared with the data from the assay with 0.9 μM initial concentration of tetramers, see Figure 4.

This model however could not confirm the quick turnover of tetramers into filaments. When this condition was taken into consideration by searching for relatively high numerical values of k_1^s, k_2^s, and k_3^s (higher than 1 $\mu M^{-1}s^{-1}$), the fit of the model to the experimental data was unsuccessful ($fq = 26.00\%$), despite numerous rounds of parameter estimation. The following mathematical argument is also indicating that this model cannot be given a reasonable fit. Based on the observation that tetramers are quickly depleted (within 10 seconds) by turning them into ULFs, the model can be split into two processes separated in time: first, the formation of filaments from tetramers, i.e. $2\,\text{T} \rightarrow \text{O}$, $2\,\text{O} \rightarrow \text{H}$, $2\,\text{H} \rightarrow \text{F}$,

and second, the elongation of filaments, i.e. $\mathrm{F} + \mathrm{F} \rightarrow \mathrm{F}$. The steady state value of F in the first process is an initial value of F in the second one. The second process is described by the differential equation $\mathrm{d}\,\mathrm{F}\,/\mathrm{d}t = -k\,\mathrm{F}^2$, which has an analytical solution of the form $\mathrm{F}(t) = \mathrm{F}_0\,/(1 + k\,t\,\mathrm{F}_0)$, where F_0 is the initial value of F. The initial concentration of tetramers in the first process is c_0, hence it follows that $\mathrm{F}_0 = c_0/8$ since all tetramers are turned into ULFs. In consequence, the mean filament length can be expressed as

$$L_m(t) = l_{ULF} + \frac{k\,c_0\,t}{8}.$$

Thus, $L_m(t)$ is a linear function. By plotting the experimental data in Table 1 for time points after 30 seconds, together with their 0.95 confidence intervals one can see that there exists no k such that the model would be fitted and validated against the data.

3.2 Data fitting the extended model

In the case of the extended model we distinguished among three modes for filament elongation: (i) with a tetramer, (ii) with a ULF, or (iii) with another filament, see Figure 2 (b). We investigated all eight possible combinations of these three mechanisms and performed parameter estimation and numerical model validation for each of them, see Figure 5. Excluding any of the three modes from the investigation was done by simply setting to 0 the corresponding rate constants, i.e. k_t^e, k_u^e, and k_f^e, respectively.

3.2.1 The extended model with fast ULF formation.

In the case of fast tetramers-to-ULF turnover, both the simple model and the extended model can be reduced. Indeed, in this case, the populations of tetramers, octamers, and hexadecamers are all quickly depleted (in a matter of seconds), leaving only the filaments as the dominant species. Consequently, the longitudinal assembly of tetramers to filaments has a negligible contribution to the overall dynamics of the model: in the first few seconds the reaction is strangled by the negligible population of filaments, whereas later on the population of tetramers is depleted. This is in agreement with [15], where it was observed that this particular elongation has insignificant role. In this case we set $k_t^e = 0$ and we searched for numerical values for the kinetic rate constants k_1^e, k_2^e, and k_3^e that are greater than $3\ \mu\mathrm{M}^{-1}s^{-1}$, to ensure a fast tetramer-to-ULF turnover. It turned out that scenario VIII, where $k_u^e = k_f^e = 0$, could be immediately excluded. Indeed, in this scenario no filament containing more than two ULFs could be assembled and so, all filaments would be at most 100 nm long, contradicting the experimental data in Table 1.

Scenarios VI and VII, where the filament elongation takes place only by ULF extension ($k_f^e = 0$), or only by filament extension ($k_u^e = 0$), respectively, could

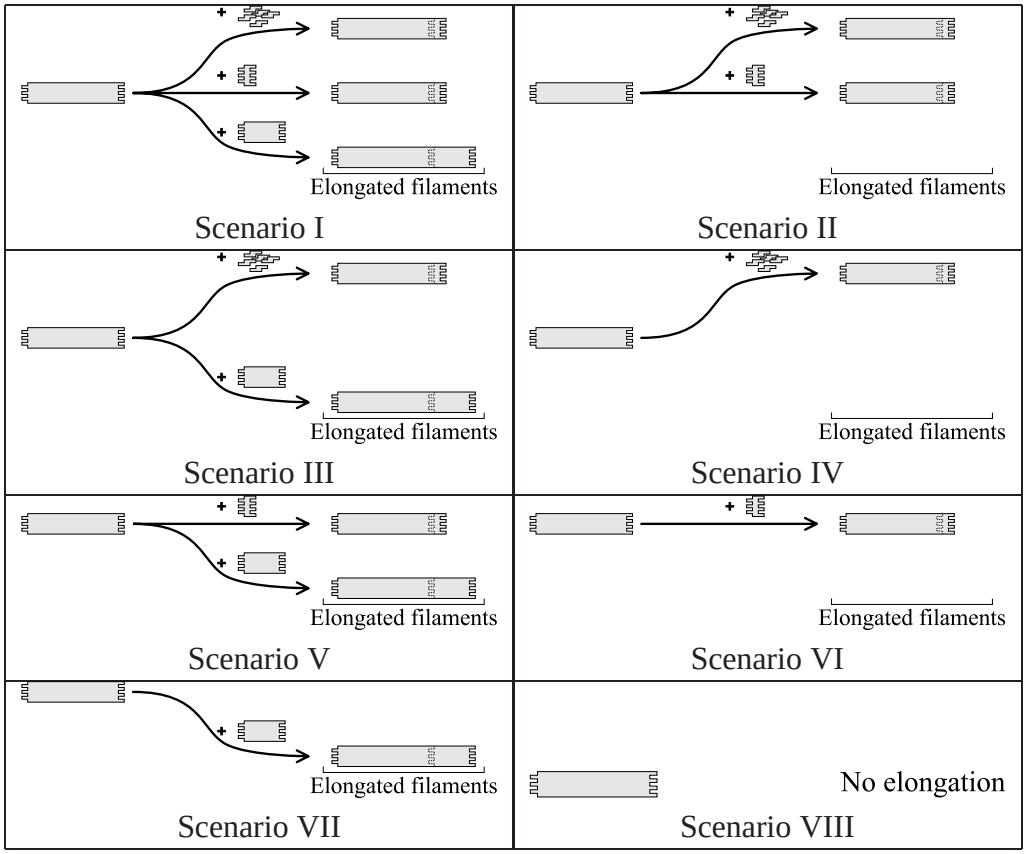


Figure 5: The eight possible scenarios for filament elongation. The tetramers/ULFs/filaments are illustrated with the same type of block as in Figure 2.

Table 3: Kinetic rate constant values in $\mu\mathrm{M}^{-1}s^{-1}$ (under fast ULF formation requirement).

k_1^e	k_2^e	k_3^e	k_4^e	k_u^e	k_f^e
3	30	30	0.25	0.95	0.11

not be fitted: for Scenario VI we obtained $fq = 22.77\%$ and for Scenario VII $fq = 14.99\%$, $vq = 16.07\%$. We concluded that these two strategies do not represent viable pathways for vimentin IFs assembly.

In the case of scenario V we were able to obtain numerical values for the parameters, see Table 3, such that the predicted mean filament length was in good agrement with the experimental data ($fq = 3.66\%$, $vq = 11.45\%$), virtually identical to that of the simple model, showed in Figure 4. We concluded that this pathway, where the filament elongation is enabled both with ULFs and with other filaments, is the only viable strategy for vimentin IFs assembly. This is in agreement with observations of [15].

Numerically fitting this scenario, we noticed that the values of the two numerical parameter k_2^e and k_3^e can be modified arbitrarily within the $[3, 30]$ interval

Table 4: Fit and validation quality measure values for scenarios I–VII (without the fast ULF formation requirement).

	I	II	III	IV	V	VI	VII
fq	1.71%	6.50%	1.98%	6.79%	2.04%	6.54%	13.01%
vq	12.70%	29.03%	12.36%	25.83%	12.65%	29.11%	19.19%

Table 5: Kinetic rate constant values in $\mu\text{M}^{-1}s^{-1}$ of scenarios I–VII (without the fast ULF formation requirement).

	I	II	III	IV	V	VI	VII
k_1^e	0.0705	30	$4.83 \cdot 10^{-3}$	$4.58 \cdot 10^{-3}$	1.24	30	30
k_2^e	30	30	30	10^{-09}	17.78	30	30
k_3^e	11.34	$4.63 \cdot 10^{-3}$	21.25	$6.06 \cdot 10^{-5}$	$2.65 \cdot 10^{-2}$	$4.67 \cdot 10^{-3}$	30
k_4^e	0.32	10.69	30	30	11.16	10.69	2.56
k_t^e	15.48	30	0.61	0.84	0	0	0
k_u^e	0.59	30	0	0	11.57	30	0
k_f^e	0.10	0	0.10	0	0.10	0	0.15

without any significant change in the mean filament length prediction. This indicates that the extended model under the fast ULF formation exhibits almost no sensitivity of mean filament length with respect to these two parameters in the mentioned interval and, in consequence, our computational model turns to have less degrees of freedom in terms of the numerical fit.

3.2.2 The extended model with slow ULF formation.

In this case, we searched for arbitrary positive numerical values for the kinetic rate constants k_1^e, k_2^e, and k_3^e. The result of fitting and validating the extended model are very different in this case. We find that three out of the eight pathways analyzed in this paper for vimentin IFs assembly can explain the experimental data, see Figures 6 and 7.

Scenario VIII could not be fitted based on similar considerations as in the case of the fast ULF formation, see Figure 7 VIII(a) and VIII(b). In the case of the other seven pathways, the model fit with respect to the EM 0.45 μM data and the model validation with respect to the EM 0.9 μM data yielded good results, summarized in Table 5, see Figures 6 and 7V-VII. We noted that in the case of scenarios II, IV, and VI the experimental MFL measurement at 1200 seconds for the EM 0.9 μM data was an outlier. In all these three scenarios, we have $k_f^e = 0$, which indicates that the process of end-to-end filament annealing plays a crucial role in the later stages of the IFs elongation process, i.e., after the first 600 seconds. In the case of scenario VII, the model left several experimental data points as outliers, see Figure 7VII(a) and (b).

We concluded that scenarios I, III, and V are similarly good in explaining the

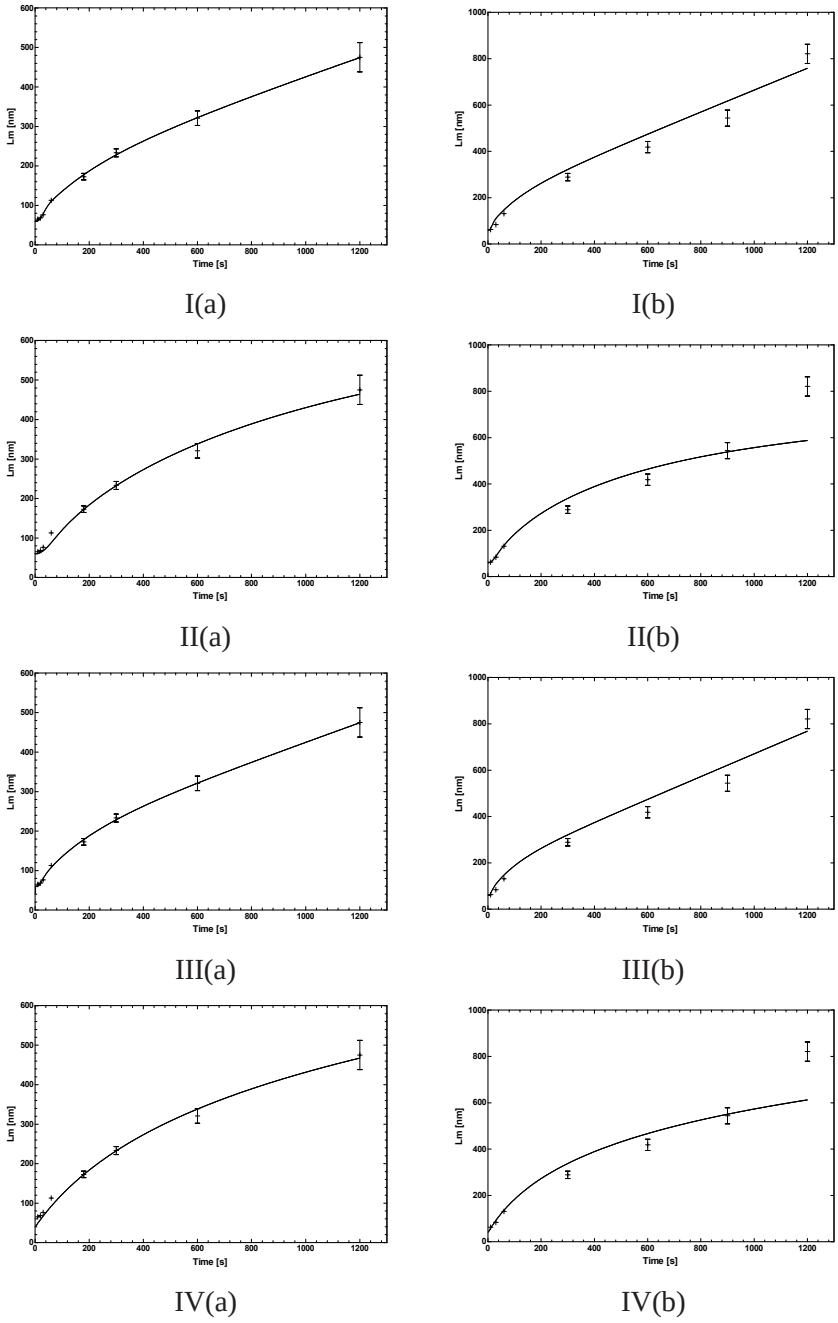


Figure 6: I(a)–IV(a) The model fit of the scenarios I to IV with respect to the EM 0.45 μM experimental data set. I(b)–IV(b) Model validation of the scenarios I to IV with respect to the EM 0.9 μM experimental data set. The continuous line is the model prediction regarding $L_m(t)$, that is compared with the experimental data showed with crossed points. The short vertical lines represent the 0.95 confidence intervals for the experimental data.

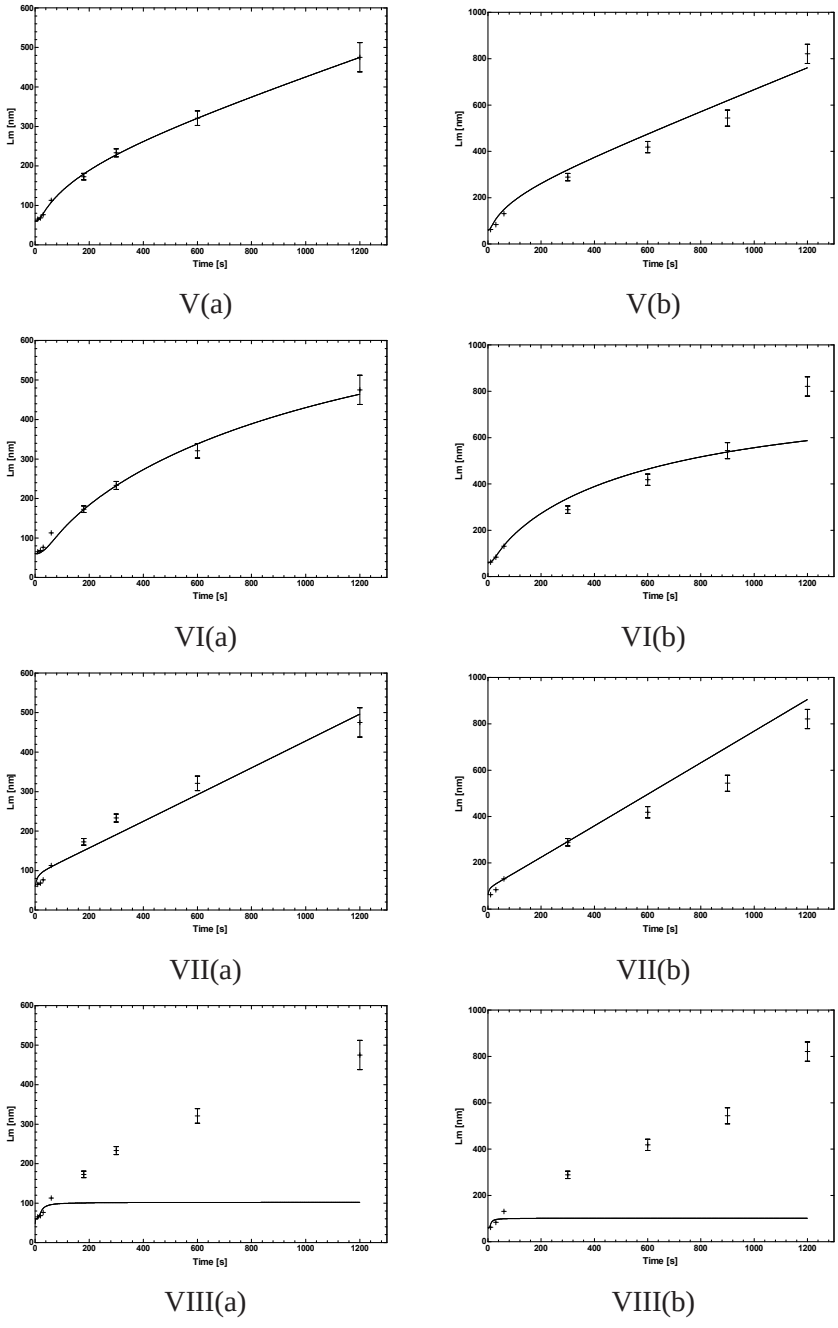


Figure 7: V(a)–VIII(a) The model fit of the scenarios V to VIII with respect to the EM 0.45 μM experimental data set. V(b)–VIII(b) Model validation of the scenarios V to VIII with respect to the EM 0.9 μM experimental data set. The continuous line is the model prediction regarding $L_m(t)$, that is compared with the experimental data showed with crossed points. The short vertical lines represent the 0.95 confidence intervals for the experimental data.

experimental data in this case. These models correspond to the following three pathways for filament elongation:

- Scenario I: by a tetramer, a ULF or another filament longitudinal elongation;
- Scenario III: by a tetramer or a filament longitudinal elongation;
- Scenario V: by a ULF or a filament longitudinal elongation.

3.3 Sensitivity analysis of the mean filament length

The effect of small variations in the model's parameters over the evolution of the entire model is estimated by the sensitivity analysis. This mathematical method consists in determining the time evolution of the partial derivatives of the solution of the system with respect to the parameters of the system. We investigated the sensitivity of the mean filament length, i.e., the $L_m(t)$ function, with respect to the parameters of the model. We compared the results of the sensitivity analysis in the case of Scenarios I-VII of the extended model in order to gain further insight into the possible pathways for IF vimentin assembly.

The concentration sensitivity coefficients are the time functions $\partial X_i/\partial \kappa_j$ for all $1 \leq i \leq 5$ and $1 \leq j \leq 7$, where $X = (X_1, \ldots, X_5)$ is the vector of the model variables ([T], [O], [H], [U], and [F], respectively) and $\kappa = (\kappa_1, \ldots, \kappa_7)$ is the vector of the model parameters ($k_1^e, k_2^e, k_3^e, k_4^e, k_t^e, k_u^e$, and k_f^e, respectively). The sensitivity of the mean filament length with respect to the parameters is obtained as follows:

$$\frac{\partial L_m(t)}{\partial \kappa_j} = \frac{\partial L_m}{\partial X}\,\frac{\partial X}{\partial \kappa_j} = \frac{\partial L_m}{\partial X_1}\,\frac{\partial X_1}{\partial \kappa_j} + \cdots + \frac{\partial L_m}{\partial X_5}\,\frac{\partial X_5}{\partial \kappa_j},$$

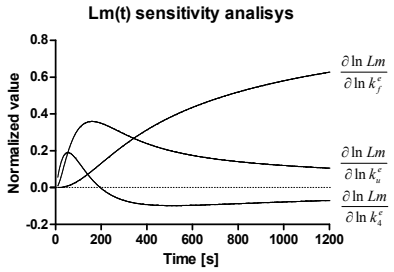
for all $1 \leq j \leq 7$.

