LUND UNIVERSITY

On gene regulatory networks and data fitting

Fogelmark, Karl

2016

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 μm 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.

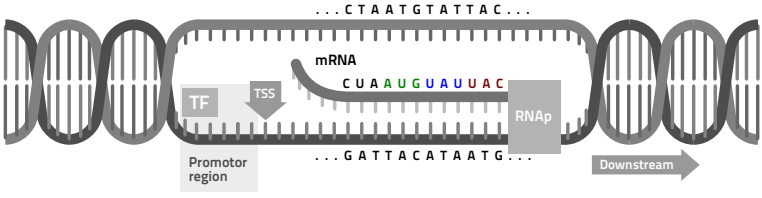

Figure 1.1 Transcription process. Transcription is initiated by transcription factors (TFs) binding to the promotor region, which recruits RNAP binding. As RNAP starts sliding downstream, from the transcriptional start site (TSS), along the uncoiled and opened double helix, it will assemble an mRNA molecule 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?

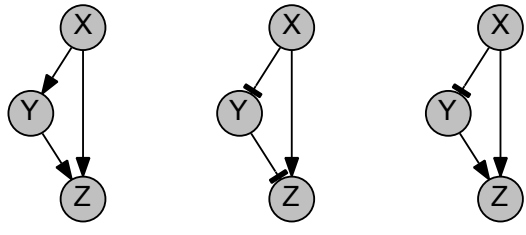

Figure 1.2 Three node network motifs. The first two graphs are coherent feed forward loop (FFL) network motifs, where the direct path from X regulating the target node Z has the same net effect on the target as the indirect path through the intermediary node Y. The rightmost motif is said to be an incoherent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

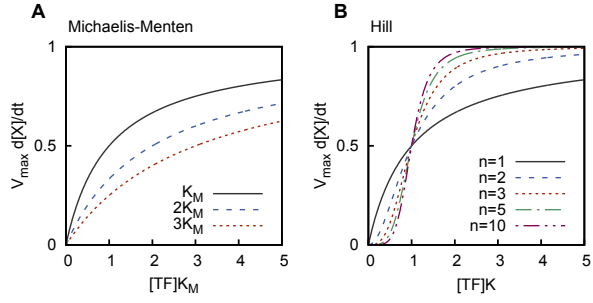

Figure 1.3 . The resulting modelled production of protein X as function of concentration of TF. (A) Michaelis-Menten kinetics, eq. (1.7), and (B) Hill equation, (1.9) for different degrees of cooperativity, n. The production rate saturates at $V_{\max}$.

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

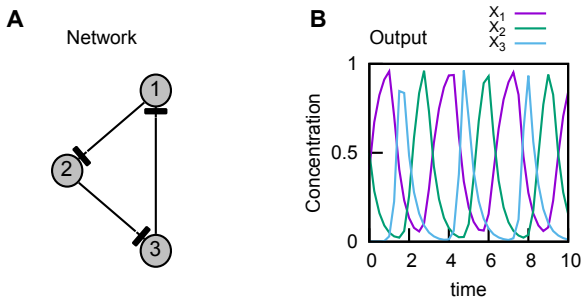

Figure 1.4 A three node network. (A) The network is connected in a loop, where each edge represses the next. (B) The output from each node, normalized to unity, oscillates with time, for suitable parameters chosen in eq. (1.11).

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 (~ 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.

PAPER I

Model parameter estimation in particle tracking

Karl Fogelmark[1], Michael A. Lomholt[2], Anders Irbäck[1] and Tobias Ambjörnsson[1]

[1] Computational Biology and Biological Physics, Department of Astronomy and Theoretical Physics, Lund University, 223 62 Lund, Sweden
[2] Department of Physics, Chemistry and Pharmacy, University of Southern Denmark, Campusvej 55, 5230 Odense M, Denmark.

Submitted, LU-TP 16-18 (2016)

Experimental super-resolution methods allow tracking of particles with an unprecedented spatial resolution. A crucial down stream objective when interpreting tracking experiments is to fit averages (typically, squared displacements at different times) with a model and extract parameters, such as diffusion constants. A commonly overlooked challenge in such fitting procedures is that fluctuations around mean values almost always exhibit temporal correlations. We show here that current methods, maximum likelihood and least squares fitting, fail at either robust parameter estimation or accurate error estimation. We remedy this deficiency by deriving a new error estimation formula for least square fitting. The new formula uses the full covariance matrix, i.e., rigorously includes correlations, but is free of the robustness issues, inherent to the maximum likelihood method. We demonstrate its accuracy in three prototypical examples of importance in cell biology: Brownian motion, fractional Brownian motion and continuous time random walks. Our correlated-corrected least squares method is general in character and will therefore be of use in

other fields where fitting to ensemble data is common, such as physics, astronomy, and finance. Our closed-form error estimation formula is well suited for standard curve fitting software packages.

I.1 INTRODUCTION

The last decade has witnessed a revolution in the ability of probing biological structures in the nanometer-range, via super-resolution fluorescence microscopy techniques, such as STED, PALM, STORM and high-accuracy localization techniques like FIONA [1–6]. These techniques have been predicted to profoundly change our understanding of the working of biological cells at a truly nanoscopic level. Super-resolution microscopy was "Method of the Year" in Nature Methods 2009 [7] and was awarded the Nobel Prize in chemistry in 2014. The method, applied on particle tracking, entails the following steps: [8] (i) label the fluorescent molecules, (ii) localize the associated "dots" in movies, (iii) connect the dots and (iv) interpret the resulting trajectories. Step (i) involves challenges such as increasing biocompatibility, brightness and photostability of fluorophores [9]. Steps (ii)-(iii) concern several theoretical and computational challenges, which have attracted considerable attention over the past ten years [10–12]. Step (iv) requires the interpretation of ensemble averages over the measured trajectories and often requires fitting the measured mean square displacements (MSD) at different sampling times to some standard model [8]. This final and crucial step is here revisited.

Fitting a model function to data is done so readily in the field of science that one seldom considers the correctness of the standard go-to solution of the least squares (LS) method (χ^2 minimization) [13]. For applications to the present problem, one of the crucial assumptions of the LS method is that the fluctuations around mean values (in tracking experiments, often the mean square displacements) are independent quantities. However, since in particle tracking experiments, the data is sampled along trajectories, this assumption is in general *never* satisfied when analyzing ensemble averages based on a set of trajectories; heuristically, if in one trajectory an observable, such as the square displacement, was smaller than its ensemble averaged value at some time, it is typically still so at the next time step (see Supplementary Figure I.S1 for an illustration). Thus, the fluctuations around the estimated MSD, or any other ensemble averaged observable, exhibit temporal correlations.

The question now arises of how severe the consequences of neglecting the temporal correlations in LS fitting are. We demonstrate that it can lead to underestimated errors for parameters (such as diffusion constants) by more than one order of magnitude for our prototype systems (see below), which can have detrimental effects when interpreting the data. To our knowledge, the only previous method for dealing fully with correlation in data is the *correlated* χ^2 minimization method [14] (maximum likelihood, ML [15]). This method is known to the lattice QCD community, but does not seem to have found wide spread use. This could partly be due to that, while statistically sound, robustness issues have been identified [16, 17]. Herein, we carefully examine the ML method and demonstrate that it only provides correct parameter estimation in a small region of the "phase space" (N, M), where N is the number of sampling times and M is the number of trajectories. Thus, it appears that the ML and LS methods are of limited general purpose use for interpretation of ensemble averages based on particle trajectories.

Here, we remedy the lack of available tools for accurate parameter estimation based on sampled trajectories. We derive a new error estimation formula for standard LS fitted parameters which takes into account the temporal correlations, intrinsic to ensemble averages based on trajectories. We compare our method, referred to as the correlation-corrected least squares (CLS) method, to both the LS method and to the ML approach. The new method has the following desirable unique features: (1) robust estimation of parameters in the full phase space (N, M); (2) estimated mean parameter values are in agreement with theory for our prototype systems; (3) the error estimation formula accurately reproduces the spread of the actual parameter values. Brownian motion (BM), fractional Brownian motion (FBM) and continuous time random walks (CTRW) are here used as prototype models. These have been identified as three important model systems for motion of fluorescently labeled particles and molecules in cells [18]. Additionally, these three model systems provide ideal test beds for our method, as they are accessible by analytical means; in particular, the MSD is known for all three cases. Like in previous tracking method evaluation studies [19] (where different methods for steps (ii) and (iii) were benchmarked) we use simulations for validation purposes.

I.2 METHODS

In what follows, we provide a ready-to-use method, which is further motivated and detailed in the Supplementary. In particle tracking experiments one records a set of trajectories, here enumerated by m. The task at hand is to fit some functional form $f(t_i;\boldsymbol{\lambda}) = f_i(\boldsymbol{\lambda})$, with K free fitting parameters $\boldsymbol{\lambda} = \lambda_1, \ldots, \lambda_K$ to some ensemble average $\overline{y}(t_i) = \overline{y}_i$ over the trajectories, i.e., to an unbiased sample mean of the form

$$\overline{y}_i = \frac{1}{M}\sum_{m=1}^{M} y_i^{(m)} \tag{I.1}$$

where the index i is over the N sampling times $\boldsymbol{T} = T_1, \ldots, T_N$ (with $N \geq K$). In all applications (see Results) we use squared displacements, i.e., $y_i^{(m)} = |\boldsymbol{x}^{(m)}(T_i) - \boldsymbol{x}^{(m)}(0)|^2$, where $\boldsymbol{x}^{(m)}(t)$ is the particle position (a vector with d elements, where d is the number of spatial dimensions) at time t for trajectory m, and the start time for the simulation/experiment is $t = 0$. We point out, however, that in the CLS method the quantity $y_i^{(m)}$ can be any observable for trajectory m at sampling time T_i. We shall consistently use single bar to denote a sample estimator and double bar for exact expectation value. The challenge of model parameter estimation [15] is to "fit" some function $f_i(\boldsymbol{\lambda})$ to data $\overline{y}_i$ and thereby extract the model parameters, $\boldsymbol{\lambda}$. This problem has previously typically been tackled using the LS or ML methods (reviewed in section I.D). We do not concern ourselves here with the model selection problem [20], i.e., how to choose the functional form for $f_i(\boldsymbol{\lambda})$.

Our approach, the CLS method, extends the standard LS procedure with a correct error estimation formula valid also for correlated data. For completeness and ease of application, we here provide the full details of the proposed CLS fitting procedure. We start by introducing a cost function, χ^2, based on the the difference between the sample average and the fitting function $\Lambda_i = \overline{y}_i - f_i(\boldsymbol{\lambda})$ for all time points, according to

$$\chi^2 = \boldsymbol{\Lambda}^T \boldsymbol{R}\, \boldsymbol{\Lambda}, \tag{I.2}$$

here in matrix (bold) notation. This cost function is to be minimized with respect to $\boldsymbol{\lambda}$ in order to determine the best parameter values, λ_a^* ($a = 1, \ldots, K$) [21]. The matrix $\boldsymbol{R}$ is any symmetric positive definite matrix. In the CLS method we use $R_{ij} = \overline{R}_{ij} = \delta_{ij}/\overline{C}_{ij}$ as in the stan-

dard LS approach, where δ_{ij} is the Kronecker delta, and the (unbiased) sample "covariance matrix of the mean" is defined as $\overline{C}_{ij} = \overline{Q}_{ij}/M$, with $\overline{\boldsymbol{Q}}$ as the sample covariance matrix

$$\overline{Q}_{ij} = \frac{1}{M-1}\sum_{m=1}^{M}(y_i^{(m)} - \overline{y}_i)(y_j^{(m)} - \overline{y}_j). \tag{I.3}$$

We note here that, while this specific choice of $\boldsymbol{R}$ is used in our applications, the results in this section, including the new error formula below, is valid for arbitrary choices of $\boldsymbol{R}$. In section I.E we elaborate on one "non-conventional" choice of $\boldsymbol{R}$ particularly adapted for BM, namely using the exact covariance matrix for BM.

The parameters, λ_a^*, obtained by minimizing the cost function χ^2 will have a (co)variance $\Delta_{ab} = \langle(\lambda_a^* - \overline{\overline{\lambda}}_a)(\lambda_b^* - \overline{\overline{\lambda}}_b)\rangle$, where $\langle\ldots\rangle$ denotes ensemble average. The variance of the fitted parameter is $\sigma_a^2 = \Delta_{aa}$. As noted in the Introduction, this covariance depends on the temporal correlations. For a stationary process, it is well-known how to estimate the variance of a mean in the presence of temporal correlations, typically by expressing the variance in terms of the sum or integral of the auto correlation function [22, 23]. In the present context, such an estimation corresponds to fitting a "horizontal line", $f_i(t) = \lambda_1$ and assuming all correlation functions only depend on time differences. We here extend these results to non-stationary processes and arbitrary fitting functions by deriving the analogous expression for Δ_{ab} by using the full multivariate probability density for their fluctuations around the mean values. We find (see section I.E.2 for details):

$$\Delta_{ab} = \frac{4}{M}\sum_{c,d}\sum_{i,j}(\boldsymbol{h}^{-1})_{ac}\left.\frac{\partial f_i}{\partial\lambda_c}\right|_{\lambda_c=\lambda_c^*}(\boldsymbol{R}^T\overline{\boldsymbol{Q}}\,\boldsymbol{R})_{ij}\left.\frac{\partial f_j}{\partial\lambda_d}\right|_{\lambda_d=\lambda_d^*}(\boldsymbol{h}^{-1})_{db}, \tag{I.4}$$

and

$$\begin{aligned} h_{ab} = {} & 2\sum_{i,j}\left.\frac{\partial^2 f_i(\boldsymbol{\lambda})}{\partial\lambda_a\partial\lambda_b}\right|_{\substack{\lambda_a=\lambda_a^*\\ \lambda_b=\lambda_b^*}} R_{ij}\Lambda_j \\ & + 2\sum_{i,j}\left.\frac{\partial f_i(\boldsymbol{\lambda})}{\partial\lambda_a}\right|_{\lambda_a=\lambda_a^*} R_{ij}\left.\frac{\partial f_j(\boldsymbol{\lambda})}{\partial\lambda_b}\right|_{\lambda_b=\lambda_b^*}, \end{aligned} \tag{I.5}$$

where the indices $a, b = 1, \ldots, K$. Eq. (I.4) gives a mathematically rigorous expression (to first order in $1/M$) for the covariance of the estimated parameters, and is a key result. It allows us to accurately

estimate the covariance of any parameter fitted using minimization of the cost function in eq. (I.2). For a good fit, one may neglect the first term on the right hand side of eq. (I.5). Notice that the correlations in fluctuations around mean values enter through the quantity $\overline{\boldsymbol{Q}}$, i.e., the covariance function of the squared displacements, and is estimated using the usual sample estimate above. Neglecting the off-diagonal elements of $\overline{\boldsymbol{Q}}$ above we recover the standard LS error estimation formula [13]. For linear $f_i(\boldsymbol{\lambda})$, eq. (I.4) reduces to previously known expression for linear LS (eq. 5.253 in van den Bos [15]). By setting $\overline{\boldsymbol{R}} = \overline{\boldsymbol{C}}^{-1}$ above we recover the covariance estimation formula for ML [15, 24]. For a stationary process one seeks to fit a horizontal line, $f_i(\lambda_1) = \lambda_1$, to the data. For such a case, the minimization procedure (solving $\partial\chi^2/\partial\lambda_1 = 0$ with $R_{ij} = \delta_{ij}$) yields $\lambda_1 = (1/N)\sum_i \overline{y}_i$, i.e. the parameter estimate is the mean of the data. The error estimation eq. (I.4), then reduces to $\Delta = (1/M)\sum_{i,j} \overline{Q}_{ij}/N^2$, which, using the fact that for a stationary process $\overline{Q}_{ij}$ only depends on time differences, reduces to the usual result [22, 23] used, for instance, in interpretation of Monte Carlo and molecular dynamics simulations. In practice our general formula, eq. (I.4), involves only matrix multiplications and is thus computationally fast and simple to implement.

I.3 RESULTS

To validate the new CLS method (see Methods), our three prototype systems were simulated as described in section I.C. For all systems the MSD is known analytically (important for validation purposes): for BM the MSD is known to behave as $\langle[\boldsymbol{x}(t) - \boldsymbol{x}(0)]^2\rangle = f_{\mathrm{BM}}(\lambda, t) = \lambda t$, where t is the sampling time and $\lambda = 2dD$, and D is the diffusion constant. For CTRW and FBM, we instead have $f(\boldsymbol{\lambda}, t) = \lambda_1 t^{\lambda_2}$, where λ_1 and λ_2 are known, see section I.B.

The integrity of the two previous standard methods, LS and ML, were evaluated together with our CLS method, by applying them to a large set of MSDs, each computed from a fixed data set of M trajectories, sampled N times. For both ML and the CLS/LS methods the 500 fitted values of a given parameter were binned to a histogram, see Figure I.1, and compared to a Gaussian centered on the mean parameter value and a variance from the average error estimate from either of the LS and CLS methods. For all parameters, the CLS/LS yield a correct parameter fit centered on the true value, but only the CLS method gives a correct error

estimation, eq. (I.4), as the narrow predicted width in the LS method, see equations (I.34) and (I.35), is much too optimistic. Clearly, the new error estimation of the CLS method performs extremely well. By contrast, the standard LS method does not provide correct errors of the estimated parameters; this result extends beyond the chosen parameters for (N,M) in Figure I.1, and holds true under general conditions, see Figure I.2.

While the parameters from the CLS and LS methods are centered on the analytical prediction, this is not true for parameters from the ML method, which show a strong bias (Figure I.1). As we show in section I.F.2 we can analytically predict the expected bias for BM when using the ML method (cost function with $\boldsymbol{R} = \overline{\boldsymbol{C}}^{-1}$), see Supplementary Figure I.S3. For large N we find that the bias for ML fitting becomes $\langle \lambda^* \rangle = \overline{\overline{\lambda}} + DG(N)/M$, where $G(N) \approx -8N/(\ln N + \gamma + 2\ln 2)$ and $\gamma \approx 0.5772$ is the Euler-Mascheroni constant. Thus, for large N, the bias increases as $N/\log N$ with the number of sampling points N (see Supplementary Figure I.S3). The strong bias in the estimated parameters for the ML procedure is a general one, appearing in all three prototype systems, as seen in Figure I.1. A similar calculation for the CLS method, see section I.F.3, yields only a minor, essentially N-independent, bias with $G(N) = -4(1 - 1/N)$.

As we have seen, the ML method gives a pronounced bias in the parameter estimate for a specific choice of the number of sampling times N and trajectories M (Figure I.1) for all three prototype systems. In order to investigate the generality of our finding, and its impact on our systems of other known issues with the ML method [16, 17], we explored an extended region of phase space (N, M), see Figure I.3. For a region determined by large N and moderate to small M the sample estimate for the covariance matrix $\boldsymbol{C}$ is ill-conditioned. In practice this means that it cannot be numerically inverted, as required in the ML parameter estimation procedure, without uncontrollable numerical errors. For parts of phase space where ill-conditioness is not an issue, we defined an acceptable fit as one where the bias is smaller than 5% (compared to the analytic value, $\overline{\overline{\lambda}}_a$). We find that for BM and FBM there is indeed a thin region defined by large M and small N, where ML works. In contrast, for CTRW the ML only yields acceptable parameter estimation in a very limited part of phase space (e.g. for $N < 25$ for $M = 1000$). The bias inherent in the ML method can be reduced by applying the common jackknife procedure [25], which removes bias terms proportional to

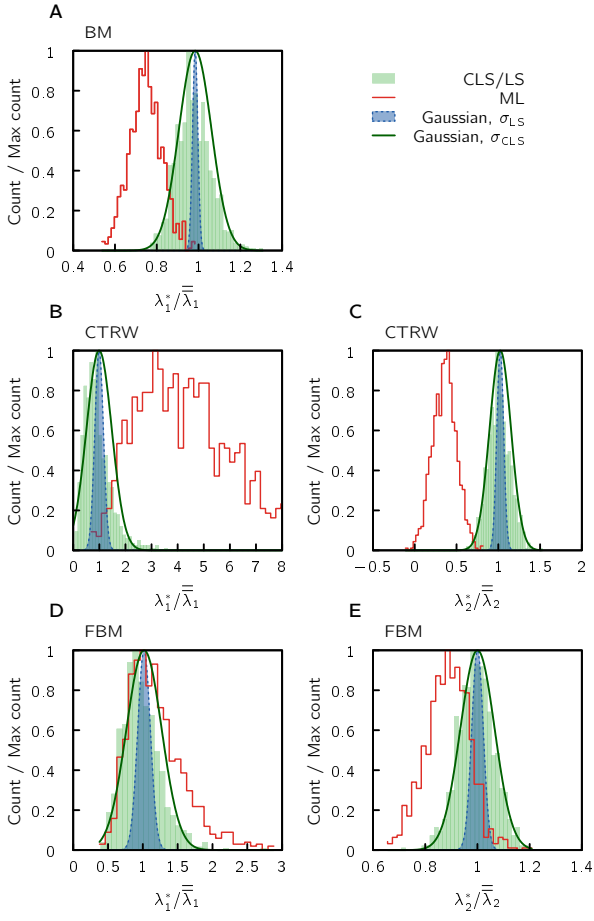

Figure I.1 Histograms of fitted parameters for CLS/LS and ML compared to theoretical predictions. Each method is tested on: (A) Brownian motion (BM), (B-C) continuous time random walk (CTRW), and (D–E) fractional Brownian motion (FBM). In each test, we generate 500 data sets, each consisting of $M = 150$ particle trajectories sampled at $N = 80$ time points (histograms). Left and right panels show histograms (500 fitted parameters) for the effective diffusion constant λ_1, and the exponent λ_2, respectively. The two Gaussian curves are centered on the mean of the fitted parameters and have a width corresponding to the parameter uncertainty estimated by the fit method (averaged over 500 fits). The ML fit exhibits a strong bias in the parameter value (not centered on the analytical prediction), and the LS fit gives an error estimation, see eq. (I.35), that is much too small. The new CLS procedure (Methods) works well, i.e. exhibits negligible bias and yields correct error estimation, eq. (I.4). The rather large number of trajectories ($M = 150$) was used in order to avoid ill-conditioness issues for the ML fitting, compare to Figure I.3. For simulation parameters, see section I.C.

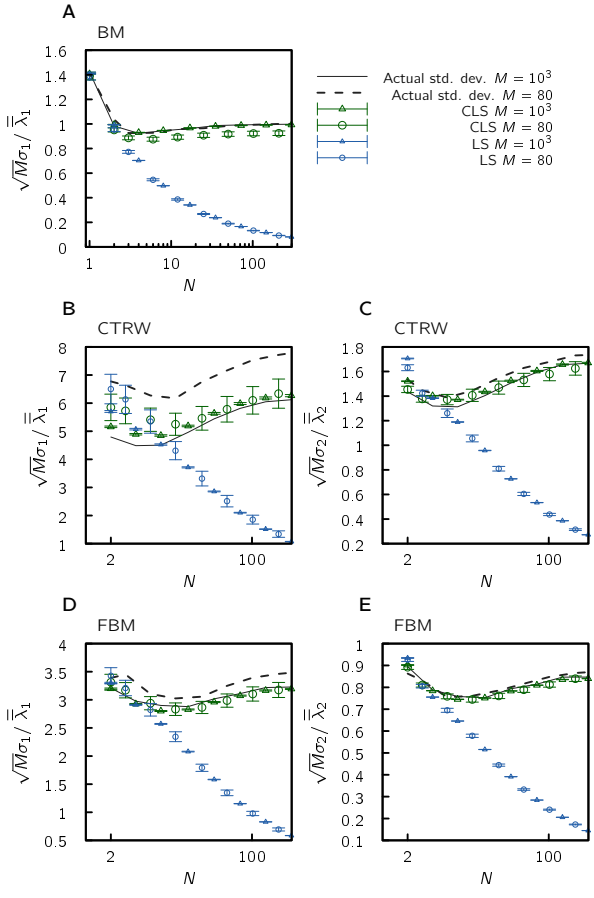

Figure I.2 Error estimation. Standard deviation from the LS and CLS parameter fits as a function of the number of sampling points, N, used in the fitting procedure (log-scale on the horizontal axis for visibility). Each method is applied to 500 realizations of data from (A) Brownian motion (BM), (B-C) continuous time random walk (CTRW), and (D–E) fractional Brownian motion (FBM). In conjunction we show the actual standard deviation of each of these methods computed from the parameters from the fit (lines), i.e. the width seen in Figure I.1, but for an extended range of N. It is evident that the standard deviation from the LS fit is far too small for almost all N. Error bars show standard error of the mean. For $M = 80$ there is a small bias in the observable σ, as compared to actual standard deviation. This bias can be removed using the jackknife procedure, see section I.G. For simulation parameters, see section I.C.

$1/M$, see section I.G. By applying the (first-order) jackknife procedure (Figure I.3) we find that the bias is reduced which expands the region of the phase space where ML method may be used reliably. However, for the prefactor λ_1, in CTRW, the jackknife procedure for ML does not reduce the bias to an acceptable level. Note that the computational time is a factor g (g is the number of groups into which the trajectories are pooled) larger for the first-order jackknife procedure compared to the non-jackknife case. Finally, in principle the jackknifing procedure can be extended to remove higher order bias terms (proportional to $1/M^n$, with $n = 2, 3, \ldots$) [25]. However, for the present case there is no guarantee that these higher order terms have this functional form with respect to M, see section I.F.1. Also, our results show that the second-order jackknife increased, rather than decreased, the bias in the parameter estimations for most parts of the phase spaces. For BM, Supplementary Figure I.S4 indicates that the reason for this is that the third order term (term proportional to $1/M^3$) is generally larger in amplitude (but of opposite sign) than the second order one. It is also important to remember that higher order bias reduction comes at a computational price. The number of numerical evaluations required for second order jackknife is $g(g+1)/2$ times that of non-jackknifed parameter estimation. Due to these findings and the lack of a formal functional form for the bias, beyond the $1/M$ term (see above), we do not recommend applying the jackknife procedure beyond first order. Finally, we point out that the new error estimation formula, eq. (I.4), remains valid also for jackknifed parameters, see section I.G.3.

In Supplementary Figure I.S5 we investigated the "goodness of fit" for the CLS and ML procedures using a standard R^2 measure, see section I.I. A good fit is characterized by $R^2 \approx 1$. We find that, in this sense, the new method provides "good" fits. In contrast, the ML method in general provides "bad" fits with $R^2 \ll 1$ for large N.

Let us finally discuss error estimation using subsampling [23], as an alternative to the CLS method. Subsampling refers to the method of choosing sampling times sufficiently sparsely in order to make the data points essentially uncorrelated (the “brute force” method in Supplementary Figure I.S1 is an extreme case of subsampling where only one data point per trajectory is kept). After subsampling, error analysis is performed using standard error analysis for independent data. In order to properly choose N within this method, N is systematically decreased until the variance saturates to a constant (this constant is assumed to be

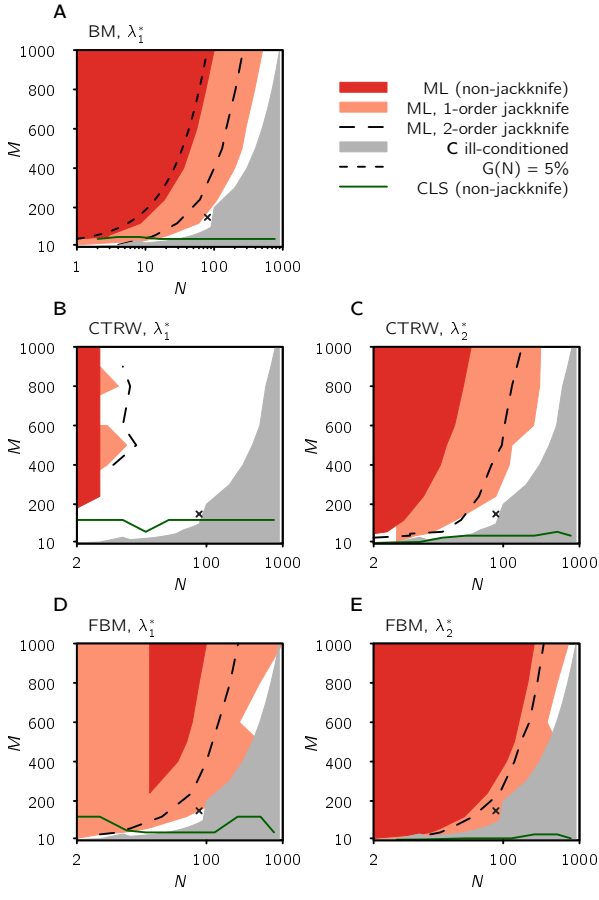

Figure I.3 Phase space of reliable ML fitting. For each of our example systems, (A) Brownian motion (BM), (B–C) continuous time random walk (CTRW), and (D–E) fractional Brownian motion (FBM), we investigate for which number of sampling times N, and number of trajectory realizations M, the fitting is more than 5% off from its analytical value, averaged over 500 MSDs. As indicated, ML is only reliable in a limited region (large M, small N), which can be extended by a first order jackknife correction. The cross marks the N, M used in Figure I.1 and Supplementary Figure I.S2. For BM we also include when the analytically predicted first order bias term for ML, $G(N)$, eq. (I.80), gives a bias that is 5% of the exact parameter value, see section I.F.2. We also show the boundary for when more than half of the 500 generated covariance matrices become ill-conditioned. Interestingly, a second order jackknife generally does more harm than good compared to the first order, which we elaborate more on in Supplementary Figure I.S4. In contrast to ML (non-jackknifed), the CLS/LS method is valid for most N, M (region above the curve), and can be extended even further using a jackknife approach (data not shown).

the true variance) [22, 23]. Figure I.2 shows how estimated errors from our LS and CLS analyses depend on the number of data points used, N. We find that temporal correlations are so strong that the LS method underestimates the errors down to very small N. Moreover, finding a sufficiently small N is difficult, since the LS does not in general saturate to a constant level as N is reduced. These challenges are completely circumvented by instead using the error estimation from the CLS method (i.e. using eq. (I.4) instead of eq. (I.35)).

Within our framework, we formulated the fitting procedure in terms of a cost function, involving a positive definite matrix $\boldsymbol{R}$. In all illustrations, we choose $\boldsymbol{R}$ consistent with a LS cost function. However, there is a great deal of flexibility in the choice of $\boldsymbol{R}$. In the Supplementary, we explore yet another choice (based on the Cholesky decomposition of the exact BM covariance matrix), which, for BM (but not for FBM and CTRW) yield comparable (slightly improved with respect to variance in parameter estimation) results to the method introduced herein (Supplementary Figure I.S6). It remains a future challenge to find $\boldsymbol{R}$ matrices which are adapted to specific classes of problems; i.e., can one tailor $\boldsymbol{R}$ to yield minimum variances? In this respect, the Cramer-Rao bound is a useful tool, providing an expression for the smallest possible variance for parameter estimators. In practice, the Cramer-Rao bound requires a knowledge of the full multivariate probability density for the process under investigation. For the cases considered here, namely, squared displacements, this probability density is not known (see section I.H), which for the present applications, may limit the practical usefulness of the Cramer-Rao bound.

I.4 DISCUSSION, CONCLUSION AND OUTLOOK

An important step in analysis of particle tracking data is that of fitting a model to the time-evolving mean of some observable and estimate the associated model parameters. Since fluctuations around observed mean values, calculated based on particle trajectories, are in general correlated in time (the particle is likely to remain at the same side of the average for two consecutive points in time), the standard least squares (LS) method provides error estimates for parameters which can be more than one order of magnitude too small. Further, we demonstrated that the maximum likelihood (ML) estimation, involving numerical inversion of a noisy covariance matrix, yields a very strong

bias, or ill-conditioning, in parameter estimation. We remedied this lack of tools for parameter estimation for ensemble averages, by deriving a new error estimation formula for LS fitting, that does not require inversion of the full covariance matrix. The new formula provides simple means for parameter estimation: (A) perform a LS fit to the data, (B) use the new error estimation formula, eq. (I.4), to estimate the variance as well as covariance among parameters. We demonstrated on three simulated prototype systems that this method provides accurate results that far outperform the standard ML and LS methods.

In this study, we emphasize the use of our fitting procedure for ensemble averages based on particle trajectories. However, we wish to point out that χ^2 minimization is ubiquitous throughout all fields of science. Therefore, we expect that our method finds its way into essentially all fields where quantitative analysis of data is used, and that the new error estimation formula will find its way into standard statistical packages. The new formula is as "easy" to compute as standard LS error estimates, and, needless to say, the formula incorporates the uncorrelated case.

ACKNOWLEDGMENTS

We are grateful to Bo Söderberg for fruitful discussions. T.A. was supported by the Swedish Research Council (grant nos 2009-2924 and 2014-4305). K.F. was supported by the Swedish Research Council (grant no 2010-5219). M.A.L. acknowledges funding from the Danish council for Independent Research-Natural Sciences (FNU), grant number 4002-00428B.

REFERENCES

1. S. W. Hell, S. J. Sahl, M. Bates, X. Zhuang, R. Heintzmann, M. J. Booth, J. Bewersdorf, G. Shtengel, H. Hess, P. Tinnefeld, *et al.*, "The 2015 super-resolution microscopy roadmap," *Journal of Physics D: Applied Physics*, vol. 48, no. 44, p. 443001, 2015.
2. A. Yildiz, J. N. Forkey, S. A. McKinney, T. Ha, Y. E. Goldman, and P. R. Selvin, "Myosin V walks hand-over-hand: single fluorophore imaging with 1.5-nm localization," *Science*, vol. 300, no. 5628, pp. 2061–2065, 2003.

3. M. Bates, B. Huang, and X. Zhuang, "Super-resolution microscopy by nanoscale localization of photo-switchable fluorescent probes," *Current opinion in chemical biology*, vol. 12, no. 5, pp. 505–514, 2008.
4. S. W. Hell, "Far-field optical nanoscopy," *Science*, vol. 316, no. 5828, pp. 1153–1158, 2007.
5. J. Lippincott-Schwartz and S. Manley, "Putting super-resolution fluorescence microscopy to work," *Nature methods*, vol. 6, no. 1, pp. 21–23, 2009.
6. A. Gahlmann and W. Moerner, "Exploring bacterial cell biology with single-molecule tracking and super-resolution imaging," *Nature Reviews Microbiology*, vol. 12, no. 1, pp. 9–22, 2014.
7. K. R. Chi, "Super-resolution microscopy: breaking the limits," *Nature methods*, vol. 6, no. 1, pp. 15–18, 2009.
8. M. J. Saxton, "Single-particle tracking: connecting the dots," *Nature methods*, vol. 5, no. 8, pp. 671–672, 2008.
9. G. Lukinavičius, K. Umezawa, N. Olivier, A. Honigmann, G. Yang, T. Plass, V. Mueller, L. Reymond, I. R. Corrêa Jr, Z.-G. Luo, *et al.*, "A near-infrared fluorophore for live-cell super-resolution microscopy of cellular proteins," *Nature chemistry*, vol. 5, no. 2, pp. 132–139, 2013.
10. S. Wolter, A. Löschberger, T. Holm, S. Aufmkolk, M.-C. Dabauvalle, S. van de Linde, and M. Sauer, "rapidSTORM: accurate, fast open-source software for localization microscopy," *Nature methods*, vol. 9, no. 11, pp. 1040–1041, 2012.
11. K. I. Mortensen, L. S. Churchman, J. A. Spudich, and H. Flyvbjerg, "Optimized localization analysis for single-molecule tracking and super-resolution microscopy," *Nature methods*, vol. 7, no. 5, pp. 377–381, 2010.
12. D. Sage, H. Kirshner, T. Pengo, N. Stuurman, J. Min, S. Manley, and M. Unser, "Quantitative evaluation of software packages for single-molecule localization microscopy," *Nature methods*, vol. 12, no. 8, pp. 717–724, 2015.
13. 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.
14. C. Michael, "Fitting correlated data," *Phys. Rev. D*, vol. 49, pp. 2616–2619, 1994.

15. A. Van den Bos, *Parameter estimation for scientists and engineers*. John Wiley & Sons, 2007.
16. D. Seibert, "Undesirable effects of covariance matrix techniques for error analysis," *Phys. Rev. D*, vol. 49, pp. 6240–6243, Jun 1994.
17. B. Yoon, Y.-C. Jang, C. Jung, and W. Lee, "Covariance fitting of highly-correlated data in lattice QCD," *Journal of the Korean Physical Society*, vol. 63, no. 2, pp. 145–162, 2013.
18. E. Barkai, Y. Garini, and R. Metzler, "Strange kinetics of single molecules in living cells," *Phys. Today*, vol. 65, no. 8, p. 29, 2012.
19. N. Chenouard, I. Smal, F. De Chaumont, M. Maška, I. F. Sbalzarini, Y. Gong, J. Cardinale, C. Carthel, S. Coraluppi, M. Winter, *et al.*, "Objective comparison of particle tracking methods," *Nature methods*, vol. 11, no. 3, p. 281, 2014.
20. B. B. Machta, R. Chachra, M. K. Transtrum, and J. P. Sethna, "Parameter space compression underlies emergent theories and predictive models," *Science*, vol. 342, no. 6158, pp. 604–607, 2013.
21. M. K. Transtrum, B. B. Machta, and J. P. Sethna, "Why are nonlinear fits to data so challenging?," *Physical Review Letters*, vol. 104, no. 6, p. 060201, 2010.
22. H. Flyvbjerg and H. G. Petersen, "Error estimates on averages of correlated data," *The Journal of Chemical Physics*, vol. 91, no. 1, pp. 461–466, 1989.
23. B. A. Berg and A. Billoire, *Markov chain monte carlo simulations*. Wiley Online Library, 2008.
24. S. Gottlieb, W. Liu, R. L. Renken, R. L. Sugar, and D. Toussaint, "Hadron masses with two quark flavors," *Phys. Rev. D*, vol. 38, pp. 2245–2265, Oct 1988.
25. R. G. Miller, "The jackknife — a review," *Biometrika*, vol. 61, no. 1, pp. 1–15, 1974.

I.A SUPPLEMENTARY FIGURES

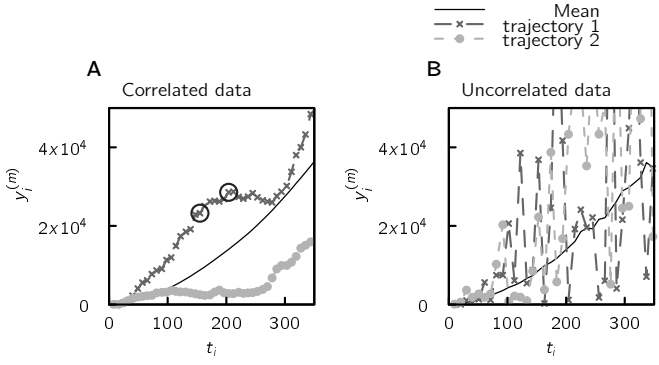

Figure I.S1 Correlated and uncorrelated (synthetic) squared displacements. The displacement squared $y_i^{(m)} = [\boldsymbol{x}^{(m)}(t_i) - \boldsymbol{x}^{(m)}(0)]^2$ for fractional Brownian motion as a function of time, t, for two trajectories, labeled by m, and the mean of a large ensemble ($M = 10^3$) of trajectories. Panel (A) shows actual trajectories which exhibit strong correlation, meaning: if we are above the mean for some time point on a trajectory, we are likely to still be for time points close to it (circled). In panel (B) we, for comparison, construct "synthetic" trajectories by only using one data point from each real trajectory, and "throw away" the rest, resulting in (computationally expensive) uncorrelated data. That is, within this "brute force" method, to generate a single uncorrelated trajectory of N sampling points, we need to use the same amount of real trajectories, and throw away all data points save one. Data was generated from a one-dimensional fractional Brownian motion simulation with Hurst parameter $H = 0.9$, see section I.B.3.

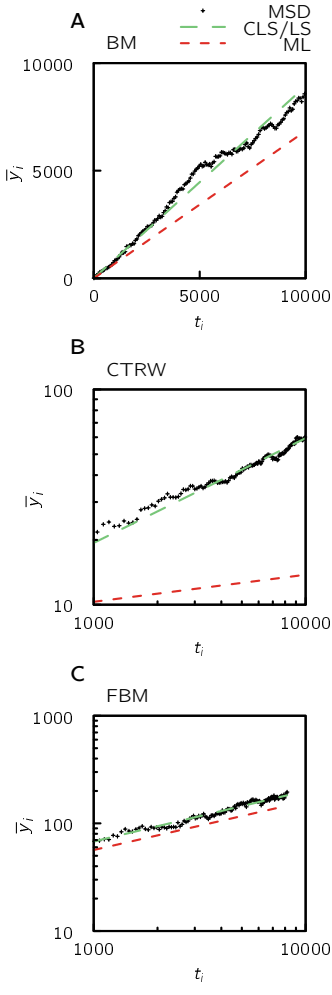

Figure I.S2 Example of a fit to MSD data for the CLS/LS and ML methods. An illustrative example of a typical fit to MSD trajectory data, based on $M = 150$ trajectories, for (A) Brownian motion (BM), (B) continuous time random walk (CTRW), and (C) fractional Brownian motion (FBM). The model parameters were fitted to the MSD using either CLS/LS, or ML fitting procedure, for $N = 80$, $M = 150$, see Fig. I.1 and cross in Fig. I.3. For ML fitting to the MSD of FBM data, we see that although the exponent is almost the same, the pre-factor is inaccurate.

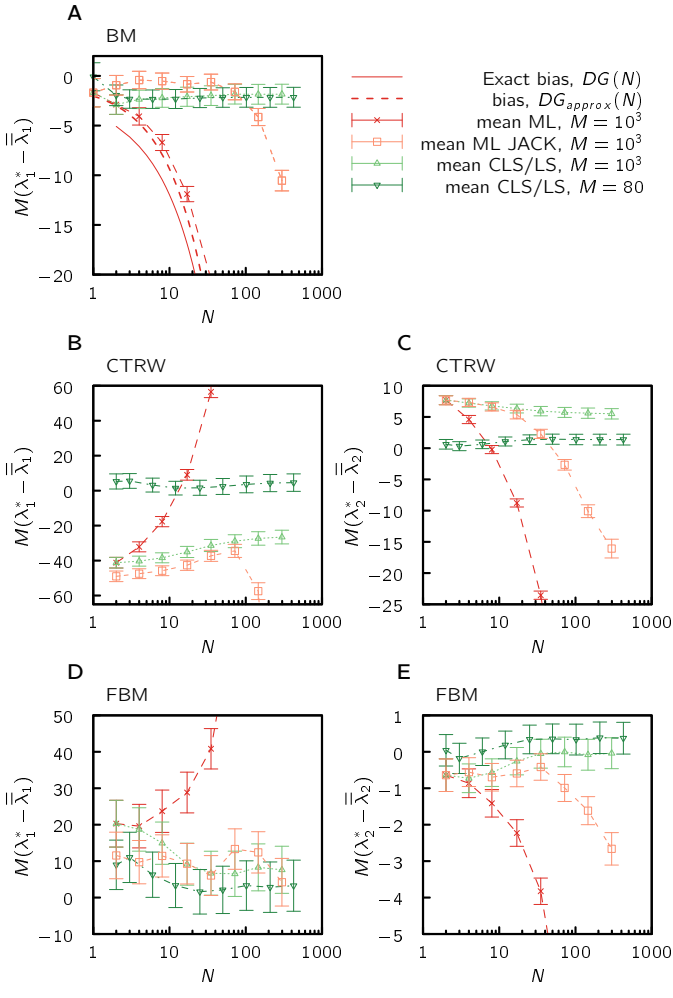

Figure I.S3 Bias in the parameter fit. The residual bias in the fit (multiplied by the number of trajectories M) as a function of sampling points, N (log-scale for the horizontal axis for visibility), averaged over parameters from fitting to 500 MSD realizations. (A) For the Brownian motion ML fit, the analytical prediction, $G(N)$, (full line) for the first order bias follows the observed bias for $M = 10^3$, data (section I.F.2). Also for (B–C) continuous time random walk, and (D–E) fractional Brownian motion, the bias term in ML is large compared to CLS. The bias can be alleviated to some degree by a computationally demanding Jackknife procedure. Error bars show standard error of the mean.

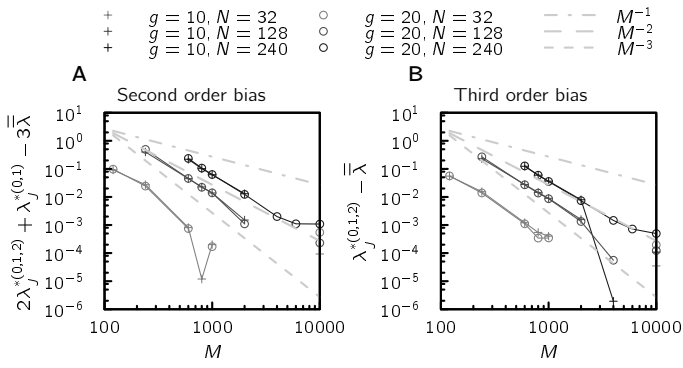

Figure I.S4 High order bias contribution. The bias in the parameter estimation is commonly assumed to be of the form $\lambda^* = \overline{\overline{\lambda}} + a/M + b/M^2 + c/M^3 + \mathcal{O}(M^{-4})$, see section I.F. In panel (A) the vertical axis shows the (negative) second order bias term $-b/M^2$, and in (B) the (positive) third order term, c/M^3, for three different number of sampling times N. Note that these are of comparable magnitude, but opposite sign. Thus a second order jackknife, which removes terms proportional to a/M and b/M^2, may yield more unfavorable results than a first order jackknife, which only removes the a/M term. We note that the slope of the second order bias term approximately corresponds to M^2, and the third order is slightly more. For panel (A) the second order bias was extracted combining eq. (I.113) and eq. (I.115), to give $-b/M^2 = 2\lambda_J^{*(0,1,2)} + \lambda_J^{*(0,1)} - 3\overline{\overline{\lambda}}$, and for panel (B) we have $(\lambda_J^{*(0,1,2)} - \overline{\overline{\lambda}}) = c/M^3$, which follows immediately from eq. (I.115).

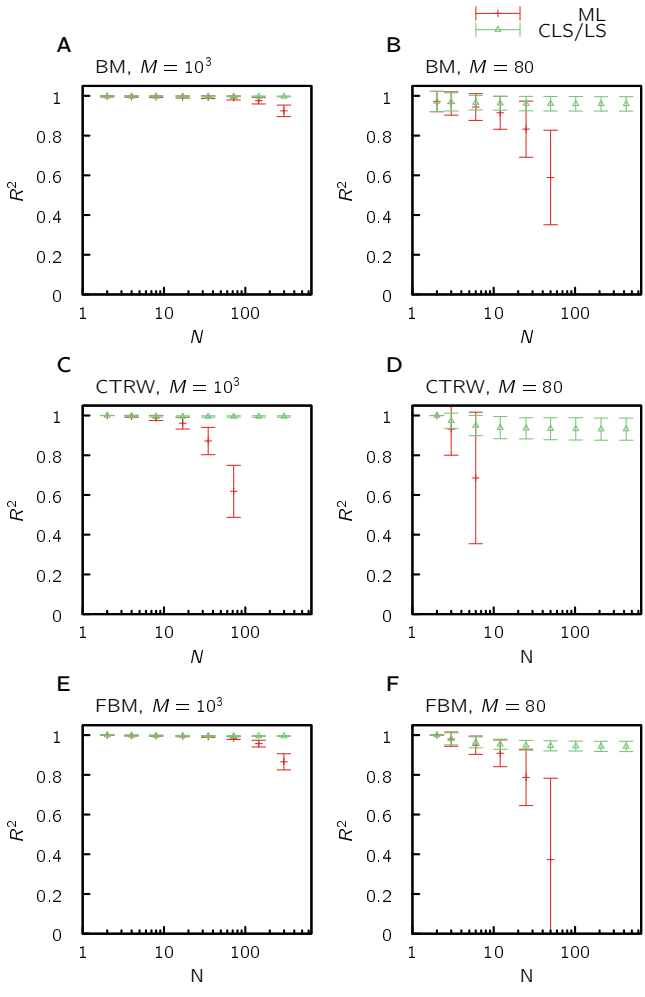

Figure I.S5 Goodness-of-fit using the the coefficient of determination. The quality of the ML and CLS/LS fit is quantified by the coefficient of determination, R^2, as a function of sampling points, N (horizontal axis on log-scale for visibility), for our three prototype systems: (A–B) Brownian motion (BM), (C–D) continuous time random walk (CTRW), and (E–F) fractional Brownian motion (FBM). A perfect fit yields unit value, while a bad fit results in $R^2 \ll 1$. The number of trajectories used in the MSD was either $M = 10^3$ (left), or $M = 80$ (right). All data was averaged over 500 realizations, with standard deviation given by the error bars. For panel (D) only a few data points could be obtained, due to numerical instability of the ML-method, and for panels (B,F) $R^2 < 0$ for larger N.

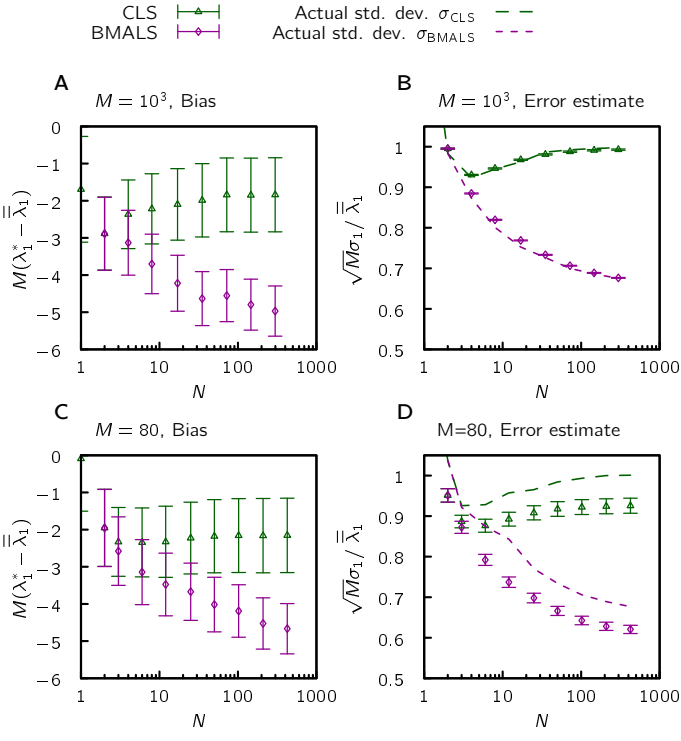

Figure I.S6 Bias and variance of Brownian motion adapted least squares (BMALS) compared to the CLS method. We show bias in parameter fit (left panels) and their variance compared to estimates from fitting procedure (right panels), as a function of the number of sampling times, N. The MSD based on two different data sizes, M (number of trajectories) was considered: (A,B) $M = 10^3$ and (C,D) $M = 80$; averaged over 500 realizations. It is evident that there is a bias-variance trade-off between the two fitting procedures used, CLS and BMALS: the lower variance in BMALS, as compared to CLS, comes at the price of a higher bias in the parameter fit. Error bars show standard errors of the mean.

I.B PROTOTYPICAL EXAMPLE SYSTEMS

In the main text we provide results for different parameter estimation procedures. As prototype systems we use three processes where the true parameter values are known, namely: (i) Brownian motion (BM), (ii) continuous time random walks (CTRW), and (iii) fractional Brownian motion (FBM). As noted in the main text, these systems are of interest as they have been shown to be of importance to motion of fluorescently labeled particles in cells [1]. For BM and CTRW in d spatial dimensions, steps in different directions are independent. Therefore, without loss of generality, all simulations are here performed in one dimension, $d = 1$, for these systems. Also, for consistency, we use $d = 1$ in our FBM simulations.

I.B.1 *Brownian motion*

Our first example is a simple Brownian motion where a single particle diffuses in one dimension. The mean square displacement (MSD) at time t, for dimension d, and diffusion constant D, is

$$\langle (\boldsymbol{x}(t) - \boldsymbol{x}(0))^2 \rangle = \langle y(t) \rangle = \lambda t, \tag{I.6}$$

where

$$\lambda = 2dD \tag{I.7}$$

and

$$y(t) = [\boldsymbol{x}(t) - \boldsymbol{x}(0)]^2. \tag{I.8}$$

In all simulations in the main text we use one-dimensional simulations, i.e., $d = 1$.

