Gene Regulatory Networks: Dynamics and Stability

Troein, Carl

2007

GENE REGULATORY NETWORKS: DYNAMICS AND STABILITY

Carl Troein

Department of Theoretical Physics
Lund University

Thesis for the degree of Doctor of Philosophy

Thesis Advisor: *Carsten Peterson*
Faculty Opponent: *Stefan Bornholdt*

To be presented, with the permission of the Faculty of Science of Lund University, for public criticism in lecture hall F of the Department of Theoretical Physics on Friday, the 13th of April 2007, at 13.15.

Organization	Document name
LUND UNIVERSITY Department of Theoretical Physics Sölvegatan 14A SE–223 62 LUND Sweden	**DOCTORAL DISSERTATION** Date of issue: March 2007 Subject designator: Project name:
Author: Carl Troein	Sponsoring organization:

Document title: Gene Regulatory Networks: Dynamics and Stability

Abstract

Life as we know it is based on cells that use proteins and RNA to carry out metabolism, self-replication, and other essential tasks. The genes that code for these molecules are encoded in DNA, and through the processes of transcription and translation the cell expresses its genes. Some proteins are transcription factors that regulate the transcription rate of genes, so genes interact and form a *gene regulatory network*.

In a *random Boolean network* the genes are modeled as being either ON or OFF, and the regulatory interactions are drawn from some ensemble that may be based on biological observations. Here, the average behavior of observables of dynamics (e.g., attractor count) and stability (e.g., robustness to perturbations) is studied, both in the original Kauffman model and in models based on data from yeast.

Signal transduction, the propagation of information about the external and internal environment of the cell, often affects transcription factors, thereby altering gene expression levels. Signaling pathway profiling is proposed as a way to reduce the complexity of microarray data and find biologically relevant signals. The core regulatory system of embryonic stem cells is a concrete example of a network where attractor basins and stability are important for biological function, and we explore its dynamics in a continuous model. Finally, the what effect transcriptional regulation has on fitness is studied in the context of metabolism in a very simple system, and the benefit of regulation is made clear.

Key words: gene regulatory networks, transcriptional regulation, random Boolean networks, Kauffman networks, nested canalyzing rules, signaling pathway profiling, stem cell regulation, metabolic pathway, fitness

Classification system and/or index terms:

Supplementary bibliographical information:

Language: English

ISSN and key title:

ISBN: 978-91-628-7118-5

Recipent's notes:

Number of pages: 169

Price:

Security class:

DOKUMENTDATABLAD enl SIS 61 41 21

Distributor

Carl Troein

Department of Theoretical Physics, Sölvegatan 14A, SE–223 62 Lund, Sweden

Signature Carl Troein Date 2007-03-05

Sammanfattning

och han talar med bönder på böndernas sätt
men med lärde män på latin.
— E. A. Karlfeldt, *Sång efter skördeanden*

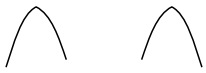
Alla levande varelser består av celler. En cell innehåller ett stort antal kemiska föreningar, och särskilt viktiga är proteiner och DNA. Proteiner utför många livsnödvändiga kemiska reaktioner, och DNA består av gener som beskriver hur dessa proteiner konstrueras. En gen kan vara uttryckt, så att det protein den beskriver – kodar för – tillverkas, men den kan också vara avstängd om proteinet inte behövs.

Förmågan att reagera på förändringar i omvärlden är en viktig egenskap hos alla former av liv. Signaler kan förmedlas från cellens yttre eller inre miljö via kedjor av kemiska reaktioner som ofta involverar proteiner. De proteiner som på så vis påverkas kan börja göra saker de inte gjorde innan. Framför allt kan vissa av dem – transkriptionsfaktorer – binda till DNA och styra vilka gener som uttrycks. Somliga styrda gener kodar själva för transkriptionsfaktorer, och därför har cellen ett helt nätverk av gener som kontrollerar varandra.

För att allmänt förstå hur sådana *genregleringsnätverk* fungerar kan man skapa modeller av dem, och studera modellernas uppförande. Mer grovhuggna modeller är enklare att arbeta med och dra slutsatser från, men kan också avvika mer från verkligheten. Särskilt enkla är booleska nätverk, där varje gen antas vara antingen PÅ eller AV, utan mellanlägen. I sådana nätverk är attraktorer, tillstånd som inte går att lämna, av intresse för sin koppling till biologins celltyper. Det är dessutom viktigt att ett genregleringsnätverk inte är alltför känsligt för störningar. I artiklarna I–III undersöker vi hur olika sorters slumpmässigt byggda booleska nätverk klarar sig i dessa båda avseenden, baserat på tanken att det genomsnittliga beteendet säger någonting om biologins nätverk, med förbehållet att riktiga nätverk har formats av evolutionen under lång tid och inte uppstått ur tomma intet.

Andra halvan av avhandlingen handlar om något mer konkreta problemställningar. I artikel IV försöker vi uppskatta hur aktiva några signalvägar är genom att mäta uttrycktsnivån hos de gener som – åtminstone i vissa fall – styrs av dem. Detta berör också frågan om hur relevanta signalvägarna är som beskrivningar av hur systemet fungerar.

Stamceller är celler som kan ge upphov till många olika sorters celler och ändå själva finnas kvar, och även om vi inte kan leva utan dem måste de hållas under sträng kontroll. Det centrala genregleringsnätverket i embryonala stamceller är helt nyligen upptäckt, och hur det styr stamcellernas öde undersöks i artikel V. Artikel VI, slutligen, handlar om en riktigt grundläggande aspekt av genregleringsnätverk: under vilka omständigheter det alls lönar sig att reglera hur mycket en gen uttrycks.



To Linda and —

This thesis is based on the following publications[1]:

I Björn Samuelsson and Carl Troein,
Superpolynomial growth in the number of attractors in Kauffman networks
Physical Review Letters **90**, 098701 (2003).

II Stuart Kauffman, Carsten Peterson, Björn Samuelsson, and Carl Troein,
Random Boolean network models and the yeast transcriptional network
Proceedings of the National Academy of Sciences of the USA **100**, 14796–14799 (2003).

III Stuart Kauffman, Carsten Peterson, Björn Samuelsson, and Carl Troein,
Genetic networks with canalyzing Boolean rules are always stable
Proceedings of the National Academy of Sciences of the USA **101**, 17102–17107 (2004).

IV Thomas Breslin, Morten Krogh, Carsten Peterson, and Carl Troein,
Signal transduction pathway profiling of individual tumor samples
BMC Bioinformatics **6**, 163 (2005).

V Vijay Chickarmane, Carl Troein, Ulrike A. Nuber, Herbert M. Sauro, and Carsten Peterson,
Transcriptional dynamics of the embryonic stem cell switch
PLoS Computational Biology **2**, e123 (2006).

VI Carl Troein, Dag Ahrén, Morten Krogh, and Carsten Peterson,
Regulating metabolic pathways to optimize fitness
LU TP 07-08, submitted (2007).

[1]For collaborations within Complex Systems / CBBP (Computational Biology & Biological Physics), or with other theory groups, authors are listed alphabetically. For collaborations with biology groups, the ordering follows other conventions.

During my time as a PhD student, I have also co-authored the following publications:

- Anders Irbäck and Carl Troein,
  Enumerating designing sequences in the HP model
  Journal of Biological Physics **28**, 1–15 (2002).

- Lao H. Saal, Carl Troein, Johan Vallon-Christersson, Sofia Gruvberger, Åke Borg, and Carsten Peterson,
  BioArray Software Environment (BASE): A platform for comprehensive management and analysis of microarray data
  Genome Biology **3**, software0003.1–0003.6 (2003).

- Dag Ahrén, Carl Troein, Tomas Johansson, and Anders Tunlid,
  PHOREST: A web-based tool for comparative analyses of expressed sequence tag data
  Molecular Ecology Notes **4**, 311–314 (2004).

- Björn Samuelsson and Carl Troein,
  Random maps and attractors in random Boolean networks
  Physical Review E **72**, 046112 (2005).

- Carl Troein, Johan Vallon-Christersson, and Lao H. Saal,
  An introduction to BASE
  In *Methods in Enzymology* **411**, DNA Microarrays, Part B: Databases and Statistics, A. Kimmel and B. Oliver (eds.), pp. 99–119, Elsevier (2006).

Contents

i

Introduction

The topic of this thesis is gene regulatory networks. Before I can even begin to describe what such networks are and do, let alone go into all the gory details, the context in which they appear must be made clear. The projects I have been involved in for the past few years are all straddling the line between biology and physics, albeit that they – and I – have a somewhat steadier footing on the physics side. Everything I say, or do not say, about biology should thus be seen for what it is: viewed from a physicist's perspective.

The publications, which are presented at the end of this short introduction, progress from the grand, sweeping, and abstract to the small, specific, and fairly concrete. At the same time, biology goes from from being an underlying idea and source of inspiration to forming a more integral part of the work. To understand the ideas presented in the six publications, and to perceive the common threads running through them, it is necessary to know a few things about biology and networks. I will therefore just briefly define and describe life, before narrowing the scope to those details and models that are pertinent to the projects represented in this thesis. Much of sections *i*.2 and *i*.3 is standard textbook material, and more complete and accurate accounts can be found in references [1–4].