Since we want to compare the MFL sensitivities of several models, we transform these coefficients into dimensionless measurements by normalizing them:

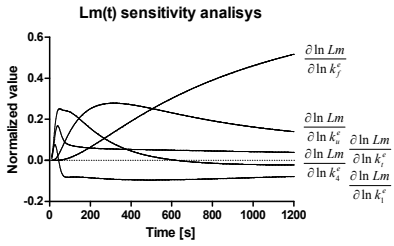
$$\frac{\kappa_j}{L_m(t)}\,\frac{\partial L_m(t)}{\partial \kappa_j} = \frac{\partial \ln L_m(t)}{\partial \ln \kappa_j}, \quad \text{for all } 1 \leq j \leq 7.$$

We can interpret these coefficients as follows: in Scenario I, an increase of 1% of the parameter k_f^e would generate at time $t = 1200\,s$ an increase of 0.5165% of the MFL, roughly as predicted by the value of $\partial \ln(L_m)/\partial \ln(k_f)$ at time $t = 1200$, see Figure 8 b).

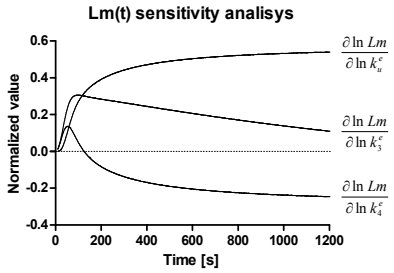
In the case of the extended model with fast ULF formation, only scenario V could be experimentally validated. The results of the sensitivity analysis in this case are presented in Figure 8 a). The most significant coefficients are with respect to the k_4^e, k_u^e, and k_f^e parameters, with the latter one being the most significant. This is consistent with the biological intuition that the mean filament length



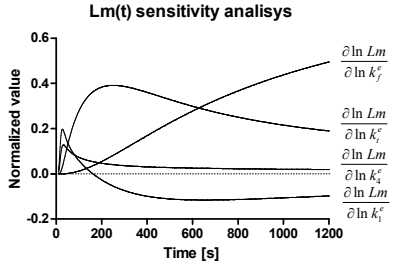
a) Scenario with fast ULF formation



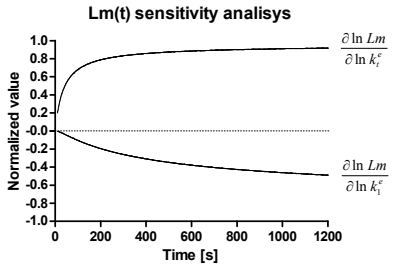
b) Scenario I



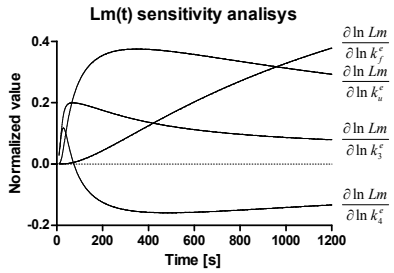
c) Scenario II



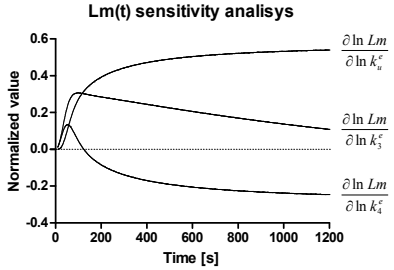
d) Scenario III



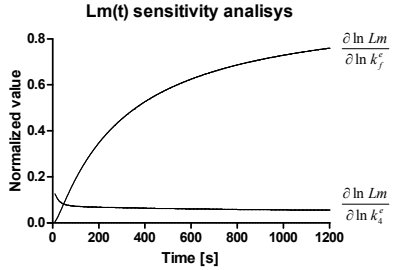
e) Scenario IV



f) Scenario V



g) Scenario VI



h) Scenario VII

Figure 8: The non-negligible sensitivity coefficients of the MFL measurement for the mathematical models corresponding to the scenario with fast ULF formation requirement and the scenarios I to VII.

is most dependent on the rate of filament formation (parameter k_4^e) and elongation (parameters k_u^e and k_f^e). Less intuitive is the fact that there is a negligible dependency of the MFL measurement on the rate constants $k_1^e, k_2^e,$ and k_3^e, which determine the fast ULF formation. The rationale for this result is that these kinetic constants play a role only in the first seconds of the assembly. Once the vast majority of tetramers are assembled into ULFs, their further contribution to the model dynamics is insignificant.

The numerical time simulation of the non-negligible normalized MFL sensitivity coefficients for scenarios I-VII without fast ULF formation requirement are presented in Figure 8 b)–h). It turned out that the mean filament length is most sensitive to k_u^e and especially to k_f^e, when these constants are non-zero. This observation helps explain why k_f^e is estimated to very similar values in most scenarios where its role is considered. Note also that while the sensitivity coefficient with respect to k_f^e increases mainly after about 200 seconds, the sensitivity coefficients for the parameters k_t^e and k_u^e have a steep increase in the first 100–200 seconds (except in scenario VII where filament elongation takes place only by longitudinal filament aggregation). The biological intuition here is that on one hand, until approximately 200 seconds the assembled filaments are relatively short and much fewer than the ULF's, while on the other hand the number of ULFs and of free tetramers becomes very low after about 200 seconds.

3.4 The length distribution of filaments in time

The models discussed so far in this paper, as well as those in [15] collect all filaments other than ULFs into one single variable denoted F, regardless of their length. This approach is indeed enough for capturing the time-dependent dynamics of the mean filament length, that could then be related to experimental data and used for parameter estimation and model validation. As pointed out also in [15], this modeling approach is however unsuitable for capturing the time-dependent distribution of the filament lengths. Indeed, the length of the assembling filaments is not directly captured in the models, which makes it impossible to reason about the time-dependant concentration of filaments of some given length. We describe in this section a refined model for the self-assembly of vimentin filaments that allows capturing the evolution of filaments of length up to n, for any given positive integer n.

For all i with $1 \leq i \leq n$, we denote by F_i the population of all filaments of length exactly i, where the length is in terms of the number of ULFs that the filament consists of. Thus, the ULFs are denoted by F_1 in the new model, the filaments formed by the longitudinal extension of a ULF with another ULF have length 2 and are denoted by F_2, etc. The population of all filaments of length higher than n is denoted by $F_{\geq n+1}$. The longitudinal extension of a filament F_i (of length $i \leq n$) with a filament F_j (of length $j \leq n$) yields a filament of length F_{i+j} if $i+j \leq n$ and a filament $F_{\geq n+1}$ if $i+j \geq n+1$. The extension of a filament

$F_{\geq n+1}$ with any other filament yields a filament $F_{\geq n+1}$.

When describing the extended model for filament self-assembly based on the populations F_i, $1 \leq i \leq n$, and $F_{\geq n+1}$, a considerable challenge is posed by the elongation of a filament with tetramers. Indeed, such a longitudinal elongation leads to a filament that ends with an incomplete ULF. Only after the lateral association of seven other tetramers would this be a complete filament of length one higher. This difficulty can be addressed by introducing a notation of the type $F_i^{j,k}$ with $1 \leq i \leq n$ and $0 \leq j, k \leq 7$ to denote filaments consisting of i complete ULFs, an incomplete ULF with j tetramers at their left end, and an incomplete ULF with k tetramers at their right end, see Figure 9. One would also denote by $F_{\geq n+1}^{j,k}$ the filaments consisting of more than n complete ULFs, an incomplete ULF with j tetramers at their left end, and an incomplete ULF with k tetramers at their right end. This approach leads however to a steep increase in the number of model variables. For example, for $n = 10$, the model would have 396 variables just to denote the different types of filaments.

To keep the size of the model manageable we can however take advantage of the kinetic observations we made on the extended model for filament assembly in Section 3.2: in the case of fast ULF formation we have demonstrated that the longitudinal elongation of filaments with tetramers has negligible kinetic influence on the dynamics of the model and that eliminating it leads to a numerically equivalent model. Consequently, we can ignore all possible filaments having incomplete ULFs at either end, since essentially all tetramers in the system assemble into ULFs within a very short period of time. In this case our model consists of the following reactions:

$$
\begin{array}{ll}
(T) & T + T \to O; \\
(O) & O + O \to H; \\
(H) & H + H \to F_1; \\
(A_{i,j}) & F_i + F_j \to F_{i+j}, \\
 & \quad \text{for all } 1 \leq i \leq j \leq n \text{ such that} \\
 & \quad i + j \leq n; \\
(B_{i,j}) & F_i + F_j \to F_{\geq n+1}, \\
 & \quad \text{for all } 1 \leq i \leq j \leq n \text{ such that} \\
 & \quad i + j \geq n + 1; \\
(C_i) & F_{\geq n+1} + F_i \to F_{\geq n+1}, \\
 & \quad \text{for all } 1 \leq i \leq n; \\
(D) & F_{\geq n+1} + F_{\geq n+1} \to F_{\geq n+1}.
\end{array}
$$

We call this a *model of resolution* n, see Figure 10 for an illustration. For example, in the case of $n = 10$, the model consists of 14 variables and 69 reactions.

The initial values of all variables except for T are set to 0, while that of T is assumed the same as in the extended model in Section 3.2. The kinetic rate constants of the new model are set in such a way that the overall number of filaments is the same as in the extended model. The kinetics of reactions (T), (O), and (H)

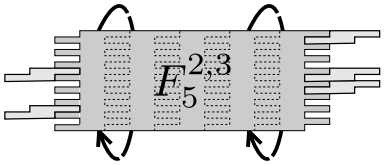


Figure 9: A filament consisting of 5 complete ULFs, an incomplete ULF with 2 tetramers at the left end, and an incomplete ULF with 3 tetramers at the right end. We denote it in our model with $F_5^{2,3}$.

are the same as in the corresponding reactions of the extended model. If $a_{i,j}$ is the kinetic rate constant of reaction ($A_{i,j}$), $b_{i,j}$ that of reaction ($B_{i,j}$), c_i that of reaction (C_i), and d that of reaction (D), then we set their values as follows:

- $a_{1,1} = k_4^e$, $a_{1,j} = k_u^e$, for all $1 < j \leq n$;
- $b_{1,j} = c_1 = k_u^e$, for all $1 \leq j \leq n$;
- $a_{i,i} = b_{i,i} = k_f^e$, for all $1 < i \leq n$, and $a_{i,j} = b_{i,j} = c_i = 2k_f^e$, for all $1 < i < j \leq n$;
- $d = k_f^e$.

Based on the corresponding ODE models, a straightforward calculation shows that with these kinetic constants, the extended model of Section 3.2 and the model of resolution n are equivalent in the following sense:

- $[F_1](t) = [U](t)$ and
- $([F_2] + \ldots + [F_n] + [F_{\geq n}])(t) = [F](t)$,

for all time points $t \geq 0$.

As an example, we have implemented in COPASI the model in the case of $n = 10$. In Figure 11 we plotted this model's prediction for the distribution in time of all filaments of length at least two. The resulting dynamics is in line with the biological expectation. For example, the number of filaments of length two, F_2, witnesses a sharp increase right after the start of the experiment, as tetramers are turned into (short) filaments. F_2 then decreases quickly as filaments start combining to each other to yield longer filaments.

4 Discussion

Related work. A recent review of the biochemistry of the intermediate filaments, including kinetic aspects of their self-assembly is in [8]. The simple and extended models for the self-assembly of vimentin proteins were originally investigated in [15]. The approach used in the fitting and the validation of the models

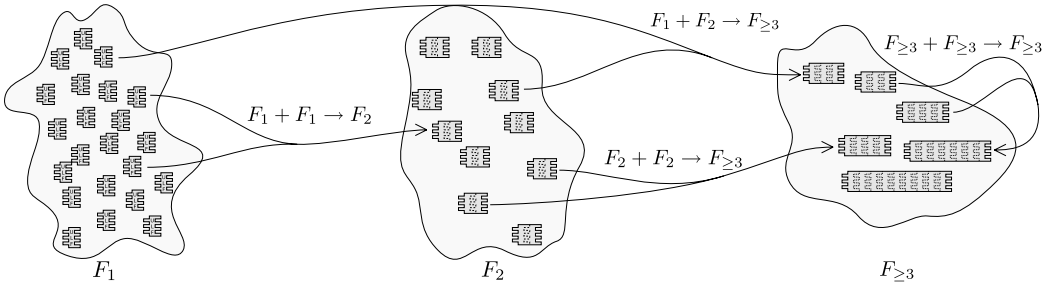


Figure 10: The scheme of a model of resolution 3 for the self-assembly of IF. We partition the population of filaments into filaments of length one (F_1), of length two (F_2), and of length at least three ($F_{\geq 3}$). The longitudinal annealing of two filaments of length one yields a filament of length two ($F_1 + F_1 \rightarrow F_2$), that of a filament of length one and another of length two yields a filament of length at least three ($F_1 + F_2 \rightarrow F_{\geq 3}$), the annealing of two filaments of length two yields a filament of length at least three ($F_2 + F_2 \rightarrow F_{\geq 3}$), and that of two filaments of length at least three results in a filament belonging to the same $F_{\geq 3}$ group ($F_{\geq 3} + F_{\geq 3} \rightarrow F_{\geq 3}$).

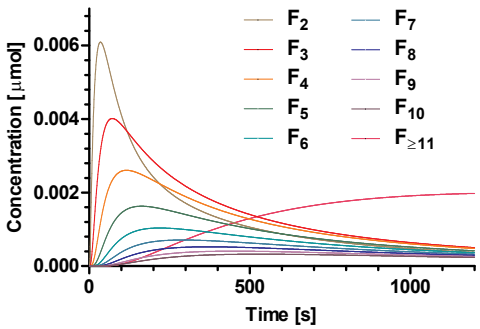


Figure 11: Model prediction for the distribution in time of all the filaments containing from two to ten ULFs.

was somewhat ad-hoc in [15], as discussed below. We made in our paper a systematic investigation of the kinetics of the two models for intermediate filament self-assembly, based on well-established techniques of model fit and model validation. Some of our results confirm those of [15], while others bring a new insight into the nature of filament assembly. We discuss in the following the main points of divergence between our approach and that of [15].

A main difference concerns the mathematical modeling of the simple and the extended models. The models in [15] assume that the lateral association of two tetramers, of two octamers, and of two hexadecamers have the same kinetic rate constants. This strong model assumption is however unsubstantiated by experimental evidence and leads to limiting the range of possible model behaviors. We assign different kinetic constants for each different reaction to allow maximum flexibility in the predictive power of the models.

Our mathematical expression for the mean filament length differs from the one presented in [15]. In there, the authors use a so-called *linear density variable* d_l, set at 43.5 nm, representing the length of a ULF inside a filament, regardless of whether the ULF is the first of the filament, or a subsequent one. This distinction is however crucial for estimating the mean filament length. Indeed, ignoring this distinction introduces an approximation error which is proportional to the length of each filament. For example, according to the formula from [15], the length of a filament consisting of only two ULFs is 2×43.5 nm $= 87$ nm, while according to the current knowledge regarding filaments measurements, see [2], its length is 63 nm $+ 42$ nm $= 105$ nm. Consequently, [15] introduces a so-called correction factor that only partially addresses the problem. Our approach for computing the MFL value is not influenced by this approximation error and leads to a correct interpretation of the experimental data.

For the experimental data fit of the models, [15] performs a so-called preassessment of the eight variants of the extended model. Based on some *fixed parameter values*, the eight variants are classified into four classes of dynamics. Three of the classes are then quickly dismissed from the analysis and only one representative of the remaining class is chosen for further assessment. This approach is however assuming that the classification of the dynamics of the eight model variants is independent of the parameter values, which is most likely not true for mathematical models with 5 or more parameters, such as those in [15]. In our case the approach was different. During parameter estimation we fitted all the variants of the extended model with respect to the EM 0.45 μM experimental data set. We then took advantage of the available data from the EM 0.9 μM experiment and performed model validation by comparing the predictions of the models with the experimental data. On the contrary, the second set of data was used in [15] in a second round of model fit, yielding different numerical values for the model parameters.

For the sake of having models of small size, in the first part of the paper we do not distinguish between filaments of different sizes and we use for the filament-

filament extensions a "generic" kinetic constant. However, in the second part of the paper we explicitly address the problem of extending the molecular model to distinguish between filaments of different sizes, recognizing that different constants may/should be used depending on the size of the filaments. We approach the problem from a numerical point of view, aiming to build the extended model in such a way that the numerical fit of the original model is preserved. On the other hand, in [18] a physical approach to estimate how the size of the complexes influences the binding rates is taken. However, this approach is based on the hypotheses that: i) reactants are shaped like balls and, especially, ii) the diameter of the balls representing larger complexes is the same as the diameter of the balls representing small complexes. Unfortunately, these assumptions make the approach of [18] unsuitable for filament-filament interactions. The approach might be developed further to suit our models by modifying the reactants-as-balls assumption and/or the assumption regarding the size of the larger complexes. This would require the recalculation of the collision probabilities in the stochastic approach to chemical kinetics. This however is a project in itself, distinct from the aim and scope of this paper.

Conclusions and further work. Our mathematical models show that if tetramers are very quickly (in just a few seconds) assembled into ULFs, then the elongation of filaments with ULFs and with other filaments both play a crucial role in the formation of long intermediate filaments. The elongation with tetramers on the other hand, has negligible quantitative contribution to the filament assembly. One reason for this is that in the case of fast ULF formation, the population of tetramers is very quickly depleted. However, this leaves open the question of the filament assembly dynamics in the case when tetramers would be continuously added to the system, i.e. by an additional synthesis mechanism. To address this problem, we investigated our mathematical models in the case when the turnover of tetramers into ULFs is slower. It turned out that in case the tetramers persist in the system for a longer time, the dynamics of the filament assembly is much richer and several different mechanisms can equally well explain the available experimental data. In fact, even the simple model discussed in [15] and in our paper could be fitted to the experimental data. An *in vitro* experiment where tetramers were added either continuously or at well-chosen time points could offer more insight into the role of tetramer longitudinal aggregation for the process of filament elongation. Choosing the time points when the additional amount of tetramers should be added to the solution could be done based on the analysis of our mathematical models. For example, one could choose the time points where the number of filaments in the solution is close to its maximum, so that the possible interplay between tetramers and filaments has maximum flux.

It is visible already from the experimental data that the system does not reach a steady state within 20 minutes, our time interval of choice. Similarly as in the study in [15], we have focused on the early dynamics of the vimentin filament

assembly, where the kinetics of the system is fast, with tetramers and ULFs being quickly replaced by emerging filaments of various lengths. During this phase, the presence of a large amount of tetramers and, a little later, of short filaments in the solution make far more likely assembly/elongation events rather than disassembly events. For this reason our models turn out to be able to explain the experimental data during the early phase of the assembly, even though they do not include any disassembly or filament breaking mechanisms. The applicability of the models is however tied to the early part of the assembly. Over longer time intervals (e.g., long enough so that the experimental data may potentially show a steady state), the lack of a disassembly mechanism in the models makes them limited in their predictive power. For example, a model with no disassembly or filament breaking mechanism would predict that the system will reach (albeit in a huge interval of time) a steady state where all initial tetramers are integrated into one single filament (of huge length).