In one-dimensional Brownian motion, the full covariance matrix for the displacement is known [2]. Choosing $x(0) = 0$ and discretizing time according to $t_i = i\epsilon$ $(i = 1, \ldots, N)$, with time step ϵ, we have

$$\overline{\overline{V}}_{ij} = \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle = 2D \min(t_i, t_j), \tag{I.9}$$

where $x_i = x(t_i)$ and D is the diffusion constant. On matrix form:

$$\overline{\overline{\boldsymbol{V}}} = 2D\epsilon \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & 2 & 2 & & 2 \\ 1 & 2 & 3 & & 3 \\ \vdots & & & \ddots & \\ 1 & 2 & 3 & & N \end{pmatrix}. \tag{I.10}$$

Of interest here is also the covariance matrix for the square displacements:

$$\overline{\overline{Q}}_{ij} = \langle (y_i - \langle y_i \rangle)(y_j - \langle y_j \rangle) \rangle. \tag{I.11}$$

Using Wick's (Isserlis') theorem for zero-mean processes, we can calculate any moment of a multivariate Gaussian according to

$$E[x_1 x_2 \cdots x_{2n}] = \sum \prod E[x_i x_j], \tag{I.12}$$

where the sum is over all distinct ways of partitioning $x_1 \dots, x_{2n}$ into pairs $x_i x_j$. Using eq. (I.12) we have the following relation between $\overline{\overline{\boldsymbol{Q}}}$ and $\overline{\overline{\boldsymbol{V}}}$:

$$\overline{\overline{Q}}_{ij} = 2\overline{\overline{V}}^2_{ij}. \tag{I.13}$$

On matrix form:

$$\overline{\overline{\boldsymbol{Q}}} = 8(D\epsilon)^2 \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & 4 & 4 & & 4 \\ 1 & 4 & 9 & & 9 \\ \vdots & & & \ddots & \\ 1 & 4 & 9 & & N^2 \end{pmatrix}. \tag{I.14}$$

The standard unbiased sample estimator $\overline{\overline{\boldsymbol{Q}}}$ is[1]

$$\overline{Q}_{ij} = \frac{1}{M-1} \sum_m (y_i^{(m)} - \overline{y}_i)(y_j^{(m)} - \overline{y}_j). \tag{I.15}$$

1 Note that the sample estimate above is unbiased, as it should:

$$\begin{aligned} \langle \overline{Q}_{ij} \rangle &= \frac{1}{M-1} \left(\sum_m \langle y_i^{(m)} y_j^{(m)} \rangle - \frac{1}{M} \langle \sum_n y_i^{(n)} \sum_p y_j^{(p)} \rangle \right) \\ &= \frac{1}{M-1} \left(M\overline{\overline{\sigma}}^2_i \overline{\overline{\sigma}}^2_j + 2M\overline{\overline{V}}^2_{ij} - M\overline{\overline{\sigma}}^2_i \overline{\overline{\sigma}}^2_j - 2\overline{\overline{V}}^2_{ij} \right) = \overline{\overline{Q}}_{ij}. \end{aligned}$$

where m labels trajectories, see main text.

For Brownian motion the inverse of the $\overline{\overline{\boldsymbol{Q}}}$ matrix is a tridiagonal matrix with column sum of zero, except the first. Explicitly

$$\overline{\overline{\boldsymbol{Q}}}^{-1} = \frac{1}{8(D\epsilon)^2}\begin{pmatrix} 1+\frac{1}{3} & -\frac{1}{3} & 0 & \dots & & \\ -\frac{1}{3} & \frac{1}{3}+\frac{1}{5} & -\frac{1}{5} & 0 & & \\ 0 & -\frac{1}{5} & \frac{1}{5}+\frac{1}{7} & -\frac{1}{7} & 0 & \\ \vdots & 0 & -\frac{1}{7} & \ddots & \ddots & \\ & & 0 & & & -\frac{1}{2N-1} \\ & & & & & \frac{1}{2N-1} \end{pmatrix}, \quad \text{(I.16)}$$

which can be written as

$$\begin{aligned}(\overline{\overline{\boldsymbol{Q}}}^{-1})_{ij} = \frac{1}{8(D\epsilon)^2}\Bigg[&\left(\frac{1}{2i-1}+\frac{(1-\delta_{i,N})}{2i+1}\right)\delta_{ij} \\ &-\left(\frac{1}{2i+1}\right)\delta_{i,j-1}-\left(\frac{1}{2i-1}\right)\delta_{i,j+1}\Bigg].\end{aligned} \quad \text{(I.17)}$$

It is straigtforward to show that indeed the matrix above satisfies $(\overline{\overline{\boldsymbol{Q}}}^{-1})\cdot\overline{\overline{\boldsymbol{Q}}} = \boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix.

For later purposes, let us also provide a (Cholesky) decomposition of the $\overline{\overline{\boldsymbol{Q}}}^{-1} = \overline{\overline{\boldsymbol{A}}}^{T}\overline{\overline{\boldsymbol{A}}}$ matrix. We have

$$\overline{\overline{\boldsymbol{A}}} = \frac{1}{\sqrt{8}D\epsilon}\begin{pmatrix} 1 & 0 & \dots & & \\ -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & 0 & & \\ 0 & -\frac{1}{\sqrt{5}} & \frac{1}{\sqrt{5}} & 0 & \\ \vdots & 0 & \ddots & \ddots & \\ & & & -\frac{1}{\sqrt{2N-1}} & \frac{1}{\sqrt{2N-1}} \end{pmatrix}, \quad \text{(I.18)}$$

which can be written as

$$\overline{\overline{A}}_{ij} = \frac{1}{\sqrt{8}D\epsilon}\left[\left(\frac{1}{\sqrt{2i-1}}\right)\delta_{ij} - \left(\frac{1}{\sqrt{2i-1}}\right)\delta_{i,j+1}\right]. \quad \text{(I.19)}$$

I.B.2 *Continuous time random walk*

Our second example uses the continuous time random walk (CTRW) in one dimension. Such a process is defined through a waiting time density

$\psi(\tau)$, and a jump length probability density, $\varphi(\ell)$ [3]. In our case we choose

$$\psi(\tau) = \frac{\alpha}{\tau^*}(1 + \tau/\tau^*)^{-1-\alpha} \tag{I.20}$$

with $0 < \alpha < 1$ so that we have infinite average waiting time $\langle\tau\rangle$.[2] The jump length probability density is chosen to be a Gaussian:

$$\varphi(\ell) = \frac{1}{\sqrt{2\pi a^2}} \exp\left(-\frac{\ell^2}{2a^2}\right) \tag{I.21}$$

with a variance a^2. For such a process, the MSD follows (for long times) [3]:

$$\langle x(t)^2\rangle = \lambda_1 t^{\lambda_2} \tag{I.22}$$

(with $x(0) = 0$) where

$$\lambda_1 = \frac{2}{\Gamma(1+\alpha)\Gamma(1-\alpha)} \frac{a^2}{2(\tau^*)^\alpha}, \tag{I.23}$$

and

$$\lambda_2 = \alpha. \tag{I.24}$$

I.B.3 *Fractional Brownian motion*

Our final example is the case of one-dimensional fractional Brownian motion (FBM), which is a zero mean Gaussian process with autocorrelation function, [4]

$$v_{ij} = \langle x(t_i)x(t_j)\rangle = c(t_i^{2H} + t_j^{2H} - |t_i - t_j|^{2H}), \tag{I.25}$$

at discrete times $t_i = i\epsilon$ and where the parameter H denotes the Hurst parameter [5]. For $H = 1/2$, fractional Brownian motion becomes standard Brownian motion. Indeed, if we set $H = 1/2$ in eq. (I.25) we find that $v_{ij} = c[(t_i + t_j) - |t_i - t_j|] = 2c\min(t_i, t_j)$ which is identical to eq. (I.9) if we choose $c = D$. The inverse covariance matrix of eq. (I.25) is (currently) not known analytically.

2 In simulations, a random waiting time is obtained by drawing a uniformly distributed random number $r \in [0, 1]$, and then calculating $\tau = \tau^*(r^{-1/\alpha} - 1)$.

From eq. (I.25) we get the MSD, for $t_i = t_j$, as $(x(0) = 0)$

$$\langle x^2(t)\rangle = \lambda_1 t^{\lambda_2}, \tag{I.26}$$

where $\lambda_1 = 2c$ and $\lambda_2 = 2H$, i.e. the MSD has, for $H < 1/2$, a sublinear (or superlinear, if $H > 1/2$) dependence on time, t.

I.C SIMULATION PROCEDURES

In this section we provide details about the methods used to generate the data for our prototypical example systems introduced in section I.B. All simulations ran to a stop time $t = 10^4$.

I.C.1 *Brownian motion*

For BM we let the particle move in one dimension with random jump length drawn from a normal distribution. We start by taking the cumulative sum of N random numbers from a Gaussian distribution with zero mean and unit variance, and square each element of the sum. Each step increments time by $\epsilon = 1$. This is repeated M times and summed and averaged. In short, the MSD was computed as:

$$\overline{y}_i = \frac{1}{M}\sum_{m=1}^{M}\left[\sum_{n=1}^{i} R_n^{(m)}\right]^2, \tag{I.27}$$

where $R_n^{(m)}$ is a random number drawn from a normal distribution, associated with the length of the nth jump for trajectory m.

I.C.2 *Continuous time random walk*

For generating the CTRW data we move a particle randomly with a step length drawn from a Gaussian probability density, eq. (I.21), at each time step and increment time with a waiting time τ from the power-law distribution in eq. (I.20). In greater detail: We let $y_i^{(m)}$ be the displacement squared at time point $\tilde{t}_i$ for trajectory m, which is to be saved, and the current time of the system is t. While t is smaller than the designated stop time we repeat the following procedure to generate one trajectory m:

1. Draw a random waiting time, τ, from the power-law in eq. (I.20).

2. While $i < N$ and $t \leq \tilde{t}_i < t + \tau$.
   a) Save displacement squared: $y_i^{(m)}$.
   b) Increase index i by one.
3. Move the particle, by increasing the current displacement by a random number R drawn from a normal distribution.
4. Update the time t by τ.

The procedure is repeated M times and averaged over, to yield the MSD. For all our simulations we chose $\alpha = 0.5$, $a = 1$ and $\tau^* = 1$. Since the prediction in eq. (I.22) is only valid for $t \gg \tau^*$, for fitting purposes, we include only time points $t \geq T_1$ with $T_1 = 200\tau^*$ in the χ^2 expression, eq. (I.2), and in the associated parameter covariance estimation formula, eq. (I.4).

I.C.3 *Fractional Brownian motion*

For FBM simulations we used an algorithm by Davies and Harte [6, 7]. In all our simulations, unless stated otherwise, the Hurst exponent was chosen as $H = 1/4$. When fitting the model in eq. (I.26), we include only time points $t \geq T_1$ with $T_1 = 200$, since this model prediction for the MSD, as for CTRW (see section I.C.2), is only valid for large simulation times.

I.D REVIEW OF STANDARD FITTING PROCEDURES

In this section we investigate the two previous ubiquitous χ^2 methods for model fitting, namely LS (uncorrelated χ^2) fitting and ML (correlated χ^2) fitting.

I.D.1 *Least squares fitting*

The previous most common method of functional fitting to data is the least squares (LS) method (uncorrelated χ^2 fitting), which is reviewed in this section.

I.D.1.1 *General fit functions*

In least squares fitting one maximizes the probability for the function $f(t_i; \boldsymbol{\lambda}) = f_i$ to have a good fit to the data:

$$P(\overline{\boldsymbol{y}}; \boldsymbol{\lambda}) \propto \prod_{i=1}^{N} \exp\left(-\frac{1}{2}\frac{(\overline{y}_i - f_i)^2}{\overline{\sigma}_i^2}\right). \tag{I.28}$$

Note that this probability is a product over the observations, $\overline{\boldsymbol{y}}$, hence the data is assumed to be statistically *independent*. Within this assumption, the unbiased estimator of variance of the mean is

$$\overline{\sigma}_i^2 = \frac{1}{M}\frac{1}{M-1}\sum_{m=1}^{M}(y_i^{(m)} - \overline{y}_i)^2. \tag{I.29}$$

Maximizing the probability P is equivalent to minimizing

$$\chi^2 = \sum_{i=1}^{N}\frac{(\overline{y}_i - f_i)^2}{\overline{\sigma}_i^2}, \tag{I.30}$$

from which we get the best parameters $\boldsymbol{\lambda}^*$, by solving

$$\left.\frac{\partial\chi^2}{\partial\lambda_a}\right|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^*} = 0 = 2\sum_i \left.\frac{\partial f_i(\boldsymbol{\lambda})}{\partial\lambda_a}\right|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^*}\frac{1}{\overline{\sigma}_i^2}(f_i(\boldsymbol{\lambda}^*) - \overline{y}_i). \tag{I.31}$$

For χ^2 close to the optimal parameter set $\boldsymbol{\lambda}^*$ we have the Taylor expansion

$$\begin{aligned}\chi^2 = \; & \chi^2\big|_{\boldsymbol{\lambda}^*} + \sum_{a=1}^{K}(\lambda_a - \lambda_a^*)\left.\frac{\partial\chi^2}{\partial\lambda_a}\right|_{\lambda_a=\lambda_a^*} \\ & + \frac{1}{2}\sum_{a,b=1}^{K}(\lambda_a - \lambda_a^*)(\lambda_b - \lambda_b^*)\left.\frac{\partial^2\chi^2}{\partial\lambda_a\partial\lambda_b}\right|_{\substack{\lambda_a=\lambda_a^*\\ \lambda_b=\lambda_b^*}},\end{aligned} \tag{I.32}$$

which we can insert back into the expression for P, eq. (I.28), to yield

$$P(\boldsymbol{\lambda}) = W\exp\left(-\frac{1}{2}\sum_{a,b=1}^{K}\boldsymbol{H}_{ab}(\lambda_a - \lambda_a^*)(\lambda_b - \lambda_b^*)\right), \tag{I.33}$$

where W is a normalization constant and

$$H_{ab} = \left.\frac{\partial^2\chi^2}{\partial\lambda_a\partial\lambda_b}\right|_{\substack{\lambda_a=\lambda_a^*\\ \lambda_b=\lambda_b^*}} \tag{I.34}$$

is the Hessian matrix, and we used $\partial\chi^2/\partial\lambda_a|_{\lambda_a=\lambda_a^*} = 0$. From eq. (I.33) we find that

$$\Delta_{ab} \equiv \langle(\lambda_a^* - \overline{\overline{\lambda}}_a)(\lambda_b^* - \overline{\overline{\lambda}}_b)\rangle = (\boldsymbol{H})_{ab}^{-1}, \tag{I.35}$$

i.e., the inverse of the Hessian matrix gives the covariance of the estimated parameters.

I.D.1.2 *Linear fit functions*

For the case that the fit function is linear, i.e. $f_i = \lambda_1 t_i$, eq. (I.31) can be solved analytically (Press *et al.* [8]). The same can be done for the variance, σ^2, in the estimated parameter. We have

$$\lambda_1^* = \frac{\sum_i \overline{y}_i t_i/\overline{\sigma}_i^2}{\sum_i t_i^2/\overline{\sigma}_i^2} \tag{I.36a}$$

$$\sigma^2 = \Delta_{11} = \frac{1}{\sum_i t_i^2/\overline{\sigma}_i^2}. \tag{I.36b}$$

I.D.2 *Maximum likelihood fitting*

In this section we review the maximum likelihood (ML) fitting procedure (correlated χ^2 fitting) [9–12].

I.D.2.1 *General fit functions*

Where least squares fit only makes use of the diagonal (variance) of the covariance matrix, we will now make use of the full matrix, defined as in eq. (I.15), where the diagonal will be the square of the standard error of the mean, $s_i^2 = \sigma_i^2/M$. The task of fitting a function $f(t_i; \boldsymbol{\lambda})$, reduces to maximizing the probability which is taken as the multi-variate Gaussian:

$$P(\overline{\boldsymbol{y}}; \boldsymbol{\lambda}) = Z \exp\left(-\frac{1}{2}(\overline{\boldsymbol{y}} - \boldsymbol{f})^T \overline{\boldsymbol{C}}^{-1} (\overline{\boldsymbol{y}} - \boldsymbol{f})\right), \tag{I.37}$$

where $\overline{\boldsymbol{C}} = \overline{\boldsymbol{Q}}/M$ is as in eq. (I.3) in the main text, and the normalization constant $Z = 1/((2\pi)^{N/2}\sqrt{\det(\overline{\boldsymbol{C}})})$ [13], $\overline{\boldsymbol{y}} = (\overline{y}_1, \ldots, \overline{y}_N)$, $\boldsymbol{f} = (f_1, \ldots, f_N)$, with $f_i = f(t_i; \boldsymbol{\lambda})$, and $(\ldots)^T$ denotes transpose. For uncorrelated data the covariance matrix estimator, $\overline{\boldsymbol{C}}$, will be diagonal and eq. (I.37) reduces to eq. (I.28), and the LS method is attained.

As for LS, maximizing P is equivalent to minimizing the cost function

$$\chi^2 = (\overline{\boldsymbol{y}} - \boldsymbol{f})^T \overline{\boldsymbol{C}}^{-1} (\overline{\boldsymbol{y}} - \boldsymbol{f}). \tag{I.38}$$

Thus, we get our optimal parameters λ_a^* $(a = 1, \ldots, K)$ by solving:

$$\begin{aligned} \frac{1}{2}\frac{\partial \chi^2}{\partial \lambda_a}\bigg|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^*} = 0 = & -\frac{1}{2}\frac{\partial \boldsymbol{f}}{\partial \lambda_a}\bigg|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^*} \overline{\boldsymbol{C}}^{-1}(\overline{\boldsymbol{y}} - \boldsymbol{f}(\boldsymbol{\lambda}^*)) \\ & + (\overline{\boldsymbol{y}} - \boldsymbol{f}(\boldsymbol{\lambda}^*))\overline{\boldsymbol{C}}^{-1}\left(-\frac{1}{2}\frac{\partial \boldsymbol{f}}{\partial \lambda_a}\bigg|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^*}\right) \\ = & \frac{\partial \boldsymbol{f}}{\partial \lambda_a}\bigg|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^*} \overline{\boldsymbol{C}}^{-1}(\boldsymbol{f}(\boldsymbol{\lambda}^*) - \overline{\boldsymbol{y}}), \end{aligned} \tag{I.39}$$

where in the last step we used the symmetry property of $\overline{\boldsymbol{C}}$, i.e., that $\overline{C}_{ij} = \overline{C}_{ji}$.

The derivation of the covariance, Δ_{ab}, of the ML estimated parameters, λ_a^* follows along identical lines as for LS (previous section). Hence, Δ_{ab} is given by eq. (I.35) where λ_a^* is now obtained by solving eq. (I.39) (instead of solving eq. (I.31) as for LS).

I.D.2.2 *Linear fit functions*

For fitting a *linear* function, $f(t_i; \boldsymbol{\lambda}) = \lambda_1 t_i$, to data one can determine the minimum of the ML χ^2 function, eq. (I.38), analytically. In particular, such a fitting function is of relevance for Brownian motion (section I.B.1). Eq. (I.39) becomes

$$0 = \frac{1}{2}\frac{\partial \chi^2}{\partial \lambda_1}\bigg|_{\lambda_1=\lambda_1^*} = (\overline{\boldsymbol{y}} - \lambda_1^*\, \boldsymbol{t})^T \overline{\boldsymbol{C}}^{-1}\boldsymbol{t}. \tag{I.40}$$

Taking the second derivative we get

$$\frac{\partial^2 \chi^2}{\partial \lambda_1^2}\bigg|_{\lambda_1=\lambda_1^*} = -\boldsymbol{t}^T \overline{\boldsymbol{C}}^{-1}\boldsymbol{t}. \tag{I.41}$$

From these results, as well as using eq. (I.34) and eq. (I.35), we get the optimal value for the parameter $\lambda_1 = \lambda_1^*$ and its variance σ^2 as

$$\lambda_1^* = \frac{\overline{\boldsymbol{y}}^T \overline{\boldsymbol{C}}^{-1}\boldsymbol{t}}{\boldsymbol{t}^T \overline{\boldsymbol{C}}^{-1}\boldsymbol{t}} \tag{I.42a}$$

$$\sigma^2 = \Delta_{11} = \frac{1}{\boldsymbol{t}^T \overline{\boldsymbol{C}}^{-1}\boldsymbol{t}}. \tag{I.42b}$$

I.E THE CORRELATION-CORRECTED LEAST SQUARE METHOD

We here describe our new fitting procedure, the CLS method, in detail.

I.E.1 *Parameter estimation*

As demonstrated in the main text, the previous standard methods for fitting of ensemble averages, the LS or ML procedures (section I.D), are of limited general applicability for fitting of correlated data: the LS method assumes data points are independent resulting in flawed error estimation, whereas the ML method (involving inversion of a noisy sample covariance matrix) provides ill-conditioned results or strong bias in the parameter estimation.

To remedy this lack of available methods for robust and accurate fitting of ensemble averages (correlated data), we here formulate the problem at hand as a minimization of a "cost function", χ^2, which can be chosen rather general. Minimizing this distance function provides an estimate, $\boldsymbol{\lambda}^*$, for the model parameters of interest. However, unlike the LS (uncorrelated χ^2) fitting procedure, where fluctuations around mean values are assumed to be independent, we use the full multivariate probability density function for the mean values, eq. (I.37) (which is Gaussian due to the multivariate central limit theorem), when estimating the standard error and covariance in the fitted parameters. This provides a mathematically rigorous way of avoiding the problems with previous fitting methods.

The cost function used herein is a χ^2 functional (eq. (I.2) in the main text) on the form:

$$\chi^2 = (\overline{\boldsymbol{y}} - \boldsymbol{f})^T \boldsymbol{R}\, (\overline{\boldsymbol{y}} - \boldsymbol{f}), \tag{I.43}$$

where, $\overline{\boldsymbol{y}} = (\overline{y}_1, \ldots, \overline{y}_N)$, $\boldsymbol{f} = (f_1, \ldots, f_N)$, with $f_i = f(t_i; \boldsymbol{\lambda})$, and $(\ldots)^T$ denotes transpose. We find the best parameters, $\boldsymbol{\lambda}^*$ by minimizing χ^2, i.e., by solving:

$$\left.\frac{\partial \chi^2}{\partial \lambda_a}\right|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^*} = 0 = 2 \left.\frac{\partial f_i(\boldsymbol{\lambda})}{\partial \lambda_a}\right|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^*} R_{ij}(f_j(\boldsymbol{\lambda}^*) - \overline{y}_j), \tag{I.44}$$

where $a = 1, \ldots, K$. As before, a bar ($\overline{Z}$) denotes sample estimator and we use double bar ($\overline{\overline{Z}}$) to denote its true value. Note that the positive definite symmetric matrix $\boldsymbol{R}$ could potentially be custom made for particular applications; for all cases considered herein (see below), $\boldsymbol{R}$ is a sample estimator based on the data. In the main text the observables $\overline{y}_i$ are mean square displacements at different sampling times, T_i. We note, however, that our CLS procedure is valid for any type of ensemble averaged observables (the matrices $\overline{\boldsymbol{C}}$ and $\overline{\boldsymbol{Q}}$ below are then the covariance matrix for those particular observables).

For the matrix $\boldsymbol{R}$, we consider three choices:

1. *Maximum likelihood (ML):* Here we make use of the full covariance matrix, (see section I.D.2):

$$\boldsymbol{R} = \overline{\boldsymbol{R}}^{[ML]} = \overline{\boldsymbol{C}}^{-1}, \tag{I.45}$$

where $\overline{\boldsymbol{C}}$ is the covariance matrix of the mean, $\overline{\boldsymbol{C}} = \overline{\boldsymbol{Q}}/M$, as defined in eq. (I.3) of the main text.

2. *Correlation-corrected and standard least squares (CLS/LS):* Here we only make use of the diagonal elements,

$$R_{ij} = \overline{R}_{ij}^{[CLS]} = \delta_{ij}/\overline{C}_{ij}, \tag{I.46}$$

where δ_{ij} is the Kronecker delta ($\delta_{ij} = 1$, if $i = j$; $\delta_{ij} = 0$, if $i \neq j$).

3. *Brownian motion adapted least squares (BMALS):* Finally we probe our fitting method by the following choice:

$$\boldsymbol{R} = \overline{\boldsymbol{R}}^{[BMALS]} = \frac{1}{M}\overline{\overline{\boldsymbol{A}}}^{T}\ \overline{\boldsymbol{d}}\ \overline{\overline{\boldsymbol{A}}}, \tag{I.47}$$

where $\overline{\overline{\boldsymbol{A}}}^{T}\overline{\overline{\boldsymbol{A}}} = \overline{\overline{\boldsymbol{Q}}}^{-1}$ gives the exact covariance matrix for observables of interest. When these observables are the squared displacements at different sampling times (case of interest in the main text), the quantity $\overline{\boldsymbol{d}}$ is a diagonal matrix with elements

$$\overline{d}_{ii} = \frac{1}{\frac{1}{M-1}\sum_m^M (\sum_j^N A_{ij}(y_j^{(m)} - \overline{y}_j))^2}. \tag{I.48}$$

For Brownian motion, the matrix $\overline{\boldsymbol{A}}$ is given in eq. (I.18). For comparison of the BMALS method to CLS, please see Supplementary Figure I.S6.

I.E.2 *Error estimation*

The covariance for the estimated parameters (i.e., the parameters $\boldsymbol{\lambda}^*$ obtained by solving eq. (I.44)) is defined

$$\Delta_{ab} = \langle (\lambda_a^* - \overline{\overline{\lambda}}_a)(\lambda_b^* - \overline{\overline{\lambda}}_b) \rangle, \tag{I.49}$$

where $\langle F(\overline{\boldsymbol{y}}) \rangle = \int F(\overline{\boldsymbol{y}}) \rho(\overline{\boldsymbol{y}}; \boldsymbol{\lambda}) d\overline{y}_1 d\overline{y}_2 \cdots d\overline{y}_N$ denotes an average over the multivariate probability density, $\rho(\overline{\boldsymbol{y}}; \boldsymbol{\lambda})$. Due to the multivariate central limit theorem (note that $\overline{\boldsymbol{y}}$ is a sum of M identically distributed

random numbers), for large M this probability density is a multi-variate Gaussian (compare to section I.D.2):

$$\rho(\overline{\boldsymbol{y}};\boldsymbol{\lambda}) = Z \exp\left(-\frac{1}{2}(\overline{\boldsymbol{y}}-\overline{\overline{\boldsymbol{y}}})^T \overline{\overline{\boldsymbol{C}}}^{-1}(\overline{\boldsymbol{y}}-\overline{\overline{\boldsymbol{y}}})\right), \tag{I.50}$$

with normalization constant $Z = 1/((2\pi)^{N/2}\sqrt{\det(\overline{\overline{\boldsymbol{C}}})})$ [13].

In order to derive an explicit expression for Δ_{ab} we follow the lines of thought of Gottlieb *et al.* [10] and make a first order Taylor series expansion of the estimated parameter values in terms of deviations of the estimated $\overline{\boldsymbol{y}}$ from their true values:

$$\lambda_a^* - \overline{\overline{\lambda}}_a = \left.\frac{\partial \lambda_a^*}{\partial \overline{y}_k}\right|_{\overline{y}_k=\overline{\overline{y}}_k} (\overline{y}_k - \overline{\overline{y}}_k) \; + \; \mathcal{O}[(\overline{y}_k-\overline{\overline{y}}_k)(\overline{y}_l-\overline{\overline{y}}_l)], \tag{I.51}$$

where repeated indices are summed over. Substituting this expression into eq. (I.49) and using the definition of the covariance matrix: $\overline{\overline{C}}_{kl} = \langle(\overline{y}_k-\overline{\overline{y}}_k)(\overline{y}_l-\overline{\overline{y}}_l)\rangle$ we find, to first order,

$$\Delta_{ab} = \left.\frac{\partial \lambda_a^*}{\partial \overline{y}_k}\right|_{\overline{y}_k=\overline{\overline{y}}_k} \overline{\overline{C}}_{kl} \left.\frac{\partial \lambda_b^*}{\partial \overline{y}_l}\right|_{\overline{y}_l=\overline{\overline{y}}_l}. \tag{I.52}$$

In order to obtain an explicit expression for $\partial\lambda_a^*/\partial\overline{y}_k$ we differentiate eq. (I.44) with respect to $\overline{y}_i$. This yields

$$0 = h_{ab}\frac{\partial \lambda_b^*}{\partial \overline{y}_i} - 2\left.\frac{\partial f_j(\boldsymbol{\lambda})}{\partial \lambda_a}\right|_{\lambda_a=\lambda_a^*} R_{jk}\delta_{ki}, \tag{I.53}$$

where we introduced

$$\begin{aligned} h_{ab} = {} & 2\left.\frac{\partial^2 f_i(\boldsymbol{\lambda})}{\partial\lambda_a\partial\lambda_b}\right|_{\substack{\lambda_a=\lambda_a^*,\\ \lambda_b=\lambda_b^*}} R_{ij}(f_j(\boldsymbol{\lambda})-\overline{y}_j) \\ & + 2\left.\frac{\partial f_i(\boldsymbol{\lambda})}{\partial\lambda_a}\right|_{\lambda_a=\lambda_a^*} R_{ij}\left.\frac{\partial f_j(\boldsymbol{\lambda})}{\partial\lambda_b}\right|_{\lambda_b=\lambda_b^*}. \end{aligned} \tag{I.54}$$

Solving eq. (I.53) we obtain:

$$\frac{\partial \lambda_b^*}{\partial \overline{y}_i} = 2(\overline{\boldsymbol{h}}^{-1})_{ab}\left.\frac{\partial f_j(\boldsymbol{\lambda})}{\partial \lambda_a}\right|_{\lambda_a=\lambda_a^*} R_{ji}, \tag{I.55}$$

which substituted into eq. (I.52) yields the following expression for the covariance of the estimated parameter, $\boldsymbol{\lambda}^*$:

$$\Delta_{ab} = \left(4(\boldsymbol{h}^{-1})_{ac} \left.\frac{\partial f_j(\boldsymbol{\lambda})}{\partial \lambda_c}\right|_{\lambda_c=\lambda_c^*} R_{jk} \overline{\overline{C}}_{kl} R_{lm} \left.\frac{\partial f_m(\boldsymbol{\lambda})}{\partial \lambda_d}\right|_{\lambda_d=\lambda_d^*} (\boldsymbol{h}^{-1})_{db} \right)_{\substack{\overline{y}_k=\overline{\overline{y}}_k \\ \overline{y}_l=\overline{\overline{y}}_l}}. \tag{I.56}$$

We finally replace all exact quantities above by the corresponding sample estimators (and use $\boldsymbol{C} = \boldsymbol{Q}/M$), giving the key result, eq. (I.4) in the main text. The replacement of exact ensemble averages by sample estimates introduces bias terms which, to first order, are proportional to $1/M$, where M is the number of trajectories, see section I.F.1. For CLS/LS procedures, we find that the bias is in practice often negligible (see main text). Just as the parameter estimates λ_a^* are typically biased, so will the error estimate Δ_{ab} in eq. (I.4) also be, as it is a nonlinear function of sample estimates, see section I.F.1. This bias in Δ_{ab} can be reduced using the jackknife procedure (see section I.G).

I.F BIAS EFFECTS IN PARAMETER ESTIMATION FOR BROWNIAN MOTION

In this section, we provide analytical expressions for the bias in parameter (diffusion constant) estimation for Brownian motion. We find that the ML method has a bias which increases strongly with the number of sampling times, N. In contrast, the LS method provides a (small) bias which is independent of N for large N.

I.F.1 *The origin of bias*

We generally expect that the bias, i.e., the expected difference between some observable, based on sample estimates and the "true" value of that observable can be written as a series expansion in terms of $1/M$, where M is the number of trajectories [14]. To understand why this is so, in the present context, we recall that any sample estimate, $\overline{Q}_{ijk....}$ (where i, j, k etc. labels sampling times), is an average (normalized sum) over the M trajectories. The multivariate central limit theorem tells us that we can, for such averages, write $\overline{Q}_{ijk...} = \overline{\overline{Q}}_{ijk....} + \gamma_{ijk...}/\sqrt{M}$, where $\gamma_{ijk...}$ is a zero-mean "noise". Therefore any observable, O, which is a function of one, or several, sample estimates (the optimal parameters $\boldsymbol{\lambda}^*$ and their associated covariance matrix $\boldsymbol{\Delta}$, see previous sections, are

examples of such observables) will (schematically) have a Taylor series expansion of the form:

$$O = \overline{\overline{O}} + \sum_{k=1}^{\infty} \frac{A_k}{\sqrt{M} M^{k-1}} + \sum_{k=1}^{\infty} \frac{B_k}{M^k} \tag{I.57}$$

for large M. The first term in the Taylor expansion is the sought quantity, $\overline{\overline{O}}$. Considering the remaining terms, we note that, by construction, we have that $\langle A_1 \rangle = 0$, and hence the first non-zero term of the expectation value of the expression above is $\langle B_1 \rangle / M \propto 1/M$. For the case that the observable, O, is function of more than one *independent* sample estimates, then we have $\langle A_k \rangle = 0$ for all k. However, note that if O is a function of several sample estimates which are *dependent*, then in general $\langle A_k \rangle \neq 0$ for $k \geq 2$. We can safely remove the first bias-term with a jackknife procedure [15], see section I.G. Also higher order bias terms can be removed formally. However, already at the second order bias reduction level computational costs becomes considerable.

I.F.2 *Bias in parameter estimation of* ML *for Brownian motion*

Consider equations (I.43) and (I.45). We write the sample estimator of the covariance matrix eq. (I.15), and the exact, $\overline{\overline{\boldsymbol{Q}}}$, as related by

$$\overline{\boldsymbol{Q}}_{ij} = \overline{\overline{\boldsymbol{Q}}}_{ij} + \eta_{ij}, \tag{I.58}$$

where $\boldsymbol{\eta}$ represents their deviation. We seek the "noise" in the inverse, $(\overline{\boldsymbol{Q}}^{-1})_{ij}$. Using the normalization condition, and writing

$$(\overline{\boldsymbol{Q}}^{-1})_{ij} = (\overline{\overline{\boldsymbol{Q}}}^{-1})_{ij} + \xi_{ij}, \tag{I.59}$$

we get

$$\begin{aligned} \boldsymbol{I} &= \overline{\boldsymbol{Q}}\ \overline{\boldsymbol{Q}}^{-1} = (\overline{\overline{\boldsymbol{Q}}} + \boldsymbol{\eta})(\overline{\overline{\boldsymbol{Q}}}^{-1} + \boldsymbol{\xi}) \\ &= \boldsymbol{I} + \boldsymbol{\eta}\overline{\overline{\boldsymbol{Q}}}^{-1} + \ \overline{\overline{\boldsymbol{Q}}}\boldsymbol{\xi} + \boldsymbol{\eta}\boldsymbol{\xi}. \end{aligned} \tag{I.60}$$

Thus, to first order $\boldsymbol{\eta}\overline{\overline{\boldsymbol{Q}}}^{-1} + \ \overline{\overline{\boldsymbol{Q}}}\boldsymbol{\xi} = 0$, and by definition $\boldsymbol{\eta} = \overline{\boldsymbol{Q}} - \overline{\overline{\boldsymbol{Q}}}$:

$$\boldsymbol{\xi} = \overline{\overline{\boldsymbol{Q}}}^{-1} - \ \overline{\overline{\boldsymbol{Q}}}^{-1}\overline{\boldsymbol{Q}}\ \overline{\overline{\boldsymbol{Q}}}^{-1}. \tag{I.61}$$

Using eq. (I.59) in eq. (I.43) and eq. (I.45) yields

$$\begin{aligned}\lambda^* = \frac{\overline{\boldsymbol{y}}^T(\overline{\overline{\boldsymbol{Q}}}^{-1}+\boldsymbol{\xi})\boldsymbol{t}}{\boldsymbol{t}^T(\overline{\overline{\boldsymbol{Q}}}^{-1}+\boldsymbol{\xi})\boldsymbol{t}} &= \frac{\overline{\boldsymbol{y}}^T\overline{\overline{\boldsymbol{Q}}}^{-1}\boldsymbol{t}}{\boldsymbol{t}^T\overline{\overline{\boldsymbol{Q}}}^{-1}\boldsymbol{t}\left(1+\frac{\boldsymbol{t}^T\boldsymbol{\xi}\boldsymbol{t}}{\boldsymbol{t}\overline{\overline{\boldsymbol{Q}}}^{-1}\boldsymbol{t}}\right)} \\ &\quad + \frac{\overline{\boldsymbol{y}}^T\boldsymbol{\xi}\boldsymbol{t}}{\boldsymbol{t}^T\overline{\overline{\boldsymbol{Q}}}^{-1}\boldsymbol{t}\left(1+\frac{\boldsymbol{t}^T\boldsymbol{\xi}\boldsymbol{t}}{\boldsymbol{t}\overline{\overline{\boldsymbol{Q}}}^{-1}\boldsymbol{t}}\right)} \\ &\approx \frac{1}{\boldsymbol{t}^T\overline{\overline{\boldsymbol{Q}}}^{-1}\boldsymbol{t}}\left(\overline{\boldsymbol{y}}^T\overline{\overline{\boldsymbol{Q}}}^{-1}\boldsymbol{t}+\overline{\boldsymbol{y}}^T\boldsymbol{\xi}\boldsymbol{t}-\frac{\overline{\boldsymbol{y}}^T\overline{\overline{\boldsymbol{Q}}}^{-1}\boldsymbol{t}}{\boldsymbol{t}\overline{\overline{\boldsymbol{Q}}}^{-1}\boldsymbol{t}}\boldsymbol{t}\boldsymbol{\xi}\boldsymbol{t}\right),\end{aligned} \tag{I.62}$$

where we did a series expansion to first order in $\boldsymbol{\xi}$. Using eq. (I.61) we get

$$\begin{aligned}\lambda^* &= \frac{\overline{\boldsymbol{y}}^T\overline{\overline{\boldsymbol{Q}}}^{-1}\boldsymbol{t}}{\boldsymbol{t}^T\overline{\overline{\boldsymbol{Q}}}^{-1}\boldsymbol{t}} + \frac{\overline{\boldsymbol{y}}^T(\overline{\overline{\boldsymbol{Q}}}^{-1}-\overline{\overline{\boldsymbol{Q}}}^{-1}\,\overline{\boldsymbol{Q}}\,\overline{\overline{\boldsymbol{Q}}}^{-1})\boldsymbol{t}}{\boldsymbol{t}^T\overline{\overline{\boldsymbol{Q}}}^{-1}\boldsymbol{t}} \\ &\quad - \frac{\overline{\boldsymbol{y}}^T\overline{\overline{\boldsymbol{Q}}}^{-1}\boldsymbol{t}}{(\boldsymbol{t}^T\overline{\overline{\boldsymbol{Q}}}^{-1}\boldsymbol{t})^2}\left(\boldsymbol{t}^T\overline{\overline{\boldsymbol{Q}}}^{-1}\boldsymbol{t}-\boldsymbol{t}^T\overline{\overline{\boldsymbol{Q}}}^{-1}\overline{\boldsymbol{Q}}\,\overline{\overline{\boldsymbol{Q}}}^{-1}\boldsymbol{t}\right) \\ &= \frac{\overline{\boldsymbol{y}}^T\overline{\overline{\boldsymbol{Q}}}^{-1}\boldsymbol{t}}{\boldsymbol{t}^T\overline{\overline{\boldsymbol{Q}}}^{-1}\boldsymbol{t}} \underbrace{-\frac{\overline{\boldsymbol{y}}^T\overline{\overline{\boldsymbol{Q}}}^{-1}\overline{\boldsymbol{Q}}\,\overline{\overline{\boldsymbol{Q}}}^{-1}\boldsymbol{t}}{\boldsymbol{t}^T\overline{\overline{\boldsymbol{Q}}}^{-1}\boldsymbol{t}} + \frac{\overline{\boldsymbol{y}}^T\overline{\overline{\boldsymbol{Q}}}^{-1}\boldsymbol{t}}{(\boldsymbol{t}^T\overline{\overline{\boldsymbol{Q}}}^{-1}\boldsymbol{t})^2}\boldsymbol{t}^T\overline{\overline{\boldsymbol{Q}}}^{-1}\overline{\boldsymbol{Q}}\,\overline{\overline{\boldsymbol{Q}}}^{-1}\boldsymbol{t}}_{\text{bias}=B}.\end{aligned} \tag{I.63}$$

Note that the expectation value of the first term on the right hand side evaluates to $\overline{\overline{\lambda}}$, hence the additional terms yield the bias, whose expectation value, $\langle B\rangle$, we now seek. We write eq. (I.63) on component form (repeated indices are summed over, as before). We have

$$B_1 = -\frac{\overline{y}_k(\overline{\overline{\boldsymbol{Q}}}^{-1})_{ki}\overline{Q}_{ij}(\overline{\overline{\boldsymbol{Q}}}^{-1})_{jl}t_l}{\boldsymbol{t}^T\overline{\overline{\boldsymbol{Q}}}^{-1}\boldsymbol{t}} \tag{I.64a}$$

$$B_2 = \frac{\overline{y}_i(\overline{\overline{\boldsymbol{Q}}}^{-1})_{ik}t_k t_j(\overline{\overline{\boldsymbol{Q}}}^{-1})_{jm}\overline{Q}_{ml}(\overline{\overline{\boldsymbol{Q}}}^{-1})_{ln}t_n}{(\boldsymbol{t}^T\overline{\overline{\boldsymbol{Q}}}^{-1}\boldsymbol{t})^2} \tag{I.64b}$$

(the component form of the quantity appearing in the denominators above is $\boldsymbol{t}^T\overline{\overline{\boldsymbol{Q}}}^{-1}\boldsymbol{t} = t_p(\overline{\overline{\boldsymbol{Q}}}^{-1})_{pq}t_q$). Consider first the expectation value of B_1. To that end we seek the quantity ($a, b, c\ldots$ label trajectories):

$$\langle\overline{y}_k\overline{Q}_{ij}\rangle = \frac{1}{M(M-1)}\Big\langle\sum_{a=1}^{M}y_k^{(a)}\left[\sum_{b=1}^{M}y_i^{(b)}y_j^{(b)} - \frac{1}{M}\sum_{b=1}^{M}y_i^{(b)}\sum_{c=1}^{M}y_j^{(c)}\right]\Big\rangle. \tag{I.65}$$

We also have:

$$\langle y_k^{(a)} \rangle = \langle \left[x_k^{(a)} - x^{(a)}(0) \right]^2 \rangle = \overline{\overline{\sigma}}_k^2 = \overline{\overline{V}}_{kk}, \tag{I.66}$$

where we in the last step used eq. (I.9). Also $\langle x_i^{(a)} - x^{(a)}(0) \rangle = 0$, and since different realizations (trajectories) are independent we have

$$\langle x_i^{(a)} x_j^{(b)} \rangle = \delta_{ab} \overline{\overline{V}}_{ij}. \tag{I.67}$$

Higher order terms can be calculated using Wick's theorem, eq. (I.12) (for large i, $x_i^{(a)}$ is a sum of many small increments, from the central limit theorem it follows that $x_i^{(a)}$ are Gaussian). We have

$$\begin{aligned} \langle y_i^{(a)} y_j^{(b)} \rangle &= \langle (x_i^{(a)})^2 (x_j^{(b)})^2 \rangle = \langle x_i^{(a)} x_i^{(a)} x_j^{(b)} x_j^{(b)} \rangle \\ &= \langle x_i^{(a)} x_i^{(a)} \rangle \langle x_j^{(b)} x_j^{(b)} \rangle + \langle x_i^{(a)} x_j^{(b)} \rangle \langle x_i^{(a)} x_j^{(b)} \rangle \\ &\quad + \langle x_i^{(a)} x_j^{(b)} \rangle \langle x_i^{(a)} x_j^{(b)} \rangle \\ &= \overline{\overline{\sigma}}_i^2 \overline{\overline{\sigma}}_j^2 + 2 \overline{\overline{V}}_{ij}^2 \delta_{ab}. \end{aligned} \tag{I.68}$$

Now, in the same way for higher order terms, we get

$$\begin{aligned} \langle y_k^{(a)} y_i^{(b)} y_j^{(c)} \rangle &= \langle x_k^{(a)} x_k^{(a)} x_i^{(b)} x_i^{(b)} x_j^{(c)} x_j^{(c)} \rangle \\ &= [\text{tedious enumeration of all cases}] = \\ &= \overline{\overline{\sigma}}_k^2 \overline{\overline{\sigma}}_i^2 \overline{\overline{\sigma}}_j^2 + 2 \overline{\overline{\sigma}}_k^2 \overline{\overline{V}}_{ij}^2 \delta_{bc} + 2 \overline{\overline{\sigma}}_j^2 \overline{\overline{V}}_{ki}^2 \delta_{ab} \\ &\quad + 2 \overline{\overline{\sigma}}_i^2 \overline{\overline{V}}_{kj}^2 \delta_{ac} + 8 \overline{\overline{V}}_{ki}^2 \overline{\overline{V}}_{kj}^2 \overline{\overline{V}}_{ij}^2 \delta_{ab} \delta_{bc} \delta_{ac}, \end{aligned} \tag{I.69}$$

(no sum over repeated indices). Eq. (I.65) now becomes

$$\begin{aligned} \langle \overline{y}_k \overline{Q}_{ij} \rangle = &\frac{1}{M(M-1)} \underbrace{\sum_{a=1}^{M} \sum_{b=1}^{M} \langle y_k^{(a)} y_i^{(b)} y_j^{(b)} \rangle}_{U_1} \\ &- \frac{1}{M^2(M-1)} \underbrace{\sum_{a,b,c} \langle y_k^{(a)} y_i^{(b)} y_j^{(c)} \rangle}_{U_2}. \end{aligned} \tag{I.70}$$

Using eq. (I.69) we get:

$$\begin{aligned} U_1 = \sum_{a,b} \langle y_k^{(a)} y_i^{(b)} y_j^{(b)} \rangle &= M^2 \overline{\overline{\sigma}}_k^2 \overline{\overline{\sigma}}_i^2 \overline{\overline{\sigma}}_j^2 + 8M \overline{\overline{V}}_{ki} \overline{\overline{V}}_{ij} \overline{\overline{V}}_{kj} \\ &+ 2M^2 \overline{\overline{\sigma}}_k^2 \overline{\overline{V}}_{ij}^2 + 2M \overline{\overline{\sigma}}_j^2 \overline{\overline{V}}_{ki}^2 + 2M \overline{\overline{\sigma}}_i^2 \overline{\overline{V}}_{kj}^2 \end{aligned} \tag{I.71a}$$

$$U_2 = \sum_{a,b,c} \langle y_k^{(a)} y_i^{(b)} y_j^{(c)} \rangle = M^3 \overline{\overline{\sigma}}_k^2 \overline{\overline{\sigma}}_i^2 \overline{\overline{\sigma}}_j^2 + 8M \overline{\overline{V}}_{ki} \overline{\overline{V}}_{ij} \overline{\overline{V}}_{kj} + 2M^2 \left[\overline{\overline{\sigma}}_k^2 \overline{\overline{V}}_{ij}^2 + \overline{\overline{\sigma}}_j^2 \overline{\overline{V}}_{ki}^2 + \overline{\overline{\sigma}}_i^2 \overline{\overline{V}}_{kj}^2 \right]. \tag{I.71b}$$

Combining eq. (I.71) with eq. (I.70) results in:

$$\begin{aligned} \langle \overline{y}_k \overline{Q}_{ij} \rangle &= \frac{1}{M(M-1)} \left[(2M^2 - 2M) \overline{\overline{\sigma}}_k^2 \overline{\overline{V}}_{ij}^2 + (8M-8) \overline{\overline{V}}_{ki} \overline{\overline{V}}_{ij} \overline{\overline{V}}_{kj} \right] \\ &= \overline{\overline{\sigma}}_k^2 2 \overline{\overline{V}}_{ij}^2 + \frac{8}{M} \overline{\overline{V}}_{ki} \overline{\overline{V}}_{ij} \overline{\overline{V}}_{kj}. \end{aligned} \tag{I.72}$$

Using eq. (I.72) in eq. (I.64a) we find

$$\langle B_1 \rangle = - \frac{\overline{\overline{\sigma}}_k^2 (\overline{\overline{\boldsymbol{Q}}}^{-1})_{ki} \delta_{il} t_l + \frac{8}{M} \overline{\overline{V}}_{ki} \overline{\overline{V}}_{ij} \overline{\overline{V}}_{kj} (\overline{\overline{\boldsymbol{Q}}}^{-1})_{ki} (\overline{\overline{\boldsymbol{Q}}}^{-1})_{jl} t_l}{\boldsymbol{t}^T \overline{\overline{\boldsymbol{Q}}}^{-1} \boldsymbol{t}}, \tag{I.73}$$

where we used that $\overline{\overline{Q}}_{ij} (\overline{\overline{\boldsymbol{Q}}}^{-1})_{jl} = \delta_{il}$. Now consider B_2, eq. (I.64b). We write eq. (I.72) according to (also see eq. (I.13))

$$\langle \overline{y}_i \overline{Q}_{ml} \rangle = \overline{\overline{\sigma}}_i^2 \overline{\overline{Q}}_{ml} + \frac{8}{M} \overline{\overline{V}}_{im} \overline{\overline{V}}_{ml} \overline{\overline{V}}_{li}. \tag{I.74}$$

Eq. (I.64b) now becomes

$$\begin{aligned} \langle B_2 \rangle &= \frac{(\overline{\overline{\boldsymbol{Q}}}^{-1})_{ik} t_k t_j (\overline{\overline{\boldsymbol{Q}}}^{-1})_{jm} \left[\overline{\overline{\sigma}}_i^2 \overline{\overline{Q}}_{ml} + \frac{8}{M} \overline{\overline{V}}_{im} \overline{\overline{V}}_{ml} \overline{\overline{V}}_{li} \right] (\overline{\overline{\boldsymbol{Q}}}^{-1})_{ln} t_n}{(\boldsymbol{t}^T \overline{\overline{\boldsymbol{Q}}}^{-1} \boldsymbol{t})^2} \\ &= \frac{\overline{\overline{\sigma}}_i^2 (\overline{\overline{\boldsymbol{Q}}}^{-1})_{ik} t_k}{\boldsymbol{t}^T \overline{\overline{\boldsymbol{Q}}}^{-1} \boldsymbol{t}} + \frac{8}{M} \frac{(\overline{\overline{\boldsymbol{Q}}}^{-1})_{ik} t_k t_j (\overline{\overline{\boldsymbol{Q}}}^{-1})_{jm} \overline{\overline{V}}_{im} \overline{\overline{V}}_{ml} \overline{\overline{V}}_{li} (\overline{\overline{\boldsymbol{Q}}}^{-1})_{ln} t_n}{(\boldsymbol{t}^T \overline{\overline{\boldsymbol{Q}}}^{-1} \boldsymbol{t})^2}. \end{aligned} \tag{I.75}$$

Combining B_1 and B_2 we arrive at an expression for the predicted first order bias (eq. (I.63)) for the suggested matrix, $\overline{\boldsymbol{R}}^{[ML]}$; (notice the cancellations of the first terms):

$$\begin{aligned} \langle B \rangle = \frac{1}{M} \frac{8}{\boldsymbol{t}^T \overline{\overline{\boldsymbol{Q}}}^{-1} \boldsymbol{t}} \Bigg(& \frac{(\overline{\overline{\boldsymbol{Q}}}^{-1})_{ik} t_k t_j (\overline{\overline{\boldsymbol{Q}}}^{-1})_{jm} \overline{\overline{V}}_{im} \overline{\overline{V}}_{ml} \overline{\overline{V}}_{li} (\overline{\overline{\boldsymbol{Q}}}^{-1})_{ln} t_n}{\boldsymbol{t}^T \overline{\overline{\boldsymbol{Q}}}^{-1} \boldsymbol{t}} \\ & - \overline{\overline{V}}_{ki} \overline{\overline{V}}_{ij} \overline{\overline{V}}_{jk} (\overline{\overline{\boldsymbol{Q}}}^{-1})_{ki} (\overline{\overline{\boldsymbol{Q}}}^{-1})_{jl} t_l \Bigg), \end{aligned} \tag{I.76}$$

which can be analytically evaluated. With this in mind we use eq. (I.9), with $t_i = i\epsilon$, and eq. (I.17), in eq. (I.76). When evaluating the associated sums over repeated indices in eq. (I.76), one uses:

$$\min(i,j) = \begin{cases} i, & \text{if } i \leq j \\ j, & \text{if } i > j \end{cases} \tag{I.77}$$

and then splits the sums accordingly. This splitting leads to sums on the form

$$I(m,p) = \sum_k \frac{k^m}{(2k-1)^p}, \tag{I.78}$$

where m and p are positive integers. These sums are rewritten according to

$$\begin{aligned} I(m,p) &= \frac{1}{2^m} \sum_k \frac{1}{(2k-1)^p} \left((2k-1)+1\right)^m \\ &= \frac{1}{2^m} \sum_{q=1}^{m} \binom{m}{q} \sum_k (2k-1)^{q-p}, \end{aligned} \tag{I.79}$$

where we used the binomial theorem. The full calculation is tedious but straightforward. The final result is:

$$\langle B \rangle = \frac{D}{M} G(N) \tag{I.80a}$$

$$G(N) = -\frac{a}{d} + \frac{b}{d^2} \tag{I.80b}$$

$$a = \frac{N}{2} + s_1 - \frac{s_2}{2} \tag{I.80c}$$

$$b = \frac{1}{16}(3s_1 - s_3) \tag{I.80d}$$

$$d = \frac{s_1}{8} \tag{I.80e}$$

$$s_n = \sum_{k=1}^{N} \frac{1}{(2k-1)^n}. \tag{I.80f}$$

I.F.2.1 *Asymptotic expansion*

Let us now investigate eq. (I.80) for large N. To that end, we write s_n, defined above, according to

$$s_n = \sum_{k=1}^{N} \left(\frac{1}{(2k-1)^n} + \frac{1}{(2k)^n} - \frac{1}{(2k)^n} \right) = \sum_{k=1}^{2N} k^{-n} - \frac{1}{2^n} \sum_{k=1}^{N} k^{-n}.$$

(I.81)

In eq. (I.80), there are three sums, s_1, s_2 and s_3. Out of these sums, s_1 decays most slowly with N and hence this sum is the only one which needs to be kept for large N. From eq. (0.131) in Gradshteyn *et al.* [16] we have

$$\sum_{k=1}^{N} \frac{1}{k} = \gamma + \ln N + \frac{1}{2N} + \mathcal{O}(\frac{1}{N^2}), \tag{I.82}$$

where $\gamma \approx 0.5772$ is the Euler-Mascheroni constant. Combining the result above with eq. (I.81) and eq. (I.80) we arrive at the asymptotic expression

$$G(N) \approx -\frac{8N}{\ln N + \gamma + 2\ln 2}, \tag{I.83}$$

where we used $\ln ab = \ln a + \ln b$. For large N, eq. (I.83) is a good approximation compared to the exact bias, eq. (I.80), see Supplementary Figure I.S3.

