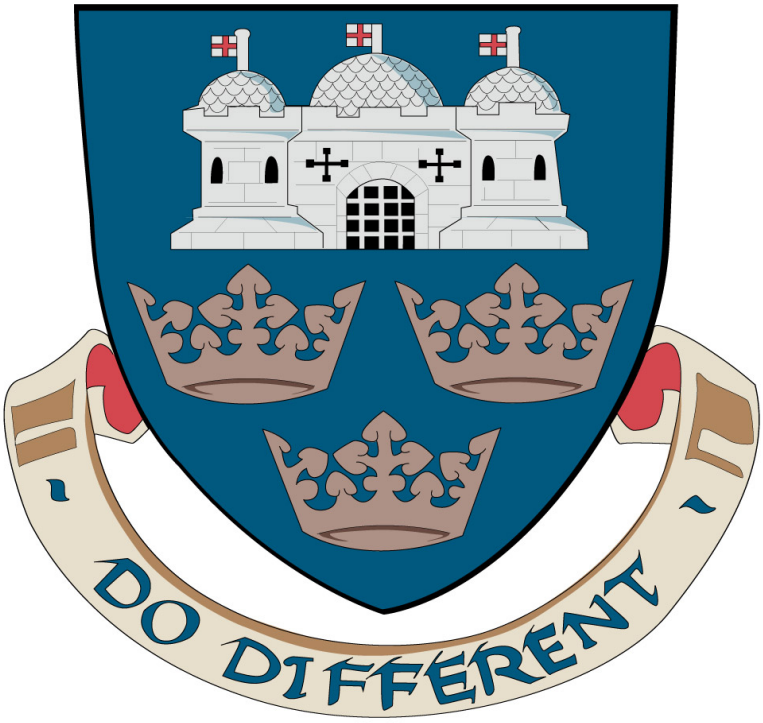
Mathematical modelling of the floral transition

— with a Bayesian flourish —

Nick Pullen

September, 2014



A thesis submitted to the University of East Anglia for the degree of Doctor of Philosophy

John Innes Centre

ABSTRACT

Flowering plants are abundant on Earth. In the model dicot plant species, *Arabidopsis thaliana*, multiple endogenous and exogenous signals converge to initiate a change from vegetative to reproductive growth in optimal environmental conditions. Much genetic and experimental research has gone into elucidating the biological mechanisms controlling the floral transition. However there has been little mathematical modelling of this process.

The aim of this thesis was to gain an understanding of the essential features and dynamic properties underlying this developmental phase change from a systems and computational biology perspective. Combining mathematical modelling with experimental results a core regulatory network was defined with multiple feedback loops. Simplified models inevitably miss finer details of the biological system yet they provide a route to understanding the overall system behaviour. This reductionist path allowed a tractable genetic regulatory network to be investigated without large numbers of parameters.

Not overfitting to data and parameter inference are two current challenges in systems biology. Treating all unknowns as a probability within the setting of Bayes' theorem as a statistical framework allows for a solution to both of these issues. This thesis investigates the use of a contemporary Bayesian inference algorithm, nested sampling, for inference problems typically found in systems biology where the data are few and noisy. Nested sampling simultaneously calculates the key term for model comparison and also produces parameter inferences allowing uncertainty in models and predictions to be robustly quantified.

Network models are developed that can accurately reproduce experimental leaf number data, show important properties of the floral transition such as the ability to filter environmental noise and provide a clue on spatial patterning of an Arabidopsis shoot apex. Incorporating network knowledge into a plant breeding program is an exciting goal for future developments addressing global food security.

CONTENTS

LIST OF PUBLICATIONS

This thesis includes material from the published works below.
* Indicates that these authors contributed equally.

A. Bentley, E. Jensen, I. Mackay, H. Hönicka, M. Fladung, K. Hori, M. Yano, J. Mullet, I. Armstead, C. Hayes, D. Thorogood, A. Lovatt, R. Morris, N. Pullen, E. Mutasa-Göttgens, and J. Cockram. **Flowering Time**. In: *Genomics and Breeding for Climate-Resilient Crops: Vol. 2 Target Traits* (2013), 1–66.

K. E. Jaeger*, N. Pullen*, S. Lamzin, R. J. Morris, and P. A. Wigge. **Interlocking feedback loops govern the dynamic behavior of the floral transition in *Arabidopsis***. *The Plant Cell* 25.3 (2013), 820–833

N. Pullen, K. E. Jaeger, P. A. Wigge, and R. J. Morris. **Simple network motifs can capture key characteristics of the floral transition in *Arabidopsis***. *Plant Signaling & Behavior* 8.11 (2013), e26149.

N. Pullen and R. J. Morris. **Bayesian model comparison and parameter inference in systems biology using nested sampling**. *PLoS ONE* 9.2 (2014), e88419.

ACKNOWLEDGEMENTS

I would like to thank my main supervisor Richard Morris for all the support he has given me throughout these long, and sometimes trying, four years. I appreciate his giving me the opportunity to study for a PhD in his group and for allowing me to attempt to combine my mathematical knowledge with my interest in plants. He has been full of useful suggestions, especially when things did not turn out as expected in this project. His good humour and scientific integrity have been sterling examples for me to follow in the future. I also thank Richard a lot for comments on draft chapters of this thesis which have improved it considerably.

I would like to thank all the post-docs and students in the Computational & Systems Biology department who have been a great group to spend my time with. In particular I'd like to acknowledge the help of David Richards in the early days for getting me up and running with learning Linux and Python. He also orchestrated much lunchtime entertainment which I think provide many of us with fond memories. I owe a lot of thanks to Matthew Hartley who joined CSB as our lab manager and was able to give me lots of hints on computing and programming.

Other members of my supervisory team deserve mentions. Phil Wigge and Katja Jaeger provided the experimental angle of the project and I thank them for the leaf number data, and helping choose me to do a PhD at JIC. Rico Coen also provided food for thought throughout the project and allowed us the benefit of his wisdom.

I am indebted to the Brassica tendency crew of Marc Jones, Rachel Wells, Judith Irwin, Martin Trick and Colin Morgan for making the final year of my PhD the most enjoyable one. We've had many fun meetings, social occasions and they have allowed me to finally enter the greenhouses to see growing plants close up. I also thank Judith for some useful comments on chapter 3.

It was also good to meet lots of people on the various training courses or conferences I attended all over Britain and Europe. Their enthusiasm showed me a good view of science and the wider world. I should also acknowledge lots of online bloggers and communities

whose various tips saved me (or sometimes caused me!) much frustration at the keyboard. Principally those who have, and continue to contribute to `StackOverflow.com` or `tex.stackexchange.com` have been incredibly valuable to my development as someone who can now program a fair bit.

In Norwich, I thank Glen and Annika Richardson for letting me lodge in their lovely house for the first nine months of my PhD. Thereafter I thank Steve Garrett for being my landlord in a friendly house very close to work. To the members of the UEA Athletics and Hockey Clubs I was a member of during my PhD years I am indebted to helping me keep it real, and reminding me life isn't all about about long hours in front of the office computer. It was an honour to serve as UEAHC Club Captain and to help coach the distance athletes from UEAAC.