In short, the broad and open-ended question behind this work is this: how can living cells control the processes behind *how* they are, *what* they do, and *when* they do it? I will attempt to answer this question, or at least outline some answers, while explaining and exploring how methods from a physicist's toolkit can be applied to such problems.

i.1 Life

"Life," said Marvin, "don't talk to me about life."
— Douglas Adams, *The Hitchhiker's Guide to the Galaxy*

The first step in discussing something as complex as life is to define what we really mean. It may seem like a trivial matter to define "life", because clearly

a cat or a mushroom or even a yeast cell is alive, while a rock or a computer or a glass of water is not. But just as a pile of grains eventually stops being a pile if you remove one tiny grain at a time, even though you could not say that any one particular grain made the difference, there are shades of gray in the spectrum of alive-ness. There is no real consensus on what constitutes life, because reasonable definitions will turn out to exclude things that some people would like to view as living, or include things that others would prefer not to see included. Various definitions have been proposed, some bolder than others [5,6]. For the purposes of this discussion, let us adopt a broad definition and consider something to be alive if it is self-replicating, responsive to its environment, and evolvable.

By self-replicating I mean that if you have one, and the conditions are right, you may soon have two. A prime example is the aphids that have been plaguing my habanero plant; only a few days after I remove all but a few of them, the plant will again be host to a sizeable population. The ability to self-replicate is not a sufficient criterion for life, because even a simple salt crystal or an open flame can reproduce if the circumstances are right.

The second criterion, responsiveness to the environment, is not a particularly well-defined concept, but what I mean is that a living thing is able to detect and react to changes in a non-trivial way. For example, a yeast cell may respond to the sudden availability of glucose by absorbing and metabolizing it, consuming the glucose to grow bigger. In contrast, sulfuric acid would react to the glucose by reacting with it, oxidizing it much like it does with just about everything else.

Evolvability, finally, is where things gets interesting [7]. This requirement on life states that a living thing has some properties that, when it reproduces, are imperfectly inherited by its offspring. In biological terms, the changes underlying the imperfect inheritance are called mutations. That some individuals, depending on their inherited properties, are more likely than others to reproduce is referred to as *natural selection*, or often just *selection*, especially if humans are directly involved in affecting the reproduction probabilities. While evolvability may seem like the most complicated of the three requirements, it is in fact the easiest to fulfill, simply because pobody's nerfect – most forms of self-replication will typically result in an imperfect copy.

For an example of evolvability consider again the aphids, who inherit their brown hue from their parent(s)[1]. There is a small chance that a mutation will give an aphid a lighter color that it would otherwise have inherited. Dark aphids are easier for me to find and squash, so an aphid that happens to be born fairer stands a better chance of reproducing; it has higher fitness. The descendants if this aphid will soon be a growing fraction of the population,

[1]These wicked sap-sucking creatures (my arch-nemesises, if you will) commonly practice parthenogenesis, or virgin birth [8].

and the darker aphids will go extinct. Thus, after a number of generations, the aphids living on my plant are likely to differ in color from their ancestors and be harder for me to kill, only because I tried to get rid of them in the first place. This is evolution at work.

i.2 Real life

Seen in the light of evolution, biology is, perhaps, intellectually the most satisfying and inspiring science. Without that light it becomes a pile of sundry facts – some of them interesting or curious but making no meaningful picture as a whole.

— Theodosius Dobzhansky

Having defined life, let us turn to the implementation of it. On earth there is, at least in any traditional and reasonable sense, only one form of life: the good old-fashioned biological kind that encompasses everything from the puniest bacterium and slimiest mold to the mightiest whale and the loftiest redwood. There is good reason to hold this view, that all organisms great and small are but variations on a single theme. Even between the most dissimilar species, as one looks beyond superficial appearances, the differences are small compared to the similarities [9–11].

All living things consist of cells, small sacks or bubbles whose foremost function is to keep their contents safely together, much like the function of an office building is to keep its workers gathered and able to work. Inside the plasma membrane, the lipid bilayer which encloses the cell, is a thick soup of organic macromolecules mixed with a rich broth of smaller molecules. Cells come in an assortment of shapes and sizes. Most are microscopically small, but some are large enough to be seen by the unaided eye. Only a tiny fraction of all cells are lumped together to form multicellular organisms such as animals and plants. The rest make up a wide variety of single-celled organisms that move around and reproduce on their own, without strong ties to other cells. Many of these are prokaryotes, comparatively simple cells that are often only a micrometer or less in diameter. Based on details that do differ, such as chemical composition of their plasma membrane, prokaryotes can be classified into two distinct groups: bacteria and archaea. Those cells that are not prokaryotes, including the cells that make up all multicellular life, are called eukaryotes. What earns them their name, and most obviously distinguishes them from prokaryotes, is the presence inside the cell of a membrane-enclosed, cell-like compartment called the nucleus. Almost all eukaryotic cells contain two or more types of compartments of similar description, or organelles, which are indeed believed to be the descendants of prokaryotes that at some point got stuck in other cells.

There is, of course, far more that could be said about what cells look like, what structures they contain, who begot whom, and so forth. But consider instead the molecules that are contained in the cells. Truly central to how cells function, and the basis for my previous claim that all cells are very similar, is a trinity of macromolecules found in absolutely all cells. These are, in no particular order, proteins, DNA, and RNA.

i.2.1 A trinity of heteropolymers

Three for the Kings
Of the elves high in light
— Blind Guardian, *Lord of the Rings*

Proteins are long chains of amino acids, intricately folded into three-dimensional structures. These amino acids are small molecules with a common backbone structure but different side chains and therefore quite different chemical properties. Each protein has a well-determined sequence of amino acids that are usually counted in the hundreds or thousands, and with twenty different amino acids to choose from, the combinatorial possibilities are virtually endless. The amino acid sequence by and large determines the structure of the protein, because the protein folds so as to shield hydrophobic amino acids from water, bring together oppositely charged amino acids, and in other ways arrange itself in a conformation that minimizes free energy. Examples of what folded proteins may look like are shown in figure *i*.1.

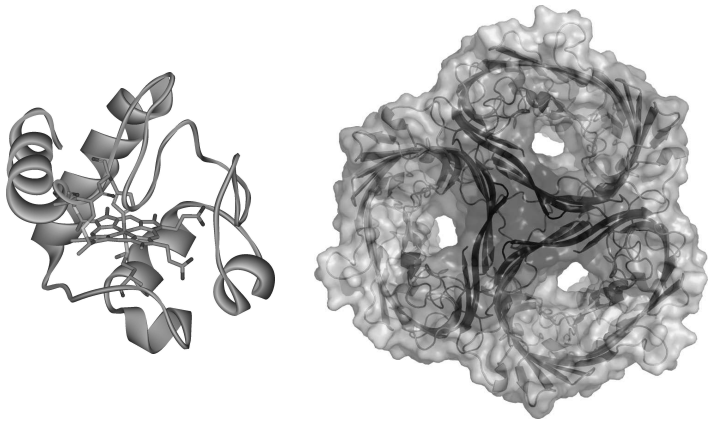


Figure *i*.1: The structures of two proteins, cytochrome c (left) and a sucrose specific porin from *Salmonella* (right), visualized in two somewhat different ways.
Images by Klaus Hoffmeier (left) and Richard Wheeler (right, GFDL).

The structure of a protein is closely linked to its function. By virtue of their extraordinary versatility, proteins are the workhorses of the cell, and there are several broad classes that serve radically different purposes. Some stick together to form webs, tails, stiff pipes, and other large structures that a cell may need, while others actively transport material into, out of, and around the interior of the cell, or even team up to move the cell about in its environment. Still others are enzymes, catalysts that facilitate specific chemical reactions, including those that create the very amino acids that make up all proteins. The chemical processes that involve breaking down, building up, and modifying chemical compounds used by the cell are collectively known as metabolism, and the compounds in question, and then especially the smaller molecules, are called metabolites. Many of them, such as glucose and ATP, are common to virtually all life, and the metabolic pathways through which they are transformed are highly conserved (that is, they vary little between species), as are the enzymes.

Something is missing from this picture if a cell is to be self-replicating. Amino acids must be assembled into proteins, and they must be so according to an inheritable blueprint. That blueprint could not be the proteins themselves, because not all of the proteins that a cell can manufacture are constantly present in the cell. Instead, it is the DNA molecule that fills this role. DNA, or deoxyribonucleic acid, consists of a potentially macroscopically long sugar–phosphate backbone, to which a small molecule is attached at each metaphorical vertebra. These small molecules are called bases, and there are but four of them: adenine (A), cytosine (C), guanine (G), and thymine (T). Just as proteins are heteropolymers of amino acids, so are DNA molecules heteropolymers of nucleotides formed by sugar, phosphate, and the four bases. The bases A, C, G, and T carry information, much like the ones and zeros in a digital computer.