I.F.3 *Bias in parameter estimation of* CLS *and* LS *for Brownian motion*

Let us now consider the second case, eq. (I.46), of choosing $\boldsymbol{R}$. According to eqs. (I.43) and (I.46) we have the following:

$$\lambda^* = \frac{\overline{\boldsymbol{y}}^T \overline{\boldsymbol{Q}}_{\text{new}}^{-1} \boldsymbol{t}}{\boldsymbol{t}^T \overline{\boldsymbol{Q}}_{\text{new}}^{-1} \boldsymbol{t}}, \tag{I.84}$$

where

$$\overline{Q}_{\text{new},ij} = \overline{Q}_{ij}\delta_{ij} \tag{I.85}$$

$$\overline{\overline{Q}}_{\text{new},ij} = \overline{\overline{Q}}_{ij}\delta_{ij} \tag{I.86}$$

$$(\overline{\boldsymbol{Q}}_{\text{new}}^{-1})_{ij} = \delta_{ij}/\overline{Q}_{ij} \tag{I.87}$$

$$(\overline{\overline{\boldsymbol{Q}}}_{\text{new}}^{-1})_{ij} = \delta_{ij}/\overline{\overline{Q}}_{ij}. \tag{I.88}$$

The calculation starting from eq. (I.61) to eq. (I.63) is identical to before, just replace $\overline{\boldsymbol{Q}}$ with $\overline{\boldsymbol{Q}}_{\text{new}}$, and same for exact (double bar). Since our new matrices are diagonal, eq. (I.64) becomes (we here reintroduce explicit sums for the sake of clarity)

$$B_1 = -\frac{\sum_k \overline{y}_k \frac{1}{(\overline{\overline{Q}}_{kk})^2} \overline{Q}_{kk} t_k}{\sum_q t_q^2 / \overline{\overline{Q}}_{qq}} \tag{I.89a}$$

$$B_2 = \frac{\sum_{j,k} \overline{y}_k \frac{1}{\overline{\overline{Q}}_{kk}} t_k \cdot t_j^2 \frac{1}{(\overline{\overline{Q}}_{jj})^2} \overline{Q}_{jj}}{\left(\sum_q t_q^2 / \overline{\overline{Q}}_{qq}\right)^2}. \tag{I.89b}$$

Also the calculation from eq. (I.65) which leads up to eq. (I.72) is identical. From eq. (I.89) we see that we need

$$\langle \overline{y}_k \overline{Q}_{jj} \rangle = 2\overline{\overline{\sigma}}_k^2 \overline{\overline{V}}_{jj}^2 + \frac{8}{M} \overline{\overline{V}}_{kj}^2 \overline{\overline{V}}_{jj} \qquad (j,k \text{ fixed}), \tag{I.90a}$$

$$\langle \overline{y}_k \overline{Q}_{kk} \rangle = 2\overline{\overline{\sigma}}_k^6 + \frac{8}{M} \overline{\overline{\sigma}}_k^6 \qquad (k \text{ fixed}). \tag{I.90b}$$

Substituting eq. (I.90b) into eq. (I.89a), and using eq. (I.13) $\overline{\overline{Q}}_{kk} = 2\overline{\overline{V}}_{kk}^2 = 2\overline{\overline{\sigma}}_k^4$, and $\overline{\overline{\sigma}}_k^2 = 2Dt_k$ we get (with sums explicitly written)

$$\begin{aligned} \langle B_1 \rangle &= -\frac{\sum_k \frac{1}{(\overline{\overline{Q}}_{kk})^2} \left(2\overline{\overline{\sigma}}_k^6 + \frac{8}{M}\overline{\overline{\sigma}}_k^6\right) t_k}{\sum_q t_q^2 / \overline{\overline{Q}}_{qq}} = -\frac{(1+\frac{4}{M})\sum_k 1/2D}{\sum_k 1/(2D)^2} \\ &= -2D\left(1+\frac{4}{M}\right). \end{aligned} \tag{I.91}$$

In much the same way, we insert eq. (I.90a) into eq. (I.89b)

$$\begin{aligned} \langle B_2 \rangle &= \frac{\sum_{j,k} \left(2\overline{\overline{\sigma}}_k^2 \overline{\overline{\sigma}}_j^4 + \frac{8}{M}\overline{\overline{\sigma}}_j^2 \overline{\overline{V}}_{kj}^2\right) \frac{1}{2\overline{\overline{\sigma}}_k^4} \frac{t_k t_j^2}{4\overline{\overline{\sigma}}_j^8}}{\left(\sum_q t_q^2 / 2\overline{\overline{\sigma}}_q^4\right)^2} \\ &= \frac{\sum_{j,k} \left(\frac{1}{4}\frac{1}{(2D)^3} + \frac{1}{M}\frac{1}{(2D)^5}\frac{\overline{\overline{V}}_{kj}^2}{t_k t_j}\right)}{1/64D^4 \left(\sum_k 1\right)^2} \\ &= 2D + \frac{2}{MD}\frac{1}{N^2} \underbrace{\sum_j \sum_k \frac{\overline{\overline{V}}_{kj}^2}{t_k t_j}}_{I}. \end{aligned} \tag{I.92}$$

Consider the double sum, I, in eq. (I.92). We have time step $t_j = \epsilon j$ and separate the sums into $j = k$ and $j \neq k$, which gives $\overline{\overline{V}}_{ij} = 2D\epsilon \min(i,j)$

$$\begin{aligned} I &= 4D^2 \sum_{k=1}^N \sum_{j=1}^N \frac{[\min(i,j)]^2}{jk} = 4D^2 \left(\sum_{k=1}^N 1 + 2\sum_{k=1}^N \sum_{j=1}^{k-1} \frac{[\min(i,j)]^2}{jk}\right) \\ &= 4D^2 \left(N + 2\sum_{k=1}^N \frac{1}{k} \sum_{j=1}^{k-1} j\right) = 4D^2 \left(N + 2\sum_{k=1}^N \frac{1}{k}\frac{k(k-1)}{2}\right) \end{aligned}$$

$$= 4D^2 \sum_{k=1}^{N} k = 2D^2 N(N+1), \tag{I.93}$$

which inserted in eq. (I.92) yields

$$\langle B_2 \rangle = 2D + \frac{4D}{M}\left(\frac{1}{N} + 1\right), \tag{I.94}$$

from which we get the complete full bias together with eq. (I.91):

$$\langle B \rangle = \langle B_1 \rangle + \langle B_2 \rangle = \frac{4D}{M}\left(\frac{1}{N} - 1\right). \tag{I.95}$$

Thus,

$$\lambda^* - \overline{\overline{\lambda}} = -\frac{4D}{M}\left(1 - \frac{1}{N}\right). \tag{I.96}$$

Note that the bias is independent of N for large N. As a final step we will now turn to the bias term of the Brownian motion adapted least squares.

I.F.4 *Bias of Brownian motion adapted LS*

We now consider our third and final choice of $\boldsymbol{R}$-matrix. We write eq. (I.47), where we make use of the decomposition of the exact covariance matrix, eq. (I.18), $\overline{\overline{\boldsymbol{A}}}^T \overline{\overline{\boldsymbol{A}}} = \overline{\overline{\boldsymbol{Q}}}^{-1}$, as

$$\overline{\boldsymbol{R}} = \frac{1}{M} \overline{\overline{\boldsymbol{A}}}^T \overline{\boldsymbol{b}}^{-1} \overline{\overline{\boldsymbol{A}}}, \tag{I.97}$$

where $\overline{\boldsymbol{b}}$ is the inverse of $\overline{\boldsymbol{d}}$, in other words, $\overline{\boldsymbol{b}}$ is diagonal with elements

$$\overline{b}_{ii} = \frac{1}{\overline{d}_{ii}} = \frac{1}{M-1} \sum_{m}^{M} \left(\sum_{j}^{N} \overline{\overline{A}}_{ij} (y_j^{(m)} - \overline{y}_j) \right)^2 \tag{I.98}$$

on the diagonal. Thus eq. (I.43) now becomes

$$\lambda^* = \frac{\overline{\boldsymbol{Y}}^T \overline{\boldsymbol{b}}^{-1} \boldsymbol{T}}{\boldsymbol{T}^T \overline{\boldsymbol{b}}^{-1} \boldsymbol{T}}, \tag{I.99}$$

where

$$\overline{\boldsymbol{Y}} \equiv \overline{\overline{\boldsymbol{A}}} \overline{\boldsymbol{y}} \tag{I.100}$$

$$\boldsymbol{T} \equiv \overline{\overline{\boldsymbol{A}}} \boldsymbol{t}. \tag{I.101}$$

Eq. (I.99) now has the same form as eq. (I.84), (case 2, CLS/LS). If we replace $\overline{y}_i$ with $\overline{Y}_i$, t_i with T_i, and $\overline{Q}_{ii}$ with $\bar{b}_{ii}$ (and same for double bar), eq. (I.89) becomes (introducing explicit sums for clarity):

$$B_1 = -\frac{\sum_k \overline{Y}_k^T \frac{1}{(\bar{\bar{b}}_{kk})^2}\bar{b}_{kk}T_k}{\sum_q T_q^2/\bar{\bar{b}}_{qq}} = -\frac{\sum_{k,l,n}\overline{y}_l\overline{\overline{A}}_{lk}^T\frac{1}{(\bar{\bar{b}}_{kk})^2}\bar{b}_{kk}\overline{\overline{A}}_{kn}t_n}{\sum_{p,q,w}\frac{1}{\bar{\bar{b}}_{qq}}\overline{\overline{A}}_{qp}t_p\overline{\overline{A}}_{qw}t_w} \tag{I.102a}$$

$$\begin{aligned} B_2 &= \frac{\sum_{j,k}\overline{Y}_k\frac{1}{\bar{\bar{b}}_{kk}}T_kT_j^2\frac{1}{(\bar{\bar{b}}_{jj})^2}\bar{b}_{jj}}{(\sum_q T_q^2/\bar{\bar{b}}_{qq})^2} \\ &= \frac{\sum_{k,l,n,j,p,p'}\left(\overline{y}_l\overline{\overline{A}}_{lk}^T\frac{1}{\bar{\bar{b}}_{kk}}\overline{\overline{A}}_{kn}t_n\right)\left(\overline{\overline{A}}_{jp}t_p\overline{\overline{A}}_{jp'}t_{p'}\frac{1}{(\bar{\bar{b}}_{jj})^2}\bar{b}_{jj}\right)}{\sum_{q,p,w}(\overline{\overline{A}}_{qp}t_p\overline{\overline{A}}_{qw}t_w/\bar{\bar{b}}_{qq})^2}. \end{aligned} \tag{I.102b}$$

From eq. (I.98) we have

$$\begin{aligned} \bar{b}_{ii} &= \frac{1}{M-1}\sum_m^M\left(\sum_{j,j'}\overline{\overline{A}}_{ij}(y_j^{(m)}-\overline{y}_j)\overline{\overline{A}}_{ij'}(y_{j'}^{(m)}-\overline{y}_{j'})\right) \\ &= \sum_{j,j'}\overline{\overline{A}}_{ij}\overline{\overline{A}}_{ij'}\underbrace{\frac{1}{M-1}\sum_m^M(y_j^{(m)}-\overline{y}_j)(y_{j'}^{(m)}-\overline{y}_{j'})}_{\overline{Q}_{jj'}} \\ &= \sum_{j,j'}\overline{\overline{A}}_{ij}\overline{Q}_{jj'}\overline{\overline{A}}_{j'i}^T, \end{aligned} \tag{I.103}$$

from which it follows that

$$\langle\bar{b}_{ii}\rangle = \left(\overline{\overline{\boldsymbol{A}}}\langle\overline{\boldsymbol{Q}}\rangle\overline{\overline{\boldsymbol{A}}}^T\right)_{ii} = \left(\overline{\overline{\boldsymbol{A}}}\,\overline{\overline{\boldsymbol{Q}}}\,\overline{\overline{\boldsymbol{A}}}^T\right)_{ii} = 1 = \bar{\bar{b}}_{ii}, \tag{I.104}$$

which used with eq. (I.102a), gives

$$B_1 = -\frac{\sum_{k,l,n}\overline{y}_l\overline{\overline{A}}_{lk}^T\bar{b}_{kk}\overline{\overline{A}}_{kn}t_n}{\sum_{p,q,w}\overline{\overline{A}}_{pq}^T\overline{\overline{A}}_{qw}t_pt_w} = -\frac{\sum_{k,l,n}\overline{y}_l\overline{\overline{A}}_{lk}^T\bar{b}_{kk}\overline{\overline{A}}_{kn}t_n}{\sum_{p,w}t_p(\overline{\overline{\boldsymbol{Q}}}^{-1})_{pw}t_w}. \tag{I.105a}$$

Substituting $\bar{b}_{ii}$ in the same way for eq. (I.102b) gives

$$\begin{aligned} B_2 &= \frac{\sum_{l,k,n,j,p,p'} \left(\overline{y}_l \overline{\overline{A}}^T_{lk} \overline{\overline{A}}_{kn} t_n \right) \left(\overline{\overline{A}}_{jp} \overline{\overline{A}}_{jp'} \bar{b}_{jj} t_p t_{p'} \right)}{(\sum_{p,w} t_p (\overline{\overline{\boldsymbol{Q}}}^{-1})_{pw} t_w)^2} \\ &= \frac{\sum_{l,n,j,p,p'} \left(\overline{y}_l (\overline{\overline{\boldsymbol{Q}}}^{-1})_{ln} t_n \right) \left(\overline{\overline{A}}_{jp} \overline{\overline{A}}_{jp'} \bar{b}_{jj} t_p t_{p'} \right)}{(\sum_{p,w} t_p (\overline{\overline{\boldsymbol{Q}}}^{-1})_{pw} t_w)^2}. \end{aligned} \tag{I.105b}$$

As a sanity check we note that the total bias, $B_1 + B_2$, is zero if bar is replaced with double bar and $\overline{\overline{y}}_l = 2Dt_l$. From eq. (I.105) we see that we now need to find the following, by route of eq. (I.103),

$$\langle \overline{y}_l \bar{b}_{kk} \rangle = \sum_{i,j} \overline{\overline{A}}^T_{ik} \overline{\overline{A}}_{kj} \langle \overline{y}_l \overline{Q}_{ij} \rangle = \sum_{i,j} \overline{\overline{A}}^T_{ik} \overline{\overline{A}}_{kj} \left(\overline{\overline{\sigma}}^2_l \overline{\overline{Q}}_{ij} + \frac{8}{M} \overline{\overline{V}}_{li} \overline{\overline{V}}_{ij} \overline{\overline{V}}_{jl} \right) \tag{I.106}$$

(no sum over k), where we in the last step used eq. (I.72). Eq. (I.105a) now becomes (repeated indices are summed over)

$$\begin{aligned} \langle B_1 \rangle &= -\frac{\overline{\overline{A}}^T_{lk} \overline{\overline{A}}^T_{ik} \overline{\overline{A}}_{kj} \left(\overline{\overline{\sigma}}^2_l \overline{\overline{Q}}_{ij} + \frac{8}{M} \overline{\overline{V}}_{li} \overline{\overline{V}}_{ij} \overline{\overline{V}}_{jl} \right) \overline{\overline{A}}_{kn} t_n}{\boldsymbol{t}^T \overline{\overline{\boldsymbol{Q}}}^{-1} \boldsymbol{t}} \\ &= -\frac{\overline{\overline{\sigma}}^2_l (\overline{\overline{\boldsymbol{Q}}}^{-1})_{ln} t_n + \frac{8}{M} \overline{\overline{A}}_{kl} \overline{\overline{A}}_{ki} \overline{\overline{A}}_{kj} \overline{\overline{A}}_{kn} \overline{\overline{V}}_{li} \overline{\overline{V}}_{ij} \overline{\overline{V}}_{jl} t_n}{\boldsymbol{t}^T \overline{\overline{\boldsymbol{Q}}}^{-1} \boldsymbol{t}}, \end{aligned} \tag{I.107}$$

where for the first term we used that $\overline{\overline{\boldsymbol{A}}}\,\overline{\overline{\boldsymbol{Q}}}\,\overline{\overline{\boldsymbol{A}}}^T = \boldsymbol{I}$ (for the term $\overline{\overline{A}}^T_{ik} \overline{\overline{A}}_{kj} \overline{\overline{Q}}_{ij}$), followed by $\overline{\overline{A}}^T_{lk} \overline{\overline{A}}_{kn} = (\overline{\overline{\boldsymbol{Q}}}^{-1})_{ln}$ which gives the final step in eq. (I.107). Now consider eq. (I.105b)

$$\begin{aligned} \langle B_2 \rangle &= \frac{(\overline{\overline{\boldsymbol{Q}}}^{-1})_{ln} \overline{\overline{A}}_{kp} \overline{\overline{A}}_{kp'} \langle \overline{y}_l \bar{b}_{kk} \rangle t_p t_{p'} t_n}{(\boldsymbol{t}^T \overline{\overline{\boldsymbol{Q}}}^{-1} \boldsymbol{t})^2} \\ &= \frac{(\overline{\overline{\boldsymbol{Q}}}^{-1})_{ln} \overline{\overline{A}}_{kp} \overline{\overline{A}}_{kp'} \overline{\overline{A}}_{ki} \overline{\overline{A}}_{kj} \left(\overline{\overline{\sigma}}^2_l \overline{\overline{Q}}_{ij} + \frac{8}{M} \overline{\overline{V}}_{li} \overline{\overline{V}}_{ij} \overline{\overline{V}}_{jl} \right) t_p t_{p'} t_n}{(\boldsymbol{t}^T \overline{\overline{\boldsymbol{Q}}}^{-1} \boldsymbol{t})^2} \\ &= \frac{t_p \overline{\overline{A}}_{kp} \overline{\overline{A}}_{kp'} t_{p'} (\overline{\overline{\boldsymbol{Q}}}^{-1})_{ln} \overline{\overline{\sigma}}^2_l t_n}{(\boldsymbol{t}^T \overline{\overline{\boldsymbol{Q}}}^{-1} \boldsymbol{t})^2} \\ &\quad + \frac{\frac{8}{M} (\overline{\overline{\boldsymbol{Q}}}^{-1})_{ln} \overline{\overline{A}}_{kp} \overline{\overline{A}}_{kp'} \overline{\overline{A}}_{ki} \overline{\overline{A}}_{kj} \overline{\overline{V}}_{li} \overline{\overline{V}}_{ij} \overline{\overline{V}}_{jl} t_p t_{p'} t_n}{(\boldsymbol{t}^T \overline{\overline{\boldsymbol{Q}}}^{-1} \boldsymbol{t})^2}, \end{aligned} \tag{I.108}$$

where we in the middle step used that $\overline{\overline{A}}_{ki}\,\overline{\overline{A}}_{kj}\,\overline{\overline{Q}}_{ij'} = \delta_{kj'}$. Let us use that $\overline{\overline{A}}_{kp}\,\overline{\overline{A}}_{kp'} = (\overline{\overline{\boldsymbol{Q}}}^{-1})_{ij}$ and arrive at

$$\langle B_2\rangle = \frac{\overline{\overline{\sigma}}_l^2(\overline{\overline{\boldsymbol{Q}}}^{-1})_{ln}t_n}{\boldsymbol{t}^T\overline{\overline{\boldsymbol{Q}}}^{-1}\boldsymbol{t}} + \frac{8}{M}\frac{(\overline{\overline{\boldsymbol{Q}}}^{-1})_{ln}\overline{\overline{A}}_{kp}\overline{\overline{A}}_{kj}\overline{\overline{A}}_{ki}\overline{\overline{A}}_{kj}\overline{\overline{V}}_{li}\overline{\overline{V}}_{ij}\overline{\overline{V}}_{jl}t_pt_{p'}t_n}{(\boldsymbol{t}^T\overline{\overline{\boldsymbol{Q}}}^{-1}\boldsymbol{t})^2}. \tag{I.109}$$

Our arduous journey has now come to its end, and the total bias $\langle B_1\rangle + \langle B_2\rangle$ is

$$\langle B_1\rangle = -\frac{8}{M}\frac{\overline{\overline{A}}_{ki}\overline{\overline{V}}_{ij}\overline{\overline{A}}_{kj}\overline{\overline{V}}_{jl}\overline{\overline{A}}_{kl}\overline{\overline{V}}_{li}\overline{\overline{A}}_{kn}t_n}{\boldsymbol{t}^T\overline{\overline{\boldsymbol{Q}}}^{-1}\boldsymbol{t}} \tag{I.110a}$$

$$\langle B_2\rangle = \frac{8}{M}\frac{\overline{\overline{A}}_{ki}\overline{\overline{V}}_{ij}\overline{\overline{A}}_{kj}\overline{\overline{V}}_{jl}(\overline{\overline{A}}_{kp}t_p)^2(\overline{\overline{\boldsymbol{Q}}}^{-1})_{ln}t_n}{(\boldsymbol{t}^T\overline{\overline{\boldsymbol{Q}}}^{-1}\boldsymbol{t})^2}. \tag{I.110b}$$

I.G JACKKNIFE BIAS REDUCTION

Through data resampling, bias in data-fitting can often be significantly reduced. Let O be the parameter estimator, based on some data set with M trajectories, to the true parameter $\overline{\overline{O}}$. Herein, we choose O as either the parameters $\boldsymbol{\lambda}^*$ or their associated covariance matrix $\boldsymbol{\Delta}$. As outlined in section I.F.1, one often expects such a finite data set to yield a bias contribution of the form

$$O = \overline{\overline{O}} + \frac{a}{M} + \frac{b}{M^2} + \frac{c}{M^3} + \mathcal{O}\left(\frac{1}{M^4}\right). \tag{I.111}$$

The bias terms can be reduced by increasing the data samples, M, or by using the jackknife method [15]. Let us split the sample into g groups, each of size h, and define $O_{[-j]}$ as the parameter fitted to a data sample with the jth group removed.

I.G.1 *First order jackknife*

The first order bias term can be removed through repeated fitting and averaging over the sampled data set:

$$O^{(1)} = \frac{1}{g}\sum_{j=1}^{g} O_{[-j]} \tag{I.112a}$$

$$O_J^{(0,1)} = gO - (g-1)O^{(1)}. \tag{I.112b}$$

By using eq. (I.111) which has bias terms proportional to $M = hg$ for the full fitting, O, and $h(g-1)$ for the reduced sample estimator in eq. (I.112), we see that we are left with

$$\begin{aligned} O_J^{(0,1)} &= \overline{\overline{O}} - \frac{b}{h^2}\frac{1}{g(g-1)} - \frac{c}{h^3}\left(\frac{1}{(g-1)^2} - \frac{1}{g^2}\right) + \mathcal{O}(g^{-3}) \\ &\approx \overline{\overline{O}} - \frac{b}{M^2} - 2\frac{c}{M^3}, \end{aligned} \tag{I.113}$$

lacking the first order bias term. Although the higher order terms remain, their contribution is expected to be lower than the first order term.

I.G.2 *Second order jackknife*

For further bias reduction we can apply a second order correction. In a similar spirit to what is done in the first order jackknife, we split the data into g groups, and define $O_{[-j,-j']}$ as the parameter estimator based on a data set with the jth and j'th group removed, each of size h. This follows Schucany *et al.* [17]. We get

$$O^{(2)} = \frac{2}{g(g-1)}\sum_{j<j'}^{g} O_{[-j,-j']} \tag{I.114a}$$

$$O_J^{(1,2)} = (g-1)O^{(1)} - (g-2)O^{(2)} \tag{I.114b}$$

$$O_J^{(0,1,2)} = \frac{g}{2}O_J^{(0,1)} - \frac{g-2}{2}O_J^{(1,2)}. \tag{I.114c}$$

If we combine our result with eq. (I.111), we are only left with the third order term and the ones that follows it,

$$\begin{aligned} O_J^{(0,1,2)} &= \overline{\overline{O}} + \frac{c}{h^3}\frac{1}{g(g-1)(g-2)} + \mathcal{O}(g^{-4}) \\ &\approx \overline{\overline{O}} + \frac{c}{M^3}. \end{aligned} \tag{I.115}$$

I.G.3 *Variance for jackknife estimators*

In this section, we use eq. (I.51) to show that $\lambda_a^* - \overline{\overline{\lambda}}_a$ is insensitive (to lowest order in $1/M$) to the jackknifing procedure. As a consequence, the covariance estimation formula, eq. (I.4) in the main text, remains valid also for jackknifed parameter estimations.

For later convenience, we define the derivative in eq. (I.51) as

$$A_{a,i} = \left.\frac{\partial \lambda_a^*}{\partial \overline{y}_i}\right|_{\overline{y}_i = \overline{\overline{y}}_i}, \tag{I.116}$$

which we will use in the following.

I.G.3.1 *First order jackknife*

To first order the jackknife estimator is obtained by dividing the M trajectories into g groups of size h. Define the observable $\overline{O}_{[-j],i}$ as the estimate for observable O, in point i, with group j removed. In particular,

$$\overline{y}_{[-j],i} = \frac{1}{M-h} \sum_{m \neq m_j} y_i^{(m)} = \frac{1}{M-h} \left(\sum_{m=1}^{M} y_i^{(m)} - \sum_{m_j} y_i^{(m)} \right). \tag{I.117}$$

The corresponding non-jackknifed estimator is

$$\overline{y}_i = \frac{1}{M} \sum_{m=1}^{M} y_i^{(m)}. \tag{I.118}$$

The bias of the first order jackknife estimator of $\overline{\overline{\lambda}}_a$ within the CLS method (see section I.E) is

$$\begin{aligned}
&\lambda_{J,a}^{(0,1)} - \overline{\overline{\lambda}}_a = g\lambda_a^* - (g-1)\left[\frac{1}{g}\sum_{j=1}^{g} \lambda_{[-j],a}\right] - \overline{\overline{\lambda}}_a \\
&= \frac{1}{h}\left[M\lambda_a^* - (M-h)\frac{1}{g}\sum_{j=1}^{g}\lambda_{[-j],a}\right] \\
&= \frac{1}{h}\sum_i A_{a,i}\left(M(\overline{y}_i - \overline{\overline{y}}_i) - (M-h)\frac{1}{g}\sum_{j=1}^{g}\left(\overline{y}_{[-j],i} - \overline{\overline{y}}_i\right)\right) \\
&= \frac{1}{h}\sum_i A_{a,i}\left(\sum_{m=1}^{M}(y_i^{(m)} - \overline{\overline{y}}_i)\right. \\
&\qquad \left. -\frac{1}{g}\sum_{j=1}^{g}\left(\sum_{m=1}^{M}(y_i^{(m)} - \overline{\overline{y}}_i) - \sum_{m_j}(y_i^{(m_j)} - \overline{\overline{y}}_i)\right)\right) \\
&= \sum_i A_{a,i}\left(\frac{1}{gh}\sum_{j=1}^{g}\sum_{m_j}(y_i^{(m_j)} - \overline{\overline{y}}_i)\right) \\
&= \sum_i A_{a,i}\left(\frac{1}{M}\sum_{m=1}^{M}(y_i^{(m)} - \overline{\overline{y}}_i)\right)
\end{aligned}$$

$$= \lambda_a^* - \overline{\overline{\lambda}}_a, \tag{I.119}$$

where we used eq. (I.51) to get to the second and last (fifth) row, and eq. (I.117)-(I.118) for the third row. Thus

$$\lambda_{J,a}^{(0,1)} - \overline{\overline{\lambda}}_a = \lambda_a^* - \overline{\overline{\lambda}}_a. \tag{I.120}$$

Hence, jackknife does not change the (co)variance:

$$(\lambda_{J,a}^{(0,1)} - \overline{\overline{\lambda}}_a)(\lambda_{J,b}^{(0,1)} - \overline{\overline{\lambda}}_b) = (\lambda_a^* - \overline{\overline{\lambda}}_a)(\lambda_b^* - \overline{\overline{\lambda}}_b). \tag{I.121}$$

I.G.3.2 *Second order jackknife*

For the second order bias removal, the M trajectories are again divided into g groups. We define, as before, $\overline{O}_{[-j,-j'],i}$ as the estimate for observable O, in point i, with group j and j' removed. In particular

$$\begin{aligned} \overline{y}_{[-j,-j'],i} &= \frac{1}{M-2h} \sum_{m \neq m_j, m_{j'}} y_i^{(m)} \\ &= \frac{1}{M-2h} \left(\sum_m^M y_i^{(m)} - \sum_{m_j} y_i^{(m_j)} - \sum_{m_{j'}} y_i^{(m_{j'})} \right). \end{aligned} \tag{I.122}$$

The average over all groups for λ_a is

$$\lambda_a^{(2)} = \frac{1}{g(g-1)} \sum_{j \neq j'} \lambda_{[-j,-j']}. \tag{I.123}$$

The second order jackknife is now (as given by eq. (I.114c))

$$\lambda_{J,a}^{(0,1,2)} = \frac{g}{2} \lambda_{J,a}^{(0,1)} - \frac{g-2}{2} \lambda_{J,a}^{(1,2)}. \tag{I.124}$$

Using eq. (I.114b) we note

$$\begin{aligned} &\lambda_{J,a}^{(1,2)} - \overline{\overline{\lambda}}_a = \frac{1}{h} \Bigg((M-h) \left[\frac{1}{g} \sum_{j=1}^{g} \lambda_{[-j],a} \right] \\ &-(M-2h) \left[\frac{1}{g(g-1)} \sum_{j \neq j'} \lambda_{[-j,-j'],a} \right] \Bigg) - \overline{\overline{\lambda}}_a \\ &= \frac{1}{h} \sum_i A_{a,i} \Bigg(\frac{1}{g} \sum_{j=1}^{g} \sum_{m=1}^{M} (y_i^{(m)} - \overline{\overline{y}}_i) - \frac{1}{g} \sum_{j=1}^{g} \left(\sum_{m_j} y_i^{(m_j)} - \overline{\overline{y}}_i \right) \\ &- \Bigg[\frac{1}{g(g-1)} \sum_{j,j'} \sum_{m=1}^{M} (y_i^{(m)} - \overline{\overline{y}}_i) \end{aligned}$$

$$-\frac{1}{g(g-1)}\sum_{j,j'}\sum_{m_j}^{M}(y_i^{(m_j)}-\overline{\overline{y}}_i)-\frac{1}{g(g-1)}\sum_{j,j'}\sum_{m_{j'}}^{M}(y_i^{(m_{j'})}-\overline{\overline{y}}_i)\Big]\Big)$$

$$=\frac{1}{h}\sum_i A_{a,i}\Big(-\frac{1}{g}\sum_{j=1}^{g}\sum_{m_j}(y_i^{(m_j)}-\overline{\overline{y}}_i)$$

$$+\frac{1}{g-1}\sum_{j'}\frac{1}{g}\sum_j\sum_{m_j}(y_i^{(m_j)}-\overline{\overline{y}}_i)+\frac{1}{g-1}\sum_j\frac{1}{g}\sum_{j'}(y_i^{(m_{j'})}-\overline{\overline{y}}_i)\Big)$$

$$=\frac{1}{h}\sum_i A_{a,i}\left(-\frac{1}{g}+\frac{1}{g}+\frac{1}{g}\right)\sum_j\sum_{m_j}(y_i^{(m_j)}-\overline{\overline{y}}_i). \tag{I.125}$$

Thus

$$\lambda_{J,a}^{(1,2)}-\overline{\overline{\lambda}}_a=\sum_i A_{a,i}\frac{1}{M}\sum_j\sum_{m_j}(y_i^{(m_j)}-\overline{\overline{y}}_i)=\lambda_a^*-\overline{\overline{\lambda}}_a \tag{I.126}$$

and

$$(\lambda_{J,a}^{(0,1,2)}-\overline{\overline{\lambda}}_a)=\lambda_a^*-\overline{\overline{\lambda}}_a. \tag{I.127}$$

Thus the second order jackknife estimator has the same variance and covariance as non-jackknifed estimators.

I.H CRAMER-RAO LOWER BOUND

The Cramer-Rao lower bound puts a precise bound on how well we can estimate a parameter. More precisely it is a bound on the variance of any observable [9]. We here state the bound for the case of a single parameter to be estimated (as for Brownian motion). In general we can write:

$$\lambda^*=\overline{\overline{\lambda}}+B(\overline{\overline{\lambda}}), \tag{I.128}$$

where λ is our estimated parameter, $\overline{\overline{\lambda}}$ is the exact parameter value and $B(\overline{\overline{\lambda}})$ is the bias. The Cramer-Rao lower bound is then

$$\text{variance}(\lambda^*)\geq\frac{\left(1+B'(\overline{\overline{\lambda}})\right)^2}{I(\overline{\overline{\lambda}})}, \tag{I.129}$$

where a prime denotes derivative and

$$I(\overline{\overline{\lambda}})=\Big\langle\frac{\partial l(y_1,\ldots,y_N;\overline{\overline{\lambda}})}{\partial\overline{\overline{\lambda}}}\Big\rangle \tag{I.130}$$

is the Fisher information, with the likelihood function defined as

$$l(y_1,\ldots,y_N;\overline{\overline{\lambda}}) = \log\left(\rho(y_1,\ldots,y_N;\overline{\overline{\lambda}})\right), \tag{I.131}$$

with $\rho(y_1,\ldots,y_N;\overline{\overline{\lambda}}))$ being the joint probability density for $\{y_i\}_{i=1}^N$. As previously discussed, in general we have that $B(\overline{\overline{\lambda}}) = a(\overline{\overline{\lambda}})/M$, and therefore we can write eq. (I.129) according to

$$\text{variance}(\lambda^*) \geq \frac{1}{I(\overline{\overline{\lambda}})}\left(1 + \frac{2a'(\overline{\overline{\lambda}})}{M} + \mathcal{O}(\frac{1}{M^2})\right). \tag{I.132}$$

Evaluating the Fisher information is a complicated task. The one exception is for the case that $\rho(y_1,\ldots,y_N;\overline{\overline{\lambda}})$ is a multivariate Gaussian, in which case $I(\overline{\overline{\lambda}})$, and hence the Cramer-Rao bound to lowest order in $1/M$, can be explicitly evaluated.

In the cases of interest here, namely, "squared processes" in d dimensions, where $y_i = \boldsymbol{x}_i^2$, the quantity $p(y_1,\ldots,y_N;\overline{\overline{\lambda}})$ is formally:

$$\rho(y_1,\ldots,y_n;\overline{\overline{\lambda}}) = \int \prod_{i=1}^{N} \delta(y_i - |\boldsymbol{x}_i|^2) p(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N;\overline{\overline{\lambda}}) d^d x_1 \cdots d^d x_N, \tag{I.133}$$

where $p(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N;\overline{\overline{\lambda}})$ is the probability density for the displacements, assuming the particle's initial position is at the origin. For BM and FBM in one-dimension, we have the explicit form:

$$p(x_1,\ldots,x_N;\overline{\overline{\lambda}}) = \frac{1}{(2\pi)^{N/2}\det(\boldsymbol{v})^{1/2}} \exp\left(-\frac{1}{2}\sum_{\text{B},j} x_i (\boldsymbol{v}^{-1})_{ij} x_j\right) \tag{I.134}$$

i.e., a multivariate Gaussian with covariance $\boldsymbol{v}$. For such a probability it is not possible (nor for squared displacements from CTRW) to express the Fisher information in a simple closed-form, for general $\boldsymbol{v}$.

I.I COEFFICIENT OF DETERMINATION

We determine the goodness of fit by using the R^2 coefficient of determination, defined as

$$R^2 = 1 - \frac{S_{\text{res}}}{S_{\text{tot}}}. \tag{I.135}$$

The method is based on a sum of squares over the N sampling points of, in our case, the MSD $\overline{\boldsymbol{y}}$; hence, measuring the deviation from the sample mean in *time*,

$$\hat{\overline{y}} = \frac{1}{N}\sum_{i}^{N} \overline{y}_i \tag{I.136}$$

$$S_{\mathrm{tot}} = \sum_{i}^{N} (\overline{y}_i - \hat{\overline{y}})^2 \tag{I.137}$$

$$S_{\mathrm{res}} = \sum_{i}^{N} \left(f(t_i; \boldsymbol{\lambda}) - \overline{y}_i\right)^2 . \tag{I.138}$$

A model that fits data perfectly has an $R^2 = 1$, while if it does not fit at all, $R^2 \ll 1$, see Supplementary Figure I.S5.

REFERENCES

1. E. Barkai, Y. Garini, and R. Metzler, "Strange kinetics of single molecules in living cells," *Phys. Today*, vol. 65, no. 8, p. 29, 2012.
2. M. Chaichian and A. Demichev, "Path integrals in physics, vol. 1: Stochastic processes and quantum mechanics," *IOP, Bristol, UK*, 2001.
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. H. Qian, "Fractional Brownian motion and fractional Gaussian noise," in *Processes with Long-Range Correlations*, pp. 22–33, Springer, 2003.
5. B. B. Mandelbrot and J. W. Van Ness, "Fractional Brownian motions, fractional noises and applications," *SIAM review*, vol. 10, no. 4, pp. 422–437, 1968.
6. R. B. Davies and D. Harte, "Tests for hurst effect," *Biometrika*, vol. 74, no. 1, pp. 95–101, 1987.
7. M. Chambers, "The simulation of random vector time series with given spectrum," *Mathematical and Computer Modelling*, vol. 22, no. 2, pp. 1–6, 1995.
8. 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.

9. A. Van den Bos, *Parameter estimation for scientists and engineers*. John Wiley & Sons, 2007.
10. S. Gottlieb, W. Liu, R. L. Renken, R. L. Sugar, and D. Toussaint, "Hadron masses with two quark flavors," *Phys. Rev. D*, vol. 38, pp. 2245–2265, Oct 1988.
11. D. Seibert, "Undesirable effects of covariance matrix techniques for error analysis," *Phys. Rev. D*, vol. 49, pp. 6240–6243, Jun 1994.
12. C. Michael, "Fitting correlated data," *Phys. Rev. D*, vol. 49, pp. 2616–2619, 1994.
13. N. G. Van Kampen, *Stochastic processes in physics and chemistry*, vol. 1. Elsevier, 1992.
14. M. H. Quenouille, "Notes on bias in estimation," *Biometrika*, vol. 43, no. 3/4, pp. 353–360, 1956.
15. R. G. Miller, "The jackknife — a review," *Biometrika*, vol. 61, no. 1, pp. 1–15, 1974.
16. I. Gradshteyn and I. Ryzhik, "Table of integrals, series and products (corrected and enlarged edition prepared by A. Jeffrey and D. Zwillinger)," *Academic Press, New York*, 2000.
17. W. Schucany, H. Gray, and D. Owen, "On bias reduction in estimation," *Journal of the American Statistical Association*, vol. 66, no. 335, pp. 524–533, 1971.

II

PAPER II

Selection shapes transcriptional logic and regulatory specialization in genetic networks

Karl Fogelmark, Carsten Peterson and Carl Troein

Computational Biology and Biological Physics, Department of Astronomy and Theoretical Physics, Lund University, 223 62 Lund, Sweden.

PLoS ONE, vol. **11**, e0150340 (2016)
doi: 10.1371/journal.pone.0150340

Background
Living organisms need to regulate their gene expression in response to environmental signals and internal cues. This is a computational task where genes act as logic gates that connect to form transcriptional networks, which are shaped at all scales by evolution. Large-scale mutations such as gene duplications and deletions add and remove network components, whereas smaller mutations alter the connections between them. Selection determines what mutations are accepted, but its importance for shaping the resulting networks has been debated.

Methodology
To investigate the effects of selection in the shaping of transcriptional networks, we derive transcriptional logic from a combinatorially powerful yet tractable

model of the binding between DNA and transcription factors. By evolving the resulting networks based on their ability to function as either a simple decision system or a circadian clock, we obtain information on the regulation and logic rules encoded in functional transcriptional networks. Comparisons are made between networks evolved for different functions, as well as with structurally equivalent but non-functional (neutrally evolved) networks, and predictions are validated against the transcriptional network of *E. coli*.

Principal findings
We find that the logic rules governing gene expression depend on the function performed by the network. Unlike the decision systems, the circadian clocks show strong cooperative binding and negative regulation, which achieves tight temporal control of gene expression. Furthermore, we find that transcription factors act preferentially as either activators or repressors, both when binding multiple sites for a single target gene and globally in the transcriptional networks. This separation into positive and negative regulators requires gene duplications, which highlights the interplay between mutation and selection in shaping the transcriptional networks.

AUTHOR SUMMARY

The living cell responds to internal and external cues, altering its activity and composition to maximize survival and reproduction. Many biological processes are regulated at the transcriptional level: the expression of individual genes is activated or repressed by proteins whose own levels depend on similar regulation. These interactions form a transcriptional network, which dynamically processes information from the environment.

The networks are shaped by mutations that rewire connections between genes, and by selection that accepts changes in relation to their impact on network function. As in other areas of biology, it is not clear whether similarities between networks found in nature reflect strongly beneficial adaptations or merely result from exposing the different networks to similar types of mutations.

So what distinguishes functional networks from non-functional ones with similar architecture? By computer simulated evolution of transcriptional networks, we examine differences in the regulation of individual genes that results from interactions between transcription factors at gene regulatory regions on the DNA. We extract Boolean logic functions from the gene regulation, and investigate how they differ between functional and non-functional networks.

II.1 INTRODUCTION

The living cell can be viewed as a decision-making system that needs to respond appropriately to a wide range of external and internal signals in order to survive and maximize its reproductive success. Interacting components such as genes and proteins form networks that control the flow of information. Such biological networks can be described at different scales, ranging from communication between large functional modules down to biochemically detailed models that include e.g. protein modifications. Here we consider genetic networks where the nodes are genes connected by edges that represent transcriptional regulation [1].

At all scales, evolutionary pressures shape networks under constraints imposed by their function. Biological functions may require specific network architectures and logic; for instance, an oscillator cannot work without a negative feedback loop. Some solutions are favourable because of greater evolvability and/or mutational stability; for instance, the modularity of evolved networks may resemble that of their engineered counterparts [2].

The structure of evolved biological networks can be replicated *in silico* through a combination of selection and large-scale duplication events [3]. On the other hand, the importance of selection has been questioned on the grounds that frequent and largely neutral rewiring events are able to explain common features of evolved transcriptional networks [4, 5]. In this view, gene duplications and other large mutation events are drivers of the exploration of the vast space of possible networks, with selection acting as a guide.

To study the balance between selection and neutral evolution *in silico*, as well as the role of gene duplications and other mutations, we need a model of evolvable transcriptional networks, including both network topology and transcriptional dynamics. Gene expression levels can be modelled either as continuous or discrete variables, each with its own advantages. Describing transcriptional regulation in terms of logic rules that govern gene expression is straightforward when the networks are modelled as discrete systems using Boolean functions [6]. However, real gene expression is not an all-or-nothing process, and a continuous model gives a more accurate representation of the transcriptional dynamics. Even so, the regulation of a gene is easier to understand in a Boolean description. We will therefore discretize the continuous expression levels only when analyzing the transcriptional logic encoded in networks.

A continuous dynamical model for the binding of transcription factors (TFs) to gene regulatory regions was derived by Banzhaf in [7]. Starting from a simplistic description of DNA and binding motifs as sequences of bits, the expression of a gene was determined by the sequence mismatch between the TFs and two binding regions, one activating and one repressing. However, limiting the regulatory region to only two non-interacting binding sites severely restricted the possibilities for transcriptional logic in that model.

In order to generate transcriptional logic functions, Buchler *et al.* constructed a model with competitive and cooperative binding of TFs at nearby binding sites. The recruitment or blocking of RNA polymerase (RNAP) depended on the positions of bound TFs, leading to activation or repression of the gene [8]. Cooperative binding, which in nature occurs through several possible molecular mechanisms, favours network connectivity and rewiring [5]. Even though Boolean terminology was used to describe the rich set of logic generated by this binding model, the underlying rules were continuous functions of TF concentrations.

Inspired by this earlier work, we construct a dynamical model of transcriptional regulation with combinatorial interactions of TFs on the DNA, which allows us to grow and evolve networks to explore the effects of selection. The representation of genes as strings of bits is borrowed from Banzhaf [7], but we allow multiple binding sites so that we can derive complex logic from the interactions of TFs within the regulatory region, as suggested by Buchler *et al.* [8]. In addition, we formulate dynamics for the production and degradation of proteins.

To perform network selection, we need to compute a biologically relevant measure of fitness based on the dynamics of the system. Ideally, we would like to simulate an entire organism, but this is not feasible. Instead, as targets for the simulated evolution we choose two well-defined computational problems with the potential to generate a variety of large circuits. The first is a relatively simple artificial problem: the majority rule, a Boolean decision problem which can be expected to yield networks with few feedback loops and mostly positive regulation. The second is a more complex problem which is directly linked to biology: the circadian clock, whose inherent properties include oscillations, internal feedbacks and input from the environment [9]. Simple oscillators have previously been used as a target for evolution in more detailed models of transcriptional and posttranslational regulation [10, 11].

Even if neutral mutations dominate the evolutionary process, the effects of selection must in some way be reflected in the structure of biological networks. The aim of this work is to identify markers of biological function in transcriptional networks evolved under selection, as compared to neutrally evolved ones, at the level of individual genes and their connections. To address this issue, we investigate properties such as the degree of specialization in transcription factors towards either activation or repression. We identify properties of regulation in functional networks, validate the results against the transcriptional network of *E. coli*, and show that large-scale mutations are necessary for reproducing the observed separation between activators and repressors.

II.2 METHODS

We have implemented a dynamical model of transcriptional regulation where proteins and regulatory regions are represented as sequences of bits, and transcription rates are determined by the interactions between TFs and DNA. An individual network consists of a variable number of genes, each represented as strings of bits. In the current model, all proteins are TFs, which may bind to the regulatory region of a gene to modify its expression, by either facilitating or inhibiting the recruitment of RNAP. A regulatory region may contain binding sites for many different TFs, giving rise to complex logic through the combinatorics of cooperative and mutually exclusive binding, as explained in the following section. This regulatory model is shown as miniature example in Figures II.1A and II.1B, together with an overview of the simulation process in Figure II.1C.

II.2.1 *Transcriptional regulation*

At the heart of the model, transcriptional regulation is derived from regulatory regions and binding motifs which are described as sequences of ones and zeroes. Typical binding motifs are about 5–20 basepairs in size, which motivates us to represent motifs by 32 bits to serve as 16 nucleotides, considering that DNA carries two bits of information per basepair. Regulatory regions are represented by 256 bits surrounding the transcriptional start site (TSS). Extending the regulatory regions to 512 bits produced similar results (Figures II.S3 and II.S4), suggesting that the smaller size provides sufficient combinatorics at lower computational cost.

To form a network, every TF motif is compared with all positions along the DNA sequences, and binding is considered to occur where the Hamming distance (the number of mismatching bits) is below a threshold, $H_{\max}= 6$ (with similar results at $H_{\max}= 8$). A regulatory region may thus contain a large number of binding sites, and the same TF may bind to several sites. If a pair of binding sites have any overlap, they cannot be occupied simultaneously, and this exclusion limits the number of bound TFs at any given moment. For our choice of parameters, there are $256 - 32 + 1 = 225$ possible binding sites and no more than $256/32 = 8$ simultaneously bound TFs.

Transcriptional regulators may have properties that make them predominantly activating or repressing. Examples include the potent activator Gal4 in yeast [12] and the KRAB domain which is strongly associated with repression in eukaryotes [13]. To be able to investigate whether the separation of positive and negative regulation is a fundamental principle of gene networks, we based the model on the more flexible arrangement in *E. coli*, where the sign of TF regulation is primarily determined by the positions of binding sites relative to the TSS [14]. In the model, TFs that bind to the regulatory region upstream of the TSS act as activators to initiate or enhance transcription, whereas TFs that are located downstream of the TSS are assumed to block RNAP and act as repressors, disallowing any transcription of the gene. The TSS is located near the middle of the regulatory region, such that half of the possible binding sites are activating and half are repressing.

A TF may thus act as an activator for some genes and as a repressor for others, depending on where it finds a matching binding pattern on the DNA sequence. It may also regulate ambiguously, with binding sites of opposite signs in a single regulatory region.

Negative interactions are created by overlapping binding sites, but the model also includes cooperative binding between TFs at nearby binding sites, partly to capture the effects of complex formation and other protein–protein interactions. As motivated in the Buchler *et al.* model, two TFs that occupy closely spaced binding sites lower their binding energy by $\beta = 3\ k_B T$ [8]. Presently, our model includes this cooperative interaction between all pairs of binding sites within a distance of 10 bits end-to-end, regardless of the identity of the TFs. This nonspecificity is a simplification aimed at capturing cooperative binding regardless of its mechanisms (cf. ref. [5]).

The transcription rate of gene g is determined by the recruitment of RNAP to its promotor region, which depends on the number of bound activating TFs, which in turn follows a distribution computed from the statistical weights of all possible binding states.

Given that n TFs are bound as activators, the probability of finding RNAP bound at gene g is

$$P_g(n) = \frac{e^{-(b_g - \lambda n)}}{1 + e^{-(b_g - \lambda n)}}, \tag{II.1}$$

where the RNAP binding energy is lowered by $\lambda = 3\ k_B T$ by each additional bound TF [8], and where the gene-specific ground state energy, $-3\ k_B T < b_g < 9\ k_B T$, allows for a wide range of basal transcription rates.

Assume that TF i can bind to site j with a mismatch of $H_{ij} < H_{\max}$ bits. The statistical weight of site j being occupied by any TF at time t is

$$z_j(\boldsymbol{c}(t)) = \sum_i \alpha e^{-\gamma H_{ij}} c_i(t), \tag{II.2}$$

where the dimensionless TF concentration levels, $\boldsymbol{c}(t)$, typically peak in the range $0.001 < c_i < 10$ due to the dynamics (see eq. (II.6)). The binding affinity drops by $\gamma = \frac{1}{2} \ln 10\ k_B T \approx 1.15\ k_B T$ for every mismatching bit, in agreement with experimentally measured mismatch energies of about 1–3 $k_B T$ per nucleotide [15]. The factor $\alpha = 1000$ is required to convert the concentration levels into a realistic range of site occupancies.

