# On gene regulatory networks and data fitting

Fogelmark, Karl

2016

*Document Version:*
Other version

# ON GENE REGULATORY NETWORKS AND DATA FITTING

*Karl Fogelmark*

LUNDS UNIVERSITET

2016

Thesis for the degree of Doctor of Philosophy

Department of Astronomy and Theoretical Physics
Faculty of Science
Lund University

Thesis advisor: *Carl Troein*
Faculty opponent: *Ala Trusina*

To be presented, with the permission of the Faculty of Science of Lund University, for public criticism in Lundmarksalen at the Lund Observatory, on the 19th of May 2016 at 10:15.

Organization
LUND UNIVERSITY
Department of Astronomy and Theoretical Physics
Sölvegatan 14A
SE–223 62 LUND
Sweden

Document name
DOCTORAL DISSERTATION

Date of issue
May 2016

Sponsoring organization

Author(s)
Karl Fogelmark

Title and subtitle
On gene regulatory networks and data fitting

Abstract
Living organisms can be viewed as complex biological machines. In order to function, they must regulate their internal mechanism to do the right thing, at the right time, and in the right amount. Part of this regulation is encoded in gene regulatory networks. These are built up of genes which produce special proteins (transcription factors, TF) that regulate other TF-producing genes. Thus a network is formed with genes (nodes) linked together by their mutual regulation (edges).

By constructing simplified models, we investigate such gene networks. The models allow us to probe general principles behind what shapes these networks (paper II), as well as specific networks such as that which endows the plant *Arabidopsis thaliana* with the ability to predict dawn and dusk (paper III). We also present a model for dynamically generating transcriptional networks which encode function from a single variable-length binary representation of DNA (string of ones and zeroes). This gives a natural way for the network to evolve by mutations. However, performing a meaningful and efficient crossover operation on two DNA strings of different length becomes a challenge. We address this by introducing a heuristic algorithm, which we compare against existing methods (paper IV).

Additionally, we present a correct error estimation for the popular least squares method that is valid also for nonlinear functions applied to highly correlated data (paper I). For model fitting to correlated data, one has previously been constrained to use either a maximum likelihood approach, which leads to strong bias in the estimated parameters, or a least squares approach, which gives an incorrect error estimate. We also derive the first order contribution of the bias for both the maximum likelihood and the least squares method, and introduce a minimum variance function fitting method suited for Brownian motion.

Key words:
Circadian rhythms, gene regulation, transcription networks, correlated data

Classification system and/or index terms (if any):

Supplementary bibliographical information:

Language
English

ISSN and key title:

ISBN
978-91-7623-699-4

Recipient's notes

Number of pages
238

Price

Security classification

# ON GENE REGULATORY NETWORKS AND DATA FITTING

*Karl Fogelmark*


LUNDS
UNIVERSITET

Printed in Sweden by Media-Tryck, Lund 2016

ISBN 978-91-7623-699-4 (print)
ISBN 978-91-7623-700-7 (pdf)

Cover illustration:
*Svartedauen — Pesta i trappen* (1900),
by Theodor Kittelsen,
courtesy of Nasjonalmuseet.

> The reasonable man adapts himself to the world; the unreasonable one persists in trying to adapt the world to himself. Therefore, all progress depends on the unreasonable man.
>
> George Bernard Shaw


## SAMMANFATTNING

Världen är föränderlig. För att kunna överleva måste allt liv kunna anpassa sig till rådande förhållanden. För cellen, livets minsta enhet, sker detta bland annat genom reglering av produktionstakten av proteiner, vilka är de molekyler som utför de flesta grundläggande funktioner.

En speciell klass av proteiner utgörs av så kallade transkriptionsfaktorer. Dessa slår av eller på en gens produktion av proteiner, genom att binda till gens position på DNA-molekylen. Eftersom dessa transkriptionsfaktorer också själva är proteiner, som produceras av gener som regleras av andra transkriptionsfaktorer, bildas komplexa nätverk där gener som producerar denna proteinklass kan sägas interagera med varandra. Dessa transkriptionsnätverk av genreglering ligger till grund för hur, till exempel, en växt kan stänga av klorofyllproduktion i avsaknad av ljus.

I praktiken har genregleringsnätverken gått än längre och kan — givet dagsljusets periodicitet — förutsäga solens upp- och nedgång. I två artiklar undersöker vi dessa gennätverk med hjälp av matematiska modeller. I artikel III undersöker vi ett nätverk, specifikt för växten backtrav, som fungerar som en klocka, med vilken gryning och skymning kan förutsägas genom oscillationer i specifika proteinkoncentrationer. I artikel II undersöks mer generella nätverk utan direkt anknytning till någon specifik organism. I dessa nätverk lagras den genetiska informationen i en sträng av ettor och nollor, vilken representerar DNA-kedjan. Denna binära sträng tillåts i artikel IV att vara av variabel längd, vilket försvårar den matchning som är av biologisk relevans vid reproduktion. Vi undersöker därför olika metoder för att effektivt jämföra två olika långa binära strängar.

Orelaterat till genreglering ovan, presenteras i artikel I en korrigerad feluppskattningsformel för parameteranpassning till korrelerad data. När datapunkter sägs vara *korrelerade* avses att dessa inte är oberoende av varandra. Det vill säga, att addera fler punkter, t.ex. genom att göra fler mätningar, innebär inte nödvändigtvis att vi får mer information om systemet. Den vanligaste metoden för att anpassa en funktion till data, minsta kvadratmetoden, kommer däremot att ge sken av att så är

fallet, och således ge en allt för optimistisk uppskattning av felet. Detta avhjälper vi genom att introducera en korrigerad feluppskattningsformel för minsta kvadratmetoden, vars giltighet vi demonstrerar på tre system där data är benägen att vara korrelerad.

PUBLICATIONS

The thesis is based on the following publications:

I Karl Fogelmark, Michael Lommholt, Anders Irbäck and Tobias Ambjörnsson.
**Model parameter estimation in particle tracking**.
Submitted, LU-TP 16-18 (2016).

II Karl Fogelmark, Carsten Peterson and Carl Troein.
**Selection Shapes Transcriptional Logic and Regulatory Specialization in Genetic Networks**.
*PLoS ONE* **11**, e0150340 (2016).

III Karl Fogelmark and Carl Troein.
**Rethinking transcriptional activation in the *Arabidopsis* circadian clock**.
*PLoS Computational Biology* **10**, e1003705 (2014).

IV Karl Fogelmark, Adriaan Merlevede, Carl Troein and Henrik Åhl.
**An efficient crossover algorithm by global alignment for evolution of variable length genomes**.
Manuscript, LU-TP 16-11 (2016).

During my time as PhD-student, I have also co-authored the following publications that are not included in the thesis.

- Lloyd P Sanders, Michael A Lomholt, Ludvig Lizana, Karl Fogelmark, Ralf Metzler and Tobias Ambjörnsson.
**Severe slowing-down and universality of the dynamics in disordered interacting many-body systems: ageing and ultraslow diffusion**.
*New Journal of Physics* **16**, 113050 (2014).

- Ralf Metzler, Lloyd Sanders, Michael A Lomholt, Ludvig Lizana, Karl Fogelmark and Tobias Ambjörnsson.
**Ageing single file motion**.
*European Physical Journal* **223**, 3287–3293 (2014).

> Orm sade på gamla dagar om denna tid, att den var lång att leva men kort att berätta om; ty den ena dagen var den andra lik, så att det på ett sätt var som om tiden stått stilla.
> Frans G. Bengtsson, *Röde orm, sjöfarare i västerled*

ACKNOWLEDGMENTS

Systems tend to equilibrate with their surroundings. If this also is true for humans, then I could not wish for a more rewarding working environment to interact with than that of the department of theoretical physics, where people are always eager to help just for the sheer joy of solving an interesting problem, and where anything can be discussed. In the following, I shall make an attempt at mentioning a subset of the numerous persons who have influenced this work.

First and foremost, I would like to sincerely thank my supervisor Carl Troein, whom I could ask anything at any time, and without whose guidance this thesis would not have come to be. Not only has his inexhaustible energy, often running into the office to try something out, proved to be a great inspirational source, but his many crazy antics has made me look like an almost normal person by comparison.

I was first introduced to the wonders and woes of research during my masters project by my previous supervisor Tobias Ambjörnsson, to whom I would like to express my heartfelt gratitude. Paper I stands as a testament of his clear supervision and seemingly infinite patients for my many intrusions into his office. Thanks for always laughing at my bad jokes, but never at my stupid questions.

During my masters project, I was also introduced to Carsten Peterson, who encouraged me to pursue a career in science, and to focus more on its "wonders" than its "woes". Since then, he has provided useful insights, and entertained me greatly with many anecdotes, for which I am thankful.

Yet, when dark clouds do gather, I have had the good fortune to be able to rely on my fellow PhD-students for support. Countless are the lunches where burdens and laughter were shared alike, over discussions of varying philosophical, existential, and cultural depth. I am grateful to Christian Holtzgräfe, Iskra Staneva and André Larsson for helping me maintain my (in)sanity over the many Govindas lunches; as well as the rest of the old "PhD-gang": Lloyd Sanders, Michaela Reiter-Schad, and Sigurður Ægir Jónsson, with whom much spare time has been spent.

My former office mates Behruz Bozorg, Victor Olariu, and Jeremy Gruel, deserve recognition for putting up with me, but judging from the things uttered in that room, I think I was in good company.

In addition, I would like to (again) express my sincere appreciation to Iskra and André, for meticulous proofreading of my Introduction and providing useful suggestions and corrections; the remaining mistakes are my own.

A thank you goes to Anders Irbäck for many interesting conversations, Mattias Ohlsson for helping with computers (and a toaster!), and to the "brain trust": Bo Söderberg and Patrik Edén, for letting me bathe in their reflected brilliance. Their many brief, but always sharp, suggestions have lead to direct improvements of this thesis. Also, thanks to Adriaan Merlevede, who brought a fresh perspective to our project, and to Najmeh Abiri for many discussions on what truly matters: 80s movies.

When nothing works and the eyes go weary from reading too much C++ code, I have found refuge in the free software project of *Pioneer*, where I can read *other* C++ code. From one of my GNU Emacs IRC buffers, I have gotten to know my fellow development team members, whom I would like to acknowledge, especially the project's art lead Bálint Szilárd for helping me realize my vision for Figure 1.1.

Needless to say, GNU Emacs has been instrumental in all work and non-work related activities, as it is that which gives the universe beauty and meaning, for which not only I, but all of mankind, is forever indebted to Richard Stallman.

But what makes life bearable is friends of old, who stood me by, never faltering, with whom merry times have been shared.

Last, but certainly not least, I am grateful to my mother and father for helping me when I needed it the most, but realized it the least.

No thanks at all to posers, fashionable sheepeople in need of herding, or trendy designers riding high on their "graphical profile", now forbidding the classic blank thesis cover.

Up the hammers & down the nails!

# *Contents*

*Tillägnat det som en gång var...*

It is possible to believe that all the past is but the beginning of a beginning, and that all that is and has been is but the twilight of the dawn. It is possible to believe that all that the human mind has ever accomplished is but the dream before the awakening. We cannot see, there is no need for us to see, what this world will be like when the day has fully come. We are creatures of the twilight. But it is out of our race and lineage that minds will spring, that will reach back to us in our littleness to know us better than we know ourselves, and that will reach forward fearlessly to comprehend this future that defeats our eyes. All this world is heavy with the promise of greater things, and a day will come, one day in the unending succession of days, when beings, beings who are now latent in our thoughts and hidden in our loins, shall stand upon this earth as one stands upon a footstool, and shall laugh and reach out their hands amid the stars.

H.G. Wells, *The discovery of the future* (1902)

There is no such things as magic, though there is such
a thing as knowledge of the hidden ways of Nature.
H. Rider Haggard, *She* (1887)

# *Introduction*

Nature can be understood. This is a realization that we in large part owe to Aristotle (384–322 BC), a student of Plato. He fathered the field of biology and made significant contributions to all fields of science of the era, including physics. The two fields of biology and physics, where the former is devoted to the study of the living, and the latter to the inanimate laws of our universe, have generally been kept separated.

In this thesis we investigate biological systems by applying the methods which have proven so lucrative in the field of physics [1]. This entails constructing mathematical models which reproduce the observed behaviour of the system under investigation. To this effort we strive to "make things as simple as possible, but not simpler" [2], which might leave a reader with a background in biology wanting for a less idealized description of the biological systems addressed in this thesis. However, if we are to understand the inner workings of a (metaphorical) fine mechanical clock, we have to start with pendulums.

This introduction aims to give the reader a firm footing of the key concepts touched upon in this thesis, from which he can leap into any of the articles which are to follow. Our first step illustrates how the marriage of a biologist's discovery and a physicist's endeavours born the revelation of the smallness of matter, that is necessary for life.

### 1.1.1 *Physics and flowers*

In 1827 the Scottish botanist Robert Brown observed, through his microscope, the irregular motion of particles enclosed by micrometer sized pollen grains suspended in water [3].[1] He initially attributed this to

1 It is worth pointing out that he was not the first to describe the phenomenon that now bears his name. Dutch physician Jan Ignenhousz observed it with coal

"the vitality of pollen" [5]; however, the motion persisted undiminished in the absence of nutrients. Brown found that even ground down inanimate particles from the Sphinx behaved in this peculiar fashion [6], thus ruling out the discovery of living "animalcules" [7].

It was shown by theoretical physicist Albert Einstein, in one of his *annus mirabilis* papers of 1905 [8], that this was the result of the thermal motion of the hypothesized molecules, acting in conjunction to displace the pollen grain at random. He derived the mean square displacement of a particle undergoing what he coined "Brownian motion", and provided a relation which connected the macroscopic observable (diffusion constant) with the microscopic world, allowing a numerical value to be determined for both Boltzmann's constant, and Avogadro's number. This not only proved the existence of molecules, but also gave an experimental way to determine their size, for which the french experimentalist Jean Baptiste Perrin was awarded the Nobel prize in 1926 [3, 6].