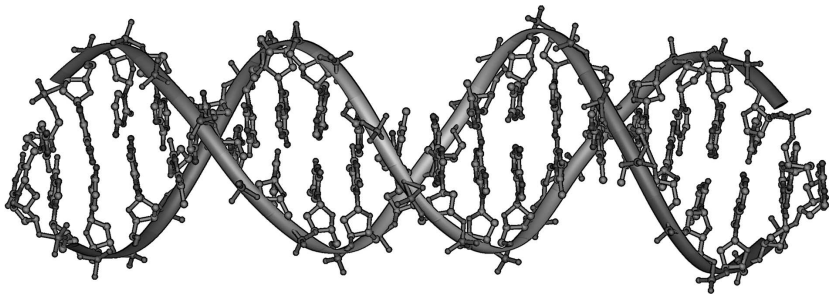


Figure *i*.2: Double-stranded DNA twisted into its normal double helix structure. The deoxyribose backbone is represented by ribbons, while the bases and phosphates are shown in more detail.
Image by Michael Ströck/Wikipedia (GFDL).

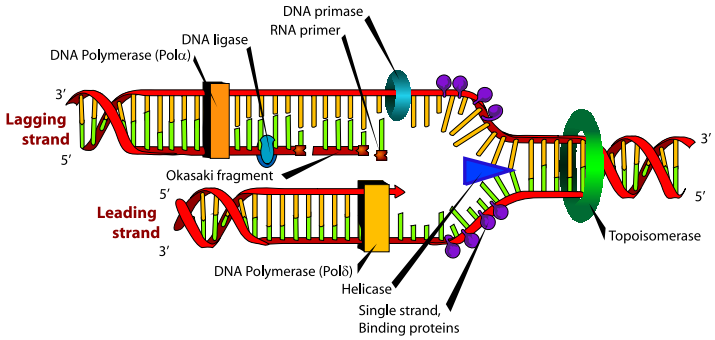


Figure *i*.3: Replication of DNA. Although conceptually simple, the actual process involves a large number of proteins.
Image by Mariana Ruiz.

Crucial to the role of DNA as a blueprint is the concept of base pairing. When two strands of DNA line up in opposite directions, they can be stuck together by hydrogen bonds between their bases, but only where adenine meets thymine or cytosine meets guanine. Other pairings do not leave room for all atoms, or are otherwise disfavored. Thus every DNA sequence has a complement, which is the sequence reversed and with bases substituted. For example, the complement of GAGAACAT is ATGTTCTC. Base pairing makes it possible to duplicate DNA by first assembling the complement of a sequence and then the complement of that. The DNA in cells is normally found in a double-stranded form, a pair of complementary sequences paired up and twisted into a double helix by forces between the nucleotides, as illustrated in figure *i*.2. Making a copy of the DNA is a matter of temporarily (and locally) separating the two strands and copying both. Figure *i*.3 shows this process in some detail.

Biologists make a distinction between an organism's phenotype, the set of traits that describe how the organism looks and functions, and its genotype, which is a set of genes, units of hereditary information usually carried by DNA. In modern parlance, a gene is a stretch of DNA that codes for a protein or, as we shall soon see, something else. The words *gene expression* loosely sum up the processes that lead from DNA to gene products (such as proteins).

To assemble proteins from the genetic information in DNA, the cell makes use of base pairing, but not between DNA and DNA. Ribonucleic acid, or RNA for short, is universally the mediator for protein sequence information from DNA to proteins. It is a molecule quite similar to DNA, but the difference in its backbone composition means that RNA does not form long double helices, and it is less stable than DNA over time. RNA primarily uses the same

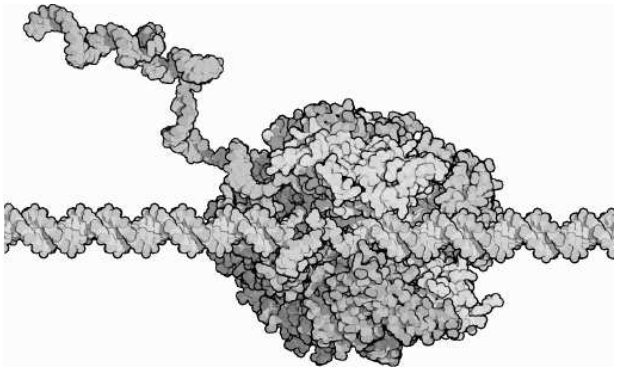


Figure *i*.4: The structure of RNA polymerase during the process of transcription. DNA is the rope-like double helix, and the long molecule exiting to the upper left is RNA. Part of the RNAP protein complex is not shown, to allow the DNA and RNA inside it to be seen.
Image by David S. Goodsell.

four bases as DNA, except that uracil replaces thymine. The process of producing an RNA molecule from a DNA template is called *transcription*, and the protein that nucleotide by nucleotide carries out the transcription is known as RNA polymerase (RNAP). The process is illustrated in figure *i*.4.

A piece of RNA that describes the amino acid sequence of a protein is referred to as messenger RNA (mRNA), to distinguish it from RNA with other functions. In mRNA (and DNA), a protein build from twenty different amino acids must be represented using only four different bases. The amino acid sequence is encoded in a straightforward way, with the bases of mRNA organized into triplets called codons. With four bases, there is a total of 64 different codons that can be used to code for amino acids and flow control (start/stop signals). The *genetic code* describes how the codons are mapped to amino acids. As there are more codons than there are amino acids, there is some redundancy in the code, and the third base of each triplet on average carries less information than the first two. Furthermore, codons that are similar often code for amino acids with similar properties, which invites to speculation over how the genetic code has evolved [12]. Not all species have identical genetic codes, but only one quarter of all codons are known to ever take on a different meaning [13].

RNA has other functions, beyond being a passive messenger. Of particular interest is ribosomal RNA (rRNA), which is at the heart of the ribosomes, large RNA–protein complexes in which all proteins are assembled from the information carried by mRNA. During translation, the process of protein synthesis by the ribosomes, the physical mapping from codons to amino acids

Figure *i*.5: A transport RNA (tRNA) molecule, used by the cell to translate mRNA into proteins. All tRNAs have the same overall shape, but the bases of the codon-matching anticodons are different, as are the amino acid binding parts.
Image by Neil Voss.

is performed by yet another type of RNA, dubbed transfer RNA (tRNA). Its structure is shown in figure *i*.5. Each variety of tRNA transports a specific amino acid, and contains an anticodon, a triplet of bases that can pair with one or a few codons. When a strand of mRNA passes through a ribosome, its codons are one by one exposed to tRNAs, and the correct tRNA will engage and thereby bring its amino acid into position to be attached to the growing protein chain.

This brings us full circle. A variety of proteins make the building blocks of macromolecules, and produce DNA and RNA from DNA. Meanwhile, RNA and proteins together make proteins from information carried from DNA by RNA. This fits the description of a self-replicating system on our checklist for life. Evolvability, too, is fulfilled, because DNA will inevitably suffer mutations from time to time, in particular during replication. The final criterion, responsiveness to the environment, will be given more attention in the next section.

i.3 Signaling and gene regulation

Spring into action!
— Topato Potato

It is primarily through proteins that the cell senses and responds to its environment. Receptor proteins inside the cell or embedded in its surface can be activated by the binding of specific molecules (ligands) and other changes in the immediate environment, and will in the activated form have chemical properties that differ in small but significant ways. A change in ligand concentration constitutes a signal that a receptor may pass on by changing its ability to catalyze reactions that modify other proteins. These reactions frequently involve the addition or removal of small chemical groups such as phosphate. More generally expressed, a signaling event may cause a protein or protein complex to generate another signaling event. By chaining several events together into *signaling pathways*, it is possible to get very strong non-linear responses, but also to launch the same response to different external signals, make a response conditional on multiple input signals, and so forth.

To be useful, signaling pathways must accomplish something more than just the propagation of signals. In some cases it may suffice to activate or deactivate existing proteins, but in many cases it is necessary to synthesize more proteins to deal with whatever condition prompted the signal, or, conversely, to shut down the production of proteins. Regulating the rate of protein production can be done in a number of ways, for example by altering the translation rate or the degradation rate of mRNA. Yet the primary way to regulate gene expression, especially when switching between zero and non-zero rather than just fine-tuning the expression level, is to target the transcription rate. Proteins are, as previously noted, a highly versatile class of heteropolymers. It should therefore come as no surprise that there is a group of proteins that can bind to DNA at more or less well-defined nucleotide sequences. These are *transcription factors*, which can in detail control the transcription rates of genes, and thereby change the corresponding protein production rates. Figure *i*.6 shows part of a signaling pathway that goes all the way from a receptor protein to transcriptional regulation of genes.

RNA polymerase cares about the DNA sequence when it starts transcribing DNA into RNA, and will only bind and start at certain sites, called promoters. By binding in the vicinity of promoters, upstream of the genes along the DNA, transcription factors regulate the rate of transcription of their target genes, either by recruiting RNAP to the promoter or by blocking it from binding. This description of transcriptional regulation is, like everything else I have said, somewhat simplified. In reality, DNA is heavily coiled up around proteins called histones, and different parts may be more or less accessible. Both histones and DNA can be modified (e.g., by methylation) in ways that further

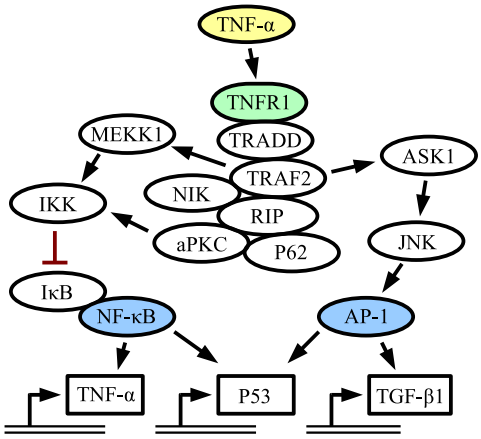


Figure *i.6*: A schematic view of a real signaling pathway. When a ligand (TNF-*α*) binds to a receptor at the cell surface, a series of biochemical events are triggered, eventually leading to activation or repression of the expression of a set of genes.

affect transcription rates. The packing of DNA also means that points far apart along its length can end up close in space, so some transcription factors can act over long distances [14].

