From Cellular Components to Living Cells (and Back) – Evolution of Function in Biological Networks

seit 1558

Dissertation

zur Erlangung des akademischen Grades
doctor rerum naturalium (Dr. rer. nat)

vorgelegt dem Rat der Fakultät für Mathematik und Informatik
der Friedrich-Schiller-Universität Jena

von Thorsten Lenser
geboren am 11. Januar 1982 in Oberhausen

Gutachter

1. PD Dr. Peter Dittrich

2. Prof. Dr. Stefan Schuster

3. PD Dr. Peter Hemmerich

Tag der öffentlichen Verteidigung: 18. Februar 2011

Abstract

Network models pervade modern biology. From ecosystems down to molecular interactions in cells, they provide abstraction and explanation for biological processes. Thus, the relation between structure and function of networks is central to any comprehensive attempt for a theoretical understanding of life.

Just as any living system, biological networks are shaped by evolutionary processes. In reverse, artificial evolution can be employed to reconstruct networks and to study their evolution. To this end, I have implemented an evolutionary algorithm specifically designed for the evolution of network models. It couples genetic programming of the network topology with an evolution strategy applied to the kinetic parameters. By this separation of structural from parameter evolution, the kinetic parameters can adapt to new topologies before these are evaluated. Using exemplary fitness functions, it is shown that this approach smoothens the path in topology search space, prevents premature convergence, and has superior performance compared to alternative procedures.

With the developed evolutionary framework, a study of the evolution of information-processing networks was performed. Applying the concept of chemical organisation theory, it is shown that selection favours an organisational structure that is related to function, such that computations can be visualised as transitions between organisations. Mutations of the network topology subsequently change the organisational structure towards a lattice representing the desired computations. This approach presents a novel and useful analysis of the evolution of computation in networks.

Fluorescence imaging techniques such as fluorescence recovery after photobleaching (FRAP) and fluorescence correlation spectroscopy (FCS) provide quantitative and dynamic data about protein kinetics in the cell. In this thesis, mathematical modelling is applied to extract knowledge about reaction-kinetic constants from this data, and the model is presented and discussed in detail. Using this approach, a detailed quantitative model of exchange dynamics at PML nuclear bodies (NBs) is created, showing that PML NB components exhibit highly individual exchange kinetics. For the first time, PML NB assembly dynamics in cells lacking endogenous PML are described.

The FRAP data for PML NBs is additionally used as a test-case for automatic model inference using evolutionary methods. While there exists no single best model to fit the data, a set of necessary and sufficient criteria for a good model fit is revealed. A detailed investigation of the search space shows large plateaus with identical fitness values, hampering the progress of the evolutionary algorithm.

When analysing evolution, a comparison of different evolutionary stages can often be facilitated by modelling. In the last part of this thesis, a stochastic analysis of the genetic regulatory system of *DEF*-like and *GLO*-like class B floral homeotic genes provides an explanation for their intricate regulatory wiring. The different potential regulatory architectures are investigated using Monte Carlo simulation, a simplified master-equation model, and fixed-point analysis. It is shown that a positive autoregulatory loop via obligate heterodimerisation of transcription factor proteins reduces noise in cell-fate organ identity decisions. Furthermore, the comparison of potential regulatory systems establishes an “evolutionary motif”, which might have more general applications.

Zusammenfassung

Netzwerkmodelle sind weit verbreitet in der modernen Biologie. In allen Teilgebieten - von der Ökologie bis hin zur Molekularbiologie - bieten sie die Möglichkeit, untersuchte Prozesse und Phänomene zu abstrahieren und damit auf theoretischer Ebene zugänglich zu machen.

Wie alle lebenden Systeme resultieren auch biologische Netzwerke aus evolutionären Abläufen. Im Umkehrschluss kann man künstliche Evolution verwenden, um solche Netzwerke zu rekonstruieren und ihre Entstehung zu untersuchen. Vor diesem Hintergrund wird hier ein evolutionärer Algorithmus vorgestellt, der speziell für die Erzeugung von Netzwerkmodellen angepasst ist. Dafür wurde eine Genetische Programmierung der Netzwerkstruktur mit einer Evolutionsstrategie auf den kinetischen Parametern verknüpft. Durch diese Trennung von Struktur- und Parameterevolution können sich die Parameter einer neu erzeugten Netzwerkstruktur erst anpassen, bevor diese bewertet wird. Dieser Ansatz glättet den evolutionären Pfad durch den Suchraum und verhindert vorzeitige Konvergenz, wie durch einen Vergleich mit alternativen Methoden gezeigt wird.

Mit dem neu entwickelten Evolutionären Algorithmus wurde dann eine Studie zur Evolution von informationsverarbeitenden Netzwerken durchgeführt. Mithilfe der Chemischen Organisationstheorie wird gezeigt, dass die Selektion eine funktionale Organisationsstruktur erzeugt, in welcher eine Berechnung als Transition zwischen Organisationen abgebildet werden kann. Mutationen der Netzwerkstruktur erzeugen Änderungen in der Organisationsstruktur des Netzes, so dass die evolvierende Berechnungskapazität visuell dargestellt werden kann. Dieser Ansatz bietet eine neue und nützliche Analyse der Evolution von signalverarbeitenden Netzwerken.

Bildgebende Verfahren mittels fluoreszierender Proteine - z.B. FRAP (fluorescence recovery after photobleaching) und FCS (fluorescence correlation spectroscopy) - erzeugen quantitative und dynamische Daten über Proteinkinetiken in der Zelle. In dieser Arbeit wird gezeigt, wie man mathematische Modellierungsmethoden verwenden kann, um kinetische Reaktionskonstanten aus diesen Daten zu gewinnen. Die verwendete Methode wird im Detail vorgestellt und diskutiert. Auf diese Weise entstand ein detailliertes Modell des Proteinaustauschs an PML nuclear bodies (NBs), in welchem die Komponenten der PML NBs sehr differenzierte Austauschverhalten zeigen. Darüber hinaus wird - zum ersten Mal - die Dynamik von PML NB Komponenten in Zellen ohne endogene PML Proteine beschrieben.

Darüber hinaus werden die durch FRAP gewonnenen Daten genutzt, um die automatische Evolution von Netzwerkmodellen in einer realistischen Fallstudie zu testen. Auf diese Weise werden notwendige und hinreichende Bedingungen aufgezeigt, welche die „passenden“Modelle erfüllen müssen. Eine detailierte Erkundung des Suchraums ergibt Plateaus mit gleichen Fitnesswerten, welche den Evolutionären Algorithmus stark beeinflussen und behindern.

Modellierung kann oftmals helfen, verschiedene Stufen in der Evolution eines Systems zu vergleichen. In dieser Arbeit wird eine stochastische Analyse des Zusammenspiels der *DEF*- und *GLO*-Gene in der Blütenentwicklung gezeigt, welche eine Erklärung für ihre überraschend komplexe Verschaltung liefert. Die verschiedenen möglichen Regulationsmechanismen werden mithilfe von Monte-Carlo-Simulation, einem Master-Equation-Ansatz und der Fixpunktanalyse verglichen. Es wird gezeigt, dass positive Autoregulation durch obligatorische Heterodimerisierung den Einfluss des Zufalls auf die Organidentität reduziert. Dieser Vergleich regulatorischer Systeme enthüllt ein „evolutionäres Motiv“, das auch über die Blütenentwicklung hinaus Anwendung finden könnte.

Acknowledgements

My first and very deep thanks go to Peter Dittrich, who contributed countless ideas and generous guidance for this work and my career. Furthermore, Thomas Hinze, Naoki Matsumaru, Pietro Speroni di Fenizio, Florian Centler, and numerous other members of our research group contributed with their ideas and friendliness. The group provided extraordinary support, and I am especially grateful that they covered for me whenever I did not have the time to participate adequately in group duties.

Large parts of this thesis would have been impossible without the support of my biological collaborators: Peter Hemmerich, Günter Theißen, and their respective groups. Their willingness to undertake the (often difficult) interdisciplinary adventure with me is highly appreciated, and I have profited immensely from them.

Special thanks go to the reviewers, who have taken onto themselves the hard task of reading and actually rating everything I wrote for this thesis.

Additional thanks are extended to the departmental staff - especially Conny Müsse for invaluable help - and colleagues in the groups of Stefan Schuster and Sebastian Böcker. On the international level, the ESIGNET project brought me in contact with a great bunch of people in Birmingham, Dublin, and Eindhoven. Special thanks go to Anthony Liekens, Chrisantha Fernando, George Mitchell and James Decraene for their inspiration and the warm welcome they provided in their home-countries. Going back to my graduation days, Jon Pitchford, Richard Law, and Andrew Constable deserve my thanks for waking my interest in the interdisciplinary work of mathematical biology and bioinformatics.

Besides great colleagues and a superb working environment, money is also needed to do research. Therefore, I thank the EU (ESIGNET, project no.

12789), the BMBF (TassDB project), and the DFG for funding. This thesis - and much of the work in it - was developed with free software. I want to thank the developers for their time and effort, especially in the GNU/Linux, Emacs, and LaTeX projects.

Last but not least, friends and family have helped a long way in completing this thesis. I especially want to name Anke, Mathias, and Martina, for the many opportunities to forget about my work. Tim, Dirk, Eva, Pia, and Malte provided a much welcomed link to my old hometown. Additionally, my volleyball team provided a sportive challenge whenever I needed one.

All of this would not have been possible without the support of my family. Special thanks go to my parents, parents-in-law, and grandparents, who kept up a constant interest in my work and provided moral (and financial...) support. I am especially proud of the deep friendship with my brother Christian, who always learns first about new turns in my life.

Two special results of my PhD time are infinitely more valuable than anything presented in this thesis: My daughters Malin and Frieda. I am deeply thankful for the opportunity to see and help them grow up, and I look forward to every single adventure with them that is still to come. At the very end, words simply fail me to thank my wife Teresa, without whom I certainly would not be where I am now.

Contents

1
Introduction

Chapter summary. *In modern biology, network models are all over the place, from ecosystem scales down to cellular processes. Due to their non-linear nature, the relation between structure and function of networks is - at the same time - difficult to study and important to understand. This chapter provides a brief overview of structure, dynamics and function of networks, and characterises motifs, modules and global organisation of cell-biological networks. Furthermore, the evolutionary processes shaping biological networks are sketched, and several in silico attempts to model and study the evolution of networks are presented. In a complementary approach, in silico evolution is used to automatically infer network models from data. As an important tool for all this work, the field of Evolutionary Algorithms is presented, with special emphasis on Genetic Programming of network models and CMA Evolution Strategies for parameter fitting.*

Chapter contents

1.1 Biological networks and modelling

The 20th century has brought a paradigm-shift to all biological sciences, from ecology to molecular biology: away from a pure description of the living world, towards the quantification and ultimately prediction of living phenomena. This new paradigm was brought about by the combined ascent of more fine-grained, larger scale experimental techniques and computing technologies that enable a comprehensive understanding of complex data sets. Meanwhile, our understanding of biology has changed: where the focus was traditionally centred on linear cause-effect relationships, it has now become clear that things are much more complicated, that "everything is connected with everything", and that networks are the more appropriate way to describe relations between entities in the biological sciences.

This finding is true on every scale: in ecosystems, networks (food-webs) start to replace the traditional food-chains (Bascompte, 2007; Montoya et al., 2006; Pimm, 2003), while on the molecular level, the computer-guided analysis and integration of data has created a whole new field: systems biology (Alon, 2006; Kitano, 2002; Klipp et al., 2006; Ventura et al., 2006). The fact that networks are at the core of many of the new aspects in biology has lead to an increasing interest by theoretical biologists into their formal treatment and analysis (e.g. Barabási and Oltvai, 2004; Milo et al., 2002; Strogatz, 2001).

The rapidly increasing importance of networks in (molecular) biology provides the background for this thesis: I have looked, from different angles, at the relationship between structure and function of biological networks. Even though the work was focused on applications in molecular biology, most of the results can also be applied to other kinds of networks describing dynamical systems.

1.2 Structure and function in evolution

Colloquial and technical definition of evolution There are at least two different meanings to the term "evolution" that need to be discussed here. In colloquial language, it usually denotes the mechanism that shaped life on earth as we know it, leaving palaeontological traces that help us to decipher it. In a more technical usage, it can denote any process of changing a collection of entities over time, using random changes (mutations) and selection on the mutated entities. In this thesis, I use the second, more abstract definition, which includes the first one as a special case.

Evolution can act on any changeable (evolvable) entity. As the title suggests, this thesis deals with the evolution of networks, specifically networks in systems biology. Special care has to be taken in defining where the processes of mutation and selection apply. In the evolution of life, mutation usually acts on genetic information encoded in DNA, while selection happens at the level of the organism. The focus in this thesis lies on evolution that directly acts on networks. Mutation and selection thus happen on the level of individual networks.

1.2.1 Relation between structure and function in biological networks

Molecular networks can be described at different levels When looking at molecular networks, we find at least three levels that are influenced by evolutionary forces:

- network structure
- network behaviour
- network function

The structural level is the most basic one, composed of the network's nodes (vertices) and their connections (edges). Typically, the network nodes denote molecular species such as genes, proteins, RNA segments, and so on. On the intermediate level, we find the dynamic behaviour of the network, i.e. how the concentrations of its participating proteins vary over time under different conditions. Therefore, the second level emerges from the first one in a mechanistic fashion. The topmost level consists of the function of the network, i.e. the action it performs in the cell. This level is usually created as an interpretation of the network behaviour by an (internal or external) observer. As an example, consider a specific version of the Lac operon (see Prill et al., 2005). The system consists of a positive feedback loop (structural description) that realises a genetic switch (dynamic behaviour), which is used by the cell as a primitive memory mechanism (function).

The need for network models Before going into details on the relation between network structure and function, we need to ask one fundamental question: Why do we need network models in biology? To see this, let us review that in recent years, high-throughput experimental approaches have started to create large amounts of data that cannot be intuitively understood any more. Network models offer a formalised way of integrating this data into relationships between biological entities (Han, 2008). To make this integrated knowledge available to the scientific community, the models themselves can in turn be stored in databases (Le Novère et al., 2006), ideally together with annotations linking them to biological descriptions (Le Novère et al., 2005).

The function of a network A second question, not less fundamental: What actually is the function of a biological network? Network function is generally understood to be the "desired system characteristics" or "performance" (Carlson and Doyle, 2002). For a more specific definition, the answer depends on context, mainly on the type of network one is dealing with. For genetic networks, especially those in developmental processes, it is the steady-state attractors and transitions between them that are of consequence to the cell (Alvarez-Buylla et al., 2008; Huang et al., 2005). In metabolic networks, the flux of metabolites through the system constitutes its function (Heinrich and Schuster, 1996; Schilling et al., 2000; Schuster et al., 1999; Stelling et al., 2002), whereas in signalling networks, the reliable short-time propagation and integration of signals plays this role (Bhalla and Iyengar, 1999; Eungdamrong and Iyengar, 2004; Kholodenko, 2006; Tyson et al., 2003).

More generally speaking, the concept of *biological function* is a key to the philosophy of biology (see e.g. Diaz-Herrera, 2006, and references therein). The general notion seems to be that biological function describes, in the context of natural selection, *why* a certain biological system exists, i.e. it specifies the way the system contributes to the fitness of the organism. This is in contrast to the *mechanism*, which describes *how* the systems functions. Importantly, this definition of function is not a teleological one,

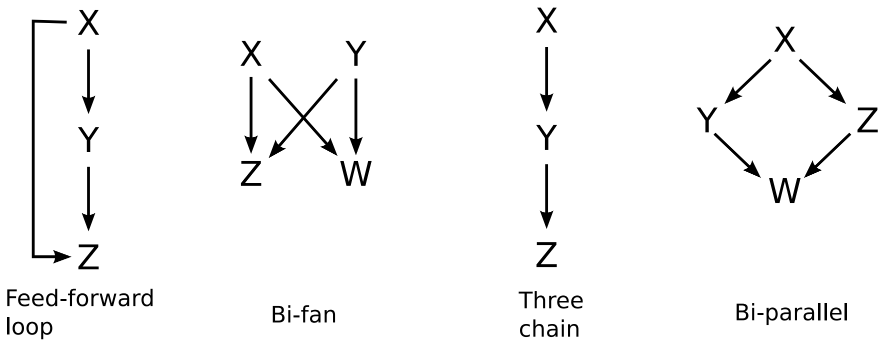


Figure 1.1: Four examples of network motifs discussed in the text. Adapted from Milo et al. (2002).

since it does not invoke the purpose of the system's existence, but rather the fitness advantage it provides to the organism.

The dynamic behaviour and function of network models is a complicated issue in itself. Different temporal and spatial scales, feedback loops, and nonlinear kinetics lead to complex dynamics that can have unforeseen consequences. Thus, there is a clear need to establish closer links between network structure and function.

Structure - function is a many to many mapping The relation between structure and function of biological networks is clearly not a simple mapping. In one direction, a single network structure does not have a uniquely defined function, but can show a range of behaviours (and thus functions) depending on its kinetic rate laws and parameters. For example, Mangan and Alon (2003) show at least two fundamentally different functions for the feed-forward loop, while Ingram et al. (2006) demonstrate five types of computations that can be performed by the bi-fan motif (see Figure 1.1 for the motif definitions). Wetlab experimental evidence for this phenomenon was described by Guet et al. (2002). In the other direction, a defined function does not imply a unique representation as a network structure. For example, Tsong et al. (2006) have shown how different network topologies can lead to the same regulatory behaviour. Additionally, in this thesis (Section 5), I show how of the 64 possible three-node network topologies with only first-order reactions, at least 18 are able to equally well fit a time-series data set from FRAP experiments of PML-bodies.

Network motifs link structure and function One approach of linking network structure and function is by small characteristic subnetworks that are significantly overrepresented in the whole network, compared to randomised networks with the same degree distribution. These subnetworks are called network motifs (Alon, 2006; Milo et al., 2002). Due to the apparent simplicity they introduce into complex networks, motifs have been at the heart of many theoretical (Mangan and Alon, 2003; Milo et al., 2004) and experimental studies (reviewed by Alon, 2007) in recent years.

In a groundbreaking study, Milo et al. (2002) have analysed networks from biochemistry, neurobiology, ecology, and engineering for significantly overrepresented three-node and four-node motifs. They found that only a few motifs characterise each network, and the characterising motifs are similar in networks with similar function (Milo et al., 2004). For example, information processing networks (gene regulatory networks, neural networks, forward logic chips) typically contain the "feed-forward loop" and the "bi-fan" motif (Figure 1.1). In contrast, networks of energy flow are dominated by "chain" and "bi-parallel" motifs. Interestingly, metabolic networks contain feed-forward loops as well as motifs with feedback (Zhu and Qin, 2005), indicating a larger capacity for information processing than typically found in food webs.

Four families of network motifs exist in transcriptional networks Because they are easier to access and better described than other types of networks, many subsequent studies that build on the work of Milo et al. (2002) have focused on transcriptional networks (reviewed by Alon, 2007). They have identified - both in theory and in experiments - four families of network motifs in transcriptional networks. Grouped in the first family are motifs of simple regulation, such as direct regulation of a gene by a transcription factor, positive and negative autoregulation (Figure 1.2(a)-(c)). Negative autoregulation decreases the transcriptional response time, while positive autoregulation leads to larger response times and ultimately to self-sustaining expression of a protein. The second family is composed of variants of the feed-forward loop, which often function as delay elements, persistence detectors, pulse generators or response accelerators (Figure 1.2(d)). "Single-input modules" define the third family, in which one regulating factor acts simultaneously on a group of target genes (Figure 1.2(e)). Finally, the fourth family is composed of "dense overlapping regulons", in which a set of regulators combinatorically controls a set of target genes (Figure 1.2(f)). These complex gate-arrays, in which many inputs regulate many outputs, usually are responsible for a significant biological function in the cell.

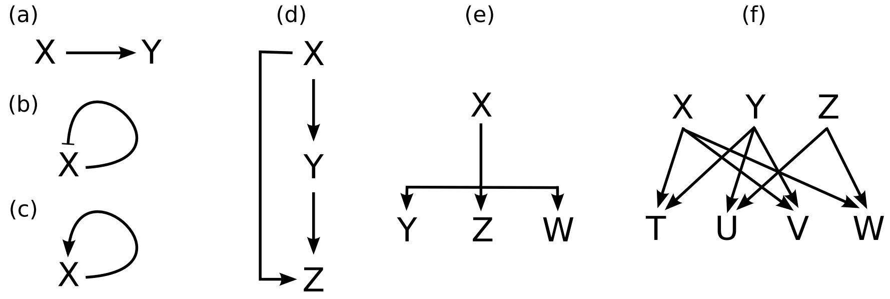


Figure 1.2: The four motif families discussed in the text, adapted form Alon (2007). (a)-(c) simple regulation: (a) direct regulation, (b) negative autoregulation, (c) positive autoregulation. (d) feed-forward loop. (e) single-input motif. (f) dense overlapping regulon. In all figures except for (b) and (c), the arrows indicate either positive or negative regulation.

Motifs often involve feedback, especially in networks that process information (Brandman and Meyer, 2008). Feedback is vital for modulation of cellular signals, for example for decision making (bistable switching by positive feedback) or smoothing of irregular signals (negative feedback). Due to its inherent nonlinearity, it typically complicates the analysis of network behaviour and function. One feedback motif with special relevance to developmental processes is the double-positive feedback loop (Alon, 2007), whose function under the influence of noise is studied in Chapter 7.

Motifs give hints about network function, but the relation is not clear Different environmental conditions require different cellular functions. If a link between network function and motifs exists, it is to be expected that different motifs are favoured by different conditions in the cell. Indeed, Luscombe et al. (2004) found that in yeast, networks that regulate the internal processes of the cell (e.g. the cell cycle) contain more feed-forward loops, which may stabilise these long-term processes by suppressing noise and damping amplifications. In contrast, modules dealing with external signals (e.g. DNA damage, stress response) favour single-input motifs, which can quickly generate a cellular response to the stimulus (Luscombe et al., 2004).

It is, however, important to note that most motifs in molecular networks do not represent a unique function isolated from the rest of the network, but are tightly integrated, as demonstrated in *E. coli* (Dobrin et al., 2004) and *S. cerevisiae* (Mazurie et al., 2005). Therefore, it is not immediately obvious if the function they display in isolation is really their contribution to the network's function. In addition to that, Knabe et al. (2008) show an evolution study on artificial gene regulatory networks in which networks evolved for different functions did not feature significantly different motif distributions. Both points indicate that the link between network function and motifs might not be as tight as suggested by studying the motifs alone, and caution is advised when the overrepresentation of certain motifs is used to infer network function.

Functional robustness Many studies (reviewed by Stelling et al., 2004) have highlighted the pivotal role of local network topology for functional robustness in the cell. Here, "robustness" can be understood in (at least) two ways: either with regard to external perturbations represented by changes to the kinetic parameters of a network, or with regard to noise introduced by the low copy number of key regulatory molecules in the cell. In this thesis, the first aspect comes up in the kinetic analysis of FRAP data (Chapter 6), and the second one is the main topic of our investigation of the B-gene regulatory circuit in flowering plants (Chapter 7).

Directly relating functional robustness to network motifs, Prill et al. (2005) demonstrated that the abundance of motifs in several biological networks is remarkably well correlated with the stability of their function to perturbations in the kinetic parameters. Thus this type of stability may either be a by-product of the evolutionary process or the chemical substrate, or it may be of evolutionary benefit to the organism. Given a more robust function in the presence of noise and an enhanced evolvability (since the parameters do not have to be fine-tuned), it seems likely that functional robustness is indeed an evolutionary advantage, at least to some extend. In analogy, Klemm and Bornholdt (2005) have found that when looking for motifs whose computations are "reliable" in the presence of noise in the inputs, those two-node and three-node motifs that actually appear in biological networks tend to be more reliable (in the sense of staying close to their deterministic behaviour) than those that do not.

A popular example for functional robustness is bacterial chemotaxis: in the network governing tumbling frequency, some key parameters can be changed by orders of magnitude, and the tumbling frequency stays the same (Alon et al., 1999; Barkai and Leibler, 1997), a principle called "robust adaptation" or "homeostasis". Similar robustness has been found in the development of segment polarity in *Drosophila melanogaster* (von Dassow et al., 2000). Moreover, the chemotaxis network in *E. coli* seems to be the smallest network that is sufficiently robust in this sense, thus minimising the cost of maintaining robust adaptation throughout evolution (Kollmann et al., 2005).

Modules - functional units in networks While motifs are relatively small, local patterns in the topology of a network, *modules* constitute larger functional units that may be composed of several motifs and that realise a certain sub-function of the network (e.g., see Hartwell et al., 1999a; Kholodenko et al., 2002; Qi and Ge, 2006). It is interesting to note that while the definition of a motif is entirely dependent on network structure, modules are defined by their function, together with potential clustering of module components. Not all modules that make up a network are always active at the same time. Depending on time and location, only some modules of a network may be activated in the cell (Han et al., 2004), for example by "just in time" assembly of regulating protein complexes (de Lichtenberg et al., 2005).

Functional modules in protein networks can often be defined by protein complexes, i.e. the proteins involved in one complex are part of one module. Of these complexes, a significant proportion seems to have evolved by stepwise duplication of their components (rather than whole genome duplications), followed by divergence of binding characteristics (Pereira-Leal and Teichmann, 2005). For example, a large protein complex with only partially characterised function, relevant to human disease, is the PML body (Chapter 6).

Global organisation: many networks are scale-free On a global scale, the understanding has emerged that many networks in the cell are scale-free, i.e. the degree distribution of their nodes is well approximated by a power-law (Albert, 2005; Barabási and Oltvai, 2004). This is most likely due to growth by preferential attachment, in which new nodes are more likely to form connections with well-connected nodes. Direct support for this hypothesis comes from the fact that in yeast, evolutionary older proteins tend to have more connections than younger ones (Berg et al., 2004; Wagner, 2003). Growth by preferential attachment can be well explained by gene duplications (Barabási and Oltvai, 2004). However, Babu et al. (2006) found that in transcriptional networks, the main regulators (hubs) are often evolutionary unrelated, suggesting a convergent evolution towards a scale-free architecture rather than hub duplication. In the same line of reasoning, Teichmann and Babu (2004) found that key regulators in one organism often have only few homologous target genes in other organisms. Convergent evolution implies a functional advantage of the scale-free architecture, rather than being a mere by-product of the evolutionary process. One potential advantage could be the topological robustness of scale-free networks to random failure of nodes, since apart from the key hubs, failure of single nodes will not disrupt the function of the whole network (Albert et al., 2000; Li et al., 2006).

Chemical organisation theory provides an algebraic link between network structure and behaviour In our group, an approach to link the global network structure to network behaviour has been developed, termed chemical organisation theory (COT). In COT, the stoichiometry of a network is used to compute the organisations of the network, i.e. sets of species that are closed and self-maintaining (Centler et al., 2008; Dittrich and Speroni di Fenizio, 2007; Kaleta et al., 2008), and which represent species combinations that are potentially able to form an attractor. In Chapter 4, I show how the concept of COT has been applied to investigate the structural evolution of chemical computation networks (Lenser et al., 2008).

With Chemical Reaction Network Theory (CRNT), Feinberg (1987) presents an alternative approach to the connection of reaction network structure and function. CRNT focuses on a relatively simple classification of (mass-action governed) networks called deficiency (Feinberg, 1987), which yields information about the potential behaviour of the network, especially with respect to multistationarity. Recently, Craciun et al. (2006) have specialised CRNT to the understanding of bistability, detailing necessary conditions for bistability in reaction networks. Due to computational limitations, CRNT is confined to small systems. Conradi et al. (2007) have bypassed this limitation by analysing subnetworks of larger networks, defined by elementary flux modes (Price et al., 2004; Schuster et al., 1999, 2007). A more mathematical analysis of mass-action systems with respect to stability and bifurcations is given by Gatermann et al. (2005).

1.2.2 Evolution of networks

Network evolution acts on different levels What does evolution act on, structure or function? To answer this question, one needs to distinguish between the different aspects of the evolutionary process. On the one hand, mutations act on DNA, which in turn changes network structure. On the other hand, only the function of a network confers a selective advantage to the organism. It is important to note that there are two highly complicated processes involved: *(1)* The DNA determines network structure, but this mapping in itself is highly non-linear. For example, most point mutations will not have an effect on network structure, but if a single point mutation occurs in just the “right” place, destroying a key binding site, it can disrupt the whole network. *(2)* Network structure, in turn, leads to network function. This process, as discussed above, is again highly non-linear by itself. For the sake of focus, this thesis is mainly concerned with the second mapping, from structure to function.

Evolution of local network structure

Molecular networks evolve primarily through gene duplication and rewiring of connections Even though evolutionary changes in network structure are solely driven by mutation and recombination in the DNA, emergent higher-level mechanisms that shape network structure can and have been observed. The current consensus is that molecular networks evolve by two processes: the gain and loss of interactions between species, and gene duplication leading to duplication of species and interactions (Babu et al., 2006; Wagner, 2003). How fast are these processes? For protein interaction data in yeast, Wagner (2003) calculates an effective gene duplication rate (i.e. without the duplications that are quickly lost again) of 10^{-3} per gene and per million year, which scales up to about 5-6 duplicated genes for the yeast genome per million years. Gain and loss of interactions each occur at approximately 5×10^{-4} per protein pair per million years, leading to more than 100 added interactions for yeast (Wagner, 2003). Thus, gain and loss of interactions happens at a much faster rate than gene duplication.

For the evolution of transcriptional networks, Ward and Thornton (2007) offer evidence from the phylogeny of yeast that the main evolutionary driving forces are rewiring of connections by mutations in promoter regions, and duplication of transcription factors. A very illustrating experimental example from *Saccharomyces cerevisiae* and *Kluyveromyces lactis* is presented by Hittinger and Carroll (2007). Babu et al. (2006) show that transcription factors evolve faster than and independent of target genes, and that local mutations in the regulatory interactions are the dominant mechanism of transcription network evolution. Additionally, transcription factors with similar roles seem to evolve independently in different organisms (Babu et al., 2006; Teichmann and Babu, 2004). Significant features that are commonly detected in distinct networks (like network motifs) are therefore likely to be the result of convergent evolution, hinting at their functional relevance to the cell.

An interesting side note: It has been observed that transcription factor evolution is closely linked to the complexity of the organism. In general, more complex organisms have a higher proportion of transcription factors in their regulatory networks (Levine and Tjian, 2003). This is in contrast to the number of genes, which does not necessarily increase with organism complexity.

Gene duplication can fulfil the need for new functionality From the perspective of population genetics, evolution of new genes by gene duplication presents a problem: the newly copied gene has to be maintained for long enough, and actually to spread far enough in the population, to acquire the advantageous mutation that will ultimately lead to its establishment as a new gene with a new function. Bergthorsson et al. (2007) propose a mechanism in which the parent gene acquires a secondary function before the duplication event. When environmental changes make this secondary function more important, this lends a selective advantage to organisms with multiple copies of the gene, thus favouring gene duplication. The multiple copies of the original gene are now free to lose some of their original dual functionality, which leaves the organism with two functionally different genes.

The role of non-adaptive processes While many of the arguments presented so far invoke the functional - and thus selectional - relevance of either network parts or mutation mechanisms, Lynch (2007a,b) makes an impassionate case for the role of non-adaptive processes in the evolution of DNA organisation. He argues that phenomena like the modular organisation of the eukaryotic genome, genome complexity in eukaryotes and even multicellularity might be (at least in part) the result of mutation and genetic drift, and do not have to be actively favoured by natural selection. A population genetics backing for this argument can be described, mainly by invoking the relative weakness of selection in small populations, as opposed to large populations that are typical for prokaryotes (Lynch, 2007b). In line with this, Wang and Zhang (2007) argue that the modular organisation of protein-protein interaction networks might be due to experimental biases in determining these networks, or the growth process by gene duplication, or both. They present evidence from phylogenetic analysis that modules (in the clustering sense) do not necessarily correspond to biological function (Wang and Zhang, 2007). However, this “neutralist” claim does not reach to network function, which is probably under constant selection pressure in all phyla. From all this, it can be argued that network structure, being situated in-between DNA organisation and network function, might at least in part be shaped by non-adaptive forces (Wagner, 2003).

Evolution of higher-order network characteristics

Evolution of network motifs Considering the impact of the motif concept on the study of network topologies (Section 1.2.1), it is interesting to turn to the mechanisms by which network motifs can evolve. Given our current knowledge, it seems that the widespread appearance of such motifs is not the result of “motif duplication and divergence”, which could have happened by partial or whole genome duplications (Babu et al., 2006). Motifs are usually not copied, which can be seen by observing that the interactions in motifs are not more conserved than those outside motifs, and orthologous genes can be part of different motifs in different organisms (Babu et al., 2006; Mazurie et al., 2005). Rather, it seems that network motifs evolved in a convergent manner, from different evolutionary origins, to the same topologies (Alon, 2007; Amoutzias et al., 2004; Conant and Wagner, 2003; Teichmann and Babu, 2004). Additionally, even though the prevalence of some motifs is not the result of copies of one particular successful motif, it seems to be the case that similar lifestyles of organisms (e.g. generalists vs. specialists) lead to similar network motifs and local interactions (Babu et al., 2006; Teichmann and Babu, 2004).

Some studies have indicated that selection actively favours network motifs. Wuchty et al. (2003) have shown that in yeast, proteins that are organised in motifs are evolutionary more conserved than other proteins. This finding indicates that via the strong link of network motifs and network function, selection on function might be “passed through” to selection on network structure.

Network motifs can arise from duplication and divergence alone However, there is also a different take to the evolution of network motifs. Banzhaf and Kuo (2004) use artificial regulatory networks to show that many network motifs can arise by a gene duplication and divergence process alone, without natural selection on the function of the network. Thus the possibility cannot be ruled out that the significant overrepresentation of motifs in cellular networks in comparison to random networks is not (only) due to functional aspects of these motifs, but (at least in part) a natural consequence of the process by which these network were created (Kuo et al., 2006b). In subsequent work by Leier et al. (2007), the authors make inferences about the specific nature of mutation and preservation processes after duplications that could be at work in genetic networks. Along the same line of reasoning, it has been shown that gain and loss of interactions, together with gene duplication, is able to maintain the scale-free topology of the yeast protein interaction network, without the need to invoke natural selection (Wagner, 2003).

Protein domains - a mediator between DNA and network structure Looking closer at the relation between DNA and network structure, one finds that proteins are arranged of domains, functional units that define the interactions of proteins with each other. Domains can be duplicated, deleted and reshuffled in domain rearrangements (Bornberg-Bauer et al., 2005). Besides gene duplications, domain arrangements form the second major driving force in the evolution of protein interaction networks (Vogel et al., 2005), and specifically transcriptional networks (Amoutzias et al., 2004). They lead to an efficient rewiring of the network, bypassing the time-consuming co-evolution of protein (and/or DNA) sequences that would otherwise be necessary to gain and rewire connections.

Whole genome duplications in plants In contrast to the findings described so far, a study on the MADS-domain MIKC-type transcription factor proteins in plants has shown that their evolution was in large parts driven by whole genome duplications (Veron et al., 2007). These proteins have, in addition to the MADS domain (or MADS box), three more interaction domains, termed I (intervening), K (keratin-like) and C (C-terminal). These domains given the proteins an unusual ability to form heteromultimers (i.e. complexes composed of different proteins, all of MIKC-type), leading to the establishment of a regulatory module that is most likely the reason for their dominant role in flower development (Kaufmann et al., 2005). Veron et al. (2007) argue that the sub- and neofunctionalisation of these higher-order complexes (one specific case is presented by Winter et al., 2002b, and in Chapter 7) is the reason why the MIKC-type proteins were largely retained after each round of whole-genome duplications.

Evolution of network function

Natural selection acts on function Evolutionary selection purely acts on the fitness of the organism, which is in part defined by the function of its cellular networks. Therefore, a top-down view of network evolution must start from selection on function and derive consequences for the structure and dynamics of networks.

Modular networks can more easily adapt to changing environments From a functional perspective, networks can be decomposed into modules with separate functions. There is evidence from bacterial genetic networks that modularity is higher in the genetics of generalists, whose networks need to perform a larger array of functions (Kreimer et al., 2008; Parter et al., 2007). This supports an interesting hypothesis that varying environments lead to more modular genetic networks (Kashtan and Alon, 2005), although varying environments are not a necessary prerequisite for the evolution of modular networks (Hintze and Adami, 2008; Solé and Valverde, 2008).

Modularly varying fitness functions speed up evolution An interesting result for evolutionary biology as well as for evolutionary optimisation algorithms is the suggestion that varying fitness functions speed up evolution, especially when the fitness goal is varied in a modular way (Kashtan et al., 2007). A modular varying fitness function is composed of different sub-functions, which can individually be altered when the fitness function changes. This way, evolved solutions appear to take advantage of the modular structure in the fitness landscape (see also Kashtan and Alon, 2005). However, some speedup is also reported for non-modular changes in the fitness function, hinting at a more general phenomenon in which fitness variation helps to prevent the population from lingering in local optima. Additionally, a dynamic fitness landscape can lead to networks that are "adapted for adaptation", i.e. networks in which mutations in a few key hubs can change the network behaviour in a consistent way, thus changing from one function to another (Crombach and Hogeweg, 2008; Draghi and Wagner, 2009). Papa et al. (2008) demonstrate this effect on the evolution of wing patterns in butterfly.

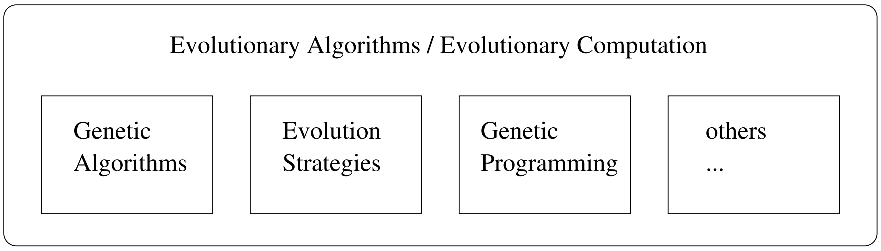


Figure 1.3: Broad structuring of the field of Evolutionary Algorithms.

1.3 Evolutionary algorithms

EAs are popular heuristic optimisation algorithms In any technical and scientific discipline, there are plenty of difficult optimisation problems for which globally optimal solutions can usually not be found. In these cases, the researcher and practitioner is interested in finding an approximate solution, using as few resources as possible. One particularly successful branch of these "heuristic" optimisation algorithms are Evolutionary Algorithms (EAs), i.e. algorithms based on evolutionary principles such as a population of candidate solutions, mutation, recombination, and selection.

As a research area, the field of Evolutionary Algorithms is already decades old, and a variety of different branches and flavours has been developed. To lay the foundation for the artificial evolution of network models, this section compares different approaches to optimisation by evolutionary algorithms. After a formal definition of optimisation and specifically evolutionary algorithms, I give a short review on Genetic Algorithms, Evolution Strategies and Genetic Programming - the main branches of EA research (Figure 1.3) - loosely based on the book by Weicker (2002, in German).

1.3.1 Optimisation problems

Optimisation, loosely speaking, is concerned with finding the best solution, from a set of possible solutions, to a defined problem. In practise, a rigorous and useful description of the optimisation problem often poses a challenge, sometimes the hardest in the whole process. From a mathematical standpoint, however, the general optimisation problem is simple to describe. In the typical case, it deals with a function $f : \Omega \longrightarrow \mathbb{R}$ that maps the search space Ω onto the real line $\mathbb{R}$. Without loss of generality, we can say that f is to be minimised.

Global and local optima A global optimum is now defined as a point $x \in \Omega$ with

$$\forall y \in \Omega : f(x) \leq f(y).$$

In many cases, we can define a metric $d : \Omega \times \Omega \longrightarrow \mathbb{R}$ with

$$\begin{aligned} d(x,y) &\geq 0 \\ d(x,y) &= 0 \Longleftrightarrow x = y \\ d(x,y) &= d(y,x) \\ d(x,z) &\leq d(x,y) + d(y,z) \end{aligned}$$

with $x, y, z \in \Omega$. In this case, a local optimum is a point $x \in \Omega$ with

$$\exists \epsilon > 0 \forall y, d(x,y) < \epsilon : f(x) \leq f(y).$$

We say that solution x is "better" than solution y if $f(x) < f(y)$. Thus, optimisation algorithms aim to find the best solution, either locally or globally. In cases where this is impossible with the available computing resources, the "best possible" solution is sought, often a local optimum in the vicinity of the starting solution.

Genotype-phenotype mapping In analogy to biological evolution, EAs are often (but not always) concerned with two views on the solution candidates: a genotype and a phenotype. The genotype is the representation of the solution in search space $\Omega = \mathbb{G}$, and the phenotype is its "functional" representation determined by the genotype-phenotype mapping $G : \mathbb{G} \longrightarrow \mathbb{P}$. The fitness F is then defined on the phenotype, $F : \mathbb{P} \longrightarrow \mathbb{R}$. Of course, the previously given fitness definition is now defined by composition: $f : \mathbb{G} \longrightarrow \mathbb{R}, f = F \circ G$.

EAs work with almost any fitness function In contrast to many other optimisation techniques, EAs do not impose any conditions on the fitness function (such as linearity, differentiability, or smoothness). However, in order to work properly, they (and any other optimisation algorithm) need at least some degree of strong causality in the fitness function, as discussed in Section 2.2.1.

Moreover, it is important to note that EAs are applicable to a broader class of problems than outlined above, since they do not need an explicitly defined fitness function. For a subclass of EAs, it is enough to have a comparison measure between solutions, i.e. to be able to determine which one of two solutions is the better one. This already enables tournament selection, in which a subset of of the population is compared in a tournament-like fashion and the best one is selected to reproduce.

1.3.2 The general structure of an EA

After this discussion on fitness, we can now turn to a brief sketch of the general evolutionary algorithm. For more detail, consult the books by Bäck et al. (1997), Weicker (2002, in German) or Eiben and Smith (2003).

Mutation, recombination and selection shape a population of solutions Each EA utilises a population, which is a multiset P of elements of Ω (the same solution can appear multiple times in the population). Mutation is defined as an operator

$$M : \Omega \longrightarrow \Omega,$$

turning one solution into a different one. Recombination, in a general definition, turns a vector of n solutions into a new one, i.e.

$$R : \Omega^n \longrightarrow \Omega.$$

Of course, this contains the special case of recombining two solutions into a new one. The only operator influenced by fitness is selection, which takes a sample of the population and determines a subset of that sample that survives or reproduces (by cloning, mutation and/or recombination), depending on the specific algorithm:

$$S : \Omega^n \longrightarrow \Omega^m, m < n.$$

Basic terminology Let us denote the size of the population as μ, while the number of offspring per generation is λ. In each generation, some (or potentially all) members of the population are selected as parents for the λ offspring, either at random or by a selection operator biased by fitness. From these, offspring are created by copying, mutation and recombination. After a fitness evaluation of the offspring, selection happens again to determine which solutions survive to the next generation. Depending on the algorithm, the latter selection acts on either the old population and offspring ($\mu + \lambda$ solutions) or only on the λ offspring.

1.3.3 Genetic algorithms

The evolution of bitstrings The most widely known form of Evolutionary Algorithms are Genetic Algorithms (GAs) (Goldberg, 1989), pioneered by Holland (1973, 1975). Primarily conceived to deal with bitstring-coded solution candidates, they can be distinguished from other EA approaches by their probabilistic (fitness-proportional) selection of parents for the next generation. Heavy emphasis is placed on recombination as the main evolutionary operator, while mutation is used at a low rate to guarantee that all points in the search space are reachable. For bitstrings, mutation can be implemented as the inversion of one bit, while a k-point crossover is created by aligning the parent strings, breaking them at k points, choosing one from each pair and recombining them into a new string.

Most theoretical work on GAs is based on the *schema theorem* (Holland, 1975), which states that in bitstring-coded GAs, short substrings that are highly fit will quickly spread through the population. Therefore, the optimisation does not have to search through all possible bitstrings, but rather through combinations of schemes. However, the consequences of this theorem as well as its application to problems with non-bitstring solution candidates present a heavily debated topic (see Altenberg, 1995, where also the Price Theorem is discussed as an alternative approach to GA theory.). For example, Radcliffe (1992) shows how in specific problems, a non-linear mapping between genotype and phenotype can prevent the application of the schema theorem.

1.3.4 Evolution strategies

Real-valued vectors and Gaussian mutation This flavour of Evolutionary Algorithms (Rechenberg, 1973, 1994a; Schwefel, 1977) usually deals with real-valued vectors $x = (x_1, ..., x_k)$ as solution candidates, taken from some interval $\mathbb{G} = [lb_1, ub_1] \times ... \times [lb_k, ub_k] \subset \mathbb{R}^k$. Therefore, Evolution Strategies (ES) in their conventional form are a special technique for numeric optimisation (see Beyer and Schwefel, 2002, for a comprehensive introduction). The major difference to the procedure in GAs is that ES use a deterministic selection called elitist, in which only the best individuals of a population survive the selection step and are selected as parents for the next generation. In (μ, λ, ρ)-ES, the new generation (of size μ) is selected solely from the λ offspring, whereas in $(\mu + \lambda, \rho)$-ES, the new generation is selected from both parents and offspring

(ρ denotes the number of offspring generated via recombination, the rest is generated using mutation).

Evolution Strategies employ mutation and recombination as evolutionary operators, with the emphasis clearly on mutation. Since the solution candidates in ES consist of real-valued vectors, mutation is realised by the addition of a random value u_i to the component x_i, in the simplest form taken from a Gaussian distribution with density function

$$\phi(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}x^2}.$$

The 1/5 success rule Theoretical work in the field of ES has led to various adaptation strategies for the mutation strength σ for Gaussian mutations, the earliest popular of which is the 1/5 success rule. This rule states that on average, 1/5 of all mutations should have an advantageous effect on the fitness. If the success rate is higher than 1/5, it is assumed that the population is still far from the optimal solution and the mutation strength can be increased. When the success rate drops below 1/5, the mutation strength is decreased to allow local search around the optimum (Beyer and Schwefel, 2002; Rechenberg, 1973).

Adapting a k-dimensional Gaussian mutation: CMA evolution strategy With the drastic rise in computer power over the last two decades, researchers have been able to extend the framework of self-adaption in ES to its fullest and use a full covariance matrix ($\frac{1}{2}k(k+1)$ parameters) for the k-dimensional Gaussian mutation. This way, not only can differently scaled problem parameters be handled (by adapting the parameters on the diagonal of the covariance matrix), but the algorithm can even adapt to linear dependencies between parameters (by adapting parameters off the diagonal, which encode correlations between the mutations for the different parameters). A particularly successful variant of this kind of ES is the “Covariance Matrix Adaption” (CMA) evolution strategy (Hansen and Ostermeier, 2001). I will sketch the CMA-ES here, since it is used in the EA framework described in Section 2, and also in the kinetic analysis of FRAP data (Section 6). For full details, the reader is referred to (Hansen and Ostermeier, 2001).

One global mutation distribution CMA-ES is based on one (global) mutation distribution that is used for all members of the population. It uses a completely derandomised update of the covariance matrix in such a way that previously successful

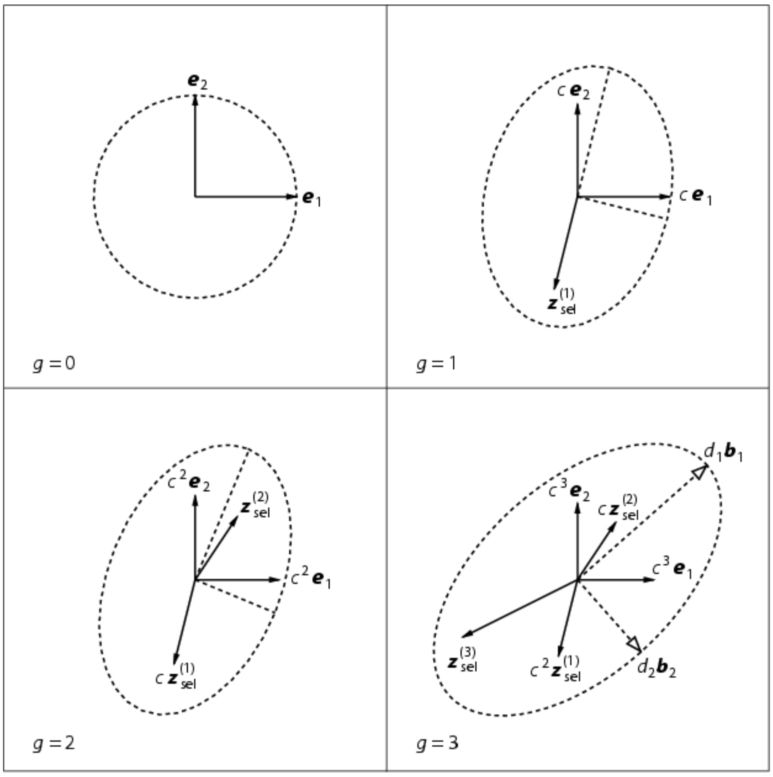


Figure 1.4: Construction of the mutation distribution in CMA, in two dimensions. The initial configuration ($g = 0$) consists of the two unit vectors e_1 and e_2. They produce an isotropic mutation distribution (dashed). The vectors $z_{\text{sel}}^{(1)}$, $z_{\text{sel}}^{(2)}$ and $z_{\text{sel}}^{(3)}$ are added successively at generations $g = 1, 2, 3$, while old vectors are reduced by a factor of $q = 0.91$. Figure reproduced with permission from (Hansen and Ostermeier, 2001).

mutation steps are more likely to happen again in future. In addition, information from a number of previous steps is accumulated to control the step-width.

In order to increase the probability of producing successful mutation steps again, CMA implements a principle component analysis of the previously selection steps and strengthens the direction of the most important components in the covariance matrix (Figure 1.4, for the formulas see Hansen and Ostermeier, 2001). This way, the shape of the mutation probability distribution adapts to the local fitness landscape, effectively performing a linear transformation of the problem representation into (locally) equally scaled, decoupled parameters.

The evolution path controls mutation step length CMA-ES uses an elegant idea to control the length of mutation steps: the evolution path. The idea is the following: Looking at a couple of selected (i.e. successful) mutation steps from previous generations, these can be positively correlated (similar direction), uncorrelated (random direction) or negatively correlated (opposite direction). The evolution path is now defined as the sum of these steps. In case of positive (negative) correlation, the evolution path is longer (shorter) than the one expected for no correlation. If the path is longer, the steps could be replaced by larger steps, thus the stepwidth can be increased. Accordingly, for an evolution path shorter than expected for uncorrelated successful mutations, the stepwidth can be decreased. Actually, CMA-ES employs this procedure both for the global step size (cumulative path length control) and for each dimension independently (covariance matrix update). Again, the reader is referred to Hansen and Ostermeier (2001) for a full description including the update formulas for the covariance matrix.

1.3.5 Genetic programming

Evolution of computer programmes As the name implies, Genetic Programming (GP) deals with the automated design of computer programs in an evolutionary process. A popular introduction to GP is given by Koza et al. (2003), while the reader is referred to Banzhaf et al. (1998), Koza (1992) and Langdon and Poli (2002) for a detailed account of the technique. In contrast to GA and ES approaches, the size of the solution candidates is not fixed in advance, but rather determined through mutation and selection. In its most widely used form, the computer programs under evolution are represented as syntax trees, in which leaves contain values and internal nodes contain commands or mathematical functions. Besides these, a variety of different program representations has been considered, from assembler languages to graph representations (for graphs, see e.g. Teller and Veloso, 1996). The latter is especially suited to biochemical networks, since these can be understood as graphs with two types of nodes, molecular species and reactions (or as hypergraphs, in which multiple species are connected via reactions).

Recombination is the standard variation operator on syntax trees In a syntax tree based GP setup, the recombination operator can be easily implemented by exchanging subtrees of two parents to obtain two offspring individuals. Mutation can be achieved by two operations: the replacement of a subtree by a randomly generated new tree, or the replacement of a single node by a new node. Recombination within one individual, i.e. the exchange of two subtrees, can be seen as another form of mutation. Analog to subprograms and functions in manual software engineering, the technique of automatically defined functions (ADF) evolves a set of separate trees simultaneously to the main tree (Koza, 1994). These are represented by special nodes and can be plugged into the main tree of the program.

Special care is required to cope with the brittle solution candidates In contrast to the binary problems in GA and the numeric optimisations in ES, most mutations and recombinations on programs in GP will have a deleterious effect. Even when the representation (instruction set) and mutation operators are designed to allow only valid solutions, mutations tend to have a drastic impact on fitness. To counter this, a small part of the population is copied to the next generation without modification, so that good solutions are not lost again. Additionally, the average population sizes used in GP are much larger, which is also due to the belief that an initially diverse population is needed for a successful GP run. Typical population sizes used in GP are between 500 and 10000 individuals. The major selection mechanism used in GP research is tournament selection, in which q individuals are randomly selected from the population, and the winner is chosen for the next generation, following recombination or mutation.

1.4 Artificial evolution of cellular networks

Artificial evolution is a toy version of the real thing Over the approximately 4 billion years of evolution on Earth, the genetic, metabolic and signalling networks in living organisms have changed their constituents, connections and functions over and over again. Thus, the networks encountered in today's biological systems have a long evolutionary history and many aspects about them cannot be understood without taking this history into account. While it is certainly impossible to "replay" this evolution in any larger scale, the dramatic rise in computing power over last decades has put small-scale toy models of evolution into our reach.

The *in silico* evolution of cellular network models is studied (primarily) for two reasons: *(1)* Network models are used to represent causal and mechanistical relationships between biological entities gain from laboratory data, and thus the evolution of models is one way of automatic inference (or learning) from this data (Section 1.4.1). *(2)* Artificial evolution has also been used to study the actual process of cellular network evolution, deriving hypothesis about structure and function of evolved networks as well as the evolutionary process itself (Section 1.4.2).

1.4.1 Network reconstruction from data

When looking at approaches to reconstruct networks from experimental data, we can distinguish two general types: pure post-processing approaches that take experimental data and turn it into a network hypothesis, and iterative approaches that propose new laboratory experiments while updating their hypothesis about the network. Below, I give a brief review on the work devoted to both approaches.

Genetic regulation In the field of genetic regulatory networks, network reconstruction is especially advanced and well established (de Jong, 2002; Hecker et al., 2009). For example, Guthke et al. (2005) used clustering of 18432 gene expression profiles followed by dynamic modelling using a linear ordinary differential equation framework to derive a six-dimensional model of immune response to bacterial infection. To determine the network structure of the model, they use a forward-strategy, i.e. adding new nodes and connections to an initially small network. To prevent too large networks, new parts are only added if they significantly increase fitting quality.

In contrast to dynamical models using differential equations, Bayesian networks are aimed at uncovering the chain of causality in regulatory processes. Geier et al. (2007) argue that with the currently available data, Bayesian networks with linear Gaussian dynamics are the optimal tool to capture the essence of gene expression time series, outperforming the commonly used Boolean networks. To account for the effect of hidden (i.e. missing from the data) upstream regulators, Nachman et al. (2004) avoid modelling these directly by using dynamic Bayesian networks that include a time-dependence of some regulatory processes, which can be learned from time-resolved expression data.

A commonly encountered problem with structural optimisation is that often, many structures can fit the same data (Section 1.2.1, 5). Thus, structural optimisation often leads to models that are hard to relate to the already existing knowledge about the system under study. Using the well-studied S-systems modelling formalism (Savageau, 1976), Spieth et al. (2005) try to overcome this problem by employing a multi-objective evolutionary algorithm to fit a network model to microarray data while keeping the evolved network “close” to pathways already described in databases. Thus, they aim to compensate for errors and omissions in the databases, possibly discovering new components and connections. Later, they extended their work to a full software framework for gene regulatory network inference, mainly using evolutionary algorithms (Spieth et al., 2006).

Signalling and metabolic systems Network inference has also been done for systems other than genetic regulation. In one such study, Sachs et al. (2005) use single-cell data on phosphorylated and non-phosphorylated proteins to infer Bayesian network models of cellular signalling, which they verify again in wetlab experiments. For metabolic networks, Wahl et al. (2006) describe a modelling framework based on stimulus-response experiments and statistical analysis with model selection methods. Kholodenko et al. (2002) use “sensitivity-” or “response-coefficients” to capture the reaction of a system to external perturbations. From these, they derive connections between modules or even individual components of the network.

Evolving networks with GP Because of its ability to evolve computer programs capable of specific tasks, the suggestive idea to tweak GP (Section 1.3.5) to infer network models of biological systems has been taken up by several researchers. One of the first approaches, coming from the core of the GP community, is presented by

Koza et al. (2001a,b). Their approach develops a tree representation of biochemical networks, and evolved this using standard GP methods to fit time-series data of a metabolic network.

In a contrasting approach, Sakamoto and Iba (2001) directly evolve the ordinary differential equation system that is used to fit the data. Additionally, they replace the mutation of numeric constants with an optimisation of equation parameters for each fitness evaluation, calculated by linearising the ODE system and solving the linear optimisation problem. Later, they supplied results indicating a robustness of their estimation to noise inherent in the GP process (Ando et al., 2002). In contrast to the linearised fitting procedure used by Sakamoto and Iba (2001), Sugimoto et al. (2005) use numeric mutations of the equation constants in GP to evolve the equation system of metabolic reactions.

Comparison of methods In an exemplary comparison, van Someren et al. (2001) rank different fitness functions and different structural optimisation methods on the task of gene network inference. They find that for the fitness function, best results are achieved by the mean-squared error of models prediction compared to training data. However, the other two investigated fitness functions (maximum squared error and leave-one-out cross-validation) performed only slightly worse. They also find that a structural optimisation method based on greedy addition of connections followed by parameter fitting outperforms a genetic algorithm on finding the correct network topology. As a note of caution, it has to be stated that van Someren et al. (2001) use a linear model of the genetic network with discrete timesteps, making parameter fitting very efficient. Thus, the results may be different when more complex models are used. In this light, an interesting proposal has been made by Gennemark and Wedelin (2009), who develop a set of benchmark problems, especially aimed at the automatic inference of ordinary differential equation systems.

Iterative approaches In comparison to post-processing of experimental data, less work has been done on iterative approaches, potentially due to the more time-consuming nature of these investigations. For example, Ideker et al. (2000) employ a cycle in which first all Boolean networks consistent with the current set of experimental results are computed, and then the most promising new experiment is proposed by an entropy-based measure. On a larger scale and again with a Boolean model, Barrett and Palsson (2006) use the known transcriptional and metabolic network of *E. coli* together with information on all previously performed experiments to calculate the experiment that has the highest potential to discover new knowledge about regulatory interactions.

In a manner analog to Ideker et al. (2000), Bongard and Lipson (2004) use a two-stage evolutionary process (they use the term "coevolution") to first evolve a (time-discrete, concentration-continuous) model of the system based on data, and then evolve the next experiment which is expected to bring the most new information about the system. They apply their algorithm both for genetic network inference and for self-diagnosis of a damaged robot. In a similar study but focusing only on parameter fitting (with a pre-determined ODE model), Gadkar et al. (2005) use a criterion of parameter-identifiability to propose new experiments.

In a follow-up to their previous work, Bongard and Lipson (2007) amend a simple ODE-inference GP system using hill-climbing selection with a technique called "partitioning", which allows to fit equations for each variable separately. During evaluation of the synthesised equation, values of the other variables are not computed, but approximated from the given time-series dataset. This way, the effort for equation synthesis scales only linearly with the number of variables, leading to increased scalability compared to approaches which infer the whole ODE system at once.

1.4.2 Artificial evolution studies

In contrast to knowledge extraction from data, which is the central goal of network reconstruction approaches, the studies of artificial evolution summarised in this section are concerned with evolutionary mechanisms and how they influence the process of network evolution.

Early work In probably one of the earliest studies in this direction, Kacser and Beeby (1984) utilise evolutionary techniques to make inferences on the evolution of enzymes. A decade later, Bray and Lay (1994) use an EA to fit the parameters of a simple network consisting of a ligand, two receptors, a signal molecule and a target molecule. They are able to evolve a large variety of responses to inputs represented by waves of ligand concentration, thus demonstrating that evolution of function in network models is possible.