The methodology introduced in this paper for increasing the resolution of the filament assembly model helps provide a deep insight into the dynamics of filament self-assembly. Details on the assembly of filaments of various lengths will help in designing finer grained experimental assays that would focus on filaments of different lengths at different time points. In terms of model complexity, increasing the resolution of the model implies a considerable increase in the size of the model, linear in the number of variables and quadratic in the number of reactions. We showed however that the kinetic rate constants can be set from a model of low resolution to one of higher resolution in such a way that the model predictions on the dynamics of the total amount of filaments, regardless of their length, are preserved. In particular, this implies that given generic data on, for example, the mean filament length, the model fit and the model validation problems can be solved on the (smaller) model of low resolution and then extrapolated to the models of higher resolution.

Acknowledgements. We are grateful to Robert Kirmse for the EM data on measurements of the mean filament length of vimentin intermediate filaments. The work of Eugen Czeizler, Elena Czeizler, Andrzej Mizera and Ion Petre was supported by Academy of Finland, grants 129863, 108421, and 122426. Andrzej Mizera is on leave of absence from the Institute of Fundamental Technological Research, Polish Academy of Sciences, Warsaw, Poland.

References

[1] M. M. Ali, C. Storey, and A. Törn. Application of stochastic global optimization algorithms to practical problems. *J. Optim. Theory Appl.*, 95(3):545–563, 1997.

[2] S. Ando, K. ichiro Nakao, R. Gohara, Y. Takasaki, K. Suehiro, and Y. Oishi. Morphological analysis of glutaraldehyde-fixed vimentin intermediate filaments and assembly-intermediates by atomic force microscopy. *Biochimica et Biophysica Acta*, 1702(1):53–65, 2004.

[3] S. M. Baker, K. Schallau, and B. H. Junker. Comparison of different algorithms for simultaneous estimation of multiple parameters in kinetic metabolic models. *J Integr Bioinform*, 7(3):133, 2010.

[4] G. Bonne, M. R. D. Barletta, S. Varnous, H.-M. Bécane, E.-H. Hammouda, L. Merlini, F. Muntoni, C. R. Greenberg, F. Gary, J.-A. Urtizberea, D. Duboc, M. Fardeau, D. Toniolo, and K. Schwartz. Mutations in the gene encoding lamin A/C cause autosomal dominant Emery-Dreifuss muscular dystrophy. *Nat. Genet.*, 21:285–288, 1999.

[5] D. Fatkin, C. MacRae, T. Sasaki, M. R. Wolff, M. Porcu, M. Frenneaux, J. Atherton, H. J. Vidaillet, S. Spudich, U. D. Girolami, J. G. Seidman, C. Seidman, F. Muntoni, G. Müehle, W. Johnson, and B. McDonough. Missense mutations in the rod domain of the lamin A/C gene as causes of dilated cardiomyopathy and conduction-system disease. *N. Engl. J. Med.*, 341(23):1715–1724, 1999.

[6] I. E. Grossmann. *Global optimization in engineering design*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996.

[7] C. Guus, E. Boender, and H. E. Romeijn. Stochastic methods. In R. Horst and P. M. Pardalos, editors, *Handbook of Global Optimization*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1995.

[8] H. Herrmann and U. Aebi. Intermediate filaments: molecular structure, assembly mechanism, and integration into functionally distinct intracellular scaffolds. *Ann Rev Biochem*, 73:749–789, 2004.

[9] H. Herrmann, H. Bär, L. Kreplak, S. V. Strelkov, and U. Aebi. Intermediate filaments: from cell architecture to nanomechanics. *Nature Reviews Mol Cell Biol*, 8:562–573, 2007.

[10] H. Herrmann, M. Häner, M. Brettel, N.-O. Ku, and U. Aebi. Characterization of distinct early assembly units of different intermediate filament proteins. *J. Mol. Biol.*, 286(5):1403–1420, 1999.

[11] K. C. Holmes, D. Popp, W. Gebhard, and W. Kabsch. Atomic model of the actin filament. *Nature*, 347:44–49, 1990.

[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] R. Horst and H. Tuy. *Global optimization: Deterministic approaches*. Springer-Verlag, Berlin, 1990.

[14] R. Kirmse. Personal communication, 2008.

[15] R. Kirmse, S. Portet, N. Mücke, U. Aebi, H. Herrmann, and J. Langowski. A quantitative kinetic model for the in vitro assembly of intermediate filaments from tetrameric vimentin. *J. Biol. Chem.*, 282(52):18563–18572, 2007.

[16] M. Kühnel, L. S. Mayorga, T. Dandekar, J. Thakar, R. Schwarz, E. Anes, G. Griffiths, and J. Reich. Modelling phagosomal lipid networks that regulate actin assembly. *BMC Systems Biology*, 2:107, 2008.

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

[18] L. Lok and R. Brent. Automatic generation of cellular reaction networks with moleculizer 1.0. *Nat. Biotechnol.*, 23:131–136, 2005.

[19] P. Mendes, S. Hoops, S. Sahle, R. Gauges, J. O. Dada, and U. Kummer. Computational modeling of biochemical networks using COPASI. *Methods Mol Biol*, 500:17–59, 2009.

[20] P. Mendes and D. Kell. Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. *Bioinformatics*, 14(10):869–883, 1998.

[21] C. G. Moles, P. Mendes, and J. R. Banga. Parameter estimation in biochemical pathways: A comparison of global optimization methods. *Genome Res*, 13:2467–2474, 2003.

[22] E. Nogales and K. H. Downing. Tubulin and microtubule structures. In T. Fojo, editor, *The Role of Microtubules in Cell Biology, Neurobiology, and Oncology*. Humana Press, 2008.

[23] S. Sahle, R. Gauges, J. Pahle, N. Simus, U. Kummer, S. Hoops, C. Lee, M. Singhal, L. Xu, and P. Mendes. Simulation of biochemical networks using COPASI: a complex pathway simulator. In *WSC '06 Proceedings of the 38th conference on Winter simulation*, pages 1698–1706, 2006.

[24] M. Schliwa. *The Cytoskeleton*, volume 13 of *Cell Biology Monographs*. Springer-Verlag, Vienna, Austria, 1986.

[25] M. O. Steinmetz, D. Stoffler, A. Hoenger, A. Bremer, and U. Aebi. Actin: from cell biology to atomic detail. *J. Struct. Biol.*, 119:295–320, 1997.

[26] H. J. Worman and J. C. Courvalin. The nuclear lamina and inherited disease. *Trends Cell. Biol.*, 12:591–598, 2002.

Paper VIII

Self-assembly models of variable resolution

Andrzej Mizera, Eugen Czeizler, and Ion Petre

TUCS Technical Report number 1014, June 2011.

Self-assembly models of variable resolution

Andrzej Mizera
Department of Information Technologies, Åbo Akademi University,
FI-20520 Turku, Finland
amizera@abo.fi
Eugen Czeizler
Department of Information and Computer Science, Aalto University,
FI-00076 Aalto, Finland
eugen.czeizler@aalto.fi
(Work done while the author was at Åbo Akademi University.)
Ion Petre
Department of Information Technologies, Åbo Akademi University,
FI-20520 Turku, Finland
ipetre@abo.fi

TUCS Technical Report
No 1014, June 2011

Abstract

Model refinement is an important aspect of the model-building process. It can be described as a procedure which, starting from an abstract model of a system, performs a number of refinement steps in result of which a more detailed model is obtained. At the same time, in order to be correct, the refinement mechanism has to be capable of preserving already proven systemic quantitative properties of the original model, e.g. model fit, stochastic semantics, etc. In this study we concentrate on the refinement in the case of self-assembly models. Self-assembly is a process in which a disordered ensemble of basic components forms an organized structure as a result of specific, local interactions among these components, without external guidance. We develop a generic formal model for this process and introduce a notion of model resolution capturing the maximum size up to which objects can be distinguished individually in the model. All bigger objects are treated homogenously in the model. We show how this self-assembly model can be systematically refined in such a way that its resolution can be increased and decreased while preserving the original model fit to experimental data, without the need for tedious, computationally expensive process of parameter refitting. We demonstrate how the introduced methodology can be applied to a previously published model: we consider the case-study of *in vitro* self-assembly of intermediate filaments.

Keywords: Model refinement — Model resolution — Self-assembly — Model fit — Intermediate filaments

TUCS Laboratory
Computational Biomodelling

1 Introduction

The great complexity of biological systems enforces the need for representing them in formal models in order to investigate them and make specific predictions about their behaviour that can be tested in subsequent experiments. Starting from a model abstracting a biological system, the iterative process of hypothesis generation, experimental design, experimental analysis, and model refinement lies at the core of systems biology ([4, 16, 22]). Even more, this approach is proposed as the only logical way for biology to advance ([19]). Development and refinement of a mathematical model of a biochemical process proceeds, in general, in accordance with the following scenario. First, an abstraction of the process is made by identifying a relatively small set of biochemical reactions which are capturing the main features of the process' machinery. The chosen biochemical reactions may be very abstract themselves, i.e. one reaction may in fact encapsulate many real reactions which constitute a whole subprocess in a living organism. Second, the molecular model formed of the chosen reactions is transformed into an associated mathematical model. This usually involves two steps: obtaining equations describing the dynamics of the system by assuming some proper kinetic law, e.g. mass-action law, Michaelis-Menten kinetics, etc., and then identifying the model parameter values so that the model fits some experimental data.

During the process of model development some simplifications and abstractions are introduced. With time, there may be a necessity for them to be refined and modelled in a more detailed, accurate way. However, some carefulness is required on this stage. For example, one could take all the intended changes into consideration while simply repeating the whole model development procedure. But this solution involves repeating from scratch the time-consuming, computationally-intensive model fitting, see [5]. Another approach, not much investigated in the literature, is to refine the model in such a way that the previously obtained fit is preserved. This basically implies deriving the parameter values of the refined model from the ones of the original model.

In this study we concentrate on the step of model refinement in the iterative cycle of systems biology, which is an important aspect of the model-building process. In particular, we develop a refinement procedure for a family of ordinary differential equation (ODE) models describing the process of self-assembly. Self-assembly is a process in result of which some pronounced structures emerge out of an ensemble of scattered basic elements. Important is the fact that the arrangements take place based just on local interactions between the building blocks, without any external guidance. In our work we develop a generic formal model for self-assembly. It consists of an ensemble of all possible objects that can potentially appear in the course of the self-assembly, a composition operation and a mapping from objects of the ensemble to positive integers. The number is interpreted as the size of the considered object. The generic model allows us to further introduce the notion of model resolution. We continue by discussing the refine-

ment of such models, i.e. we formally show how the resolution of a self-assembly model can be increased and decreased while preserving the original model fit to experimental data. We demonstrate how our methodology of self-assembly model refinement can be applied to an existing model. To this aim we utilize the previously published model of the *in vitro* assembly of intermediate filaments from tetrameric vimentin, see [6, 15].

Our methodology of self-assembly model refinement is a particular instance of *formal model refinement*. This topic has been extensively studied in Computer Science, see, e.g., [3, 23, 24], especially in connection to formal software specifications. The method we propose is an instance of *data refinement*, where one replaces a variable with a set of other variables in a way that introduces more details into the model, while keeping the model constraints unchanged.

The paper is organized as follows. First, a general, formal characterization of the self-assembly process is presented. Then, the notion of model resolution is introduced and the model refinement procedure consisting in increasing and decreasing the model resolution while preserving the fit to experimental data is described. Finally, the technique is applied in a case study where the self-assembly of intermediate filaments is considered.

2 A generic model for self-assembly

Self-assembly is a term coined to name processes in which a disordered ensemble of basic components forms an organized structure as a result of specific, local interactions among these components, without external guidance. In a general case, the process of self-assembly can be formalized as follows. We consider an ensemble $\mathcal{E}$ of all possible objects that can potentially appear in the course of the self-assembly process, including the initial ones. Each object O from the ensemble has a scalar value $size(O)$ associated with it and determined through a mapping $size : \mathcal{E} \to \mathbb{N}_+$. Moreover, the objects from $\mathcal{E}$ can combine with each other to form another object from $\mathcal{E}$ in such a way that the sum of the sizes of the objects equals the size of the resulting object. More formally, if we denote the composition operation with $+$, then

$$O_1 + O_2 = O_r \quad \Rightarrow \quad size(O_1) + size(O_2) = size(O_r), \tag{1}$$

where O_r is the object assembled from component objects O_1 and O_2. The ensemble $\mathcal{E}$ together with the binary operation $+$ forms a structure $(\mathcal{E}, +)$, which in abstract algebra is named a *semigroup*. Furthermore, this structure is homomorphic with the $(\mathbb{N}_+, +)$ semigroup by the $size$ map.

Our generic model for self-assembly is on a high level of abstraction, focusing on the $size$ of the emerging structures, while ignoring all details of the topology of such structures. *Size* here can mean any semigroup homomorphism between $(\mathcal{E}, +)$ and $(\mathbb{N}_+, +)$, as noted above. Intuitively, the $size$ map would count the

number of elementary blocks forming the self-assembled structure under consideration. This approach is applicable to any type of self-assembly processes: unidimensional (such as the elongation of intermediate filaments, the case-study investigated in this paper), branched two-dimensional structures, three-dimensional assemblies, etc. However, extending the dynamics of the $size$ distribution of the self-assembled structures with some of their topological details would require a very different type of modelling, which goes beyond the scope of our approach.

Through the map $size$, for a fixed $n \in \mathbb{N}_+$ we define a family of object classes $\mathcal{S}^{(n)} = \{\mathcal{S}_1^{(n)}, \ldots, \mathcal{S}_n^{(n)}, \mathcal{S}_{\geq n+1}^{(n)}\}$: $\mathcal{S}_i^{(n)}$ contains all the objects from $\mathcal{E}$ with size i for $i = 1, \ldots, n$ and $\mathcal{S}_{\geq n+1}^{(n)}$ consists of all objects with size greater than n. Each object from $\mathcal{E}$ belongs to exactly one of these classes. Notice that for $m > n$ we have $\mathcal{S}_k^{(n)} = \mathcal{S}_k^{(m)}$ for all $k \in \{1, \ldots, n\}$.

The composition of objects in $\mathcal{E}$ is described by a system of rules. For the general characterization of self-assembly we will assume that the rules are at the level of abstraction of $\mathcal{S}^{(n)}$, i.e. that the system of rules is of the form

$$\begin{aligned}
\mathcal{S}_i^{(n)} + \mathcal{S}_j^{(n)} &\to \mathcal{S}_{i+j}^{(n)}, && \text{for all } 1 \leq i \leq j \leq n,\ i+j \leq n; \\
\mathcal{S}_i^{(n)} + \mathcal{S}_j^{(n)} &\to \mathcal{S}_{\geq n+1}^{(n)}, && \text{for all } 1 \leq i \leq j \leq n,\ i+j \geq n+1; \\
\mathcal{S}_i^{(n)} + \mathcal{S}_{\geq n+1}^{(n)} &\to \mathcal{S}_{\geq n+1}^{(n)}, && \text{for all } 1 \leq i \leq n; \\
\mathcal{S}_{\geq n+1}^{(n)} + \mathcal{S}_{\geq n+1}^{(n)} &\to \mathcal{S}_{\geq n+1}^{(n)}. &&
\end{aligned} \tag{2}$$

In the case of biochemical systems these rules are usually referred to as (biochemical) reactions and we will use this terminology in the following. The semantics of the reactions in the above form can be described as: an object from class $\mathcal{S}_i^{(n)}$ combines with an object from class $\mathcal{S}_j^{(n)}$ to form an object of class $\mathcal{S}_{i+j}^{(n)}$ if $i+j \leq n$ or $\mathcal{S}_{\geq n+1}^{(n)}$ if $i + j \geq n + 1$. Notice that any reaction of this form automatically satisfies the self-assembly condition (1).

In mathematical modelling it is common to associate a variable (understood as a function) $F : \mathbb{R}_+ \to \mathbb{R}_+$ with each of the sets in $\mathcal{S}^{(n)}$. We denote with $F_i^{(n)}$ the variable corresponding to the set $\mathcal{S}_i^{(n)}$ for $i \in \{1, \ldots, n, \geq n+1\}$. The value of the variable $F_i^{(n)}$ is interpreted as the concentration of objects from the associated set $\mathcal{S}_i^{(n)}$, present in the system undergoing self-assembly at a particular point in time. Further, we assume that the kinetics of the reactions is based on the *law of mass action* ([17]). This law is a mathematical model of reaction dynamics: it states that the reaction rate is proportional to the probability of collision of the reactants, while the probability itself is proportional to the product of concentrations of reactants raised to the number in which they enter the reaction ([17]). We use $k_{i,j}$, $1 \leq i \leq j \leq n+1$ to denote the respective proportionality factor, the so-called *rate constant*, of the reaction with the left-hand side containing $\mathcal{S}_i^{(n)}$ (or $\mathcal{S}_{\geq n+1}^{(n)}$ if $i = n+1$) as one and $\mathcal{S}_j^{(n)}$ (or $\mathcal{S}_{\geq n+1}^{(n)}$ if $j = n+1$) as the other term. For

example,

$$\mathcal{S}_2^{(n)} + \mathcal{S}_3^{(n)} \xrightarrow{k_{2,3}} \mathcal{S}_5^{(n)}$$

and

$$\mathcal{S}_2^{(n)} + \mathcal{S}_{\geq n+1}^{(n)} \xrightarrow{k_{2,n+1}} \mathcal{S}_{\geq n+1}^{(n)}.$$

The change of concentrations in time of the objects undergoing self-assembly can be described using ordinary differential equations (ODEs). By the law of mass action, the system of ODEs associated with the self-assembly system determined by the reactions in (2) is

$$\begin{cases} \dfrac{\mathrm{d}F_i^{(n)}}{\mathrm{d}t} = -\displaystyle\sum_{j=1}^{n} k_{i,j}\, F_i^{(n)}\, F_j^{(n)}\, [i \neq j] - 2\, k_{i,i}\, {F_i^{(n)}}^2 - k_{i,n+1}\, F_i^{(n)}\, F_{\geq n+1}^{(n)} \\ \qquad\qquad + \displaystyle\sum_{j=1}^{\lceil \frac{i-1}{2} \rceil} k_{j,i-j}\, F_j^{(n)}\, F_{i-j}^{(n)} \qquad \text{for all } 1 \leq i \leq n, \\ \dfrac{\mathrm{d}F_{\geq n+1}^{(n)}}{\mathrm{d}t} = \displaystyle\sum_{\substack{1 \leq i \leq j \leq n, \\ i+j \geq n+1}} k_{i,j}\, F_i^{(n)}\, F_j^{(n)} - k_{n+1,n+1}\, {F_{\geq n+1}^{(n)}}^2, \end{cases} \tag{3}$$

where $[\ldots]$ are used as the Iverson brackets ([14, 18]), i.e. $[i \neq j]$ is 1 if $i \neq j$ and 0 otherwise. The negative term in the equation for $\mathrm{d}F_{\geq n+1}^{(n)}/\mathrm{d}t$ originates from the last rule in (2), where two objects from the set $\mathcal{S}_{\geq n+1}^{(n)}$ combine to form a bigger object belonging to the same class. In consequence, in $\mathcal{S}_{\geq n+1}^{(n)}$ two objects are consumed and one is produced, thus the net result is that one object disappears from $\mathcal{S}_{\geq n+1}^{(n)}$.

