Self-Regulating Gene Switches and Molecular Evolution

DISSERTATION
zur Erlangung des akademischen Grades
Doctor rerum naturalium

Vorgelegt der
Fakultät für Naturwissenschaften und Mathematik
der Universität Wien

von
Mag. Stefanie Widder

im Dezember 2003

Nach drei Stunden Fragen, theoretischen Erörterungen und konkreten Erkundigungen [...] kehrte er zurück in sein Büro, niedergedrückt von der Gewissheit, keinerlei Lösung für so viele Probleme gefunden zu haben, sondern vielmehr neue und vielfältige Probleme für keine Lösung.
Gabriel Garcia Marquez, "Die Liebe in den Zeiten der Cholera."

Danksagung

Zuallererst möchte ich gerne den Stefan umarmen.
Ich möchte meiner Mama und meinem Papa danken, daß sie mir als Referenz für Wichtiges und Unwichtiges dienen, jeder auf seine Art und mit seinem jeweiligen menschlichen und beruflichen Hintergrund.
Außerdem natürlich der Melanie für hervorragenden Realitätsausgleich.
Ich habe Euch sehr lieb.

Peter Schuster danke ich als meiner persönlichen wissenschaftlichen Authorität.
Bei Christoph und Ivo möchte ich mich für ihre kompetente Unterstützung bedanken.
Mein Dank gilt auch Andreas Novak, der in einem brenzligen Moment mit mathematischer Hilfestellung zugegen war.
Weiters möchte ich mich bei Berni, Svrci und Bärbel Krackhofer für Inspiration, sowie bei Stefan, Christina und Roman, Camille, Ulli und Uli, MTW, Caro, Sonja und Claudia für Unterhaltung und die Schaffung eines so angenehmen Klimas am Institut bedanken.
Ohne Euch allen wäre diese Dissertation nicht möglich gewesen.

Abstract

Networks are common themes of molecular evolution. Interacting circuitries of regulatory genes manage the smooth execution of organisms' functions at very high complexity levels. The understanding of such networks relies on the analysis of simpler sub-modules or building blocks. In the framework of the study presented, a model for an auto-regulatory, minimal gene switch was developed. The model consists of two genes either activating or silencing each other. A focus on a realistic implementation was set using as molecular players genes with specific enhancer and promoter regions, transcripts and proteins, whereby both sequential as well as structural properties of the individuals were taken into account. The interaction of the transcription factors with the promoter is based on a cooperative, structural change of the DNA. The dynamic behavior of the gene network, caused by the underlying reaction rates, was analyzed and depending on the parameter set, global stability or bistability could be assessed.

In the fine tradition of *in silico* flow reactors, we developed a software termed 'RegNet', capable to accept genetic networks as individuals for the reactor population. For the simulations carried out, we used the analyzed gene switch as predefined unit constituting the evolving population. The individual fitness consists of a multi-objective measure involving multi-layered molecular properties of the network species responsible for the dynamic behavior of the specific individual, and the dynamics itself.

We simulated evolutionary adaptation of the reactor population under artificially enhanced selective pressure, thereby accentuating selection and reducing genetic drift. This setting is particularly well suited to resemble populations adapting to e.g. rapidly changing environmental conditions, strongly compromising parasites or noise-like genetic perturbations such as bursts of transposons. Two different setups using noise to perturb the reaction rates of transcription and translation were investigated: (i) Exposure to noise enhancing the chance for bistable dynamics and (ii) the evolution of noise tolerance. The populations of both types of simulation develop a specific strategy to deal with the stochastic fluctuations. In the former setting we observe a structured broadening of the genetic variability of the population due to noise-dependent effects. The ability for quick adaptation decreases with the degree of genetic uniformity. Thresholds for the pivotal point of fertile and devastating noise effects are observed.

In the latter setup, the population 'invents' co-evolution to succeed in the

optimization. Two subpopulations relevant for either their fitness or the providing of alternative starting points in solution space, co-evolve towards noise tolerance. Again, the genetic variability of the reactor population plays a major role. In both setups stochastic perturbations principally act at the level of the population. Only at enhanced noise intensities influences on the individual can be observed. Based on these data, we conclude a principal function of stochastic mechanisms for the generation of genetic variability. The model represents a base for further investigation of genetic networks and their evolvability.

Zusammenfassung

Netzwerke sind im Gebiet der molekularen Evolution weit verbreitet. Sich gegenseitig beeinflussende, komplexe Verschaltungen aus regulatorischen Genen garantieren z.B. den reibungslosen Ablauf von Funktionen des Organismus. Das Verstehen und die Interpretation solcher komplizierter Netze beruht jedoch auf der Analyse von einfacheren Submodulen. Im Rahmen der hier vorgelegten Arbeit, wurde dementsprechend das Modell eines selbst-regulierenden Genschalters entwickelt. Das Modell besteht aus zweien, sich gegenseitig entweder induzierenden oder deaktivierenden Genen, wobei ein spezielles Augenmerk auf Wirklichkeitsnähe gelegt wurde. Zu diesem Zweck gehen sowohl Gene mit spezifischen Promotor- und Enhancer-Strukturen, sowie Transkripte und Proteine mit sequenz- und strukturbedingten Eigenschaften in das Modell ein. Das Zusammenspiel der Transkriptionsfaktoren mit dem Promotor beinhaltet eine kooperativ motivierte, strukturelle Änderung der DNA. Das auf den zugrundeliegenden Reaktionsraten beruhende dynamische Verhalten des Gennetzwerks wurde analysiert und je nach Parameterset konnten Stabilität oder Bistabilität veranschlagt werden.

In der Tradition von *in silico* Flußreaktoren, entwickelten wir eine Software namens 'RegNet', welche die beschriebenen genetischen Netze als Individuen der Reaktorpopulation verwendet. Die jeweilige Fitness der Individuen, setzt sich aus ihren molekularen Eigenschaften, die letztendlich für die Dynamik verantwortlich sind, sowie aus deren Dynamik selber zusammen.
Wir führten Experimente zur evolutionären Anpassung von Populationen unter künstlich verstärktem selektivem Druck durch, wobei der Mechanismus der Selektion gegenüber dem genetischer Drift stark bevorzugt wurde. Dieser Versuchsansatz eignet sich besonders zur Beschreibung von evolvierenden Populationen in sich rasch verändernder Umwelt, aber auch unter dem Einfluß von schädigenden Parasiten oder von stochastischen Mechanismen zur Erhaltung genetischer Vielfalt.
Zwei prinzipiell verschiedene Ansätze wurden untersucht, beide unter Einsatz von Rauschen in den Reaktionsraten von Transkription und Translation: (i) die Steigerung der Chance Bistabilität bei einem Netzwerk zu finden durch Applikation von Rauschen und (ii) die Evolution von Toleranz gegen stochastische Störungen. Spezielle Strategien für die Bewältigung von Störungen wurden von beiden evolvierenden Populationen entwickelt. Beim ersten Ansatz konnte eine strukturierte Verbreiterung der genetischen

Variabilität auf Grund von störungsabhängigen Effekten festgestellt werden. Die Fähigkeit zur schnellen Adaptierung der Population veringerte sich mit verstärkter genetischer Gleichheit. Weiters konnten Grenzen zwischen förderndem und störendem Einfluß von Rauschen bestimmt werden.
Im zweiten Ansatz "erfand" die Population Co-Evolution um mit der schwierigen Aufgabenstellung fertig zu werden. Es entstanden zwei Untergruppen, die jeweils entweder wegen ihrer Fitness oder weil sie neue Ansatzpunkte für die Suche im Lösungsraum boten, relevant für die Gesamtbevölkerung waren. Wieder spielte hierbei der Grad der genetischen Verschiedenheit eine große Rolle.
Bei beiden Ansätzen wirkte die stochastische Störung zuerst auf Populationsebene und erst bei entsprechender Intensivierung auf der Ebene des Individuums. Aus den Daten schließen wir eine grundlegende Funktion von stochastischen Mechanismen bei der Erhaltung von genetischer Vielfalt.
Das vorgelegte Modell bietet eine ausgezeichneten Grundlage für weitere Versuche im Bereich von Gennetzwerken und deren Evolvierbarkeit.

Contents

1 Introduction

1.1 Of Flowers and Bees

Complex lifeforms, such as bacteria, plants and animals are built from few classes of molecular species. Among these the bio-polymers are the most prominent. They form the ultimate basis of inheritance and determine traits. According to current knowledge bio-polymers are involved in any aspect of present-day life, from molecular to macroscopic biology.

The century of molecular genetics started with the discovery of DNA structure, first published by James Watson and Francis Crick with the pivotal support of Rosalind Franklin and Maurice Wilkins. On this basis Crick [16] formulated the central dogma of molecular biology describing the bio-polymer based information flux in living systems: Originally, the flow of genetic information was viewed as essentially unidirectional. DNA as information storage molecule is being copied to both itself and to RNA. RNA is subsequently decoded during protein synthesis. Proteins were assumed to be the only biomolecules with catalytic properties.

Within the last decades, this view has given way to a more detailed understanding due to several important discoveries. Especially the misestimation of the role of RNA could be revised (see also figure 1): (i) The study of RNA viruses showed that RNA, like DNA, can serve as a primary information-encoding molecule, i.e. the genome.(ii) Various types of RNA molecules with catalytic properties were found, the first ones independently by the groups of Cech and Altman in the 1980s [13, 31, 32]. Recently, DNA molecules exhibiting catalytic behavior, deoxyribozymes, have also been designed and synthesized [11]. Last but not least, aptamers, RNA molecules able to specifically bind to substrate molecules have been evolved by *in vitro* selection [21, 56].

These discoveries helped to formulate a greater general picture of the interplay between nucleic acids and proteins, as can be seen in figure 2. The sketch displayed there, represents the information flux between the molecules within the cell. With impressing simplicity, this model manages to represent the complex processes of (i) information decoding and (ii) catalysis and control. One has to note though, that exactly this simplicity also provides the model's limitations. This is because the picture misses out a description of the multi-layered structure of exactly the featured interactions. There exist multiple forms of interplay between and among the molecular species, creat-

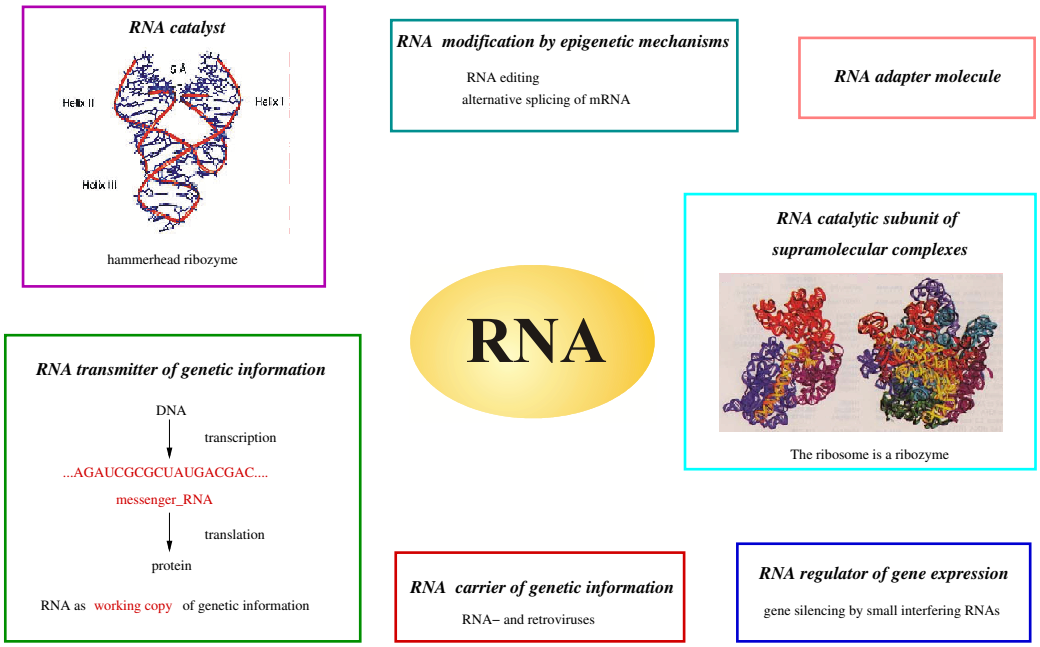


Figure 1: RNA is central to cellular processes: Multiple roles of RNA in information storage, information transmission, catalytic processes and in gene expression are sketched.

ing network-like circuitries with usage and re-usage of molecules and motifs and sophisticated regulatory processes to optimize the resources. These networks underlying the simple scheme are not even touched, let alone explained or indicated in their approximate complexity.

Impressive progress in genetic engineering techniques and the availability of a vast amount of reliable sequence data, expanded the general view once more, and allowed for the following reasoning to be widely accepted: The relatively small number of genes in organisms does not suffice to account for the unconceivable multitude of different reactions, implemented with biomolecules, if it was in a linear fashion. On the contrary, the different gene products are used and reused in different reactions and pathways by that building networks of mutual dependencies. Indeed, numerous examples have been found in *in vivo* systems ranging from simple feedback/feed-forward dependencies of molecular players as in the bacterial lactose operon [61, 75, 82], to enormous interaction networks, e.g. the segment polarity module in *Drosophila* [79]. The net-like architecture is especially

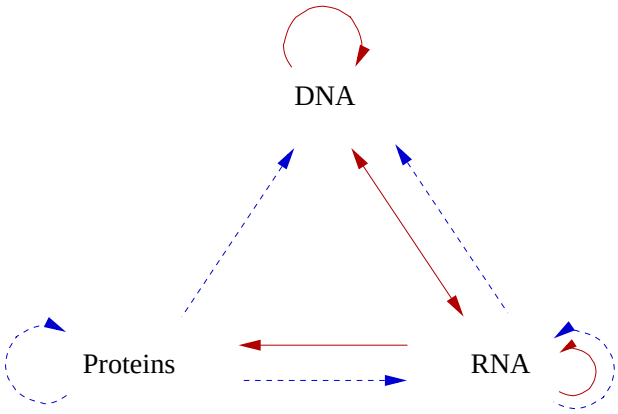


Figure 2: Central dogma with additions: The scheme shows the bio-polymers and their interactions. Red connections represent the information flow, blue dashed lines show the regulatory and catalytic flow. DNA ↔ RNA: transcription and reverse transcription; RNA → proteins: translation; DNA → DNA: replication.

prominent in processes such as metabolism, signalling pathways and gene expression regulation.

Eventually, even the picture of heredity based exclusively on the transduction of the DNA sequence had to be enlarged. The study of heritable changes in gene function that occur without changes in the DNA sequence namely epigenetics, paraphrases the role of proteins and RNA in biological systems. Epigenetic mechanisms such as DNA methylation, histone acetylation and RNA interference, and their effects in gene activation and silencing, are increasingly understood to be more than solely 'bit players' in phenotype transmission and development.

The central dogma has blurred and the genome can no longer be seen as a strict, vertically operating system, but must be understood as fluid and highly networking. A new standard in the progress of the life-sciences has been reached.
With this work we make an attempt to account for this dynamic view.

1.2 Organization of the Thesis

The study presented here features the complex dependencies of genetic regulatory networks.
Genetic regulatory networks describe the net-like interconnection of genes and their products, thereby constituting a level of gene expression regulation. They are highly sensitive to changes in the concentrations of their mediator molecules and very effective in the execution of their tasks. Currently, various studies are dedicated to the determination of their interaction patterns and building blocks.

In the field of technology, networks are frequently build of switches, capable to change between 'YES' and 'NO' states. Solely the combination of these simple components allows to form highly complex interaction patterns and the mastering of difficult tasks. Analogously in biology, it has been proposed that gene regulatory nets with virtually any desired property can be constructed from networks of simple regulatory elements, [26, 49]. The search, however, for recurrent, basic motifs *in vivo*, has just begun [6, 14, 70]. In this study we will contribute to this quest with the detailed investigation of such an elementary module. We describe the construction of a genetic switch model and the simulation of its molecular adaptation in a population of its kind.

In the introductory chapter 1 we shortly describe the contextual theory behind our study. The sections include background about nucleic acids and proteins, a short review on artificial design and modeling of biological networks and on the topic molecular evolution. All of these fields are vital for the formulation and the simulation of evolutionary development of our model switch. In chapter 2 we will be engaged in the construction of the genetic switch network and the mathematical analysis of its dynamics. Such 'rational network design' may lead both to the engineering of new cellular behaviors and to an improved understanding of naturally occurring networks [18]. In chapter 3 the developed gene switch will be implemented in a self-designed software, which simulates evolutionary network formation. The program is inspired by the *in silico* flow reactor elaborated by Schuster and Fontana in the 1980s, [23]. By means of the program, steps in evolution of artificial gene circuits can be visualized, network formation and extension can be modeled. Relevant theoretical bits and pieces are given during the progression of the sections. The numerical results are described in the following chapter 4. They cover the multi-objective calibration of the fitness criteria, a detailed

analysis of the dynamic behavior of the population during the simulation experiments and the astonishing effects of noise, when applied to the evolving population.

With our approach we make a contribution to the exploration of genetic networks and their common underlying principles and to the understanding of their dynamics and evolution.

1.3 Nucleic Acids

1.3.1 Genetic Information

Genetic information is transmitted from cell to cell and generation to generation. This process is understood as inheritance or 'vertical' transfer of genetic information. The storage of information is provided by DNA, a heteropolymer formed of monomeric building blocks, the nucleotides. Each nucleotide consists of three molecular components: A cyclic sugar core (desoxyribose), a phosphate, forming the backbone and a heterocyclic base. In DNA we find four classes of bases, adenine (**A**), guanine (**G**), cytosine (**C**) and thymine (**T**). The actual encoding of the information is performed through a one-dimensional progression of digits chosen from this four-letter alphabet of the nucleotides. The obtained string constitutes the primary structure of DNA, the sequence $s = a_1 a_2 ... a_n$ with $a_i \in \{\text{A,T,G,C}\}$.

In the eukaryotic cell, DNA molecules form chromosomes, highly organized structures complexed with shape-inducing proteins called histones. The genes, structured substreches of the chromosome, provide exact building plans for proteins which are to be processed at the ribosome. The linkage between the DNA sequence and the amino acid sequence as obtained by translation is provided by two classes of RNA molecules, messenger and transfer RNA.

RNA is built from the same constituents as is DNA, with two exceptions: the sugar units contain 2' -OH groups (ribose) and one of the bases, **T**, is replaced by the functionally equivalent uracil (**U**). A detailed picture of the molecular structure of RNA can be seen in figure 3. It should be noted here that several naturally occurring modified nucleotides exist besides those containing the four conventional bases. Such modified nucleotides stabilize for example the structure of transfer-RNA.
Aside from messenger function (**mRNA**), RNA molecules conduct various other functions in the cell. These comprise the transfer of amino acids to the ribosome (**tRNA**), the processing of mRNA with so called **snRNAs** complexed to protein, furthermore RNA forms part of the ribosome (**rRNA**) and determines sites of rRNA modifications in the nucleolus (**snoRNA**). Recently, small interfering RNAs (**siRNAs**) were discovered. They are involved in the decay of specific ssRNAs, most likely in order to defend the organism against virus attacks, modulate transposons activity and eliminate aberrant transcription products.

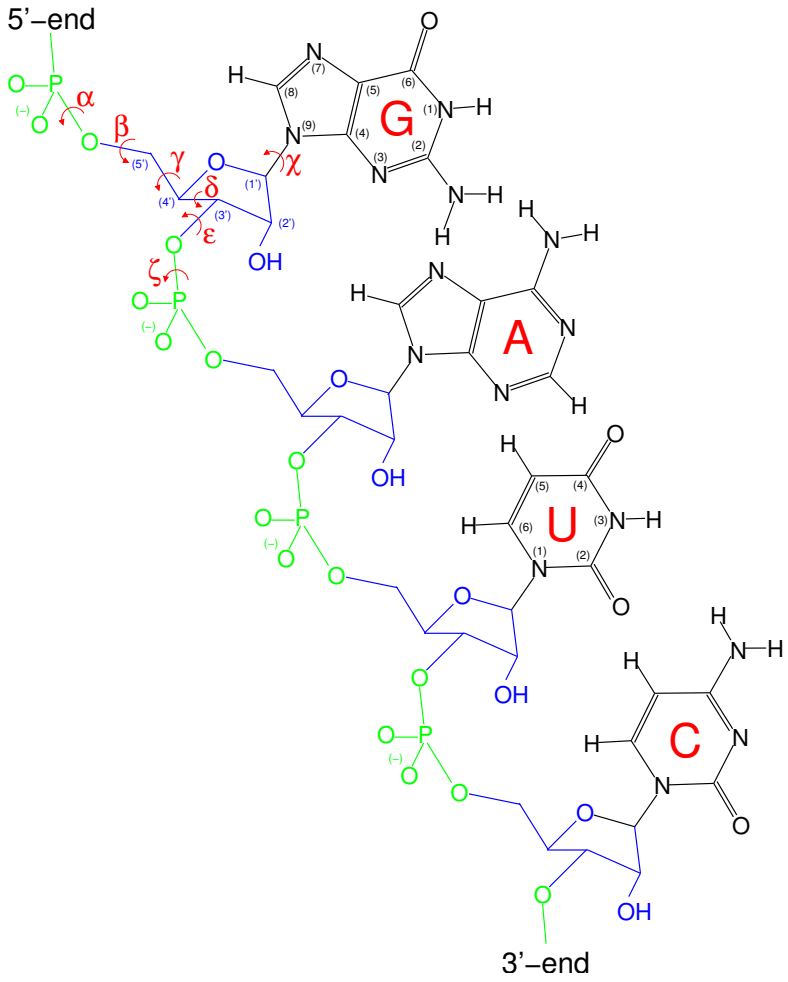


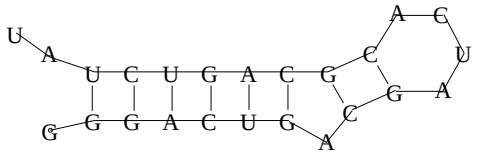
Figure 3: Atomic structure of RNA: green - phosphate, blue - ribose, black - purine (G, A) and pyrimidine (C, U) bases, the flexible torsion angles α to χ allow for tertiary interactions

1.3.2 RNA Structure

The function of biomolecules is determined by their structure. This fact led to a growing interest in the shapes of RNA molecules, see also figure 4. During

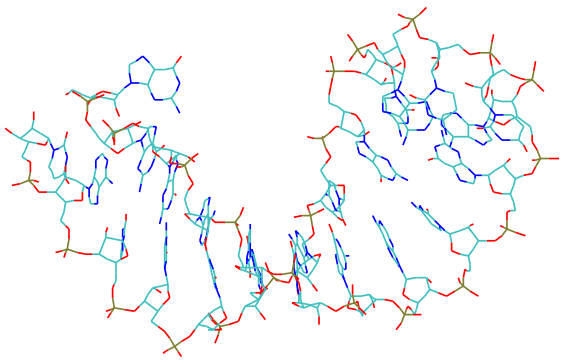
5′-GGGACUGACGAUCACGCAGUCUAU-3′

(a) Primary structure



. ((((((. ((. . . .)))))))) . .

(c) Secondary structure



(d) Tertiary structure

Figure 4: (a) Primary structure (sequence) extracted from `1tfn.pdb`, (c) mfe secondary structure calculated with RNAfold from the Vienna RNA package [37, 45, 85] in graph representation (upper) and bracket dot notation (lower), (d) tertiary structure from the NMR model with PDB identifier `1tfn`.

the formation process of a three-dimensional, thermodynamically most stable RNA structure, unfavorable interactions of hydrophobic bases with the

polar solvent are minimized. Accordingly, the molecule folds back onto itself. Furthermore, the bases form so called stacks by organizing themselves one above the other driven by Van der Waal and π interactions.

Even though a range of such driving forces for the shape finding of an RNA sequence is known, the accurate prediction of RNA tertiary structures stays from very difficult to impossible. The intent of prediction mostly collapses due to the high complexity of the folding process and the limited computational resources which still pose the major hindrances toward a reliable and general structure prediction success.

The necessary restriction of RNA structure prediction to secondary structure shapes does not mean a second quality choice, because secondary structures have features that allow for very detailed analysis. The concept of secondary structure is indeed a very useful, coarse grained representation, since it accounts for most of the free energy of folding anyway. Moreover, at this level the interaction is binary, in the sense that two bases either do or do not form a base pair. Taking advantage of this simplification, the minimum free energy (mfe) structures, can be efficiently calculated by the use of *dynamic programming* algorithms [85] and its independence of structural subunits.

The secondary structure is determined by the sequence of bases, particularly because not all combinations of bases can form base pairs that fit into a Watson-Crick helix. The hydrogen bonds, which determine the geometry of pairing, can be formed between the complementary Watson-Crick pairs $G \equiv C$ and $A = U$, as well as the less stable $G = U$ wobble pair. The secondary structure can be defined as a planar graph, see also figure 4b, with a set of vertices, the nucleotides, $V = \{1, 2, ...i, ...n\}$ and a set of edges, the linkage between nucleotides, $E = \{1, 2, ...j, ...N\}$ with an edge being $i \cdot j$, $1 \leq i < j \leq N$. Furthermore, the following assumptions have to be made:

(i)(backbone) $\forall i < n$ holds $i \cdot (i + 1) \in E$
(ii)(binary pair) for each i maximally exists one $j \neq i - 1, i + 1$ with $i \cdot j \in E$
(iii)(no pseudoknots are allowed) if $i \cdot j \in E$ and $k \cdot l \in E$ and $i < k < j$, then $i < l < j$ follows

RNA secondary structures can be built up from a small set of structural subunits called motifs shown in figure 5. The motifs are loops and external elements. The canonical loops are specified by their degree, i.e. the number of stack-terminating base pairs they contain. External elements are unpaired

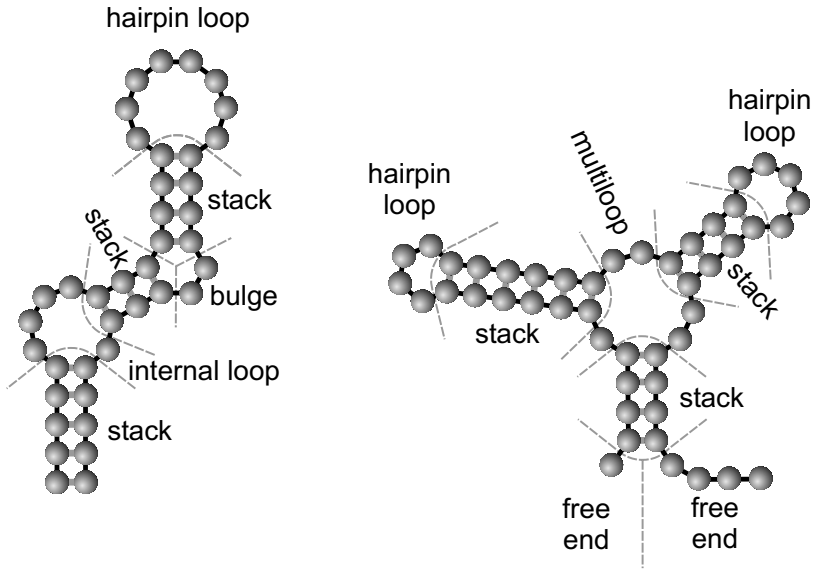


Figure 5: RNA secondary structure motifs.

regions that are not part of a loop, for example segments connecting substructures or free 5' and 3' dangling ends. According to this definition, hairpin loops are loops of degree one. Internal loops are loops of degree two and bulges can be viewed as internal loops with unpaired bases occurring only on one side of a stem. Multiloops are internal loops of degree larger than 2 and arbitrary size. A stacking pair consist of consecutive base pairs and form a loop of degree two and size 0. Based on these definitions, each nucleotide of a secondary structure can uniquely be assigned to one of these motifs, as shown in figure 5. Apart from the graph representation, there are various more ways to sketch RNA secondary structure. In this work we will mainly be using the so called 'bracket-dot-notation' as can be seen in figure 4c lower. The single nucleotides are represented by one of the three symbols '.' , '(' or ')'. A dot stands for an unpaired nucleotide, the opening bracket means that a base is paired with a downstream and a closing bracket with an upstream complementary base.
Yet another way to represent RNA secondary structure is the circular nota-

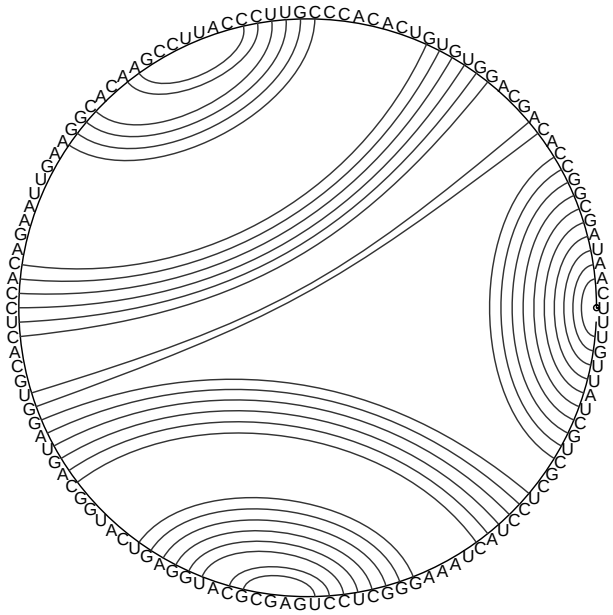


Figure 6: Circular notation of RNA secondary structure. The 5s RNA of *Methanospirillum hungatii* is depicted. Chords between bases represent basepairs. The boundaries of the substructures are clearly identified.

tion as in figure 6. It serves well to easily identify the boundaries of RNA substructures, which are defined by the first, or last, closing basepair.

Beyond the insights into structure formation and the progress in structure prediction efforts, the concept of secondary structure is also highly valuable for applications in theoretical biology. This is because primary and secondary structure provide a tractable model system to investigate the genotype-phenotype relationships, a concept of molecular evolution. Intensive investigations with this model system have provided great knowledge of the process of evolutionary adaptation [23, 24, 63, 73].