It should also be noted that transcription factors would be of little use if they were always active. At least some of them need to be at the receiving end of signaling pathways in order to be involved in responses to the environment. The mechanisms by which transcription factors can be activated (or deactivated) are numerous, and include the usual set of modifications: phosphorylation, methylation, dimerization, and so on. In eukaryotes it also matters whether the transcription factors can enter the nucleus, where the cell's DNA resides.

This all paints a picture of transcriptional regulation as something exceedingly complicated. In most, if not all, cases, we can nevertheless view the transcription rate of a gene merely as some function of the concentration levels of a set of transcription factors. What is important to observe is that this applies also to those genes that themselves code for transcription factors. Hence, we may regard the expression of all genes in the cell as being governed by a network of interacting genes. The entire set of genes form a *gene regulatory network*, a directed graph where every node is a gene and every edge represents a regulatory interaction.

Gene regulatory networks have been studied from many angles and with a wide range of tools. Attention has been lavished on structural aspects of the networks, both in terms of statistics on local properties such as small network

motifs [15] and degree distributions [16], and global properties such as fractal dimension [17]. See reference [18] for a recent review.

Making sense of existing networks requires understanding of their evolutionary origins. At the very smallest scale, it is feasible to watch components evolve under high selective pressure, as was recently done with a small regulatory module, the *lac* operon [19, 20] of *Escherichia coli*, for over 500 generations at different lactose concentrations [21]. In an even longer-running experiment, several strains of *E. coli* were allowed to evolve for the better part of two decades, or some 20 000 generations [22]. Clearly it would be impossible to test all hypotheses with experiments of this magnitude, and more so if the aim is to understand how the entire genome of an organism has evolved from nothing. The natural course of action is to turn to computer simulations. Evolution performed by computers has been applied to realistic models of signaling pathways [23] and gene regulatory networks [24], but it has also been used to create complex organisms that live in a digital world very different from ours [25, 26].

Artificial chemistries [27], and in particular autocatalytic sets [28], have a long history as models of the basic processes of life and evolution, but are less suitable for studying higher-level biological systems where complex mechanisms, for example for DNA replication and protein synthesis, are already firmly established. The application of artificial life concepts to systems where such mechanisms are woven into the model is a fairly recent development; see references [23, 29, 30]. Going beyond answering questions about evolutionary processes [26] and fitness properties [31], such studies may shed light on the emergence of gene regulatory networks and other biological networks.

i.3.1 Measurement techniques

There ain't no ant bigger'n a bee's knee for a thousand miles...
— a farmer in *It Came From The Desert*

A little something should be said about how gene expression and regulation can be measured. The amount of mRNA from a very large number of genes can be measured simultaneously using microarrays [32], which are small glass slides covered with tiny but individually identifiable spots that contain single-stranded DNA. Through the process of reverse transcription, mRNA is used to create DNA, which is then fluorescently labeled and poured onto the microarray where it binds only to those spots which its sequence matches. The end result is a rather noisy estimate of expression levels, possibly relative to some reference sample, of up to several tens of thousands of genes.

There will inevitably be gaps and errors in the data. Furthermore, there will be so many measured values that no human can possibly view them all

at the same time or hope to see the big picture, at least not without the help of computers to wash out the noise and purify the useful information. There are many data-driven approaches to reduce the complexity of microarray data. Supervised clustering methods, for example, can be used to build classifiers for identifying new samples as belonging to one of several medically relevant groups, to assist in diagnosing and treating patients. Unsupervised clustering methods can find sets of samples or genes that behave similarly in a set of experiments, to aid in the discovery of new biological subdivisions between samples or causal relationships in gene expression.

Gene expression data reveal transcriptional regulation [33], but microarrays can also be used to more directly infer regulatory interactions. In ChIP–chip experiments [34], one measures how strongly a transcription factor binds to DNA in the regulatory regions of a large number of genes. The "ChIP" step, chromatin immunoprecipitation, aims at extracting those pieces of DNA that the transcription factor binds to. The DNA is fragmented with the transcription factor still attached, and as the transcription factor-bound pieces are plucked out of the solution, the rest of the DNA is discarded. Then, in the "chip" step, the pieces of DNA are applied to a microarray, which will indicate what genes are regulated by the transcription factor under the conditions of the experiment.

These high throughput technologies are immensely efficient tools for revealing the big picture, but the individual measurements are not all that reliable. For investigating smaller details of the regulatory system, more accurate measurements are possible with other methods, and the high throughput results may then provide starting points for further investigations.

There exists an enormous corpus of literature on all things biological, with whole papers dedicated to what I would view as tiny parts of regulatory networks. This is not to say that such studies are needlessly detailed, only that a full description of a biological system is terribly complicated. Luckily, there have been some efforts to gather these little pieces of information from the literature and arrange them into more or less well-structured databases of transcription factors [35], signaling pathways [36], and so forth. To make better use of these database, it is possible to make predictions of what genes a transcription factor regulates, based on its known binding motifs and the the DNA sequence of the studied organism. Hence, genome-wide searches for binding sites can quickly uncover regulatory interactions that involve a great number of transcription factors [37]. This can be one component of integrative approaches that combine data from several sources to expand our knowledge of gene regulatory networks [38–40].

i.3.2 Cell types

And the mice were squealing in my prison cell
— Brendan Behan, *The Auld Triangle*

Depending on how one counts, mammals have around a few hundred different types of cells, such as smooth muscle cells, helper T lymphocytes, and olfactory receptor neurons. Cell type is closely linked with what genes are expressed, and cells will on the whole retain their type for long periods of time without the need for external signals. This is largely the work of transcriptional regulation. If a set of transcription factors are expressed in a given cell type, and they together promote their own expression while preventing the expression of others, the cell type may remain unchanged.

Starting from a single cell, a multicellular organism grows by dividing and specializing cells stepwise until it consists of a large number of tissues and cell types. Dividing and being active carries risks for a cell, and its regulatory system may be damaged in ways that cause it to start dividing unchecked, creating a tumor or other form of cancer. Therefore, numerous safeguards have evolved to prevent this from happening, including triggers for apoptosis (programmed cell death) when errors are detected, and limits on the number of times a cell can divide. Stem cells are those rare cells that even in an adult retain the potential to keep dividing, and to differentiate into cells of various types [41]. They form a hierarchy of cell development, where only a small fraction of the cells live sluggish lives in the least differentiated states, and from there give rise to more specialized stem cells, which in turn beget progenitors to mature cells. Thus, when cells dies, whether by accident, old age, or shedding, their ranks can be replenished promptly.

Stem cells are a way to create many types of cells from a single fertilized egg, a means for the body to fill vacancies when cells die, and a part of the solution to the problem of cancer. However, the properties of stem cells, such as the ability to divide endlessly, make them dangerous if gone bad, so the number of stem cells is kept low. That the population is divided into a smaller number of long-term stem cells and more numerous short-term stem cells, which are more active and prone to differentiate, reflects the need to keep the pool of stem cells healthy and under control. It is crucial that cells do not become stem cells by accident if their regulatory networks are broken, so the transcriptional regulation of stem cell-ness and differentiation can be expected to be extremely robust.

i.4 Boolean networks

```
#define MAYBE 2
```

— a *CodeSOD* on thedailywtf.com

Only by idealizing and approximating can we hope to understand anything. This does not, of course, imply that we should always try to oversimplify things, but the more details we strip away, the more universal may our conclusions be. From a practical point of view, we may need to simplify to be able to perform simulations and other mundane tasks. For a great many genes, a significant portion of the information on their expression levels lies in whether they are at all expressed. In other words, it may be sufficient for our understanding to know whether or not the genes are expressed, without differentiating between levels for those genes that are "on". Reducing gene expression to discrete values greatly simplifies the description of interactions between genes. This is the basis for modeling gene regulatory networks as Boolean networks.

A Boolean variable is, as every computer geek knows, one that can only take the values TRUE and FALSE. We may equally well call the two states 1 and 0 or ON and OFF. If we claim that the expression level of every gene can be reasonably well described using just those two words, then the state of each node of the gene regulatory network is a Boolean variable, and we call the network itself a *Boolean network*. A more in-depth discussion of how Boolean models can apply to gene expression is found in reference [42].