3 A notion of model resolution

When considering the dynamics of the self-assembly process, one of the main concerns is the distribution of the number of components of different sizes in time. To this aim we introduce the notion of *model resolution* in the context of self-assembly. We say that *a self-assembly model is of resolution* n if it consists of the set of reactions describing the interactions between the classes of objects $\mathcal{S}_1^{(n)}, \ldots, \mathcal{S}_n^{(n)}, \mathcal{S}_{\geq n+1}^{(n)}$, i.e. the set of rules of the form in (2). The associated mathematical model (ODE-based or not), comprising variables $F_1^{(n)}, \ldots, F_n^{(n)}, F_{\geq n+1}^{(n)}$ is also referred to as an n-resolution model. Thus, the system in (3) is a self-assembly ODE model of resolution n. Intuitively, a self-assembly mathematical model is of resolution n if it allows for capturing the dynamics of the number (or concentration) of components that are exactly of size i, where $0 \leq i \leq n$.

In light of this definition the superscript (n) obtains a new meaning: it indicates the resolution of the considered model, i.e. $F_j^{(n)}$ determines the concentration of objects of size j in time in the model of resolution n and $\mathcal{S}_j^{(n)}$ refers to the class of objects of size j which appears in the set of reactions of the n-resolution self-assembly model. This will be useful when considering the relationships between models of various resolutions in the subsequent subsections.

When setting the resolution of our generic self-assembly model we effectively partition the set of possible emerging structures into two, depending on their size:

(i) the set of *visible assemblies* whose size is at most the resolution level, and

(ii) the set of *invisible assemblies* whose size is larger than the resolution level.

The self-assembly process can be modelled in all of its combinatorial details on the set of visible assemblies, including the assembly of all possible pairs of visible assemblies and even their disassembly (disassembly is however not covered in our case-study). For the invisible assemblies (size larger than the resolution level) we only specify a number of generic reactions covering their elongation. The idea here is that the details of the dynamics of such assemblies are beyond the scope (or beyond the experimental measuring capabilities) of our current model.

Choosing the resolution of a self-assembly model should be done in a careful way, so that it includes in its visible assemblies that part of the species space that is important for the model. Changing the resolution of a model may be needed during the modelling process, depending on the application. For example, a model of relatively low resolution may be enough in the early stage of the process, when no (or very few) assemblies of large size exist. Later on however, as the size of the existing self-assembled structures grows, the modeller may need to increase the resolution level to be able to track the details of the interactions involving larger structures. We discuss in the next section a method to increase the model resolution in such a way that the model's numerical fit to experimental data is preserved. Note also that the resolution may be fixed *a priori* to a level that is higher than the number of available molecules, thus making the whole species space visible, with the price that the manipulation of the model (such as the model fit and validation) may become computationally expensive.

3.1 Increasing the model resolution while preserving the model fit

In this section we concentrate on the refinement in the case of the self-assembly models. The aim is to increase the range of component sizes for which the distribution in time is captured by the model, i.e. increase the model resolution, while preserving the data fit of the original model. In the context of the associated mathematical models, we say that a model of resolution $n + 1$ is a *quantitative*

refinement of a model of resolution n if and only if the following quantitative refinement conditions are satisfied:

$$F_i^{(n+1)}(t) = F_i^{(n)}(t), \quad 1 \leq i \leq n \tag{4}$$

and

$$F_{n+1}^{(n+1)}(t) + F_{\geq n+2}^{(n+1)}(t) = F_{\geq n+1}^{(n)}(t), \tag{5}$$

for all $t \geq 0$.

In the case of the self-assembly ODE models of the form in (3), the quantitative refinement from resolution n to $n+1$ involves appropriate setting of the rate constants and the initial values of the model of resolution $n+1$ given the rate constants and the initial values of the model of resolution n. We show in the following how this should be performed.

We start our considerations with the statement of a lemma concerning the existence and uniqueness of solutions of the self-assembly ODE system of any fixed resolution.

Lemma 1. *The system of ODEs for a self-assembly model of resolution n, where $n \in \mathbb{N}$, admits exactly one solution for any fixed initial condition.*

Proof. Let us rewrite (3) in the form

$$\mathbf{F}' = \mathcal{F}(\mathbf{F}),$$

where $\mathbf{F}(t) = [F_1^{(n)}(t), \ldots, F_n^{(n)}(t), F_{\geq n+1}^{(n)}(t)]^{\mathrm{T}}$ and $\mathcal{F} : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ defines a vector field on $\mathbb{R}^{n+1}$. A solution of this system is a function $\mathbf{F} : J \to \mathbb{R}^{n+1}$ defined on some interval $J \subset \mathbb{R}$ such that, for all $t \in J$, $\mathbf{F}'(t) = \mathcal{F}(\mathbf{F}(t))$. Now, it is enough to observe that the right-hand sides of the equations in (3) are continuously differentiable with respect to the coordinates of $\mathbf{F}$. Thus, the mapping $\mathcal{F}$ is Lipschitz continuous on a bounded domain ([8]) and by the Picard-Lindelöf theorem ([8]) it follows that for any initial conditions the considered system has a unique solution $\mathbf{F}(t)$. □

Equipped with Lemma 1, we continue to show how the refinement of a self-assembly model can be effectively achieved. This is the content of the following theorem, where $l_{i,j}$, $1 \leq i \leq j \leq n+2$ denote the rate constants of the $(n+1)$-resolution model and $k_{p,q}$, $1 \leq p \leq q \leq n+1$ are the rate constants of the n-resolution model. A discussion about the biological basis for the numerical choices made in Theorem 1 is included after its proof.

Theorem 1. *Setting the kinetic rate constants of the $(n+1)$-resolution model in the following way*

$$\begin{cases} l_{i,j} \;\; := k_{i,j} & 1 \leq i \leq j \leq n, \\ l_{i,n+1} := k_{i,n+1} & 1 \leq i \leq n, \\ l_{i,n+2} := k_{i,n+1} & 1 \leq i \leq n, \\ l_{n+1,n+2} := 2\, k_{n+1,n+1}, & \\ l_{n+1,n+1} := k_{n+1,n+1}, & \\ l_{n+2,n+2} := k_{n+1,n+1}, & \end{cases} \tag{6}$$

and its initial values so that they satisfy

$$F_i^{(n+1)}(0) = F_i^{(n)}(0), \quad 1 \le i \le n, \tag{7}$$

$$F_{n+1}^{(n+1)}(0) + F_{\ge n+2}^{(n+1)}(0) = F_{\ge n+1}^{(n)}(0) \tag{8}$$

ensures that the self-assembly ODE model of resolution $n+1$ is a quantitative refinement of the self-assembly ODE model of resolution n.

Proof. Let us write the system of ODEs for the model of resolution $n+1$:

$$\begin{cases} \dfrac{\mathrm{d}F_i^{(n+1)}}{\mathrm{d}t} = -\displaystyle\sum_{j=1}^{n} l_{i,j}\, F_i^{(n+1)}\, F_j^{(n+1)}\,[i \ne j] - 2\, l_{i,i}\, F_i^{(n+1)^2} \\ \qquad\qquad - l_{i,n+1}\, F_i^{(n+1)}\, F_{n+1}^{(n+1)} - l_{i,n+2}\, F_i^{(n+1)}\, F_{\ge n+2}^{(n+1)} \\ \qquad\qquad + \displaystyle\sum_{j=1}^{\lceil \frac{i-1}{2} \rceil} l_{j,i-j}\, F_j^{(n+1)}\, F_{i-j}^{(n+1)} \qquad \text{for all } 1 \le i \le n, \\ \dfrac{\mathrm{d}F_{n+1}^{(n+1)}}{\mathrm{d}t} = -\displaystyle\sum_{j=1}^{n} l_{j,n+1}\, F_j^{(n+1)}\, F_{n+1}^{(n+1)} - 2\, l_{n+1,n+1}\, F_{n+1}^{(n+1)^2} \\ \qquad\qquad - l_{n+1,n+2}\, F_{n+1}^{(n+1)}\, F_{\ge n+2}^{(n+1)} + \displaystyle\sum_{j=1}^{\lceil \frac{n}{2} \rceil} l_{j,n+1-j}\, F_j^{(n+1)}\, F_{n+1-j}^{(n+1)} \\ \dfrac{\mathrm{d}F_{\ge n+2}^{(n+1)}}{\mathrm{d}t} = \displaystyle\sum_{\substack{1 \le i \le j \le n, \\ i+j \ge n+2}} l_{i,j}\, F_i^{(n+1)}\, F_j^{(n+1)} + \sum_{j=1}^{n} l_{j,n+1}\, F_j^{(n+1)}\, F_{n+1}^{(n+1)} \\ \qquad\qquad + l_{n+1,n+1} F_{n+1}^{(n+1)^2} - l_{n+2,n+2} F_{\ge n+2}^{(n+1)^2}. \end{cases} \tag{9}$$

Let us further denote by $G^{(n+1)}$ the sum of $F_{n+1}^{(n+1)}$ and $F_{\ge n+2}^{(n+1)}$, i.e.

$$G^{(n+1)}(t) = F_{n+1}^{(n+1)}(t) + F_{\ge n+2}^{(n+1)}(t).$$

With use of the expressions for $\mathrm{d}F_{n+1}^{(n+1)}/\mathrm{d}t$ and $\mathrm{d}F_{\ge n+2}^{(n+1)}/\mathrm{d}t$ in (9), we can compute the derivative of $G^{(n+1)}$

$$\begin{aligned} \frac{\mathrm{d}G^{(n+1)}}{\mathrm{d}t} &= \frac{\mathrm{d}F_{n+1}^{(n+1)}}{\mathrm{d}t} + \frac{\mathrm{d}F_{\ge n+2}^{(n+1)}}{\mathrm{d}t} = \\ &= \sum_{i=1}^{\lceil \frac{n}{2} \rceil} l_{i,n+1-i}\, F_i^{(n+1)}\, F_{n+1-i}^{(n+1)} + \sum_{\substack{1 \le i \le j \le n, \\ i+j \ge n+2}} l_{i,j}\, F_i^{(n+1)}\, F_j^{(n+1)} \\ &\quad - l_{n+1,n+1}\, F_{n+1}^{(n+1)^2} - l_{n+1,n+2}\, F_{n+1}^{(n+1)}\, F_{\ge n+2}^{(n+1)} - l_{n+2,n+2} F_{\ge n+2}^{(n+1)^2}. \end{aligned} \tag{10}$$

By substituting the rate constants in the above expression for $\mathrm{d}G^{(n+1)}/\mathrm{d}t$ in accordance with (6) we obtain that

$$\begin{aligned}\frac{\mathrm{d}G^{(n+1)}}{\mathrm{d}t} &= \sum_{\substack{1\le i\le j\le n,\\ i+j\ge n+1}} k_{i,j}\, F_i^{(n+1)}\, F_j^{(n+1)} - k_{n+1,n+1}\,(F_{n+1}^{(n+1)} + F_{\ge n+2}^{(n+1)})^2 = \\ &= \sum_{\substack{1\le i\le j\le n,\\ i+j\ge n+1}} k_{i,j}\, F_i^{(n+1)}\, F_j^{(n+1)} - k_{n+1,n+1}\, G^{(n+1)^2}. \end{aligned} \tag{11}$$

Now, by substituting the rate constants also in the equations for $\mathrm{d}F_i^{(n+1)}/\mathrm{d}t$ in (9) for all $1 \le i \le n$ and combing with (11) we have that

$$\begin{cases} \dfrac{\mathrm{d}F_i^{(n+1)}}{\mathrm{d}t} = -\displaystyle\sum_{j=1}^{n} k_{i,j}\, F_i^{(n+1)}\, F_j^{(n+1)}\,[i\neq j] - 2\,k_{i,i}\, F_i^{(n+1)^2} \\ \qquad\qquad - k_{i,n+1}\, F_i^{(n+1)}\, G^{(n+1)} + \displaystyle\sum_{j=1}^{\lceil\frac{i-1}{2}\rceil} k_{j,i-j}\, F_j^{(n+1)}\, F_{i-j}^{(n+1)} \\ \qquad\qquad \text{for all } 1\le i\le n, \\ \dfrac{\mathrm{d}G^{(n+1)}}{\mathrm{d}t} = \displaystyle\sum_{\substack{1\le i\le j\le n,\\ i+j\ge n+1}} k_{i,j}\, F_i^{(n+1)}\, F_j^{(n+1)} - k_{n+1,n+1}\, G^{(n+1)^2}. \end{cases} \tag{12}$$

The above system is identical with (3) modulo the renaming of variables, i.e. $F_i^{(n+1)}$ is in place of $F_i^{(n)}$ for all $1 \le i \le n$ and $G^{(n+1)}$ is in place of $F_{\ge n+1}^{(n)}$. Hence, if the initial values are set up as stated in the theorem, then (3) and (12) constitute the same initial value problem. By the existence and uniqueness stated in Lemma 1, there exists exactly one solution to this problem and thus we have that $F_i^{(n)}(t) = F_i^{(n+1)}(t)$ for all $1 \le i \le n$ and $G^{(n+1)}(t) = F_{n+1}^{(n+1)}(t) + F_{\ge n+2}^{(n+1)}(t) = F_{\ge n+1}^{(n+1)}(t)$. □

Notice that what is important for the refinement is that the initial values of the $(n+1)$-resolution model satisfy (8), however how the initial value of $F_{\ge n+1}^{(n)}$ is split into $F_{n+1}^{(n+1)}(0)$ and $F_{\ge n+2}^{(n+1)}(0)$ is irrelevant, i.e. any partition of this value in accordance with (8) leads to a quantitative refinement of the model of resolution n into a model of resolution $n+1$.

The choice of the kinetic rate constants in Theorem 1 for the refined model is consistent with the following basic principle:

> *by distinguishing several subtypes of a reactant, we do not change the kinetics of the reactions they participate in.*

In other words, whenever we distinguish several subspecies A_1, A_2, ..., A_m of a species A, we consider in the refined model that each subspecies A_i participates in the same reactions in which A was participating in the original model and moreover, their kinetics is unchanged. (Extra biological knowledge about kinetic differences among A_1, ..., A_m may be included in the model in a subsequent step; we only focus here on setting up the more detailed model as a quantitative refinement of the original model.) Our reasoning about the model refinement is discrete, in terms of a finite number of subspecies of a given species. Consequently, our reasoning about the reaction kinetics and its changes is also discrete, in terms of collision-based reactions.

When seen as the result of a collision between the reactants, the kinetics of a reaction depends on a biochemical constant (whose value depends on the specifics of the reactants and of the environment) and on the number of possible combinations of reactant molecules, see [9, 10] for a detailed presentation of this approach. The number of such combinations in the case of a collision $A + B$ (say, type 1) is $[A] \cdot [B]$, but in the case of a collision $A + A$ (say, type 2), it is $[A] \cdot ([A] - 1)/2$, where $[A]$, $[B]$ denote the number of molecules of species A and B, respectively. This is the fundamental reason why $l_{n+1,n+2}$ is set in Theorem 1 to a value that is twice as large as the kinetic rate constant of its corresponding reaction in the original model. Indeed, reaction

$$\mathcal{S}^{(n)}_{\geq n+1} + \mathcal{S}^{(n)}_{\geq n+1} \xrightarrow{k_{n+1,n+1}} \mathcal{S}^{(n)}_{\geq n+1} \tag{13}$$

is replaced in the refined model with reactions

$$\mathcal{S}^{(n+1)}_{n+1} + \mathcal{S}^{(n+1)}_{n+1} \xrightarrow{l_{n+1,n+1}} \mathcal{S}^{(n+1)}_{\geq n+2}, \tag{14}$$

$$\mathcal{S}^{(n+1)}_{n+1} + \mathcal{S}^{(n+1)}_{\geq n+2} \xrightarrow{l_{n+1,n+2}} \mathcal{S}^{(n+1)}_{\geq n+2}, \tag{15}$$

$$\mathcal{S}^{(n+1)}_{\geq n+2} + \mathcal{S}^{(n+1)}_{\geq n+2} \xrightarrow{l_{n+2,n+2}} \mathcal{S}^{(n+1)}_{\geq n+2}. \tag{16}$$

When reasoning about the kinetic rate constants of the refined reactions, we preserve the same biochemical constants as in the case of the original reaction (no changes in the biochemical details of the subspecies as compared to the original species, as formulated in our basic principle). The number of combinations of reactants in the various reactions is however different: whereas reactions (13), (14), and (16) are of type 2 (as defined above), reaction (15) is of type 1. If we chose a discrete mathematical model formulation in terms of stochastic processes, then the kinetic rate constants of reactions (14)-(16) would be set to be equal to that of reaction (13). Translating such a model into a continuous, ODE-based model involves a change in the kinetic rate constants, where that of reaction (15) is set to twice that of reactions (13), (14), and (16) to account for the different way of reasoning about collisions in discrete and in continuous terms. Indeed, an ODE-based model considers the kinetic of a reaction of type 2 to be proportional to $[A]^2$,

unlike in the case of a discrete model, where it is proportional to $[A] \cdot ([A]-1)/2$. We refer to [9] for a detailed discussion on the relationship between the stochastic and the deterministic version of a biomodel. We also note that similar choices for the kinetic rate constants were made in [7] when dealing with the refinement of rule-based models. Finally, we remark that the calculations in the proof of Theorem 1 show that our choice of kinetic rate constants, justified by the biochemical arguments above, lead to a numerically-correct quantitative model refinement.

Now, let us consider a more general case, namely the refinement of a model of resolution n to a model of resolution $n+m$. In this case the refinement conditions that need to be satisfied for all $t \geq 0$ are the following:

$$F_i^{(n+m)}(t) = F_i^{(n)}(t), \quad 1 \leq i \leq n$$

and

$$\sum_{j=1}^{m} F_{n+j}^{(n+m)}(t) + F_{\geq n+m+1}^{(n+m)}(t) = F_{\geq n+1}^{(n)}(t).$$

We start our considerations by a simple lemma.

Lemma 2. *The property of a self-assembly ODE model to be the quantitative refinement of another model of lower resolution is transitive, i.e. if the model $\mathcal{M}^{(n+m)}$ of resolution $n+m$ is the refined version of the model $\mathcal{M}^{(n)}$ of resolution n and $\mathcal{M}^{(n+m+k)}$ of resolution $n+m+k$ is the refined version of the model $\mathcal{M}^{(n+m)}$, then $\mathcal{M}^{(n+m+k)}$ is a quantitative refinement of $\mathcal{M}^{(n)}$, where n, m, k are positive integers.*

Proof. By the refinement conditions we have that for all $t \geq 0$

$$\begin{cases} F_i^{(n)}(t) = F_i^{(n+m)}(t), \quad 1 \leq i \leq n, \\ \sum_{i=1}^{m} F_{n+i}^{(n+m)}(t) + F_{\geq n+m+1}^{(n+m)}(t) = F_{\geq n+1}^{(n)(t)} \end{cases}$$

and

$$\begin{cases} \forall_{1 \leq i \leq n+m} \quad F_i^{(n+m)}(t) = F_i^{(n+m+k)}(t), \\ \sum_{i=1}^{k} F_{n+m+i}^{(n+m+k)}(t) + F_{\geq n+m+k+1}^{(n+m+k)}(t) = F_{\geq n+m+1}^{(n+m)(t)}. \end{cases}$$

This implies that

$$F_i^{(n)}(t) = F_i^{(n+m+k)}(t), \quad 1 \leq i \leq n$$

and

$$\begin{aligned} F_{\geq n+1}^{(n)}(t) &= \sum_{i=1}^{m} F_{n+i}^{(n+m)}(t) + \sum_{i=1}^{k} F_{n+m+i}^{(n+m+k)}(t) + F_{\geq n+m+k+1}^{(n+m+k)}(t) = \\ &= \sum_{i=1}^{m} F_{n+i}^{(n+m+k)}(t) + \sum_{i=1}^{k} F_{n+m+i}^{(n+m+k)}(t) + F_{\geq n+m+k+1}^{(n+m+k)}(t) = \\ &= \sum_{i=1}^{m+k} F_{n+i}^{(n+m+k)}(t) + F_{\geq n+m+k+1}^{(n+m+k)}(t). \end{aligned}$$

Thus it follows that the model of resolution $n+m+k$ constitutes a refinement of the model of resolution n. □

In the next theorem we show how the quantitative refinement of the model of resolution n to the one of resolution $n+m$ can be effectively achieved. We denote by $l_{i,j}$, $1 \leq i \leq j \leq n+m+1$ the rate constants of the $(n+m)$-resolution self-assembly model $\mathcal{M}^{(n+m)}$ and by $k_{p,q}$, $1 \leq p \leq q \leq n+1$ the ones of the n-resolution self-assembly model $\mathcal{M}^{(n)}$.