The binding state a is defined as a set of occupied binding sites. Its statistical weight, w_a, is a product of the weights of the individual sites and the interactions of all pairs of sites:

$$w_a(\boldsymbol{c}(t)) = \prod_{j \in a} z_j(\boldsymbol{c}(t)) \prod_{j_1 < j_2 \in a} C_{j_1 j_2}, \tag{II.3}$$

where the contribution from the pair of sites j_1, j_2 is

$$C_{j_1 j_2} = \begin{cases} 0, & \text{if exclusive} \\ e^{\beta} \approx 20, & \text{if cooperative} \\ 1, & \text{otherwise.} \end{cases} \tag{II.4}$$

Putting these pieces together, we arrive at an expression for the transcription rate of gene g, as a weighted mean over all binding states:

$$T_g(\boldsymbol{c}(t)) = \frac{\sum_{a \subset A_g} \left(P_g(|a|) w_a(\boldsymbol{c}(t)) \prod_{j \in a} R_j \right)}{\sum_{a \subset A_g} w_a(\boldsymbol{c}(t))}, \tag{II.5}$$

where the sums run over all subsets of A_g, the set of all identified binding sites for gene g, and $|a|$ is the number of sites in subset a. There is no transcription if RNAP binding is blocked by any TF, as R_j is 0 if binding site j is in the repressing region and 1 otherwise.

As an implementation detail, the fact that λ in eq. (II.1) does not depend on the identity of the TF (see [8]) enabled us to implement eq. (II.5) with a dynamical programming algorithm that scales far better with the number of binding sites than a naive enumeration of all $2^{|A_g|}$ states. This algorithm considers the sites in order and tracks the statistical weight of having up to k TFs bound, with the last being at site j; at worst it thus scales as 225 sites times 8 bound TFs, which is crucial when the transcription rate is evaluated at every time step. The algorithm can be extended to discrete TF-specific RNAP binding energies (λ), both positive and negative.

II.2.2 *Network dynamics*

The scheme for transcriptional logic defines transcription rates for all genes, given the TF concentration levels, $\boldsymbol{c}(t)$. To model the time development of these TF levels, transcription and translation are treated as a single step, a simplification motivated by the shorter typical timescale for turnover of mRNA compared to proteins [16]. The time derivative of the protein level for gene g is

$$\frac{dc_g}{dt} = p_g T_g(\boldsymbol{c}(t)) - d_g c_g(t), \tag{II.6}$$

where the transcription rate is modified by a gene-dependent translation efficiency $1 < p_g < 10$, which allows fine-tuning of protein levels. Protein degradation follows mass action kinetics at a protein-specific rate $0.1 < d_g < 10$. We simulated the production and degradation of TFs deterministically, as a set of ordinary differential equations.

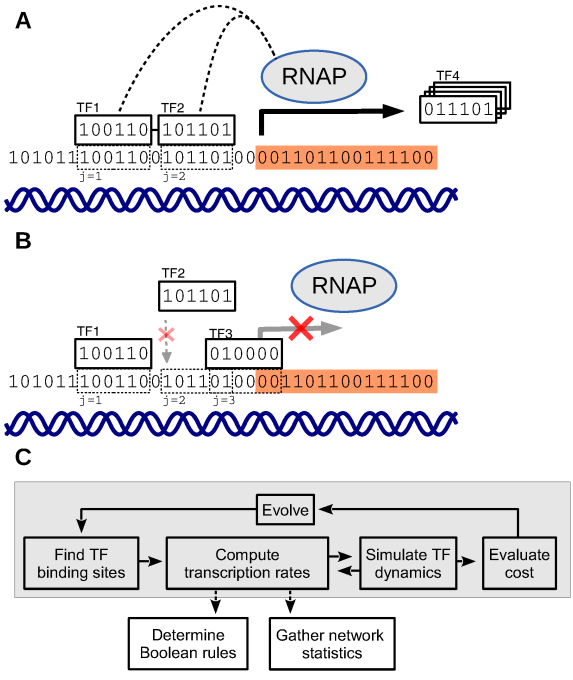

Figure II.1 Transcriptional logic and networks derived from bitstrings. (A-B) A miniature example of possible interactions between 6-bit transcription factors at a 36-bit gene regulatory region. There are in this example binding sites for three different TFs (TF1, TF2, and TF3) which can bind to regulate the transcription rate of a fourth (TF4), whose properties are determined by information outside of the 36 bit regulatory region. The model computes the transcription rate as a mean over *all* possible states of TF–DNA binding, weighted by probability. In the notation of eq. (II.5), $A_g = \{1, 2, 3\}$. (A) Example of activation ($a = \{1, 2\}$ in eq. (II.5)): TF1 and TF2 are bound cooperatively ($C_{12} = e^{\beta}$) and promote RNAP recruitment to the transcriptional start site. (B) Example of repression ($a = \{1, 3\}$): When TF2 is not bound to site $j = 2$, TF3 is free to bind to site $j = 3$ and block the recruitment of RNAP, disabling transcription ($R_3 = 0$). Binding by TF2 and TF3 is mutually exclusive ($C_{23} = 0$). Of all possible binding sites, half lead to repression when a TF blocks RNAP at or downstream of the transcriptional start site (orange region). In the full model, the TFs and DNA are longer than here (32 and 256 bits), and binding is possible also when sequences partially match ($H_{\max}= 6$). (C) The simulation process. Each generation starts with the indexing of TF binding sites for all genes, which defines the transcriptional network. Transcription rates are computed according to eq. (II.5) and TF levels are updated according to eq. (II.6). The resulting dynamics are evaluated by a cost function whose value is used by the selection step of an evolutionary algorithm. When the evolution loop has completed after a number of generations, Boolean rules and other statistics are extracted for further analysis.

II.2.3 *Cost functions*

The fitness of a network was evaluated in terms of its ability to produce a suitable response to a range of inputs. Two different systems were implemented using our model of transcriptional regulation.

In the Boolean majority decision system seven TFs were used only as binary inputs, held at constant low or high levels (0 or 1, respectively). The task of the system was to determine whether a majority of the inputs were high or low; see Figure II.2. A target profile was defined as 0 or 1 according to the majority of input bits, for each of the $2^7 = 128$ combinations of binary inputs. For each combination, we simulated the system for up to 480 time units, or until a fixed point was found. For each gene, the final expression levels were recorded and normalized to the same sum as the the target profile, to which they were then compared. The cost was defined as the mean square deviation of the best matching gene, normalized to 0 for a perfect match and 1 for a flat expression profile.

For the circadian clock system, we used a cost function that strives to focus the expression of a set of genes to specific times of the day. To encourage the emergence of a circadian clock, the network should find the correct timing of gene expression over a range of light conditions [17]. A large number of transcription and degradation rates are light-dependent in models of the plant clock [18]. We therefore simulated the input of light into the system through a 24 h periodic binary signal which selected between two independent sets of degradation rates for all TFs, d_g and d'_g; see Figure II.3.

For each light input, the network dynamics were run for a maximum of 20 days or until convergence to a limit cycle, with 6, 12 or 18 hours of light centred at noon. The expression level of each gene was integrated in six 4 h time windows over 24 h and normalized to a sum of 1. For each time window, the gene with the highest expression when averaged over the three different light conditions was chosen as the output in that window. The cost of the network was one minus the mean of the six output genes in their respective windows. Thus the cost function measured how well the system divided the 24 hours into six equal parts, reminiscent of the consecutively expressed *PRR* genes in the plant circadian clock [19].

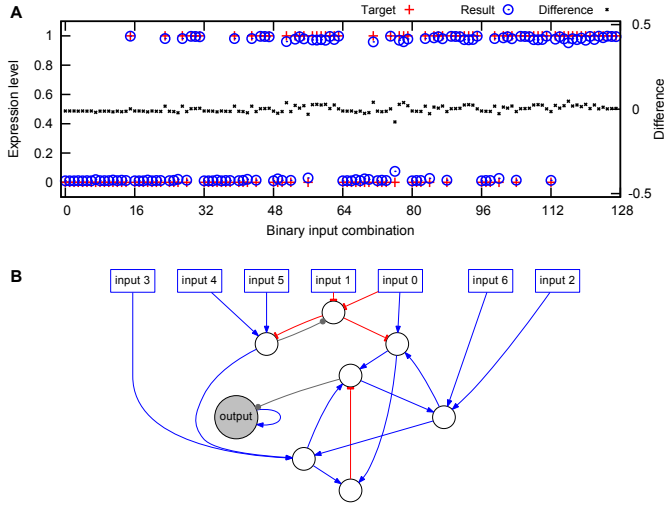

Figure II.2 Evolving networks with majority function. Networks were selected for their ability to determine if the majority of seven binary inputs were on or off. (A) For all 128 possible input combinations, the network output (blue circles) should be as close as possible to the target (red dots), as measured by a cost function based on the deviations (black crosses, right hand y-axis). (B) The evolved network used to generate the output in (A). The input nodes (squares) take binary signals, and the output is the steady state level of the output node (grey). Blue edges with arrows represent activation, red edges with bars represent repression, and grey edges with circles represent ambiguous regulation. This example network was evolved for $2 \cdot 10^6$ generations with a non-zero link cost in order to become suitably small for publication.

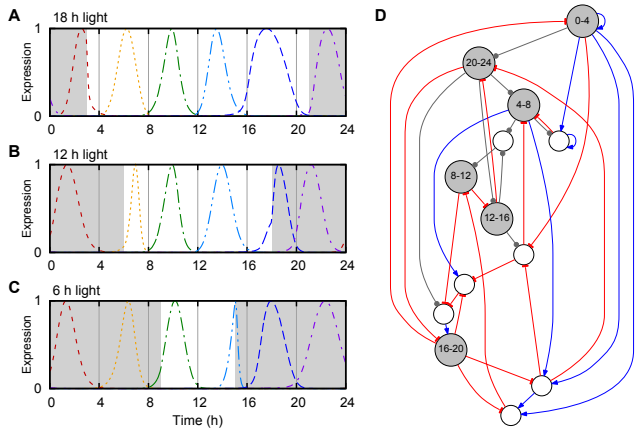

Figure II.3 Evolving networks with clock function. Networks were selected for having each 4 h time window marked by the temporally focused expression of one TF. The six output TFs should be expressed at the correct time in 24 h periodic light conditions with 18 h (A), 12 h (B) and 6 h (C) of light. The coloured lines show the normalized expression levels of the output TFs, and the background is white/grey for light/dark. (D) The evolved network used to generate the output in panels A-C, with the output nodes (grey) labelled by time window. Light acts as input to all nodes through the protein degradation rates. Blue edges with arrows represent activation, red edges with bars represent repression and grey edges with circles represent ambiguous regulation. This example network was evolved for $2.5 \cdot 10^6$ generations with a non-zero link cost in order to be suitably small for publication.

II.2.4 *Evolution of fitness*

To evolve the networks using an evolutionary algorithm, we defined mutations and crossover operations. Possible point mutations included alterations of the bitstrings for TFs and regulatory regions, the TF production rate and degradation rate (p_g and d_g in eq. (II.6)) and the affinity of RNAP for a specific promoter region (b_g in eq. (II.5)). Genes could also be deleted, and new genes were produced by duplication of a whole gene, from a recombination of a regulatory region and a TF, or, more rarely, *de novo* (which may be interpreted as an influx of genes into the system, e.g. from unrelated parts of the organism's genome). The probabilities for the different types of mutations were chosen arbitrarily, and were not expected to greatly affect the results.

The initial network consisted only of a single randomized gene. Instead of imposing a small cost for each additional gene, which may impose an undue pressure on networks early in their evolution, we capped the total number of genes at 40, which is considerably more than the number of genes in, e.g., models of the *Arabidopsis* circadian clock. When

studying the effects of evolution without gene duplication, we started the simulations with 40 random genes and disabled duplication events but allowed deletions and *de novo* gene creation.

For each generation of the evolutionary algorithm, we replaced the least fit individual out of a population of 20 networks. With 90% probability, the best of two randomly chosen individuals was duplicated and mutated. In the remaining cases, two parents were chosen in the same way (tournament selection) and used for crossover, where each gene of the offspring was picked from a random parent. After the final generation, the network with the lowest cost was saved. The networks were evolved for $7 \cdot 10^4$ or $1 \cdot 10^5$ generations for the majority rule or clock cost function, respectively. This process was iterated to create a sample of 100 independently evolved networks for each of our two cost functions. The resulting distributions in cost is shown in Figure II.S1.

To make the networks more comparable with data on real transcriptional networks, we designed a pruning process to remove interactions that left the fitness nearly unchanged. The individual network links were sorted according to the fitness cost of removing them, and links were then removed one at a time until the total change in cost would have exceeded 0.01 (clocks) or 0.001 (majority rule).

II.2.5 *Neutrally evolved networks*

To study the effects of selection, we generated networks without selection for function but with the same structural characteristics. The in-degree of a node (gene) was defined as the number of distinct TFs that bound to its regulatory region, and the out-degree was similarly defined as the number of distinct genes to which a TF bound. We constructed a cost function which compared two networks, such that a value of zero corresponded to the new network having the same number of nodes and edges, and the same distributions of in-degree and out-degree as the target network. Aside from this selection towards similar structure, the networks were allowed to evolve neutrally, using the same mutation mechanisms as the functional networks. From each functional network, we created 5 neutrally evolved networks.

II.2.6 *Extracting Boolean rules*

The interactions of TFs that bind to a gene regulatory region result in a multivariate function, with input and output levels that need to be discretized if we are to extract a Boolean representation of the

transcriptional logic rule. To build the truth table of a k-input rule, we applied all 2^k combinations of high and low input TF levels, where "low" was defined as 0 and "high" as the peak concentration level of the respective input TF in the dynamics. The gene was considered to be on for combinations where it reached at least half of its peak transcription rate observed in the dynamics, and off otherwise.

Many of the possible Boolean truth tables describe rules that do not depend on all of their inputs. This is a general problem when discretizing expression levels: TFs that only weakly affect the transcription rate may not be sufficient to push the output across the binarization threshold in any of the input conditions. For the analysis of Boolean rules, we removed all such "unused" inputs.

II.3 RESULTS

As described in Methods, we evolved functional networks with selection either for performing the Boolean majority rule (Figure II.2) or for circadian clock function (Figure II.3), with and without gene duplications as a possible mutation step, and compared these networks with structurally similar but non-fuctional networks created by neutral evolution. The functional networks showed considerable variation in fitness, with gene duplications improving the rate of convergence towards the selection target for the same number of generations (Figure II.S1). As expected, selection for network function led to an enrichment in strong TF–DNA binding relative to random sequences, particularly for circadian clocks (Figure II.S2).

II.3.1 *Low ambiguity of transcriptional regulation*

A TF may act as an activator at one binding site and as a repressor at another, even within the same regulatory region. We refer to this case as ambiguous regulation of the target gene. The *E. coli* transcriptional network database RegulonDB includes information on this level, with nearly all TF–DNA interactions described as either activating or repressing. Disregarding a small number of binding sites with unknown or dual function, we could thus compare the prevalence of ambiguous regulation between the model and the *E. coli* data.

Considering only cases where a TF bound to exactly two sites in a regulatory region, we defined n_{++}, n_{--} and n_{+-} as the number of activators, repressors and ambiguous regulators, respectively. If the sign

of regulation is random, we expect that the ratio $n_{+-}/(2\sqrt{n_{++}n_{--}}) = 1$. As shown in Figure II.4, this relative ambiguity was indeed close to 1 in neutrally evolved networks. In networks evolved with selection, the ratio was much lower, between 0.2 and 0.3, regardless of the network function and whether gene duplications were allowed or not. This trend is in qualitative agreement with data from the *E. coli* network, where the relative ambiguity was only about 0.04.

The relative ambiguity was decreased by pruning unimportant links in the networks as described in Methods, whereas altering the TF–DNA binding cutoff to include many weaker binding sites (setting $H_{\max}$=8) had the opposite effect (Figure II.S3). Strong and important links were thus associated with lower ambiguity, which could explain why we observed so few ambiguous interactions among those that have been found worthy of study and inclusion in RegulonDB.

II.3.2 *Binding site interactions*

Depending on how they interact, every pair of binding sites in a regulatory region may be classified as competitive, cooperative or independent (Figure II.1 and eq. (II.4)). A further division can be made into pairs of sites where either identical or different TFs bind. In the case of cooperative binding, these homogeneous and heterogeneous pairs of sites may represent binding by homo- and heterodimers, respectively.

Regardless of the selection target used to evolve functional networks, competitive binding was considerably less likely between homogeneous pairs of binding sites than between heterogeneous ones (Figure II.5A). It appears that networks have little use for components that directly counteract themselves. Conversely, cooperative binding was most likely between identical TFs (Figure II.5B). Homodimer-like regulators were thus particularly favoured, but a comparison with the random expectation shows that cooperativity caused by heterodimer-like regulators was also significantly overrepresented.

To test these model predictions against data from a real transcriptional network, we collected statistics on the distance between midpoints of binding sites in *E. coli* from RegulonDB. For simplicity and comparability with the model, we classified sites within 16 basepairs as competitive and those within 32 basepairs as cooperative. As shown in Figure II.5, a comparison between heterogeneous and homogeneous pairs of binding sites verified the model prediction that cooperative binding is preferentially associated with homodimers.

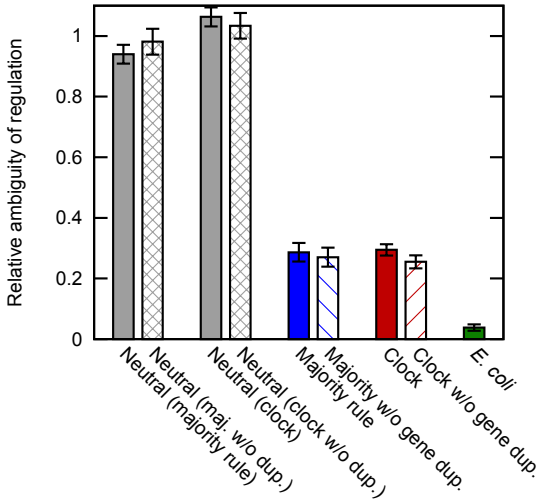

Figure II.4 Prevalence of ambiguous gene regulation. The probability of a TF having exactly one activating and one repressing binding site in a regulatory region, relative to the null hypothesis that the sign of regulation is independent between sites (see main text). In neutrally evolved networks (gray), regulation was as ambiguous as expected by chance. In contrast, genes in functional evolved networks were predominantly regulated unambiguously, regardless of the network function (blue and red for majority rule and clock networks, respectively). Restricting the evolutionary paths by disabling gene duplications had little effect on this ambiguity (hashed bars). The model qualitatively predicted the situation in the transcriptional network of *E. coli* (green), where TFs almost always regulate their targets with a clearly defined sign. Binding site data were pooled from all networks; error bars indicate standard errors based on the total counts. See also Figure II.S3, which explores the effects of some model parameter choices on the relative ambiguity.

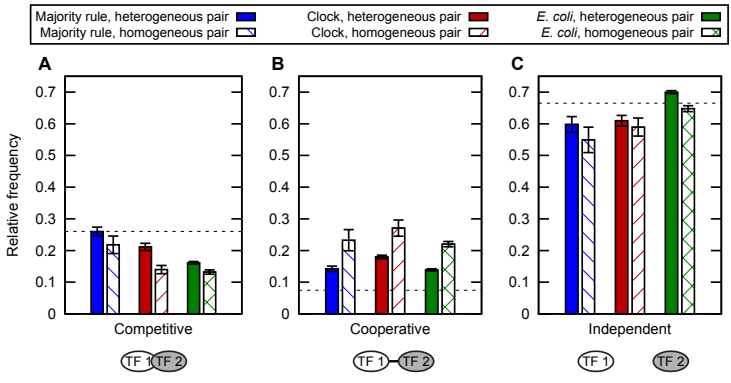

Figure II.5 Pairwise interactions of TF binding sites. The fraction of pairs of TF binding sites that were (A) mutually exclusive because of competitive, overlapping binding, (B) cooperatively binding due to proximity, or (C) neither. Binding sites were defined as specific to one TF and may thus overlap completely on the DNA sequence. We further distinguished between homogeneous interaction (TF1 and TF2 are the same) and heterogeneous (TF1 and TF2 are different). Horizontal black dotted lines show the expectation values for random DNA and TF sequences. All pairs of binding sites (within the same regulatory region) were counted in 100 networks evolved as either majority rule (blue) or clock (red). *E. coli* data from RegulonDB [20] (green). Error bars indicate standard errors.

II.3.3 *Dominant sign of regulation*

When a TF could bind to multiple sites at a gene, the regulation of that gene was predominantly either activating or repressing. This preference for a clear sign of regulation at the level of individual genes does not necessarily mean that TFs regulate all their targets in equal direction. We hypothesized that TFs may be divided into activators and repressors by selection if such a division constitutes a design principle for successful transcriptional networks.

We define the activator-repressor status of a TF as the fraction of a TF's individual binding sites that are classified as activating. When all binding sites for a TF are considered, the expectation with random DNA and TF sequences is that the activator-repressor status follows a binomial distribution; it is rare to see mostly activation or repression by chance (Figure II.6A).

The activator-repressor status of TFs in networks evolved neutrally without gene duplication was found to closely follow the expected binomial distribution. Similar results for networks selected for function indicated that selection did not encourage the separation of TFs into activators and repressors. The only deviation from the random expectation was a small bias towards negative regulation in clock networks (Figure II.6A).

When gene duplication events were allowed, selection for either of the two network functions produced networks where the TFs were clearly separated into activators and repressors (Figure II.6B). Neutral evolution produced similar results, which suggested that the separation between activators and repressors was not selected for, but rather a consequence of evolution following paths opened up by gene duplications. The results from the simulations were remarkably similar to data from *E. coli*. A majority of the TFs still had some binding sites of opposing signs, but pure activators and pure repressors were far more abundant than expected by chance (Figure II.6C).

As before, clock networks showed a preference for negative regulators, while in the *E. coli* data we could see an overrepresentation of TFs without any activating interactions; these may represent inherent inhibitors of transcription.

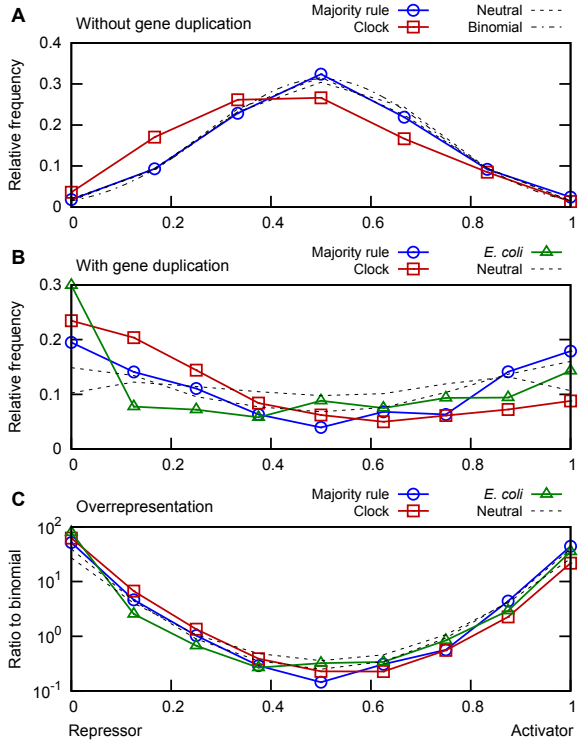

Figure II.6 TF proclivity towards positive or negative regulation. For each TF, we computed the fraction of binding sites that enhanced rather than blocked transcription. The distribution of this quantity is presented for functional networks evolved either for majority rule or clock function (blue squares and red squares, respectively), compared with structurally similar neutrally evolved networks (black dashed lines) and with *E. coli* data [20] (green triangles) where applicable. (A) Without gene duplication, both functional and neutrally evolved networks followed the binomial distribution expected when individual target genes are randomly regulated (black dot-dashed line). A small bias towards negative regulation was observed in clock networks. (B) With gene duplications, TFs separated into mostly activating or repressing. The same pattern was observed in *E. coli*. (C) The relative overrepresentation of positive/negative regulators in functional networks, shown as the ratio between the data from panel B and the expectation for random networks. The number of pure activators or repressors was twentyfold to hundredfold higher than expected by chance, both in the simulated networks and in *E. coli*. The graphs are based on all TFs with at least 6 (A) or 8 (B-C) binding sites, with rebinning applied to those with additional sites. The lower limit in (A) is due to a scarcity of highly connected TFs in the majority rule networks; aside from statistical noise, the number of bins does not affect the shape of the curves. Data from 100 functional and 500 neutrally evolved networks of each kind.

II.3.4 *Transcriptional logic*

The function of a transcriptional network is determined not only by its structure but also by the logic rules that govern the expression of its genes. These rules are defined by the binding and mutual interactions of TFs, and the transcriptional logic is thus constrained: only a subset of the Boolean functions can practically be realized [8]. As we have shown, our model predicts that the pairwise interactions between TFs follow similar patterns in networks evolved to perform different tasks. However, we would expect the resulting logic to differ.

The transcription rates, which are described by eq. (II.5) as continuous functions of the TF concentrations levels, were discretized into Boolean functions based on typical expression levels observed in the dynamics of the system. This procedure, which is explained in further detail in Methods, works only for functional networks; the neutrally evolved networks lack meaningful dynamics, and the resulting rules would depend strongly on arbitrary assumptions about expression levels. Hence, we have only compared Boolean rules extracted from functional evolved networks.

Binary Boolean rules such as AND and OR accounted for about 20% of the roughly 3000 multivariate rules extracted from each ensemble of networks (Figure II.7C). The distribution of these rules is shown in Figure II.7A, from which it can be seen that the presence or absence of gene duplication had little effect on the types of rules formed by rewiring of the transcriptional regulation. In contrast, the selection target strongly affected the number of rules of the most common types.

The most common logic rules were AND-like functions that tie together the presence of activators and the absence of repressors. The corresponding OR-like functions were relatively uncommon, accounting only for a few percent of the total. None of the evolved networks contained any cases of EQ or XOR, though tests with random sequences showed that such rules are possible to express in this model (not shown). The six binary rules that were realized are all *canalizing*, i.e., they have at least one input value that renders the other inputs irrelevant for determining the output. When the canalizing input is set to its non-canalizing value, the rule may be canalizing on additional inputs. Rules that are defined completely by such recursion are referred to as *nested canalizing*. These occur frequently in biology and generally lead to stable network dynamics [21].

A majority of the genes were regulated by more than two TFs. To characterize the resulting Boolean functions, we determined their *nested canalizing depth*. This number is maximal for nested canalizing rules such as $a \wedge (b \vee (c \wedge \neg d))$ but zero for non-canalizing rules [22]. Two patterns emerged, one concerning biological function and one concerning evolutionary mechanisms: Clocks make use of more canalization than networks that solve the majority rule problem, and gene duplications favour the emergence of canalizing rules; see Figure II.7B.

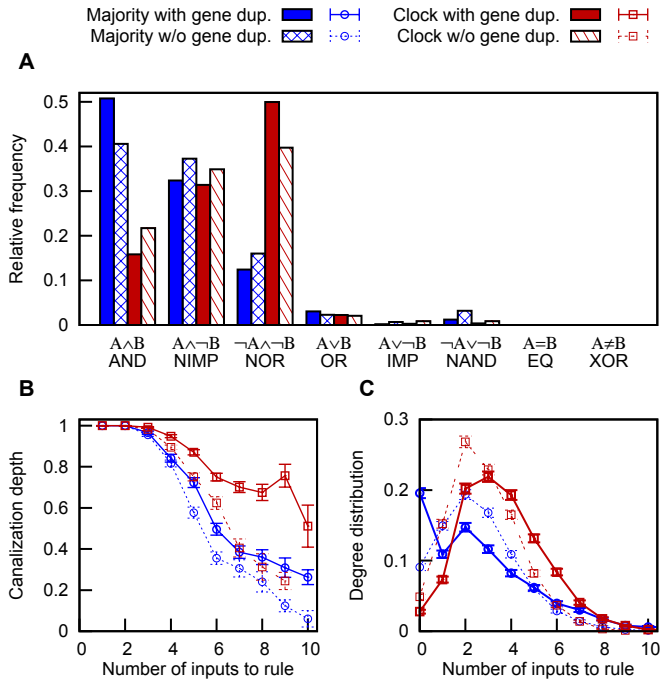

Figure II.7 Transcriptional logic of evolved networks. Comparison between the logic rules in majority rule networks (blue) and clock networks (red), evolved with (solid) or without (hashed/dashed) gene duplications. (A) The relative frequencies among the eight rules that have only two inputs. (B) The structure of logic rules with up to 9 inputs, computed as the mean *nested canalizing depth* [22] normalized by the number of inputs. (C) The degree distribution of panel B. All three panels indicate that network function is a major factor in determining the distribution of transcriptional logic rules. The transcription rates, which were modelled as continuous functions of TF concentration levels, were discretized into Boolean logic rules as described in Methods. Data from 100 networks of each kind.

II.4 DISCUSSION

By applying *in silico* evolution to a combinatorial model of transcriptional regulation, we have explored how mutation and selection together shape interactions between genes in transcriptional networks. We have compared two specific network functions with the transcriptional network of an entire organism, partly for practical reasons but motivated by the philosophy that grand principles should be generic and robust.

The model demonstrates that TFs separate into activators and repressors even when they carry no inherent predisposition towards either sign of regulation. There are two aspects to this. First, the regulation of an individual gene by a TF is mostly unambiguously positive or negative; ambiguous regulation is relatively rare in nature as well as in our simulations of functional networks. Presumably, ambiguity leads to weak interactions that are unlikely to persist. Second, the model reproduces the observed broad separation of TF function into predominantly activating or repressing, but only in simulations of evolution that include gene duplications. These two points differ in one important regard: The former is specific to networks with functional dynamics, whereas the latter is also found in neutrally evolved networks. However, both are surprisingly insensitive to the choice of network function used as the target for selection.

In contrast, we found that the function of a network largely determined its distribution of logic rules, regardless of large-scale mutations such as gene duplications. Networks that were evolved to act as circadian clocks depended on negative interactions and distinctly canalizing logic rules. Furthermore, they were rich in cooperative binding and strong repression, in agreement with the view that oscillations are favoured by negative and nonlinear regulation of gene expression [23]. The importance of repression in circadian clocks is corroborated by modelling work in organisms such as *Arabidopsis thaliana* [24, 25]. In contrast, networks evolved to solve a simple decision problem made use of less canalizing logic and showed no bias towards repression.

Despite these differences, we observed strong similarities between all functional networks compared with neutrally evolved ones. This indicates that networks evolved with selection adhere to certain "design principles" for transcriptional regulation. TFs regulate their targets with less ambiguity and a higher incidence of cooperative binding than expected by chance, but the most striking feature is the polarization of the sign of regulation.

Four main conclusions can thus be drawn:

- The model produces testable hypotheses that agree with data from *E. coli*, such as the overrepresentation of cooperative binding among homogeneous pairs of binding sites (Figure II.5).
- Selection leaves clear marks on gene regulation in functional networks; neutrally evolved networks do not regulate their genes in a coherent, unambiguous way (Figure II.4).
- The evolutionary shortcuts created by gene duplications favour the approximate division of TFs into activators and repressors (Figure II.6).
- The choice of target function for network selection has impact on the resulting logic rules, whereas the choice of allowed evolutionary paths appears to be of less importance; for the clock system, nested canalizing rules are dominant (Figure II.7).

The importance of selection to the transcriptional regulation contrasts with earlier results on local structure by Kuo *et al.*, who found that the prevalence of network motifs depended only on gene duplication and was unaffected by selection [3]. On the other hand, Kashtan *et al.* reported that selection did affect motif formation in modular information-processing networks compared with networks evolved only for modularity [26]. These conflicting results seem to be sensitive to model assumptions that affect the evolved non-functional networks. We expect greater robustness when conclusions are based on comparisons between functional networks, evolved either for different biological functions or using different mutational steps.

The effects of gene duplication on TF specialization requires further study. It appears that when target genes are duplicated, TFs retain their initial bias towards specialization despite subsequent rewiring of binding sites. Further work will be needed to clarify the robustness of this result. For instance, by tracking the evolutionary history of every binding site, we could quantify the extent and impact of the rewiring.

In the model presented here, transcriptional binding sites arise from a bit-based genotype, without prior assumptions about how the genes are connected. Network structure and function are created by the combinatorial interactions between transcriptional regulators. Models that include such a dynamical mapping from genotype to phenotype are well suited to mimic the complexity of natural systems and explore

the mutation space to improve fitness [7]. The present model describes the core parts of transcriptional regulation in a simplified form, which is robust towards changes in model parameters (see e.g. Figure II.S4). Depending on the context, the model could be extended to include explicit RNA and translation, as well as post-translational interactions such as protein complex formation. The model represents a simplified picture compared to the situation in *E. coli*, where repressors may bind upstream of the TSS [14], and to generic models that explicitly include longer-range interactions [8]. Future developments may thus include nonlocal interactions (DNA looping) as well as more TF-specific binding strengths and cooperative interactions.

ACKNOWLEDGMENTS

We thank Patrik Edén and Adriaan Merlevede for fruitful suggestions.

REFERENCES

1. 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.
2. U. Alon, "Biological networks: the tinkerer as an engineer," *Science*, vol. 301, no. 5641, pp. 1866–1867, 2003.
3. 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.
4. 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.
5. T. R. Sorrells and A. D. Johnson, "Making sense of transcription networks," *Cell*, vol. 161, no. 4, pp. 714–723, 2015.
6. S. A. Kauffman, "Metabolic stability and epigenesis in randomly constructed genetic nets," *J. Theor. Biol.*, vol. 22, no. 3, pp. 437–467, 1969.
7. W. Banzhaf, "On the dynamics of an artificial regulatory network," in *Advances in Artificial Life*, pp. 217–227, Springer, 2003.
8. 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.

9. H. G. McWatters and P. F. Devlin, "Timing in plants—a rhythmic arrangement," *FEBS Lett.*, vol. 585, no. 10, pp. 1474–1484, 2011.
10. P. François and V. Hakim, "Design of genetic networks with specified functions by evolution in silico," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 101, no. 2, pp. 580–585, 2004.
11. S. Paladugu, V. Chickarmane, A. Deckard, J. Frumkin, M. McCormack, and H. Sauro, "In silico evolution of functional modules in biochemical networks," *IEE Proceedings-Systems Biology*, vol. 153, no. 4, pp. 223–235, 2006.
12. A. Traven, B. Jelicic, and M. Sopta, "Yeast Gal4: a transcriptional paradigm revisited," *EMBO reports*, vol. 7, no. 5, pp. 496–499, 2006.
13. J. F. Margolin, J. R. Friedman, W. K. Meyer, H. Vissing, H.-J. Thiesen, and F. J. Rauscher, "Krüppel-associated boxes are potent transcriptional repression domains," *Proc. Natl. Acad. Sci. USA*, vol. 91, no. 10, pp. 4509–4513, 1994.
14. M. M. Babu and S. A. Teichmann, "Functional determinants of transcription factors in *Escherichia coli*: protein families and binding sites," *Trends Genet*, vol. 19, no. 2, pp. 75–79, 2003.
15. U. Gerland, J. D. Moroz, and T. Hwa, "Physical constraints and functional characteristics of transcription factor–DNA interaction," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 99, no. 19, pp. 12015–12020, 2002.
16. B. Schwanhäusser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, and M. Selbach, "Global quantification of mammalian gene expression control," *Nature*, vol. 473, no. 7347, pp. 337–342, 2011.
17. C. Troein, J. C. W. Locke, M. S. Turner, and A. J. Millar, "Weather and seasons together demand complex biological clocks," *Curr. Biol.*, vol. 19, no. 22, pp. 1961–1964, 2009.
18. A. Pokhilko, P. Más, and A. J. Millar, "Modelling the widespread effects of TOC1 signalling on the plant circadian clock and its outputs," *BMC Syst Biol*, vol. 7, no. 1, pp. 1–12, 2013.
19. 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.
20. H. Salgado, S. Gama-Castro, M. Peralta-Gil, A. Santos-Zavaleta, L. Muniz-Rascado, H. Solano-Lira, V. Jimenez-Jacinto, V. Weiss,

J. S. Garcia-Sotelo, A. Lopez-Fuentes, L. Porron-Sotelo, S. Alquicira-Hernandez, A. Medina-Rivera, I. Martinez-Flores, K. Alquicira-Hernandez, R. Martinez-Adame, C. Bonavides-Martinez, J. Miranda-Rios, A. M. Huerta, A. Mendoza-Vargas, L. Collado-Torres, B. Taboada, L. Vega-Alvarado, M. Olvera, L. Olvera, R. Grande, E. Morett, and J. Collado-Vide, "RegulonDB v8.0: Omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more," *Nucleic acids research*, vol. 41, no. D1, pp. D203–D213, 2013.

21. S. Kauffman, C. Peterson, B. Samuelsson, and C. Troein, "Genetic networks with canalyzing boolean rules are always stable," *Proc. Natl. Acad. Sci. USA*, vol. 101, no. 49, pp. 17102–17107, 2004.
22. L. Layne, E. Dimitrova, and M. Macauley, "Nested canalyzing depth and network stability," *Bull. Math. Biol.*, vol. 74, no. 2, pp. 422–433, 2012.
23. M. B. Elowitz and S. Leibler, "A synthetic oscillatory network of transcriptional regulators," *Nature*, vol. 403, no. 6767, pp. 335–338, 2000.
24. I. Carré and S. R. Veflingstad, "Emerging design principles in the *Arabidopsis* circadian clock," *Semin. Cell. Dev. Biol.*, vol. 24, no. 5, pp. 393–398, 2013.
25. K. Fogelmark and C. Troein, "Rethinking transcriptional activation in the *Arabidopsis* circadian clock," *PLoS Comput. Biol.*, vol. 10, no. 7, p. e1003705, 2014.
26. N. Kashtan and U. Alon, "Spontaneous evolution of modularity and network motifs," *Proc. Natl. Acad. Sci. USA*, vol. 102, no. 39, pp. 13773–13778, 2005.

II.A SUPPLEMENTARY FIGURES

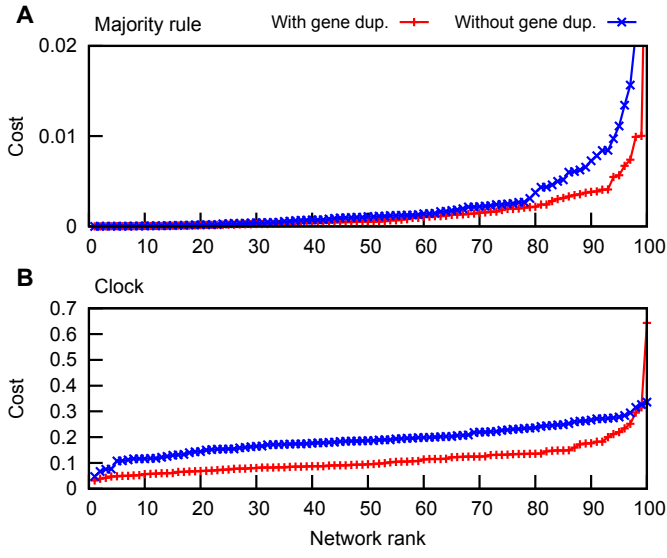

Figure II.S1 Distribution of fitness in evolved networks. Final cost function values for networks evolved with or without gene duplication events, sorted by cost. (A) Majority rule networks evolved for $7 \cdot 10^4$ generations. (B) Clock networks evolved for $1 \cdot 10^5$ generations. Data from 100 simulations of each kind.

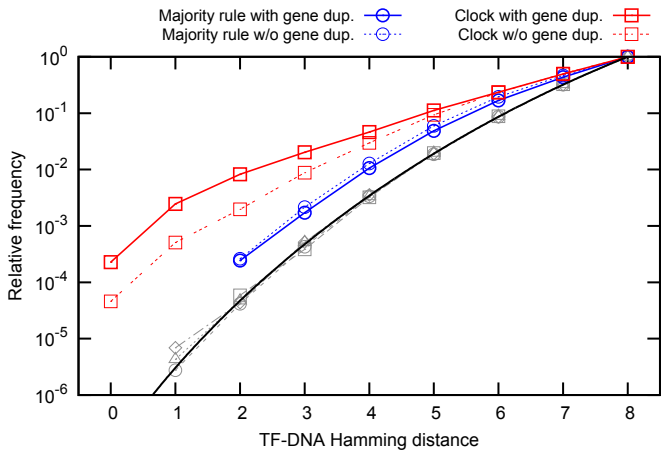

Figure II.S2 Distribution of TF–DNA binding strengths. The relative frequency of Hamming distances between TF binding motifs and DNA sequences, for all binding sites in networks evolved at binding strength cutoff $H_{\max}= 8$. Data from 100 networks selected for function as the majority rule (blue circles) or circadian clock (red squares), either with or without gene duplications (thick solid and thin dashed, respectively). The enrichment in strong binding sites is due to selection; this is demonstrated by a comparison with 500 non-functional networks with similar structure (gray symbols and lines), which match the binomial distribution expected for random sequences (solid black line).

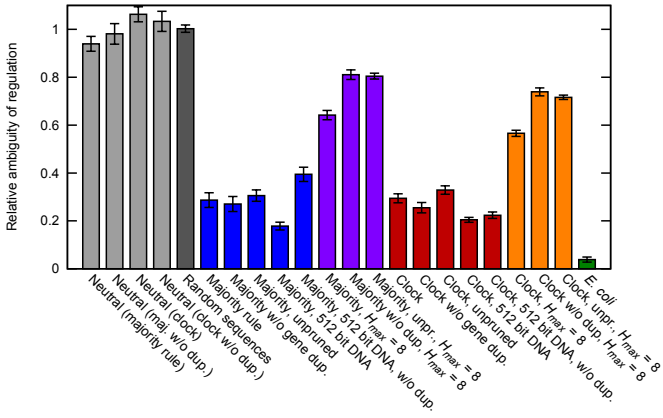

Figure II.S3 Prevalence of ambiguous gene regulation. The probability of a TF having exactly one activating and one repressing binding site in a regulatory region, relative to the probability under the null hypothesis that the sign of regulation is independent between sites. This figure expands on Figure II.4 by including four model choices for neutrally evolved networks (gray), random regulatory regions and binding sequences (dark gray), majority rule networks (blue) evolved either normally, without gene duplications, without the pruning of unimportant interactions or with a larger regulatory region (512 bits), majority rule networks evolved with weaker binding mismatch cutoff ($H_{\max}= 8$) with or without gene duplications or pruning, circadian clock networks for the same cases as the majority rule networks (red and orange) and data for *E. coli*. Unless otherwise stated, the model used $H_{\max}= 6$ and 256-bit regulatory regions. As expected, $H_{\max}= 8$ resulted in more ambiguous regulation due to a larger number of weak binding sites; this is reflected in the greater effect of pruning these networks. Increasing the size of the regulatory region leads to a decrease in ambiguity in the clock networks, which suggests that clocks sometimes use one negative and one positive binding site to implement their preferred cooperative negative regulation. Error bars indicate standard errors.

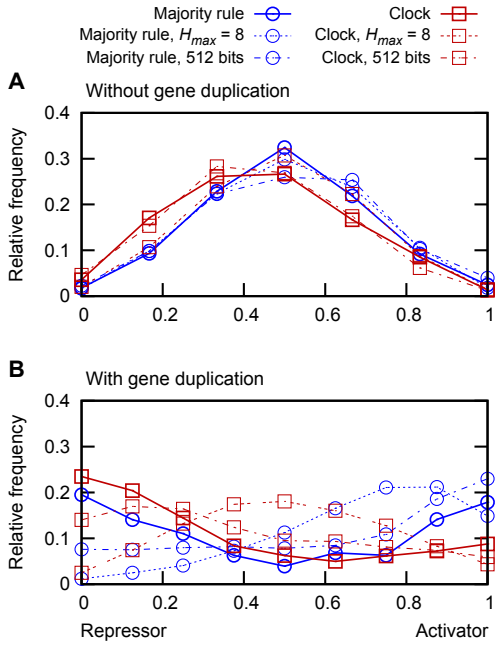

Figure II.S4 TF proclivity towards positive or negative regulation. This figure is identical to Figure II.6A-B ($H_{\max}$= 6, 256 bits), except that we here include data from networks with twice as large (512 bit) regulatory regions (dashed lines), or simulations with lower binding strength cutoff ($H_{\max}$= 8, dot-dashed lines). The inclusion of weaker binding sites shifts the results towards the random expectation, but otherwise the results are qualitatively unchanged by these parameter changes. Data from 100 functional and 500 neutrally evolved networks of each kind.

PAPER III

Rethinking transcriptional activation in the Arabidopsis circadian clock

Karl Fogelmark and Carl Troein

Computational Biology and Biological Physics, Department of Astronomy and Theoretical Physics, Lund University, 223 62 Lund, Sweden.

PLoS Computational Biology, vol. **10**, e1003705 (2014)
doi: 10.1371/journal.pcbi.1003705

Circadian clocks are biological timekeepers that allow living cells to time their activity in anticipation of predictable daily changes in light and other environmental factors. The complexity of the circadian clock in higher plants makes it difficult to understand the role of individual genes or molecular interactions, and mathematical modelling has been useful in guiding clock research in model organisms such as *Arabidopsis thaliana*.

We present a model of the circadian clock in *Arabidopsis*, based on a large corpus of published time course data. It appears from experimental evidence in the literature that most interactions in the clock are repressive. Hence, we remove all transcriptional activation found in previous models of this system, and instead extend the system by including two new components, the morning-expressed activator RVE8 and the nightly repressor/activator NOX.

Our modelling results demonstrate that the clock does not need a large number of activators in order to reproduce the observed gene expression patterns. For example, the sequential expression of the PRR genes does not

require the genes to be connected as a series of activators. In the presented model, transcriptional activation is exclusively the task of RVE8. Predictions of how strongly RVE8 affects its targets are found to agree with earlier interpretations of the experimental data, but generally we find that the many negative feedbacks in the system should discourage intuitive interpretations of mutant phenotypes. The dynamics of the clock are difficult to predict without mathematical modelling, and the clock is better viewed as a tangled web than as a series of loops.

III.1 AUTHOR SUMMARY

Like most living organisms, plants are dependent on sunlight, and evolution has endowed them with an internal clock by which they can predict sunrise and sunset. The clock consists of many genes that control each other in a complex network, leading to daily oscillations in protein levels. The interactions between genes can be positive or negative, causing target genes to be turned on or off. By constructing mathematical models that incorporate our knowledge of this network, we can interpret experimental data by comparing with results from the models. Any discrepancy between experimental data and model predictions will highlight where we are lacking in understanding. We compiled more than 800 sets of measured data from published articles about the clock in the model organism thale cress (*Arabidopsis thaliana*). Using these data, we constructed a mathematical model which compares favourably with previous models for simulating the clock. We used our model to investigate the role of positive interactions between genes, whether they are necessary for the function of the clock and if they can be identified in the model.

III.2 INTRODUCTION

The task of the circadian clock is to synchronize a multitude of biological processes to the daily rhythms of the environment. In plants, the primary rhythmic input is sunlight, which acts through photoreceptive proteins to reset the phase of the clock to local time. The expression levels of the genes at the core of the circadian clock oscillate due to mutual transcriptional and post-translational feedbacks, and the complexity of the feedbacks makes it difficult to predict and understand the response of the system to mutations and other perturbations without the use of mathematical modelling [1].

Early modelling of the system by Locke *et al.* demonstrated the feasibility of gaining new biological insights into the clock through the use of model predictions [2]. The earliest model described the system as a negative feedback loop between the two homologous MYB-like transcription factors CIRCADIAN CLOCK ASSOCIATED 1 (CCA1) and LATE ELONGATED HYPOCOTYL (LHY) [3, 4] on one hand and TIMING OF CAB EXPRESSION 1 (TOC1/PRR1) [5] on the other. Over the past decade, models have progressed to describing the system in terms of multiple interacting loops, still centred around LHY/CCA1 (treated as one component) and TOC1. The latest published model by Pokhilko *et al.* (2013) describes transcriptional and post-translational interactions between more than dozen components. We refer to that model as P2012 [6], in keeping with the tradition of naming the *Arabidopsis* clock models after author and submission year (cf. L2005 [2], L2006 [7], P2010 [8] and P2011 [9]).

The clock depends on several genes in the PSEUDO RESPONSE REGULATOR (PRR) family: *PRR9*, *PRR7*, *PRR5*, *PRR3* and *TOC1/PRR1* are expressed in a clear temporal pattern, with *PRR9* mRNA peaking in the morning, *PRR7* and *PRR5* before and after noon, respectively, and *PRR3* and *TOC1* near dusk [10]. PRR9, PRR7 and PRR5 act to repress expression of *CCA1* and *LHY* during the day [11], but, until recently, TOC1 was thought to be a nightly activator of *CCA1* and *LHY*, acting through some unknown intermediate. However, TOC1 has firmly been shown to be a repressor of both *CCA1* and *LHY*, and it now takes its place in the models as the final repressor of the "PRR wave" [9, 12–14]. PRR3 has yet to be included in the clock models and the roles of the other PRRs are being reevaluated following the realization that TOC1 acts as a repressor [15].

The GIGANTEA (GI) protein has long been thought to form part of the clock [16], whereas EARLY FLOWERING 3 (ELF3) was known to affect clock function [17] but was only more recently found to be inside the clock, rather than upstream of it [18, 19]. GI and ELF3 interact with each other and with other clock-related proteins such as the E3 ubiquitin-ligase COP1 [20]. GI plays an important role in regulating the level and activity of ZEITLUPE (ZTL) [21], which in turn affects the degradation of TOC1 [22] and PRR5 [23] but not of the other PRRs [24]. The clock models by Pokhilko *et al.* include GI and ZTL; GI regulates the level of ZTL by sequestering it in a GI-ZTL complex during the day and releasing it at night [8].

Together with EARLY FLOWERING 4 (ELF4) and LUX ARRHYTHMO (LUX), ELF3 is necessary for maintaining rhythmicity in the clock [25–27]. The three proteins are localized to the nucleus, and ELF3 is both necessary and sufficient for binding ELF4 and LUX into a complex termed the evening complex (EC) [19]. In recent models, EC is a major repressor; it was introduced in P2011 to repress the transcription of *PRR9*, *LUX*, *TOC1*, *ELF4* and *GI* [9].

We here present a model (F2014) of the circadian clock in *Arabidopsis*, extending and revising the earlier models by Pokhilko *et al.* (P2010–P2012). To incorporate as much as possible of the available knowledge about the circadian clock into the framework of a mathematical model, we have compiled a large amount of published data to use for model fitting. These curated data are made available for download as described in Methods.

The aim of this work is to clarify the role of transcriptional activation in the *Arabidopsis* circadian clock. Specifically, we use modelling to test whether the available data are compatible with models with and without activation. There is no direct experimental evidence for any of the activators postulated in earlier models, and as a crucial step in remodelling the system we have removed all transcriptional activation from the equations. Instead, we have added a major clock component missing from earlier models: the transcription factor REVEILLE 8 (RVE8), which positively regulates the expression of a large fraction of the clock genes [28, 29]. A further addition is the nightly transcription factor NOX/BROTHER OF LUX ARRHYTHMO (NOX/BOA), which is similar to LUX but may also act as an activator of *CCA1* [30]. By examining transcriptional activation within the framework of our model, we have clarified the relative contributions of the activators to their different targets.

III.3 RESULTS

Based on available experimental data and interpretations in the published literature, we have developed a revised model of the *Arabidopsis* circadian clock. The new model is presented in Figure III.1, and a comparison with the most recently published model, P2012 [6], is shown in Figure III.S1. Five major alterations are discussed below: remodelling of EC, addition of the LUX homologue NOX, removal of sequential activation in the PRR wave, repression of the PRRs by CCA1, and ad-

dition of RVE8 as the main transcriptional activator. For brevity, we refer to Section III.B for further details and results concerning nuclear localization of TOC1 by PRR5, splitting of LHY/CCA1 and removal of unmotivated components and light inputs.

To increase the robustness of the conclusions drawn from the modelling, all our model simulations are presented as eight curves, derived from an ensemble of eight independent parameter sets as described in Methods.

III.3.1 *A remodelled evening complex*

Overexpression of ELF3 rescues clock function in the otherwise arrythmic *elf4-1* mutant [27]. This suggests that the function of ELF4 is to amplify the effects of ELF3 through the ELF3-ELF4 complex, which led us to consider an evening complex (EC) where free ELF3 protein can play the role of ELF3-ELF4, albeit with highly reduced efficacy. This, together with our aim to add the NOX protein in parallel with LUX, as described in the next section, prompted us to rethink how to model this part of the clock.