In addition to the network architecture – how the nodes are connected – the behavior of a network depends on how the connected nodes interact. The links (graph edges) that point to a node represent regulatory interactions. In the Boolean world, the expression level of a gene is still some function of the expression levels of those genes that regulate it, but it will necessarily be a Boolean function. As a simple example, a gene might only be expressed when induced by two transcription factors, and this corresponds to a node with two incoming links that feed into the Boolean function AND. Figure *i*.7 shows a larger example, a small Boolean network where one of the nodes has three inputs. At every node there is a Boolean function that determines what the state of the node will be, depending on what the state of its input nodes are. We often refer to these functions as rules for updating the nodes, which makes it clear that it must be specified how, precisely, the nodes are updated.

We are interested in the dynamics of the network, which is to say that we want to know how it behaves as time goes by. Given a network architecture, an assignment of rules, and a network state, it is easy to compute the next state of any node. The problem is that if we update the states of the nodes one by one, the behavior of the network will depend on the updating order. Unlike in a continuous model, state changes are never gradual and we cannot take

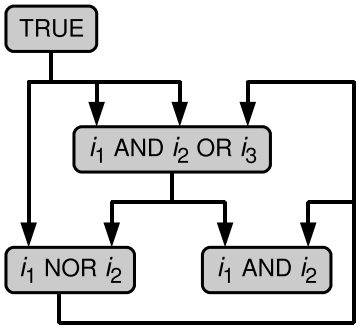


Figure *i*.7: A Boolean network of four nodes, where the number of inputs to the nodes varies from none to three, as does, coincidentally, the number of outputs per node.

shorter steps in time. There is a variety of ways to update the network state asynchronously, or one node at a time: draw a node at random each time, or go through all nodes and then re-randomize the order, or just use the same arbitrary order over and over. However, there is a conceptually simpler way to update the network: all nodes are updated at the same time, so that at time $t + 1$ the state of each node is given by its input nodes at time t.

Regardless of the updating order, there are some simple cases where a Boolean network does not behave like the corresponding continuous model. A gene which negatively regulates itself, as to stabilize its expression at an intermediary level, may be represented in a Boolean model as a node that changes its state every time it is updated. It will look as though the gene expression oscillates, which in reality is a possible but less likely behavior. For larger networks, the synchronous updating scheme, which represents an idealized, noise-free view of the dynamics, tends to suffer the most from artefacts of this type. However, this scheme has the advantage of being deterministic while treating all nodes as equal. It is possible to keep these positive attributes and yet reduce the artefacts, by actively considering the effects of delayed updating of some nodes [43].

The synchronous updating scheme, which is all I will consider from here on, deterministically assigns a successor state to each of the 2^N states of a network with N nodes. Consequently, every state eventually leads to a state that has already been visited, and then round and round in a loop. In other words, every trajectory in the state space ends in an attractor, be it a fixed point or a cycle of some finite length. Figure *i*.8 illustrates this concept for six of the sixteen states in a network of four nodes. One of the states points directly to itself, so it is a fixed point, an attractor of size one. The other five states are part of its basin of attraction, the set of states that lead to the attractor.

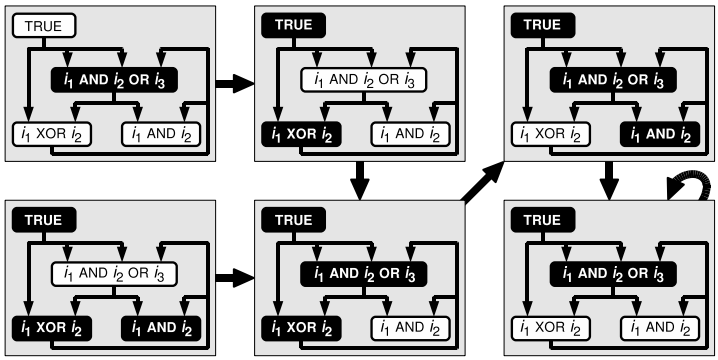
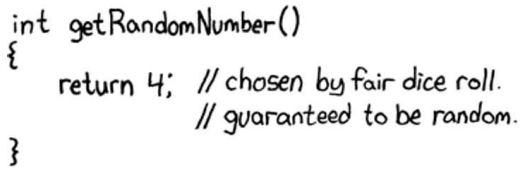


Figure *i*.8: An example of the dynamics of a Boolean network, namely the one in figure *i*.7, under the synchronous updating scheme. Six network states are shown, and the nodes are colored according to their states, with dark for TRUE and white for FALSE. Each node is updated according to the states of its inputs in the previous time step, which leads to the trajectories indicated by the thick arrows. The lower right state points to itself, and is therefore a fixed point of the dynamics.

i.4.1 Random Boolean networks



— XKCD, `xkcd.com`

Boolean networks in the context of biology trace back to the work of Stuart Kauffman in the late 1960s, as do the ideas for what to do with such networks [44]. Much of what I have said about gene regulation was known back then [45], albeit that some of the details were hazy. What was not known, however, was what a complete gene regulatory network looked like. But as always when there is something we do not know in detail, a reasonable approach is to consider the whole ensemble of possible cases, to see what that can tell us. Applied to Boolean networks, this reasoning leads us to generate networks where the network architecture and the rules are chosen randomly. We may then proceed to investigate these *random Boolean networks* [46], in the hope that properties common to most such networks are also shared by the real gene regulatory networks. However, we should keep in mind that lacking the selective pressure to perform a useful task, random networks could potentially be very different from real networks.

There is a class of Boolean networks, called NK models, which is defined as networks with N nodes, all of which have exactly K inputs. The number of distinct states that the input to a node can assume is 2^K, and as the output of a rule can be either TRUE or FALSE for each of those states, there are 2^{2^K} different K-input rules. It was randomly connected networks of this class that Kauffman originally studied, with $K = 2$ and all 16 of the 2-input rules equally probable [44]. To generate a random network – a *realization* of the random network model – each of the N nodes is assigned a rule and two input nodes at random (it may even be the same input node twice).

If we can measure some observable of interest on a network reasonably quickly, we can generate a large number of network realizations and get an estimate of what distribution the observable follows. For example, if we want to know how common it is with nodes that have 25 outputs in networks of size 100, we can generate a large number of such networks and count how many of them contain 25-output nodes. In this case it is fairly obvious that we can also calculate the expected value without actually generating any networks, but far from all observables have that property.

Robustness is an important aspect of any regulatory network. In random Boolean networks it is of interest to know what the average effect of changing the state of a single node is. Depending on the network model such a perturbation may die out, remain the same size, or grow to affect many nodes, and the network dynamics are then referred to as ordered, critical, or chaotic, respectively. The $K = 2$ Kauffman networks are critical, because the {mean number of outputs per node} times {the chance that a rule will give a different output when one input changes} is exactly $2 \cdot \frac{1}{2} = 1$. That is, a small perturbation will on average neither grow to affect the whole network, nor rapidly die down. Papers II and III give examples of networks that are not critical, and introduce other measures of robustness.

It is natural to discuss the dynamics of a Boolean network in terms of its attractor structure, starting from the question of how many attractors there are and how their sizes are distributed. The number of attractors is an observable with some coupling to biology, following the notion that the cell types of multicellular organisms correspond to different attractors of the gene regulatory network. One could hope that Boolean networks with as many nodes as we have genes, just over 25 000 by today's best estimates [47], would on average have an attractor count comparable to the number of cell types. It may seem strange that a gene regulatory network, whose workings are heavily restricted by the need for fitness, would behave essentially like a random network, but selection can only work with the variations that spontaneously arise, and it is conceivable that at large scales the network is mainly shaped by such events. This does not mean that the network will necessarily look like a random network, but it will at least improve the odds that it does so in some respects.

Unfortunately, the number of attractors in a large Boolean network turns out to be very difficult to find, as it is easy to overlook those cycles and fixed points that have small basins of attraction. This problem of *biased undersampling* is implicated in the original, erroneous finding for Kauffman networks that the mean number of attractors grows with the number of nodes as $\sqrt{N}$, which would have neatly matched the number of cell types. I will get back to the true scaling with N, as this is the topic of paper I.

To further complicates things, there are some very rare networks with an extreme number of attractors, for example those few where *all* states are fixed points. If such pathological networks are not sufficiently rare, then the mean number of attractors will be dominated by exceedingly rare networks whose contributions are unlikely to be included even when many network realizations are examined. In these cases, the median may be a more suitable average than the arithmetic mean for characterizing the typical network.

i.5 Continuous models

Year still after year flows
down the Seven Rivers;
cloud passes, sunlight glows,
reed and willow quivers
— J. R. R. Tolkien, *The Last Ship*

An important part of modeling a system such as the living cell is to choose the right level of detail for the model. Depending on the goals of the modeler and the sensitivity of the system to the assumptions made, different types of models will be appropriate. In some cases Boolean networks may be sufficient to describe signaling and gene regulation, whereas in other cases it is necessary to track the movement of individual molecules in the cell. Although spatial effects inside the cell can play a role, for example in chemotaxis, it is often sufficient to know the concentration levels of the different molecular species. For example, the typical time scales in DNA transcription are far longer than the time needed for diffusion to evenly distribute transcription factors in the cell.

When the number of molecules in a cell is small, stochastic fluctuations must in general be taken into account, if the model is to capture the qualitative behavior of the system. A classical example of this is the lysis–lysogeny switch of phage λ [48]. However, when the number of molecules is large, it is possible to work with continuous concentration levels and ignore the stochastic effects. The model will then consist of a set of chemical rate equations, deterministic differential equations that describe how concentrations change with time.