1.4 Proteins

Proteins may be visualized as the executing force in living systems which is encoded by nucleic acids. As enzymes, proteins catalyze most of the biochemical reactions in the cell with high specificity and efficiency, leading to a multitude of different products as simple as methane or as complex as alkaloids. They play the key role in light harvesting during photosynthesis by converting electromagnetic energy to a gradient in chemical potential that is used for the synthesis of key molecules in cellular metabolism. They form the cytoskeleton and represent the main constituents of hair or the silk of spiders, which have properties that are not easily obtained by other materials. This enormous versatility is achieved by heteropolymers composed of 20 different classes of monomers, the amino acids. All proteins share the same basic outlay, i.e. the single amino acids are connected by a so called peptide bond, see figure 7. The three-dimensional structure of the peptide chain is determined by its sequence. The question of how exactly the sequence of amino acids assesses the tertiary interactions leading to the spatial architecture of proteins is one of the most difficult to answer in contemporary bioscience.

The mapping from primary structures into protein folds is not invertible in the sense that many sequences form the same three-dimensional shape and thus there is substantial neutrality of sequences with respect to the structures they encode. It was estimated in 1991 that for every tertiary structure measured by NMR or X-ray crystallography, there existed 50 corresponding

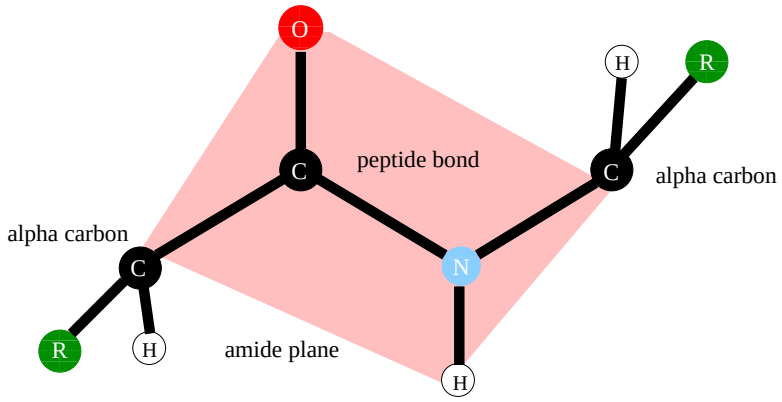


Figure 7: The peptide bond joins two amino acids to a peptide chain. The fixed angles of the bond involved atoms result in a planar structure, the amid plain. The Cα of each amino acid is shown.

primary structures [8]. Due to data obtained from the extensive genome projects in recent years, one can expect that this proportion is even more in favor of a huge degree of neutrality.
In contrast to nucleic acids, where at least the secondary structure can be predicted reliably, we find that for proteins even this task is hard to accomplish. This is mainly attributed to the more or less unspecific hydrophobic interactions involved in protein folding, lacking the all-or-none nature of Watson-Crick base pairing. Furthermore, the secondary structure of proteins is defined locally and does not contain long distance interactions in most of the cases. Therefore it is not as meaningful as RNA secondary structure.
In principle, it is assumed that the native fold of proteins corresponds to the minimum of its free energy function $W(s)$. If this function is known, any sequence s can be assigned to its fold $\psi(s)$. Alas, the number of terms contributing to this function is enormous, depending on the sequence of amino acids as well as on the natural environment, such as pH, temperature, nature of ions and more.

Various approaches to solve the prediction problem are used today. The most promising ones try to find a protein with known structure and similar sequence, followed by the alignment of the new sequence to this structure. We know that although there is an astronomic number of possible amino acid sequences, i.e. 20^n distinct sequences at chain length n, [69], the number of tertiary structures occurring in nature seems to be limited, with 90% of the native sequences sharing only 930 stable folds [29] at a total number of 4000 to 8000 folds estimated to date [2,29,83]. We also know that there are neutral networks in the sequence space [5, 39], which can lead to proteins with high structural but no conceivable sequential homology. It is obvious that trying to find a tertiary structure for a new amino acid sequence is highly dependent on what we already know. If little data is available mostly a different ansatz is used. So called 'knowledge based' protein potentials are designed from statistical mechanics information and used for prediction.

Another interesting question is the opposite one: Which sequences will fold into a given tertiary structure? The universality of proteins is shown in many different applications, e.g. catalysts in organic synthesis or cures for diseases. Exactly this striking property makes them a primary target for artificial design. To be able to tailor proteins for any use, to change, for example substrate specificity of an enzyme or the paratope of an antibody, it is preferable to know *a priori* whether a mutant protein will retain the desired structure or not, since the three-dimensional structure determines

the function of the protein. Knowledge based potentials should be able to fulfill this function, making it possible to save time and resources in protein design.

Other questions not answered to any extent until now, concern the theoretical background of proteins. How are they able to fold so quickly? How do they exactly interact with each other and with other molecules of their native environment? While there is a lot of work invested in answering these questions, they are still far from being solved. Thus, new computational methods and models to help understanding proteins are required.

1.4.1 The Z-Score Evaluation of a Protein

In the field of protein structure prediction, the approaches are divers and rare are ideal conceptions. The decisive factor for success or failure of prediction is, last but not least, dependent on the exact question one wants to clarify. While the prediction of a protein structure from scratch is practically impossible, the prediction whether a sequence can fold into a given structure or not can be carried out reliably. The latter problem gains specific importance in protein design, where the function of a protein is defined by its tertiary structure and hence, the knowledge of the phenotype of a modified sequence is essential. Naturally there is a range of possibilities to overcome the question, here we will present the z-score evaluation solely, mainly because it is used at later stages of this work.

In order to determine if the tertiary fold of a protein sequence is similar to a given structure of the same sequence length and to what extend, we introduce the quality measure z so called z-score. The z-score *per se* is defined as the energy barrier between the native fold of the molecule and the average of an ensemble of mis-folds in the units of standard deviations from the ensemble. This measure can subsequently be used to construct an energy scale by means of which the conformations and their differences between protein sequences can be compared. According to Sippl [12, 66–68], we introduce the z-score being

$$z(s,\psi) = \frac{W(s,\psi) - \overline{W}(s)}{\sigma_{W(s)}} \ , \tag{1}$$

where $W(s,\psi)$ is the energy of sequence s in the conformation ψ, $\overline{W}(s)$ is the average energy of s in all conformations ψ of a generated data set and

$\sigma_W(s)$ denotes the standard deviation. The mentioned data set consists of alternative conformations of the sequence s with length l.
Though, if the size of the data set is set to a fixed value for feasibility reasons, the number of possible mistakes scales with l of s. Hence, if $s \to \infty$ the number of decoys is proportional, too and subsequently the data set becomes insignificant. This trouble is circumvented by the construction of a so called polyprotein. The polyprotein has to be imagined as a big aggregate of linked proteins and hence forms a structural library.

Now, the sequence s of the protein to be tested is slid along the polyprotein from the N- to the C-terminus, amino acid by amino acid. For each aligned structure a z-score is calculated and counted as mis-fold to the ensemble. If n is the length of the polyprotein, $n - l$ mis-folds can be constructed. Since $n \gg l$, the number of sequence-structure pairs is of the same order of magnitude as the polyprotein length. This computational brute force attack allows to bypass the weakness of the data set size. An experimental test of the z-score using thermodynamic data could demonstrate the definite significance of the scale [84].

1.4.2 Zinc Fingers and the GAGA Factor

Complex sets of regulatory elements control the initiation of transcription in eukaryotes. Upstream of the initiation site, the binding site for the polymerase, various combinations of small sequence motifs allow a site specific recognition by corresponding DNA-binding proteins, the transcription factors. Viewed from the structure/function point of view, these proteins have two functionally different domains: the DNA-binding domain and the transactivating domain, both necessary for the transcription process. Investigations in the field of protein structure determination clarified the shape of the DNA-binding module. Results from NMR and X-ray cristallography showed that three recurring themes are responsible for DNA binding: the so called zinc fingers, leucine zippers and helix-turn-helix motifs.

The zinc finger domain, firstly described in 1985 by Klug and coworkers [62], is a small, functional, independently folding unit that requires the coordination of one or more Zn^{++} ions to stabilize its three-dimensional structure. The motif itself consists of two anti-parallel β-strands followed by an α-helix. A single Zn^{++} ion is tetrahedrally coordinated with two conserved histidine and cysteine residues each, responsible for the stabilization of the shape. In figure 8 the structure of a zinc finger module is shown. DNA-

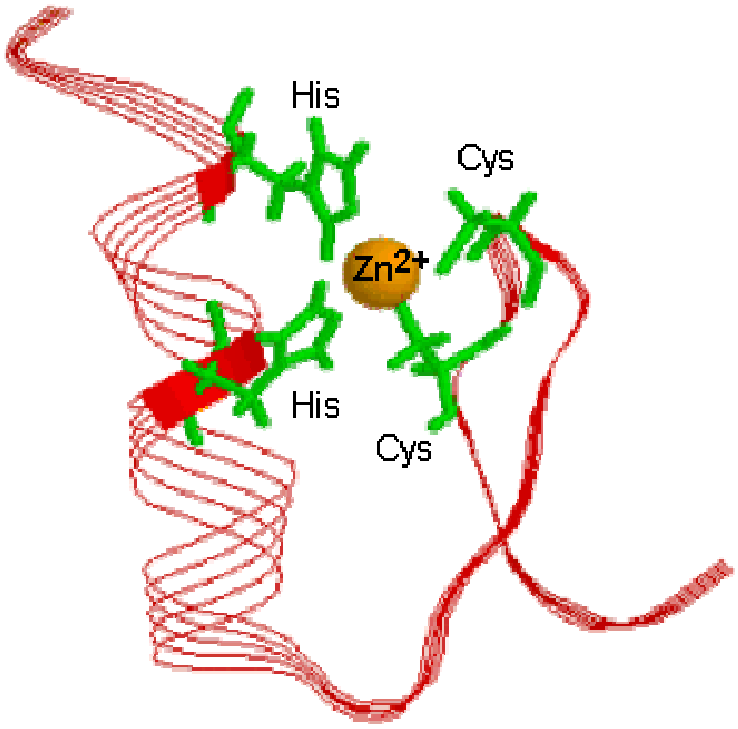


Figure 8: The structure of a zinc finger module is shown. The Zn^{++} ion, as well as the residues responsible for its cooperations are indicated. Two anti-parallel β strands and a α helix are sketched. The corresponding PDB identifier is 1sp1. The picture is taken from www.web-books.com.

binding proteins contain one or more of such zinc fingers in a single peptide chain to create their respective specificity. Each of the fingers of such a DNA-binding protein identifies a three base pair long sub-element of the transcription factor recognition site. The logic of multiple sub-elements to a larger super-element allows the relatively simple zinc finger to generate a great variety of highly specific interactions with DNA.

The *Drosophila* transcription activator GAGA, also called Trithorax-like, is a prominent member of the zinc finger family. The protein owes its current presentation to the fact, that it will be used as model protein structure in the study presented here. According to the bauplan mentioned earlier, GAGA also has two major structural domains: a single $C_2 - H_2$ zinc finger and an N-terminal trans-activation domain, which is split into the so called POZ and Q elements. The GAGA factor is ubiquitous in *Drosophila*

and highly conserved. The protein plays multiple roles in the fly system, in essence though, its functions are connected with its capability for selective DNA-binding. Firstly it is required for the regular expression of homeotic genes. It is known to enhance transcription of genes with promoters containing the core consensus sequence 'GAGAGAG' [57]. The protein alters the accessibility of specific promoter parts by a modification of the nucleosome structure. By this it facilitates the approximation and binding process of distinct transcription factors [80] essential for the start of the transcription process. Furthermore, its binding property allows for GAGA to be crucially involved in chromatin remodeling processes. The respective chromatin disruption of the promoter, has been shown to counteract proteins responsible for gene-silencing, specifically the Polycomb group proteins and its action. Another property of GAGA warrants mention: Many transcription factors dissociate from the genome during mitosis, but GAGA remains associated and thus interferes with the continuity of gene activation states. It allows to preserve the activity of GAGA dependent genes over generations of cell divisions [52,60]. Finally, the factor exhibits strong tendencies to self-oligomerize both *in vivo* and *in vitro*. The cooperative interaction of such an oligomer with promoters containing several GAGA binding sites versus promoters with single recognition sites, is of higher affinity and specificity [20].
For reviews on the detailed role of GAGA see [7, 30].

1.5 Biological Networks

Cells use complex networks of interacting molecular components to transfer and process information [6]. Already in the 1960s Monod and Jacob predicted that such fundamental cellular processes as differentiation and protein regulation are accomplished through 'signaling pathways' resident at the level of the gene [36]. Yet not later than today, at the beginning of the postgenomic era with its wealth of data genetic players and pathways, one striking insight stands out: Living cells do not alone use such gene network, but their regulatory impact on ontogeny and on fundamental functions is mission crucial. A central focus of postgenomic research will thus be, to understand how these cellular phenomena arise from the connectivity of genes and proteins. Especially the underlying design principles of such circuitry modules must be studied to eventually gain a deeper insight into the great variety of complex interactions.

1.5.1 Biological Models and Artificial Networks

Inspired by the field of electrical engineering with its well established techniques, the advances in the theory of nonlinear dynamics and the concurrent gain of significant computing power, researchers started to model biological, as well as design artificial networks. In the former case scientists build models of natural scenarios, such as the regulatory network involving the tumor suppressor protein p53 [1,78], figure 9, a network characterized by the interplay of pathways responsible for cell cycle control, stress response and genomic stability.

Such a model is built from knowledge of basic regulatory units and should predict the effects of genetic perturbation to the system [36]. From this concept, the undertaking's complicacy becomes obvious. Part of the difficulty is the high degree of complexity inherent in natural systems and the hassle of carrying out experiments on them *in vivo*. This is surely one of the basic reasons, why so far, relatively few computational modeling studies have involved tight coupling between model and experiment. The complexity indicates that modeling attempts should center on describing relatively simple systems and should be, despite the efforts, closely linked with experiments.

Accordingly, the λ-bacteriophage has been a system of choice for modeling studies. The biochemical reactions constituting the phage's control are well characterized and the fundamental biochemical reactions are understood [40,

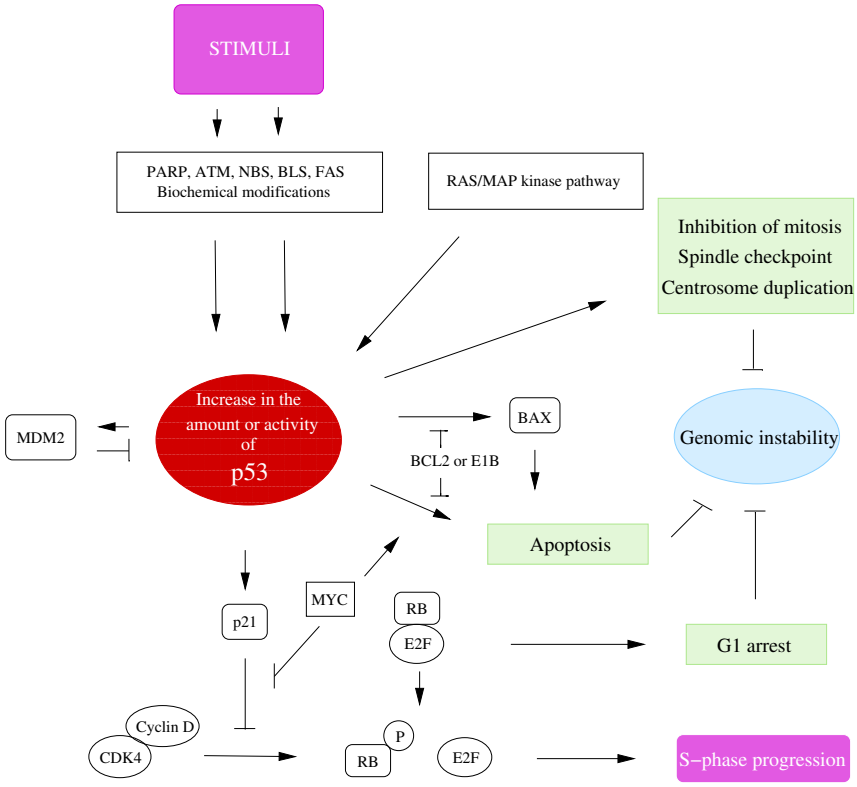


Figure 9: Components of the p53 network. Genotoxic stress activates p53 via PARP, ATM, NBS, BLS, and FAS. Pathways leading to modification of p53 are largely unknown. The RAS/MAP kinase pathway is involved in establishing basal p53 levels. Some of the cellular functions affected by p53 can be compromised by deregulation of Myc, Bcl2, E1B or E2F. Furthermore, an auto-regulatory loop involving MDM2 contributes to p53 regulation. p53-dependent pathways mantain genomic integrity by contorolled cell cycle arrest or apoptosis. p53 is furthermore involved in the regulation of entry into mitosis, spindle formation and centrosome integrity. Loss of functional p53 is found in half of all human cancer types. After [1].

47,53,59]. From the pool of many fertile studies, one representative is chosen here [36]. Hasty and coworkers used the λ control of the pivot between the lytic growth state and the dormant lysogenic state for their examinations. A set of differential equations (ODEs) describing the control process was used to investigate the transcription regulation of the λ repressor cI and the effects of intrinsic and extrinsic noise to the lysis/lysogeny control. The

authors conclude that stochastic fluctuations in specific reactions are capable to drive the system under investigation from one stable state to the other, thus is able to govern the decision between the two pathways.

A different scale of problem was approached in [79]. The authors used a set of rate equations to examine the regulation of segmentation in *Drosophila melanogaster*. The large parameter space of the model was searched numerically to find solutions matching experimental data. The initially assumed pattern of interaction, predominant in the literature, made it very difficult to find solutions with the desired behavior. After the introduction of several key connections, however, such patterns appeared to be relatively common. The authors predict these connections to exist also *in vivo*.

The design of artificial networks, offers the chance to study carefully chosen example models. These examples can be understood either as subsystems of naturally occurring circuitries or as attempt of constructing engineered control of cellular function. The big advantage is the possibility to test the qualities and the behavior of these small networks in a tractable experimental system. The long range goal would be to assemble increasingly complete models of the behavior of natural systems.

Co-repressive genetic switches are commonly known to be one of the frequent gene regulatory mechanisms [49]. The synthetic toggle switch developed by Gardner and coworkers [26] facilitates the study of such regulation motifs. The proposed network consists of two repressors and two constitutive promoters connected in such a way that one gene silences the other gene's expression. The pattern is depicted in figure 10. The model can be described in a set of ODEs and the dynamics can be numerically predicted. Yet, this synthetic gene motif can be brought to live, thanks to genetic engineering and tested in *E. coli*. The realization of an operating switch however, vitally depends on the choice of parameter, which lead the system into bistability. Such parameter combination are facilitated by the usage of effective transcriptional repression, the formation of protein multimers and similar protein degradation rates for different protein types.

At the same time, Elowitz and Leibler [18] designed and constructed a synthetic oscillatory network of transcriptional regulators. Their 'repressilator' consists of three promoters and corresponding genes with cyclic repressibility. It showed controlled oscillations in its protein production, mirrored in fluctuations of GFP, also included into the architecture. Their results provide valuable information on design principles of oscillatory systems, such as circadian clocks. Specifically the effects of stochastic fluctuations in these

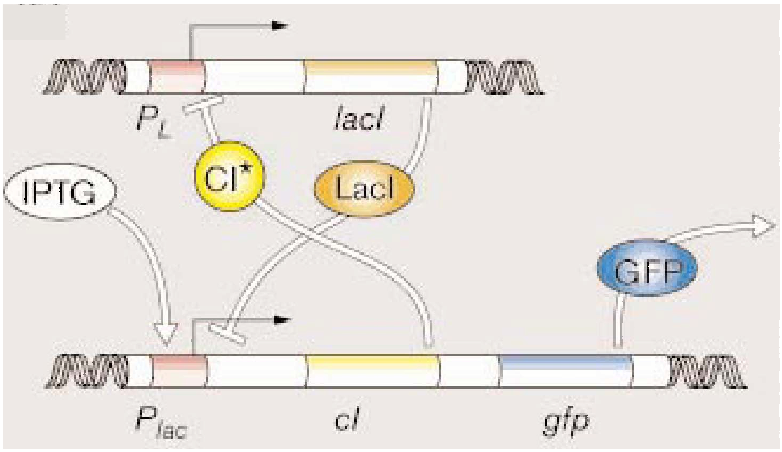


Figure 10: The model of a genetic toggle switch according to [26]. The two constructs are designed to silence each other's expression rates. Promoter regions, genes and products are indicated. The expression of cI is additionally controlled by a IPTG sensitivity of its promoter $Plac$. Moreover the tandem architecture of the lower construct allows the second gene on the plasmmid, GFP, to be expressed under the exact same conditions as is cI. The GFP fluctuations can be visualized and measured. The picture is taken from McAdams and Arkin (2000), Curr. Biol. 10.

systems seem to be problematic. The authors predict some kind of additional control *in vivo* to either circumvent or integrate noise influences.

Guet and coauthors [33] even designed a library of synthetic networks. It is composed of well characterized genetic units, such as the *LacI*, *TetR* and λ *cI* modules with the corresponding promoter sites. A combinatorial synthesis resulted in a great variety of possible interaction patterns displayed in *E. coli*. The different circuitries of the genetic elements showed phenotypic behaviors resembling mostly binary logic circuits. However, a group of networks formed the exception to the rule. The authors conclude that these unpredictable phenomena can be caused by the effects of subtle additional regulation or stochastic effects. They state that genetic network dynamics are non linear and stochastic processes. These properties facilitate undetected details in the interplay between components, which might be crucially important. The accessibility of a library of simple networks provides an alternative ansatz for the study of biological networks, as well as an efficient method for coupling

theoretical work with experiments *in vivo*.

Another approach is the venture for the artificial cell. The field ranges from the creation of specific-type artificial cells [58] over the rational design of a molecular scale cell machine to use for computation [65], to 'synthetic cell' models [4] for the better understanding of cellular behavior.

All these different approaches and strategies are driven by one concerted target: the craving for a more complete and deeper insight into networking processes at molecular level.

1.6 Molecular Evolution

1.6.1 Survival of the Fittest

Since Charles Darwin's 'The Origin of Species' was published in 1859, great scientific advances and profound insights have been made in the field of evolutionary theory. No later than with the establishment of modern molecular genetics in the middle of the last century, the deflecting of this field from a descriptive investigation of the phenotype to the study of the genotype at molecular level began. Despite various completions, further developments and alternative postulates, 'the survival of the fittest' forms a core monument not to be overlooked.

The power of Darwinian evolution is based on the dichotomy of genotype and phenotype. The genotype is the object under variation, whereas the phenotype is the target of selection. Two counteracting forces act on the heterogeneity of populations, these are mutation and the coupled increase of diversity versus selection with its diversity decreasing properties.
This knowledge allowed the development of the phylogenetic order of the species by means of the molecular clock hypothesis.

To date phylogenetic relations are created on the base of a genetic distance measure and the theory that there is a positive linear relationship from the point of time when two species diverge and the amount of genetic difference between these two species. The phylogenetic split of eubacteria, archaebacteria and eukaryota is shown in figure 11
There has been discussion, though, about the rate of DNA evolution and its continuity. The observations of phenomena coupled with mechanisms that account for the so called fluid genome, form the bases of the criticism.

Genomic reconstructions between species seem to be more often caused by different events as simply equally distributed random mutation: Descriptions of gene duplications in order to alter expression or to 'free' a gene copy of genetic pressure to evolve new functions, bursts of transposons, concerted evolution, gene shuffling or simply the encounter of hotspots of mutation, enforce the picture of a fluid genome with unsimilar evolution probabilities within, as well as between species. No doubt, the capability of genomes has gone far beyond what we had imagined.

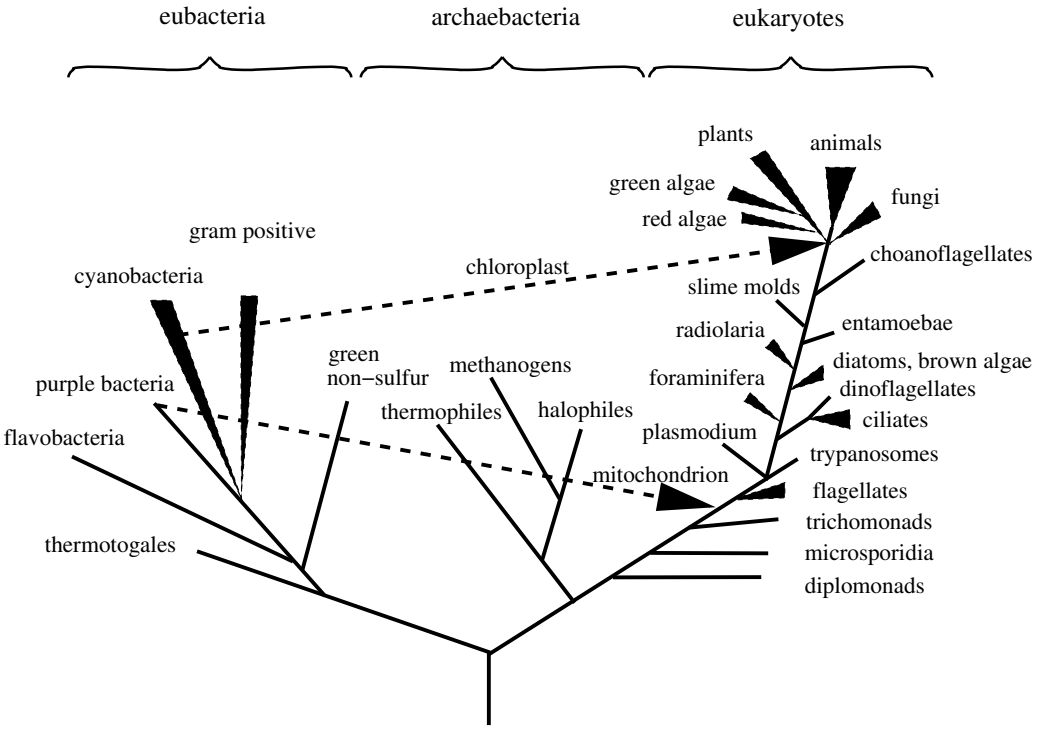


Figure 11: The phylogenetic tree of the three domains based on rRNA. The origin of chloroplasts and mitochondria is indicated. Modified after [9].

1.6.2 Genetic Algorithms

While the lifesciences slowly rearrange our view of the genome and its evolution at molecular level, theoreticians are developing great skills to model new insights and reuse them in topic-unrelated applications. Evolutionary principles found their way into mathematics and informatics. They find application in the field of non-trivial problems, problems that are computationally hardly solved. These problems have normally a huge space of possible solutions that can never be fully calculated, see also figure 12. The success of this approach is based on the assumption, that better solutions can be found in the neighborhood of good ones, similar to biological reasoning. In the mid-1970 John Holland first presented the concept of genetic algorithms (GA) in his pioneering book [38]. A genetic algorithm is an optimization procedure based on the Darwinian principle of the survival of the fittest.

A set of starting conditions, each of them a potential solution to the

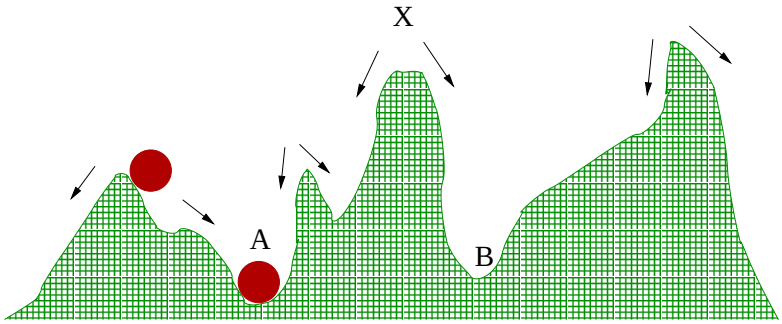


Figure 12: A two dimensional solution space. The height of position corresponds to its potential or lack of fitness. The valley A is, thus, the global minimum (highest fitness), the valley B is a local optimum. The valleys can also be understood as attractors in the solution space. The peak X delimits the basin of attraction of B.

problem of choice, is encoded and processed in a program. Each solution proposition needs a computed fitness value, which forms the basis for the stochastic selection and replication process. Since replication is supposed to be error-borne and hence the source of variation, it is one of the crucial steps in a GA, too. The implementation in form of recombination of two items, insertion, deletion, or a point-mutation of the encoded solution chunks is conceivable [42]. Possible types of changes in the items are called moveset. Within this moveset the concept of neighborhood is created. Important properties of the solutions have to be preserved and improvements, increasing fitness, must be possible between neighbors.

Because of the probabilistic selection, the best individual is not necessarily selected for replication and the worst can still remain in the population. Nevertheless, in general better solutions are favored. This gives GAs an advantage over pure hill climbing methods, which often fail with nontrivial problems.

1.6.3 *In vitro* Evolution

While the usage of evolutionary principles helps to find solutions to complex problems, which origin far from biological background, a core application is the simulation of evolution itself in order to gain a deeper insight into this highly complex process. The decisive *in vitro* experiment for *in silico* biology was a serial transfer experiment by Spiegelman and coworkers in the 1960s. Serial transfer describes the iterative transfer of a subsample of virus

or bacteria culture into a fresh environment of stock solution. Spiegelman simulated molecular evolution by the serial transfer of a $Q\beta$ replicase and the evolution of a super-fast replicase species. The fitness criterion, velocity of replication, was implemented by the observation that quick replicators were able to gain a higher proportion of the necessary raw material, the nucleotides, just because they were faster in using them. With each transfer step the chances to transfer a majority of quick replicators increased. The result replicase exhibited an increased RNA synthesis rate by more than one order of magnitude [71].