In a study on the stoichiometry of ATP-producing networks, Stephani et al. (1999) present an EA to design stoichiometry models of glycolysis coupled to an external ATP-consuming reaction. They show that while a direct enumeration of all network variants is impossible beyond a certain network size, an EA is able to produce a variety of networks with a common function, thus enabling the comparison of these networks and extraction of common features.

Evolution of mathematical functions In analogy to Chapter 3, but historically earlier, Deckard and Sauro (2004) evolved biochemical network models capable of performing arithmetic functions such as square root, cube root and natural logarithm. They give a detailed account of the evolutionary procedure used and the difficulties encountered. In contrast to Chapter 3, the EA evolves network structure and parameters together, with parameter mutations occurring more frequently than structural mutations. While they were able to (easily) evolve networks for the square root task, and analytically proved that these networks compute a square root function, the evolution for cube root and logarithm functions was much more difficult, and the most successful networks merely gave polynomial approximations to the desired solution.

Evolution of biologically relevant behaviour Besides the more abstract, mathematical functions that were evolved in the work described above, artificial evolution can be targeted towards more complex, biologically realistic functionality. One popular example is chemotactic behaviour, usually simulated by controlling the tumbling frequency of an *in silico* bacteria. For rather generic models of signalling networks, Soyer et al. (2006) evolve the connection strengths in fully connected networks of three, four and five nodes. They show that a three-node network is already capable of controlling chemotaxis by acting as a derivative-sensor, but an increase in network size leads to better performance. Using genetic programming of an artificial chemistry, Ziegler and Banzhaf (2001) evolve a control metabolism for a robot and relate their findings to motile bacteria.

Based on models of genetic networks, François and Hakim (2004) evolve and mathematically analyse bistable switches and oscillating networks. Building on the work of Deckard and Sauro (2004), Paladugu et al. (2006) evolve networks capable of performing more advanced functions, such as a bistable switch, oscillators, homeostasis, and frequency filtering. Their networks are based either on mass-action, Michaelis-Menten or Hill kinetics.

Artificial evolution can be used to study the origin of network complexity To investigate the evolution of complexity in cellular networks, Holland (2002) uses the concept of classifier systems to evolve networks capable of responding to signals.

Decraene et al. (2007) build on this work to develop a model in which computational aspects of the evolution of cell signalling networks can be studied.

One intriguing question in the evolution of biochemical pathways is why almost all networks found in molecular biology are more complex than they need to be, i.e. more complex than the minimal network capable of performing the same function. In an attempt to answer this question, Soyer and Bonhoeffer (2006) present an artificial evolution model that displays an increase in network complexity driven by neutral drift rather than selection, even though selection acts on the level of network function. The intuitive explanation presented is that while destructive mutations tend to diminish network function, mutations adding proteins to the network are less likely to do so. Thus, selection for network function leads to a bias in the selected mutations that increases network size (Soyer and Bonhoeffer, 2006). Of course, this conclusion does not rule out a role for robustness and other characteristics in the evolution of network complexity.

A word of caution In general, one has to be cautious when transferring results from artificial evolution to real biological networks. Chu (2007) argues that when reviewing artificial evolution experiments aimed at evolving oscillators, one cannot find any "typical" network structure that is evolved. This makes it very hard to compare the evolved networks to examples of oscillating networks found in cellular systems.

1.4.3 Other approaches

Novel algorithmic ideas inspired by biological networks An additional research direction in the artificial evolution of networks is aimed at novel computational procedures rather than biological insight. For example, Banzhaf and Lasarczyk (2005) use genetic programming to construct algorithmic chemistries, i.e. sets of molecules that each represent register machine instructions. At runtime, the instructions are randomly drawn from the instruction multiset, and the specific mix of instructions in this "program" ensures that the correct result is approximated. GP can act on these instruction sets in a natural way: mutation and recombination can be implemented by addition/removal of instructions and mixing of instruction sets.

A major field: The evolution of artificial neural networks In a parallel effort to the artificial evolution of cellular networks, much work has been done on the evolutionary construction of artificial neural networks (see e.g. Haykin, 1999). Here, we find striking parallels to the evolution of network models in systems biology: both structure and parameters have to be determined (Yao, 1999). Quite often, evolution and learning are combined in training ANNs, or an evolutionary structure optimisation method is combined with a faster (but more local) gradient-based method for training of the ANN.

1.5 Concluding remarks

This chapter gave a very brief review of the literature on structure and function in network biology, evolution of networks, evolutionary algorithms, and artificial evolution. It is by no means designed to be complete, but to give a taste of the fields involved in this thesis, and to name the most relevant developments and publications.

As with any interdisciplinary work, the background for this thesis is wide and diverse. Each of the topics mentioned above is important for the following chapters, but each chapter has its own focus. Chapters 2 and 3 are based heavily around the application of Evolutionary Algorithms to the artificial evolution of networks. Chapter 4 then deals with the application of the created evolutionary framework to the evolution of computing networks, analysed with the concept of Organisation Theory. Chapter 5 has a more real-world focus, applying the evolutionary concept to data from FRAP experiments, and thus highlighting interesting aspects in cellular network biology. Shifting the focus from algorithmic issues to the biological application, Chapter 6 describes the application of CMA-ES to ordinary differential equation models of PML nuclear body kinetics and HP1 binding to chromatin. Chapter 7, the only one without any evolutionary algorithms in it, presents a model for an evolutionary transition in genetic network architecture, and highlights how comparative stochastic modelling can be used to gain new insights into this evolutionary motif.

2

SBMLevolver: An Evolutionary Algorithm for Network Models

Chapter summary. *To gain a practical understanding of the artificial evolution of cellular network models, this chapter describes the design of an evolutionary algorithm for this task, and its underlying rationale. The central design idea is a separation of structural evolution of the network from kinetic parameter evolution, aiming at stronger causality on the structural level by smoothing fitness transitions between network structures. In this way, network parameters can adapt to a new topology before this topology is evaluated. Results are presented which show that this two-level approach yields a pronounced increase in the algorithm's fitness performance. More specifically, our studies show that this separation helps to prevent premature convergence when evolving networks performing arithmetic calculations.*

Chapter contents

2.1 Introduction

The artificial evolution of networks *in silico* is an important, highly timely topic in bioinformatics research. It is used primarily in two forms (see Section 1.4): (1) for the automated inference of network models from a variety of data sources, and (2) for novel insights into the evolutionary process that shaped and shape dynamical networks. However, recent investigations of this topic were either pioneering studies that opened up a concept rather than investigated the evolutionary procedure in detail (Bray and Lay, 1994; Koza et al., 2001a,b; Sakamoto and Iba, 2001), or have focused on the evolved networks and their behaviour (Deckard and Sauro, 2004; François and Hakim, 2004; Paladugu et al., 2006; Soyer et al., 2006), not on the EA itself. Of course there are exceptions, but many target specific network formalisms (Spieth et al., 2005; Sugimoto et al., 2005) and may not be generally applicable.

Investigating the process of evolving dynamical networks requires a modular, transparent framework in which different factors influencing the process can be clearly distinguished. To this end, the SBMLevolver software was developed. This chapter will present the software, its design and implementation. Results on the effect of the design decisions are given in Chapter 3.

In the introductory chapter, Section 1.3 gave a brief introduction into some of the main categories of Evolutionary Algorithms. This list is by no means exhaustive, but is meant to highlight the main ideas in the field. For a specific problem, the standard methods outlined in Section 1.3 will be inferior to a tailor-made solution, which takes the specific problem structure into account[1]. Therefore, a design for an EA specific for the evolution of dynamical networks was conceived.

The SBMLevolver software, together with a description and usage instructions, is publicly available on `http://users.minet.uni-jena.de/~biosys/esignet/`.

[1] In a more formal language, this statement is called the *No free lunch* theorem and can be proven mathematically (Wolpert and Macready, 1997)

2.2 Design of a two-level evolutionary algorithm

2.2.1 Causality

Strong causality is important for optimisation In this discussion, two variants of causality come into play: Firstly, weak causality (or just causality) denotes that two identical actions lead to identical consequences. For our discussion of EAs, this simply means that the fitness function is deterministic, always yielding the same fitness for the same model. Thus, two identical mutations of one model will always have identical effects. Secondly, strong causality means that small changes to the solution lead to small changes in fitness. In other words, the neighbourhood structure of the search space is carried over to the fitness space.

It is intuitive that some degree of strong causality is needed for any type of optimisation algorithm to work. Without it, the fitness function would - from the optimisation point of view - just be a random mapping, and the search would be equivalent to testing of randomly generated solution candidates for their fitness.

A measure for strong causality In an interesting theoretical treatment, Sendhoff et al. (1997) have defined strong causality based on the transfer of neighbourhood structure from genotype to phenotype space. Directly applying their measure, any evolutionary system with an identical genotype and phenotype would be strongly causal (this is the reasoning used for numerical optimisation with ES). However, a strongly causal genotype-phenotype mapping has no advantages to the search process if the phenotype-fitness mapping is not strongly causal. To my mind, it seems much more natural to apply the strong causality definition of Sendhoff et al. (1997) to the genotype-fitness mapping, as this is what the EA ultimately acts on.

In analogy to Sendhoff et al. (1997), I now give a definition of strong causality for the mutation operation in EAs. For this, a distance measure (metric) on genotype space has to be defined. An intuitive distance between two network models m_1 and m_2 is given by the probability $P(m_1 \to m_2)$ of creating m_2 from m_1 by mutation. Taking the logarithm of this value to make the measure additive rather than multiplicative, we get:

$$d(m_1, m_2) = -\log P(m_1 \to m_2).$$

Note that

$$\begin{aligned} d(m_1, m_2) < d(m_1, m_3) &\iff -\log P(m_1 \to m_2) < -\log P(m_1 \to m_3) \\ &\iff P(m_1, m_2) > P(m_1, m_3). \end{aligned}$$

For simplicity's sake, the parametrisation of the mutation operator is neglected, but it is straightforward to formulate this idea with a parametrised mutation (Sendhoff et al., 1997). A genotype-fitness mapping f is termed "strongly causal" if it fulfils the following condition:

$$\begin{aligned} &\forall m_1, m_2, m_3 \in \Omega : \\ &\quad \|f(m_1) - f(m_2)\| < \|f(m_1) - f(m_3)\| \iff P(m_1 \to m_2) > P(m_1 \to m_3) \end{aligned} \tag{2.1}$$

Usually, we do not require this condition for all triplets of models, but rather for triplets taken from a local environment, thus enforcing "local" strong causality. In the case of structure evolution of networks, this locality is defined by the number of mutation steps allowed per generation.

In the same publication, Sendhoff et al. (1997) also introduce a quantitative variant of Equation 2.1: If m_1, m_2 and m_3 are taken as random variables (with the locality condition in mind), then the inequalities in 2.1 become Boolean random variables:

$$\begin{aligned} A &:= \|f(m_1) - f(m_2)\| < \|f(m_1) - f(m_3)\| \\ B &:= P(m_1 \to m_2) > P(m_1 \to m_3) \end{aligned}$$

Now we can quantify strong causality as the conditional probability $P(A|B)$ that the model m_2 is nearer to m_1 than m_3 is to m_1, given that the fitness difference between m_2 and m_1 is smaller than between m_3 and m_1.

While I did not use this exact formalism to investigate the evolutionary system, preliminary results with fitness-distance correlation (Jones and Forrest, 1995; Tomassini et al., 2005; Vanneschi et al., 2003) indicate that strong causality is present in the EA system (Section 2.2.5).

Network evolution is not (always) strongly causal The relationship between a network model and its behaviour is certainly not strongly causal. To see this, look at the smallest mutation possible, the slight change of one kinetic parameter. In most cases, this change does not effect network behaviour very much. Thus, two identical networks with just one slight change in one kinetic parameter must lie closely together in any intuitive representation of the search space. However, in rare but important cases, their behaviour might be totally different due to a bifurcation in the network's dynamics (see e.g. Bindschadler and Sneyd, 2001; Swat et al., 2004). Bifurcations are an inherent property of the dynamical systems describing network models, and thus this difficulty cannot be overcome by choosing a different representation.

Network evolution is not random On the other hand, the evolution of a network model with a specified behaviour is by far superior to a random search for such behaviour (see Section 3.2), so that the model-behaviour relationship is certainly far from random. It lies in the intermediate zone between a random mapping and a strongly causal one, and the apparent difficulty of evolving networks with a non-trivial function hints at a position closer to the random pole than many other optimisation problems.

Strong causality is difficult to enforce Other approaches that deal with the evolution of dynamical systems have similarly realised the need for strong causality in the fitness function (Hirche et al., 2002; Santibáñez Koref et al., 2001). However, to my knowledge, no general procedure to enforce strong causality has been proposed (Weise et al., 2009), and it seems very likely that this is indeed impossible for the general task. Weak causality is an inherent property of the problem representation and the attached variation operators, and thus has to be tackled at this level.

Inferring the degree of strong causality from EA success If the assumption is right that the degree of strong causality in an EA system to a large part determines the EAs success in finding a good solution (Sendhoff et al., 1997), we can also use the EA's quality to draw conclusions about the degree of strong causality in the system. An appropriate measure is the comparison of the best fitness found by the EA against the best fitness found by a random search with the same resources, which corresponds to a system with locally uncorrelated fitness values (weakly but not strongly causal).

2.2.2 Network evolution on different levels

Network evolution acts on three levels In the evolution of cellular network models, three different levels have to be considered. The highest is the network topology, containing the participating proteins and their interactions. On an intermediate level,

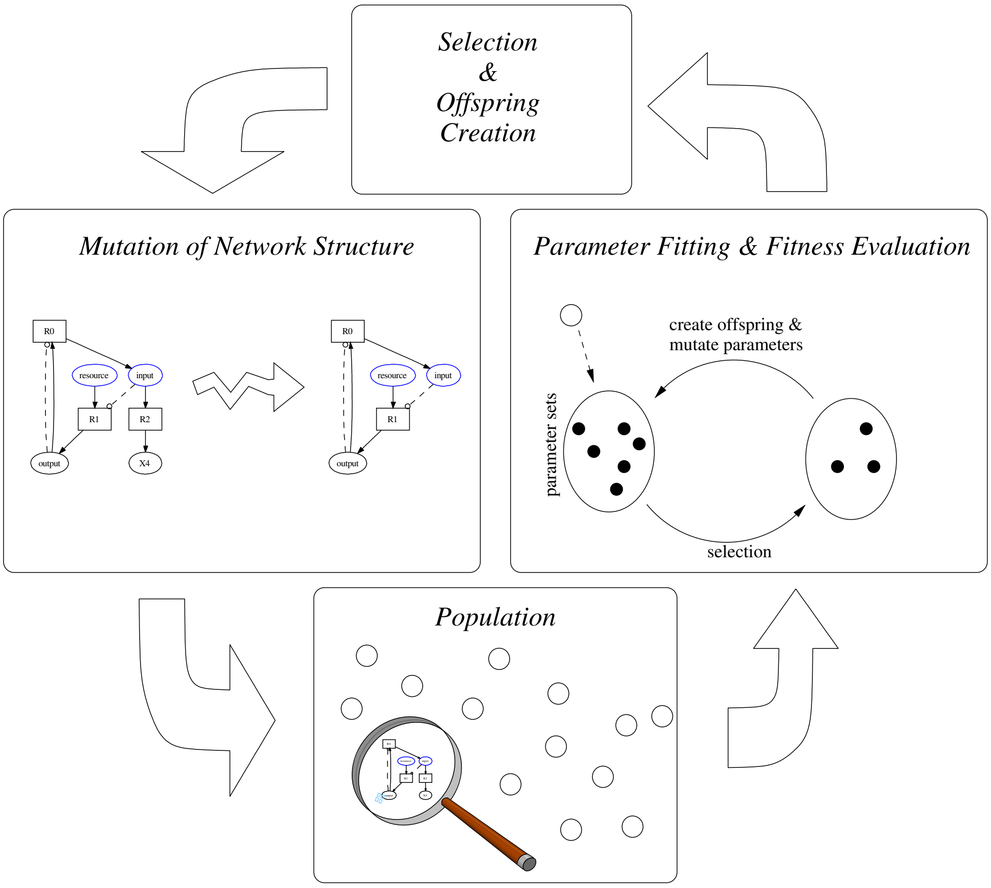


Figure 2.1: Illustration of the dual-layer Evolutionary Algorithm. The outer loop optimises the network structure, while the inner loop (inside the fitness evaluation) fits the network parameters to the desired behaviour.

the reaction kinetics contain information about the type of interaction, while the lowest level consists of the exact numeric parameters (note that these levels are independent of the structural, dynamical and functional level used in the description of networks). For the SBMLevolver software, I have chosen to apply different strategies at each level, in the hope of making maximal use of the advantages of the appropriate evolutionary concepts.

Genetic Programming can evolve the structure and kinetics of networks The presented approach understands the evolution of network topology (i.e. species and their connections via reactions) as a form of Genetic Programming, since GP concepts seem to be the most adequate to the evolution of graph structures. Different rate laws of reactions are represented as different interactions, which integrates the dynamic formulation of reaction kinetics (second level) with network topology (first level). In contrast to programs evolved in classical GP, network models cannot be represented as trees, as feedback and cycles are integral components in many of them. Therefore, the GP variant employed here has to be graph based, considering molecular species as nodes (with their concentrations as values) and reactions as directed connections between nodes. This immediately implies that we are dealing with hypergraphs, as a reaction can have more than one source and destination, respectively.

Kinetic parameters are treated with Evolution Strategies Apart from its topology, a network model description also involves a variety of kinetic parameters (the third level), which have to be tuned to the given problem. This is essentially a numeric optimisation problem in a highly nonlinear setting, for which CMA-ES (Hansen and Ostermeier, 2001) constitutes a state-of-the-art heuristic optimisation algorithm. Therefore, the SBMLevolver couples the CMA-ES algorithm with graph-based GP, optimising the network parameters for each evolved topology. To this end, a two-level evolutionary algorithm was implemented (Figure 2.1), where the upper level evolves the model structure in analogy to GP, while the lower level takes care of the parameters with a CMA-ES.

Nested Evolution Strategies are a known concept It has to be noted that the idea of a two-level evolutionary algorithm is not new. In essence, it is equivalent to the nested ES described by Rechenberg (1994b) (with different representations and mutation operators on the two levels), and has been applied to structure evolution by Lohmann (1993). In formal ES notation, we are dealing with a $[\mu_S + \lambda_S(\mu_P, \lambda_P)]$ strategy. However, to my knowledge, this is the first application to the evolution of cellular network models.

2.2.3 Network representation and search space

SBML is the current standard representation language for biomolecular networks The SBMLevolver system was designed with the goal to enable a comprehensive investigation of the artificial evolution of network models. For this reason, it was decided not to use a specialised representation involving a genotype of some sort that would be mapped to the phenotype (a dynamical system, usually a system of ODEs). Rather, I chose to use the current standard in the systems biology community, the Systems Biology Markup Language (SBML, Hucka and Finney, 2005; Hucka et al., 2004, 2003), to represent network models directly. As of today, more than 180 software packages are able to work with SBML model specifications (`http://sbml.org/SBML_Software_Guide`).

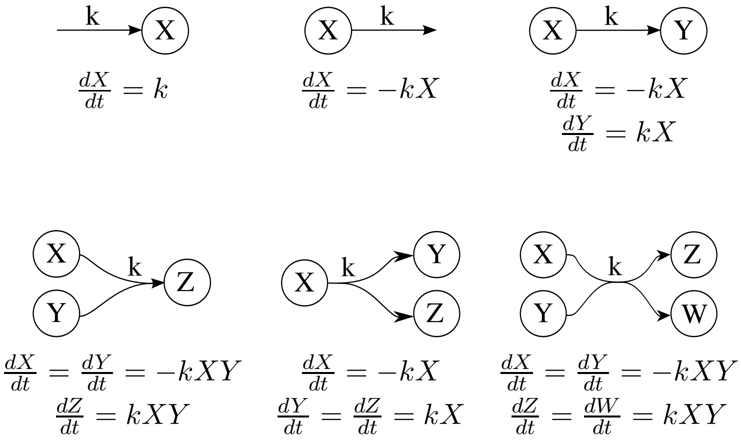
From a preliminary review of the literature available at the start of this work, it seemed that no other system was able to significantly reduce the causality problem mentioned above, so there was no reason to go with a representation different from the standard one. This choice of a direct representation also ensures that the results generated with the SBMLevolver system do carry some weight beyond the specific evolutionary system under study.

Search space limitation SBML allows a nearly unlimited freedom in the model representation. Kinetic rates are specified by arbitrary mathematical expressions, and reactions can have as many reactants, products or modifiers (enzymes, activators, inhibitors) as desired. The same freedom applies to the SBMLevolver software. To keep things simple, however, the network models evolved here contain of reactions of second order or less (Table 2.1), and use pure mass-action kinetics, unless stated otherwise.

In SBML, a model is represented by a list of species and a list of reactions, where each reaction is associated with a number of reactants (species that are consumed by the reaction), products (species produced), and modifiers (species which control the reaction but are neither consumed or produced, such as enzymes). Additionally, each reaction has its own kinetic law, determining the reaction rate.

Search space grows very fast and contains many similar regions To approximate the size of the search space, we can make a straightforward assessment. To this end, we assume that all species and reactions are of generic type, i.e. only placeholder species and no specific species names are used in both species and reactions. In our typical case, the reaction involving the most species is of bi-bi type, and so in a network of n species, essentially n^4 different reactions are possible. More specifically, with the reaction types of Table 2.1, $n^4 + 2n^3 + n^2 + 2n$ different reactions are possible. In a network with k reactions, each reaction can be drawn from this pool of possible reactions, yielding $(n^4)^k = n^{4k}$ network topologies (using only the highest order term).

Usually, the EA deals with a search space describing networks that contain between 0 and N species and between 0 and R reactions. Thus, the total size of the search



$$\frac{dX}{dt} = k \qquad \frac{dX}{dt} = -kX \qquad \frac{dX}{dt} = -kX,\; \frac{dY}{dt} = kX$$

$$\frac{dX}{dt} = \frac{dY}{dt} = -kXY,\; \frac{dZ}{dt} = kXY \qquad \frac{dX}{dt} = -kX,\; \frac{dY}{dt} = \frac{dZ}{dt} = kX \qquad \frac{dX}{dt} = \frac{dY}{dt} = -kXY,\; \frac{dZ}{dt} = \frac{dW}{dt} = kXY$$

Table 2.1: Types of reactions typically used to evolve models in SBMLevolver runs, together with the associated mass-action kinetics. Note that production, decay and conversion reactions with a catalyst are special cases of these reactions where one educt and one product are identical.

space is approximated by

$$SS(n,k) = \sum_{n=0}^{N} \sum_{k=0}^{K} n^{4k}.$$

Even though there exist formulas for the sum of the k-th power of the first n integers (e.g. Faulhaber's formula), applying them here does not make this approximation more intuitive. Instead, I will give a few instances to supply the reader with a feeling for the search space size:

$$\begin{aligned} SS(2,2) &= 257 \\ SS(3,10) &\approx 10^{19} \\ SS(10,10) &\approx 10^{40} \\ SS(30,30) &\approx 10^{177} \end{aligned}$$

Even if we take into account that many of these topologies are actually identical due to permutations in the species and reactions, this does not reduce the size of the search space for the EA. Thus, we are looking at a vast search space, in which many regions are mirrored images of others with permuted dimensions.

2.2.4 Design of mutation and recombination operators

Desired characteristics of variation operators In natural evolution, mutations that change a cellular network actually happen on the DNA level, which constrains the effect of these mutational operators. In contrast, network mutations in artificial systems can take any form. Therefore, it is of interest to gather desirable characteristics of mutational operators on networks, so that optimal operators can be designed.

To describe these desired characteristics, one needs to define a suitable metric on the search space Ω. A metric is considered "suitable" if it creates a smooth fitness landscape over the search space, i.e. candidate solutions that lie closely together in the search space have similar fitness values. Such a metric corresponds to the concept of strong causality (see Section 2.2.1) in an EA system. Obviously, a metric that violates this criterion cannot be used in this context.

Along this line, Droste and Wiesmann (2000, 2003) propose a set of guidelines to design mutation and recombination operators. To allow a smooth movement of the population through the search space, they require three points from the operators:

- Every point in the search space should be reachable from any initial population
- "Small" mutations should have a larger probability than "large" ones
- The operators themselves should not have any preferred direction in search space

Mathematically exact formulations of these criteria, and specialised formulations for mutation and recombination, are straightforward to derive (Droste and Wiesmann, 2000, 2003). In an interesting attempt to enforce these criteria, Wiesmann (2002) applies these guidelines to design a semantical mutation operator for Evolutionary Programming of deterministic finite automata, i.e. the mutation operator takes information on the solution behaviour into account.

The first criterion given above is self-explanatory, an absolute must-have for any optimisation algorithm. In the same way, the third criterion is quite obvious, as any preferred direction in an ideal EA should come from selection according to the fitness function, and not from any other fitness-blind operator. The most interesting criterion is the second one. This requirement, together with the smooth metric, ensures that small fitness changes occur more often than large ones, aiding the population on a smooth path towards to optimum.

In addition to the second criterion, many mutation operators in EAs are designed to be scalable, such that the average effect of the operator can be increased or decreased according to the search situation (Beyer and Schwefel, 2002). In exploratory phases of the search, when not much is know about the location of the optimum, the stepsize can be increased for faster exploration. When an optimum is located, the stepsize can be decrease to allow a more fine-tuned approach to the final solution.

Essentially, the work on strong causality by Sendhoff et al. (1997) and the operator design guidelines Droste and Wiesmann (2000, 2003) look at the same thing from different perspectives: While Sendhoff et al. (1997) look at a combination of representation and mutations to determine criteria for strong causality, Droste and Wiesmann (2000, 2003) use the notion of strong causality to create such combinations.

Applicability of design guidelines For the above design guidelines, the existence of a smooth metric, i.e. one that leads to neighbouring solutions having similar fitness, is crucial. However, it can be argued that in any hard optimisation problem, such a metric does not necessarily exist. Rather, one characteristic feature of hard optimisation problems is the existence of ridges in the fitness landscape, prohibiting the use of simple hill-climbing techniques.

The objection can be raised that such a metric is straightforward to construct: One can simply define the distance of two network models by the absolute difference of their fitness values. However, such a metric needs explicit knowledge about the fitness values of every solution. Additionally, in order to gain insight into the design of variational operators from this knowledge, one would need an understanding of the structure-fitness mapping. To see this, imaging a scenario in which we know the fitness value of every possible solution. However, the fitness function assigns fitness values purely at random, i.e. the fitness value of a solution does not depend on any characteristic of the solution. Now we can design the aforementioned metric, but this does not help us at all in designing proper variational operators.

The structure-function mapping of dynamical systems is typically very complex, poorly understood, and not very robust to changes in the structure of the system. Therefore, we cannot expect to define mutation and recombination operators that strictly follow the guidelines by Droste and Wiesmann (2000, 2003). In the following, I discuss how in spite of this, it was tried to implement operators following the guidelines as much as possible.

Mutation operators The evolutionary algorithm acting on network structure employs seven different mutations:

- Addition / deletion of a species. When a species is added to the network, it is initially not connected to any reaction, but has to be connected via a later mutation. When a species is deleted, all reactions are checked for occurrence of that species, and it is replaced by a species picked at random from the remaining ones.

- Addition / deletion of a reaction. When a new reaction is added, its reactants, products and modifiers can either be placeholders or specific species IDs. If they are placeholders, these are replaced by randomly picked species from the model. In case of specific species (which have IDs that are already represented in the model), these are kept.

- Replacement of a reaction with a different one. When replacing a reaction by a new one, the participants are kept to the largest degree possible, and necessary adjustments in the participating species numbers happen by random addition and removal of species.

- Duplication of a species with all its reactions. This leaves the duplicated species and reactions free to mutate further, while the original function is not perturbed.

- Cloning of a network without changing its structure

Note: In earlier versions of the SBMLevolver, the replacement of a reaction with a different one was not implemented. Instead, single connections between a species and a reaction could be added and deleted, accompanied by an appropriate change in the kinetic law of the reaction. Later, it was felt that these mutations led to a high proportion of invalid models, since often the addition or removal of a reaction participant led to a reaction for which no kinetic law was available. The replacement of a reaction is a slight generalisation of these mutations, and leads to a higher percentage of valid mutations.

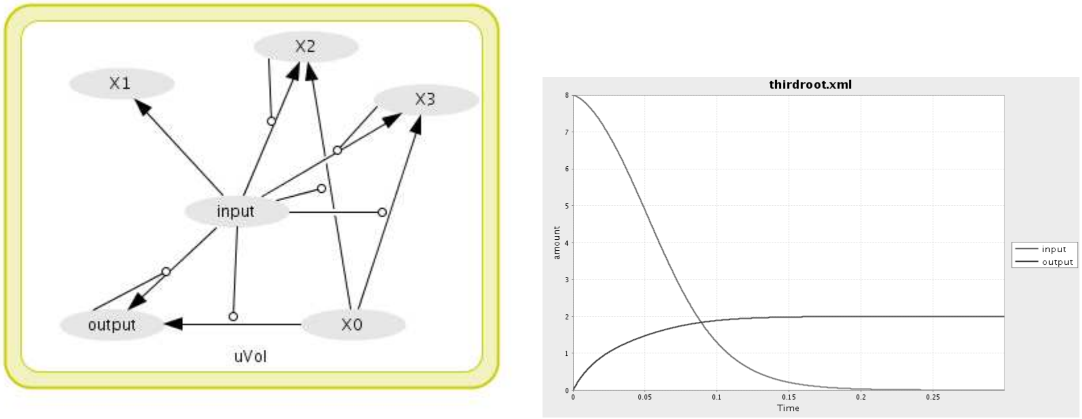


Figure 2.2: Example solution and corresponding time series of *input* and *output* species for the third root network, produced using the CellDesigner Funahashi et al. (2003) tool.

Crossover It is generally very difficult to perform a network crossover operation that does not have a very high chance of destroying the network function (Deckard and Sauro, 2004; François and Hakim, 2004; Paladugu et al., 2006). An "intelligent" crossover operator would involve matching of one graph onto another one, in order to identify common subgraphs. This matching is difficult since it suffers from the permutation problem (Hancock, 1992; Yao, 1999), i.e. species not bound to a specific name ("X_1" rather than "ATP") can be permuted in different networks, yielding effectively different structures with the same function.

In spite of this, a simple crossover operator has been implemented that creates an offspring from two parent models, by merging all species from both models and copying each reaction with probability one-half. This is a very simple implementation since it avoids the need to match parts of one parent to the other. However, it does not guarantee that crossover of one parent with itself will reproduce that parent. In fact, this is very unlikely.

Even though crossover of unrelated models probably has disastrous effects on their behaviour, it may very well have interesting consequences later in the search. When the models in the population are closely related, crossover may lead to "genetic repair" effects, extracting advantageous features from individual members of the population (Beyer, 1997).

Initial solutions and model backbone At the beginning of the run, initial solutions can be given as SBML model. In these, specific species and reactions that are deemed necessary for the model can be marked as undeletable, which prevents their deletion by any structural variation operator. This feature is useful if one wants to incorporate knowledge about the system that should by built upon by the evolutionary process.

2.2.5 Two-level approach

Parameter fitting after structural mutations amplifies strong causality In the discussion above, it became clear that the degree of strong causality in the evolutionary systems significantly determines the expected chance of success for the artificial evolution of cellular network models. Following this reasoning, I now outline an approach that tries to strengthen strong causality on the structural level.

Typically, a structural mutation - such as the deletion of a reaction - has a drastic effect on the behaviour of the mutated network. Even in cases where the mutated network is still potentially able to produce a behaviour similar to the parents, its parameters will not be fine-tuned for this. Looking at this from a different angle, a parameter adaptation after a structural mutations gives the network the possibility to get the best possible fitness value. If the parent network already has a relatively good fitness, this will increase the chances of the offspring behaving similarly (good) to the parent, strengthening strong causality (Yao, 1997). In other words, parameter fitting after structural mutations softens their effect, and can have a positive effect on the EAs trajectory through the structural search space (Emmerich et al., 2001).

Preliminary results indicate some strong causality For a first glimpse at the causality behaviour of the designed EA system, I computed the fitness correlation between individual models and their offspring under mutation. Both direct offspring and higher-order siblings (e.g. grandchildren) with up to ten generations distance were considered. The results (Figure 2.3) clearly show that there is a positive fitness correlation between a model and its direct offspring. For later generations, correlation depends on the method of measuring it. While Pearson's linear correlation coefficient (Figure 2.3 left) indicates no further correlation after more than one generation, Spearman's rank correlation coefficient suggests that correlation decreases in a linear way as the generational distance increases (Figure 2.3 right).

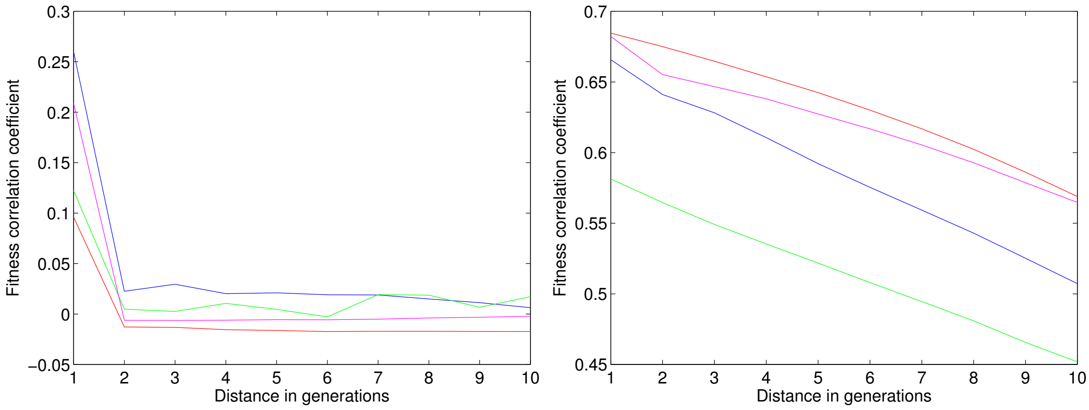


Figure 2.3: Estimates of fitness distance correlation for the evolution of networks computing a logarithm and a polynomial (blue and green, respectively), a network with oscillating behaviour (red), and a network with a discrete switching threshold (magenta). Shown are the number of generations between parent and offspring vs. the correlation in fitness values. The left panel shows Pearsons's linear correlation coefficient, the right one shows Spearman's rank correlation coefficient.

This difference can be explained by noting that Pearson's method measures *linear* correlation, which can be greatly disturbed by outliers. In our case, outliers are a common phenomenon, since the function of a network can easily be disrupted, causing a drastic fitness change. The rank correlation coefficient is much less affected by outliers, and can therefore detect correlation over more generations in the EA. These results indicate that there is some clear positive fitness correlation between parents and offspring in the SBMLevolver system, which indicates strong causality and explains why the EA clearly outperforms random search (Section 3.2).

What would Baldwin and Lamarck say? In the context of artificial neural networks, Yao (1997) links the idea of a two-level evolution to the Baldwin effect. However, I feel that this analogy is flawed, because the two-level adaptation described is different from Baldwin's idea. Baldwin proposed that a learning capability can be of evolutionary advantage, i.e. provides higher fitness. In contrast, the “learning” (parameter fitting) in the two-level EA happens only once, and subsequent parameter fitting in offspring is done on top of that. Note, however, that the SBMLevolver approach does favour network models that are suitable for parameter fitting, which can be related to the Baldwin effect. If one really wanted to relate this algorithmic approach to a biological concept, the better analogy would be Lamarckian inheritance: Traits acquired during the lifetime of an individual (parameters learned in the fitting process) are passed down to the offspring (see below). However, I feel that it is equally correct to regard both structural and parameter changes in this approach as evolution and not to speak of “learning” at all.

2.2.6 Parameter fitting

Cumulative round of CMA-ES As mentioned above, the parameter fitting loop is realised using the CMA-ES algorithm, and more specifically using an adapted version of the C implementation by Nikolas Hansen[1]. The technique of Exponentially Scaled Search Steps (ESSS), presented in Section 2.4, was added to the code, as well as handling of parameter boundaries via sticky boundary conditions. In contrast to a standalone run of CMA-ES, limited computational resources restrict the population size and number of generations used in the parameter fitting loop. As a partial compensation for this effect, the parameter values of a network are inherited by its descendants under topological mutations (to the possible extend), in the hope that they might provide more favourable initial solution for the next round of CMA-ES than a random initial solution.

An alternative: extensive parameter fitting and a fitness database It has to be noted that this procedure might introduce a bias: a network topology that appears twice in the course of evolution is likely to get a better fitness values for the later appearance, as the parameter values are potentially more adapted by that time. If one wanted to evaluate fitness only once for each structure, several starts of the CMA-ES algorithm would be needed in order to reduce the variation in the result of parameter fitting. Since this type of fitness evaluation would be very costly, fitness values for the evaluated structures should be stored in a graph-indexed database (Niehaus et al., 2007) using a canonical graph representation to avoid isomorphism problems (McKay, 1981; Stagge and Igel, 2001). However, I still consider the computational resources for this approach to be much larger, because the bad model structures would consume a large amount of parameter fitting time, i.e. the approach is not dynamic in recognising inapt model structures.

[1] downloaded from `http://www.lri.fr/~hansen/cmaes_inmatlab.html`.

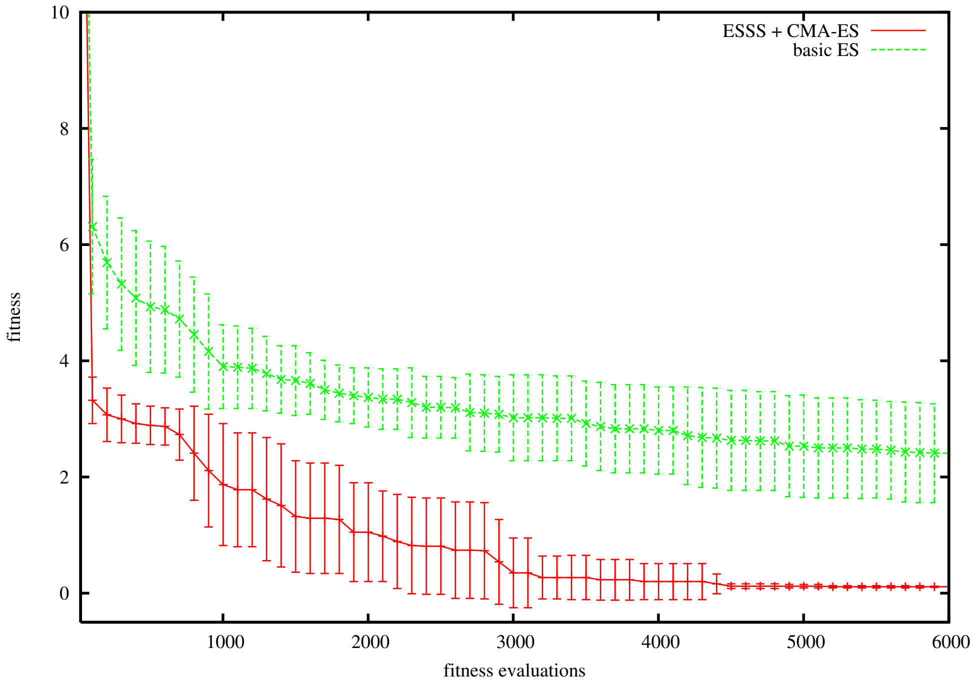


Figure 2.4: Improvement of the parameter fitter with CMA and ESSS compared to the one with a basic ES using one strategy parameter per model parameter. The example problem used here is parameter fitting for conditional learning (see Section 2.5). Each line is the mean of 15 runs with the CMA-ES parameter fitter (red) and the simple one (green). The error bars show standard deviation. Starting network is the handmade conditional learning network by Fernando et al. (2009) (for initial parameters, see (Beck, 2008)). The initial step size is 0.05 times the parameter range. Whole population (structural level) $\mu = 10$, $\lambda = 10$, no structural mutations are allowed; parameter fitter: (2+2)-ES, 10 generations. Only the best individual of every 10th generation is plotted.

> **Note:** In earlier versions of the SBMLevolver, and especially in the version used to conduct the experiments summarised in Figures 3.1, 3.2, 3.3, and 3.9, a simpler version of ES - using one strategy parameter per model parameter - was implemented for parameter fitting. It was replaced by CMA-ES with ESSS, because experiments (Figure 2.4, 2.5) showed a significant performance gain of the latter method.

2.2.7 Fitness evaluation and selection

Fitness evaluation in the algorithm is done by integrating the ODE system resulting from an individual model using the SBML ODE Solver Library Machne et al.

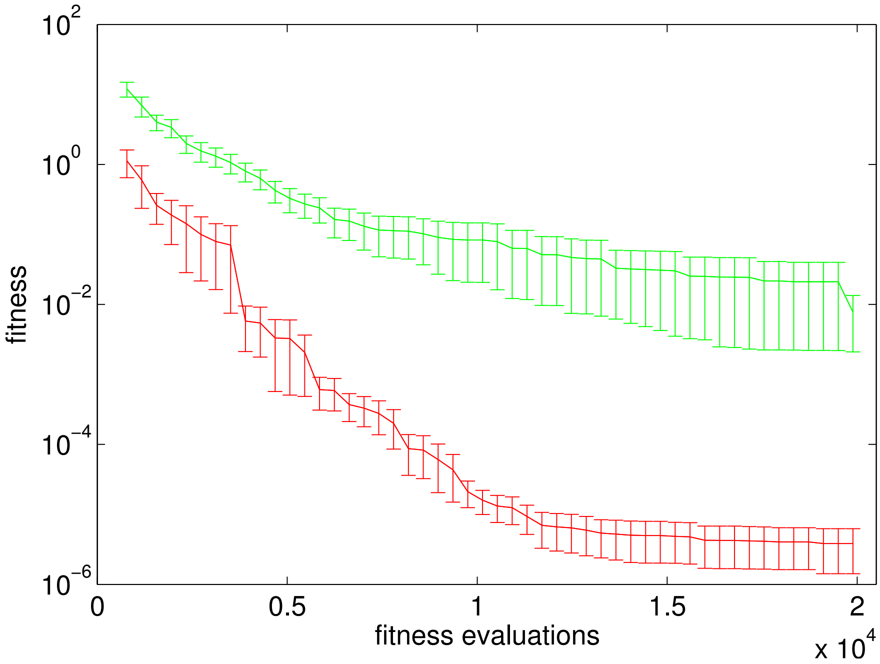


Figure 2.5: Improvement of the parameter fitter with CMA and ESSS compared to the one with a the basic ES using one strategy parameter per model parameter and model. The example problem used here is the simple addition of two numbers (note the log-scale of the fitness values, nearly all runs were able to find a perfect solution). Each line is the mean of 50 runs with the CMA-ES parameter fitter (red) and the simple one (green). The error bars show standard error. Whole population (structural level) $\mu = 10$, $\lambda = 10$, 50 generations, structural mutations are allowed; parameter fitter: (3,6)-ES, 6 generations.

(2006). The resulting multidimensional time series is then compared to a target, and the weighted quadratic difference

$$f \quad = \quad \frac{1}{C}\sum_{c=1}^{C}\frac{1}{S}\sum_{i=1}^{S}\sum_{j=1}^{N}(x_{c,i}(t_j) - y_{c,i}(t_j))^2$$

between the resulting time series x and the target time series y defines the fitness. Here, $i = 1, \ldots, S$ runs over the set of evaluated species, $c = 1, \ldots, C$ runs over the fitness cases, and $j = 1, \ldots, N$ runs over the evaluation timesteps, which do not have to start from $t = 0$. Thus, fitness values are minimised, 0 being the absolute lower bound. If a steady state value is regarded as the desired result of a network simulation, a constant time series is the target and the first few timesteps are discarded.

Selection at the structural level is elitist, with a certain percentage of the population surviving to the next generation, which is filled by offspring of survivors, generated by mutation and recombination.

2.3 Optimal algorithmic parameter settings

> **Note:** I include here the results from the Diploma thesis by Hendrik Rohn (2008), which I supervised. All results in this section are his, not mine. Motivation, outline and supervision are my contributions to this work.

2.3.1 Experimental setup

The SBMLevolver algorithm is controlled by 17 parameters, such as population size, number of offspring, etc. To find optimal settings for these parameters, Sequential Parameter Optimisation (Bartz-Beielstein et al., 2005) was employed. To asses the effect of the parameters, we used a performance measure calculated from the best fitness reached and the speed of evolution, which is estimated by averaging the best fitness values of all generations. Problem settings were taken from different problem domains:

1. steady-state computation of a a polynomial of degree seven (analog to the third-root and logarithm computation above) ("Polynomial (SS)")
2. a Boolean parity function with three inputs ("Parity")
3. time-course approximation of a polynomial ("Polynomial")
4. time-course approximations of the behaviour of a given biological network ("Legewie06").

2.3.2 Results

The results show that with the given implementation and limited computation time, only one population instead of parallel ones should be used. This might be an effect of the limited run-time granted to the EA and might change for longer runs. For the population parameters on the structural level, we find optimal values of $\mu \approx 200$ and $\lambda \leq 50$ for a $(\mu + \lambda)$ strategy with overlapping generations, and $\mu \approx 200$, $\lambda \geq 450$ for a (μ, λ) strategy with non-overlapping generations.

Against expectations, the simple crossover implemented in the SBMLevolver did not diminish performance, but was rather able to improve it. However, both crossover probability and the maximal number of mutations in one offspring creation only have

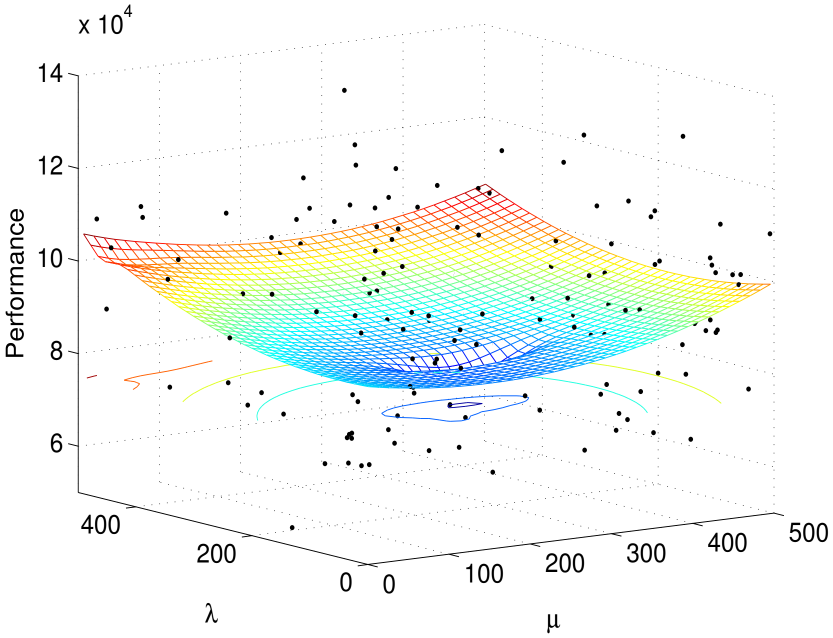


Figure 2.6: Example results for parameters μ and λ, for a plus-strategy and the problem "Legewie06".

a very slight effect on performance. The investigation of the relative mutation probabilities gave only very noisy and inconsistent results.

Optimal settings for the parameter fitting loop are more clear, and the influence of these parameters on performance is quite large: population size should be five or less, number of offspring should be between 5 and 50. The parameter fitter should be run for 10-100 generations.

2.3.3 Discussion

Due to the large computational demand of a single run of the SBMLevolver, not many repetitions could be carried out for each sample point evaluation in parameter space. However, EAs are stochastic search algorithms, so measuring their performance is an extremely noisy business. Additionally, it seems that most parameters only have small effect on system performance, which is very hard to detect in such a noisy setting. Therefore, all diagnosis except for the number of populations (use only one), population size and number of offspring, are very uncertain and are not to be blindly trusted.

Parameter	Polynomial (SS)	Parity	Polynomial	Legewie06
	——— Plus - Strategy ———			
Structural optimisation				
Populations size μ	100-200	150-450	250-350	100-300
Number of offspring λ	<50	<100	<50, >450	50-300
Crossover probability	0.2-0.6	0.4-0.6	0.8-0.99	0.9-0.99
Number of mutations	2-3	5	6-9	3-5
Mutation probabilities				
Add species	0.99	0.01-0.3	0.99	0.01
Delete species	0.2-0.4	0.01-0.1	0.99	0.01
Add species to reaction	0.99	0.9-0.99	0.01	0.3-0.5
Delete species fr. reaction	0.7-0.99	0.9-0.99	0.7-0.8	0.7
Add reaction	0.99	0.01-0.25	0.9-0.99	0.3-0.99
Delete reaction	0.6-0.8	0.2-0.4	0.6-0.7	0.3-0.8
Species duplication	0.01	0.8-0.99	0.1-0.99	0.8-0.99
Parameter fitting				
Population size μ_{PF}	<10	5	15	<5
Number of offspring λ_{PF}	<5, >45	20-25	>45	<5
Generations PF	80-100	<50	15	10 -100
	——— Comma - Strategy ———			
Structural optimisation				
Population size μ	300-400	170-250	>30	200-250
Number of offspring λ	<100	>450	>350	>450
Crossover probability	0.9-0.99	0.01-0.5	0.1-0.3	0.3-0.6
Number of mutations	<5	5-6	2-3	3-6

Table 2.2: Overview of parameter values that led to good performance. Values were quantified by visual inspection of the parameter-performance landscapes, see e.g. Figure 2.6 and and Figure 2.7.

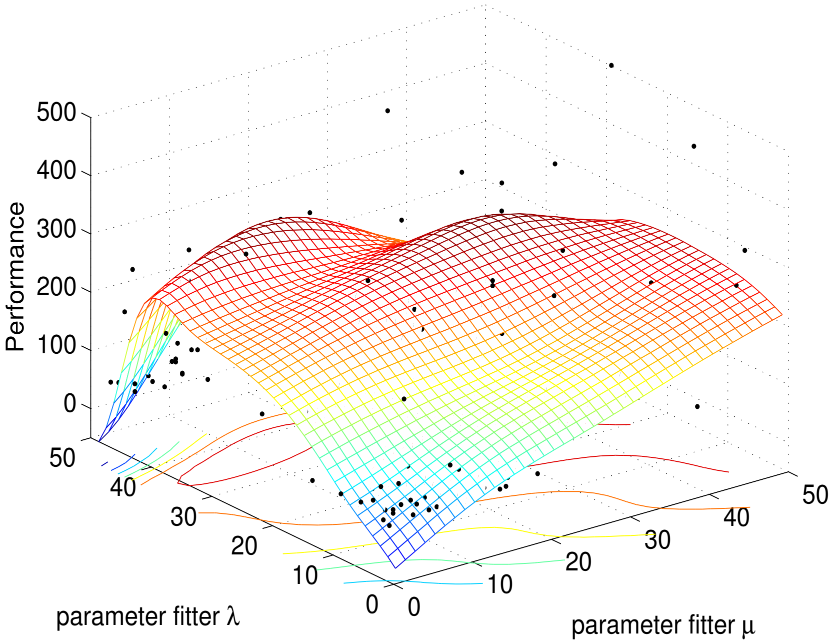


Figure 2.7: Example results for population parameters μ_{PF} and λ_{PF} of the CMA-ES algorithm, for the problem "Polynomial (Steady State)".

Hendrik Rohn proposed and tested five hypotheses to explain the insufficient quality of some of the results:

1. Not enough repetitions per experiment.

2. Not enough computation time per run.

3. Presentation of measured performance values is difficult, since the parameter-performance space is usually of dimension > 3. Thus, more than the two displayed parameters play a role in the performance of a sampling point, making the sampled points in one picture effectively incomparable.

4. Interpolation by kringing method does not work sufficiently well, especially at the borders of the sampled region.

5. Many parameters simply have very little influence on performance.

For a detailed presentation of the results and a discussion of their associated difficulties, see the excellent thesis by Rohn (2008).

2.4 Exponentially scaled search steps in parameter fitting

> **Note:** The work reported on in this section arose from the Diploma thesis of Hendrik Rohn, supervised by the author. It was published in Rohn et al. (2008).

2.4.1 Motivation

In a multidimensional parameter fitting problem, the optimal parameters (and thus the parameter values at runtime) will differ by orders of magnitude (see for example Ibrahim et al., 2009; Marwan, 2003; Tyson, 1991). While many modern numerical optimisation algorithms utilise techniques to adapt the search steps for each dimension individually, the question arises whether prior knowledge of the magnitude of parameters can be used to improve their search behaviour. In the case of biochemical network models, different parameters may play different roles (rate constants, saturation constants, etc.), each with their own typical range of values.

Here, we introduce a technique called exponentially scaled search steps (ESSS), which adapts the search steps automatically to the assumed parameter range. We extensively investigate the influence of exponentially scaled search steps on the performance of two evolutionary and one deterministic optimisation technique; namely CMA-Evolution Strategy (CMA-ES, Hansen and Kern, 2004), Differential Evolution (DE, Storn and Price, 1997), and the Hooke-Jeeves algorithm (HJ, Hooke and Jeeves, 1961), respectively. As test cases, we use 12 different models from the BioModels database (Le Novère et al., 2006). For details on the experimental setup, see the corresponding publication by Rohn et al. (2008). We find that in most test cases, exponential scaling of search steps significantly improves the search performance for all three methods.

2.4.2 Description and implementation of ESSS

We compared the results of parameter optimisation on models with unscaled and scaled parameters. By scaling the parameters, the search steps were automatically scaled as well (see below). In the unscaled case, the optimisation algorithms allow parameter values p_i in the range between 10^{-5} to 10^{5}. In the scaled case, a hidden parameter $\tilde{p_i}$ with values between 0 and 1 was used to compute the original parameter p_i according to

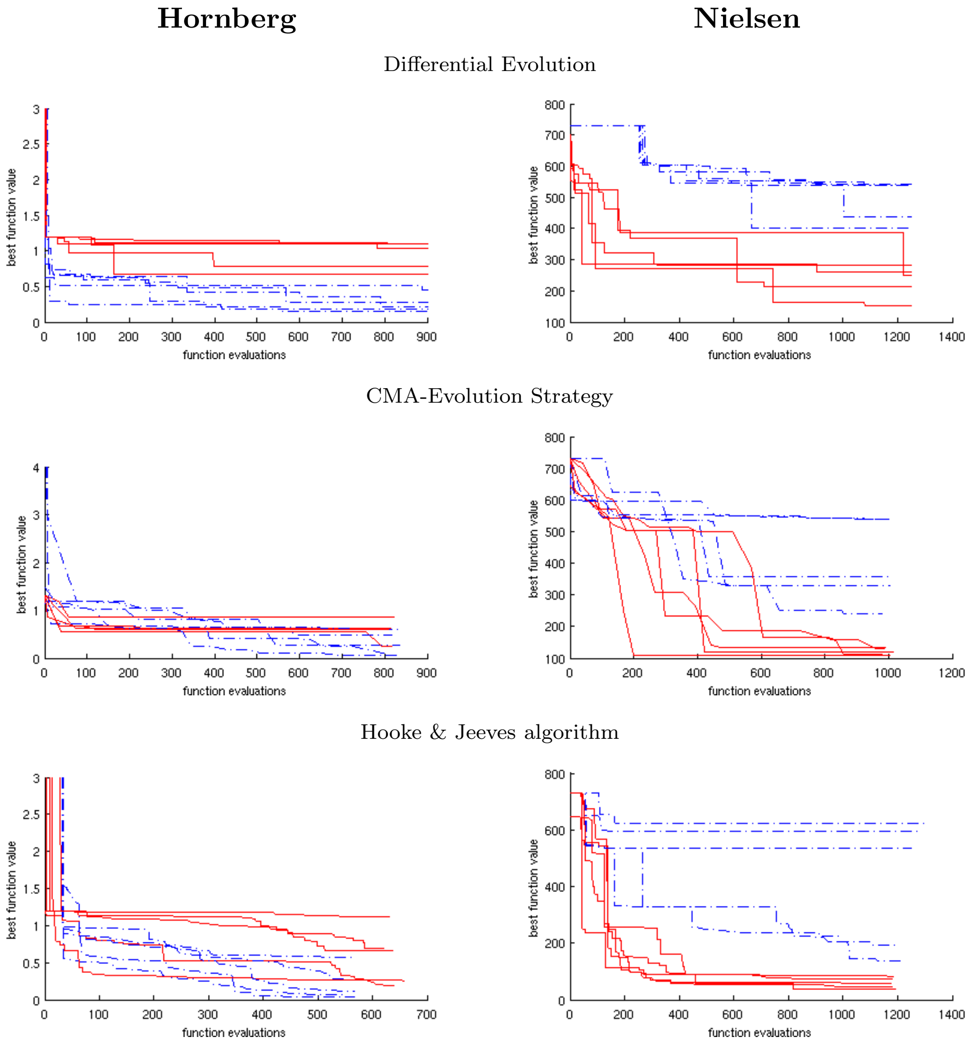
$$p_i = 10^{-5+10\tilde{p_i}}.$$

Optimisation was then carried out on the hidden parameters $\tilde{p_i}$, resulting in search steps that are larger for high values and smaller for low values of the original parameters p_i, independent of the optimisation procedure. This allows an investigation of small parameter values with high resolution and large parameter values with low resolution even without an additional adaptation mechanism for the search step-size.

2.4.3 Results and discussion

For each network, each optimisation algorithm was run 50 times, and the quality of the solution together with the speed of reaching it was recorded. We define the quality as the value of the objective function at the end of the run, while we take the average value of the objective function over the whole run (except for the initial phase which depends on the starting points) as a measure of speed. Figure 2.8 contains six example runs, and Table 2.3 shows a compact overview of the results. (see Rohn et al. (2008) for details including the measured values for quality and speed).

The experimental results (Table 2.3, Table 2 in Rohn et al. (2008)) clearly confirm an advantage of using scaled search steps, in particular when the search interval for the parameters is large. When the interval to be searched is not too large, our results show that exponential scaling at least does not deteriorate performance in most cases. Some test-cases show only a small difference (Fisher, Fung, Lenser) and a few give worse results (Hornberg, Huang). All other test-cases show an improvement in speed and/or solution quality for the optimisation procedure. The reason for the different effects of scaled search steps still remains to be examined. One likely point is that the behaviour of some networks might be quite sensitive to changes in the parameters with large values, in which case the scaling may prevent adequate fine-tuning of these parameters. However, it also has to be noted that in cases where performance is worse, this effect is not too pronounced, giving more weight to the partially drastic performance increase in most other cases.



Hornberg

Nielsen

Differential Evolution

CMA-Evolution Strategy

Hooke & Jeeves algorithm

Figure 2.8: Typical runs of six exemplary fitting cases of the networks by Hornberg and Nielsen, optimised by all three algorithms; the dashed lines indicate the unscaled approach, solid lines show the scaled one.

	—DE—		—ES—		—HJ—	
Network	better	faster	better	faster	better	faster
Fisher et al. (2006)	-	-	(-)	(-)	(+)	+
Fung et al. (2005)	-	+	(+)	(+)	+	+
Hornberg et al. (2005)	-	-	-	-	-	-
Huang and Ferrell (1996)	-	-	-	-	-	-
Kofahl and Klipp (2004)	+	+	+	+	+	+
Kongas and van Beek (2001)	(+)	+	+	+	+	+
Martins and Boekel (2003)	+	+	+	+	+	(+)
Marwan (2003)	+	+	+	+	+	+
Nielsen et al. (1998)	+	+	+	+	+	+
Lenser et al. (2007)	-	(-)	-	(-)	(-)	(-)
Tyson (1991)	+	+	+	+	-	(+)
Yildirim and Mackey (2003)	+	+	+	+	+	+

Table 2.3: Advantage (plus) or disadvantage (minus) of the scaled approach in comparison to the unscaled approach. Brackets indicate a non-significant result according to the Mann-Whitney-Wilcoxon-Test Wilcoxon (1945)with $P \geq 10^{-3}$. Based on 50 independent repetitions for each test-case. DE = Differential Evolution, ES = Evolution Strategy, HJ = Hooke-Jeeves. Columns "better" indicate quality of the fit, "faster" indicate speed.