Even a seemingly simple reaction can consist of many steps, and the corresponding rate equations can in principle be quite complicated. However,

when studying enzyme-driven reaction in metabolism and gene regulation, such a detailed view is usually not necessary, as the equations to a good approximation result in simple Michaelis–Menten kinetics [49]. In general, enzyme reactions where the final step is irreversible can be described by Michaelis–Menten equations [2]. It is not difficult to derive similar straightforward approximations for other types of reactions, given that the relative magnitudes of the reaction constants are know.

Based on the conclusions of Shea and Ackers [50], the transcription rate of a gene, and thus its protein production rate, can be modeled as being proportional to the binding affinity of RNAP to the promoter. This affinity is largely determined by what transcription factors occupy binding sites in the regulatory region, and by how they interact with RNAP. In real systems where the regulatory interactions are known and the transcription rates under different condition have been measured, it may be possible to write down an expression for the Boltzmann weights of the states where RNAP is bound or not bound to the promoter. When such details are not known, or the system exists in a computer simulation only, a practical approach is that of Buchler et al. [51], which is based on the thermodynamics of bacterial transcription regulation and was shown to produce a rich set of transcriptional logic. There, transcription factors may bind to one or more sites in the regulatory region, each with its own binding energy. As some sites are closely spaced or overlapping, the binding is subject to cooperativity and mutual exclusivity, and it is in principle easy to calculate how the transcription factors together recruit or block RNAP.

i.6 The Publications

If every PhD student changed the world,
everyone would get a migraine.
— Andy Hopper

As I said at the very beginning of this introduction, the six papers in this thesis span a range of abstraction and generality. The first three papers revolve around models that do not have immediate biological counterparts. Instead, they describe whole classes of regulatory networks that include the biological ones. In contrast, paper V describes a specific network, and makes predictions about its behavior. The modeling in paper VI is closer to paper V in this respect, but it does make some wider generalizations. Paper IV is the odd man out; instead of modeling the regulatory networks, it introduces a data mining algorithm that relies on previous knowledge about them.

i.6.1 About Boolean networks

Papers I–III are concerned with random Boolean networks in various forms. In paper I, we explore how the numbers of fixed points and cycles of different length grow with the number of nodes in the critical $K = 2$ Kauffman networks. The background is that the quoted $\sqrt{N}$ scaling law for the mean number of attractors [44] had been called into question [52–55]. Bilke and Sjunneson reported that improved sampling showed at least linear scaling [54], and Socolar and Kauffman went on to demonstrate a super-linear relationship [55]. In paper I we show that biased undersampling (see section *i*.4.1) can lead to observations that suggest scaling as a power law such as $\sqrt{N}$. A fixed number of trajectories from random states will not always find all attractors in a set of network realization, and more will be missed the larger the networks are. Over a limited range of network sizes, the number of found attractors can then look like a power law with N.

The main result of paper I is an expression for the mean number of fixed points or cycles of length L in networks of size N. This expression, eq. (I.10), is not easy to evaluate for cycles longer than a few steps, but it is easily shown that for any given power law, it is possible to find an L such that the number of L-cycles grows faster than that power. Hence, the total number of attractors grows faster than any power law with N. The reasoning used to arrive at the expression is based on the idea that rather than try to count the number of attractors in networks with a given architecture and assignment of rules by looking at network states, one may count the number of rule assignments and architectures that let a state be a fixed point.

In more detail, the reasoning goes like this: All network states are equal, in the sense that the rule distribution does not favor some states over others, so we only need to consider one state. If that state is chosen to have all nodes set to FALSE (or all TRUE), then every architecture is equivalent if the state is a fixed point. Left to consider is only the chance that the rules of N identical and independent nodes really make the state a fixed point. For cycles it gets a bit more complicated, but the problem is readily mapped to a similar fixed point problem.

We have proved that the mean number of attractors in $K = 2$ Kauffman networks grows extremely rapidly with the network size, but it is not clear how this relates to biological function. For a large network, long cycles with hopelessly small attractor basins would not seem to have any bearing on the system's normal operations, or they might represent pathological cell types that are rarely realized. Furthermore, Klemm and Bornholdt have demonstrated that in large networks only a small fraction of all attractors are stable under infinitesimal perturbations in time, but a disproportionally large fraction of all trajectories end in stable attractors [43].

In papers II and III we apply some of the knowledge gained over the past

decades to random Boolean networks in a biological context. The architecture of the gene regulatory network of yeast, *Saccharomyces cerevisiae*, was in part determined by Lee et al. in 2002, in a set of ChIP–chip experiments in which the genes regulated by 106 transcription factors were identified with some level of confidence [56]. Disregarding genes that are not transcription factors, we extract a core transcriptional network from the data, and use it as the architecture for otherwise random Boolean networks. To better mimic real regulatory interactions we introduce *nested canalyzing rules*, a class of Boolean functions that includes many of the rules that we would consider "simple" and seemingly easy to implement biochemically. There is a distribution of nested canalyzing rules that agrees well with rules collected and inferred from the literature by Harris et al. [57]. A caveat is that not all the possible input states are experimentally tested for rules with many inputs, as this would require hundreds of experiments. This presumably introduces a bias towards thinking the rules simpler than they really are.

Aside from introducing nested canalyzing rules, paper II describes the dynamics of the yeast network with random such rules, and then especially the response to perturbations. The nested canalyzing rules confer far greater stability than a flat distribution of all Boolean rules, and as a result the network almost always finds its way to a very stable fixed point. When the states of several nodes are flipped at once, the typical behavior of the network is to relax back to the same fixed point within a few time steps.

Self-couplings, links directly from nodes to themselves, present something a problem in Boolean models, as they can behave quite unlike their continuous counterparts. In the continuous world, negative autoregulation pushes something up when it is low and down when it is high, so that it is stabilized at an intermediate level. The Boolean version instead drives the state back and forth between TRUE and FALSE, even without any external signals. For this reason we remove such interactions from the network in paper II, except to investigate their effects on the dynamics.

Paper III goes on to describe the stability properties of networks where the nested canalyzing rules are applied to random network architectures. Instead of returning to the $K = 2$ networks, we study networks where the in-degrees, i.e., the number of inputs per node, are drawn from a power law distribution. This is closer, but not identical, to the distribution seen in real transcriptional networks [56, 58]. How robust a network is towards perturbations can be quantified as the number of nodes that are immediately affected if the state of a single node is changed when the network is at equilibrium. This number, which we call r, is always less than 1 in the studied class of networks, which means that the nested canalyzing rules confer great stability to the network dynamics. The networks are well inside the ordered regime, and only approach criticality when the power in the power law grows very large. In that

limit, all nodes have a single input, and the network consists of long loops where signals circulate forever.

Through some rather convoluted mathematical jiggery-pokery, it is possible to show that the number of attractors in large networks is highly sensitive to r. The results in this direction in paper III were subsequently expanded on in reference [59] by Samuelsson and myself.

i.6.2 About pathway profiling

What we set out to do in paper IV is to combine microarray data with prior knowledge about signaling pathways and transcriptional regulation. By mapping microarray data to the activity of pathways, it should be possible to summarize the main trends of a biological sample in a concise and intuitive manner. The pathways, as described in the TRANSPATH database [36], start at one or more receptors, proceed through multiple reaction steps, and end at a set of transcription factors. With the information in the TRANSFAC [35] database, it is possible to find out what genes the individual transcription factors regulate. If some genes are controlled by the same signaling pathway, then their expression levels should change in unison, provided that the activity level of the signaling pathway shows some variation across the samples. However, the databases do not reveal whether activity in a pathway can be expected to result in the up- or down-regulation of a gene.

We define a score based on squares of the correlations between those genes that are regulated by the same transcription factor. Using data from some publicly available cancer microarray data sets, we verify that this score identifies more transcription factors as regulating their target genes than would be expected by chance. That many transcription factors do not show a significant signal is not a failure, as they can not all be expected to be differentially activated between samples in the data sets. Entire pathways, too, are shown to be significant. Hence, the data sets carry some information about the activity levels of pathways.

We do not know whether a pathway up- or down-regulates each of the genes it is thought to affect, but if most target genes go the same way, a signal should still be seen if the expression data for all of them are combined. This reasoning leads us to a procedure for finding a sign and a p-value for the activity of each pathway in each individual sample. To demonstrate that these scores capture some real biological signal in the data, we make a comparison between the predicted status for a few pathways and the medical annotations of the samples, and find a significant level of agreement. This does not imply that our method presently would be useful for diagnosing patients, but it does indicate that a pathway-level view on gene expression data may find uses in such contexts.

In a recent paper, Liu and Ringnér present a related method of pathway profiling. Starting from lists of differentially expressed genes, and using transcription factor binding motifs to identify target genes, they identified pathways that were abnormally regulated in experiments on mice or in cancer in humans [40].

i.6.3 About the stem cell switch

Stem cell regulation is a vital component of mammalian life. During development, embryonic stem cells must differentiate at the right times and into the right types of cells for tissues to form as they should. Later in life, stem cells play a critical role in damage repair, and there is great hope for practical medical applications in fields like restorative neurology. Unfortunately, stem cells are difficult to maintain *in vitro*, as they spontaneously differentiate in the absence of various chemical signals that are normally present *in vivo* [60]. Such tendencies can be understood in the light of the potentially dire consequences of mutations in stem cells; they already possess the mobility and unlimited replication that are two of the hallmarks of cancer [61]. Similarly, differentiation is essentially a one-way process, but one that it would be greatly useful to be able to reverse, as the least differentiated stem cells are very rare and difficult to extract. Understanding stem cell regulation is a big deal.