This epoch making experiment opened new frontiers for *in silico* biology. Especially the simple setting of the replicase experiment emphasized the advantages of alike experiments carried out on the computer. The absolute knowledge of 'what goes in and what comes out', the possibility of a detailed analysis and last but not least the up-speeding of what usually happens in years or decades to hours or days, makes the computer the favored device for such undertakings.

1.6.4 The Flow Reactor

The first theoretical model of molecular evolution was proposed by Manfred Eigen [17]. This article discusses the kinetics of replication, mutation and selection in populations of asexually reproducing species, in this case polynucleotides. The model of polynucleotide replication was based on ODEs derived from chemical kinetics. The simple reaction network is set up in an environment called flow reactor, a huge holding tank with in- and out-flux control. The evolving culture is constantly mixed to avoid mal-distributions of source material, degradation products are removed by a size specific filter and the population size is kept constant. A cartoon of a flow reactor is shown in figure 13.

The system writes as

$$\begin{aligned} (A) + I_j &\xrightarrow{\mathrm{Q_{kj}A_j}} I_j + I_k \\ I_j &\xrightarrow{\mathrm{D_j}} (B) \\ I_i &\xrightarrow{\Phi(\mathrm{t})} 0 \ . \end{aligned}$$

An RNA sequence I_j is replicated with the rate constant A_j, Q_{kj} is the probability of erroneous replication generating the offspring individual I_k. Q_{jj}

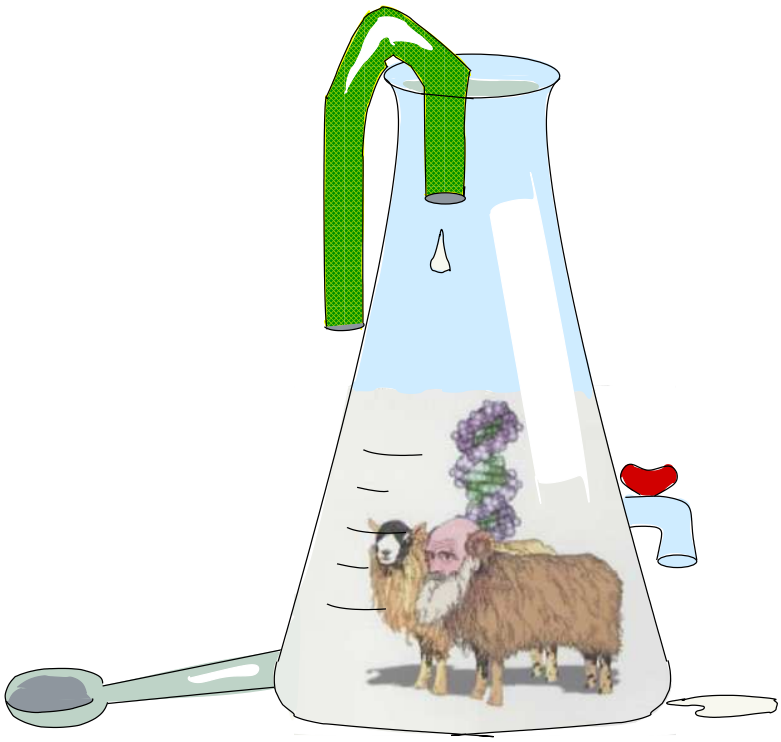


Figure 13: The flow reactor. The result population of a simulation is shown. Specific in- and out-flow channels assure a constant environment, as well as population size. The reaction medium is permantently mixed to avoid mal-distribution of the source material.

would be the probability for error-free replication of sequence I_j. Buffered source nucleotides are represented by A. The degradation products of the sequence I_j, which are produced with the rate D_j are constantly removed. Furthermore included in this model is an unspecific dilution flux $\Phi(t)$, removing templates from the system to keep the total population size constant.

The kinetic analysis of this replication-mutation system showed that there exists a sharply defined minimum replication accuracy, the error threshold, below which populations become unstable and drift randomly through sequence space. Stationary states of a population are characterized by distributions of sequences, so called quasispecies, around a master sequence, the fittest individual among them. However, not a single fittest type, but the entire quasispecies is selected by the evolutionary process. Such quasistationary states can be destabilized by rare advantageous mutations and subsequently

the population moves toward a new quasistationary sequence distribution.

Another computer simulation of evolutionary optimization was carried out by Walter Fontana and Peter Schuster [23]. They analyzed a replication-mutation model using hydrolytic degradation. Selection values for replication and degradation were derived from phenotype properties, the secondary structure of the RNA sequence. Rate constants are calculated explicitly for all individuals in the reactor. Thereby attention is turned to secondary structure elements like stacked regions, which are assumed to slow down replication, and to increasement in hydrolytic degradation in unpaired regions.
All typical features of evolution of populations, like error thresholds and quasistationary sequence distribution could be observed.

In early computer simulations of evolutionary adaptation, fitness values were computed from kinetic constants derived from secondary structure folds according to [22]. Recent work on this subject focused on the distance between an evaluated phenotype and a target secondary structure, as well as on the implications of neutral networks in sequence space with identical structure in adaptive evolution [24, 25].

2 Development and Mathematical Analysis of a Genetic Switch Model

The synthesis of biological data into formal models of cellular functions has rapidly assumed vast proportions [4]. The complexity of interactions among cellular constituents and the quantity of data available impede the intuitive interpretation and prediction of such networks' behaviors. Hence, the advantage of systematization and formalization in models is availed for the description and analysis of such complex, dynamic processes.

Of course, data are the precursor to any *biological* model. The minimal basis of a cellular network model is a list of molecular players and a list of the proposed interactions among them. This is equally true for the *rational* design of networks and their respective models. Even though they are no reproduction of a natural reality, they are still supposed to exist in a similar environment and to involve the same sort of *natural* players. Further, the long term target of artificial design must always be the connection or re-connection with the 'wet' experiment and ideally the transfer to *in vivo* systems. Hence, a strong resemblance to real systems is desirable, even though the conventionalized design only mimics biology.

The part of work presented in this chapter deals with the artificial design of a minimal gene switch. The concept for an auto-regulatory switch was inspired by the particularly frequent usage of switches as building blocks of electronic circuitries, circuitries of virtually any possible dynamic behavior. Analogously, ubiquitous biological building blocks are of special interest, because of their conceptual, descriptive simplicity and the universal applicability in the control of cellular functions. Basic patterns such as for example the analogues to the logic 'AND' or 'OR' gates, and, not to be forgotten, switches between two or more possible cellular conditions apply for this challenge. Such a designed 'bi-switch' module is characterized by its bistable dynamic behavior. In terms of dynamics, the network's protein production should exhibit specific characteristics with respect to its flow behavior, which are depending on the initial concentration state of the network.

We imagine a scenario with two different protein types. The net production of protein type 'A' is significantly higher than of protein type 'B' at the equilibrium of the reaction flows. A shift of the protein concentrations should result in a switch of the maximal protein production between the two protein species. A different, second equilibrium is reached. The concentration

of protein type 'B' out-competes type 'A' at this distinct steady state.

We designed the module under the condition of the greatest possible nearness to nature. This setting includes the interactions of genes, their transcripts and proteins for the construction of a genetic regulatory network. Furthermore, the interactions of the listed molecule species and the probabilities of reactions were modeled carefully with a main focus on sequential, structural and kinetic properties of the molecules. Additionally, the model features cooperativity of the proteins, the usage of strong promoters and constant number of genes. Natural processes that consist of several coupled reactions are often target of stochastic fluctuations [19, 35]. To account for this natural phenomenon we want to introduce the possibility for noise acting on transcription and translation reactions in our model.

The models presented here are formulated as chemical reaction schemes, which allow the deduction of a set of ODEs for a qualitative as well as a quantitative description. The mathematical analysis of these sets allows to deduce respective network properties in analytical or semi-analytical expressions depending on the number of involved players. On that basis, we can decide if the model under consideration resembles a gene switch or not and statements over its 'mathematical environment' can be formulated.

In later stages of this study we can take advantage of this formal framework. The simulation of evolution of RNA molecules *in silico* has so far been limited to the adaptation of single molecules [22–25]. On the basis of these studies, we want to take a step ahead and combine the molecular approach used in these experiments and the mathematical formalization of the switch model to be able to study the adaptation processes of complex molecule networks. The detailed description see chapter 3.

2.1 PLUM Network

In order to construct a genetic switch, which is capable to fall into two different states, as for a 'YES' and a 'NO' bit, we choose a minimal model consisting of a 'YES' and a 'NO' component itself, designated PLUM (**PLU**s-**M**inus). PLUM is sketched in figure 14. The idea for a genetic minimal switch is based on two molecular players: Gene A acts as inductor, whereas gene B is an inhibitor of transcriptional activity. The activity of both genes is mediated by a conformational change of their structures induced by the interaction of the gene and its antagonist gene product. The transcript of gene A forms a complex with gene B, whereby the gene's promoter is struc-

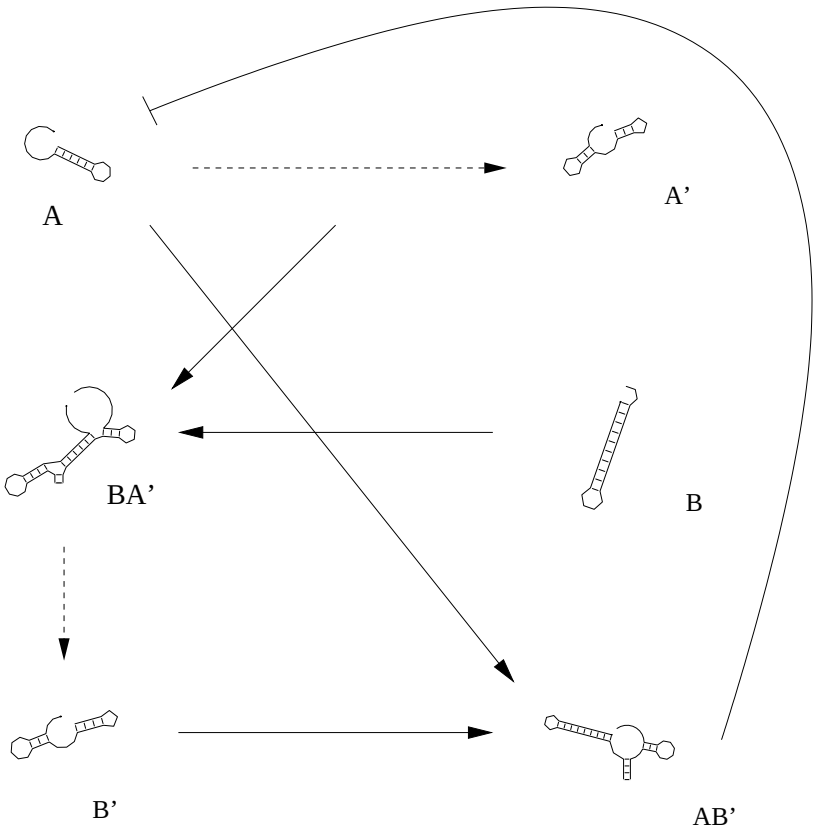


Figure 14: The PLUM network. The genes A and B act antagonistically as inducer or inhibitor of transcription. These effects are caused by a conformational change in the promoter region induced by the forming of a complex between gene and the analogous gene product. A, B ...genes, A', B'...transcripts.

turally opened and transcription of gene B is enabled. Here, DNA and RNA molecules are assumed to be present as single strands in order to be capable to form hybrid secondary structure. We assume that RNA molecules posses half-lives according to the real-life situation in a functional cell. The second half-cycle of the network consists of gene product B, which forms a complex with gene A. This complex formation silences the promoter region of gene A, no further transcripts are generated.

The chemical reaction scheme The above model can be converted into a chemical reaction scheme. The kinetic steps forming and removing

RNA molecules are assumed to be irreversible, whereas complex formation is supposed to be reversible. DNA molecules, the genes, are neither formed nor degraded and thus, the total amount is constant, denoted a_0 and b_0, respectively.

$$\begin{aligned}
A &\xrightarrow{\mathrm{t_1}} A + A' \\
A + B' &\xrightarrow{\mathrm{k_1}} AB' \\
AB' &\xrightarrow{\mathrm{k'_1}} A + B' \ ; \ K_A = k_1/k'_1 \\
A' &\xrightarrow{\mathrm{d_a}} 0 \\
B + A' &\xrightarrow{\mathrm{k_2}} BA' \\
BA' &\xrightarrow{\mathrm{k'_2}} B + A' \ ; \ K_B = k_2/k'_2 \\
BA' &\xrightarrow{\mathrm{t_2}} BA' + B' \\
B' &\xrightarrow{\mathrm{d_b}} 0
\end{aligned}$$

According to the reaction scheme, we are dealing with six molecular species, A, A', B, B', and AB', A'B. The dynamical system is constrained by the conservation relations for genes, $a_0 = [A] + [AB'] = const.$ and $b_0 = [B] + [BA'] = const.$, respectively. Furthermore, we assume that the reversible reactions k_1, k'_1, k_2 and k'_2 are at an equilibrium and find $K_A = k_1/k'_1 = [AB']/[A][B']$ and $K_B = k_2/k'_2 = [A'B]/[A'][B]$. From this we compute at the equilibrium that,

$$\begin{aligned}
[A] &= \frac{a_0}{(1 + K_A y)} \\
[B] &= \frac{b_0}{(1 + K_B x)} \\
[AB'] &= \frac{K_A a_0 y}{(1 + K_A y)} \\
[A'B] &= \frac{K_B b_0 x}{(1 + K_B x)} \ ,
\end{aligned}$$

given $x = [A']$ and $y = [B']$.

PLUM's ODEs and fixed points We find PLUM to be fully described by 2 states variables, namely x and y. The proposed reaction mech-

anism yields the following two differential equation to describe the dynamics of the network.

$$\begin{aligned} \frac{dx}{dt} &= \frac{t_1 a_0}{1 + K_A y} - d_a x \\ \frac{dy}{dt} &= \frac{t_2 b_0 K_B x}{1 + K_B x} - d_b y \end{aligned}$$

In order to find the stationary states $(\overline{x}, \overline{y})$ of the system we have to search for solutions characterized by vanishing time derivatives. They are easily computed from the solution of the two dimensional equation system $\frac{dx}{dt} = 0$ and $\frac{dy}{dt} = 0$ and write as,

$$\begin{aligned} \overline{x}_{1,2} &= \frac{\gamma K - d}{2K(\gamma K + d)} \pm \sqrt{\frac{1}{4K^2} + \frac{\gamma^2}{(\gamma K + d)^2}} \\ \overline{y}_{1,2} &= \frac{\gamma K \overline{x}}{d(1 + K\overline{x})} \ , \end{aligned}$$

where $\gamma \equiv t_1 \cdot a_0 = t_2 \cdot b_0$, $K \equiv K_A = K_B$ and $d \equiv d_a = d_b$. We find that the root expression will always be greater than the first term, hence we obtain two solutions for $\overline{x}$, one positive and one negative. Since the network lives in the world of concentrations the positive solution for $\overline{x}$ and the corresponding $\overline{y}$ denote the 'legal' fixed point of the PLUM system.

Dynamic behavior Before we start analyzing PLUM's behavior, we first need to clarify which behavior is to be expected at the fixed points. The chronologically first question to ask would be, what is stability? A non wandering set (i.e. a fixed point, a limit cycle, a quasi periodic or chaotic orbit) may be stable or unstable. Moreover, it is either asymptotically stable, marginally stable (see also Lyapunov stable) or unstable. Asymptotically stable non-wandering sets are so called attractors. The basin of attraction is the set of all initial states approaching the attractor in the long time limit.

The determination of the stability of a fixed point is carried out by the numerical analysis of its eigenvalues, as will be demonstrated later. The eigenvalues λ_i are the roots of the characteristic polynomial $p_J(\lambda) = det(J - \lambda E) = 0$, with E being the unitary matrix and J the Jacobian matrix. The fixed point is asymptotically stable, if all eigenvalues λ_i reside inside

a stability area of the complex plane spanned by the Eigenvectors $\vec{\lambda_i}$ or in other words, if all real parts of the eigenvalues are negative.

If at least one eigenvalue is outside the stability area, i.e. positive, the corresponding solution of the ODE would increase exponentially, thus, the fixed point is unstable. Figure 15 shows a classification scheme of fixed points of two dimensional phase spaces. The notion spiral and node are inspired by the flow near the fixed points. A pair of conjugate complex eigenvalues cause a spiral, whereas a node is characterized by two real eigenvalues of the same sign. Real eigenvalues of different sign account for a saddle. Generally spoken, a saddle is a fixed point with at least one eigenvalue having a positive real part, but also at least one eigenvalue having a negative real part. Near a saddle, an orbit is usually attracted at first, but repelled later on. There are points in the phase space that approach the fixed point for $t \to \infty$, they form the stable manifold. The unstable manifolds are built by all points approaching the fixed point for $t \to -\infty$. Saddles and their stable manifolds

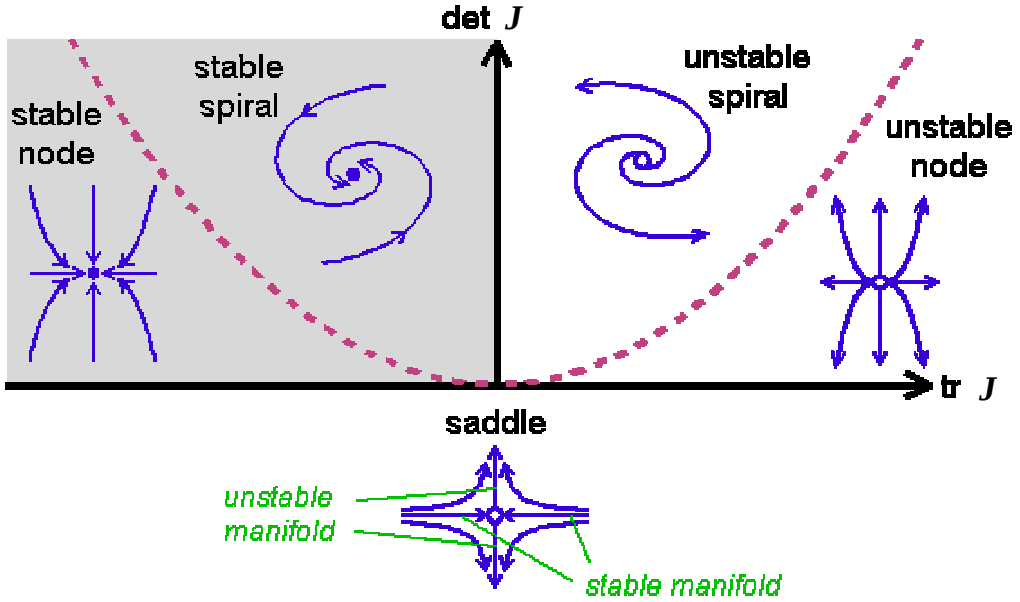


Figure 15: Classification of fixed points in two dimensions. The behavior at a fixed point is dependent on the eigenvalues' sign and whether the eigenvalues are all real or complex numbers. A saddle point is caused by one negative and one positive eigenvalue. Picture taken from monet.physik.unibas.ch/ẽlmer/pendulum/

are usually the boundaries of the basin of attraction.

In order to determine PLUM's dynamics at the stationary point, we compute the Jacobian matrix J and solve through normal mode analysis. The Jacobian is built up by the partial derivatives of the differential equations.

$$J = \begin{pmatrix} \frac{\partial}{\partial x}(\frac{dx}{dt}) - \lambda & \frac{\partial}{\partial y}(\frac{dx}{dt}) \\ \frac{\partial}{\partial x}(\frac{dy}{dt}) & \frac{\partial}{\partial y}(\frac{dy}{dt}) - \lambda \end{pmatrix}$$

For our network we find the following expression:

$$J = \begin{pmatrix} -d - \lambda & \frac{-\gamma K}{(1+K\overline{y})^2} \\ \frac{\gamma K}{(1+K\overline{x})^2} & -d - \lambda \end{pmatrix} .$$

The eigenvalues of the system are found by solving $Det(J) = 0$. For the PLUM system we find the following expressions:

$$\lambda_{1,2} \;=\; -d \pm i\left(\frac{K\gamma}{(1+K\overline{x})(1+K\overline{y})}\right) .$$

The eigenvalues $\lambda_{1,2}$ are complex conjugate solutions and have a negative real part. This fact shows that the PLUM network is globally stable and the concentrations of x and y spiral into the stationary point as can be seen in figure 16. In other words, after a time of disequilibrium the concentrations of x and y adopt fixed values and keep them for all times. As for the completion of the task to construct a genetic switch, the analysis of the eigenvalues $\lambda_{1,2}$ yields that by choosing any given combination of parameters $\alpha \in \mathbb{R}^+$, the system will always behave as a stable focus.

2.2 Cooperative Binding and Dynamic Systems

The PLUM model is a stable network as has been shown earlier. Straightforwardly, the reason for that property is, that the set of differential equations is highly linear. In order to find more complex behavior in a model, such as

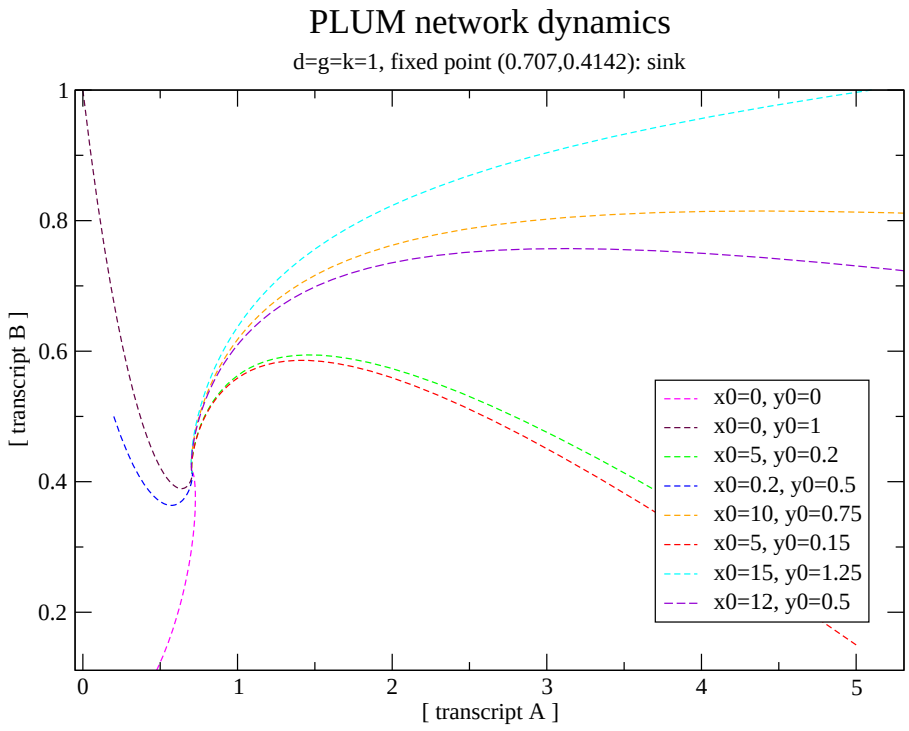


Figure 16: The dynamics of the PLUM network at default parameter setting. Default parameter are $\gamma = d = k = 1$, start values of x and y are varying as indicated. All time trajectories spiral into the sink-like stationary point $(1/\sqrt{2}, \sqrt{2}-1)$.

bistability, oscillations or deterministic chaos, nonlinear terms have to be introduced. It is common sense that complex dynamic patterns can be found in systems of equations containing nonlinear terms of degree three and higher.

In the literature we find multi-stability and oscillations with larger networks. Sticking with the idea of a minimal model though, without extending the number of genes, cooperativity comes along handy. Cooperativity is a well described phenomenon in enzyme kinetics. The binding of a ligand to a macromolecule causes a conformational change in the macromolecule, whereby it leads to increasing binding affinity and kinetic rates with increasing number of ligands. Without changes in the macromolecular structure

anticooperativity is more common. Further ligands are less strongly bound because of ligand-ligand repulsion and other bond weakening phenomena. Accordingly, the introduction of cooperativity in the PLUM minimal model might provide the missing nonlinear terms to exhibit complex behavior.

2.2.1 Induced Fit Model with Two Ligands

The induced fit model with two ligands is set up as a minimal model for cooperativity. However, it is not clear whether sufficient self-enhancement can be achieved.

The model considers a nucleic acid molecule with two binding sites for a ligand S. In context with the PLUM model it is straightforward to think of a dimer of transcription factors binding to the promoter site of the gene. Binding of the first monomer S induces a structural change, e.g. a bend of the DNA, that results in a higher binding affinity of the second ligand. The binding equilibria fulfill the following reaction scheme

$$E + 2S \overset{K_1}{\Longleftrightarrow} ES + S \overset{K_2}{\Longleftrightarrow} ES_2 \ .$$

Cooperativity implies $K_1 < K_2$ or $K = K_1$ and $K\gamma = K_2$ with $\gamma > 1$. Accordingly, $\gamma = 1$ implies the independent binding of the two ligands S and $\gamma < 1$ represents the anti-cooperative case. For the two equilibria we find

$$K = \frac{[ES]}{[E][S]} \quad \text{and} \quad K\gamma = \frac{[ES_2]}{[ES][S]}$$

Conservation of mass yields the two equations

$$\begin{aligned} e_0 &= e + e_1 + e_2 \quad \text{and} \\ s_0 &= s + e_1 + 2e_2 \ , \end{aligned}$$

where $e = [E]$, $s = [S]$, $e_1 = [ES]$ and $e_2 = [ES_2]$. We are interested in the fraction of free and bound sites, as a function dependent on the total concentration of ligand molecules, s_0 and total enzyme e_0. This binding coefficient is given by the expression $\theta = (e_1 + 2e_2)/2e_0$.

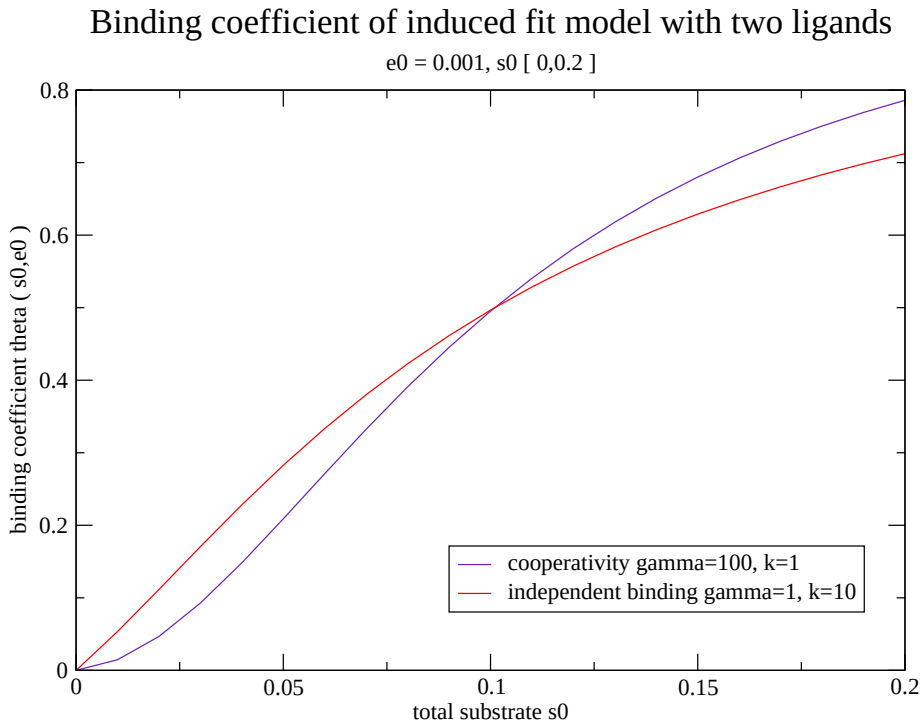


Figure 17: Binding curve of the induced fit model with two ligands. Cooperativity and an independent binding curve is shown. The parameter settings were determined empirically. The parameter for the sigmoid cooperativity curve are total enzyme $e_0 = 0.001$, total substrate $s_0 = [0, 0.2]$, $K = 1$ and $\gamma = 100$. The settings for the independent binding vary in $K = 10$ and $\gamma = 1$.

By partly solving the four dimensional equation system, we find an expression for the free substrate s as a function of total substrate s_0 and total enzyme e_0,

$$s^3 + s^2\left(\frac{K}{H} + (2e_0 - s_0)\right) + s\left(\frac{1}{H} + \frac{K}{H}(e_0 - s_0)\right) - \frac{s_0}{H} = 0 \ ,$$

where $H = \gamma K^2$. The 'legal' concentration $s(e_0, s_0)$ is obtained as the real, positive root of this equation of third degree. Accordingly, we find the con-

centrations of the complexes e_1 and e_2 as

$$\begin{aligned} e_1 &= \frac{Ke_0 s(e_0, s_0)}{1 + Ks(e_0, s_0) + Hs(e_0, s_0)^2} \\ e_2 &= \frac{He_0 s(e_0, s_0)^2}{1 + Ks(e_0, s_0) + Hs(e_0, s_0)^2} \ . \end{aligned}$$

Numerical analysis of $\theta(e_0, s_0)$ yields a sigmoid graph for $\gamma > 1$ and a hyperbolic curve for $\gamma = 1$ as plotted in figure 17.

Nonlinear self-enhancement of biological activity induced by cooperativity requires parameter settings in the range of those, producing sigmoid binding curves. Accordingly, it is possible to obtain cooperativity phenomena in an induced fit model with only two ligands. We propose to obtain nonlinear dynamics by introducing this minimal cooperativity model into a linear network.

2.3 PLOOP Network

2.3.1 Analytical Approach

Extending the PLUM network by cooperativity, we propose a network designated PLOOP (**PL**us-minus c**OOP**erativity).