2.5 Application: a chemical network with learning capability

> **Note:** This section gives a brief summary of the Diploma thesis by Christian Beck, which I supervised. All results in this section are his work, not mine.

2.5.1 Introduction

This case study aims at evolving and testing a gene regulatory network that is capable of classical conditioning, a basic form of learning. Here, the network learns the correlation of different, simultaneously occurring external factors and adapts its behaviour accordingly. Although conditioning is very common in multi-cellular organisms, there is no evidence so far for evolution of any learning systems in prokaryotes. The endowment of bacteria like *E. coli* with learning could open up new possibilities in several scientific fields such as medicine, technology, and biology.

In 2003, Thomas Knight presented a simple way to equip living organisms with

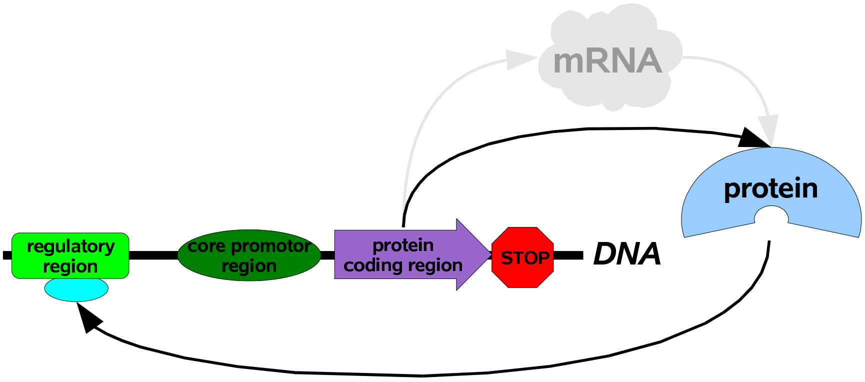


Figure 2.9: General structure of a gene with its promoter terminator. The picture also shows the transcription into messenger RNA and the translation into a protein. These two steps are greyed due to the fact that in this thesis all kinds of post-transcriptional regulation and modification are ignored. The protein can act as a regulator for the expression of its own or another gene by binding to the promoter region.

additional features by using standardised DNA sequences called BioBricks (Shetty et al., 2008). In this case study, models of these BioBricks were used as building blocks for the evolution of learning networks. Their mathematical representation is based on Hill kinetics, where the production of a protein y is regulated by the concentration of transcription factor (TF) x as follows:

$$\frac{dy}{dt} = \alpha * \frac{x^n}{K_d + x^n}.$$

While this is the formula for positive regulation by a single TF, negative regulation and co-regulation by two TFs were also considered here. Notably, the evolutionary process focuses on realistic networks by exclusively employing gene regulatory mechanisms that can be found in nature.

2.5.2 Results and discussion

One of the most complicated parts of this work turned out to be the description of the fitness function, which needed to be general enough to characterise the learning phenomenon, but specific enough to provide the EA with the information to evolve the network. In the end, a weighted combination of the desired response to different stimuli was used. For details, the reader is referred to Beck (2008).

Parameter fitting of a Hebbian learning network In a first step, the parameters of a hand-crafted reaction network were fine-tuned for the learning task. The network itself was inspired as a synthetic implementation of the learning task using specific biological modules. For details, see the publication by Fernando et al. (2009), which contains this part of the results. Repeating the evolutionary procedure, we learned that many qualitatively different parameter sets led to the same desired behaviour. Using a principle component analysis, we were able to show that the ensemble of evolved (28-dimensional) “working” parameter sets were determined by eight underlying parameters, and how the major principal components related to the parameters of the network.

Structural evolution of learning networks In a second step, the structure and the parameters of a conditional learning network were to be evolved. For this, the fitness function had to be adapted. Christian Beck (2008) was able to evolve both simple conditional learning networks and Hebbian learning networks. Figure 2.10 shows a network that can perform conditional learning, but not of the Hebbian type. It has to be noted that the structural optimisation task was a lot more difficult than pure parameter optimisation of a specialised structure, and most runs ended in local optima not showing the desired behaviour. Beck (2008) also analyses the typical local optima in the fitness landscape.

Outlook: a potential wetlab implementation One goal of this case study was to evolve realistic networks that could in principle be transferred into biological cells using the BioBricks approach. Figure 2.11 shows such an implementation of a Hebbian learning network. It would be very interesting and informative to see this network implemented *in vivo*, but this was not in the scope of this work. For the future, it would be interesting to extend the SBMLevolver in such a way that the evolved gene regulatory networks are completely built up of these standard biological parts. This would make the software a reasonable tool for synthetic biologists to develop networks with complex behaviour that can easily be transferred into living organisms.

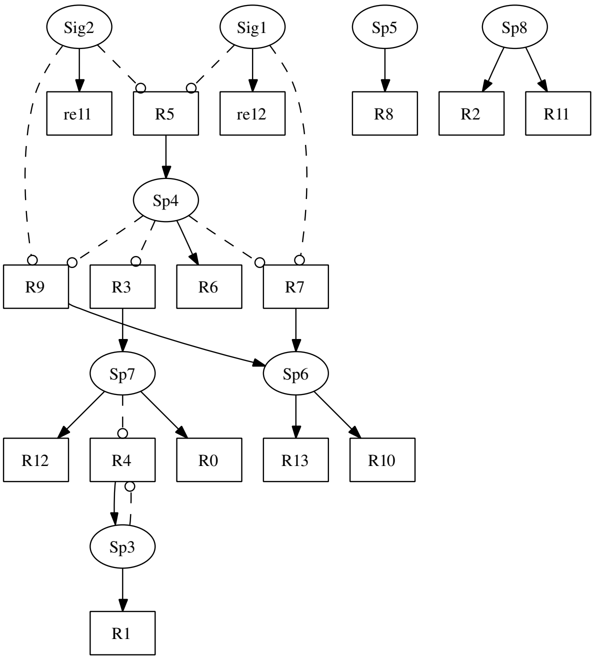


Figure 2.10: Structure of a conditional learning network. The species Sp6 is the output species that represents the learning. It is activated by the unconditioned stimulus (Sig1) through reaction R7. This reaction is so fine tuned that even the smallest amount of Sp4 is enough to produce Sp6. Reaction R9 on the other hand needs a higher concentration of Sp4 to get activated. Thus before pairing, the concentration of Sp4 is too small to produce Sp6 together with Sig2. The activation of both signal species produces a higher amount of Sp4 and so R9 produces Sp6 while Sig2 is active. Sp4 can be seen as a species which memorised the pairing of both stimuli. Although this network shows conditional learning, this is no Hebbian learning. For details and behaviour of the network, see the thesis by Beck (2008).

2.6 Concluding remarks

In this chapter, I introduced an evolutionary algorithm design and its implementation, aimed at the artificial evolution of networks in molecular biology. The problem is hard to tackle, since both the structural evolution of a network and the fitting of kinetic parameters provide enough difficulties to require specialised algorithmic solutions. In combining those solutions, an algorithm was created that is specifically tailored for the problem at hand. Several rounds of parameter fitting after each structural mutation are used to soften the effect of these structural mutations and to provide a smoother path through the search space of network structures. To provide a quantitative backing

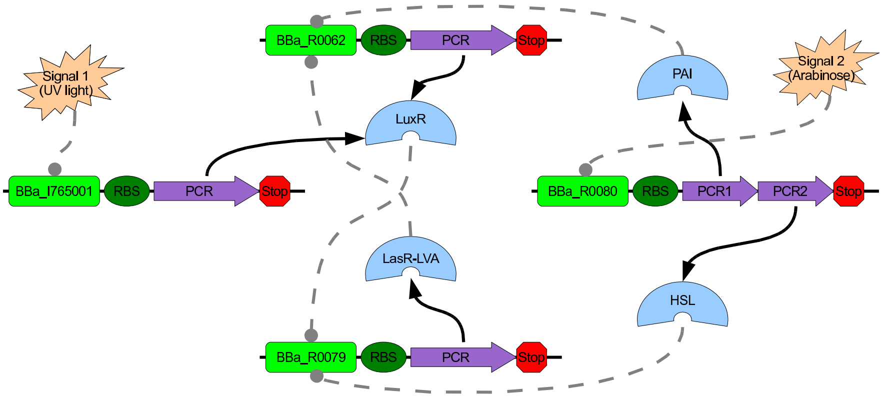


Figure 2.11: The depicted network shows one theoretical possibility how an evolved Hebbian learning network could be transferred into living cells using BioBricks. For easier handling, ultra-violet light and the monosaccharide arabinose were used as signals. The production of output species LuxR is directly regulated by signal one and also by LasR - which is the weight species - and PAI. PAI and HSL are directly produced by signal two. Signal two (via HSL) activates the production of LasR in the presence of LuxR. For a slower decay, the weight species LasR is tagged with LVA. The regulatory regions are labeled with their identification numbers in the BioBrick database. For details see Beck (2008).

for the claim that a nested EA approach helps in evolving cellular networks, Chapter 3 provides extensive comparisons between the two-level setup and two one-level variants.

This chapter also presented a few ideas of how to formalise the strong causality requirement of EAs, and how to measure the strength of causality in an EA system. Because of time constraints, these ideas have not yet been consequently applied to the SBMLevolver software system. Even though they have been used as guidelines in the design of the two-level approach, a quantitative and detailed investigation of these aspects is very likely to reveal further insights into the work of this evolutionary system.

3
Evaluation of the SBMLevolver Approach

Chapter summary. *Evaluating the approach to network evolution presented in the last chapter, I show that a separation of model-structure evolution from parameter-fitting helps to prevent premature structural convergence, presumably by strengthening causality and thus smoothing the path of the population through the search space. This justifies the two-level approach to network evolution implemented in the SBMLevolver software. Comparisons with random search on the network level and with pure parameter fitting of fully connected networks reveal that the combination of structural and parameter evolution is indeed advantageous. Further results indicate clearly that the deletion of network parts is less likely to have a large effect on network behaviour than the addition of new parts. The chapter is concluded by a case study on the evolutionary improvement of a cell cycle spindle checkpoint model.*

3. EVALUATION OF THE SBMLEVOLVER APPROACH

Chapter contents

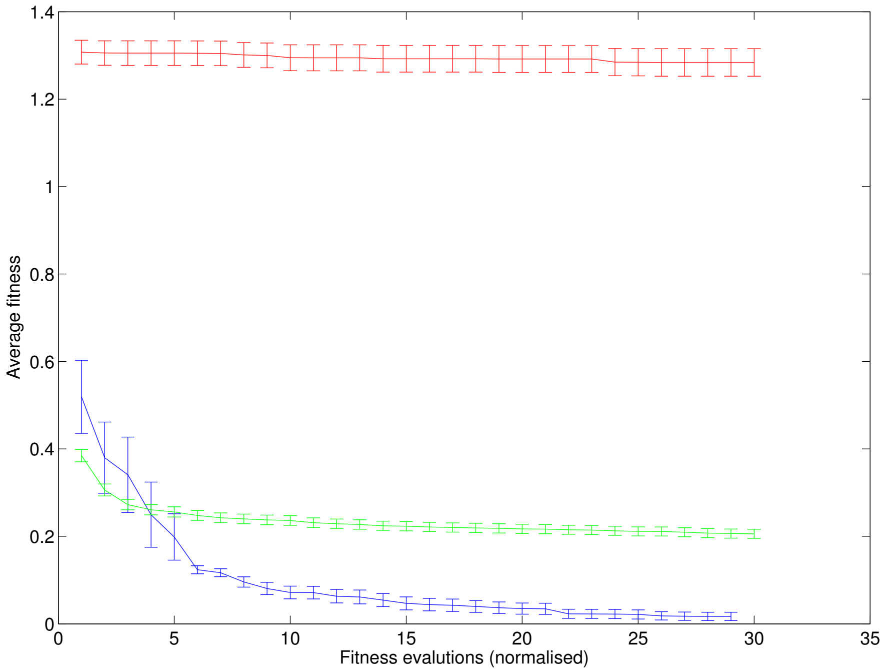


Figure 3.1: Average best-fitness with standard error over 10 (blue) and 100 (red, green) (see note) runs of the evolution of logarithm networks. Blue: two-level EA with structural (25+25) loop for 29 generations and (3,10) parameter fitter loop for 99 generations; Red: one-level EA with (25+25) loop for 29999 generations (note the different scale); Green: one-level EA with (2500+2500) loop and 299 generations. Fitness evaluations are given in thousands.

3.1 Evaluation of the two-level approach

3.1.1 Experimental setup

Arithmetic calculations as test functions In order to test the effect of the separation between structural evolution and parameter fitting (see Chapter 2) on performance, we evolved networks supposed to perform two tasks: calculating the third root and logarithm of a positive real number. Here, “calculating” means that the input is set as initial concentration of species *input*, while the output is read from the steady state concentration of species *output*. Therefore, the target time series for the *output* species is simply the desired output value, constant over a period of time, where the first few timepoints are excluded from the fitness evaluation. An example solution for a third root network is shown in Figure **??**. While the third root has been observed to

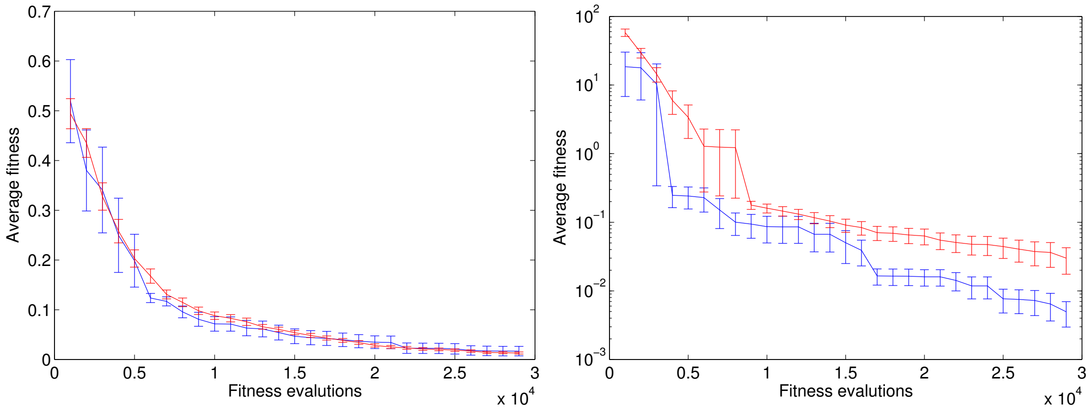


Figure 3.2: "Accidental" result: A slight change in the ratio between structural and parameter mutations does not change the results qualitatively (see note). Blue line: Correct two-level setting with 10 repetitions; Red line: Incorrect two-level setting (with (5,15)-ES instead of (3,10) for the parameter fitting loop) and 100 repetitions. Left: logarithm task; Right: third root task. Error bars denote standard error.

be solvable but substantially more difficult than a square root network (Deckard and Sauro, 2004), no precise solution to the logarithmic problem is known yet. Therefore, the best possible approximation to the logarithm is sought. In this work, the main focus is not put on the evolved networks, but rather on the evolutionary process.

Two-level vs. one-level approaches To compare the effectiveness of the two-level approach, we designed two one-level settings, where the structural and parameter mutations were used together in one evolutionary loop. Thus, three different strategies were used:

1. Two-level evolution using ES for local fitting (upper level: (25+25)-elitist selection, 29 generations, only structural mutations; lower level: (5,10)-ES, 99 generations, only parameter mutations)

2. One-level evolution running for more generations ((25+25)-elitist selection, 29999 generations, structural and parameter mutations)

3. One-level evolution employing a larger population ((2500+2500)-elitist selection, 299 generations, structural and parameter mutations)

The parameter settings were chosen such that the number of fitness evaluations and the ratio of structural vs. parameter mutations are identical, enabling an objective

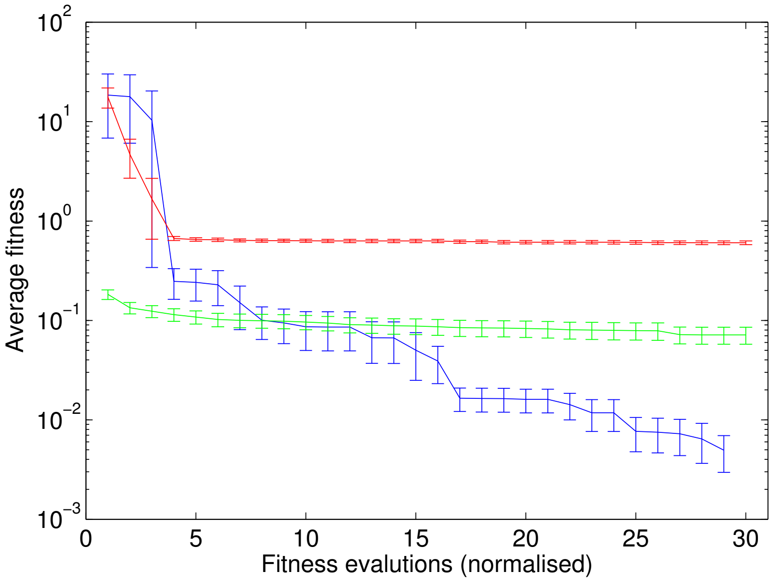


Figure 3.3: Average best-fitness with standard error over 10 (left) and 100 (middle, right) runs of the evolution of third root networks. Left: two-level EA with structural (25+25) loop for 29 generations and (3,10) parameter fitter loop for 99 generations (note different scale), middle: one-level EA with (25+25) loop for 29999 generations (note the different scale), right: one-level EA with (2500+2500) loop and 299 generations.

comparison. The one-level strategies invested the saved fitness evaluations into more generations (2) or more individuals (3). In the ES, adaptive stepsizes were disabled to make the results comparable. Computations were carried out as single-processor runs on a cluster of workstations equipped with two Dual Core AMD Opteron(tm) 270 processors running Rocks Linux.

> **Note:** In the original experiments published in (Lenser et al., 2007), there was an error in the experimental design of the two-level strategy, with a (5,15)-ES for local parameter adaptation, instead of the (3,10)-ES correctly used. This resulted in a larger number of fitness evaluations and a different ratio of structural vs. parameter mutations. When the error was discovered, the computational resources needed for a repetition of the experiment were not available any more. Therefore, it was repeated with only 10 repetitions instead of 100. The results from the correct and incorrect setting are compared in Figure 3.2: the difference is negligible, thus the exact settings do not significantly influence the quality of the EA.

3.1.2 Results

Two-level approach yields significant advantages Results of the evolution of a logarithm-network (Figure 3.1) show that the two-level structure of the algorithm improves fitness development drastically in comparison to a larger number of generations, while it prevents the premature convergence seen with a larger population. A large population seems to enable the algorithm to guess a good initial network, but it is unable to improve upon this. In contrast, the two-level approach improves the network continuously, yielding significantly better results in the end.

Figure 3.3 shows the average fitness development for the third root task. Results are similar to Figure 3.1, although not as pronounced. Again, the two-level strategy drastically outperforms the setting with more generations, while its initial progress is slower than for a large population. However, the large-population approach converges too early, while the two-level setting continues to improve in a smooth fashion. In this task, the networks were also required to be mass-conserving, i.e. it was demanded that a feasible configuration of molecular masses for the different species exists. This constraint might explain the slower rate of convergence in comparison to the logarithm-trials.

More successful mutations produce larger networks Since a separation of structural evolution from parameter fitting leaves newly created topologies time to adapt their parameters before they are evaluated, the probability for a successful structural mutation is higher. Moreover, the structural evolution starts from small networks, so that in the two-level setting, larger network solutions are expected. Indeed, this is the case: For the logarithm task, the networks resulting from the two-level setting have 6.9 ± 0.03 species on average, whereas for the one-level settings, networks with 3.0 ± 0.002 and 3.1 ± 0.0001 species on average evolve. The picture is similar for the number of reactions: 7.5 vs. 3.0 and 3.1, respectively (the standard error is always relatively small and thus omitted). For the third root task, the networks are of comparable size. While this increase in network size for the two-level approach helps to explain its success, it certainly comes with increased computation times for this approach. Unfortunately, runtimes were not directly measured at the time the experiments were carried out.

3.2 Comparison with random search

A fair comparison must not sample the whole search space As a control test, a random search (RS) with the same parameters was performed, replacing mutations in the structural evolutionary algorithm with creation of new random individuals. For this, it has to be considered that the search space is vastly larger than the space actually explored by the EA (networks were allowed up to 100 species and reactions, but all successful models had 4-9 species and 2-19 reactions). Thus, a random search sampling the whole search space would not be a fair comparison to the EA trials. To restrict the random search to the utilised search area, the size of random individuals was determined by probability distributions gained from the final populations of the two-level EA, separately for both tasks. The slight advantage that sampling from sizes of the final population lends to RS only strengthens the conclusion that the EA outperforms RS.

Parameter inheritance in random search As discussed in the description of the algorithm, the SBMLevolver approach uses inheritance of parameter settings from a parental model structure to its offspring, in order to provide the best possible starting values for the new round of parameter fitting associated with the creation of a new structure. While it is impossible to apply this direct inheritance to the random search algorithm (since the offspring bears no structural similarity to its parent), the seemingly best possible approach was to randomly select the offspring's parameters from the parental parameter values. This at least allowed an adaptation of the magnitude of the parameters, e.g. if a task requires parameters in the 10^{-2} range, the parameter fitting loop did not have to tune to that range over and over again.

RS cannot narrow its search While the development of the best fitness is similar for the SBMLevolver EA and structural RS in the initial phase of the run, the EA drastically outperforms RS in finding a precise solution, resulting in fitness values one order smaller after 750000 fitness evaluations (Figure 3.4). The random search approach clearly lacks the ability to narrow its search and fine-tune the network for the desired calculation. However, it is not totally clear how much of this advantage is due to the structural similarity between parent and offspring, and how much is caused by the parameter inheritance, which leads to a better adaptation of the kinetic parameters for the EA.

3.3 Comparison with pure parameter fitting

3.3.1 Motivation

SBMLevolver as a CMA-ES with macromutations In comparison to random search on the network level, pure parameter fitting of a fully connected network model lies at the other end of the spectrum of possible model fitting approaches. In the SBMLevolver approach, topological mutations and the CMA-ES parameter fitting algorithm are heavily interwoven. In effect, one could argue that the SBMLevolver represents a variant of CMA-ES, augmented with macro-mutations on the system structure.

From an a-priori point of view, such a variation of the CMA-ES idea has potential benefits and drawbacks. On the positive side, the SBMLevolver does not need to fit a full system containing all possible reactions, but can limit itself to smaller sub-systems. This leads to drastically decreased computation times (see below). On the negative side, the idea of macro-mutations is opposed to the design of Evolution Strategies, since they can violate strong causality (Section 2.2.1) and cannot be arbitrarily scaled to accommodate the search situation. These different effects of macro-mutations motivate a more thorough investigation of the SBMLevolver in comparison to pure CMA-ES on a fully connected model, by looking at a gradual transition between the two approaches.

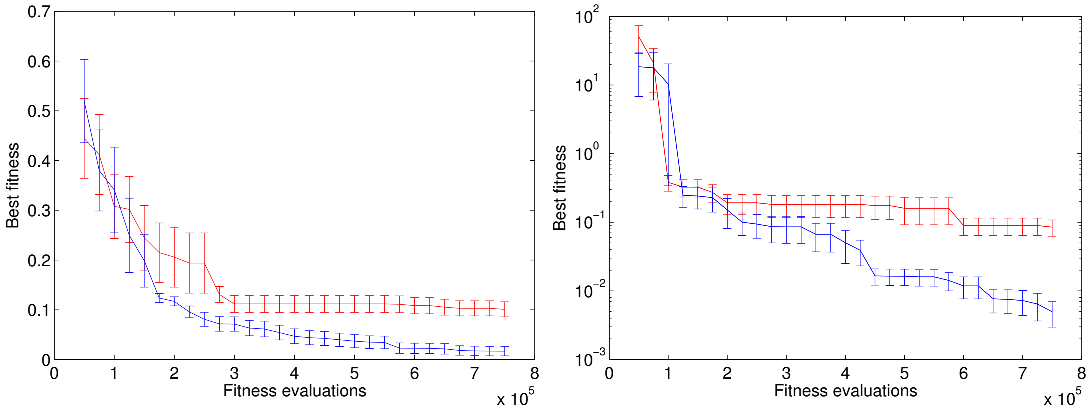


Figure 3.4: Comparison of two-level EA (blue) with random search in structures with appropriate size (red). Average best-fitness with standard error over 10 runs for the logarithm (left) and third root (right) task. Structural EA (25+25) for 29 generations, parameter fitter (3,10) for 99 generations.

3.3.2 Experimental setup

Ratio of structural vs. parameter mutations determines influence of macromutations To this end, an experimental setup was designed in which the ratio of structural mutations vs. parameter mutations could be varied, and the resulting algorithmic behaviour could be investigated. First of all, all parameters except the number of generations spend in the structural EA and in the parameter fitting CMA-ES were kept fixed: Structural evolution was done in a (30+30) setup, while the parameter fitter used a (3,6) strategy. On the structural level, the number of species was fixed to three (to facilitate comparison with a fully connected network model), and the number of reactions was allowed to vary between 0 and 100. All reactions from Table 2.1 were allowed.

Premature convergence of parameter fitting leads to differing numbers of fitness evaluations To vary the ratio of structural vs. parameter mutations, the number N_S of structural EA generations and the number N_P of parameter fitting generations were varied in such a way that the overall number of fitness evaluations stayed constant. It has to be noted that this setup does not guarantee a constant number of fitness evaluations, since the parameter fitting loop for every network of the structural loop can terminate prematurely if the numerical integration fails repeatedly or the CMA-ES reaches a fitness plateau. Indeed, the average number of fitness evaluations varies significantly between the experimental setups (see Results section).

Details on the experimental setup The following combinations of N_S and N_P were used for the experiments, sorted by a decreasing amount of structural mutations: ($N_S = 100$,$N_P = 5$), ($N_S = 50$,$N_P = 10$), ($N_S = 25$,$N_P = 20$), ($N_S = 10$,$N_P = 50$), ($N_S = 100$,$N_P = 5$), ($N_S = 1$,$N_P = 500$), and ($N_S = 0$,$N_P = 15000$). The last two settings present a special case: no structural mutations were carried out, and only parameter fitting was performed. For this, two alternatives were explored: In the first one (termed ($N_S = 1$,$N_P = 500$)), the population of the structural EA consisted of the fully connected model plus 29 randomly generated models, in the same way as the initial population of all other runs. In the second alternative (termed ($N_S = 0$,$N_P = 15000$)), only one fully connected model was fitted, which means that instead of the usual $\mu = \lambda = 30$, the population of the structural EA only consisted of the one fully

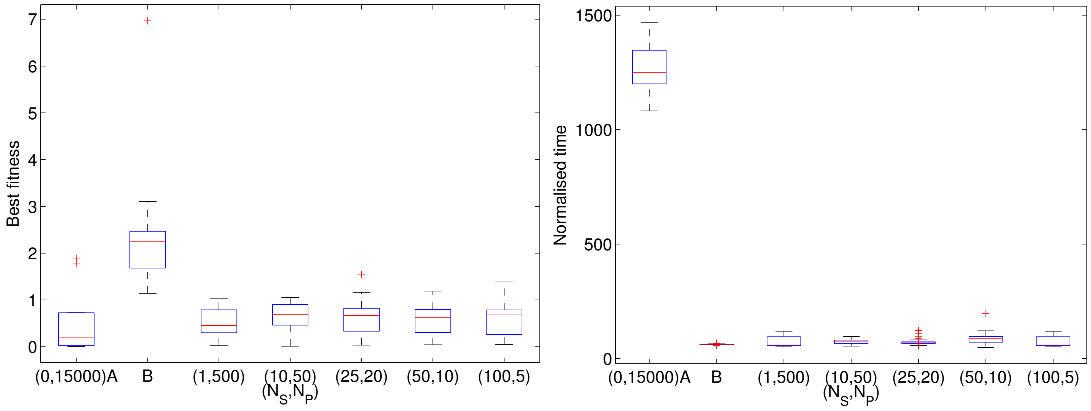


Figure 3.5: Exemplary results for the influence of the ratio between structural and parameter mutations on the evolutionary success for the logarithm task. For a given number of generations (see text, (0,15000)A and B denote the (0,15000) setting run for the whole 15000 parameter mutations (A) or for 1150 parameter mutations (B)), the best fitness reached (left) and the utilised computation time of all settings except (0,15000) do not differ significantly for the different combinations of structural EA generations (N_S) and parameter fitter generations (N_P). The (0,15000) setting, when run to the end (A), leads to a drastically increased computation time (right). When the computation time is limited (B), fitness results are inferior to the other settings (left). Since the experiments were carried on a number of differing hardware configurations, computation times were normalised by the performance of the respective machine on a given task ('normalised time').

connected model, which underwent $30 \times 500 = 15000$ generations in the parameter fitting loop.

3.3.3 Results and discussion

Ratio of structural and parameter mutations is not significant to evolutionary success Unfortunately, the results show that the influence of chance on the evolutionary success is a lot higher than that of parameters. For a given number of total fitness evaluations, the exact parameter configuration does not significantly influence either the best fitness reached or the utilised computation time, except for the special case of pure parameter fitting of a fully connected network, which is treated below (Figure 3.5 shows the results for the logarithm task, the corresponding results for the third root task are qualitatively similar). For the trails shown, only 30 repetitions were carried out, so that it may be possible to reach significant differences in the mean values for a larger number of repetitions. However, the standard deviations indicate that any such difference in the mean value would not be of consequence to individual runs, which are mainly determined by chance, not by the parameter choice.

Putting all fitness evaluations into a fully connected model is not fair Initially, the fully connected model was put into the parameter fitting procedure for the full 15000 generations (setting (0,15000)A in Figure 3.5). While this leads to a fair comparison in terms of fitness evaluations, it is highly unfair in terms of computation time, since the effort for numerical simulation of the fully connected model is about 15 times higher compared to the sparse models that typically appear in the other setups (Figure 3.5 right). For a practitioner, results should be available as fast as possible, so that computation time seems to be a more important criterion than the number of fitness evaluations. For this reason, the number of generations for fitting the full model was reduced by a factor of 15, leading to computation times that are comparable to the other optimisation setups (setting (0,15000)B in Figure 3.5). It should be noted that the fully connected model here has only three species. When larger models are considered, the above effect will increase exponentially.

With comparable computation time, fitting a full model is not competitive When the computation times are made comparable, the fitting algorithm for the fully connected model can only use around 6900 fitness evaluations, in contrast to the approximately 90000 that are used in runs with sparse models. This leads to a drastic decrease in fitness performance (Figure 3.5 left). These results clearly show that with the given time limitation, the performance of parameter fitting on a fully connected model is significantly worse compared to structural evolution with a free model structure.

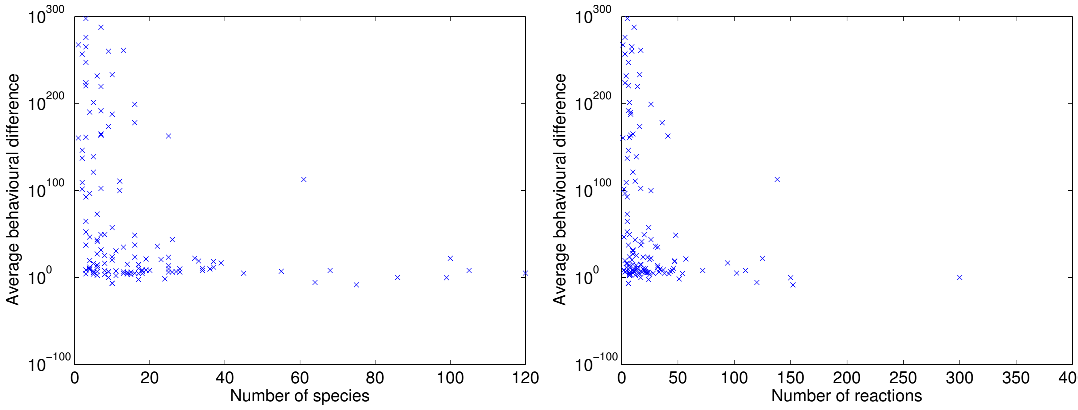


Figure 3.6: The influence of the number of species (left) and reactions (right) on the mean behavioural difference over all mutations and repetitions. The correlation is not very strong (significance values around p=0.1) but clearly visible. Note the log-scale on the y-axis.

3.4 Measuring the effect of mutation operators

3.4.1 Experimental setup

Mutation operators change network behaviour, but how much? When considering the network mutation operators designed in Section 2.2.4, an important question to ask is: "What is the expected (or average) effect of this operator when applied to a biochemical network model?" In a slightly more formal phrasing, we want to know how much the behaviour (i.e. the simulated time course) of a network changes when the operator is applied to it. To this end, I have applied the mutational operators to models taken from the BioModels database (Le Novère et al., 2006), and recorded the difference in the simulated time courses before and after the mutation.

Sampling models of different sizes and purposes At the time the experiment was conducted, the BioModels database (release 11) contained 184 curated model, which were downloaded in SBML format. The models were not modified, keeping their kinetic rate constants and initial concentrations. The models contained between 0 and 120 species and between 0 and 352 reactions, with an average of 13.3 species and 25.4 reactions.

Behavioural difference is measured as the average difference per species To compute the difference in model behaviour induced by the mutation, the time-

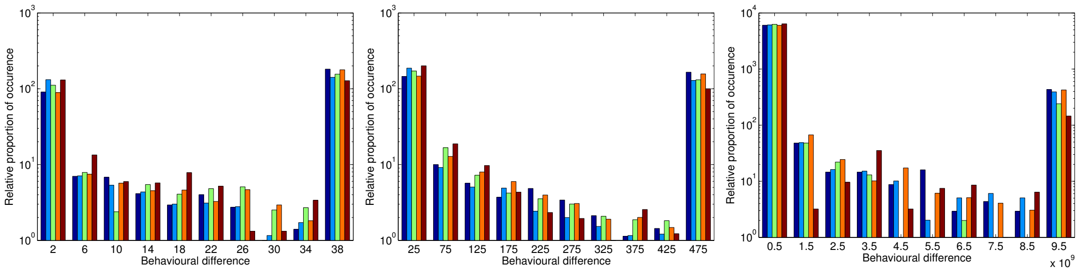


Figure 3.7: Effects of the different mutation operators on network behaviour, displayed at three different scales. Mutations are species deletion (dark blue), addition (light blue), deletion (green) and replacement (orange) of reactions, and species duplication (brown). The different scales are created by summarising all values greater than the histogram range in the rightmost bin. The rightmost bins roughly correspond to the 40%, 60% and 95% quantiles of the behavioural difference distributions.

courses before and after the mutation were computed, and the average sum-of-squares difference between corresponding species (neglecting added or deleted species) was taken as the summarising difference measure. To this end, the model was simulated once before the mutation, then the mutation was applied by the SBMLevolver software, and the mutated model was simulated again. In a large proportion of cases (∼50% for deletion of a species, ∼30% for the other mutations), the numerical integration of the mutated model failed, and the result from these runs was discarded. By design, the addition of a species does not cause a change in network behaviour (as the species is initially not connected by a reaction), and thus the experiment was only carried out for the deletion of a species, addition, deletion, and replacement of a reaction, and species duplication.

3.4.2 Results and discussion

Mutation effects vary immensely and decrease with network size The first aspect to notice about the mutation-induced behavioural differences is that they vary greatly, between difference values of 0 and 10^{299} (Figure 3.7). The very large values are caused by mutations that lead to infinitely increasing concentration values in the network, making the behavioural difference infinitely large. In many cases, the numerical integration of such models fails, but if it does succeed, large difference values are recorded. On the other hand, most mutations only induce small differences in network behaviour, with median values of 280, 18, 50, 176 and 12 for the five investigated mutations (Figure 3.7).

Comparing the size of networks (measured in number of species and reactions, respectively) with the average effect of a mutation, a very weak negative correlation can be discovered (Figure 3.6). The intuitive assumption that a mutation of a large network will lead to a relatively smaller effect seems to be correct. However, this correlation is much weaker for smaller networks, which constitute the majority of models in the investigated BioModels release.

The model, not mutation, determines behavioural difference Looking at the whole collection of behavioural difference measurements, the very large values obscure any trends that might be visible in the data, even in log-scale. Therefore, the dataset was visualised at different scales, by limiting it to maximum values approximately at the 40%, 60% and 95% quantiles (Figure 3.7). On all three scales, the effects of the different mutations seem pretty similar. Closer investigation of the data (not shown) revealed that it is mainly the mutated model that determines the mutation effect, not the type of applied mutation.

On first examination, the histogram plots in Figure 3.7 look like a scale-free graph. However, when plotted in a log-log plot, the slop changes at the different scales, so the graph is not purely scale-free.

Addition and replacement of reactions have the strongest influence on network behaviour To reveal differences in the effect of the mutation operators, the strong effect of the mutated network had to be removed from the behavioural difference graphs. To this end, I normalised the effect of the mutations on each model to the accumulated effect of all five mutations on that model (Figure 3.8). These *relative* mutation effects clearly classify the mutation operators in two groups. While deletion of a species, deletion of a reaction, and species duplication most often has a relatively weak effect, the addition and replacement of reactions tend to have a relatively strong effect. Species duplication is especially designed for this (see below). For the rest, the removal of model parts is less likely to have a strong effect on network behaviour than the addition of new parts (replacement is deletion plus addition).

Species duplication is a soft mutation operator A major problem with evolving biochemical networks seems to be the often deleterious effect of structural mutations on network behaviour. As seen above, the addition of new parts usually changes the resulting time series drastically, especially for smaller networks. Therefore,

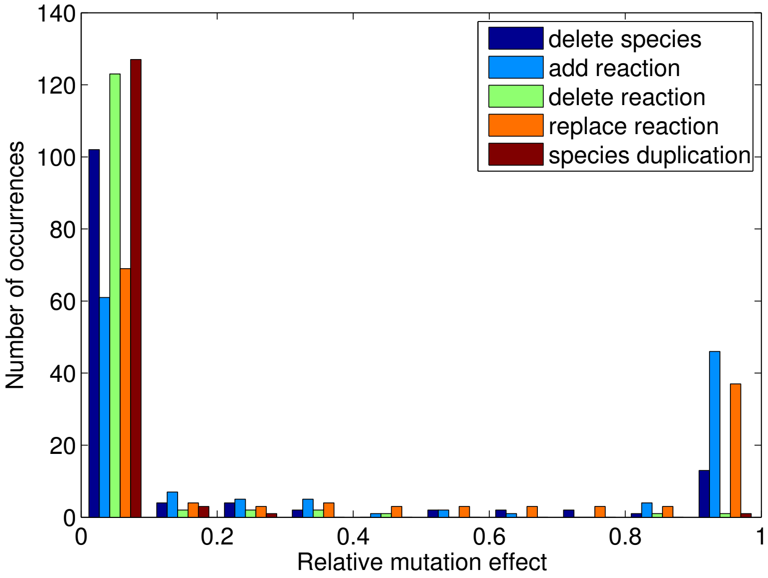


Figure 3.8: Histogram of the effects of the mutation operators, relative to their combined effect on each model. Out of the 184 models, 132 are represented here, for the rest the numerical integration of the mutated model failed. Strong effects are mainly caused by the addition and replacement of a reaction, while the other three mutations primarily cause weak effects.

we are looking for a “softer” mutation operator. Inspired by biology, one such operator is the duplication of a species and all the reactions it participates in. When the rate constants of all reactions producing the species are halved, this operator does not affect the concentrations of non-mutated species. Later on, deletions and rate mutations can exploit the additional freedom gained by duplication.

The relative mutation effect (Figure 3.8) reveals species duplication to be the softest mutation operator in the repertoire of the SBMLevolver. Going further, I examined the influence of this operator on the evolutionary performance for a logarithm network fitness function (see Section 3.1.1). The results (Figure 3.9) show that even though species duplication alone is inferior to addition of new species with random reactions, combining both operators does not yield an inferior result. However, it is still open under which conditions the combined approach improves the random addition of new species.

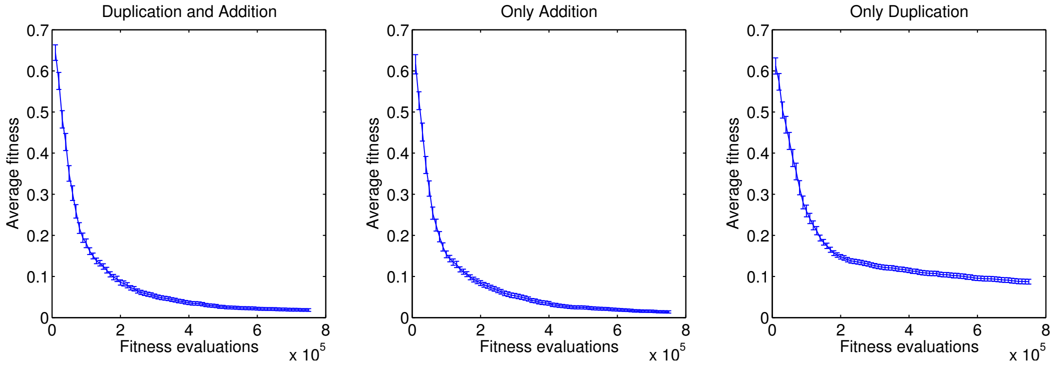


Figure 3.9: Average fitness development with standard error for logarithm networks, results from 100 independent runs. Left: species addition and duplication together, middle: only addition, right: only duplication. Also shown are the best 5 runs per setting (Grey). Global selection is (50+100)-elitist, local fitting is a (1+10)-ES, and 50 generations were calculated.

3.5 Case study: the human cell cycle spindle checkpoint

> **Note:** This case study resulted from a collaboration with Bashar Ibrahim, who provided the spindle checkpoint model and guided the fitness function design.

Segregation of newly duplicated sister chromatids into daughter cells during anaphase is a critical event in each cell division cycle. Any mishap in this process gives rise to aneuploidy that is common in human cancers and some forms of genetic disorders (Chung and Chen, 2002). Eukaryotic cells have evolved a surveillance mechanism for this challenging process known as the spindle checkpoint. The spindle checkpoint monitors the attachment of kinetochores to the mitotic spindle and the tension exerted on kinetochores by microtubules and delays the onset of anaphase until all the chromosomes are aligned at the metaphase plate (Fang, 2002).

To demonstrate the usefulness of our approach in systems biology, we applied combined structural- and parameter-optimisation to a recent model by Ibrahim et al. (2008) of the mitotic spindle checkpoint. This model, which is originally crafted by hand according to literature and laboratory data, describes in details the concentration dynamics of 17 species, namely Mad2, Mad1, BubR1, Bub3, Mad2*, Mad1*, BubR1*, BubR1:Bub3, APC, Cdc20, MCC, MCC:APC, Cdc20:Mad2, and APC:Cdc20, CENPE, Mps1 together with Bub1, and the kinetochore as a pseudospecies. Different kineto-

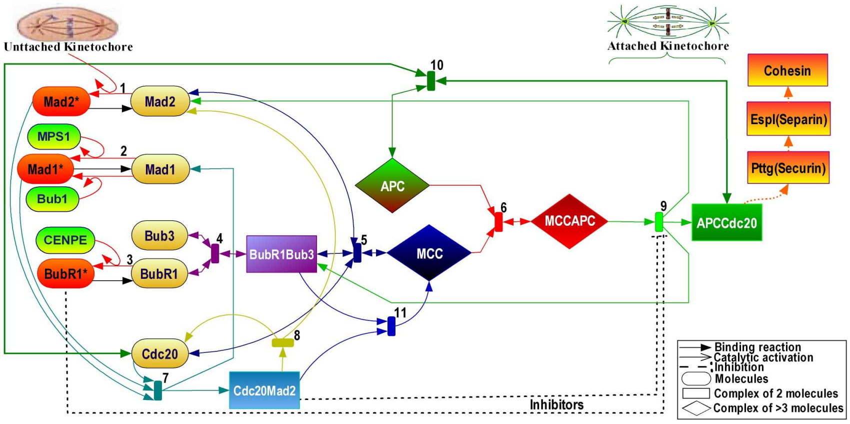


Figure 3.10: Schematic network model of mitotic spindle checkpoint. Figure taken from Ibrahim et al. (2008).

chores are represented by three compartments coupled by diffusion, each with the same 11 reaction rules. The last four species represent input signals to the model, reflected in the rate constants of certain reactions. The model corresponds to biological experimental results, which characterise the main components of the mitotic checkpoint.

As an optimisation target, the concentration of the central species APC:Cdc20 is supposed to be low as long as not all kinetochores are attached, but to rise to a higher value when they all are. In the paper by Ibrahim et al. (2008), this target has been combined with behaviour from knockout-experiments (which we do not consider here) to fit the rate constants. Here, we test which results can be achieved when the algorithm is allowed to add additional reactions. Any reaction given in the original model cannot be deleted. In future, it will be interesting to loosen this, which could lead to an evolutionary model reducer isolating only the important parts of a given model.

Our results, summarised in Table 3.1, show that performance of the model has improved compared to the optimisation of rate constants alone. In principle, this could also help to uncover additional structure in the data and to propose additional features of the system which can then be verified experimentally. To achieve biological plausibility, these additional features have to be constrained, which has not been observed in this proof of concept. The best model evolved contains only two additional reactions

	Level 1	Level 2	Level 3	Level 4	Fitness
Desired value	0	0	0	0.3	0
Without optimisation	0.086154	0.0865285	0.0869323	0.0872706	27.0122
Parameter optimised	0.010924	0.011051	0.011359	0.298768	0.470562
Structure optimised	0.000106	0.000051	0.000037	0.29972	0.000077

Table 3.1: Steady-state concentrations of APC:Cdc20 for four settings of the cell: All kinetochores unattached (1), one attached (2), two attached (3), all three attached (4). The unoptimised model has all parameters set to 0.1, the parameter optimised one has the values from Ibrahim et al. (2008), and the structurally optimised model is the result of the procedure described here. Note that the fitness function used is different to the one used by Ibrahim et al. (2008).

compared to the original,

$$\begin{aligned}\mathrm{BubR1}_Z &\rightarrow \mathrm{Mad1}^*_X + \mathrm{BubR1}^*_Y\\ \mathrm{BubR1}^*_X + \mathrm{Cdc20}_Y &\rightarrow \mathrm{Mad2}_X + \mathrm{Cdc20}_Y,\end{aligned}$$

but these require species from different compartments to interact, which is not intended in the model. After these experiments were performed, the SBMLevolver software was extended to allow for a compartmental structure of the evolved models. Conversion reactions are now only allowed to take place between species inside the same compartment, and designated transfer reaction take care of material transport between compartment. Unfortunately, due to time constraints, no results can be reported yet.

3.6 Conclusion

After a thorough presentation of the proposed two-level approach to the evolution of network models in Chapter 2, results from the systematic investigation of this approach and comparison to other algorithms are presented in this chapter. The main findings indicate that the SBMLevolver algorithm is superior to different one-level approaches, to random search on networks, and to pure parameter fitting of a fully connected network. When using the two-level approach, the exact composition of structural and parameter mutation is not too important, as the very strong influence of chance on the evolutionary outcome blurs any distinction.

A significant and valid criticism of the results in this chapter is the small range of evolutionary targets that were used to assess the algorithm. The main reason for this is the experienced difficulty in the evolution of networks with more complicated fitness functions. When the EA was presented with more difficult fitness functions, like the evolution of oscillatory behaviour, most runs did not even come close to a solution. For example, when evolving an oscillating network, the usual outcome would be a network that approaches the mean value of the desired oscillation at steady state. This is clearly a very suboptimal local minimum, since a manual construction of oscillating networks is possible. However, I clearly have to admit that I did not put enough effort in these problems to rule a possibility of solving them within the presented framework. Therefore, more work in this direction is encouraged and will certainly lead to an increased insight into the artificial evolution of network models.

In the literature, the evolution of network models has usually been treated from an outcome-oriented perspective. In this chapter, the focus is on the algorithm itself, and interesting results on its working are derived. However, this field is still largely open, and improvements are to be expected (and hoped for!) from variants of the two-level approach as well as from more unconventional approaches.

4

Evolution of Organisations in Chemical Computing Networks

Chapter summary. *How do chemical reaction networks that process information evolve? This is not only a fundamental question in the study of the origin of life, but also in diverse fields like molecular computing, synthetic biology, and systems biology. Here, we study the evolution of chemical flip-flops by means of chemical organisation theory. Additionally, we compare evolved circuits with manually constructed ones. We found that evolution selects for an organisational structure that is related to function. That is, the resulting computation can be explained as a transition between organisations. Furthermore, an evolutionary process can be tracked as a change of the organisational structure, which provides a fundamentally different view than looking at the structural changes of the reaction networks. In our experiments, 90% of evolutionary improvements coincide with a change in the organisational structure. We conclude that our approach provides a novel and useful perspective to study evolution of chemical information processing systems.*

> **Note:** The work presented in this section was done in collaboration with Naoki Matsumaru, and resulted in a publication at the ALife 2008 conference (Lenser et al., 2008). Naoki contributed the hand-designed network and its dynamic simulation, the rest of the chapter is my work.

4. EVOLUTION OF ORGANISATIONS IN CHEMICAL COMPUTING NETWORKS

Chapter contents

4.1 Introduction

While a bottom-up approach to design chemical computing networks has been pursued (Guido et al., 2006), top-down approaches, specifically evolutionary algorithms, have gained growing interests recently in order to design or program reaction systems. Efforts have been undertaken to evolve simple computational units (Deckard and Sauro, 2004), small biological networks (François and Hakim, 2004; Koza et al., 2001a; Soyer et al., 2006), genetic regulatory networks (Kuo et al., 2006a) or components thereof (Paladugu et al., 2006). Most of this work, however, has been focused on the final product, that is, the networks evolved to reproduce a certain specified behaviour. Here, we rather concentrate on the process of evolution. For that purpose, new methods are required that can deal with constructive systems (Fontana and Buss, 1994), that is, systems where new components (molecular species) or new interactions between existing components appear so that the network topology changes dynamically. Matsumaru et al. (2006) used chemical organisation theory (Dittrich and Speroni di Fenizio, 2007) in order to study the evolutionary dynamics of (artificial) chemical systems. In this paper, we analyse the trajectory of evolving chemical reaction networks that compute. In particular, we focus on networks that function as flip-flops.

4.2 Methods

4.2.1 Reaction networks and chemical organisations

Here, we utilised the notation of a chemical organisation developed by Dittrich and Speroni di Fenizio (2007) to analyse reaction networks. Following Fontana and Buss (1994), an organisation is defined as a set of molecular species that is closed and self-maintaining. The hierarchy of all organisations of a reaction network represents its organisational structure, which can be used to describe the dynamical (qualitative) behaviour of a reaction system as a movement between organisations (Speroni di Fenizio et al., 2001). Choosing a proper coding scheme, the organisational structure can be interpreted as a repertoire of behaviour patterns of the reaction system. For example, Dittrich and Speroni di Fenizio (2007) have shown that only species that form an organisation can make up a stationary state.

Algebraic description of a chemical reaction network The basic concepts needed here are now described more formally: A **reaction network** $\langle \mathcal{M}, \mathcal{R} \rangle$ consists of a set of (molecular) species $\mathcal{M}$ and a set of reaction rules $\mathcal{R}$. A **reaction rule** $\rho \in \mathcal{R}$ can be written according to the chemical notation:

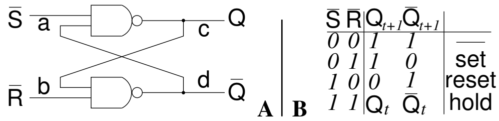
$$l_{a_1,\rho}\, a_1 + l_{a_2,\rho}\, a_2 + \dots \rightarrow r_{a_1,\rho}\, a_1 + r_{a_2,\rho}\, a_2 + \dots$$

The **stoichiometric coefficients** $l_{a,\rho}$ and $r_{a,\rho}$ describe the amount of molecular species $a \in \mathcal{M}$ in reaction $\rho \in \mathcal{R}$ on the lefthand and righthand side, respectively. Together, the stoichiometric coefficients define the **stoichiometric matrix**

$$\mathbf{S} = (s_{a,\rho}) = (r_{a,\rho} - l_{a,\rho}).$$

An entry $s_{a,\rho}$ of the stoichiometric matrix denotes the net amount of molecules of type a produced in reaction ρ. We also define mappings $\mathrm{LHS}(\rho) \equiv \{a \in \mathcal{M} : l_{a,\rho} > 0\}$ and $\mathrm{RHS}(\rho) \equiv \{a \in \mathcal{M} : r_{a,\rho} > 0\}$, returning the species with a positive coefficient on the lefthand and righthand side, respectively. Reaction ρ can take place in $A \subseteq \mathcal{M}$ only when $\mathrm{LHS}(\rho) \subseteq A$.

Organisational structure Given a reaction network $\langle \mathcal{M}, \mathcal{R} \rangle$ with $m = |\mathcal{M}|$ species and $r = |\mathcal{R}|$ reactions, the organisational structure is derived with respect to the following two criteria: closure and self-maintenance. A set of species $A \subseteq \mathcal{M}$



$\bar{S}$	$\bar{R}$	Q_{t+1}	$\bar{Q}_{t+1}$	
0	*0*	*1*	*1*	—
0	*1*	*1*	*0*	set
1	*0*	*0*	*1*	reset
1	*1*	Q_t	$\bar{Q}_t$	hold

Figure 4.1: Circuit diagram and operation mode of flip-flop.

is **closed**, if for all reactions ρ with LHS(ρ) $\subseteq A$, the products are also contained in A, that is, RHS(ρ) $\subseteq A$. This closure property ensures that there exists no reaction in A producing new species not yet present in the organisation using only species of that organisation. The other property is a theoretical capability of an organisation to maintain all of its members. Since the maintenance possibly involves complex reaction pathways, the stoichiometry of the whole reaction network must be considered in general. A set of molecules $C \subseteq \mathcal{M}$ is **self-maintaining**, if there exists a flux vector $\mathbf{v} \in \mathbb{R}^r$ such that the following three conditions apply: (1) for all reactions ρ that can take place in C (i.e., LHS(ρ) $\subseteq C$) the flux $v_\rho > 0$; (2) for all remaining reactions ρ (i.e., LHS(ρ) $\nsubseteq C$), the flux $v_\rho = 0$; and (3) for all molecules $a \in C$, the production rate $(\mathbf{Sv})_a \geq 0$. v_ρ denotes the element of $\mathbf{v}$ describing the flux (i.e. rate) of reaction ρ. $(\mathbf{Sv})_a$ is the production rate of molecule a given flux vector $\mathbf{v}$.

We visualise the set of all organisations with a Hasse diagram, in which organisations are arranged vertically according to their size in terms of the number of their members (cf. Figure 4.6). Two organisations are connected by a line if the upper organisation contains all species of the lower organisation and there is no other organisation between them. The Hasse diagram represents the hierarchical **organisational structure** of the reaction network under study.

4.2.2 Evolving flip-flop networks

Evolutionary process We employ an evolutionary algorithm that instantiates a natural selection process on chemical reaction networks (Fernando and Rowe, 2007). The algorithm can mutate the reaction rules $\mathcal{R}$ of a reaction network with a fixed predefined set of molecular species $\mathcal{M} = \{a^0, a^1, b^0, b^1, c^0, c^1, d^0, d^1\}$. As mutational operators, the algorithm can add or delete a reaction, or replace a reaction with a different one, keeping as many of the previous participants as possible. To keep things simple, we employ a (1+1)-EA. That is, one parent generates one offspring, while the better of the two survives.

To enable neutral mutations and thus search space exploration, the offspring is kept if both have the same fitness. No parameter fitting is done, so that a change in parameters can only be realised through a replacement of a reaction with the same reaction, which has a different (randomly chosen) reaction constant. Only mass-action kinetics of first and second order are used in the evolution.

Flip-flop implementation When speaking of a flip-flop logic gate in this work, we specifically mean an RS (Reset and Set) flip-flop, with a behaviour according to the truth table in Figure 4.1. To represent the four binary variables `a`, `b`, `c` and `d` making up this flip-flop in a chemical format, we employ two opposing species x^0 and x^1 for each binary variable `x`, where the presence of x^0 denotes the value `x = 0`, and x^1 denotes `x = 1` (cf. Matsumaru et al., 2007). To help maintain a valid state inside the system, we fix four destructive reactions $x^0 + x^1 \to \emptyset$ for all four species pairs $x^i = a^i, b^i, c^i, d^i$. These reactions cannot be changed or deleted by the evolutionary algorithm.

The ideal flip-flop that is the target of the artificial evolution works in the following way: The set operation $(\bar{S}, \bar{R}) = (0, 1)$ changes the state Q to 1, while the reset $(\bar{S}, \bar{Q}) = (1, 0)$ changes Q to 0. To hold the previous state, both inputs are set to 1. The forbidden input $(\bar{S}, \bar{Q}) = (0, 0)$ is not considered in the fitness function. In chemical form, the input $(\bar{S}, \bar{R}) = (0, 1)$ is represented by defining an inflow for a^0 and b^1, that is, $\{\emptyset \to a^0, \emptyset \to b^1\} \subseteq \mathfrak{R}$; and the other two cases are treated similarly. The initial concentrations of c^i and d^i are set according to the previous state Q_t. Taking this together, we get six different test cases, coming from three different operations with two initial conditions each.

Fitness evaluation For each case, we specify either the presence or the absence of each species as desired, measured in steady state after simulating the reaction system for 1000 seconds. Numerical integration is done using the SBML ODE Solver Library (Machne et al., 2006). The classification as present or absent is decided by a concentration threshold of 10^{-9} (arbitrary units). For example, in the reset case, the following steady state concentrations are considered as correct: $a^1 = 1, a^0 = 0, b^1 = 0, b^0 = 1, c^1 = 0, c^0 = 1, d^1 = 1, d^0 = 0$. The fitness value is then calculated by counting the number of wrong presence / absence measurements, with 0 being the best possible fitness value. Once a fitness of 0 is reached, the evolution stops.

4.3 Results

To analyse the evolution of reaction networks acting as flip-flops, we performed 30 independent runs in order to evaluate properties of a “typical” run. Additionally, we also looked at one run in more detail.

Since the distinction between the three different input settings is realised by enabling or disabling inflow reactions for a^1, a^0, b^1 and b^0, we need to compute three lattices of organisations for each analysed candidate solution, one for each input setting.

4.3.1 Statistical analysis of many runs

Fitness development over normalised evolutionary time The average fitness development (Figure 4.2) shows a stronger gain in fitness at the beginning, while the convergence towards zero is slower later on. Eventually, all runs reached a fitness of zero, i.e. the networks behaved as specified in the fitness function. Since a run stops exactly when the fitness of the current individual is 0, the number of generations usually differs between runs. In order to be able to average over these runs, we had to resample the data on fitness and number of organisations, such that a common number of measurements for each run is achieved. To this end, we constructed a timescale of “normalised evolutionary progress”, defined by its endpoints 0.0 at the beginning of the evolution and 1.0 at the end when the final solution is found. The MATLAB function `resample`, which applies an anti-aliasing low-pass FIR filter during the resampling process, was used to create new data points at 1001 equally space points between 0.0 and 1.0.

Evolution of the number of organisations Looking at the number of organisations for the three different input cases, we can see from Figure 4.3 that starting from around four to five organisations on average, the numbers diverge between the set/reset operations and the hold operation. While the number of organisations for the set/reset organisation converges between two and three, the hold operation yields around seven organisations on average.