I feel very lucky to have met Yan Ma who has shown much kindness and support to me. As a dedicated and hard-working PhD student she has been a great example of carrying on when times get tough. Her knowledge and enthusiasm have helped me to gain a far greater understanding of many aspects of biological experiments. She also provided the Arabidopsis plants I took photos of in chapter 1 and helped me with the photo editing.

Lastly, I thank my family for always being there. They have always encouraged me and seem much more convinced of my capability than I am. I wish them all the happiness for the future.

Nick Pullen
John Innes Centre, Norwich
September 2014

1 INTRODUCTION & BACKGROUND

1.1 Outline of the thesis

This thesis opens by providing a very short summary of the molecular biology the non-biologist reader may need to know to understand future sections of this thesis and also introduces the model plant that informs our study of the floral transition. More detail on the genes known to be involved in this process, as well as upstream, from the first perceived signals, and downstream, to floral organ specification, are provided which gives a sense of the scale of the network. After the literature reporting experimental biological studies is covered, published mathematical models of flowering time in model species and crop species are reviewed. Thereafter the introduction will cover the basics of Bayesian inference, and why this statistical framework is used as opposed to optimisation and maximum likelihood approaches. This will be needed for the following chapter on nested sampling which is quite a new technique for a proper Bayesian treatment of an inference problem. It allows one to calculate the key quantity for model comparison and perform parameter inference for mathematical models. We are amongst the first to apply nested sampling to the field of systems biology. Following initial testing and tuning of the algorithm its output will be measured against that of the current workhorse of Bayesian inference. System dynamics will be recovered and models of biological oscillators compared. Our own model for the floral transition is developed in the next chapter. A reductionist approach will be taken to help us understand key features of the network by boiling the multi-gene network from the literature review down to a few key hubs. A simple linear model of these hubs will be compared with an ordinary differential equation (ODE)-based model. This ODE model has at its core well-studied network motifs for enabling a system to reduce noise levels and confer irreversibility. Using nested sampling all the developed models are compared in a robust fashion and how accurately they predict experimental data is studied. That chapter is concluded with a de-

tailed discussion of the strengths and limitations of the model. Lastly the thesis ends with an overall discussion of the work presented, the themes considered and gives an outlook on possible future developments.

1.2 Basic biology and *Arabidopsis thaliana*

Plants are eukaryotes which means their cells have a nucleus[1]. Deoxyribonucleic acid (DNA) is found in the nucleus and is comprised of four bases: adenine (A), cytosine (C), guanine (G) and thymine (T) and a sugar-phosphate backbone [1]. Hydrogen bonds between the complementary base pairs A:T and C:G give DNA its famous double helix structure. The process of transcription takes place in the nucleus. Double stranded DNA is opened by enzymes and pre-messenger ribonucleic acid (pre-mRNA) is transcribed. This contains exons and introns, regions that do or do not code for a protein respectively. The introns are spliced out and exons joined together so that mature mRNA contains only coding regions. Nuclear export follows from the cell's nucleus to the cytoplasm which is the location of the process of translation. This means the mRNA is translated from its coding sequence and a protein is formed, facilitated by ribosomes. Certain proteins called transcription factors can bind to the promoter sequence of a gene to activate or inhibit the transcription of that specific gene. Further control processes such as post-transcriptional modifications or micro-RNAs also affect the level of a gene's expression. Gene regulation is a highly complex and intricate process only briefly touched on here. Understanding the interactions between genes, and the proteins they code for, is a major aim for scientists across the world. Fortunately for the systems biologist tackling problems in plant biology there is a model organism which over two decades of detailed genetic studies have revealed many components of its genetic regulatory networks (GRNs).

[1] *Eukaryote derives from the Greek meaning true kernel.*

Arabidopsis thaliana (L.) Heynh.[2] is a model plant species in the Brassica family that was the first plant to have its sequenced genome published [2]. A common weed, its relatively small diploid genome, short life cycle and small physical size provide a good testbed for understanding many biological processes. Post-germination, the life cycle of an Arabidopsis plant can be simply described as a vegeta-

[2] *Hereafter* Arabidopsis *is used as the common name.*

Figure 1.1: The phenotype of short-day grown Arabidopsis. Top) Rosette and early flower bolt. Lower left) The main stem of a bolting Arabidopsis. Cauline leaves are visible. Lower right) A branching Arabidopsis. Siliques are visible on the main stem.

Figure 1.2: An Arabidopsis inflorescence. Note siliques forming from the oldest flower and young buds still developing. This is because wildtype Arabidopsis are indeterminate—they will keep producing new growth from the shoot apex.

tive phase, where leaves are formed in a rosette on the ground, followed by a transition to reproductive development where flowers are formed. Morphologically around the time of this transition the plant bolts. This means it develops vertical stems attached to which are cauline leaves, and when the transition is complete the apical primordia that would otherwise have become leaves become flowers. After pollination and fertilisation the seed will set in pods (siliques) before, as an annual, the plant will die. The floral transition (as it's known) is therefore important for correct timing of flower and seed production to enable the parent plant's progeny to germinate and develop. Understanding the floral transition is an active area of research globally. Excellent genetic studies have revealed many genes involved in this crucial developmental phase but there have been few attempts to give a mathematical understanding to the transition — the focus of this thesis. Next we give an overview of the experimental literature for

multiple pathways that converge to effect flowering time, and genes that affect the floral transition and the development of floral organs.

1.3 The genetics of flowering time

The floral transition in Arabidopsis has been a well-studied developmental progression over the past 25 years. Many key genes and signalling pathways have been elucidated from experimental studies. Microarrays, which allow for the analysis of thousands of genes genome-wide, have revealed that hundreds of transcripts are specifically affected in their expression in the apex upon floral induction [3, 4].

There are thought to be at least six pathways [5, 6] (see Figure 1.3) that stimulate flowering in Arabidopsis: ageing, photoperiod, vernalisation, ambient temperature, autonomous and gibberellin (GA). Some of the most important elements in these pathways have been revealed to be genes that are integrators for multiple pathway signals. These floral pathway integrators activate floral meristem identity genes to facilitate meristem changes that lead to the activation of organ identity genes which control flower development.

Vernalisation is the process whereby prolonged exposure to winter cold increases the competence of a number of species to flower in the spring. This has resulted in a number of agricultural crops like wheat and beans being bred to establish in the autumn before the next summer's harvest, in contrast to lines that can be planted in the spring and harvested just a few months later. Vernalisation is also an important pathway in certain accessions of Arabidopsis. Repression of the MADS-box transcription factor *FLOWERING LOCUS C* (*FLC*) has been established as the main target of the vernalisation pathway in Arabidopsis [7] but there are also *FLC*-independent mechanisms of vernalisation [8]. In particular, the related genes *AGAMOUS-LIKE 19* (*AGL19*) [9] and *AGL24* [10] promote flowering which contrasts with *FLC* which acts as a major repressor of floral initiation in some accessions. FLC represses some of the photoperiod pathway genes such as *FLOWERING LOCUS T* (*FT*), *SUPPRESSOR OF OVEREXPRESSION OF CONSTANS1* (*SOC1*) and *FD* before vernalisation by direct binding [11]. Profound investigations into complex molecular biology and epigenetic silencing after plants