The PLOOP system consists of two genes, gene A and B, working as antagonists as can be seen in figure 18. Gene A is transcribed and the messenger is processed to protein A.
Gene B is available only in supercoiled form, the promoter site is inaccessible, gene B is silenced. In order to alter the expression rate of gene B, a dimer of protein A has to bind to its upstream region. Protein A interacts with the promoter site of gene B, whereby the binding affinity of a second monomer of protein A enhances, mediated by a structural change in the nucleic acid molecule. Once the protein dimer is located properly, transcription of gene B is induced. B mRNA is processed to the corresponding protein. Both mRNAs and proteins are produced and degraded at a certain rate, the number of genes is constant, complexes are formed reversibly.

THE PLUS–MINUS–COOPERATIVITY NETWORK – PLOOP

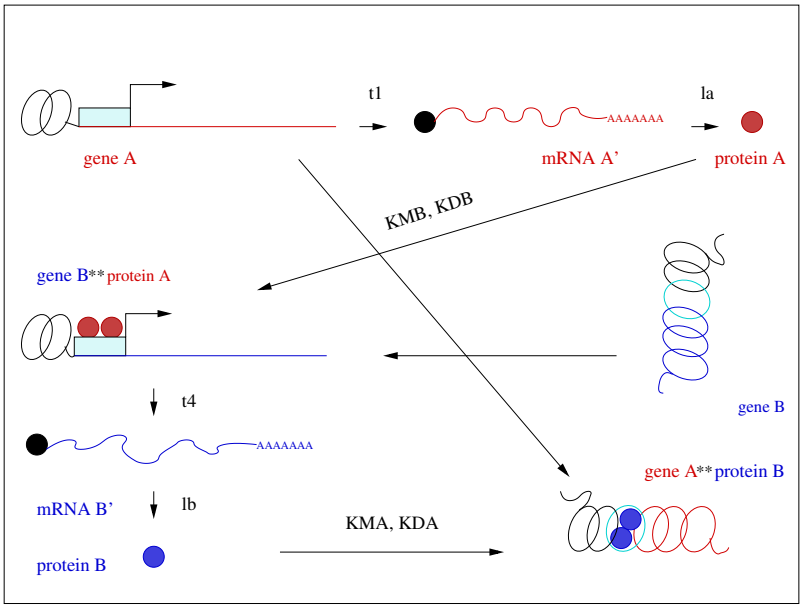


Figure 18: The PLOOP network. Gene A is transcribed and translated with the respective rates, t1 and la. A dimer of protein A bound to the promoter site of gene B relaxes the supercoiled structure, processing becomes possible. A dimer of protein B bound to gene A inhibits transcription by conformational change. Species of the 'A' half cycle are denoted in red, of the 'B' half cycle in blue. The notions for the reaction rates are taken from the reaction scheme.

The model reactions write as

$$
\begin{aligned}
A &\xrightarrow{t_1} A + A' \\
A + P_B &\xrightarrow{k_1} AP_B \\
AP_B &\xrightarrow{k_1'} A + P_B \ ; \ K_{MA} = k_1/k_1' \\
AP_B + P_B &\xrightarrow{k_2} A(P_B)_2 \\
A(P_B)_2 &\xrightarrow{k_2'} AP_B + P_B \ ; \ K_{DA} = k_2/k_2' \ , \ K_{DA} > K_{MA} \\
AP_B &\xrightarrow{t_2} AP_B + A' \ ; \ t_1 >> t_2 \\
A' &\xrightarrow{l_a} A' + P_A \\
A' &\xrightarrow{d_a} 0 \\
P_A &\xrightarrow{d_{pa}} 0 \ ; \ d_a >> d_{pa}
\end{aligned}
$$

$$
\begin{aligned}
B + P_A &\xrightarrow{k_3} BP_A \\
BP_A &\xrightarrow{k'_3} B + P_A \ ; \ K_{MB} = k_3/k'_3 \\
BP_A + P_A &\xrightarrow{k_4} B(P_A)_2 \\
B(P_A)_2 &\xrightarrow{k'_4} BP_A + P_A \ ; \ K_{DB} = k_4/k'_4 \ , \ K_{DB} > K_{MB} \\
BP_A &\xrightarrow{t_3} BP_A + B' \\
B(P_A)_2 &\xrightarrow{t_4} B(P_A)_2 + B' \ ; \ t_3 << t_4 \\
B' &\xrightarrow{l_b} B' + P_B \\
B' &\xrightarrow{d_b} 0 \\
P_B &\xrightarrow{d_{pb}} 0 \ ; \ d_b >> d_{pb} \ .
\end{aligned}
$$

From conservation of the number of genes, we obtain the expression for the concentrations of

$$
\begin{aligned}
[A] &= \frac{a_0}{1 + K_{MA} w + K_{DA} K_{MA} w^2} \\
[B] &= \frac{b_0}{1 + K_{MB} v + K_{DB} K_{MB} v^2} \ .
\end{aligned}
$$

The reversible reactions are supposed to be at equilibrium using $K_{M(A,B)}$ as binding rate for the first protein monomer and $K_{D(A,B)}$ for the second. At the equilibrium the system is fully described by four state variables, i.e. $x = [A']$, $y = [B']$, $v = [P_A]$ and $w = [P_B]$, the concentrations for the complexes are

$$
\begin{aligned}
[AP_B] &= \frac{K_{MA}[A]w}{1 + K_{DA} w} \\
[A(P_B)_2] &= \frac{K_{MA} K_{DA} [A] w^2}{1 + K_{DA} w} \\
[BP_A] &= \frac{K_{MB}[B]v}{1 + K_{DB} v} \\
[B(P_A)_2] &= \frac{K_{MB} K_{DB} [B] v^2}{1 + K_{DB} v}
\end{aligned}
$$

.

From the chemical reaction scheme, the following differential equations are

deduced:

$$\begin{aligned}
\frac{dx}{dt} &= t_1[A] + t_2[AP_B] - d_a x \\
\frac{dy}{dt} &= t_3[BP_A] + t_4[B(P_A)_2] - d_b y \\
\frac{dv}{dt} &= l_a x - d_{pa} v \\
\frac{dw}{dt} &= l_b y - d_{pb} w \ .
\end{aligned}$$

We introduce the biological constraint that transcription and inhibition of transcription are induced only after the protein dimer has caused the conformational change in the 5' UTR of the gene by setting $t_2 = t_3 = 0$. Accordingly, the fully rewritten set of equations looks as follows:

$$\begin{aligned}
\frac{dx}{dt} &= \frac{t_1 a_0}{1 + K_{MA} w + K_{DA} K_{MA} w^2} - d_a x \\
\frac{dy}{dt} &= \frac{t_4 b_0 K_{MB} K_{DB} v^2}{(1 + K_{DB} v)(1 + K_{MB} v + K_{DB} K_{MB} v^2)} - d_b y \\
\frac{dv}{dt} &= l_a x - d_{pa} v \\
\frac{dw}{dt} &= l_b y - d_{pb} w
\end{aligned}$$

For demonstration, we calculate stationary states of the PLOOP system by simplifying the parameters $l_a = l_b = d_a = d_b = d_{pa} = d_{pb} = K_{MA} = K_{MB} = t_1 = t_4 = 1$ and $K \equiv K_{DA} = K_{DB}$ and find

$$\begin{aligned}
\overline{x} &= \overline{v}(K) \\
\overline{y} &= \overline{w} \\
\overline{w} &= \frac{K\overline{v}(K)}{(1 + K\overline{v}(K))(1 + \overline{v}(K) + K\overline{v}(K)^2)}
\end{aligned}$$

and $\overline{v}(K)$ as a polynomial of degree seven,

$$\begin{aligned}
&-\overline{v}^7 K^4 + \overline{v}^6 K^3 (K-5) + \overline{v}^5 K^2 (K-8) + \overline{v}^4 K (2K^2 - K - 5) \\
&+\overline{v}^3 (5K^2 - 3K - 1) + \overline{v}^2 (K^2 + 4K - 1) + \overline{v}(1 + 2K) + 1 = 0 \qquad (2)
\end{aligned}$$

The expressions for the fixed points without any simplifications look structurally similar, but contain all given parameters. They are not shown here.

Calculating the roots of a polynomial analytically is impossible if its order is greater than five and tedious if the order is greater than two. The solution for $\overline{v}(K)$ and all further steps of analysis are therefor, properly calculated numerically.

According to numerous calculations of the solution to expression 2, we find two different scenarios: The system has one or three 'legal' stationary points. Thus, the underlying polynomial of degree seven, shown in expression 2, can be actually simplified to a cubic polynomial.

2.3.2 Semi-Analytical Analysis

Since we have determined the expressions for the fixed states dependent on $\overline{v}(K)$, we can now built the Jacobian at the stationarity dependent on $\overline{v}(K)$. The partial derivatives are readily calculated and plugged into the matrix.

$$J = \begin{pmatrix} -d, 0, 0, D(\overline{v}) \\ 0, -d, G(\overline{v}), 0 \\ l, 0, -dp, 0 \\ 0, l, 0, -dp \end{pmatrix} ,$$

where small letters denote the parameters, degenerated in A and B, i.e. for example $l \equiv l_a = l_b$ and capital letters stand for the partial derivatives at the respective entry in J with

$$\begin{aligned} D &= \frac{\partial}{\partial w}\Big(\frac{dx}{dt}\Big) = \frac{-\big(a_0 t_1 (K_M + 2K_D K_M \overline{w}(\overline{v}))\big)}{(1 + K_M \overline{w}(\overline{v})(1 + K_D \overline{w}(\overline{v})))^2} \\ G &= \frac{\partial}{\partial v}\Big(\frac{dy}{dt}\Big) = \frac{b_0 K_D K_M t_4 \overline{v}(P)\big(2 + (K_D + K_M)\overline{v}(P) - K_D^2 K_M \overline{v}(P)^3\big)}{(1 + K_D \overline{v}(P))^2 (1 + K_M \overline{v}(P)(1 + K_D \overline{v}(P)))^2} \text{ and} \\ P &= a_0, b_0, t_1, t_4, d, dp, l, K_M, K_D \end{aligned}$$

It is possible to obtain an analytical expression for the eigenvalues λ_i due to the empirical reduction of the parameter set. We therefor find, after the diagonalization of the Jacobian matrix J, the following terms:

$$\lambda_{1,2,3,4} = \frac{1}{2}\Big(- dp - d \pm \sqrt{(dp - d)^2 \pm 4l\sqrt{DG}}\Big) \qquad (3)$$

As described earlier the system shifts between one or three fixed points. Using equation 3 we find for the former case, the eigenvalues λ_i have negative real parts and two dimensions are additionally complex conjugate. For the latter case two stable fixed points comport according to the same scheme and one being an unstable saddle with one positive real eigenvalue. This circumstance is shown in figure 19. The figure plots the course of the fixed point in $\overline{x}$ along the bifurcation axis $\Gamma = K_D$. From this result we predict PLOOP to exhibit two different behavior types: stability and bistability, which will be to be verified in the later analysis steps.

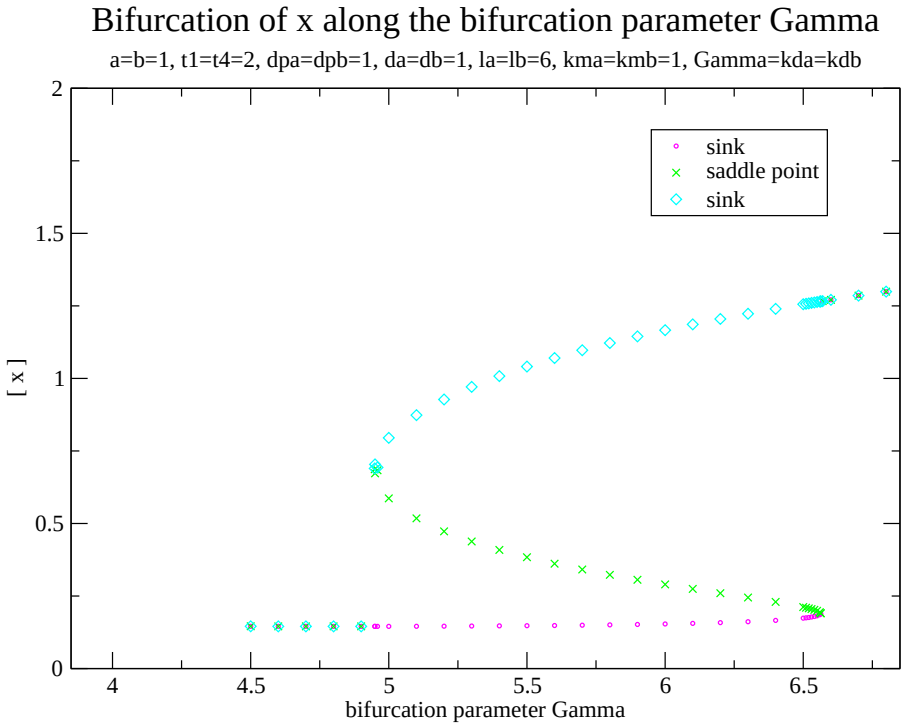


Figure 19: Fixed points along the axis of bifurcation parameter Γ. The course of the fixed point induced by tuning a single parameter, $\Gamma = K_D$ is shown.

Time-series Using the numerical integrator package *CVODE* [1], we can construct the time-series of PLOOP for arbitrary control parameter sets and initial conditions. For the time-series in figure 20 we selected a parameter set that induces three fixed points. We chose two different sets of initial conditions. Depending on these external stimuli, the trajectory ends in stationarity one or two. By shifting the initial conditions, the system shows a maximum concentration of a species in the network, either for protein A or

[1] CVODE is a solver for stiff and non-stiff initial value problems for systems of ODEs. It is freely available on the web, among others at *robotics.stanford.edu/users/scohen/bio.html*

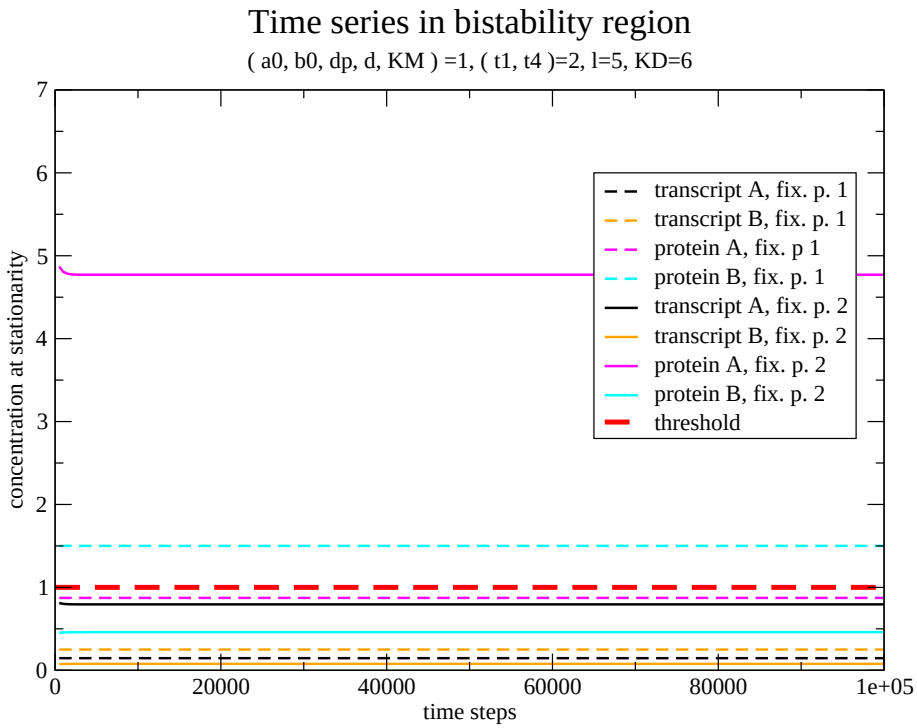


Figure 20: Time-series of PLOOP in a bistability region. Depending on initial conditions the system falls into attractor one or two with either gene A or gene B at max. expression. Introducing a threshold points out the switch property of the system.

protein B. According to this result, PLOOP behaves as a switch in the bistability region. The saddle point acts as a boundary between the embedding stable fixed points.

Bifurcation analysis We now turn to the question of the bifurcation analysis. Until now we have been engaged with problems of quantities and qualities of the fixed points and we have observed nonlinear phenomenons in the time-series. However, we are not aware of what causes the shift between one and three fixed points and in which ranges it occurs.

Bifurcations are common phenomenons in every day life. We use a two state light switch as example. Unlike the continuous control light switches, they have a sharp point of transition turning it 'on' or 'off', a bifurcation. Moreover, this point changes depending on whether one is going from 'on' to 'off' or reverse. This is a nonlinear phenomenon called hysteresis, which we will deal with later on.

The bifurcation theory deals with dynamic systems changing their behavior as some parameters of the systems are altered. Bifurcation theory tries to predict the system's behavior while crossing critical values in the parameter set, just like the point of transition in the light switch example. Bifurcations are accompanied by the shifting of stability of a fixed point. At a bifurcation point at least one eigenvalue of the Jacobian has a zero real part.

A *transcritical bifurcation* occurs, if in the combined space consisting of phase space and control parameter space, two different manifolds of fixed points cross each other, see figure 21. At the crossing point the fixed points exchange their stability properties, i.e. the unstable fixed point becomes stable and vice versa. The number of fixed points stays the same.

A *super-* or *subcritical bifurcation* takes place, if two fixed points with a broken symmetry bifurcate at once in a so called *pitchfork* or *double point bifurcation* as shown in figure 22. Both are either stable (supercritical pitchfork) or unstable (subcritical pitchfork).

The bifurcation diagram of a *Hopf bifurcation*, depicted in 24, looks similar to the diagram of a pitchfork bifurcation. However, passing the critical point, periodic solutions bifurcate. These solutions are, again, either sub- or supercritical. A Hopf bifurcation is furthermore characterized by a conjugate complex pair of eigenvalues crossing the boundary of stability. The real part of the pair becomes zero and a limit cycle forks at this critical point. A Hopf bifurcation accounts for the oscillatory solutions of the famous predator-prey

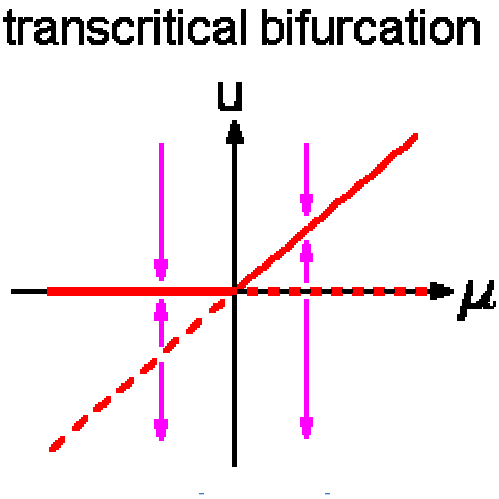


Figure 21: A transcritical bifurcation. A stable and an unstable fixed point change their stability properties at the critical point. U denotes the phase space variable and μ the bifurcation parameter.

model.

Analyzing the bifurcation PLOOP is undergoing, we find half a subcritical bifurcation causing hysteresis after the first shift of the net, as can be seen in 19. Whereas five fixed points are necessary to cause a subcritical pitchfork bifurcation, PLOOP is determined by only three, thus, half of the scenario is observed. Hence, a bistability region is created in parameter space inducing the switch-like behavior of PLOOP.
Once the second shift has occurred, PLOOP regains a single, stable stationarity. The network left the bistability regime and a different level of gene expression is reached.

In figure 23 we show the stationarity landscape of PLOOP with respect to two tunable parameters, i.e. the cooperativity factor K_D and translation. The odd nine parameters are tuned according to the specifications given in the figure. An intuitive impression of the bifurcation in parameter space is transmitted using calculated data. Originating in a stable section of the surface, the system passes a critical parameter value and flaps over, thereby creating a bistable region. A sketch of this scenario is given in 26. It is important to note however, that the respective figures show only a cutout of

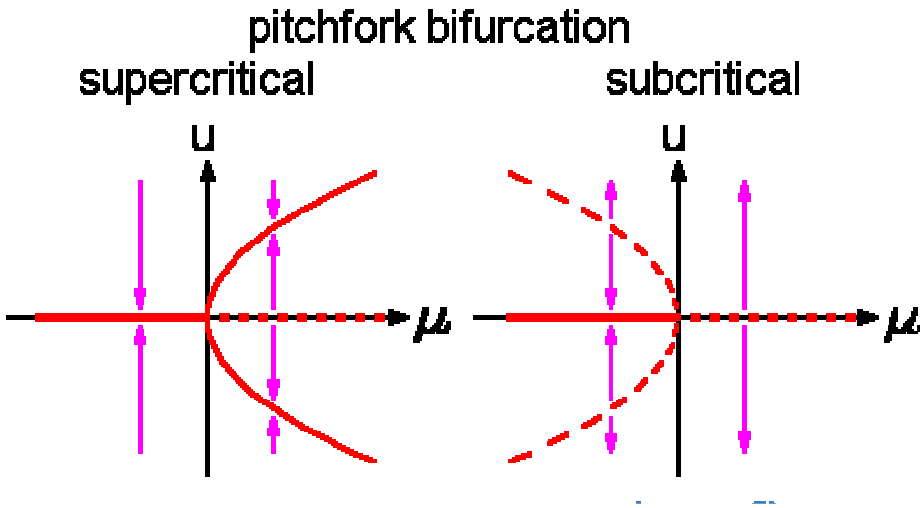


Figure 22: A supercritical and a subcritical pitchfork bifurcation. Supercritical: A stable point forks into an unstable and two embracing stable fixed points. Subcritical: A stable point and two unstable points collapse at the critical point forming one unstable equilibrium. U denotes the phase space variable and μ the bifurcation parameter.

the stationarity landscape.

Oscillations and Hopf bifurcation We were interested to know if PLOOP also exhibits oscillatory behavior in addition to its bistability. As has been described in 2.3.2, Hopf bifurcations cause such periodic results, see also 24. To gain insight into the question whether or not a Hopf bifurcation is theoretically possible, we used the following approach: The critical point of a Hopf bifurcation is characterized by a complex conjugate pair of eigenvalues $\lambda_{1,2}$ of the Jacobian J with real part equal to zero. Assuming that eigenvalues consist of complex numbers, we define $\lambda_{1,2} = \pm\alpha I$, with $\alpha \in \mathbb{R}^+$ and conclude: If a solution $\alpha \in \mathbb{R}^+$ to $p_J(\alpha) = det(J - \alpha IE) = 0$ can be found, the complex conjugate pair with zero real part is found, thus, the critical point for a Hopf bifurcation is determined. We derive the characteristic

Steady state of manifold x in parameter space – 360 deg. panorama

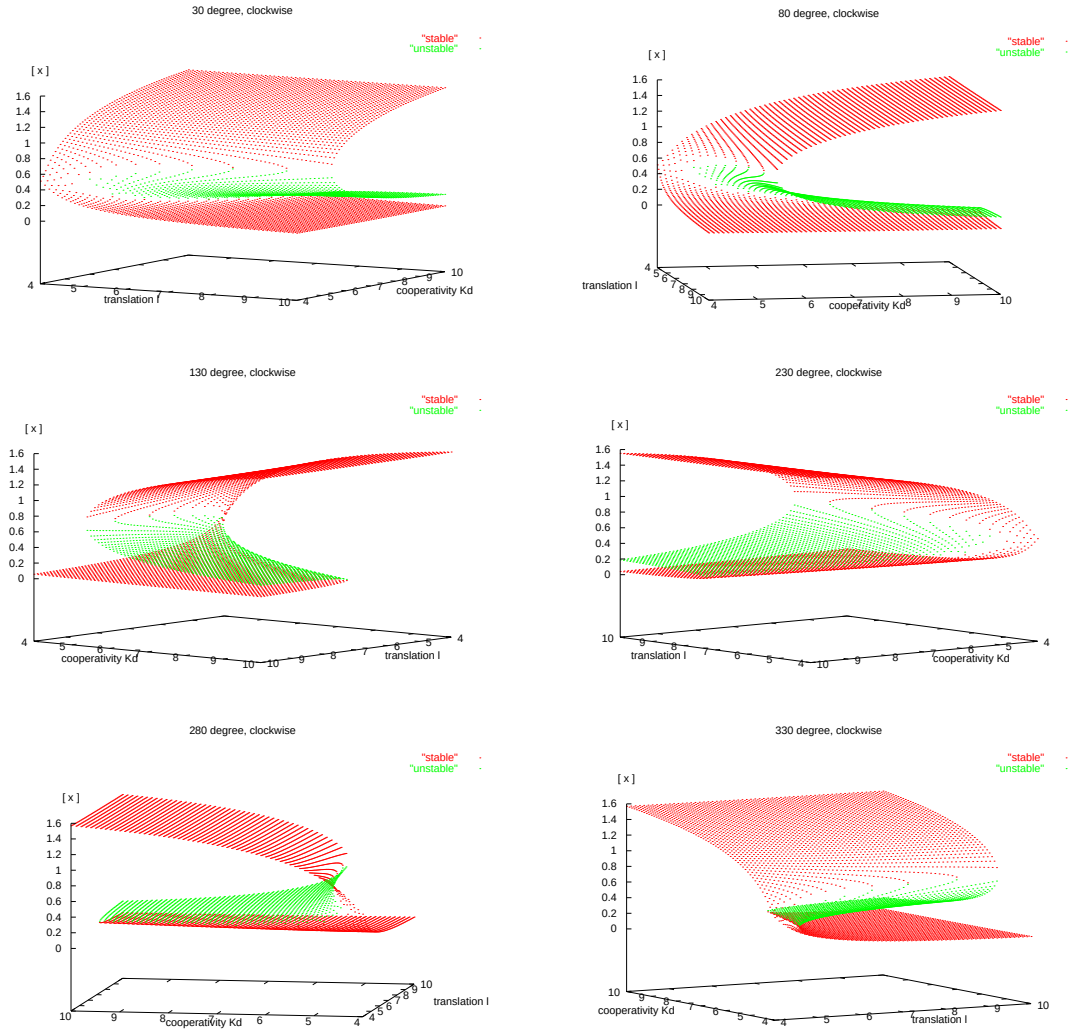


Figure 23: Steady state of manifold $\overline{x}$. The following parameter settings were used: $(a_0, b_0, d, dp, K_M) = 1$, translation and cooperativity were chosen flexibly $K_D = l = [4, 10]$. This cutout of parameter space represents the progression from stable to bistable equilibrium.

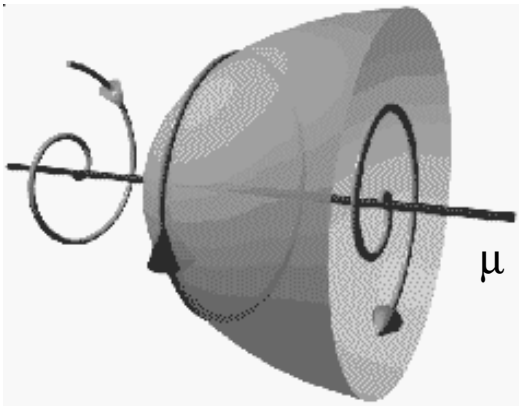


Figure 24: A supercritical Hopf bifurcation. At the critical point periodic solutions fork along the bifurcation axis. μ denoted the bifurcation parameter.

polynomial p_J for α_i and obtain

$$\alpha_{1,2,3,4} \;=\; \frac{1}{2}i\Big(\pm(dp+d)+\sqrt{(dp-d)^2\pm 4l\sqrt{DG}}\Big)\;.$$

The expression within brackets can never adopt a purely imaginary value. From this result we deduce straightforwardly, that a positive, real solution can not be found for α_i within the positive, real parameter space. Accordingly, PLOOP can neither show a Hopf bifurcation nor exhibit oscillatory behavior.

Hysteresis and predictability Hysteresis is the history dependency of the fixed points under an external strain, the change of the bifurcation parameter. Hysteresis yields a 's-shaped' graph as demonstrated in figure 19. As has been described in section 2.3.2, the figure shows the change in quantity and quality of the fixed points. Once the area of bistability is entered, the system has to deal with three fixed points, three valid solutions of the system. The outer two are stable, whereas the inner one is unstable. Depending on the initial conditions with respect to the unstable manifolds of the saddle, the system approaches stable fixed point one or two.

PLOOP is only fully described by four states variables, accordingly, the fixed points are four dimensional, too. If split up, the single dimensions show roughly the same graph, however, two of them start with high values for the

single fixed point region ending in a lower one and vice versa. Furthermore, the fixed point concentrations of the four states variables vary considerably. Data are not shown with exception of $\overline{x}$ depicted in figure 19.

In order to cover the question of predictability, we will be forced to involve all four dimensions in an illustration, as can be found in figure 25. The scheme depicts the course of fixed points along the bifurcation axis in all 4 dimensions. Stable and unstable courses are marked. The parameter set under investigation consists of $(a_0, b_0, dp_{a,b}, d_{a,b}, K_{MA,MB}) = 1$, $(t_1, t_4) = 2$, $l_{a,b} = 6$ and the bifurcation parameter $\Gamma \equiv K_{DA,DB} = [4.5, 6.7]$. Considering the

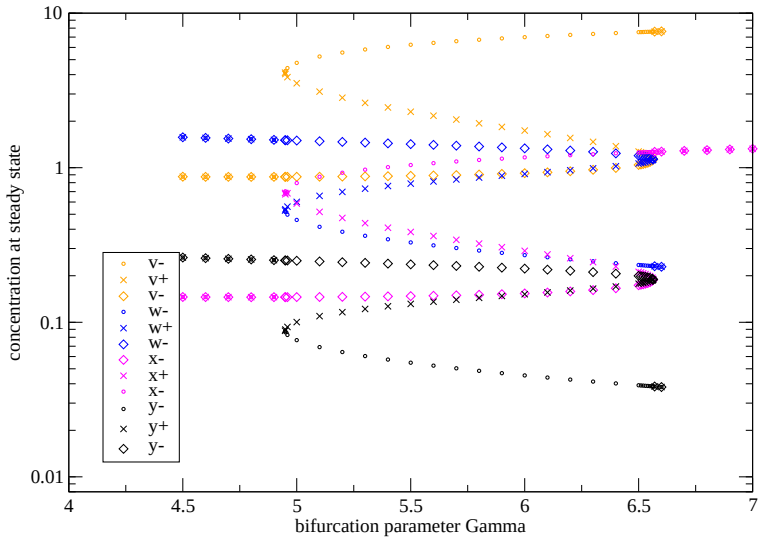


Figure 25: Steady state concentrations of PLOOP. The figure denotes all four dimensions of the fixed points along the bifurcation parameter axis. The following parameter settings were used: $(a_0, b_0, d, dp, K_M) = 1$, $t_1 = t_4 = 2$, $l = 6$ and $\Gamma = K_D$. The plot clarifies the interaction of the attracting versus the repelling manifolds in a bistability area. For enhanced perceivability a log scale was used for the y axis. For comparison see figure 19.

nearest stable manifold of each dimension, the prediction of the initial conditions' choice of attractor is possible. In case of doubt, the strengths of the eigenvalues give further information about the system's tendency. As a rule of thumb, if more than two dimensions of the initial conditions tend toward the same fixed point, the system will most probably end in the respective stationarity.