Indeed, it is the very smallness of the molecules, allowing them to act in enormous numbers, that permits life. The deterministic physical and chemical laws that are relevant to life rely on the statistical laws that are valid only for large ensembles. So does the irregular heat movement of particles give rise to the regular phenomenon of diffusion [9]. However, in stark contrast to the microscopic disorder, we find the DNA molecule. It contains the recipe for life, held in the hereditary unit of *genes*. These give rise to organized events, in spite of the disordered thermal motion around it.

### 1.1.2 *What is life?*

Brown's experiment with the ground down Sphinx particles raises an important and difficult question (beyond that of the ethics of archaeological desecration): what is alive, and what is dead? At one end of the spectrum we find the inanimate stone statue of aeons past, at the other we may place our animate selves; we must clearly be alive to pose this ultimate question to begin with.

If life is the outcome of a continuous process of evolution, then the boundary between the living and the non-living is a difficult one to distinguish [10]. A growing crystal or a replicating virus is by most definitions not considered to be alive, yet they exhibit traits which we associate with the living [11]. Anyone who has been chased by an

---

particles on alcohol in 1785 [3], and before him the Roman Lucretius (c. 99 – 55 BC) described it in a poem [4], see appendix 3.A, p. 45.

angry bee would consider it to be most alive, even if it is incapable of reproducing or replicating. However, we can attempt to identify a "least common denominator" of living systems.

Life is an ordered process which adheres to a set of common requirements. For order to persist, there needs to be an organized plan, a *program*, that implements instructions for the parts needed for maintaining life and how they interact. For the system to be self-sustaining it needs *energy* to drive its chemical and physical movement that act to reverse entropy and keep the system from its equilibrium state of death. Finally, the system needs to be *self-regenerating*, and replenish, to counteract the thermodynamic losses of the processes that instil order [11]. However, the regeneration does not restore the system to the exact original state. As we look upon the previous generation, whether it be our own species or bacteria, we see the cost of time: We age.

Death is a necessity for life, and evolution is its direct consequence. With time the cumulative changes cause ageing which inches the individual ever closer towards its end. The cure is for life to reset itself by starting over through reproduction. This introduces the need for the life-instructing program to be passed to the next generation. The information transfer will be perceptible to imperfections (mutations) which combined with selection will optimize the species to better serve the genes as "survival-machines" [12]. We are but vessels for the immortal genes. To this end life comes in many forms, both as single celled organisms and as multicellular.

All living organisms can be categorized into two main branches based on cell structure. At the simplest we find the small *prokaryotes* (typically 1-10 $\mu$m in size), such as bacteria, which all lack a membrane enveloped cell nucleus. The other class is the *eukaryotes*, which make up all multicellular life, but does not exclude single cell organisms. Scientist have adopted a particularly keen liking to a set of *model organisms* with desirable traits that are well suited for their probing minds, such as the organism having short generations, small genetic material, being in abundant supply, as well as being subjected to the whimsical disdain of human society, giving scientists free rein. In the following we will touch upon the prokaryote *Escherichia coli* (bacteria), as well as the eukaryotes *Arabidopsis thaliana* (plant, thale cress), *Mus musculus* (mammal, mouse), *Neurospora crassa* (fungus), and *Drosophila melanogaster* (insect, fruit fly). The first mentioned from each respective domain shall also play a part in the papers that are to follow.

## 1.2 THE GENE AS THE FUNDAMENTAL INFORMATION UNIT OF LIFE

The information that is necessary to maintain and replicate life needs a representation for encoding and a reliable system for storage and copying. At its core, information is stored by simply stringing together different entities that are not all the same, just like the letters of the alphabet making up words, or the base two system used by digital computers, usually represented as ones and zeroes. The cell uses a similar system where four nucleotides, A (adenine), T (thymine), C (cytosine), and G (guanine), make a base four system. By attaching the bases to the sugarphosphate backbone of deoxyribonucleic acid a long polymer is formed: the DNA molecule. The nucleotide bases pair up by forming hydrogen bonds between A-T (adenine-thymine) and C-G (guanine-cytosine), thereby creating a complementary cDNA strand which stabilizes the structure and, in addition, acts as a backup copy [13]. The two strands combine to form a long double helix, which coils and loops itself multiple time into a *chromosome* if in a eukaryote, or a single closed loop if in bacterial prokaryote [13, 14]. In eukaryotes the entire DNA code is contained within the cell nucleus. For humans the DNA packing allows two meters of DNA, ($3.2 \cdot 10^9$ nucleotides), with 1 nm diameter to fit into the micro meter sized cell nucleus [13]. The chromosomes are collectively referred to as the *genome*, as it contains all the genes, which are the discrete units of hereditary information, as well as the non-coding regions.

The genome sequence is used as a blueprint to generate the long chains of *amino acids* that constitute the protein molecules. The genetic sequence is read in triplets. A triplet in a coding region is referred to as a *codon*, and is interpreted as a "word" that instructs the cell which amino acid should come next. The amino acids come in twenty different flavours, and are linked together to a long chain, in the order specified by the codons, into a protein. With four nucleotides, read in triplets, there are $4^3 = 64$ possible codons which map to the 20 different possible amino acids, thus there is a degeneracy: generally several codons map to the same amino acid. Codons that are similar typically map to the same amino acid. This redundancy acts as a safeguard against mutations. However, not all codons are reserved for coding amino acids, as the boundaries of the coding region are marked by special start and stop codons.

A gene is a well defined region on the DNA, where the genetic information between the start codon and stop codon encodes a protein (*gene*

*product*). The start codon is unique, and defines the reference frame of the genetic code. The triplet following the start codon corresponds to the first amino acid of the protein to be. If there is a shift of one base pair, the meaning of all codons following it will subsequently change, thus we have entered a new *reading frame*. This means that there are three distinct reading frames on the DNA strand, and an additional three in the opposite direction on the complementary chain. In theory, one section of a single DNA strand could therefore encode three different proteins, and its complement yet another three, making in total six overlapping genes. In reality, the information content of the genome is sparse, genes are separated by large non-coding intergenic regions, and only rarely do overlapping reading frames occur.

The information in the DNA chain can be read through two different processes, each serving a different purpose. When a cell divides, the entire DNA is read and copied, resulting in a new identical DNA molecule. This is equivalent to copying a program on the hard drive of a modern computer. However, if we want to execute the genetic program, the "wetware", in order to synthesize a protein, only the region of the DNA chain containing the gene in question needs to be accessed, and loaded into "memory". This process of *gene expression* entails many steps and differs between prokaryotes and eukaryotes [13], but can be described in the following (see Figure 1.1):

1. A large protein, RNA polymerase (RNAP), attaches at a specific DNA-sequence. The double helix is locally uncoiled and opened by the RNAP molecule. As RNAP slides downstream, it *transcribes* the DNA code (80 bp/sec [14]) to a single stranded short lived ($\sim$ 10 minutes) complementary "working copy" of the DNA sequence, through a 1:1 base pair alignment — except where base T (thymine) is replaced by U (uracil), and ribose is used as backbone instead of deoxyribose as in the DNA molecule — resulting in the aptly named messenger RNA molecule (mRNA) [14]. The genetic program is now loaded into the "memory". Transcription stops when RNAP reaches the *transcriptional terminator* which triggers a release of the mRNA and RNAP from the DNA-strand [13].

2. The mRNA transcript is transported from the nucleus (if in eukaryote) to the *ribosome*, a large protein complex in the cytoplasm of the cell. Here each codon, between the start codon (AUG) and the degenerate stop codon (UAA, UGA, or UAG), is *translated* to an amino acid which are all chained together to form a protein. In

*E. coli* the speed of this process is about 40 amino acids per second, allowing a full protein to be translated in minutes [14]. The one dimensional four-letter information stored in the transcript has now been mapped to a base twenty amino acid sequence that defines the protein.

3. The protein then folds by exposing its hydrophilic part and enveloping its hydrophobic, giving it a complex three dimensional structure, which defines its function. The nanometer sized protein is now free to perform its function.

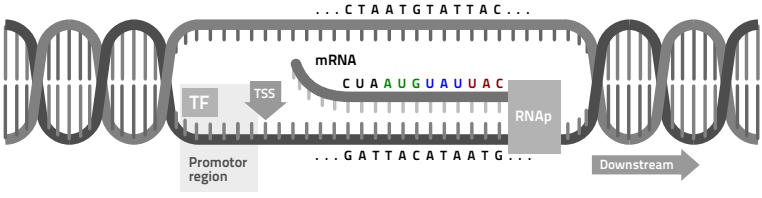


**Figure 1.1 Transcription process.** Transcription is initiated by transcription factors (TFs) binding to the promotor region, which recruits RNAP binding. As RNAP starts sliding downstream, from the transcriptional start site (TSS), along the uncoiled and opened double helix, it will assemble an mRNA molecule with complementary base pairs, except T is replaced by U. The process stops when RNAP reaches the transcriptional terminator (not shown) and releases mRNA and itself from the strand. The mRNA will be transported to the ribosome where each base triplet (codon), will be translated into a specific amino acid, that will be assembled into a protein. In the example sequence shown, the two codons following the start codon (AUG) both code for the same amino acid *Tyrosine*. The complementary DNA can also be transcribed in the same way, but in the opposite direction. For example, in order for the cDNA sequence to be expressed, a promotor region would be needed upstream of it, and a start codon that would define a second reading frame. The description is simplified compared to present understanding, where the process differs between eukaryotes and prokaryotes, but the main characteristics are conserved.

A large part of the genome does not contain any genetic information and is never expressed. This also applies to the transcribed gene sequence, as only a subset of the mRNA sequence, the *exons*, are expressed. The *introns*, the region between the exons, is removed, through splicing, from the transcript prior to translation [13]. Thus the sequence of the introns have no bearing on the final synthesized gene product.

The genome length and fraction of unexpressed code differs between species. The genome of prokaryotes, such as *E. coli* (1 Mbp, i.e. $10^6$ base pairs), typically holds a few thousand genes, while eukaryotes, like *Arabidopsis* (142 Mbp) or human (3200 Mbp) both hold some

30,000 genes [13]. The difference in length is mainly due to the larger amount of introns and intergenic regions, e.g. only 11% of the genome is unexpressed in *E. coli* while the same holds true for 98.5% of the human genome [13]. This unexpressed code is often referred to as "junk DNA", but this is a misnomer as it serves as a playground for evolution of the species by allowing the emergence of new functional genes. For eukaryotes there does not seem to be any great disadvantage to have a long genome. The length does not necessarily mean the organism is more "advanced". Some species of amoeba have a genome 200 times longer than that of humans [13].

### 1.2.1 *Mutation and fidelity of base pairs*

Stagnation means death. The ability to adapt to the changes in the environment is a requirement for survival. Through accumulating mutations of the DNA a species can evolve to better suit its environment, thereby improving its survival *fitness*. The genes are not selected for directly, but rather through their effect on the *phenotype* — the resulting traits and properties of the underlying *genotype* of the organism [15].

The replication of DNA shows a remarkable high fidelity. For life to be possible, the genetic information must be preserved over generational time, and at the same time be able to adapt to changing conditions, by incremental trial-and-error through small changes to the code [16]. The mutation rate of *E. coli* is $10^{-9}$ per bp and replication, and similar in eukaryotes [16]. Since most mutations are harmful and lower the fitness of the organism, the mutation rate is also under evolution. It is lowered by proof-reading mechanisms [17].

Through a *point mutation* a single base in the genome is changed. A point mutation is often *neutral*, not having any effect on the phenotype, due to the extent of non-coding regions, as well as the degeneracy of the codons — similar codons map to the same amino acid. A point mutation through *substitution*, (e.g. A to G, C or T), can result in a *missense mutation*, meaning that the codon will map to another amino acid. This is most likely to happen if the first or second base in the codon is mutated, as the last base pair holds the least information [18]. A mutation can also lead to the creation of a stop codon in the middle of the gene causing an abrupt stop of transcription.

A point mutation in the form of deletion or insertion of a base can be a highly intrusive point mutation as in an exon it leads to a *frame*

*shift*, which will change the reading frame of all codons following it, as they are defined from their first position.

## 1.3 REGULATION THROUGH TRANSCRIPTION NETWORKS

The cell is continuously affected by its external and internal environment and in order to function it must correctly regulate its gene expression (protein production) in response to different input signals so that the right genes are expressed at the right time and in the correct tissue.

For a gene to be transcribed, RNAP must first bind upstream of it, to a *promotor site*. However, the expression rate of an individual gene is regulated by special DNA binding proteins, so called *transcription factors* (TFs). Through *facilitated diffusion* — a combination of a diffusive three-dimensional random walk in the cytoplasm followed by a one-dimensional diffusion along the DNA — they quickly locate and bind to their target binding site in the promotor region [19, 20]. From there their presence modulates the probability of RNAP binding to the promotor, resulting in either less mRNA being transcribed (*repression*) or more (*activation*), which will affect the overall concentration of the protein species in the cell. Repression of the gene expression can be achieved by a TF blocking RNAP from binding to the promotor site, and activation by a TF recruiting RNAP to the promotor site, by lowering the binding energy of RNAP. Usually, transcriptional networks have comparable number of positive (activating) and negative (repressing) edges (the interactions connecting two nodes) [14].

The TFs are proteins themselves, and are regulated by each other, thereby forming a *gene regulatory network*, where the genes (nodes) are connected by their transcriptional interaction (edges) into a directed graph, see Figure 1.2. The network can receive environmental input signals in the form of small molecules, or protein modifications, which changes the activity of a TF. This can happen on timescales of $\sim$ 1 msec [14]. Thus a signal feeding into the transcription network changes a TF causing a modification in the rate of transcription/translation of the gene products which in turn changes the overall concentration of the proteins ($\sim$ 1 h) in the cell. Some of the proteins carry out vital functions like DNA repair, metabolite synthesis, etc. while others, being TFs themselves, feed back to some node (gene) [14].