EC is not given its own variable in the differential equations, unlike in the earlier models. Instead, EC activity is seen as rate-limited by LUX and NOX on one hand and by ELF3-ELF4 and free ELF3 on the other. In either pair, the first component is given higher importance, in accordance with previous knowledge. For details, see the equations in Section III.C. This simplified description requires few parameters, which was desirable because the model had to be constrained using time course data for the individual components of EC, mainly at the mRNA level.

The effects of our changes to EC are illustrated in Figure III.2, which shows EC and related model components in the transition from cycles of 12 h light, 12 h dark (LD 12:12) to constant light (LL). ELF3, which is central to EC in our model, behaved quite differently at the mRNA level compared with the P2011 and P2012 models, and more closely resembled the available experimental data, with a broad nightly peak and a trough in the morning at zeitgeber time (ZT) 0–4 (Figure III.2A).

The differences in the dynamics of the EC components between our eight parameter sets demonstrate an interesting and more general point: The components that are most reliably constrained are not always those that were fitted to measured data. In our case, the model was fitted to data for the amount of *ELF3* mRNA (Figure III.2A) and total

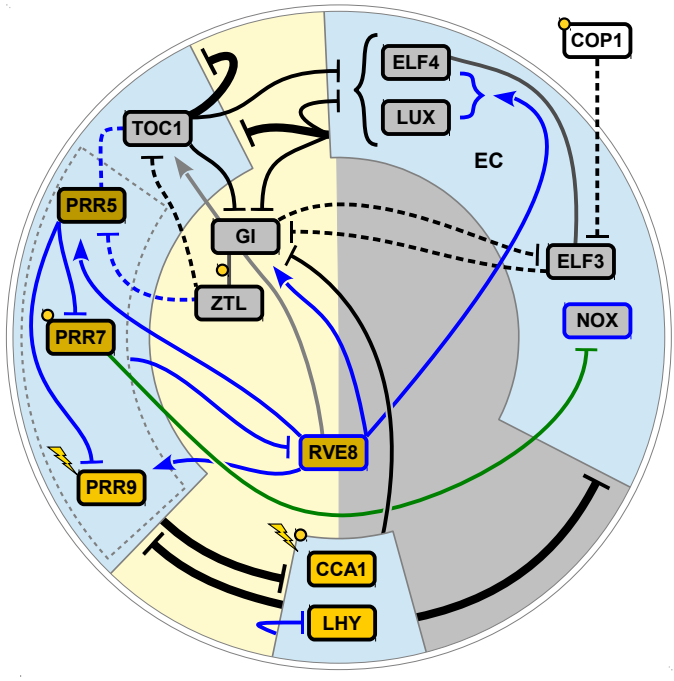

Figure III.1 The F2014 model of the *Arabidopsis* circadian clock. Components of the clock are laid out according to approximate time of peak mRNA expression, clockwise with zeitgeber time 0 (lights on) at the bottom. Yellow and grey boxes indicate proteins that are active primarily during the day and night, respectively. Solid lines indicate transcriptional regulation and dashed lines indicate protein–protein interactions, with arrows for activation and bars for repression or degradation. Additions to the model relative to P2012 are shown in blue. The green line indicates a hypothetical interaction, and the light grey line indicates an interaction that the model predicts to be extremely weak. The light blue boxes show three main modules of the clock, and interactions between them are shown with thick black lines. EC is the evening complex between ELF3, ELF4 and LUX or NOX, and the dark grey line indicates the ELF3-ELF4 complex. Lightning and yellow circles symbolize light input at the transcriptional and post-transcriptional level, respectively. For an alternative version comparing F2014 with P2012 [6], see Figure III.S1.

ELF3 protein (not shown), but the distribution between free ELF3 and ELF3 bound in the ELF3-ELF4 complex was not directly constrained by any data. As expected, the variation between parameter sets was indeed greater for the levels of free ELF3 protein and the ELF3-ELF4 complex, as shown in Figure III.2B-C. However, the predicted level of EC (Figure III.2D) showed less variation than even the experimentally constrained *ELF3* mRNA. This indicates that the shape and timing of EC were of such importance that the EC profile was, in effect, tightly constrained by data for the seven EC repression targets (*PRR9*, *PRR7*, *PRR5*, *TOC1*, *GI*, *LUX* and *ELF4*).

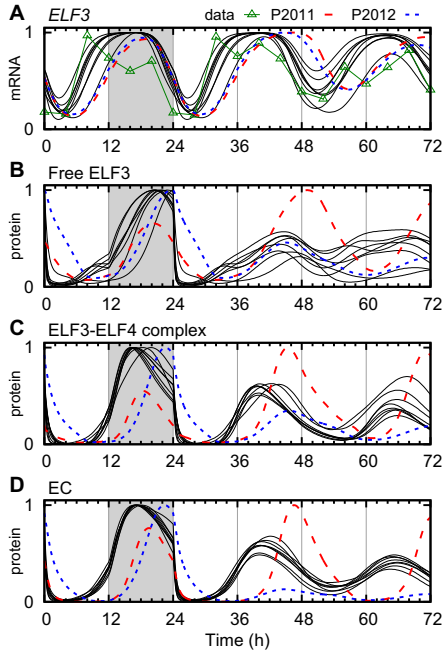

Figure III.2 The evening complex and its components. Concentration levels of a selection of model components relevant to EC, in the transition from LD 12:12 (light/dark cycles) to LL (constant light), comparing our ensemble of models (eight parameter sets, black lines), to the previous models P2011 (dashed red line) and P2012 (dotted blue line). (A) *ELF3* mRNA in wild type (wt), compared with a typical experiment (green triangles, data from [31]). (B) ELF3 protein in the nucleus, not counting complexes. (C) The ELF3-ELF4 protein complex. (D) The resulting evening complex. Each curve was normalized to a peak level of 1. Grey background signifies the night of the last day of LD before the transition to LL at ZT 24.

III.3.2 *NOX as a brother of LUX*

NOX is a close homologue of LUX, with a highly similar DNA-binding domain and a similar expression pattern which peaks in the evening. Like LUX, NOX can form a complex with ELF3 and ELF4, but it is only partially redundant with LUX, which has a stronger clock phenotype [32]. The recruitment of ELF3 to the *PRR9* promoter is reduced in the *lux-4* mutant and abolished in the LUX/NOX double amiRNA line [33]. To explain these findings, we introduced NOX into the model as a component acting in parallel with LUX; we assumed that NOX and LUX play similar roles as transcriptional repressors in the evening complex.

There is evidence that NOX binds to the promoter of *CCA1* (and possibly *LHY*) *in vivo* and activates its transcription. Accordingly, the peak level of *CCA1* expression is higher when NOX is overexpressed, and the period of the clock is longer [30]. This possible role of NOX as an activator fits badly with its reported redundancy with LUX as a repressor. In an attempt to resolve this issue, we first modelled the system with NOX only acting as a repressor in EC, and then investigated the effects of adding the activation of *CCA1* expression.

Figure III.3 illustrates the role of NOX in the model in comparison with LUX. The differences in their expression profiles (Figure III.3A-B) reflect the differences in their transcriptional regulation (cf. Figure III.1). *CCA1* expression is decreased only marginally in the *nox* mutant (Figure III.3C-D) but more so in *lux* (Figure III.3E). Because of the redundancy between NOX and LUX, the model predicted that the double mutant *lux;nox* has a stronger impact on circadian rhythms, with *CCA1* transcription cut at least in half compared with *lux* (Figure III.S2A). According to the model, the loss of LUX and NOX renders the evening complex completely ineffective, which in turn allows the *PRR* genes (including *TOC1*) to be expressed at high levels and thereby repress *LHY* and *CCA1*.

A comparison with the P2011 and P2012 models, which include LUX but not NOX, is shown in Figure III.3B, C and E. Here, the most noticeable improvement in our model was the more accurate peak timing after entry into LL, where in the earlier models the clock phase was delayed during the first subjective night [34].

Period lengthening and increased *CCA1* expression was observed in NOX-ox only for some of the parameter sets (Figure III.3F). The four parameter sets with increased *CCA1* all had a very weakly repressing NOX whose main effect was to counter LUX by taking its place in EC. Removing NOX from EC in the equations and reoptimizing a relevant

subset of the parameters worsened the fit to the data (Figure III.S3). These results support the idea of NOX acting through EC in manner that makes it only partially redundant with LUX.

The possibility that NOX is a transcriptional activator of *CCA1* and *LHY* was probed by adding an activating term to the equations (see Section III.C) and reoptimizing the parameters that control transcription of *CCA1* and *LHY*. The resulting activation was very weak in all parameter sets, and had negligible effect on the expression of *CCA1* in NOX-ox (Figure III.S2B-C). Accordingly, the addition of the activation term did not improve the fit to data as measured by the cost function described in Methods (Figure III.S3).

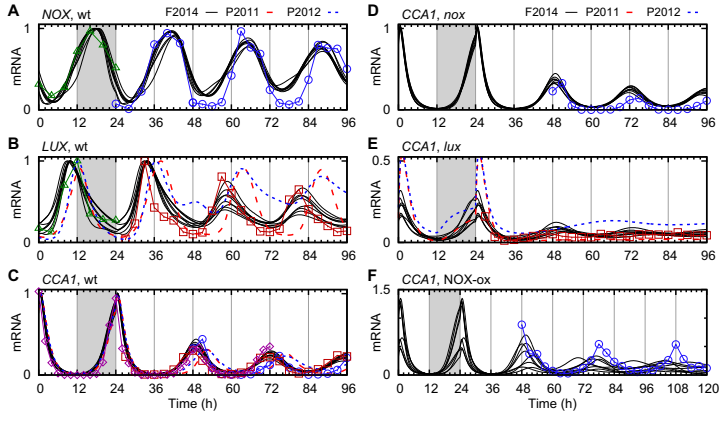

Figure III.3 NOX and its interaction with CCA1. Comparison between the F2014 model (eight parameter sets, black lines) and experimental data (green triangles [32], blue circles [30], red squares [35] and purple diamonds [36]), and the earlier models P2011 (dashed red lines) and P2012 (dotted blue lines), where applicable, in the transition from LD to LL. (A) *NOX* mRNA in wt. (B) *LUX* mRNA in wt. (C-F) *CCA1* mRNA in (C) wt, (D) *nox* mutant (*boa-1*), (E) *lux* mutant (*pcl1-1*), and (F) NOX-ox. The peak mRNA levels for the models were normalized to 1 in wt, and the same normalization was kept for the mutants. Experimental data were scaled to match the model in panel C, and the same normalization was used in panels D-F. Note the different vertical scales.

III.3.3 *Sequential PRR expression without activation*

In earlier models that included the PRR genes, the PRRs were described as a series of activators; during the day, PRR9 activated the transcription of *PRR7*, which similarly activated *PRR5*. These interactions improved the clock's entrainability to different LD cycles [8]. However, this sequential

activation disagrees with experimental data for *prr* knockout mutants, which indicate that loss of function of one PRR leaves the following PRR virtually unaffected. For instance, experiments have shown that the expression levels of *PRR5* and *TOC1* (as well as *LHY* and *CCA1*) are unaffected in both *prr9-1* and *prr7-3* knockout mutants [11, 37].

Instead, direct interactions between the PRRs have been found to be negative and directed from the later PRRs in the sequence to the earlier ones [15, 38]. A strong case has been made for TOC1 as a repressor of the PRR genes [9, 14]. As in P2012, we modelled transcription of *PRR9*, *PRR7* and *PRR5* as repressed by TOC1, but we also included negative auto-regulation of TOC1, as suggested by the ChIP-seq data that identified the TOC1 target genes [14]. Likewise, PRR5 directly represses expression of *PRR9* and *PRR7* [38], and we have added these interactions to the model.

As illustrated in Figure III.4A-C, this reformulation of the PRR wave is compatible with correct timing of the expression of the PRRs in the wild type, and the timing and shape of the expression curves were improved compared with the P2012 model. An earlier version of our model gave similar profiles despite missing the repression by PRR5, which suggests that such repression is not of great importance to the clock.

A nightly repressor appears to be acting on the *PRR7* promoter, as seen in the rhythmic expression of *PRR7* in LD in the *cca1-11;lhy-21;toc1-21* mutant [39]. An observed increase in *PRR7* expression at ZT 0 in the *lux-1* mutant relative to wild type [29] points to EC as a possible candidate. Although Helfer *et al.* report that LUX does not bind to the LUX binding site motif found in the *PRR7* promoter [32], we included EC among the repressors of *PRR7*. This interaction was confirmed by Mizuno *et al.* while this manuscript was in review [40], demonstrating the power of modelling and of timely publication of models.

We further let EC repress *PRR5*. We are not aware of any evidence for such a connection, but the parameter fitting consistently assigned a high value to the connection strength, as was also the case with *PRR7*. This result hints that nightly repression of *PRR5* is of importance, whether it is caused by EC or some related clock component.

The real test of the model came with knocking out members of the PRR wave. Here, the model generally outperformed the P2012 model, as judged by eye, but we are missing data for some important experiments such as PRR7 in *prr9*. As an example, Figure III.4D shows the level of PRR5 protein in the *prr9;prr7* double mutant, where half of our

parameter sets predict the correct profile and peak phase. In the earlier models, the only remaining inputs to PRR5 were LHY_{mod} (a hypothetical delayed LHY/CCA1), TOC1 (in P2012 only) and light (which stabilized the protein), and these were unable to shape the PRR5 profile correctly. The crucial difference in our model was the repression of *PRR5* by CCA1 and LHY, as described in the next section.

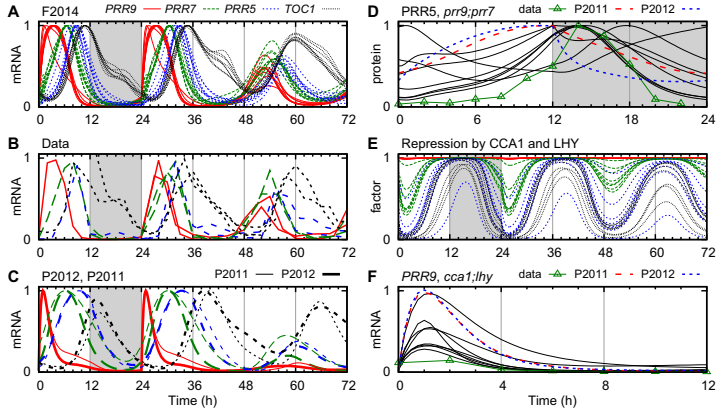

Figure III.4 Expression and regulation of the PRR genes. (A-C) The mRNA levels of *PRR9* (solid red), *PRR7* (long dashed green), *PRR5* (short dashed blue) and *TOC1* (dotted black) in the transition from LD to LL. (A) The F2014 model with eight different parameter sets. (B) Experimental data: *PRR9* [38, 39, 41], *PRR7* [38, 41, 42], *PRR5* [29, 42, 43] and *TOC1* [36, 44, 45]. (C) The P2012 and P2011 models (thick and thin lines, respectively). (D) Total PRR5 protein level in *prr9;prr7* in LD in F2014 (solid black), P2011 (dashed red), P2012 (dotted blue) and experimental data (green triangles [41]). (E) The predicted repression of *PRR* transcription by CCA1 and LHY, as a multiplicative factor, with colours as in (A-C). (F) *PRR9* mRNA in *cca1-11;lhy-21* in LD, normalized to the corresponding wt curves in (A-C); colours as in (D) but data from [11]. The peak levels in (A), (C) and (D) were normalized to 1, whereas the levels in (B) were adjusted manually.

III.3.4 *Regulation of the PRRs by CCA1 and LHY*

CCA1 and LHY appear to work as transcriptional repressors in most contexts in the clock (see e.g. [46]), but knockdown and overexpression experiments seem to suggest that they act as activators of *PRR9* and *PRR7* [37]. Accordingly, previous models have used activation by LHY/CCA1, combined with an acute light response, to accomplish the rapid increase observed in *PRR9* mRNA in the morning. However, with the misinterpretation of TOC1 regulation of CCA1 [12] in mind, we were reluctant to assume that the activation is a direct effect.

To investigate this issue, we modelled the clock with CCA1 and LHY acting as repressors of all four PRRs. If repression was incompatible with the data for any of the PRRs, parameter fitting should reduce the strength of that repression term to near zero. As is shown in Figure III.4E, the model consistently made CCA1 and LHY strongly repress *PRR5* and *TOC1*. *PRR7* was also repressed, but in a narrower time window that acted to modulate the phase of its expression peak. In contrast, *PRR9* was virtually unaffected; CCA1 and LHY do not directly repress *PRR9* in the model.

Even though CCA1 and LHY were not modelled as activators, the model reproduced the reduction in *PRR9* expression observed in the *cca1-11;lhy-21* double mutant (Figure III.4F and Figure III.S4). *PRR7* behaved similarly to *PRR9* in both experiments and model. Conversely, in the P2011 and P2012 models, where LHY/CCA1 was supposed to activate *PRR9*, there was no reduction in the peak level of *PRR9* mRNA in *cca1;lhy* compared to wild type (Figure III.S5A).

To explore whether CCA1 and LHY may be activating *PRR9* transcription, we temporarily added an activation term to the equations (see Section III.C) and reoptimized the relevant model parameters. The activation term came to increase *PRR9* expression around ZT 2 at least twofold in two of the eight parameter sets, and by a smaller amount in several (Figure III.S5B). This would seem to suggest that activation improved the fit between data and model. Surprisingly, there was no improvement as measured by the cost function (Figure III.S3). With the added activation, *PRR9* was reduced only marginally more in *cca1;lhy* than in the original model (Figure III.S5C). A likely explanation is that feedbacks through EC and TOC1, which repress *PRR9*, almost completely negate the removed activation of *PRR9* in the *cca1;lhy* mutant. Thus the model neither requires nor rules out activation of *PRR9* by CCA1 and LHY.

III.3.5 *Transcriptional activation by RVE8*

Like CCA1 and LHY, RVE8 is a morning expressed MYB-domain transcription factor. However, unlike CCA1 and LHY, RVE8 functions as an activator of genes with the evening element motif, and its peak activity in the afternoon is strongly delayed in relation to its expression [28]. Based on experimentally identified targets, we introduced RVE8 into our model as an activator of the five evening expressed clock compo-

nents *PRR5*, *TOC1*, *GI*, *LUX* and *ELF4*, as well as the morning expressed *PRR9* [29].

PRR5 binds directly to the promoter of *RVE8* to repress its transcription [38], and it is likely that PRR7 and PRR9 share this function [28, 29]. Using only these three PRRs as repressors of *RVE8* was sufficient to capture the expression profile and timing of RVE8, both in LL and LD (Figure III.5A).

RVE8 is partially redundant with RVE4 and RVE6 [28], which led us to model the *rve8* mutant as a 60% reduction in the production of RVE8. To clearly see the effects of RVE8 in the model, we instead compared with the *rve4;rve6;rve8* triple mutant, which we modelled as a total knockout of RVE8 function. The phase of the clock was delayed in LD, and the period lengthened by approximately two hours in LL in the simulated triple mutant, in agreement with data for LHY (Figure III.5B-C), though we note that CAB::LUC showed a greater period lengthening in experiments [29].

To investigate the significance of RVE8 as an activator in the model, we made a version of the model without RVE8. The model parameters were reoptimized against the time course data (excluding data for RVE8 and from *rve* mutants). As with NOX, we found that removing the activation had no clear effect on the costs of the parameter sets after refitting (Figure III.S3). It appears that activators such as RVE8 are not necessary for clock function. Still, the effects of the *rve* mutants can only be explained when RVE8 is present in the model, motivating its inclusion.

The model used RVE8 as an activator for four of its targets in a majority of the parameter sets (Figure III.5D-F). The exceptions were *TOC1* and *ELF4*. Although *TOC1* is a binding target of RVE8 *in vivo*, *TOC1* expression is not strongly affected by *RVE8*-ox or *rve8-1* [28, 47]. This was confirmed by our model, where the parameter fitting disfavoured the activation of *TOC1* in most of the parameter sets (Figure III.5E). The eight parameter sets may not represent an exhaustive exploration of the parameter space, but the results nevertheless support the notion that the effect of RVE8 on *TOC1* is of marginal importance.

III.4 METHODS

As with previous models of the *Arabidopsis* clock, our model consists of a set of ordinary differential equations (ODEs) with parameters that need

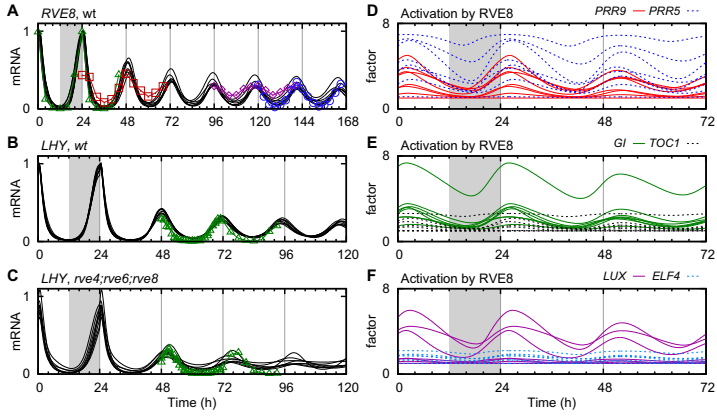

Figure III.5 The effects of RVE8 in the model. (A-C) Expression levels in the transition from LD to LL, comparing the model (eight parameter sets, solid black lines) with experimental data (green triangles [29], red squares [48], blue circles [28] and purple diamonds [49]). (A) *RVE8* mRNA in wt, (B) *LHY* in wt, and (C) *LHY* in *rve4;rve6;rve8*. (D-F) The effect of RVE8 on each of its target genes, as a time-dependent multiplicative factor, in the eight parameter sets. (D) *PRR9* (solid red) and *PRR5* (dotted blue), (E) *GI* (solid green) and *TOC1* (dotted black), and (F) *LUX* (solid purple) and *ELF4* (dotted light blue).

to be fitted against experimental observations. The final F2014 model consists of equations for 35 variables, with a total of 119 parameters. The number of variables has increased compared with previous models (see Table III.1), but the number of parameters has been reduced relative to P2012, due to the simplifications described in Results and Section III.B.

III.4.1 *Data collection*

Constraining the many parameters in our model requires a cost function based on a large number of experiments. To this end, we compiled time course data from the published literature, mainly by digitizing data

Model	Parameters	Variables
L2006 [7]	60 (+8)	16
P2010 [8]	80 (+17)	19
P2011 [9]	107 (+6)	28
P2012 [6]	123 (+10)	28 (+4)
F2014	119 (-)	35

Table III.1 The number of parameters and variables in different *Arabidopsis* clock models. Parameter counts in parentheses refer to constant integer Hill coefficients, which are written explicitly into the F2014 equations. Variables in parentheses for P2012 refer to ABA related variables.

points from figures using the free software package `g3data` [50]. We extracted more than 11,000 data points from 800 time courses in 150 different mutants or light conditions, from 59 different papers published between 1998 and 2013. The median time resolution was 3 hours. The list of time courses and publications can be found in section III.G, and the raw time course data and parameter values are available for download from `http://cbbp.thep.lu.se/activities/clocksim`.

Most of the compiled data refer to the mRNA level, from measurements using Northern blots or qPCR, but there are also data at the protein level (67 time courses) and measurements of gene expression using luciferase assays (12 time courses). About one third of the time courses can be considered as replicates, mainly from wild type plants in the most common light conditions. Many of these data are controls for different mutants. Where wild type and mutant data were plotted with the same normalization, we made note of this, as their relative levels provide crucial information that is lost if the curves are individually normalized.

III.4.2 *Model fitting and constraining*

To find suitable values for the model parameters, we constructed a minimalistic cost function based on the mean squared error between simulations and time course data. This approach was chosen to allow the model to capture as many features of the gene expression profiles as possible, with a minimum of human input.

The cost function consists of two parts, corresponding to the profiles and levels of the time course data, respectively. For each time course i with n_i experimental data points x_{ij}, the corresponding simulated data y_{ij} were obtained from the model. The simulations were performed with the mutant background represented in the model equations, with entrainment for up to 50 days in light/dark cycles followed by measurements, all in the experimental light conditions. The cost for the concentration profile was computed as

$$E_i^{(\mathrm{p})} = w_i \frac{1}{n_i} \sum_{j=1}^{n_i} \left(\frac{y_{ij}}{\bar{y}_i} - \frac{x_{ij}}{\bar{x}_i} \right)^2, \text{ where } \bar{u}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} u_{ij}, \; u \in \{x, y\}. \tag{III.1}$$

Since the profile levels are thus normalized, eq. (III.1) is independent of the units of measurements. The parameters w_i (see Section III.G for values) allowed us to weight time courses to reflect their relative

importance, e.g. where less data was available to constrain some part of the model.

Where several experimental time courses had the same normalization, e.g. in comparisons between wild type and mutants, the model should reproduce the relative changes in expression levels between the time courses. For each group of time courses, G_k, we could minimize the sum

$$\sum_{i \in G_k} w_i \left(\frac{\bar{y}_i}{\langle \bar{y} \rangle_k} - \frac{\bar{x}_i}{\langle \bar{x} \rangle_k} \right)^2, \text{ where } \langle \bar{u} \rangle_k = \frac{1}{\sum_{i \in G_k} w_i} \sum_{i \in G_k} w_i \bar{u}_i. \quad \text{(III.2)}$$

Unlike eq. (III.1), the nominators in this sum are guaranteed to be non-zero, which allows us to operate in log-space where fold changes up or down from the mean will be equally penalized. Replacing $\bar{x}_i / \langle \bar{x} \rangle_k$ with $\ln \bar{x}_i - \langle \ln \bar{x} \rangle_k$, and likewise for y, we write the final scaling cost for group k as

$$E_k^{(\mathrm{s})} = \sum_{i \in G_k} w_i \left(\ln \frac{\bar{y}_i}{\bar{x}_i} - \langle \ln \frac{\bar{y}}{\bar{x}} \rangle_k \right)^2. \quad \text{(III.3)}$$

This cost term thus penalizes non-uniform scaling between experiment and data within the group.

The total cost to minimize was

$$E = \sum_i E_i^{(\mathrm{p})} + \lambda \sum_k E_k^{(\mathrm{s})}, \quad \text{(III.4)}$$

where λ sets the balance between fitting the simulation to the profile or the level of the data. We used $\lambda = 0.1$.

A downside to our approach is that period and phase differences between different data sets result in fitting to a mean behaviour that is more damped than any individual data set. To reduce this problem, we removed the most obvious outliers from the fitting procedure. We also considered distorting the time axis (e.g. dynamic time warping) to normalize the period of oscillations in constant conditions, in order to better capture the effects of mutants relative to the wild type. This process would be cumbersome and arbitrary, which is why it was deemed outside the scope of our efforts.

Compared to previous models by Pokhilko *et al.*, fewer parameters were manually constrained in our model. In the P2010–P2012 models, roughly 40% of the parameters were constrained based on the experimental data [6, 8, 9], and the remaining free parameters were fitted to

mRNA profiles in LD and the free running period in LL and DD (constant dark) in wild type and mutants [9]. For the F2014 model, we completely constrained 16 parameters in order to obtain correct dynamics for parts of the system where we lacked sufficient time course data. Specifically, the parameters governing COP1 were taken from P2011 where they were introduced, whereas the parameters for the ZTL and GI proteins (except the GI production and transport rates) were fitted by hand to the figures in [51]. All other parameters were fitted to the collected time course data through the cost function.

The eight parameter sets presented here were selected from a group of 30, where each was independently seeded from the best of 1000 random points in parameter space, then optimized using parallel tempering for $> 10^4$ iterations at four different temperatures which were gradually lowered. The resulting parameter values, which are listed in Section III.F, typically span at least an order of magnitude between the different parameter sets (Figure III.S11). The sensitivity of the cost function to parameter perturbations is presented in Figure III.S6 and further discussed in Section III.D. Plots of the single best parameter set against all experimental data is shown in Figure III.S7.

To simulate the system and evaluate the cost function rapidly enough for parameter optimization to be feasible, we developed a C++ program that implements ODE integration and parameter optimization using the GNU Scientific Library [52]. Evaluating the cost function for a single point in parameter space, against the full set of experiments and data, took about 10 seconds on a 3 GHz Intel Core i7 processor. Our software is released under the GNU General Public License (GPL) [53] and is available from `http://cbbp.thep.lu.se/activities/clocksim/`.

III.5 DISCUSSION

III.5.1 *Modelling and data*

Accurately modelling the circadian clock as a network of a dozen or more genes is challenging. Previous modelling work (e.g. P2010–P2012) [6, 8, 9] has drawn on existing data and knowledge to constrain the models, but as the amount of data increases it becomes ever more difficult to keep track of the effects of mutations and other perturbations. For a system as large as the plant circadian clock, it is desirable to automate the parameter search as much as possible, but encoding the uncertainties

surrounding experimental data in a computer-evaluated cost function is not trivial.

Our modelling demonstrates the feasibility of fitting a model of an oscillating system against a large set of data without the construction of a complicated cost function based on qualitative aspects of the model output, such as entrainability, free-running period or amplitude. Instead, we relied on the large amount of compiled time course data to constrain the model, using a direct comparison between simulations and data. This minimalistic cost function had the additional advantage of allowing the use of time courses that span a transition in environmental conditions, e.g. from rhythmic to constant light, where the transient behaviour of the system may contain valuable information. Consequently, our model correctly reproduces the phase of the clock after such transitions (see e.g. Figure III.3C).

Our approach makes it easy to add new data, at the price of ignoring previous knowledge (e.g., clock period) from reporters that are not represented in the model. Accordingly, our primary modelling goal was not to reproduce the correct periods of different clock mutants, but rather to capture the profiles of mRNA and protein curves, and the changes in amplitude and profile between mutants and different light conditions. Compiling a large amount of data from different sources has allowed us to see patterns in expression profiles that were not apparent without independent replication. For example, the *TOC1* mRNA profile shows a secondary peak during the night in many data sets (see examples in Figure III.4B).

All collected time course data were used in fitting the parameters. To validate the model, we instead used independently obtained period data from clock period mutants. The results are shown in Section III.E. In brief, most predictions in LL are in good agreement with experiments, with the exception of *elf4* where the period changes in the wrong direction.

To experimentally measure a specific parameter value, such as the nuclear translocation rate of a protein, is exceptionally challenging. Hence, constraining a model with measured parameters can introduce large uncertainties in the model predictions, especially when the understanding of the full system is incomplete. Fitting the model with free parameters can instead give a large spread in individual parameter values, but result in a set of models that make well constrained predictions. For this reason, we have based our results on an ensemble of independently

optimized parameter sets, as recommended by Gutenkunst *et al.* [54]. At the cost of computational time, this approach gives a more accurate picture of the uncertainties in the model and its predictions, rather than focusing on individual parameter values.

Based on our experience of curation of time course data, we offer some suggestions for how data can be compiled and treated to be more useful to modellers. These points arose in the context of the circadian clock, but they apply to experiments that are to be used for modelling in a broader context.

- If the raw data contain information about the relative levels between experiments, for example between mutant and wild type, do not discard this information by normalizing the peak levels of the curves individually.
- If possible, provide data from both before and after treatment, preferably as one uninterrupted time course, so that changes in expression levels become clear. In clock experiments, this would entail including data from the last day of entrainment before a shift into constant light.
- Increase the time resolution of measurements where expression levels are expected to change rapidly, as this adds valuable information about timing. This is especially important around light/dark transitions to distinguish between acute light responses and circadian rhythms.
- Be clear about the conditions during entrainment, especially if they were varied between experiments.
- If possible, apply background correction so that the data reflect the true ratio between peak and trough levels. Alternatively, be clear about whether background correction has been applied.
- Use supplementary figures or files to present data that were not included in the figures and that would otherwise be lost to the research community.

Two of these suggestions concern the preservation of information about the relative expression levels between experiments. One example of the value of such information comes from the dramatic reduction in *PRR9* expression in *cca1;lhy* (Figure III.4F). As implied in the section on *PRR9* activation in Results, clock models ought to be able to explain

both shape and level of expression curves in such mutant experiments, but this is only possible if that information is present in the data.

III.5.2 *RVE8 as an activator*

Based on the current knowledge of the clock, most clock components are exclusively or primarily repressive, and RVE8 sets itself apart by functioning mainly (or solely) as an activator. According to our model, RVE8 has only a marginal effect on the expression of *TOC1*, but activates *PRR5* and other genes more strongly, in agreement with earlier interpretations of the experimental data [29].

We note that all six targets of RVE8 in the model (*PRR9*, *PRR5*, *TOC1*, *GI*, *LUX* and *ELF4*) are also binding targets of TOC1 [14]. This may be a coincidence, because TOC1 is a repressor of a majority of the genes in the model. It is conceivable, however, that activation by RVE8 around noon is gated by TOC1 to confer sensitivity to the timing of RVE8 relative to TOC1 in a controlled fashion.

We were surprised by the ease with which we could remove RVE8 from the model. After reoptimization of the parameters, the cost was decreased in three of the eight parameter sets compared with the original model (Figure III.S3). Thus, the clock is not dependent on activation for its function (although it should be noted that the model without RVE8 lost the ability to explain any RVE8-related experiments). This result indicates that the model possesses a high degree of flexibility, whereby the remaining components and parameters are able to adjust and restore the behaviour of the system. Such flexibility challenges our ability to test hypotheses about individual interactions in the model, but we argue that predictions can also be made based on entropy.

Even if an alteration to the model, such as the addition of RVE8, does not result in a significant change in the cost function, it may open up new parts of the high-dimensional parameter space. If, following local optimization, most parameter sets indicate that a certain interaction is activating, we may conclude that the activation is likely to be true. The parameter space is sampled in accordance with the prior belief that the model should roughly minimize the cost function, and the same reasoning motivates the use of an ensemble of parameter sets to explore the model. The conclusion about activation is indeed strengthened by the use of multiple parameter sets, because we learn whether it is valid in different areas of the parameter space.

III.5.3 *Problems and predictions*

Our model agrees with a majority of the compiled data sets, but like earlier models it also fails to fit to data for some mutants. This indicates that important clock components or interactions may yet be unknown or misinterpreted. We here give a few examples.

NOX expression is rhythmic in the short period double mutant *cca1;lhy* [30], but our model predicts a constant high NOX level in constant light (Figure III.S4F). If NOX is repressed by PRR7 as assumed in the model (see Section III.B.2), the rhythmicity can only be explained if PRR7 is also rhythmic and drives the NOX oscillations. Unfortunately, the model predicts that PRR7 oscillates only for a single cycle in *cca1;lhy*, before going to a constant low level (Figure III.S4B). This is a prediction shared with the P2012 model; we are not aware of any data that invalidate the prediction, but given that PRR7 is only slightly reduced in *cca1;lhy* in light/dark cycles [39], we believe that PRR7 may be rhythmic in constant light in this mutant.

The addition of NOX as a component partly redundant with LUX leads to an untested prediction regarding *CCA1* and *LHY*. Their peak expression levels are reduced only marginally in *nox* but roughly by half in *lux* compated with wt. In the *lux;nox* double mutant, the model predicts that their expression is cut by at least half again, to nearly zero even in light/dark cycles (see Figure III.3 and Figure III.S2).

The modelling suggests that nightly repression of *PRR5* and *PRR7* is of importance. The evening complex (EC) is thought to repress *PRR9* and *TOC1*, and our prediction that EC also represses *PRR7* was experimentally confirmed while this manuscript was in review [40].

Several known clock components were not included in the model, partly due to a lack of suitable data. Examples of genes that could be included in future models are *CHE* [55] and *EBI* [56]. More experiments and data are also needed to clarify the differences between CCA1 and LHY, the role of NOX as a part of the evening complex, and how PRR5 affects the localization of TOC1.

Additional non-transcriptional interactions should also be considered in future work. This includes protein interactions such as the regulation of LHY degradation by DET1 [57, 58]. Most importantly, the recently discovered and highly conserved redox-related circadian oscillator is linked to the transcriptional clock [59, 60]. Understanding that link may help explain why some clock components more easily remain rhythmic in experiments than in simulations.

III.5.4 *The complexity of the clock*

The insensitivity of *PRR9* to LHY/CCA1 in the P2011–P2012 models, as illustrated by its unchanged level in the *cca1;lhy* mutant (Figure III.S5A), shows one of the problems of constructing and fitting large models: The transcriptional activation of *PRR9* by LHY/CCA1 looks like an important term in the model equations, but the effects of this term are small. To reduce the prevalence of such "dead" terms and parameters in the equations, we recommend examining their effects in isolation, as was done with the corresponding repression terms in Figure III.4E.

The ability of our model to reduce *PRR9* expression in *cca1;lhy* (Figure III.4F) can only be explained by indirect effects. CCA1 and LHY repress *TOC1*, which in turn represses *PRR7* and *PRR9*, and the resulting indirect activation may be sufficient to counteract the direct repression by CCA1 and LHY. In general, in a highly interconnected system such as the circadian clock, it is perilous to draw conclusions about whether interactions are activating or repressing based only on altered expression levels in mutants.

Previous models (L2006–P2012) described the *Arabidopsis* circadian clock as primarily divided into two interacting feedback loops, the "morning loop" and the "evening loop". In contrast, we describe the clock in terms of three main modules linked by transcriptional repression and many additional connections (Figure III.1). Our results and experiences support an important point formulated by Hsu *et al.* [29]: The plant clock is best viewed as a highly interconnected, complex regulatory network, in which discrete feedback loops are virtually impossible to identify.

ACKNOWLEDGMENTS

We are grateful to James Locke for suggesting the inclusion of *RVE8*, to Andrew Millar for encouragement, to Carsten Peterson for proofreading and comments, to Patrik Edén for improvements to the cost function and to Maria Eriksson for bioinformatics analyses and fruitful discussions.

REFERENCES

1. 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.
2. J. Locke, A. Millar, and M. Turner, "Modelling genetic networks with noisy and varied experimental data: the circadian clock in *Arabidopsis thaliana*," *J of Theor Biol*, vol. 234, no. 3, pp. 383–393, 2005.
3. 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.
4. 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.
5. D. E. Somers, A. Webb, M. Pearson, and S. A. Kay, "The short-period mutant, *toc1-1*, alters circadian clock regulation of multiple outputs throughout development in *Arabidopsis thaliana*," *Development*, vol. 125, no. 3, pp. 485–494, 1998.
6. A. Pokhilko, P. Más, and A. J. Millar, "Modelling the widespread effects of TOC1 signalling on the plant circadian clock and its outputs," *BMC Syst Biol*, vol. 7, no. 1, pp. 1–12, 2013.
7. J. C. Locke, L. Kozma-Bognár, P. D. Gould, B. Fehér, E. Kevei, F. Nagy, M. S. Turner, A. Hall, and A. J. Millar, "Experimental validation of a predicted feedback loop in the multi-oscillator clock of *Arabidopsis thaliana*," *Mol Syst Biol*, vol. 2, no. 1, p. 59, 2006.
8. 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.
9. 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.
10. 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.

11. 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.
12. D. E. Somers, "The *Arabidopsis* clock: time for an about-face?," *Genome Biol*, vol. 13, no. 4, p. 153, 2012.
13. 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.
14. 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.
15. I. Carré and S. R. Veflingstad, "Emerging design principles in the *Arabidopsis* circadian clock," *Semin Cell Dev Biol*, vol. 24, no. 5, pp. 393–398, 2013.
16. D. H. Park, D. E. Somers, Y. S. Kim, Y. H. Choy, H. K. Lim, M. S. Soh, H. J. Kim, S. A. Kay, and H. G. Nam, "Control of circadian rhythms and photoperiodic flowering by the *Arabidopsis* GIGANTEA gene," *Science*, vol. 285, no. 5433, pp. 1579–1582, 1999.
17. K. A. Hicks, A. J. Millar, I. A. Carré, D. E. Somers, M. Straume, D. R. Meeks-Wagner, and S. A. Kay, "Conditional circadian dysfunction of the *Arabidopsis early-flowering 3* mutant," *Science*, vol. 274, no. 5288, pp. 790–792, 1996.
18. B. Thines and F. G. Harmon, "Ambient temperature response establishes ELF3 as a required component of the core *Arabidopsis* circadian clock," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 107, no. 7, pp. 3257–3262, 2010.
19. 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.
20. 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.

21. 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.
22. 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.
23. 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.
24. 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.
25. 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.
26. E. Kolmos and S. J. Davis, "*ELF4* as a central gene in the circadian clock," *Plant Signal Behav*, vol. 2, no. 5, pp. 370–372, 2007.
27. 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.
28. 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.
29. 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.
30. 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.

31. S. X. Lu, C. J. Webb, S. M. Knowles, S. H. Kim, Z. Wang, and E. M. Tobin, "CCA1 and ELF3 interact in the control of hypocotyl length and flowering time in *Arabidopsis*," *Plant Physiol*, vol. 158, no. 2, pp. 1079–1088, 2012.
32. 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.
33. 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.
34. A. N. Dodd, N. Dalchau, M. J. Gardner, S.-J. Baek, and A. A. Webb, "The circadian clock has transient plasticity of period and is required for timing of nocturnal processes in *Arabidopsis*," *New Phytol*, vol. 201, no. 1, pp. 168–179, 2014.
35. K. Onai and M. Ishiura, "*PHYTOCLOCK 1* encoding a novel GARP protein essential for the *Arabidopsis* circadian clock," *Genes Cells*, vol. 10, no. 10, pp. 963–972, 2005.
36. K. D. Edwards, O. E. Akman, K. Knox, P. J. Lumsden, A. W. Thomson, P. E. Brown, A. Pokhilko, L. Kozma-Bognar, F. Nagy, D. A. Rand, and A. J. Millar, "Quantitative analysis of regulatory flexibility under changing environmental conditions," *Mol Syst Biol*, vol. 6, no. 1, p. 424, 2010.
37. 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.
38. N. Nakamichi, T. Kiba, M. Kamioka, T. Suzuki, T. Yamashino, T. Higashiyama, H. Sakakibara, and T. Mizuno, "Transcriptional repressor PRR5 directly regulates clock-output pathways," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 109, no. 42, pp. 17123–17128, 2012.
39. 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.
40. T. Mizuno, Y. Nomoto, H. Oka, M. Kitayama, A. Takeuchi, M. Tsubouchi, and T. Yamashino, "Ambient temperature signal feeds into the circadian clock transcriptional circuitry through the EC night-

time repressor in *Arabidopsis thaliana*," *Plant Cell Physiol*, vol. 55, pp. 958–976, 2014.

41. 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.
42. 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.
43. A. Matsushika, A. Imamura, T. Yamashino, and T. Mizuno, "Aberrant expression of the light-inducible and circadian-regulated *APRR9* gene belonging to the circadian-associated APRR1/TOC1 quintet results in the phenotype of early flowering in *Arabidopsis thaliana*," *Plant Cell Physiol*, vol. 43, no. 8, pp. 833–843, 2002.
44. 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.
45. E. Sato, N. Nakamichi, T. Yamashino, and T. Mizuno, "Aberrant expression of the *Arabidopsis* circadian-regulated *APRR5* gene belonging to the APRR1/TOC1 quintet results in early flowering and hypersensitiveness to light in early photomorphogenesis," *Plant Cell Physiol*, vol. 43, no. 11, pp. 1374–1385, 2002.
46. 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.
47. B. Farinas and P. Mas, "Functional implication of the MYB transcription factor RVE8/LCL5 in the circadian control of histone acetylation," *Plant J*, vol. 66, no. 2, pp. 318–329, 2011.
48. W. Gong, K. He, M. Covington, S. Dinesh-Kumar, M. Snyder, S. L. Harmer, Y.-X. Zhu, and X. W. Deng, "The development of protein microarrays and their applications in DNA–protein and protein–protein interaction analyses of *Arabidopsis* transcription factors," *Mol Plant*, vol. 1, no. 1, pp. 27–41, 2008.
49. P. Y. Hsu and S. L. Harmer, "Circadian phase has profound effects on differential expression analysis," *PLoS One*, vol. 7, no. 11, p. e49853, 2012.

50. J. Frantz, "g3data," 2009. version 1.5.2.
51. J.-Y. Kim, H.-R. Song, B. L. Taylor, and I. A. Carré, "Light-regulated translation mediates gated induction of the *Arabidopsis* clock protein LHY," *EMBO J*, vol. 22, no. 4, pp. 935–944, 2003.
52. M. Galassi, J. Davies, J. Theiler, B. Gough, G. Jungman, P. Alken, M. Booth, and F. Rossi, *GNU Scientific Library Reference Manual*. Network Theory Ltd., third ed., 2009.
53. "GNU General Public License."
54. R. N. Gutenkunst, J. J. Waterfall, F. P. Casey, K. S. Brown, C. R. Myers, and J. P. Sethna, "Universally sloppy parameter sensitivities in systems biology models," *PLoS Comput Biol*, vol. 3, no. 10, p. e189, 2007.
55. 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.
56. 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.
57. H.-R. Song and I. A. Carré, "DET1 regulates the proteasomal degradation of LHY, a component of the *Arabidopsis* circadian clock," *Plant Mol Biol*, vol. 57, no. 5, pp. 761–771, 2005.
58. O. S. Lau, X. Huang, J.-B. Charron, J.-H. Lee, G. Li, and X. W. Deng, "Interaction of *Arabidopsis* DET1 with CCA1 and LHY in mediating transcriptional repression in the plant circadian clock," *Mol Cell*, vol. 43, no. 5, pp. 703–712, 2011.
59. J. S. O'Neill and A. B. Reddy, "Circadian clocks in human red blood cells," *Nature*, vol. 469, no. 7331, pp. 498–503, 2011.
60. 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.

III.A SUPPLEMENTARY FIGURES

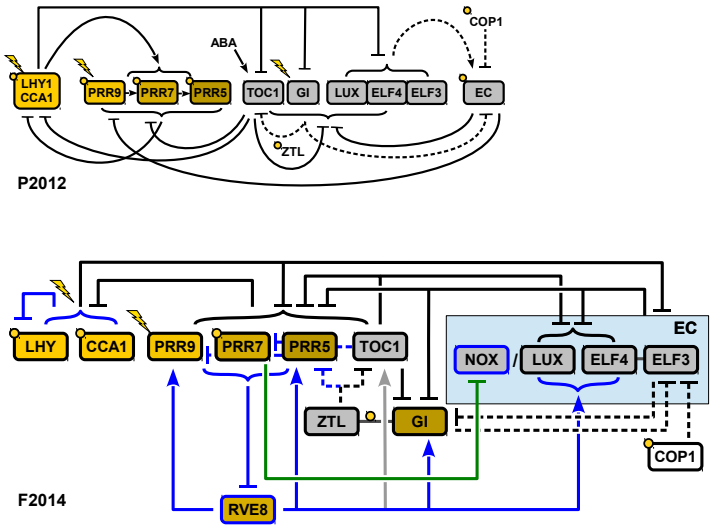

Figure III.S1 Model comparison. An alternative representation of the F2014 model (bottom), allowing easier comparison with the P2012 model (top), adapted from [1]. Symbols as in Figure III.1.

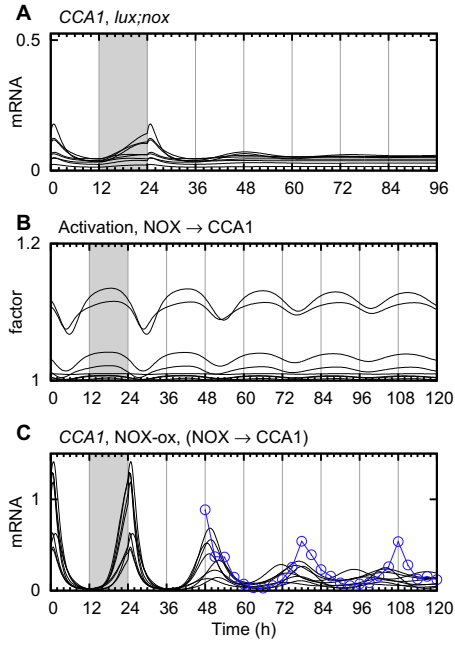

Figure III.S2 NOX interaction with CCA1. (A) The predicted *CCA1* expression level in the *lux;nox* double mutant, in the transition from LD to LL in F2014. The peak levels were normalized to 1 in wt, as in Figure III.3. (B) The activation of *CCA1* expression by NOX in a variant of the model, expressed as a multiplicative factor. (C) *CCA1* mRNA in NOX-ox in same model variant as (B), shown as in Figure III.3F.

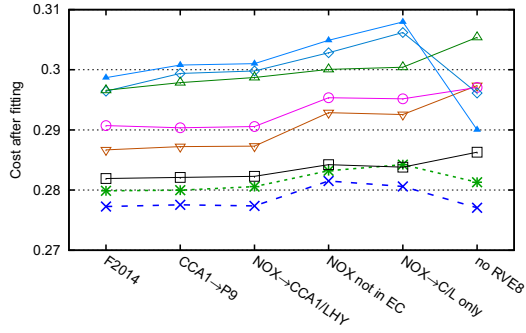

Figure III.S3 Cost function values. The value of the cost function for the eight best parameter sets in the six different model variants discussed in the text. Note that all parameters were reoptimized in the model without RVE8, whereas only a subset of the parameters were reoptimized in the variants with CCA1 activating *PRR9* or with different NOX function. Furthermore, the original model improved somewhat as it was optimized in parallel with the other variants.

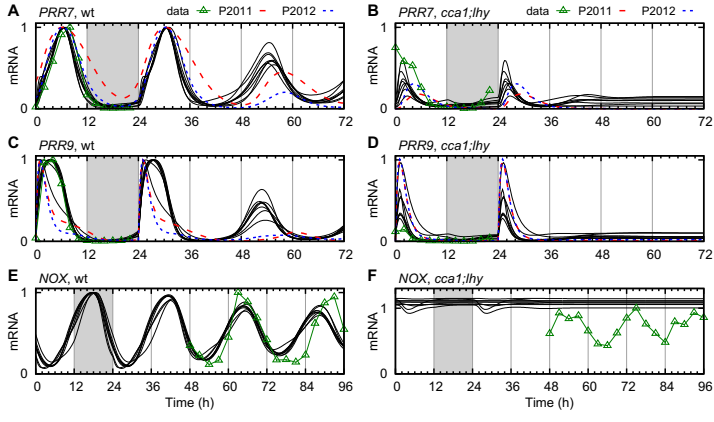

Figure III.S4 *PRR7*, *PPR9* and *NOX* mRNA in wt and *cca1;lhy*. Comparison between our model (solid black lines), P2011 (dashed red lines), P2012 (dotted blue lines) and data (green triangles) between wt (left panels) and *cca1;lhy* (right panels), in the transition from LD to LL. (A-B) *PPR7*, (C-D) *PPR9*, and (E-F) *NOX*. Data from [2] (A-D) and [3] (E-F). Peak levels were normalized to 1 in wt.

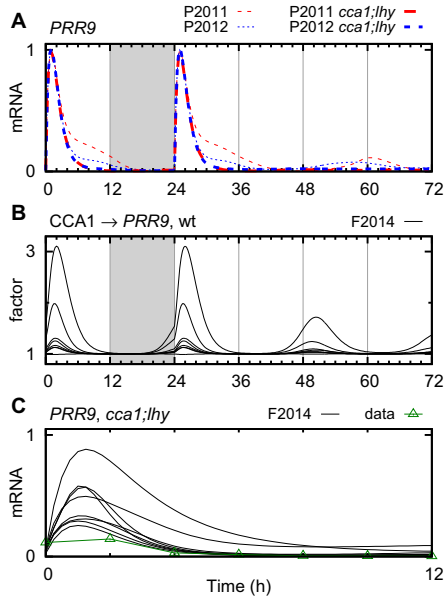

Figure III.S5 Effects of activation of *PRR9* by CCA1. (A) *PRR9* mRNA in P2011 (dashed red) and P2012 (dotted blue) in wt (thin lines, higher) and *cca1;lhy* (thick lines, lower), in the transition from LD to LL. Activation by LHY/CCA1 affects the expression of *PRR9* in the afternoon, but the peak level is unaffected in the double mutant. (B) The activation of *PRR9* by CCA1, after refitting our model with such an activation term. The activation is shown as a multiplicative factor, whose peak is > 1.2 for half of the eight parameter sets. (C) Expression of *PRR9* in *cca1;lhy* in the day (LD or first day of LL), in the model where CCA1 activates *PRR9* transcription, with peak levels normalized to 1 in wt. The difference between the model (black lines) and data (green triangles [2]) is comparable to the difference without the activation term (Figure III.4F).