Sampling the parameter space - a stochastic approach We are now turning to the question of the distribution of bistability and stability in parameter space. In order to gain a quantitative impression of the frequency of bistability given a random combination of parameters, we used the *Mathematica* software package. We analyzed over 200000 random combinations of the nine parameters $(a_0, b_0, t_1, t_4, dp, d, l, K_M, K_D)$ in the range of $r_p = [0.1, 100]$. Note that the random numbers provided by *Mathematica* are uniformly distributed pseudo random numbers. We obtained 37649 cases with three solutions to the fixed point problem. This result equals to $Prob = 0.188$ or approximately 20% to encounter a bistability in parameter space. This result reflects bistability as a rather frequent event.

Parameter influence analysis It is useful to investigate the influence of the particular parameters on the dynamics of a system in detail. Nonetheless, in a system which is defined by stability and bistability, the scope of changes in parameter space will be restricted to the onset of the dynamics. Once the attractor is found no further changes are expected, thus potential influences cease. However, the robustness against small changes in the parameter set of known dynamic behavior, is probably a better measure with respect to PLOOP. The manual analysis of several settings causing bistability shows more highly robust combinations than labile ones. Considering the type of bifurcation PLOOP shows, we conclude that the robustness of specific settings depends on the distance from the critical point. As has been shown in figure 26, the region of bistability originates in the critical bifurcation point and enlarges in the form of a 2-dimensional funnel. Based on these results, we deduce that robustness of the parameter combinations increases with the distance from this point and the probability of being positioned somewhere within the bistability funnel in parameter space.

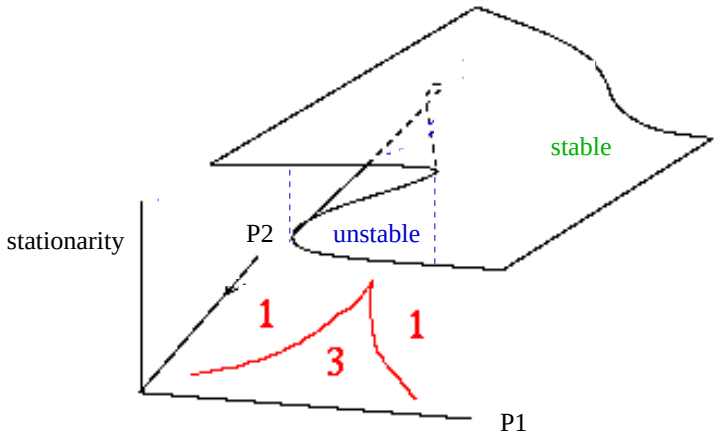


Figure 26: Sketch of PLOOP's stationarity landscape. The landscape is stable until it reaches a critical point at which it flaps over and forms a bistability region. This bistability section is projected onto a v-like cusp. The amount of fixed points with respect to the cusp is depicted. Hysteresis is drawn in.

Principal components analysis The concerted action of the parameter settings and the underlying equations determine the dynamics of our system. As has been shown earlier, the answer to how many fixed points are conceivable, lies in the solution of a polynomial of degree seven, which can not be estimated intuitively. We were trying to gain a simplification of that question, such that interdependencies of the parameter might be revealed and a prediction becomes easily possible.

In order to get a graspable idea of the dependencies, that result in stability or in bistability, we find that the large number of parameters cause the major problem for the simplification of the relevant expressions. Considering the PLOOP system we are dealing with fourteen parameters at most. Empirically we can cut them down to nine i.e. a_0, $t1$, $b0$, $t2$ plus the rest, which exhibits similar values for the A and B half-cycle, i.e. d, dp, l, K_M, K_D. Still a whole lot to deal with.

From the field of three-dimensional molecular structure prediction, we learn how to analyze high dimensional systems with respect to their 'relevant' dimensions, in order to reduce the complexity of the problem. The correlation between coordinates, in our case parameters, can be used to re-

duce the number of relevant parameters, that are necessary to describe the dynamics, such that if one parameter is known the others can be deduced from it. In other words, we want to find characteristics in the dependencies of the parameters that cause bifurcation or stability.

We will now describe a mathematical method for determining such correlations. In order to describe the system in question in the reduced, relevant dimensions, we first have to find out, which they are. We assume we have a set of n dimensional data vectors of type $\vec{d_s}$ containing the parameters $p_i(s)$, $i = 1 \rightarrow n$ and $s = 0 \rightarrow$ set size:

$$\vec{\mathbf{d}}(\mathbf{s}) = \begin{pmatrix} p_1(s) \\ p_2(s) \\ . \\ . \\ . \\ p_n(s) \end{pmatrix}$$

If we have s vectors of n dimensions we can use them to construct a $n \times n$ correlation or covariance matrix C. The entry at the i^{th} row and j^{th} column of C is given by

$$C_{ij} \;=\; < (p_i - < p_i >)(p_j - < p_j >) > \;\;,$$

where $<>$ indicates averages. If the covariance matrix C is diagonalized, the eigenvalues λ_i are obtained. Each λ_i has an Eigenvector $\vec{\lambda_i}$ associated with it.

The eigenvalues λ_i are in fact the mean square displacements (msd) of p_i in the direction of the corresponding Eigenvector $\vec{\lambda_i}$. It is now possible to transform the data-vector $\vec{d_s}$ of the Cartesian vector-system into a data-point d_s of the Eigenvector-system by computing the inner product:

$$d_s \;=\; \vec{d_s} \cdot \vec{\lambda_i} \tag{4}$$

In the case of high interdependencies of the parameters, we find a great numerical difference between the calculated eigenvalues, hence in the msd of the parameters. The highest numerical values indicate highest priority for

the dynamics, the correlated Eigenvectors can be used to form a new set of basis vectors. The data vectors are transformed and reduced. Thus, it is possible to depict the data vectors in a reduced form in the parameter dimensions most relevant for the system.

Our data consist of two data vector sets of $s = 37500$, each, one accounting for bistability the other for stability. The parameter vectors include $i = 9$ dimensions, accordingly, we build a 9×9 covariance matrix and calculate the eigenvalues for each set independently and the combined data sets. We practically do not find a numerical difference for the set having one fixed point, whereas a higher interdependence for the parameter settings causing bistability is observed. From the combination of the two data sets we undertake a transformation of the data vectors to data points according to equation 4 and plot them in the vector space of the first principal component. In figure 27 right the distribution of the data points accounting for one and in 27 left for three fixed points is shown. The histogram of the distribution of the obtained data is shown. We observe roughly an equipartition of values in parameter space causing stability, whereas in the case of bistability a paraboloid dis-

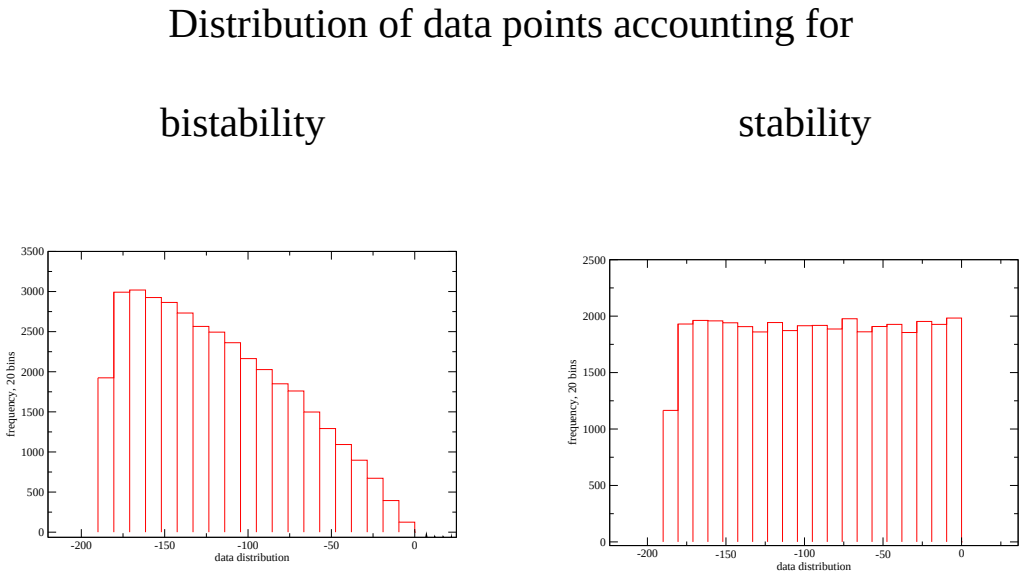


Figure 27: Principal component analysis of $2 \cdot 10^5$ random parameter settings. The figures show the distribution of parameter vectors transformed to data points in the vector space of the first principal component.

tribution of values is calculated. We interpret a certain set of restrictions in parameter space that reduces the number of combinations accounting for bistability. It is not possible, however, to explicitly name them in the context of this analysis. Regardlessly we find a characteristic quality to predict the behavior of PLOOP in statistical terms given a specific parameter set.

It can generally be concluded, that parameter combinations causing stability of PLOOP are equally distributed over the parameter space, such that account for bistability are not. The interdependencies of the parameter comport accordingly, in terms of reducibility of dimensions. If the parameter sets are equally distributed, the single parameter values are, presumably, too. This is mirrored by the eigenvalues of the covariance matrix. The opposite holds for the bistability case.

3 Evolution *in silico*

3.1 Implementation of the Program 'RegNet'

'RegNet' is a program for the simulation of evolutionary adaptation of genetic regulatory networks. The program is designed in the style of the *in silico* flow reactor developed by [23], described in greater detail in section 1.6.4. In the following sections 3.1.1 - 3.1.8 we describe the theory and the rules implemented in 'RegNet' followed by the numerical results obtained by respective simulations in section 4.

3.1.1 Reaction Rates and ODEs

The formal description of the single network module and its automation form the bases for a comparatively simple handling during the simulation of network evolution. The representation of the network by chemical reactions, their reaction rates and involved reactants allows the deduction of a set of ODEs and the numerical evaluation of the dynamic behavior.

To allow a set of ODEs to be constructed, firstly the set of chemical reactions has to be determined. We use the network architecture of the bistable gene switch PLOOP, described in 2.3, as draft. This means we are dealing with a set of reactions consisting of transcription, translation, association, dissociation and decay. The involved molecules stem from the population in the reactor, the reaction rates have yet to be determined.

Naturally, the potential difference between the educt and the product species accounts for the production rate of a reaction. This concept goes beyond the scope of our model. We deduce the reaction rates additively from various properties of the educt molecules as tabulated in 1. For each type of reaction a special key has been developed, a schematic is shown in table 2: The transcription rate is determined by the gene's characteristically qualities. The individual under evaluation has to live up to a predefined sequence pattern as well as to a structural motif. The degree of accomplishment is translated into the reaction rate. The detailed evaluation of the molecule individuals is described in section 3.1.3.

The translation rate and the rate for the transcript decay is given by kinetic properties of the mRNA under evaluation. The idea is that the degree of accessibility of its shape will determine at which rate the transcript is processed. Note, however, the numerical harmonization of the rates for both

reaction type	example	rate expression	reaction rate
transcription	$G_1 \xrightarrow{t_1} G_1 + T_1$	$[G_1]t_1$	G_1
	$G_2 \bullet PD_1 \xrightarrow{t_2} G_2 \bullet PD_1 + T_2$	$[G_2 \bullet PD_1]t_2$	G_2
translation	$T_1 \xrightarrow{l_1} T_1 + P_1$	$[T_1]l_1$	T_1
association	$G_2 + P_1 \xrightarrow{a_1} G_2 \bullet P_1$	$[G_2][P_1]a_1$	G_2, P_1
	$G_2 \bullet P_1 + P_1 \xrightarrow{a_2} G_2 \bullet PD_1$	$[G_2 \bullet P_1][P_1]a_2$	fixed value
dissociation	$G_2 \bullet P_1 \xrightarrow{c_1} G_2 + P_1$	$[G_2 \bullet P_1]c_1$	fixed value
	$G_2 \bullet PD_1 \xrightarrow{c_2} G_2 \bullet P_1 + P_1$	$[G_2 \bullet PD_1]c_2$	fixed value
decay	$T_1 \xrightarrow{d_1} 0$	$[T_1]d_1$	T_1
	$P_1 \xrightarrow{dp_1} 0$	$[P_1]dp_1$	P_1

Table 1: The five different types of reactions used in 'RegNet' are shown. The representation consists of the reaction type, an example, its rate expression and molecules consulted for the assignment of the reaction rate. • denotes the complexation between the respective individuals.

reaction types.

Coupled association and dissociation reactions of proteins and DNA are assumed to occur quickly. They are at equilibrium. We therefor construct an equilibriums constant for these bidirectional reactions. The forward reaction rate is formed by adding the quality values of the protein molecule and the gene. For the backward reaction a fixed value is introduced for reasons of simplification. This saves us from an explicit evaluation of the complexed molecule species. PLOOP furthermore assumes cooperative binding of the proteins to the DNA. This fact implies that even though the binding rate of the first monomer is to be calculated freshly with every new 'first-level' association, the gain of binding affinity for the second monomer can be fixed to a certain, higher value.

Once the reactions are formed completely, the set of ODEs can be deduced. The expression for each species is modularly build from the rate expressions of each reaction, where the individual is involved. The rate expression rex of a reaction r describes the probability of some educts to react forming the products of a reaction. It is a function of its educts' concentrations and the corresponding reaction rate. The reaction expression of an example reaction $r : 2A + B \xrightarrow{k_1} C$ writes as $rex : [A]^2[B]k_1$. For the con-

mol. species	sequence, structure
G_1	C U G U C C C G C C C G U G G G G U
mfe	. . . (((((. . . .))))) .
G_2	G G C U G C G A C C A G U C G U C U
mfe	(((((. . . .))))) . . . .
T_1	A C C C C A C G G G C G G G A C A G
mfe	. . (((. . . . . .))) . . . .
P_1	T P R A G Q
target structure	

Table 2: Examples for the reactant species are given. The construction of a quality measure includes sequence and structure motifs (boxes) for genes, a kinetic criterion for the transcript's mfe structure and a structure comparison for proteins.

struction of the ODE all reactions have to be scanned for involvement of species X. The respective rate expressions are added or subtracted depending on which side of the reaction species X is found, to form the differential equation.

3.1.2 The Genotype-Phenotype Mapping Problem

In silico evolution is naturally built to follow Darwin's principles of variation and selection. Molecules are reproduced by erroneous replication having variation as a result. Because of fitness dependent selection, fitter individuals have a greater probability to reproduce and pass their properties to their descendants. The assignment of the specific fitness values is far from trivial, though.

With regard to the fitness evaluation, the first basic problem is the determination of a phenotype for each genotype. In the case of RNA molecules, the 'traditional' object of *in silico* evolution, the sequence constitutes the genotype, whereas the shape of its secondary structure represents the phenotype. To succeed in this genotype-phenotype mapping, firstly the properties of the geno- as well as the phenotype space have to be clarified.

The genotype space $\mathcal{I}$ of sequences is structured by a natural metric. The so called Hamming distance d_{ij}^h between the sequences i and j [34] provides this metric, which specifies the distance between two individuals by the number of differing positions in their sequence. That way a net of single mutant

steps can be drawn over sequence space. The phenotype space $\mathcal{S}$ can also be assumed as a metric space by defining a distance measure d^s_{ij}, even though a simple measure is far less natural. The annotation of distance doesn't reflect the accessibility through Darwinian evolution, which is based on mutations on sequence level.

Caused by the dissimilarity of the distance measures in both spaces, the mapping of the genotype onto the phenotype space is highly complex and can not be expressed in analytical terms.

$$\psi : \{\mathcal{I}; d^h_{ij}\} \Rightarrow \{\mathcal{S}; d^s_{ij}\}$$

However, there exists an algorithm, that assigns a phenotype S_k to every genotype I_k. Once a corresponding phenotype has been determined, it is evaluated and a non negative, real fitness value is assigned. The evaluation procedure relies on a fitness function which is based on phenotype inherent properties in such a way, that a simple distance measure can be applied to determine the distance of an individual to a predefined target property.

The objects of selection in our simulations are genetic regulatory network modules. The mapping onto a fitness value is even more complex, because of the multi-layered structure of dependent parameter spaces as can be seen in figure 28. The degeneration of the 'mapping states' in single spaces is caused by the fact that multiple networks map on one fitness value, multiple networks involve a certain individual molecule and multiple transcript sequences encode the same protein.

The network space in figure 28 shows the greatest reduction of 'mapping lines' which represents the greatest degree of degeneration of complexity while trying to map onto fitness values. The prior augmentation of 'mapping lines' also means that network space joins the ancestor spaces to sub-spaces in a super-space of highest complexity. Changes in the molecule spaces, genes, transcripts and proteins, result in another point in network space, hence produce different networks. The erroneous replication of a network as used in our simulations, is thus based on changes in either gene and transcript spaces or in all three of them, gene, transcript and protein space. The properties of the single constituents of the network provide the reaction parameters, which cause the network's dynamic behavior. The dynamic is evaluated and added while mapping on a fitness value. This is why the dynamic behavior is drafted as a filter-like object in figure 28.

The multi-layered architecture of the mapping process and the varying degrees of complexity impede the intuitive, analytical representation of a

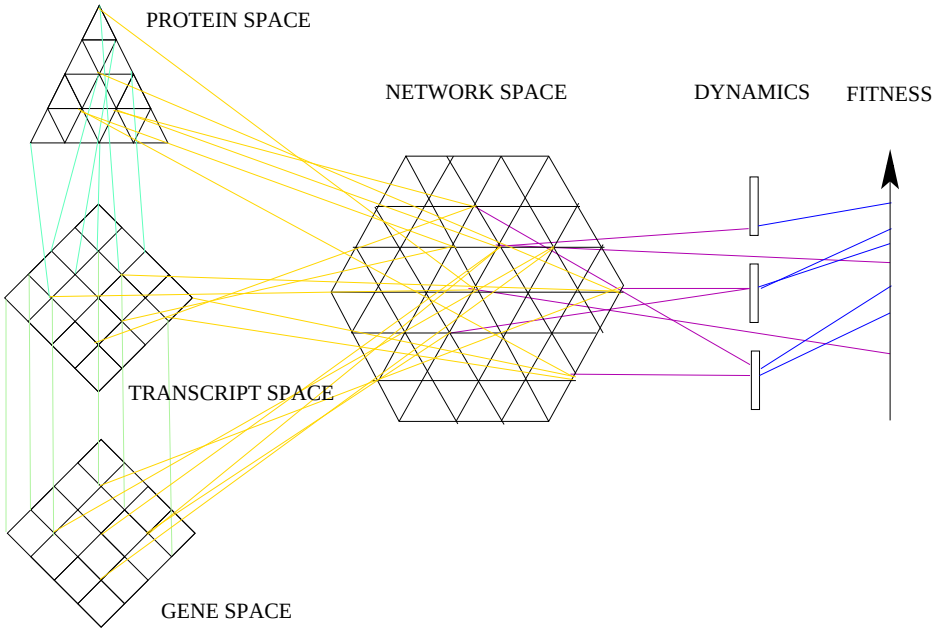


Figure 28: Mapping genetic networks onto fitness values. The schema depicts the mapping between and onto different spaces. Genes transcribe to transcripts, which are translated to proteins. All three species form genetic networks, which based on the interacting individuals' properties, show certain types of dynamic behavior. The dynamic adds to the networks' qualities when mapping onto a fitness value and is hence depicted as a filter-like object. The degree of complexity increases with the number of mappings. Each mapping onto a point in the following space includes the properties of the ancestor individuals. Based on [63].

mapping function. Yet the feasibility of the determination of a fitness value for each network allows an optimization process. In this way the simulation of molecular evolution is successful and the mean fitness increase will lead the path to a possible target.

3.1.3 Quality Evaluation of the Network Players

We developed an elaborate procedure to evaluate the different molecule species forming the gene network. Genes, transcripts and proteins fulfill distinct functions in a gene regulatory network and hence, provide various possible starting points for a quality evaluation. The exact criteria used are to be described in the following paragraph.

Genes are physically involved in three types of reactions in our model system: transcription, association and dissociation reactions with a protein. Firstly, we define a property decisive for the gene's general transcription 'potential'. We chose a sequence motif, the GCC box, upstream of the gene sequence. GCC boxes are eukaryotic enhancer elements that, among others, facilitate the transcription process. The extent to which the sequence criterion is fulfilled, is translated into a value for the 'quality' of the gene. Next, we rate the gene with respect to association and dissociation. The model assumes that the complex-forming proteins act like transcription factors, thereby regulating the gene's transcriptional activity. We therefor use a criterion that involves the structural circumstances in the promoter region. The accessibility of this section and its adjustability is investigated by the program while checking for a defined opened or closed structure. The 'quality values' ascertained are regarded independently for the respective reactions during the experiment.

According to the PLOOP architecture, transcripts engage in translation and in decay reactions. We decided to evaluate the mfe structure of the transcript with respect to its accessibility for other mediators involved in these reactions. Highly packed structures, for example, impede a quick decay due to their inaccessibility for RNAses, on the other hand they also complicate the translation process. The program positively selects for molecules with mfe structures in a certain energy range. The energy range is sequence length dependent and favors molecules with intermediate structure stability. Here, structure stability is an indirect measure for the tendency to open or maintain its shape, to facilitate or inhibit reactivity.

Last, but not least proteins come into play. Proteins are involved in association, dissociation and degradation processes. According to the assumption of the model, that proteins exhibit a cooperative association behavior when complexing to the promoter sections of a gene, we are actually dealing with two different levels of association-dissociation reactions: The first step is the binding of the protein monomer to the DNA molecule, thereby changing the structure of the target slightly to facilitate the binding of the second monomer. The formation of the protein dimer is the second level of association. Dissociation reactions run parallel, but in reverse order.

While it is difficult to find a parameter for the first monomer to bind to the target, the rate for its dimerization is obvious. As described in section 3.1.1, we fix the dimerization rate to a value higher than the maximal monomer association rate, thereby providing cooperative behavior. Thus,

the evaluation of the monomer is simply dispensed. In order to determine the monomer association rate, on the other hand, we need to find a protein specific criterion. The class of DNA binding proteins exhibit recurring motifs for the functional domains. The subclass of zinc finger proteins is thoroughly examined and documented. We use the structure of the DNA binding domain as a draft for the evaluation of the protein factors. The precise paradigm is the GAGA factor of *Drosophila*, pdb identifier 1YUJ. Details on the role of zinc finger proteins and the gaga factor can be found in section 1.4.2.

In order to assess a nearness measure for any protein sequence of the simulation, we implemented a z-score evaluation routine. The z-score forms the basis for a comparison of a given protein sequence and a predefined protein shape. For a profound explanation of this topic look up chapter 1.4.1. All novel protein sequences are tested for their specific probability to fold into the target structure. A value for doing so is deduced from the probability distribution generated on the basis of an artificial polyprotein to fold into the given target. In other words, a sequence data set is evaluated with respect to its structure distribution compared with the target shape and the protein of question is ranked within this distribution. In this framework a value for nearness can be calculated and assigned. Note that the nearness of the target to its shape is maximal, hence we normalize to one.

We do not introduce a specific criterion to contribute to the decay rate of the protein, but reuse the calculated structure value, mostly because of the computationally costly z-score evaluation.

3.1.4 Dynamic Behavior and Noise

For the functionality of a regulatory network the actual architecture is less important than its dynamic behavior. More than elsewhere the end justifies the means. In a context of feedback inhibition, for example, it is important that the network of involved proteins manages to silence its own production, equally how many straight steps or web-like interactions are used. The evolution of network architecture is obviously driven by a differently motivated type of evolutionary pressure. To account increasingly for the functionality of a regulatory network, we extend the fitness evaluation from purely architectural to dynamic aspects.

Dynamic behavior The dynamics of interacting reactions are described by their set of ODEs. The solution of the numerical integration of the ODEs

represents the changes in concentration flows of the involved species. The space of possible dynamic solutions in the parameter space P must be imagined as n-dimensional with n being the number of the parameters used. It consists of two exclusive subspaces S and B such that,

$$\begin{aligned} S &\subset P \\ B &\subset P \\ S &\cup B = P \\ S &\cap B = \{\} \end{aligned}$$

and S being the stable and B the bistable solutions. Within the framework of the model two further types of dynamic solutions are introduced, these are noise-dependent bistable NB and double bistable individuals DB. They relate to solution space P according to

$$NB = \{x \mid x \in S, f_N(x) \in B\} \quad (5)$$

$$DB = \{x \mid x \in B, f_N(x) \in B\} \ , \quad (6)$$

with x being a solution in P and $f_N(x)$ is noise applied to the network. The reaction parameters used in the ODEs are deduced from the underlying quality determination of the network players. They are predefined in such a way that a network accomplishing all tasks automatically exhibits bistability. The dynamics of less fit networks naturally depend on their site in solution space P or in other words on the combination of parameters and the topology of the fixed points' landscape. The program distinguishes between the two possible types of behavior: stable and bistable, with the latter dynamics being the target behavior.

For the evaluation of the network behavior in our computer experiments, 'RegNet' calls the numerical solver CVODE, available among others at *robotics.stanford.edu/users/scohen/bio.html* and explicitly calculates the dynamics for every new individual.

Noise Natural systems are prone to noise. Recent studies [19, 35, 76] concerning this topic teach us, that natural systems are either noise tolerant or flexible enough to 'domesticate' it for their own good. Both solutions are highly interesting with respect to gene networks. How much do they

tell us about the equilibrium landscapes these systems live in? In other words if we apply a simplified perception, is an exclusively stable system always noise tolerant and do systems with higher order dynamics have to evolve either 'around' noise or domesticate it? Up to which extend does noise influence genetic networks and thus, influence the evolution of novel types of architecture? Insights to these questions would be rather exciting, so far we confine ourselves to complete our evolution simulations with the phenomenon noise.

Since transcription and translation both are reactions that really consist of a series of reactions involving quite a number of implicit co-players, we anticipate that these reactions are most frequently subject to noise [19]. Thus, we implement stochastic fluctuations in the parameters of these reactions according to $f_N(x)$ used in equations 5 and 6. 'RegNet' provides two different sources for noisy parameters. These are white noise and a wavy field of noise. The values, which are added to the genuine network parameters in a multiplicative way, are either equally distributed or based on a sinus function, respectively. The amplitude and hence the amount of noise applied to the networks is tunable. The behavior of the noisy network is determined by a second integration run.

3.1.5 Fitness Evaluation of the Network

The evaluation of network players, the determination of the reaction parameters, the construction of the network and its dynamics - all of these efforts focus in one point, the transformation into an individual fitness value for each net.

The fitness value of a network is constructed from two components: the quality of the involved players translated into specific reaction rates $F_M(R)$ and a value representing its dynamics $F_D(R)$ with R being the set of all reactions r of the network. $F_M(R)$ is the normalized sum over all $F^R(r)$, the fitness of each reaction.

$$F_M(R) := \frac{1}{|R|} \Sigma_{r \in R} F^R(r)$$

The fitness value for each reaction $F^R(r)$ is build from the normalized sum over all educt fitness values $f(e)$ of the reaction r.

$$F^R(r) := \frac{1}{|E(r)|}\Sigma_{e\in E(r)} f(e)$$

The single educts e belong to the set of educts $E(r)$, which is itself a subset of the list of molecules M with size k. The set of educts E is set of all possible subsets 2^M of M.

$$\begin{aligned} E &: R \to S^M \\ E(r) &= \{m_1, .., m_k\} \subset M \end{aligned}$$

The introduced sets are summarized as follows,

$$\begin{aligned} M &.. \text{ set of all molecules} \\ R &.. \text{ set of all reactions} \\ 2^M &.. \text{ set of all subsets of M} \\ E(r) &.. \text{ set of all educts of reaction r .} \end{aligned}$$

In order to determine a specific fitness for the educts we split them in three classes, genes f_G, transcripts f_T and proteins f_P.

$$f(e) := \begin{cases} f_G & \text{if } e = gene \\ f_T & \text{if } e = transcript \\ f_P & \text{if } e = protein \end{cases}$$

Each of the classes fulfill different criteria for the evaluation of their fitness. Genes have to cover a sequence motif and exhibit a certain substructure, the accomplishment is measured gradually above a minimum min. Both values f_{GS} and f_{GX} are added and normalized to one.

$$\begin{aligned} f_G &= (f_{GS} + f_{GX}) \cdot \frac{1}{2} \\ f_{GS} &= f_{GX} = \frac{\text{motif}_{test} - \min}{\text{motif}_{target} - \min} \end{aligned}$$

Proteins are compared to a target structure. The z-score zs_i of both shapes is compared.

$$f_P = zs_{test}/zs_{target}$$

The energy of the mfe structure of the transcript has to locate within a given range around the average mfe in the ensemble.

$$\begin{aligned} f_T &= \begin{cases} 0 & \text{if } \mathrm{mfe}_{test} > z \\ \frac{\mathrm{mfe}_{test}-\mathrm{min}}{\mathrm{mfe}_{target}-\mathrm{min}} & \text{otherwise} \end{cases} \\ z &= < \mathrm{mfe}_{ensemble} > \cdot(1+0.4) \end{aligned}$$

The calculation of $F_D(R)$ is less complex. If the system exhibits two stable steady states $ss(R)$ causing bistable behavior it is rewarded, any other dynamics are ignored.

$$F_D(R) := \begin{cases} 1 & \text{if } ss(R) \geq 2 \\ 0 & \text{otherwise} \end{cases}$$

The molecular part and the dynamic part can be variously weighted, subsequently they are combined and normalized to one, such that:

$$\begin{aligned} F(R) &:= (\alpha_D F_D(R) + \alpha_M F_M(R)) \cdot \frac{1}{2} \\ \alpha_M &+ \alpha_D = 1 \end{aligned}$$

with α_i being the weights for the different objectives. The evolution simulation as carried out with 'RegNet' represents a maximization of $F(R)$.