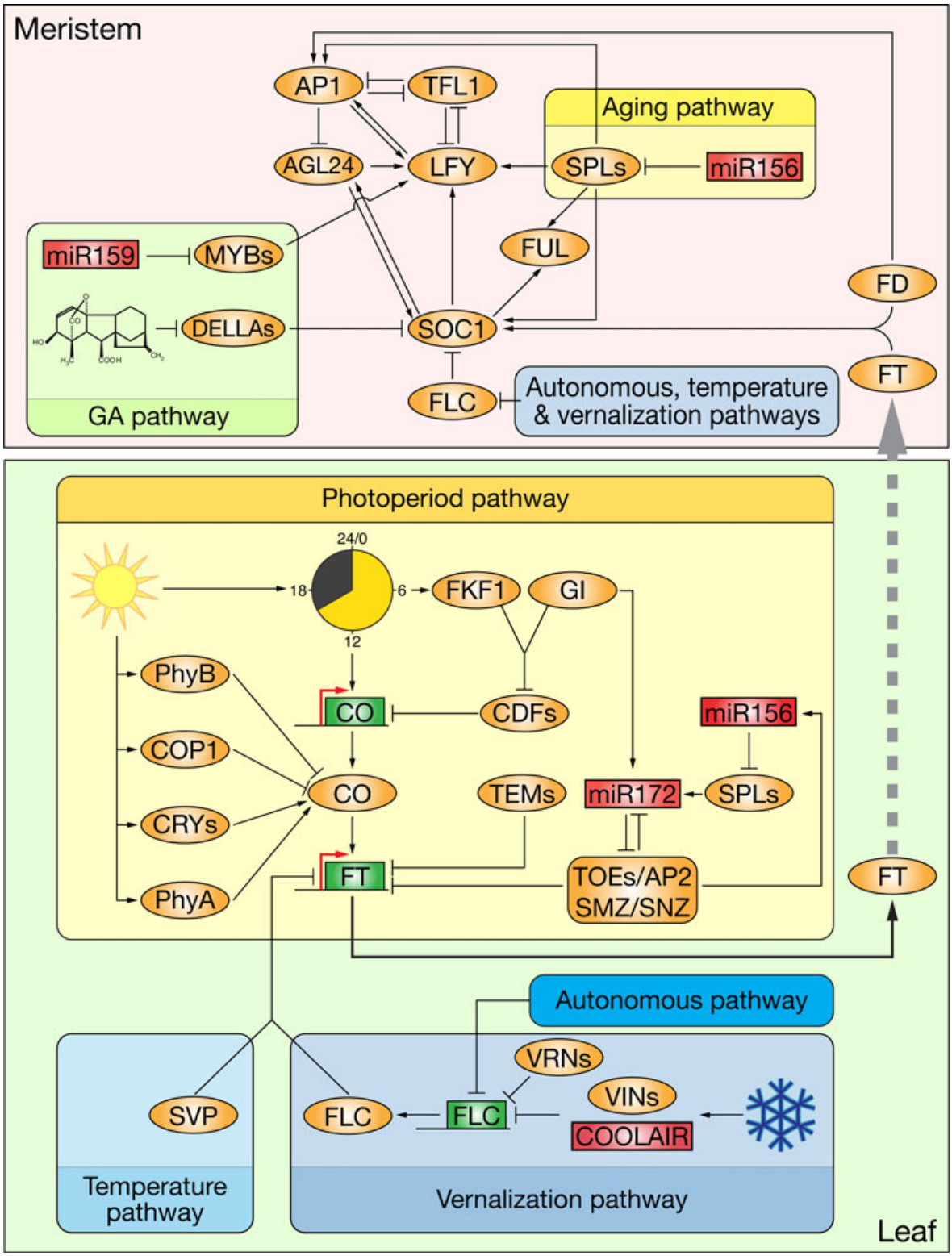


Figure 1.3: The flowering time network of Arabidopsis. Multiple environmental and endogenous signals, some travelling over a long distance, converge at the apex. At the shoot apical meristem a complex network integrates these signals to robustly initiate flowering time. Figure taken from Srikanth & Schmid [6].

return to the warm have revealed how *FLC* is stably repressed after enough cold exposure at the cell level [12]. Li et al. [13] showed that another repressor, *SHORT VEGETATIVE PHASE* (*SVP*), interacts with FLC to delay flowering by associating to the *FT* and *SOC1* promoter regions. *SVP* is in addition regulated by the ambient temperature, autonomous and gibberellin pathways [13, 14] and is but one example of the degree of overlap between the autonomous and temperature pathways.

Autonomous pathway proteins such as FCA, FPA, FVE and LUMINIDEPENDENS, promote flowering independently of photoperiod but the late flowering mutants respond to vernalisation if *FLC* is present [8]. This suggests they act upstream of *FLC* and promote flowering by inhibiting *FLC* expression [8]. Like many genes, *FCA* and *FVE* have a dual role in flowering time control. They are redundant in the autonomous pathway but act together in a temperature-dependent pathway [15]. SOC1 expression is also affected by the autonomous pathway [16, 17].

In the era of global warming the effect of temperature on flowering is important to understand for breeding heat resilience into crops. The ambient temperature pathway [15, 18] is thus of growing interest. The surrounding temperature in a plant's environment can induce flowering at warmer temperatures under otherwise non-inductive short-day light conditions [18]. A small shift from 23 °C to 27 °C was enough to reduce the time to flower in many accessions and mutant lines [18]. Kumar et al. [19] showed that increasing temperature causes PHYTOCHROME INTERACTING FACTOR4 (PIF4), a transcription factor, to activate *FT* and is necessary for floral induction in short-day photoperiods with temperatures of 27 °C. Overexpression of *PIF4* causes premature flowering at 22 °C but when grown at 12 °C this early flowering is strongly suppressed [19].

Two recent similar papers deal with another road to ambient temperature response in parallel to the route of *PIF4*. These articles [20, 21] focus on the control of temperature-dependent flowering by *SVP* and *FLOWERING LOCUS M* (*FLM*) and deliver a number of interesting results. As mentioned above SVP is a floral repressor that interacts with FLC. FLM, related to FLC, also interacts with SVP to control flowering, yet in two opposing ways. This is because of the

process of alternative splicing whereby certain exons or introns that make up the gene are either included or excluded from the mature mRNA before being translated into a protein. This method can therefore give extra levels of molecular control. The two main forms of *FLM* in Col-0 wildtype are labelled *FLM-β* and *FLM-δ*. At 16 °C *FLM-β* is dominant and at 27 °C *FLM-δ* is prevalent [20]. The FLM-β isoforms form a complex with SVP to repress flowering whereas the SVP-FLM-δ complex promotes flowering [20, 21]. Hence the temperature-dependent splicing of these FLM variants has an antagonistic effect at different temperatures. The repressor complex can bind to the promoters of various floral genes, for example *SOC1* or *FT*, to affect their transcription [20, 21]. Furthermore at higher temperatures the stability of SVP protein is decreased [21] and *svp* mutants flowered earlier than wildtype at temperatures from 5 °C to 27 °C [21].