In this way the network architecture encodes how to perform computational tasks: it takes an input and processes the information according

to how nodes are connected and gives an output. This allows the organism to shut down redundant processes to conserve resources or direct them where they are needed.

An effective means for the gene to accomplish this is by regulating its own expression. The most common form of this *autoregulation* is negative repression, which allows the transcript level to quickly increase to its steady state value, and remain stable there. This works much like the mechanical equivalent to James Watt's centrifugal governor for steam regulation [14, 15].

Most genes are regulated by more than one TF. The gene expression resulting from the interaction at the promotor site, where TFs can block or promote each other, lends itself to a Boolean description of logic rules. We can imagine an AND-gate, where both TFs are required in order to switch the gene from an off-state to on-state, or an OR-gate where either one will suffice for the gene to be expressed [21]. Furthermore, one can have non-Boolean gates such as SUM-gate, where each TF binding to the promotor will increase the transcription rate of the gene [14].

Most TFs regulate more than one gene. The sign of the regulation mediated by a TF is highly correlated. The TF is either predominantly repressing or activating its targets. However, the sign of the incoming edges regulating the TF are less so [14]. This gives valuable information about how networks are shaped, as we soon shall see.

### 1.3.1 *The structure of functional networks*

The different networks of the cell exhibit similarities in both global as well as local structure. In parallel with the previously described protein–DNA transcription network, there is also an additional protein–protein and a protein–metabolite network. On a global scale, all three networks share the same type of *out-degree distribution* — the number of edges going out from a node — which follows an approximate power-law, where a few nodes are more important to the network and have many edges, while many nodes have only a few [14, 22]. Concerning TF–DNA networks, these show common features across function and species, such as a high degree of cooperative binding, overlapping gene function, as well as encompassing a large set of nodes [23].

Biological networks also bear a strong resemblance to engineered circuits, as they share common design criteria. They must be robust to random deletion of nodes, as well as be able to operate in noisy conditions, and manage all conceivable input ranges the network might

be subjected to [24, 25]. Furthermore, both biological and engineered networks show strong modularity, with only a few input and output nodes exposed to the wider network, but high degree of connectivity among the nodes of the module [24, 26]. This allows a network to adapt more readily to changing design specifications [26]. Also on the local scale of the biological network there is similarity to engineered circuits, by recurring elements, of so called *network motifs* [25].

Network motifs are small patterns that are found in evolved networks in far greater abundance than what would be expected from simple random connections [27]. The motifs are nature's recurring solution to frequent regulatory problems. These subgraphs can be though of as the building blocks of networks. Different network motifs are found in networks that have different function. Information processing networks, such as transcriptional networks, have a high frequency of the three node *feed forward loop* (FFL) motif [25], where node $Z$ is regulated directly through $X \rightarrow Z$ and indirectly through $X \rightarrow Y \rightarrow Z$ (see Figure 1.2). If the direct and indirect paths have the same effect on the target node $Z$ this *coherent* FFL acts as a noise filter, capable of ignoring either brief on-signals, or off-signals, depending on whether $X$ and $Y$ interact with node $Z$ as AND or OR gates, respectively [27]. When the direct and indirect paths differ in net sign (odd number of negative edges) this *incoherent* FFL can act as a pulse generator, as the indirect path will counteract the direct but with a delay [14]. But by what mechanism have these observed local patterns and global structure of networks emerged?

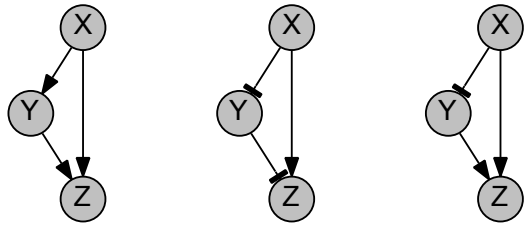


**Figure 1.2 Three node network motifs.** The first two graphs are coherent feed forward loop (FFL) network motifs, where the direct path from $X$ regulating the target node $Z$ has the same net effect on the target as the indirect path through the intermediary node $Y$. The rightmost motif is said to be an incoherent FFL, where the flat arrow represents repression counteracting the other activating triangular arrows.

### 1.3.2 *The construction of a network*

The common structure shared by the different networks of the cell, across a multitude of species, betray the forces by which they were shaped. The similarity can not be attributed to a common ancestor, as many of the studied networks are younger than the time of divergence from the ancestor [23]. It is warranted to ask if the over-abundance of network motifs and common large scale properties, shared in biological networks, are a result of their function, or are they simply the outcome of the evolutionary process? In the case of network motifs, it has been argued that they might exist due to being the optimal solution given the functional requirements of the network [14]. However, there are also indications that motifs are not strongly linked to network function [28].

The evolution of the networks follows the most probable path of least resistance through evolutionary space. Neutral evolution, that does not affect the phenotype, can open up new possibilities and remove fitness barriers, allowing new regions to be explored, under the constraints of what is permitted by biochemical and physical reactions [23].

The process of *gene duplication* is the main method for creating new genes [29]. It allows the original gene to maintain necessary function while its copy is free to diverge and explore new possibilities. If the gene has bifunctionality, the duplicates can subfunctionalize, by dividing the functions of the ancestral gene among them, and in that way become more specialized [30].

The sheer duplication of genes leads to an inherent high probability of network motifs [23, 31]. For instance, a FFL motif (Figure 1.2) could arise from a duplication event of node $Y$, followed by divergence where it turns into the new node $Z$ and receives an extra edge. Indeed, even in networks with no function, but evolved by duplication, motifs do appear [32]. However, since the TF binding sites are short ($\sim$ 10 bp [14, 19]) they are easily lost to mutational drift if not explicitly selected for, as a single point mutation in the binding site can abolish an edge. Gene duplications offer a conceivable explanation for how almost all genes in eukaryotes are regulated by more than two TFs, resulting in the high degree of connectivity observed [23]. Furthermore, through a neutral process of repeated gene duplication and removal, an approximate power-law degree distribution can emerge naturally [22]. Duplication of a whole genome is often followed by divergence and large gene loss [33].

The DNA is susceptible to mutations during duplication events. In the course of cell division, when the cell creates an identical copy

of itself, the DNA is replicated (*mitosis*), but imperfections can arise. Duplication errors can be introduced by misalignment during *crossover* events, which is the process where two chromosomes, one from each parent, are "blended" into a single copy (*meiosis*), lest the number of chromosomes of a species would double with each new generation. This is done by creating a copy that, at random *crossover points* along the sequence, changes which of the two chromosomes it is duplicating. The two "parent" chromosomes are aligned at the beginning of the crossover process, resulting in the blended offspring having the same length and a complete set of genes, from either parent [13, 34].

## 1.4 MODELLING OF GENETIC NETWORKS

Gene networks quickly become highly complex structures with increasing number of nodes, too complicated to intuitively understand. Through experiments we can start to unravel their intricacies. But to understand a fine mechanical clock we should not stop at prying it open and investigating its gears and springs; we must venture further by reconstructing it ourselves. This has been done experimentally, by building small synthetic gene networks in living cells [35, 36]. Although these systems are, in themselves, remarkable feats of experimental techniques, they are limited to a small size and by the currently available experimental methods. Instead, using mathematical reconstruction and modelling of gene networks, we shall know no such limitation.

By describing a network mathematically the dynamics of its interactions can be modelled and compared to known experimental data, followed by model experimentation that yield falsifiable predictions that can be verified or disproved by experiments. Even though the model is constructed manually, with preassigned input, the outcome can often be surprising.

The concentration level of each TF can be seen as describing the current state of the cell. Through a set of coupled ordinary differential equations (ODEs) that describe the change of state variables (TF concentration levels), $\boldsymbol{X} = (X_1, \ldots, X_n)$, the dynamics can be solved if the update function $\boldsymbol{f}(\boldsymbol{X})$, which describes the interactions, is known:

$$\frac{d\boldsymbol{X}}{dt} = \boldsymbol{f}(\boldsymbol{X}). \tag{1.1}$$

Here each component of $\boldsymbol{X}$ can describe the concentration of a protein at the current time step. The update function can model the gene

expression either as a binary Boolean function, being on or off, or as a continuous process.

The coupled equation system can be solved through numerical integration, where the system in next time step $t + \Delta t$ is computed from a simple Euler step, $X(t) - X(t + \Delta t) \approx \Delta t f(\boldsymbol{X})$, which follows from a series expansion of $X(t + \Delta t)$ [37]. In practice one typically uses higher order methods, with accuracy equivalent to a 4th order Runge-Kutta, or better [38].

### 1.4.1 *Law of mass action*

We now turn our attention to find the updating function that describes the system. Through the pioneering work of Norwegian chemist Peter Waage and his brother-in-law Cato Maximilian Guldberg, the *law of mass action* was derived at the end of the 19th century [39]. It describes a system in dynamical equilibrium such that the forward and backward reaction rates, $k_f$ and $k_b$ respectively, are in balance, in the following

$$\mathrm{A} + \mathrm{B} \underset{k_b}{\overset{k_f}{\rightleftharpoons}} \mathrm{C}. \tag{1.2}$$

The probability of the reactants colliding depends on their concentration, thus the chemical reaction rate is proportional to the product of (the mass of) the reactants,

$$\begin{aligned} \frac{d[A]}{dt} &= -k_f[A][B] + k_b[C] = \frac{d[B]}{dt} \\ \frac{d[C]}{dt} &= k_f[A][B] - k_b[C], \end{aligned}$$

where quantity $[X]$ in square brackets denote the concentration of $X$ in some arbitrary unit. This can be generalized to a system with $m$ reactants and $n - m$ products

$$\nu_1 X_1 + \ldots + \nu_m X_m \underset{k_b}{\overset{k_f}{\rightleftharpoons}} \nu_{m+1} X_{m+1} + \ldots + \nu_n X_n, \tag{1.3}$$

with *stoichiometric coefficients* $\nu_i$ defining the number of molecules of each reactant $X_i$ which is needed for the reaction to occur. The generalized chemical reaction in eq. (1.3) forms an ODE system:

$$\begin{aligned} \frac{d[X_i]}{dt} &= -k_f \nu_i X_1^{\nu_1} \cdot \ldots X_m^{\nu_m} + k_b \nu_i X_{m+1}^{\nu_{m+1}} \cdot \ldots X_n^{\nu_n} \;\; i = 1, \ldots, m \\ \frac{d[X_j]}{dt} &= k_f \nu_j X_1^{\nu_1} \cdot \ldots X_m^{\nu_m} - k_b \nu_j X_{m+1}^{\nu_{m+1}} \cdot \ldots X_n^{\nu_n} \;\; j = m+1, \ldots, n. \end{aligned}$$

For chemical equilibrium the ratio of the reaction rates must equal the chemical equilibrium, thus

$$k_{\text{eq}} = \frac{k_f}{k_b} = \frac{[X_{m+1}]^{\nu_{m+1}} \cdot \ldots \cdot [X_n]^{\nu_n}}{[X_1]^{\nu_1} \cdot \ldots \cdot [X_m]^{\nu_m}}.$$

However, in our transcription networks we are concerned with reactions where TFs bind to a site on the DNA to regulate the production of some protein, $X$, without itself being consumed. If the binding TF is an activator it acts as an enzyme catalysing the reaction, although during the time it is bound to the DNA it can not partake in any other reaction. We get Michaelis-Menten kinetics [14, 40]:

$$\text{TF} + \text{DNA} \underset{k_b}{\overset{k_f}{\rightleftharpoons}} \text{TF–DNA} \overset{k_c}{\to} \text{TF} + \text{DNA} + X \tag{1.4}$$

This gives the equation system:

$$\frac{d[\text{TF}]}{dt} = -k_f[\text{TF}][\text{DNA}] + (k_b + k_c)[\text{TF–DNA}] \tag{1.5a}$$

$$\frac{d[\text{TF–DNA}]}{dt} = k_f[\text{TF}][\text{DNA}] - (k_b + k_c)[\text{TF–DNA}] \tag{1.5b}$$

$$\frac{d[\text{DNA}]}{dt} = -\frac{d[\text{TF–DNA}]}{dt} \tag{1.5c}$$

$$\frac{d[X]}{dt} = k_c[\text{TF–DNA}]. \tag{1.5d}$$

We assume the first reaction is much faster than the last ($k_f, k_b \gg k_c$), so the reaction is in quasi-equilibrium.[2] From the chemical equilibrium of the intermediate, rate limiting, process and the observation that the total amount of DNA is constant $[\text{DNA}_\text{T}] = [\text{DNA}] + [\text{TF–DNA}]$, we get

$$[\text{TF–DNA}] = k_{eq}[\text{DNA}][\text{TF}] = (k_b + k_c)[\text{DNA}][\text{DNA}_\text{T} - \text{TF–DNA}],$$

from which we get the probability of the TF being bound to the DNA

$$P_{\text{bound}} = \frac{[\text{TF–DNA}]}{[\text{DNA}_\text{T}]} = \frac{[\text{TF}]}{\frac{k_b + k_c}{k_f} + [\text{TF}]}, \tag{1.6}$$

which is known as the *Michaelis-Menten equation*, and is useful for describing many process in biology [14]. Inserted in eq. (1.5d) this gives the *gene activity*, through its production rate of $[X]$

$$\frac{d[X]}{dt} = \frac{V_{\max}[\text{TF}]}{K_M + [\text{TF}]} \tag{1.7}$$

2 Typically, TF binding to DNA reaches equilibrium in seconds [14].

where we have introduced the Michaelis-Menten constant $K_M = (k_b + k_c)/k_f$, and $V_{\max} = k_c[\mathrm{DNA_T}]$ which is the maximum production rate when [TF] has saturated the system, see Figure 1.3A.