3.1.6 The Replication-Selection Procedure

The flow reactor simulations in this work are based on the algorithm developed by Daniel Gillespie [27, 28]. At the beginning of an experiment the reactor contains a start population of genes, transcripts and proteins, grouped together in interacting circuitries. The goal of the simulation is to find a

possible solution for a target network, whereby the target's architecture is predefined. The interacting individuals, as well as the dynamic of the network have to evolve and fulfill certain fitness criteria to be accepted as a possible target solution. The evolution is realized by the application of point mutations as variation's driving force and a fitness-dependent replication and decay procedure.

The first step in this circular simulation is the determination, which step to take next, i.e. replication or decay. The decision depends on the overall activity $A(t)$ of the reactor, a measure which implicitly contains the fitness f_i of each individual i and the size $N(t)$ of the current reactor population. It writes as

$$A(t) \;=\; R^{rep}(t) + R^{out}(t) = R^{rep}(t) \cdot \big(1 + \frac{N(t)}{N_{set}}\big) \; .$$

A random number rn of the interval $]0, A(t)]$ is drawn and the decision is made according to $rn \cdot A(t) \leq R^{out}(t)$.

The next step is the selection of the prey individual, which will undergo replication or decay, from the entire population. The selection procedure for replication is different from the procedure for outflow. The replication rate r_i^{rep} of each individual is a function of the respective fitness f_i and the size of the corresponding subpopulation n_i.

$$\begin{aligned} r_i^{rep} &= f_i \cdot n_i \\ R^{rep}(t) &= \Sigma_{i=1}^{k} r_i^{rep} \end{aligned}$$

The sum over all replication rates is called R^{rep}.
The outflow rate of each individual is given by

$$r_i^{out} \;=\; r_i^{rep} \cdot \frac{n_i}{N_{set}} \; .$$

Given that the current population size $N(t) = \Sigma_{i=1}^{k} n_i$ and $R^{out}(t) = \Sigma_{i=1}^{k} r_i^{out}$, we can write

$$\begin{aligned} R^{out}(t) &= \frac{N(t)}{N_{set}} \cdot \Sigma_{i=1}^{k} r_i^{rep} \\ R^{out}(t) &= \frac{N(t)}{N_{set}} \cdot R^{rep}(t) \; . \end{aligned}$$

As can be seen from $R^{out}(t)$ the population size is meant to fluctuate around a predefined value $N_{set}(t)$ with a standard deviation of $\sigma = \sqrt{N_{set}}$.

According to the expression for $A(t)$, we find that if $N(t)$ is larger or smaller than N_{set} decay or replication is favored respectively. If at a time t $N_{set} = N(t)$ the probability for replication or outflow is equal.

Eventually the algorithm provides also an internal clock to determine the current time $t_n = t_{n-1} + \Delta t$ as a function of $A(t)$. We write

$$\Delta t \;=\; \frac{log(1/rn)}{A(t)} \;,$$

with rn being a second random number out of $]0, 1]$. During the initial phase of a simulation, when the reactivity and fitness values are relatively small, time moves faster, whereas in a reactor with either a higher population size or at the end of a run time elapses much slower. This assures that time passes relatively similar with respect to the probability to gain an improving mutation.

3.1.7 The Neutral and the Nearly Neutral Theory

The neutral theory of molecular evolution was introduced with provocative effect in the late 1960s by Motoo Kimura [41]. Although the theory was received by some as an argument against Darwin's theory of evolution by natural selection, Kimura and most evolutionary biologists today persist that the two theories are compatible.

The neutral theory attributes a large role in evolution to genetic drift. A certain fraction of new mutations in a population is free of constraint, thus they are neither subject to nor explicable by natural selection, they are selectively neutral. The rest of the possible mutations have deleterious effects and are eliminated from the population. Accordingly, Kimura assumes that the motor of molecular evolution is the random fixations of neutral mutations in the population rather than natural selection fixing advantageous mutations. A second assertion is, that most evolutionary change is the result of genetic drift acting on neutral alleles. These novel alleles drift through the population and eventually perish or in rare cases become fixed in the population. In this way neutral mutations tend to accumulate and genomes tend to evolve.

The molecular clock hypothesis is based on the neutral theory of molecular evolution, but is nowadays preferentially used for null hypothesis testing. For details see section 1.6.1.

The nearly neutral theory of molecular evolution is Tomoko Ohta's [54] extension of the neutral theory. It enlarges the theory in such a way that both minor deleterious and advantageous mutations are allowed to occur. Ohta postulates that random genetic drift, as well as selection both influence the behavior of very weakly selected mutations, with drift predominating in small populations and selection in large ones. Such mutations are selected against in large populations, but behave as if neutral in small ones. They are called nearly neutral mutations and have a negative correlation between evolutionary rate and population size.

In the simulations of evolution presented in context of this work, the selection procedure was carried out according to the nearly neutral theory. We reason that in a target oriented process of molecular adaptation the usage of Darwin's 'pure' principles would only result in an augmentation of simulation length, but not in the generation of novel evolution patterns. The disapproval of deleterious mutations on the contrary, might force the evolving system into local optima without the chance to recover from there ever after. The nearly neutral theory, finally, provides our simulations with bits and parts from all extremes.

3.1.8 Network Acceptance Procedure

In section 3.1.1 - 3.1.7 we reviewed the fitness evaluation procedure for PLOOP-like networks used in our evolution simulations. We anticipated well-formed networks with correctly chosen molecular species. Yet, the formation of these modules deserves a closer look.

During the evolution process of the population, candidates for replication are selected according to Gillespie. Replication is erroneous and hence the offspring looks often different from its parent. The sources for variation are manifold though. For the simulations presented in the following chapter we used gene duplication and subsequent point mutation with a probability $p = 0.01$. For alternative scenarios, we implemented a gene transduction mechanism, where a gene of the original net is substituted by any gene available in the entire reactor population. A third possibility is the capture of an entirely novel gene sequence. If a gene is changed, so has to be its transcript and most often the protein.

The new network must fulfill a certain set of requirements to assure that it operates properly and to become integrated into the reactor population. Firstly the promoter structure of the new gene has to stay intact despite

mutation. This means, the promoter has to be either accessible or inaccessible. This general functionality must not be destroyed. Secondly the potential interactions of the new block of individuals with the residing network and vice versa are determined. For the latter scenario we developed the following idea: A candidate gene comes, merely by chance, under the control of a network-residual protein. The candidate's gene environment is not considered thereby, but it might be part of another regulatory network or a house-keeping gene. Accordingly, we use a random number to decide whether this initial approach succeeds or not. A positive execution of this tasks does not determine whether the gene is silenced or activated, but this decision originates in the gene's initial promoter structure.

Once the first step is accomplished, the chance to be permanently integrated into the residing network increases with the capability of the candidate protein itself to form interactions with the network's members. On the other hand, a protein out of a group of functional equivalent, will assert oneself only if its affinity and dynamics are higher and quicker than the others'. Hence, in our model the new protein is solely accepted if the rate of association with a gene under consideration is equal or better than the mean of all original interactions of the gene.

If all demands can be met, the new gene-transcript-protein block is included into the network and a fitness value can be asserted. According to the nearly neutral theory of molecular evolution, the new individual's molecular fitness $F_M(R)$ must reside in a range $F_M(R) \geq F_{Mparent}(R) \cdot (1 - 0.02)$. If the candidate fails at one of the requirements, the chance of innovation is gambled away and a clone of the replicating network is introduced into the reactor.

4 Computational Results

In the following sections we present numerical results obtained from the simulation of molecular adaptation using the software 'RegNet'. We split them into three aspects. In section 4.1 we deal with the choice and the calibration of the fitness criteria.
Section 4.2 is dealing with the documentation and analysis of the networks' dynamic behaviors during a simulation. Furthermore, the strategies in progression of molecular evolution *in silico* are described in greater detail.
Finally, in section 4.3 we present the effects of noise on molecular evolution. Specifically the influence on the characteristics of the evolving population and its reaction to perturbation are highlighted.

4.1 Multi-Objective Optimization or The Difficulty of Creating a Fitness Measure

The simulation of a genetic network's evolutionary adaptation, as is presented here, is based on the evaluation of several fitness criteria, as has been described in chapter 3 in detail. The single fitness parts are evaluated independently and subsequently summed up and 'mapped' onto a unique scalar fitness value. Optimization procedures of such nature are known as multi-objective optimization problems (MOP) and have their roots in the works of Pareto and Edgeworth.

The evolutionary solution must fulfill several criteria, but it is unlikely that these different objectives can be optimized by the same, often mutually limiting parameter choices. Hence some trade-off between the objectives $f_i(\vec{x})$ is needed to ensure a satisfactory solution $minF(\vec{x})$,

$$\mathbf{minF}(\vec{\mathbf{x}}) = \begin{pmatrix} f_1(\vec{x}) \\ f_2(\vec{x}) \\ \cdot \\ \cdot \\ \cdot \\ f_n(\vec{x}) \end{pmatrix} ,$$

with $n \geq 2$ and $\vec{x} \in C$,

$$C = \{\vec{x} : h(\vec{x}) = 0, g(\vec{x}) \leq 0, a \leq \vec{x} \leq b\} \ .$$

The space, in which the objective vector $\vec{x}$ resides, is called the *objective space*. C is the feasible set in the *objective space* confined by equality and inequality constraints and explicit variable bounds a and b. The image of the feasible set under F is called the *attained set*.

With respect to an optimal solution, the scalar concept of 'optimality' does not apply directly in the multi-objective setting. An adequate replacement is the notion of 'Pareto optimality'. Essentially, a vector $\vec{x}* \in C$ is globally Pareto optimal for the MOP if all other vectors $\vec{x} \in C$ have a higher value for at least one $f_i(\vec{x}*)$ of the objective functions $f_j(\vec{x})$:

$$\exists i : f_i(\vec{x}*) < f_i(\vec{x})$$
$$\forall j \neq i : f_j(\vec{x}*) \leq f_j(\vec{x}) \ .$$

This assumes that for a global Pareto optimal $minF(\vec{x})$ the objective functions $f_i(\vec{x})$ need to be minimized.

There exist several standard techniques to obtain solutions for this class of problem. In the case of evolutionary adaptation a weighted sum of objective functions,

$$\Sigma_{i=1}^{n} \alpha_i f_i(\vec{x}) \ , \ \alpha_i > 0, \ i = 1, 2, .., n \ ,$$

is optimized by the usage of Darwin's principles. Yet, the determination of the weights α_i for the single objective functions and the optimization of the objectives themselves is tricky.

4.1.1 Reaction Fitness and Mapping

According to the fitness function defined in section 3.1.5, we use sequence and structure characteristics of the network players to describe the molecular properties of the network and determine a corresponding fitness value.

In detail, we use a sequential as well as a structural motif f_{GS} and f_{GX} for genes, a criterion f_T applied on the energy range of the transcript's mfe structure and a structure comparison f_P for the proteins. The simulation of evolutionary adaptation reflects a multi-objective optimization for the described objectives f_i. Under the premise to be as close to nature as possible, we avoid bias and demand equality for each f_i. Based on this, the weights x_i have to be similar, too. If neither the desired properties nor the weights can be adjusted empirically, the possibilities of exerting influence on the simulation progression is strongly reduced. Yet, there are alternatives for the

calibration of the multi-layered fitness demands, especially in the area of the selection mechanism: (i) application of constraints for the acceptance of novel networks into the population, (ii) a stronger accentuation of fitter individuals via non-linear mapping to avoid genetic drift effects.

In the first phase of calibration, in which the dynamic aspects of the net are not yet considered, we focus on the velocity of the evolution simulation. We want to facilitate a practicable mean simulation length, even when using the calculation-intensive determination of the dynamic behavior later-on. In the course of the first experiments it turned out that the energy criterion

f_T	id	q	s	K cycles to target
no constraints on $F_M^{x+1}(R)$				
no	yes			> 1000
no		yes		> 1000
no			yes	207 ± 128.6
$F_M^{x+1}(R) \geq F_M^x(R)(1 - 0.02)$				
no	yes			> 1000
yes	yes			> 1000
no		yes		107 ± 48.6
yes		yes		180.6 ± 49.6
no			yes	66.3 ± 26.8
yes			yes	74 ± 33.4
$F_M^{x+1}(R) \geq F_M^x(R)$				
no	yes			99 ± 55.2
yes	yes			186 ± 186.3
no		yes		54.3 ± 7.7
yes		yes		70 ± 28.3
no			yes	40.83 ± 20.0
yes			yes	47.2 ± 17.5

Table 3: Table showing the effects of several independent criteria on the simulation velocity. The following aspects are depicted: The energy criterion for transcripts f_T, the mapping functions $id(x) = x$, $q : [0,1] \rightarrow [0,1]$ with $q(x) = x^2$ and $s : [0,1] \rightarrow [0,17]$ with $s = (2x)^4 + x$, and constraints for the acceptance of novel networks. Criteria denoted with 'yes' are applied in the simulation runs. Simulations of length $l > 10^6$ cycles are not included. R is the set of all network reactions.

f_T, was particularly difficult to fulfill. This effect basically results from the conflicting trends standing behind this criterion. On one hand, the transcript is supposed to show an accessible structure to allow for translation, on the other hand, with enhancement of accessibility the decay rate of the molecule increases, too. Accordingly, we carried out a number of experiments with a specific focus on f_T, applying two different degrees of constraints for the acceptance of novel networks and two different mapping functions $q(x)$ and $s(x)$ for the total fitness depicted in table 3.

In the basic simulation setting, the offspring of a replicating network η_x is accepted if it is functional. No constraints are applied to the fitness of the new individual η_{x+1}. For the up-speeding of the evolution process, we tested restrictions for the fitness of a novel net to be accepted: Firstly, the offspring is admitted, if its fitness $F_M^{x+1}(R) \geq F_M^x(R)(1-0.02)$, with $F_M^x(R)$ being the fitness for the molecular properties of the parent. Else a clone is generated. Secondly, only offspring obeying $F_M^{x+1}(R) \geq F_M^x(R)$ is tolerated. R is the set of all reactions in the network.

In order to weaken the effects of genetic drift and hence accelerate the simulation, we tested the mapping of the fitness values according to

$$\begin{aligned} id : [0.1] &\rightarrow [0,1] \ , \ id(x) = x \\ q : [0,1] &\rightarrow [0,1] \ , \ q(x) = x^2 \\ s : [0,1] &\rightarrow [0,17] \ , \ s(x) = (2x)^4 + x. \end{aligned}$$

Even though the co-domain of the mapped values is unlike, the effects of the functions are directly comparable, because linear changes in the fitness values do not affect the results of the replication-flowout process.

Without the claim to present a complete statistic, we want to summarize the following trends of the solution finding behavior obtained from the series of experiments shown in table 3: The velocity of the target finding increases with enhanced fitness constraints for novel networks and with an emphasized accentuation of fitter networks with respect to less fit individuals.

4.1.2 Dynamic Fitness

Upon having chosen (i) the combination of fitness criteria $F_M(R)$ and (ii) suitable constraints, we want to extend the obtained fitness measure for the network's dynamic aspects $F_D(R)$. In addition, noise comes into play for

the first time. Stochastic fluctuations in the rates of predefined reactions account for noise in the network's dynamics. Noise application can be viewed as generally observed stochastic mechanisms acting on evolving populations. The underlying theory is described in section 3.1.4, the calibration of the two objectives $F_D(R)$ and $F_M(R)$ is presented here.

In accordance with the proceedings for the evaluation of the single components of $F_M(R)$ in section 4.1.1, both objectives are added and normalized to one, such that $max\{\alpha_D \cdot F_D(R) + \alpha_M \cdot F_M(R)\} = 1$. In order to determine the appropriate weights α_i, we use settings in the ranges $\alpha_D = [0, 1]$ and $\alpha_M = [1, 0]$, respectively.

We tested eight distinct combinations as can be seen in table 4 using population size $s = 1000$, mutation rate $m = 0.01$ and a noise range $n = \ ]0, 1]$ exclusively affecting transcription and translation reactions. The statistic of these runs shows that there exists an overall tendency of speeding up the solution finding process along with an increasing α_M.
It is not intuitively insightful, why the gain of importance for one objective, $F_M(R)$, accelerates the simulation, while this is not true for the inverse case. This asymmetric result is caused by the underlying evaluation of the fitness criterion and is aggravated by the lack of size of the solution subspace for bistable behavior, as we could show in section 2.3.2.
Accordingly, in the simulations with $\alpha_M >> \alpha_D$, we obtain results akin to results from single-objective optimizations. This is, because the target net-

sample size	α_M	α_D	K cycles to target
10	1/4	3/4	68.2 ± 31.7
29	1/3	2/3	57.0 ± 21.0
30	1/2	1/2	49.7 ± 20.4
30	2/3	1/3	45.0 ± 23.0
10	3/4	1/4	39.6 ± 16.0
41	4/5	1/5	43.6 ± 23.4
46	16/17	1/17	37.5 ± 17.1
30	99/100	1/100	39.1 ± 17.4

Table 4: Multi-objective weight calibration. The data set for the calibration of α_M vs. α_D is shown. The simulations were carried out using mean population size $s = 1000$, mutation rate $m = 0.01$ and a noise range $n = \ ]0, 1]$. A tendency of increasing velocity for the finding of the solution with increasing weight α_M can be seen.

work *per definitionem* exhibits bistable behavior. In other words, in the most extreme scenario the evolving population does not have to acquire bistable behavior at all, but optimizes the molecular properties solely. Once $F_M(R)$ is minimized, bistability is given away as 'addition'.

On the other hand, in the simulations using more weight on $F_D(R)$, we force the evolution run to proceed via the network of bistable behavior by putting a high weight onto the dynamic aspects. Preferably networks exhibiting bistability are selected, because a loss of bistability causes also a tremendous loss in fitness. Accordingly, a change in the proportions of the objectives does influence the frequency and the total number of bistable networks, as well as the characteristics of the dynamics of the simulation: In figure 29 the change of the ratio $bistable\ NW/(bistable + stable\ NW)$ over time for three distinct weight settings are shown. We show the two extreme and one median data set.
The progressions of the three groups exhibit obvious differences. Especially

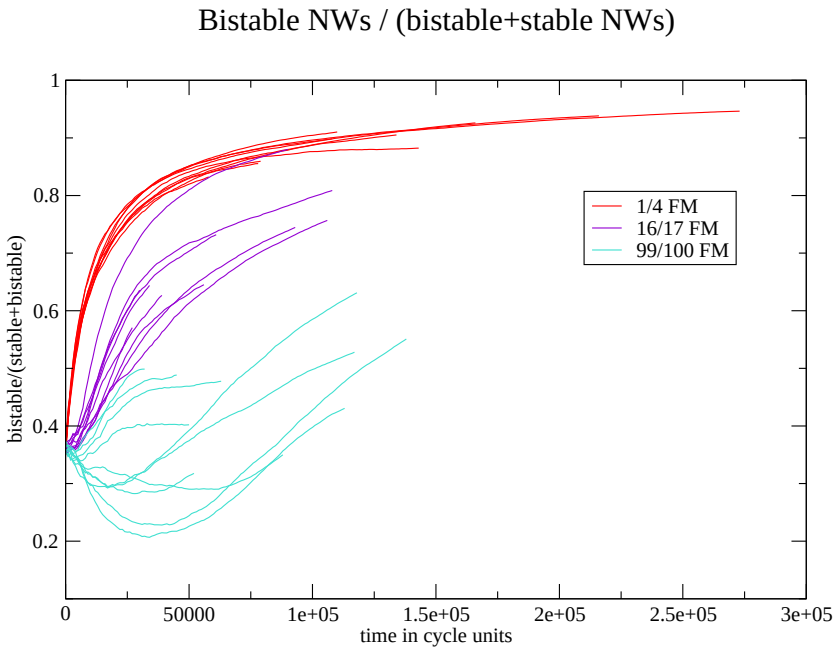


Figure 29: Progression of bistable networks in the course of a multi-objective analysis. The settings for the weight α_M are shown. The weight influence on the total number of bistable and stable networks, as well as on the characteristics of the progression during a simulation can be easily recognized.

the difference in the ratio of bistable versus stable networks is significant. While the number of bistable rises rapidly up to 90% of the current population for $max\{\alpha_M \cdot F_M(R)\} = 1/4$, the ratio stays roughly around 40% for $max\{\alpha_M \cdot F_M(R)\} = 99/100$. The characteristics of the progression show distinct properties, too. While the data sets $\alpha_M = 1/4$ and $\alpha_M = 16/17$ consistently gain bistable individuals, $\alpha_M = 99/100$ exhibits 'ups and downs' in its acquisition. These differences are in good accordance with the respective decrease of mean simulation length as discussed earlier.

In the study presented here, we lay stress on the role of noise during the evolution of genetic networks, among other topics. With the step from evolution simulation of single molecules to reactions and networks of reactions, stochastic fluctuations in the reaction rates become relevant for the first time.

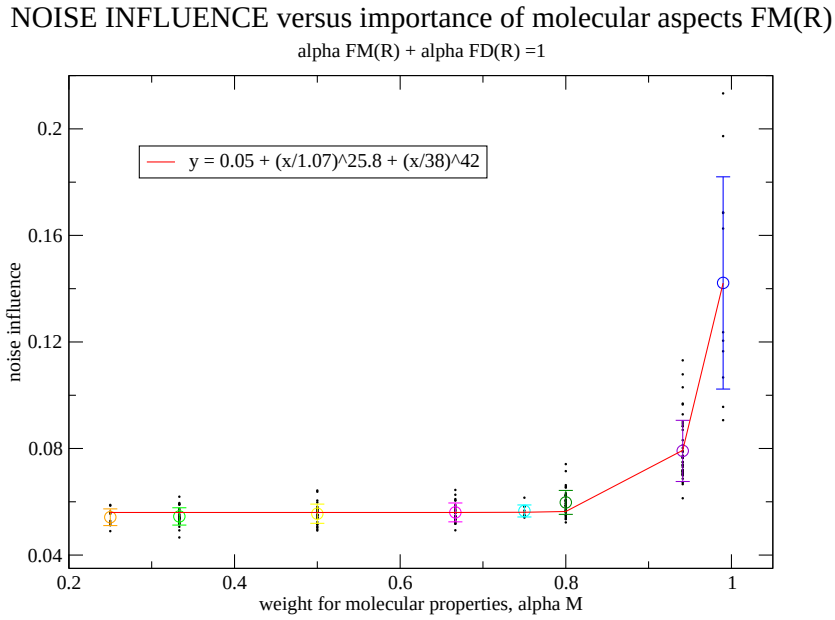


Figure 30: Multi-objective weight optimization. Eight data sets with distinct weights α_i for the dynamic objective $F_D(R)$ and the molecular objective $F_M(R)$ were calculated, such that $max\{\alpha_D \cdot F_D(R) + \alpha_M \cdot F_M(R)\} = 1$. The data sets are sorted by the importance of $F_M(R)$. Noise values for each single run $r_i(x)$ are collected at the point of intersection $y : x \cdot 2.3e-06 = r_i(x)$, with x being the number of cycles and y the noise influence on the population. The colored circles correspond to the mean value of the respective data set, the standard deviation is drafted.

As discussed in 3.1.4, solely transcription and translation reactions are susceptible to noise. Numerical noise values are either chosen equally distributed from within a specific range or, if not specified differently, generated by a sinus wave of certain amplitude.

Noise influence is measured by the number of networks that exclusively exhibit bistable dynamic behavior with noisy parameters. The genuine reaction rates do not account for bistability.
Generally, we observe noise influence changing with time: The investigated data sets exhibit a steep decrease in noise influence with increasing progression in simulation length. A sharp bend of the noise influence curve is followed by a flat decrease until the respective solutions are found. The sole outliers to this behavior are data of setting $max\{\alpha_M \cdot F_M(R)\} = 99/100$. The progressions of this set show maximal and maximally persisting noise influence on their pathway to find the target, even though a vague overall tendency towards the behavior described earlier is still visible. This effect is caused by the need to exhibit the desired property without being strongly selected for it. Noise is utilized to overcome this antagonism. Yet, the strength of noise influence at large is also dependent on the different weight setting α_i.
In figure 30 we depict different weight settings versus noise influence in a graph. We select the noise influence on single populations at a certain point during the simulation. The values are collected from the point of time were the noise influence curves cross the function $y = x \cdot 2.3e - 06$. The function was estimated empirically, yet designed to find the point of time were each single population enters a stage of constant noise influence. We calculated the mean value and its standard deviation for each data set. The data sets' sizes vary, because simulations which already found their target solution at the point of noise measurement are naturally not included.

We find that the noise impact with varying weight α_M behaves according to the fitted function $y = A0 + (x/A1)^{A2} + (x/A3)^{A4}$ with $A0 = 0.0559922$, $A1 = 1.08861$, $A2 = 25.8238$, $A3 = 38$ and $A4 = 42$. However, we predict this function to be true only in the boundaries]0,1[, for the selective pressure must collapse if $max\{\alpha_M \cdot F_M(R)\} = 0$ or $max\{\alpha_M \cdot F_M(R)\} = 1$. In the former case the first bistable network 'wins', in the latter the one which manages to fulfill the molecular criteria first. Noise is no longer of significance.

The data presented here suggest the use of $max\{\alpha_M \cdot F_M(R)\} = 4/5$ and $max\{\alpha_D \cdot F_D(R)\} = 1/5$ as standard setting. In the frame of all tested weight proportions, the mean simulation length is median and both, noise

influence and its time-dependent progression, are located at the edge between non-controllability and triviality.

4.2 Dynamic Behavior of Genetic Networks

Currently, the scientific community is challenged to process biological data achieved by means of high-throughput methods. The amount of experimental data obtained from such conceptual searches is huge and so is the need for their analysis. Mostly, we deal with what proves to be an inverse problem. Given a parameter θ dependent dynamic system

$$\frac{dx}{dt} = g(t, x, \theta), \text{ with } x(t_0) = x_0$$

where g is known and $\theta \in \Theta$. We face a forward problem, if the parameter θ and the initial condition x_0 are given and we want to solve the dynamics $z(t)$ for $t \geq t_0$. The inverse problem is posed, if we have $z(t)$ for $t \geq t_0$ and need to find the parameter θ.

Taken for example the expression data of a group of certain genes as described in e.g. [72], we can reconstruct their dynamics, we know they are interacting in a complex dynamic pattern, but we do not know the pattern structure nor the reaction rates. We need a predictive model of the network's structure.

From a mathematical point of view, often even simple forward solutions demand sophisticated mathematical, statistical and computational methods. Inverse problem methodology poses an even greater challenge, especially for non-deterministic systems. In particular applications in the context of stochastic nonlinear dynamics, as the fitting of expression data on interaction networks, are largely unexplored. Thereby, one of the major problems poses the limited computational power. Hence, the community urges for a new, intelligent framework composed from core mathematical components, such as PDE theory, optimization, approximation theory, functional analysis and computational algorithms, combined with probabilistic foundations, as e.g. empirical process theory, and statistical methodology, as e.g. computational fitting algorithms, nonparametric functions and density estimation. For a detailed description of inverse problems in life sciences and others see for example *www.samsi.info*.

Also *in silico* evolution can be understood from an inverse point of view. Accordingly, the search for the right set of parameters causing a given dynamic behavior represents an inverse problem as well. Based on evolutionary criteria the population searches the parameter subspace to find solutions to the posed fitness demands.
This quest is illustrated by the changing trends of the reactor population to adopt certain dynamic behavior types that can be visualized in a socalled 'relay series'.

4.2.1 Dynamic Aspects of Evolution under Noise Exposure

Simulation of evolutionary adaptation would be half as exploitable without the relay series. Due to the usage of a target solution,which shows maximum fitness, it becomes plausible to trace back the generations of individuals to the first ancestor that lead to the target.

The relay series is, hence, a list of networks beginning with the target network and ending with a network of the initial population. It can be reconstructed in retrospective only, searching for the network η_{x-1} that gave rise to the individual η_x. These steps are repeated until network η_0 is reached.

The progression of the fitness changes, as well as the degree of accomplishment of the single demands and the characteristics of the target approximation can be illustrated excellently in such a graph. It is important, however, to bear in mind that we visualize exclusively the trajectory of a single evolutionary solution, the quickest in our case, but we do not make a statement about alternative, slower trajectories, nor about the total population in the reactor.