SVP is also involved in the gibberellin pathway. Gibberellin is a phytohormone[3] and is required for short-day flowering in Arabidopsis [22]. The strong GA-deficient mutant *ga1-3* never flowered at 21 °C or 25 °C in 8 hour short-day photoperiods [22]. Exogenous application of GA rescues the flowering phenotype and speeds up wildtype flowering in short-days [22]. A number of enzymes are involved in gibberellic acid biosynthesis. One of these enzymes, GA20-OXIDASE 2, is rate-limiting and, because it's reduced in its gene expression levels by SVP, lower levels of gibberellic acid ensue [23]. Floral integrators are also implicated as having a function in the GA pathway. *SOC1* is regulated by gibberellins [16, 24] as, for example, in the *ga1-3* mutant grown in short-days it was shown that with GA treatment *SOC1* expression significantly increased after six weeks [24]. Blázquez et al. demonstrated that *LEAFY* (*LFY*) levels were far lower in *ga1-3* mutants compared to wildtype and overexpression of *LFY* can partially overcome the failure of these mutants to flower in short-days [25]. More detailed experiments using tissue specific promoters have shown that in long-days GA can increase *FT* transcipt levels in the phloem, and this was likely to be independent of FLC [26]. A recent report found that while gibberellin promotes the transition from vegetative to inflorescence development it surprisingly inhibits flower formation [27]. GA mutants, such as *ga1-3*,

[3] *Phytohormone derives from the Greek for plant stimulus.*

grown in long-days developed more rosette leaves but fewer cauline leaves and exhibited reduced branching [27]. On the other hand after applying a GA treatment the plants formed fewer rosette leaves and more cauline leaves [27]. Thus *LFY* expression is increased by gibberelin levels, which promote the floral transition, but *LFY* then indirectly aids catabolism of GA triggering the onset of flower development [27].

Other hormones are known to influence floral development. *LFY* is directly induced by auxin-activated MONOPTEROS in incipient primordia [28]. In short-days a supply of cytokinin was sufficient to induce flowering that required *SOC1* for this functionality as *soc1* mutants did not flower after hormone treatment [29]. The spatial location of *SOC1* expression and the cross-talk between cytokinin and auxin in the shoot apical meristem (SAM), in particular at the stem cell niche [30], has led some to wonder on the connection between stem cell maintenance, cytokinins and floral integrators at the time of floral induction [31].

When there are no inductive floral signals a plant must still attempt to produce seeds before dying. As a fallback mechanism the ageing pathway ensures that Arabidopsis will flower eventually. The main players so far elucidated in this respect are micro-RNAs (short non-coding sequences around 21 nucleotides in length that silence mRNA) and *SQUAMOSA PROMOTER BINDING PROTEIN LIKE* (*SPL*) transcriptions factors. Early flowering in *miR-172* overexpressing lines is caused by reduced levels of *APETALA2* (*AP2*)-like floral repressors such as *TARGET OF EAT1* (*TOE1*) and *TOE2* [32]. An important finding was that *miR-172*, a floral promoter, increases over time [32] whereas *miR-156*, a floral repressor, decreases as the plant ages [33–35]. *SPL3* is directly targeted by *miR-156* [33] and *SPL3* expression increases more than 10-fold in a week after a shift to long-days [4]. *SPL4* and *SPL5* are also regulated by *miR-156* [33] and this somewhat redundant clade is required for upregulation of meristem identity genes such as *LFY*, *FRUITFULL* (*FUL*) and *AP1* [36]. Similarly *SPL9* binds *SOC1*, an important integrator of the floral pathways, as well as *AGL42*, both MADS box family members [35]. *SPL9* also directly activates *miR-172*, and probably does this by overlap-

ping with *SPL10*, thus regulating vegetative phase change in Arabidopsis [34].

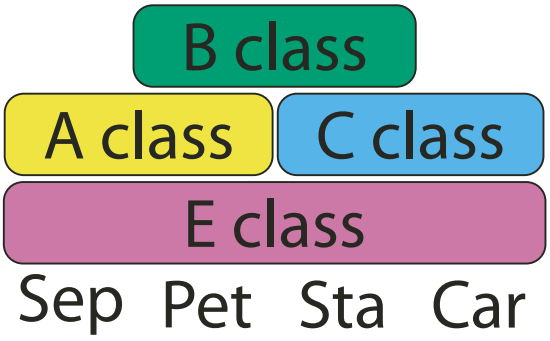
Taken together these results show the plant has an insurance policy for flowering in non-inductive conditions. Interactions between *miR-156*, which promotes juvenile development, and *miR-172*, which is more highly expressed in the adult growth phase, regulate developmental growth and abundance of *SPL* genes. Binding of these transcription factors to regulatory regions of genes involved in the reproductive phase transition can thus eventually stimulate inflorescence development before the plant is too old.

CONSTANS (*CO*) is at the helm of the light-dependent pathway in Arabidopsis [37]. It is regulated by circadian clock genes [37], is expressed in the SAM during floral induction [38] amongst many others tissues during the plant's life [39] and acts in the phloem [39]. The main target of CO is *FT* [17, 40, 41], one of the most important regulators of flowering in higher plants [42]. For a number of species Arabidopsis *FT* homologues are a core element of the photoperiod pathway [43, 44]. *FT* mRNA is transcribed in leaves when CO protein is stabilised in long-day light conditions (for example 16 hours light, 8 hours dark), it being unstable in dark [45]. Through this mechanism flowering is only activated when the days get longer to ensure pollination in the correct season. Short inductions of *FT* are sufficient to cause floral commitment if the plant is old enough [46]. The FT protein is translated in the leaves and moves through the phloem to the SAM in Arabidosis [47–49], and rice [44], and thus is a major component of the "florigen" signal that intrigued early naturalists [42]. This mobile protein provides the timing of flowering and spatial specificity is conferred by the transcription factor FD, with which it functions at the apex [50, 51]. This is evident because *fd-2* can partially suppress the early flowering phenotypes of *FT* overexpressing plants [50, 51]. The FT-FD complex activates various floral meristem genes such as *FUL* and *AP1* [50–52]. In rice, the interaction of homologues of FT and FD, Hd3a and OsFD1 respectively, is mediated by a 14-3-3 protein [53] but this has not yet been determined in Arabidopsis. *TWIN SISTER OF FT* (*TSF*) acts redundantly with *FT* [54]. The double mutant flowers later than either single mutant, but in short-day conditions the *tsf-1* mutant is

severely delayed compared to its effect in long-days [54]. D'Aloia and colleagues show that *TSF* is required for a flowering response to cytokinin in Arabidopsis but *FT* is not [29]. FT is able to activate *LFY* expression through the transcription factor *SOC1* [55, 56]. Thus SOC1 expression is in part controlled by the photoperiod pathway of floral induction [16] as well as integrating gibberellin and vernalisation signals [24] as described above. AGL24 and SOC1 directly bind to each other and cause mutual upregulation during the floral transition, yet they are affected in different ways by upstream elements and affect different downstream genes, for example SOC1 bound the *LFY* genomic sequence but AGL24 does not [57]. This could be because AGL24 works to maintain inflorescence identity as opposed to a floral fate, hence it is targeted by AP1 and LFY for repression [58].