Central to maintaining embryonic stem cells in human and mouse are the genes OCT4, SOX2, and NANOG [62]. These genes code for three transcription factors whose exact interactions are still being mapped out. In paper V, we use the available knowledge about these players to make a model of their transcriptional network. For this we use rate equations and the Shea–Ackers approach to expressing the transcription rates. How the three genes regulate each other has been fairly well characterized, but how they together control their target genes less so. What is known is that the transcription factors positively regulate each other and themselves, and that they together keep genes that are involved in stem cell self-renewal expressed, while genes that initiate differentiation are heavily repressed.

What our model reveals is that due to the multiple positive regulations between OCT4, SOX2, and NANOG, the system is strongly bistable. The three genes can be regarded as forming a switch with two distinct states, ON and OFF, and external signals help move the system between these states. It is shown that the transcription factors can regulate all their target genes in qualitatively similar ways, and yet achieve completely different expression of the differentiation genes and the stem cell self-renewal genes, but only if NANOG is positively autoregulating. Thus, the model can help fill gaps in our knowledge of the regulatory interactions.

One set of external signals can cause the expression of NANOG to be repressed, and it is predicted that this irreversibly causes the switch to be shut OFF. When this happens, the stem cell self-renewal genes are repressed, and differentiation is triggered. This matches the response of the cell to DNA damage and similar disasters. Another signal causes SOX2 and OCT4 to be up-regulated, and depending on the basal transcription rate of NANOG the switch may either be turned ON by the signal and then stay ON when the signal is removed, or require the signal to be present to avoid being irreversibly turned OFF. If this signal is lost, a similar same chain of events will unfold as above. The cell may use this mechanism to start differentiation when it is not needed as an embryonic stem cell, or as a fail-safe if the signal is lost for some other reason.

i.6.4 About metabolic pathway regulation

For all that has been said about regulatory networks, it remains to be asked why transcriptional regulation is so common. In the light of evolution, why do gene regulatory networks exist, and why are they so complex? Only if a regulatory interaction confers some fitness benefit will it be retained once it has arisen by mutation. This is true even though the network is mostly built though the duplication and subsequent modification of existing sets of genes. To understand how evolution shapes entire gene regulatory networks, we would do well to understand when and how a single regulatory interaction can be of use.

In paper VI we develop a highly simplified model of a microorganism that does nothing but eat and grow. Every detail that can possibly be removed or simplified is given the axe. What is left is a small set of differential equations that describe how a nutrient diffuses into a cell and is converted into a useful form by the action of a single enzyme. From this useful form, which we may picture as representing pyruvate or some other small, energy-carrying metabolite, the cell can both manufacture more of the enzyme and grow larger. Cell division itself is simplified into an average dilution effect, and all that remains for gene regulation to decide is how to balance the investments in enzyme production and growth.

Consider first how the growth rate of the modeled organism depends on the enzyme production in a perfectly static environment. There is no need for fancy transcriptional regulation, provided that the basal transcription rate of the enzyme can be adjusted. Evolution will effectively tune the expression level of the enzyme to maximize the growth rate. This point was experimentally demonstrated in *E. coli* that were grown in a constant lactose concentration [21]. It is conceivable that transcriptional regulation would still fill a function here, if it can reduce detrimental stochastic fluctuations in the en-

zyme level. However, a major enzyme in a metabolic pathway would likely be expressed at comparatively high levels, and then the relative fluctuations would be small.

A reasonable definition of fitness is the expected growth rate of an organism's genetic material over a period of time, taking into account not only growth and reproduction, but also the risk of death. If cells live in environments where the nutrient abundance is always high enough for sustained growth, and individuals are siphoned off at random so as to keep the total population constant, as in the example above, then the average growth rate alone is a good measure of fitness.

Rather than simulate an environment where the nutrient level changes with time, which certainly can be done in this context [24], we assume that changes are rare, which means that the fitness of an organism can be computed from a set of fixed environments. As a positive practical consequence of this approach, the fitness becomes easy to compute with high precision, and for different variants of the model we can find the parameter values that maximize it. The first such variant we consider is the most central: a transcription factor is added and given the power to up- or down-regulate the expression of the enzyme.

What we find is that when the transcriptional regulation is tuned to maximize fitness, the resulting fitness is far greater than for any unregulated case. In fact, there is a substantial benefit in regulating the enzyme production even when the range of nutrient concentrations is narrowed down. This goes some way towards explaining what is so great about gene regulation.

When we extend the model to include two nutrients and two metabolic pathways, we see that the activation of a pathway ideally is a continuous, if steep, function of the nutrient level. This fits with the *in silico* observation that the *lac* system in *E. coli* shows a graded response to lactose [24]. This lack of hysteresis in the system is also indicative of the stability of those regulatory networks that govern metabolic pathways.

Acknowledgments

Bunnies aren't just cute like everybody supposes! They got them hoppy legs and twitchy little noses. And what's with all the carrots? What do they need such good eyesight for anyway? Bunnies! Bunnies, it must be bunnies!

— Joss Whedon, *Once More, With Feeling*

First and foremost, I am infinitely grateful to Linda, my beloved wife, for her unwavering support, patience, and understanding.

I have greatly enjoyed the relaxed yet scientific atmosphere at the department, and would like to thank all who have contributed to it over the years. I will name a few, and blame a lack of space for not dropping more names. You're not forgotten.

I am especially grateful to my supervisor, Carsten Peterson, who firmly has proven that being silly is a workable option. Even though he has kept me on a loose rein, he has always had time for serious discussions, tasteless jokes, and everything in between. May you live in interesting times!

The influence of Björn Samuelsson on this thesis stretches beyond papers I–III and the mathematical horrors in section III.C, not only because I got the template for the thesis itself from him. By his attention to details, Björn has inspired me to aspire excellence in typesetting, textual clarity, figures and the like. But more importantly, I have greatly enjoyed our discussions (in the widest sense, and then some) on everything from how to evolve regulatory networks to combinatorial games and Islay whisky.

Duller would the department be without Michael Green and Simon Mitternacht. You guys is the shit! Ditto for everyone else I routinely disturb when I've found something noteworthy on teh Intarwub: Pontus Melke, Patrik Sahlin, Mattias Ohlsson, and so forth. For these interruptions I offer my apologies to those who feel that they deserve them.

I also want to thank the biology people I have collaborated with – especially Johan Vallon-Christersson, Lao Saal, and Dag Ahén – for providing a different perspective and teaching me new things. From Morten Krogh I have learned a thing or two about playing the devil's advocate. The code monkeys in the basement deserve thanks for relieving me of the burden of BASE.

And then there are the people in the sauna and on #skåne (simultaneously only if AFK, luckily), who have provided me with much-needed weekly social contact. And of course the regulars at Rydbergs, on either side of the bar.

This thesis was typeset with LaTeX in Palatino, and printed by the helpful people at Media-Tryck. I thank Pontus Melke and Carsten Peterson for proofreading and commenting on this introduction. The images of biomolecules come from Wikipedia and Wikimedia Commons, and I am grateful to the peo-

ple who allowed me to use them. Figures *i*.1(right) and *i*.2 are licensed under the GNU FDL, `http://gnu.org/licenses/fdl.txt`, and are reproduced here with the permission of their creators; the others are in the public domain.

i References

[1] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walte, *Molecular Biology of the Cell*, Garland Science, New York, 1996.

[2] P. Nelson, *Biological Physics*, W. H. Freeman, New York, 2004.

[3] *Wikipedia, the free encyclopedia,* `http://en.wikipedia.org/`.

[4] A. Brazma, H. Parkinson, T. Schlitt, and M. Shojatalab, *A quick introduction to elements of biology - cells, molecules, genes, functional genomics, microarrays*, `http://www.ebi.ac.uk/microarray/biology_intro.html`.

[5] P. L. Luisi, *About various definitions of life*, Orig. Life Evol. Biosph. **28**, 613–622 (1998).

[6] J. Koshland, D. E., *The seven pillars of life*, Science **295**, 2215–2216 (2002).

[7] R. Dawkins, *The Blind Watchmaker: Why the evidence of evolution reveals a universe without design*, W. W. Norton, New York, 1986.

[8] F. Delmotte, N. Leterme, J. Bonhomme, C. Rispe, and J. C. Simon, *Multiple routes to asexuality in an aphid species*, Proc. Biol. Sci. **268**, 2291–2299 (2001).

[9] T. Dobzhansky, *Nothing in biology makes sense except in the light of evolution*, Am. Biol. Teach. **35**, 125–129 (1973).

[10] R. Dawkins, *Climbing Mount Improbable*, W. W. Norton, New York, 1996.

[11] N. R. Pace, *A molecular view of microbial diversity and the biosphere*, Science **276**, 734–740 (1997).