The reference setting Firstly we will describe results obtained by simulations with the reference setting, later-on we will specify changes in the behavior along with changes in the parameter settings. The reference setting for the simulation of network evolution is as follows:

- size of the start-population of networks $s = 1000$
- mutation rate $m = 0.01$
- the sequence length of genes and transcripts is 162 bp and 54 aa for proteins

- the range of sinus noise used is $n =]0, 2]$
- the optimal shape for the network proteins is the structure of the GAGA [55], a DNA-binding protein, pdb-identifier 1YUI
- the maximum score for the dynamic objective equals $\alpha_D F_D(R) = 1/5$, with $\alpha_D F_D(R) + \alpha_M F_M(R) = 1$
- based on the reaction rates, either the genuine network or the noisy network must exhibit bistability to obtain the maximum dynamic score ('single bistability')

In figure 31 we show the trajectory of a simulation using the reference setting. The birth dates of the respective individuals are denoted by small red bars

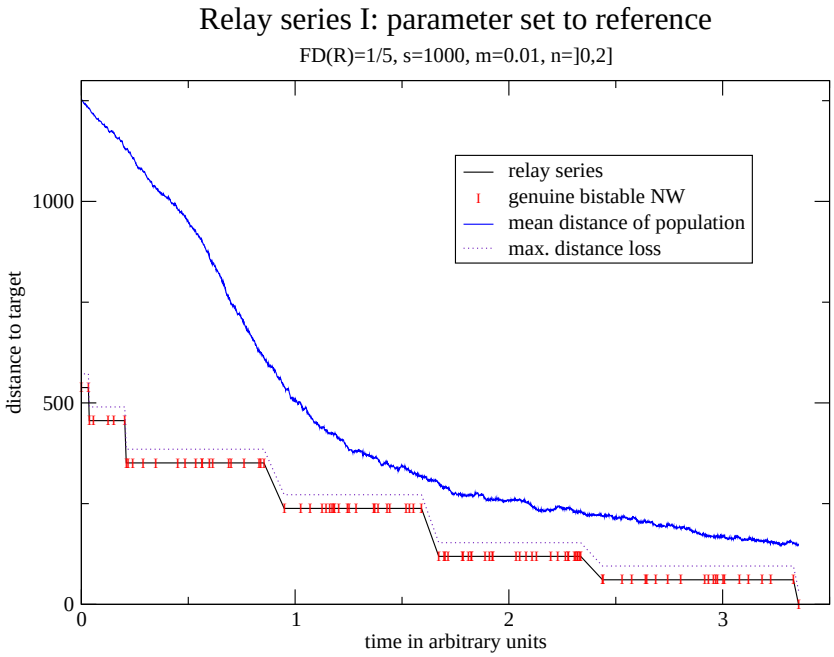


Figure 31: Relay series of simulation with reference parameter setting. The steps of fitness gain toward the target network can be seen. The single individuals constituting the relay series are plotted, all of them are genuinely bistable. All fitness criteria are met completely at $t = 3.4$. The dotted line denotes the potential accepted fitness loss $F_M(R)$ during the evolutionary progress. The mean fitness of the entire population is plotted in blue.

in the reconstruction. If the evolution of the individuals involves generations of cloned networks η_x with $x = \{0, .., latest\ clone\}$, only the birth date of the oldest η_0 is considered for the plot. The dotted line visualizes the 2% limit for the permitted fitness loss $F_M(R)$ in the offspring generation with respect to its parent. We furthermore display the progression of the mean fitness of the entire reactor population in the plot. The relay series of the solution trajectory usually exhibits a fitness progression which resides higher than the mean population fitness, as can be seen in the plot.

Comparing network and single molecule evolution In homology to relay series obtained by the simulation of single molecule evolution [23,24,81], the progression of the relay series in figure 31 proceeds in defined steps of fitness gain. Likewise, we observe a distinction between an (i) initial phase, with fast and a (ii) second period with slow increase of mean fitness. Yet, the second phase looks qualitatively different from the examples in [81]. The mean fitness never increases in steps as was found in single molecular simulations.
The reason lies in the behavior of the population and the mapping of the fitness values. In the second phase of the simulation, the population in [81] regularly steadies at certain fitness levels, which are met by a major fraction of individuals, i.e. 80% to 95%. Once fitness-homogenized, the population suffers from genetic drift. A sudden gain of fitness allows for selection of the fittest pseudo-species - until the rest of populations adjusts and sinks into drift again.
In our approaches we avoid genetic drift by accentuating the differences between better and even better individuals. At the level of the population, this leads characteristically to a continuous mean fitness gain, while the progression of the relay series proceeds via steps. Based on this observation, we conclude a reactor population consisting of many different and small subpopulations. This assumption also explains the relatively short relay steps compared to simulations in [81]. The artificially enhanced selective pressure forces the subpopulation to improve quickly or to go extinct. There is simply no time to build and explore extensive neutral networks of similar fitness. Yet, we can confirm the existence of neutral networks in the chosen parameter space, even though their sizes are blurred.

The classification of transitions of the relay series into continuous and discontinuous as proposed in [81] based on the the types of change in the

shape of the molecule, the frequency of such changes and the individual history of the novel shape, is not followed up in this work for obvious reasons. The multi-objective approach anticipates a unique classification of novelties and a clear-cut measure for major changes within the network and its players.

A novelty to the relay series is the explicit notation of the dynamic behavior of the individual networks. It is denoted by a color code as indicated in the figure's legend. In plot 31, the trajectory proceeds exclusively via genuinely bistable networks, i.e. the reaction rates obtained from the properties of the molecules cause bistable behavior of the net by themselves, noise influence is neglectable for the relay series. Yet, this is in contrast to the influence noise exerts over the entire population. A detailed analysis of this aspect can be found in section 4.3.

The analysis of 41 relay series of independent simulations under reference conditions using the same initial population and target network, showed that a loss of genuine bistability, which happened in 12.2% of the cases at least once per simulation, is a rare event under this parameter setting. Whenever bistability is lost in favor of stability, it is regained within one or two generations and a higher fitness level with respect to the last bistable level is reached. An example for this progression is shown in figure 32.

Moreover, two different reasons for the loss of bistability can be observed. They empirically depend on the duration of the simulation at the respective point of time and the mean fitness of the reactor population: If the loss happens very early in evolution, the trajectory resides generally at much higher fitness levels than the mean population. Even a harsh loss of fitness does not cause the individual to range among the least fit of the ensemble and thus, it is not explicitly selected against.

If the loss happens later in evolution the situation differs considerably. As can be seen in figure 32, the mean fitness and the fitness level of the trajectory get very close. The trajectory individuals suffer an enhanced selective pressure, because of the high competition between the subpopulations. A shortcut via the larger subspace of stable networks taken at the right time saves the situation.

Another effect caused by the severe selective pressure is the regular loss of fitness $F_M(R)$ within the permitted range, whereas the general dynamic behavior is maintained. We are dealing with a 2% fitness loss according to the nearly neutral theory of evolution. In simulations of single molecule evolution as in [81], loss of fitness is never observed, even though any possible fitness change is permitted theoretically. The selective pressure inherent in

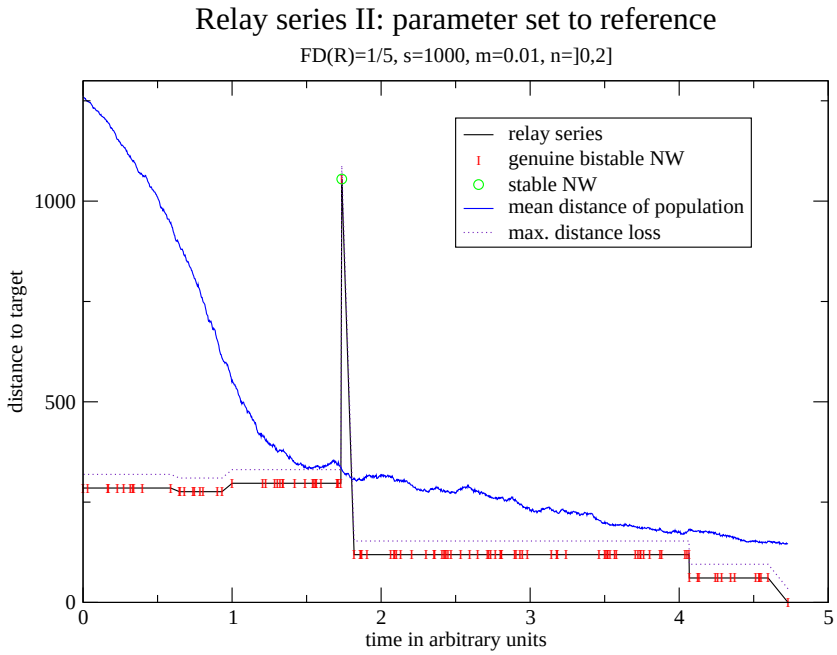


Figure 32: Relay series with parameters set to reference. The trajectory shows a distance loss at time step=1, the loss lies within the permitted range for $F_M(R)$. At time step=1.7 the network η_x individual loses its genuine bistability and generation η_{x+1} falls into stable behavior. This reduces its fitness value by $F_D(R) = 1/5$ of the maximum fitness possible plus the permitted fitness loss for $F_M(R)$. The following generation η_{x+2} regains bistability and enters a higher fitness level than generation η_x. Note the closeness of the mean fitness and the fitness level of generation η_x at the point of bistability loss.

our model prevents the formation of only few subpopulation at the same fitness level. This causes the continuous gain of mean fitness and make regular losses of fitness for the relay trajectory possible.

Statistics We calculated detailed statistics of the trajectory's dynamic behavior for different settings in table 5, whereby the overall parameter set is fixed to the reference and the noise amplitude N is varied between $N = \{0, 0.5, 1, 2, 3\}$. We analyzed distinct effects of the noise strength on the dynamic behavior of the relay series, such as on the fraction of runs constantly showing gain of the total fitness $F(R)$, on the fraction exhibiting loss of fitness $F_M(R)$ and $F_D(R)$ during their relay series. We classify

	units	N=0	N=0.5	N=1	N=2	N=3
cont. gain of $F(R)$	%	42.9	20	24.4	24.3	25.7
const. bistab.	%	100	85	87.8	84.8	85.7
loss of $F_M(R)$	%	57.1	80	68.3	72.7	71.4
bistab.$\leftrightarrow$ stab.	%	0	15	12.2	15.2	8.6
bistab.$\leftrightarrow$ noisy bistab.	%	0	0	0	3.1	8.6
<simulation length>	cycles	74 ± 33	109 ± 53	99 ± 57	97 ± 47	91 ± 47
sample size		28	20	41	33	35

Table 5: Statistics of the dynamic behavior of relay series at different noise ranges. The parameter set is fixed, the noise amplitude N is varied as indicated. The column denoted by $N = 1$ represents the reference setting. A detailed analysis of various aspects of the relay series dynamics is shown, multiple entries are possible. The section 'constantly bistable' accounts for trajectories showing exclusively genuine bistability. Loss of $F_D(R)$ is split into two parts, loss of bistability in favor of stability and in favor of noise-dependent bistability. The simulation length is represented in K cycles needed to reach a target solution. Abbreviated 'cont.' stands for 'continuous'.

'loss of $F_D(R)$' into loss of genuine bistability in favor of stability and loss in favor of noise-dependent bistability. The latter is enabled exclusively by noisy parameters. Furthermore, we were interested in the frequency of exclusively genuinely bistable relay series and eventually, in effects on the mean simulation length.

Basically, the results show two distinct tendencies with respect to runs with and without noise application. Accordingly, we find runs with constant gain of $F(R)$ to be 7-13% more frequent in simulations without noise. Loss of $F_M(R)$ and $F_D(R)$ increases in samples with noise application, i.e. in average by $16 \pm 4\%$ and by $15 \pm 2\%$ respectively. Since multiple entries are possible and frequent, such that $F_M(R) + F_D(R) \approx 17\%$, noise-dependent $F(R)$ gain and loss directly add up to 100%, as is expected.

Relay series loosing their genuine bistability, caused exclusively by the molecular conditions of the networks, usually regain this property very quickly. In 81.25% of the cases the regeneration is completed within two generations. A different trend is observed if noisy parameters account for the network's bistability, while the genuine parameter set lost this property. The length of such 'noise intermezzi' varies from 1-18 generations in the runs obtained from the samples presented. The enhanced persistence of such periods is partly

caused by the fact that noise-dependent bistability is equally rewarded as genuine bistability, partly because such by-passing of dynamic deficiencies avoids a negative selection of the respective individuals.
The subpopulation responsible for the last relay step suffers from an augmented selective pressure caused by the competitive composition of the reactor population at the point of bistability loss. Therefor it changes to a bigger solution subspace, i.e. takes a shortcut via a less fit evolutionary path. Yet, selective stress forces a quick regaining of the lost properties in order to avoid extinction.

Interestingly, the trend to fall into noise-dependent bistability is proportional to the size of the noise range applied. While small N show no influence on the individuals in the relay series, settings with $N >= 2$ certainly do, as can be seen in table 5.

For the sake of completeness, we do not forget to mention the tendency of simulations with noise to show extended mean simulation lengths compared to runs without noise. This counterintuitive result is explained in section 4.3 in detail.

We recapitulate the results obtained as follows: In experiments using the reference setting we find that the mean fitness of the reactor population increases continuously, while the relay series gains fitness in steps. These steps are interspersed by sudden, short losses of $F_D(R)$ and longer periods of lost $F_M(R)$. Thereby $F_M(R)$ represents the part of fitness based on the molecular properties of the individual and $F_D(R)$ the fitness based on its dynamic behavior. We deduce that the reactor population evolves many small subpopulations in the sense of quasispecies, which span a broad range of fitness levels. These effects are primarily caused by an artificially enhanced selective pressure used in order to reduce genetic drift.
Furthermore, we observe a qualitative change in the dynamic behavior of the relay series with changing noise amplitudes with respect to losses of $F_D(R)$. We find a frequency shift from stability to noise-dependent bistability, which is proportional to the applied noise range.
The fraction of runs exhibiting $F_M(R)$ loss increases considerably comparing noise-free and noise simulations, respectively.
The evolvability of the proposed model for the individual networks, as well as for the variation-selection mechanism, is thence provided.

4.2.2 Evolution of Noise-Tolerant Networks

In order to study a different aspect of noise in evolution *in silico*, namely the evolution of noise tolerance, we used the following, slightly altered experimental setup: Exclusively networks exhibiting bistability with *and* without noise application are fully rewarded. Individuals showing single bistability receive half of the maximum $F_D(R)$ value possible. This setting amounts to the attempt to evolve networks via the solution subspace of noise-tolerant switches.

In figure 33, we show a typical relay series constructed from the simulation of noise-tolerance evolution. Noise-tolerant individuals, which exhibit bistability genuinely and under noise application, are depicted in red. Noise-susceptible networks, which solely show genuine bistability, but fail under noise influence, are encoded in green.

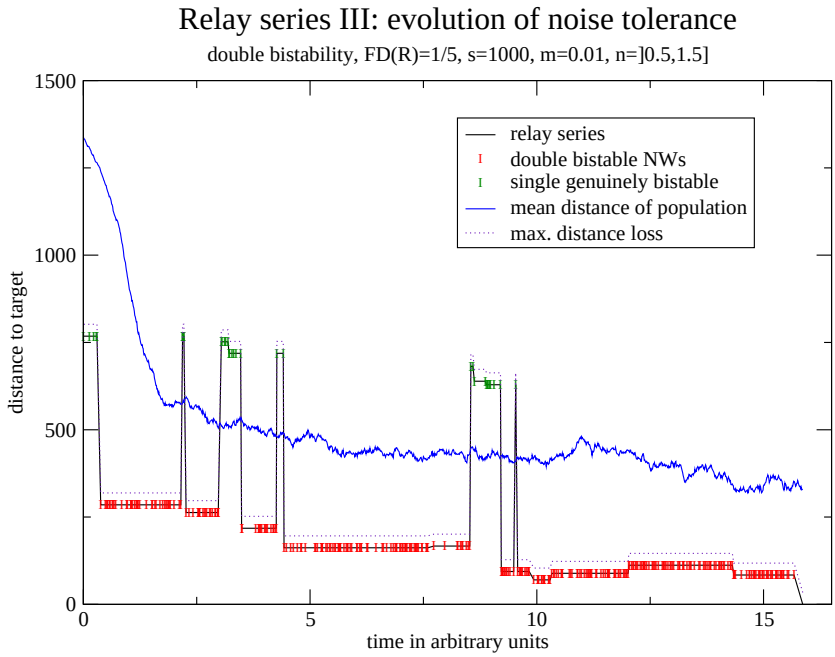


Figure 33: Relay series showing the evolution of noise-tolerance. Noise-tolerant individuals are depicted in red, noise-susceptible switches in green. The mean reactor fitness runs mainly between the fitness fronts of the noise-tolerant and the noise-susceptible individuals of the series. Steps of fitness gain are found in both fronts.

Interestingly, the mean reactor fitness mainly runs between the two fitness fronts of double bistable and single bistable networks. Based on this observation and on the progression of the reactor population composition (not shown), we conclude that the reactor population is constituted from subpopulations showing two principal trends in their dynamic behavior: noise-tolerant bistability and individuals of genuine bistability without the property to tolerate noise.

This result is highly significant, specifically because of the artificially enhanced selective pressure conditions applied. Without a decisive role in the quest for good noise-tolerant solutions, the single bistable subpopulation would suffer from negative selection and greatly reduce in size. The mean reactor fitness would tend towards the fitness levels of the noise-tolerant switches. The fitness steps in the noise-susceptible relay progression in figure 33 must be understood according to this insight. We are dealing with two parallel evolving subpopulations classified by their dynamic behavior, though only one of them lives up to the fitness demands.
Yet, the strengthened negative selection against noise-susceptible individuals can not be without effect, for details see section 4.3.3.

Statistics In table 6 we present a statistic over the qualitative properties of 87 independent relay series.
The simulations presented are calculated under reference conditions, yet, to gain maximum $F_D(R)$ the networks must constitute noise-tolerant switches as described earlier. The range of stochastic fluctuations of the reaction rates is varied as indicated. The runs without noise application $N = 0$ resemble the generic reference runs without noise.

We can anticipate the degree of difficulty of the task by interpreting the three first rows of table 6. None of the 87 relay series proceeded towards the target either via a continuous gain of total fitness or via a series of pure noise-tolerant switches. Even the number of series using both noise-tolerant and susceptible individuals is not very dominant, especially at higher noise amplitudes.
Searching a solution, the mean relay series resides enhancingly in solution subspaces other than of noise-resistant switches, i.e. up to 50% of the generations in the series. The respective trend increases with noise amplitude.
On the other hand, the number of runs, which exhibit loss of $F_M(R)$ at least once during the relay series decreases. We speculate that the population

	units	N=0	N=0.5	N=0.7	N=1
cont. gain of $F(R)$	%	42.9	0	0	0
const. d bistab.	%	100	0	0	0
const. bistab. (s/d)	%	100	86.7	64.3	66.7
loss of $F_M(R)$	%	57.1	43.4	28.6	20
<loss d bistab./ simulation>	%	0	32 ± 17	49 ± 20	50 ± 22
d bistab.↔ s genuine bistab.	%	0	100	100	100
d bistab.↔ s noisy bistab.← stab.	%	0	10	7.1	13.4
s/d bistab.↔ stab.	%	0	13.4	35.7	33.4
<simulation length>	cycles	74 ± 33	94 ± 67	90 ± 66	108 ± 42
sample size		28	30	14	15

Table 6: Statistics of relay series for the evolution of noise-tolerant networks. In order to gain max. $F_D(R)$ networks must be noise-tolerant switches, i.e. show bistability with and without noise application. The experimental setup is otherwise set to reference, the noise amplitudes N are varied as indicated. Minuscule 's' and 'd' denote single and double bistability, noise susceptibility and tolerance respectively. Abbreviated 'cont.' stands for 'continuous'. Mean simulation length is shown in K cycles. Loss of $F_D(R)$ is split into the three possible subgroups: double bistability changing to single bistability, genuinely or noise-dependent, or to stability. Multiple entries are possible.

increasingly shifts towards a tactic of co-evolution, partly because of the reducing size of the respective solution subspace with enhancing perturbation, partly because of the great competitivity among the fitter solutions.

We present the qualitative interaction pattern between the different subspaces of dynamic behavior in row six to eight of table 6. The proportions stay basically the same for all experiments using noise exposure. Yet, we want to pointout the difference in importance of the distinct solution subspaces for the relay series. The principal shortcut leads via the subspace of noise-susceptible switches. The next frequent choice, followed with major distance, is the subspace of globally stable solutions and last in priority are single, noise-dependent solutions.

The mean simulation length grows with the degree of noise applied, thus, with the difficulty of the task.

We summarize that the evolution of noise-tolerant switches poses a more difficult problem than evolution of switches under noise exposure, solely. We observe a change in the strategy for the solution finding, i.e. the augmented

supply of new search starting points deduced from hierarchically lower dynamic subspaces. These lower dynamic subspaces co-evolve with the principal solution subspace, as could be shown for the subspace of noise-susceptible switches in figure 33.

4.3 Noise and Evolution

4.3.1 A Model of Noise Impact for Evolution

The impact of noise in simulations of network evolution is multi-layered and not mandatorily intuitively insightful. In order to clarify its influence we used the following experimental setting: During a simulation, network units have to meet a range of criteria that disembogue in a dynamic behavior of the unit. Preferentially, this behavior is bistability, thus, the network represents a functional switch. Alternatively, the network can exhibit global stability, which is not rewarded with respect to the fitness value of the net. Subsequently, noise comes into play. In simulations with noise application, the networks get a 'second chance' in case they only show stability. The reaction rates for transcription and translation reactions suffer from stochastic fluctuations within a given range and the dynamics for the 'noisy network' is re-evaluated. The concept was to assist evolution to live up to the difficult fitness criteria, the results narrate the impact of noise for evolution in general.

Especially the detailed analysis of the data presented in table 5 permits to interpret the effects of noise and alike stochastic mechanisms in evolution simulations. In order to elucidate dependencies in the sense of the following argumentation and in the sense of the noise impact at the levels of individuals and of populations, we want to introduce two notions, these are 'primary' and 'secondary noise-dependent effects'.

Primary effects simply denote the occurrence of noise in the reaction rates at the level of the individual networks. Interestingly, the primary effects do not show in the relay series until considerable amplitudes of noise are applied as can be seen in table 5, row five. This allows for the reasoning that noise-dependent bistability is rare in the population. The primary effects grow with the size of noise range applied and are identified in the shape of noise-dependent bistability. They start to affect the individual solution pathway after their frequency has passed a certain threshold T_N in the population, in our case based on $T_N \propto N = 2$. The explicit primary effects are thence of

weak nature.

The secondary noise-dependent effects describe a subsequent reshuffling at the level of populations. Based on the frequency of fitness loss in favor of stable dynamics, already at low noise rates, and the progression of the mean fitness during a simulation, we presume that the primary effects cause noise-dependent bistable individuals to primarily produce stable offspring. This is facilitated, because the parents themselves are *de facto* genuinely stable.
Hence, the secondary noise-dependent effects cause broadening of the population from a genotypic point of view and a broadening of the strategies for a solution pathway. Hence, the relay series tilts into stability, because the broad population and its respectively lower mean fitness of the reactor establish the possibility.

These effects account for various phenomena observed in the simulation runs, e.g. the counterintuitive extension of the mean simulation length under noise application. Due to the secondary effects the population actually proceeds via more than one solution subspace towards the target. This prevents the extensive search for a pathway in one particular subspace, the relay series is driven by a more stochastic sampling of the respective subspaces in selective stress situations, the mean fitness gain proceeds slower, the mean simulation length extends.

Natural evolutionary processes are devoid of predefined targets, hence there exists no focus on the time period necessary to evolve a certain species. These presumptions exclusively result from the model of the simulation of such processes. Accordingly, the genetic broadening of evolving populations and networks of populations is not under negative constraints. To the contrary, stochastic breaking up of otherwise too streamlined populations by noise-like mechanisms [35,43,44,46] facilitates quick adaptation of the species under changing conditions [74], whereas genetically uniform populations become very fragile or go extinct [50, 51].

4.3.2 Development-Dependent Noise Aspects

Even though noise is applied constantly during the runs of the 'optional noise' experiments, we found specific periods of importance and periods quasi without influence on the evolving population. These correlate with the initial phase of the simulation and the second phase respectively, notions termed in context of the mean fitness progression.

In figure 34 we show the correlation between the noise influence and the

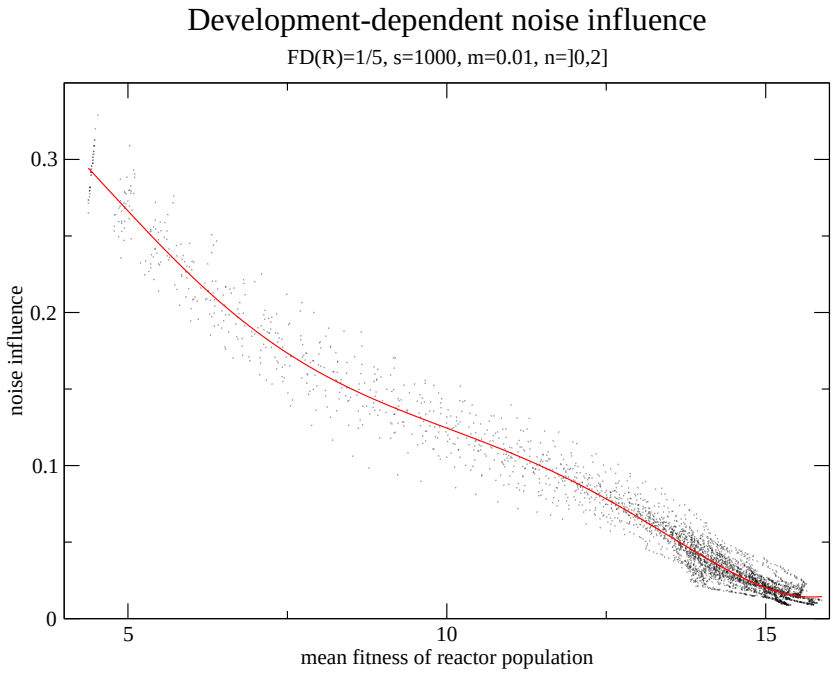


Figure 34: Correlation of noise influence and mean fitness of the reactor population. The data set consists of 41 simulations, a data regression of 5^{th} degree was carried out and is shown in red. Two bends intersperse the otherwise linear correlation. The density of the data points correlates with the change over time.

mean fitness of the population in simulations using the reference setting. Both parameters are time dependent, hence a higher density of data points represents a longer time period. The data set consists of 41 independent samples, a data regression of 5^{th} degree was carried out and is depicted as red line. The decrease of noise influence is direct proportional with the increase of the reactor population's mean fitness, while noise is applied constantly at the same range. The population 'accepts' noise constrained by its stage of development. According to the model proposed in section 4.3.1, the reactor population can not afford to keep too many noise-dependent subpopulations, while advancing in mean fitness. This effect reduces the genotypic variability of the population and leads to an increasingly focused search for the target dispersing from several starting points in solution space.

Furthermore, we find the correlation between noise influence and the mean fitness flattens, specifically for high mean fitness values with increas-

ing noise amplitudes, data not shown. We interprete that the population is forced to reject noise more strictly with higher maximum noise amplitudes, because the chance to find a noise-dependent bistable network increases statistically with an enhanced noise range. This trend stands in conflict with the focusing on the search behavior at higher mean fitness levels and driven by selective pressure its trade-off shifts in favor of the mean fitness.

We recapitulate that profitable influence of noise reduces with enhanced genetic uniformity generated under strengthened selective pressure. This comportment is caused by two conflicting trends: (i) the need for targeted adaptation of the population and (ii) the general tendency of noise to enhance genetic variability.
In our simulations this fact is reflected by the population's diminished usage of noise at developmental stages close to the optimum solution.

4.3.3 Noise Tolerance and Co-Evolution

In experiments simulating the evolution of noise-tolerant switches, noise can no longer be understood as stochastic perturbations, but as yet another fitness demand. By which means the reactor population succeeds to incorporate an unforseeable force is, however, worth a look.

In figure 35, we show the ratio between the subpopulations of different dynamic behavior during the optimization of noise tolerance. The bistable subpopulations must be understood as follows: Double bistable individuals represent the successful noise incorporation, they are noise-tolerant switches as for our mini-system. Genuinely bistable networks are switches that failed to tolerate noise with respect to their dynamics. Noise-dependent bistable individuals are actually stable networks exhibiting switch-like behavior under noise exposure. Their genuine reaction rates do not account for bistable behavior.

At the beginning of the simulation the proportions are in good accordance with the natural distribution of stable and bistable in solution space of our model network, i.e. $stable : bistable = 4 : 1$ with noise-dependent bistable counting for stable solutions. For details see also 2.3.2.
In the course of the simulation the number of noise-susceptive switches rises dominantly in the population to the disadvantage of stable individuals. Furthermore, the number of noise-tolerant switches rises to disembogue into a constant gain rate of individuals, but at a much lower level compared to the noise-susceptive switches. The latter subpopulation is evolving at higher

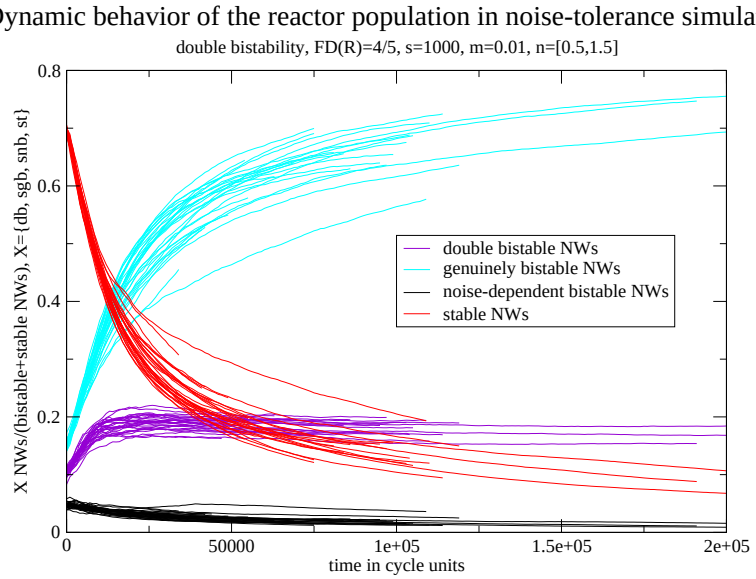


Figure 35: Ratio between subpopulations of different dynamic behavior during evolution of noise-tolerance. The single values are normalized over the entire population. The graph consists of 30 independent simulation runs under the same parameter set.

speed. Two reasons for this cause: Firstly, the respective solution subspace is bigger. Secondly, the discrepancy between the size of the noise-susceptive subpopulation and the actual demand for noise-tolerance makes them target to negative selection and accordingly to a quicker turnover. Noise-susceptible individuals exhibit a higher evolution rate than noise-tolerant.