Figure III.S6 Parameter sensitivity analysis. (External data, available online through journal site) The relative change in cost function in each of the eight best parameter sets (eight different symbols) when each parameter is altered. Symbols above (below) the zero cost line refer to multiplication (division) of the parameter by 1.1.

Figure III.S7 Model simulations compared with all data. (External data, of 370 plots, available online through journal site) Simulations with the single best parameter set, plotted against all 800 time courses used for fitting the model. As described in Methods, simulations and data are normalized to the same mean. Time courses with identical normalization are shown on the same page ("Scale group G_k"), with the total scaling cost in the title. The profile and scaling costs ($E^{(p)}$ and $E^{(s)}$) for each individual time course is shown in the legend. The time courses are named after the data files used; these are available for download as described in Methods. The naming convention is as follows: initial letters denote light condition, dd (constant dark), ll (constant light), rr (constant red light), bb (constant blue light), ld (light dark LD 12:12), lgd (long day LD 16:8), and shd (short day LD 8:16); followed by gene name, C (CCA1), L (LHY), T or P1 (TOC1), G (GI), P5,7,9 (PRR5,7,9), LUX (LUX), NOX (NOX), R8 (RVE8), E3 (ELF3), E4 (ELF4), Z (ZTL); suffixed by "_m" for mRNA data and an arbitrary number for uniqueness, or just the number for protein data. The last part of the filename is "-ox" for overexpression, and/or lower case gene names for mutants. A combination of LL and another light condition indicates entrainment in something other than LD 12:12, followed by LL. Where all data come from the same light conditions, the background is shaded for night; exceptions include scaling groups with data from different photoperiods.

III.B ADDITIONAL RESULTS

Herein, we provide further information about the modelling, covering details about the evening complex, the regulation of NOX, the splitting of CCA1 and LHY into two variables, the localization of TOC1 and PRR5, and the removal of the ABA circuit, LHY_{mod} and some light inputs. We also include the differential equations of the model, a table of periods comparing model to experiments, and the parameter values of the eight best fitted parameter sets. The equations are presented in their wild type forms, which do not include modifications used when simulating the many different mutants.

III.B.1 *Evening complex modelling details*

The equations describing EC are practically identical between P2011 and P2012. A minor difference is that several parameters for ELF3 degradation by COP1 are merged in the latter model. In the P2011–P2012 equations, EC is formed in two steps: ELF3 and ELF4 form a complex, which then binds with LUX to form EC. The models also describe the formation of a complex between GI and ELF3, whereby GI facilitates the degradation of ELF3 through COP1, based on observed interactions between the three proteins [4].

To reformulate the EC equations in order to take into account the possible redundancy between NOX and LUX and the ability of ELF3-ox to rescue EC function in *elf4*, as described in the main text, we had to simplify the description of EC formation to avoid a combinatorial explosion of reaction paths between sub-complexes.

In our model, the formation of EC begins with the homodimerization of ELF4 [5]. This homodimer, which is given its own variable, is in turn bound in the ELF3-ELF4 complex [6]. We assumed that delays caused by e.g. the time needed for LUX to bind to ELF3 (with or without ELF4) can at least partly be transferred to other steps in the model. Thus we modelled EC activity directly as a function of the levels of ELF3, ELF3-ELF4, LUX and NOX.

The equations for COP1, which regulates degradation of ELF3 and GI, were left unchanged from the P2012 model, but the action of COP1 was by necessity adapted to the altered EC. ELF3 is strongly localized to the nucleus [7] and unlike P2012 our model only considers its nuclear fraction. In our model, COP1 acts on ELF3 through its nuclear "day" and "night" forms, and in addition ELF3 degradation is directly increased by nuclear GI; the ELF3-GI complex of P2012 has been removed. Cytosolic

and nuclear GI are no longer handled as being in a quasi-steady state, but instead are given their own variables. The degradation of GI by COP1 is mediated by ELF3 [4]. This is reflected in the equations, where nuclear GI is degraded by the two forms of COP1 only when ELF3 is present. This is qualitatively similar to the structure of the P2012 model, even though the equations are different and may allow different dynamics.

In simulations of the *elf3-4* knockout mutant, the clock loses rhythmicity in LL, and the expression levels of *PRR9*, *PRR7*, *GI* and *TOC1* peak at the wrong time in LD compared with experiments (Figure III.S8). This is true for F2014 as well as P2011–P2012, even though the predicted expression profiles are different. Simulating the mutants as having ELF3 function at 20% of its normal value led to expression profiles in better agreement with the data for *GI* and *TOC1* and possibly also *PRR9*. However, the *elf3-4* mutant has an early stop codon which is expected to lead to a total loss of function [8].

One way to resolve this conflict could be to assume that LUX and NOX retain some function in the absence of ELF3. However, when we allowed LUX and NOX to act as EC in the absence of ELF3, with or without ELF4, the resulting expression of GI was low and out of phase with experiments. When instead we assumed that ELF4 on its own is able to interact with LUX and NOX, the level of EC became constant rather than oscillating. We conclude that if, as expected, the *elf3-4* mutant leads to a total loss of function, EC function may be rescued by some other clock component which is partly redundant with ELF3.

Furthermore, strong rhythmicity was seen in *ELF3* mRNA in LD in the loss-of-function *elf3-1* and *elf3-2* mutants [8], even though *CCA1* and *LHY* are repressed and only weakly oscillating in *elf3-1* [9]. Even simulations with ELF3 at 20% function showed only weak *ELF3* rhythms, which suggests that *ELF3* transcription is regulated by a clock component other than CCA1 and LHY, probably one with daytime expression and preserved rhythms in *elf3*. The best candidate represented in the model is PRR9. Making PRR9 a repressor of *ELF3* transcription did not work well in the current model, possibly because the predicted timing of the weakly constrained PRR9 protein was incorrect, but we think this predicted connection is worth exploring in future work.

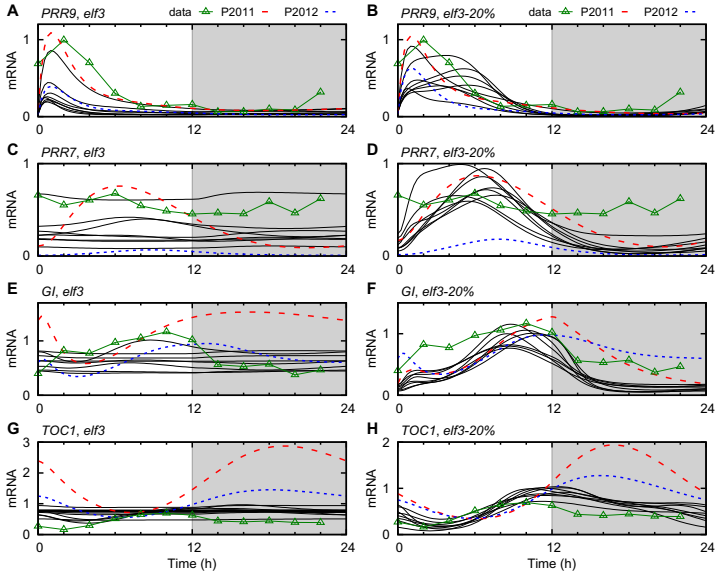

Figure III.S8 Retained ELF3 function in *elf3* mutants. Comparison between modelling the *elf3-4* mutant as a complete loss of function (left panels) and as lowering ELF3 production to 20% of its normal value (right panels). The F2014 model (solid black lines) is compared with data from Dixon *et al.* [2] (green triangles) and the models P2011 (dashed red lines) and P2012 (dotted blue lines) in LD. (A-B) *PRR9* expression, (C-D) *PRR7* expression, (E-F) *GI* expression, and (G-H) *TOC1* expression. Levels were normalized to a peak value of 1 in wt.

III.B.2 *Additional input into NOX*

Having only CCA1 and LHY as inputs to *NOX* was not sufficient to reproduce all *NOX* expression data; *NOX* is rhythmic in constant light in the *cca1-11;lhy-21* double mutant [3]. Our interpretation is that *NOX* should have at least one more repressor in the model. Among the clock components in our model, the expression profiles of PRR7 or PRR9 in LD provided the closest match to what we expected of an additional repressor. For computational reasons, we did not fully explore the difference between using PRR7 or PRR9 as the repressor of NOX transcription in the equations, but our initial attempts suggested that PRR7 may lead to a better fit. Hence, we included PRR7 as a transcriptional repressor in the equation for NOX. Figure III.S4C, in the main text, shows the resulting expression profile, where the input from PRR7 is seen to modulate the shape and peak phase of *NOX* expression by reducing transcription around ZT 10.

III.B.3 *CCA1 and LHY are modelled separately*

Although CCA1 and LHY are closely related, highly coexpressed, and have some overlap in function, they are not redundant [10–12]. There are noticeable differences in their regulation, as only the *CCA1* promoter interacts with CHE (which is not represented in the current model due to a lack of experimental data) [13], which may also be true for NOX [3]. Furthermore, CCA1 is more important than LHY at lower temperatures for regulating the period, and vice versa at higher temperatures [14].

These facts, in conjunction with access to significant amounts of separate data for CCA1 and LHY, and for their mutants, led us to split the LHY/CCA1 module of previous models (L2005 to P2012) into two separate parts. Both parts contribute to the repression of the targets of the previous LHY/CCA1 in P2012. The difference between the two parts in our model lies only in the transcriptional regulation of *CCA1* and *LHY* themselves, not in their binding targets. In contrast to the *CCA1* promoter, the *LHY* promoter contains two predicted specific CCA1 binding elements [15, 16]. For this reason, we modelled only *LHY* as repressed by CCA1 and LHY. However, we cannot rule out that the interaction is activating, but the model agrees with experimental data that *CCA1* expression is lower in *lhy* whereas *LHY* expression is lower in *cca1* [17].

By modelling CCA1 and LHY separately we were able to include the interaction between the two proteins in the model. CCA1 and LHY are

single MYB-domain transcription factors [18] and form both homodimers and heterodimers *in vivo* in order to bind DNA, which most likely requires two MYB-domains [19, 20]. In the model, we assumed that CCA1 and LHY may differ in their overall binding affinities, but not in any target-specific way. The heterodimer could be more or less active than the homodimers, but parameter fitting indicated that this freedom was not needed; hence, we removed it from the equations.

With the separation of *CCA1* and *LHY*, we hope to set forth a process of better distinguishing what the differences between them actually are. It is usually taken for granted that one trait of LHY must probably also be true for CCA1. Only on rare occasions is it explicitly said that it is not so, as in the case of CHE and NOX in relation to CCA1 and LHY. Another example relates to *CCA1* and *LHY* mRNA stability in dark/light, where results by Yakir *et al.* [21] and Kim *et al.* [22] are in direct contradiction.

III.B.4 *Localization of TOC1 and PRR5*

PRR5 plays a major part in translocating TOC1 to the nucleus, in addition to its role as a transcriptional repressor. In the absence of functional PRR5, the level of TOC1 is lower in the nucleus and higher in the cytosol than in the wild type. The total TOC1 protein level is lower, even though the *TOC1* mRNA level is unchanged, suggesting that PRR5 both localizes and stabilizes TOC1 [23].

Like TOC1, PRR5 is targeted for degradation by ZTL [24, 25], which is localized only to the cytosol [26]. Thus the model must include cycling of both TOC1 and PRR5 between the cytosol and the nucleus.

Due to the small amount of data at the protein level, we modelled this part of the system in a relatively simple way: TOC1 diffuses freely into the nucleus, but diffusion back into the cytosol is inhibited by nuclear PRR5. The PRR5 protein may diffuse between nucleus and cytosol, unaffected by TOC1. The model encourages stabilization of TOC1 by also allowing PRR5 to inhibit nuclear degradation of TOC1. However, if TOC1 is more stable in the nucleus than in the cytosol, such a mechanism may be unnecessary.

For some parameter sets, the model reproduces the qualitative level changes in *prr5* compatered with wt, but there is a great variation between the parameter sets and the fit to data for total TOC1 protein is bad (Figure III.S9). Likewise, neither the nuclear degradation rate nor the diffusion rate of TOC1 shows any clear pattern between the parameter sets. This difficulty in fitting the model was likely due to

both the relatively small amount of data relevant to TOC1 localization and the large discrepancies in TOC1 peak timing between different data sets.

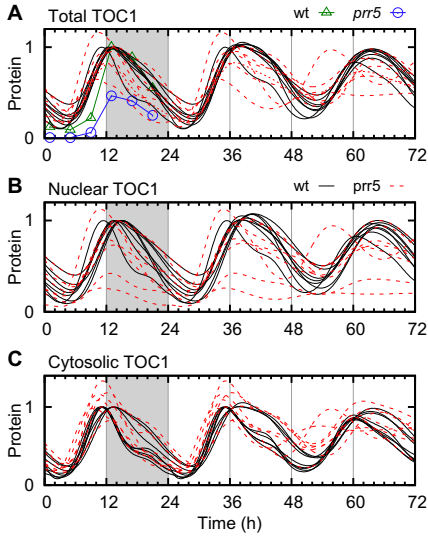

Figure III.S9 Localization of TOC1 protein. TOC1 protein simulated in wt (solid black lines) and *prr5* (dashed red lines), compared with data [23] for wt (green triangles) and *prr5* (blue circles). (A) Total TOC1 protein. (B) Nuclear TOC1 protein. (C) Cytosolic TOC1 protein. In each panel, the curves were normalized to a peak level of 1 in wt.

III.B.5 *Removal of light inputs and components*

We discarded several experimentally unmotivated or computationally unnecessary components and interactions compared with the P2012 model. This includes the removal of several light inputs for which we could find no convincing evidence. Specifically, we removed the direct light dependence in the degradation rates of *CCA1* and *LHY* mRNA and of the PRR9, PRR5 and TOC1 proteins. In the case of PRR5, the light input was replaced by ZTL-dependent degradation [24, 25]. The direct transcriptional light response of *GI* was also removed, since the degradation of EC by COP1 was sufficient to explain the experimentally observed rise in *GI* transcription in the morning.

We removed the hypothetical modified form of LHY/CCA1, LHY_{mod}. Its purpose in P2010–P2012 was to give a delayed positive input into PRR5, which proved to be redundant in our model where the rise in

PRR5 in the afternoon is instead due to ceasing repression by CCA1 and LHY (see Figure III.4A and E in the main text).

An additional difference between our model and P2012 is our exclusion of equations related to ABA. The primary purpose of the ABA circuit was to introduce an output from the clock, and although this circuit feeds back into *TOC1*, is has very little impact on the dynamics of the clock when the ABA input level is kept at its normal value (Figure III.S10).

The removal of unmotivated parameters and addition of new clock components balanced out. In spite of the inclusion of *NOX* and *RVE8* and the separation of *CCA1* and *LHY*, our model reduces the number of parameters compared with P2012, as shown in Table III.1 in the main text.

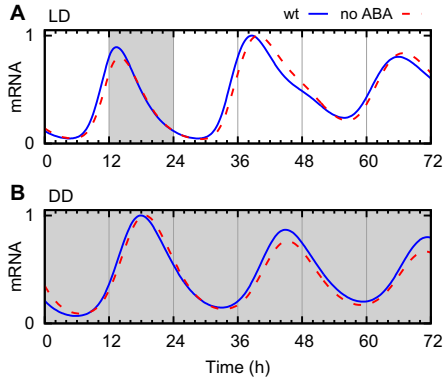

Figure III.S10 Limited feedback from ABA circuit to the clock in P2012. Normalized *TOC1* transcription in the P2012 model, with (solid blue lines) and without (dashed red lines) the ABA circuit connected to *TOC1* transcription. (A) in the transition from LD 12:12 to LL, and (B) in DD.

Short	Component
P	dark accumulator
R	RVE8
C	CCA1
L	LHY
$P9$	PRR9
$P7$	PRR7
$P5$	PRR5
T	TOC1
$E3$	ELF3
$E4$	ELF4
$E4d$	ELF4 dimer
$E34$	ELF3-ELF4 complex
LUX	LUX
NOX	NOX
ZTL	ZTL
G	GI
ZG	ZTL-GI complex
$COP1$	COP1

Table III.2 Symbols used in the equation system.

III.C MODEL EQUATIONS

We here describe the system of ordinary differential equations of the F2014 model. The dimensionless concentration levels of mRNA and protein of clock component X are denoted $c_X^{(m)}$ and c_X, respectively, where X is an abbreviated component name explained in Table III.2. Non-subscript L and D denote light and darkness, respectively, where one is 0 when the other is 1. When localization of a protein X is included in the model, it is either nuclear, Xn, or cytosolic, Xc. However, with the nomenclature inherited from P2012, *COP1*n and *COP1*d both denote nuclear COP1 protein, in its day and night forms. For ELF4, d indicates a dimer.

In order to simplify the equations, eqs. (III.5), (III.6), (III.7), (III.18), (III.21), (III.27), (III.30), (III.39) and (III.44) we define some recurring expressions. LC is a weighted sum of CCA1 and LHY concentrations, used where both CCA1 and LHY repress transcription. LC_{com} is the common term in the regulation of *CCA1* and *LHY* transcription. $P5_{trans}$, T_{trans} and G_{trans} describe the cytosolic/nuclear translocation of their respective proteins. $E34_{prod}$ and ZG_{prod} are complex formation rates, and $E3_{deg}$ is the $E3$ degradation rate that also applies to $E34$.

Parameters are named according to function. Parameters that govern transcriptional activation and repression are denoted by a and r, respectively. The symbol q is used for light-activated transcription, t for

protein transport rates, m for degradation rates (protein and mRNA), and n for protein production (for COP1 only). Weights between components that play similar roles (in EC and LC) are denoted by f, and p is used for various parameters from P2012 for protein production, transport, degradation and complex formation.

The mRNA production terms are all based on the same general assumptions about how repressors and activators bind to DNA to regulate transcription. CCA1 and LHY are assumed to share binding sites, as are the PRR proteins, but otherwise the binding sites for different proteins are assumed to be independent. As in P2012, all repression terms are squared in the denominators to represent the unknown degree of cooperativity. Activators have a corresponding term in the numerators, with a parameter describing the maximum level of activation relative to the unactivated state. The degradation rates of mRNAs always follow mass-action kinetics.

With few exceptions, the levels of mRNAs and proteins are arbitrary in the model as a change in the production rate could equally well be described as an opposite change in all binding affinities of the protein. The exceptions are those proteins that are involved in complexes, where the model has parameters to set the relative production rates. For other proteins, the maximum levels are determined by degradation and the regulation of production.

The expression for EC is designed such that it is limited by ELF3 and ELF3-ELF4 when LUX and NOX are high, and vice versa. What "high" means is defined by f_3 and f_4. It is assumed that LUX and ELF3-ELF4 are the most important players in the complex, and f_1, f_2 and f_6 allow NOX and ELF3 to also participate. For the difference between NOX and LUX, we separate the activity (numerator, f_6) from the saturation (denominator, f_2) to allow for the possibility that EC with NOX is a weaker repressor than EC with LUX. In contrast, the same expression with f_1 is used in both numerator and denominator because ELF3 is supposed to act like more dilute ELF3-ELF4.

The equations and parameter values are also available for download[1].

$$LC = (c_L + f_5 c_C) \tag{III.5}$$

$$LC_{com} = \frac{q_1 L c_P + 1}{1 + (r_1 c_{P9})^2 + (r_2 c_{P7})^2 + (r_3 c_{P5n})^2 + (r_4 c_{Tn})^2} \tag{III.6}$$

$$EC = \frac{(c_{LUX} + f_6 c_{NOX})(c_{E34} + f_1 c_{E3})}{1 + f_3(c_{LUX} + f_2 c_{NOX}) + f_4(c_{E34} + f_1 c_{E3})} \tag{III.7}$$

1 http://cbbp.thep.lu.se/activities/clocksim/

$$\frac{dc_L^{(m)}}{dt} = \frac{LC_{com}}{(1+(r_{11}LC)^2)} - m_1 c_L^{(m)} \tag{III.8}$$

$$\frac{dc_L}{dt} = (L + m_4 D)c_L^{(m)} - m_3 c_L \tag{III.9}$$

$$\frac{dc_C^{(m)}}{dt} = LC_{com} - m_1 c_C^{(m)} \tag{III.10}$$

$$\frac{dc_C}{dt} = (L + m_4 D)c_C^{(m)} - m_3 c_C \tag{III.11}$$

$$\frac{dc_P}{dt} = p_7 D(1 - c_P) - m_{11} c_P L \tag{III.12}$$

$$\begin{aligned}\frac{dc_{P9}^{(m)}}{dt} &= q_3 c_P L + (1 + a_3 r_{33} c_R)\frac{1}{(1+r_{33}c_R)}\frac{1}{(1+(r_5 LC)^2)} \\ &\quad \times \frac{1}{(1+(r_6 EC)^2)}\frac{1}{(1+(r_7 c_{Tn})^2)} \\ &\quad \times \frac{1}{(1+(r_{40}c_{P5n})^2)} - m_{12} c_{P9}^{(m)}\end{aligned} \tag{III.13}$$

$$\frac{dc_{P9}}{dt} = c_{P9}^{(m)} - m_{13} c_{P9} \tag{III.14}$$

$$\begin{aligned}\frac{dc_{P7}^{(m)}}{dt} &= \frac{1}{(1+(r_8 LC)^2)}\frac{1}{(1+(r_9 EC)^2)} \\ &\quad \times \frac{1}{(1+(r_{10}c_{Tn})^2)}\frac{1}{(1+(r_{40}c_{P5n})^2)} - m_{14} c_{P7}^{(m)}\end{aligned} \tag{III.15}$$

$$\frac{dc_{P7}}{dt} = c_{P7}^{(m)} - (m_{15} + m_{23} D)c_{P7} \tag{III.16}$$

$$\begin{aligned}\frac{dc_{P5}^{(m)}}{dt} &= (1 + a_4 r_{34} c_R)\frac{1}{(1+r_{34}c_R)}\frac{1}{(1+(r_{12}LC)^2)} \\ &\quad \times \frac{1}{(1+(r_{13}EC)^2)}\frac{1}{(1+(r_{14}c_{Tn})^2)} - m_{16} c_{P5}^{(m)}\end{aligned} \tag{III.17}$$

$$P5_{trans} = t_5 c_{P5c} - t_6 c_{P5n} \tag{III.18}$$

$$\frac{dc_{P5c}}{dt} = c_{P5}^{(m)} - (m_{17} + m_{24} c_{ZTL})c_{P5c} - P5_{trans} \tag{III.19}$$

$$\frac{dc_{P5n}}{dt} = P5_{trans} - m_{42} c_{P5n} \tag{III.20}$$

$$T_{trans} = t_7 c_{Tc} - \frac{t_8}{1 + m_{37} c_{P5n}} c_{Tn} \tag{III.21}$$

$$\begin{aligned}\frac{dc_T^{(m)}}{dt} &= (1 + a_5 r_{35} c_R)\frac{1}{(1+r_{35}c_R)}\frac{1}{(1+(r_{15}LC)^2)} \\ &\quad \times \frac{1}{(1+(r_{16}EC)^2)}\frac{1}{(1+(r_{17}c_T n)^2)} - m_5 c_T^{(m)}\end{aligned} \tag{III.22}$$

$$\frac{dc_{Tn}}{dt} = T_{trans} - \frac{m_{43}}{1 + m_{38} c_{P5n}} c_{Tn} \tag{III.23}$$

$$\frac{dc_{Tc}}{dt} = c_T^{(m)} - (m_8 + m_6 c_{ZTL})c_{Tc} - T_{trans} \tag{III.24}$$

$$\frac{dc_{E4}^{(m)}}{dt} = (1 + a_6 r_{36} c_R)\frac{1}{(1 + r_{36} c_R)}\frac{1}{(1 + (r_{18} EC)^2)} \times \frac{1}{(1 + (r_{19} LC)^2)}\frac{1}{(1 + (r_{20} c_{Tn})^2)} - m_7 c_{E4}^{(m)} \quad \text{(III.25)}$$

$$\frac{dc_{E4}}{dt} = p_{23} c_{E4}^{(m)} - m_{35} c_{E4} - {c_{E4}}^2 \quad \text{(III.26)}$$

$$E34_{prod} = p_{25} c_{E3} c_{E4d} \quad \text{(III.27)}$$

$$\frac{dc_{E4d}}{dt} = {c_{E4}}^2 - m_{36} c_{E4d} - E34_{prod} \quad \text{(III.28)}$$

$$\frac{dc_{E3}^{(m)}}{dt} = \frac{1}{1 + (r_{21} LC)^2} - m_{26} c_{E3}^{(m)} \quad \text{(III.29)}$$

$$E3_{deg} = (m_{30} c_{COP1d} + m_{29} c_{COP1n} + m_9 + m_{10} c_{Gn}) \quad \text{(III.30)}$$

$$\frac{dc_{E3}}{dt} = p_{16} c_{E3}^{(m)} - E34_{prod} - E3_{deg} c_{E3} \quad \text{(III.31)}$$

$$\frac{dc_{E34}}{dt} = E34_{prod} - m_{22} c_{E34} E3_{deg} \quad \text{(III.32)}$$

$$\frac{dc_{LUX}^{(m)}}{dt} = (1 + a_7 r_{37} c_R)\frac{1}{(1 + r_{37} c_R)}\frac{1}{(1 + (r_{22} EC)^2)} \times \frac{1}{(1 + (r_{23} LC)^2)}\frac{1}{(1 + (r_{24} c_{Tn})^2)} - m_{34} c_{LUX}^{(m)} \quad \text{(III.33)}$$

$$\frac{dc_{LUX}}{dt} = c_{LUX}^{(m)} - m_{39} c_{LUX} \quad \text{(III.34)}$$

$$\frac{dc_{COP1c}}{dt} = n_5 - p_6 c_{COP1c} - m_{27} c_{COP1c}(1 + p_{15} L) \quad \text{(III.35)}$$

$$\frac{dc_{COP1n}}{dt} = p_6 c_{COP1c} - (n_{14} + n_6 Lc_P) c_{COP1n} \quad \text{(III.36)}$$

$$- m_{27} c_{COP1n}(1 + p_{15} L) \quad \text{(III.37)}$$

$$\frac{dc_{COP1d}}{dt} = (n_{14} + n_6 Lc_P) c_{COP1n} - m_{31}(1 + m_{33} D) c_{COP1d} \quad \text{(III.38)}$$

$$ZG_{prod} = p_{12} c_{ZTL} c_{Gc} - (p_{13} D + p_{10} L) c_{ZG} \quad \text{(III.39)}$$

$$\frac{dc_{ZTL}}{dt} = p_{14} - ZG_{prod} - m_{20} c_{ZTL} \quad \text{(III.40)}$$

$$\frac{dc_{ZG}}{dt} = ZG_{prod} - m_{21} c_{ZG} \quad \text{(III.41)}$$

$$\frac{dc_G^{(m)}}{dt} = (1 + a_8 r_{38} c_R)\frac{1}{(1 + r_{38} c_R)}\frac{1}{(1 + (r_{25} EC)^2)} \times \frac{1}{(1 + (r_{26} LC)^2)}\frac{1}{(1 + (r_{27} c_{Tn})^2)} - m_{18} c_G^{(m)} \quad \text{(III.42)}$$

$$c_{E3tot} = c_{E3} + c_{E34} \quad \text{(III.43)}$$

$$G_{trans} = p_{28} c_{Gc} - \frac{p_{29}}{1 + t_9 c_{E3tot}} c_{Gn} \quad \text{(III.44)}$$

$$\frac{dc_{Gc}}{dt} = p_{11} c_G^{(m)} - ZG_{prod} - G_{trans} - m_{19} c_{Gc} \quad \text{(III.45)}$$

$$\frac{dc_{Gn}}{dt} = G_{trans} - m_{19}c_{Gn} - m_{25}c_{E3tot}(1 + m_{28}c_{COP1d} + m_{32}c_{COP1n})c_{Gn} \quad \text{(III.46)}$$

$$\frac{dc_{NOX}^{(m)}}{dt} = \frac{1}{(1 + (r_{28}LC)^2)(1 + (r_{29}c_{P7})^2)} - m_{44}c_{NOX}^{(m)} \quad \text{(III.47)}$$

$$\frac{dc_{NOX}}{dt} = c_{NOX}^{(m)} - m_{45}c_{NOX} \quad \text{(III.48)}$$

$$\frac{dc_R^{(m)}}{dt} = \frac{1}{1 + (r_{30}c_{P9})^2 + (r_{31}c_{P7})^2 + (r_{32}c_{P5n})^2} - m_{46}c_R^{(m)} \quad \text{(III.49)}$$

$$\frac{dc_R}{dt} = c_R^{(m)} - m_{47}c_R \quad \text{(III.50)}$$

III.C.1 *Model variants*

In the model without RVE8, $c_R^{(m)}$ and c_R were set to 0, and all data for RVE8 and the *rve* mutants were removed from the cost function. For testing NOX as an activator of *CCA1* and *LHY*, an activation term, $a_1 c_{NOX}/(1 + r_{39}c_{NOX})$, was added to the numerator of eq. (III.6). Similarly, the activation of *PRR9* transcription by CCA1 and LHY was implemented by the addition of $a_2(r_5 LC)^2$ to the numerator in eq. (III.13).

III.D PARAMETER SENSITIVITY ANALYSIS

The sensitivity of the cost function to perturbations in the parameter values are presented in Figure III.S6, which shows that the parameter sets generally agree on which parameters are sensitive to perturbations. However, parameters with high sensitivity are not necessarily constant between parameter sets.

Figure III.S11 shows that there is only a very weak correlation between the variability of a parameter between parameter sets and the robustness of the model to changes in that parameter. Thus, parameter sensitivity cannot be used to estimate how widely a parameter can vary between alternative parameter sets. Even though we have removed any obviously redundant parameters from the equations, the model is likely to be constraining many nonlinear functions of several parameters rather than the individual parameters. That is, the parameter values are often meaningful only in the context of their respective parameter sets.

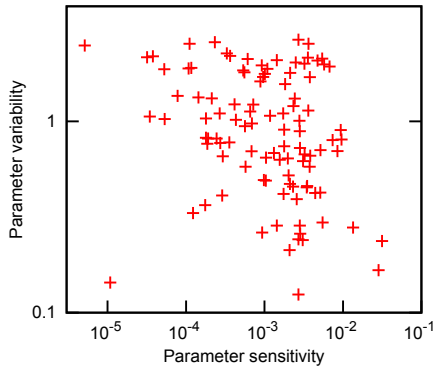

Figure III.S11 Parameter sensitivity and variability. For each parameter, the variability between parameter sets is plotted against the sensitivity of the parameter. Variability is defined as the standard deviation of the logarithm of the parameter value across the eight parameter sets. Sensitivity is defined as the mean relative change in the cost function when the parameter is increased and decreased by 10%, averaged over the parameter sets.

III.E MODEL PERIOD PREDICTIONS

To independently verify the output of the model, we compared with experimental data for the relative period change between wild type and mutants.

Experiment	Light cond	wt (exp)	Mutant (exp)	Change (exp)	Change (sim)	Source
toc1 RNAi	LL	24.27	20.51	−3.72	−1.50	[27]
toc1-1	LL	24.5	21	−3.43	−1.50	[28]
toc1-1	LL	24.82	22.46	−2.28	−1.50	[3]
cca1-1	LL	26.41	24.77	−1.49	−0.76	[19]
cca1-1	LL	25.31	23.82	−1.41	−0.76	[12]
cca1-11	LL	26.02	23.25	−2.55	−0.76	[3]
cca1-11;lhy-21	RR	24.5	18.2	−6.17	−5.90	[29]
cca1-11;lhy-21	LL	26.02	17.4	−7.95	−5.90	[3]
cca1-11;lhy-21	LL	26.41	19.73	−6.07	−5.90	[19]
cca1-1;lhy-R	LL	23.99	arr		−1.82	[12]
lhy	LL	22.71	23.64	0.98	−1.07	[10]
lhy^{TN104}	LL	22.71	24.67	2.07	−1.07	[10]
prr7-3	LL	24.3	25.0	0.69	1.27	[30]
prr7-3;prr9-1	LL	24.3	36.2	11.75	0.30	[30]
prr9-1	LL	24.3	24.8	0.49	1.26	[30]
PRR5-ox	LL	23.41	22.66	−0.77	−0.30	[31]
NOX-ox	LL	25.15	29.95	4.58	2.16	[3]
nox-1	LL	24.47	23.23	−1.22	−0.05	[3]

Experiment	Light cond	wt (exp)	Mutant (exp)	Change (exp)	Change (sim)	Source
lux	LL		arr		1.93	[32]
elf3-1	LL		arr		0.90	[8]
elf4	LL	23.8	22.3	−1.51	2.42	[33]
ELF4-OX	LL	27.308	30.89	3.15	−0.63	[34]
ELF4-OX	LL	25.1	28.75	3.49	−0.63	[35]
ELF3-OX	LL	26.25	26.85	0.55	1.25	[35]
ELF3-OX	LL	24.28	26.41	2.11	1.25	[36]
elf4-1;ELF3-OX	LL	26.25	27.05	0.73	3.23	[35]
ztl-22	LL	26.1	33.0	6.34	2.59	[37]
ztl-21	LL	26.3	27.7	1.28	2.59	[37]
ztl-21	RR	24.5	27.1	2.55	2.59	[29]
ztl-1	BB	24.9	28.95	3.90	2.59	[38]
ztl-3	RR	24.9	29.4	4.34	2.59	[38]
ztl-1 (Bx4)	LL	27.3	32.0	4.13	2.59	[39]
ztl-2 (Bx1)	LL	27.3	32.8	4.84	2.59	[39]
RVE8-OX	LL	24	22.16	−1.84	−1.07	[40]
rve8	LL	24	25.68	1.68	0.15	[40]
gi-11	LL	24.4	23.4	−0.98	−0.83	[14]
gi-201	LL	25.12	24.44	−0.65	−0.83	[41]
toc1-1	DD	27.5	22.3	−4.54	0.54	[28]
gi-201	DD	27.48	arr		−0.29	[41]
elf3-1	DD	25.07	25.41	0.33	1.46	[36]
elf4-1;ELF3-OX	DD	28.0	30.3	1.97	1.12	[35]
elf4	DD	26.4	27.1	0.64	0.88	[33]
ELF3-OX	DD	28.0	29.5	1.29	1.06	[35]
ELF3-OX	DD	25.07	25.10	0.03	1.06	[36]
ztl-22	DD	27.05	33.56	5.78	3.82	[37]
ztl-27	DD	27.05	36.43	8.32	3.82	[37]
prr7-3	DD	25.7	25.8	0.09	−0.16	[30]
NOX-OX	DD	24.95	26.85	1.83	0.01	[3]

Table III.3 Period change in mutants, compared between experiments and the F2014 model. The change in period between mutant, x, and wild type, y, is computed as $(x-y)\frac{24}{y}$. Experimental data were averaged where replicates were available within a publication (e.g. *toc1* RNAi [27], *ztl-1/3* [38], ELF4-ox[34], ELF3-ox [35], NOX-ox [3], RVE8-ox, *rve8* [40], and *cca1-1* [12]). The periods from the model were taken as the mean, across the eight parameters sets, of the median of the period of *TOC1*, *CCA1* and *PRR5* mRNA. Simulations were run in LD 12:12 and then transferred to constant light (LL) or darkness (DD) for four days. Some experiments were performed in constant red (RR), or blue (BB) light; these were simulated as LL. The experimental periods were largely based on luciferase data which were not used to fit the model. Mutants marked with "arr" were found to be arrhythmic.

III.F THE EIGHT BEST FITTED PARAMETER SETS

	1	2	3	4	5	6	7	8
a_1	5.3	1	7	4.8	9.9	9.7	10	10
a_4	7.2	9.9	4.7	9.8	8.4	7.9	7.1	1.2
a_5	2	5.4	1	5	4.9	10	2.7	1.1
a_6	8.9	2.3	2.5	1.4	2.2	1.8	1	1
a_7	1.5	5.9	1	9.2	1.7	1	9.3	8.1
a_8	3	4.4	6.9	9.4	7	9.9	2.7	9.9
r_1	5.7	3.7	11	1.2	3.8	3.5	1	1.2
r_2	1.4	2.4	2.6	5.2	3.1	2.8	4.9	3.5
r_3	8.6	4.7	1.2	7.6	9.8	4	0.019	9.4
r_4	3.4	16	5.8	95	0.1	0.1	0.11	0.1
r_5	0.15	0.1	0.1	0.12	0.1	0.1	0.1	0.1
r_6	26	0.47	97	280	41	31	1.3	180
r_7	12	36	23	230	2.8	0.45	14	12
r_8	1.8	2.2	1.9	1.6	1.4	1.8	0.32	1
r_9	31	0.59	150	360	48	39	2.5	400
r_{10}	10	22	12	82	3.6	0.011	27	11
r_{11}	1.9	2.1	1.6	1.1	1.2	2.4	1.6	1.1
r_{12}	5	6	3.4	3.5	4.9	11	15	1.3
r_{13}	45	1.1	140	440	59	36	4.4	330
r_{14}	3.9	13	8.9	0.12	1.8	0.34	94	17
r_{15}	5.3	6.7	19	4.4	5	10	6.8	11
r_{16}	9.7	0.15	52	82	13	8.8	0.91	84
r_{17}	2.2	5.2	42	0.38	1.8	0.85	61	15
r_{18}	42	1.2	290	490	64	22	3.1	390
r_{19}	16	16	26	12	12	11	7.6	30
r_{20}	0.11	0.15	0.15	0.43	0.63	0.013	0.25	12
r_{21}	3.3	5.1	3.1	2.5	1.7	13	2.5	4
r_{22}	56	2	180	150	33	94	6.6	730
r_{23}	3.3	7.1	5	4.9	2.6	4.8	4.4	3.9
r_{24}	4.3	16	11	54	4.5	0.11	18	0.1
r_{25}	41	1	160	370	43	39	2.8	370
r_{26}	4.8	5.5	4.5	3.6	4.3	7	3.4	4.7
r_{27}	1.3	6.9	4.3	1.1	2.1	0.43	17	7.2
r_{28}	5.9	8.4	8.7	4.2	3.9	12	4.7	7.3
r_{29}	0.27	0.14	0.1	1.7	0.75	0.1	0.1	0.1
r_{30}	1.4	2.7	6.7	0.52	0.55	3.9	0.21	0.12
r_{31}	0.04	0.01	0.21	0.2	1.9	0.4	2.5	2
r_{32}	7	4.8	1.1	7.8	11	3.5	0.11	0.04
r_{33}	0.71	0.9	0.035	0.59	0.064	0.12	0.11	0.89
r_{34}	2.9	0.057	1.3	0.19	1	1	10	6.8
r_{35}	0.06	0.029	0.26	0.14	0.039	0.026	9.9	0.029
r_{36}	0.089	0.49	0.51	2.5	8.7	1.3	0.021	0.01
r_{37}	10	0.55	0.011	0.52	0.14	0.042	0.17	0.08
r_{38}	3.9	0.051	0.32	0.13	0.2	0.61	0.9	0.37
r_{40}	1.4	1.1	0.28	1	2.3	1.2	1.7	5.3

	1	2	3	4	5	6	7	8
r_{41}	1.9	0.33	0.39	0.61	2.3	1	9.9	3
f_1	0.06	0.41	0.13	0.44	0.42	0.095	0.35	0.045
f_2	0.027	2	0.01	0.01	0.023	0.074	7.6	2.7
f_3	0.27	0.033	4.8	49	1.9	6.9	0.011	4
f_4	0.27	0.1	0.18	5.6	0.1	0.19	0.66	1.9
f_5	0.29	0.39	0.28	0.44	0.35	0.25	0.21	0.45
f_6	0.09	0.25	0.035	0.081	1.7	0.017	0.093	0.25
t_5	0.25	1.1	1.5	0.63	3.1	6.6	0.51	0.18
t_6	0.11	0.59	0.61	1.1	3	1.8	0.21	0.95
t_7	3.8	0.23	13	0.12	0.22	0.36	0.61	3.3
t_8	28	0.15	1.2	3.4	0.13	0.53	0.14	2.6
t_9	2.5	0.85	1.9	0.95	0.1	1.2	3	3.3
m_1	0.61	1	0.8	0.65	0.44	0.64	0.5	0.72
m_3	0.62	0.59	0.47	0.61	0.6	0.53	0.38	0.45
m_4	0.43	0.38	0.47	0.5	0.55	0.54	0.45	0.54
m_5	1.9	2.3	0.5	0.93	0.81	0.7	0.5	0.96
m_6	0.45	0.013	0.19	0.059	2.8	0.036	0.01	0.01
m_7	1.3	0.65	0.6	0.61	0.6	0.78	0.72	0.6
m_8	0.1	5.4	3.4	5.5	2.4	2.3	4.8	0.53
m_9	0.19	0.12	0.33	0.29	0.072	0.14	0.01	0.033
m_{10}	0.01	0.01	0.01	0.011	0.23	0.011	3.9	0.24
m_{11}	0.76	0.68	0.61	0.75	1	1.2	0.51	0.88
m_{12}	2.6	2	3	2.1	1.1	2.7	1.7	1.6
m_{13}	0.67	0.38	0.67	0.21	0.3	0.61	0.27	0.22
m_{14}	0.5	4.9	0.5	3.5	0.51	0.5	0.64	0.5
m_{15}	0.18	0.093	0.18	0.2	0.24	0.22	0.23	0.19
m_{16}	0.54	0.58	0.41	0.65	2.6	2.9	1.4	0.28
m_{17}	0.075	0.047	0.11	0.16	0.2	0.2	0.071	0.1
m_{18}	4.5	2.4	1.5	2.1	2.2	1.4	1.2	1.2
m_{22}	0.3	0.3	0.3	0.3	0.3	2.4	0.38	0.43
m_{23}	0.085	0.18	0.095	0.12	0.065	0.039	0.08	0.074
m_{24}	1.5	2.8	4.2	1.9	2.6	5.9	1.2	7.7
m_{25}	0.65	0.42	0.26	0.46	0.086	0.39	0.1	0.069
m_{26}	1	1.6	0.79	0.84	1	0.3	0.75	0.6
m_{28}	0.038	0.028	0.023	0.45	7.6	8.1	10	1.7
m_{29}	0.01	0.01	0.01	0.06	0.17	0.01	0.01	0.032
m_{30}	4.9	5.5	3.8	5.2	7.2	1.8	9.9	3.9
m_{32}	10	5.7	6.3	8.6	7	3.6	10	5.1
m_{34}	0.16	0.11	0.21	0.85	0.23	0.11	0.11	0.16
m_{35}	0.94	0.92	1.2	6.9	6.8	1.6	0.73	5.4
m_{36}	0.51	0.57	0.51	0.51	0.5	9.8	0.5	0.5
m_{37}	0.01	0.94	0.039	0.01	0.7	2.8	0.031	0.01
m_{38}	1.8	8	0.076	25	10	44	0.24	0.19
m_{39}	0.2	0.22	0.2	0.29	0.2	1.6	0.2	0.45
m_{42}	0.92	0.38	0.23	0.46	0.1	0.096	0.35	0.32
m_{43}	1.1	0.52	0.033	0.054	0.078	0.068	0.022	0.0015
m_{44}	0.68	0.36	0.43	1.4	4.4	0.38	0.47	0.45
m_{45}	0.8	0.79	10	1.6	5.1	0.88	6.2	8.5

	1	2	3	4	5	6	7	8
m_{46}	5.3	0.75	0.97	0.84	0.73	0.5	0.7	2.6
m_{47}	0.25	0.13	0.17	0.18	0.23	0.15	0.25	0.24
p_{11}	1.9	1.8	0.66	0.68	3.2	0.45	1.4	1.9
p_{16}	0.12	0.4	0.16	0.23	0.13	0.3	0.19	0.12
p_{23}	1	1.5	4	15	5.2	30	11	30
p_{25}	1	1.1	1	1.2	4.4	4.3	4.8	1
p_{28}	1.1	2.1	1.2	1.6	8.1	1.1	1.7	3.7
p_{29}	10	25	5.8	24	3.1	19	3.9	5.9
q_1	0.26	0.12	0.55	0.3	0.97	1.5	1.4	0.54
q_3	0.47	0.29	0.31	1.1	1.2	1.7	9.7	2.5

Table III.4 The eight best parameter sets. The values of the parameters after optimization with parallel tempering from random initial starting points in parameter space, as described in Methods.

III.G TABLE OF EXPERIMENTAL DATA SOURCES

An overview of the compiled time course data. Table of the roughly 800 experimental data sets that were compiled and used for fitting the model.

Ref.	Fig.	Reporter	Light	Mut.	w_i
[18] Alabadi, 2001	1A	*TOC1*	LL 12:12	wt, Ler	
[18] Alabadi, 2001	1A	*TOC1*	LL 12:12	lhy (lhy^{TN104})	
[18] Alabadi, 2001	1B	*TOC1*	LL 12:12	CCA1-ox	
[18] Alabadi, 2001	1B	*TOC1*	LL 12:12	wt, Col	
[18] Alabadi, 2001	1C	*TOC1*	LL 12:12	*elf3-1*	
[18] Alabadi, 2001	1C	*TOC1*	LL 12:12	wt, Col	
[31] Baudry, 2010	3D	*CCA1*	LL 12:12	*ztl-4*	
[31] Baudry, 2010	3D	*CCA1*	LL 12:12	wt	
[31] Baudry, 2010	3E	*LHY*	LL 12:12	*ztl-4*	
[31] Baudry, 2010	3E	*LHY*	LL 12:12	wt	
[31] Baudry, 2010	3F	*PRR9*	LL 12:12	*ztl-4*	
[31] Baudry, 2010	3F	*PRR9*	LL 12:12	wt	
[31] Baudry, 2010	S1A	ZTL pr	LL 12:12	wt	
[31] Baudry, 2010	S2	*PRR5*	LL 12:12	wt	
[31] Baudry, 2010	S2	*PRR5*	LL 12:12	*ztl-4*	
[31] Baudry, 2010	S2A	*PRR7*	LL 12:12	*ztl-4*	
[31] Baudry, 2010	S2A	*PRR7*	LL 12:12	wt	
[31] Baudry, 2010	S2A	*PRR7*	LL 12:12	*lhy-20*	
[31] Baudry, 2010	S5	TOC1 pr	LL 12:12	*ztl-4;fkf1;lkp2*	
[31] Baudry, 2010	S5	TOC1 pr	LL 12:12	*ztl-4*	
[31] Baudry, 2010	S5	TOC1 pr	LL 12:12	wt	
[31] Baudry, 2010	S5	*TOC1*	LL 12:12	*ztl-4*	
[31] Baudry, 2010	S5	*TOC1*	LL 12:12	wt	
[31] Baudry, 2010	S5	PRR5 pr	LL 12:12	*ztl-4;fkf1;lkp2*	
[31] Baudry, 2010	S5	PRR5 pr	LL 12:12	*ztl-4*	
[31] Baudry, 2010	S5	PRR5 pr	LL 12:12	wt	
[3] Dai, 2011	1	*NOX*	LL 12:12	wt, Col-0	
[3] Dai, 2011	1	*NOX*	LL 12:12	35S:BOA-8	
[3] Dai, 2011	5C	*NOX*	LL 12:12	wt, Ws	
[3] Dai, 2011	5C	*NOX*	LL 12:12	*cca1-11*	
[3] Dai, 2011	5C	*NOX*	LL 12:12	*lhy-21*	
[3] Dai, 2011	5C	*NOX*	LL 12:12	*lhy-21;cca1-11*	
[3] Dai, 2011	6E	*CCA1*	LL 12:12	wt, Col-0	
[3] Dai, 2011	6E	*CCA1*	LL 12:12	35S:BOA-8	

Ref.	Fig.	Reporter	Light	Mut.	w_i
[3] Dai, 2011	7A	*LHY*	LL 12:12	wt, Col-0	
[3] Dai, 2011	7A	*LHY*	LL 12:12	35S:BOA-8	
[3] Dai, 2011	7B	*CI*	LL 12:12	wt, Col-0	
[3] Dai, 2011	7B	*CI*	LL 12:12	35S:BOA-8	
[3] Dai, 2011	7C	*TOC1*	LL 12:12	wt, Col-0	
[3] Dai, 2011	7C	*TOC1*	LL 12:12	35S:BOA-8	
[3] Dai, 2011	8A	*NOX*	LL 12:12	wt, C24	
[3] Dai, 2011	8A	*NOX*	LL 12:12	*toc1-1*	
[3] Dai, 2011	8B	*CCA1*	LL 12:12	wt, C24	
[3] Dai, 2011	8B	*CCA1*	LL 12:12	*toc1-1*	5
[3] Dai, 2011	8C	*NOX*	LL 12:12	wt, Ler-0	
[3] Dai, 2011	8C	*NOX*	LL 12:12	*gi-3*	
[3] Dai, 2011	8C	*NOX*	LL 12:12	*gi-4*	
[3] Dai, 2011	S1	*CCA1*	LL 12:12	wt, Col-0	
[3] Dai, 2011	S1	*CCA1*	LL 12:12	*boa-1*	
[3] Dai, 2011	S1	*NOX*	LL 12:12	wt, Col-0	
[3] Dai, 2011	S1	*NOX*	LL 12:12	*boa-1*	
[3] Dai, 2011	S3	*NOX*	LL 12:12	wt, Col-0	
[3] Dai, 2011	S3	*NOX*	LL 12:12	CCA1-ox38	
[3] Dai, 2011	S6	*CCA1*	DD 12:12	wt, Col-0	
[3] Dai, 2011	S6	*CCA1*	DD 12:12	35S:BOA-8	
[42] David, 2006	1C	CI pr	LD 16:8	HA-CI protein	
[42] David, 2006	1C	CI pr	LD 8:16	HA-CI protein	
[42] David, 2006	5C	CI pr	LD 16:8	wt	
[42] David, 2006	5C	CI pr	LD 8:16	wt	
[43] Ding, 2007	2A	*CI*	LL 12:12	*cca1-11;lhy-21;toc1-21*	
[43] Ding, 2007	2A	*CI*	LL 12:12	wt	
[43] Ding, 2007	2B	*LUX*	LL 12:12	*cca1-11;lhy-21;toc1-21*	
[43] Ding, 2007	2B	*LUX*	LL 12:12	wt	
[43] Ding, 2007	2C	*PRR9*	LL 12:12	*cca1-11;lhy-21;toc1-21*	
[43] Ding, 2007	2C	*PRR9*	LL 12:12	wt	
[43] Ding, 2007	5A	*PRR9*	LD 8:16	*lhy-21*	
[43] Ding, 2007	5A	*PRR9*	LD 8:16	*cca1-11;toc1-21*	
[43] Ding, 2007	5A	*PRR9*	LD 8:16	*cca1-11;lhy-21;toc1-21*	
[43] Ding, 2007	5A	*PRR9*	LD 8:16	*cca1-11;lhy-21*	
[43] Ding, 2007	5A	*PRR9*	LD 8:16	*cca1-11*	
[43] Ding, 2007	5A	*PRR9*	LD 8:16	wt	
[43] Ding, 2007	5C	*PRR9*	LD 8:16	*toc1-21*	
[43] Ding, 2007	5C	*PRR7*	LD 8:16	*lhy-21;toc1-21*	
[43] Ding, 2007	5C	*PRR7*	LD 8:16	*lhy-21*	
[43] Ding, 2007	5C	*PRR7*	LD 8:16	*cca1-11;toc1-21*	
[43] Ding, 2007	5C	*PRR7*	LD 8:16	*cca1-11;lhy-21;toc1-21*	
[43] Ding, 2007	5C	*PRR7*	LD 8:16	*cca1-11;lhy-21*	
[43] Ding, 2007	5C	*PRR7*	LD 8:16	*cca1-11*	
[43] Ding, 2007	5C	*PRR7*	LD 8:16	wt	
[2] Dixon, 2011	S3	*TOC1*	LD 12:12	*elf3-4*	3
[2] Dixon, 2011	S3	*TOC1*	LD 12:12	*cca1-11;lhy-21;elf3-4*	3
[2] Dixon, 2011	S3	*TOC1*	LD 12:12	*cca1-11;lhy-21*	3
[2] Dixon, 2011	S3, 2	*TOC1*	LD 12:12	wt, Ws	
[2] Dixon, 2011	S3	*PRR9*	LD 12:12	*elf3-4*	3
[2] Dixon, 2011	S3	*PRR9*	LD 12:12	*cca1;lhy-21;elf3-4*	3
[2] Dixon, 2011	S3	*PRR9*	LD 12:12	*cca1-11;lhy-21*	3
[2] Dixon, 2011	S3	*PRR9*	LD 12:12	wt, Ws	
[2] Dixon, 2011	S3	*PRR7*	LD 12:12	*elf3-4*	3
[2] Dixon, 2011	S3	*PRR7*	LD 12:12	*cca1;lhy-21;elf3-4*	3
[2] Dixon, 2011	S3	*PRR7*	LD 12:12	*cca1-11;lhy-21*	3
[2] Dixon, 2011	S3	*PRR7*	LD 12:12	wt, Ws	
[2] Dixon, 2011	S3	*CI*	LD 12:12	*elf3-4*	3
[2] Dixon, 2011	S3	*CI*	LD 12:12	*cca1;lhy-21;elf3-4*	3
[2] Dixon, 2011	S3	*CI*	LD 12:12	*cca1-11;lhy-21*	3
[2] Dixon, 2011	S3	*CI*	LD 12:12	wt, Ws	
[44] Edwards, 2010	2A	*CCA1*	LL 9:15	wt	