Relation of fitness improvements and organisational changes By comparing the organisational structures between successive candidate solutions, we calculated that 90% of all fitness improvements are accompanied by a change in the organisational structure for at least one input case. In contrast, only 18% of organisational changes

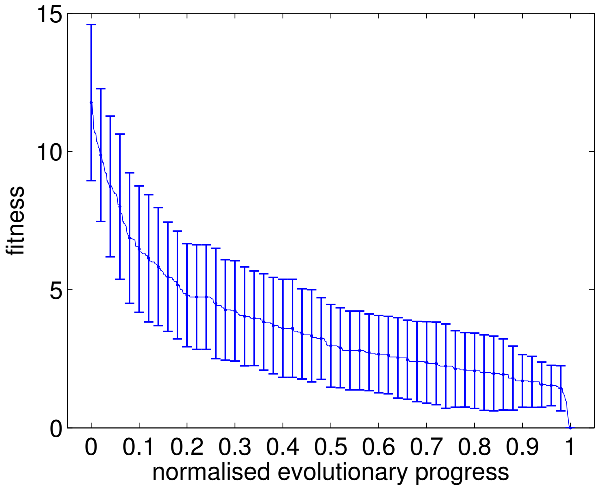


Figure 4.2: Average fitness value from beginning to end of evolutionary runs, from 30 independent repetitions. The x-axis denotes the normalised evolutionary progress from the random initial solution ($x = 0$) to finding a solution with fitness 0 ($x = 1$). For this, the different runs were resampled to 1000 samples, as described in the text. Error bars indicate standard deviation.

also come with a fitness improvement. When looking at the lineage of networks that led to the final solution, disregarding unsuccessful candidates, we find that 35% of all mutations changed the organisational structure for at least one input.

4.3.2 Detailed analysis of one run

For an in-depth analysis, we pick the first evolutionary run that we performed for this problem. We will describe how the fitness improvements correlate with changes in the organisational structure, and give details on one specific mutational event and its consequences for the organisational structure of the network.

Analysed run is a typical representative Comparing the average fitness development shown in Figure 4.2 with the single run analysed here (Figure 4.4 upper part), we can conclude that the fitness of the individual run progressed in a fairly standard way. This is especially true given that the behaviour of the 30 runs is quite diverse, as indicated by the large standard deviations. Also the length of this run (162 generations) is in the usual region, with an average run taking 221 generations with a standard deviation of 119. Also the number of organisations (Figure 4.4 lower part) is in agreement with the average number (Figure 4.3), even if the number of organisations

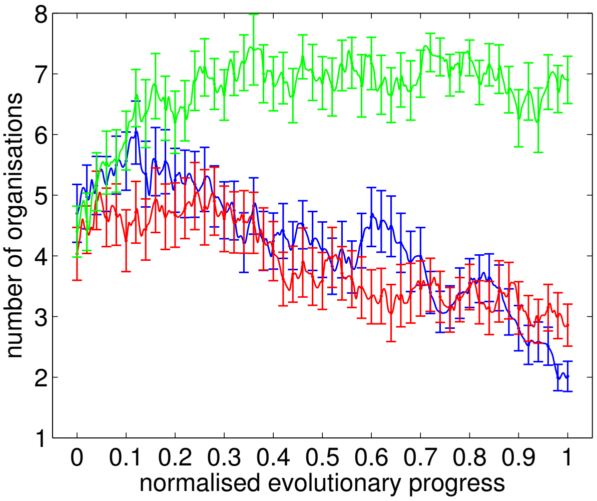


Figure 4.3: Average number of organisations from 30 independent runs of the evolution. Colours denote the (a_0, b_1) input (blue), the (a_1, b_0) input (red), and the (a_1, b_1) input (green). Error bars indicate the standard error. Unit of x-axis as in Figure 4.2.

for the set/reset operations are at the outer limits of the typical range (five and one, respectively).

Fitness jumps come with organisational changes, but not vice versa Looking at the fitness increases in the course of the evolution (Figure 4.4 upper part) and the organisational structures of all networks that appear during the run (not shown), it can be observed that all but one of the eight fitness jumps are accompanied by changes in the organisational structure for at least one of the three input cases. However, taking the number of organisations at any point in the evolution into account, one can also see that not every change in organisational structure leads to a fitness change, in fact, most do not.

An example mutation We now relate the fitness change in one successful mutation to the change in organisational structure incurred by that mutation. As an example, we pick the fitness jump from generation 112 to 113, which improved the fitness from seven to four wrong presence/absence values. Looking at the reaction networks before and after the mutation (Figure 4.5), we see that the mutation added one reaction, which converts b^0 into d^1.

This additional reaction does not change the organisational structure for input cases (a^0, b^1) and (a^1, b^1) (set and hold, not shown), but reduces the lattice of organisations

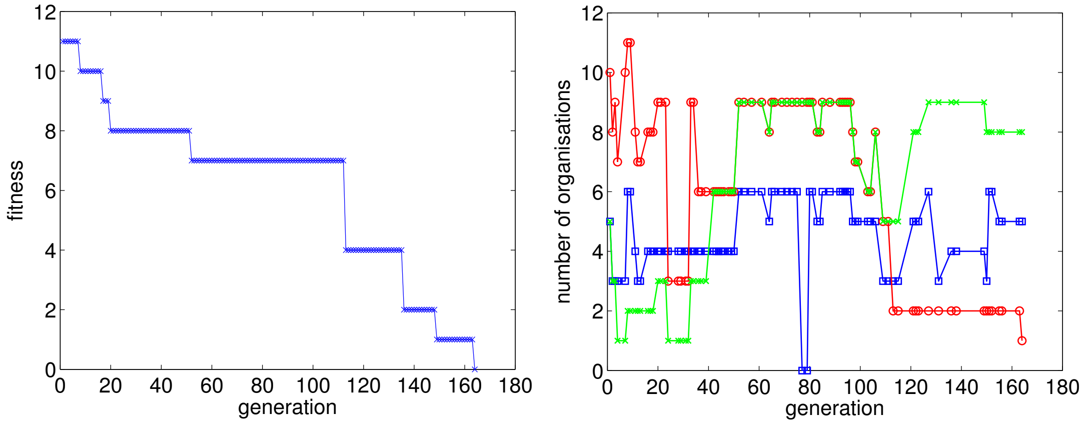


Figure 4.4: One exemplary run. Given are fitness (upper plot) and number of organisations for all three input cases (lower plot). In the lower plot, the three input cases are shown in blue $((a^0, b^1)$ input), red $((a^1, b^0)$ input), and green $((a^1, b^1))$. Each mark (square, cross or circle) denotes a new network structure in the evolutionary trajectory.

for input case (a^1, b^0) (reset) from five organisations to two (Figure 4.6). Looking at the behaviour of both networks for all input cases and initial configurations (i.e. all six test-cases), one can observe that the change occurs only in input case (a^1, b^0) with an initial configuration in which c^1 and d^0 are present (data not shown). For this case, the steady-state before the mutation has a^1, b^0, and c^1 present, and d^0 is still present after 1000 seconds even though its concentration is still decreasing at that time. This yields four wrong presence/absence values, since c^0 and d^1 should be present and c^1 and d^0 should be absent, but the total opposite is the case. After the mutation, d^1 is present and c^1 and d^0 are absent, but also c^0 is absent, so there is one wrong value left.

On the organisational level, the mutation removes three organisations (Figure 4.6), among which is also the organisation (a^1, b^0, c^1) responsible for the wrong behaviour of the network. After the mutation, the dynamics take the steady-state into organisation (a^1, b^0, d^1), resulting in a better behaviour. However, it is interesting to note that both organisational structures also contain organisation (a^1, b^0, c^0, d^1), which is the “correct” one that is also used in the final solution. Even though this organisation is present, the dynamics of both reaction systems are such that the steady-state does not lie inside it. We had to wait for another 49 generations for this to happen.

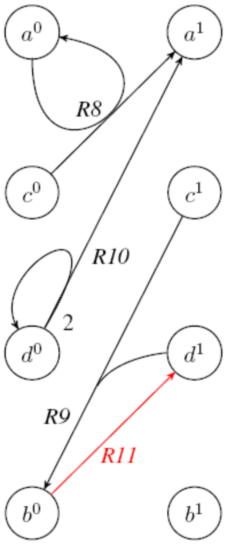


Figure 4.5: The reaction network of the candidate solution analysed in the text, after the mutation adding reaction *R11*. The added reaction is shown in red.

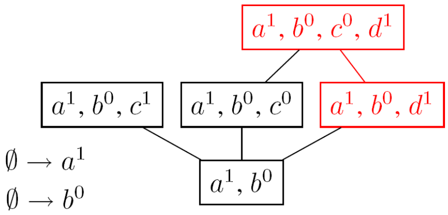


Figure 4.6: Organisational structure of the networks from Figure 4.5 for input (a^1, b^0), before (whole structure) and after addition of reaction *R11* (only red part).

4.3.3 An evolved chemical flip-flop

An outcome of the evolutionary process described above is analysed. The reaction network considered here has a fitness value of 0, i.e. it solves the given task. The network structure is shown in Figure 4.7.

There are seven reactions, labelled as *rea1* to *rea7* in the figure, in addition to four reactions of cooperative decay (not shown in the figure), $a^1 + a^0 \to \emptyset$, $b^1 + b^0 \to \emptyset$, $c^1 + c^0 \to \emptyset$, $d^1 + d^0 \to \emptyset$. This base reaction network is extended to include inflow reactions, representing the inputs to the flip-flop circuit, depending on the operations. Organisational structures of the reaction system for each operational mode are shown in Figure 4.8.

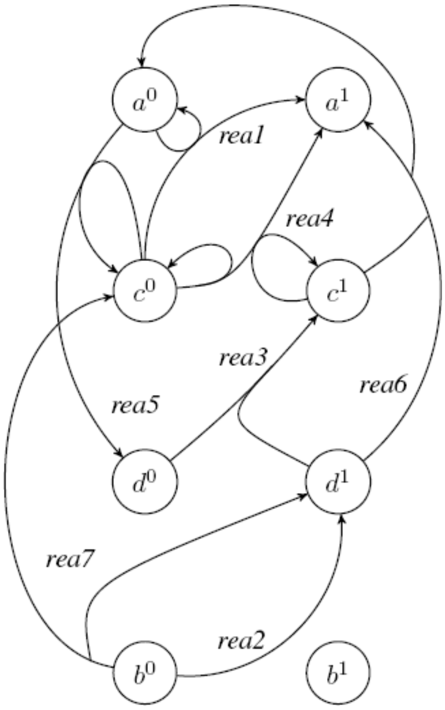


Figure 4.7: Chemical reaction network implementing flip-flop circuits, designed through an evolutionary process. Cooperative decay reactions ($a^1+a^0 \rightarrow \emptyset$, $b^1+b^0 \rightarrow \emptyset$, $c^1+c^0 \rightarrow \emptyset$, $d^1 + d^0 \rightarrow \emptyset$) are omitted.

Organisational analysis Analysing the organisational structure of the reaction network, it becomes evident that the reaction system based on this reaction network is surely usable for the flip-flop computation. Including the two inflows $\emptyset \rightarrow a^1$ and $\emptyset \rightarrow b^0$ in the reaction network, as shown in Figure 4.8 A, only one set of species $\{a^1, b^0, c^0, d^1\}$ satisfies the conditions to be the organisation. It implies that only this species combination can be found in the dynamical reaction system in equilibrium states. Therefore, the reset operation can be realised in the evolved reaction system. The network with the inflows of $\emptyset \rightarrow a^0$ and $\emptyset \rightarrow b^1$ contains five organisations as shown in Figure 4.8 B, and one of those $\{a^0, b^1, c^1, d^0\}$ corresponds to the set operation.

Changing inflow reactions to $\emptyset \rightarrow a^1$ and $\emptyset \rightarrow b^1$ achieves the hold operation. In terms of the organisations, as shown in Figure 4.8 C, the two organisations *orgHR*= $\{a^1, b^1, c^0, d^1\}$ and *orgHS*= $\{a^1, b^1, c^1, d^0\}$ in the reaction network with those inflows reflect the bistability of the flip-flop circuit. Depending on the state at the previous time step, the hold operation results in a different state, namely the previous one.

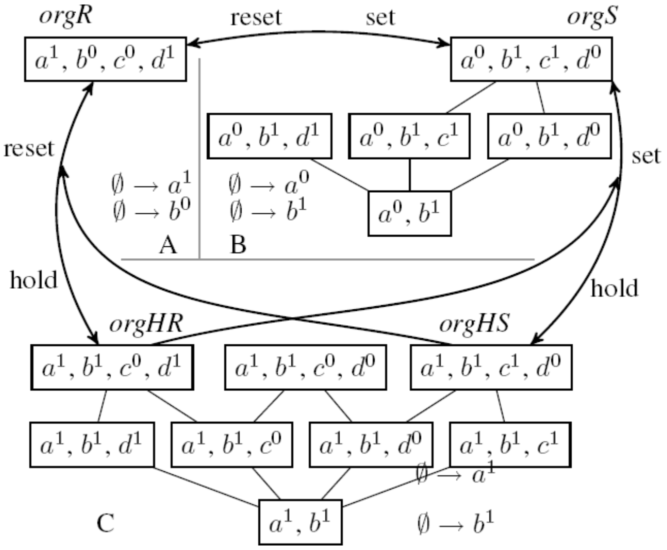


Figure 4.8: Organisational structure in the reaction network shown in Figure 4.7.

When the reaction system has been in the state after the set operation, (i.e., *orgS*), the hold operation brings the system to the state of *orgHS*, keeping the output species unchanged as c^1 and d^0. Holding the information that the system has been reset can be achieved by moving the system state from *orgR* to *orgHR*.

The last operation of setting both inputs to be zero ($\mathtt{a} = \mathtt{b} = 0$) is forbidden for the flip-flop circuit. If adding two inflows of $\emptyset \to a^0$ and $\emptyset \to b^0$ to the base reaction network, only one organisation arises: $\{a^1, a^0, b^0, c^1, c^0, d^1, d^0\}$.

If no inflow reaction is present, there are 42 organisations in the base reaction network. The smallest organisation is the empty set $\emptyset$. The sets containing four species form the largest organisations, and there are four organisations of that size. The organisations with the size of four in Figure 4.8 are also found to be the organisation without inflows, except the organisation labelled as *orgR*. In fact, all organisations in Figure 4.8 except *orgR* are also organisations without inflows.

4.3.4 Dynamical behaviour

Mathematical description of network dynamics To validate the organisational analysis of the reaction network, a dynamical reaction system is constructed and simulated with *Copasi* (Hoops et al., 2006), a biochemical reaction system simulator. Agreeing to the fitness calculation of the evolutionary design process, mass action kinetics is assumed for every reaction, if applicable. The ordinary differential equations (ODEs) for the input species read:

$$\begin{aligned}
d[a^1]/dt &= k_1[a^0][c^0] + k_4[c^1][c^0] + k_6[c^1][d^1] - d_a[a^1][a^0] + I_{a^1}(1 - [a^1]) \\
d[a^0]/dt &= -k_5[a^0][c^0] + k_6[c^1][d^1] - d_a[a^1][a^0] + I_{a^0}(1 - [a^0]) \\
d[b^1]/dt &= -d_b[b^1][b^0] + I_{b^1}(1 - [b^1]) \\
d[b^0]/dt &= -k_2[b^0] - k_7[b^0] - d_b[b^1][b^0] + I_{b^0}(1 - [b^0])
\end{aligned}$$

where a kinetic parameter for a reaction rea_id is denoted as k_{rea_id}. Kinetic parameters for the cooperative decay reactions are represented by d, and the subscript specifies the pair. For example, the decay rate of the cooperative decay reaction $a^0 + a^1 \rightarrow \emptyset$ is denoted as d_a.

Inflow reactions representing the operation of reset, set, and hold are controlled by the four parameters: I_{a^1}, I_{a^0}, I_{b^1}, and I_{b^0}. These parameters are binary variables, accepting only 0 or 1. For example, when the chemical flip-flop is set, I_{a^0} and I_{b^1} are set to one and the other pair of parameters I_{a^1} and I_{b^0} are set to zero. Inflows are assumed to be constant fluxes. Furthermore, the inflows are coupled to decay reactions such as $a^1 \rightarrow \emptyset$ in order to avoid endless increase of the input species concentration. The resulting term of the ODE is $I_{a^1}(1 - [a^1])$, for example.

The ODEs for the output species read:

$$\begin{aligned}
d[c^1]/dt &= k_3[d^1][d^0] - k_6[d^1][c^1] - d_c[c^1][c^0] - I_{b^0}[c^1] \\
d[c^0]/dt &= -k_1[a^0][c^0] + k_7[b^0] - d_c[c^1][c^0] - I_{b^0}[c^0] \\
d[d^1]/dt &= k_2[b^0] - k_3[d^1][d^0] - k_6[d^1][c^1] + k_7[b^0] - d_d[d^1][d^0] - I_{b^0}[d^1] \\
d[d^0]/dt &= -k_3[d^1][d^0] + k_5[a^0][c^0] - d_d[d^1][d^0] - I_{b^0}[d^0]
\end{aligned}$$

Modifications for repeatable computations Kinetic parameter values are also provided as the outcome of the evolutionary design, but we manually adjusted the values so that the operations can be continuously repeated. When the fitness of the reaction system was calculated during the evolution process, three of the operations were evaluated separately and the reaction system was reinitialised for each case. This re-initialisation step between operations is prevented here, so that the end state of the previous operation becomes the initial state of the next operation. For that purpose, the outflows of the input species are added as described above in order to restrict the increase of the concentration. For the output species, the outflows are also added as shown above, activated only when the inflow of b^0 is present. This modification is also to restrict the increase of the concentrations of the output species, specially, when the system is reset. These modifications are purely restricted to the dynamical simulation of this one model and do not influence the organisational analysis above.

The last modification is the kinetic parameter of the reaction *rea1*, k_1, from 4.44231 to 0.5. The rational of this adjustment is: under the input condition "set", the system is observed to converge to the organisation of $\{a^0, b^1, d^1\}$, instead of *orgS*. This behaviour results from the fast extinction of species c^0 so that the generation of d^0 by *rea5* is insufficient. Slowing down the reaction speed of *rea1*, species c^0 stays in the system longer and produces d^0 enough to neutralise d^1.

With these modifications, the network is able to continuously act as a flip-flop, which is more than the original fitness function demanded. The reason for the simpler fitness function is straightforward: Determining whether a system can continuously work as a flip-flop requires long-time monitoring, which makes fitness evaluations very costly.

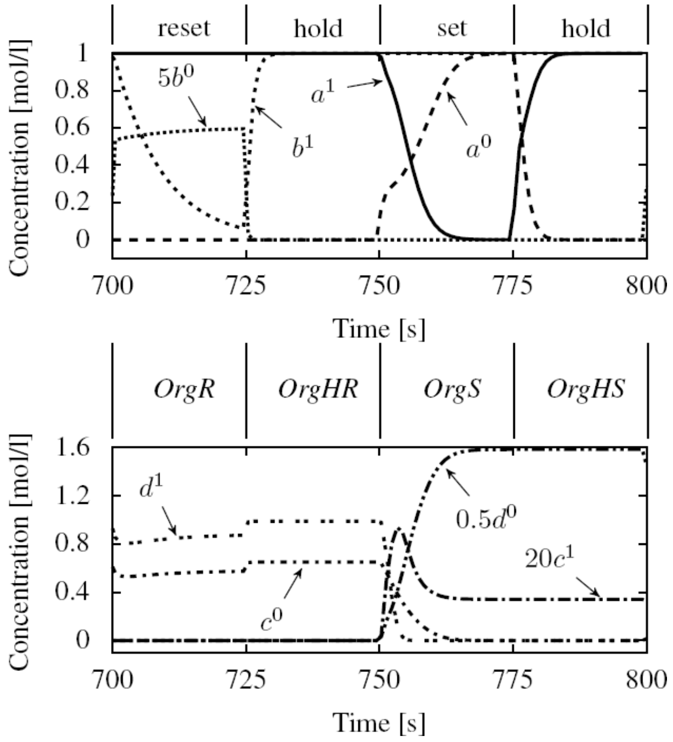


Figure 4.9: Dynamical simulation of chemical flip-flop designed by evolution. Parameters are set as follows: $d_a = d_b = 0.1$, $k_4 = 2.33941$, $k_6 = 2.83745$, $k_1 = 4.44231$, $k_5 = 3.62963$, $k_7 = 4.82838$, $k_3 = 1.0$, $k_2 = 0.1$, $d_c = 0.001$, $d_d = 1.0$. Additionally, for each operation of reset, set, and hold, inflow reactions are activated. For the set operation, the parameters are set such that $I_{a^0} = I_{b^1} = 0$ and $I_{a^1} = I_{b^0} = 1$ to activate inflows of a^1 and b^0 species and to deactivate the others. The reset operation is initiated by setting $I_{a^0} = I_{b^1} = 1$ and $I_{a^1} = I_{b^0} = 0$. The hold operation is achieved with the parameter settings of $I_{a^1} = I_{b^1} = 1$ and $I_{a^0} = I_{b^0} = 0$.

4.4 Conclusion

Organisation theory helps to understand network evolution, but it is not enough We found that most fitness improvements come together with change in organisational structure (90%), showing that organisation analysis indeed yields insight into the evolutionary process. On the other hand, most organisational changes are fitness-neutral (82%), indicating that a lot of the information given in the lattice of organisations does not directly relate to the measured function of the networks. We have also seen mutations where the replacement of a reaction with the same type of reaction led to a fitness increase caused purely by the changing of a kinetic parameter, as well as changes of network structure not reflected in the organisations (but improving fitness). All this implies that while organisational analysis can give us many indications regarding the function of a reaction network, sometimes it does not tell the whole story of the network's dynamics.

Evolved networks contain more organisations than hand-constructed ones We have also seen that the number of organisations for the set and reset states is substantially smaller than the number for the hold state, in analogy to the hand-constructed flip-flop by Matsumaru et al. (2007). In comparison to their solution, the evolved networks show a larger number of organisations for each input case. To realise the flip-flop behaviour in the reaction system, the minimum number of organisations in the reaction network is one for the set and reset operation and three for the hold operation. The hand-designed flip-flop implementation shown by Matsumaru et al. (2007) has two organisations for set and reset, respectively, and three for hold. In comparison, the evolved networks have more organisations, on average between two and three each for set and reset, and seven for hold. This implies that even though the function of the flip-flop networks is reflected in their organisational structure, this structure contains more information than only the operational modes specified in the fitness function, and additionally potential "junk" in the form of unused organisations.

Extensions: organisation-oriented evolution, differential effects of mutation operators As an interesting extension to this work, one could use organisational analysis to direct the evolution of reaction networks. By first designing the perfect organisational structure and then evolving networks with this structure, it would be possible to study whether these networks have the desired functionality. A key step in this direction is certainly the design of an appropriate fitness function based on a network's lattice of organisations.

In an additional investigation on top of the results shown here, one should look at the effect of different mutational operators on the organisational structure of a network. This will lead to helpful insights on how the mutations affect the lattice of organisations, and also on how specific organisational changes are related to changes in the fitness function.

Organisational structure of evolved networks favours flip-flop operation In our opinion, the most important lesson to be learned from this work is that the evolutionary process investigated here produces reaction networks with an organisational structure that reflects their flip-flop functionality. Even though our choice of representation format of the binary information in chemical form may favour this, we believe that this phenomenon is mainly caused by the structure of the fitness function, i.e. by the task that is required of the networks.

5

Network Reconstruction From FRAP Data

Chapter summary. *Network models of biological systems often contain "white spots", where the available knowledge about components and their connections is not complete. In these cases, automatic network inference from data is often used to fill these gaps. Evolutionary algorithms, such as the one described in Chapter 2, provide a popular method for automatic inference of networks. In this chapter, we rigorously investigate the inference of dynamic models for binding and unbinding of PML I, DAXX and Sp100 to PML nuclear bodies (cf. Chapter 6). We show that while no single best model exists, the EA reveals a set of necessary and sufficient criteria that successful models need to fulfil. Additionally, an enumeration of all possible models for PML I shows that the search space consists of three large plateaus, preventing an effective application of heuristic optimisation techniques on the network level.*

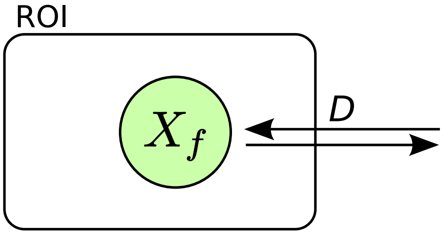
Chapter contents

5.1 Introduction

Automatic generation of alternative models through evolution In Chapter 6, a mathematical model is developed that captures the diffusion and binding behaviour of a number of molecular components of the promyelocytic leukaemia nuclear bodies (PML NBs). The topological design of this model was guided by biological insight, and free parameters were subsequently fitted to the FRAP data. However, a natural question to pose is whether this is the only reasonable model able to explain the data, or whether other explanations (models) exist. In this chapter, we employ the SBMLevolver software (detailed in Chapter 2) to create multiple alternative models for a subset of the data used in Chapter 6, namely the FRAP time curves for PML I, DAXX and Sp100. The collection of resulting model topologies is then statistically analysed in order to assess the uncertainty in explaining the data with a specific kinetic model.

Artificial network evolution is relatively unexplored Beyond the specific biological study, this chapter deals with the feasibility of automatic network reconstruction in general, especially in the non-linear case, using evolutionary algorithms (EAs). While EAs have been used extensively to fit kinetic parameters of a given network to time-series data (e.g., see Dräger et al., 2007; Moles et al., 2003), its role in the evolution of the whole network structure *ab initio* is relatively unexplored (but see Deckard and Sauro, 2004; Koza et al., 2001a; Paladugu et al., 2006). Here, an EA approach is applied to a real-world time-series dataset, and its feasibility is evaluated.

An application of the SBMLevolver This work also provides an important use case for the SBMLevolver, by showing how it can be employed to automatically construct models from time series data, and highlights some of the problems in this process. This leads to valuable insight for a possible further development of this tool and for its practical use.



$$\frac{dX_f}{dt} = D(C - X_f)$$

Figure 5.1: Basic framework from which the evolved models start. One freely moving species X_f diffuses inside and out of the region of interest (ROI) with diffusion constant D, where D incorporates both the measured diffusion constant as well as the circumference of the ROI.

5.2 Experimental setup

5.2.1 Search space

Evolved models need a connection to biological reality Preliminary experiments (not shown) gave the impression that if the topology search space is large enough, the evolution of network models that fit the desired FRAP time courses is easy. However, these models were generic networks, with no connection between the components of the model and of the corresponding biological system (In the AI community, this is referred to as the "grounding problem"). For example, if the resulting model consists of ten species with mixed kinetic reactions between them, it is very hard to decide which species in the evolved model corresponds to which biological entity in the cell. Therefore, a certain degree of grounding - i.e. pre-specified external knowledge built into the model - is needed in order to make the automatically evolved part interpretable.

The initial model consists only of the inflow reaction For this reason, the decision was taken to include *a priori* a diffusion-reaction of one species into the region of interest (ROI), and make this reaction untouchable by the evolutionary algorithm. Species duplication was disabled in the EA and no diffusion reaction could be added, so this is the only species with a diffusive inflow throughout the evolutionary run. Apart from that, the initial model contains no more reactions and only this one species (Figure 5.1).

Allowed species and reactions The number of species in the models is allowed to vary between one (the initial) species and three species, while the number of reactions can vary between zero and ten. As we are only interested in binding and unbinding processes, the only allowed type of reaction is a uni-uni reaction with mass-action kinetics and a rate constant between 0.0 and 5.0.

5.2.2 Artificial evolution

EA setup In all evolutionary runs, a (30+30) strategy was pursued on the network topology level, and 100 replications were computed for each dataset. All mutational operators of the SBMLevolver except for species duplication (see Chapter 2) were enabled and operated with equal probability. For the parameter fitting loop, a (1,5) CMA-ES was employed with maximally ten generations per evaluation of one network structure, always starting from the parameters values of the ancestor structure (which may in fact have been the same structure, since one of the structural operators is the copy operator).

Fitness function In analogy to Chapter 6, the evolved models deal with concentrations inside the ROI, considering only species fused to a fluorescent molecular tag (usually GFP). Therefore, the fluorescence measurement in the ROI is defined as the sum over all species concentration curves obtained from numerical integration of the respective model, and the fitness function is given by the quadratic distance of this summary curve to the FRAP measurements for PML I, DAXX or Sp100. All measurements and simulations are quantified in relative fluorescence intensity (RFI), with 0 representing no fluorescence and 1 the steady state value before the photobleach.

As in the sequential binding model in Chapter 6, called the *null model* from here on, the initial concentrations of the three species were determined by starting the integration from a concentration of zero for all species, then integrating until the sum of all concentrations reached the first FRAP measurement. This approach assumes that the bleach pulse is 100% effective, even though the first measurement already yields a non-zero RFI, due to the time delay of the fluorescence measurement (see discussion in Chapter 6).

Collection of evolved models To relate all alternative evolved models to the null model of PML NB assembly, all evolved model structures that were better or not more than 10% worse than the null model were collected. For the PML I and DAXX

	PML I	**DAXX**	**Sp100**
considered models	138	105	107
selected by fitness	138	105	44
mean fitness of selected	0.095	0.080	0.047
fitness variance of selected	10^{-6}	$4.1 * 10^{-5}$	10^{-4}
mean fitness of non-selected	-	-	0.097
null model fitness	0.127	0.134	0.061

Table 5.1: Overview of the ensemble of models evolved for the FRAP data. Listed are all evolution experiments performed, i.e. for PML I, DAXX, and Sp100. Shown are the number of considered models (the best models from each run, sometimes one run produced more than one best model), from these the number of models that were not more than 10% worse than the null model, the mean fitness of these, the variance between the fitnesses, the mean fitness of the non-selected models, and the fitness of the null model for each experiment.

datasets, this procedure selected all models, while for the Sp100 dataset, only 44 out of 107 models were good enough to be included in the analysis (Table 5.1). It is not clear why not all network topologies for Sp100 pass the fitness criterion, while all solutions for PML I and DAXX pass it. It seems that the time-course for Sp100 favours network topologies which can quickly be tuned to a mediocre fit, but which do not allow a perfect fit and cannot easily be escaped by structural mutations.

Network processing and matching The reactions in these models were filtered from the SBML files, neglecting trivial ones (of type $X \longrightarrow X$). Note that in contrast to the null model, the parameter k_{on} in the evolved models is not determined by specific values for the other three parameters, since an asymptotic return of the RFI to 1 is not automatically enforced (cf. Chapter 6). This additional constraint in the null model probably explains why its fitness is worse than the average fitness of all selected models in all three datasets.

To assess the difference between the null model and each evolved model, each species in an evolved model has to be matched to a species in the null model. As the freely diffusing species with inflow (X_f) is fixed in all evolved models, only two possibilities for assigning the other two species exist. For each evolved model, the assignment with less differing reactions was selected for further consideration.

Determination of operator bias The evolutionary operators used in the search as well as the permutation of species lead to a bias in the ensemble of evolved models. E.g., the reaction $X_f \longrightarrow X_1$ occurs more frequently than $X_f \longrightarrow X_0$, simply because the former reaction actually appears in the null model, and thus X_1 and X_0 are usually permuted in such a way that $X_f \longrightarrow X_1$ is included in the evolved model as well.

To compensate for that effect, another 100 control runs of the EA were conducted under neutral fitness, i.e. all models appearing in the course of evolution were assigned a fitness of 1.0, independent of their structure or parameters. For the resulting set of models (from the final generation), the same analysis as described above was performed. Now, the effect of selection can be distinguished from the bias introduced by the mutational operators, by comparing the models evolved under selection with those evolved without selection. The three models that occurred more than four times in the evolution under neutral fitness are shown in Figure 5.3. Comparing these to the networks that were evolved for the three datasets, it is obvious that the search operators clearly introduce a bias towards the $X_f \longrightarrow X_1$ and $X_1 \longrightarrow X_f$ reactions, but do not introduce any other effect in the results.

5.3 Results

5.3.1 Criteria for successful model fits

The procedure outlined above was performed for the PML I, DAXX and Sp100 FRAP datasets. For all three datasets, a variety of network topologies (up to eleven for PML I) leads to near-perfect fits of the data. However, these topologies share a number of common traits that can be found in all three runs. Basically, these traits allow us to precisely specify the conditions under which a model can provide a satisfactory fit of the data.

Two-way reaction between X_f and X_1 is necessary The most striking trait is that all successful models feature a two-way reaction between X_f and X_1 (again, the reason that this reaction always uses X_1 instead of X_0 is that the X_1-reaction also exists in the null model). This reaction is the main controller of the fluorescence recovery, as it directly regulates the concentration of X_f in the ROI, which in turn determines the inflow of fluorescent material into the ROI. The feedback of X_1 on X_f is used to shut off the inflow once enough fluorescent material is in the ROI. Without this feedback, the RFI would increase linearly without end.

All successful models employ two binding mechanisms However, the two-way reaction between X_f and X_1 is clearly not enough to fit the data, as no successful model topology contains only this reaction. On its own, this reaction together with the diffusive inflow would create a negative-exponential FRAP curve of the form $A(1-e^{-kt})$, which apparently cannot fit any of the three datasets accurately. Therefore, a second characteristic trait of all successful topologies is that there is always a reaction producing X_0, either from X_f or from X_1. In biological terms, this implies that the observed behaviour cannot be explained by a single fraction of bound molecules, but that a second fraction (or binding mechanism) must be involved.

Feedback from X_0 is only needed for DAXX and Sp100 In all evolved models for DAXX and Sp100, X_0 feeds back into either of the other two species, while for PML I, 112 of the 138 topologies do not contain a feedback from X_0. Apparently, this feedback reaction in not needed for an optimal fit of the PML I data, in contrast to DAXX and Sp100. Hence, a third criterion (the first one not met by all solutions) to characterise the topologies is the presence of reactions going from X_0 to either X_f

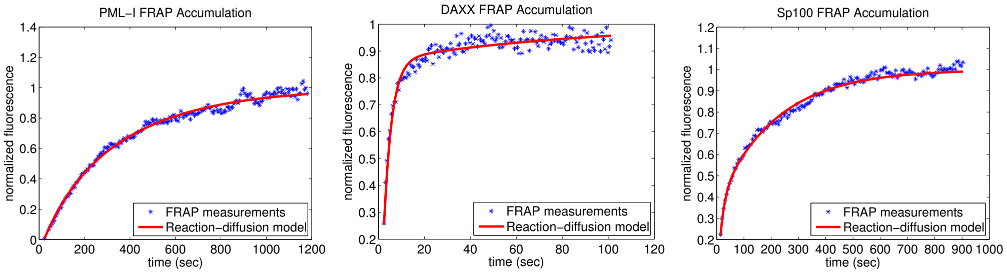


Figure 5.2: Time series data and original fits from Chapter 6 for the PML I (left), DAXX (middle) and Sp100 (right) datasets.

	DAXX		Sp100	
	$X_0 \longrightarrow X_f$	$X_0 \longrightarrow X_1$	$X_0 \longrightarrow X_f$	$X_0 \longrightarrow X_1$
$X_f \longrightarrow X_0$	45 (p=0.408)	55 (p=0.043)	22 (p=0.216)	24 (p=0.074)
$X_1 \longrightarrow X_0$	32 (p=0.065)	41 (p=0.792)	9 (p=0.040)	15 (p=0.679)

Table 5.2: Correlation of reactions in successful models for DAXX and Sp100. The count of common appearances of each pair of reactions involving X_0 show that there is no significant preference for any of the four topologies. Note that many evolved solutions contain more than one pair of X_0-reactions, e.g. a fully connected network contains all four pairs and is counted four times in this table. Details on the p-values are given in the text.

or X_1. This criterion seems to suggest that the PML I data is “easier” to explain, i.e. requires a less complex model to fit the data.

5.3.2 Building blocks and search space enumeration

Successful models are composed of four topologies The first criterion outlined above is fulfilled by all evolved solutions for all three datasets. However, the other two criteria outline four basic model topologies, which make up the building blocks of all solutions to the DAXX and Sp100 data. Besides the two-way reaction between X_f and X_1, these topologies consist of two additional reactions each, one producing X_0 either from X_f or X_1, and one consuming X_0 to produce either X_f or X_1. Many of the evolved solutions are actually superpositions of these building blocks, but the data shows that these topologies alone can explain the FRAP measurements.

No model topology is clearly preferred When counting the number of appearances of the four building blocks, it becomes clear that none of them is significantly

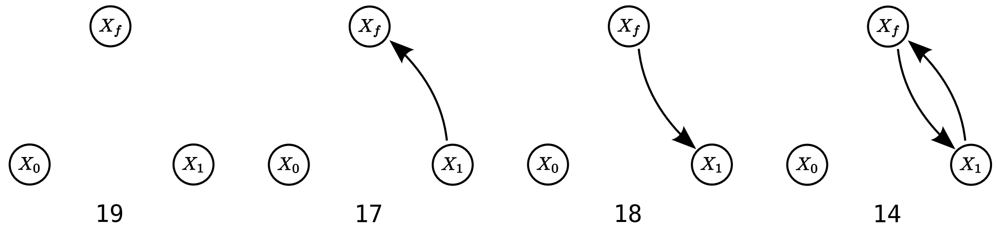


Figure 5.3: The four dominating network topologies result from 100 runs with neutral fitness. All other topologies appeared less than five times.

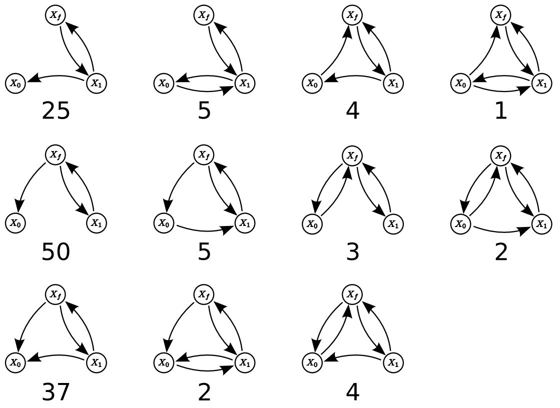


Figure 5.4: The network topologies resulting from 100 runs for the PML I dataset.

overrepresented (Table 5.2). For statistical testing, we employed the null hypothesis that all four pairs of reactions have equal probability, and calculated the probability for an observation that is equally or further away from the expected value than the observed values (the p-value). Interestingly, the patterns for DAXX and Sp100 are similar. While the pairings ($X_f \longrightarrow X_0$, $X_0 \longrightarrow X_f$) and ($X_1 \longrightarrow X_0$, $X_0 \longrightarrow X_1$) are certainly neither over- nor underrepresented, the pairing ($X_f \longrightarrow X_0$, $X_0 \longrightarrow X_1$) may be slightly overrepresented and the last pairing, ($X_1 \longrightarrow X_0$, $X_0 \longrightarrow X_f$), might be underrepresented. However, in no case does the statistical data allow a waterproof conclusion. Moreover, for both datasets, there seems to exist a preference for the reaction $X_f \longrightarrow X_0$ over $X_1 \longrightarrow X_0$, and for $X_0 \longrightarrow X_1$ over $X_0 \longrightarrow X_f$. Again, the statistic allows no certain conclusions. Summarising, there is no single preferred model topology that matches the data best.

Network enumeration divides search space into three fitness plateaus Restricting our search to networks with at most three species, we are in the fortunate situation to be able to enumerate the whole search space, which has been done

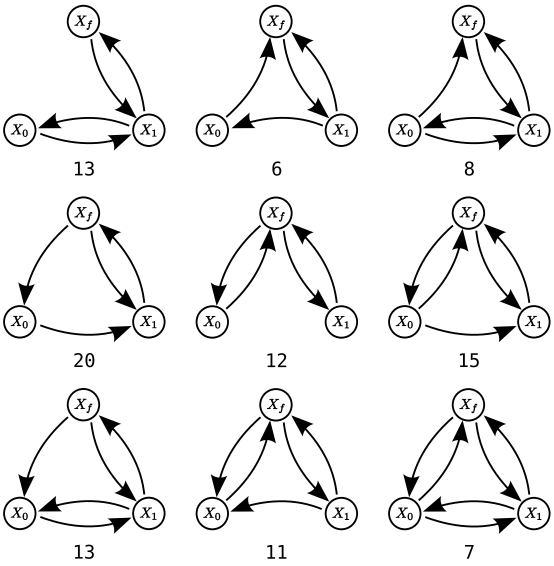


Figure 5.5: The network topologies resulting from 100 runs for the DAXX dataset.

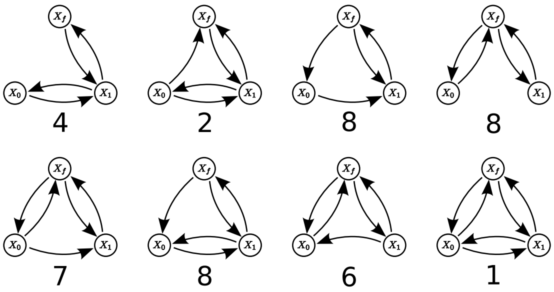


Figure 5.6: The network topologies resulting from 44 runs for the Sp100 dataset that passed the fitness criterion.

exemplary for PML I (Figure 5.8). Looking at the best possible fitness values (after parameter fitting) for all networks, the search space can apparently be divided into three classes: models with a very bad fitness (99.42) that do not contain a reaction converting X_f into either X_0 or X_1, and which thus simulate only the diffusional inflow into the ROI; models of mediocre fitness (5.69) which do contain reactions starting from X_f, but do not contain a feedback reaction into X_f that shuts off the inflow after a fast transient phase; and very good models (≤ 0.15) that always contain a (direct or indirect) feedback on X_f. In the last class, the models containing only a two-way reaction from X_f to either X_0 or X_1 provide a slightly worse fit than the models involving all three species and a feedback on X_f, in accordance to the result of the evolution-

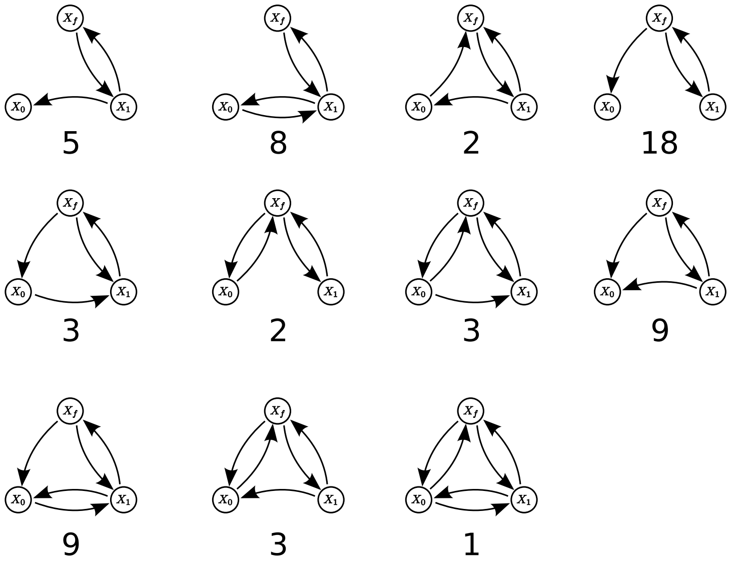


Figure 5.7: The network topologies resulting from 63 runs for the Sp100 dataset that did not pass the fitness criterion.

ary runs. Since no clear distinction in fitness can be made between the models of one (sub-)class, these classes form fitness plateaus in the search space that are difficult to traverse for an EA.

5.4 Discussion

Network design is paramount in modelling Modelling a biological system involves a number of design choices, mostly concerning the type of modelling strategy as well as the right model setup. Of course, the optimal modelling strategy depends on the available knowledge about the system. For systems in which time-course data is available, differential equations are a popular choice to describe the average dynamics - both in a stochastic and in a spatial sense - of the system. Once the modelling strategy is chosen, the specific model has to be created, based on both data and expert knowledge. Most of the time, the species to be included in the model and the reactions between them are determined by the available knowledge in the specific application area. The task of the modeller is then to represent this knowledge in the chosen framework.

Incomplete knowledge can be bridged by assumptions or inference However, nearly no system in molecular and cell biology is so thoroughly investigated that all involved species and reactions are known. On the contrary, a popular application of modelling is to explore "what-if" scenarios in the face of uncertain and incomplete knowledge. Therefore, the idea of automatic inference of a (minimal and realistic) model of a system carries a lot of appeal to modellers and has been explored in a variety of ways (Alves and Sorribas, 2007; Barrett and Palsson, 2006; Gadkar et al., 2005; Guthke et al., 2005; Koza et al., 2001a; Nachman et al., 2004; Steffen et al., 2002).

Goodness-of-fit is often insufficient to discriminate model topologies For the FRAP data of PML-related molecules that are the topic of Chapter 6, our initial investigation involved models with two populations of bound molecules, one tightly bound fraction and one loosely bound (top-right in Figure 5.5). However, after some discussion, this approach was replaced by a model in which the already bound molecules could move further inside the PML body, thus becoming more tightly bound (top-left in Figure 5.5). This choice was made for reasons of biological plausibility, not because one model provided a better fit than the other one. The investigations in this chapter reveal that both model topologies, together with several others, are equally suitable to fit the data. Since both models are also equally parsimonious, our results show conclusively that a goodness-of-fit criterion alone cannot decide between the two models, and that expert knowledge has to be included in deciding how to model the system.

Automatic inference can reveal necessary building blocks and alternatives Even if automatic inference of model topologies is often not able to provide the *one* best solution, it can provide hints about necessary components of the model which are required to fit the data. For the three datasets investigated in this chapter, the artificial evolution process revealed a two-way reaction that is absolutely necessary for a good fit, as well as four reaction pairs of which one has to be included in the model. Such knowledge, coming from the data only, can then be combined with additional biological aspects to create the most likely, parsimonious model that fits the data.

Fitness plateaus prevent effective structural evolution An exhaustive exploration of the whole search space for PML I (Figure 5.8) has shown two aspects: that the fitness landscape in the space of all possible networks consists basically of three levels, very bad, mediocre, and very good models, and that no distinction in terms of fitness is possible inside these classes. Of course, this is a bad prerequisite for an optimisation algorithm acting on the networks, since it cannot do much better than probing the search space randomly for models from a better level than the current one, and then sticking with the randomly found one. Keeping in mind that the search space in this chapter consisted of only three species, care has to be taken when transferring this result to higher-dimensional search spaces. More species - and thus, many more reactions between them - might break up the plateaus into smaller parts.

A network cannot be uniquely identified from its function Due to the potentially large classes of networks with similar function, identifying a network from its function without further knowledge is usually not possible. However, there are ways to overcome this problem. First of all, biological knowledge about the system under study can be converted to constraints on the search space, which would also limit the number networks with correct function. Additionally, networks can be evolved around a "core" network that is experimentally validated and assumed to be present *a priori* (see Section 3.5). After biological constraints and *a priori* knowledge have been incorporated, the goal must be to evolve as many different networks as possible with the required functionality, presenting the wetlab biologist with a set of alternative explanations for his or her data that can lead to further experiments and thus more data, which can in turn be used to reduce the number of appropriate network models.

5.5 Conclusion

In conclusion, it has to be stated that goodness-of-fit to time-series data is not a sufficient criterion to discriminate different model topologies, at least for the FRAP data used here. Since this data is already quite well-resolved in comparison to many other data sources, this conclusion carries some weight beyond the specific case under study. However, this does not prevent the use of structural optimisation techniques to find a set of suitable models, which can then be further investigated in the light of expert knowledge. Additionally, the search space of all possible networks was very restricted in this study, since we only allowed three species. It is not clear from our investigations how these results scale to larger search spaces. Even thought the plateau-like nature of the fitness landscape is probably preserved, plateaus might be comparatively smaller and the combined application of structure- and parameter-evolution could additionally help to prevent the EA from stalling in these regions.

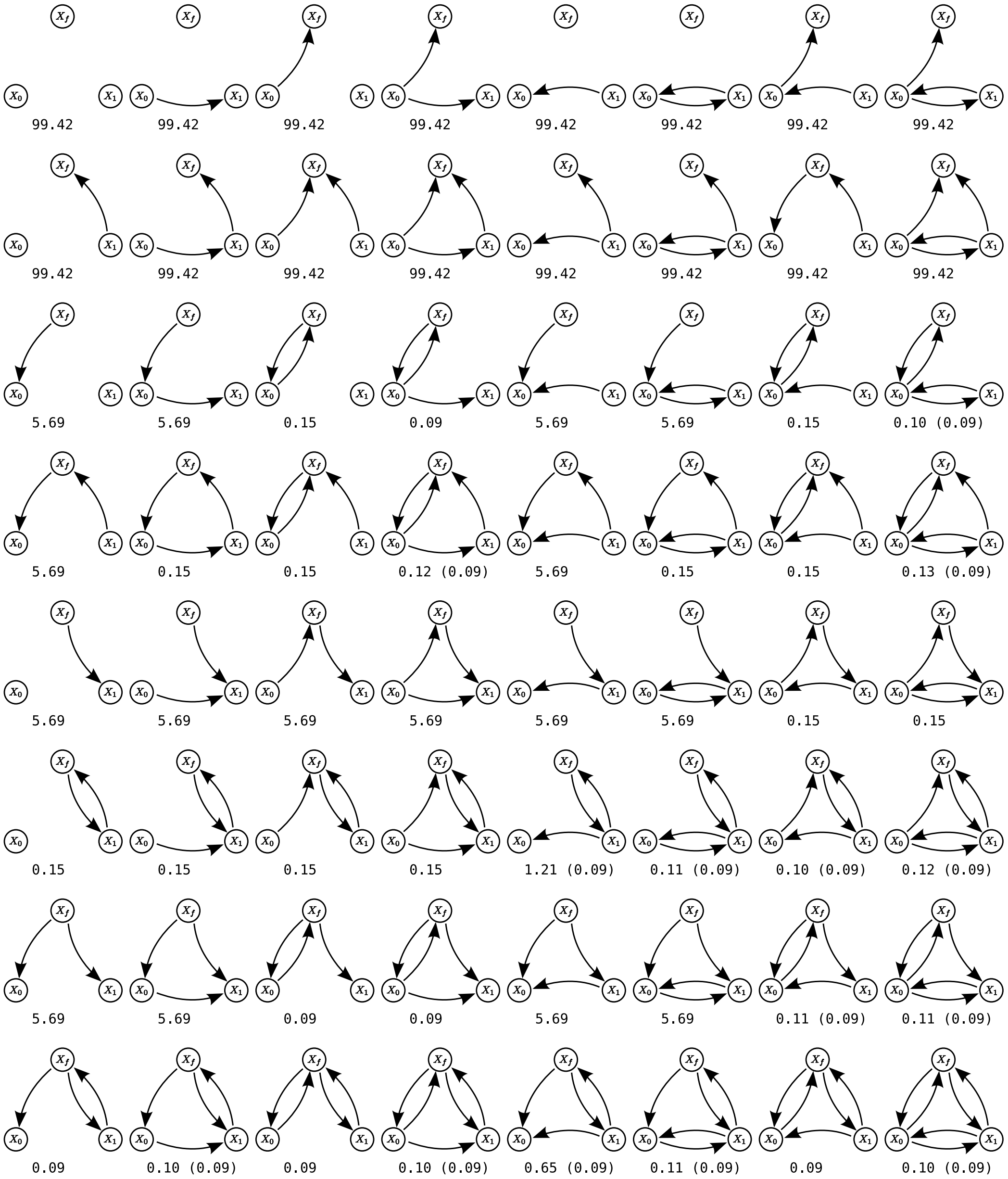


Figure 5.8: All possible network topologies with three species. The parameters of each network have been fit to the PML I data using a (30,30) CMA-ES, and the average fitness of ten runs is given under each network (together with the minimal fitness if it is different from the average).

6 Kinetic Analysis of Fluorescence Recovery Experiments

Chapter summary. *Fluorescence imaging methods provide a very powerful and attractive way of observing processes inside the nucleus of living cells. In this work, we have used fluorescence recovery after photobleaching (FRAP) and fluorescence correlation spectroscopy (FCS) to get a better understanding of the kinetics processes involved in PML nuclear body (NB) formation. We show that all six nuclear PML isoforms exhibit individual exchange rates at NBs and identify PML V as a scaffold subunit. Moreover, we detail the specific exchange kinetics of SP100, PML-RARα, HIPK2, DAXX and BLM at PML NBs. For the first time, dynamics of PML NB formation are measured in cells lacking endogenous PML, and we show that all six PML isoforms are able to form nuclear bodies under such conditions. Additionally, we show how the developed analysis framework can be extended to FRAP data of heterochromatin protein 1 (HP1) binding processes. The underlying mathematical model plays a crucial role in determining kinetic parameters. We present the specific model used in this contribution, and discuss possible shortcomings and extensions.*

> **Note:** The work presented in this chapter results from a collaboration between the author and PD Dr. Peter Hemmerich at the FLI Jena, and has been described in two publications (Brand et al., 2010; Weidtkamp-Peters et al., 2008). A third one is in preparation (Klement et al., 2010). The author's contribution to this work is the mathematical analysis and interpretation of the data. All cell-biological experiments were performed at the FLI and not by the author.

6. KINETIC ANALYSIS OF FLUORESCENCE RECOVERY EXPERIMENTS

Chapter contents

6.1 Introduction

Aim of this study In modern cellular biology, the understanding has emerged that many cellular processes cannot be understood in isolation, but must be viewed in their natural surrounding, the living cell. One popular way for such *in vivo* observations is fluorescence microscopy, in which fluorescent marker proteins such as green fluorescent protein (GFP) are excited with a laser beam and their photon emissions are recorded to yield an image of the protein distribution in the cell. Mathematical models of the underlying biological processes can then be used to assess their different quantities and characteristics, such as molecular concentrations, diffusion coefficients, and kinetic parameters. Such an analysis of fluorescence microscopy data forms the general aim of the work presented in this chapter.

6.1.1 Protein dynamics in the cell nucleus

The mammalian cell nucleus The nucleus is a dynamic, crowded environment that hosts the main cellular processes, such as cell fate decision, cell division, and apoptosis. Its spatial structure is highly ordered, containing the chromosomal territories and other internal domains such as nucleoli, speckles, Cajal bodies and promyelocytic leukaemia nuclear bodies (PML NBs) (Figure 6.1; Lamond and Earnshaw, 1998). This spatial organisation can change over time, depending on cell cycle state and additional external stimuli. Most of its functionality is understood to regulate biochemical activities on chromatin (Lanctôt et al., 2007).

PML bodies PML NBs are large protein assemblies that are found in the nucleus of mammalian cells. Typically, they have a sphere-like shape (Ascoli and Maul, 1991; Maul et al., 1995), their size ranges from 0.3-1.0 μm in diameter (Maul et al., 2000), and average copy numbers range between 10 and 20 depending on cell cycle status and cell type (Ascoli and Maul, 1991; Maul et al., 1995). In contrast to other macromolecules in the nucleus, the exact biochemical function of PML NBs is still unclear today (Bernardi and Pandolfi, 2007; Hemmerich and Diekmann, 2005), but they are implicated in apoptosis, cellular senescence, angiogenesis, and DNA damage response pathways (Dellaire and Bazett-Jones, 2004).

The name PML nuclear body is derived from the promyelocytic leukaemia gene product PML (Daniel et al., 1993), the protein characterising and distinguishing PML

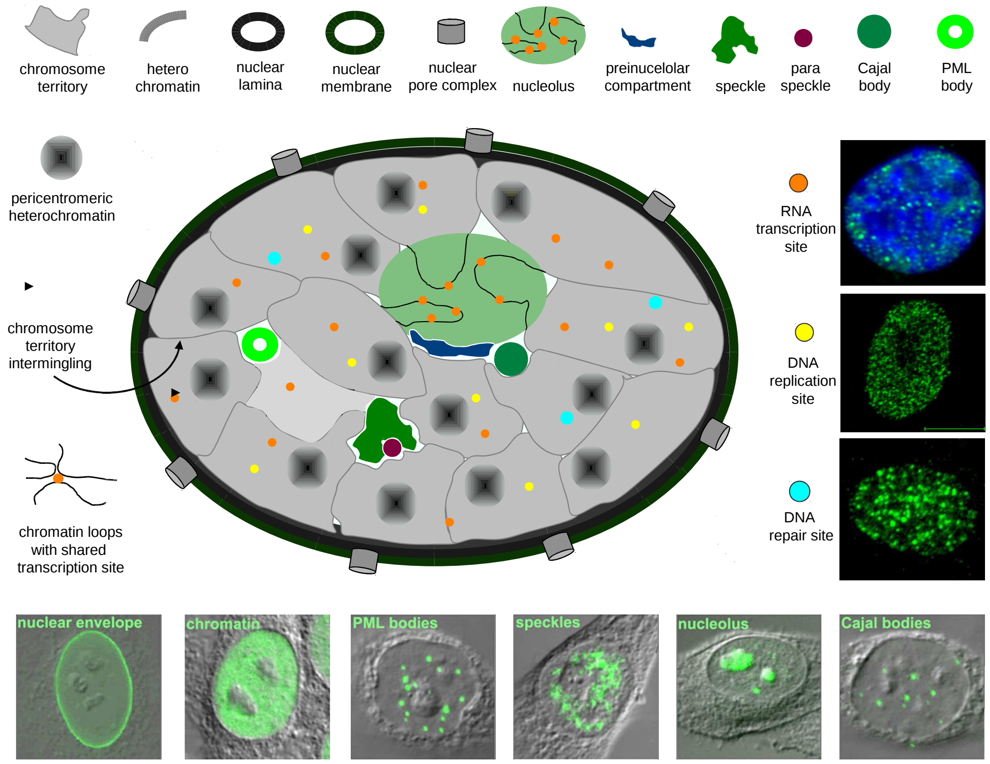


Figure 6.1: Structure and function in the mammalian cell nucleus. The centre image shows a schematic depiction of a typical mammalian cell nucleus. Focal sites of biochemical activity in the nucleus, such as mRNA transcription, DNA replication and DNA repair can be identified and visualised by immunofluoresence and confocal imaging (images on the right). In case of RNA transcription sites, the nuclear DNA was simultaneously stained with a DNA-binding dye (blue). A similar approach also reveals subnuclear domains, as shown in the bottom panels. Here, the confocal immunofluorescence signal was merged with the corresponding differential interference contrast (DIC) which detects cellular entities. The structures were visualised using specific antibodies against a nuclear pore protein (nuclear envelop), histones (chromatin), promyelocytic leukemia protein (PML bodies), splicing factor SC-35 (speckles), fibrillarin (nucleolus), and p80 coilin (Cajal bodies). Image courtesy of Peter Hemmerich, FLI Jena.

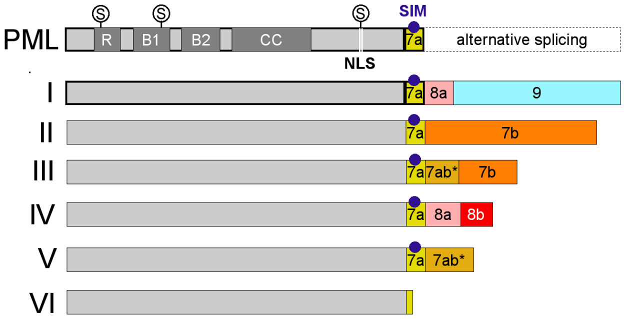


Figure 6.2: Schematic depiction of the domain structure of PML isoforms (data taken from Jensen et al. (2001)). All PML isoforms share a common N terminus but differ in their C termini, attributable to the alternative splicing of exons 7 to 9. Protein domains of all PML isoforms include the RING finger (R), the B1 and B2 boxes, the coiled-coil motif (CC), a nuclear localisation signal (NLS) and three SUMOylation sites (S). PML VI does not contain the SUMO interacting motif (SIM) within exon 7a.

NBs. Several isoforms exist (Figure 6.2), generated by alternative splicing, that are treated separately in this work (PML I-VI, Jensen et al., 2001). In acute promyelocytic leukaemia (APL), the name-giving disease, the PML gene is fused to the retinoic acid receptor α (RARα). Expression of PML-RARα induces leukaemia and disrupts PML NBs (reviewed in Melnick and Licht, 1999). Medical intervention with retinoic acid counters this effect, leading to PML reformation and recovery of APL patients (Koken et al., 1994).

Several proteins, such as SP100, DAXX, and BLM, are associated with PML NBs at endogenous expression levels (Negorev and Maul, 2001). Therefore, these proteins are included in this study, on top of the PML I-VI isoforms. Additionally, PML proteins contain three SUMOylation sites, which are essential for the self-assembly of PML NBs (Ishov et al., 1999; Zhong et al., 2000), and thus PML proteins with mutated SUMOylation sites were included as well. The current model proposes that specific nuclear proteins are titrated and/or posttranslationally modified at PML nuclear bodies, thereby modulating key nuclear biochemical activities (Bernardi and Pandolfi, 2007). It is therefore essential to acquire the biophysical properties of protein exchange at these nuclear subdomains.