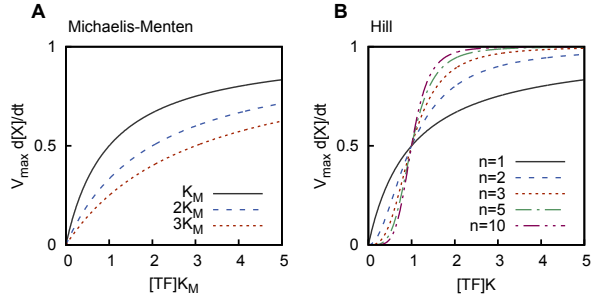


**Figure 1.3 .** The resulting modelled production of protein $X$ as function of concentration of TF. (A) Michaelis-Menten kinetics, eq. (1.7), and (B) Hill equation, (1.9) for different degrees of cooperativity, $n$. The production rate saturates at $V_{\max}$.

For gene transcription networks, cooperativity can be a key player. To model this we require several transcription factors, $n$ in total, to interact for a reaction to happen,

$$n\mathrm{TF} + \mathrm{DNA} \underset{k_b}{\overset{k_f}{\rightleftharpoons}} n\mathrm{TF}\text{–}\mathrm{DNA} \xrightarrow{k_c} n\mathrm{TF} + \mathrm{DNA} + X. \tag{1.8}$$

resulting in

$$\frac{d[X]}{dt} = \frac{V_{\max}[\mathrm{TF}]^n}{K^n + [\mathrm{TF}]^n} \tag{1.9}$$

with Hill coefficient $n$ and Hill constant $K$, which is the dissociation equilibrium constant, giving the rate between DNA-binding ratio and DNA-unbinding ratio [40]. If cooperativity is not required but merely assisted, or otherwise not fully understood, the Hill coefficient need not be integer [40].

Hill functions can describe the production (and its regulation) of a gene product. If the interactions are not fully understood one usually fits $n$ and $K$ to experimental data. For this purpose, a least squares method is commonly used, which we will have reason to get back to in Section 1.5.

### 1.4.2 *A three-node network*

As an instructive example we now consider the small network in Figure 1.4A. It consists of three nodes connected in a loop by the same number of edges. Each component represses the next and is in turn itself being repressed by the previous. While giving an overview of the system, the graph representation does not reveal much information on the exact mechanism of the interactions. Unlike eq. (1.6), the interaction is now repressive, instead of activating. If $X_1$ is being repressed by $X_3$, its production will depend on the probability of $X_3$ *not* being bound:

$$P_{\text{not-bound}} = 1 - \frac{X_3^n}{K^n + X_3^n} = \frac{K^n}{K^n + X_3^n}. \tag{1.10}$$

Thus, with a linear degradation term, the three coupled ODE equations can be describe by:

$$\frac{dX_i}{dt} = k_i \frac{K_i^{n_i}}{K_i^{n_i} + X_{i-1}^{n_i}} - d_i X_i, \qquad i = 1, 2, 3. \tag{1.11}$$

Here, the first term is our Hill function, where the production is repressed as motivated in eq. (1.10). The second term represents the degradation of $X_i$. In the absence of production, we are left with simple exponential decay. We can interpret each component $X_i$ as the concentration of a TF. Thus eq. (1.11) includes transcription, transport to/from the nucleus (if in a eukaryote) and translation as a single step.

The output concentration over time of each component, for a set of parameters (see table 3.1, p. 47), can be made to oscillate (Figure 1.4B). We shall have cause to return to the fundamental traits needed for a system to exhibit such properties. A similar network, consisting of three proteins in a closed loop, each repressing the next, was built in a real cell and borough to oscillate in a similar manner [36].

## 1.5 MODEL FITTING

In order to evaluate a model, we compare its prediction to data representing the very system that the model aims to describe. Models often have free parameters that need to be determined by fitting them to data. This involves minimizing the deviation of the observations $\boldsymbol{y} = (y_1, \ldots, y_N)^T$, at corresponding measurement points $\boldsymbol{x} = (x_1, \ldots, x_N)^T$, with the estimating function $\boldsymbol{f}(\boldsymbol{x}; \boldsymbol{\lambda}) = (f(x_1; \boldsymbol{\lambda}), \ldots, f(x_N; \boldsymbol{\lambda}))^T$, with respect to

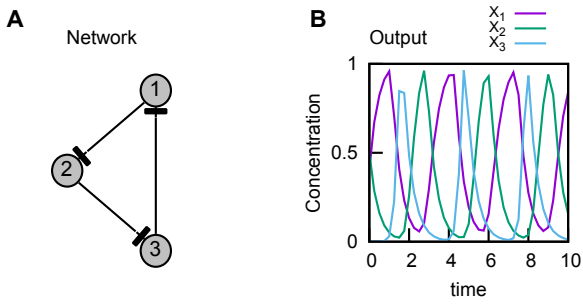


**Figure 1.4 A three node network.** (A) The network is connected in a loop, where each edge represses the next. (B) The output from each node, normalized to unity, oscillates with time, for suitable parameters chosen in eq. (1.11).

the $K$ parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)^T$. This can be summarized as minimizing the residuals

$$\boldsymbol{\Delta}(\boldsymbol{\lambda}) = \boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}; \boldsymbol{\lambda}). \tag{1.12}$$

The two main methods for determining the optimal model parameter estimators are the least squares method and the maximum likelihood method. The following derivations are adapted from van den Bos [41].

### 1.5.1 *Least squares method*

One of the standard methods for fitting a model to data is the *least squares method*. It can be defined from the *weighted* least squares minimization criterion [41]

$$\chi^2(\boldsymbol{\lambda}) = \boldsymbol{\Delta}^T(\boldsymbol{\lambda}) \boldsymbol{R} \boldsymbol{\Delta}(\boldsymbol{\lambda}), \tag{1.13}$$

where $\boldsymbol{R}$ is a known positive definite $(N \times N)$ weighting matrix. If this matrix is diagonal, eq. (1.13) is reduced to $\chi^2(\boldsymbol{\lambda}) = \sum_{i=1}^{N} r_{ii} \Delta_i^2(\boldsymbol{\lambda})$, which becomes an *ordinary* least squares method if $r_{ii} = 1 \ \forall i$, with minimization criterion: $\chi^2 = \boldsymbol{\Delta}^T \boldsymbol{\Delta}$.

At the stationary point, where $\lambda = \overline{\lambda}$ is the estimator of the unknown true parameters $\overline{\overline{\lambda}}$ that we seek, the gradient of eq. (1.13) is the null vector and defines $K$ *normal equations* for the least squares criterion:

$$\frac{\partial \chi^2(\boldsymbol{\lambda})}{\partial \lambda_k} = -2 \frac{\boldsymbol{f}^T(\boldsymbol{x}; \boldsymbol{\lambda})}{\partial \lambda_k} \boldsymbol{R} \boldsymbol{\Delta}(\boldsymbol{\lambda}) = 0, \quad k = 1, \ldots, K, \tag{1.14}$$

and likewise for the ordinary least squares, but with weights given by the unit matrix.

When the expectation model is linear, the expectation of the observable may be written as

$$\langle \boldsymbol{y} \rangle = \boldsymbol{f}(\boldsymbol{x}; \boldsymbol{\lambda}) = \boldsymbol{X}\overline{\overline{\boldsymbol{\lambda}}}, \tag{1.15}$$

where $\boldsymbol{X}$ is a known nonsingular $(N \times K)$ matrix independent of $\boldsymbol{\lambda}$. From this it follows that the least squares criterion, eq. (1.13), becomes

$$\begin{aligned} \chi^2(\boldsymbol{\lambda}) &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\lambda})^T \boldsymbol{R}\, (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\lambda}) \\ &= \boldsymbol{y}^T \boldsymbol{R}\boldsymbol{y} - \boldsymbol{\lambda}^T \boldsymbol{X}^T \boldsymbol{R}\boldsymbol{y} - \boldsymbol{y}^T \boldsymbol{R}\boldsymbol{X}\boldsymbol{\lambda} + \boldsymbol{\lambda}^T \boldsymbol{X}^T \boldsymbol{R}\boldsymbol{\lambda} \\ &= \boldsymbol{y}^T \boldsymbol{R}\boldsymbol{y} - 2\boldsymbol{\lambda}^T \boldsymbol{X}^T \boldsymbol{R}\boldsymbol{y} + \boldsymbol{\lambda}^T \boldsymbol{X}^T \boldsymbol{R}\boldsymbol{\lambda}, \end{aligned} \tag{1.16}$$

which leads to the normal equations

$$\frac{\partial \chi^2(\boldsymbol{\lambda})}{\partial \lambda} = -2\boldsymbol{X}^T \boldsymbol{R}\boldsymbol{y} + 2\boldsymbol{X}^T \boldsymbol{R}\boldsymbol{X}\boldsymbol{\lambda} = 0, \quad k = 1, \ldots, K. \tag{1.17}$$

Thus we get $\boldsymbol{X}^T \boldsymbol{R}\boldsymbol{X}\boldsymbol{\lambda} = \boldsymbol{X}^T \boldsymbol{R}\boldsymbol{y}$ from which we find our estimating parameters

$$\overline{\boldsymbol{\lambda}} = (\boldsymbol{X}^T \boldsymbol{R}\boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{R}\boldsymbol{y} \equiv \boldsymbol{A}\boldsymbol{y}, \tag{1.18}$$

where in the last step we defined, for convenience, the matrix $\boldsymbol{A}$. Next, taking the expectation value of our parameter estimator, results in

$$\langle \overline{\boldsymbol{\lambda}} \rangle = \langle \boldsymbol{A}\boldsymbol{y} \rangle = \boldsymbol{A}\langle \boldsymbol{y} \rangle = \boldsymbol{A}\boldsymbol{X}\overline{\overline{\boldsymbol{\lambda}}} = \overline{\overline{\boldsymbol{\lambda}}}, \tag{1.19}$$

where we used eq. (1.15), and from eq. (1.18) we note that $\boldsymbol{A}\boldsymbol{X}$ is the unit matrix. Thus, if the assumption of the linearity of the estimating model is correct, and that the weighting matrix is know, the weighted least squares estimator is an *unbiased* estimator, free of systematic errors.

To get an estimate of the nonsystematic errors in the parameter fit, we can determine its covariance matrix. First we note: $\overline{\boldsymbol{\lambda}} - \langle \overline{\boldsymbol{\lambda}} \rangle = \boldsymbol{A}(\boldsymbol{y} - \langle \boldsymbol{y} \rangle)$, thus

$$\begin{aligned} \operatorname{cov}(\overline{\boldsymbol{\lambda}}, \overline{\boldsymbol{\lambda}}) &= \langle (\overline{\boldsymbol{\lambda}} - \langle \overline{\boldsymbol{\lambda}} \rangle)(\overline{\boldsymbol{\lambda}} - \langle \overline{\boldsymbol{\lambda}} \rangle)^T \rangle \\ &= \langle \boldsymbol{A}(\boldsymbol{y} - \langle \boldsymbol{y} \rangle)(\boldsymbol{y} - \langle \boldsymbol{y} \rangle)^T \boldsymbol{A}^T \rangle \\ &= \boldsymbol{A}\langle (\boldsymbol{y} - \langle \boldsymbol{y} \rangle)(\boldsymbol{y} - \langle \boldsymbol{y} \rangle)^T \rangle \boldsymbol{A}^T \\ &= \boldsymbol{A}\boldsymbol{C}\boldsymbol{A}^T, \end{aligned} \tag{1.20}$$

or when written explicitly, from eq. (1.18), and using the symmetry of the matrices $\boldsymbol{R}$ and $(\boldsymbol{X}^T\boldsymbol{R}\boldsymbol{X})^{-1}$:

$$\operatorname{cov}(\overline{\boldsymbol{\lambda}}, \overline{\boldsymbol{\lambda}}) = (\boldsymbol{X}^T\boldsymbol{R}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{R}\boldsymbol{C}\boldsymbol{R}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{R}\boldsymbol{X})^{-1}. \tag{1.21}$$

We see that the parameter (co)variance depends on the measurement points $\boldsymbol{X}$, the covariance $\boldsymbol{C}$ of the observable $\boldsymbol{y}$ and the choice of weighting matrix $\boldsymbol{R}$.[3] The variance for the weighted linear least squares method is minimized by the choice $\boldsymbol{R} = \boldsymbol{C}^{-1}$, which yields a covariance of the estimated parameters as [41]:

$$\operatorname{cov}(\overline{\boldsymbol{\lambda}}, \overline{\boldsymbol{\lambda}}) = (\boldsymbol{X}^T\boldsymbol{C}^{-1}\boldsymbol{X})^{-1}, \tag{1.22}$$

with error of the estimated parameters as the diagonal elements.

### 1.5.2 *Maximum likelihood method*

Provided that the probability density function of the observable $\boldsymbol{y}$ and its dependence on the parameters $\boldsymbol{\lambda}$ are known, then the *maximum likelihood* method is applicable. The method has several desirable traits, such as, under general conditions, $\overline{\lambda} - \overline{\overline{\lambda}}$ tending to a normal distribution with increasing observations, with zero mean and minimal (co)variance [41]. The *likelihood function* is based on the joint probability distribution of the observations where the fixed exact parameters $\overline{\overline{\boldsymbol{\lambda}}}$ are replaced with independent variables $\boldsymbol{\lambda}$, and the probability is parametric in the observations,

$$p(\boldsymbol{y}; \boldsymbol{\lambda}). \tag{1.23}$$

The maximum likelihood estimator of $\overline{\overline{\boldsymbol{\lambda}}}$ are the parameters, $\overline{\boldsymbol{\lambda}}$, that maximizes the likelihood function, or alternatively, that maximizes the *log-likelihood function*:

$$q(\boldsymbol{y}; \boldsymbol{\lambda}) = \ln p(\boldsymbol{y}; \boldsymbol{\lambda}). \tag{1.24}$$

For the most probable parameters, $\boldsymbol{\lambda} = \overline{\boldsymbol{\lambda}}$, the gradient of $q$ is equal to the null vector, and we get $K$ *likelihood equations*:

$$\frac{\partial q(\boldsymbol{y}; \boldsymbol{\lambda})}{\partial \lambda_k} = 0, \quad k = 1, \ldots, K. \tag{1.25}$$

3 The result of eq. (1.21) is alluded to in paper I as "eq. 5.253 of van den Bos [41]", which we there extend into the nonlinear regime.