Ref.	Fig.	Reporter	Light	Mut.	w_i
[44] Edwards, 2010	2A	*CCA1*	LL 6:18	wt	
[44] Edwards, 2010	2A	*CCA1*	LL 3:21	wt	
[44] Edwards, 2010	2A	*CCA1*	LL 18:6	wt	
[44] Edwards, 2010	2A	*CCA1*	LL 12:12	wt	
[44] Edwards, 2010	2C	*CI*	LL 9:15	wt	
[44] Edwards, 2010	2C	*CI*	LL 6:18	wt	
[44] Edwards, 2010	2C	*CI*	LL 3:21	wt	
[44] Edwards, 2010	2C	*CI*	LL 18:6	wt	
[44] Edwards, 2010	2C	*CI*	LL 12:12	wt	
[44] Edwards, 2010	2E	*TOC1*	LL 9:15	wt	
[44] Edwards, 2010	2E	*TOC1*	LL 6:18	wt	
[44] Edwards, 2010	2E	*TOC1*	LL 3:21	wt	
[44] Edwards, 2010	2E	*TOC1*	LL 18:6	wt	
[44] Edwards, 2010	2E	*TOC1*	LL 12:12	wt	
[44] Edwards, 2010	2B	*CCA1*	DD 12:12	wt	
[44] Edwards, 2010	2B	*CCA1*	DD 18:6	wt	
[44] Edwards, 2010	2B	*CCA1*	DD 6:18	wt	
[44] Edwards, 2010	2D	*CI*	DD 12:12	wt	
[44] Edwards, 2010	2D	*CI*	DD 18:6	wt	
[44] Edwards, 2010	2D	*CI*	DD 6:18	wt	
[44] Edwards, 2010	2E	*TOC1*	DD 18:6	wt	
[44] Edwards, 2010	2F	*TOC1*	DD 12:12	wt	
[44] Edwards, 2010	2F	*TOC1*	DD 6:18	wt	
[44] Edwards, 2010	S3	*CCA1*	DD 12:12	wt	
[44] Edwards, 2010	S3	*CI*	DD 12:12	wt	
[44] Edwards, 2010	S3	*TOC1*	DD 12:12	wt	
[44] Edwards, 2010	S3	*CCA1*	DD 3:21	wt	
[44] Edwards, 2010	S3	*CI*	DD 3:21	wt	
[44] Edwards, 2010	S3	*TOC1*	DD 3:21	wt	
[44] Edwards, 2010	S3	*CCA1*	DD 9:15	wt	
[44] Edwards, 2010	S3	*CI*	DD 9:15	wt	
[44] Edwards, 2010	S3	*TOC1*	DD 9:15	wt	
[40] Farinas, 2011	1A	*CCA1*	LL 12:12	wt	
[40] Farinas, 2011	1B	*CCA1*	LL 8:16	wt	
[40] Farinas, 2011	1C	*CCA1*	LL 16:8	wt	
[40] Farinas, 2011	1A	*RVE8*	LL 12:12	wt	
[40] Farinas, 2011	1B	*RVE8*	LL 8:16	wt	
[40] Farinas, 2011	1C	*RVE8*	LL 16:8	wt	
[40] Farinas, 2011	1D	*TOC1*	LL 12:12	wt	
[40] Farinas, 2011	1F	*TOC1*	LL 16:8	wt	
[40] Farinas, 2011	1E	*TOC1*	LL 8:16	wt	
[40] Farinas, 2011	2C	*RVE8*	LL 16:8	wt	
[40] Farinas, 2011	2C	*RVE8*	LL 16:8	CCA1-ox	
[40] Farinas, 2011	2F	*RVE8*	LL 16:8	*cca1;lhy*	
[40] Farinas, 2011	3C	*CCA1*	LL 16:8	wt	
[40] Farinas, 2011	3C	*CCA1*	LL 16:8	RVE8-ox	
[40] Farinas, 2011	3F	*CCA1*	LL 16:8	*rve8*	
[40] Farinas, 2011	4C	*TOC1*	LL 16:8	wt	
[40] Farinas, 2011	4C	*TOC1*	LL 16:8	RVE8-ox	
[40] Farinas, 2011	4F	*TOC1*	LL 16:8	*rve8*	
[30] Farré, 2005	3F	*TOC1*	LL 12:12	*prr7-3;prr9-1*	3
[30] Farré, 2005	3E	*TOC1*	LL 12:12	*prr9-1*	
[30] Farré, 2005	3E	*TOC1*	LL 12:12	*prr7-3*	
[30] Farré, 2005	3E	*TOC1*	LL 12:12	wt, Col	
[30] Farré, 2005	3D	*LHY*	LL 12:12	*prr7-3;prr9-1*	3
[30] Farré, 2005	3C	*LHY*	LL 12:12	*prr9-1*	
[30] Farré, 2005	3C	*LHY*	LL 12:12	*prr7-3*	
[30] Farré, 2005	3C	*LHY*	LL 12:12	wt, Col	
[30] Farré, 2005	3B	*CCA1*	LL 12:12	*prr7-3;prr9-1*	3
[30] Farré, 2005	3A	*CCA1*	LL 12:12	*prr9-1*	
[30] Farré, 2005	3A	*CCA1*	LL 12:12	*prr7-3*	
[30] Farré, 2005	3A	*CCA1*	LL 12:12	wt, Col	

Ref.	Fig.	Reporter	Light	Mut.	w_i
[30] Farré, 2005	4A	*PRR9*	LD 12:12	wt, Ws	
[30] Farré, 2005	4A	*PRR9*	LD 12:12	*cca1-1*	
[30] Farré, 2005	4A	*PRR9*	LD 12:12	*cca1-1*;*lhy*-R	
[30] Farré, 2005	4B	*PRR9*	LD 12:12	CCA1-ox	
[30] Farré, 2005	4B	*PRR9*	LD 12:12	wt, Col	
[30] Farré, 2005	4C	*PRR9*	LD 12:12	*lhy-1* (-ox)	
[30] Farré, 2005	4C	*PRR9*	LD 12:12	wt, Ler	
[30] Farré, 2005	4D	*PRR7*	LD 12:12	wt, Ws	
[30] Farré, 2005	4D	*PRR7*	LD 12:12	*cca1-1*, Ws	
[30] Farré, 2005	4D	*PRR7*	LD 12:12	*cca1-1*;*lhy*-R	
[30] Farré, 2005	4E	*PRR7*	LD 12:12	CCA1-ox	
[30] Farré, 2005	4E	*PRR7*	LD 12:12	wt, Col	
[30] Farré, 2005	4F	*PRR7*	LD 12:12	*lhy-1* (-ox)	
[30] Farré, 2005	4F	*PRR7*	LD 12:12	wt, Ler	
[45] Farré, 2007	3G	*TOC1*	LL 12:12	P7-ox	3
[45] Farré, 2007	3G	*TOC1*	LL 12:12	wt	
[45] Farré, 2007	3J	*PRR7*	LL 12:12	P7-ox	3
[45] Farré, 2007	3J	*PRR7*	LL 12:12	wt	
[45] Farré, 2007	3D	*LHY*	LL 12:12	P7-ox	3
[45] Farré, 2007	3D	*LHY*	LL 12:12	wt	
[45] Farré, 2007	3A	*CCA1*	LL 12:12	P7-ox	3
[45] Farré, 2007	3A	*CCA1*	LL 12:12	wt	
[45] Farré, 2007	4B	*CCA1*	LL 12:12	P7-ox	3
[45] Farré, 2007	4B	*CCA1*	LL 12:12	wt	
[45] Farré, 2007	4B	*CCA1*	LL 12:12	*prr7-3*	
[45] Farré, 2007	5A	PRR7 pr	LL 12:12	wt	
[45] Farré, 2007	5B	PRR7 pr	DD 12:12	wt	
[45] Farré, 2007	5C	PRR7 pr	LL 12:12	P7-ox	
[45] Farré, 2007	5D	PRR7 pr	DD 12:12	P7-ox	
[46] Fowler, 1999	3A	*GI*	DD 18:6	wt	
[46] Fowler, 1999	3A	*GI*	LD 18:6	wt	
[46] Fowler, 1999	3A	*GI*	LL 18:6	wt	
[46] Fowler, 1999	3B	*GI*	LD 10:14	wt, 18:6 to 10:14	
[46] Fowler, 1999	3B	*GI*	LD 18:6	wt	
[46] Fowler, 1999	3B	*GI*	LD 10:14	wt	
[46] Fowler, 1999	3B	*GI*	LD 18:6	wt	
[46] Fowler, 1999	4A	*GI*	LD 18:6	wt, LDL 18:5:1	
[46] Fowler, 1999	4A	*GI*	LD 18:6	*elf3*, LDL 18:5:1	
[46] Fowler, 1999	4C	*GI*	LL 18:6	*elf3*	
[46] Fowler, 1999	4C	*GI*	LL 18:6	wt	
[46] Fowler, 1999	5A	*GI*	LL 18:6	C-ox	
[46] Fowler, 1999	5A	*GI*	LL 18:6	wt	
[46] Fowler, 1999	5B	*GI*	LD 18:6	LHY-ox	
[46] Fowler, 1999	5B	*GI*	LD 18:6	wt	
[46] Fowler, 1999	5C	*CCA1*	LD 18:6	LHY-ox	
[46] Fowler, 1999	5C	*CCA1*	LD 18:6	wt	
[46] Fowler, 1999	6A	*LHY*	LD 18:6	wt, LDL 18:5:1	
[46] Fowler, 1999	6A	*LHY*	LD 18:6	*gi-3*, LDL 18:5:1	
[46] Fowler, 1999	6B	*CCA1*	LD 18:6	wt, LDL 18:5:1	
[46] Fowler, 1999	6B	*CCA1*	LD 18:6	*gi-3*, LDL 18:5:1	
[25] Fujiwara, 2008	1A	*PRR9*	LL 12:12	wt	
[25] Fujiwara, 2008	1B	*PRR7*	LL 12:12	wt	
[25] Fujiwara, 2008	1C	*PRR5*	LL 12:12	wt	
[25] Fujiwara, 2008	1E	*TOC1*	LL 12:12	wt	
[25] Fujiwara, 2008	3A	*PRR9*	LL 12:12	*ztl-1*	
[47] Hazen, 2005	4A	*CCA1*	LL 12:12	wt	
[47] Hazen, 2005	4A	*CCA1*	LL 12:12	*lux-1*	
[47] Hazen, 2005	4A	*CCA1*	LL 12:12	*lux-2*	
[47] Hazen, 2005	4B	*LHY*	LL 12:12	wt	
[47] Hazen, 2005	4B	*LHY*	LL 12:12	*lux-1*	
[47] Hazen, 2005	4B	*LHY*	LL 12:12	*lux-2*	
[47] Hazen, 2005	4C	*TOC1*	LL 12:12	wt	

Ref.	Fig.	Reporter	Light	Mut.	w_i
[47] Hazen, 2005	4C	*TOC1*	LL 12:12	*lux-1*	
[47] Hazen, 2005	4C	*TOC1*	LL 12:12	*lux-2*	
[48] Helfer, 2011	3A	*LUX*	LD 12:12	wt	
[48] Helfer, 2011	3A	*PRR9*	LD 12:12	wt	
[48] Helfer, 2011	3B	*LUX*	LL 12:12	wt	
[48] Helfer, 2011	3B	*PRR9*	LL 12:12	wt	
[48] Helfer, 2011	3C	*PRR9*	LL 12:12	wt	
[48] Helfer, 2011	3C	*PRR9*	LL 12:12	*lux-4*	
[48] Helfer, 2011	S1B	*LUX*	LD 12:12	wt	
[48] Helfer, 2011	S1B	*NOX*	LD 12:12	wt	
[48] Helfer, 2011	S1C	*LUX*	LL 12:12	wt	
[48] Helfer, 2011	S1C	*NOX*	LL 12:12	wt	
[49] Herrero, 2011	3.11	*PRR9*	LD 8:16	E4-ox	
[49] Herrero, 2011	3.11	*PRR9*	LD 8:16	E3-ox	
[49] Herrero, 2011	3.11	*PRR9*	LD 8:16	wt	
[49] Herrero, 2011	3.11	*PRR7*	LD 8:16	E4-ox	
[49] Herrero, 2011	3.11	*PRR7*	LD 8:16	E3-ox	
[49] Herrero, 2011	3.11	*PRR7*	LD 8:16	wt	
[49] Herrero, 2011	3.11	*PRR9*	LL 12:12	E4-ox	
[49] Herrero, 2011	3.11	*PRR9*	LL 12:12	E3-ox	
[49] Herrero, 2011	3.11	*PRR9*	LL 12:12	wt	
[49] Herrero, 2011	3.11	*PRR7*	LL 12:12	E4-ox	
[49] Herrero, 2011	3.11	*PRR7*	LL 12:12	E3-ox	
[49] Herrero, 2011	3.11	*PRR7*	LL 12:12	wt	
[49] Herrero, 2011	3.2A	*LHY*	LL 12:12	*elf3-4*, LHY::LUC	
[49] Herrero, 2011	3.2A	*LHY*	LL 12:12	E4-ox, LHY::LUC	
[49] Herrero, 2011	3.2B	*LHY*	LL 12:12	*elf4-1*, LHY::LUC	
[49] Herrero, 2011	3.2B	*LHY*	LL 12:12	E3-ox, LHY::LUC	
[49] Herrero, 2011	3.2B	*LHY*	LL 12:12	wt, LHY::LUC	
[8] Hicks, 2001	4B	*ELF3*	LD 12:12	wt	
[8] Hicks, 2001	4D	*ELF3*	LD 12:12	*elf3-1*	
[8] Hicks, 2001	4D	*ELF3*	LD 12:12	*elf3-2*	
[8] Hicks, 2001	4E	*ELF3*	LD 18:6	wt	
[8] Hicks, 2001	4E	*ELF3*	LD 9:15	wt	
[8] Hicks, 2001	6A	*ELF3*	LL 12:12	wt, Ler	
[8] Hicks, 2001	6A	*ELF3*	LL 12:12	lhy	
[50] Hsu, 2012	2H	*CCA1*	LL 12:12	Col	
[50] Hsu, 2012	2H	*CCA1*	LL 12:12	*rve8-1*	
[50] Hsu, 2012	1C	*RVE8*	LL 12:12	Col	
[50] Hsu, 2012	1C	*RVE8*	LL 12:12	*rve8*	
[50] Hsu, 2012	2G	*CCA1*	LL 12:12	Col	
[50] Hsu, 2012	2G	*CCA1*	LL 12:12	*rve8-1*	
[51] Hsu, 2013	5A	*PRR5*	LD 12:12	wt, Col	
[51] Hsu, 2013	5A	*PRR5*	LD 12:12	*rve4,rve6,rve8*	
[51] Hsu, 2013	5B	*TOC1*	LD 12:12	wt, Col	
[51] Hsu, 2013	5B	*TOC1*	LD 12:12	*rve4,rve6,rve8*	
[51] Hsu, 2013	5C	*CCA1*	LD 12:12	wt, Col	
[51] Hsu, 2013	5C	*CCA1*	LD 12:12	*rve4,rve6,rve8*	
[51] Hsu, 2013	5D	*LHY*	LD 12:12	wt, Col	
[51] Hsu, 2013	5D	*LHY*	LD 12:12	*rve4,rve6,rve8*	
[51] Hsu, 2013	5I	*PRR5*	LL 12:12	wt Col	
[51] Hsu, 2013	5I	*PRR5*	LL 12:12	*rve4,rve6,rve8*	
[51] Hsu, 2013	5J	*TOC1*	LL 12:12	wt Col	
[51] Hsu, 2013	5J	*TOC1*	LL 12:12	*rve4,rve6,rve8*	
[51] Hsu, 2013	5K	*CCA1*	LL 12:12	wt Col	
[51] Hsu, 2013	5K	*CCA1*	LL 12:12	*rve4,rve6,rve8*	
[51] Hsu, 2013	5L	*LHY*	LL 12:12	wt Col	
[51] Hsu, 2013	5L	*LHY*	LL 12:12	*rve4,rve6,rve8*	
[51] Hsu, 2013	5M	*TOC1*	LL 12:12	wt Col	
[51] Hsu, 2013	5M	*TOC1*	LL 12:12	*rve4,rve6,rve8*	
[51] Hsu, 2013	5O	*LHY*	LL 12:12	wt Col	
[51] Hsu, 2013	5O	*LHY*	LL 12:12	*rve4,rve6,rve8*	

Ref.	Fig.	Reporter	Light	Mut.	w_i
[51] Hsu, 2013	6A	*RVE8*	LD 12:12	wt Col	
[51] Hsu, 2013	6A	*RVE8*	LD 12:12	*toc1-4*	
[51] Hsu, 2013	6A	*RVE8*	LD 12:12	*lux-1*	
[51] Hsu, 2013	6A	*RVE8*	LD 12:12	CCA1-ox	
[51] Hsu, 2013	6B	*PRR5*	LD 12:12	wt Col, 1 point	
[51] Hsu, 2013	6B	*PRR5*	LD 12:12	*toc1-4*, 1 point	
[51] Hsu, 2013	6B	*PRR5*	LD 12:12	*lux-1*, 1 point	
[51] Hsu, 2013	6B	*PRR5*	LD 12:12	CCA1-ox, 1 point	
[51] Hsu, 2013	6B	*PRR7*	LD 12:12	wt Col, 1 point	
[51] Hsu, 2013	6B	*PRR7*	LD 12:12	*toc1-4*, 1 point	
[51] Hsu, 2013	6B	*PRR7*	LD 12:12	*lux-1*, 1 point	
[51] Hsu, 2013	6B	*PRR7*	LD 12:12	CCA1-ox, 1 point	
[51] Hsu, 2013	6B	*PRR9*	LD 12:12	wt Col, 1 point	
[51] Hsu, 2013	6B	*PRR9*	LD 12:12	*toc1-4*, 1 point	
[51] Hsu, 2013	6B	*PRR9*	LD 12:12	*lux-1*, 1 point	
[51] Hsu, 2013	6B	*PRR9*	LD 12:12	CCA1-ox, 1 point	
[52] Huang, 2012	2L	*ELF4*	LD 12:12	wt	
[52] Huang, 2012	2L	*ELF4*	LD 12:12	TOC1-ox	
[52] Huang, 2012	2K	*GI*	LD 12:12	wt	
[52] Huang, 2012	2K	*GI*	LD 12:12	TOC1-ox	
[52] Huang, 2012	2I	*PRR9*	LD 12:12	TOC1-ox	
[52] Huang, 2012	2I	*PRR9*	LD 12:12	wt	
[52] Huang, 2012	2J	*PRR7*	LD 12:12	TOC1-ox	
[52] Huang, 2012	2J	*PRR7*	LD 12:12	wt	
[52] Huang, 2012	4A	*LHY*	LL 12:12	wt	
[52] Huang, 2012	4A	*LHY*	LL 12:12	TOCRNAi	
[52] Huang, 2012	4B	*PRR7*	LL 12:12	wt	
[52] Huang, 2012	4B	*PRR7*	LL 12:12	TOCRNAi	
[52] Huang, 2012	4C	*PRR9*	LL 12:12	wt	
[52] Huang, 2012	4C	*PRR9*	LL 12:12	TOCRNAi	
[52] Huang, 2012	4D	*GI*	LL 12:12	wt	
[52] Huang, 2012	4D	*GI*	LL 12:12	TOCRNAi	
[52] Huang, 2012	S10	*PRR7*	LD 12:12	wt	
[52] Huang, 2012	S10	*LHY*	LD 12:12	wt	
[52] Huang, 2012	S10	*LHY*	LD 12:12	*toc1-2*	
[52] Huang, 2012	S10	*PRR7*	LD 12:12	*toc1-2*	
[52] Huang, 2012	S7	*PRR9*	LL 12:12	TOC1-ox	
[52] Huang, 2012	S7	*PRR9*	LL 12:12	wt	
[52] Huang, 2012	S7	*PRR7*	LL 12:12	TOC1-ox	
[52] Huang, 2012	S7	*PRR7*	LL 12:12	wt	
[52] Huang, 2012	S7	*LHY*	LL 12:12	TOC1-ox	
[52] Huang, 2012	S7	*LHY*	LL 12:12	wt	
[52] Huang, 2012	S7	*CCA1*	LL 12:12	TOC1-ox	
[52] Huang, 2012	S7	*CCA1*	LL 12:12	wt	
[52] Huang, 2012	S8B	*LUX*	LD 12:12	wt, LUX::LUC	
[52] Huang, 2012	S8C	*LUX*	LL 12:12	wt, LUX::LUC	
[52] Huang, 2012	S8D	*LUX*	LD 12:12	TOC1-ox, LUX::LUC	
[52] Huang, 2012	S8D	*LUX*	LD 12:12	wt, LUX::LUC	
[53] Ito, 2008	2C	*GI*	LL 12:12	*toc1-2;prr5-11*	
[53] Ito, 2008	2C	*GI*	LL 12:12	*toc1-2*	
[53] Ito, 2008	2C	*GI*	LL 12:12	wt	
[53] Ito, 2008	2C	*CCA1*	LL 12:12	*toc1-2;prr5-11*	
[53] Ito, 2008	2C	*CCA1*	LL 12:12	*toc1-2*	
[53] Ito, 2008	2C	*CCA1*	LL 12:12	wt	
[53] Ito, 2008	4	*GI*	LD 16:8	*toc1-2;prr5-11*	
[53] Ito, 2008	4	*GI*	LD 16:8	*toc1-2*	
[53] Ito, 2008	4	*GI*	LD 16:8	*prr5-11*	
[53] Ito, 2008	4	*GI*	LD 16:8	wt	
[53] Ito, 2008	4G	*LHY*	LD 16:8	*prr5-11*	
[53] Ito, 2008	4G	*LHY*	LD 16:8	wt	
[53] Ito, 2008	4H	*LHY*	LD 16:8	*toc1-2;prr5-11*	
[53] Ito, 2008	4H	*LHY*	LD 16:8	*toc1-2*	

Ref.	Fig.	Reporter	Light	Mut.	w_i
[54] Kaczorowski, 2003	6A	*CCA1*	LL 12:12	wt, Red l. entr.	
[54] Kaczorowski, 2003	6A	*CCA1*	LL 12:12	*prr7-1*, Red l. entr.	
[54] Kaczorowski, 2003	6B	*LHY*	LL 12:12	wt, Red l. entr.	
[54] Kaczorowski, 2003	6B	*LHY*	LL 12:12	*prr7-1*, Red l. entr.	
[54] Kaczorowski, 2003	6C	*TOC1*	LL 12:12	wt, Red l. entr.	
[54] Kaczorowski, 2003	6C	*TOC1*	LL 12:12	*prr7-1*, Red l. entr.	
[24] Kiba, 2007	1B	*PRR5*	LD 12:12	wt	
[24] Kiba, 2007	1B	PRR5 pr	LD 12:12	wt	
[24] Kiba, 2007	1C	*PRR5*	LL 12:12	wt	
[24] Kiba, 2007	1C	PRR5 pr	LL 12:12	wt	
[24] Kiba, 2007	1D	PRR5 pr	DD 12:12	wt	
[24] Kiba, 2007	1D	*PRR5*	DD 12:12	wt	
[24] Kiba, 2007	1E	*PRR5*	LD 16:8	wt	
[24] Kiba, 2007	1E	PRR5 pr	LD 16:8	wt	
[24] Kiba, 2007	1F	*PRR5*	LD 8:16	wt	
[24] Kiba, 2007	1F	PRR5 pr	LD 8:16	wt	
[55] Kikis, 2005	3A	*TOC1*	RR 12:12	wt, Ws (DD to RR)	
[55] Kikis, 2005	3A	*TOC1*	RR 12:12	*cca1-1;lhy-12* (DD to RR)	
[55] Kikis, 2005	4A	*CCA1*	RR 12:12	wt, Col (DD to RR)	
[55] Kikis, 2005	4A	*CCA1*	RR 12:12	*toc-101* (DD to RR)	
[55] Kikis, 2005	4C	*LHY*	RR 12:12	wt, Col (DD to RR)	
[55] Kikis, 2005	4C	*LHY*	RR 12:12	*toc-101* (DD to RR)	
[55] Kikis, 2005	5A	*CCA1*	RR 12:12	wt, Col (DD to RR)	
[55] Kikis, 2005	5A	*CCA1*	RR 12:12	*elf4-101* (DD to RR)	
[55] Kikis, 2005	5C	*LHY*	RR 12:12	wt, Col (DD to RR)	
[55] Kikis, 2005	5C	*LHY*	RR 12:12	*elf4-101* (DD to RR)	
[55] Kikis, 2005	5E	*TOC1*	RR 12:12	wt, Col (DD to RR)	
[55] Kikis, 2005	5E	*TOC1*	RR 12:12	*elf4-101* (DD to RR)	
[55] Kikis, 2005	6A	*CCA1*	RR 12:12	wt, Col (DD to RR)	
[55] Kikis, 2005	6A	*CCA1*	RR 12:12	*elf3-1* (DD to RR)	
[55] Kikis, 2005	6C	*LHY*	RR 12:12	wt, Col (DD to RR)	
[55] Kikis, 2005	6C	*LHY*	RR 12:12	*elf3-1* (DD to RR)	
[55] Kikis, 2005	6E	*TOC1*	RR 12:12	wt, Col (DD to RR)	
[55] Kikis, 2005	6E	*TOC1*	RR 12:12	*elf3-1* (DD to RR)	
[55] Kikis, 2005	7A	*ELF4*	RR 12:12	wt, Col (DD to RR)	
[55] Kikis, 2005	7A	*ELF4*	RR 12:12	*elf3-1* (DD to RR)	
[55] Kikis, 2005	7C	*ELF3*	RR 12:12	wt, Col (DD to RR)	
[55] Kikis, 2005	7C	*ELF3*	RR 12:12	*elf4-101*, (DD to RR)	
[55] Kikis, 2005	8A	*ELF4*	RR 12:12	wt, Ws (DD to RR)	
[55] Kikis, 2005	8A	*ELF4*	RR 12:12	*cca1-1;lhy-12* (DD to RR)	
[22] Kim, J-Y, 2003	3C	*LHY*	LD 12:12	LHY-ox	
[22] Kim, J-Y, 2003	3C	*LHY*	LD 12:12	wt	
[22] Kim, J-Y, 2003	3C	LHY pr	LD 12:12	LHY-ox	
[22] Kim, J-Y, 2003	3C	LHY pr	LD 12:12	wt	
[56] Kim, W-Y, 2005	5A	*CI*	LD 16:8	wt, Col	
[56] Kim, W-Y, 2005	5A	*CI*	LD 16:8	*elf3-1*	
[56] Kim, W-Y, 2005	5A	*CI*	LD 16:8	E3-ox	
[56] Kim, W-Y, 2005	5A	*CI*	LD 16:8	*elf3-1* ZTL-ox	
[56] Kim, W-Y, 2005	5A	*CI*	LD 16:8	*elf3-1;ztl-3*	
[56] Kim, W-Y, 2005	5A	*CI*	LD 16:8	*ztl-3*	
[56] Kim, W-Y, 2005	5A	*CI*	LD 16:8	ZTL-ox	
[26] Kim, W-Y, 2007	1B	ZTL pr	LD 12:12	CI-ox	10
[26] Kim, W-Y, 2007	1B	ZTL pr	LD 12:12	*gi-1*	10
[26] Kim, W-Y, 2007	1B	ZTL pr	LD 12:12	wt	10
[26] Kim, W-Y, 2007	1C	CI pr	LD 12:12	wt	
[26] Kim, W-Y, 2007	1C	CI pr	LD 12:12	*ztl-103*	
[26] Kim, W-Y, 2007	1C	CI pr	LD 12:12	wt, cyt	
[26] Kim, W-Y, 2007	1C	CI pr	LD 12:12	*ztl-103*	
[26] Kim, W-Y, 2007	S5	TOC1 pr	LD 12:12	wt, nuc	
[26] Kim, W-Y, 2007	S5	TOC1 pr	LD 12:12	wt, cyt	
[26] Kim, W-Y, 2007	S5	CI pr	LD 12:12	wt, nuc	
[26] Kim, W-Y, 2007	2F	ZTL pr	LD 12:12	*ztl-21*	10

Ref.	Fig.	Reporter	Light	Mut.	w_i
[26] Kim, W-Y, 2007	2F	ZTL pr	LD 12:12	wt	10
[26] Kim, W-Y, 2007	4A	TOC1 pr	LD 12:12	*gi-2*	
[26] Kim, W-Y, 2007	4A	TOC1 pr	LD 12:12	wt	
[26] Kim, W-Y, 2007	4B	*TOC1*	LD 12:12	*gi-2*	
[26] Kim, W-Y, 2007	4B	*TOC1*	LD 12:12	wt	
[57] Kim, Y. 2013	2A	*LHY*	LD 16:8	wt	
[57] Kim, Y. 2013	2A	*LHY*	LD 16:8	*gi-2*	
[57] Kim, Y. 2013	2A	*LHY*	LD 16:8	CI-NLS	
[57] Kim, Y. 2013	2A	*LHY*	LD 16:8	CI-NES	
[57] Kim, Y. 2013	2B	*LHY*	LL 12:12	wt	
[57] Kim, Y. 2013	2B	*LHY*	LL 12:12	*gi-2*	
[57] Kim, Y. 2013	2B	*LHY*	LL 12:12	CI-NLS	
[57] Kim, Y. 2013	2B	*LHY*	LL 12:12	CI-NES	
[57] Kim, Y. 2013	4F	CI pr (N)	LD 16:8	wt	
[57] Kim, Y. 2013	4F	CI pr (C)	LD 16:8	wt	
[57] Kim, Y. 2013	4F	CI pr (N)	LD 16:8	CI-constitutive	
[57] Kim, Y. 2013	4F	CI pr (C)	LD 16:8	CI-constitutive	
[57] Kim, Y. 2013	4G	*LHY*	LD 16:8	CI-constitutive	
[57] Kim, Y. 2013	S1B	*CI*	LD 16:8	wt	
[57] Kim, Y. 2013	S1B	*CI*	LD 16:8	CI-NES (CI cyt)	
[57] Kim, Y. 2013	S1B	*CI*	LD 16:8	CI-NLS (CI nuc)	
[57] Kim, Y. 2013	S1C	CI pr	LD 16:8	CI-NLS	
[57] Kim, Y. 2013	S1C	CI pr	LD 16:8	CI-NES	
[57] Kim, Y. 2013	S2A	*TOC1*	LD 16:8	wt	
[57] Kim, Y. 2013	S2A	*TOC1*	LD 16:8	*gi-2*	
[57] Kim, Y. 2013	S2A	*TOC1*	LD 16:8	CI-NLS	
[57] Kim, Y. 2013	S2A	*TOC1*	LD 16:8	CI-NES	
[57] Kim, Y. 2013	S6	*LHY*	LD 8:16	wt Ler	
[57] Kim, Y. 2013	S6	*LHY*	LD 8:16	*lhy*-mutant	
[5] Kolmos, 2009	4A	*CCA1*	LD 12:12	wt	
[5] Kolmos, 2009	4A	*CCA1*	LD 12:12	*elf4-207*	
[5] Kolmos, 2009	4A	*CCA1*	DD 12:12	wt, Col-0	0.1
[5] Kolmos, 2009	4A	*CCA1*	DD 12:12	*elf4-207*	0.1
[5] Kolmos, 2009	4A	*CCA1*	LL 12:12	*elf4-207*	
[5] Kolmos, 2009	4A	*CCA1*	LL 12:12	Col-0, wt	
[5] Kolmos, 2009	4B	*LHY*	LD 12:12	*elf4-207*	
[5] Kolmos, 2009	4B	*LHY*	LD 12:12	Col-0, wt	
[5] Kolmos, 2009	4B	*LHY*	DD 12:12	wt, Col-0	0.1
[5] Kolmos, 2009	4B	*LHY*	DD 12:12	*elf4-207*	0.1
[5] Kolmos, 2009	4B	*LHY*	LL 12:12	*elf4-207*	
[5] Kolmos, 2009	4B	*LHY*	LL 12:12	Col-0, wt	
[5] Kolmos, 2009	4C	*PRR9*	LD 12:12	*elf4-207*	
[5] Kolmos, 2009	4C	*PRR9*	LD 12:12	Col-0, wt	
[5] Kolmos, 2009	4C	*PRR9*	DD 12:12	wt	0.1
[5] Kolmos, 2009	4C	*PRR9*	DD 12:12	*elf4-207*	0.1
[5] Kolmos, 2009	4C	*PRR9*	LL 12:12	Col-0, wt	
[5] Kolmos, 2009	4C	*PRR9*	LL 12:12	*elf4-207*	
[5] Kolmos, 2009	4D	*PRR7*	LD 12:12	*elf4-207*	
[5] Kolmos, 2009	4D	*PRR7*	LD 12:12	Col-0, wt	
[5] Kolmos, 2009	4D	*PRR7*	DD 12:12	wt, Col-0	0.1
[5] Kolmos, 2009	4D	*PRR7*	DD 12:12	*elf4-207*	0.1
[5] Kolmos, 2009	4D	*PRR7*	LL 12:12	*elf4-207*	
[5] Kolmos, 2009	4D	*PRR7*	LL 12:12	Col-0, wt	
[5] Kolmos, 2009	4E	*CI*	LD 12:12	*elf4-207*	
[5] Kolmos, 2009	4E	*CI*	LD 12:12	Col-0, wt	
[5] Kolmos, 2009	4E	*CI*	DD 12:12	wt, Col-0	0.1
[5] Kolmos, 2009	4E	*CI*	DD 12:12	*elf4-207*	0.1
[5] Kolmos, 2009	4E	*CI*	LL 12:12	*elf4-207*	
[5] Kolmos, 2009	4E	*CI*	LL 12:12	Col-0, wt	
[5] Kolmos, 2009	4F	*TOC1*	LD 12:12	*elf4-207*	
[5] Kolmos, 2009	4F	*TOC1*	LD 12:12	Col-0, wt	
[5] Kolmos, 2009	4F	*TOC1*	DD 12:12	wt	0.1

Ref.	Fig.	Reporter	Light	Mut.	w_i
[5] Kolmos, 2009	4F	*TOC1*	DD 12:12	*elf4-207*	0.1
[5] Kolmos, 2009	4F	*TOC1*	LL 12:12	*elf4-207*	
[5] Kolmos, 2009	4F	*TOC1*	LL 12:12	Col-0, wt	
[5] Kolmos, 2009	4G	*LUX*	LD 12:12	*elf4-207*	
[5] Kolmos, 2009	4G	*LUX*	LD 12:12	Col-0, wt	
[5] Kolmos, 2009	4G	*LUX*	DD 12:12	wt	0.1
[5] Kolmos, 2009	4G	*LUX*	DD 12:12	*elf4-207*	0.1
[5] Kolmos, 2009	4G	*LUX*	LL 12:12	*elf4-207*	
[5] Kolmos, 2009	4G	*LUX*	LL 12:12	Col-0, wt	
[58] Lau, 2011	3C	*TOC1*	LL 16:8	Col	
[58] Lau, 2011	3C	*TOC1*	LL 16:8	*det1-1*	
[58] Lau, 2011	4A	*TOC1*	LL 16:8	CCA1-ox	
[58] Lau, 2011	4A	*TOC1*	LL 16:8	*det1-1*;CCA1-ox	
[58] Lau, 2011	3D	*CI*	LL 16:8	Col	
[58] Lau, 2011	3D	*CI*	LL 16:8	*det1-1*	
[58] Lau, 2011	4B	*CI*	LL 16:8	CCA1-ox	
[58] Lau, 2011	4B	*CI*	LL 16:8	*det1-1*;CCA1-ox	
[58] Lau, 2011	S2B	LHY pr	LL 16:8	Col wt	
[58] Lau, 2011	S2B	LHY pr	LL 16:8	*det1-1*	
[58] Lau, 2011	S2C	*LHY*	LL 16:8	Col wt	
[58] Lau, 2011	S2C	*LHY*	LL 16:8	*det1-1*	
[58] Lau, 2011	S2D	*CCA1*	LL 16:8	Col wt	
[58] Lau, 2011	S2D	*CCA1*	LL 16:8	*det1-1*	
[59] Li, 2011	3D	*ELF4*	LL 12:12	Col-0, wt	
[59] Li, 2011	3D	*ELF4*	LL 12:12	CCA1-ox	
[59] Li, 2011	5E	*ELF4*	LD 12:12	Ws	
[59] Li, 2011	5E	*ELF4*	LD 12:12	*cca1;lhy*	
[59] Li, 2011	5E	*ELF4*	LL 12:12	Ws	
[59] Li, 2011	5E	*ELF4*	LL 12:12	*cca1;lhy*	
[7] Liu, 2001	2C	ELF3 pr	LL 12:12	first plot	
[7] Liu, 2001	2C	ELF3 pr	LL 12:12	second plot	
[60] Locke, 2005	6	*CI*	LL 12:12	wt	
[60] Locke, 2005	6	*CI*	LL 12:12	*cca1;lhy*	
[9] Lu, 2012	1	*ELF3*	LL 12:12	CCA1-ox	
[9] Lu, 2012	1	*ELF3*	LL 12:12	*cca1-1*	
[9] Lu, 2012	1	*ELF3*	LL 12:12	wt	
[9] Lu, 2012	1	*CCA1*	LL 12:12	ELF3-ox	
[9] Lu, 2012	1	*CCA1*	LL 12:12	*elf3-1*	
[9] Lu, 2012	1	*CCA1*	LL 12:12	wt	
[9] Lu, 2012	6	*CI*	LD 16:8	wt	
[9] Lu, 2012	6	*CI*	LD 16:8	*elf3-1*;C-ox	
[9] Lu, 2012	6	*CI*	LD 16:8	*elf3-1*	
[9] Lu, 2012	6	*CI*	LD 16:8	C-ox	
[9] Lu, 2012	7	*CI*	LD 8:16	*elf3*;C-ox	
[9] Lu, 2012	7	*CI*	LD 8:16	*elf3*	
[9] Lu, 2012	7	*CI*	LD 8:16	C-ox	
[9] Lu, 2012	7	*CI*	LD 8:16	wt	
[41] Martin-Tryon, 2007	5A	*CCA1*	LL 12:12	Col	
[41] Martin-Tryon, 2007	5A	*CCA1*	LL 12:12	*gi-201*	
[41] Martin-Tryon, 2007	5A	*CCA1*	LL 12:12	*toc1-2*	
[41] Martin-Tryon, 2007	5B	*LHY*	LL 12:12	Col	
[41] Martin-Tryon, 2007	5B	*LHY*	LL 12:12	*gi-201*	
[41] Martin-Tryon, 2007	5B	*LHY*	LL 12:12	*toc1-2*	
[41] Martin-Tryon, 2007	5C	*TOC1*	LL 12:12	Col	
[41] Martin-Tryon, 2007	5C	*TOC1*	LL 12:12	*gi-201*	
[41] Martin-Tryon, 2007	5C	*TOC1*	LL 12:12	*toc1-2*	
[41] Martin-Tryon, 2007	5D	*CI*	LL 12:12	Col	
[41] Martin-Tryon, 2007	5D	*CI*	LL 12:12	*gi-201*	
[41] Martin-Tryon, 2007	5D	*CI*	LL 12:12	*toc1-2*	
[61] Mas, 2003	2A	TOC1 pr	LD 12:12	TMC	
[61] Mas, 2003	2B	*TOC1*	LD 12:12	TMC	
[61] Mas, 2003	2C	TOC1 pr	LD 12:12	*ztl-1* TMC	

Ref.	Fig.	Reporter	Light	Mut.	w_i
[61] Mas, 2003	2D	*TOC1*	LD 12:12	*ztl-1* TMC	
[61] Mas, 2003	3A	TOC1 pr	LL 12:12	wt, TMC	
[61] Mas, 2003	3B	*TOC1*	LL 12:12	wt, TMC	
[62] Matsushika, 2000	7A	*PRR9*	LD 16:8	wt	
[62] Matsushika, 2000	7A	*PRR7*	LD 16:8	wt	
[62] Matsushika, 2000	7A	*PRR5*	LD 16:8	wt	
[62] Matsushika, 2000	7A	*PRR3*	LD 16:8	wt	
[62] Matsushika, 2000	7A	*PRR1*	LD 16:8	wt	
[62] Matsushika, 2000	7B	*PRR9*	LD 12:12	wt	
[62] Matsushika, 2000	7B	*PRR7*	LD 12:12	wt	
[62] Matsushika, 2000	7B	*PRR5*	LD 12:12	wt	
[62] Matsushika, 2000	7B	*PRR3*	LD 12:12	wt	
[62] Matsushika, 2000	7B	*PRR1*	LD 12:12	wt	
[62] Matsushika, 2000	7C	*PRR9*	LD 8:16	wt	
[62] Matsushika, 2000	7C	*PRR7*	LD 8:16	wt	
[62] Matsushika, 2000	7C	*PRR5*	LD 8:16	wt	
[62] Matsushika, 2000	7C	*PRR3*	LD 8:16	wt	
[62] Matsushika, 2000	7C	*PRR1*	LD 8:16	wt	
[63] Matsushika, 2002	3B	*PRR9*	LL 12:12	P9-ox	3
[63] Matsushika, 2002	3B	*PRR9*	LL 12:12	wt, Col	
[63] Matsushika, 2002	3B	*PRR7*	LL 12:12	P9-ox	3
[63] Matsushika, 2002	3B	*PRR7*	LL 12:12	wt, Col	
[63] Matsushika, 2002	3B	*PRR5*	LL 12:12	P9-ox	3
[63] Matsushika, 2002	3B	*PRR5*	LL 12:12	wt, Col	
[63] Matsushika, 2002	3B	*PRR1*	LL 12:12	P9-ox	3
[63] Matsushika, 2002	3B	*PRR1*	LL 12:12	wt, Col	
[63] Matsushika, 2002	4A	*LHY*	LL 12:12	P9-ox	3
[63] Matsushika, 2002	4A	*LHY*	LL 12:12	wt, Col	
[63] Matsushika, 2002	4B	*ELF3*	LL 12:12	P9-ox	3
[63] Matsushika, 2002	4B	*ELF3*	LL 12:12	wt, Col	
[63] Matsushika, 2002	4A	*CCA1*	LL 12:12	P9-ox	3
[63] Matsushika, 2002	4A	*CCA1*	LL 12:12	wt, Col	
[34] McWatters, 2007	6G	*ELF4*	LD 12:12	wt C24	
[34] McWatters, 2007	6G	*ELF4*	LD 12:12	*toc1-1*	
[34] McWatters, 2007	S2A	*CCA1*	LL 12:12	Ws	
[34] McWatters, 2007	S2A	*CCA1*	LL 12:12	*elf4-1*	
[34] McWatters, 2007	S2B	*LHY*	LL 12:12	Ws	
[34] McWatters, 2007	S2B	*LHY*	LL 12:12	*elf4-1*	
[34] McWatters, 2007	S2D	*TOC1*	LL 12:12	Ws	
[34] McWatters, 2007	S2D	*TOC1*	LL 12:12	*elf4-1*	
[34] McWatters, 2007	S1A	*ELF4*	LL 12:12	Ws	
[34] McWatters, 2007	S1A	*ELF4*	LL 12:12	ELF4-ox-11	
[17] Mizoguchi, 2002	7C	*GI*	LD 16:8	Ler	
[17] Mizoguchi, 2002	7C	*GI*	LD 16:8	*lhy-12*	
[17] Mizoguchi, 2002	7C	*GI*	LD 16:8	*cca1-1*	
[17] Mizoguchi, 2002	7D	*GI*	LD 16:8	*cca1-1;lhy-12*	
[17] Mizoguchi, 2002	7E	*TOC1*	LD 16:8	Ler	
[17] Mizoguchi, 2002	7E	*TOC1*	LD 16:8	*lhy-12*	
[17] Mizoguchi, 2002	7E	*TOC1*	LD 16:8	*cca1-1*	
[17] Mizoguchi, 2002	7F	*TOC1*	LD 16:8	*cca1-1;lhy-12*	
[17] Mizoguchi, 2002	7K	*LHY*	LL 16:8	Ler	
[17] Mizoguchi, 2002	7K	*LHY*	LL 16:8	*gi-3*	
[17] Mizoguchi, 2002	7L	*CCA1*	LL 16:8	Ler	
[17] Mizoguchi, 2002	7L	*CCA1*	LL 16:8	*gi-3*	
[17] Mizoguchi, 2002	6C	*GI*	LL 16:8	Ler	
[17] Mizoguchi, 2002	6C	*GI*	LL 16:8	*lhy-12*	
[17] Mizoguchi, 2002	6C	*GI*	LL 16:8	*cca1-1*	
[17] Mizoguchi, 2002	6G	*TOC1*	LL 16:8	Ler	
[17] Mizoguchi, 2002	6G	*TOC1*	LL 16:8	*lhy-12*	
[17] Mizoguchi, 2002	6G	*TOC1*	LL 16:8	*cca1-1*	
[17] Mizoguchi, 2002	6D	*GI*	LL 16:8	*lhy-12;cca1-1*	
[17] Mizoguchi, 2002	6H	*TOC1*	LL 16:8	*lhy-12;cca1-1*	

Ref.	Fig.	Reporter	Light	Mut.	w_i
[17] Mizoguchi, 2002	2C	*CCA1*	LL 16:8	wt	
[17] Mizoguchi, 2002	2C	*CCA1*	LL 16:8	*lhy-12*	
[17] Mizoguchi, 2002	2D	*LHY*	LL 16:8	wt	
[17] Mizoguchi, 2002	2D	*LHY*	LL 16:8	*cca1-1*	
[64] Nakamichi, 2003	1	*PRR9*	LD 12:12	wt	
[64] Nakamichi, 2003	1	*PRR7*	LD 12:12	wt	
[64] Nakamichi, 2003	1	*PRR5*	LD 12:12	wt	
[64] Nakamichi, 2003	1	*PRR1*	LD 12:12	wt	
[64] Nakamichi, 2003	3	*PRR1*	DD 12:12	wt	
[64] Nakamichi, 2003	3	*PRR5*	DD 12:12	wt	
[64] Nakamichi, 2003	3	*PRR7*	DD 12:12	wt	
[64] Nakamichi, 2003	3	*PRR7*	DD 12:12	wt	
[64] Nakamichi, 2003	3	*PRR9*	DD 12:12	wt	
[64] Nakamichi, 2003	4	*CCA1*	DD 12:!2	wt	
[64] Nakamichi, 2003	4	*LHY*	DD 12:12	wt	
[65] Nakamichi, 2005	4A	*CCA1*	LL 12:12	wt, Col	
[65] Nakamichi, 2005	4A	*CCA1*	LL 12:12	*prr5-11;prr7-11*	
[65] Nakamichi, 2005	4B	*LHY*	LL 12:12	wt, Col	
[65] Nakamichi, 2005	4B	*LHY*	LL 12:12	*prr5-11;prr7-11*	
[65] Nakamichi, 2005	4C	*TOC1*	LL 12:12	wt, Col	
[65] Nakamichi, 2005	4C	*TOC1*	LL 12:12	*prr5-11;prr7-11*	
[65] Nakamichi, 2005	5A	*CCA1*	DD 12:12	wt, Col	
[65] Nakamichi, 2005	5A	*CCA1*	DD 12:12	*prr5-11;prr7-11*	
[65] Nakamichi, 2005	5B	*LHY*	DD 12:12	wt, Col	
[65] Nakamichi, 2005	5B	*LHY*	DD 12:12	*prr5-11;prr7-11*	
[65] Nakamichi, 2005	6A	*CCA1*	LD 12:12	wt, Col	
[65] Nakamichi, 2005	6A	*CCA1*	LD 12:12	*prr5-11;prr7-11*	
[65] Nakamichi, 2005	6B	*LHY*	LD 12:12	wt, Col	
[65] Nakamichi, 2005	6B	*LHY*	LD 12:12	*prr5-11;prr7-11*	
[65] Nakamichi, 2005	6C	*TOC1*	LD 12:12	wt, Col	
[65] Nakamichi, 2005	6C	*TOC1*	LD 12:12	*prr5-11;prr7-11*	
[65] Nakamichi, 2005	6E	*CI*	LD 12:12	wt, Col	
[65] Nakamichi, 2005	6E	*CI*	LD 12:12	*prr5-11;prr7-11*	
[66] Nakamichi, 2005b	5A	*CCA1*	LL 12:12	wt, Col	
[66] Nakamichi, 2005b	5A	*CCA1*	LL 12:12	*prr9-10;prr7-11;prr5-11*	
[66] Nakamichi, 2005b	5B	*TOC1*	LL 12:12	wt, Col	
[66] Nakamichi, 2005b	5B	*TOC1*	LL 12:12	*prr9-10;prr7-11;prr5-11*	
[66] Nakamichi, 2005b	5C	*CI*	LL 12:12	wt, Col	
[66] Nakamichi, 2005b	5C	*CI*	LL 12:12	*prr9-10;prr7-11;prr5-11*	
[66] Nakamichi, 2005b	6C	*TOC1*	LD 12:12	wt	
[67] Nakamichi, 2010	6A	*CCA1*	LD 12:12	wt	
[67] Nakamichi, 2010	6B	PRR9 pr	LD 12:12	*prr5;prr7*	
[67] Nakamichi, 2010	6A	PRR9 pr	LD 12:12	wt	
[67] Nakamichi, 2010	6A	PRR7 pr	LD 12:12	wt	
[67] Nakamichi, 2010	6C	PRR5 pr	LD 12:12	*prr7;prr9*	
[67] Nakamichi, 2010	6A	PRR5 pr	LD 12:12	wt	
[67] Nakamichi, 2010	6C	*LHY*	LD 12:12	*prr7;prr9*	
[67] Nakamichi, 2010	6B	*LHY*	LD 12:12	*prr5;prr7*	
[67] Nakamichi, 2010	6A	*LHY*	LD 12:12	wt	
[67] Nakamichi, 2010	6B	*CCA1*	LD 12:12	*prr5;prr7*	
[67] Nakamichi, 2010	6C	*CCA1*	LD 12:12	*prr7;prr9*	
[68] Nakamichi, 2012	S4B	*LHY*	LL 12:12	wt	
[68] Nakamichi, 2012	S4B	*LHY*	LL 12:12	P5-ox	3
[68] Nakamichi, 2012	S4B	*PRR9*	LL 12:12	wt	
[68] Nakamichi, 2012	S4B	*PRR7*	LL 12:12	wt	
[69] Niwa, 2007	6L	*CI*	LD 10:14	wt	
[69] Niwa, 2007	6L	*CI*	LD 10:14	*toc1-2;cca1-1*	
[69] Niwa, 2007	6R	*CI*	LD 10:14	wt	
[69] Niwa, 2007	6R	*CI*	LD 10:14	*toc1-2;cca1-1*	
[69] Niwa, 2007	8	*CI*	LD 10:14	*toc1-2;cca1-1;lhy-11*	
[69] Niwa, 2007	8	*CI*	LD 10:14	*cca1-1;lhy-11*	
[69] Niwa, 2007	8	*CI*	LD 10:14	wt	