Chromatin organisation and HP1 One of the earliest notions about the organisation of the nucleus was that there are two types of regions with different chromatin density in some cell cycle phases. During interphase, gene-rich euchromatin is de-condensed, whereas gene-poor heterochromatin stays condensed throughout the cell cycle. A prominent marker of the latter regions is an enrichment in heterochromatin protein 1 (HP1) (James and Elgin, 1986; Maison and Almouzni, 2004). This is a small and highly conserved protein that has homologues in a large range of organisms, from fission yeast to mammals (Singh et al., 1991). In mammals, three isoforms (HP1α, HP1β, HP1γ) are known. HP1 proteins consist of three domains: the chromodomain which binds H3K9 (histone 3 methylated at lysine 9, Lachner et al., 2001) and thus chromatin; a hinge region interacting (at least *in vitro*) with RNA, DNA, and chromatin; and a chromoshadow domain for dimerisation and interaction with other proteins.

This modular structure of HP1 has led to the hypothesis that HP1 is a structural modulator, able to modify chromatin structure and facilitate the assembly of larger macromolecular chromatin structures. Via its interaction partners, HP1 has been implicated in transcriptional regulation, telomere maintenance, DNA replication and repair, and nuclear organisation (Vermaak and Malik, 2009). Dimers of HP1 molecules can bind to two neighbouring histones in chromatin, thus strengthening and modifying the structural organisation of chromatin. However, HP1 is not stably bound, but subject to constant exchange with freely diffusing HP1 molecules (Cheutin et al., 2003; Festenstein et al., 2003; Schmiedeberg et al., 2004). By binding chromatin and many different nuclear proteins, some of which have also been implied in nuclear organisation, HP1 provides a means of self-organisation in the nucleus, in which the stable nuclear domains emerge from the transient binding interactions of their constituent proteins (Misteli, 2001).

6.1.2 Exploring the nucleus using fluorescence imaging techniques

Fluorescence recovery after photobleaching A widely used technique in modern molecular biology is the tagging of proteins of a specific type with fluorescent markers such as green fluorescent protein (GFP), mostly by fusing the GFP gene to the protein's cDNA in a vector construct. In a laser-scanning microscope, a laser-beam is then applied to excite the fluorescent material, thus providing a visual image of the cellular distribution of the tagged protein.

When subjected to repeated cycles of excitation and emission, fluorescent molecules eventually lose their ability to emit fluorescence, enabling the creation of photobleached spots by repeated application of a strong laser beam. Since nearly all proteins in the nucleus are highly mobile, fluorescence eventually returns to these dark areas. Measuring the flux of fluorescence into this region then yields the FRAP recovery curve, which can be analysed using mathematical models to yield kinetic parameters of the proteins under study (Carrero et al., 2003). FRAP itself has been established for more than three decades (Peters et al., 1974). However, only recently has the availability of cotranslational fluorescent marker proteins lead to a drastic increase in the use of this technique to measure protein dynamics in vivo.

In the first mathematical analysis of FRAP, Axelrod et al. (1976) developed an effective-diffusion model (without binding and unbinding processes) based on a 2D photobleach with a Gaussian intensity profile. For this case, they are able to provide an explicit solution for the FRAP curve, which can be fitted to the measurement in order to get the effective diffusion constant D_{eff}. Using this approach, Phair and Misteli (2000) measured effective diffusion coefficients for several molecules involved in the cell nucleus. Today, many FRAP models of processes in the cell nucleus assume that the proteins undergo diffusion as well as binding / unbinding events, both contributing to their spatial dynamics (Sprague and McNally, 2005).

Fluorescence correlation spectroscopy When the proteins of interest are tagged with a fluorescent marker, their concentration in the confocal volume of the microscope system can be tracked over time by illuminating this small volume with a laser and recording the fluorescence emission. Since the temporal resolution of these measurements is high, their temporal autocorrelation function can be computed:

$$G(\tau) = \frac{\langle \Delta I(t+\tau)\Delta I(t)\rangle}{\langle I(t)\rangle^2} = \frac{\langle I(t+\tau)I(t)\rangle}{\langle I(t)\rangle^2} - 1$$

where $I(t)$ is the fluorescence intensity recorded at time t, $\langle I(t)\rangle$ is its average value with respect to time t, and $\Delta I(t) = I(t) - \langle I(t)\rangle$ is the deviation from the average value. With this normalisation, the number of particles in the confocal volume is inversely related to the intercept value $G(0)$.

For obvious reasons, the analysis of correlation in confocal volume fluorescence is termed fluorescence correlation spectroscopy (FCS) (Elson and Magde, 1974; Magde et al., 1972). Using appropriate models, FCS can be used to extract kinetic information such as diffusion coefficients, not only in the case of standard diffusion, but also for anomalous diffusion or diffusion consisting of different particle fractions with different diffusion coefficients (Elson, 2004). Additionally, FCS can be used to extract binding parameters (Elson, 1985).

Raster image correlation spectroscopy The images taken by fluorescence microscopy are not snapshots taken at one precise moment (as in a CCD camera), but rather result from a laser beam linearly scanning the whole image area, typically line-by-line. Raster image correlation spectroscopy (RICS) exploits the fact that the time the beam spends at individual pixels, the movement time between pixels, and the movement time between lines are known.

Using these times, a comparison of adjacent pixels in a horizontal line yields correlation information on the microsecond time scale, while adjacent pixels on the vertical axis or the comparison of whole scan lines lead to a millisecond time scale (Brown et al., 2008; Digman et al., 2005). Moreover, a correlation analysis of subsequent images (temporal image correlation spectroscopy or t-ICS, Wiseman et al., 2000) operates on a timescale of seconds or more. From these correlations, the RICS procedure estimates protein concentrations and diffusion coefficients, but it can also be queried for information about binding / unbinding dynamics (Digman and Gratton, 2009). Special treatment is applied to deal with immobile supermolecular structures that dominate the correlation functions, as well as with slow cellular movement.

6.1.3 Mathematical modelling and analysis

"The eukaryotic nucleus and the control of chromatin function pose greater challenges for a systems biology modelling approach than many (of the) other cellular networks..." (Visser and Fell, 2007). *"The ultimate goal of a systems biology view of the cell nucleus is to understand genome function within the architectural framework of the nucleus. [...] Simulation is becoming indispensable for the analysis of the kinetics of nuclear processes."* (Gorski and Misteli, 2005). These two quotes exemplify the emerging consensus in the community that a further understanding of nuclear processes in the cell is only possible through a combination of in vivo experimental techniques and computer modelling.

A kinetic model is essentially a mathematical description of the hypothesised biological processes. The model is characterised by biophysical parameters, such as binding and release constants, residence times and diffusion coefficients. Using dynamic fluorescence microscopy data, the set of parameters is determined that results in the best model fit to the data, which serves as a test of the model as well as quantitative information about the parameters under study (Phair and Misteli, 2001). This approach has been successfully employed for example to set up a kinetic framework model for the RNA polymerase I transcription machinery in the nucleolus (Dundr et al., 2002) or the ordered recruitment of DNA repair factors (Politi et al., 2005).

Well-developed mathematical techniques are available to extract kinetic information from FRAP data (Beaudouin et al., 2006; Carrero et al., 2004; Lele and Ingber, 2006; Phair and Misteli, 2001; Sprague et al., 2004, Section 6.2). Given a kinetic model with experimentally derived parameter settings, questions can be asked about functions and interactions of different parts of the cellular machinery included in the model (e.g. Klingauf et al., 2006).

6.2 Methods: A mathematical model for FRAP experiments

> **Note:** The biochemical methods that were used and described in the original publication (Weidtkamp-Peters et al., 2008) are left out here, since none of them were applied by the author of this thesis.

6.2.1 Fluorescence correlation spectroscopy measurements

Fluorescence correlation spectroscopy (FCS) measurements were performed at 37°C on a LSM 510Meta/ConfoCor2 combi system using a C-Apochromat infinity-corrected 1.2 NA 40× water objective (Carl Zeiss, Jena, Germany) as described in detail elsewhere (Schmiedeberg et al., 2004; Weidtkamp-Peters et al., 2009). Briefly, GFP-tagged proteins were spot-illuminated with the 488 nm line of a 20 mW Argon laser at 5.5 A tube current attenuated by an acousto-optical tunable filter (AOTF) to 0.1%. The detection pinhole had a diameter of 70 μm and emission was recorded through a 505-nm long-path filter. For the measurements, 10×30 time series of 10 seconds each were recorded with a time resolution of 1 μsecond and then superimposed for fitting to an anomalous diffusion model in three dimensions with triplet function (Saxton, 2001) using Origin Software (OriginLab, Northhampton, MA). The D values and anomaly parameters were extracted from fit curves as previously described (Hemmerich et al., 2008).

6.2.2 Fluorescence recovery after photobleaching

Fluorescence recovery after photobleaching (FRAP) experiments were carried out on a Zeiss LSM 510Meta confocal microscope (Carl Zeiss, Jena, Germany) essentially as described before (Hemmerich et al., 2008). Five to ten images were taken before the bleach pulse and 50-200 images after the bleaching of regions of interest (ROIs) that contained one nuclear body (NB) each at 0.05% laser transmission to minimise scan bleaching. Image-acquisition frequency was adapted to the recovery rate of the respective GFP fusion protein. The pinhole was adjusted to 1 airy unit. Quantification of relative fluorescence intensities was done according to Schmiedeberg et al. (2004) using Excel (Microsoft, Redmond, WA) and Origin software (OriginLab, Northhampton, MA).

6.2.3 Initial conditions for the mathematical model

Post-bleach fluorescence distribution For the initial conditions of the model, we assume that at some time $t = 0$, the fluorescence intensity inside the ROI is photobleached to 0. Due to inevitable diffusion during the bleach process, the actual fluorescence in the ROI certainly does not reach 0. However, the raw data from the microscope has been normalised by subtracting background fluorescence, leaving an initial value of 0 as our best guess. When FRAP was applied on fixed cells, fluorescence within the ROI was indeed 0 (P. Hemmerich, unpublished observations), indicating that values above 0 originate from diffusion in the short time window after the end of the bleach pulse and the acquisition of the first image.

Starting from this assumption, we numerically solve the model equations until the sum of fluorescence from freely diffusing, loosely bound and tightly bound molecules in the ROI reaches the value of the first FRAP measurement. Due to a temporary "blinding" of the detector by the photobleach, the first FRAP measurement is always taken some time after the bleach and thus always truly positive. For this first point, we now have an approximation of the fluorescence distribution between the three fractions of molecules included in the model. Standard numeric integration (Section 6.2.5) continues from here to yield the rest of the estimated FRAP recovery curve.

Recently, the criticism has been applied that the initial fluorescence distribution is Gaussian in shape, not constant (Mueller et al., 2008). Even though this might principally enhance our results, we do not have enough data on this at the moment in order to get a more accurate post-bleach estimation. Thus, we apply the commonly used argument that when the bleached area is small enough, the Gaussian profile is approximated by a constant (Carrero et al., 2003; McGrath et al., 1998; Tardy et al., 1995). Braeckmans et al. (2003) call this the uniform-disc model, which can be applied in cases where the bleached disc is large in comparison to the laser-beam's point-spread-resolution, but not too large so that diffusion during bleaching does not play a significant role.

No significant diffusion during bleach phase In order to avoid errors in the parameter estimation, no significant recovery should occur during the bleach. In general, bleach time should be at least 15 times shorter than the characteristic recovery time (Braga et al., 2004; Meyvis et al., 1999). This is the case for the measurements treated here, as the binding / unbinding processes yield a very slow effective diffusion constant and thus a high characteristic recovery time. Therefore, we can safely assume that bleaching is instantaneous and no significant diffusion takes place during the bleach process.

The rapid, diffusion-dominated phase is independent of FRAP in our model In contrast to other approaches reported in the literature, we do not use FRAP data to estimate the diffusion coefficient D, but rather utilise FCS to measure D and subsequently include this measurement into our FRAP model. Since the first phase of the recovery curve is diffusion-dominated, this protocol reduces the sensitivity to the initial FRAP measurements. These may be more imprecise than the later ones due to interference effects of the bleach beam, diffusion during the bleach process, and the approximation of the initial fluorescence distribution.

6.2.4 Relation of local diffusion coefficient to net flux into ROI

In this section, we show how the diffusion coefficient D in Fick's Law (measured by FCS here) is related to the the net flux (exchange rate) between compartments, which is needed to quantify the inflow of fluorescent material into the region of interest (ROI) of the FRAP model. The solution of the diffusion equation on a disk is adapted from the MathWorld website [1].

Problem statement We assume a circular compartment with radius r_0 in 2D (a disk), which is embedded in an infinite space with constant concentration (or relative fluorescence intensity, RFI) of c_0. We will solve the diffusion equation (more commonly called heat equation or Fick's second law)

$$\frac{\partial C(x,y,t)}{\partial t} = D\Delta C(x,y,t).$$

Since we are interested in the solution on a disk with radial symmetry, we formulate the problem in terms of polar coordinates r and θ with

$$x = r\cos(\theta),$$

$$y = r\sin(\theta).$$

[1] Weisstein, Eric W. Heat Conduction Equation–Disk. From *MathWorld*–A Wolfram Web Resource. `http://mathworld.wolfram.com/HeatConductionEquationDisk.html`, last accessed 26.11.08

Separable solution We will search for a separable solution using

$$C(r,\theta,t) = R(r)\Theta(\theta)T(t).$$

The Laplacian operator can be written in polar coordinates

$$\Delta = \frac{d^2R}{dr^2} + \frac{1}{r}\frac{dR}{dr} + \frac{1}{r^2}\frac{d^2\Theta}{d\theta^2},$$

so that the diffusion equation has the form

$$\frac{R\Theta}{D}\frac{dT}{dt} = \frac{d^2R}{dr^2}\Theta T + \frac{1}{r}\frac{dR}{dr}\Theta T + \frac{1}{r^2}\frac{d^2\Theta}{d\theta^2}RT.$$

We assume that the bleach pulse annihilates all fluorescence in the ROI, so the initial condition inside the disk is

$$C(r,\theta,0) = 0, r < r_0.$$

Since the surrounding area has constant RFI c_0, we additionally have

$$C(r_0,\theta,t) = c_0.$$

Following the derivation at the Mathworld website, we arrive at the solution

$$C(r,\theta,t) = C(r,t) = c_0 - 2c_0 \sum_{n=1}^{\infty} \frac{J_0(\frac{\alpha_n r}{r_0})}{\alpha_n J_1(\alpha_n)} e^{-\alpha_n^2 Dt/r_0^2}, \qquad \boxed{6.1}$$

where J_0 and J_1 are Bessel functions of the first kind and α_n is the nth positive root of $J_0(x)$ (Bowman, 1958, pp. 37-39).

Net flux into compartment Our aim now is to get an expression for the net flux of concentration at the whole compartment boundary in terms of the average concentration inside and outside the compartment. To allow for an analytical treatment, in the following we assume that no other reaction takes place inside the compartment, and thus $C(r,t)$ accurately describes the concentration at time t and distance r from the centre of the disk. We first derive an expression for the average concentration inside the compartment by integration in polar coordinates.

$$\bar{C}(t) = \frac{1}{\pi r_0^2} \int_0^{2\pi} \int_0^{r_0} (c_0 - 2c_0 \sum_{n=1}^{\infty} \frac{J_0(\frac{\alpha_n r}{r_0})}{\alpha_n J_1(\alpha_n)} e^{-\alpha_n^2 Dt/r_0^2}) r dr d\theta$$

Note that the additional factor r in the above equation comes from the transformation from Cartesian to polar coordinates. The integrand is independent of θ, so we arrive at

$$\begin{aligned}\bar{C}(t) &= \frac{2}{r_0^2}\int_0^{r_0}(c_0 - 2c_0\sum_{n=1}^{\infty}\frac{J_0(\frac{\alpha_n r}{r_0})}{\alpha_n J_1(\alpha_n)}e^{-\alpha_n^2 Dt/r_0^2})rdr \\ &= \frac{2}{r_0^2}\int_0^{r_0} c_0 r dr - \frac{4c_0}{r_0^2}\int_0^{r_0}\sum_{n=1}^{\infty}\frac{rJ_0(\frac{\alpha_n r}{r_0})}{\alpha_n J_1(\alpha_n)}e^{-\alpha_n^2 Dt/r_0^2}dr \\ &= c_0 - \frac{4c_0}{r_0^2}\sum_{n=1}^{\infty}\frac{e^{-\alpha_n^2 Dt/r_0^2}}{\alpha_n J_1(\alpha_n)}\int_0^{r_0} rJ_0(\frac{\alpha_n r}{r_0})dr\end{aligned}$$

Using the integral identity $\int_0^u u'J_0(u')du' = uJ_1(u)$, and substituting $u' = \alpha_n r/r_0$, we get

$$\int_0^{r_0} rJ_0(\frac{\alpha_n r}{r_0})dr = \frac{r_0^2}{\alpha_n}J_1(\alpha_n).$$

Thus, we have

$$\begin{aligned}\bar{C}(t) &= c_0 - \frac{4c_0}{r_0^2}\sum_{n=1}^{\infty}\frac{e^{-\alpha_n^2 Dt/r_0^2}}{\alpha_n J_1(\alpha_n)}\frac{r_0^2}{\alpha_n}J_1(\alpha_n) \\ &= c_0 - 4c_0\sum_{n=1}^{\infty}\frac{e^{-\alpha_n^2 Dt/r_0^2}}{\alpha_n^2}.\end{aligned}$$

Since we have a solution for $C(r,t)$, we can also calculate the first derivative at the boundary of the disk.

$$\frac{\partial C(r,t)}{\partial r} = -2c_0\sum_{n=1}^{\infty}\frac{e^{-\alpha_n^2 Dt/r_0^2}}{\alpha_n J_1(\alpha_n)}\frac{d}{dr}J_0(\frac{\alpha_n r}{r_0})$$

Using the derivative identity for Bessel functions $\frac{d}{dx}[x^m J_m(x)] = x^m J_{m-1}(x)$ as well as $J_{-m}(x) = (-1)^m J_m(x)$ and substituting $\alpha_n r/r_0$ for x, we get

$$\frac{d}{dr}J_0(\frac{\alpha_n r}{r_0}) = -\frac{\alpha_n}{r_0}J_1(\frac{\alpha_n r}{r_0}),$$

and thus

$$\frac{\partial C(r,t)}{\partial r} = 2c_0\sum_{n=1}^{\infty}\frac{e^{-\alpha_n^2 Dt/r_0^2}}{r_0 J_1(\alpha_n)}J_1(\frac{\alpha_n r}{r_0}).$$

At the boundary $r = r_0$, we thus have

$$\frac{\partial C}{\partial r}(r_0,t) = \frac{2c_0}{r_0}\sum_{n=1}^{\infty}e^{-\alpha_n^2 Dt/r_0^2}.$$

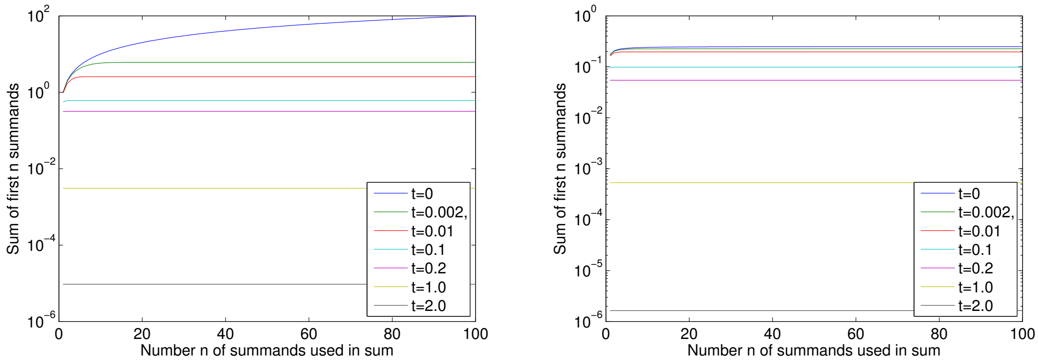


Figure 6.3: Cumulative contribution of the first n summands in the sums S_1 (left) and S_2 (right), for different points in time. After $t = 0.2$, the contribution of any summands beyond the first one is negligible.

Since we want to relate $\partial C/\partial r(r_0, t)$ with $\bar{C}(t)$, we need to compare the infinite sums in both terms. We define

$$S_1 = \sum_{n=1}^{\infty} e^{-\alpha_n^2 Dt/r_0^2}$$

and

$$S_2 = \sum_{n=1}^{\infty} \frac{e^{-\alpha_n^2 Dt/r_0^2}}{\alpha_n^2}.$$

The summands in S_1 and S_2 differ, so a closed form solution of their ratio S_1/S_2 cannot be expected. Luckily, for values of t such that $Dt/r_0^2 > 0.2$, the only significant contribution to the sums comes from their first summand, respectively (Fig. 6.3). When only the first summand of each respective sum is considered, we arrive at an approximate ratio of

$$S_1/S_2 \approx \alpha_1^2.$$

Now we can write

$$\bar{C}(t) = c_0 - 4c_0 S_2(t) \approx c_0 - 4c_0 S_1/\alpha_1^2$$

and thus

$$S_1 \approx \frac{\alpha_1^2}{4c_0}(c_0 - \bar{C}(t)),$$

$$\frac{\partial C}{\partial r}(r_0, t) = \frac{2c_0}{r_0} S_1 \approx \frac{\alpha_1^2}{2r_0}(c_0 - \bar{C}(t)).$$

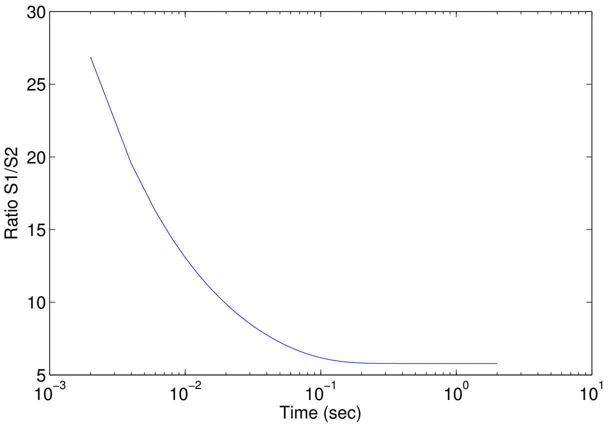


Figure 6.4: The ratio of sums S_1/S_2, approximated using the first 100 summands. Note the logarithmic x-axis. After 0.2 seconds, the ratio is constant with a value of α_1^2.

It is important to remember that the approximation is almost exact for $Dt/r_0^2 > 0.2$, but is increasingly incorrect for $t \to 0$. This is expected, since for small t, the derivative naturally tends to infinity, while the average concentration converges to zero.

Using Green's theorem, we can now derive an approximate expression for the net flux into the disk:

$$\begin{aligned}\frac{1}{\pi r_0^2}\iint D\Delta C(x,y,t)dxdy &= \frac{1}{\pi r_0^2}\int_0^{2\pi r_0} D\frac{d}{dr}C(r_0,t)ds\\ &\approx \frac{D\alpha_1^2}{r_0^2}(c_0-\bar{C}(t)) \qquad \boxed{6.2}\end{aligned}$$

Error estimation Now we will estimate the error we make with this approximation. From Figure 6.4, we see that $S_1/S_2 > \alpha_1^2$. From this, it is straightforward to derive

$$\frac{\partial C}{\partial r}(r_0,t) = \frac{2c_0}{r_0}S_1 > \frac{\alpha_1^2}{2r_0}(c_0-\bar{C}(t),$$

an thus our approximation for the derivative in normal direction is a systematic underestimation of the actual derivative value. From a comparison of the analytical solution $\boxed{6.1}$ for $\bar{C}$ with the integral of the approximate net flux $\boxed{6.2}$ (Figure 6.5), it again is obvious that the approximation underestimates the diffusive inflow of fluorescent material.

When we use the actual concentration inside the compartment $c(t)$ in the expression for the net flux approximation, and integrate this expression with respect to time, we get the solution

$$c(t) = c_0 - c_0 e^{-\alpha_1^2 D/r_0^2 t}$$

for the concentration resulting from pure diffusive inflow. Now we can calculate the error $E(t)$ between the approximated concentration $c(t)$ and the exact solution of the diffusion equation $\bar{C}(t)$:

$$\begin{aligned} E(t) &= \bar{C}(t) - c(t) \\ &= c_0 - 4c_0 \sum_{n=1}^{\infty} \frac{e^{-\alpha_n^2 Dt/r_0^2}}{\alpha_n^2} - c_0 + c_0 e^{-\alpha_1^2 D/r_0^2 t} \\ &= c_0 e^{-\alpha_1^2 D/r_0^2 t} - 4c_0 \sum_{n=1}^{\infty} \frac{e^{-\alpha_n^2 Dt/r_0^2}}{\alpha_n^2} \end{aligned}$$

Now we can derive the total error up to time T:

$$\int_0^T E(t)dt = \frac{c_0 r_0^2}{D\alpha_1^2}(1 - e^{-\alpha_1^2 DT/r_0^2}) - \frac{4c_0 r_0^2}{D} \sum_{n=1}^{\infty} \frac{1}{\alpha_n^4}(1 - exp^{-\alpha_n^2 DT/r_0^2})$$

With $T \to \infty$, we get

$$\tilde{E} = \int_0^{\infty} E(t)dt = \frac{c_0 r_0^2}{D}(\frac{1}{\alpha_1^2} - 4\sum_{n=1}^{\infty} \frac{1}{\alpha_n^4})$$

showing that the error increases linearly with c_0 and quadratically with r_0, but decreases inversely proportional to D.

Introducing a correction factor Finally, we will modify the diffusion constant D in the net flux gradient to $D' = mD$, such that the accumulated error $\tilde{E} = 0$. Replacing D with mD in the terms for $c(t)$ above, we derive an expression for the error

$$\tilde{E} = \frac{c_0 r_0^2}{D}(\frac{1}{m\alpha_1^2} - 4\sum_{n=1}^{\infty} \frac{1}{\alpha_n^4}).$$

For $\tilde{E} = 0$, we thus determine

$$m = (4\alpha_1^2 \sum_{n=1}^{\infty} \frac{1}{\alpha_n^4})^{-1} \approx 1.3856$$

It is interesting to note that this correction factor is independent of c_0, r_0 and D, although it is only valid to disk-shaped compartments.

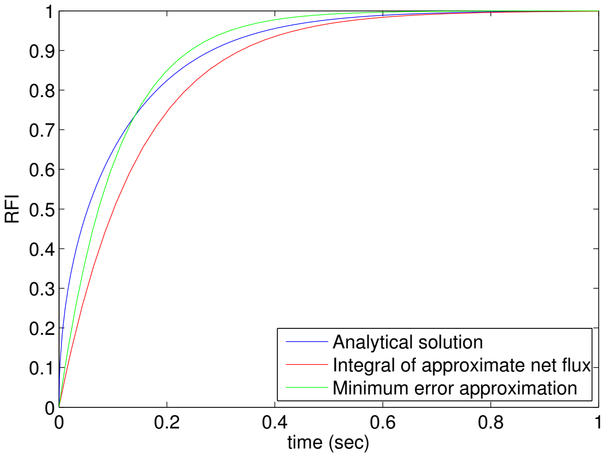


Figure 6.5: Comparison of the analytical solution for the RFI inside the compartment (blue) with the integral of the approximate net flux $D\alpha_1^2/r_0^2(c_0 - \bar{C}(t))$ (red) and the modified approximation with minimum error (green). Parameter values are as for PML I: $D = 1.85$, $r_0 = 1.25$, $c_0 = 1.0$.

To test the validity of our mathematical derivation, we have compared the analytical solution (6.1) to a partial differential equation (PDE) simulation of the spherical diffusion problem using Matlab$^{\mathrm{TM}}$(2008a, The MathWorks, Natick, MA). The results show that the analytical solution using the first 1000 terms of the sum of Bessel functions accurately describes the behaviour of the PDE system (Figure 6.6).

The diffusion term in the FRAP model In this section, we have given a rigorous derivation of the linear diffusion term in the compartmental FRAP model, and detailed its relation to the actual solution of the diffusion equation on a sphere. Of course, with modern computing power it would be possible to use a much more accurate approximation in the compartmental model, involving more than just the first components of the sums of Bessel functions. However, there are two reasons not to do so.

Firstly, when fitting data to an ODE model rather than simulating it once, computational constraints do play a role indeed. When hundreds or thousands of fitness evaluations - and thus model simulations - are needed, every possible speed-up is welcome. Secondly, a FRAP model involves reaction processes inside the ROI, in addition to diffusion. With these processes, the analytical solution derived above becomes invalid and has to be replaced by a differential equation description. The linear formulation is

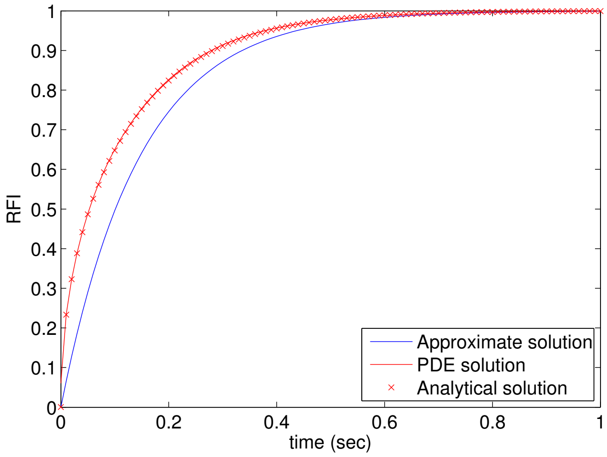


Figure 6.6: Comparison of the analytical solution for the RFI inside the compartment (red crosses) with result of a PDE simulation using Matlab (red line) and the approximate solution (blue). Parameter values are as for PML I: $D = 1.85$, $r_0 = 1.25$, $c_0 = 1.0$.

both obviously straightforward and - following the reasoning above - mathematically justified.

When comparing the approximate solution (6.2) to the analytical one (6.1), it is important to recognise that the actual error might be smaller than discussed above. Rather than bleaching the ROI 100% and not changing the directly adjacent area, the bleach profile of a laser pulse is better approximated by the exponential of a Gaussian function (Blonk et al., 1993; Braeckmans et al., 2003; Braga et al., 2004; Mazza et al., 2007), and thus does not have a sharp transition between zero and one at the boundary of the ROI. Therefore, the instantaneous diffusional inflow after the bleach pulse is in reality not as sharp as the analytical solution (6.1).

Finally, it is interesting to speculate about the behaviour of the diffusion equation on non-spherical domains. It can be expected that - whenever mathematically tractable - first-order approximations to analytical solutions will also be of the linear form $kD(x_{\text{outside}} - x_{\text{inside}}(t))$, but the constant k is most likely a different one. A simulation study of the dependence of this constant on the shape and size of the compartment would be very informative here, but does not fit in the scope of this chapter.

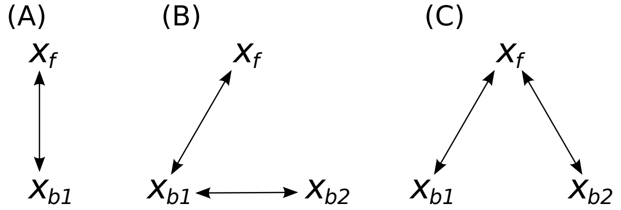


Figure 6.7: The three models used to analyse the FRAP data. (A) one-step binding / unbinding process. (B) two-step process with sequential steps. (C) two-step process with parallel, independent steps.

6.2.5 Numerical reaction-diffusion models for FRAP analysis

To analyse the FRAP recovery curves and to relate them to kinetic parameters of the proteins under study, mathematical models of the hypothesised binding behaviour have been been fitted to the FRAP curves. For these purposes, we used three model structures: a simple one-step binding / unbinding process, a two-step process with sequential binding steps, and a two-step process with parallel, independent binding steps (Figure 6.7).

Different simplifications for FRAP models, such as reaction-dominant and diffusion-dominant models, are discussed in the literature (Sprague et al., 2004). However, the vast array of different proteins under study here leads to the expectation that only the full model can adequately describe their dynamics, and thus no reduced model variant is used. Erroneously neglecting diffusion can lead to parameter estimation errors of several orders of magnitude (Sprague et al., 2006).

Observable fluorescence and diffusion from outside the ROI The experimental set-up provided that bleached ROI and FRAP ROI are similar and, so, are assumed to be the same for modelling purposes. Diffusion inside and out of the ROI was modelled as a linear two-way process with a modified diffusion constant D' (according to Section 6.2.4) based on the diffusion constant D measured by FCS. This constant yields the effective exchange rate through the boundary of a circular area.

The model system describes three variables representing the different fractions of fluorescent protein: free diffusion (x_f), bound in the first binding state (x_b1), and bound in the second binding state (x_b2). The one-step model only contains x_f and x_b1. Since these different fractions cannot be distinguished directly in FRAP, the observable amount of fluorescence is given by

$$x_{\mathrm{obs}}(t) = x_{\mathrm{f}}(t) + x_{\mathrm{b1}}(t)$$

or

$$x_{\mathrm{obs}}(t) = x_{\mathrm{f}}(t) + x_{\mathrm{b1}}(t) + x_{\mathrm{b2}}(t),$$

respectively.

The ratio between background fluorescence and fluorescence inside the PML-body (p) was determined individually for each type of molecule by using confocal microscopy and pixel intensity evaluation with MetaMorph software (Molecular Devices, Sunnyvale, CA). Since the FRAP values inside the ROI are normalised to 1.0 in steady state, the p values can be used to compute the relative concentration of free protein outside the ROI, namely $x_{\mathrm{f}}^{outside} = 1/p$.

Binding / unbinding processes The binding / unbinding processes are represented with mass-action kinetics, which leads to the following ODE representations: For the one-step model (Figure 6.7A)

$$\begin{aligned}
\frac{dx_{\mathrm{f}}}{dt} &= k_{\mathrm{off1}}x_{\mathrm{b1}} - k_{\mathrm{on1}}x_{\mathrm{f}} + D'(x_{\mathrm{f}}^{SS} - x_{\mathrm{f}}) \\
\frac{dx_{\mathrm{b1}}}{dt} &= k_{\mathrm{on1}}x_{\mathrm{f}} - k_{\mathrm{off1}}x_{\mathrm{b1}}
\end{aligned}$$

For the sequential two-step model (Figure 6.7B)

$$\begin{aligned}
\frac{dx_{\mathrm{f}}}{dt} &= k_{\mathrm{off1}}x_{\mathrm{b1}} - k_{\mathrm{on1}}x_{\mathrm{f}} + D'(x_{\mathrm{f}}^{SS} - x_{\mathrm{f}}) \\
\frac{dx_{\mathrm{b1}}}{dt} &= k_{\mathrm{on1}}x_{\mathrm{f}} - k_{\mathrm{off1}}x_{\mathrm{b1}} + k_{\mathrm{off2}}x_{\mathrm{b2}} - k_{\mathrm{on2}}x_{\mathrm{b1}} \\
\frac{dx_{\mathrm{b2}}}{dt} &= k_{\mathrm{on2}}x_{\mathrm{b1}} - k_{\mathrm{off2}}x_{\mathrm{b2}}.
\end{aligned}$$

And for the parallel two-step model (Figure 6.7C)

$$\begin{aligned}
\frac{dx_{\mathrm{f}}}{dt} &= k_{\mathrm{off1}}x_{\mathrm{b1}} - k_{\mathrm{on1}}x_{\mathrm{f}} + k_{\mathrm{off2}}x_{\mathrm{b2}} - k_{\mathrm{on2}}x_{\mathrm{f}} + D'(x_{\mathrm{f}}^{SS} - x_{\mathrm{f}}) \\
\frac{dx_{\mathrm{b1}}}{dt} &= k_{\mathrm{on1}}x_{\mathrm{f}} - k_{\mathrm{off1}}x_{\mathrm{b1}} \\
\frac{dx_{\mathrm{b2}}}{dt} &= k_{\mathrm{on2}}x_{\mathrm{f}} - k_{\mathrm{off2}}x_{\mathrm{b2}}.
\end{aligned}$$

In steady state, i.e. with $x_{\mathrm{f}} + x_{\mathrm{b1}} + x_{\mathrm{b2}} = 1$ and $dx_{\mathrm{f}}/dt = dx_{\mathrm{b1}}/dt = dx_{\mathrm{b2}}/dt = 0$, it is straightforward to solve for the relative proportions of the three binding states (Table 6.1).

	one-step	two-step parallel	two-step sequential
$x_{\mathrm{f}}^{SS} =$	$\frac{k_{\mathrm{off1}}}{k_{\mathrm{on1}}+k_{\mathrm{off1}}}$	$\left(1+\frac{k_{\mathrm{on1}}}{k_{\mathrm{off1}}}+\frac{k_{\mathrm{on2}}}{k_{\mathrm{off2}}}\right)^{-1}$	$\left(1+\frac{k_{\mathrm{on1}}}{k_{\mathrm{off1}}}+\frac{k_{\mathrm{on1}}}{k_{\mathrm{off1}}}\frac{k_{\mathrm{on2}}}{k_{\mathrm{off2}}}\right)^{-1}$
$x_{\mathrm{b1}}^{SS} =$	$\frac{k_{\mathrm{on1}}}{k_{\mathrm{on1}}+k_{\mathrm{off1}}}$	$\left(1+\frac{k_{\mathrm{off1}}}{k_{\mathrm{on1}}}+\frac{k_{\mathrm{off1}}k_{\mathrm{on2}}}{k_{\mathrm{on1}}k_{\mathrm{off2}}}\right)^{-1}$	$\left(1+\frac{k_{\mathrm{off1}}}{k_{\mathrm{on1}}}+\frac{k_{\mathrm{on2}}}{k_{\mathrm{off2}}}\right)^{-1}$
$x_{\mathrm{b2}}^{SS} =$		$\left(1+\frac{k_{\mathrm{off2}}}{k_{\mathrm{on2}}}+\frac{k_{\mathrm{on1}}k_{\mathrm{off2}}}{k_{\mathrm{off1}}k_{\mathrm{on2}}}\right)^{-1}$	$\left(1+\frac{k_{\mathrm{off2}}}{k_{\mathrm{on2}}}+\frac{k_{\mathrm{off1}}}{k_{\mathrm{on1}}}\frac{k_{\mathrm{off2}}}{k_{\mathrm{on2}}}\right)^{-1}$

Table 6.1: Equilibrium distribution of the proteins into the three binding states, summarised for all three models.

Residence time An important quantity to describe the binding behaviour in the FRAP models is the residence time, i.e. the average time until a newly bound molecule returns to the state of free diffusion. For the one-step and the parallel model, the residence time is

$$Rt = 1/k_{\mathrm{off1}}$$

and

$$Rt = \frac{k_{\mathrm{on1}}}{k_{\mathrm{on1}}+k_{\mathrm{on2}}}\frac{1}{k_{\mathrm{off1}}} + \frac{k_{\mathrm{on2}}}{k_{\mathrm{on1}}+k_{\mathrm{on2}}}\frac{1}{k_{\mathrm{off2}}},$$

respectively. For the sequential model, this is not straightforward to calculate, since a bound molecule can switch back and forth between the weakly bound and strongly bound state before going back to free diffusion.

To calculate the residence time Rt for the sequential model, a three-state Markov chain was considered, consisting of states f (free diffusion), w (weakly bound) and s (strongly bound). For a single, newly bound molecules (automatically in state w), the binding/unbinding rates lead to the following transition probabilities:

$$\begin{aligned} P(w \to f) &= \frac{k_{\mathrm{off1}}}{k_{\mathrm{off1}}+k_{\mathrm{on2}}} \\ P(w \to s) &= \frac{k_{\mathrm{on2}}}{k_{\mathrm{off1}}+k_{\mathrm{on2}}} \end{aligned}$$

A molecule in state w can move into state f directly, or it can visit state s before returning to state w. The time until the first transition occurs follows an Exponential distribution with rate parameter $k_{\mathrm{off1}}+k_{\mathrm{on2}}$ (Gillespie 1977), thus the mean time until any of the two transitions occur is given by $1/(k_{\mathrm{off1}}+k_{\mathrm{on2}})$. For a molecule in state s, the mean time until the (only possible) transition back to w is given by $1/k_{\mathrm{off2}}$. Thus, the residence time Rt is recursively defined by

$$Rt = \frac{k_{\mathrm{off1}}}{k_{\mathrm{off1}}+k_{\mathrm{on2}}}\frac{1}{k_{\mathrm{off1}}+k_{\mathrm{on2}}} + \frac{k_{\mathrm{on2}}}{k_{\mathrm{off1}}+k_{\mathrm{on2}}}\left(\frac{1}{k_{\mathrm{off1}}+k_{\mathrm{on2}}} + \frac{1}{k_{\mathrm{off2}}} + Rt\right).$$

Eliminating the recursion, we find

$$Rt = \frac{1}{k_{\text{off1}}} \left(1 + \frac{k_{\text{on2}}}{k_{\text{off2}}} \right).$$

It agrees with common sense that if k_{on2} is zero, the formula collapses to the residence time for a single binding/unbinding reaction, while in the case that either k_{off1} or k_{off2} approaches zero, the residence time becomes infinitely large.

Numerical integration, initial conditions, parameter fitting The mathematical model is numerically solved using the ode45 method in MATLAB (2008a, The MathWorks, Natick, MA), which uses an explicit Runge-Kutta (4,5) formula, the Dormand-Prince pair (Dormand and Prince, 1980). Assuming an ideal bleach, all three variables are initially set to zero. The timelines of the recorded data and the model predictions are then synchronised by aligning the time at which the sum of all three variables in the model prediction reaches the first measured fluorescence value.

Parameter fitting is done by minimising the sum (over time) of squared deviations between the fluorescence measurements and the predictions provided by the model, i.e. the sum of all three variables. Even though the model is not too large, parameter fitting is difficult due to potential local optima and the non-linear relation between parameters and model predictions. Therefore, an evolution strategy with covariance matrix adaptation (CMA-ES) was employed (Hansen and Ostermeier, 2001), using the MATLAB implementation by Nikolaus Hansen[1] with default parameters.

> **Note:** In our original paper (Weidtkamp-Peters et al., 2008), we used a slightly different model description. There, the protein concentrations in the model were normalised by the fluorescence outside of the ROI, which is not the case here. However, this difference in representation does not lead to different results. Additionally, after we published the PML results (Weidtkamp-Peters et al., 2008), we introduced a new method to compute the diffusive influx into the ROI (Section 6.2.4). This new method is used here, leading to slightly altered results compared to the original paper.

Model specialisation for PML FRAP data For the mathematical model of PML NB kinetics, the structural complexity of a PML NB has been approximated

[1] Downloaded from `http://www.lri.fr/~hansen/cmaes_inmatlab.html`

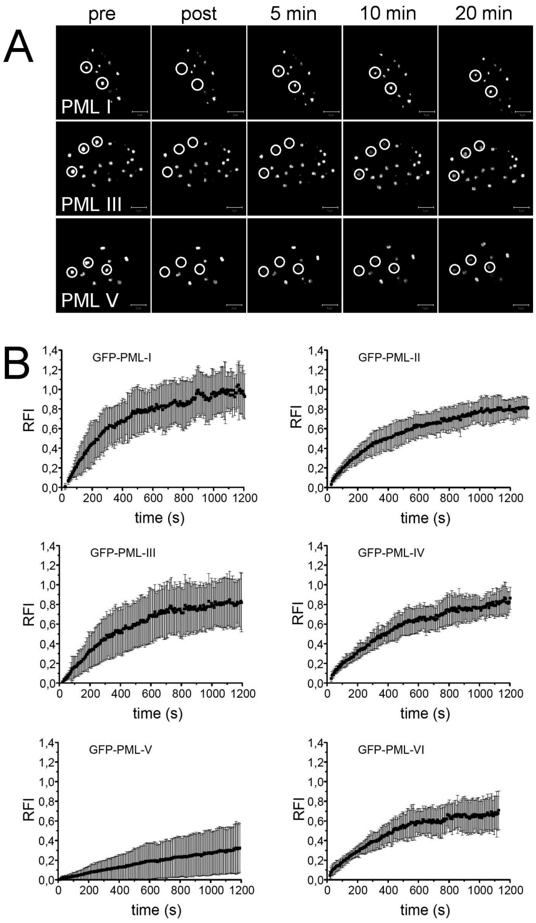


Figure 6.8: Exchange of PML isoforms at NBs. (A) FRAP experiments were performed on U-2 OS cells that express the indicated GFP-tagged PML isoforms by bleaching circled areas that contain a NB (pre, before bleach pulse; post, immediately after the bleach pulse) and monitoring fluorescence recovery for 20 minutes. Scale bars, 5 μm. (B) Quantification of FRAP experiments for each isoform. The graphs show mean values (± s.d. from at least 20 FRAP experiments each) as relative fluorescence intensity (RFI) after normalisation to prebleach levels.

assuming that molecules that undergo binding and unbinding to and from the body do so at the surface, and molecules situated more towards the inside of the body cannot unbind before moving to the surface. There is thus a reservoir of tightly bound "inner" molecules and one of loosely bound "outer" ones. Exchange between these reservoirs is modelled by linear kinetics, i.e. the more molecules there are, the more will move inside or out. These considerations lead to the sequential two-step model described above, which has been used in the kinetic analysis of the FRAP data for PML and related molecules. In accordance to the meaning of the variables, the kinetic parameters k_{on1}, k_{off1}, k_{on2}, and k_{off2} get assigned more explicit names: k_{on}, k_{off}, k_{in}, and k_{out} (Figure 6.10C).

To quantify the concentration of free proteins outside the ROI, the following p values were obtained for the GFP fusion proteins: all PML isoforms, p = 20; PML-RARα,

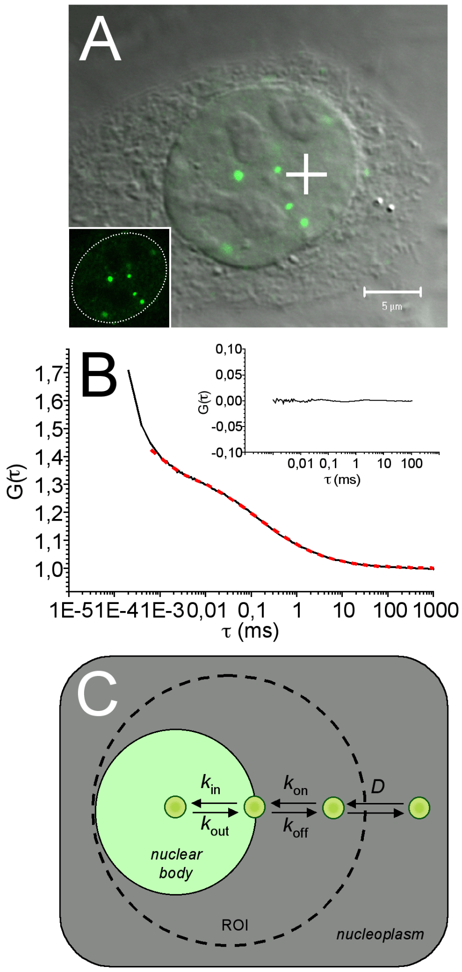


Figure 6.9: Diffusional behaviour of PML protein isoforms outside NBs. (A) U-2 OS cell expressing GFP-PML-IV (green) merged with the respective DIC image before FCS measurement. The inset shows only the GFP signals within the nucleus (indicated by dotted line) of this cell. +, position of the FCS laser beam. Scale bar, 5 μm. (B) Mean autocorrelation data obtained from FCS count-rate traces for GFP-PML-IV-expressing cells (solid black line). Data were fitted using an anomalous diffusion model (dashed red line). The inset graph displays a residual plot from the fit. (C) Kinetics modelling of PML NB assembly according to a diffusion-binding model. Molecules with the potential to accumulate at PML NBs move by diffusion (D) in the nucleoplasm outside NBs. Upon stochastic encounter, molecules associate and dissociate from the periphery of the NB (k_{on} and k_{off}, respectively) and penetrate into and out of the core of the NB (k_{in} and k_{out}, respectively). Dashed circle, ROI for bleaching and recovery measurements employed in FRAP experiments.

$p = 10$; all SP100 constructs, $p = 20$; DAXX, $p = 3$; HIPK2, $p = 2$; HIPK2(K221A), $p = 2$; PML-SUMO-2K, $p = 1.5$; PML-SUMO-3K, $p = 100$, BLM, $p = 5$.

For the PML data, we decided to remove one degree of freedom from the model by assuming that the fluorescence $x_{\text{obs}}(t)$ will always asymptotically return to 1.0 for large values of t, i.e. $x_{\text{f}}^{SS} + x_{\text{b1}}^{SS} + x_{\text{b2}}^{SS} = 1$. This assumption effectively denies the existence of an “immobile fraction”. Using this assumption, the relation

$$k_{\text{on}} = \frac{k_{\text{off}}(p-1)}{1 + k_{\text{in}}/k_{\text{out}}}$$

can be derived and was built into the model for PML FRAP data.

Model specialisation for HP1 FRAP data From preliminary experiments (not shown), it is clear that the kinetic analysis of the FRAP data for HP1 variants in eu- and heterochromatin (see introduction) requires at least two binding steps, as a model with only one binding step cannot adequately explain the data (Cheutin et al., 2004; Schmiedeberg et al., 2004). In contrast to the PML case, however, the literature suggests different binding modes that could lead to the sequential or the parallel model (Figure 6.7), or a combination of both.

The two scenarios are: *(1)* Freely diffusing HP1 (x_{f}) transiently attaches to chromatin. Once attached (x_{b1}), it can interact with other proteins that are already bound, and thus form a stronger bound (x_{b2}). This is the sequential binding model. *(2)* Alternatively, diffusing HP1 (x_{f}) can exist as a monomer or as a dimer/oligomer. These different states have different binding kinetics, and are modelled as different binding populations (x_{b1} and x_{b2}, respectively), where the higher-order complexes can attain stronger binding to chromatin. This is the parallel binding model. Of course, it is quite likely that in reality, both binding processes exist. However, this cannot be decided by FRAP data alone, as both models yield nearly-perfect fits to the data.

Again, diffusion is modelled as a two-way transport process according to Section 6.2.4. The ROI radius is $2.5\mu m$ for euchromatin measurements and $1.5\mu m$ for heterochromatin, and a basal diffusion constant of $D = 12\mu m^2/s$ is assumed for HP1.

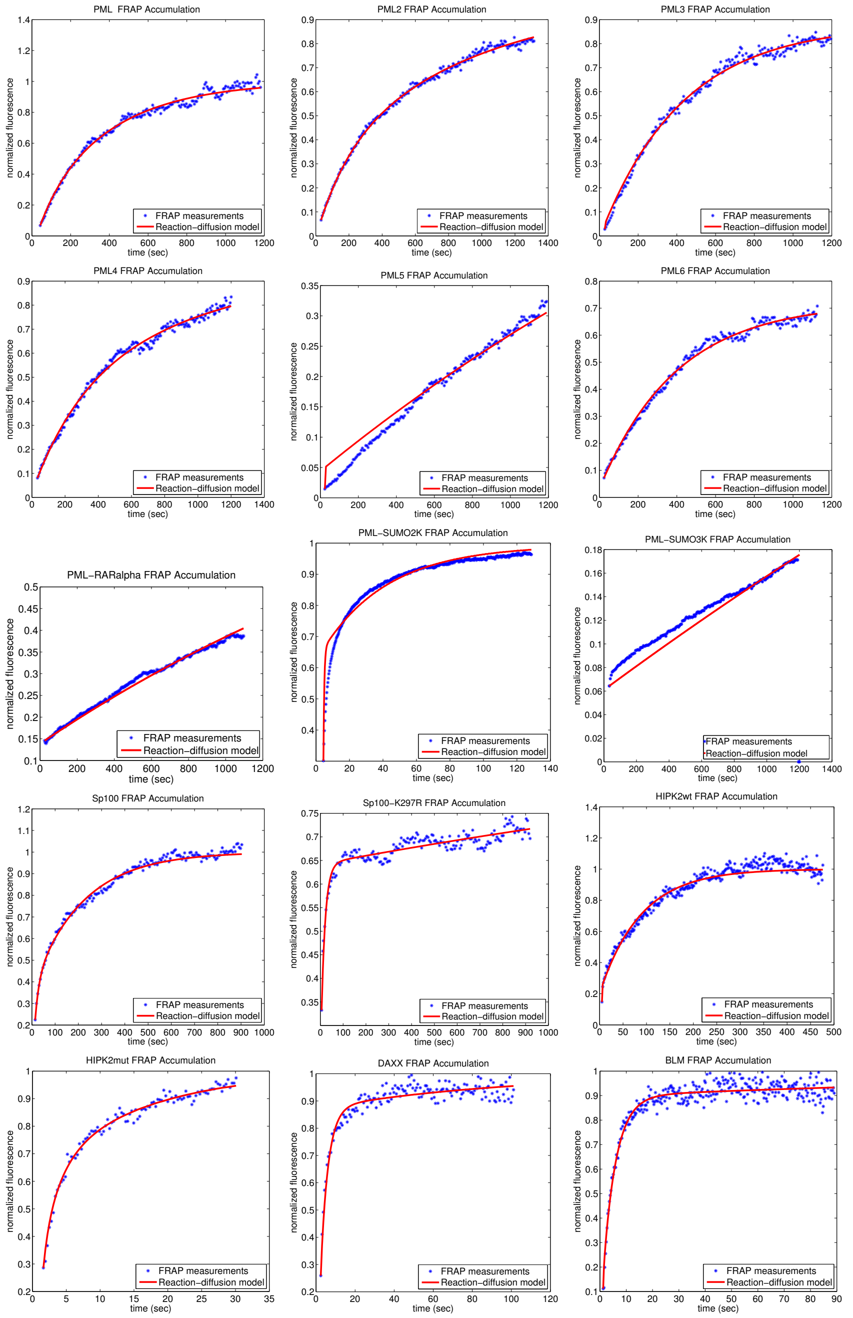


Figure 6.10: Fitting of PML FRAP data with the diffusion-binding model. The mean values of FRAP curves for the indicated GFP-tagged proteins (blue dots) were fitted using the diffusion-binding model depicted in Figure 6.9C. Fit curves are shown as solid red lines. Note, that these fits also consider the individual D values derived from FCS measurements.

6.3 Results

GFP-tagged PML isoforms are functional In order to enable visual detection of PML proteins, the genetic sequence encoding the six PML isoforms was fused to the sequence coding for green fluorescent protein (GFP), and the whole construct was transiently transfected into U-2 OS, HEp-2 or HeLa cells (for details, see Weidtkamp-Peters et al., 2008). Here, the construct is not inserted into the cells' genome, but is moved into the cells as additional DNA units using plasmid vectors. This way, the constructed DNA is transcribed inside the nucleus, but it is not copied in reproduction, hence the qualification as "transient" transfection. Using a variety of cell-biological means, it was shown that the GFP-PML constructs are functional, i.e. the addition of GFP to PML did not impair its function or significantly alter its biophysical parameters (Weidtkamp-Peters et al., 2008).

Individual isoforms are able to form NBs In addition to the analysis of PML NB in human cell lines, we performed biophysical analyses in living mouse cells, using PML-GFP fusion proteins both in PML knock-out (3T3-PML$^{-/-}$) and in control cells (3T3-PML$^{+/+}$) (Brand et al., 2010). This allowed us to study the nuclear body formation abilities of individual PML isoforms in a live cell setting. Individual diffusion constants of the isoforms were taken to be the same as in the human cell lines, determined by FCS as described above. For the kinetic analysis, the same reaction-diffusion model was used (Figure 6.9C).

In 3T3-PML$^{-/-}$ cells lacking endogenous PML, each individual PML isoform was able to form nuclear bodies (Figure 6.11). This confirms that the NB formation capacity of PML depends on the common part of all three isoforms, not on the C-terminal extension. In analogy to human cells, all PML isoforms display individual exchange kinetics, both in knock-out and control cells (Figure 6.12, Table 6.2). In the absence of endogenous PML, the isoforms PML I to IV form stable aggregates that allow only very little exchange (Table 6.2). However, for PML II and IV we observed a minor population of cells showing protein exchange, indicating a cell cycle dependence of these processes. The dynamics of PML V and PML VI were nearly identical between PML positive and negative cells.

Protein	Cells	k_{on}	k_{off}	k_{in}	k_{out}	Rt	F_{out}	F_{in}
PML I	PML$^{+/+}$	0.0633	0.0051	0.0002	0.0004	300	0.65	0.35
PML I	PML$^{-/-}$	0.0015	0.0015	<0.0001	0.0049	648	1.00	0.00
PML II	PML$^{+/+}$	0.0856	0.0069	<0.0001	<0.0001	222	0.65	0.35
PML II^{immob}	PML$^{-/-}$	0.0000	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.
PML II^{mobile}	PML$^{-/-}$	0.0288	0.0031	<0.0001	<0.0001	659	0.49	0.51
PML III	PML$^{+/+}$	0.0692	0.0055	<0.0001	0.0007	274	0.67	0.33
PML III	PML$^{-/-}$	0.0023	0.0013	<0.0001	0.5592	744	1.00	0.00
PML IV	PML$^{+/+}$	0.0565	0.0034	<0.0001	<0.0001	337	0.87	0.13
PML IV^{immob}	PML$^{-/-}$	0.0040	0.0005	<0.0001	<0.0001	4777	0.38	0.62
PML IV^{mobile}	PML$^{-/-}$	0.0571	0.0037	<0.0001	<0.0001	333	0.81	0.19
PML V	PML$^{+/+}$	0.0199	0.0024	<0.0001	<0.0001	957	0.43	0.57
PML V	PML$^{-/-}$	0.0188	0.0016	<0.0001	<0.0001	1010	0.61	0.39
PML VI	PML$^{+/+}$	0.0351	0.0021	<0.0001	<0.0001	541	0.90	0.10
PML VI	PML$^{-/-}$	0.0385	0.0036	0.0008	0.0010	493	0.56	0.44

Table 6.2: The diffusion-binding model was used to extract kinetic data from the FRAP and FCS experiments for all PML isoforms in PML positive and PML negative Mouse cells. k_{on}: association rate at the surface of the nuclear body; k_{off}: dissociation rate at the surface of the nuclear body; k_{in}: penetration rate into the nuclear body; k_{out}: penetration rate out of the core of the nuclear body; Rt: mean residence time at nuclear bodies; F_{out}: fraction of molecules residing at the surface of the nuclear body; F_{in}: fraction of molecules residing in the core of the nuclear body.

6.3.1 Quantification of protein dynamics

Exchange dynamics are specific to each protein To investigate the turnover of individual PML isoforms at NBs we employed FRAP. A spherical region containing one PML NB was bleached to background levels and fluorescence recovery was monitored for 20 minutes (Figure 6.8). FRAP at bleached NBs was substantially different between the isoforms. Whereas, for example, a considerable amount of GFP-PML-I fluorescence had already recovered after 5 minutes, there was only little recovery in GFP-PML-V-expressing cells after that time (Figure 6.8A). FRAP curves confirmed that all PML proteins exhibited individual exchange dynamics at NBs (Figure 6.8B). Most strikingly, GFP-PML-V showed an almost linear increase of fluorescence, and recovery reached only 32% after 20 minutes (Figure 6.8E). Thus, PML V is the most

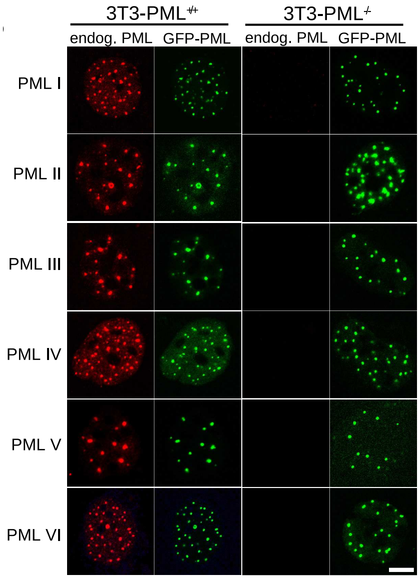


Figure 6.11: Mouse 3T3 cells with (3T3-PML$^{+/+}$) or without (3T3-PML$^{-/-}$) endogenous PML expression were transfected with expression vectors encoding human PML isoforms I to VI as GFP fusion proteins. Cells on coverslips were fixed and processed for immunofluorescence staining to detect endogenous mouse PML protein (red) and the exogenous GFP-tagged human PML isoforms (green). Images show mid-nuclear confocal sections. Note that the anti-mouse-PML antibody does not cross-react with human PML. Bar, 5 μm.

stable isoform, suggesting that it serves as a scaffold component of PML NBs.