If the observations $\boldsymbol{y}$ are independent stochastic variables their likelihood function may be written on the form

$$p(\boldsymbol{y};\boldsymbol{\lambda}) = \prod_i^N p_i(y_i;\boldsymbol{\lambda}) \tag{1.26}$$

and log-likelihood

$$q(\boldsymbol{y};\boldsymbol{\lambda}) = \sum_i^N q_i(y_i;\boldsymbol{\lambda}). \tag{1.27}$$

If the observables are normally distributed, as often is the case due to the central limit theorem [42, 43], the log-likelihood function is

$$\begin{aligned} q(\boldsymbol{y};\boldsymbol{\lambda}) &= \ln\left(\frac{1}{(2\pi)^{N/2}\sqrt{\det \boldsymbol{C}}}\exp\left(-\frac{1}{2}\boldsymbol{\Delta}^T(\boldsymbol{\lambda})\boldsymbol{C}^{-1}\boldsymbol{\Delta}(\boldsymbol{\lambda})\right)\right) \\ &= -\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln(\det \boldsymbol{C}) - \frac{1}{2}\boldsymbol{\Delta}^T(\boldsymbol{\lambda})\boldsymbol{C}^{-1}\boldsymbol{\Delta}(\boldsymbol{\lambda}), \end{aligned} \tag{1.28}$$

from which we get $K$ likelihood equations by demanding that the gradient is equal to the null vector at the stationary point

$$\frac{\partial \boldsymbol{f}^T(\boldsymbol{\lambda})}{\partial \lambda_k}\boldsymbol{C}^{-1}\boldsymbol{\Delta}(\boldsymbol{\lambda}) = 0, \quad k = 1,\ldots,K. \tag{1.29}$$

For jointly normally distributed observations, the weighted least squares estimator is the same as the maximum likelihood estimator, eq. (1.14), with the inverse covariance of the observables as weighting matrix, provided $\boldsymbol{C}$ does not depend on the unknown parameters. From eq. (1.29) it follows, in the same way as for the weighted linear least squares, that the estimator of a linear model is unbiased.

## 1.6 THE CIRCADIAN CLOCK

We live in a world of periodic change. Hence, most life has evolved endogenous mechanisms which can accurately predict the diurnal cycle, and respond in anticipation of dawn and dusk, rather than react to the periodic environmental changes after they occur [44].

By predicting when a cell function is needed, resources can be directed towards that aim, and likewise conserved when not needed, thereby improving the survival ability. Both mammals and plants show improved health and survival fitness when their internal clock is synchronized with the environment [45–47]. It is also found that arrhythmic plants

grow far worse than plants with a clock with the wrong period [47]. In this thesis we will focus on the *circadian clock* (from Latin: circa diem, meaning approximately daily) of plants.

The earliest written observation of circadian clocks originates from the fourth century BC. At that time, Aristotle had encouraged his student, Alexander the Great, to defeat Persia, and to be "a hegemon (leader) of Greeks and a despot to the barbarians, to look after the former as after friends and relatives, and to deal with the latter as with beasts or plants" [48]. It was during Alexander's the campaign in Tylos (modern Bahrain) that Androsthenes made note of the leaf movement of the Tamarind tree which tracked the motion of the sun. Close to two millennia came to pass before the discovery, in 1729 by french astronomer de Marian, that the rhythmic leaf movement persisted also for plants held in constant darkness. Yet another century later came the realization that these are not exactly 24 h periodic, but *circadian*, indicating that the plant is not just using external environmental signals but indeed has an internal clock [44].

The plant circadian clock is remarkably robust despite the many challenges it faces. It relies on biochemical reactions, yet it is able to operate under a wide range of temperature fluctuations ($\sim 20$ degrees) [49]. The clock is *entrained* by using the light as its main *zeitgeber* (German: time giver) to match its phase with the environment. Usually, one measures the state of the clock from the *zeitgeber time* (ZT), marking the time of when light is turned on. To prevent the clock from resetting in the middle of the day, the response to the light input is time-dependent, or *gated*; meaning its importance is primarily during dawn and dusk, since there is no seasonal information in light variation in the middle of the day [44, 49]. In the absence of its main input the clock can be entrained by as little as a two degree temperature fluctuation, or even by changes in the sugar solution it grows on in the laboratory [50, 51].

The importance of the clock is demonstrated by the sheer scope of genes that are regulated by it. In *Arabidopsis* roughly a third of the genes are directly regulated by the clock and up to 89% show diurnal rhythm, be that from cyclic external environmental stimuli, like light or temperature, or independent of environment [49, 52]. Among the many processes controlled by the clock we find both photosynthesis and enzyme activity. There is also a strong overrepresentation of genes regulating stress response as well as hormones like *auxin*, which is a plant growth hormone [44, 49, 53]. The clock predicts seasonal changes

by comparing the external photoperiod with its internal state. This allows the clock to control fragrance emission, germination [44], and flowering [54–57], furthermore, at the onset of winter the plant can pre-treat its cells to withstand cold [58].

To investigate the direct benefit of a clock, experimentalists have created mutant plants, by removing genes to partially change the clock mechanism. Plants with a normal 24 hour period (T24) clock grow better (fixate more carbon, and contain more chlorophyll) when subjected to a matching period of light/dark cycle [47]. Likewise, both short-period mutants (T20), and long-period mutants (T28) perform best when their respective environment matches their *free running period* — their intrinsic period when subjected to constant light or constant dark, in order to not be reset by dawn and dusk [47].

### 1.6.1 *What makes the clock tick?*

The circadian clock stems from oscillations of protein concentrations in cells. A three-node system, e.g. Figure 1.4, is the smallest network that exhibits stable oscillations [59]. There are several additional requirements on a network for oscillations to emerge. First, a negative feedback loop is required for the system to bring itself back to its starting point. This makes the system converge to a *limit cycle*, where the variable set is repeated in a cyclic manner, forming a closed loop in phase space. Additionally, the system needs to retain a memory of its past states, to avoid convergence to a steady state. This is achieved by introducing a time delay by components acting indirectly on their targets, together with balancing the timescales of the processes. Furthermore, the rate laws must be sufficiently non-linear to destabilize the system from its stable state [59].

Oscillations of protein concentrations can be experimentally resolved for individual cells, each having its own autonomous clock, needing no external input to persist [40]. The genes of each cell are rhythmically expressed as a result of the regulatory interactions encoded in the transcription network. The cells need not share phase information between each other [60]; different tissues can have different phase, but the main clock in mammals stem from the protein oscillation in cells of the *hypothalamus* [52].

The circadian gene network is diverse across different domains of life. The transcription factors which constitute the core clock genes in eukaryotes like the fungus *Neurospora crassa* (*FRQ* and *WC*), the

plant *Arabidopsis thaliana* (*CCA1* and *TOC1*), the insect *Drosophila melanogaster* (*PER* and *TIM*), and the mammal *Mus musculus* (*BMAL1* and *PERIOD*) are not shared, indicating the clock has developed independently across taxa [61, 62].[4]

Although different in execution, the gene networks share common design principles. Through the trial-and-error process of rewiring and tinkering nature seem to converge on the same solution [24]. Each implementation of a period predicting circuit consist of a gene network with transcriptional and translational interaction with feedback loops (TTFL) for generating robust oscillations with correct period, phase and amplitude [59, 61]. The multiple feedback loops and light input of the TTFL network allows it to track both dawn and dusk, as well as withstand seasonal changes in day length, and input noise [49, 63].

However, it has been shown that the clock of prokaryotic cyanobacteria does not only rely on a TTFL, but also on a post transcription-translation oscillator (PTO). The two oscillators are mainly independent of each other, but combined give a robust clock [60]. Even more intriguing is the discovery of circadian oscillations in eukaryote cells such as found in human red blood cells [64], which lack a cell nucleus and therefore have no means for a TTFL circuit. Alternative means for oscillations have also been identified in algae [65].

Recent investigations indicate that a PTO proto-clock is preserved across all probed phylogenetic domains. It has been found that a separate post translational clock is shared in prokaryote bacteria, as well as in eukaryotes such as mouse, fruit fly, and fungus. It manifests itself through oscillations in the oxidation level of a protein (peroxiredoxin). If either the TTFL or PTO clock of the organism is disabled, the remaining one will continue unabated, although at a different phase [61, 62]. The advantage of having two separate clocks could be higher resistance to stochastic molecular noise, and a PTO based clock gives stability during the metabolic stress and dilution at high cell division rates [52, 60].

### 1.6.2 *The transcriptional clock in Arabidopsis*

The clock in the plant *Arabidopsis thaliana*, known under the common name "thale cress",[5] or the more descriptive one: "mouse-ear cress", has been the focus of much research over the past decades. Through an

4 It is worth pointing out that although the *PERIOD* gene is homologous in mouse and fruit fly, they appear to have different functions [62].

5 Known as *Backtrav* in Swedish, *Vårskrinneblom* in Norwegian, *Gåsemad* in Danish, and *Schaumkressen* in German.

iterative process of experimentation and modelling, its inner workings has been probed ever further. The models recreate existing data, and make predictions for where no data yet exists, that the experimentalists then can verify or refute. The experimentalists typically measure time series of clock gene expression in *wild type* (wt) plants, which have all genes fully functional, and compare these to mutant plants where one, or several, genes have been "knocked out" rendering them effectively non-functional [66]. Also partially working mutant plants can yield important clues to decipher the intricate workings of the gene regulatory network.

The initial *Arabidopsis* circadian clock model started as a simple system with two genes, each having three components (mRNA, cytosolic and nucleic protein), connected in a loop with feedback.[6] This first model, conceived in 2005 by Locke *et al.* [67, 68], treated the two closely related morning expressed genes *CIRCADIAN CLOCK ASSOCIATED 1* (*CCA1*) and *LATE ELONGATED HYPOCOTYL* (*LHY*) as a single node [69–71], which represses the evening expressed gene *TIMING OF CAB EXPRESSION 1* (*TOC1*), which in turn regulates *CCA1/LHY* and thus closes the loop [66]. It is believed that *CCA1* and *LHY* need to form a homodimer or heterodimer in order to bind to DNA [72] where they typically act as repressors [73]. In spite of the close relation of the two morning genes, they are only partially redundant, as loss of either one will affect the clock by shortening the period, in an additive manner [74, 75].

It was long believed that TOC1 activates *CCA1* transcription [67, 68, 76]. In a *toc1* loss-of-function mutant[7] the levels of mRNA of both *CCA1* and *LHY* is low; however, this is also the case for when *TOC1* is over-expressed, resulting in a drastic increase of the *TOC1* mRNA concentration, and consequently the TOC1 protein [66]. The confusion was cleared when it was found that TOC1 binds to DNA and can regulate the *CCA1/LHY* expression directly [77], by repression [78].

The early two-component clock model, consisting of *CCA1/LHY* and *TOC1*, was extended by including more genes to account for period lengthening and shortening by mutants of genes defined in the model [66]. Among them were the *PSEUDO RESPONSE REGULATORS 9,7,5* (*PRR9*, *PRR7*, *PRR5*), which, together with *TOC1* (also known as *PRR1*), form

6 For a schematic overview, see Figure III.S1, p. 161.

7 We here use the same notation as used for *Arabidopsis* where genes are written in cursive and its gene product (protein) in upright; in addition, if it is the (functional) wild type it is written in uppercase, and if mutant in lowercase.

a "*PRR* wave" by their sequential expression starting with *PRR9* in the morning [79]. Each component in the wave can bind to DNA [77] to repress *CCA1/LHY* [80], thereby helping to turn off the earlier expressed morning genes [81]. Since CCA1/LHY regulate the *PRR*s, the loop is closed [81].

The multiple feedback loops confer the clock redundancy against gene loss. In order to render the clock arrhythmic, multiple genes need to be knocked-out, such as the triple mutant *prr5;prr7;prr9* [82], or *cca1;lhy;toc1* [83]. Nonetheless, a non-functional *EARLY FLOWERING 4* (*ELF4*) gene stops all oscillation of *TOC1*, *CCA1*, and *LHY* in the absence of rhythmic light, as this evening expressed gene is required for activating the morning genes [84, 85]. The ELF4 transcript represses *TOC1* and another gene, *LUX ARRHYTHMO* (*LUX*), which is required for the expression of *ELF4* itself [86]. If either *LUX* or the gene *EARLY FLOWERING 3* (*ELF3*) is over-expressed, they can counteract the detrimental effect of the *elf4* mutant [87]. Both ELF3 and ELF4 target the promotor region of *PRR9* [87, 88], where also LUX has a binding site [87, 89]. The three genes have similar phenotypic effects [87], and are believed to form a multiprotein evening complex (EC), where ELF3 tether ELF4 and LUX together, as they do not interact directly [86]. Through EC, ELF3 represses many genes together with ELF4 during the night, among them *PRR9*, to which LUX helps it bind [87, 89]. Furthermore, it is found that both *LUX* and the gene *NOX* help the formation of the EC [90]. The latter is regulated negatively by CCA1 [91], as is the former [92, 93].

In addition, there are yet other genes that play a part in regulating components of the clock, but are not yet included in any models, such as *CCA1 HIKING EXPEDITION* (*CHE*) which binds to the promotor region of *CCA1* and decreases its activity when in high concentration [94], and *EARLY BIRD* (*EBI*) which interacts with another clock controlled protein, ZEITLUPE (ZTL), through a not yet fully understood mechanism [95].

### 1.6.3 *Post translational circadian regulation in Arabidopsis*

There are several components of the clock in *Arabidopsis* that are subject to post translational modifications. An early gene to be included in the models was *GIGANTEA* (*GI*) [68]. It is not regarded to encode for a transcription factor, but it is believed to be cyclically regulated by TOC1, and stabilize the oscillation of *ZTL* [96], that in turn will regulate both TOC1 and PRR5 proteins [97, 98] (but no other *PRR* [99]), by marking TOC1 [97] and PRR5 [100] for degradation. The GI protein is

also repressed by LHY [93] and ELF4 [85], and degraded by the protein CONSTITUTIVE PHOTOMORPHOGENIC 1 (COP1), which acts in this regard with ELF3 [101].