Theorem 2. *Setting the kinetic rate constants of the $(n+m)$-resolution self-assembly ODE model $\mathcal{M}^{(n+m)}$ in accordance with the rate constants of the n-resolution self-assembly ODE model $\mathcal{M}^{(n)}$ in the following way*

$$\begin{cases} l_{i,j} := k_{i,j} & 1 \leq i \leq j \leq n+1, \\ l_{i,n+j} := k_{i,n+1} & 1 \leq i \leq n,\ 2 \leq j \leq m+1, \\ l_{n+i,n+i} := k_{n+1,n+1} & 2 \leq i \leq m+1, \\ l_{n+i,n+j} := 2\,k_{n+1,n+1} & 1 \leq i < j \leq m+1, \end{cases} \tag{17}$$

and its initial values so that they satisfy

$$F_i^{(n+m)}(0) = F_i^{(n)}(0), \quad 1 \leq i \leq n, \tag{18}$$

$$\sum_{i=1}^{m} F_{n+i}^{(n+m)}(0) + F_{\geq n+m+1}^{(n+m)}(0) = F_{\geq n+1}^{(n)}(0) \tag{19}$$

ensures that $\mathcal{M}^{(n+m)}$ is a quantitative refinement of $\mathcal{M}^{(n)}$.

Proof. The proof is by induction on m. The basis of the induction which is the step from resolution n to $n+1$ ($m=1$) is given by Theorem 1. The statement of Theorem 2 clearly holds in this case and we proceed to the inductive step. We assume that the statement is true for $m=z$ for some $z \geq 2$ and we consider the case where $m = z+1$. Theorem 1 assures that setting

$$\begin{cases} l_{i,j}^{(n+z+1)} := l_{i,j}^{(n+z)} & 1 \leq i \leq j \leq n+z, \\ l_{i,n+z+1}^{(n+z+1)} := l_{i,n+z+1}^{(n+z)} & 1 \leq i \leq n+z, \\ l_{i,n+z+2}^{(n+z+1)} := l_{i,n+z+1}^{(n+z)} & 1 \leq i \leq n+z, \\ l_{n+z+1,n+z+1}^{(n+z+1)} := l_{n+z+1,n+z+1}^{(n+z)}, & \\ l_{n+z+1,n+z+2}^{(n+z+1)} := 2\,l_{n+z+1,n+z+1}^{(n+z)}, & \\ l_{n+z+2,n+z+2}^{(n+z+1)} := l_{n+z+1,n+z+1}^{(n+z)} & \end{cases} \tag{20}$$

and the initial values of $F_{n+z+1}^{(n+z+1)}$ and $F_{\geq n+z+2}^{(n+z+1)}$ in such a way that

$$F_{n+z+1}^{(n+z+1)}(0) + F_{\geq n+z+2}^{(n+z+1)}(0) = F_{\geq n+z+1}^{(n+z)}(0) \tag{21}$$

is satisfied results in a refinement from the self-assembly model $\mathcal{M}^{(n+z)}$ of resolution $n+z$ to the model $\mathcal{M}^{(n+z+1)}$ of resolution $n+z+1$ (the subscripts of the

kinetic rate constants in (20) indicate the reactions and the superscripts the models in terms of their resolution). By the induction hypothesis setting

$$\begin{cases} l_{i,j}^{(n+z)} := k_{i,j} & 1 \le i \le j \le n+1, \\ l_{i,n+j}^{(n+z)} := k_{i,n+1} & 1 \le i \le n,\ 2 \le j \le z, \\ l_{n+i,n+i}^{(n+z)} := k_{n+1,n+1} & 2 \le i \le z, \\ l_{n+i,n+j}^{(n+z)} := 2\,k_{n+1,n+1} & 1 \le i \le j \le z, \\ l_{i,n+z+1}^{(n+z)} := k_{i,n+1} & 1 \le i \le n, \\ l_{n+i,n+z+1}^{(n+z)} := 2\,k_{n+1,n+1} & 1 \le i \le z, \\ l_{n+z+1,n+z+1}^{(n+z)} := k_{n+1,n+1} \end{cases} \tag{22}$$

and the initial values of $F_{n+i}^{(n+z)}$ and $F_{\ge n+z+1}^{(n+z)}$, where $1 \le i \le z$ in such a way that

$$\sum_{i=1}^{z} F_{n+i}^{(n+z)}(0) + F_{\ge n+z+1}^{(n+z)}(0) = F_{\ge n+1}^{(n)}(0) \tag{23}$$

is satisfied gives a refinement of $\mathcal{M}^{(n)}$ to $\mathcal{M}^{(n+z)}$. Combining (20) with (22) results in

$$l_{i,j}^{(n+z+1)} := k_{i,j} \qquad 1 \le i \le j \le n+1, \tag{24}$$

$$l_{i,n+j}^{(n+z+1)} := k_{i,n+1} \qquad 1 \le i \le n,\ 2 \le j \le z, \tag{25}$$

$$l_{n+i,n+i}^{(n+z+1)} := k_{n+1,n+1} \qquad 2 \le i \le z, \tag{26}$$

$$l_{n+i,n+j}^{(n+z+1)} := 2\,k_{n+1,n+1} \qquad 1 \le i < j \le z, \tag{27}$$

$$l_{i,n+z+1}^{(n+z+1)} := k_{i,n+1} \qquad 1 \le i \le n, \tag{28}$$

$$l_{n+i,n+z+1}^{(n+z+1)} := 2\,k_{n+1,n+1} \qquad 1 \le i \le z, \tag{29}$$

$$l_{i,n+z+2}^{(n+z+1)} := k_{i,n+1} \qquad 1 \le i \le n, \tag{30}$$

$$l_{n+i,n+z+2}^{(n+z+1)} := 2\,k_{n+1,n+1} \qquad 1 \le i \le z, \tag{31}$$

$$l_{n+z+1,n+z+1}^{(n+z+1)} := k_{n+1,n+1}, \tag{32}$$

$$l_{n+z+1,n+z+2}^{(n+z+1)} := 2\,k_{n+1,n+1}, \tag{33}$$

$$l_{n+z+2,n+z+2}^{(n+z+1)} := k_{n+1,n+1}. \tag{34}$$

Putting together (25), (28) and (30) gives $l_{i,n+j}^{(n+z+1)} := k_{i,n+1}$ for $1 \le i \le n$ and $2 \le j \le z+2$; combining (26), (32) and (34) results in $l_{n+i,n+i}^{(n+z+1)} := k_{n+1,n+1}$ for $2 \le i \le z+2$; finally, (27), (29), (31) and (33) can be simply written as $l_{n+i,n+j}^{(n+z+1)} := 2\,k_{n+1,n+1}$ for $1 \le i \le j \le z+2$. Together with (24) this coincides with (17). Moreover, (23) together with (21) gives (19). By Lemma 2, since $\mathcal{M}^{(n+z)}$ refines $\mathcal{M}^{(n)}$ and $\mathcal{M}^{(n+z+1)}$ refines $\mathcal{M}^{(n+z)}$, we have that $\mathcal{M}^{(n+z+1)}$ is a refinement of $\mathcal{M}^{(n)}$. This proves the induction hypothesis. □

3.2 Decreasing the model resolution while preserving the model fit

Let us now consider the reverse problem. Given a self-assembly model of certain resolution, say $n+1$, we want to obtain a self-assembly model of resolution n such that the model of resolution $n+1$ constitutes its quantitative refinement. We refer to this problem as the problem of decreasing model resolution. As in the case of increasing model resolution, the ODE systems of these two models are (3) and (9). However, now the known rate constants are the ones of the model of resolution $n+1$, i.e. $l_{i,j}$ for all $1 \leq i \leq j \leq n+2$, and the task is to set appropriately the values of the rate constants $k_{i,j}$, $1 \leq i \leq j \leq n+1$ of the model of resolution n.

In this presentation we restrict our considerations to the particular case where $k_{i,j} := l_{i,j}$ for all $1 \leq i \leq j \leq n$. This is in accordance with the biological motivation of the model: species that were modelled explicitly in the original model and continue to be so in the new model should not see their kinetics changed. From a mathematical point of view, one could also consider a general approach where the constants $k_{i,j}$, $1 \leq i \leq j \leq n$ are part of the unknowns. In this case, a similar approach would be applicable, leading however to more complicated equations.

We investigate how to set the remaining constants, i.e. $k_{i,n+1}$, $1 \leq i \leq n+1$, so that the quantitative refinement conditions are satisfied. Since we want for the two models to satisfy (4) and (5), based on (3) and the fact that $k_{i,j} := l_{i,j}$ for all $1 \leq i \leq j \leq n$ the derivatives of $F_i^{(n+1)}$, $1 \leq i \leq n$ and $(F_{n+1}^{(n+1)} + F_{\geq n+2}^{(n+1)})$ can be expressed as

$$
\begin{cases}
\dfrac{\mathrm{d}F_i^{(n+1)}}{\mathrm{d}t} = -\displaystyle\sum_{j=1}^{n} l_{i,j}\, F_i^{(n+1)}\, F_j^{(n+1)}\,[i \neq j] - 2\, l_{i,i}\, {F_i^{(n+1)}}^2 \\
\qquad\qquad - k_{i,n+1}\, F_i^{(n+1)}\,(F_{n+1}^{(n+1)} + F_{\geq n+2}^{(n+1)}) + \displaystyle\sum_{j=1}^{\lceil \frac{i-1}{2} \rceil} l_{j,i-j}\, F_j^{(n+1)}\, F_{i-j}^{(n+1)} \\
\qquad\qquad \text{for all } 1 \leq i \leq n, \\
\dfrac{\mathrm{d}(F_{n+1}^{(n+1)} + F_{\geq n+2}^{(n+1)})}{\mathrm{d}t} = \displaystyle\sum_{\substack{1 \leq i \leq j \leq n, \\ i+j \geq n+1}} l_{i,j}\, F_i^{(n+1)}\, F_j^{(n+1)} \\
\qquad\qquad - k_{n+1,n+1}\,(F_{n+1}^{(n+1)} + F_{\geq n+2}^{(n+1)})^2.
\end{cases}
$$

Now, we equalize the right-hand sides in the above system with the respective right-hand sides in the model of resolution $n+1$, i.e. (9), where the expressions for the derivatives of $F_{n+1}^{(n+1)}$ and $F_{\geq n+2}^{(n+1)}$ are added up to obtain an expression for $\mathrm{d}(F_{n+1}^{(n+1)} + F_{\geq n+2}^{(n+1)})/\mathrm{d}t$. After simplifying we obtain that the rate constants $k_{i,n+1}$,

$1 \leq i \leq n+1$ have to satisfy

$$\begin{gathered} l_{i,n+1}\, F_i^{(n+1)}\, F_{n+1}^{(n+1)} + l_{i,n+2}\, F_i^{(n+1)}\, F_{\geq n+2}^{(n+1)} \\ = \\ k_{i,n+1}\, F_i^{(n+1)}\, (F_{n+1}^{(n+1)} + F_{\geq n+2}^{(n+1)}) \end{gathered} \tag{35}$$

and

$$\begin{gathered} l_{n+1,n+1}\, {F_{n+1}^{(n+1)}}^2 + l_{n+1,n+2}\, F_{n+1}^{(n+1)}\, F_{n+2}^{(n+1)} + l_{n+2,n+2}\, {F_{\geq n+2}^{(n+1)}}^2 \\ = \\ k_{n+1,n+1}\, (F_{n+1}^{(n+1)} + F_{\geq n+2}^{(n+1)})^2 \end{gathered} \tag{36}$$

independently of time, i.e. at any time point t, where $t \geq 0$. We do not reduce (35) by dividing its sides by $F_i^{(n+1)}$ since the variable for a particular i may be identically zero. In such case the rate constant $k_{i,n+1}$ can admit an arbitrary value. At the same time we notice that if for all $1 \leq i \leq n$ the variables $F_i^{(n+1)}$ are not identically zero, then such reduction can be done without loss of generality and in this case all $k_{i,n+1}$ admit the same value.

The variables $F_i^{(n+1)}$s are in fact functions of time which constitute a solution to the system of nonlinear, first-order differential equations in (9). Having the explicit solutions, one could easily check whether there exist $k_{i,n+1}$, $1 \leq i \leq n+1$ such that (35) and (36) are satisfied at any time point $t \geq 0$. However, to the best of our knowledge, obtaining an analytical solution to (9) in a general case, i.e. for arbitrary n, is infeasible. Thus, we consider numerical integration of the system and propose the following procedure for checking whether the reduction of resolution in the discussed case can be performed and, if yes, how the rate constants should be set. First, we numerically integrate the ODE system for the model of resolution $n+1$ in (9) to identify all i, $1 \leq i \leq n$, for which the product $F_i^{(n+1)}\, (F_{n+1}^{(n+1)} + F_{\geq n+2}^{(n+1)})$ is identically zero. In all these cases any arbitrary value of the rate constant $k_{i,n+1}$ satisfies (35). For the remaining is we pick a time point at which the product is non-zero and simply solve (35) for $k_{i,n+1}$ at the chosen time point. Similarly, we solve (36) for the value of $k_{n+1,n+1}$ at a time point at which $F_{n+1}^{(n+1)} + F_{\geq n+2}^{(n+1)}$ is non-zero. Second, in order to be correct, the values of the rate constants have to satisfy the refinement conditions without exception at any arbitrary time point. The correctness can be checked numerically by setting the initial values of the n-resolution model as follows

$$\begin{cases} F_i^{(n)}(0) := F_i^{(n+1)}(0), \quad 1 \leq i \leq n, \\ F_{\geq n+1}^{(n)}(0) := F_{n+1}^{(n+1)}(0) + F_{\geq n+2}^{(n+1)}(0) \end{cases}$$

and investigating whether the dynamics of the two considered models satisfy (4) and (5). The numerical check provides the ultimate answer whether the resolution decrease is realizable or not in the discussed case. Notice that if the values of

the rate constants of the model of resolution $n+1$, say $\mathcal{M}^{(n+1)}$, are such that $l_{n+1,n+1} = l_{n+2,n+2}$, $l_{n+1,n+2} = 2\, l_{n+1,n+1}$ and $l_{i,n+1} = l_{i,n+2}$, for all $1 \leq i \leq n$, then the decrease of resolution can be simply achieved by changing the sides of the assignments in (6). In particular, if $\mathcal{M}^{(n+1)}$ were the result of applying Theorem 1 to a model of resolution n $\mathcal{M}^{(n)}$, then this way of decreasing the resolution of $\mathcal{M}^{(n+1)}$ recovers $\mathcal{M}^{(n)}$.

4 A case study: the self-assembly of intermediate filaments

One of the characteristics of eukaryotic cells is the existence of the cytoskeleton – an intricate network of protein filaments that extends throughout the cytoplasm. It enables the cells to adopt a variety of shapes, interact mechanically with the environment, organize the many components in their interior, carry out coordinated and directed movements. It also provides the machinery for intracellular movements, e.g. transport of organelles in the cytoplasm and the segregation of chromosomes at mitosis ([1, 2]). There are three kinds of protein filaments that form the cytoskeleton: actin filaments, intermediate filaments (IFs) and microtubules. Each kind has different mechanical properties and is assembled from an individual type of proteins. Actin filaments and microtubules are formed from *globular* proteins (*actin* and *tubulin* subunits, respectively), whereas *fibrous proteins* are the building blocks of intermediate filaments ([2, 11]). Thousands of these basic elements assemble into a construction of girders and ropes that spreads throughout the cell.

One of the main functions of intermediate filaments is to provide cells with mechanical strength and they are especially prominent in the cytoplasm of cells that are exposed to such conditions. For example, IFs are abundantly present along nerve cells axons where they provide crucial internal reinforcement of these long cell extensions. They can also be observed in great number in muscle cells and epithelial cells. IFs are characterized by great tensile strength. By stretching and distributing the effect of locally applied forces, they protect cells and their membranes against breaking due to mechanical shear. Compared with microtubules and actin filaments, IFs are more stable, tough and durable, e.g. remain intact during exposure of cells to salt solutions and nonionic detergents, while the rest of the cytoskeleton is mostly destroyed ([1]).

Intermediate filaments can be grouped into four classes: (1) *keratin filaments* in epithelial cells; (2) *vimentin filaments* in connective-tissue cells, muscle cells and supporting cells of the nervous system; (3) *neurofilaments* in nerve cells; and (4) *nuclear lamins*, which strengthen the nuclear membrane of all eukaryotic cells, see [1]. In [15] a quantitative kinetic model for the *in vitro* self-assembly of intermediate filaments from tetrameric vimentin was considered. The authors introduced two molecular models (the so-called *simple* and *extended* models) of this

process. In general, the *in vitro* assembly of vimentin IF proteins can be described as a process consisting of three major phases: (i) formation of the unit-length filaments (ULFs); (ii) longitudinal annealing of ULFs and growing filaments; (iii) radial compaction of immature (16 nm diameter) filaments into mature (11 nm diameter) IFs ([12, 13]). However, in both models of [15] the last, third phase was excluded from consideration.

In the case of the simple model from [15], ULFs are treated as ordinary filaments. Moreover, as discussed in [6, 15], the extension of filaments with tetramers plays an insignificant numerical role. This correlates with an experimental observation that *in vitro*, starting from an initial pool of tetramers, tetramers quickly turn into ULFs. Thus, the filament elongation by tetramers is inhibited in the beginning by the lack of filaments and later by the lack of free tetramers. In consequence, the assembly process is described through the following sequence of molecular events:

$$\begin{gathered} 2\,T_1 \xrightarrow{k_1} T_2 \qquad 2\,T_2 \xrightarrow{k_2} T_4 \qquad 2\,T_4 \xrightarrow{k_3} T_{\geq 8} \\ 2\,T_{\geq 8} \xrightarrow{k_f} T_{\geq 8} \end{gathered} \tag{37}$$

where T_1 is interpreted as a tetramer, T_2 as an octamer, T_4 as a hexadecamer and, finally, $T_{\geq 8}$ is an emerging filament, having at least one ULF.