[12] F. H. Crick, *The origin of the genetic code*, J. Mol. Biol. **38**, 367–79 (1968).

[13] R. D. Knight, S. J. Freeland, and L. F. Landweber, *Rewiring the keyboard: evolvability of the genetic code*, Nat. Rev. Genet **2**, 49–58 (2001).

[14] M. Bulger and M. Groudine, *Looping versus linking: toward a model for long-distance gene activation*, Genes Dev. **13**, 2465–2477 (1999).

[15] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, *Network motifs in the transcriptional regulation network of Escherichia coli*, Nat. Genet. **31**, 64–68 (2002).

[16] A. L. Barabási and Z. N. Oltvai, *Network biology: understanding the cell's functional organization*, Nat. Rev. Genet. **5**, 101–113 (2004).

[17] C. Song, S. Havlin, and H. A. Makse, *Self-similarity of complex networks*, Nature **433**, 392–395 (2005).

[18] M. M. Babu, N. M. Luscombe, L. Aravind, M. Gerstein, and S. A. Teichmann, *Structure and evolution of transcriptional regulatory networks*, Curr. Opin. Struct. Biol. **14**, 283–291 (2004).

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

[20] E. M. Ozbudak, M. Thattai, H. N. Lim, B. I. Shraiman, and A. van Oudenaarden, *Multistability in the lactose utilization network of Escherichia coli*, Nature **427**, 737–740 (2004).

[21] E. Dekel and U. Alon, *Optimality and evolutionary tuning of the expression level of a protein*, Nature **436**, 588–592 (2005).

[22] T. F. Cooper, D. E. Rozen, and R. E. Lenski, *Parallel changes in gene expression after 20,000 generations of evolution in Escherichia coli*, Proc. Natl. Acad. Sci. USA **100**, 1072–1077 (2003).

[23] O. S. Soyer, T. Pfeiffer, and S. Bonhoeffer, *Simulating the evolution of signal transduction pathways*, J. Theor. Biol. **241**, 223–232 (2006).

[24] M. J. A. van Hoek and P. Hogeweg, *In silico evolved lac operons exhibit bistability for artificial inducers, but not for lactose*, Biophys. J. **91**, 2833–2843 (2006).

[25] A. K. Dewdney, *Computer recreations: In the game called Core War hostile programs engage in a battle of bits*, Sci. Am. **250**, 14–22 (1984).

[26] R. E. Lenski, C. Ofria, R. T. Pennock, and C. Adami, *The evolutionary origin of complex features*, Nature **423**, 139–144 (2003).

[27] P. Dittrich, J. Ziegler, and W. Banzhaf, *Artificial chemistries – a review*, Artif. Life **7**, 225–275 (2001).

[28] S. A. Kauffman, *Autocatalytic sets of proteins*, J. Theor. Biol. **119**, 1–24 (1986).

[29] W. Banzhaf and P. D. Kuo, *Network motifs in natural and artificial transcriptional regulatory networks*, J. Biol. Phys. Chem. **4**, 85–92 (2004).

[30] T. Pfeiffer, O. S. Soyer, and S. Bonhoeffer, *The evolution of connectivity in metabolic networks*, PLoS Biol. **3**, e228 (2005).

[31] C. O. Wilke, J. L. Wang, C. Ofria, R. E. Lenski, and C. Adami, *Evolution of digital organisms at high mutation rates leads to survival of the flattest*, Nature **412**, 331–333 (2001).

[32] P. Jaluria, K. Konstantopoulos, M. Betenbaugh, and J. Shiloach, *A perspective on microarrays: current applications, pitfalls, and potential uses*, Microb. Cell Fact. **6**, 4 (2007).

[33] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, et al., *Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data.*, Nat. Genet. **34**, 166–176 (2003).

[34] M. J. Buck and J. D. Lieb, *ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments*, Genomics **83**, 349–360 (2004).

[35] E. Wingender, X. Chen, E. Fricke, R. Geffers, R. Hehl, I. Liebich, et al., *The TRANSFAC system on gene expression regulation*, Nucleic Acids Res. **29**, 281–283 (2001).

[36] M. Krull, N. Voss, C. Choi, S. Pistor, A. Potapov, and E. Wingender, *TRANSPATH: an integrated database on signal transduction and a tool for array analysis*, Nucleic Acids Res. **31**, 97–100 (2003).

[37] S. Aerts, P. V. Loo, G. Thijs, H. Mayer, R. de Martin, Y. Moreau, et al., *TOUCAN 2: the all-inclusive open source workbench for regulatory sequence analysis*, Nucleic Acids Res. **33**, W393–W396 (2005).

[38] Z. Bar-Joseph, G. K. Gerber, T. I. Lee, N. J. Rinaldi, J. Y. Yoo, F. Robert, et al., *Computational discovery of gene modules and regulatory networks*, Nat. Biotechnol. **21**, 1337–1342 (2003).

[39] A. Kel, N. Voss, R. Jauregui, O. Kel-Margoulis, and E. Wingender, *Beyond microarrays: finding key transcription factors controlling signal transduction pathways*, BMC Bioinformatics **7**, S13 (2006).

[40] Y. Liu and M. Ringnér, *Revealing signaling pathway deregulation by using gene expression signatures and regulatory motif analysis*, LU TP 06-36, in *Statistical and Functional Analysis of Genomic and Proteomic Data*, PhD thesis (2006).

[41] E. Fuchs and J. A. Segre, *Stem cells: A new lease on life*, Cell **100**, 143–155 (2000).

[42] I. Shmulevich, E. R. Dougherty, and W. Zhang, *From Boolean to probabilistic Boolean networks as models of genetic regulatory networks*, Proc. IEEE **90**, 1778–1792 (2002).

[43] K. Klemm and S. Bornholdt, *Stable and unstable attractors in Boolean networks*, Phys. Rev. E **72**, 055101 (2005).

[44] S. A. Kauffman, *Metabolic stability and epigenesis in randomly constructed genetic nets*, J. Theor. Biol. **22**, 437–467 (1969).

[45] F. Jacob and J. Monod, *Genetic regulatory mechanisms in the synthesis of proteins*, J. Mol. Biol **3**, 1 (1961).

[46] M. Aldana-Gonzalez, S. Coppersmith, and L. P. Kadanoff, *Boolean dynamics with random couplings*, in *Perspectives and Problems in Nonlinear Science*, E. Kaplan, J. E. Marsden, and K. R. Sreenivasan (eds.), pp. 23–89, Springer, 2003.

[47] P. Flicek, E. Keibler, P. Hu, I. Korf, and M. R. Brent, *Leveraging the mouse genome for gene prediction in human: from whole-genome shotgun reads to a global synteny map*, Genome Res. **13**, 46–54 (2003).

[48] A. Arkin, J. Ross, and H. H. McAdams, *Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected Escherichia coli cells*, Genetics **149**, 1633–1648 (1998).

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

[50] M. A. Shea and G. K. Ackers, *The OR control system of bacteriophage lambda. a physical-chemical model for gene regulation*, J. Mol. Biol. **181**, 211–230 (1985).

[51] N. E. Buchler, U. Gerland, and T. Hwa, *On schemes of combinatorial transcription logic*, Proc. Natl. Acad. Sci. USA **100**, 5136–5141 (2003).

[52] U. Bastolla and G. Parisi, *A numerical study of the critical line of Kauffman networks*, J. Theor. Biol. **187**, 117–133 (1997).

[53] U. Bastolla and G. Parisi, *The modular structure of Kauffman networks*, Physica D **115**, 219–233 (1998).

[54] S. Bilke and F. Sjunnesson, *Stability of the Kauffman model*, Phys. Rev. E **65**, 016129 (2002).

[55] J. E. S. Socolar and S. A. Kauffman, *Scaling in ordered and critical random Boolean networks*, Phys. Rev. Lett. **90**, 068702 (2003).

[56] T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, et al., *Transcriptional regulatory networks in Saccharomyces cerevisiae*, Science **298**, 799–804 (2002).

[57] S. E. Harris, B. K. Sawhill, A. Wuensche, and S. A. Kauffman, *A model of transcriptional regulatory networks based on biases in the observed regulation rules*, Complexity **7**, 23–40 (2002).

[58] A. M. Huerta, H. Salgado, D. Thieffry, and J. Collado-Vides, *RegulonDB: a database on transcriptional regulation in Escherichia coli*, Nucleic Acids Res. **26**, 55–59 (1998).

[59] B. Samuelsson and C. Troein, *Random maps and attractors in random Boolean networks*, Phys. Rev. E **72**, 046112 (2005).

[60] N. Sato, L. Meijer, L. Skaltsounis, P. Greengard, and A. H. Brivanlou, *Maintenance of pluripotency in human and mouse embryonic stem cells through activation of Wnt signaling by a pharmacological GSK-3-specific inhibitor*, Nat. Med. **10**, 55–63 (2004).

[61] D. Hanahan and R. Weinberg, *The hallmarks of cancer*, Cell **100**, 57–70 (2000).

[62] L. A. Boyer, T. I. Lee, M. F. Cole, S. E. Johnstone, S. S. Levine, J. P. Zucker, et al., *Core transcriptional regulatory circuitry in human embryonic stem cells*, Cell **122**, 947–956 (2005).