The transcription factor LFY [59, 60] plays a key role in the integration of flowering signals in parallel with FT to activate floral meristem identity genes [40]. As one of the master integrators it functions in multiple pathways as the *LFY* promoter is also a target of GA-dependent signalling [25]. LFY confers floral meristem identity [59] and has a separate role in activating subsequent organ identity genes [61]. AP1 and LFY are direct mutual transcriptional activators [62–64] acting in a positive feedback loop. AP1 has the function of regulating floral primordia growth genes such as *FUL* and *CAULIFLOWER* (*CAL*) [65] whilst preventing reversion in the floral meristem by inhibiting a number of genes including *AGL24*, *SVP*, *SOC1* and *FD* [64, 66]. Furthermore AP1 orchestrates organ specification genes for correct sepal and petal development by binding to MADS-box proteins like SEPALLATA3 (SEP3) [64, 67]. Other homeotic genes are initially controlled by LFY. For instance LFY activates *AG* through direct binding [68].

Organ patterning and cell fate determination in the SAM concerns distinct spatial gene expression patterns that were originally published in the now famous ABC model [69, 70]. In this model certain classes of proteins interact in specific spatial domains to give rise to a distinct floral organ identity. In the outer whorl, where sepals are made, only A class genes are expressed. The next whorl, which gives rise to petals, has A and B class proteins active. Inside are the male reproductive organs, the stamens, and these require the func-



(a) The ABC model of flower development. A class but not B or C class genes are required for sepal (Sep) development. A and B are required for petals (Pet); B and C for stamens (Sta); and only C for carpels (Car). Genes in the A and C classes inhibit each other. The four E class SEPALLATA genes are required for correct development of all organs.

(b) A dissected Arabidopsis flower. From inside out are the carpel, some stamens, two petals and two sepals. A few petals and sepals were removed for clarity of presentation during dissection.

Figure 1.4: Arabidopsis floral organ development.

tion of B and C class genes. Finally the inner whorl is where the carpels, the female reproductive organs, are located. These rely on the presence of the C class proteins. Importantly genes acting in the A (e.g. AP2) and C (e.g. AG) domains are mutually inhibiting [71, 72]. The ABC model was extended by the discovery that SEP1–3 are needed for the B and C class genes to function correctly — all organs are sepal-like [73] otherwise — and SEP4 is required for the correct formation of all organs as leaf-like organs result in quadruple mutants [74]. Figure 1.4a gives the basic conceptual idea.

From the interactions described a general picture emerges of LFY and AP1 at the core of a highly complex network specifying floral fate that operate by repressing inflorescence identity genes and activating downstream patterning genes. This process is most commonly started through the integrators of many signalling pathways—*FT* and *SOC1*. Nevertheless floral repressors play important roles. One major example is *TERMINAL FLOWER1* (*TFL1*) [75] which is

a member of the same gene-family as *FT* and *TSF*, along with three other homologues [41]. It is thus interesting that it features as a strong floral repressor because a single nucleotide base change can give the opposite function — conversion of TFL1's inhibitory function to floral-promoting FT function and vice versa [76]. The shoot apical meristem converts to a terminal flower in *tfl1* mutants [75] and hence *TFL1* maintains the indeterminacy of Arabidopsis. Its expression is noticeably increased after entering the floral transition [77] with very low levels during the vegetative growth phase. *TFL1* does not enter incipient floral primordia due to repression from LFY and AP1 with whom it has a mutually antagonistic relationship [78–80]. A thorough investigation by Conti & Bradley shows TFL1 protein moves outside of its mRNA expression domain in the centre of the mature apex [81]. TFL1 movement in the inflorescence meristem is coordinated by LFY but not AP1 or CAL [81].

In summary the genetics of flowering time are complex and highly interwoven (Figure 1.3). Activators and inhibitors of flowering converge from many pathways to a few pivotal integrators who influence the development of floral primordia which leads to meristem shape changes and organ development. As a dynamic and growing system, mechanical forces will also play a part. To this end Hamant and colleagues [82] proposed that two parallel and (at least partially) independent processes control plant morphogenesis, one depending on microtubules and the other depending on auxin patterning at the shoot apex [83–85]. Thus the web of connections between genetics, hormones, environmental and endogenous factors all combine to govern the correct timing of floral development.

1.4 Mathematical modelling of flowering time in crops and Arabidopsis

High-throughput technologies such as microarrays, deep sequencing, transcriptomics and proteomics have revolutionised plant biology. Progress in these areas has been rapid and has transformed the way biologists tackle new problems, providing a wealth of easily accessible and searchable information such as annotated genomes, phlyogenetic relationships, function and structure predictions, expression and co-expression patterns, metabolic profiles and more.

Sequence-based bioinformatics has become a key part of plant biology and one that is likely to gain importance with the ever-increasing ease and speed of genome sequencing. Mechanistic modelling has played second fiddle to the wave of genetic and bioinformatics discoveries that have been prevalent in the field of plant biology in the past two decades and our understanding is lagging behind the data accumulation rate. More recently, however, there has been recognition that systems approaches, including computational modelling, will have a key role to play in elucidating many aspects of plant development and the interactions between a plant and its local environment [86, 87]. In Arabidopsis, some areas are already well advanced such as circadian clock modelling (reviewed by Bujdoso & Davis [88]). Flowering time control is another area that has benefited from modelling. With the development and availability of computers and software packages, early simple theoretical models have been generally superseded by much more complicated systems of many variables, which are solved numerically.

Whilst modelling of floral development in Arabidopsis was just evolving a decade ago, the modelling of flowering time in crops was in full bloom. However the types of modelling used in crop and model species are different [86] and it is interesting to compare these approaches. Crop modelling [89] has been goal oriented in terms of making useful predictions for agriculture [87], whereas model species approaches have targeted a more gene based understanding of the system. Crop modelling has very much been based on empirical studies, using data such as observed time to flowering or fruit production, to restrain predictions that were built using regression models. With regression statistics and analysis of variance (ANOVA) it is possible to account for factors like CO_2 emissions, location and light intensity that can vary hugely and are of importance to plant breeders and growers alike. Correct timing of crop production is essential for many producers, and the efficacy of these models is testament to their strength.

Many crop models use quantitative trait loci (QTL) analysis for traits of interest, for example yield or days to flower. This operates at a level above genes by linking phenotypic and genotypic data. Genetic markers on a chromosome are tested for association with a trait

that is scored for (quantified), often through field work. This commonly relies upon parent plants being genetically different which allows for identification of recombination effects in the offspring. Then a relatively simple statistical model for the phenotype can be created using the sum of various genetic effects. Markers that segregate with a trait of interest are likely to be near a QTL. Key questions that are attempted to be answered by QTL mapping include: how many QTL are there controlling a trait of interest? where in the genome are they located? and, what is the relative weighting each of them has on the trait?

There are many climate change models for CO_2, water availability, and temperature for the years ahead. These are key factors for plant development and the challenge is to incorporate these predictions into plant breeding tools [90]. A number of examples of modelling in crops and Arabidopsis are now discussed and chapter 4 provides a reflection on how a multiscale approach could lead to the combining of the phenotype-based work in crops with molecular level research. This may be a crucial step in ensuring future plant breeding can successfully incorporate knowledge of climate change at the same time as supporting a growing global population.