Localization of a protein in the cell can provide the means of regulating transcription. This can be achieved by controlling how much transcript is released from the nucleus into the cytoplasm, where it would be translated into a working protein [13]. Conversely, if a protein is a TF, it will not be able to function (if in eukaryote) unless it is located in the nucleus where the DNA molecule resides. In *Arabidopsis* TOC1 is transported into the nucleus by PRR5 [102], by forming a dimer which helps TOC1 accumulate in the nucleus [103], where it is protected from degradation from ZTL, which is only found in the cytosol [96].

REFERENCES

1. E. P. Wigner, "The unreasonable effectiveness of mathematics in the natural sciences," *Communications on pure and applied mathematics*, vol. 13, no. 1, pp. 1–14, 1960.
2. A. Einstein, *The ultimate quotable Einstein*. Princeton University Press, 2010.
3. R. Metzler and J. Klafter, "The random walk's guide to anomalous diffusion: a fractional dynamics approach," *Physics Reports*, vol. 339, no. 1, pp. 1–77, 2000.
4. Lucretius, *On The Nature of Things*, vol. 785 of *Project Gutenberg*. P.O. Box 2782, Champaign, IL 61825-2782, USA: Project Gutenberg, 1997.
5. R. E. Kennedy, *A student's guide to Einstein's major papers*. Oxford University Press, 2012.
6. P. Nelson, *Biological physics*. WH Freeman New York, 2004.
7. E. Barkai, Y. Garini, and R. Metzler, "Strange kinetics of single molecules in living cells," *Phys. Today*, vol. 65, no. 8, p. 29, 2012.
8. A. Einstein, "Über die von der molekularkinetischen theorie der wärme geforderte bewegung von in ruhenden flüssigkeiten suspendierten teilchen," *Annalen der Physik*, vol. 17, no. 8, pp. 549–560, 1905.
9. E. Schrödinger, *What is life?: With mind and matter and autobiographical sketches*. Cambridge University Press, 1992.
10. P. L. Luisi, "About various definitions of life," *Origins of Life and Evolution of the Biosphere*, vol. 28, no. 4-6, pp. 613–622, 1998.
11. D. E. Koshland, "The seven pillars of life," *Science*, vol. 295, no. 5563, pp. 2215–2216, 2002.
12. R. Dawkins, *The Selfish Gene: 30th Anniversary Edition*. ISSR library, OUP Oxford, 2006.
13. B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell*. Garland Science, 1997.
14. U. Alon, *An introduction to systems biology: design principles of biological circuits*. CRC press, 2006.
15. R. Dawkins, *The Blind Watchmaker: Why the Evidence of Evolution Reveals a Universe Without Design*. National bestseller. Science, Norton, 1986.
16. T. A. Kunkel, "DNA replication fidelity," *Journal of Biological Chemistry*, vol. 279, no. 17, pp. 16895–16898, 2004.

17. J. W. Drake, B. Charlesworth, D. Charlesworth, and J. F. Crow, "Rates of spontaneous mutation," *Genetics*, vol. 148, no. 4, pp. 1667–1686, 1998.
18. F. H. Crick, "The origin of the genetic code," *Journal of molecular biology*, vol. 38, no. 3, pp. 367–379, 1968.
19. G.-W. Li, O. G. Berg, and J. Elf, "Effects of macromolecular crowding and DNA looping on gene regulation kinetics," *Nature Physics*, vol. 5, no. 4, pp. 294–297, 2009.
20. P. H. von Hippel and O. Berg, "Facilitated target location in biological systems.," *Journal of Biological Chemistry*, vol. 264, no. 2, pp. 675–678, 1989.
21. N. E. Buchler, U. Gerland, and T. Hwa, "On schemes of combinatorial transcription logic," *Proc. Natl. Acad. Sci. USA*, vol. 100, no. 9, pp. 5136–5141, 2003.
22. S. R. Proulx, D. E. L. Promislow, and P. C. Phillips, "Network thinking in ecology and evolution," *Trends Ecol. Evol.*, vol. 20, no. 6, pp. 345–353, 2005.
23. T. R. Sorrells and A. D. Johnson, "Making sense of transcription networks," *Cell*, vol. 161, no. 4, pp. 714–723, 2015.
24. U. Alon, "Biological networks: the tinkerer as an engineer," *Science*, vol. 301, no. 5641, pp. 1866–1867, 2003.
25. R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: simple building blocks of complex networks," *Science*, vol. 298, no. 5594, pp. 824–827, 2002.
26. N. Kashtan and U. Alon, "Spontaneous evolution of modularity and network motifs," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 102, no. 39, pp. 13773–13778, 2005.
27. S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, "Network motifs in the transcriptional regulation network of *Escherichia coli*," *Nature genetics*, vol. 31, no. 1, pp. 64–68, 2002.
28. J. F. Knabe, C. L. Nehaniv, and M. J. Schilstra, "Do motifs reflect evolved function? — no convergent evolution of genetic regulatory network subgraph topologies," *Biosystems*, vol. 94, no. 1, pp. 68–74, 2008.
29. S. A. Teichmann and M. M. Babu, "Gene regulatory network growth by duplication," *Nature genetics*, vol. 36, no. 5, pp. 492–496, 2004.
30. A. L. Hughes, "Gene duplication and the origin of novel proteins," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 102, no. 25, pp. 8791–8792,

2005.

31. P. D. Kuo, W. Banzhaf, and A. Leier, "Network topology and the evolution of dynamics in an artificial genetic regulatory network model created by whole genome duplication and divergence," *Biosystems*, vol. 85, no. 3, pp. 177–200, 2006.
32. W. Banzhaf and P. D. Kuo, "Network motifs in natural and artificial transcriptional regulatory networks," *Journal of Biological Physics and Chemistry*, vol. 4, pp. 85–92, 2004.
33. R. De Smet and Y. Van de Peer, "Redundancy and rewiring of genetic networks following genome-wide duplication events," *Current opinion in plant biology*, vol. 15, no. 2, pp. 168–176, 2012.
34. B. Hutt and K. Warwick, "Synapsing variable-length crossover: Meaningful crossover for variable-length genomes," *Evolutionary Computation, IEEE Transactions on*, vol. 11, no. 1, pp. 118–131, 2007.
35. D. Sprinzak and M. B. Elowitz, "Reconstruction of genetic circuits," *Nature*, vol. 438, no. 7067, pp. 443–448, 2005.
36. M. B. Elowitz and S. Leibler, "A synthetic oscillatory network of transcriptional regulators," *Nature*, vol. 403, no. 6767, pp. 335–338, 2000.
37. N. J. Giordano and H. Nakanishi, *Computational physics*. Pearson Education India, 2006.
38. W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. New York, NY, USA: Cambridge University Press, 3rd ed., 2007.
39. C. M. Guldberg and P. Waage, "Studies concerning affinity," *CM Forhandlinger: Videnskabs-Selskabet i Christiana*, vol. 35, no. 1864, p. 1864, 1864.
40. P. Nelson, *Physical Models of Living Systems*. W. H. Freeman and Company, 2015.
41. A. Van den Bos, *Parameter estimation for scientists and engineers*. John Wiley & Sons, 2007.
42. N. Van Kampen, *Stochastic Processes in Physics and Chemistry*. Elsevier, 2nd ed., 2004.
43. K. F. Riley, M. P. Hobson, and S. J. Bence, *Mathematical methods for physics and engineering: a comprehensive guide*. Cambridge University Press, 2006.
44. C. R. McClung, "Plant circadian rhythms," *The Plant Cell*, vol. 18, no. 4, pp. 792–803, 2006.

45. A. B. Reddy and J. S. O'Neill, "Healthy clocks, healthy body, healthy mind," *Trends in cell biology*, vol. 20, no. 1, pp. 36–44, 2010.
46. L. K. Barger, S. W. Lockley, S. M. Rajaratnam, and C. P. Landrigan, "Neurobehavioral, health, and safety consequences associated with shift work in safety-sensitive professions," *Current neurology and neuroscience reports*, vol. 9, no. 2, pp. 155–164, 2009.
47. A. N. Dodd, N. Salathia, A. Hall, E. Kévei, R. Tóth, F. Nagy, J. M. Hibberd, A. J. Millar, and A. A. Webb, "Plant circadian clocks increase photosynthesis, growth, survival, and competitive advantage," *Science*, vol. 309, no. 5734, pp. 630–633, 2005.
48. P. Green, *Alexander of Macedon, 356-323 BC Berkeley.* University of California Press, 1991.
49. H. G. McWatters and P. F. Devlin, "Timing in plants — a rhythmic arrangement," *FEBS lett*, vol. 585, no. 10, pp. 1474–1484, 2011.
50. M. J. Haydon, L. J. Bell, and A. A. Webb, "Interactions between plant circadian clocks and solute transport," *Journal of experimental botany*, pp. 1–16, 2011.
51. M. J. Haydon, O. Mielczarek, F. C. Robertson, K. E. Hubbard, and A. A. Webb, "Photosynthetic entrainment of the *Arabidopsis thaliana* circadian clock," *Nature*, vol. 502, no. 7473, pp. 689–692, 2013.
52. G. van Ooijen and A. J. Millar, "Non-transcriptional oscillators in circadian timekeeping," *Trends in biochemical sciences*, vol. 37, no. 11, pp. 484–492, 2012.
53. C. R. McClung and R. A. Gutiérrez, "Network news: prime time for systems biology of the plant circadian clock," *Current opinion in genetics & development*, vol. 20, no. 6, pp. 588–598, 2010.
54. S. Fowler, K. Lee, H. Onouchi, A. Samach, K. Richardson, B. Morris, G. Coupland, and J. Putterill, "*GIGANTEA*: a circadian clock-controlled gene that regulates photoperiodic flowering in *Arabidopsis* and encodes a protein with several possible membrane-spanning domains," *EMBO J*, vol. 18, no. 17, pp. 4679–4688, 1999.
55. A. Matsushika, M. Kawamura, Y. Nakamura, T. Kato, M. Murakami, T. Yamashino, and T. Mizuno, "Characterization of circadian-associated pseudo-response regulators: Ii. the function of PRR5 and its molecular dissection in *Arabidopsis thaliana*," *Biosci Biotechnol Biochem*, vol. 71, no. 2, pp. 535–544, 2007.

56. S. Ito, Y. Niwa, N. Nakamichi, H. Kawamura, T. Yamashino, and T. Mizuno, "Insight into missing genetic links between two evening-expressed pseudo-response regulator genes *TOC1* and *PRR5* in the circadian clock-controlled circuitry in *Arabidopsis thaliana*," *Plant Cell Physiol*, vol. 49, no. 2, pp. 201–213, 2008.
57. N. Nakamichi, M. Kita, K. Niinuma, S. Ito, T. Yamashino, T. Mizoguchi, and T. Mizuno, "*Arabidopsis* clock-associated pseudo-response regulators PRR9, PRR7 and PRR5 coordinately and positively regulate flowering time through the canonical CONSTANS-dependent photoperiodic pathway," *Plant and cell physiology*, vol. 48, no. 6, pp. 822–832, 2007.
58. M. E. Eriksson and A. A. Webb, "Plant cell responses to cold are all about timing," *Current Opinion in Plant Biology*, vol. 14, no. 6, pp. 731–737, 2011.
59. B. Novák and J. J. Tyson, "Design principles of biochemical oscillators," *Nature reviews Molecular cell biology*, vol. 9, no. 12, pp. 981–991, 2008.
60. C. H. Johnson, T. Mori, and Y. Xu, "A cyanobacterial circadian clockwork," *Current Biology*, vol. 18, no. 17, pp. R816–R825, 2008.
61. A. S. Loudon, "Circadian biology: a 2.5 billion year old clock," *Current Biology*, vol. 22, no. 14, pp. R570–R571, 2012.
62. R. S. Edgar, E. W. Green, Y. Zhao, G. van Ooijen, M. Olmedo, X. Qin, Y. Xu, M. Pan, U. K. Valekunja, K. A. Feeney, *et al.*, "Peroxiredoxins are conserved markers of circadian rhythms," *Nature*, vol. 485, no. 7399, pp. 459–464, 2012.
63. C. Troein, J. C. Locke, M. S. Turner, and A. J. Millar, "Weather and seasons together demand complex biological clocks," *Current Biology*, vol. 19, no. 22, pp. 1961–1964, 2009.
64. J. S. O'Neill and A. B. Reddy, "Circadian clocks in human red blood cells," *Nature*, vol. 469, no. 7331, pp. 498–503, 2011.
65. J. S. O'Neill, G. Van Ooijen, L. E. Dixon, C. Troein, F. Corellou, F.-Y. Bouget, A. B. Reddy, and A. J. Millar, "Circadian rhythms persist without transcription in a eukaryote," *Nature*, vol. 469, no. 7331, pp. 554–558, 2011.
66. N. Bujdoso and S. J. Davis, "Mathematical modeling of an oscillating gene circuit to unravel the circadian clock network of *Arabidopsis thaliana*," *Frontiers in Plant Science*, vol. 4, 2013.
67. J. Locke, A. Millar, and M. Turner, "Modelling genetic networks with noisy and varied experimental data: the circadian clock in

*Arabidopsis thaliana*," *Journal of theoretical biology*, vol. 234, no. 3, pp. 383–393, 2005.