In [6] and [15] the model is fit to experimental data of [15]. The raw data consists of four sets, each containing the length distributions of growing filaments at distinct time points up to 20 min. The data sets are obtained for two initial concentrations of tetramers, i.e. 0.45μM and 0.9μM, in two cases: first, with adsorption onto carbon-coated copper grids and second, with adsorption onto mica support. The filament length distributions are determined from electron microscopy (EM) images and atomic force microscopy (AFM) images in the first and second case, respectively. For each set the time-dependent mean filament length (MFL) is calculated and only the processed data are reported in [15]. The models in [6, 15] are capable of reproducing the experimental data on time-dependent dynamics of the mean filament length, however are unsuitable for capturing the time-dependent distribution of the filament lengths. In consequence, the information carried by the available experimental data is not utilized to the full extent. The high resolution of the data is not incorporated into the models, the predictive power of the models is significantly limited since no predictions about the length distributions in time are possible, and the models cannot be fully validated against the available biological knowledge. This highlights the necessity for high-resolution models as a tool for better understanding of the still little-known process of filament self-assembly. In order to meet this requirement, we apply our methodology of quantitative model refinement to (37). By increasing the resolution with two in two steps we get the

following models: first

$$\begin{array}{cc} 2\,T_1 \xrightarrow{k_1} T_2 & 2\,T_2 \xrightarrow{k_2} T_4 \\ 2\,T_4 \xrightarrow{k_3} T_8 & 2\,T_8 \xrightarrow{k_f} T_{\geq 9} \\ T_8 + T_{\geq 9} \xrightarrow{2\,k_f} T_{\geq 9} & 2\,T_{\geq 9} \xrightarrow{k_f} T_{\geq 9} \end{array} \tag{38}$$

and next

$$\begin{array}{ccc} 2\,T_1 \xrightarrow{k_1} T_2 & 2\,T_2 \xrightarrow{k_2} T_4 & 2\,T_4 \xrightarrow{k_3} T_8 \\ 2\,T_8 \xrightarrow{k_f} T_{\geq 10} & T_8 + T_9 \xrightarrow{2\,k_f} T_{\geq 10} & T_8 + T_{\geq 10} \xrightarrow{2\,k_f} T_{\geq 10} \\ 2\,T_9 \xrightarrow{k_f} T_{\geq 10} & T_9 + T_{\geq 10} \xrightarrow{2\,k_f} T_{\geq 10} & 2\,T_{\geq 10} \xrightarrow{k_f} T_{\geq 10}. \end{array}$$

Note that T_9 is not a product of any reaction and it will not become one in any further refinement of the model. Since in our experimental set-up we have $T_9(0) = 0$, it follows that $T_9(t) = 0$ for all $t \geq 0$, i.e. reactions $T_8 + T_9 \to T_{\geq 10}$, $2\,T_9 \to T_{\geq 10}$ and $T_9 + T_{\geq 10} \to T_{\geq 10}$ can be eliminated. Thus, the model of resolution 8 coincides with the model of resolution 9. With the same reasoning, all models of resolution between 8 and 15 are identical. The model of resolution 16 is however different:

$$\begin{array}{ccc} 2\,T_1 \xrightarrow{k_1} T_2 & 2\,T_2 \xrightarrow{k_2} T_4 & 2\,T_4 \xrightarrow{k_3} T_8 \\ 2\,T_8 \xrightarrow{k_f} T_{16} & T_8 + T_{16} \xrightarrow{2\,k_f} T_{\geq 17} & 2\,T_{16} \xrightarrow{k_f} T_{\geq 17} \\ T_{16} + T_{\geq 17} \xrightarrow{2\,k_f} T_{\geq 17} & 2\,T_{\geq 17} \xrightarrow{k_f} T_{\geq 17}. & \end{array}$$

Thus, in a model of resolution n, for some arbitrary $n \geq 8$, the variables of the model are T_1, T_2, T_4, T_8, T_{16}, T_{24}, ..., T_{8k}, $T_{\geq n+1}$, where $k = \lfloor n/8 \rfloor$. The biological interpretation of the variable T_{8i}, $1 \leq i \leq k$, is the species of filament consisting of i complete ULFs. Using the terminology of [6] and [15], these are the filaments of length i. Thus, our model of resolution n is in fact the model of resolution $\lfloor n/8 \rfloor$ in terms of the number of complete ULFs included in the filament. This can be seen by rewriting the model (38) as follows (with some of the rate constants renamed):

$$\begin{array}{cc} 2\,T_1 \xrightarrow{k_1} T_2 & 2\,T_2 \xrightarrow{k_2} T_4 \\ 2\,T_4 \xrightarrow{k_3} F_1 & 2\,F_1 \xrightarrow{k_4} F_{\geq 2} \\ F_1 + F_{\geq 2} \xrightarrow{k_u} F_{\geq 2} & 2\,F_{\geq 2} \xrightarrow{k_f} F_{\geq 2}, \end{array} \tag{39}$$

where F_1 stands for filament of length 1 (denoted as T_8 in (38)), and $F_{\geq 2}$ stands for the longer filaments (denoted as $T_{\geq 9}$ in (38)). The refinement of this model to

Rate constant	k_1	k_2	k_3	k_4	k_u	k_f
Value	3	30	30	0.25	0.95	0.11

Table 1: Kinetic rate constant values of the extended IF self-assembly model with fast ULF formation (39). The unit is $\frac{1}{\mu M \cdot s}$.

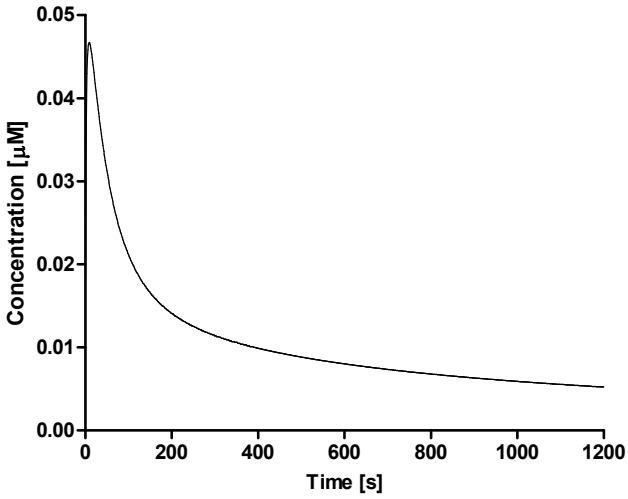
a higher resolution level, say $n \geq 2$, can be done as follows:

$$\begin{aligned}
2\,T_1 &\xrightarrow{k_1} T_2 \\
2\,T_2 &\xrightarrow{k_2} T_4 \\
2\,T_4 &\xrightarrow{k_3} F_1 \\
F_1 + F_1 &\xrightarrow{k_4} F_2 \\
F_1 + F_i &\xrightarrow{k_u} F_{i+1} \qquad i \in \{2, \ldots, n, \geq n+1\} \\
F_i + F_i &\xrightarrow{k_f} F_{2i} \qquad 2 \leq i \leq n \\
F_i + F_j &\xrightarrow{2\,k_f} F_{i+j} \qquad 2 \leq i < j \leq n \\
F_i + F_{\geq n+1} &\xrightarrow{2\,k_f} F_{\geq n+1} \qquad 2 \leq i \leq n \\
F_{\geq n+1} + F_{\geq n+1} &\xrightarrow{k_f} F_{\geq n+1},
\end{aligned} \tag{40}$$

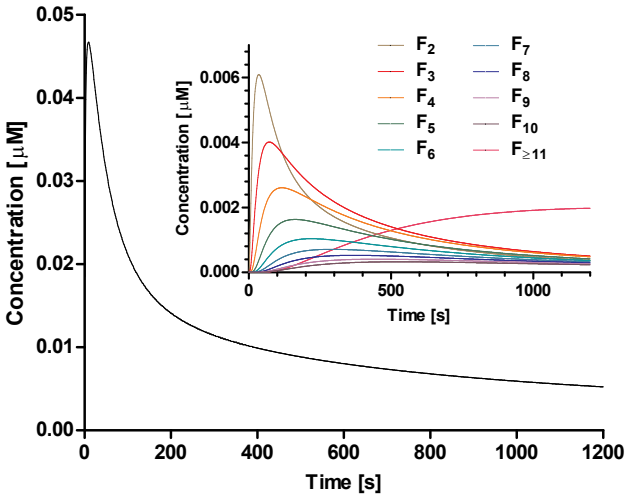
where we adopt the convention that all Fs with indices greater than n are identified with $F_{\geq n+1}$. Model (39) has been experimentally validated in [6]. Using the kinetic constants in Table 1, the numerical behaviour of the model correlates very well with experimental data in [15] on the *in vitro* assembly process of recombinant vimentin at $37\,^{\circ}\mathrm{C}$. Next, we refine the model in (39) by setting $n = 10$ in (40). In result we obtained a model of resolution 10 for the process of *in vitro* intermediate filament self-assembly that preserves the experimental data fit of the original model. In Figure 1 the dynamics of the overall concentration of filaments predicted by (39) and the model of resolution 10 are presented. Notice that the results are identical, which is in complete agreement with the theoretical deliberations, and there is no need for tedious parameter estimation during the construction of the high-resolution model.

5 Discussion

In this work we concentrated on model refinement, an important aspect of the model-building process. In general, the concept of model refinement can be described as a procedure which, starting with an abstract model of a system, carries out a number of refinement steps which lead to the construction of a more detailed



(a)



(b)

Figure 1: Comparison between the dynamics of the extended model of IFs self-assembly with fast ULF formation originally introduced in [6] and the refined version of resolution 10. (a) The original extended model with fast ULF formation introduced in [6]. The curve shows the concentration of the intermediate filaments of any length in time. (b) The refined version with resolution 10. The colour curves of the subplot show the dynamics of IFs of lengths from the set $\{1, \ldots, 10\}$ and the overall concentration of filaments of length greater than 10. The black curve in the main plot is obtained by summing the concentrations in time of filaments of length 1 to 10 and those of length greater than 10. Notice that the two models predict identical overall concentration of IFs in time.

model. At the same time, in order to be correct, the refinement mechanism should be capable of preserving already proven system properties of the original model, e.g. model fit, stochastic semantics, etc. In particular, in our study we focused on the issue of refining an ODE model describing the process of self-assembly. We introduced the notion of model resolution and showed how the resolution can be both increased and decreased while satisfying the condition of preserving the model fit. Moreover, we showed how the technique can be applied to an existing model: we considered the case-study of self-assembly of intermediate filaments.

Restricted sets of reactions There are two ways of restricting the set of reactions of a generic self-assembly model: either by considering just the intended subset of all possible reactions or by setting to zero the kinetic rate constants for those reactions that are not taking place. It is worth noticing that in both cases the refinement procedure will lead to the correct, expected model: in the first case none of the unwanted reactions will be introduced to the new model and in the second case all the new reactions related through the refinement to the original reactions with the rate constant set to zero will remain inactive, i.e. their rate constants will be zero as well.

Models of infinite resolution In this study we discussed the refinement of a self-assembly model of resolution n to the model of resolution $n+m$, where n and m are some fixed positive integers. One could however think of a refinement to the model of infinite resolution. Although we believe that our methodology would work also in this case, formal theoretical considerations of this issue are much more intricate. Already at the stage of writing the differential equations of the model one needs to make sure that the appearing infinite function series are convergent. For example, let us consider a model of resolution 0, i.e $F + F \xrightarrow{k} F$, and refine it to a model of infinite resolution by assuming in accordance with our methodology that $k_{i,j} := 2k$ for $1 \leq i < j \leq \infty$ and $k_{i,i} := k$ for $1 \leq i \leq \infty$. The solution to the ODE model associated with the 0-resolution model, i.e. $\mathrm{d}F/\mathrm{d}t = -kF^2(t)$, can be obtained analytically: $F(t) = F(0)/(1 + k\,t\,F(0))$. In the case of the infinite resolution model one already faces a problem of function series convergence while writing the differential equations for F_is. For each fixed i, the expression for the derivative $\mathrm{d}F_i/\mathrm{d}t$ contains a finite number of terms $k_{l,j}F_iF_j$ where $l+j=i$ with $1 \leq l \leq j < i$, and an infinite number of terms $-k_{i,j}F_iF_j$ where $j \geq 1$. The trouble is whether the infinite series $\sum_{j=1}^{\infty} k_{i,j}F_iF_j$ is convergent for all $t \geq 0$ or whether the terms can be reordered in such a way that the requirement of convergence is satisfied. The difficulty is increased by the fact that the explicit formulas for F_is are unknown. Further, in order for the refinement to be correct, the infinite function series $\sum_{i=1}^{\infty} F_i(t)$ has to be convergent to $F(t)$, i.e. $\sum_{i=1}^{\infty} F_i(t) = F(t)$. If $\sum_{i=1}^{\infty} \mathrm{d}F_i(t)/\mathrm{d}t$ were uniformly convergent, one could

write

$$\mathrm{d}F/\mathrm{d}t = \sum_{i=1}^{\infty} \mathrm{d}F_i(t)/\mathrm{d}t. \tag{41}$$

In order to check whether the refinement condition is satisfied, it would be enough to verify (41) and make sure that $\sum_{i=1}^{\infty} F_i(0) = F(0)$. To this aim, by the refinement condition, the left-hand side in (41) could be written as

$$\mathrm{d}F/\mathrm{d}t = -k(\sum_{n=1}^{\infty}\sum_{i=1}^{n} F_{n-i}F_i),$$

where the Cauchy product of $(\sum_{i=1}^{\infty} F_i(t))^2$ is considered. Now, satisfiability of (41) could be checked by proper reordering of the terms on the both sides of (41). However, prior to this, one would need to make sure that all the convergence conditions required by such reorderings are fulfilled. We just signal this issue here without providing a solution to this interesting problem and leave it for further investigation.

Related work The discussed methods for decreasing and increasing the resolution of self-assembly ODE models can be viewed as examples of adaptations of formal model refinement techniques from the field of computer science to systems biology. To the best of our knowledge, formal model refinement has not been explored much in the context of systems biology and this is the first time that it is considered in relation to computational ODE-based models. Some attempts have been made previously in the case of the rule-based formalism, see [7, 21], where the authors consider a process called the *rule refinement*. It is a method to refine rule sets in such a way that the stochastic semantics, dictated by the number of different ways in which a given rule can be applied to a system, is preserved. It is shown how to refine rules and how to choose the refined rates so that the global dynamics of the original and refined systems are the same. For more details we refer to [7, 21].

In Section 3.1, we discussed the numerical choices for the rate constants of the refined self-assembly model and we presented the biological basis for them. However, in general, when considering refinement of reactions describing assembly of larger and lager complexes, one could think of deriving the rate constants based on physical deliberations, i.e. try to estimate how the size of the complexes influences the binding rates. Such an attempt was originally made in [20], where the collision probabilities in the stochastic approach to chemical kinetics were recalculated with taking into account the change in the masses of complexes under formation. However, the solution presented in [20] is not completely satisfactory due to the following two assumptions it is based on: i) reactants are shaped like balls, and, especially, ii) the diameter of the balls representing larger complexes is the same as the diameter of the balls representing small complexes. Nevertheless, this approach seems to have the potential to be developed further to correctly

address the problem of relationship between rate constants of reactions involving reactants of same type but different sizes. We leave this interesting problem for further investigation.

Acknowledgments. The work of Eugen Czeizler, Andrzej Mizera and Ion Petre was supported by Academy of Finland, grants 129863, 108421, and 122426. Andrzej Mizera is on leave of absence from the Institute of Fundamental Technological Research, Polish Academy of Sciences, Warsaw, Poland.

References

[1] B. Alberts, D. Bray, K. Hopkin, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Essential Cell Biology*. Garland Science, New York, 2nd edition, 2004.

[2] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*. Garland Science, New York, 4th edition, 2002.

[3] R.-J. Back and J. von Wright. *Refinement Calculus*. Springer, 1998.

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

[5] W. W. Chen, B. Schoeberl, P. J. Jasper, M. Niepel, U. B. Nielsen, D. A. Lauffenburger, and P. K. Sorger. Input–output behavior of ErbB signaling pathways as revealed by a mass action model trained against dynamic data. *Molecular Systems Biology*, 5(239), 2009.

[6] E. Czeizler, A. Mizera, E. Czeizler, R.-J. Back, J. E. Eriksson, and I. Petre. Quantitative analysis of the self-assembly strategies of intermediate filaments from tetrameric vimentin. *Manuscipt*, 2010.

[7] V. Danos, J. Feret, W. Fontana, R. Harmer, and J. Krivine. Rule-based modelling, symmetries, refinements. In J. Fisher, editor, *Formal Methods in Systems Biology. First International Workshop, FMSB 2008, Proceedings*, volume 5054 of *Lecture Notes in Bioinformatics*, pages 103–122, Berlin Heidelberg, 2008. Springer-Verlag.

[8] G. de Vries, T. Hillen, M. Lewis, J. Müller, and B. Schönfisch. *A Course in Mathematical Biology: Quantitative Modelling with Mathematical and Computational Methods*. Monographs on Mathematical Modeling and Computation. SIAM, 2006.

[9] D. T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22:403–434, 1976.

[10] D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry*, 81(25):2340–2361, 1977.

[11] R. C. Henrikson, G. I. Kaye, and J. E. Mazurkiewicz. *NMS Histology*. National Medical Series for Independent Study. Lippincott Williams & Wilkins, 1997.

[12] H. Herrmann, M. Häner, M. Brettel, N.-O. Ku, and U. Aebi. Characterization of distinct early assembly units of different intermediate filament proteins. *Journal of Molecular Biology*, 286(5):1403–1420, 1999.

[13] H. Herrmann, M. Häner, M. Brettel, S. A. Müller, K. N. Goldie, B. Fedtke, A. Lustig, W. W. Franke, and U. Aebi. Structure and assembly properties of the intermediate filament protein vimentin: the role of its head, rod and tail domains. *Journal of Molecular Biology*, 264(5):933–953, 1996.

[14] K. E. Iverson. *A Programming Language*. Wiley, New York, 4th edition, 1962.

[15] R. Kirmse, S. Portet, N. Mücke, U. Aebi, H. Herrmann, and J. Langowski. A quantitative kinetic model for the in vitro assembly of intermediate filaments from tetrameric vimentin. *Journal of Biological Chemistry*, 282(52):18563–18572, 2007.

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

[17] E. Klipp, R. Herwig, A. Kowald, C. Wierling, and H. Lehrach. *Systems Biology in Practice*. Wiley-VCH, 2006.

[18] D. Knuth. Two notes on notation. *American Mathematical Monthly*, 99(5):403–422, 1992.

[19] A. D. Lander. The edges of understanding. *BMC Biology*, 8:40, 2010.

[20] L. Lok and R. Brent. Automatic generation of cellular reaction networks with moleculizer 1.0. *Nat. Biotechnol.*, 23:131–136, 2005.

[21] E. Murphy, V. Danos, J. Feret, J. Krivine, and R. Harmer. Rule-based modeling and model refinement. In H. M. Lodhi and S. H. Muggleton, editors, *Elements of Computational Systems Biology*. John Wiley & Sons, Inc., Hoboken, NJ, USA, 2010.

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

[23] W. L. Scherlis and D. S. Scott. First steps towards inferential programming. In R. E. A. Mason, editor, *Information Processing 83: Proceedings of the IFIP 9th World Computer Congress*, 1983.

[24] N. Wirth. Program development by stepwise refinement. *Communications of the ACM*, 14(4):221–227, 1971.

Paper IX

Modelling of ultrasound therapeutic heating and numerical study of the dynamics of the induced heat shock response

Andrzej Mizera and Barbara Gambin

Originally published in *Communications in Nonlinear Science and Numerical Simulation*, 16(5):2342–2349, 2011.