An early approach to modelling flowering time was developed by Thornley over 40 years ago [91]. This symmetric model was based on biochemical interactions between two enzymes that catalyse a substrate into two morphogens, the relative concentration of which leads to a switch between either vegetative or flowering steady stable states. With an elegant derivation of the equations this work emphasises minimal modelling and the benefits of a reductionist approach to gain an understanding of a system. Although the application of this work is to flowering plants because of the simplicity of the equations it could be used to describe any system with different developmental pathways. The states of the system are interpreted qualitatively and there is discussion of how perturbations to the system could affect the final outcome. Like many models it is parameter dependent, because the two stable states become one with a change in parameter value. With only one stable state a perturbation of a variable would return the system state to the dividing line between the two states, which is difficult to interpret. Thornley though does

discuss how perhaps in the developing system the parameters could be a function of time, thus at a certain stage the plant may develop a competency to flower corresponding to the system with an unstable steady state and two stable ones. At this point a perturbation from the dividing unstable state, such as a flowering signal, would be sufficient for the plant to switch to a reproductive phase of development. If a vegetative signal was perceived then the plant would enter the vegetative state and hence there would need to be a far larger flowering signal for the plant to switch growth states.

This very early non-species-specific theoretical work is an excellent example of why, although unrealistically simple, it can be good to use mathematical modelling to assist in understanding system structure and provide tractable insights that can lead to the development of more data-based models.

1.4.1 *Crop species*

White et al. [92] include two major flowering regulators of bread wheat in their approach that used genetic information from 29 spring and winter wheat strains. Data from multiple locations worldwide were split into either a calibration or evaluation set for the gene-based model parameters. A linear regression approach was used to estimate the genetic effects of the *Vrn-1* loci on vernalisation requirements and the *Ppd-D1* locus on photoperiod sensitivity. The use of a specific simulation environment is common to these types of models and this work uses CSM-Cropsim-CERES-Wheat [93] which can simulate the development of many stages of wheat growth and also incorporate strain-specific factors. The conventionally estimated parameters in the simulations predicted almost all of the variation in time to flowering for the calibration data, with a modest reduction for the evaluation set, as expected. Results from using gene-based coefficients reduced the accuracy only slightly indicating the possibility that using genetic information in wheat modelling together with the more conventional phenotypic data has potential. The quantity and quality of data is a constraining factor at present especially in terms of understanding the loci effects. Uncertainty is also present in the environmental data, for example the accuracy of the reported weather conditions, which can have large effects.

A similar approach was used in soybean by Messina and others [94]. A simulation model, CROPGRO-Soybean [95] was used with linear functions to predict cultivar-specific parameters which inform estimates of flowering time, as well as post-flowering developmental stages and yield. The model was evaluated using field trial data from other locations, and was shown to predict maturity date particularly well for most varieties. Interestingly the results are stated to be comparable to those from common bean, which is encouraging for the development of gene-based modelling across species.

Yin et al. [96] develop a model for spring barley using reciprocal photoperiod transfer experiments to estimate genotype-specific parameters which are evaluated in independent field trials. Additionally a sensitivity analysis was performed on their four parameters, and the authors show they are all important for predicting inter-genotype differences in flowering time. They also find that the importance of their four parameters can be ranked, with the minimum number of days to flowering at optimal temperature and photoperiodic conditions being relatively the most important, with the photoperiod sensitivity next. This regression based model gave a reasonably good prediction of variation in time to flowering across both genotypes and environments. The original model [96] is then further developed by adding a QTL-base to a new model [97]. The accuracy of this model is reduced by 9% (to 72%) of the overall variation, caused by genotypes and environments, with changing the parameters to QTL effects from the genotype-specific parameters used previously.

Wurr and coworkers [98] consider the effects of climate change on winter cauliflower production using simulations of four different scenarios for future global greenhouse gas emissions. All forecasts predicted a rise in temperature. In the model this increase in temperature led to shorter juvenile and curd growth phases, but longer curd induction in most cases. Importantly location effect was found to dominate the time to maturity, raising questions for both breeders and growers.

A recent model by Uptmoor and colleagues [99], based on a previous crop model [100], uses genotype-specific parameters and QTL effects as the inputs to a model for predicting flowering time in *Bras-*

sica oleracea. In this model the predictability of flowering time using genotype-specific parameters was reduced by unfavourably high temperatures. This suggests that noisy environmental conditions, which can be filtered by using an integrated network approach[4], are not fully taken into account with this modelling framework. Using QTL effects as the parameters instead further reduced the ability of the model to capture inter-genotype variability under both low and high temperatures. Incorporating QTL effects into models does at present seem to produce unsatisfactory results but the exact reasons are not yet clear. This could be because of undetected minor QTL [97] or poor estimation of their effects [101]. Sampling more plants, and at a finer resolution, should result in data that can give a more precise idea of the effects of QTL. Nevertheless the results using genotype-specific model parameters can give good predictions of flowering time but the use of more complex models should, for an extra computational cost, give consistently better predictions.

[4] *See subsection 3.2.4.*

These data-driven approaches can be very successful and highlight the need to reduce the inherent complexity of the system in order to use the power of data to guide predictions. In Arabidopsis research there has been a wide range of methods used to elucidate the underlying biology through simplifying assumptions. The goal is often to gain an understanding of genetic control elements and infer molecular mechanisms. The availability of greater quantities of genetic data in Arabidopsis allows for a more detailed description of processes connected with flowering as discussed below.

1.4.2 *Arabidopsis*

Welch et al. [102] employed a neural network approach to quantifying flowering time in Arabidopsis for a number of genotypes. Neural networks are composed of interlinked nodes, each with a number of inputs, and an output to a subsequent node. This network structure is decided by the modeller. The links between nodes have an associated weight which adjusts the value between the output and input nodes. The weights are established through a training procedure using experimental data, typically using a least squares residual. Welch et al. look at the inflorescence transition in Arabidopsis and how it is specifically controlled by the autonomous and photoperiod path-

ways. Their network can reproduce the floral transition of many mutant genotypes at both 16 °C and at 24 °C. At the lower temperature the rate of Arabidopsis development is much reduced. Intriguingly they find the order of inflorescence transition between two loss-of-function genotypes switches between the two temperatures. Many crop simulation models would not be able to show this result, which demonstrates how using network-based methods could hopefully do more than just predict flowering time.

Prusinkiewicz et al. [103] describe the building of a model to try and understand the development and evolution of inflorescence architectures. The main types of inflorescence architectures observed in nature are panicles, racemes and cymes. This paper relies on the suggestion that these are only a few of the theoretically possible structures that, because of an iterative pattern of development, are available to nature. This iterative pattern is elegantly visualised using L-Systems. The authors introduce the idea of a meristematic continuum that gives rise to shoots at one end, and flowers at the other. In a generalising leap, the authors state this continuum can be characterised by an abstract variable, *veg*, which declines with age. High levels of this correspond to shoot meristems and low levels to floral meristems. It is shown that if *veg* decline is uniform across all meristems a panicle is the result. This is as far as this model will go, so the authors provide further extensions to make the model account for the other main inflorescence architectures. *LFY* and *TFL1* are introduced into their model because mutants in these genes have different phenotypic effects in Arabidopsis. Modelled architectures of mutant and transgenic *LFY* and *TFL1* phenotypes are shown and said to agree with experimental data although photographs of real plants are not included but can be found elsewhere [78, 79]. The authors also discuss the potential evolutionary origins of floral phenotypes. It is interesting that, because not all meristems flower at the same time, racemes and cymes may have evolved to have higher fitness than panicles in a variable growth season. Hence panicles are shown to be relatively more frequent in the tropics. An explanation is also offered that could explain the existence of only these particular architectures. By using layers of 2D fitness landscapes to build a 3D fitness space, the authors capture relationships between

architectures and season duration/plant longevity to show that the angiosperms are only likely to have evolved along high fitness paths that connect racemes, panicles and cymes. The level of abstraction in this work requires further validation to elucidate the biology behind the *veg* factor[5] yet it is an interesting attempt at explaining the evolution and development of diverse inflorescence architectures.