Based on data shown in figure 33 and table 6, we conclude that the evolution of noise-tolerance proceeds via two levels: the search in the appropriate solution subspace, as well as via the supply of new starting points in the latter subspace by the evolution of noise-susceptible individuals.

We find that the enhanced degree of difficulty of the evolutionary task, which is reflected in a smaller, potentially less connected solution subspace, forces the population to change the evolution strategy. An 'evolutionary muse' becomes essential.

The total population becomes split into two relevant subgroups, these are the noise-tolerant switches relevant for their fitness and the noise-susceptible switches for their subpopulation size. These subpopulations are mutually dependent on each other in terms of successful species development, which

clearly shows the situation of co-evolution. Co-evolution is frequent with organisms that are ecologically intimate, such as predator and prey or host and parasite. In our case we observe a more symbiotic type of co-evolution, i.e. the individual development of each species is supported by the respective other. For further reading on co-evolution see also [3, 10, 15, 48, 77].

4.3.4 Noise Exposure and Adaptation

Whether noise is understood as actual stochastic fluctuations in the reaction rates of complex systems or as genetic mechanisms to ensure genetic variability even at the expense of innocent deaths, such as bursts of transposons, gene shuffling, gene duplications or simply recombination, the question for plasticity and tolerance limits of affected populations must be posed.

In this context and restricted by the limits of *in silico* experiments we were interested to study different patterns of noise exposure and the corresponding reaction of the evolving population. Two essentially different setups with respect to the noise pattern were used: Continuous and population covering noise application versus exposure of only 20% of the population at different stages of evolution.
The reaction of the population was measured by the maximum size of the noise-tolerant subpopulation and the evolutionary velocity to reach this size. As can be seen in table 7 the relative maximum size of a noise tolerant subpopulation is independent of the pattern of noise exposure, i.e. total versus partial. The absolut numbers naturally differ with the amount of individuals affected. Furthermore, we conclude that the potential maximum number of noise-tolerant individuals results directly from the trade-off between noise-susceptibles and noise-tolerants. In other words, the potential amount of good solutions results from the degree of difficulty of the posed problem. As

N	total	partial
0.5	$\approx 1/5$	$\approx 1/5$
0.7	$\approx 1/10$	
1	$\approx 1/20$	

Table 7: Max. size of noise-tolerant subpopulation. The relative fraction of noise exposed individuals are shown. Noise amplitude is varied as indicated. Results of noise exposure to the total population versus exposure to a part (20%) of the population is presented.

N onset	total	partial
begin	4.4e-6	4.4e-6
middle		3e-7
end		1.2e-7

Table 8: Evolutionary velocity towards the max. noise-tolerant subpopulation. Noise was applied to total or partial (20%) population at varying stages of development as indicated. Noise onset in the middle of the simulation means after 25000 cycles, at the end means after 75000 cycles.

shown in table 7, increasing the noise amplitude two fold results in four fold less valid solutions.

In table 8 we present the mean velocity of the population towards a maximum size of noise-tolerant individuals. The velocity is approximated in a linear fashion. The evolutionary velocity is independent of the mode of noise application based on on a detailed analysis of the proportions of the subpopulation during *in silico* evolution and on table 8. We furthermore conclude that it is not independent of the time of the noise onset with respect to the total population's stage of development. The composition of the population, in other words the genetic variability, influences the reaction time of a population to a posed perturbation.

The experiments presented here show that the ability of a reactor population to respond to an external perturbation depends on its genetic variability, as does the amount of solutions on the degree of difficulty of the task. Both measures are independent from the fashion of noise application. We want to note however, that the exposure of the entire population to such an external perturbation leads to a bottleneck situation for the population due to the amount of failing individuals. The partial exposure strategy, though, permits a population to adapt continuously without causing a dangerous squeeze for the species.

5 Conclusions

Networks appear as common theme at all levels of molecular evolution [64]: The structure of biopolymers and their folding kinetics is determined by a network of metastable states and their connecting saddle points. Interacting replicators form complicated ecological networks. Neutral networks in sequence space explain the evolvability of both nucleic acids and polypeptides by linking Darwinian selection and neutral drift. The reaction networks of metabolism and last but not least the network of genetic regulation, all of them provide the picture of highly dynamic complexity.

In order to establish the possibility for insights into such systems, it is necessary to find simplified frameworks facilitating the access and interpretation of their intricate properties. The study presented here, is based on the assumption that knowledge of the behavior of simple basic prototypes of networks is helpful for this process. Not only the enhanced conceivability of such small artificial systems, but also the cognition of the modularity of large natural circuitries contribute to their understanding.

The quest for such generic building blocks proves to be proportionally intensive, in particular because potential candidates have to be determined empirically. Thus, in the framework of this context we studied the properties of a minimal switch. Switches are known to be modules for e.g. electronic circuitries of any conceivable finesse. In analogy we developed an auto-regulatory gene switch, studied its dynamic behavior and simulated the evolution of such modules *in silico*.

The design of a gene switch comprises firstly the choice for the degree of naturalism of the model and subsequently the selection of the molecular players and their interactions, eventually the mathematical description of the interacting network and the detailed analysis.
The model proposed here consists of two genes and their products, transcripts and proteins, mutually regulating their activities, whereby gene A acts as inductor and gene B as inhibitor of gene expression. Higher order dynamics in minimal gene networks are implemented by strong constitutive promoters, enhancement of the number of interacting levels, i.e. genes, transcripts and proteins, and by introducing cooperative auto-regulative dependencies.

Accordingly, the conformational change of the promoter region, and thereby the respective enhancing or silencing of gene expression, depends on the cooperative binding and oligomerization of the transcription factors in the upstream regulatory region of the gene. The molecular properties of the in-

teracting individuals translate into binding coefficients and reaction rates for the reactions possible under the constraints of the model.

The potential dynamic behavior of this predefined framework of interactions was investigated, partly analytically, partly computer-assisted, and the network's capability for stability and bistability under respective parameter settings was assessed. The frequency of bistability versus stability in solution space is 1:4, respectively, the distribution of the two types of behavior results from the progression of the bifurcation of the system's stationarity landscape.

Naturally, as well genetic network modules such as oscillators, switches between multiple states or others are conceivable for the role of the universal basic unit.

In the fine tradition of evolutionary simulations *in silico*, we developed a software capable to simulate the adaptation of gene networks, in particular switches. The analyzed gene circuitry is used as a single unit in a population of alike individuals. A fitness value is assigned to each network that results from an elaborate combination of molecular properties of the involved molecular individual, such as the mandatory presence of an intact promoter structure, the strength of an enhancer element, the degree of accessibility of the transcripts for both translation and degradation and the transcription factors' probability for binding to the promoter based on the similarity to a zinc-finger, with the dynamic characteristics of the net caused by the pertinently generated, underlying fluxes.

The simulation proceeds in accordance with Darwin's principles via fitness-motivated, erroneous replication and selection. The decision, whether mutated offspring is viable or not, depends on the changes in the molecular interplay and a maximum tenable fitness loss caused by this molecular interplay if compared to the parent. This assumption is made following the nearly neutral theory of evolution.

We simulated the evolutionary progress of a population under artificially enhanced selective pressure, thereby genetic drift is reduced to a minimum. This setting is among other reasons motivated by the computational intensity of an approach considering explicitly the dynamics of each individual and the limitations for disposable resources. The enhanced selective pressure is realized by an 'unbiological' mapping of the fitness values into a different domain, accentuating in particular the fitness difference between good and better individuals. These conditions result in a highly competitive behavior of small subpopulations fighting for the fitness-based, numerical hegemony

in the reactor. The extensive search for evolutionary solutions on neutral networks is competitively limited and genetic drift is reduced.
Naturally, the results obtained from such simulations must be viewed under the aspect of these 'environmental' conditions. In other words we simulate populations suffering from a certain level of selective stress, situations such as rapidly changing environmental conditions, but also the vital adaptation to strongly compromising parasites, diseases or noise-like perturbations.

Under these premises our model serves especially well to study the impact of noise on evolving populations. Two fundamentally different setups were investigated under this aspect: (i) the influence of 'optional' noise and (ii) the evolution of noise-tolerance.
For the maximum fitness obtained from the dynamic objective of a network, the individuals must exhibit switch behavior, in other words bistability. For the first approach, the maximum fitness was assigned if the net showed bistability of its own accord *or* under noise influence. For the latter approach the individuals had to be functional switches with *and* without noise exposure.

The results obtained from the former setup seem counterintuitive at first sight, particularly with respect to their mean simulation length. Noise conceived in this context as 'evolutionary help' does not speed up the optimization, but re-structures the evolving population resulting in a deceleration of the adaptation.
We proposed the notions primary and secondary noise-dependent effects in this connection. The primary effects are the optional acceptance of noise in order to fulfill the high standard of fitness criteria at the level of the individual. Interestingly, these effects rarely show up in the relay series, which is the reconstruction of the quickest trajectory towards the fitness-target, if at all then only at great noise strength. To the contrary, the secondary noise-dependent effects show immediately at any noise strength and result in the observed deceleration of the simulation. As was learned from the simulations, these effects cause the enhancing of the genetic variability in the reactor population by preventing that networks genuinely showing stable behavior become target of negative selection when showing bistability under noise influence. We observe a general tendency of noise to primarily act at the level of populations, before at high noise strengths, trespassing towards exerting a statistically relevant influence on individuals.
Furthermore, we observe that optional noise-influence depends on the stage of development of the reactor population, in other words the degree of genetic uniformity. Influence ceases with enhancing mean fitness of the reactor

population. We interpret that with enhancing genetic uniformity noise cause more harm than benefit with respect to fitness gain. In a genetically variable population on the other hand, noise provides alternative starting points for a targeted search in solution space and, thus, provide superior means in the struggle to survive.

In the simulation of noise-tolerance evolution, the successful handling of noise becomes compulsory. Then the degree of difficulty increases drastically. Yet, that fact is, again, not reflected by the mean computer time needed for the simulations, but in the evolutionary strategy required for success.
The analysis of our sample set revealed a generally enhanced tendency of fitness losses along the relay series, particularly in favor of less fit dynamic behaviors. Though, the fact that two relevant mutually dependent subpopulations constitute in the reactor during each simulation, one relevant for its fitness, the other for its size, which undergo co-evolution makes this setup particularly interesting. The subpopulation of noise-tolerant individuals residing at the fitness edge gains new starting points in solution space from the large noise-sensitive community in the reactor, the latter group on the other hand would collapse under negative selection without the existence of noise-tolerant networks. The suffering from negative selection of the noise-sensitive subpopulation results in an augmented evolutionary turnover of this specific group.

Eventually we were interested, whether the mode of noise application is relevant for the evolving reactor population. Total exposure versus statistically equally distributed partial exposure of the total population were investigated for this purpose. We could show that neither the velocity of the reactor population to develop a maximum subpopulation of fittest solutions, nor the relative maximum size of this subgroup depends on the pattern of noise application.
We found, however, that the size of the maximum subpopulation is negatively proportional to the noise-strength or the degree of difficulty of the task and that the evolutionary velocity towards its maximum size is dependent on the genetic variability of the total population.

Whether we understand noise as actual stochastic fluctuations in the reaction rates of complex systems, as mechanism to ensure genetic variability of a population, such as bursts of transposons, gene shuffling and gene duplication or simply recombination, or as enhanced environmental demand on the evolving population, the following can be concluded: Based on the simulation experiments carried out in the context of this study, we state that noise

has multi-layered, even contra-intuitive effects when applied to evolving populations under enhanced selective pressure. Yet, noise phenomena and the evolutionary reaction of a population can only be interpreted in context of its genetic variability and plasticity. The mutual dependencies of noise and genetic plasticity characterize the evolutionary strategy of a population. Since we observe a variety of effects even in highly simplified setups as presented here, we speculate that the impact of stochastic perturbations on naturally evolving populations is even more diverse and exciting.

6 Outlook

The model of a gene switch presented here, as well as the developed software allow to cover the demands of evolution *in silico* as has been shown. In particular the usage of regulatory gene networks constituting a reactor population, allows for investigations on a whole range of further interesting questions.

Among them, are the simulations with networks capable to extend their degree of interplay, as well as their number of players. Basically, the software offers all necessary sub-parts to expand the model in this direction.

Under the selective pressure of an ideal dynamic behavior, the study of emerging interaction patterns is conceivable. It will be interesting to see whether recurrent network parts or units appear or whether novel types of network architecture are randomly distributed.

In the former case, the collection and the detailed analysis of such recurring themes may provide a deeper insight into the structural evolution of regulatory gene networks, especially when compared with naturally occurring, biological networks.

Here, the usage of detailed information concerning this topic, stored in public databases seems a fruitful approach. Eventually, the re-connection of theoreticians and experimentalists is tremendously important, if not crucial, for an innovative biology of the future.

List of Figures

List of Tables

References

[1] M. L. Agarwal, W. R. Taylor, M. C. Chernov, O. B. Chernova, and G. R. Stark. The p53 network. *J. Biol. Chem.*, 273:1–4, 1998.

[2] N. N. Alexandrov and N. Go. Biological meaning, statistical significance and classification of local spatial similarities in non-homologous proteins. *Protein Sci.*, 3:866–875, 1994.

[3] R. Antia and M. Lipsitch. Mathematical models of parasite responses to host immune defenses. *Parasitology*, 115:S155–S167, 1997.

[4] A. P. Arkin. Synthetic cell biology. *Curr. Opin. Biotechnol.*, 12:638–644, 2001.

[5] A. Babajide, I. L. Hofacker, M. J. Sippl, and P. F. Stadler. Neutral networks in protein space: A computational study based on knowledge-based potentials of mean force. *Folding & Design*, 2:261–269, 1997.

[6] N. Barkai and S. Leibler. Robustness in simple biochemical networks. *Nature*, 387:913–917, 1997.

[7] P. B. Becker. *Drosophila* chromatin and transcription. *Semin. Cell. Biol.*, 6:185–190, 1995.

[8] J. U. Bowie, R. Luthy, and D. Eisenberg. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 235:164–70, 1991.

[9] T. D. Brock. *Biology of Micro-organisms*. Madigan, Martinko, Parker and Stull, Prentice Hall. Englewood Cliffs, NJ., 1994.

[10] P. Capy, G. Gasperi, C. Biemont, and C. Bazin. Stress and transposable elements: Co-evolution or useful parasites? *Heredity*, 85:101–106, 2000.

[11] N. Carmi, B. R. Shameelah, and R. R. Breaker. Cleaving DNA with DNA. *Proc. Natl. Acad. Sci. USA*, 95:2233–2237, 1998.

[12] G. Casari and M. J. Sippl. Structure-derived hydrophobic potential: Hydrophobic potentials derived from X-ray structures of globular proteins is able to identify native folds. *J. Mol. Biol.*, 224:725–732, 1992.

[13] T. Cech. RNA as an enzyme. *Scientific American*, 11:76–84, 1986.

[14] O. Cinquin and J Demongeot. Positive and negative feedback: Striking balance between necessary antagonists. *J. theor. Biol.*, 216:229–241, 2002.

[15] P. S. Covello and M. W. Gray. On the evolution of RNA editing. *Trends Genet.*, 9:265–268, 1993.

[16] F. Crick. The biological replication of macromolecules. *Symp. Soc. Exp. Biol.*, 12:138, 1958.

[17] M. Eigen. Selforganization of matter and the evolution of biological macromolecules. *Die Naturwissenschaften*, 10:465–523, 1971.

[18] M. B. Elowitz and S. Leibler. A synthetic oscillatory network of transcriptional regulators. *Nature*, 403:335–338, 2000.

[19] M. B. Elowitz, A. Levine, E. Soggia, and P. Swain. Stochastic gene expression in a single cell. *Science*, 297:1183–1186, 2002.

[20] M. L. Espinas, E. Jimenez-Garcia, A. Vaquero, S. Canudas, J. Bernues, and F. Azorin. The N-terminal POZ domain of GAGA mediates the formation of oligomers that bind DNA with high affinity and specificity. *J. Biol. Chem.*, 23:16461–16469, 1999.

[21] M. Famulok. Oligonucleotide aptamers that recognize small molecules. *Curr. Opin. Struct. Biol.*, 9:324–329, 1999.

[22] W. Fontana, W. Schnabl, and P. Schuster. Physical aspects of evolutionary optimization and adaptation. *Phys. Rev. A*, 40:3301–3321, 1989.

[23] W. Fontana and P. Schuster. A computer model of evolutionary optimization. *Biophys. Chem.*, 26:123–147, 1987.

[24] W. Fontana and P. Schuster. Continuity in evolution. On the nature of transitions. *Science*, 280:1451–1455, 1998.

[25] W. Fontana and P. Schuster. Shaping space. the possible and the attainable in RNA genotype-phenotype mapping. *J. Theor. Biol.*, 194:491–515, 1998.

[26] T. S. Gardner, C. R. Cantor, and J. J. Collins. Construction of a genetic toggle switch in *E. coli*. *Nature*, 403:339–342, 2000.

[27] D. T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comp. Phys.*, 22:403–434, 1976.

[28] D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, 81:2340–2361, 1977.

[29] S. Govindarajan, R. Recabarren, and R. A. Goldstein. Estimating the total number of protein folds. *Proteins*, 35:408–414, 1999.

[30] H. Granok, B. A. Leibovitch, C. D. Shaffer, and S. C. Elgin. Chromatin. Ga-ga over GAGA factor. *Curr. Biol.*, 3:238–241, 1995.

[31] C. Guerrier-Takada and S. Altman. Catalytic activity of an RNA molecule prepared by transcription *in vitro*. *Science*, 223:285–286, 1984.

[32] C. Guerrier-Takada, K. Gardiner, T. Marsh, N. Pace, and S. Altman. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell*, 35:849–857, 1983.

[33] C. C. Guet, M. B. Elowitz, W. Hsing, and S. Leibler. Combinatorial synthesis of genetic networks. *Science*, 296:1466–1470, 2002.

[34] R. W. Hamming. Error detecting and error correcting codes. *Bell. Syst. Tech. J.*, 29:147–160, 1950.

[35] J. Hasty, J.Pradines, M. Dolnik, and J. J. Collins. Noise-based switches and amplifiers for gene expression. *Proc. Natl. Acad. Sci. USA*, 97:2075–2080, 2000.

[36] J. Hasty, D. McMillen, F. Isaacs, and J. J. Collins. Computational studies of gene regulatory networks: *in numero* molecular biology. *Nature*, 2:268–279, 2001.

[37] I. L. Hofacker. *The rules for the evolutionary game for RNA: A statistical characterization of the sequence to structure mapping in RNA*. PhD thesis, University of Vienna, 1994.

[38] J. H. Holland. *Adaptation in natural and artificial systems: An introductory analysis with application to biology.* University of Michigan Press, 1975.

[39] M. A. Huynen, P. F. Stadler, and W. Fontana. Smoothness within ruggedness: The role of neutrality in adaptation. *Proc. Natl. Acad. Sci. USA*, 93:397–401, 1996.

[40] A. D. Johnson, A. R. Poteete, G. Lauer, R. T. Sauer, G. K. Ackers, and M. Ptashne. Lambda repressor and cro-components of an efficient molecular switch. *Nature*, 294:217–223, 1981.

[41] M. Kimura. *Neutral theory of molecular evolution.* Cambridge University Press, 1983.

[42] J. R. Koza, F. H. Bennett, M. A. Keane, and D. Andre. *Genetic programming III: Darwinian invention and problem solving.* Morgan Kaufmann, 1998.

[43] D.C. Krakauer and A. Sasaki. Noisy clues to the origin of life. *Proc. R. Soc. Lond. B. Biol. Sci.*, 269:2423–8, 2002.

[44] W. E. Lonnig and H. Saedler. Plant transposons: Contributors to evolution? *Gene*, 205:245–53, 1997.

[45] J. S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structures. *Biopolymers*, 29:1105–1119, 1990.

[46] J. F. McDonald. Evolution and consequences of transposable elements. *Curr. Opin. Genet. Dev.*, 3:855–64, 1993.

[47] B. J. Meyer, R. Maurer, and M. Ptashne. Gene regulation at the right operator (OR) of bacteriophage lambda. II. OR1, OR2, and OR3: their roles in mediating the effects of repressor and cro. *J. Mol. Biol.*, 139:163–194, 1980.

[48] G. F. Mitchell. Co-evolution of parasites and adaptive immune responses. *Immunol. Today*, 12:A2–5, 1991.

[49] J. Modod and F. Jacob. General conclusions: Teleonomic mechanisms in cellular metabolism, growth and differentiation. *Cold Spring Harb. Symp. Quant. Biol.*, 26:389–401, 1961.

[50] S. J. O'Brien, M. E. Roelke, L. Marker, A. Newman, C. A. Winkler, D. Meltzer, L. Colly, J. F. Evermann, M. Bush, and D. E. Wildt. Genetic basis for species vulnerability in the cheetah. *Science*, 227:1428–1434, 1985.

[51] S. J. O'Brien, D. E. Wildt, M. Bush, T. M. Caro, C. FitzGibbon, I. Aggundey, and R. E. Leaky. East african cheetahs: Evidence for two population bottlenecks? *Proc. Natl. Acad. Sci. USA*, 84:508–511, 1987.

[52] T. O'Brien, R. C. Wilkins RC, C. Giardina, and J. T. Lis. Distribution of GAGA protein on *Drosophila* genes *in vivo*. *Genes Dev.*, 9:1098–1110, 1995.

[53] D. H. Ohlendorf and B. W. Matthews. Structural studies of protein-nucleic acid interactions. *Annu. Rev. Biophys. Bioeng.*, 12:259–284, 1983.

[54] T. Ohta. The nearly neutral theory of molecular evolution. *Annu. Rev. Ecol. Syst.*, 23:263–286, 1992.

[55] J. G. Omichinski, P. V. Pedone, G. Felsenfeld, A. M. Gronenborn, and G. M. Clore. The solution structure of a specific GAGA factor-DNA complex reveals a modular binding mode. *Nat. Struct. Biol.*, 4:1072–8368, 1997.

[56] D. J. Patel. Structural analysis of nucleic acid aptamers. *Curr. Opin. Chem. Biol.*, 1:32–46, 1997.

[57] P. V. Pedone, R. Ghirlando, G. M. Clore, A. M. Gronenborn, G. Felsenfeld, and J. G. Omichinski. The single Cys2-His2 zinc finger domain of the GAGA protein flanked by basic residues is sufficient for high-affinity specific DNA binding. *Proc. Natl. Acad. Sci. USA*, 7:2822–2826, 1996.

[58] A. Pohorille and D. Deamer. Artificial cells: Prospects for biotechnology. *Trends Biotechnol.*, 20:123–128, 2002.

[59] M. Ptashne, A. Jeffrey, A. D. Johnson, R. Maurer, B. J. Meyer, C. O. Pabo, T. M. Roberts, and R. T. Sauer. How the lambda repressor and cro work. *Cell*, 19:1–11, 1980.

[60] J. W. Raff, R. Kellum, and B. Alberts. The *Drosophila* GAGA transcription factor is associated with specific regions of heterochromatin throughout the cell cycle. *EMBO J.*, 24:5977–5983, 1994.

[61] W. S. Reznikoff. The lactose operon-controlling elements: A complex paradigm. *Mol. Microbiol.*, 6:2419–2422, 1992.

[62] D. Rhodes and A. Klug. Zinc fingers. *Sci. Am.*, 2:56–59,62–65, 1993.

[63] P. Schuster. A testable genotype-phenotype map: Modeling evolution of RNA molecules. In Michael Lässig and Angelo Valleriani, editors, *Biological evolution and statistical physics*. Springer-Verlag, Berlin, 2002.

[64] P. Schuster and P. F. Stadler. Networks in molecular evolution. *Complexity*, 8:34–42, 2002.

[65] M. L. Simpson, G. S. Sayler, J. T. Fleming, and B. Applegate. Whole-cell biocomputing. *Trends Biotechnol.*, 19:317–323, 2001.

[66] M. J. Sippl. Calculation of conformational ensembles from potentials of mean force - An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.*, 213:859–883, 1990.

[67] M. J. Sippl. Boltzmann's principle, knowledge-based mean fields and protein folding. an approach to the computational determination of protein structures. *J. Computer-Aided Mol. Design*, 7:473–501, 1993.

[68] M. J. Sippl. Recognition of errors in three-dimensional structures of proteins. *Proteins*, 17:355–362, 1993.

[69] J. M. Smith. Natural selection and the concept of protein space. *Nature*, 225:563–564, 1970.

[70] R. V. Sole, I. Salazar-Ciudad, and J. Garcia-Fernandez. Common pattern formation, modularity and phase transitions in a gene network model of morphogenesis. *Physica A*, 305:640–654, 2002.

[71] S. Spiegelman. An approach to the experimental analysis of precellular evolution. *Quart. Ref. Biophys.*, 4:213–253, 1971.

[72] A. V. Spirov and A. B. Kazansky. The inverse problem for dynamical models of genetic networks: Fast coarse-grained solutions by evolutionary computations. 3rd Int. Conf. on Soft Computing and Measurements, St.Petersburg, 2000.

[73] P. F. Stadler. The genotype-phenotype map. Submitted, 2002.

[74] S. Suerbaum. Genetic variability within *Helicobacter pylori*. *Int. J. Med. Microbiol.*, 290:175–81, 2000.

[75] J. M. Vilar, C. C. Guet, and S. Leibler. Modeling network dynamics: The lac operon, a case study. *J. Cell. Biol.*, 161:471–476, 2003.

[76] J. M. Vilar, H. Y. Kueh, N. Barkai, and S. Leibler. Mechanisms of noise-resistance in genetic oscillators. *Proc. Natl. Acad. Sci. USA*, 99:5988–5992, 2002.

[77] T. U. Vogel, D. T. Evans, J. A. Urvater, D. H. O'Connor, A. L. Hughes, and D. I. Watkins. Major histocompatibility complex class I genes in primates: Co-evolution with pathogens. *Immunol. Rev.*, 167:327–337, 1999.

[78] B. Vogelstein, D. Lane, and A. J. Levine. Surfing the p53 network. *Nature*, 408:307–310, 2000.

[79] G. von Dassow and G. M. Odell. Design and constraints of the *Drosophila* segment polarity module: Robust spatial patterning emerges from intertwined cell state switches. *J. Exp. Zool.*, 294:179–215, 2002.

[80] G. Wall, P.D. Varga-Weisz, R. Sandaltzopoulos, and P. B. Becker. Chromatin remodeling by GAGA factor and heat shock factor at the hypersensitive *Drosophila* hsp26 promoter *in vitro*. *EMBO J*, 8:1727–1736, 1995.

[81] A. Wernitznig. *RNA optimization in flow reactors: A study in silico*. PhD thesis, University of Vienna, 2001.

[82] N. Yildirim and M. C. Mackey. Feedback regulation in the lactose operon: A mathematical modeling study and comparison with experimental data. *Biophys. J.*, 84:2841–2851, 2003.

[83] C. T. Zhang. Relations of the number of protein sequences, families and folds. *Protein Eng.*, 10:757–761, 1997.

[84] L. Zhang and J. Skolnick. What should the Z-score of native protein structures be ? *Prot. Sci.*, 7:1201–1207, 1998.

[85] M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamic and auxiliary informations. *Nucl. Acid. Res.*, 9:133–148, 1981.

Mag. Stefanie Widder

Institute for
Theoretical Chemistry and
Structural Biology
Währingerstr. 17
1090 Vienna, Austria

Stefanie Widder

Personal data

Date of birth	January 10, 1975
Marital status	single
Nationality	Austrian
Languages	german (mother tongue), english, spanish, some danish

Education

2000 – 2003	Graduate studies at the Institute for Theoretical Chemistry and Structural Biology in the group of Prof. Peter Schuster, University of Vienna
1993 – 2000	Studies in Genetics/Biology at Vienna University, Austria and Aarhus University, Denmark
2000	Master of Science (Mag.) Diploma thesis: *The structural organization and expression of* P*-derived neogene cluster in* D. guanche with Prof. Dieter Schweizer, University of Vienna Diploma exam passed with distinction
1997 – 1998	Characterization of 2'-5' oligoadenylate synthetase in the group of Prof. Just Justesen, Aarhus University
1993	Graduation (Matura) with distinction
1985 – 1993	Highschool (Gymnasium)

Professional activities

2000 – 2003	Dissertation at the Institute for Theoretical Chemistry and Structural Biology, University of Vienna, *Self-Regulating Gene Switches and Molecular Evolution*
2001 – 2003	Participation at "Women's mentoring project" University of Vienna, Mentor Prof. Immanuel Bomze
2001	Participation at the Oberwolfach School "Mathematical Challenges in Molecular Biology" Mathematisches Forschungsinstitut Oberwolfach, Oberwolfach, Germany
2001	Participation at the Conference "Theoretical biology of adaptation" Tihany, Hungary
2001	Participation at the Complex Systems Summer School 2001 at the Santa Fe Institute, Santa Fe, NM, USA
1999	Participation at the Annual Drosophila Meeting, Seattle, USA
1999	Participation at the Regional Drosophila Meeting, Hannover, Germany

Publications

Widder, S., S. Marsland, C. Witwer and A. Janulevicius, 2001, Transposition - the elixir of life? Proceedings of CSSS 2001

Widder, S., The structural organization and expression of the *P*-derived neogene cluster in *D. guanche*. Master thesis 2000

Hartmann, R., H. S. Olsen, **S. Widder**, R. Joergensen and J. Justesen, 1998, p59OASL, a 2'- 5' oligoadenylate synthetase like protein: a novel human gene related to the *2'- 5' oligoadenylate synthetase* family. Nucleic Acids Research **26**: 4121 – 4128

Widder, S., I. Hofacker, C. Flamm and P. Stadler, Vienna RNA package - RNA secondary structure prediction and comparison, poster, TBA Tihany, Hungary

Widder, S., The "*P*-repressor-like" neogene cluster in *D. guanche*, poster, Annual Drosophila Meeting 1999, Seattle, USA