Commun Nonlinear Sci Numer Simulat 16 (2011) 2342–2349

Contents lists available at ScienceDirect

Commun Nonlinear Sci Numer Simulat

journal homepage: www.elsevier.com/locate/cnsns

Modelling of ultrasound therapeutic heating and numerical study of the dynamics of the induced heat shock response

Andrzej Mizera [a,b,*], Barbara Gambin [a]

[a] Institute of Fundamental Technological Research, Polish Academy of Sciences, Pawińskiego 5B, 02-106 Warsaw, Poland
[b] Department of Information Technologies, Åbo Akademi University & Turku Centre for Computer Science, Joukahaisenkatu 3-5A, FIN-20520 Turku, Finland

ARTICLE INFO

Article history:
Available online 15 June 2010

Keywords:
Hyperthermia
Heat shock response dynamics
Ultrasound therapeutic treatment
Mathematical modelling

ABSTRACT

In this presentation we consider hyperthermia, a procedure of raising the temperature above 43 °C, as a treatment modality. To this purpose, a numerical model of *in vivo* soft tissue ultrasound heating is proposed by extending a previously presented *in vitro* model. Based on the numerical simulations, a heating scheme satisfying some constraints related to potential clinical applications is established, and the resulting temperature time-course profile is composed with the temperature-dependent protein denaturation formula of a recently published mathematical model for the eukaryotic heat shock response. The obtained simulation results of the combined models are discussed in view of potential application of ultrasound soft tissue heating in clinical treatment.

1. Introduction

The heat shock response (HSR) is a highly evolutionarily conserved defence mechanism allowing the cell to promptly react to elevated temperature and other forms of environmental, chemical or physical stress. Exposure to shock conditions leads to misfolding of proteins, which in turn accumulate and form aggregates with disastrous effect for the cell. However, damage to cells can initiate one of two opposite responses: either apoptosis, the process of programmed cell death which prevents inflammation in multicellular organisms, or heat shock response which enables recovery and survival of the cell. Thus, these two pathways and the interplay between them have the decisive influence on the biological consequences of the stress. At least two main reasons why the heat shock response has been subject to intense research recently (see [3,19,22]) should be mentioned. First, as a well-conserved mechanism, it is considered a promising candidate for deciphering the engineering principles being fundamental for any regulatory network. Second, regardless of their regulatory functions in HSR, heat shock proteins have fundamental importance to many key biological processes. Therefore, profound understanding of the HSR mechanism is hoped to have far-reaching consequences for the cell biology and to contribute to the development of new treatment methods for a number of diseases, e.g. neurodegenerative and cardiovascular disorders, cancer, ageing, see [1,12,13,15,25].

The key part of the heat shock response is an abrupt upregulation of the heat shock proteins which prevent the accumulation and aggregation of misfolded proteins. Two groups of heat shock proteins can be distinguished. Some heat shock proteins are constitutively and ubiquitously expressed in all eukaryotic cells. These proteins are called *heat-shock cognates* and are involved in house-keeping roles, e.g. assist nascent proteins in the establishment of proper conformation, transport (shuttle) other proteins between different compartments inside the cell and participate in signal transduction. The second

* Corresponding author at: Department of Information Technologies, Åbo Akademi University, Joukahaisenkatu 3-5A, FIN-20520 Turku, Finland. Tel.: +358 2 2154045.
E-mail addresses: amizera@abo.fi, amizera@ippt.gov.pl (A. Mizera), bgambin@ippt.gov.pl (B. Gambin).

1007-5704/$ - see front matter © 2010 Published by Elsevier B.V.
doi:10.1016/j.cnsns.2010.04.056

group contains those which expression is induced by stress. They act as *chaperones*, i.e. help proteins to maintain their structural integrity or assist the damaged proteins in re-establishment of the functional structure. Moreover, some of them can either act as negative regulators of the apoptotic cascade [2] or aid the apoptotic machinery through their chaperone functions, see [20] for the review of this issue. These two functions fulfilled by the heat shock proteins, i.e. protein chaperoning and modulation of survival and death-signaling pathways, make them an attractive therapeutic target, for example in the case of neurodegenerative diseases [8,15] or cancer [12,13,25]. Furthermore, the heat-induced expression of heat shock protein genes is itself a mechanism of particular interest as it enables the design of heat-responsive gene therapy vectors, cf. [23].

In this study we consider hyperthermia, procedure of raising the temperature above 37 °C, as a treatment modality both on the tissue and cellular levels. Theoretically, a properly tuned tempo-spatial temperature distribution in a tissue would lead to a desired heat shock response in the tissue forming cells and, in consequence, enhanced expression of heat shock proteins which are important from the therapeutic point of view. One of the most relevant problems which arise in this context is related to the question whether in the considered type of tissue a controlled and effective application of hyperthermia is practically feasible. The application has to be strictly controlled since it is important to assure that the temperature itself is kept within the therapeutic range, i.e. up to 43 °C. Furthermore, the tissue area and exposure time to heating must be precisely defined in order to activate the finely tuned heat shock response, on which the effectiveness of the treatment depends. The utilization of ultrasonic technique for hyperthermia seems a very promising approach capable of meeting such requirements, cf. [7,9,21]. Ultrasound irradiation does not stimulate ion activity within the cells, which is an undesired side effect of other irradiation techniques, and is non-invasive, i.e. does not require surgical intervention. Technical improvements of the focused ultrasound ensure the non-invasive and strictly controlled heating of the target tissue volumes. As mentioned before, the control over the spatial temperature distribution in a tissue is of essential importance for the appropriate induction of gene expression on the cellular level. By adjusting the ultrasound beam's intensity, frequency, pulse duration, duty-cycle and exposure time, the proper ultrasonic regime can be tuned. It is now crucial that the research is extended towards the establishment of safe protocols for inducing heat shock response by ultrasound irradiation, which could be applied in clinical treatment.

In [4], a very simple Finite Element Method (FEM) model of soft tissue ultrasound heating was introduced. Based on it, a heating scheme satisfying the requirement that the temperature induced by the ultrasound transducer in the focal area does not exceed 43 °C was proposed in [14]. Further, the influence of the tissue heating scheme on the heat shock response measured by the levels of induced free heat shock proteins and misfolded proteins in the cells was discussed. The construction of the soft tissue heating model in [4] was based on an *in vitro* experiment performed in order to investigate the possibilities of inducing temperature fields in soft tissues by the use of focused ultrasound. Hence, the heating process only with respect to the material properties was considered and neither perfusion nor metabolic heat generation were incorporated into the numerical model. For a more detailed discussion on the experimental setup and the soft tissue heating model we refer the reader to [4,14].

In this presentation, we extend the numerical tissue heating model from [4] by additionally taking into account both perfusion and metabolic heat generation (Section 2). The extended model is utilized to establish an ultrasound heating scheme that meets the requirement of not exceeding the temperature of 43 °C at the transducer's focal point. Next, in Section 3, the resulting temperature time-course profile is combined with the heat-induced protein denaturation formula of the basic HSR mathematical model presented in [18]. Further, based on the numerical simulations of the combined models, the dynamics of the response is compared with the outcomes of the model in [14] and the obtained results are discussed in view of potential application of ultrasound induced soft tissue heating for therapeutic purposes. Finally, in Section 4, we end with some conclusions and suggestions for further work.

2. Numerical model of the soft tissue ultrasound heating

A very simple numerical model of tissue ultrasound heating was presented in [4,14] and used to compute tempo-spatial temperature fields generated in soft tissues by ultrasound treatment. The model was constructed in accordance with an *in vitro* experiment discussed in [4]. The schematic illustration of this experiment is given in Fig. 1. In this presentation we extend the model by considering not only the heating process with respect to material properties, but also by taking into account perfusion and metabolic heat generation in a soft tissue. These modifications make the extended model to reflect the *in vivo* conditions rather than *in vitro*, which was the case of the original model described in [4,14].

As stated in [4], the general bioheat transfer equation in an inhomogeneous thermally anisotropic medium, occupying domain V in the 3D real space, may be written as:

$$\rho(\mathbf{x})C(\mathbf{x})\frac{\partial T(\mathbf{x},t)}{\partial t} = \nabla\cdot K(\mathbf{x})\cdot\nabla T(\mathbf{x},t) + Q_p(\mathbf{x},t) + Q_{int}(\mathbf{x},t) + Q_{ext}(\mathbf{x},t) \quad \text{for} \quad \mathbf{x}\in V, \tag{1}$$

where T, t, ∇, ρ, C, K, Q_p, Q_{int}, Q_{ext} denote temperature, time variable, gradient vector, density, specific heat, thermal conductivity of a medium (second order tensor in our case), heat sources due to perfusion, internal heat generation and external heating (e.g. by irradiation processes), respectively (see [16]). The bioheat equations are present in the literature in many different forms, see, e.g. [24].

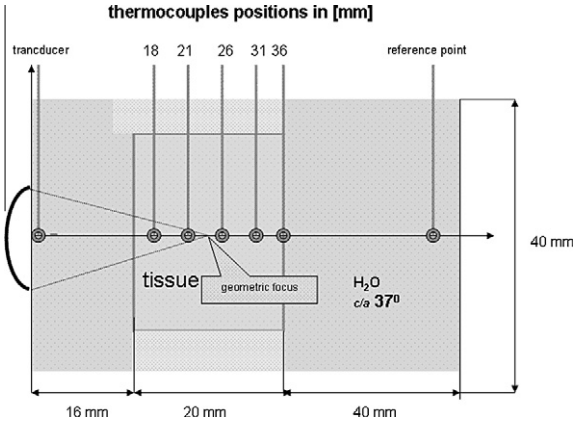


Fig. 1. Schematic illustration of the experiment presented in [4]. Seven thermocouples were used to measure the temperature induced by ultrasound irradiation in various field points along the acoustic axis. The positions are shown in relation to the transducer. In this presentation the temperature in the neighbourhood of the transducer's focal point is considered for establishing the therapeutic heating scheme presented in Fig. 3.

We state the initial boundary-value problem of the Pennes' bioheat equation (Eq. (1)) as follows. The medium under consideration consists of two kinds of material occupying domain $V = V_w \cup V_t$, where V_w and V_t are the volumes occupied by water and tissue, respectively (Fig. 2(a)). The coefficients in Eq. (1) depend on $\mathbf{x}$ in the following way:

$$\rho(\mathbf{x}) = \begin{cases} \rho_w & \text{for } \mathbf{x} \in V_w \\ \rho_t & \text{for } \mathbf{x} \in V_t \end{cases}, \quad C(\mathbf{x}) = \begin{cases} C_w & \text{for } \mathbf{x} \in V_w \\ C_t & \text{for } \mathbf{x} \in V_t \end{cases}, \qquad K = \begin{cases} K_w & \text{for } \mathbf{x} \in V_w \\ K_t & \text{for } \mathbf{x} \in V_t \end{cases}, \quad K(\mathbf{x}) = K\mathbf{I}, \quad \text{for } \mathbf{x} \in V, \tag{2}$$

where $\mathbf{I}$ denotes the unit second order tensor. The temperature on the boundary ∂V of the domain V is assumed to be constant, namely

$$T(\mathbf{x}, t) = 37\ ^\circ\mathrm{C}, \quad \mathbf{x} \in \partial V. \tag{3}$$

Perfusion and metabolic heat generation have a significant influence on the heating process of a soft tissue *in vivo*. Taking into account these two elements is the main difference between the model presented in [4,14], which reflects the *in vitro* conditions, and the one discussed in this presentation. We assume in our numerical computations that the perfusion is given by

$$Q_p(\mathbf{x}, t) = w_b C_b (T_0 - T), \tag{4}$$

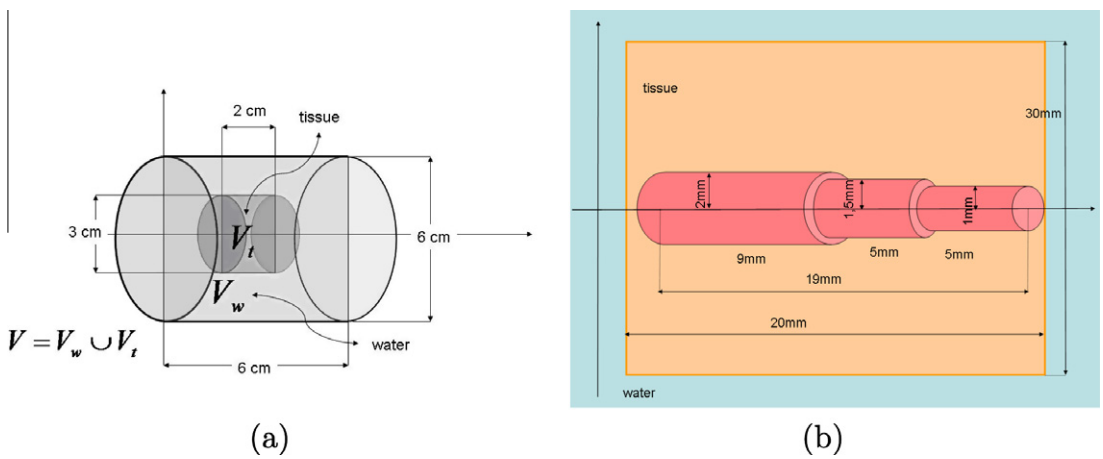


Fig. 2. (a) Two domains occupied by water and tissue considered in numerical computations. (b) The heat sources geometry assumed in numerical calculations (adopted from [4]). The total power of the heat sources is 0.16 W. The power is assumed to be uniformly distributed over the volume occupied by the heat sources ($\approx 10^6$ W/m^3).

where w_b is the blood perfusion rate per unit volume of a tissue and C_b is the specific heat capacity of blood (cf. [26,27]). Q_{int}, the metabolic heat generation per unit volume is assumed to be constant, i.e.

$$Q_{int}(\mathbf{x}, t) = q_m. \tag{5}$$

Finally, the external heat Q_{ext} is modelled by heat sources of the total power 0.16 W. The heat sources are assumed to be produced by the focused acoustic beam and their arrangement inside the tissue, depicted in Fig. 2(b), is adopted from [4], where it was optimized to fit the experimental results. The total power is assumed to be uniformly distributed over the volume occupied by the heat sources, which results in the power density of $\approx 10^6$ W/m^3. The numerical values of the constants appearing in the model are presented in Table 1.

Eqs. (1)–(5) together with the heat sources geometrical distribution provide a well defined boundary-value problem. The solution to this problem was obtained numerically by utilizing standard Finite Element Method approach. The simulations were performed with use of the Abaqus 6.9 software (DS Simulia Corp.) and the temperature time-course profiles in the neighbourhood of the ultrasound transducer physical focus point (the place of maximal temperature) were considered. Based on these results, a heating scheme satisfying the previously discussed requirement was obtained. First, the heat sources were turned on at the initial temperature of 37 °C ($t = 0$ s). The heating was turned off when the temperature at the considered point reached 43 °C ($t = 130$ s) and the tissue was left to cool to 38 °C. Subsequently, the cooling process was interrupted by turning on the heating again ($t = 201$ s). The last two phases, i.e. cooling and heating, where repeated periodically in order to obtain a temperature time-course profile for 4 h. The initial heating phase followed by one periodic phase is depicted in Fig. 3.

It is worth noticing that, although the experiment in [4] was performed *in vitro*, its schematic illustration (Fig. 1) remains valid in the *in vivo* case. For example, if the tissue that undergoes the treatment is part of an organ in the abdomen, the surrounding water in Fig. 1 can represent the peritoneal fluid, which covers the organ.

3. The dynamics of the ultrasound induced heat shock response

In order to investigate how the obtained temperature time-course profile influences the heat shock response on the cellular level, the basic mathematical model of the heat shock response in eukaryotic cells, recently presented in [18], was exploited. The biochemical model consists of three main modules: the dynamic transactivation of the HSP-encoding genes, their backregulation and the chaperone activity of the heat shock proteins. At elevated temperatures proteins tend to misfold and create aggregates. This has disastrous effects on the cell. Hence, in order to survive, the cell under stress has to promptly increase the level of heat shock proteins (HSP) which act as chaperones by interacting with the misfolded proteins (MFP) and helping them to regain the native conformation (PROT). The control over this defence mechanism is exercised through the regulation of the transactivation of the HSP-encoding gene. In order to transactivate transcription, heat shock factors (HSF) trimerize (by transitory forming dimers (HSF_2)) and in this form (HSF_3) bind to the heat shock element (HSE), i.e. the promotor element of the HSP-encoding gene. Once bound (HSF_3:HSE), the gene is transactivated and new heat shock proteins are synthesized. When the amount of chaperones is big enough to cope with the stress, the mechanism is turned off by free HSPs which bind to free HSFs and HSFs that are in complex forms (HSF_2, HSF_3, HSF_3:HSE) by previously breaking the complexes. In consequence, the production of new HSPs is switched off and no new HSF_3s can be formed. The full list of biochemical reactions is given in Table 2. The biochemical model takes into account only well-documented reactions and does not include any "artificial" elements such as experimentally unsupported components or reactions.

An associated mathematical model is obtained by assuming the *law of mass-action* [5,6] for the all considered biochemical reactions. The resulting model is in terms of ordinary, first order differential equations, which form the nonlinear dynamical system presented in Table 3. The heat-induced protein denaturation is modelled by adapting the temperature-dependent formula for fractional protein denaturation originally introduced in [17]. It is incorporated into the mathematical model in the form of the rate coefficient of protein misfolding (reaction (R_{14})), which is given by the following expression:

$$\varphi(T) = \left(1 - \frac{0.4}{e^{T-37}}\right) \cdot 1.4^{T-37} \cdot 1.45 \cdot 10^{-5}\ \mathrm{s}^{-1}, \tag{6}$$

Table 1
Numerical values for the constants appearing in the tissue model discussed in Section 2.

Material	Water	Soft tissue
Density [kg/m^3]	$\rho_w = 1000$	$\rho_t = 1060$
Specific heat [J/(kg K)]	$C_w = 4200$	$C_t = 3800$
Conductivity [W/(m K)]	$K_w = 0.6$	$K_t = 0.5$
Parameter		Value
Blood perfusion [kg/(m^3 s)]		$w_b = 0.9$
Blood specific heat [J/(kg K)]		$C_b = 3800$
Metabolic heat generation [W/m^3]		$q_m = 1085$

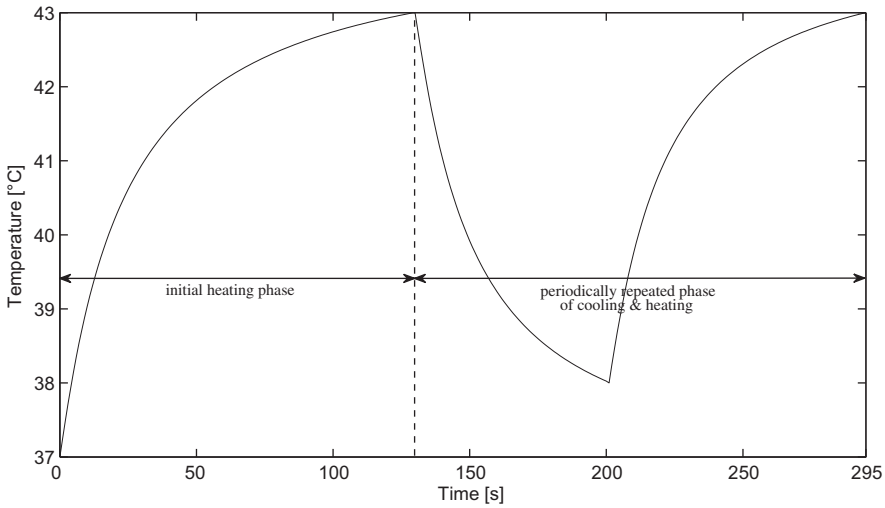


Fig. 3. The initial heating phase (0–130 s) followed by cooling and heating phase (130–295 s). The last phase has been repeated periodically in order to obtain a heating scheme of 4 h.