Diffusion rates of NB components in nucleoplasm To assess the mobility of PML isoforms and the other NB components under study we applied FCS (Weidtkamp-Peters et al., 2009). Using this technique we were able to directly assess the diffusion coefficient (D) of the nucleoplasmic pools of GFP-tagged PML protein isoforms and all other NB components analysed in this study (Figure 6.9A,B; Table 6.3). The D value of GFP alone in the nucleus was 9.5 ± 1.5 μm^{2}sec^{-1}, which is in perfect agreement with independent measurements (Wachsmuth et al., 2000). This D value is well below the one of GFP measured by FCS or FRAP in solution ($D \approx 90$ μm^{2}sec^{-1}) (Hink et al., 2000; Swaminathan et al., 1997), hence reflecting anomalous diffusion of untagged GFP in the crowded nuclear environment. The D values of PML isoforms, however, were significantly smaller than for GFP alone. They ranged from 1.02 ± 0.11 μm^{2}sec^{-1}for PML II to 2.79 ± 0.11 μm^{2}sec^{-1}for PML V (Table 6.3). These D values are too low to account only for diffusion barriers based on the size of the fusion proteins compared with GFP. Evaluation of the FCS data also delivered the anomaly parameter (α), which describes the degree of obstruction for diffusing particles by the medium (Saxton, 2001). Under conditions of free diffusion, i.e. in buffer solutions, α equals 1, but decreases continuously to a limit value of α=0.75 for GFP with increasing crowding conditions and obstacle concentration (Banks and Fradlin, 2005). Compatible with this

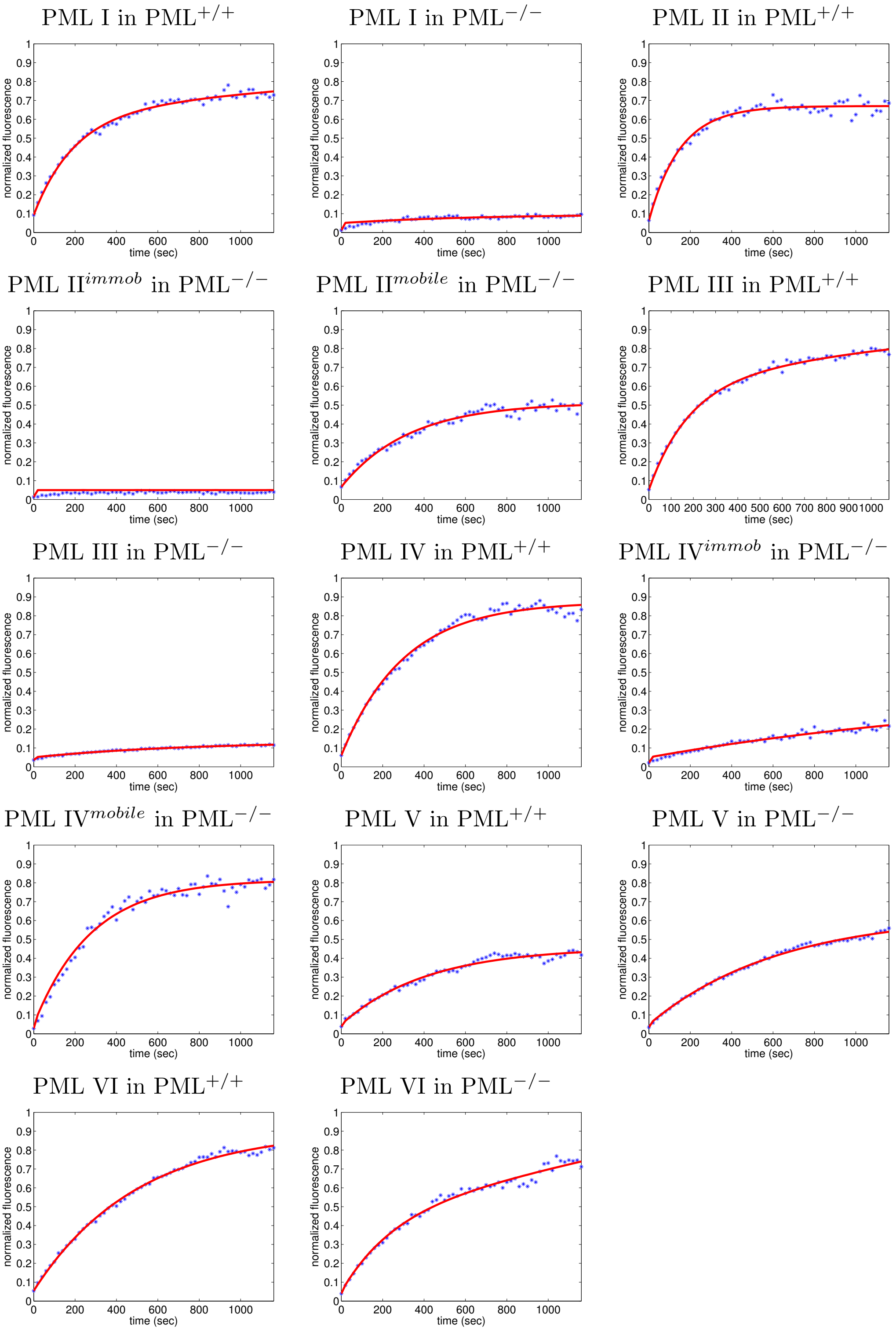


Figure 6.12: Fitting of the measured FRAP curves (blue dotted line) in mouse cells with (PML$^{+/+}$) and without (PML$^{-/-}$) endogenous PML. The binding-diffusion model shown in Figure 6.9C results in good fits (red lines).

Protein	D	α	k_{on}	k_{off}	k_{in}	k_{out}	Rt	F_{out}	F_{in}
PML I	1.85	0.47	0.0717	0.0056	0.0022	0.0046	265.07	0.68	0.32
PML II	1.03	0.44	0.0451	0.0041	0.0010	0.0014	421.0	0.59	0.41
PML III	1.63	0.53	0.0397	0.0024	<0.0001	<0.0001	478.3	0.87	0.13
PML IV	1.04	0.50	0.0377	0.0028	0.0004	0.0010	503.5	0.72	0.28
PML V	2.79	0.57	0.0051	0.0015	0.1571	0.0338	3695.2	0.18	0.82
PML VI	1.49	0.51	0.0343	0.0026	<0.0001	<0.0001	554.4	0.71	0.29
PML-IV-2K	1.19	0.55	0.0125	0.0256	<0.0001	0.0001	39.9	0.98	0.02
PML-IV-3K	2.23	0.61	0.0111	0.0055	0.2027	0.0042	8954.2	0.02	0.97
PML-RARα	0.93	0.45	0.0032	0.0069	0.1307	0.0071	2829.7	0.05	0.94
Sp100	1.23	0.62	0.5968	0.0702	0.0065	0.0053	31.8	0.45	0.55
Sp100K297R	0.84	0.58	1.5180	0.1276	0.0002	0.0003	12.5	0.63	0.37
HIPK2	2.60	0.66	0.0344	0.0115	0.0	0.0051*	87.2	1.0	0.0
HIPK2mut	0.66	0.50	0.2240	0.3905	0.0608	0.0819	4.5	0.57	0.43
DAXX	0.95	0.60	0.6597	0.3994	0.0022	0.0105	3.0	0.83	0.17
BLM	3.10	0.61	0.8563	0.2439	0.0007	0.0049	4.7	0.88	0.12
nls-GFP	9.50	0.73	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.

Table 6.3: Summary of the kinetic parameter results for the PML analysis. Diffusion coefficient D [$\mu\text{m}^2\text{sec}^{-1}$], anomality parameter α, rate of binding / unbinding to PML body k_{on}/ k_{off}, rate of incorporation into PML body core and exit from core k_{in}/ k_{out}, average residence time at PML NB Rt [s], fraction of molecules bound to outer / inner part of PML NB F_{out}/F_{in}. The fits for PML V, PML-IV-2K and PML-IV-3K are not fully satisfactory (see Figure 6.10) and their kinetic parameter thus need to be treated with extra caution. *the k_{out} value of HIPK2 is of no importance, since $k_{\text{in}} = 0$.

assumption we found α=0.73 for nls-GFP alone, but substantially lower values for PML isoforms and all other PML NB components (Table 6.3). The low D values and anomaly parameters, therefore, strongly indicate transient binding events of PML components with chromatin or protein-based complexes throughout the nuclear volume, not only at PML NBs.

A binding-diffusion model to describe component exchange at PML NBs On the basis of the FRAP and FCS results we developed a binding-diffusion model for the quantitative description of protein exchange at these structures (Sprague and McNally, 2005) (Figure 6.9C). Diffusion inside and out of bleached regions was modelled as a linear two-way process using the measured D values from the FCS experiments.

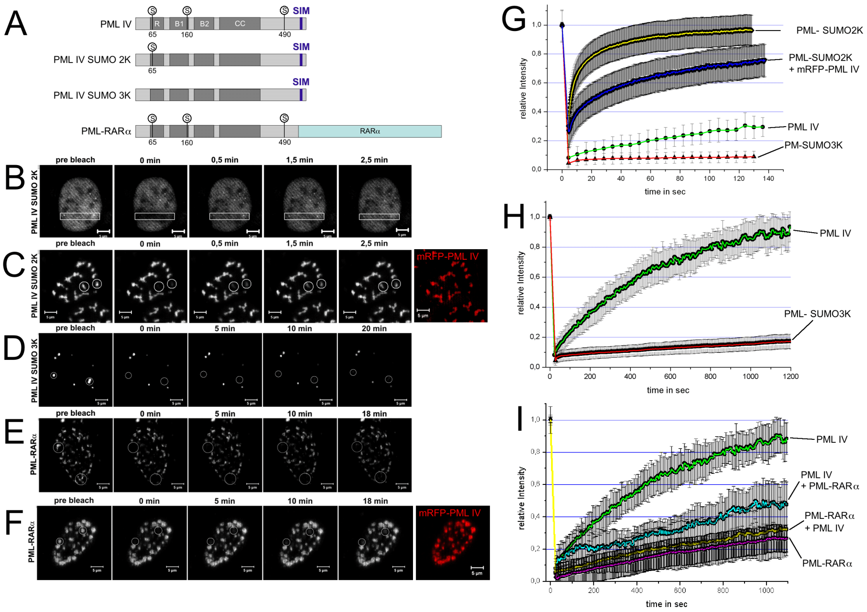


Figure 6.13: Exchange of PML protein variants at NBs. (A) Schematic depiction of PML-fusion proteins used for FRAP experiments. The domain structure is as described in Figure 6.2. SUMO-modifiable lysine residues at positions 65, 160 and 490 in PML are also shown. (B-F) U-2 OS cells transfected with the indicated GFP-tagged PML protein constructs were used in FRAP analyses of areas that contain NBs and fluorescence recovery was monitored for various times as indicated. Scale bars, 5 μm. In C and F, cells were co-transfected with PML IV tagged to mRFP in order to analyse the impact of wild-type PML on the exchange dynamics of PML-IV-SUMO-2K (C) and PML-RARα (F). (G-I) Graphs show mean values (± s.d.) from at least 20 FRAP experiments each, of the indicated proteins or protein combinations as relative fluorescence intensity (RFI) after normalisation to pre-bleach levels.

Table 6.3 contains for each of the PML NB components the D values and anomaly parameters α, the binding and unbinding rates k_{on} and k_{off}, and the rates of movement to the inner core and to the outer surface of the body k_{in} and k_{out}, respectively. From these, we computed the residence time (Rt), i.e. the mean time a molecule spends bound to the NB, and the fraction of molecules bound in the inner and outer region of the body, F_{in} and F_{out}, respectively. Conceptually, this model can also be applied to proteins with short Rt values, since these may split into two populations, one which immediately dissociates after surface contact with rate constants k_{on} and k_{off}, and a second which encounters additional binding partners at or in the nuclear body with rate constants k_{in} and k_{out}. We therefore applied this binding-diffusion model to all PML NB components analysed in this study. This model provided good fits to the measured FRAP curves of all PML components (Figure 6.10). The one-step model, assuming that all molecules exchange with the same rate, did not result in good fits (data not shown, c.f. Weidtkamp-Peters et al., 2008). Moreover, FRAP curves could never be fitted well with one-component exponential functions (T.L. and P.H., unpublished results). These observations confirmed the existence of at least two differently mobile populations of molecules at NBs.

6.3.2 Consequences for biological function

> **Note:** The measured kinetic parameters imply a large number of results about the biological function of specific NB components investigated in this study. A detailed account of these results is not in the focus of this thesis, and the reader is referred to the original publication (Weidtkamp-Peters et al., 2008) for details. For the sake of completeness, a short summary is given below.

SUMOylation affects kinetic exchange rates Small Ubiquitin-like Modifier proteins (SUMO) can attach to PML isoforms at three different sites. To test whether SUMOylation changes the exchange rates, we applied FRAP to PML mutations lacking two or three of these sites (Figure 6.13A,B). When two SUMOylation sites were mutated, the recovery kinetics were much faster (Table 6.3, Figure 6.13B,G). In contrast, proteins in which all three sites were mutated showed very stable binding behaviour, with an increased fraction of proteins bound in the inner region of the PML NB (Figure 6.13D,H).

The PML-RARα oncoprotein shows hyperstable binding In our FRAP experiments, the PML-RARα oncoprotein localised in a microspeckled pattern containing PML and SP100 (as expected, see Dong et al., 2004). FRAP revealed only very slow recovery of GFP-PML-RARα when observed over a period of 20 minutes (Fig. 6.13E,I). In comparison to the most similar PML isoform, PML VI, retention time of PML-RARα is increased fourfold and the stably bound fraction is doubled. GFP-PML-RARα also reduced the exchange kinetics of other GFP-tagged isoforms upon coexpression (data not shown). These observations demonstrate that the PML-RARα oncoprotein has a dominant NB-retention effect on wild-type PML proteins.

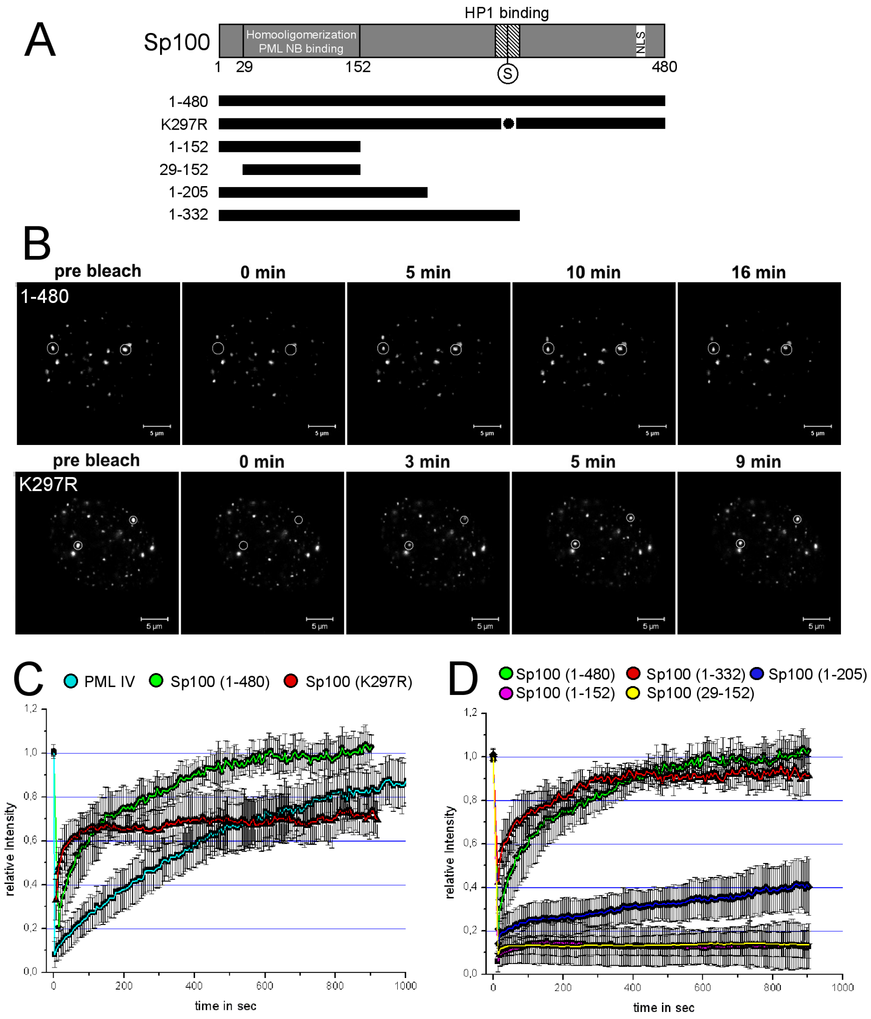


Figure 6.14: Exchange dynamics of Sp100 and Sp100 mutant constructs at NBs. (A) Schematic depiction of Sp100 protein and mutants used for FRAP analyses. The oligomerisation and PML NB targeting resides within a region spanning amino acids 29-152 of Sp100 (Negorev and Maul, 2001). Sp100 has a SUMOylation site at Lys297 (S) within the HP1-binding motif and a nuclear localisation signal at the C-terminus (NLS). Note, that Sp100 variants that lack the endogenous NLS were engineered to contain one at their N-termini (Negorev and Maul, 2001). (B) FRAP experiments were performed in U-2 OS cells that express GFP-tagged Sp100 wildtype (aa 1-480) and the indicated variants, by bleaching areas that contain NBs and by monitoring fluorescence recovery for various intervals as indicated. Scale bars, 5 μm. (C,D) Quantification of FRAP experiments as shown in B. Graphs show mean values (± s.d.) from at least 20 FRAP experiments. For comparison, the FRAP curve of GFP-PML-IV at NBs was included in C.

Sp100 turns over rapidly at PML NBs and requires SUMOylation for tight binding SP100 is a PML NB constituent with potential functions in transcriptional regulation at promoters of specific genes (Sternsdorf et al., 2005). Exchange rates of these proteins were analysed by FRAP analysis (Figure 6.13 and 6.14), and showed that SP100 exchanged at NBs more than five times faster than the fastest PML isoform and fluorescence recovered completely after ∼10 minutes (vs >20 minutes in the case of PML I). Additionally, mutation of the SUMOylation sites in Sp100 indicates two distinct populations of molecules: whereas the k_{on} and k_{off} values were similar to wild-type SP100, k_{in} and k_{out} were reduced ∼30 and ∼20 times, respectively (Table 6.3), indicating that SUMOylation does not influence the binding of SP100 to NBs but it affects its incorporation into the core.

HIPK2 binding requires kinase activity and is modulated by PML IV HIPK2 is a kinase involved in cell proliferation and apoptosis by phosphorylation-induced stabilisation of the tumour suppressor p53 (D'Orazi et al., 2002; Hofmann et al., 2002). Modelling of the FRAP data revealed an *Rt* value of 11 seconds for GFP-HIPK2 at PML NBs indicating a fast turnover (Figure 6.15A,F; Table 6.3). Consistent with previous observations (Hofmann et al., 2002), we observed a strong NB retention of HIPK2 when PML IV was expressed at elevated levels (Figure 6.15B,F). The recovery halftime of HIPK2 was ∼60 seconds in the absence and ∼240 seconds in the presence of excess PML IV. These observations clearly indicate that PML IV recruits HIPK2 by increasing its *Rt* at NBs. Then, GFP-tagged HIPK2(K221A) was analysed, which lacks the kinase activity. This construct showed only little accumulation at PML NBs (Figure 6.15C), demonstrating that a functional kinase domain on HIPK2 is required to retain this enzyme at PML NBs.

Rapid turnover of DAXX and BLM at PML NBs We also analysed exchange kinetics of DAXX and BLM at PML NBs. DAXX and BLM showed rapid FRAP at NBs with *Rt* values of 2.6 seconds and 4.1 seconds, respectively, and the complete pool of NB-bound molecules turned over within less than one minute (Fig. 6.15H,I). We like to point out that FRAP curves of these proteins could neither be fitted with a binding-diffusion model assuming a single exchanging population, nor with one-component exponential functions (Weidtkamp-Peters et al., 2008, and unpublished observations). Therefore, kinetics modelling using the two-component binding-diffusion model reveals at least two differently mobile populations of DAXX and BLM at PML

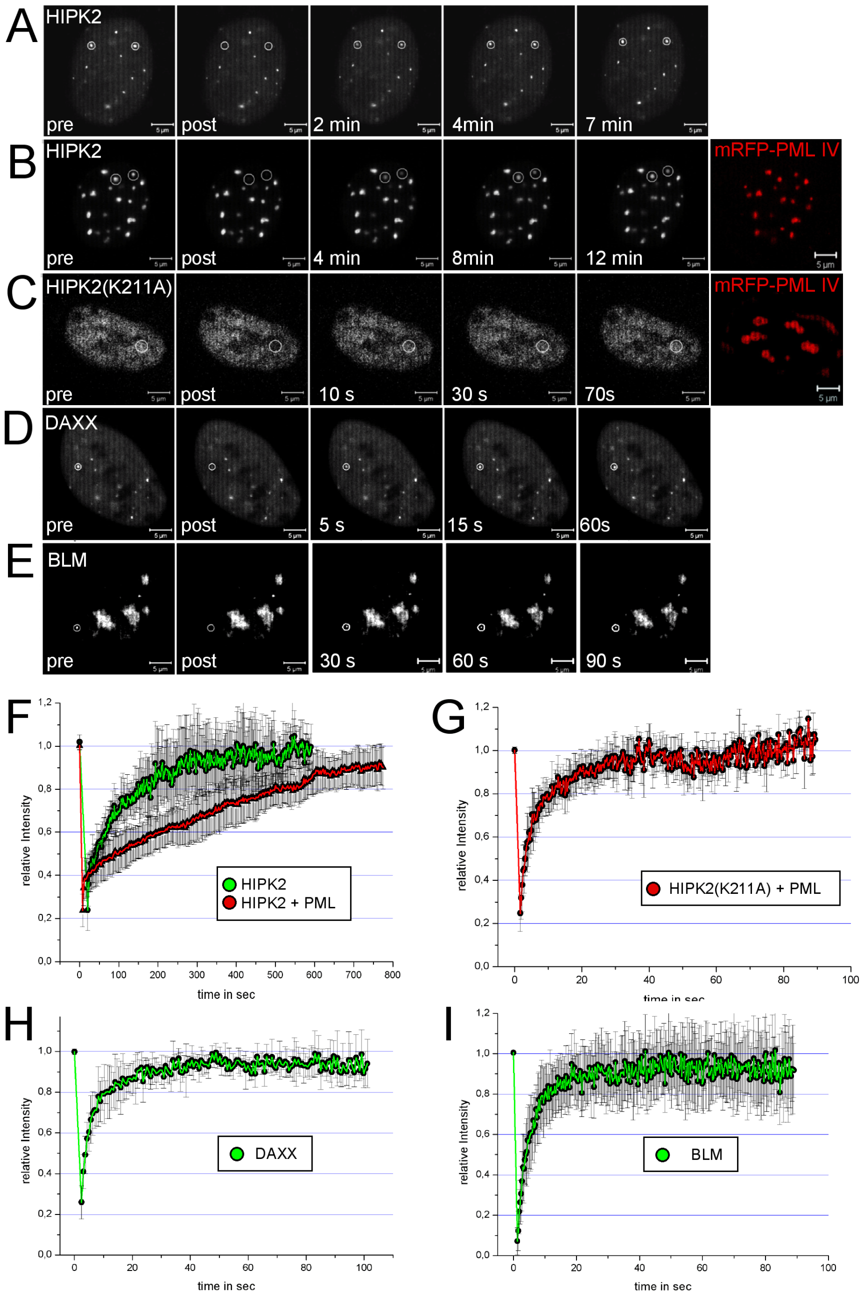


Figure 6.15: Exchange dynamics of HIPK2, DAXX and BLM at PML NBs. FRAP experiments were performed in U-2 OS cells that express (A) GFP-tagged wild-type HIPK2, (B) GFP-HIPK2 in combination with mRFP-PML-IV (red), (C) the kinase-defective GFP-HIPK2(K221A) mutant in combination with mRFP-PML-IV (red), GFP-DAXX (D), and GFP-BLM (E). Areas containing HIPK2 accumulations were bleached and fluorescence recovery was recorded for various times as indicated. Scale bars, 5 μm. (F-I) Quantification of FRAP experiments as shown in A-E. Graphs show mean values (± s.d.) from at least 20 FRAP experiments each.

NBs, a fast one that rapidly binds and unbinds at the surface, and a slower one that is retained at NBs for seconds. The majority (>80%) of GFP-tagged DAXX or BLM molecules was identified in the more loosely bound fraction by kinetics modelling (Table 6.3).

6.3.3 Further work: Analysis of HP1 binding to chromatin

In addition to the presented results on PML nuclear bodies, we also applied the described modelling framework to FRAP data of heterochromatin protein 1 (HP1). HP1 binds to chromatin, more specifically to H3K9, i.e. the histone H3 methylated on lysine 9 (see introduction). We found that in analogy to the PML data, a binding model with only one bound state cannot adequately explain the data. Thus, two different binding modi have to exist. However, it is not clear whether the appropriate description is given by the sequential or parallel binding model, or a combination of both (Section 6.2.5). Both models fit the FRAP time courses for all three HP1 variants equally well, so that a distinction based on this data is not possible.

Further insights into this issue are expected from FRAP experiments with HP1 mutants, in which one or two of the three domains (chromodomain, hinge-region, chromoshadow-domain) are deleted. At the time of writing, this data has not been analysed thoroughly, and it seems inappropriate to present it in depth before this analysis has been carried out. Therefore, the interested reader is referred to the manuscript in preparation for further details.

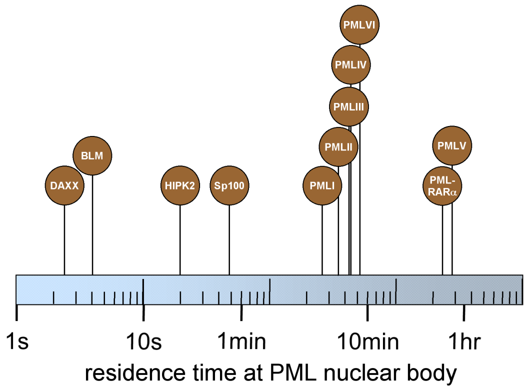


Figure 6.16: The residence times of the indicated PML nuclear body components as deduced from our kinetic modelling approach is shown on a logarithmic seconds scale.

6.4 Discussion

6.4.1 Lessons for PML NB assembly

Understanding PML NB assembly and function requires detailed knowledge of its organisation in space and time. Using live-cell microscopy and kinetics modelling we have determined the intranuclear dynamics of eleven components of PML NBs. These data reveal a wide range of assembly plasticity of PML NBs (Figure 6.16).

PML traffic at NBs Previously, only PML IV had been analysed in FRAP experiments, (Boisvert et al., 2001; Everett and Murray, 2005; Wiesmeijer et al., 2002). In this study, we have analysed all six nuclear PML isoforms, all of them showing protein-specific exchange dynamics at NBs. Comparing the exchange rates of the PML isoforms to each other, it becomes apparent that the SUMO-interacting motif (SIM, lacking in PML VI) is not the main determinant for exchange dynamics, even though it is required for NB formation (Shen et al., 2006). Rather, it seems that diverse functional domains at the C-terminus of PML isoforms determine their kinetics.

Different PML isoforms show different binding behaviour with certain proteins (Fogal et al., 2000). It thus seems suggestive that freely diffusing, highly mobile PML molecules “scan” the nucleoplasm for interaction partners and subsequently transport those to NBs. There, the proteins brought together by PML can then interact in a strictly confined space.

PML V as a scaffold subunit of NBs? It came as a surprise that PML V is a very stable isoform at NBs. The reaction-diffusion model revealed an *Rt* value of 48 minutes (Table 6.3), whereas typical *Rt* values of proteins within speckles, Cajal bodies and nucleoli range between seconds to a few minutes (Chen and Huang, 2001; Dundr et al., 2004; Handwerger et al., 2003; Kruhlak et al., 2000; Phair and Misteli, 2000). This high *Rt* value very likely indicates a structural role of this isoform within PML NBs, acting as a long-lasting but not static scaffold subunit. Such a metastable PML scaffold has the advantage to be able to rapidly respond to external stimuli, while at the same time serving as a stable platform for the assembly of functional complexes built from faster-exchanging molecules, such as DAXX, BLM, HIPK2 and CBP, with *Rt* values of several seconds (Table 6.3) (Boisvert et al., 2001; Everett and Murray, 2005).

Targeting and release mechanisms at PML NBs Combining the results of this chapter (for details, see Weidtkamp-Peters et al., 2008), the time a component stays at PML NBs might be modulated by (1) its own SUMOylation status, (2) the presence of a SUMO-interaction motif, (3) the PML isoform it binds to, (4) the SUMOylation pattern of this isoform, and (5) the relative abundance of other isoforms in the same NB. Different SUMO paralogs, polymeric SUMO chains and phosphorylation could provide additional layers. PML-RARα is suggested to interfere with normal myeloid differentiation by inhibiting wild-type RARα transcriptional activity (Melnick and Licht, 1999). On the basis of FRAP analyses it has recently been proposed that PML-RARα expression interferes with normal RARα functions by reducing the intranuclear mobility of RARα and its co-regulator SMRT, suggesting a crucial role for the reduced intranuclear mobility of PML-RARα in the pathogenesis of APL (Dong et al., 2004; Huang et al., 2007). Here, we demonstrated that the PML-RARα oncogene can efficiently immobilise wild-type PML proteins, thus drastically reducing the amount of available wild-type PML in the nucleoplasm.

PML NBs: active or passive nuclear sites? Several functions have been postulated for PML NBs, such as (1) nuclear storage sites for the accumulation or sequestration of proteins, (2) catalytic surfaces where proteins accumulate to be post-translationally modified, or (3) active sites for specific nuclear functions such as transcriptional and chromatin regulation (reviewed in Bernardi and Pandolfi, 2007). During a stress response, HIPK2-mediated phosphorylation of p53 and acetylation by CBP is believed to occur at PML NBs (D'Orazi et al., 2002; Fogal et al., 2000; Guo et al., 2000; Hofmann et al., 2002). Here, we have shown that, even in the presence of excess PML IV, HIPK2 binding to NBs was almost absent when its kinase is inactive. Furthermore, quantitative expression of PML IV increased the *Rt* of wild-type HIPK2 at NBs several fold. These observations strengthen the catalytic-platform theory for PML NBs, which suggests that specific processes (phosphorylation, acetylation, SUMOylation), occur on the surface of the body (Kentsis and Borden, 2004). BLM and DAXX turned over at NBs rapidly and completely. Since these proteins function directly at DNA-synthesis sites or at specific promoters, respectively (Lin et al., 2007; Wu, 2007), their low *Rt* values are likely to reflect the function of PML NBs to titrate the concentration of these factors in the nucleus. SP100 is also a transcriptional co-regulator at promoters but showed dissociation kinetics that were one order of magnitude slower at PML NBs than those of BLM or DAXX. Although this does not rule out a titration effect by NBs, the *Rt* value for SP100 also suggests a scaffold function for the regulated retention of SP100-interacting proteins at NBs, such as NBS1 and HP1 (Naka et al., 2002; Seeler et al., 1998).

All in all, our kinetics data do not favour a particular model for PML NB function but, instead, provide evidence for all of them. We therefore suggest that PML NBs are integrated into many nuclear pathways as both biophysical and biochemical hubs, which provide at the same time structural stability at specific chromatin loci, titration sites for specific factors, and catalytic surfaces for post-translational modifications or complex assembly.

6.4.2 Lessons for the mathematical modelling of FRAP

FRAP modelling seems simple, but it is not! At the onset, the mathematical modelling of FRAP experiments seems straightforward: the models involved are relatively low-dimensional and simple, and the relation between data and model seems clear. Digging deeper, it becomes obvious that things may not be as clear as they first seem to be. For example, the range of models that can be considered is actually quite wide: from the simple exponential function, via a compartmental model with multiple binding states (as in this work), to spatially explicit models containing a detailed representation of the nuclear architecture (Beaudouin et al., 2006). From a different perspective, many potential problems with the data are unearthed: How is the fluorescence decay caused by laser scanning taken into account? Is there a “blinding” effect of the photobleach pulse on the first image acquisitions? How much noise is there in the data? And how much variation between repeated measurements?

Taking together the uncertainties about model and data, one is forced to conclude that the mathematical analysis of FRAP experiments is limited to rough estimations of the kinetic parameters (Mueller et al., 2008). Specifically, there is at the moment no possibility to arbitrarily increase the resolution and reliability of these estimations, no matter how much time and money is invested. Fundamental advances in the methodology, both *in vivo* and *is silico*, are needed for this.

Two potential errors In a recent publication, Mueller et al. (2008) describe two principal sources of systematic error found in many approaches to FRAP analysis: a neglect of the role of diffusion, and an incorrect approximation of the initial fluorescence distribution after the photobleach. They show that when applying different FRAP approaches that do not correct for these errors, the choice of the FRAP analysis method determines the kinetic parameters, not the type of protein under study (cf. Hinow et al., 2006; Phair et al., 2004; Sprague et al., 2004).

The study presented here was designed from the onset to take diffusion into account. Diffusion is measured separately (by FCS) and then included into the FRAP model, which is used to estimate the binding coefficients. Moreover, the problem of the initial photobleach distribution does not seem too severe here (Section 6.2), even though improvements in this matter are likely to add more precision to the results.

No gold standard There cannot be enough emphasis on the argument that in the current state, FRAP analysis by mathematical modelling can only yield predictions of the magnitude of kinetic parameters, not precise values (Mueller et al., 2008; Sprague and McNally, 2005). The main reason for this is that there is no “gold standard” for FRAP experiments, i.e. no experimental setup where the “correct” values are precisely known. Such a standard could be used to calibrate and test FRAP analysis procedures. Without it, systematic errors that underlie the different methods can easily go undetected, and the only possible safeguard is the comparison of results from different methodologies (Mueller et al., 2008).

The best possible approach Wetlab biology is a messy business, and so is its mathematical analysis. In contrast to dead matter under study in physics or chemistry, living entities always come with a high degree of noise, variability and sensitivity to laboratory conditions. Thus, there is always a certain level of uncertainty implied in biological measurements. In analogy, mathematical models of biological systems always have to be more abstract than the real thing, and thus simplifications and approximations are unavoidable.

In spite of all this, the strongest argument for the kind of scientific work presented in this section is that it is simply the best we have. When systems are inherently noisy and variable, measuring them can never be done without uncertainty. When systems are complex, modelling has to simplify and approximate. Therefore, I believe that with the methodologies currently available, the approach presented here is well worth the effort. The kinetic quantification of FRAP measurements, in this detail only possible through mathematical modelling, has led to significant insights that would be unavailable by pure inspection of the FRAP curves.

7

Investigating an Evolutionary Motif: Developmental Robustness by Obligate Interaction of Floral Homeotic Genes

Chapter summary. *DEF-like and GLO-like class B floral homeotic genes encode closely related MADS-domain transcription factors that act as developmental switches involved in specifying the identity of petals and stamens during flower development. Class B gene function requires transcriptional upregulation by an autoregulatory loop that depends on obligate heterodimerisation of DEF-like and GLO-like proteins. Since switch-like behaviour of gene expression can be displayed by single genes already, the functional relevance of this complex circuitry has remained enigmatic. Based on a stochastic in silico model of class B gene and protein interactions we suggest that obligate heterodimerisation of class B floral homeotic proteins is not simply the result of neutral drift, but enhanced the robustness of cell-fate organ identity decisions in the presence of stochastic noise. This finding strongly corroborates the view that the appearance of this regulatory mechanism during angiosperm phylogeny led to a canalisation of flower development and evolution. The transition from positive autoregulatory feedback of one gene to obligate heterodimerisation can suggestively be termed an "evolutionary motif". We discuss the generality of our approach, especially in the light of potential applications to other such motifs.*

> **Note:** The work described in this chapter has been carried out in collaboration with Günter Theißen at the Department of Genetics, FSU Jena, and has been published in PLoS Computational Biology (Lenser et al., 2009).

7. INVESTIGATING AN EVOLUTIONARY MOTIF: DEVELOPMENTAL ROBUSTNESS BY OBLIGATE INTERACTION OF FLORAL HOMEOTIC GENES

Chapter contents

7.1 Introduction

Self-regulating genes can act as genetic switches The development of organs, their position and boundaries in multicellular organisms is defined by genes that can sustain their own activation over long periods of time, termed genetic switches. A good case in point is provided by the genetic machinery controlling the development of flowers in higher plants. Depending on the nature of the interactions of their constituents, gene regulatory circuits can display a variety of dynamical behaviours ranging from simple steady states, to switching and multistability, to oscillations. Temporal or spatial patterning during development requires activation of genes at a particular time or position, respectively, and the inhibition in the remaining time or part (Lewis, 2008).

Regulatory genes involved in such processes often show a switch-like temporal or spatial dynamics, which requires a direct or indirect positive non-linear feedback of the genes on their own expression, e.g. via dimers of their own product (Pigolotti et al., 2007). Switch-like behaviour can be displayed by a single gene (Ferrell, 2002; Wolf and Arkin, 2003), but many gene regulatory switches have a more complex structure. Due to the small number of molecules involved, these switches are inherently stochastic and their behaviour under noisy conditions can strongly depend on their genetic architecture (Elowitz et al., 2002; Kaern et al., 2005; McAdams and Arkin, 1997; Raser and O'Shea, 2005). In some cases the complex regulatory interactions have been quite well documented, but the functional implications of the corresponding regulatory circuitry have remained enigmatic. A good example is provided by some floral homeotic (or organ identity) genes from model plants such as *Arabidopsis thaliana* (thale cress; henceforth termed *Arabidopsis*) and *Antirrhinum majus* (snapdragon; henceforth called *Antirrhinum*).

Floral homeotic genes determine organ identity in flower development Floral homeotic genes act as developmental switches involved in specifying organ identity during flower development. According to the 'ABC model', three classes of floral organ identity (or homeotic) genes act in a combinatorial way to specify the identity of four types of floral organs, with class A genes specifying sepals in the first floral whorl, A+B petals in the second whorl, B+C stamens (male reproductive organs) in the third whorl, and C alone carpels (female organs) in the fourth floral whorl (Coen and

Meyerowitz, 1991). The combinatorial genetic interaction of floral homeotic genes may involve the formation of multimeric transcription factor complexes that also include class E (or SEPALLATA) proteins, as outlined by the 'floral quartet' model (Theißen and Saedler, 2001).

The enigma of obligate heterodimerisation in class B genes In *Arabidopsis thaliana* and other plants, a particular class of these genes - *DEF*-like and *GLO*-like floral homeotic genes - regulates the development of petals and stamens. These genes are self-activating via a heterodimer of their protein products, making the activity of each one of them fully bound to the activity of the other one. The reason for their total functional interdependence has long remained unclear, as the expression of both genes is jointly controlled by shared transcription factors in addition to the heterodimer. In principle, one gene alone could provide their switching functionality.

In *Antirrhinum*, there are two different class B genes termed *DEFICIENS* (*DEF*) and *GLOBOSA* (*GLO*). In *Arabidopsis* these genes are represented by *APETALA3* (*AP3*), the putative orthologue of *DEF*, and *PISTILLATA* (*PI*), the putative *GLO* orthologue. For simplicity, we will refer to *DEF*-like and *GLO*-like genes from here on. *DEF*-like and *GLO*-like genes represent paralogous gene clades that originated by the duplication of a class B gene precursor 200 – 300 million years ago (Kim et al., 2004; Winter et al., 2002a). All class B genes identified so far, like most other floral homeotic genes, belong to the family of MADS-box genes, encoding MADS-domain transcription factors (Kaufmann et al., 2005; Theißen et al., 2000).

Class B genes determine petal and stamen identity Mutant phenotypes reveal that *DEF*-like and *GLO*-like genes are essential for the development of petals and stamens, since *def* and *glo* loss-of-function mutants all produce flowers with petals converted into sepals and stamens transformed into carpels (Goto and Meyerowitz, 1994; Jack et al., 1992; Sommer et al., 1990; Tröbner et al., 1992; Zahn et al., 2005). When co-expressed in the context of a flower, DEF and GLO are not only required, but even sufficient for specifying petal and stamen identity, as revealed by transgenic studies (e.g. Krizek and Meyerowitz, 1996).

Obligate heterodimerisation provides positive autoregulation to class B genes Induction and stable maintenance of switch-gene expression are typically two independent processes, depending on a transient external signal and autoregulation, respectively (Schwarz-Sommer et al., 1992). Whenever a transient activating signal is

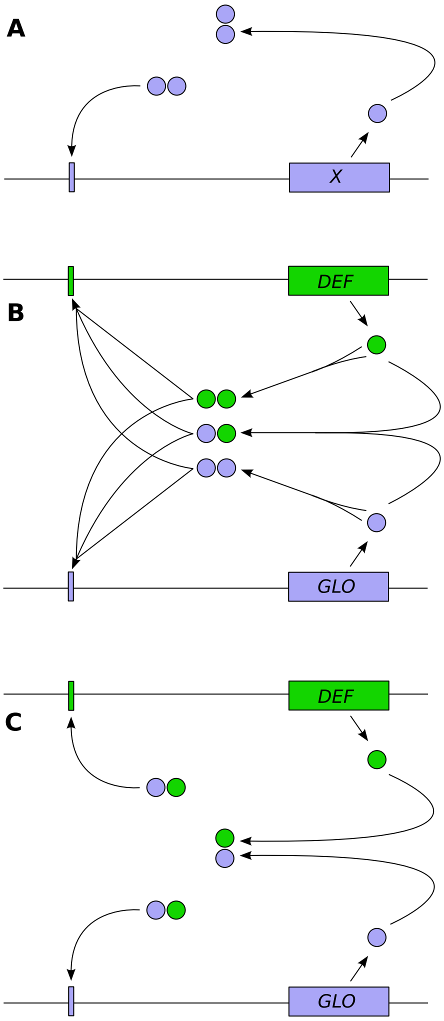


Figure 7.1: The three types of regulatory mechanisms that are investigated. "DEF" and "GLO" denote *DEF*-like and *GLO*-like genes. Large boxes represent the coding regions of genes (neglecting intron-exon structure), small boxes symbolise cis-regulatory elements (CArG-boxes). (A) Ancestral state. A single gene *X* is positively regulated by a dimer of its gene product. (B) Intermediate state. After gene duplication, three types of protein dimers can be formed, since both homo- and heterodimerisation are possible. All three dimers regulate both genes. (C) Final state. After having lost the homodimerisation ability, the remaining heterodimer regulates both genes.

above a threshold, the gene activity switches from the OFF- to the ON-state. The signal is required only for initiation, but not for maintenance of gene activity. Due to the autoregulation, the gene's response becomes in a wide range independent of the exact strength of the input signal. During later stages of flower development (in *Arabidopsis* from stage 5 on), mRNA of *DEF*- and *GLO*-like genes is detected only in whorls 2 and 3 (Goto and Meyerowitz, 1994; Jack et al., 1992). This is so because upregulation and maintenance of class B gene expression in *Arabidopsis* and *Antirrhinum* during later stages of flower development depends on both DEF and GLO, due to an autoregulatory loop involving these proteins (Figure 7.1, bottom). The proteins encoded by class B genes of *Arabidopsis* and *Antirrhinum* are stable and functional in the cell only as heterodimers, i.e. DEF-GLO complexes, because both nuclear lo-

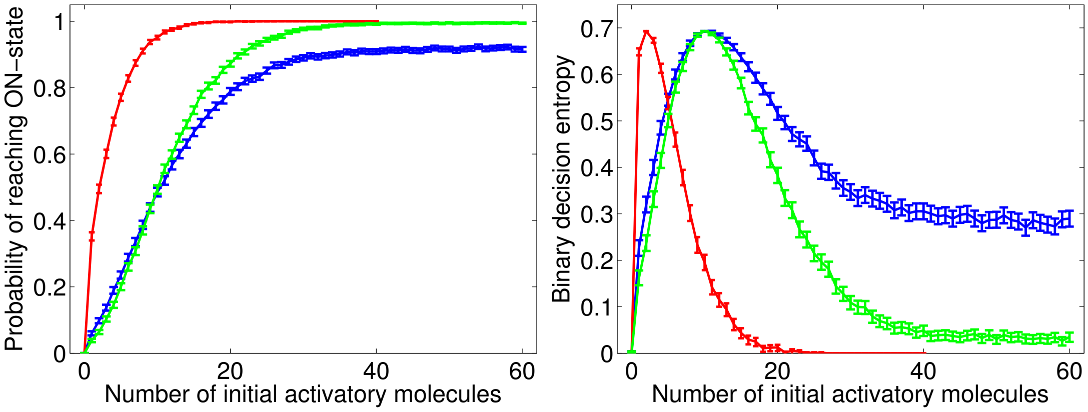


Figure 7.2: Statistical analysis of 10,000 independent runs for each given number of regulatory input molecules. Probability of reaching ON-state (left) and uncertainty of ON-OFF decision (right) for one autoregulatory gene (blue), the two-gene circuit immediately after duplication (red) and with obligate heterodimerisation (green). Shown are estimated values and 99% confidence intervals. Parameters are as in Table 7.2.

calisation and sequence-specific DNA-binding depend on obligate heterodimerisation (McGonigle et al., 1996; Schwarz-Sommer et al., 1992). Class B protein heterodimers bind to specific *cis*-regulatory DNA sequence elements termed 'CArG-boxes' (consensus 5'-CC(A/T)6GG-3'). Except *PI*, the promoter regions of all class B genes of *Arabidopsis* and *Antirrhinum* contain CArG-boxes that are involved in positively regulating class B gene expression (Chen et al., 2000; Hill et al., 1998; Tilly et al., 1998). These data, together with the total functional interdependence of the two class B gene paralogues, strongly corroborate the hypothesis that positive autoregulatory control of class B genes involves heterodimers of class B proteins that bind to CArG-boxes in the promoters of class B genes (Figure 7.1, bottom) (Tröbner et al., 1992). Since *PI* lacks CArG-boxes in a minimal promoter region, the autoregulatory feedback may work indirectly in this case (Chen et al., 2000; Honma and Goto, 2000).

Evolution of obligate heterodimerisation Obligate heterodimerisation of their encoded products involved in positive autoregulation explains why *DEF*-like and *GLO*-like genes are functionally non-redundant and totally interdependent. This raises the question as to how and why such a regulatory system originated in evolution. Studies on the interaction of class B protein orthologues from diverse gymnosperms and angiosperms suggested that, following a gene duplication within the class B gene clade, obligate heterodimerisation evolved in two steps from homodimerisation via facultative

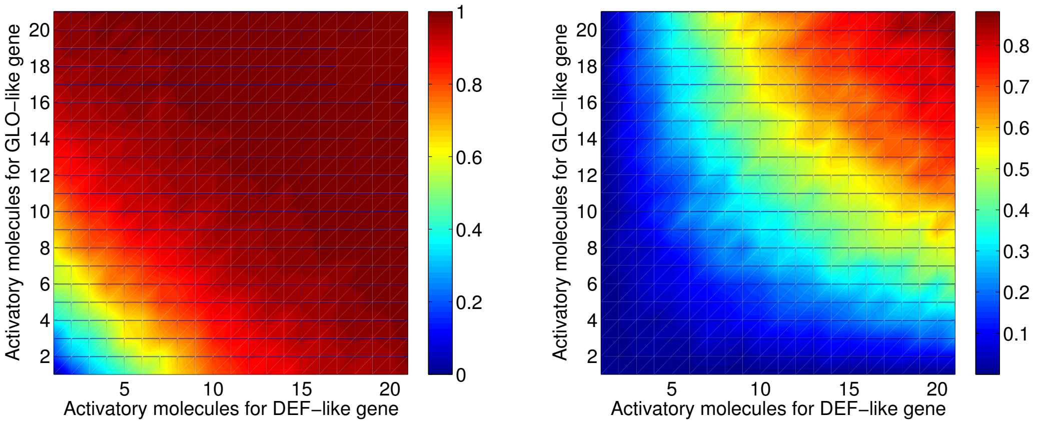


Figure 7.3: Switching behaviour of the two-gene circuits simulating two independent inputs, under facultative heterodimerisation (left) and obligate heterodimerisation (right). X- and Y-axis denote the number of available input molecules for *DEF* and *GLO*, respectively. The probability of ending in the ON state is indicated by colour: blue is low, red is high. Parameters are as in Table 7.2.

heterodimerisation (Winter et al., 2002b). Meanwhile obligate heterodimerisation of DEF-like with GLO-like proteins has also been observed outside of the eudicots *Arabidopsis* and *Antirrhinum* in diverse groups of monocots, suggesting that it originated quite early or several times independently during angiosperm evolution (Whipple and Schmidt, 2006).

So why then did obligate heterodimerisation evolve? In principle, it could represent a neutral change in protein-protein interactions that occurred by random genetic drift (Winter et al., 2002b). This cannot be excluded at the moment, but for several reasons, it appears not very likely. Even though obligate heterodimerisation originated early or several times independently within class B proteins, it did not occur in any other class of floral homeotic proteins, suggesting some kind of functional specificity. Moreover, it occurs within evolutionary especially 'successful' (e.g., species-rich) groups of angiosperms, suggesting that it might provide some selective advantage.

Obligate heterodimerisation canalises flower development Winter et al. (2002b) suggested that obligate heterodimerisation in combination with autoregulation may have provided a selective advantage because of the fixation of class B gene expression patterns and thus the spatial domain of the floral homeotic B-function within the flower during evolution. Mutational changes in the promoter region of only one class B gene that expand the gene's expression domain may leave the late and functionally especially relevant expression domain of the class B genes unchanged, because expression of the other partner would be missing in the ectopic expression domain. Only parallel changes in both types of class B genes, which are much less likely than changes in single genes, could lead to ectopic expression of the B-function under the assumption of obligate heterodimerisation and strong autoregulation. Thus obligate heterodimerisation may have evolved in parallel, or even as a prerequisite, of the canalisation of floral development and thus standardisation of floral structure in some groups of flowering plants (Winter et al., 2002b).

Obligate heterodimerisation provides robustness against developmental noise Amending this 'evolutionary' explanation of obligate heterodimerisation, this contribution tests the hypothesis that obligate heterodimerisation also provides advantages during development by providing robustness against wrong cell-fate decisions caused by stochastic noise. To this end, we put forward and examine a set of stochastic *in silico* models of class B gene and protein interactions as shown in Figure 7.1. The models enabled us to study the influence of noise in isolation from other factors, and allowed the comparison of three major stages in the envisioned path of evolutionary transitions (Figure 7.1): (A) One ancestral gene positively regulates its transcription via a homodimer of its own gene product; (B) Two genes positively regulate their transcription via homo- and heterodimers of both types of products; this very likely represents the situation directly after duplication of the ancestral gene; (C) Obligate heterodimerisation of the two products for regulation, i.e. the situation in extant *Arabidopsis* and *Antirrhinum.* Since only a small number of individual transcription factors is actually in the nucleus at any time (Honma and Goto, 2000; Winter et al., 2002b), stochastic fluctuations play a large role in the behaviour of gene regulatory circuits, and may have an influence on their evolutionary dynamics (Kaneko, 2007; Raser and O'Shea, 2005).

7.2 Methods

7.2.1 Model description and analysis

Model description Each model consists of a set of reactions for transcription factor binding, transcription, dimerisation, and decay (Table 7.1), where translation is modelled in one step together with dimerisation for efficiency. Transcription factor binding and unbinding are simple reaction processes, where we assume that exactly one functional copy of both *DEF* and *GLO* genes are available. For simplicity, we assume that only DNA activated by transcription factor binding is transcribed; however, experiments with basal transcription rates have led to qualitatively similar results. The decision to model translation and dimerisation in one step was taken to simplify the model while keeping the focus on transcriptional rather than translational regulation. This entails that we only model *DEF* and *GLO* mRNA and the dimerised proteins, but not the single DEF and GLO proteins. The slight loss of accuracy here has been unavoidable, as we needed to keep the model computationally tractable for the large numbers of replicated experiments. Each reaction is associated with a propensity function (Tables 7.2 and 7.3), which yields the probability of an occurrence of that reaction in a time step. A list of all modelled reactions is given in Table 7.1, and the full model is shown in Figure 7.5. All constituents of the model decay with a linear rate.

Corresponding ODE system contains two stable fixed points Linear stability analysis of the corresponding differential equation system reveals that both the active and the inactive state constitute stable fixed points in all three systems, with an unstable fixed point in between. The ODE system describing a deterministic approximation of the obligate heterodimerisation system can be described as follows, where x and y given the fraction (probability) of inactive *def* and *glo* genes, D and G given the concentration of *DEF* and *GLO* transcription factor proteins, and H is the concentration of *DEF-GLO* heterodimers:

$$\begin{aligned}
\frac{dx}{dt} &= k_{\text{off}}(1-x) - k_{\text{on}}xH && \boxed{7.1} \\
\frac{dy}{dt} &= k_{\text{off}}(1-y) - k_{\text{on}}yH && \boxed{7.2} \\
\frac{dD}{dt} &= \beta(1-x) - d_D D && \boxed{7.3} \\
\frac{dG}{dt} &= \beta(1-y) - d_G G && \boxed{7.4} \\
\frac{dH}{dt} &= 1/2k_{12}DG - d_H H - k_{\text{on}}(x+y)H + k_{\text{off}}(2-x-y) && \boxed{7.5}
\end{aligned}$$

The basal transcription rate β_0 is set to 0. The rate k_{12} determining the conversion of *DEF* and *GLO* proteins into heterodimers has to be halved, since it is now a deterministic rate constant. Similar ODE systems can be set up for the case of one gene and facultative heterodimerisation.

With the parameter values provided in Table 7.2 and 7.3, the ODE system above has three fixed points: $(x, y, D, G, H) = (1, 1, 0, 0, 0), (0.8, 0.8, 10, 10, 2.5), (0.2, 0.2, 40, 40, 40)$. The first one corresponds to the OFF-state: the probability that the genes are inactive is 1, and the concentration of all proteins is zero. The last case corresponds to the ON-state, both genes are activated 80% of the time, and 40 copies of each protein and the heterodimer are in the system. Note the agreement with Figure 7.4. The second fixed point corresponds to the threshold between OFF- and ON-state.

By the looks of it, we expect the first and third fixed point (OFF- and ON-state) to be stable, and the second one (threshold) to be unstable. Linear stability analysis shows that this is indeed the case. We can determine the Jacobian matrix J of the ODE system:

$$J = \begin{bmatrix} -k_{\text{off}} - k_{\text{on}}H & 0 & 0 & 0 & -k_{\text{on}}x \\ 0 & -k_{\text{off}} - k_{\text{on}}H & 0 & 0 & -k_{\text{on}}y \\ -\beta & 0 & -d_D & 0 & 0 \\ 0 & -\beta & 0 & -d_G & 0 \\ -k_{\text{off}} - k_{\text{on}}H & -k_{\text{off}} - k_{\text{on}}H & k_{12}GLO & k_{12}DEF & -d_H - k_{\text{on}}(x+y) \end{bmatrix}$$

The eigenvalues of this matrix, determined at the fixed points, are all negative for the first and third fixed point, thus characterising them as asymptotically stable. The Jacobian for the second fixed point (the threshold) contains a positive entry, characterising it as a saddle point and thus as unstable. Equivalent results hold for the one-gene system and for the system with facultative heterodimerisation.

7.2.2 Simulation and evaluation

Stochastic simulation The model is simulated using the Gillespie algorithm (Gillespie, 1977), implemented as a C++ function linked to MATLAB (The MathWorks, Inc. 2008). This method, which simulates an exact instance of the stochastic master equation, explicitly accounts for each reaction event and thus represents stochastic effects in full detail. Using the Gillespie algorithm, the exact order and timing of reactions is then stochastically determined, based on the propensities. To model transient activation of the circuits, we simulate an inflow of activating molecules (summarising all different activating transcription factors other than DEF/GLO that act on the respective genes) over 50 minutes of simulated time. After this time, the inflow is switched off and the system equilibriates, i.e. reaches a state in which no change occurs except for stochastic fluctuations (always reached after 72 hours of simulated time). If at this point gene product dimers are still present, the circuit is considered as active (full expression), otherwise it is inactive (no expression of class B genes). We conducted 10,000 experiments for each parameter combination.

Simulating different modes of regulation The different types of regulation are achieved by enabling or disabling the binding and activation of one type of gene by either a transcription factor homodimer produced by itself, a heterodimer of the products of both genes, or a homodimer of the proteins encoded by the other gene (see Table 7.3).

Transient initial activation of transcription The class B floral homeotic genes are regulated by a number of (possibly interacting) transcription factors, some of which are still unknown. Since the aim of this contribution is to investigate the effect of autoregulation on gene activity, we summarise the effects of all upstream transcription factors in two specific input factors, I_{DEF} and I_{GLO}, and a common input factor, I_C.

As developmental switches, the B-genes are transiently activated by their inputs, which are switched off after activation. Depending on the level of gene activity reached by that time, this activity either stays high or decays to a low value again, corresponding to on- and off-states of the genes. To model the transient activation, an inflow of (on average) N activatory molecules (of type I_{DEF}, I_{GLO} or I_C, respectively) over a period of T minutes was simulated using a Poisson process. After time T, the inflow is switched

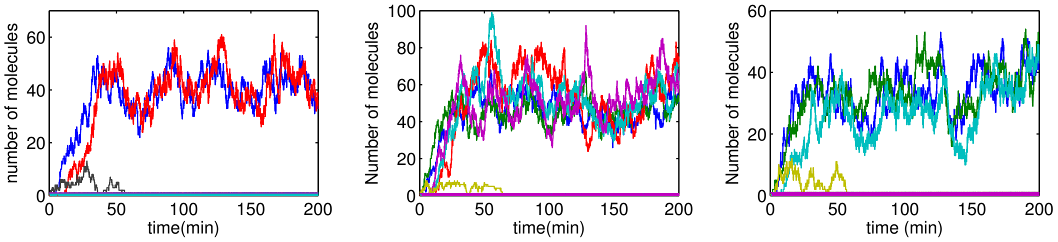


Figure 7.4: Single runs from all three modes of regulation. Left: one single gene; middle: two genes directly after duplication; right: obligate heterodimerisation of the transcription factors. Lines in yellow and black show the inputs transcription factors, which are switched off after 200 sec. Blue and green lines show the DEF and GLO transcripts, and red, purple and cyan denote the DEF-DEF, GLO-GLO and DEF-GLO dimers, respectively. The seemingly solid line at the bottom shows the activity level of the genes, which changes very rapidly and cannot be distinguished at this resolution.

off and the system is left alone, reaching steady state. Figure 7.4 shows example time courses for all three modes of regulation considered here.

Binary entropy as a measure of decision uncertainty All three systems investigated in this work represent autoactivatory circuits, which are used by the plant to establish the expression (ON-state) or non-expression (OFF-state) of homeotic genes in certain floral whorls. Therefore, a decision has to be made, depending on the number of activatory input molecules initially coming into the system. For low numbers of input molecules, the decision should be 'OFF', for higher numbers it should be 'ON'.

To measure the uncertainty of this decision, we use the binary entropy function. Let X be a random variable that takes value 1 with probability p, value 0 with probability $1-p$, i.e. a Bernoulli trial. The entropy of X is defined as

$$H(X) = -p * log(p) - (1-p) * log(1-p).$$

In our case, X taking value 1 means that the system reaches ON-state, value 0 means OFF-state. For each specific number N of activatory input molecules I_C, we repeated the simulation 10,000 times and determined the probability p (Figure 7.2, left). Using the formula above, this translates to the binary entropy, or decision uncertainty (Figure 7.2, right).