Ref.	Fig.	Reporter	Light	Mut.	w_i
[70] Nusinow, 2011	1	*ELF3*	LL 12:12	wt	
[70] Nusinow, 2011	1	*ELF4*	LL 12:12	wt	
[70] Nusinow, 2011	1	*LUX*	LL 12:12	wt	
[70] Nusinow, 2011	S11	*ELF4*	LD 8:16	wt	
[70] Nusinow, 2011	S11	*ELF4*	LD 8:16	*lhy-1* (-ox)	
[70] Nusinow, 2011	S11	*ELF3*	LD 8:16	wt	
[70] Nusinow, 2011	S11	*ELF3*	LD 8:16	*lhy-1* (-ox)	
[70] Nusinow, 2011	S11	*LUX*	LD 8:16	wt	
[70] Nusinow, 2011	S11	*LUX*	LD 8:16	*lhy-1* (-ox)	
[32] Onai, 2005	1D	*GI*	LL 12:12	wt	
[32] Onai, 2005	1D	*GI*	LL 12:12	*lux* (*pcl1-1*)	
[32] Onai, 2005	1F	*TOC1*	LL 12:12	wt	
[32] Onai, 2005	1F	*TOC1*	LL 12:12	*lux* (*pcl1-1*)	
[32] Onai, 2005	1G	*ELF4*	LL 12:12	wt	
[32] Onai, 2005	1G	*ELF4*	LL 12:12	*lux* (*pcl1-1*)	
[32] Onai, 2005	1H	*CCA1*	LL 12:12	wt	
[32] Onai, 2005	1H	*CCA1*	LL 12:12	*lux* (*pcl1-1*)	
[32] Onai, 2005	1I	*LHY*	LL 12:12	wt	
[32] Onai, 2005	1I	*LHY*	LL 12:12	*lux* (*pcl1-1*)	
[32] Onai, 2005	6A	*LUX*	LL 12:12	wt	
[32] Onai, 2005	6A	*LUX*	LL 12:12	*lux* (*pcl1-1*)	
[32] Onai, 2005	6G	*LUX*	LL 12:12	wt	
[32] Onai, 2005	6G	*LUX*	LL 12:12	*lux* (*pcl1-1*)	
[32] Onai, 2005	6G	*LUX*	LL 12:12	LUX-ox	
[32] Onai, 2005	7A	*LUX*	LL 12:12	wt	
[32] Onai, 2005	7A	*LUX*	LL 12:12	*lux* (*pcl1-1*)	
[32] Onai, 2005	7A	*LUX*	LL 12:12	LUX-ox	
[32] Onai, 2005	7B	*GI*	LL 12:12	wt	
[32] Onai, 2005	7B	*GI*	LL 12:12	*lux* (*pcl1-1*)	
[32] Onai, 2005	7B	*GI*	LL 12:12	LUX-ox	
[32] Onai, 2005	7C	*TOC1*	LL 12:12	wt	
[32] Onai, 2005	7C	*TOC1*	LL 12:12	*lux* (*pcl1-1*)	
[32] Onai, 2005	7C	*TOC1*	LL 12:12	LUX-ox	
[32] Onai, 2005	7D	*ELF4*	LL 12:12	wt	
[32] Onai, 2005	7D	*ELF4*	LL 12:12	*lux* (*pcl1-1*)	
[32] Onai, 2005	7D	*ELF4*	LL 12:12	LUX-ox	
[32] Onai, 2005	7E	*CCA1*	LL 12:12	wt	
[32] Onai, 2005	7E	*CCA1*	LL 12:12	*lux* (*pcl1-1*)	
[32] Onai, 2005	7E	*CCA1*	LL 12:12	LUX-ox	
[32] Onai, 2005	7F	*LHY*	LL 12:12	wt	
[32] Onai, 2005	7F	*LHY*	LL 12:12	*lux* (*pcl1-1*)	
[32] Onai, 2005	7F	*LHY*	LL 12:12	LUX-ox	
[71] Park, 1999	3A	*GI*	LL 12:12	wt, Col	
[71] Park, 1999	3A	*GI*	LL 12:12	*gi-1*	
[71] Park, 1999	3A	*GI*	LL 12:12	*gi-2*	
[71] Park, 1999	3A	*CCA1*	LL 12:12	wt, Col	
[71] Park, 1999	3A	*CCA1*	LL 12:12	*gi-1*	
[71] Park, 1999	3A	*CCA1*	LL 12:12	*gi-2*	
[71] Park, 1999	3A	*LHY*	LL 12:12	wt, Col	
[71] Park, 1999	3A	*LHY*	LL 12:12	*gi-1*	
[71] Park, 1999	3A	*LHY*	LL 12:12	*gi-2*	
[71] Park, 1999	3B	*LHY*	DD 12:12	wt, Col	
[71] Park, 1999	3B	*LHY*	DD 12:12	*gi-1*	
[71] Park, 1999	3B	*LHY*	DD 12:12	*gi-2*	
[71] Park, 1999	3B	*GI*	DD 12:12	wt, Col	
[71] Park, 1999	3B	*GI*	DD 12:12	*gi-1*	
[71] Park, 1999	3B	*GI*	DD 12:12	*gi-2*	
[72] Pokhilko, 2012	2A	*TOC1*	LD 12:12	wt	
[72] Pokhilko, 2012	2A	*TOC1*	LD 12:12	*cca1;lhy*	
[72] Pokhilko, 2012	2B	*LUX*	LD 12:12	wt	
[72] Pokhilko, 2012	2B	*LUX*	LD 12:12	*cca1;lhy*	
[72] Pokhilko, 2012	3C	*TOC1*	LD 12:12	*cca1;lhy;gi*	

Ref.	Fig.	Reporter	Light	Mut.	w_i
[72] Pokhilko, 2012	3C	*TOC1*	LD 12:12	*cca1;lhy*	
[72] Pokhilko, 2012	3D	*LUX*	LD 12:12	*cca1;lhy;gi*	
[72] Pokhilko, 2012	3D	*LUX*	LD 12:12	*cca1;lhy*	
[72] Pokhilko, 2012	5A	*CCA1*	LD 12:12	wt	
[72] Pokhilko, 2012	5A	*CCA1*	LD 12:12	*toc1*	
[72] Pokhilko, 2012	5A	*CCA1*	LD 12:12	T-ox	
[72] Pokhilko, 2012	5B	*LHY*	LD 12:12	wt	
[72] Pokhilko, 2012	5B	*LHY*	LD 12:12	*toc1*	
[72] Pokhilko, 2012	5B	*LHY*	LD 12:12	T-ox	
[73] Rawat, 2011	4B	RVE8 pr	LL 12:12	wt	
[73] Rawat, 2011	6A	*CCA1*	LL 12:12	wt Col	
[73] Rawat, 2011	6A	*CCA1*	LL 12:12	*rve8-1*	
[73] Rawat, 2011	6A	*CCA1*	LL 12:12	RVE8-ox	
[73] Rawat, 2011	6B	*LHY*	LL 12:12	wt Col	
[73] Rawat, 2011	6B	*LHY*	LL 12:12	*rve8-1*	
[73] Rawat, 2011	6B	*LHY*	LL 12:12	RVE8-ox	
[73] Rawat, 2011	6C	*TOC1*	LL 12:12	wt Col	
[73] Rawat, 2011	6C	*TOC1*	LL 12:12	*rve8-1*	
[73] Rawat, 2011	6C	*TOC1*	LL 12:12	RVE8-ox	
[73] Rawat, 2011	6D	*CCA1*	LL 12:12	*rve8-1*;RVE8-ox	
[73] Rawat, 2011	6E	*LHY*	LL 12:12	*rve8-1*;RVE8-ox	
[73] Rawat, 2011	6F	*TOC1*	LL 12:12	*rve8-1*;RVE8-ox	
[73] Rawat, 2011	8C	*PRR5*	LL 12:12	wt Col	
[73] Rawat, 2011	8C	*PRR5*	LL 12:12	*rve8-1*	
[73] Rawat, 2011	8C	*PRR5*	LL 12:12	RVE8-ox	
[73] Rawat, 2011	8D	*TOC1*	LL 12:12	wt Col	
[73] Rawat, 2011	8D	*TOC1*	LL 12:12	*rve8-1*	
[73] Rawat, 2011	8D	*TOC1*	LL 12:12	RVE8-ox	
[73] Rawat, 2011	8G	*RVE8*	LL 12:12?	Col	
[73] Rawat, 2011	8G	*RVE8*	LL 12:12?	*prr9;prr7;prr5*	
[73] Rawat, 2011	8E	*PRR7*	LL 12:12	wt Col	
[73] Rawat, 2011	8E	*PRR7*	LL 12:12	*rve8-1*	
[73] Rawat, 2011	8E	*PRR7*	LL 12:12	RVE8-ox	
[73] Rawat, 2011	S1A	*RVE8*	LL 12:12	Col	
[73] Rawat, 2011	S1A	*RVE8*	LL 12:12	RVE8-ox	
[73] Rawat, 2011	S1B	*RVE8*	LL 12:12	*rve8-1*	
[74] Sato, 2002	2B	*CI*	LL 12:12	wt, Col	
[74] Sato, 2002	2B	*CI*	LL 12:12	P5-ox	3
[74] Sato, 2002	3	*CCA1*	LL 12:12	wt, Col	
[74] Sato, 2002	3	*CCA1*	LL 12:12	P5-ox	3
[74] Sato, 2002	3	*LHY*	LL 12:12	wt, Col	
[74] Sato, 2002	3	*LHY*	LL 12:12	P5-ox	3
[74] Sato, 2002	4	*PRR9*	LL 12:12	wt, Col	
[74] Sato, 2002	4	*PRR9*	LL 12:12	P5-ox	3
[74] Sato, 2002	4	*PRR7*	LL 12:12	wt, Col	
[74] Sato, 2002	4	*PRR7*	LL 12:12	P5-ox	3
[74] Sato, 2002	4	*PRR5*	LL 12:12	wt, Col	
[74] Sato, 2002	4	*PRR5*	LL 12:12	P5-ox	3
[74] Sato, 2002	4	*PRR1*	LL 12:12	wt, Col	
[74] Sato, 2002	4	*PRR1*	LL 12:12	P5-ox	3
[38] Somers, 2004	9A	*CCA1*	LD 12:12	wt	
[38] Somers, 2004	9A	*CCA1*	LD 12:12	Z-ox	
[38] Somers, 2004	9A	*CCA1*	LD 12:12	*ztl-1*	
[38] Somers, 2004	9A	*CCA1*	LD 12:12	*ztl-2*	
[38] Somers, 2004	9A	*CCA1*	LD 12:12	*ztl-3*	
[38] Somers, 2004	9A	*CCA1*	LD 12:12	wt	
[38] Somers, 2004	9A	*CCA1*	LL 12:12	wt, Col	
[38] Somers, 2004	9A	*CCA1*	LL 12:12	wt, C24	
[38] Somers, 2004	9A	*CCA1*	LL 12:12	*ztl-1*, C24	
[38] Somers, 2004	9A	*CCA1*	LL 12:12	*ztl-2*, C24	
[38] Somers, 2004	9A	*CCA1*	LL 12:12	*ztl-3*, Col	
[38] Somers, 2004	9A	*CCA1*	LL 12:12	Z-ox	

Ref.	Fig.	Reporter	Light	Mut.	w_i
[38] Somers, 2004	9B	*TOC1*	LL 12:12	wt, C24	
[38] Somers, 2004	9B	*TOC1*	LL 12:12	wt, Col	
[38] Somers, 2004	9B	*TOC1*	LL 12:12	*ztl-1*, C24	
[38] Somers, 2004	9B	*TOC1*	LL 12:12	*ztl-2*, C24	
[38] Somers, 2004	9B	*TOC1*	LL 12:12	*ztl-3*, Col	
[38] Somers, 2004	9B	*TOC1*	LL 12:12	Z-ox	
[38] Somers, 2004	9C	*ELF3*	LD 12:12	wt,	
[38] Somers, 2004	9C	*ELF3*	LD 12:12	Z-ox	
[38] Somers, 2004	9C	*ELF3*	LD 12:12	*ztl-1*	
[38] Somers, 2004	9C	*ELF3*	LD 12:12	*ztl-2*	
[38] Somers, 2004	9C	*ELF3*	LD 12:12	*ztl-3*	
[38] Somers, 2004	9C	*ELF3*	LL 12:12	wt, Col	
[38] Somers, 2004	9C	*ELF3*	LL 12:12	wt, C24	
[38] Somers, 2004	9C	*ELF3*	LL 12:12	*ztl-1*, C24	
[38] Somers, 2004	9C	*ELF3*	LL 12:12	*ztl-2*, C24	
[38] Somers, 2004	9C	*ELF3*	LL 12:12	*ztl-3*, Col	
[38] Somers, 2004	9C	*ELF3*	LL 12:12	Z-ox	
[75] Song, 2005	1B	*LHY*	LD 12:12	wt LHY:LUC	
[75] Song, 2005	1B	*LHY*	LD 12:12	*det1-1* LHY:LUC	
[75] Song, 2005	1D	LHY pr	LD 12:12	wt	
[75] Song, 2005	1D	LHY pr	LD 12:12	*det1-1*	
[75] Song, 2005	4B	LHY pr	LD 12:12	*lhy-1*	
[75] Song, 2005	4B	LHY pr	LD 12:12	*lhy-1;det1-1*	
[76] Song, 2012	4C	*LHY*	LD 16:8	Col	
[76] Song, 2012	4C	*LHY*	LD 16:8	*elf3-1*	
[76] Song, 2012	6C	LHY pr	LD 16:8	*elf3-1*	
[76] Song, 2012	6D	LHY pr	LD 16:8	wt	
[11] Wang, 1998	6	*CCA1*	LL 12:12	wt	
[11] Wang, 1998	6	*CCA1*	LL 12:12	C-ox	
[11] Wang, 1998	6	*LHY*	LL 12:12	wt	
[11] Wang, 1998	6	*LHY*	LL 12:12	C-ox	
[23] Wang, 2010	3B	TOC1 pr	LD 12:12	*prr5-1*	
[23] Wang, 2010	3B	TOC1 pr	LD 12:12	wt	
[23] Wang, 2010	3D	*TOC1*	LD 12:12	*prr5-1*	
[23] Wang, 2010	3D	*TOC1*	LD 12:12	wt, TMC	
[20] Yakir, 2009	1C	CCA1 pr	LL 14:10	(quantified in GIMP)	
[20] Yakir, 2009	S2	CCA1 pr	LL 14:10		
[20] Yakir, 2009	2E	CCA1 pr	LL 14:10?	CCA1::CCA1-HA-YFP *cca1-1*	
[20] Yakir, 2009	2E	CCA1 pr	DD 14:10?	CCA1::CCA1-HA-YFP *cca1-1*	
[77] Yamashino, 2008	4C	*CCA1*	LD 12:12	wt, Col	
[77] Yamashino, 2008	4C	*CCA1*	LD 12:12	*prr9;prr7;prr5*	
[77] Yamashino, 2008	4D	*GI*	LD 12:12	wt, Col	
[77] Yamashino, 2008	4D	*GI*	LD 12:12	*cca1;lhy;toc1*	
[77] Yamashino, 2008	4D	*GI*	LD 12:12	*prr9;prr7;prr5*	
[77] Yamashino, 2008	5B	*CCA1*	LD 12:12	wt, Col	
[77] Yamashino, 2008	5B	*CCA1*	LD 12:12	*prr9;prr7;prr5;toc1*	
[77] Yamashino, 2008	5C	*GI*	LD 12:12	wt, Col	
[77] Yamashino, 2008	5C	*GI*	LD 12:12	*cca1;lhy;toc1*	
[77] Yamashino, 2008	5C	*GI*	LD 12:12	*prr9;prr7;prr5;toc1*	
[4] Yu, 2008	3A	*ELF3*	LD 8:16	*cop1-4*	
[4] Yu, 2008	3A	*ELF3*	LD 8:16	wt	
[4] Yu, 2008	3A	*ELF3*	LD 16:8	*cop1-4*	
[4] Yu, 2008	3A	*ELF3*	LD 16:8	wt	
[4] Yu, 2008	3B	*GI*	LD 8:16	*cop1-4*	
[4] Yu, 2008	3B	*GI*	LD 8:16	wt	
[4] Yu, 2008	3B	*GI*	LD 16:8	*cop1-4*	
[4] Yu, 2008	3B	*GI*	LD 16:8	wt	
[4] Yu, 2008	5C	E3 pr	LD 8:16	*cop1-4*, Col, nuc	
[4] Yu, 2008	5C	E3 pr	LD 8:16	wt, Col, nuc	
[4] Yu, 2008	5C	E3 pr	LD 16:8	*cop1-4*, Col, nuc	
[4] Yu, 2008	5C	E3 pr	LD 16:8	wt, Col, nuc	
[4] Yu, 2008	5F	*LHY*	LD 12:12	wt,	

Ref.	Fig.	Reporter	Light	Mut.	w_i
[4] Yu, 2008	5F	*LHY*	LD 12:12	*cop1-4*	
[4] Yu, 2008	5F	*LHY*	LD 12:12	E3-ox	
[4] Yu, 2008	5F	*LHY*	LD 12:12	*cop1-4;E3-ox*	
[4] Yu, 2008	6E	CI pr	LD 8:16	*cop1-4*	
[4] Yu, 2008	6E	CI pr	LD 8:16	*elf3-8*	
[4] Yu, 2008	6E	CI pr	LD 8:16	wt, left panel	
[4] Yu, 2008	6E	CI pr	LD 8:16	wt, right panel	

Table III.6 Data used for model fitting. Each row corresponds to a time course of either mRNA (XYZ) or protein (XYZ pr) concentration in some light condition, extracted from the listed publications. All time courses were given a weight $w_i = 1$ unless otherwise specified here. Increased weights were used where important results were otherwise found to be difficult to reproduce: *cca1* oscillations in *toc1* [3], ZTL protein level in wt and mutants [26], interaction between *elf3* and *cca1;lhy* [2], period lengthening in *prr7;prr9* [30], small effects of PRR7-ox [45], small period shortening in PRR9-ox [63], and level changes and period preservation in PRR5-ox [68, 74]. Decreased weights were used for some DD data where the system became arrhythmic in wt [5]. Experimental data was obtained in many different light conditions, which were simulated in the model. LD, LL and DD refers to light/dark, constant light and constant dark, respectively. Numbers such as 12:12 refer to hours of light and dark per period during entrainment, and for LD also during measurements. RR is constant red light, which was simulated as LL. Data were not manipulated in any way (e.g. stitched, joined or normalized) before entering the costfunction described in Methods in the main text. All raw data are available from the website mentioned in main text.

Parameter	Value
m_{19}	0.2
m_{20}	1.8
m_{21}	0.1
m_{27}	0.1
m_{31}	0.3
m_{33}	13
n_5	0.23
n_6	20
n_{14}	0.1
p_6	0.6
p_7	0.3
p_{10}	0.2
p_{12}	8
p_{13}	0.7
p_{14}	0.3
p_{15}	3

Table III.5 Constant parameters. These parameters control c_P, COP1, ZTL and the ZTL-GI complex, and were not included in the optimization process. Instead, they were taken from P2012 (c_P and COP1) or fitted manually (ZTL and ZTL-GI).

REFERENCES

1. A. Pokhilko, P. Más, and A. J. Millar, "Modelling the widespread effects of TOC1 signalling on the plant circadian clock and its outputs," *BMC Syst Biol*, vol. 7, no. 1, pp. 1–12, 2013.
2. 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.
3. 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.
4. 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.
5. 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.
6. 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.
7. X. L. Liu, M. F. Covington, C. Fankhauser, J. Chory, and D. R. Wagner, "*ELF3* encodes a circadian clock–regulated nuclear protein that functions in an *Arabidopsis PHYB* signal transduction pathway," *Plant Cell*, vol. 13, no. 6, pp. 1293–1304, 2001.
8. K. A. Hicks, T. M. Albertson, and D. R. Wagner, "*EARLY FLOWERING 3* encodes a novel protein that regulates circadian clock function and flowering in *Arabidopsis*," *Plant Cell*, vol. 13, no. 6, pp. 1281–1292, 2001.
9. S. X. Lu, C. J. Webb, S. M. Knowles, S. H. Kim, Z. Wang, and E. M. Tobin, "CCA1 and ELF3 interact in the control of hypocotyl length and flowering time in *Arabidopsis*," *Plant Physiol*, vol. 158, no. 2, pp. 1079–1088, 2012.
10. 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.

11. 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.
12. 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.
13. 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.
14. P. D. Gould, J. C. Locke, C. Larue, M. M. Southern, S. J. Davis, S. Hanano, R. Moyle, R. Milich, J. Putterill, A. J. Millar, and A. Hall, "The molecular basis of temperature compensation in the *Arabidopsis* circadian clock," *Plant Cell*, vol. 18, no. 5, pp. 1177–1187, 2006.
15. K. Higo, Y. Ugawa, M. Iwamoto, and T. Korenaga, "Plant cis-acting regulatory DNA elements (PLACE) database: 1999," *Nucleic Acids Res*, vol. 27, no. 1, pp. 297–300, 1999.
16. D. S. Prestridge, "SIGNAL SCAN: a computer program that scans DNA sequences for eukaryotic transcriptional elements," *Computer applications in the biosciences: CABIOS*, vol. 7, no. 2, pp. 203–206, 1991.
17. 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.
18. D. Alabadı, T. Oyama, M. J. Yanovsky, F. G. Harmon, P. Más, and S. A. Kay, "Reciprocal regulation between *TOC1* and *LHY/CCA1* within the *Arabidopsis* circadian clock," *Science*, vol. 293, no. 5531, pp. 880–883, 2001.
19. 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.
20. 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.

21. E. Yakir, D. Hilman, M. Hassidim, and R. M. Green, "*CIRCADIAN CLOCK ASSOCIATED 1* transcript stability and the entrainment of the circadian clock in *Arabidopsis*," *Plant Physiol*, vol. 145, no. 3, pp. 925–932, 2007.
22. J.-Y. Kim, H.-R. Song, B. L. Taylor, and I. A. Carré, "Light-regulated translation mediates gated induction of the *Arabidopsis* clock protein LHY," *EMBO J*, vol. 22, no. 4, pp. 935–944, 2003.
23. 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.
24. 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.
25. 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.
26. 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.
27. P. Más, D. Alabadí, M. J. Yanovsky, T. Oyama, and S. A. Kay, "Dual role of TOC1 in the control of circadian and photomorphogenic responses in *Arabidopsis*," *Plant Cell*, vol. 15, no. 1, pp. 223–236, 2003.
28. C. Strayer, T. Oyama, T. F. Schultz, R. Raman, D. E. Somers, P. Más, S. Panda, J. A. Kreps, and S. A. Kay, "Cloning of the *Arabidopsis* clock gene *TOC1*, an autoregulatory response regulator homolog," *Science*, vol. 289, no. 5480, pp. 768–771, 2000.
29. 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.

30. 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.
31. 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.
32. K. Onai and M. Ishiura, "*PHYTOCLOCK 1* encoding a novel GARP protein essential for the *Arabidopsis* circadian clock," *Genes Cells*, vol. 10, no. 10, pp. 963–972, 2005.
33. M. R. Doyle, S. J. Davis, R. M. Bastow, H. G. McWatters, L. Kozma-Bognár, F. Nagy, A. J. Millar, and R. M. Amasino, "The *ELF4* gene controls circadian rhythms and flowering time in *Arabidopsis thaliana*," *Nature*, vol. 419, no. 6902, pp. 74–77, 2002.
34. 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.
35. 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.
36. M. F. Covington, S. Panda, X. L. Liu, C. A. Strayer, D. R. Wagner, and S. A. Kay, "ELF3 modulates resetting of the circadian clock in *Arabidopsis*," *Plant Cell*, vol. 13, no. 6, pp. 1305–1316, 2001.
37. E. Kevei, P. Gyula, A. Hall, L. Kozma-Bognár, W.-Y. Kim, M. E. Eriksson, R. Tóth, S. Hanano, B. Fehér, M. M. Southern, R. M. Bastow, A. Viczián, V. Hibberd, S. J. Davis, D. E. Somers, F. Nagy, and A. J. Millar, "Forward genetic analysis of the circadian clock separates the multiple functions of *ZEITLUPE*," *Plant Physiol*, vol. 140, no. 3, pp. 933–945, 2006.
38. D. E. Somers, W.-Y. Kim, and R. Geng, "The F-box protein ZEITLUPE confers dosage-dependent control on the circadian clock, photomorphogenesis, and flowering time," *Plant Cell*, vol. 16, no. 3, pp. 769–782, 2004.

39. D. E. Somers, T. F. Schultz, M. Milnamow, and S. A. Kay, "*ZEITLUPE* encodes a novel clock-associated PAS protein from *Arabidopsis*," *Cell*, vol. 101, no. 3, pp. 319–329, 2000.
40. B. Farinas and P. Mas, "Functional implication of the MYB transcription factor RVE8/LCL5 in the circadian control of histone acetylation," *Plant J*, vol. 66, no. 2, pp. 318–329, 2011.
41. E. L. Martin-Tryon, J. A. Kreps, and S. L. Harmer, "*GIGANTEA* acts in blue light signaling and has biochemically separable roles in circadian clock and flowering time regulation," *Plant Physiol*, vol. 143, no. 1, pp. 473–486, 2007.
42. K. M. David, U. Armbruster, N. Tama, and J. Putterill, "*Arabidopsis* GIGANTEA protein is post-transcriptionally regulated by light and dark," *FEBS lett*, vol. 580, no. 5, pp. 1193–1197, 2006.
43. 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.
44. K. D. Edwards, O. E. Akman, K. Knox, P. J. Lumsden, A. W. Thomson, P. E. Brown, A. Pokhilko, L. Kozma-Bognar, F. Nagy, D. A. Rand, and A. J. Millar, "Quantitative analysis of regulatory flexibility under changing environmental conditions," *Mol Syst Biol*, vol. 6, no. 1, p. 424, 2010.
45. 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.
46. 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.
47. 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.
48. 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.

49. E. Herrero Serrano, *A molecular basis of ELF3 action in the Arabidopsis circadian clock*. PhD thesis, Universität zu Köln, 2011.
50. P. Y. Hsu and S. L. Harmer, "Circadian phase has profound effects on differential expression analysis," *PLoS One*, vol. 7, no. 11, p. e49853, 2012.
51. 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.
52. 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.
53. 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.
54. K. A. Kaczorowski and P. H. Quail, "*Arabidopsis* PSEUDO-RESPONSE REGULATOR 7 is a signaling intermediate in phytochrome-regulated seedling deetiolation and phasing of the circadian clock," *Plant Cell*, vol. 15, no. 11, pp. 2654–2665, 2003.
55. E. A. Kikis, R. Khanna, and P. H. Quail, "ELF4 is a phytochrome-regulated component of a negative-feedback loop involving the central oscillator components CCA1 and LHY," *Plant J*, vol. 44, no. 2, pp. 300–313, 2005.
56. W.-Y. Kim, K. A. Hicks, and D. E. Somers, "Independent roles for EARLY FLOWERING 3 and ZEITLUPE in the control of circadian timing, hypocotyl length, and flowering time," *Plant Physiol*, vol. 139, no. 3, pp. 1557–1569, 2005.
57. Y. Kim, S. Han, M. Yeom, H. Kim, J. Lim, J.-Y. Cha, W.-Y. Kim, D. E. Somers, J. Putterill, H. G. Nam, and D. Hwang, "Balanced nucleocytosolic partitioning defines a spatial network to coordinate circadian physiology in plants," *Dev Cell*, vol. 26, no. 1, pp. 73 – 85, 2013.
58. O. S. Lau, X. Huang, J.-B. Charron, J.-H. Lee, G. Li, and X. W. Deng, "Interaction of *Arabidopsis* DET1 with CCA1 and LHY in mediating transcriptional repression in the plant circadian clock," *Mol Cell*, vol. 43, no. 5, pp. 703–712, 2011.

59. G. Li, H. Siddiqui, Y. Teng, R. Lin, X.-y. Wan, J. Li, O.-S. Lau, X. Ouyang, M. Dai, J. Wan, P. F. Devlin, X. W. Deng, and H. Wang, "Coordinated transcriptional regulation underlying the circadian clock in *Arabidopsis*," *Nat Cell Biol*, vol. 13, no. 5, pp. 616–622, 2011.
60. J. Locke, A. Millar, and M. Turner, "Modelling genetic networks with noisy and varied experimental data: the circadian clock in *Arabidopsis thaliana*," *J of Theor Biol*, vol. 234, no. 3, pp. 383–393, 2005.
61. 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.
62. 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.
63. A. Matsushika, A. Imamura, T. Yamashino, and T. Mizuno, "Aberrant expression of the light-inducible and circadian-regulated *APRR9* gene belonging to the circadian-associated APRR1/TOC1 quintet results in the phenotype of early flowering in *Arabidopsis thaliana*," *Plant Cell Physiol*, vol. 43, no. 8, pp. 833–843, 2002.
64. N. Nakamichi, A. Matsushika, T. Yamashino, and T. Mizuno, "Cell autonomous circadian waves of the APRR1/TOC1 quintet in an established cell line of *Arabidopsis thaliana*," *Plant Cell Physiol*, vol. 44, no. 3, pp. 360–365, 2003.
65. N. Nakamichi, M. Kita, S. Ito, E. Sato, T. Yamashino, and T. Mizuno, "The *Arabidopsis* pseudo-response regulators, PRR5 and PRR7, coordinately play essential roles for circadian clock function," *Plant Cell Physiol*, vol. 46, no. 4, pp. 609–619, 2005.
66. 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.
67. 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.
68. N. Nakamichi, T. Kiba, M. Kamioka, T. Suzuki, T. Yamashino, T. Higashiyama, H. Sakakibara, and T. Mizuno, "Transcriptional

repressor PRR5 directly regulates clock-output pathways," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 109, no. 42, pp. 17123–17128, 2012.

69. Y. Niwa, S. Ito, N. Nakamichi, T. Mizoguchi, K. Niinuma, T. Yamashino, and T. Mizuno, "Genetic linkages of the circadian clock-associated genes, *TOC1*, *CCA1* and *LHY*, in the photoperiodic control of flowering time in *Arabidopsis thaliana*," *Plant Cell Physiol*, vol. 48, no. 7, pp. 925–937, 2007.
70. 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.
71. D. H. Park, D. E. Somers, Y. S. Kim, Y. H. Choy, H. K. Lim, M. S. Soh, H. J. Kim, S. A. Kay, and H. G. Nam, "Control of circadian rhythms and photoperiodic flowering by the *Arabidopsis GIGANTEA* gene," *Science*, vol. 285, no. 5433, pp. 1579–1582, 1999.
72. 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.
73. 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.
74. E. Sato, N. Nakamichi, T. Yamashino, and T. Mizuno, "Aberrant expression of the *Arabidopsis* circadian-regulated *APRR5* gene belonging to the APRR1/TOC1 quintet results in early flowering and hypersensitiveness to light in early photomorphogenesis," *Plant Cell Physiol*, vol. 43, no. 11, pp. 1374–1385, 2002.
75. H.-R. Song and I. A. Carré, "DET1 regulates the proteasomal degradation of LHY, a component of the *Arabidopsis* circadian clock," *Plant Mol Biol*, vol. 57, no. 5, pp. 761–771, 2005.
76. H.-R. Song, "Interaction between the late elongated hypocotyl (*LHY*) and early flowering 3 (*ELF3*) genes in the *Arabidopsis* circadian clock," *Genes Genomics*, vol. 34, no. 3, pp. 329–337, 2012.
77. T. Yamashino, S. Ito, Y. Niwa, A. Kunihiro, N. Nakamichi, and T. Mizuno, "Involvement of *Arabidopsis* clock-associated pseudo-response regulators in diurnal oscillations of gene expression in the presence of environmental time cues," *Plant Cell Physiol*, vol. 49,

no. 12, pp. 1839–1850, 2008.

PAPER IV

IV

An efficient crossover algorithm by global alignment for evolution of variable length genomes

Karl Fogelmark, Adriaan Merlevede, Carl Troein and Henrik Åhl

Computational Biology and Biological Physics, Department of Astronomy and Theoretical Physics, Lund University, 223 62 Lund, Sweden

Manuscript, LU-TP 16-11 (2016)

In silico evolution has applications in computer science and evolutionary biology. Although most implementations use genomes of constant length, variable-length genomes are a natural choice when modelling evolutionary mechanisms such as copy number and structural variations, or traverse search spaces of variable or unknown dimensionality. However, such genomes are costly to manipulate and interpret, especially for performing crossover.

Here, we compare different crossover methods for variable-length linear genomes. Qualities used for comparison are the ability of the crossover to retain homologous features in the parental genomes, CPU time consumption and performance in a toy evolutionary model.

We find that existing methods are not fully optimized, neither in terms of quality of the offspring nor computation time. Crossover of variable-length genomes is computationally expensive, but can accelerate evolution when other steps such as fitness evaluation are also expensive, which is often the case. We show that simple heuristics can improve the overall performance compared to earlier methods, and outline directions for further improvements.

IV.1 INTRODUCTION

Evolutionary experiments are costly and time-consuming. One alternative way to perform experiments in evolutionary biology is a computational approach, where instead of biochemical organisms, instantiations of a virtual model are subjected to iterative reproduction, mutation and selection. This simulated evolution allows us to study the general process of evolution, of which life on earth is a special case.

When generalizing results obtained from *in silico* experiments to biological evolution, or to evolution in general, one has to be mindful of the limitations and general properties of the model. For example, many existing models use genome representations with high information density. While computationally efficient and easily interpretable, it does not leave room for the exploration of neutral networks, inhibiting neutral evolution and thus changing the evolutionary dynamics of the system. In addition, in order to investigate the evolution of genome structure, the model must be rich enough to accommodate insertions and deletions as well as concepts such as synteny, modularity, sequence motifs, and copy numbers. Linear genomes of variable length do not impose fixed information densities or structures, and are a natural setting to model genome structure dynamics close to how it is understood in biology. They can also be of use in computer science, to traverse search spaces with variable or unknown dimensionality [1, 2]. Despite their potential, few past experiments have featured genomes of variable length, possibly due to their computationally expensive reading and manipulation, especially when performing crossover.

Herein, we compare existing and new methods for crossover of variable-length linear genomes consisting of binary digits, both in an artificial setting of randomly generated genomes with a given similarity, and in an experimental setting during a toy evolution. The comparison is based on three properties: the ability of the crossover to match real homologous sequences, CPU time consumption, and the success of the algorithm to produce sensible high-fitness offspring (i.e. the number of generations needed to reach a certain fitness level).

We hope that our method, and general approach, can be used in future research seeking to unravel the mechanisms governing the evolution of genomic structure and other evolutionary concepts which require variable-length genome representations. In addition, we hope that it can be of use in novel approaches to evolutionary computation.

IV.2 METHODS

IV.2.1 *Crossover algorithms*

Several crossover methods exist for digital genomes of constant size. In evolutionary algorithms, popular choices include one-point, two-point, multiple-point and uniform crossover for genomes [3]. Each algorithm creates two new complementary sequences that together contain all the sequence information from their parents. Usually, only one of the two siblings is retained for selection. Similar strategies can be used for genomes of variable length, but the added difficulty for variable-length genomes is to know where to align the crossover points. The crossover should construct offspring with a high expected fitness by recombining genomic structures unique to each parent, while keeping the homologous information present in both.

One solution is to align the two parental genomes using a sequence alignment algorithm, prior to choosing the crossover points. Aligned locations (not including gaps) can then serve as possible sites for crossover points in the same way as in sequences of constant length. Because the appropriate number of crossover points should increase with genome length, it is natural to give each aligned bit the same probability of acting as a crossover point. Many alignment methods exist (for review, see [4]). Because it is simple and theoretically well-founded, we chose to use the Hirschberg algorithm with an affine gap penalty as described by Myers and Miller [5, 6]. This is an adaptation of the traditional Needleman-Wunsch that lowers the memory complexity from $\mathcal{O}(N^2)$ to $\mathcal{O}(N)$ while retaining the $\mathcal{O}(N^2)$ time complexity. Our implementation uses a binary alphabet, where alignments are scored $+1$ for a match, -5 for a mismatch, and an affine gap penalty of -20 to open and -3 to extend. These numbers, though largely arbitrary, were selected to produce results in agreement with a manual alignment.

To lower the time consumption of the alignment process for crossover, we implemented a simple heuristic method for quickly breaking down the alignment problem, under the assumption that the parental genomes are usually highly similar. This recursive heuristic algorithm extracts three substrings of length 64 bits, centred at one, two and three quarters of one of the genomes. For each of these three subsequences, the other genome is scanned for matching regions with a Hamming distance ≤ 20. If at least two of the three yield a well-defined best match (lowest Hamming distance), and if the matches occur in the correct order without overlap, the two genomes are cut in the corresponding positions

and the pieces are aligned by recursion. If these conditions are not met, or if the genomes are shorter than 256 bits, the method falls back on the Hirschberg algorithm.

An alternative method was proposed by Hutt and Warwick, inspired by the chi structure of chromosomes during meiosis in biology [2]. Their synapsing method finds and aligns the longest common subsequence of the two parental genomes (a synapse). This is recursively repeated on the left and right sides of the synapse, until no longest common subsequence above a specified threshold length can be found. This results in an alternating pattern of synapses, where both parental genomes are identical, and unaligned regions, where they are not. The bits contained in synapses are used as possible crossover points in a multi-point crossover, exchanging the unaligned regions in between. In our implementation, the minimum synapse length was set to 3 bits. For both the synapsing crossover and the alignment-based methods, each aligned bit had a probability of 0.02 to serve as a crossover point.

IV.2.2 *Benchmark evolution*

Ultimately, what defines a good crossover is its ability to produce high-fitness offspring and accelerate evolution. While the potential of any crossover to do so is strongly dependent on the genome structure, we chose to compare the evolutionary dynamics resulting from different crossover methods in a toy evolution model that we think is representative of many interesting and useful situations. Similar models have been used in literature (e.g., [7, 8]). The genomes and their relation to the phenotype and evolution target are structured enough to allow crossover to combine useful building blocks of both parents' genomes, and complex enough to present a complex dynamical genomic structure [7].

In the model, we choose to represent each individual's phenotype as a function $f : [0, 1] \to \mathbb{R}$, described by a sum of triangular basis functions. Each gene in the genome codes for an isosceles triangle of height $h \in [-1, 1]$ and base $b \in (0, 1]$, which rests on the x axis centred at a point $s \in [0, 1]$. A gene is defined by a start sequence (110011), followed by three sets of ten bits each for the parameters h, b and s. These are encoded as binary integers which are then rescaled to the relevant ranges. A gene can be recognized anywhere in the genome, except inside another gene. The phenotype is calculated by summing the triangles represented by all genes. The fitness function, F, is defined by

the distance of the phenotype to the target function $g(x) = \sin(6 \cdot 2\pi x)$, measured as the L^2 norm, $F = -\int_0^1 (f(x) - g(x))^2 dx$.

Selection is effected by a tournament method. In each generation, two random individuals are chosen, and the one with the lowest fitness is removed from the population. It is replaced by a new individual generated either through crossover (with probability p_X) or mutation (with probability $1 - p_X$). In either case, parents are picked by taking the individual with the highest fitness from a random sample of two. For crossover, the two parents must be distinct.

In the case of mutation, one of three operators is executed: substitution (probability 0.8), deletion (probability 0.1), or insertion (probability 0.1). During substitution, each bit in the genome is flipped with probability 0.001. During deletion or insertion, a random sequence section of random length l, drawn from a power law distribution proportional to l^{-2} (not longer than the length of the genome), is removed from the sequence or is repeated in a random genomic location. A similar power law distribution for the size of insertions and deletions has been observed in nature [9].

IV.3 RESULTS

As an initial comparison, we view each crossover method as a sequence alignment. Figure IV.1A-B shows the alignment score thus obtained. By design, the Hirschberg method gives the globally optimal value. Our heuristic algorithm also performs optimally, unless the two parents are highly divergent; thus the heuristic is highly similar to the Hirschberg method in most cases. However, the alignment score is not a good measure of success for crossover. Aside from being arbitrary, it is a poor proxy for the evolutionary history of divergent sequences, and it does not measure the properties of the resulting offspring. Specifically, the synapsing method is not optimized as an alignment algorithm and produces an excessive number of small gaps, resulting in low alignment scores.

In order to better compare the ability of the crossovers to propagate genetic information shared by the two parents, we measured the fraction of homologous bits that were consistently inherited after crossover. More precisely, for each method, we performed a large set of crossovers, each resulting in a pair of complementary offspring. In each case, one of the parents was a copy of the other, mutated to some degree.

With the exception of bits involved in insertion or deletion, each bit in either parental genome can be matched uniquely to an unchanged or substituted bit in the other, resulting in a set of homologous bit pairs. Here, homology is used in the biological sense, i.e. sequence features that are shared because they descend from the same ancestral original.[1] Figure IV.1C-D shows the fraction of such homologous pairs that were divided evenly among the two complementary offspring. That is, both homologous offspring should have exactly one of the two homologous bits.

All methods preserve a large fraction of the existing homology when the parental sequences are similar. In general, the Hirschberg method outperforms the other two methods in this regard. The heuristic method performs similarly unless the parents are highly dissimilar; in particular, insertions of duplicated genetic material will increase the risk of misidentifying the cutting points for the heuristic algorithm. For dissimilar sequences, the synapsing method preserves much less homology than the alignment-based methods. It should be noted that 10% sequence divergence is high and unlikely to occur often during most evolutionary experiments; in nature, organisms with such dissimilarity are unlikely to produce fertile offspring.

To assess the performance of these methods in practical computation, we compared their CPU usage. Figure IV.2 shows CPU cost as a function of genome length and sequence dissimilarity. The heuristic and synapsing methods are much faster than Hirschberg, which has quadratic complexity. The computational cost of the Hirschberg and synapsing alignment methods do not depend strongly on the similarity between the parental sequences. In contrast, the heuristic method is fast for similar genomes, but breaks down when the parents are highly dissimilar, as it increasingly falls back on the Hirschberg method. For most applications, sequence divergence is usually low and the heuristic method is approximately twice as fast as the synapsing method.

Finally, we compared the ability of the crossovers to perform successful sexual reproduction in a model evolutionary experiment. Figure IV.3A shows the progression of the increasing fitness of the population over time. The results of the heuristic alignment and synapsing methods are similar. Note that, in this experiment, evolution is faster with crossover than without. Figure IV.3B gives the speed of evolution for different

1 In our system, insertions can also result in paralogy, i.e. sequences that share a common history through duplication inside the same genome. This kind of homology is not considered here.

crossover probabilities. Again, the two methods perform similarly, with the convergence time minimized around $p_X = 0.3$. However, faster convergence in number of generations is offset by the computational cost of the crossover (Figure IV.3C). In our simple benchmark evolution, most calculations are trivial and crossover is the most time-consuming step. In contrast, most applications have complex fitness functions that are often much more costly to compute. In that case, faster convergence speed in number of generation means fewer fitness evaluations and lower computational cost overall. This is illustrated in Figure IV.3D, where the heuristic alignment method is the fastest by a narrow margin.

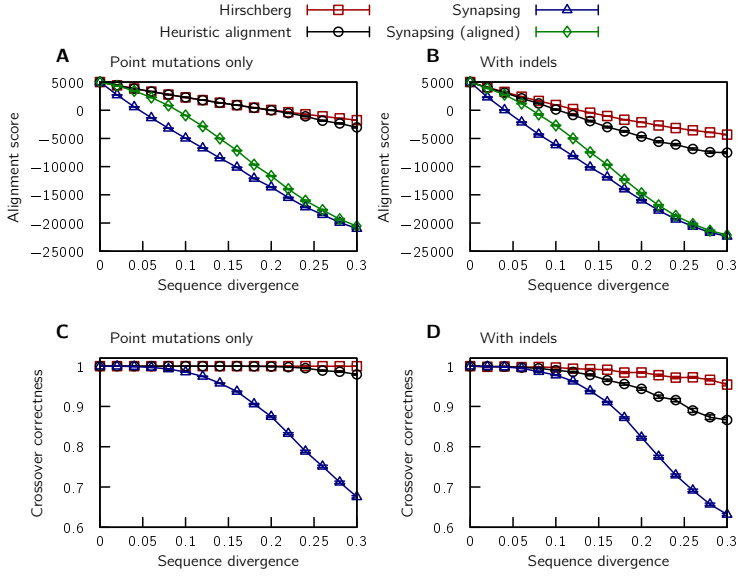

Figure IV.1 Comparison of alignments and crossovers. (A-B) The quality of alignment between a random genome of length 5000 bits and a mutated version thereof, for global alignment with the Hirschberg method (red squares) and the heuristic described in the text (black circles). For comparison, the synapsing method is also included, treating unsynapsed regions as gaps (blue triangles), or as aligned (with mismatches) when they are the same length and shorter than 21 bits (green diamonds). (C-D) The fraction of homologous pairs of bits in the parent genomes that are present in both offspring genomes, for various levels of sequence dissimilarity between the parents. The methods were examined using only point mutations (A, C) or a 40:1:1 mix of point mutations, insertions and deletions (B, D). Sequence divergence here measures the fraction of bits affected by mutation. Data from 400 genomes and mutated partners, each crossed over 100 times (in C-D). Error bars indicate the standard error of the mean.

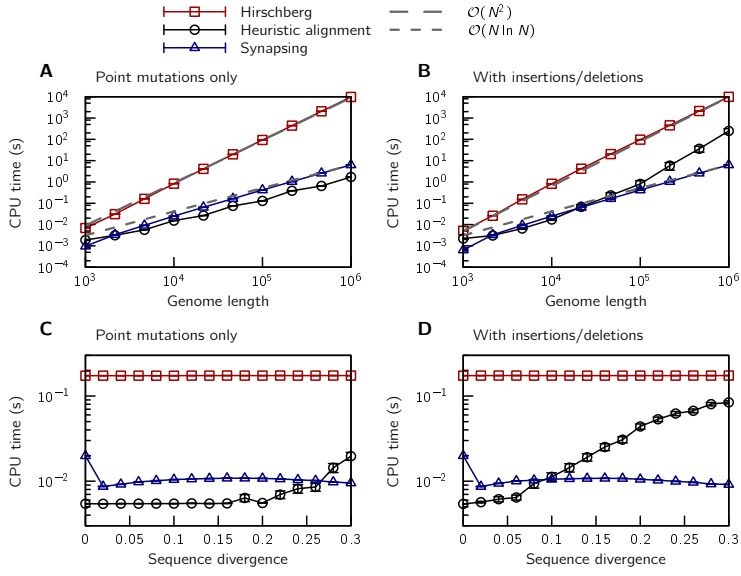

Figure IV.2 Run time of genome alignment. (A-B) The time required to align two genomes of given length for crossover using the Hirschberg method (red squares), the heuristic alignment method (black circles) and the synapsing crossover method (blue triangles) on an Intel Core i7 processor. The two genomes were separated by point mutations affecting 5% of the bits. $\mathcal{O}(N^2)$ and $\mathcal{O}(N \ln N)$ scaling is indicated by short and long dashed gray lines, respectively. (C-D) The relationship between sequence similarity and the required CPU time for the different crossover methods. The sequence divergence between the two parental genomes is defined as the fraction of bits affected by mutation. The mutations used were only point mutations (A, C) or a 40:1:1 mixture of point mutations, insertions and deletions (B, D). Error bars indicate the standard error of the mean.

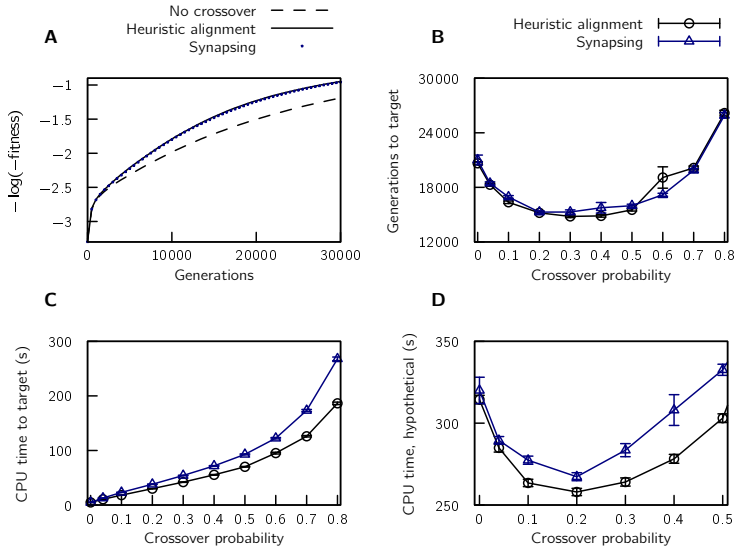

Figure IV.3 The effect of crossover in optimization. (A) Fitness of the best individual in each generation, using the heuristic alignment crossover method (solid line) or synapsing crossover (blue points) with crossover probability $p_X = 0.3$, compared with evolution without crossovers (dashed line). (B) The number of generations needed to reach a high fitness ($F = -30$), as a function of the crossover probability, p_X, for the heuristic alignment method (black circles) and synapsing crossover (blue triangles). The Hirschberg method was excluded due to its computational cost. For this specific system, the optimal crossover rate for optimization is around $p_X = 0.3$. (C) The amount of CPU time needed to reach fitness $F = -30$ on an Intel Core i7 processor. Despite lowering the number of generations needed, crossover increases the total required CPU time. (D) The same as (C), for a hypothetical experiment where the computational cost of fitness evaluation is drastically increased from 0.1 to 15 msec per evaluation. In this case, the optimal crossover rate is a compromise between the cost of crossovers and the decreased number of generations needed to reach the target fitness.

IV.4 DISCUSSION

From these results, we conclude that alignment-based methods are a good basis for creating crossover algorithms. The Hirschberg alignment produces high-quality offspring that retain the homologous features present in both parents, while also recombining unique features.

Synapsing is outperformed in several ways by the alignment-based methods. The results from Figure IV.1 suggest that synapsing is much less likely to retain the original properties of the parental genomes. In our benchmark evolution experiment, synapsing is also slower than the heuristic method. One potential problem with synapsing is that it only considers local comparisons between the two parental genomes, not the greater context. As the number of mutations separating the two genomes grows, the longest common subsequence is shortened and there is an increased risk of synapsing two unrelated parts. However, different crossover can also confer different qualitative behaviour on the evolutionary dynamics, other than simply influencing the speed, which may be useful or interesting in some contexts. For example, the inability of the synapsing crossover to retain shared homologous sequences from both parents when they are dissimilar may result in a spontaneous similarity selection, reproductive isolation, or genomic restructuring events.

Both approaches leave a lot of room for improvement. The quick and dirty heuristic presented here has a significantly lower computation time than the Hirschberg method, at low cost in performance. It is likely that other heuristics can further improve on alignment-based crossover. For the synapsing method, we expect that it can be improved by using a less rigid and faster local alignment algorithm to find suitable synapses, rather than the longest common subsequence.

In the future, we believe that research in new types of evolutionary dynamics and new genome representations, together with the continued increase in computation power, will make more complex *in silico* evolution possible. Such experiments may use our findings to select or develop a suitable crossover algorithm.

ACKNOWLEDGMENT

A.M., C.T. and K.F. were supported by grant 621-2010-5219 from the Swedish Research Council, `http://vr.se/`.

REFERENCES

1. C. Y. Lee and E. K. Antonsson, "Variable length genomes for evolutionary algorithms.," in *GECCO*, vol. 2000, p. 806, 2000.
2. 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.
3. X. Yu and M. Gen, *Introduction to Evolutionary Algorithms*. Decision Engineering, London: Springer London, 2010.
4. W. Haque, A. Aravind, and B. Reddy, "Pairwise sequence alignment algorithms: A survey," in *Proceedings of the 2009 Conference on Information Science, Technology and Applications*, ISTA '09, (New York, NY, USA), pp. 96–103, ACM, 2009.
5. D. S. Hirschberg, "A linear space algorithm for computing maximal common subsequences," *Commun. ACM*, vol. 18, pp. 341–343, June 1975.
6. E. W. Myers and W. Miller, "Optimal alignments in linear space," *Computer applications in the biosciences: CABIOS*, vol. 4, no. 1, pp. 11–17, 1988.
7. B. Batut, D. P. Parsons, S. Fischer, G. Beslon, and C. Knibbe, "In silico experimental evolution: a tool to test evolutionary scenarios," *BMC Bioinformatics*, vol. 14, no. 15, pp. 1–11, 2013.
8. G. Beslon, D. P. Parsons, J.-M. Pena, C. Rigotti, and Y. Sanchez-Dehesa, "From digital genetics to knowledge discovery: Perspectives in genetic network understanding," *Intelligent Data Analysis*, vol. 14, no. 2, pp. 173–191, 2010.
9. R. A. Cartwright, "Problems and solutions for estimating indel rates and length distributions," *Molecular Biology and Evolution*, vol. 26, no. 2, pp. 473–480, 2009.

Index

All this he saw, for one moment breathless and intense, vivid on the morning sky; and still, as he looked, he lived; and still, as he lived, he wondered.

Kenneth Grahame, *The Wind in the Willows* (1908)