Table 2
The simplified model for the eukaryotic heat shock response originally discussed in [18].

$2HSF \rightarrow HSF_2$	(R_1)
$HSF_2 \rightarrow 2HSF$	(R_2)
$HSF + HSF_2 \rightarrow HSF_3$	(R_3)
$HSF_3 \rightarrow HSF + HSF_2$	(R_4)
$HSF_3 + HSE \rightarrow HSF_3 : HSE$	(R_5)
$HSF_3 : HSE \rightarrow HSF_3 + HSE$	(R_6)
$HSF_3 : HSE \rightarrow HSF_3 : HSE + HSP$	(R_7)
$HSP + HSF \rightarrow HSP : HSF$	(R_8)
$HSP : HSF \rightarrow HSP + HSF$	(R_9)
$HSP + HSF_2 \rightarrow HSP : HSF + HSF$	(R_{10})
$HSP + HSF_3 \rightarrow HSP : HSF + 2HSF$	(R_{11})
$HSP + HSF_3 : HSE \rightarrow HSP : HSF + HSE + 2HSF$	(R_{12})
$HSP \rightarrow$	(R_{13})
$PROT \rightarrow MFP$	(R_{14})
$HSP + MFP \rightarrow HSP : MFP$	(R_{15})
$HSP : MFP \rightarrow HSP + MFP$	(R_{16})
$HSP : MFP \rightarrow HSP + PROT$	(R_{17})

where T is the numerical value of the temperature of the environment in Celsius degrees. It is valid for $37 \leqslant T \leqslant 45$ and is based on experimental investigations presented in [11,10]. For a detailed description of the model we refer the reader to [18].

Instead of setting the temperature to a constant value as in [18], we composed the time-dependent temperature profile obtained from the numerical tissue model from Section 2 with the protein denaturation formula (Eq. (6)). In this way, the basic model from [18] was adapted for simulation of the cellular defence against ultrasound induced heating. The simulation results in the form of the number concentration variations in time of the heat shock proteins and misfolded proteins are depicted in Figs. 4 and 5, respectively.

The obtained results for the new *in vivo* model coincide with the outcomes of the model presented in [14]. The ultrasound induced free HSP level (Fig. 4) is significantly higher than the HSP level under the physiological conditions (37 °C, black dashed line in Fig. 4), which is desired from the therapeutic point of view. Moreover, in the new model the average free HSP level, computed alternatingly as the mean of two consecutive top and bottom or bottom and top peak values (red

Table 3
The simplified mathematical model of the heat shock response originally presented in [18]. The model is obtained from the biochemical model shown in Table 2 by assuming the law of mass-action. It is formulated in terms of a system of 10 ordinary, first order, nonlinear differential equations. The numerical values of the rate constants, the relationship between the model variables and the metabolites, and initial values of the variables are presented is Table 4.

$$dX_1/dt = -2k_1^+X_1^2 + 2k_1^-X_2 - k_2^+X_1X_2 + k_2^-X_3 - k_5^+X_1X_6 + k_5^-X_7 + k_6X_2X_6 + 2k_7X_3X_6 + 2k_8X_5X_6 \quad (7)$$

$$dX_2/dt = k_1^+X_1^2 - k_1^-X_2 - k_2^+X_1X_2 + k_2^-X_3 - k_6X_2X_6 \quad (8)$$

$$dX_3/dt = k_2^+X_1X_2 - k_2^-X_3 - k_3^+X_3X_4 + k_3^-X_5 - k_7X_3X_6 \quad (9)$$

$$dX_4/dt = -k_3^+X_3X_4 + k_3^-X_5 + k_8X_5X_6 \quad (10)$$

$$dX_5/dt = k_3^+X_3X_4 - k_3^-X_5 - k_8X_5X_6 \quad (11)$$

$$dX_6/dt = k_4X_5 - k_5^+X_1X_6 + k_5^-X_7 - k_6X_2X_6 - k_7X_3X_6 - k_8X_5X_6 - k_{11}^+X_6X_8 + (k_{11}^- + k_{12})X_9 - k_9X_6 \quad (12)$$

$$dX_7/dt = k_5^+X_1X_6 - k_5^-X_7 + k_6X_2X_6 + k_7X_3X_6 + k_8X_5X_6 \quad (13)$$

$$dX_8/dt = \varphi(T)X_{10} - k_{11}^+X_6X_8 + k_{11}^-X_9 \quad (14)$$

$$dX_9/dt = k_{11}^+X_6X_8 - (k_{11}^- + k_{12})X_9 \quad (15)$$

$$dX_{10}/dt = -\varphi(T)X_{10} + k_{12}X_9 \quad (16)$$

Table 4
The numerical values of the rate constants and the initial values of the variables in the simplified mathematical HSR model presented in [18]. The tissue model from Section 2 was combined with the HSR model by composing the protein denaturation coefficient $\varphi(T)$ with the time-dependent temperature profile obtained from the tissue model (Fig. 3). # denotes the number of molecules, V is the cell volume and s is second.

Param.	Reaction	Value	Unit	Metabolite	Var.	Init. no.
k_1^+	(R_1)	3.49	$\frac{V}{\#\cdot s}$	HSF	X_1	0.669
k_1^-	(R_2)	0.19	s^{-1}	HSF_2	X_2	8.73×10^{-4}
k_2^+	(R_3)	1.07	$\frac{V}{\#\cdot s}$	HSF_3	X_3	1.23×10^{-4}
k_2^-	(R_4)	10^{-9}	s^{-1}	HSE	X_4	29.733
k_3^+	(R_5)	0.17	$\frac{V}{\#\cdot s}$	HSF_3:HSE	X_5	2.956
k_3^-	(R_6)	1.21×10^{-6}	s^{-1}	HSP	X_6	766.875
k_4	(R_7)	8.3×10^{-3}	s^{-1}	HSP:HSF	X_7	1403.13
k_5^+	(R_8)	9.74	$\frac{V}{\#\cdot s}$	MFP	X_8	517.352
k_5^-	(R_9)	3.56	s^{-1}	HSP:MFP	X_9	71.648
k_6	(R_{10})	2.33	$\frac{V}{\#\cdot s}$	PROT	X_{10}	1.15×10^8
k_7	(R_{11})	4.31×10^{-5}	$\frac{V}{\#\cdot s}$			
k_8	(R_{12})	2.73×10^{-7}	$\frac{V}{\#\cdot s}$			
k_9	(R_{13})	3.2×10^{-5}	s^{-1}			
k_{10}	(R_{14})	$\varphi(T)$	s^{-1}			
k_{11}^+	(R_{15})	3.32×10^{-3}	$\frac{V}{\#\cdot s}$			
k_{11}^-	(R_{16})	4.44	s^{-1}			
k_{12}	(R_{17})	13.94	s^{-1}			

dashed line), is higher than the corresponding average of the outcomes of the model in [14], where neither perfusion nor metabolic heat generation was considered (blue dashed line). This shows that *in vivo* ultrasound induced heating may be even more efficient than indicated by *in vitro* experimental results.

However, in therapeutic applications, it is very important to control the level of misfolded proteins and keep it low during the treatment. Otherwise, the heating could cause the cells' death rather than stimulate them to self-repair. Hence, in order to assess a heating protocol in view of therapeutic applicability, it is crucial to examine the induced MFP level. The obtained results (Fig. 5) show that under the discussed heating scheme the level of misfolded proteins evenly oscillates around the reference level obtained under constant 42 °C heating (black dashed line), i.e. except for the initial phase of less than 20 min, the reference line coincides with the average calculated as the mean of two consecutive top and bottom (or *vice versa*) MFP time-course peaks (red dashed line). As in [14], the response to constant 42 °C is chosen as the reference on, since the cells are usually capable of surviving in such conditions. Again, although the difference is not as clear as in the case of the HSP level time-course, the obtained results for the new *in vivo* model are slightly better than in the case of the *in vitro* model in [14]. After about 20 min of treatment, the average for the *in vitro* model (blue dashed line) is above the average of the model with perfusion and metabolic heat generation taken into account. However, as in [14], alarming is the protein misfolding at the beginning of the treatment. The only improvement which can be observed here with respect to the previous model is that the peak value of the whole response in the case of the *in vivo* model is lower (4.5×10^4 instead of 4.7×10^4 misfolded protein molecules).

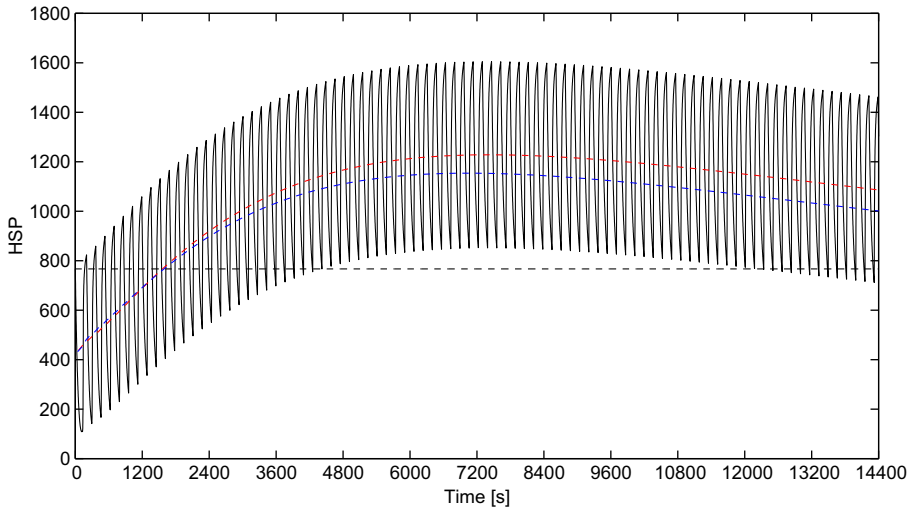


Fig. 4. Number of molecules in time of the free heat shock proteins induced by the ultrasound irradiation. The simulation results were obtained by exploiting the basic mathematical model from [18]. The black dashed line indicates the HSP level at physiological conditions (37 °C). The red dashed line is the average obtained by computing the mean values of two consecutive HSP time-course peak values (top and bottom or bottom and top, alternatively). Each mean value is placed in the middle of the time interval determined by the two peaks from which the mean value was obtained. The blue dashed line indicates the analogous average for the *in vitro* model presented in [14]. (For interpretation of color mentioned in this figure the reader is referred to the web version of the article.)

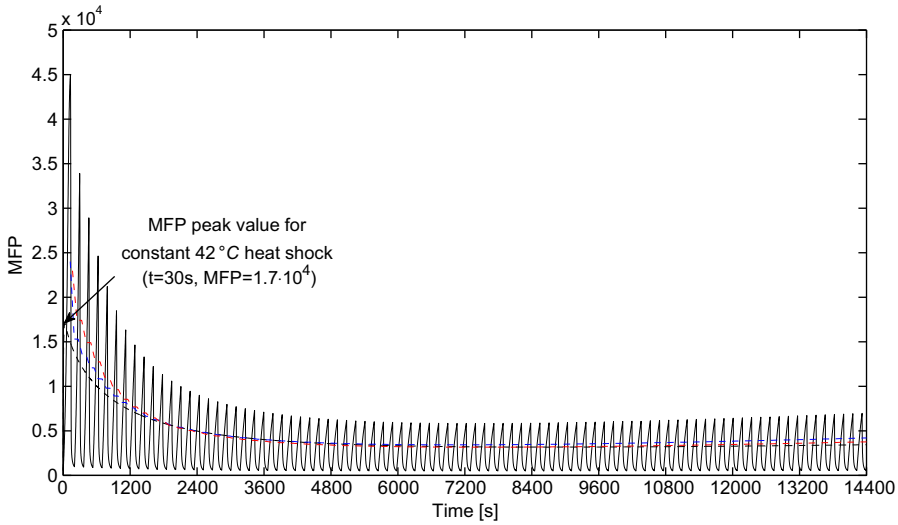


Fig. 5. Number of molecules in time of the misfolded proteins induced by the ultrasound irradiation. The simulation results were obtained by exploiting the basic mathematical model from [18]. The black dashed line indicates the MFP level at constant 42 °C heat shock. The red dashed line is the average obtained by computing the mean values of two consecutive MFP time course peak values (top and bottom or bottom and top, alternatively). Each mean value is placed in the middle of the time interval determined by the two peaks from which the mean value was obtained. The blue dashed line indicates the analogous average for the *in vitro* model presented in [14]. (For interpretation of color mentioned in this figure the reader is referred to the web version of the article.)

4. Conclusions and further research

In this presentation hyperthermia was considered as a treatment method. A soft tissue heating model based on the Pennes' bioheat equation presented in [4,14] was extended by considering two additional elements: perfusion and metabolic heat generation. Further, it was combined with a new mathematical model of the heat shock response in eukaryotic cells recently presented in [18]. The HSR model is formulated in terms of a system of 10 ordinary, first order, nonlinear differential equations. Based on the performed simulations, an ultrasound heating scheme has been proposed.

The obtained heating regime on the tissue level is capable of inducing a rather reasonable, in view of therapeutic application, heat shock response on the cellular level. The assessment of the heating scheme is based on the time-course

behaviour of the induced levels of free heat shock proteins and misfolded proteins. However, alarming with respect to the MFP level are the first 20 min of the response. An improvement could potentially be achieved by exploiting the "self-learning" property of the heat shock response mechanism in the following way. Since numerical simulations of the model in [18] indicate that the response to a consecutive heat shock is significantly weaker, the presented heating procedure could be preceded by some properly adjusted temperature increase. In consequence, the initial MFP level peaks would be reduced. However, such pre-treatment should be finely tuned in order to minimize the negative influence it would have on the induction of free heat shock proteins level increase, which is essential for the effectiveness of the therapy.

Finally, the presented simulation results reveal that the basic mathematical model from [18] might not be robust. This may be concluded based on the fact that the model drastically reacts to temperature changes of a relatively high frequency. The dynamics displayed by the HSR model might be unrealistic with respect to the energy resources it would require. Moreover, robustness is a common and rather crucial feature of all biological systems, which is a contrast with the model, that is supposed to reflect a biological mechanism. This issue asks for a more thorough investigation, potentially accompanied by some experimental verifications which would cast some more light on the problem of robustness of the heat shock response machinery.

Acknowledgement

The work was partially supported by the Polish Ministry of Science and Higher Education, Grant No. N N518 426936.

References

[1] Balch WE, Morimoto RI, Dillin A, Kelly JW. Adapting proteostasis for disease intervention. Science 2008;319:916–9.
[2] Beere HM. 'The stress of dying': the role of heat shock proteins in the regulation of apoptosis. J Cell Sci 2004;117(13):2641–51.
[3] Chen Y, Voegli T, Liu P, Noble E, Currie R. Heat shock paradox and a new role of heat shock proteins and their receptors as anti-inflammation targets. Inflamm Allergy Drug Targets 2007;6(2):91–100.
[4] Gambin B, Kujawska T, Kruglenko E, Mizera A, Nowicki A. Temperature fields induced by low power focused ultrasound in soft tissues during gene therapy. Numerical predictions and experimental results. Arch Acoustics 2009;34(4):445–59.
[5] Guldberg C, Waage P. Studies concerning affinity. C.M. Forhandlinger: Videnskabs-Selskabet i Christiana 1864;35.
[6] Guldberg C, Waage P. Concerning chemical affinity. Erdmann's J für Pract Chem 1879;127:69–114.
[7] Humphrey VF. Ultrasound and matter – physical interactions. Prog Biophys Mol Biol 2007;93:195–211.
[8] Kalmar B, Kieran D, Greensmith L. Molecular chaperones as therapeutic targets in amyotrophic lateral sclerosis. Biochem Soc Trans 2005;33:551–2.
[9] Kujawska T, Wójcik J, Filipczyński L. Possible temperature effects computed for acoustic microscopy used for living cells. Ultrasound Med Biol 2004;30(1):93–101.
[10] Lepock JR, Frey HE, Ritchie KP. Protein denaturation in intact hepatocytes and isolated cellular organelles during heat shock. J Cell Biol 1993;122(6):1267–76.
[11] Lepock JR, Frey HE, Rodahl AM, Kruuv J. Thermal analysis of chl v79 cells using differential scanning calorimetry: implications for hyperthermic cell killing and the heat shock response. J Cell Physiol 1988;137(1):14–24.
[12] Liu B, DeFilippo AM, Li Z. Overcoming immune tolerance to cancer by heat shock protein vaccines. Mol Cancer Ther 2002;1:1147–51.
[13] Lukacs KV, Pardo OE, Colston M, Geddes DM, Alton EW. Heat shock proteins in cancer therapy. In: Nagy A. Habib, editor. Cancer gene therapy: past achievements and future challenges. Kluwer; 2000. p. 363–8.
[14] Mizera A, Gambin B. The dynamics of heat shock response induced by ultrasound therapeutic treatment. In: 10th conference on dynamical systems – theory and applications, DSTA-2009, Proceedings, vol. 2. Poland: Łódź; 2009. pp. 847–852 .
[15] Morimoto RI. Proteotoxic stress and inducible chaperone networks in neurodegenerative disease and aging. Genes Dev 2008;22:1427–38.
[16] Pennes HH. Analysis of tissue and arterial blood temperatures in the resting human forearm. J Appl Physiol 1948;1(2):93–122.
[17] Peper A, Grimbergent C, Spaan J, Souren J, van Wijk R. A mathematical model of the hsp70 regulation in the cell. Int J Hyperthermia 1997;14:97–124.
[18] Petre I, Mizera A, Back R-J. Computational heuristics for simplifying a biological model. In: Ambos-Spies K, Löwe B, Merkle W. editors. Mathematical theory and computational practice: 5th conference on computability in Europe, CiE 2009, Proceedings. LNCS, vol. 5635. Germany, Springer: Heidelberg; 2009. pp. 399–408.
[19] Powers M, Workman P. Inhibitors of the heat shock response: biology and pharmacology. FEBS Lett 2007;581(19):3758–69.
[20] Takayama S, Reed JC, Homma S. Heat-shock proteins as regulators of apoptosis. Oncogene 2003;22:9041–7.
[21] ter Haar G. The resurgence of therapeutic ultrasound – a 21st century phenomenon. Ultrasonics 2008;48(4):233.
[22] Voellmy R, Boellmann F. Chaperone regulation of the heat shock protein response. Adv Exp Med Biol 2007;594:89–99.
[23] Walther W, Stein U. Heat-responsive gene expression for gene therapy. Adv Drug Delivery Rev 2009;61:641–9.
[24] Weinbaum S, Jiji LM. A new simplified bioheat equation for the effect of blood flow on local average tissue temperature. J Biomech Eng 1985;107(2):131–9.
[25] Workman P, de Billy E. Putting the heat on cancer. Nat Med 2007;13(12):1415–7.
[26] Yuan P. Numerical analysis of an equivalent heat transfer coefficient in a porous model for simulating a biological tissue in a hyperthermia therapy. Int J Heat Mass Transfer 2009;52:1734–40.
[27] Yue K, Zhang X, Yu F. An analytic solution of one-dimensional steady-state Pennes bioheat transfer equation in cylindrical coordinates. J Therm Sci 2004;13(3):255–8.

Turku Centre for Computer Science
TUCS Dissertations

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 Säntt**i, 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 Methods 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 Communication 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.

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

Turku School of Economics
- Institute of Information Systems Science

Åbo Akademi University
Division for Natural Sciences and Technology
- Department of Information Technologies

ISBN 978-952-12-2616-8
ISSN 1239-1883

Andrzej Mizera

Methods for Construction and Analysis of Computational Models in Systems Biology