Calculating all probabilities: the stochastic master equation To get a better understanding of the system behaviour, and to support the conclusions drawn from its simulated realisations, we developed a master equation approximation of the

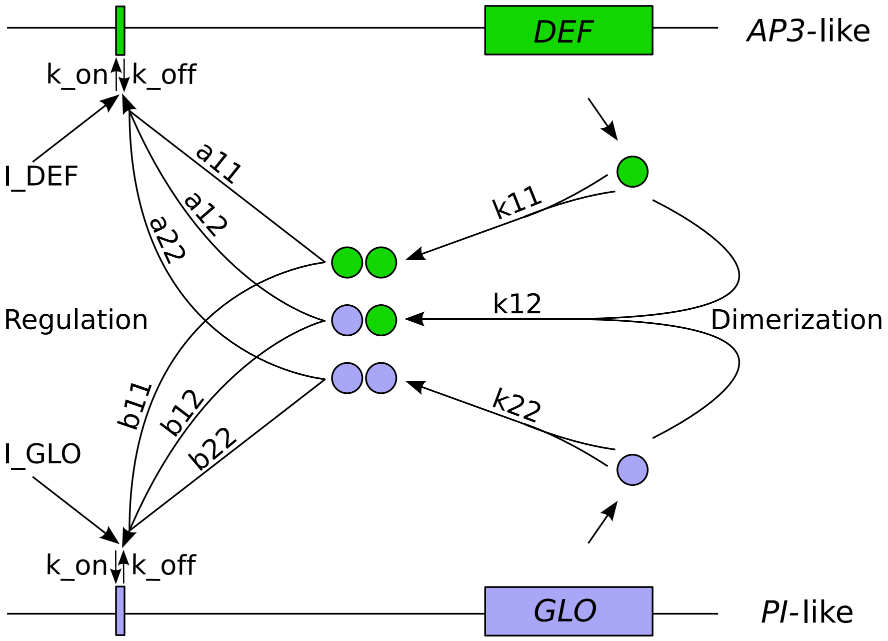


Figure 7.5: The full model showing all regulatory parameters. The three different model instances are generated by setting the rate constants according to Table 7.3

system under all three types of regulation. This enabled us to complement the statistical results from a large number of individual simulations with an analytical description of the probability distribution for different numbers of activatory transcription factors. Specifically, the master equation approach provides a system of ordinary differential equations, in which each equation describes the rate of change in the probability of one specific system state. We constructed a system with up to 150 molecules of both DEF and GLO-like type, leading to 22.500 variables which are sparsely coupled. To be able to solve this system numerically, two approximations had to be made. On the one hand, transcription factor binding is assumed to be fast in comparison to the transcription/translation process, so that the binding/unbinding process which was explicitly incorporated in the full model could be replaced by a probability for each gene to be bound to a transcription factor. On the other hand, we also assumed the dimerisation to be fast (i.e. in equilibrium), such that the number of dimers available for regulation is a function of the current number of monomers in the system. We are aware that the latter approximation is quite crude, and the results gained from this should be treated with care. With these two assumptions, we could reduce the dynamics of the full system to a system containing only the numbers of the two gene products as variables, which was tractable by numerical solvers in MATLAB.

7.2.3 Details on kinetic parameters

Stochastic and deterministic rate constants differ for higher-order reactions It is important to note that since we are using a stochastic framework, the rate parameters given in Table 7.2 and 7.3 are stochastic rate constants rather than deterministic ones. As already shown by Gillespie (Gillespie, 1977), a consequence of this is that the rate constants for first-order reactions (transcription, decay) are given in "1/min on average", but are otherwise equivalent to deterministic rates, whereas the rate constants for dimerisation (second-order) are given in "1/(min * l) on average", since the volume in which the reaction takes place (given in l) plays a role. All except the interaction parameters describe stochastic rates, i.e. they characterise the average rate of the corresponding reactions. In other words, if reaction $R : M1+M2 \longrightarrow 2*M1$ has rate constant c, then $c \times dt$ is the probability that a specific pair of molecules of types $M1$ and $M2$ will react in the short time-step dt.

Dimerisation parameters summaries a complex process of transport, translation and dimerisation Parameters k_{ij} summarise the processes of transport out of the nucleus, translation, dimerisation, and transport back into the nucleus into one number, which yields a process occurring in the time-scale of one hour, in accordance with estimations for plant cells (Günter Theißen, personal correspondence) (with given parameters, it takes $\sim$ 69 min to transcribe 50% of a given amount of mRNA in the nucleus to gene products, dimerise outside of the nucleus, and move back in).

Ratio between k_{on} and k_{off} characterises activation threshold Most important for the behaviour of the model are the transcription factor binding and unbinding rates k_{on} and k_{off}. Their ratio $K_D = k_{off}/k_{on}$, often called the "dissociation constant", characterises the approximate threshold between ON-state and OFF-state of the whole system. This can be seen by noting that in a bimolecular binding reaction

$$A + B \leftrightarrow AB,$$

the dissociation constant KD denotes the equilibrium ratio between the concentrations of unbound A and B and the dimer AB:

$$K_D = \frac{[A] \times [B]}{[AB]}$$

(where [X] denotes the equilibrium concentration of chemical X). In our case, A is the gene (of which only one copy exists), B is the transcription factor, and AB the activated form of the gene. Since we are dealing with discrete particle numbers, [AB] denotes the probability that the gene is activated by the transcription factor. Also note that [A] + [AB] = 1 holds. Combining all this yields

$$K_D = \frac{(1 - [AB]) \times [B]}{[AB]}$$

and thus

$$[AB] = \frac{[B]}{KD + [B]}.$$

Since [AB] is the activation probability of the gene, we conclude that K_D is exactly the concentration of the transcription factor that is needed to activate the gene with probability 0.5, and therefore the threshold for activation.

Interaction parameters define specific model topology In contrast to the chemical parameters above, the interaction parameters a_{ij} and b_{ij} are purely phenomenological. They describe the proportion of dimers that are involved in the regulation of the two transcription sites, according to the full model shown in Figure 7.5. In this work, they are limited to the states “dimer can/cannot regulate site”, i.e., they are set to either one or zero. Parameter a_{11} represents the influence of the DEF-DEF homodimer, a_{12} the influence of the DEF-GLO heterodimer, and a_{22} the one of the GLO-GLO homodimer, all on the *DEF*-like gene. In an analogous fashion, b_{ij}s denote the influences on the *GLO*-like gene. For example, if $a_{12} = 1$, the DEF-GLO heterodimer can bind and activate the *DEF*-like gene, while this binding and activation is disabled for $a_{12} = 0$.

Process	Reaction	Propensity
TF-binding	DEF + $\mathrm{TF}_{\mathrm{DEF}}$ → DEF*	$k_{on}[\mathrm{TF}_{\mathrm{DEF}}]$
	GLO + $\mathrm{TF}_{\mathrm{GLO}}$ → GLO*	$k_{on}[\mathrm{TF}_{\mathrm{GLO}}]$
TF-unbinding	DEF* → DEF + $\mathrm{TF}_{\mathrm{DEF}}$	k_{off}
	GLO* → GLO + $\mathrm{TF}_{\mathrm{GLO}}$	k_{off}
Translation	def + def → def + def + DEF-DEF	k_{11}[def]([def]-1)/2
+ Dimerisation	glo + glo → glo + glo + GLO-GLO	k_{22}[glo]([glo]-1)/2
	def + glo → def + glo + DEF-GLO	k_{12}[def][glo]
Decay	def → ∅	d[def]
	glo → ∅	d[glo]
	DEF-DEF → ∅	d[DEF-DEF]
	GLO-GLO → ∅	d[GLO-GLO]
	DEF-GLO → ∅	d[DEF-GLO]

Table 7.1: A summary of all reactions in the model. Given are the reaction equations and their associated propensity functions. In the one gene model, only gene *DEF* is considered, standing as a surrogate for the ancestral gene of both *DEF* and *GLO*. [X] denotes the number of particles of chemical X in the system. Genes are specified in italics, mRNA in small letters, and proteins in capitals. $\mathrm{TF}_{\mathrm{DEF}}$ and $\mathrm{TF}_{\mathrm{GLO}}$ summarise the transcription factors acting on the genes in the specific model, e.g. in the obligatory heterodimerisation model, $\mathrm{TF}_{\mathrm{DEF}} = \mathrm{TF}_{\mathrm{GLO}}$ = DEF-GLO, while in the system after duplication $\mathrm{TF}_{\mathrm{DEF}}$ = {DEF-DEF, GLO-GLO, DEF-GLO}.

Parameters are biologically plausible Even though the model gives a quantitative description of the reaction processes, the results can only be taken qualitatively since the exact kinetic parameters have not yet been measured in plants. The parameters used in this study have been chosen in accordance with experimental evidence from other systems (Alon, 2006; Kaern et al., 2005; Smolen et al., 1998; Tian and Burrage, 2006), especially the overview table 2.1 from (Alon, 2006). Our parameter choices ensure that the model replicates the time-scales involved in homeotic gene regulation, i.e. the main processes involved happen on a timescale of hours. We also varied the parameters in a range of 10% up or down and did not observe a significant change in the model behaviour.

Note: The fact that the dimerisation constant k_{12} is only half the value of k_{11} and k_{22} is not due to biological plausibility, but due to an implementation error that

Parameter	Units	Value
β	molecules/min	10.0
β_0	molecules/min	0.0
k_{on}	1/min	0.8
k_{off}	molecules/min	8.0
d	1/min	0.2

Table 7.2: Parameters that are kept constant in all experiments. β is the production propensity for gene products for both genes when they are activated, while β_0 is their base-level production rate. k_{on} and k_{off} give the binding and unbinding propensities of regulatory dimers to both genes, while d is the decay rate uniformly used for mRNA, dimers and initial activatory molecules

was only discovered while writing up this thesis. Since all parameter values in this chapter are only used as plausible approximations, I deemed it unnecessary to repeat the computations with the corrected formula, but rather adapted the rate constant.

Parameter (Unit)	One gene	After duplication	Obligatory heterodimerisation
a_{11} (proportion)	1.0	1.0	0
a_{12} (proportion)	0	1.0	1.0
a_{22} (proportion)	0	1.0	0
b_{11} (proportion)	0	1.0	0
b_{12} (proportion)	0	1.0	1.0
b_{22} (proportion)	0	1.0	0
k_{11} (1/(min*l))	0.01	0.01	0
k_{12} (1/(min*l))	0	0.005	0.005
k_{22} (1/(min*l))	0	0.01	0

Table 7.3: Parameters that are varied between the three experiments. a_{ij} and b_{ij} are binary parameters that determine which types of dimers regulate which gene, while k_{ij}s describe the stochastic rate constants in the dimerisation propensities for all combinations of monomers.

7.3 Results

The activation of the DEF and GLO genes depends on a temporally limited concerted action of many more genes and proteins besides the class B genes themselves, which have been described from an evo-devo perspective (Kaufmann et al., 2005) and by mathematical modelling (Alvarez-Buylla et al., 2007). To keep the focus on the self-regulation of the genetic switch, we summarise these in one common or two distinct activators for both genes, respectively. In the first experiment we used a common regulator to temporally activate both genes, and investigated the switching behaviour of the three circuits with regard to the number of available activatory input molecules.

Network topology determines switching threshold Looking at the probability of reaching full expression (Figure 7.2, left), the most probable state in the one-gene circuit switches from no steady-state expression (resulting in a non-class B cell identity) to full expression (class B, i.e. petal or stamen cell) at approximately 10 input molecules. Gene duplication without further mutational changes leads to a 3 times lower switching threshold (Figure 7.2, left), which may entail a drastically increased zone of class B gene expression in the flower. Mutations leading to obligate heterodimerisation again increase the activation threshold to the previous level, thus

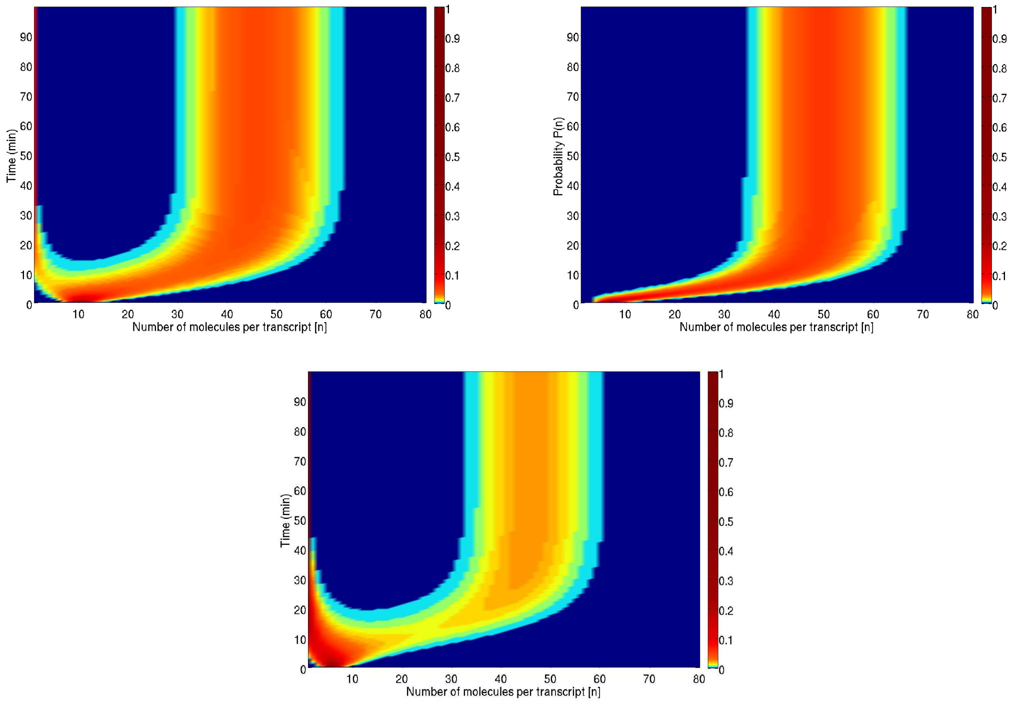


Figure 7.6: Numerical solutions to the master equations for one autoregulatory gene (top left), two genes after duplication (top right), and the two gene circuit with obligate heterodimerisation (bottom). All three graphs show the time-evolution of the probability to have n transcripts from which the activatory dimers are made

restoring the class B gene expression region (Figure 7.2, left). Therefore, in contrast to the facultative heterodimerisation circuit, obligate heterodimerisation results in the same switching threshold and thus the same domain of expression as just one autoregulatory gene. This result is in contradiction to an intuitive expectation that two genes can produce twice as many dimers as a single gene. With obligate heterodimerisation, however, the heterodimers assemble from translated products of one DEF and one GLO mRNA intermediate, while the homodimer in the one- gene system is produced from two translated proteins of the same type. Because mRNA is not used up in translation, this leads to equal production rates for the heterodimer in the obligate heterodimerisation system and the homodimer in the one-gene system.

To solidify and support the individual stochastic experiments, we numerically determined the probabilities for different numbers of molecules in the nucleus over time, i.e. we solved the stochastic master equation for simplified versions of the three models.

Figure 7.6 illustrates the result for a starting configuration with 10 activatory molecules, i.e. near the switching threshold of the regulatory circuits with one gene or obligate heterodimerisation. For these two systems, one can see how after a transient phase, the probability is high for around 80 activatory molecules (full expression) and zero molecules (no expression). For facultative heterodimerisation, probability is concentrated at the state of full expression. These results are in accordance to the simulated specific realisations and thus support the conclusions from a different perspective.

Obligate heterodimerisation provides robustness to incorrect cell-fate identity decisions To look at the robustness of the switching decision against stochastic noise, we calculated the decision uncertainty (binary entropy), thus more uncertainty implies less robustness. Focusing on the two circuits with identical expression domains, this uncertainty is nearly equal in the first and third circuit for small numbers of activatory input molecules, until the peak of uncertainty is reached. In contrast, the probability for a decision against class B gene mediated cell identity despite large numbers of activatory input molecules is significantly higher in the one-gene circuit than in the circuit with obligate heterodimerisation. With 60 activatory molecules, the probability for such a 'false negative' in the former circuit is still 10%, while the latter one achieves nearly 100% correct decisions under our conditions (Figure 7.2, right).

Hence, comparing one autoregulatory class B gene with the circuit after duplication and reduction to obligate heterodimerisation, our model suggests that an important difference lies in the response to larger numbers of activatory molecules, where the latter system exhibits a clearly reduced tendency to switch off by mistake. This is explained by the fact that although the circuit needs both DEF-like and GLO-like proteins to sustain activation, its two pools of gene products provide a buffer to temporary stochastic failure of one of the two genes. This is especially important during the initial phase of activation, where circuits that are supposed to lock themselves into permanent expression are susceptible to a run of 'bad luck', i.e. the supposedly-active genes are inactive over a longer period of time. Obligate heterodimerisation of gene products therefore provides a way to gain robustness against wrong cell identity decisions while retaining the original expression domain of one autoregulatory gene.

Obligate heterodimerisation avoids self-sensitivity We can phrase this idea of "buffering" in a more formal way: From the Jacobian matrix of the ODE system, information can be derived on the sensitivity of the concentration levels of each protein to changes in the concentration of other proteins (the *sensitivity coefficients*). Looking at those sensitivities, we can see that in the obligate heterodimerisation system, the sensitivity of [*DEF*-*GLO*] to changes in [*DEF*] is $k_{12}GLO$, and vice versa (where [ ] denotes concentration levels). In contrast, the concentration of a homodimer *DEF*-*DEF* has sensitivity $k_{11}DEF$ with respect to [*DEF*]. This means that for the obligate heterodimerisation system, the sensitivity of the heterodimer to one TF does not depend of that TF's concentration level, while for the homodimer there is a linear dependence (with respect to its own constituents). Now image a run of "bad luck", i.e. one of the genes is inactive for some time, and thus the concentration level of its TF drops. With some delay, the dimer concentration also drops. When the gene is activated again, the one-gene system has a hard time to recover, because the sensitivity of its homodimer concentration is linearly depended on the TF concentration, which is low. Even if some TF is produced, the dimer concentration takes a long time to respond. In contrast, a heterodimer production is not sensitive to the low population of that one TF (as long as the other TF is still at high levels). This explains why this increased robustness is found in both systems containing heterodimers, but not in the one-gene system.

Loss of homodimerisation ability switches from OR to AND logic Even though the mechanisms of the initial activation of *DEF*-like and *GLO*-like genes appear to be quite similar, they are very likely not identical (Chen et al., 2000), since the initial expression patterns of *DEF*- and *GLO*-like genes are slightly different. In *Arabidopsis* flowers at an early developmental stage 3, *AP3* (*DEF*-like) is expressed in the organ primordia of whorls 2 and 3, but also in parts of whorl 1, while *PI* (*GLO*-like) is expressed in whorls 2-4 at the same stage [15,16]. In contrast, the *AP3* orthologue *DEF* is expressed weakly in the organ primordia of whorl 4 (carpels) and very weakly in those of whorl 1 (sepals), while the *PI* orthologue *GLO* is expressed in sepal but not carpel primordia of early stages during Antirrhinum flower development (Schwarz-Sommer et al., 1992; Tröbner et al., 1992). To investigate the consequences of independent input into both genes, we explored a model setting in which the *DEF*-like and the *GLO*-like gene are activated independently by two input signals. Our experiments showed that immediately after gene duplication, the mode of integration represents a logical 'OR', meaning that both inputs can independently switch on the circuit (Figure 7.3, left). In this case, each input has the role of the one input present before duplication. After the transition to obligate heterodimerisation, a logic 'AND' function is achieved (Figure 7.3, right), thus both inputs are needed for activation.

7.4 Discussion

We are providing here, to the best of our knowledge, the first rationale, developmental genetic explanation for the intricate design of a genetic switch controlling class B floral homeotic gene expression in core eudicots, involving obligate heterodimerisation and positive autoregulatory feedback of two duplicate genes or their protein products, respectively. The increased robustness against unwanted deactivation by chance found in case of obligate heterodimerisation strongly suggests that this mechanism has a distinct advantage when the number of available regulatory molecules is small, leading to less cells of wrong identity in a floral organ and therefore to sharper organ identity transitions. It should be noted that since the mathematical model applies to any system with obligate heterodimerisation and positive feedback, the conclusions drawn here also transfer to any such system.

Obligate heterodimerisation with positive feedback is not common However, to the best of our knowledge, the phenomenon of obligate heterodimerisation together with positive feedback seems quite rare in genetic regulation outside of flower development, potentially due to the high cost of maintaining this system together with a strong dependence of the predicted fitness gain on external factors that might be specific for the situation depicted here.

Obligate heterodimerisation may satisfy the need for sharp expression domains and stringent control In the standard ABC model, class A and C genes are mutually antagonistic (Coen and Meyerowitz, 1991; Krizek and Fletcher, 2005), while class B genes have no floral homeotic 'repressor', possibly explaining the class-specific need for sharpened expression domains and thus obligate heterodimerisation, which is not found in the other two gene classes. However, Zhao et al. (2007) recently reported that the antagonistic expression of class A and class C genes is involved in defining the expression domain of class B genes in Arabidopsis, suggesting that our observation may not be sufficient to explain the obligate heterodimerisation of class B proteins. Taking a different perspective, the evolution of a regulatory 'AND' function out of an 'OR' function may have provided the plant with a more stringent control of the class B floral homeotic genes depending on different induction signals. The fact that there must be different inputs into *DEF*- and *GLO*-like genes is obvious from gene expression studies (see above), but its functional importance may have escaped the attention of previous investigations because of the coordinate upregulation and functional importance of *DEF*- and *GLO*-like genes in the second and third floral whorl. Our results suggest that identifying these different induction pathways, and clarifying their molecular mechanisms (e.g., *trans*-acting factors and *cis*-regulatory DNA motifs in *DEF*-like and *GLO*-like genes being involved) would enable an important step forward in understanding class B floral homeotic gene function in flowering plants.

Differential timing in the feedback loops might provide further advantages An interesting speculation about further advantages of positive feedback via obligate heterodimerisation arises if the feedback in the two loops acts on different timescales. Brandman et al. (2005); Brandman and Meyer (2008) show that for dual-positive feedback loop with OR logic, in which the stimulus acts as a catalyst (and is thus always required for the ON state), a fast feedback loop provides quick activation of the circuit in response to the stimulus, while the slow feedback prevents premature deactivation due to stochastic fluctuations in the stimulus level. Transferring this idea to the B-gene system is not straightforward, since the wiring of the circuit is drastically different. However, this idea opens up a new direction in the study of genetic networks, and will certainly be further explored in subsequent studies.

Proposal for transgenic studies with altered promoter logic The functional implication of these different input signals, and hence also of our hypothesis, could be tested by transgenic experiments. For example, *Arabidopsis* class B gene mutants in which both the *AP3* and the *PI* gene have been brought under the control of the *AP3* or the *PI* promoter rather than every gene under its own promoter (as in the wild-type) should affect the spatial or temporal development of petals or stamens, or both. Transgenic plants mutated at the *pi* locus (*pi-1*) in which wild-type PI is expressed under control of the *AP3* promoter (*5D3*) have already been reported (Lamb and Irish, 2003). These plants were used only as control for other experiments and have therefore not been described in much detail concerning the traits of interest here. However, it is clear that the *5D3::PI pi-1/pi-1* plants do not just show petals in the second floral whorl and stamens in the third floral whorl, as wild-type plants do; rather, they frequently develop sepal/petal mosaics in the second whorl, and mosaic organs or even carpels in the third whorl. These observations support our hypothesis concerning the functional importance of different induction pathways controlling the expression of *DEF*- and *GLO*-like genes for a proper development of organ identity in whorls two and three. More detailed analyses should be done to better understand how exactly the transgenic plants deviate from wild-type plants, and why. In addition, complementary transgenic studies in which AP3 is expressed under control of the *PI* gene promoter (*pPI*) should be performed in order to determine whether the *pPI::AP3 ap3/ap3* plants have also developmental defects. The construction of a transgenic plant with switched promoters (i.e. *pAP3::PI pPI::AP3 ap3/ap3 pi/pi*) would also be of great interest. Due to the apparently symmetric roles of *AP3* and *PI*, one might speculate that this phenotype shows less deviation from the wild type than the transgenic plants with both genes under the control of a single promoter.

Positive selection acts on DEF-GLO heterodimerisation domain If the origin of obligate heterodimerisation of class B proteins during evolution provided some plants with selective advantages, one may expect that this had an impact on the molecular evolution of these proteins, which indeed seems to be the case. Class B floral homeotic proteins are MIKC-type MADS-domain proteins characterised by a defined domain structure, including a MADS (M), Intervening (I), Keratin-like (K) and a C-terminal (C) domain (Kaufmann et al., 2005; Theißen et al., 2000). The K-domain mediates heterodimerisation of GLO- and DEF-like proteins and has been postulated to fold into three amphipatic α-helices termed K1, K2 and K3 (Yang et al., 2003). In accordance with the expectations mentioned above, phylogenetic data indicate that after the duplication leading to *DEF*-like and *GLO*-like gene lineages, positive selection acted on the sections of these genes encoding the K-domain (Hernández-Hernández et al., 2007). Intriguingly, one site under positive selection (Hernández-Hernández et al., 2007) is in a subdomain of K1 ("position 97-102" according to Yang et al., 2003) proposed to be critical for heterodimerisation specificity of DEF- and GLO-like proteins, as revealed by yeast two-hybrid analyses (Yang et al., 2003).

Given that the duplicates resulting from one homodimerising protein would be capable of homo- as well as heterodimerisation, our results suggest that positive selection should have enforced the loss of the homodimerisation ability, since our model with duplicated class B genes and obligatory heterodimerisation implies an expression domain of class B genes that is closer to the previous, functional domain (with one gene) than the one with facultative heterodimerisation. It has been proposed that within the subdomain of K1 mentioned above, the interaction of Glu-97 in PI and Arg-102 in AP3 facilitates specific heterodimerisation between AP3 and PI and prevents formation of homodimers (Yang et al., 2003). For these sites, however, positive selection has not been detected (Hernández-Hernández et al., 2007). Clearly, the relationships between the molecular evolution and biophysical interactions of DEF- and GLO-like proteins deserve more detailed studies in the future.

An evolutionary motif The evolutionary process which might have created the obligatory heterodimerisation system is quite simple: A transcription factor that activates itself via a homodimer of its own product undergoes gene duplication, and the two copies subsequently lose their homodimerisation ability. The atomar and clear nature of this process has invited us to term it an "evolutionary motif", i.e. a process that might have happened repeatedly in evolution. This process would take genetic circuits of a certain type (in our case developmental genetic circuits which use positive feedback via a homodimer) and improve it in a predictable way (by duplication of the TF and subsequent loss of the homodimerisation ability). Even though we cannot present any similar circuit at the moment, the idea seems suggestive enough to spark new investigations in this direction.

7.5 Conclusion

All in all, our findings strongly support the view that the unexpected complexity of the floral homeotic gene switch considered here was not simply produced by random genetic drift but evolved because it provided the plant with a clear selective advantage. This might have led to the establishment of this regulatory motif in a whole range of plant species. Theoretical support for this hypothesis stems from the idea that the requirement for robustness is the main driver for the evolution of cellular complexity (Hartwell et al., 1999b; Lauffenburger, 2000; Stelling et al., 2004).

It is intriguing that at least some basal angiosperms (which lack the obligate heterodimerisation system described here) do not have sharp, but 'fading borders' of expression of orthologues of *DEF*-like and *GLO*-like genes as well as gradual transitions in organ identity (Soltis et al., 2007). This underlines the hypothesis (Winter et al., 2002b) that the mechanism described here improves developmental robustness and thus helped to canalise the development and hence also the evolution of flowers within angiosperm evolution.

In summarising and evaluating this study, I believe that it is not only the insight about the specific obligate heterodimerisation system that is of value, but that the generality of the method needs to be pointed out as well. The application of stochastic modelling and simulation techniques to genetic circuits is not new and well-established, but in this case it is the comparative nature of the simulations which yields the result. Each of the three models investigated above, taken in isolation, behaves in the expected manner. However, by comparing them, we can take a detailed look at the behavioural differences, and this difference is much less obvious and predictable. Ultimately, it is this difference that natural selection acts upon, and so it requires as much attention as the common, predictable switch-type mode of operation of all three models.

In a recent study, Çağatay et al. (2009) have implemented the same approach *in vivo* to differentiate between a gene regulatory circuit in *Bacillus subtilis* and an artificially engineered alternative. In the same manner as here, they show that evolution has selected for a circuit design that is favourable in terms of noise handling. Their results show that the presented *in silico* approach can, in principle, be verified *in vivo*.

8
Concluding remarks

Evolution of function in biological networks In this thesis, I have looked at different perspectives on the structure-function mapping of biological network models: *(1)* Artificial evolution provides a way to synthesise network models automatically, with a given function or behaviour. A specific evolutionary algorithm was designed for this task and implemented in the SBMLevolver software (Chapter 2 and 3). *(2)* When such an evolutionary framework exists, it can be used to generate instances of network evolution, which can be analysed to shed light on general principles of network evolution. In Chapter 4, I detailed how the evolution of a network and of its organisational structure are related, based on the evolutionary trajectory of a chemical computing network. *(3)* For small network search spaces, topology enumeration and fitness screening have shown that the search space consists of large plateaus of networks with identical function, hindering the search on a evolutionary level (Chapter 5). *(4)* In the more biological parts of this thesis, evolutionary principles were used to fit kinetic parameters in dynamic network models for fluorescence microscopy data (Chapter 6). *(5)* In an opposite direction of research, stochastic network analysis was used to understand the evolutionary rationale (i.e. fitness advantage) of a certain gene-regulatory network structure (Chapter 7), proposing a specific evolutionary mechanism as an "evolutionary motif".

I present this very compressed compilation of the utilised methods in order to make one point: that the evolution of function in biological networks is a multi-faceted, complex process. It cannot be adequately described by a single equation, model, opinion, publication, or doctoral thesis. Rather, understanding functional evolution is a stepwise, iterative venture, which will continue in the future and will probably never be completed. Given the large neutral plateaus in network evolution (Banzhaf and

Leier, 2006; Gavrilets, 2004; Schuster et al., 1994; Wagner, 2008, Chapter 5), the role of chance events and genetic drift must not be underestimated. However, it is my belief that many discoveries in this field are yet to come, and that techniques such as the ones outlines in this thesis will play a role here.

Network evolution suffers from a lack of strong causality In any form of evolutionary process, strong causality is required for fitness improvements that go beyond random search. However, there is a large variety in the degree to which an evolutionary system is strongly causal. In the case of network evolution as implemented in the SBMLevolver approach (Chapter 2), strong causality is present, as proven by the comparison to random search on network topologies (Section 3.2). However, the enumeration of a complete search space for a simple problem (Chapter 5) has shown that topology search spaces contain large neutral plateaus, where the search for an improvement on the topological level is near to random search. Moreover, the structure-function mapping in networks is complex and depends on both structure and parameters (Section 1.2.1), further hindering (but not preventing!) the progress of an EA on network models.

Dynamic network models can give significant insights and quantifications that cannot be achieved otherwise In the last decade, fluorescence microscopy has opened a way to study kinetic properties of proteins in living cells. Using a combination of mathematical modelling and adequate parameter fitting algorithms, the kinetic parameters underlying this data can be revealed (Chapter 6). FRAP data has the advantage that instead of network function, it describes network behaviour (i.e. protein concentrations over time), which is closer to the level of the desired kinetic parameters. By comparing the fits of different model topologies to a dataset, one can decide which topology describes the data in a more adequate way. In the example of PML body formation, this method gave formal support to the biological hypothesis that PML proteins and most interaction partners bind to PML bodies in two different binding modes, as opposed to a single binding mode.

Stochastic dynamics reveal additional functionality The research described in Chapter 7 arose from a simple question: Why are there two class B floral homeotic genes in *Arabidopsis thaliana* and many other flowering plants? These genes are self-activatory via obligate heterodimerisation of their protein products, so that a single self-activatory gene could in principle fulfil the exact same regulatory logic. Using stochastic simulation, a comparison of the transition characteristics in these genetic switches was performed, revealing that the obligate heterodimerisation system features a significantly reduced error rate in the face of low protein copy numbers. This error reduction provides a potential fitness advantage and thus an explanation for the existence of this “overly-complicated” genetic wiring. In addition to presenting this particular hypothesis, the same method can be extended to other network comparisons, and even be implemented *in vivo*.

Network function in different stages of modelling Network function plays a large role in every part of the modelling process. In this thesis, the relation of network function to the different stages of modelling has been exemplified. Let us see: The evolutionary algorithm implemented in the SBMLevolver clearly is a tool for model creation, based on function. Model calibration (parameter fitting) is performed extensively for the PML body formation model, again under the guidelines of network function, given by FRAP data time courses. Last but not least, the analysis and comparison of models is illustrated with a stochastic simulation of the class B gene system in flower development. While these demonstrations of the “functional notion” of networks are by no means complete or exhaustive, I believe they provide insights into and valuable additions to their field of application.

References

Albert, Jeong, and Barabasi (2000). Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382. 11

Albert, R. (2005). Scale-free networks in cell biology. *Journal of Cell Science*, 118(Pt 21):4947–4957. 10

Alon, U. (2006). *An Introduction to Systems Biology: Design Principles of Biological Circuits*, volume 10 of *Mathematical & Computational Biology*. Chapman & Hall/CRC. 3, 7, 178

Alon, U. (2007). Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, 8(6):450–461. 7, 8, 14

Alon, U., Surette, M. G., Barkai, N., and Leibler, S. (1999). Robustness in bacterial chemotaxis. *Nature*, 397(6715):168–171. 9

Altenberg, L. (1995). The schema theorem and price's theorem. In Whitley, L. D. and Vose, M. D., editors, *Foundations of Genetic Algorithms*, volume 3, pages 23–49. Morgan Kaufmann. 20

Alvarez-Buylla, E. R., Benítez, M., Dávila, E. B., Chaos, A., Espinosa-Soto, C., and Padilla-Longoria, P. (2007). Gene regulatory network models for plant development. *Current Opinion in Plant Biology*, 10:83–91. 180

Alvarez-Buylla, E. R., Chaos, A., Aldana, M., Benítez, M., Cortes-Poza, Y., Espinosa-Soto, C., Hartasánchez, D. A., Lotto, R. B., Malkin, D., Santos, G. J. E., and Padilla-Longoria, P. (2008). Floral morphogenesis: stochastic explorations of a gene network epigenetic landscape. *PLoS ONE*, 3(11):e3626. 5

Alves, R. and Sorribas, A. (2007). In silico pathway reconstruction: Iron-sulfur cluster biogenesis in saccharomyces cerevisiae. *BMC Systems Biology*, 1:10. 114

Amoutzias, G. D., Robertson, D. L., Oliver, S. G., and Bornberg-Bauer, E. (2004). Convergent evolution of gene networks by single-gene duplications in higher eukaryotes. *EMBO Reports*, 5(3):274–279. 14, 15

Ando, S., Sakamoto, E., and Iba, H. (2002). Evolutionary modeling and inference of gene network. *Information Sciences*, 145(3–4):237–259. 27

Ascoli, C. A. and Maul, G. G. (1991). Identification of a novel nuclear domain. *Journal of Cell Biology*, 112(5):785–95. 121

Axelrod, D., Koppel, D. E., Schlessinger, J., Elson, E., and Webb, W. W. (1976). Mobility measurement by analysis of fluorescence photobleaching recovery kinetics. *Biophysical Journal*, 16(9):1055–1069. 125

Babu, M. M., Teichmann, S. A., and Aravind, L. (2006). Evolutionary dynamics of prokaryotic transcriptional regulatory networks. *Journal of Molecular Biology*, 358(2):614–633. 10, 12, 14

Bäck, T., Fogel, D., and Michalewicz, Z. (1997). *Handbook of Evolutionary Computation*. Oxford University Press. 19

Banks, D. and Fradlin, C. (2005). Anomalous diffusion of proteins due to molecular crowding. *Biophysical Journal*, 89:2960–2971. 148

Banzhaf, W. and Kuo, P. D. (2004). Network motifs in natural and artificial transcriptional regulatory networks. *Journal of Biological Physics and Chemistry*, 4:85–92. 14

Banzhaf, W. and Lasarczyk, C. W. G. (2005). Genetic programming of an algorithmic chemistry. In O'Reilly, U.-M., Yu, T., Riolo, R., and Worzel, B., editors, *Genetic Programming Theory and Practice II*, volume 8 of *Genetic Programming*, pages 175–190. Springer, New York. 30

Banzhaf, W. and Leier, A. (2006). Evolution on neutral networks in genetic programming. In Yu, T., Riolo, R., and Worzel, B., editors, *Genetic Programming Theory and Practice III*, pages 207–221. Springer US. 191

Banzhaf, W., Nordin, P., Keller, R. E., and Francone, F. D. (1998). *Genetic Programming - An Introduction: On the Automatic Evolution of Computer Programs and its Applications.* Morgan Kaufmann, dpunkt.verlag. 23

Barabási, A.-L. and Oltvai, Z. (2004). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5:101–112. 3, 10

Barkai, N. and Leibler, S. (1997). Robustness in simple biochemical networks. *Nature*, 387(6636):913–917. 9

Barrett, C. L. and Palsson, B. O. (2006). Iterative reconstruction of transcriptional regulatory networks: an algorithmic approach. *PLoS Computational Biology*, 2(5):e52. 27, 114

Bartz-Beielstein, T., Lasarczyk, C. W. G., and Preuss, M. (2005). Sequential parameter optimization. *Evolutionary Computation*, 1:773–780. 53

Bascompte, J. (2007). Networks in ecology. *Basic and Applied Ecology*, 8:485–490. 3

Beaudouin, J., Mora-Bermudez, F., Klee, T., Daigle, N., and Ellenberg, J. (2006). Dissecting the contribution of diffusion and interactions to the mobility of nuclear proteins. *Biophysical Journal*, 90:1878–1894. 127, 159

Beck, C. (2008). Evolution of conditional learning gene regulatory networks. Diplomarbeit, Department of Mathematics and Computer Science, Friedrich-Schiller-University Jena. 50, 61, 62, 63, 64

Berg, J., Lassig, M., and Wagner, A. (2004). Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications. *BMC Evolutionary Biology*, 4(1):51. 10

Bergthorsson, U., Andersson, D. I., and Roth, J. R. (2007). Ohno's dilemma: evolution of new genes under continuous selection. *Proceedings of the National Academy of Sciences USA*, 104(43):17004–17009. 13

Bernardi, R. and Pandolfi, P. P. (2007). Structure, dynamics and functions of promyelocytic leukaemia nuclear bodies. *Nature Reviews Molecular Cell Biology*, 8(12):1006–16. 121, 123, 158

Beyer, H. G. (1997). An alternative explanation for the manner in which genetic algorithms operate. *BioSystems*, 41(1):1–15. 46

Beyer, H.-G. and Schwefel, H.-P. (2002). Evolution strategies. *Natural Computing*, 1:3–52. 20, 21, 44

Bhalla, U. S. and Iyengar, R. (1999). Emergent properties of networks of biological signaling pathways. *Science*, 283(5400):381–387. 5

Bindschadler, M. and Sneyd, J. (2001). A bifurcation analysis of two coupled calcium oscillators. *Chaos*, 11(1):237–246. 38

Blonk, J., Don, A., Van Aalst, H., and Birmingham, J. (1993). Fluorescence photobleaching recovery in the confocal scanning light microscope. *Journal of Microscopy*, 169(3):363–374. 137

Boisvert, F., Kruhlak, M., Box, A., Hendzel, M., and Bazett-Jones, D. (2001). The transcription coactivator cbp is a dynamic component of the promyelocytic leukemia nuclear body. *Journal of Cell Biology*, 152:1099–1106. 157, 158

Bongard, J. and Lipson, H. (2007). Automated reverse engineering of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences USA*, 104(24):9943–9948. 28

Bongard, J. C. and Lipson, H. (2004). Automating genetic network inference with minimal physical experimentation using coevolution. In *Genetic and Evolutionary Computation, GECCO 2004*, LNCS, pages 333–345. Springer Berlin / Heidelberg. 28

Bornberg-Bauer, E., Beaussart, F., Kummerfeld, S. K., Teichmann, S. A., and Weiner, J. (2005). The evolution of domain arrangements in proteins and interaction networks. *Cellular and Molecular Life Sciences*, 62(4):435–445. 15

Bowman, F. (1958). *Introduction to Bessel Functions*. Courier Dover Publications, New York. 131

Braeckmans, K., Peeters, L., Sanders, N. N., De Smedt, S. C., and Demeester, J. (2003). Three-dimensional fluorescence recovery after photobleaching with the confocal scanning laser microscope. *Biophysical Journal*, 85(4):2240–2252. 129, 137

Braga, J., Desterro, J. M., and Carmo-Fonseca, M. (2004). Intracellular macromolecular mobility measured by fluorescence recovery after photobleaching with confocal laser scanning microscopes. *Molecular Biology of the Cell*, 15(10):4749–4760. 129, 137

Brand, P., Lenser, T., and Hemmerich, P. (2010). Assembly dynamics of PML nuclear bodies in living cell. *PMC Biophysics*, 3:3. 119, 146

Brandman, O., Ferrell, J. E., Li, R., and Meyer, T. (2005). Interlinked fast and slow positive feedback loops drive reliable cell decisions. *Science*, 310(5747):496–498. 186

Brandman, O. and Meyer, T. (2008). Feedback loops shape cellular signals in space and time. *Science*, 322(5900):390–395. 8, 186

Bray, D. and Lay, S. (1994). Computer simulated evolution of a network of cell-signaling molecules. *Biophysical Journal*, 66(4):972–977. 28, 35

Brown, C. M., Dalal, R. B., Hebert, B., Digman, M. A., Horwitz, A. R., and Gratton, E. (2008). Raster image correlation spectroscopy (rics) for measuring fast protein dynamics and concentrations with a commercial laser scanning confocal microscope. *Journal of Microscopy*, 229(Pt 1):78–91. 126

Carlson, J. M. and Doyle, J. (2002). Complexity and robustness. *Proceedings of the National Academy of Sciences USA*, 99 Suppl 1:2538–2545. 5

Carrero, G., Crawford, E., Hendzel, M., and de Vries, G. (2004). Characterizing fluorescence recovery curves for nuclear proteins undergoing binding events. *Bulletin of Mathematical Biology*, 66:1515–1545. 127

Carrero, G., McDonald, D., Crawford, E., de Vries, G., and Hendzel, M. J. (2003). Using frap and mathematical modeling to determine the in vivo kinetics of nuclear proteins. *Methods*, 29(1):14–28. 125, 129

Çağatay, T., Turcotte, M., Elowitz, M. B., Garcia-Ojalvo, J., and Süel, G. M. (2009). Architecture-dependent noise discriminates functionally analogous differentiation circuits. *Cell*, 139(3):512–522. 189

Centler, F., Kaleta, C., di Fenizio, P. S., and Dittrich, P. (2008). Computing chemical organizations in biological networks. *Bioinformatics*, 24(14):1611–1618. 11

Chen, D. and Huang, S. (2001). Nucleolar components involved in ribosome biogenesis cycle between the nucleolus and nucleoplasm in interphase cells. *Journal of Cell Biology*, 153:169–176. 158

Chen, X., Riechmann, J. L., Jia, D., and Meyerowitz, E. (2000). Minimal regions in the arabidopsis pistillata promoter responsive to the apetala3/pistillata feedback control do not contain a carg box. *Sexual Plant Reproduction*, 13(2):85–94. 168, 183

Cheutin, T., Gorski, S. A., May, K. M., Singh, P. B., and Misteli, T. (2004). In vivo dynamics of swi6 in yeast: evidence for a stochastic model of heterochromatin. *Mol Cell Biol*, 24(8):3157–3167. 144

Cheutin, T., McNairn, A. J., Jenuwein, T., Gilbert, D. M., Singh, P. B., and Misteli, T. (2003). Maintenance of stable heterochromatin domains by dynamic hp1 binding. *Science*, 299(5607):721–725. 124

Chu, D. (2007). Evolving genetic regulatory networks for systems biology. In *IEEE Congress on Evolutionary Computation, CEC 2007*, pages 875–882. 30

Chung, E. and Chen, R.-H. (2002). Spindle checkpoint requires mad1-bound and mad1- free mad2. *Molecular Biology of the Cell*, 13:1501–1511. 80

Coen, E. S. and Meyerowitz, E. M. (1991). The war of the whorls: genetic interactions controlling flower development. *Nature*, 353(6339):31–37. 165, 185

Conant, G. C. and Wagner, A. (2003). Convergent evolution of gene circuits. *Nature Genetics*, 34(3):264–266. 14

Conradi, C., Flockerzi, D., Raisch, J., and Stelling, J. (2007). Subnetwork analysis reveals dynamic features of complex (bio)chemical networks. *Proceedings of the National Academy of Sciences USA*, 104(49):19175–19180. 11

Craciun, G., Tang, Y., and Feinberg, M. (2006). Understanding bistability in complex enzyme-driven reaction networks. *Proceedings of the National Academy of Sciences USA*, 103(23):8697–8702. 11

Crombach, A. and Hogeweg, P. (2008). Evolution of evolvability in gene regulatory networks. *PLoS Computational Biology*, 4(7):e1000112. 16

Daniel, M. T., Koken, M., Romagné, O., Barbey, S., Bazarbachi, A., Stadler, M., Guillemin, M. C., Degos, L., Chomienne, C., and de Thé, H. (1993). Pml protein expression in hematopoietic and acute promyelocytic leukemia cells. *Blood*, 82(6):1858–67. 121

de Jong, H. (2002). Modeling and simulation of genetic regulatory systems: A literature review. *Journal of Computational Biology*, 9(1):67–103. 25

de Lichtenberg, U., Jensen, L. J., Brunak, S., and Bork, P. (2005). Dynamic complex formation during the yeast cell cycle. *Science*, 307(5710):724–727. 10

Deckard, A. and Sauro, H. (2004). Preliminary studies on the in silico evolution of biochemical networks. *ChemBioChem*, 5:1423–1431. 29, 35, 46, 68, 87, 104

Decraene, J., Mitchell, G. G., and McMullin, B. (2007). A molecular approach to complex adaptive systems. In *CS2007 - IEEE SMC UK and RI 6th Conference on Cybernetic Systems, 6-7 Sept 2007 , Dublin, Ireland*, pages 140–145. IEEE Systems, Man and Cybernetics Society. 30

Dellaire, G. and Bazett-Jones, D. P. (2004). Pml nuclear bodies: dynamic sensors of dna damage and cellular stress. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, 26(9):963–77. 121

Diaz-Herrera, P. (2006). What is a biological function? In *Proceeding of the 2006 conference on Formal Ontology in Information Systems*, pages 128–140, Amsterdam, The Netherlands, The Netherlands. IOS Press. 5

Digman, M. A., Brown, C. M., Sengupta, P., Wiseman, P. W., Horwitz, A. R., and Gratton, E. (2005). Measuring fast dynamics in solutions and cells with a laser scanning microscope. *Biophysical Journal*, 89(2):1317–1327. 126

Digman, M. A. and Gratton, E. (2009). Analysis of diffusion and binding in cells using the rics approach. *Microscopy Research and Technique*, 72(4):323–332. 126

Dittrich, P. and Speroni di Fenizio, P. (2007). Chemical organisation theory. *Bulletin of Mathematical Biology*, 69(4):1199–1231. 11, 87, 88

Dobrin, R., Beg, Q. K., Barabási, A.-L., and Oltvai, Z. N. (2004). Aggregation of topological motifs in the escherichia coli transcriptional regulatory network. *BMC Bioinformatics*, 5:10. 8

Dong, S., Stenoien, D., Qiu, J., Mancini, M., and Tweardy, D. (2004). Reduced intranuclear mobility of apl fusion proteins accompanies their mislocalization and results in sequestration and decreased mobility of retinoid x receptor alpha. *Molecular and Cellular Biology*, 24:4465–4475. 152, 158

D'Orazi, G., Cecchinelli, B., Bruno, T., Manni, I., Higashimoto, Y., Saito, S., Gostissa, M., Coen, S., Marchetti, A., and Del Sal, G. et al. (2002). . Nature Cell Biology. 4, .-. (2002). Homeodomain-interacting protein kinase-2 phosphorylates p53 at ser 46 and mediates apoptosis. *Nature Cell Biology*, 4:11–19. 154, 159

Dormand, J. R. and Prince, P. J. (1980). A family of embedded runge-kutta formulae. *Journal of Computational and Applied Mathematics*, 6:19–26. 141

Dräger, A., Supper, J., Planatscher, H., Magnus, J., Oldiges, M., and Zell, A. (2007). Comparing various evolutionary algorithms on the parameter optimization of the valine and leucine biosynthesis in corynebacterium glutamicum. In *CEC 2007. IEEE Congress on Evolutionary Computation*, pages 629–627. 104

Draghi, J. and Wagner, G. P. (2009). The evolutionary dynamics of evolvability in a gene network model. *Journal of Evolutionary Biology*, 22(3):599–611. 16

Droste, S. and Wiesmann, D. (2000). Metric based evolutionary algorithms. In *Proceedings of the European Conference on Genetic Programming*, pages 29–43, London, UK. Springer-Verlag. 43, 44

Droste, S. and Wiesmann, D. (2003). On the design of problem-specific evolutionary algorithms. In *Advances in evolutionary computing: theory and applications*, pages 153–173. Springer-Verlag New York, Inc., New York, NY, USA. 43, 44

Dundr, M., Hebert, M. D., Karpova, T. S., Stanek, D., Xu, H., Shpargel, K. B., Meier, U. T., Neugebauer, K. M., Matera, A. G., and Misteli, T. (2004). In vivo kinetics of cajal body components. *Journal of Cell Biology*, 164(6):831–42. 158

Dundr, M., Hoffmann-Rohrer, U., Hu, Q., Grummt, I., Rothblum, L. I., Phair, R. D., and Misteli, T. (2002). A kinetic framework for a mammalian rna polymerase in vivo. *Science*, 298(5598):1623–6. 127

Eiben, A. E. and Smith, J. E. (2003). *Introduction to Evolutionary Computing*. Natural Computing Series. Springer Verlag. 19

Elowitz, M. B., Levine, A. J., Siggia, E. D., and Swain, P. S. (2002). Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186. 165

Elson, E. L. (1985). Fluorescence correlation spectroscopy and photobleaching recovery. *Annual Review of Physial Chemistry*, 36:379–406. 126

Elson, E. L. (2004). Quick tour of fluorescence correlation spectroscopy from its inception. *Journal of Biomedical Optics*, 9(5):857–864. 126

Elson, E. L. and Magde, D. (1974). Fluorescence correlation spectroscopy. I. conceptual basis and theory. *Biopolymers*, 13:1–27. 125

Emmerich, M., Grötzner, M., and Schütz, M. (2001). Design of a graph-based evolutionary algorithm: A case study for chemical process networks. *Evolutionary Computation*, 9(3):329–354. 47

Eungdamrong, N. J. and Iyengar, R. (2004). Modeling cell signaling networks. *Biology of the Cell*, 96:355–362. 5

Everett, R. and Murray, J. (2005). Nd10 components relocate to sites associated with herpes simplex virus type 1 nucleoprotein complexes during virus infection. *Journal of Virology*, 79:5078–5089. 157, 158

Fang, G. (2002). Checkpoint protein bubr1 acts synergistically with mad2 to inhibit anaphase-promoting complex. *Molecular Biology of the Cell*, 13:755–766. 80

Feinberg, M. (1987). Chemical reaction network structure and the stability of complex isothermal reactors-I. the deficiency zero and deficiency one theorems. *Chemical Engineering Science*, 42(10):2229–2268. 11

Fernando, C. and Rowe, J. (2007). Natural selection in chemical evolution. *Journal of Theoretical Biology of the Cell*, 247(1):152–167. 89

Fernando, C. T., Liekens, A. M. L., Bingle, L. E. H., Beck, C., Lenser, T., Stekel, D. J., and Rowe, J. E. (2009). Molecular circuits for associative learning in single-celled organisms. *Journal of the Royal Society Interface*, 6(34):463–469. 50, 62

Ferrell, J. E. (2002). Self-perpetuating states in signal transduction: positive feedback, double-negative feedback and bistability. *Current Opinion in Cell Biology*, 14(2):140–148. 165

Festenstein, R., Pagakis, S. N., Hiragami, K., Lyon, D., Verreault, A., Sekkali, B., and Kioussis, D. (2003). Modulation of heterochromatin protein 1 dynamics in primary mammalian cells. *Science*, 299(5607):719–721. 124

Fisher, W. G., Yang, P. C., Medikonduri, R. K., and Jafri, M. S. (2006). NFAT and NFκB activation in T lymphocytes: a model of differential activation of gene expression. *Annals of Biomedical Engineering*, 34(11):1712–28. 60

Fogal, V., Gostissa, M., Sandy, P., Zacchi, P., Sternsdorf, T., Jensen, K., Pandolfi, P., Will, H., Schneider, C., and Del Sal, G. (2000). Regulation of p53 activity in nuclear bodies by a specific PML isoform. *EMBO Journal*, 19:6185–6195. 157, 159

Fontana, W. and Buss, L. (1994). 'The arrival of the fittest': Toward a theory of biological organization. *Bulletin of Mathematical Biology*, 56:1–64. 87, 88

François, P. and Hakim, V. (2004). Design of genetic networks with specified functions by evolution in silico. *Proceedings of the National Academy of Sciences USA*, 101:580–585. 29, 35, 46, 87

Funahashi, A., Tanimura, N., Morohashi, M., and Kitano, H. (2003). Celldesigner: a process diagram editor for gene-regulatory and biochemical networks. *Biosilico*, 1:159–162. 46

Fung, E., Wong, W. W., Suen, J. K., Bulter, T., Lee, S., and Liao, J. C. (2005). A synthetic gene-metabolic oscillator. *Nature*, 435(7038):118–22. 60

Gadkar, K. G., Gunawan, R., and Doyle, F. J. (2005). Iterative approach to model identification of biological networks. *BMC Bioinformatics*, 6:155. 28, 114

Gatermann, K., Eiswirth, M., and Senssea, A. (2005). Toric ideals and graph theory to analyze hopf bifurcations in mass action systems. *Journal of Symbolic Computation*, 40:1361–1382. 11

Gavrilets, S. (2004). *Fitness Landscapes and the Origin of Species*, volume 41 of *Monographs in Population Biology*. Princeton University Press. 192

Geier, F., Timmer, J., and Fleck, C. (2007). Reconstructing gene-regulatory networks from time series, knock-out data, and prior knowledge. *BMC Systems Biology*, 1(1):11. 26

Gennemark, P. and Wedelin, D. (2009). Benchmarks for identification of ordinary differential equations from time series data. *Bioinformatics*, 25(6):780–786. 27

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry*, 81(25):2340–2361. 173, 176

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA. 20

Gorski, S. and Misteli, T. (2005). Systems biology in the cell nucleus. *Journal of Cell Science*, 118(18):4083–4092. 126

Goto, K. and Meyerowitz, E. M. (1994). Function and regulation of the arabidopsis floral homeotic gene pistillata. *Genes and Development*, 8(13):1548. 166, 167

Guet, C. C., Elowitz, M. B., Hsing, W., and Leibler, S. (2002). Combinatorial synthesis of genetic networks. *Science*, 296(5572):1466–1470. 6

Guido, N., Wang, X., Adalsteinsson, D., McMillen, D., Hasty, J., Cantor, C., Elston, T., and Collins, J. (2006). A bottom-up approach to gene regulation. *Nature*, 439(7078):856–860. 87

Guo, A., Salomoni, P., Luo, J., Shih, A., Zhong, S., Gu, W., and Pandolfi, P. (2000). The function of PML in p53-dependent apoptosis. *Nature Cell Biology*, 2:730–736. 159

Guthke, R., Möller, U., Hoffmann, M., Thies, F., and Töpfer, S. (2005). Dynamic network reconstruction from gene expression data applied to immune response during bacterial infection. *Bioinformatics*, 21(8):1626–1634. 25, 114

Han, J.-D. J. (2008). Understanding biological functions through molecular networks. *Cell Res*, 18(2):224–237. 5

Han, J.-D. J., Bertin, N., Hao, T., Goldberg, D. S., Berriz, G. F., Zhang, L. V., Dupuy, D., Walhout, A. J. M., Cusick, M. E., Roth, F. P., and Vidal, M. (2004). Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, 430(6995):88–93. 10

Hancock, P. J. B. (1992). Genetic algorithms and permutation problems: A comparison of recombination operators for neural net structure specification. In Whitley, D. and Schaffer, J. D., editors, *Proceedings of the International Workshop on Combinations of Genetic Algorithms and Neural Networks (COGANN-92)*, pages 108–122. IEEE Computer Science, Los Alamitos, CA. 46

Handwerger, K., Murphy, C., and Gall, J. (2003). Steady-state dynamics of cajal body components in the xenopus germinal vesicle. *Journal of Cell Biology*, 160:495–504. 158

Hansen, N. and Kern, S. (2004). Evaluating the cma evolution strategy on multimodal test functions. In *Eighth International Conference on Parallel Problem Solving from Nature PPSN VIII*, Springer, pages 282–291. 57

Hansen, N. and Ostermeier, A. (2001). Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9:159–195. 21, 22, 23, 40, 141

Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999a). From molecular to modular cell biology. *Nature*, 402:C47–C52. 10

Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999b). From molecular to modular cell biology. *Nature*, 402(6761 Suppl):C47–C52. 189

Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation*. Prentice Hall. 30

Hecker, M., Lambeck, S., Toepfer, S., van Someren, E., and Guthke, R. (2009). Gene regulatory network inference: Data integration in dynamic models–a review. *BioSystems*, 96(1):86 – 103. 25

Heinrich, R. and Schuster, S. (1996). *The regulation of cellular systems*. Chapman & Hall, New York. 5

Hemmerich, P. and Diekmann, S., editors (2005). *Visions of the cell nucleus*. American Scientific Publishers, CA, USA. 121

Hemmerich, P., Weidtkamp-Peters, S., Hoischen, C., Schmiedeberg, L., Erliandri, I., and Diekmann, S. (2008). Dynamics of inner kinetochore assembly and maintenance in living cells. *Journal of Cell Biology*, 180(6):1101–14. 128

Hernández-Hernández, T., Martínez-Castilla, L. P., and Alvarez-Buylla, E. R. (2007). Functional diversification of b mads-box homeotic regulators of flower development: Adaptive evolution in protein-protein interaction domains after major gene duplication events. *Molecular Biology and Evolution*, 24(2):465–481. 187, 188

Hill, T. A., Day, C. D., Zondlo, S. C., Thackeray, A. G., and Irish, V. F. (1998). Discrete spatial and temporal cis-acting elements regulate transcription of the arabidopsis floral homeotic gene apetala3. *Development*, 125:1711–21. 168

Hink, M., Griep, R., Borst, J., van Hoek, A., Eppink, M., Schots, A., and Visser, A. (2000). Structural dynamics of green fluorescent protein alone and fused with a single chain Fv protein. *Journal of Biological Chemistry*, 275:17556–17560. 148

Hinow, P., Rogers, C. E., Barbieri, C. E., Pietenpol, J. A., Kenworthy, A. K., and DiBenedetto, E. (2006). The DNA binding activity of p53 displays reaction-diffusion kinetics. *Biophysical Journal*, 91(1):330–342. 160

Hintze, A. and Adami, C. (2008). Evolution of complex modular biological networks. *PLoS Computational Biolology*, 4(2):e23. 16

Hirche, S., Santibáñez-Koref, I., and Boblan, I. (2002). Design of strong causal fitness functions. In Abraham, A., del Solar, J. R., and Köppen, M., editors, *Soft Computing Systems Design, Management and Applications*, volume 87 of *Frontiers in Artificial Intelligence and Applications*, pages 183–192. IOS Press. 38

Hittinger, C. T. and Carroll, S. B. (2007). Gene duplication and the adaptive evolution of a classic genetic switch. *Nature*, 449(7163):677–681. 12