68. J. C. Locke, M. M. Southern, L. Kozma-Bognár, V. Hibberd, P. E. Brown, M. S. Turner, and A. J. Millar, "Extension of a genetic network model by iterative experimentation and mathematical analysis," *Mol Syst Biol*, vol. 1, no. 1, p. 2005.0013, 2005.
69. R. Schaffer, N. Ramsay, A. Samach, S. Corden, J. Putterill, I. A. Carré, and G. Coupland, "The late elongated hypocotyl mutation of *Arabidopsis* disrupts circadian rhythms and the photoperiodic control of flowering," *Cell*, vol. 93, no. 7, pp. 1219–1229, 1998.
70. Z.-Y. Wang and E. M. Tobin, "Constitutive expression of the *circadian clock associated 1* (*cca1*) gene disrupts circadian rhythms and suppresses its own expression," *Cell*, vol. 93, no. 7, pp. 1207–1218, 1998.
71. D. Alabadı, M. J. Yanovsky, P. Más, S. L. Harmer, and S. A. Kay, "Critical role for CCA1 and LHY in maintaining circadian rhythmicity in *Arabidopsis*," *Curr Biol*, vol. 12, no. 9, pp. 757–761, 2002.
72. E. Yakir, D. Hilman, I. Kron, M. Hassidim, N. Melamed-Book, and R. M. Green, "Posttranslational regulation of CIRCADIAN CLOCK ASSOCIATED 1 in the circadian oscillator of *Arabidopsis*," *Plant Physiol*, vol. 150, no. 2, pp. 844–857, 2009.
73. T. Mizoguchi, K. Wheatley, Y. Hanzawa, L. Wright, M. Mizoguchi, H.-R. Song, I. A. Carré, and G. Coupland, "*LHY* and *CCA1* are partially redundant genes required to maintain circadian rhythms in *Arabidopsis*," *Dev Cell*, vol. 2, no. 5, pp. 629–641, 2002.
74. S. X. Lu, S. M. Knowles, C. Andronis, M. S. Ong, and E. M. Tobin, "CIRCADIAN CLOCK ASSOCIATED 1 and LATE ELONGATED HYPOCOTYL function synergistically in the circadian clock of *Arabidopsis*," *Plant Physiol*, vol. 150, no. 2, pp. 834–843, 2009.
75. R. Green and E. Tobin, "Loss of the circadian clock-associated protein 1 in *Arabidopsis* results in altered clock-regulated gene expression," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 96, no. 7, pp. 4176–4179, 1999.
76. A. Pokhilko, S. K. Hodge, K. Stratford, K. Knox, K. D. Edwards, A. W. Thomson, T. Mizuno, and A. J. Millar, "Data assimilation constrains new connections and components in a complex, eukaryotic circadian clock model," *Mol Syst Biol*, vol. 6, no. 1, p. 416, 2010.

77. J. M. Gendron, J. L. Pruneda-Paz, C. J. Doherty, A. M. Gross, S. E. Kang, and S. A. Kay, "*Arabidopsis* circadian clock protein, TOC1, is a DNA-binding transcription factor," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 109, no. 8, pp. 3167–3172, 2012.
78. W. Huang, P. Pérez-García, A. Pokhilko, A. Millar, I. Antoshechkin, J. Riechmann, and P. Mas, "Mapping the core of the *Arabidopsis* circadian clock defines the network structure of the oscillator," *Science*, vol. 336, no. 6077, pp. 75–79, 2012.
79. A. Matsushika, S. Makino, M. Kojima, and T. Mizuno, "Circadian waves of expression of the APRR1/TOC1 family of pseudo-response regulators in *Arabidopsis thaliana*: insight into the plant circadian clock," *Plant Cell Physiol*, vol. 41, no. 9, pp. 1002–1012, 2000.
80. N. Nakamichi, T. Kiba, R. Henriques, T. Mizuno, N.-H. Chua, and H. Sakakibara, "PSEUDO-RESPONSE REGULATORS 9, 7, and 5 are transcriptional repressors in the *Arabidopsis* circadian clock," *Plant Cell*, vol. 22, no. 3, pp. 594–605, 2010.
81. E. M. Farré, S. L. Harmer, F. G. Harmon, M. J. Yanovsky, and S. A. Kay, "Overlapping and distinct roles of *PRR7* and *PRR9* in the *Arabidopsis* circadian clock," *Curr Biol*, vol. 15, no. 1, pp. 47–54, 2005.
82. N. Nakamichi, M. Kita, S. Ito, T. Yamashino, and T. Mizuno, "PSEUDO-RESPONSE REGULATORS, PRR9, PRR7 and PRR5, together play essential roles close to the circadian clock of *Arabidopsis thaliana*," *Plant Cell Physiol*, vol. 46, no. 5, pp. 686–698, 2005.
83. Z. Ding, M. R. Doyle, R. M. Amasino, and S. J. Davis, "A complex genetic interaction between *Arabidopsis thaliana* TOC1 and CCA1/LHY in driving the circadian clock and in output regulation," *Genetics*, vol. 176, no. 3, pp. 1501–1510, 2007.
84. H. G. McWatters, E. Kolmos, A. Hall, M. R. Doyle, R. M. Amasino, P. Gyula, F. Nagy, A. J. Millar, and S. J. Davis, "*ELF4* is required for oscillatory properties of the circadian clock," *Plant Physiol*, vol. 144, no. 1, pp. 391–401, 2007.
85. E. Kolmos, M. Nowak, M. Werner, K. Fischer, G. Schwarz, S. Mathews, H. Schoof, F. Nagy, J. M. Bujnicki, and S. J. Davis, "Integrating *ELF4* into the circadian system through combined structural and functional studies," *HFSP J*, vol. 3, no. 5, pp. 350–366, 2009.
86. D. A. Nusinow, A. Helfer, E. E. Hamilton, J. J. King, T. Imaizumi, T. F. Schultz, E. M. Farré, and S. A. Kay, "The ELF4-ELF3-LUX complex links the circadian clock to diurnal control of hypocotyl

growth," *Nature*, vol. 475, no. 7356, pp. 398–402, 2011.

87. E. Herrero, E. Kolmos, N. Bujdoso, Y. Yuan, M. Wang, M. C. Berns, H. Uhlworm, G. Coupland, R. Saini, M. Jaskolski, *et al.*, "EARLY FLOWERING4 recruitment of EARLY FLOWERING3 in the nucleus sustains the *Arabidopsis* circadian clock," *The Plant Cell*, vol. 24, no. 2, pp. 428–443, 2012.
88. L. E. Dixon, K. Knox, L. Kozma-Bognar, M. M. Southern, A. Pokhilko, and A. J. Millar, "Temporal repression of core circadian genes is mediated through EARLY FLOWERING 3 in *Arabidopsis*," *Curr Biol*, vol. 21, no. 2, pp. 120–125, 2011.
89. A. Helfer, D. A. Nusinow, B. Y. Chow, A. R. Gehrke, M. L. Bulyk, and S. A. Kay, "*LUX ARRHYTHMO* encodes a nighttime repressor of circadian gene expression in the *Arabidopsis* core clock," *Curr Biol*, vol. 21, no. 2, pp. 126–133, 2011.
90. B. Y. Chow, A. Helfer, D. A. Nusinow, and S. A. Kay, "ELF3 recruitment to the *PRR9* promoter requires other Evening Complex members in the *Arabidopsis* circadian clock," *Plant Signal Behav*, vol. 7, no. 2, pp. 170–173, 2012.
91. S. Dai, X. Wei, L. Pei, R. L. Thompson, Y. Liu, J. E. Heard, T. G. Ruff, and R. N. Beachy, "BROTHER OF LUX ARRHYTHMO is a component of the *Arabidopsis* circadian clock," *Plant Cell*, vol. 23, no. 3, pp. 961–972, 2011.
92. S. P. Hazen, T. F. Schultz, J. L. Pruneda-Paz, J. O. Borevitz, J. R. Ecker, and S. A. Kay, "*LUX ARRHYTHMO* encodes a MYB domain protein essential for circadian rhythms," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 102, no. 29, pp. 10387–10392, 2005.
93. S. M. Knowles, S. X. Lu, and E. M. Tobin, "Testing time: can ethanol-induced pulses of proposed oscillator components phase shift rhythms in *Arabidopsis*?," *J Biol Rhythms*, vol. 23, no. 6, pp. 463–S, 2008.
94. J. L. Pruneda-Paz, G. Breton, A. Para, and S. A. Kay, "A functional genomics approach reveals CHE as a component of the *Arabidopsis* circadian clock," *Science*, vol. 323, no. 5920, pp. 1481–1485, 2009.
95. M. Johansson, H. G. McWatters, L. Bakó, N. Takata, P. Gyula, A. Hall, D. E. Somers, A. J. Millar, and M. E. Eriksson, "Partners in time: EARLY BIRD associates with ZEITLUPE and regulates the speed of the *Arabidopsis* clock," *Plant Physiol*, vol. 155, no. 4, pp. 2108–2122, 2011.

96. W.-Y. Kim, S. Fujiwara, S.-S. Suh, J. Kim, Y. Kim, L. Han, K. David, J. Putterill, H. G. Nam, and D. E. Somers, "ZEITLUPE is a circadian photoreceptor stabilized by GIGANTEA in blue light," *Nature*, vol. 449, no. 7160, pp. 356–360, 2007.
97. P. Más, W.-Y. Kim, D. E. Somers, and S. A. Kay, "Targeted degradation of TOC1 by ZTL modulates circadian function in *Arabidopsis thaliana*," *Nature*, vol. 426, no. 6966, pp. 567–570, 2003.
98. A. Baudry, S. Ito, Y. H. Song, A. A. Strait, T. Kiba, S. Lu, R. Henriques, J. L. Pruneda-Paz, N.-H. Chua, E. M. Tobin, S. A. Kay, and T. Imaizumi, "F-box proteins FKF1 and LKP2 act in concert with ZEITLUPE to control *Arabidopsis* clock progression," *Plant Cell*, vol. 22, no. 3, pp. 606–622, 2010.
99. S. Fujiwara, L. Wang, L. Han, S.-S. Suh, P. A. Salomé, C. R. McClung, and D. E. Somers, "Post-translational regulation of the *Arabidopsis* circadian clock through selective proteolysis and phosphorylation of pseudo-response regulator proteins," *J Biol Chem*, vol. 283, no. 34, pp. 23073–23083, 2008.
100. T. Kiba, R. Henriques, H. Sakakibara, and N.-H. Chua, "Targeted degradation of PSEUDO-RESPONSE REGULATOR 5 by an $SCF^{ZTL}$ complex regulates clock function and photomorphogenesis in *Arabidopsis thaliana*," *Plant Cell*, vol. 19, no. 8, pp. 2516–2530, 2007.
101. J.-W. Yu, V. Rubio, N.-Y. Lee, S. Bai, S.-Y. Lee, S.-S. Kim, L. Liu, Y. Zhang, M. L. Irigoyen, J. A. Sullivan, Y. Zhang, I. Lee, Q. Xie, N.-C. Paekemail, and X. W. Deng, "COP1 and ELF3 control circadian function and photoperiodic flowering by regulating GI stability," *Mol Cell*, vol. 32, no. 5, pp. 617–630, 2008.
102. E. M. Farré and S. A. Kay, "PRR7 protein levels are regulated by light and the circadian clock in *Arabidopsis*," *Plant J*, vol. 52, no. 3, pp. 548–560, 2007.
103. L. Wang, S. Fujiwara, and D. E. Somers, "PRR5 regulates phosphorylation, nuclear import and subnuclear localization of TOC1 in the *Arabidopsis* circadian clock," *EMBO J*, vol. 29, no. 11, pp. 1903–1915, 2010.

Von allem Geschriebenen liebe ich nur Das, was Einer mit seinem Blute schreibt. Schreibe mit Blut: und du wirst erfahren, dass Blut Geist ist.
Freidrich Nietzsche, *Also sprach Zarathustra* (1891)

# *Summary of Publications*

The articles that follow are here presented in the context of this introduction. The articles are independent of each other, but can be divided into three fields: functional fitting to correlated data (paper I); a model of the circadian clock in the plant *Arabidopsis thaliana* (paper III), and transcriptional networks, represented as strings of bits (paper II and IV).

## 2.1 ON MODEL FITTING TO CORRELATED DATA

Despite the many years that have come to pass since the discovery and explanation of Brownian motion, it still remains an active area of both experimental and theoretical research. The advent of super-resolution microscopy, capable of resolving individual particles of the cell, with unprecedented quality [1, 2], has a great potential for increasing our understanding of biological processes, e.g. following a single mRNA from transcription to translation to a protein is almost within our reach [3]. In particle tracking experiments, one typically takes the squared displacement of the fluorescently tagged particle over time and averages over many trajectories, to get the mean square displacement (MSD) as a function of time. One then extracts model parameters such as diffusion constants, by functional fitting using some standard method like least squares (LS) which minimizes the residuals.

However, in this setting, the error estimation of the fitted parameters of the LS method will generally be orders of magnitude too optimistic, as the LS method is not valid when applied to correlated data, like trajectory data. The correlation is apparent when considering two neighboring sampling points for an individual trajectory. If the displacement is

larger than the mean at that point, it is likely to still be for the next point. Thus more frequent measurements do not necessarily increase the accuracy of the parameter estimation as much as the LS method lets on. A maximum likelihood method (ML) does little to alleviate the problems of LS fitting, as it is associated with numerical instability when inverting the covariance matrix of the observable. In addition, the parameter estimate of the ML method is also subject to a strong bias in the parameter estimation itself. In paper I we highlight this problem, that seems to have gone largely unnoticed in the particle tracking community. We provide a new correlation corrected error estimation formula for the otherwise robust LS method, making it valid also for nonlinear models. We demonstrate the improvement of the new method on three prototypical systems: one linear system describing ordinary Brownian motion, and two nonlinear subdiffusive systems with weaker time dependence than Brownian motion [3]. We also derive an expression for the bias of the ML method, valid to first order, and evaluate both first and second order jackknife bias reduction procedures applied to ML fitted parameters.

Furthermore, we introduce a Brownian motion adapted LS method, which uses the exact covariance matrix for Brownian motion as basis for its weighting matrix for the LS method. We find that the variance of the estimated parameters is smaller than what was found for the correlation corrected LS method, but at the cost of increased bias of the parameter estimation itself.