[5] *Perhaps the micro-RNA miR-156 as mentioned in section 1.3 can be considered a candidate as it decreases as the plant ages.*

The intuitively simple ABC model (Figure 1.4a) has stimulated great interest from modellers who naturally wish to provide a more quantitative understanding of the molecular interactions. Two particular studies require detailed comment.

First is a discrete model with logical rules described by Espinosa-Soto and coworkers [104]. After an exhaustive literature search for genetic interactions the authors are able to define a genetic regulatory network of 15 genes involved in cell fate determination. Some connections are hypothesised to ensure the correct expression patterns are recovered. Experiments testing these interactions could therefore provide validation or otherwise of the network structure. Eight genes are Boolean (on or off) in their expression level, but the remaining seven can have an off level, an intermediate level or a full level of activity. The logical rules are therefore based on observed experimental results and in total the network has $2^8 \times 3^7 = 139968$ possible initial conditions. From all these initial conditions the network has only 10 steady states which nicely correspond to the organ types in the apex and the inflorescence meristem where *TFL1* is high but no activity of floral marker genes such as *LFY* or *AP1*. The basins of attraction for the reproductive organs are shown to be far larger than the perianth organs suggesting their fates are less unstable, possibly because they are more important (thus under natural selection), evolutionarily older (gymnosperms have no perianth organs), or both. The final cell types are dependent on the network architecture not the logical rules for each gene as shown by small random changes to the rules. Espinosa-Soto et al. also simulate mutations in the selected genes which recover mutant phenotypes. For example the steady states in the B class *AP3* knockout mutant corresponded to only inflorescences, sepals or carpels and are in absence of petals and stamens as known experimentally [70, 105]. The approach can also be applied to petunia. The advantages of the logical

framework adopted by this paper are that there are no parameters to infer and to a first approximation it likely reflects very well the underlying genetic behaviour.

Second, an ODE model of interacting MADS-box transcription factors controlling floral organ identity was developed by van Mourik and colleagues [106]. The model is based on the demonstration that MADS proteins can form dimers or higher order complexes [107], and this is therefore explicitly included in their model. Redundant genes are assumed to have similar interactions or expression patterns and each system variable is thus representing more than one gene of each class. Triggers are incorporated to drive the system into one of four steady states: sepals, petals, stamens or carpels. In total there are 37 parameters which are optimised by a gradient-based search method. The authors change microarray data from the literature in to a format substituting for whorl-specific protein concentrations which can then be optimised against to determine the model parameters. The fit to the experimental data is reasonable but importantly the model is validated by comparing MADS protein mutants to known phenotypes from the literature. This validation method showed four out of five mutants to be correctly predicted and for the remaining mutant, ectopic AP3 expression, to be half right. Finally the authors remove certain dimers from the network to predict organ mutations. As one example, the removal of the SEP dimer predicts "no development of floral organs", and Ditta et al. [74] have found that the quadruple *sep1 sep2 sep3 sep4* mutant formed leaf-like organs in place of flowers. Thus this model has captured some of the kinetics of MADS-domain protein dimerisation leading to floral organ specification in Arabidopsis which had never been done before. Additionally as a time-dependent system it gives more dynamic information about the variables than the discrete approach taken elsewhere [104]. Furthermore, in a boon to minimal modellers everywhere, a recent follow-up study suggested that the original network could be reduced in its complexity whilst still accounting for the system behaviour under mutant conditions [108].

In combination these two works have taken different approaches to provide a quantitative understanding of the qualitative ABC model. Both routes are valid and with sensible assumptions and simplifica-

tions can predict phenotypes unknown to the models. The ability to also suggest interactions and phenotypes that are untested in the literature gives weight to the involvement of mathematical modelling studies in biology. A discrete approach loses dynamic resolution but has no need for computationally expensive parameter searches. Therefore deriving a network architecture and testing it for coherence before applying a higher level of dynamic modelling may avoid wasted time and effort on an incorrect model [104].

1.4.3 *Summary*

Complex traits are rarely transferable between species, yet genes are frequently highly similar (homologous) and likely to carry out the same functions, motivating gene models. It is often general genetic motifs that are most conserved between species. Thus the knowledge of the workings of one motif in a species is likely applicable to another species. Our increasing understanding of gene networks coupled with QTL analysis allows drilling down to individual genes or even single nucleotide polymorphisms (SNPs). Hence transferable gene-level models that cross scales and integrate up to the environment level are within grasp.

In order to make this approach tractable, many factors are excluded from such GRN based models that are relevant to those with a more agricultural interest. Modelling such large genetic regulatory networks is a complex task as even if all components are known—to perform kinetic measurements for quantities such as binding constants is rarely experimentally feasible. The limited knowledge of component concentrations and kinetic interactions results in a mathematically highly underdetermined problem. This means that the available data is not sufficient to uniquely determine the parameters in the model.

Although as we have seen in the literature different approaches exist for simplifying the parameterisation of the model, e.g. Boolean networks or neural networks, these do not allow so much for a dynamic analysis of a mechanistic model with kinetic parameters having a biological meaning. Thus the focus on differential equation-based systems allows the dynamic system of interacting components to be tracked and can provide a more detailed understanding of the

processes involved. Unfortunately this avenue can require many parameters that have to be constrained by available experimental data to some degree. In the next section we cover methods for parameter estimation and discuss how using Bayesian inference allows us to quantify the uncertainty in our model's parameters.

1.5 Parameter estimation

The scarcity of large quantities of high quality and detailed mechanistic data is a common problem faced by computational biologists seeking to model an experimental system. In all but the simplest cases a challenge to the mathematical modeller is the choice of a useful parameterisation of the problem and, often in discussion with experimentalists, devising ways of obtaining reasonable estimates for the parameters of the system. Depending on the method, these parameters may be inherent to a machine learning approach, so-called black box parameters, and of little interest to the biologist or for mechanistic models they may actually correspond to biological entities such as concentrations, dissociation constants or degradation rates that may be used for validation purposes and the design of further experiments. In this thesis the focus is on dynamic mechanistic modelling for which the parameters themselves are of interest and not merely a means to an end. Many mechanistic modelling studies in biology have employed ODEs as the mathematical framework of choice [106, 109–111]. The reasons for this include the natural way that many biological problems can be posed as the study of the behaviour of a dynamic system of interacting components over time and the well-established numerical routines for solving such systems [112]. For instance, converting a genetic regulatory network into a mathematical formalism can be achieved using established enzyme kinetics and following standard conventions [109]. This approach gives rise to a mechanistic model with (in principle) measurable, kinetic parameters. Unfortunately, however, these parameters are often unknown experimentally, or determined under *in vitro* conditions for analogous systems, and so have to be estimated from available data. This is a major hurdle that has received a lot of attention from systems biologists [113–115].