Hofmann, T., Moller, A., Sirma, H., Zentgraf, H., Taya, Y., Droge, W., Will, H., and Schmitz, M. (2002). Regulation of p53 activity by its interaction with homeodomaininteracting protein kinase-2. *Nature Cell Biology*, 4:1–10. 154, 159

Holland, J. H. (1973). Genetic algorithms and the optimal allocation of trails. *SIAM Journal on Computing*, 2(2):88–105. 20

Holland, J. H. (1975). *Adaptation in natural and artificial systems*. University of Michigan Press, Ann Arbor, MI. 20

Holland, J. H. (2002). Exploring the evolution of complexity in signaling networks. *Complexity*, 7(2):34–45. 29

Honma, T. and Goto, K. (2000). The arabidopsis floral homeotic gene pistillata is regulated by discrete cis-elements responsive to induction and maintenance signals. *Development*, 127(10):2021–30. 168, 170

Hooke, R. and Jeeves, T. A. (1961). “Direct search” solution of numerical and statistical problems. *Journal of the ACM*, 8(2):212–229. 57

Hoops, S., Sahle, S., Gauges, R., Lee, C., Pahle, J., Simus, N., Singhal, M., Xu, L., Mendes, P., and Kummer, U. (2006). Copasi - a complex pathway simulator. *Bioinformatics*, 22:3067–3074. 97

Hornberg, J. J., Bruggeman, F. J., Binder, B., Geest, C. R., de Vaate, A. J. M. B., Lankelma, J., Heinrich, R., and Westerhoff, H. V. (2005). Principles behind the multifarious control of signal transduction. erk phosphorylation and kinase/phosphatase control. *FEBS Journal*, 272:244–258. 60

Huang, C. Y. and Ferrell, J. E. J. (1996). Ultrasensitivity in the mitogen-activated protein kinase cascade. *Proceedings of the National Academy of Sciences USA*, 93(19):10078–83. 60

Huang, S., Eichler, G., Bar-Yam, Y., and Ingber, D. E. (2005). Cell fates as high-dimensional attractor states of a complex gene regulatory network. *Physical Review Letters*, 94:128701. 5

Huang, Y., Qiu, J., Chen, G., and Dong, S. (2007). Coiled-coil domain of pml is essential for the aberrant dynamics of pml-raralpha, resulting in sequestration and decreased mobility of SMRT. *Bio-*

chemical and Biophysical Research Communications, 365:258–265. 158

Hucka, M. and Finney, A. (2005). Escalating model sizes and complexities call for standardized forms of representation. *Molecular Systems Biology*, 1:2005.0011. 41

Hucka, M., Finney, A., Bornstein, B. J., Keating, S. M., Shapiro, B., Matthews, J., Kovitz, B. L., Schilstra, M. J., Funahashi, A., Doyle, J., and Kitano, H. (2004). Evolving a lingua franca and associated software infrastructure for computational systems biology: The systems biology markup language (sbml) project. *IEE Systems Biology*, 1(1):41–53. 41

Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., Arkin, A. P., Bornstein, B. J., Bray, D., Cornish-Bowden, A., Cuellar, A. A., Dronov, S., Gilles, E. D., Ginkel, M., Gor, V., Goryanin, I. I., Hedley, W. J., Hodgman, T. C., Hofmeyr, J.-H., Hunter, P. J., Juty, N. S., Kasberger, J. L., Kremling, A., Kummer, U., Novère, N. L., Loew, L. M., Lucio, D., Mendes, P., Minch, E., Mjolsness, E. D., Nakayama, Y., Nelson, M. R., Nielsen, P. F., Sakurada, T., Schaff, J. C., Shapiro, B. E., Shimizu, T. S., Spence, H. D., Stelling, J., Takahashi, K., Tomita, M., Wagner, J., Wang, J., and the SBML forum (2003). The systems biology markup language (sbml): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531. 41

Ibrahim, B., Diekmann, S., Schmitt, E., and Dittrich, P. (2008). In-silico modeling of the mitotic spindle assembly checkpoint. *PLoS ONE*, 3(2):e1555. 80, 81, 82

Ibrahim, B., Schmitt, E., Dittrich, P., and Diekmann, S. (2009). In silico study of kinetochore control, amplification, and inhibition effects in MCC assembly. *BioSystems*, 95(1):35–50. 57

Ideker, T. E., Thorsson, V., and Karp, R. M. (2000). Discovery of regulatory interactions through perturbation: inference and experimental design. *Pacific Symposium on Biocomputing*, pages 305–316. 27, 28

Ingram, P. J., Stumpf, M. P. H., and Stark, J. (2006). Network motifs: structure does not determine function. *BMC Genomics*, 7(1):108. 6

Ishov, A. M., Sotnikov, A. G., Negorev, D., Vladimirova, O. V., Neff, N., Kamitani, T., Yeh, E. T., Strauss, J. F., and Maul, G. G. (1999). Pml is critical for nd10 formation and recruits the pml-interacting protein daxx to this nuclear structure when modified by sumo-1. *Journal of Cell Biology*, 147(2):221–34. 123

Jack, T., Brockman, L. L., and Meyerowitz, E. M. (1992). The homeotic gene apetala3 of arabidopsis thaliana encodes a mads box and is expressed in petals and stamens. *Cell*, 68(4):683–697. 166, 167

James, T. C. and Elgin, S. C. (1986). Identification of a nonhistone chromosomal protein associated with heterochromatin in drosophila melanogaster and its gene. *Mol Cell Biol*, 6(11):3862–3872. 124

Jensen, K., Shiels, C., and Freemont, P. S. (2001). Pml protein isoforms and the rbcc/trim motif. *Oncogene*, 20(49):7223–33. 123

Jones, T. and Forrest, S. (1995). Fitness distance correlation as a measure of problem difficulty for genetic algorithms. In *Proceedings of the 6th International Conference on Genetic Algorithms*, pages 184–192, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. 37

Kacser, H. and Beeby, R. (1984). Evolution of catalytic proteins or on the origin of enzyme species by means of natural selection. *Journal of Molecular Evolution*, 20(1):38–51. 28

Kaern, M., Elston, T. C., Blake, W. J., and Collins, J. J. (2005). Stochasticity in gene expression: from theories to phenotypes. *Nature Reviews Genetics*, 6:451–464. 165, 178

Kaleta, C., Centler, F., di Fenizio, P. S., and Dittrich, P. (2008). Phenotype prediction in regulated metabolic networks. *BMC Systems Biology*, 2:37. 11

Kaneko, K. (2007). Evolution of robustness to noise and mutation in gene expression dynamics. *PLoS ONE*, 2(5):e434. 170

Kashtan, N. and Alon, U. (2005). Spontaneous evolution of modularity and network motifs. *Proceedings of the National Academy of Sciences USA*, 102(39):13773–13778. 16

Kashtan, N., Noor, E., and Alon, U. (2007). Varying environments can speed up evolution. *Proceedings of the National Academy of Sciences USA*, 104(34):13711–13716. 16

Kaufmann, K., Melzer, R., and Theißen, G. (2005). Mikc-type mads-domain proteins: structural modularity, protein interactions and network evolution in land plants. *Gene*, 347(2):183–198. 15, 166, 180, 187

Kentsis, A. and Borden, K. (2004). Physical mechanisms and biological significance of supramolecular protein self-assembly. *Current Protein and Peptide Science*, 5:125–134. 159

Kholodenko, B. N. (2006). Cell-signalling dynamics in time and space. *Nature Reviews Molecular Cell Biology*, 7(3):165–176. 5

Kholodenko, B. N., Kiyatkin, A., Bruggeman, F. J., Sontag, E., Westerhoff, H. V., and Hoek, J. B. (2002). Untangling the wires: a strategy to trace functional interactions in signaling and gene networks. *Proceedings of the National Academy of Sciences USA*, 99(20):12841–12846. 10, 26

Kim, S. T., Yoo, M. J., Albert, V. A., Farris, J. S., Soltis, P. S., and Soltis, D. E. (2004). Phylogeny and diversification of b-function mads-box genes in angiosperms: evolutionary and functional implications of a 260-million-year-old duplication. *American Journal of Botany*, 91:2102–2118. 166

Kitano, H. (2002). Systems biology: A brief overview. *Science*, 295(5560):1662 – 1664. 3

Klement, K., Lenser, T., and Hemmerich, P. (2010). HP1 proteins display isoform-specific binding characteristics in heterochromatin. *Molecular Biology of the Cell*, in preparation:??? 119

Klemm, K. and Bornholdt, S. (2005). Topology of biological networks and reliability of information processing. *Proceedings of the National Academy of Sciences USA*, 102(51):18414–18419. 9

Klingauf, M., Stanek, D., and Neugebauer, K. M. (2006). Enhancement of u4/u6 small nuclear ribonucleoprotein particle association in cajal bodies predicted by mathematical modeling. *Molecular Biology of the Cell*, 17(12):4972–81. 127

Klipp, E., Herwig, R., Kowald, A., Wierling, C., and Lehrach, H. (2006). *Systems Biology in Practice: Concepts, Implementation and Application*. Wiley-VCH, Weinheim. 3

Knabe, J. F., Nehaniv, C. L., and Schilstra, M. J. (2008). Do motifs reflect evolved function?-no convergent evolution of genetic regulatory network subgraph topologies. *BioSystems*, 94(1-2, Sp. Iss. SI):68–74. 7th International Workshop on Information Processing in Cells and Tissue, Oxford, ENGLAND, AUG 29-31, 2007. 9

Kofahl, B. and Klipp, E. (2004). Modelling the dynamics of the yeast pheromone pathway. *Yeast*, 21(10):831–50. 60

Koken, M. H. M., Puvion-Dutilleul, F., Guillemin, M. C., Viron, A., Linares-Cruz, G., Stuurman, N., de Jong, L., Szostecki, C., Calvo, F., and Chomienne, C. e. a. (1994). The t(15;17) translocation alters a nuclear body in a retinoic acid-reversible fashion. *EMBO Journal*, 13:1073–1079. 123

Kollmann, M., Løvdok, L., Bartholomé, K., Timmer, J., and Sourjik, V. (2005). Design principles of a bacterial signalling network. *Nature*, 438(7067):504–507. 10

Kongas, O. and van Beek, J. H. G. M. (2001). Creatine kinase in energy metabolic signaling in muscle. *Proceedings of the 2nd International Conference on Systems Biology (ICSB 2001)*, pages 198–207. 60

Koza, J., Keane, M., and Streeter, M. (2003). Evolving inventions. *Scientific American*, 288(2):52–59. 23

Koza, J., Mydlowec, W., Lanza, G., Yu, J., and Keane, M. (2001a). Automatic synthesis of both the topology and sizing of metabolic pathways using genetic programming. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2001)*, pages 57–65. Morgan Kaufmann. 27, 35, 87, 104, 114

Koza, J. R. (1992). *Genetic programming. On the programming of computers by means of natural selection*. Complex adaptive systems, Cambridge, MA: The MIT (Massachusetts Institute of Technology) Press. 23

Koza, J. R. (1994). *Genetic programming II: Automatic discovery of reusable programmes.* MIT Press, Cambridge, MA. 24

Koza, J. R., Mydlowec, W., Lanza, G., Yu, J., and Keane, M. A. (2001b). Reverse engineering of metabolic pathways from observed data using genetic programming. *Pacific Symposium on Biocomputing*, pages 434–445. 27, 35

Kreimer, A., Borenstein, E., Gophna, U., and Ruppin, E. (2008). The evolution of modularity in bacterial metabolic networks. *Proceedings of the National Academy of Sciences USA*, 105(19):6976–6981. 16

Krizek, B. and Fletcher, J. (2005). Molecular mechanisms of flower development: An armchair guide. *Nature Reviews Genetics*, 6(9):688–698. 185

Krizek, B. A. and Meyerowitz, E. M. (1996). The arabidopsis homeotic genes apetala3 and pistillata are sufficient to provide the b class organ identity function. *Development*, 122(1):11–22. 166

Kruhlak, M., Levert, M., Fischle, W., Verdin, E., Bazett-Jones, D., and Hendzel, M. (2000). The mobility of the GFP:ASF splicing factor in live cells. *Journal of Cell Biology*, 150:41–51. 158

Kuo, P., Banzhaf, W., and Leier, A. (2006a). Network topology and the evolution of dynamics in an artificial genetic regulatory network model created by whole genome duplication and divergence. *BioSystems*, 85:177–200. 87

Kuo, P. D., Banzhaf, W., and Leier, A. (2006b). Network topology and the evolution of dynamics in an artificial genetic regulatory network model created by whole genome duplication and divergence. *BioSystems*, 85(3):177–200. 14

Lachner, M., O'Carroll, D., Rea, S., Mechtler, K., and Jenuwein, T. (2001). Methylation of histone h3 lysine 9 creates a binding site for hp1 proteins. *Nature*, 410(6824):116–120. 124

Lamb, R. S. and Irish, V. F. (2003). Functional divergence within the apetala3/pistillata floral homeotic gene lineages. *Proceedings of the National Academy of Sciences USA*, 100(11):6558–6563. 186

Lamond, A. and Earnshaw, W. (1998). Structure and function in the nucleus. *Science*, 280:547–553. 121

Lanctôt, C., Cheutin, T., Cremer, M., Cavalli, G., and Cremer, T. (2007). Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. *Nature Reviews Genetics*, 8(2):104–115. 121

Langdon, W. B. and Poli, R. (2002). *Foundations of Genetic Programming.* Springer-Verlag Berlin Heidelberg. 23

Lauffenburger, D. A. (2000). Cell signaling pathways as control modules: complexity for simplicity? *Proceedings of the National Academy of Sciences USA*, 97(10):5031–5033. 189

Le Novère, N., Bornstein, B., Broicher, A., Courtot, M., Donizelli, M., Dharuri, H., Li, L., Sauro, H., Schilstra, M., Shapiro, B., L., J. S., and Hucka, M. (2006). BioModels database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Research*, 34:D689–91. 5, 57, 76

Le Novère, N., Finney, A., Hucka, M., Bhalla, U. S., Campagne, F., Collado-Vides, J., Crampin, E. J., Halstead, M., Klipp, E., Mendes, P., Nielsen, P., Sauro, H., Shapiro, B., Snoep, J. L., Spence, H. D., and Wanner, B. L. (2005). Minimum information requested in the annotation of biochemical models (MIRIAM). *Nature Biotechnology*, 23:1509–1515. 5

Leier, A., Kuo, P. D., and Banzhaf, W. (2007). Analysis of preferential network motif generation in an artificial regulatory network model created by dupliation and divergence. *Advances in Complex Systems*, 10(2):155–172. 14

Lele, T. P. and Ingber, D. E. (2006). A mathematical model to determine molecular kinetic rate constants under non-steady state conditions using fluorescence recovery after photobleaching (frap). *Biophysical Chemistry*, 120:32–35. 127

Lenser, T., Hinze, T., Ibrahim, B., and Dittrich, P. (2007). Towards evolutionary network reconstruction tools for systems biology. In Marchiori, E., Moore, J., and Rajapakse, J., editors, *Proceedings of the Fifth European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics (EvoBIO)*, volume 4447 of *LNCS*. 60, 69

Lenser, T., Matsumaru, N., Hinze, T., and Dittrich, P. (2008). Tracking the evolution of chemical computing networks. In Bullock, S., Noble, J., Watson, R., and Bedau, M. A., editors, *Artificial Life XI: Proceedings Eleventh International Conference on the Simulation and Synthesis of Living Systems*, pages 343–350. MIT Press, Cambridge, MA. 11, 85

Lenser, T., Theißen, G., and Dittrich, P. (2009). Developmental robustness by obligate interaction of class B floral homeotic genes and proteins. *PLoS Computational Biology*, 5(1):e1000264. 163

Levine, M. and Tjian, R. (2003). Transcription regulation and animal diversity. *Nature*, 424(6945):147–151. 13

Lewis, J. (2008). From signals to patterns: space, time, and mathematics in developmental biology. *Science*, 322(5900):399–403. 165

Li, D., Li, J., Ouyang, S., Wang, J., Wu, S., Wan, P., Zhu, Y., Xu, X., and He, F. (2006). Protein interaction networks of saccharomyces cerevisiae, caenorhabditis elegans and drosophila melanogaster: large-scale organization and robustness. *Proteomics*, 6(2):456–461. 11

Lin, D. Y., Huang, Y. S., Jeng, J. C., Kuo, H. Y., Chang, C. C., Chao, T. T., Ho, C. C., Chen, Y. C., Lin, T. P., and Fang, H. I. e. a. (2007). Role of SUMO-interacting motif in Daxx SUMO modification, subnuclear localization, and repression of sumoylated transcription factors. *Molecular Cell*, 24:341–354. 159

Lohmann, R. (1993). Structure evolution and incomplete induction. *Biological Cybernetics*, 69:319–326. 40

Luscombe, N. M., Babu, M. M., Yu, H., Snyder, M., Teichmann, S. A., and Gerstein, M. (2004). Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, 431(7006):308–312. 8

Lynch, M. (2007a). The evolution of genetic networks by non-adaptive processes. *Nature Reviews Genetics*, 8(10):803–813. 13

Lynch, M. (2007b). The frailty of adaptive hypotheses for the origins of organismal complexity. *Proceedings of the National Academy of Sciences USA*, 104 Suppl 1:8597–8604. 13

Machne, R., Finney, A., Muller, S., Lu, J., Widder, S., and Flamm, C. (2006). The sbml ode solver library: a native api for symbolic and fast numerical analysis of reaction networks. *Bioinformatics*, 22(11):1406–7. 50, 90

Magde, D., Elson, E., and Webb, W. W. (1972). Thermodynamic fluctuations in a reacting system: Measurement by fluorescence correlation spectroscopy. *Physical Review Letters*, 29(11):705–708. 125

Maison, C. and Almouzni, G. (2004). HP1 and the dynamics of heterochromatin maintenance. *Nature Reviews Molecular Cell Biology*, 5:296–304. 124

Mangan, S. and Alon, U. (2003). Structure and function of the feed-forward loop network motif. *Proceedings of the National Academy of Sciences USA*, 100(21):11980–11985. 6, 7

Martins, S. I. F. S. and Boekel, M. A. J. S. V. (2003). Kinetic modelling of Amadori N-(1-deoxy-D-fructos-1-yl)-glycine degradation pathways. Part II–kinetic analysis. *Carbohydrate Research*, 338(16):1665–78. 60

Marwan, W. (2003). Theory of time-resolved somatic complementation and its use to explore the sporulation control network in *Physarum polycephalum*. *Genetics*, 164(1):105–15. 57, 60

Matsumaru, N., Centler, F., Speroni di Fenizio, P., and Dittrich, P. (2007). Chemical organization theory as a theoretical base for chemical computing. *International Journal of Unconventional Computing*, 3(4):285–309. 90, 101

Matsumaru, N., Speroni di Fenizio, P., Centler, F., and Dittrich, P. (2006). On the evolution of chemical organizations. In Artmann, S. and Dittrich, P., editors, *Explorations in the complexity of possible life: abstracting and synthesizing the principles of living systems, Proceedings of the 7th German-Workshop of Artificial Life*, pages 135–146. Aka, Berlin. 87

Maul, G. G., Negorev, D., Bell, P., and Ishov, A. M. (2000). Review: Properties and assembly mechanisms of nd10, pml bodies, or pods. *Journal of Structural Biology*, 129(2-3):278–287. 121

Maul, G. G., Yu, E., Ishov, A. M., and Epstein, A. L. (1995). Nuclear domain 10 (nd10) associated proteins are also present in nuclear bodies and redistribute to hundreds of nuclear sites after stress. *Journal of Cellular Biochemistry*, 59(4):498–513. 121

Mazurie, A., Bottani, S., and Vergassola, M. (2005). An evolutionary and functional assessment of regulatory network motifs. *Genome Biology*, 6(4):R35. 8, 14

Mazza, D., Cella, F., Vicidomini, G., Krol, S., and Diaspro, A. (2007). Role of three-dimensional bleach distribution in confocal and two-photon fluorescence recovery after photobleaching experiments. *Applied Optics*, 46(30):7401–7411. 137

McAdams, H. H. and Arkin, A. (1997). Stochastic mechanisms in gene expression. *Proceedings of the National Academy of Sciences USA*, 94(3):814–819. 165

McGonigle, B., Bouhidel, K., and Irish, V. F. (1996). Nuclear localization of the arabidopsis apetala3 and pistillata homeotic gene products depends on their simultaneous expression. *Genes and Development*, 10(14):1812–21. 168

McGrath, J. L., Tardy, Y., Dewey, C. F., Meister, J. J., and Hartwig, J. H. (1998). Simultaneous measurements of actin filament turnover, filament fraction, and monomer diffusion in endothelial cells. *Biophysical Journal*, 75(4):2070–2078. 129

McKay, B. D. (1981). Practical graph isomorphism. *Congressus Numerantium*, 30:45–87. 49

Melnick, A. and Licht, J. D. (1999). Deconstructing a disease: Raralpha, its fusion partners, and their roles in the pathogenesis of acute promyelocytic leukemia. *Blood*, 93(10):3167–215. 123, 158

Meyvis, T. K. L., De Smedt, S. C., Van Oostveldt, P., and Demeester, J. (1999). Fluorescent recovery after photobleaching: a versatile tool for mobility and interaction measurements in pharmaceutical research. *Pharmaceutical Research*, 16:1153–1162. 129

Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M., and Alon, U. (2004). Superfamilies of evolved and designed networks. *Science*, 303(5663):1538–1542. 7

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs: Simple building blocks of complex networks. *Science*, 298:824–827. 3, 6, 7

Misteli, T. (2001). The concept of self-organization in cellular architecture. *J Cell Biol*, 155(2):181–185. 124

Moles, C. G., Mendes, P., and Banga, J. R. (2003). Parameter estimation in biochemical pathways: A comparison of global optimization methods. *Genome Research*, 13:2467–2474. 104

Montoya, J. M., Pimm, S. L., and Solé, R. V. (2006). Ecological networks and their fragility. *Nature*, 442:259–264. 3

Mueller, F., Wach, P., and McNally, J. G. (2008). Evidence for a common mode of transcription factor interaction with chromatin as revealed by improved quantitative fluorescence recovery after photobleaching. *Biophysical Journal*, 94(8):3323–3339. 129, 160

Nachman, I., Regev, A., and Friedman, N. (2004). Inferring quantitative models of regulatory networks from expression data. *Bioinformatics*, 20 Suppl 1:i248–i256. 26, 114

Naka, K., Ikeda, K., and Motoyama, N. (2002). Recruitment of NBS1 into PML oncogenic domains via interaction with SP100 protein. *Biochemical and Biophysical Research Communications*, 299:863–871. 159

Negorev, D. and Maul, G. G. (2001). Cellular proteins localized at and interacting within nd10/pml nuclear bodies/pods suggest functions of a nuclear depot. *Oncogene*, 20(49):7234–42. 123, 153

Niehaus, J., Igel, C., and Banzhaf, W. (2007). Reducing the number of fitness evaluations in graph genetic programming using a canonical graph indexed database. *Evolutionary Computation*, 15(2):199–221. 49

Nielsen, K., Sorensen, P. G., Hynne, F., and Busse, H. G. (1998). Sustained oscillations in glycolysis: an experimental and theoretical study of chaotic and complex periodic behavior and of quenching of simple oscillations. *Biophysical Chemistry*, 72(1-2):49–62. 60

Paladugu, S., Chickarmane, V., Deckard, A., Frumkin, J., McCormack, M., and Sauro, H. (2006). In silico evolution of functional modules in biochemical networks. *IEE Proceedings-Systems Biology*, 153(4):223–235. 29, 35, 46, 87, 104

Papa, R., Martin, A., and Reed, R. D. (2008). Genomic hotspots of adaptation in butterfly wing pattern evolution. *Current Opinion in Genetics and Development*, 18(6):559–564. 16

Parter, M., Kashtan, N., and Alon, U. (2007). Environmental variability and modularity of bacterial metabolic networks. *BMC Evolutionary Biology*, 7:169. 16

Pereira-Leal, J. B. and Teichmann, S. A. (2005). Novel specificities emerge by stepwise duplication of functional modules. *Genome Research*, 15(4):552–559. 10

Peters, R., Peters, J., Tews, K. H., and Bähr, W. (1974). A microfluorimetric study of translational diffusion in erythrocyte membranes. *Biochimica et biophysica acta*, 367(3):282–294. 125

Phair, R. and Misteli, T. (2000). High mobility of proteins in the mammalian cell nucleus. *Nature*, 404:604–609. 125, 158

Phair, R. D. and Misteli, T. (2001). Kinetic modelling approaches to in vivo imaging. *Nature Reviews Molecular Cell Biology*, 2(12):898–907. 127

Phair, R. D., Scaffidi, P., Elbi, C., Vecerová, J., Dey, A., Ozato, K., Brown, D. T., Hager, G., Bustin, M., and Misteli, T. (2004). Global nature of dynamic protein-chromatin interactions in vivo: three-dimensional genome scanning and dynamic interaction networks of chromatin proteins. *Molecular Cell Biology*, 24(14):6393–6402. 160

Pigolotti, S., Krishna, S., and Jensen, M. H. (2007). Oscillation patterns in negative feedback loops. *Proceedings of the National Academy of Sciences USA*, 104:6533–6537. 165

Pimm, S. L. (2003). *Food webs.* Univ. Chicago Press, Chicago. 3

Politi, A., Moné, M. J., Houtsmuller, A. B., Hoogstraten, D., Vermeulen, W., Heinrich, R., and van Driel, R. (2005). Mathematical modeling of nucleotide excision repair reveals efficiency of sequential assembly strategies. *Molecular Cell*, 19(5):679–690. 127

Price, N. D., Reed, J. L., and Palsson, B. O. (2004). Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nature Reviews Microbiology*, 2:886–897. 11

Prill, R. J., Iglesias, P. A., and Levchenko, A. (2005). Dynamic properties of network motifs contribute to biological network organization. *PLoS Biology*, 3(11):e343. 5, 9

Qi, Y. and Ge, H. (2006). Modularity and dynamics of cellular networks. *PLoS Computational Biology*, 2(12):e174. 10

Radcliffe, N. J. (1992). Non-linear genetic representations. In Männer, R. and Manderick, B., editors, *Parallel problem solving from nature*, volume 2, pages 259–285. North-Holland. 20

Raser, J. M. and O'Shea, E. K. (2005). Noise in gene expression: origins, consequences, and control. *Science*, 309(5743):2010–3. 165, 170

Rechenberg, I. (1973). *Evolutionsstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution.* frommann-holzbog, Stuttgart. 20, 21

Rechenberg, I. (1994a). Evolution strategy. In Zurada, J. M., Marks-II, R. J., and Robinson, C. J., editors, *Computational Intelligence - Imitating Life.* MIT Press. 20

Rechenberg, I. (1994b). *Evolutionsstrategie '94.* Frommann Holzboog. 40

Rohn, H. (2008). Systematische Untersuchung von Parametern der *in silico* - Evolution biologischer Zellsignalnetzwerke. Diplomarbeit (in German), Department of Mathematics and Computer Science, Friedrich-Schiller-University Jena. 53, 56

Rohn, H., Ibrahim, B., Lenser, T., Hinze, T., and Dittrich, P. (2008). Enhancing parameter estimation of biochemical networks by exponentially scaled search steps. In Marchiori, E. and Moore, J. H., editors, *Proceedings Fifth European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics (EvoBIO)*, volume 4973 of *LNCS*, pages 177–187. Springer Berlin / Heidelberg. 57, 58

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529. 26

Sakamoto, E. and Iba, H. (2001). Inferring a system of differential equations for a gene regulatorynetwork by using genetic programming. In *Proceedings of the 2001 Congress on Evolutionary Computation, CEC 2001*. 27, 35

Santibáñez Koref, I., Boblan, I., Lohnert, F., and Schütte, A. (2001). Causality and design of dynamical systems. In Giannakoglou, K., Tsahalis, D., Periaux, J., Papailiou, K., and Fogarty, T., editors, *Evolutionary methods for design, optimisation and control*. 38

Savageau, M. A. (1976). *Biochemical systems analysis: a study of function and design in molecular biology*. Addison-Wesley, Reading, MA. 26

Saxton, M. (2001). Anomalous subdiffusion in fluorescence photobleaching recovery: a monte carlo study. *Biophysical Journal*, 81:2226–2240. 128, 148

Schilling, C. H., Letscher, D., and Palsson, B. O. (2000). Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *Journal of Theoretical Biology*, 203(3):229–248. 5

Schmiedeberg, L., Weisshart, K., Diekmann, S., zu Hoerste, G. M., and Hemmerich, P. (2004). High- and low-mobility populations of hp1 in heterochromatin of mammalian cells. *Molecular Biology of the Cell*, 15(6):2819–2833. 124, 128, 144

Schuster, P., Fontana, W., Stadler, P., and Hofacker, I. (1994). From sequences to shapes and back: A case study in RNA secondary structures. *Proceedings of the Royal Society B: Biological Sciences*, 255:279–284. 192

Schuster, S., Dandekar, T., and Fell, D. A. (1999). Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol*, 17(2):53–60. 5, 11

Schuster, S., von Kamp, A., and Pachkov, M. (2007). Understanding the roadmap of metabolism by pathway analysis. In *Metabolomics*, volume 358 of *Methods in Molecular Biology*, pages 199–226. Humana Press. 11

Schwarz-Sommer, Z., Hue, I., Huijser, P., Flor, P., Hansen, R., Tetens, F., Lönning, W., Saedler, H., and Sommer, H. (1992). Characterization of the antirrhinum floral homeotic mads-box gene deficiens: evidence for dna binding and autoregulation of its persistent expression throughout flower development. *EMBO Journal*, 11:251–263. 166, 168, 183

Schwefel, H.-P. (1977). *Numerische Optimierung von Computer-Modellen mittels der Evolutionsstrategie*, volume 26 of *Interdisciplinary Systems Research*. Birkhäuser. 20

Seeler, J., Marchio, A., Sitterlin, D., Transy, C., and Dejean, A. (1998). Interaction of SP100 with HP1 proteins: a link between the promyelocytic leukemia-associated nuclear bodies and the chromatin compartment. *Proceedings of the National Academy of Sciences USA*, 95:7316–7321. 159

Sendhoff, B., Kreutz, M., and von Seelen, W. (1997). A condition for the genotype phenotype mapping: causality. In Bäck, T., editor, *Proceedings of the Seventh International Conference on Genetic Algorithms (ICGA'97)*. Morgan Kauffmann, San Francisco. 36, 37, 38, 44

Shen, T. H., Lin, H. K., Scaglioni, P. P., Yung, T. M., and Pandolfi, P. P. (2006). The mechanisms of pml-nuclear body formation. *Molecular Cell*, 24:331–339. 157

Shetty, R. P., Endy, D., and Knight, T. F. (2008). Engineering biobrick vectors from biobrick parts. *Journal of Biological Engineering*, 2:5. 61

Singh, P. B., Miller, J. R., Pearce, J., Kothary, R., Burton, R. D., Paro, R., James, T. C., and Gaunt, S. J. (1991). A sequence motif found in a drosophila heterochromatin protein is conserved in animals and plants. *Nucleic Acids Res*, 19(4):789–794. 124

Smolen, P., Baxter, D. A., and Byrne, J. H. (1998). Frequency selectivity, multistability, and oscillations emerge from models of genetic regulatory systems. *American Journal of Physiology*, 274:C531–42. 178

Solé, R. V. and Valverde, S. (2008). Spontaneous emergence of modularity in cellular networks. *Journal of the Royal Society Interface*, 5(18):129–133. 16

Soltis, D. E., Chanderbali, A. S., Kim, S., Buzgo, M., and Soltis, P. S. (2007). The abc model and its applicability to basal angiosperms. *Annals of Botany*, 100(2):155–63. 189

Sommer, H., Beltran, J. P., Huijser, P., Pape, H., Saedler, H., and Schwarz-Sommer, Z. (1990). Deficiens, a homeotic gene involved in the control of flower morphogenesis in antirrhinum majus: the protein shows homology to transcription factors. *EMBO Journal*, 9:605–613. 166

Soyer, O., Pfeiffer, T., and Bonhoeffer, S. (2006). Simulating the evolution of signal transduction pathways. *Journal of Theoretical Biology*, 241:223–232. 29, 35, 87

Soyer, O. S. and Bonhoeffer, S. (2006). Evolution of complexity in signaling pathways. *Proceedings of the National Academy of Sciences USA*, 103(44):16337–16342. 30

Speroni di Fenizio, P., Dittrich, P., and Banzhaf, W. (2001). Spontanous formation of cells in a universal artificial chemistry on a planar graph. In Kelemen, J. and Sosik, P., editors, *Advances in Artificial Life. Proc. 6th European Conference on Artificial Life (ECAL 2001)*, volume 2159 of *LNCS*, pages 206–215. Springer, Berlin. 88

Spieth, C., Streichert, F., Speer, N., and Zell, A. (2005). Inferring regulatory systems with noisy pathway information. In *German Conference on Bioinformatics, GCB 2005*. 26, 35

Spieth, C., Supper, J., Streichert, F., Speer, N., and Zell, A. (2006). Jcell–a java-based framework for inferring regulatory networks from time series data. *Bioinformatics*, 22(16):2051–2052. 26

Sprague, B. and McNally, J. (2005). FRAP analysis of binding: proper and fitting. *Trends in Cell Biology*, 15:84–91. 125, 150, 160

Sprague, B. L., Müller, F., Pego, R. L., Bungay, P. M., Stavreva, D. A., and McNally, J. G. (2006). Analysis of binding at a single spatially localized cluster of binding sites by fluorescence recovery after photobleaching. *Biophysical Journal*, 91(4):1169–1191. 138

Sprague, B. L., Pego, R. L., Stavreva, D. A., and McNally, J. G. (2004). Analysis of binding reactions by fluorescence recovery after photobleaching. *Biophysical Journal*, 86:3473–3495. 127, 138, 160

Stagge, P. and Igel, C. (2001). *Structure optimization and isomorphisms*, pages 409–422. Springer-Verlag, London, UK. 49

Steffen, M., Petti, A., Aach, J., D'haeseleer, P., and Church, G. (2002). Automated modelling of signal transduction networks. *BMC Bioinformatics*, 3:34. 114

Stelling, J., Klamt, S., Bettenbrock, K., Schuster, S., and Gilles, E. D. (2002). Metabolic network structure determines key aspects of functionality and regulation. *Nature*, 420(6912):190–193. 5

Stelling, J., Sauer, U., Szallasi, Z., Doyle, F. J., and Doyle, J. (2004). Robustness of cellular functions. *Cell*, 118(6):675–685. 9, 189

Stephani, A., no, J. C. N., and Heinrich, R. (1999). Optimal stoichiometric designs of atp-producing systems as determined by an evolutionary algorithm. *Journal of Theoretical Biology*, 199(1):45–61. 28

Sternsdorf, T., Gostissa, M., Sirma, H., Sal, G. D., Ruthard, M., Schmitz, M. L., Will, H., and Hofmann, T. G. (2005). Pml nuclear bodies: Cellular function and disease association. In Diekmann, S. and Hemmerich, P., editors, *Visions of the cell nucleus*. American Scientific Publishers, CA, USA. 154

Storn, R. and Price, K. (1997). Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, 11(4):341–359. 57

Strogatz, S. H. (2001). Exploring complex networks. *Nature*, 410:268–276. 3

Sugimoto, M., Kikuchi, S., and Tomita, M. (2005). Reverse engineering of biochemical equations from time-course data by means of genetic programming. *BioSystems*, 80(2):155–164. 27, 35

Swaminathan, R., Hoang, C., and Verkman, A. (1997). Photobleaching recovery and anisotropy decay of green fluorescent protein GFP-S65T in solution and cells: cytoplasmic viscosity probed by green fluorescent protein translational and rotational diffusion. *Biophysical Journal*, 72:1900–1907. 148

Swat, M., Kel, A., and Herzel, H. (2004). Bifurcation analysis of the regulatory modules of the mammalian G1/S transition. *Bioinformatics*, 20(10):1506–1511. 38

Tardy, Y., McGrath, J. L., Hartwig, J. H., and Dewey, C. F. (1995). Interpreting photoactivated fluorescence microscopy measurements of steady-state actin dynamics. *Biophysical Journal*, 69(5):1674–1682. 129

Teichmann, S. A. and Babu, M. M. (2004). Gene regulatory network growth by duplication. *Nature Genetics*, 36(5):492–496. 10, 12, 14

Teller, A. and Veloso, M. (1996). PADO: A new learning architecture for object recognition. In Ikeuchi, K. and Veloso, M., editors, *Symbolic visual learning*, pages 81–116. Oxford University Press. 23

Theißen, G., Becker, A., Rosa, A. D., Kanno, A., Kim, J. T., Munster, T., Winter, K. U., and Saedler, H. (2000). A short history of mads-box genes in plants. *Plant Molecular Biology*, 42(1):115–49. 166, 187

Theißen, G. and Saedler, H. (2001). Floral quartets. *Nature*, 409:469–71. 166

Tian, T. and Burrage, K. (2006). Stochastic models for regulatory networks of the genetic toggle switch. *Proceedings of the National Academy of Sciences USA*, 103:8372–8377. 178

Tilly, J. J., Allen, D. W., and Jack, T. (1998). The carg boxes in the promoter of the arabidopsis floral organ identity gene apetala3 mediate diverse regulatory effects. *Development*, 125(9):1647–57. 168

Tomassini, M., Vanneschi, L., Collard, P., and Clergue, M. (2005). A study of fitness distance correlation as a difficulty measure in genetic programming. *Evolutionary Computation*, 13(2):213–239. 37

Tröbner, W., Ramirez, L., Motte, P., Hue, I., Huijser, P., Lonnig, W. E., Saedler, H., Sommer, H., and Schwarz-Sommer, Z. (1992). Globosa: a homeotic gene which interacts with deficiens in the control of antirrhinum floral organogenesis. *EMBO Journal*, 11(13):4693–704. 166, 168, 183

Tsong, A. E., Tuch, B. B., Li, H., and Johnson, A. D. (2006). Evolution of alternative transcriptional circuits with identical logic. *Nature*, 443(7110):415–420. 6

Tyson, J. J. (1991). Modeling the cell division cycle: cdc2 and cyclin interactions. *Proceedings of the National Academy of Sciences USA*, 88(16):7328–32. 57, 60

Tyson, J. J., Chen, K. C., and Novak, B. (2003). Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. *Current Opinion in Cell Biology*, 15:221–231. 5

van Someren, E. P., Wessels, L., Reinders, M., and Backer, E. (2001). Searching for limited connectivity in genetic network models. In *Proceedings of the Second International Conference on Systems Biology*, pages 222–230. 27

Vanneschi, L., Tomassini, M., Collard, P., and Clergue, M. (2003). *Genetic Programming*, chapter Fitness Distance Correlation in Structural Mutation Genetic Programming, pages 455–464. Springer Berlin / Heidelberg. 37

Ventura, B. D., Lemerle, C., Michalodimitrakis, K., and Serrano, L. (2006). From in vivo to in silico biology and back. *Nature*, 443(7111):527–533. 3

Vermaak, D. and Malik, H. S. (2009). Multiple roles for heterochromatin protein 1 genes in drosophila. *Annual Reviews Genetics*, 43:467–492. 124

Veron, A. S., Kaufmann, K., and Bornberg-Bauer, E. (2007). Evidence of interaction network evolution by whole-genome duplications: a case study in mads-box proteins. *Molecular Biology and Evolution*, 24(3):670–678. 15

Visser, A. E. and Fell, D. A. (2007). Systems biology meets chromatin function. workshop on nuclear organization. *EMBO Reports*, 8(5):446–450. 126

Vogel, C., Teichmann, S. A., and Pereira-Leal, J. (2005). The relationship between domain duplication and recombination. *Journal of Molecular Biology*, 346(1):355–365. 15

von Dassow, G., Meir, E., Munro, E. M., and Odell, G. M. (2000). The segment polarity network is a robust developmental module. *Nature*, 406(6792):188–192. 9

Wachsmuth, M., W., W., and Langowski, J. (2000). Anomalous diffusion of fluorescent probes inside living cell nuclei investigated by spatially-resolved fluorescence correlation spectroscopy. *Journal of Molecular Biology*, 298:677–689. 148

Wagner, A. (2003). How the global structure of protein interaction networks evolves. *Proceedings. Biological sciences / The Royal Society*, 270(1514):457–466. 10, 12, 14, 15

Wagner, A. (2008). Neutralism and selectionism: a network-based reconciliation. *Nature Reviews Genetics*, 9:965–974. 192

Wahl, S. A., Haunschild, M. D., Oldiges, M., and Wiechert, W. (2006). Unravelling the regulatory structure of biochemical networks using stimulus response experiments and large-scale model selection. *Systems Biology*, 153(4):275–285. 26

Wang, Z. and Zhang, J. (2007). In search of the biological significance of modular structures in protein networks. *PLoS Computational Biology*, 3(6):e107. 13

Ward, J. J. and Thornton, J. M. (2007). Evolutionary models for formation of network motifs and modularity in the saccharomyces transcription factor network. *PLoS Computational Biology*, 3(10):1993–2002. 12

Weicker, K. (2002). *Evolutionäre Algorithmen*. Teubner, Stuttgart. 17, 19

Weidtkamp-Peters, S., Lenser, T., Negorev, D., Gerstner, N., Hofmann, T. G., Schwanitz, G., Hoischen, C., Maul, G., Dittrich, P., and Hemmerich, P. (2008). Dynamics of component exchange at PML nuclear bodies. *Journal of Cell Science*, 121:2731–2743. 119, 128, 141, 146, 152, 154, 158

Weidtkamp-Peters, S., Weisshart, K., Schmiedeberg, L., and Hemmerich, P. (2009). Fluorescence correlation spectroscopy to assess the mobility of nuclear proteins. *Methods in Molecular Biology*, 464:321–341. 128, 148

Weise, T., Zapf, M., Chiong, R., and Nebro Urbaneja, A. J. (2009). Why is optimization difficult? In Chiong, R., editor, *Nature-Inspired Algorithms for Optimisation*, volume 193 of *Studies in Computational Intelligence*, chapter 1, pages 1–50. Springer. 38

Whipple, C. and Schmidt, R. (2006). *Genetics of Grass Flower Development*, chapter 10, pages 385–42. Advances in Botanical Research. Academic Press. 169

Wiesmann, D. (2002). From syntactical to semantical mutation operators for structure optimization. In *PPSN VII: Proceedings of the 7th International Conference on Parallel Problem Solving from Nature*, pages 234–246, London, UK. Springer-Verlag. 43

Wiesmeijer, K., Molenaar, C., Bekeer, I., Tanke, H., and Dirks, R. (2002). Mobile foci of sp100 do not contain pml: Pml bodies are immobile but pml and sp100 proteins are not. *Journal of Structural Biology*, 140:180–188. 157

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1:80–83. 60

Winter, K. U., Saedler, H., and Theissen, G. (2002a). On the origin of class b floral homeotic genes: functional substitution and dominant inhibition in arabidopsis by expression of an orthologue from the gymnosperm gnetum. *Plant Journal*, 31:457–75. 166

Winter, K. U., Weiser, C., Kaufmann, K., Bohne, A., Kirchner, C., Kanno, A., Saedler, H., and Theissen, G. (2002b). Evolution of class b floral homeotic proteins: obligate heterodimerization originated from homodimerization. *Molecular Biology and Evolution*, 19:587–96. 15, 169, 170, 189

Wiseman, P. W., Squier, J. A., Ellisman, M. H., and Wilson, K. R. (2000). Two-photon image correlation spectroscopy and image cross-correlation spectroscopy. *Journal of Microscopy*, 200(Pt 1):14–25. 126

Wolf, D. M. and Arkin, A. P. (2003). Motifs, modules and games in bacteria. *Current Opinion in Microbiology*, 6:125–134. 165

Wolpert, D. H. and Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1:67. 35

Wu, L. (2007). Role of the BLM helicase in replication fork management. *DNA Repair*, 6:936–944. 159

Wuchty, S., Oltvai, Z. N., and Barabási, A.-L. (2003). Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nature Genetics*, 35(2):176–179. 14

Yang, Y., Fanning, L., and Jack, T. (2003). The k domain mediates heterodimerization of the arabidopsis floral organ identity proteins, APETALA3 and PISTILLATA. *Plant Journal*, 33(1):47–59. 187, 188

Yao, X. (1997). The importance of maintaining behavioral link between parents and offspring. In *Proc. 1997 IEEE Int. Conf. Evolutionary Computation (ICEC'97), Indianapolis, IN*, pages 629–633. 47, 48

Yao, X. (1999). Evolving artificial neural networks. *Proceedings of the IEEE*, 87(9):1423–1447. 31, 46

Yildirim, N. and Mackey, M. C. (2003). Feedback regulation in the lactose operon: a mathematical modeling study and comparison with experimental data. *Biophysical Journal*, 84(5):2841–51. 60

Zahn, L. M., Leebens-Mack, J., Depamphilis, C. W., Ma, H., and Theissen, G. (2005). To b or not to b a flower: The role of deficiens and globosa orthologs in the evolution of the angiosperms. *Journal of Heredity*, 96:225–240. 166

Zhao, L., Kim, Y., Dinh, T. T., and Chen, X. (2007). miR172 regulates stem cell fate and defines the inner boundary of APETALA3 and PISTILLATA expression domain in arabidopsis floral meristems. *Plant Journal*, 51(5):840–849. 185

Zhong, S., Salomoni, P., and Pandolfi, P. P. (2000). The transcriptional role of pml and the nuclear body. *Nature Cell Biology*, 2(5):E85–90. 123

Zhu, D. and Qin, Z. S. (2005). Structural comparison of metabolic networks in selected single cell organisms. *BMC Bioinformatics*, 6:8. 7

Ziegler, J. and Banzhaf, W. (2001). Evolving control metabolisms for a robot. *Artificial Life*, 7:171–190. 29

A
Curriculum vitae

Personal Data

Name	Thorsten Lenser
Address	Hausbergstraße 17, 07749 Jena, Germany
Date of Birth	11th January 1982, Oberhausen, NRW, Germany
Family Status	Married, two daughters
Telephone	+49 3641 470479 (home) | +49 175 3550757 (mobile)
Email	thorsten.lenser@googlemail.com

Education

2005 - 2010	PhD studies in the Bio Systems Analysis group of PD Dr. Peter Dittrich at the FSU Jena. Participation in EU-funded ESIGNET research project.
2005	Master's thesis at Australian Antarctic Division in Hobart, Tasmania. Title: A nonparametric algorithm to model movement between polygon subdomains in a spatially explicit ecosystem model.
2004 - 2005	MRes in Mathematics in the Living Environment, University of York, UK. Graduation with distinction.
2001 - 2004	Study of Mathematics and Computer Science, FSU Jena, Germany. Pre-Diploma with an average mark of 1.0 (2003).
1994 - 2001	Gymnasium in Bottrop, NRW, Germany. Received the Abitur (university entrance diploma) with an average mark of 1.1.

Awards

2006	Microsoft Research travel award for the Unconventional Computation conference 2006.
2005	Prize for the best result in MRes course at York. Travel bursary, Antarctic Climate and Ecosystems Cooperative Research Centre.
2004	Scholarship of the British Chamber of Commerce in Germany.
2001	Best Abitur (university-entrance diploma) in my year.

Conferences and Workshops

Aug 2008	Artificial Life XI, Winchester, UK (talk)
Mar 2008	European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics (EvoBIO), Napoli, Italy
Oct 2007	European conference on complex systems, ECCS 2007, Dresden, Germany (poster presentation)
May 2007	Workshop of the Jena Centre for Bioinformatics, Jena, Germany (talk)
Jan 2007	BioSysBio 2007, Manchester, UK (poster presentation)
Sep 2006	UC '06 - 5th International Conference on Unconventional Computation, York, UK (poster presentation)
Jul 2006	German workshop on artificial life, GWAL 7, Jena, Germany
Jul 2006	Workshop on membrane computing, Leiden, The Netherlands
2005-2008	Consortium meetings of the ESIGNET project, Jena, Eindhoven, Dublin, Birmingham.

Student supervisions and teaching

Diploma thesis Christian Beck, Hendrik Rohn, Stephan Richter (Bioinformatics, FSU Jena)

MSc thesis Co-supervision of Margriet Palm (Biomedical Engineering, TU Eindhoven)

Student assistants Supervision of a number of small projects carried out by student assistants in our group.

Teaching Seminars in Evolutionary Algoriths, Bio Systems Analysis, and Probability Theory.

B

List of Publications

Primary Publications

Thorsten Lenser, Klaus Weisshart, Tobias Ulbricht, Karolin Klement, and Peter Hemmerich. Fluorescence fluctuation microscopy to reveal 3D-architecture and function in the cell nucleus. In *Methods in Cell Biology*. Academic Press, 2010. invited.

Peter Brand, Thorsten Lenser, and Peter Hemmerich. Assembly dynamics of PML nuclear bodies in living cell. *PMC Biophysics*, 3:3, 2010. 119, 146

Thorsten Lenser, Günter Theißen, and Peter Dittrich. Developmental robustness by obligate interaction of class B floral homeotic genes and proteins. *PLoS Computational Biology*, 5(1): e1000264, 2009. 163

Stefanie Weidtkamp-Peters, Thorsten Lenser, Dmitri Negorev, Norman Gerstner, Thomas G. Hofmann, Georg Schwanitz, Christian Hoischen, Gerd Maul, Peter Dittrich, and Peter Hemmerich. Dynamics of component exchange at PML nuclear bodies. *Journal of Cell Science*, 121:2731–2743, 2008. 119, 128, 141, 146, 152, 154, 158

Thorsten Lenser, Naoki Matsumaru, Thomas Hinze, and Peter Dittrich. Tracking the evolution of chemical computing networks. In S. Bullock, J. Noble, R. Watson, and M. A. Bedau, editors, *Artificial Life XI: Proceedings Eleventh International Conference on the Simulation and Synthesis of Living Systems*, pages 343–350. MIT Press, Cambridge, MA, 2008. 11, 85

Thorsten Lenser and Andrew Constable. A nonparametric algorithm to model movement between polygon subdomains in a spatially explicit ecosystem model. *Ecological Modelling*, 206(1-2):79–92, 2007.

Thorsten Lenser, Thomas Hinze, Bashar Ibrahim, and Peter Dittrich. Towards evolutionary network reconstruction tools for systems biology. In E. Marchiori, J.H. Moore, and J.C. Rajapakse, editors, *Proceedings of the Fifth European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics (EvoBIO)*, volume 4447 of *LNCS*, 2007. 60, 69

Thorsten Lenser, Thomas Hinze, and Peter Dittrich. Evolving biological networks. Poster presentation, UC 2006, University of York, UK, 2006.

Full List of Publications

Thorsten Lenser, Klaus Weisshart, Tobias Ulbricht, Karolin Klement, and Peter Hemmerich. Fluorescence fluctuation microscopy to reveal 3D-architecture and function in the cell nucleus. In *Methods in Cell Biology*. Academic Press, 2010. invited.

Peter Brand, Thorsten Lenser, and Peter Hemmerich. Assembly dynamics of PML nuclear bodies in living cell. *PMC Biophysics*, 3:3, 2010. 119, 146

Thomas Hinze, Thorsten Lenser, Gabi Escuela, Ines Heiland, and Stefan Schuster. Modelling signalling networks with incomplete information about protein activation states: A p system framework of the KaiABC oscillator. In G. Paun, M.J. Perez-Jimenez, A. Riscos-Nunez, and A. Salomaa, editors, *Membrane Computing. Proceedings Tenth International Workshop on Membrane Computing (WMC10). Revised, Selected and Invited Papers*, volume 5957 of *LNCS*, pages 316–335. Springer, 2010.

Thorsten Lenser, Günter Theißen, and Peter Dittrich. Developmental robustness by obligate interaction of class B floral homeotic genes and proteins. *PLoS Computational Biology*, 5(1): e1000264, 2009. 163

Thomas Hinze, Raffael Fassler, Thorsten Lenser, and Peter Dittrich. Register machine computations on binary numbers by oscillating and catalytic chemical reactions modelled using mass-action kinetics. *International Journal of Foundations of Computer Science*, 20(3):411–426, 2009.

Stefanie Weidtkamp-Peters, Thorsten Lenser, Dmitri Negorev, Norman Gerstner, Thomas G. Hofmann, Georg Schwanitz, Christian Hoischen, Gerd Maul, Peter Dittrich, and Peter Hemmerich. Dynamics of component exchange at PML nuclear bodies. *Journal of Cell Science*, 121:2731–2743, 2008. 119, 128, 141, 146, 152, 154, 158

Thomas Hinze, Raffael Fassler, Thorsten Lenser, Naoki Matsumaru, and Peter Dittrich. Event-driven metamorphoses of P systems. In D. Corne, P. Frisco, G. Paun, G. Rozenberg, and A. Salomaa, editors, *Membrane Computing. Proceedings Ninth International Workshop on Membrane Computing (WMC9)*, volume 5391 of *LNCS*, pages 209–225. Springer Verlag, 2008a.

Thorsten Lenser, Naoki Matsumaru, Thomas Hinze, and Peter Dittrich. Tracking the evolution of chemical computing networks. In S. Bullock, J. Noble, R. Watson, and M. A. Bedau, editors, *Artificial Life XI: Proceedings Eleventh International Conference on the Simulation and Synthesis of Living Systems*, pages 343–350. MIT Press, Cambridge, MA, 2008. 11, 85

Naoki Matsumaru, Thorsten Lenser, Florian Centler, Pietro Speroni di Fenizio, Thomas Hinze, and Peter Dittrich. Common organizational structures within two chemical flip-flops. In Y. Suzuki, A. Adamatzky, M. Hagiya, and H. Umeo, editors, *Proceedings Third International Workshop on Natural Computing (IWNC2008)*, pages 50–59. Japan Society for Artificial Intelligence, 2008.

Hendrik Rohn, Bashar Ibrahim, Thorsten Lenser, Thomas Hinze, and Peter Dittrich. Enhancing parameter estimation of biochemical networks by exponentially scaled search steps. In E. Marchiori and J. H. Moore, editors, *Proceedings Fifth European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics (EvoBIO)*, volume 4973 of *LNCS*, pages 177–187. Springer Berlin / Heidelberg, 2008. 57, 58

Raffael Fassler, Thomas Hinze, Thorsten Lenser, and Peter Dittrich. Construction of oscillating chemical register machines on binary numbers using mass-action kinetics. In O.H. Ibarra and P. Sosik, editors, *Proceedings Prague International Workshop on Membrane Computing in conjunction with DNA14 (PIWMC2008)*, pages 11–22. Silesian University Press, 2008. ISBN 978-80-7248-468-3.

Thomas Hinze, Sikander Hayat, Thorsten Lenser, Naoki Matsumaru, and Peter Dittrich. Biosignal-based computing by AHL induced synthetic gene regulatory networks. In P. Encarnacao and A. Veloso, editors, *Proceedings of the First International Conference on Bio-Inspired Systems and Signal Processing (BIOSIGNALS2008)*, pages 162–169. IEEE Engineering in Medicine and Biology Society, Institute for Systems and Technologies of Information Control and Communication, INSTICC press, 2008b. ISBN 978-989-8111-18-0.

Thomas Hinze, Sikander Hayat, Thorsten Lenser, Naoki Matsumaru, and Peter Dittrich. Hill kinetics meets P systems. In G. Eleftherakis, P. Kefalas, G. Paun, G. Rozenberg, and A. Salomaa, editors, *Membrane Computing*, volume 4860 of *LNCS*, pages 320–335. Springer Verlag, 2008c.

Thorsten Lenser and Andrew Constable. A nonparametric algorithm to model movement between polygon subdomains in a spatially explicit ecosystem model. *Ecological Modelling*, 206(1-2):79–92, 2007.

Thorsten Lenser, Thomas Hinze, Bashar Ibrahim, and Peter Dittrich. Towards evolutionary network reconstruction tools for systems biology. In E. Marchiori, J.H. Moore, and J.C. Rajapakse, editors, *Proceedings of the Fifth European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics (EvoBIO)*, volume 4447 of *LNCS*, 2007. 60, 69

Naoki Matsumaru, Thorsten Lenser, Thomas Hinze, and Peter Dittrich. Toward organization-oriented chemical programming: A case study with the maximal independent set problem. In F. Dressler and I. Carreras, editors, *Advances in Biologically Inspired Information Systems*, volume 69 of *Studies in Computational Intelligence*, pages 147–163. Springer, Berlin, 2007.

Thomas Hinze, Raffael Fassler, Thorsten Lenser, Naoki Matsumaru, and Peter Dittrich. Efficient chemical computing using deterministic reaction systems with prioritisation of rules. In M. Droste and M. Lohrey, editors, *Proceedings 17. Theorietag Automaten und Formale Sprachen*, pages 68–73. University of Leipzig, 2007.

Thomas Hinze, Thorsten Lenser, and Peter Dittrich. A protein substructure based P system for description and analysis of cell signalling networks. In H.J. Hoogeboom, G. Paun, and G. Rozenberg, editors, *Proceedings Seventh Workshop on Membrane Computing (WMC7), Leiden*, volume 4361 of *LNCS*, pages 409–423. Springer, Berlin, 2006a.

Thomas Hinze, Thorsten Lenser, and Peter Dittrich. Zellsignalnetzwerke - biologische Computer mit universeller Berechnungsstärke. In R. Freund and M. Oswald, editors, *Proceedings 16. Theorietag Automaten und Formale Sprachen*, pages 73–77. TU Wien, 2006b.

Poster Presentations

Benedikt Schau, Thomas Hinze, Thorsten Lenser, Ines Heiland, and Stefan Schuster. Control system-based reverse engineering of circadian oscillators. Poster presentation, GCB2009, Martin-Luther University Halle-Wittenberg, 2009.

Thomas Hinze, Thorsten Lenser, Ines Heiland, and Stefan Schuster. Backtracking membrane systems unravel stable oscillations in distributed reaction networks. Poster presentation, BioSysBio 2009, University of Cambridge, UK, 2009.

Peter Dittrich, Bashar Ibrahim, Thomas Hinze, Thorsten Lenser, and Naoki Matsumaru. Hierarchically evolvable components for complex systems: Biologically inspired algorithmic design. Poster presentation, ECCS2007, Dresden University of Technology, 2007.

Naoki Matsumaru, Thorsten Lenser, Thomas Hinze, and Peter Dittrich. Designing a chemical program using chemical organization theory. Poster presentation, BioSysBio 2007, Manchester, UK (Runner-up for best poster award), BMC Systems Biology, 1(Suppl 1):P26, 2007.

Thorsten Lenser, Thomas Hinze, and Peter Dittrich. Evolving biological networks. Poster presentation, UC 2006, University of York, UK, 2006.

James Decraene, Peter Dittrich, Thomas Hinze, Thorsten Lenser, Barry McMullin, and George Mitchell. A multidisciplinary survey of modeling techniques for biochemical networks. Poster presentation, IPG 2006, INSA Lyon, France, 2006.

C
Ehrenwörtliche Erklärung

Hiermit erkläre ich,

- dass mir die Promotionsordnung der Fakultät bekannt ist,
- dass ich die Dissertation selbst angefertigt habe, keine Textabschnitte oder Ergebnisse eines Dritten oder eigenen Prüfungsarbeiten ohne Kennzeichnung übernommen und alle von mir benutzten Hilfsmittel, persönliche Mitteilungen und Quellen in meiner Arbeit angegeben habe,
- dass ich die Hilfe eines Promotionsberaters nicht in Anspruch genommen habe und dass Dritte weder unmittelbar noch mittelbar geldwerte Leistungen von mir für Arbeiten erhalten haben, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen,
- dass ich die Dissertation noch nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht habe.

Bei der Auswahl und Auswertung des Materials sowie bei der Herstellung des Manuskripts haben mich folgende Personen unterstützt: PD Dr. Peter Dittrich, Dr. Thomas Hinze und PD Dr. Peter Hemmerich.

Ich habe die gleiche, eine in wesentlichen Teilen ähnliche bzw. eine andere Abhandlung nicht bei einer anderen Hochschule als Dissertation eingereicht.

Jena, den 24.3.2010