*Contribution*

M.A.L. and T.A. conceived the idea of the project. All authors contributed to the conceptual design of the CLS method. I wrote all software and performed all simulations, under supervision by T.A. I also prepared all figures. I wrote the manuscript together with T.A., with input from A.I. and M.A.L. The new error estimation formula (with and without jackknife) was derived by T.A, and M.A.L. derived the bias correction prediction for Brownian motion with input from me and T.A. A.I. suggested the use of jackknife for ML fitting. T.A. coordinated the project.

## 2.2 ON WHAT SHAPES TRANSCRIPTIONAL NETWORKS

In paper II we set out to further our understanding of what shapes the structure of transcriptional networks. As previously touched upon, in section 1.3.1, it is currently unclear what underlying mechanisms give rise to the many structural similarities of gene regulatory networks. It can be argued that the similarities are a result of networks being exposed to similar mutations, or alternatively, that network function requires them to have certain structural properties. Selection and large-scale gene duplication events [4] can explain the shared properties of gene regulatory networks [5, 6]. In order to explore how mutation and selection together shape networks, we develop a model of transcriptional networks that we can subject to evolution, either neutral or towards some function. The evolution can be restricted to just point mutations and crossover, or also encompass gene duplication.

In greater detail, we represent gene regulatory regions and TFs as sequences of ones and zeros, 256 or 32 bits, respectively. The binding of TFs to DNA is determined by the number of mismatching bits between their sequences, and the regulatory action of the TFs depends on their position on the DNA relative to the transcriptional start site (TSS). Half of the possible TF binding site positions are downstream of the TSS and will block RNAP from binding to the DNA, effectively disallowing any expression of the gene. Any TF binding upstream of the TSS will act as an activator. The network is built up of genes (nodes) producing TFs, which bind to other genes to regulating them (edges). By the binding of multiple TF species to a regulatory region, complex logic combinatorics arise from cooperative and exclusive interactions. The model allows a variable number of genes.

The total transcription rate of a gene depends on the probability for RNAP to bind and initiate transcription. This is computed from the distribution of statistical weights for all possible binding states. This representation of gene interactions is then used to evolve networks with one of two possible functions. Either solve a majority decision task, where the network must determine the state of the majority of the seven binary input nodes, or act as an internal clock by using periodic input to generate a timely gene expression. Networks are also allowed to evolve neutrally, constrained to have the same structure (number of nodes, edges and degree distribution) as their evolved functional counterpart.

We noted differences between networks depending on their function. Networks performing the clock function were strongly biased towards

negative edges and strong cooperativity among the TFs. This is expected, as the clock needs negative feedback and nonlinearity for robust oscillations [7]. The majority decision system favoured positive regulation and AND logic in the interactions of binding sites. For TFs with two binding sites in the same regulatory region, the number that had ambiguous regulation (one repressing and one activating) behaved like expected for a random process in the neutrally evolved networks. However, in both our functional networks, and in data from *E. coli*, such ambiguity was reduced. This result holds regardless of whether we allow gene duplication or not.

When looking at the sign of each TF's regulatory action in the network as a whole, we found that both neutral and evolved networks follow the random expectation in the absence of gene duplication as an evolutionary step. However, when allowing gene duplication, TFs in both neutral and functional networks evolved to specialize to act predominately as either global repressors or activators. The main observed difference between the two different types of functional networks lies in their Boolean logic rules governing the gene regulation. The majority decision networks were rich in AND gates while the clock had comparatively many NOR gates. Furthermore, the networks differed in their distribution of number of inputs to the logic rules, as well as their typical structure.

*Contribution*

The model was conceived and developed in collaboration with C.T. and C.P. The software was developed in close collaboration with C.T., with whom I also co-wrote the manuscript. I also contributed to making plots and computer code for data analysis. Experiments and data analysis were done together with C.T.

## 2.3 ON TRANSCRIPTIONAL ACTIVATION IN THE CIRCADIAN CLOCK

In paper III we set out to model the circadian clock network of *Arabidopsis thaliana*. We used a system of ODEs that describe the transcription and translation of the genes. Our starting point was an earlier model by Pokhilko *et al.* [8], which we made heavy modifications to. For instance, we assumed most regulatory interaction to be mostly repressing [7], much like our example system in section 1.4.2 or what was found for our clock network in paper II. We also abolished the sequential activation

for generating the *PRR* wave, and instead modelled it as each component turning off its predecessor. Furthermore, we added two newly discovered clock genes, the night expressed *NOX* [9] and the morning expressed *REVEILLE 8* [10, 11]. The latter acts as the sole activator within in our clock network.

For our modelling procedure we developed a data driven approach. This meant culling time course measurement data from published experiments, resulting in over 11,000 extracted data points from 800 time courses in 150 different mutants and light conditions. Our model uses simulated annealing to minimize a cost function that fits both profile shape and level of the simulated expression of all variables to all data in all conditions simultaneously.

*Contribution*
I compiled all experimental time course data used in the fitting, by extracting 11,000 data points, by hand, from published articles. I went through the corpus of published experimental findings in the field of *Arabidopsis*. C.T. designed the software, but I made contributions, such as code for generating plots, and model optimization. I performed the simulations. I co-wrote the article with C.T., and prepared the figures.

## 2.4 ON ALGORITHMS FOR AN EFFICIENT CROSSOVER

To investigate mechanisms of evolution, we need a representation of the genome for it to act on. Therefore, we implement a model with a variable-length linear genome, that will allow relevant operations such as mutations and gene duplications. In our model, the genome is able to get longer, by insertion of duplicated sequences, or shorter, by deletion. This enables better exploration of evolutionary space by providing ample room for neutral evolution on the genome. However, using a variable-length genome makes meaningful crossover operations challenging. A viable offspring needs a complete set of the genes shared between its parents, and a combination of the features that are unique to either one. We solve this by aligning the parental genomes to identify the homologous regions, and use these shared sequences as potential crossover points.

The alignment can be made using a global alignment method, such as the Hirschberg algorithm [12], but this is computationally demanding. Another method exists for performing crossover operations: by aligning

the longest identical sequences ("synapses"), the regions in between can be exchanged [13]; however, the method assumes high sequence similarity, which might not be fulfilled in evolutionary simulations. We compared these two methods, together with our own heuristic alignment method. The methods were assessed through three different measures: CPU time consumption, the ability for the crossover algorithm to align homologous sequences, and the performance of the offspring in a simple evolutionary setting.

In more detail, our model represents the genome as a single string of bits. In the evolutionary simulations, a gene is identified by a start sequence, which is an arbitrary predetermined six bit pattern, and the following three groups of ten bits are read as integers, giving the height, width, and position of triangles whose area should sum up to approximate a sinusoidal function, which is how we map genotype to phenotype.

We find that our heuristic method aligns sequences as well as the theoretically optimal Hirschberg algorithm, as long as the parental sequences are not extremely divergent. The CPU time consumption scales more favourably for our heuristic algorithm as the genome length grows, than it does for the Hirschberg method. For low sequence divergence, the heuristic algorithm is approximately twice as fast as the synapsing method. We find that with crossover operations, the fitness increases faster with fewer generations, than it does without crossovers. Thus crossover operations are especially beneficial when evaluating time consuming fitness functions, resulting in an overall lower computational cost.

*Contribution*

I developed the model for encoding the network as a single bitstring together with C.T., and collaborated on implementing the synapsing algorithm with A.M., H.Å. and C.T. I prepared the figures, took part in discussions on sequence alignment, and contributed to the manuscript together with the co-authors. C.T. ran all simulations and generated the data.

REFERENCES

1. K. R. Chi, "Super-resolution microscopy: breaking the limits," *Nature Methods*, vol. 6, no. 1, pp. 15–18, 2009.

2. M. J. Saxton, "Single-particle tracking: connecting the dots," *Nature Methods*, vol. 5, no. 8, pp. 671–672, 2008.
3. E. Barkai, Y. Garini, and R. Metzler, "Strange kinetics of single molecules in living cells," *Phys. Today*, vol. 65, no. 8, p. 29, 2012.
4. P. D. Kuo, W. Banzhaf, and A. Leier, "Network topology and the evolution of dynamics in an artificial genetic regulatory network model created by whole genome duplication and divergence," *Biosystems*, vol. 85, no. 3, pp. 177–200, 2006.
5. R. De Smet and Y. Van de Peer, "Redundancy and rewiring of genetic networks following genome-wide duplication events," *Current opinion in plant biology*, vol. 15, no. 2, pp. 168–176, 2012.
6. T. R. Sorrells and A. D. Johnson, "Making sense of transcription networks," *Cell*, vol. 161, no. 4, pp. 714–723, 2015.
7. B. Novák and J. J. Tyson, "Design principles of biochemical oscillators," *Nature reviews Molecular cell biology*, vol. 9, no. 12, pp. 981–991, 2008.
8. A. Pokhilko, A. P. Fernández, K. D. Edwards, M. M. Southern, K. J. Halliday, and A. J. Millar, "The clock gene circuit in *Arabidopsis* includes a repressilator with additional feedback loops," *Mol Syst Biol*, vol. 8, p. 574, 2012.
9. S. Dai, X. Wei, L. Pei, R. L. Thompson, Y. Liu, J. E. Heard, T. G. Ruff, and R. N. Beachy, "BROTHER OF LUX ARRHYTHMO is a component of the *Arabidopsis* circadian clock," *Plant Cell*, vol. 23, no. 3, pp. 961–972, 2011.
10. R. Rawat, N. Takahashi, P. Y. Hsu, M. A. Jones, J. Schwartz, M. R. Salemi, B. S. Phinney, and S. L. Harmer, "REVEILLE 8 and PSEUDO-REPONSE REGULATOR 5 form a negative feedback loop within the *Arabidopsis* circadian clock," *PLoS Genet*, vol. 7, no. 3, p. e1001350, 2011.
11. P. Y. Hsu, U. K. Devisetty, and S. L. Harmer, "Accurate timekeeping is controlled by a cycling activator in *Arabidopsis*," *eLife*, vol. 2, p. e00473, 2013.
12. D. S. Hirschberg, "A linear space algorithm for computing maximal common subsequences," *Commun. ACM*, vol. 18, pp. 341–343, June 1975.
13. B. Hutt and K. Warwick, "Synapsing variable-length crossover: Meaningful crossover for variable-length genomes," *Evolutionary Computation, IEEE Transactions on*, vol. 11, no. 1, pp. 118–131, 2007.

In search of Truth the hopeful zealot goes,
but all the sadder turns, the more he knows
H.P. Lovecraft

# *Appendices*

Herein, we collect information deemed outside the scope of the main text, as we do not want to risk leading the reader astray.

## 3.A EXCERPT FROM "ON THE NATURE OF THINGS"

It has been argued by many that things were better in the days of yore. Indeed, gone are the days when science was written on verse, as was done by Roman natural philosopher Titus Lucretius Carus, c. 99 – 55 BC [1].

In his poem, *De rerum natura*, divided into six books, he describes the principles of atomism. He strives to explain the world through natural laws rather than the will of gods. In the second book, he describes how dust particles, dancing in the sunlight, are the result of collisions of many small atoms having an impact on an hierarchy of larger particles, finally resulting in the movements of objects large enough for our perception [2].

The following is an excerpt, as translated by William Ellery Leonard (1876–1944), from On the nature of things:

> For us thin air and splendour-lights of the sun.
> And many besides wander the mighty void–
> Cast back from unions of existing things,
> Nowhere accepted in the universe,
> And nowise linked in motions to the rest.
> And of this fact (as I record it here)
> An image, a type goes on before our eyes
> Present each moment; for behold whenever
> The sun's light and the rays, let in, pour down

Across dark halls of houses: thou wilt see
The many mites in many a manner mixed
Amid a void in the very light of the rays,
And battling on, as in eternal strife,
And in battalions contending without halt,
In meetings, partings, harried up and down.
From this thou mayest conjecture of what sort
The ceaseless tossing of primordial seeds
Amid the mightier void–at least so far
As small affair can for a vaster serve,
And by example put thee on the spoor
Of knowledge. For this reason too 'tis fit
Thou turn thy mind the more unto these bodies
Which here are witnessed tumbling in the light:
Namely, because such tumblings are a sign
That motions also of the primal stuff
Secret and viewless lurk beneath, behind.
For thou wilt mark here many a speck, impelled
By viewless blows, to change its little course,
And beaten backwards to return again,
Hither and thither in all directions round.
Lo, all their shifting movement is of old,
From the primeval atoms; for the same
Primordial seeds of things first move of self,
And then those bodies built of unions small
And nearest, as it were, unto the powers
Of the primeval atoms, are stirred up
By impulse of those atoms' unseen blows,
And these thereafter goad the next in size:
Thus motion ascends from the primevals on,
And stage by stage emerges to our sense,
Until those objects also move which we
Can mark in sunbeams, though it not appears
What blows do urge them.

## 3.B ON THE REPRESSILATOR

The parameter values used for generating our three-component repressilator.

| Parameter | Value | Parameter | Value |
|---|---|---|---|
| $k_1$ | 5.50 | $d_1$ | 2.23 |
| $k_2$ | 0.36 | $d_2$ | 2.32 |
| $k_3$ | 15.47 | $d_3$ | 1.00 |
| $K_1$ | 0.11 | $n_1$ | 3.46 |
| $K_2$ | 0.38 | $n_2$ | 3.84 |
| $K_3$ | 0.0027 | $n_3$ | 3.79 |

**Table 3.1 Parameter values.** The parameter set used for bringing the three component network described in section 1.4.2 to a limit cycle.

### REFERENCES

1. P. Collinder, *Nordisk familjebok, encyklopedi och konversationslexikon*, vol. 14. Förlagshuset norden AB Malmö, 4 ed., 1953.
2. Lucretius, *On The Nature of Things*, vol. 785 of *Project Gutenberg*. P.O. Box 2782, Champaign, IL 61825-2782, USA: Project Gutenberg, 1997.