A common approach is to use optimisation algorithms to find the best fit to the data [115–117]. Local optimisation is very well established and numerous high-performance computing software tools are available, often based around variants of Newton's method. Nevertheless the non-linearity of biological systems can lead to multimodal fitness landscapes [118] that require global optimisation techniques [113, 114, 119] to avoid getting trapped in local minima.

Global optimisation however remains a challenge. Despite a number of very powerful, modern techniques such as: simulated annealing [120], particle swarm [121], Kalman filters [122, 123], Bayesian approaches [124, 125], genetic algorithms [126] and, aptly-named for plant research, invasive weed optimisation [127], finding a global optimum can rarely be guaranteed in practice and in finite time. Furthermore, it has been noted that the global minimum may not result in biologically realistic parameters [128].

These methods can be motivated by invoking maximum likelihood arguments. A known problem with maximum likelihood and, in general, optimisation approaches is that without further precautions they can lead to the overfitting of a model to the data, i.e. the parameters are far more sharply defined than is justified from the information content of the data [129]. These are well-documented problems with established solutions such as Bayesian methodology and information theory-based corrective terms to the maximum likelihood value such as the Akaike information criterion (AIC) [130, 131]. A short review of these approaches applicable to systems biology is given by Kirk et al. [132]. Another issue is that the best-fit set of parameters to a model may not be representative of parameter space [133]. An optimisation algorithm may miss important solutions or contributions from other parts of parameter space. Furthermore, it has been shown that in systems biology that not all parameters are uniquely identifiable [134]. There are issues of sloppiness and correlations between parameters [134, 135]. Parameters have also been found to behave differently between corresponding deterministic and stochastic systems [136].

These issues affect reverse-engineering, which attempts to infer networks, functions and other regulatory mechanisms causing a system's output. Thus when parameters are non-identifiable or show

non-linear dependencies this can cause difficulties in understanding the real system from a mathematical model of the system. For some biological systems noisy and sparse data can bring further headaches when attempting to recover system behaviour. Accurately capturing experimental data in a model therefore suffers from structural and practical difficulties — both the model structure (connections, inputs and outputs) and lack of informative data could be limiting. With an experimental-modelling cycle, both of these will, hopefully, be at least partially addressed yet this may not be feasible due to issues of cost and time. Thus providing a mathematical description of a system that ensures parsimony and accuracy can be a challenge. A comprehensive review of reverse-engineering from different perspectives has recently been written by Villaverde et al. for systems biology [137].

The Bayesian framework [138, 139] is an attractive way of dealing with the issues just raised in a way that reduces the risk of over-fitting. As succinctly stated by Radford Neal [140],

> *"Bayesian inference is an approach to statistics in which all forms of uncertainty are expressed in terms of probability".*

The history of Bayes' theorem stretches back over 250 years to the work of the Rev. Thomas Bayes [141]. Bayes' theorem in its most introductory form is commonly presented using two sets, A and B (see Figure 1.5). The theorem follows from the definition of joint probability $\mathrm{P}(A \text{ and } B) = \mathrm{P}(A|B)\mathrm{P}(B)$ and describes the conditional probability of being in set A given that an element belongs to set B, as such

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(B|A)\mathrm{P}(A)}{\mathrm{P}(B)}, \tag{1.1}$$

where $\mathrm{P}(A|B)$ is the posterior probability, $\mathrm{P}(A)$ and $\mathrm{P}(B)$ are prior probabilities and $\mathrm{P}(B|A)$ is the conditional probability of B given A. Bayesian inference relates to degrees of belief and provides an effective way of combining information such that new data can easily be incorporated. This leads to the message that "today's posterior is tomorrow's prior" [142]. Importantly, the Bayesian approach is consistent in its treatment of inference problems regardless of the details of the questions being asked.

Bayesian inference naturally encompasses Occam's razor [143, 144] and so inherently accounts for the trade-off between the good-

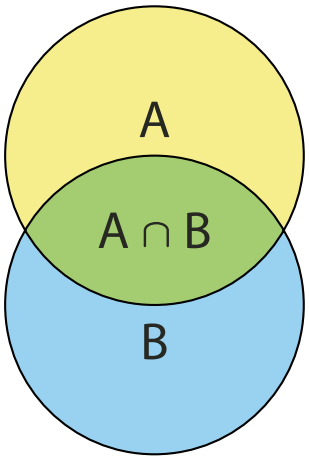


Figure 1.5: Illustration of joint probability. The probability of the intersection of two sets, A ∩ B, is the probability of A and B. The conditional probability A given B is equal to the probability of A and B normalised by dividing by P(B). As joint probability is commutative i.e. P(A ∩ B) = P(B ∩ A) Equation 1.1 follows naturally.

ness of fit of a model and its simplicity [145]. The Bayesian approach doesn't aim to produce a point estimate for quantities of interest but captures the full uncertainty of the problem that is reflected in the posterior probability distribution. In particular for non-unimodal distributions point estimates can be misleading. Bayesian techniques are gaining interest in numerous research areas and finding increased application in computational biology [146, 147] due to the availability of state-of-the-art developments [118, 124, 148–151]. Recent further advances have shown that multi-dimensional biophysical problems can be tackled successfully within the Bayesian framework; for example Markov chain Monte Carlo (MCMC) was employed for suitably approximating a prior distribution for studying the insulin secretion rate [152] and copula-based Monte Carlo sampling was used for comparing models of human zirconium processing [153]. However, the computational demands for such approaches often make them prohibitive for many problems. A main reason for this computational effort is in the calculation of high-dimensional integrals that arise through the processes of marginalisation and normalisation in Bayesian inference [133, 138]. Monte Carlo techniques are the established way to compute such integrals, however they can require many thousands of cycles to deliver adequate results and there are known issues with MCMC sample decorrelation times [154]. Nested sampling [155] (chapter 2) was put forward as a Bayesian variant of this approach and was shown to perform well for simple test examples [156]. Recently this approach has been used with success for: astronomical data analysis [157, 158], exploring configurational phase space of chemical systems [159], parameter inference of a circadian clock model [160] and for one of the most challenging problems in biophysics, namely the exploration of protein folding landscapes [161].

Having introduced the benefits of the Bayesian framework and its growing popularity amongst scientists some of the theory necessary for understanding how it works is presented next.

1.5.1 *Bayesian parameter inference*

For parameter inference the task is to infer the probability over the parameters, ω, for the hypothesis or model, $\mathcal{M}$, given some data $\mathbf{D}$

from an experiment and capturing also all relevant information I. This can be done within the setting of Bayes' Theorem which can be rewritten as

$$\mathrm{P}(\omega|\mathbf{D},\mathcal{M},I) = \frac{\mathrm{P}(\mathbf{D}|\omega,\mathcal{M},I)\cdot\mathrm{P}(\omega|\mathcal{M},I)}{\mathrm{P}(\mathbf{D}|\mathcal{M},I)}, \tag{1.2}$$

where $\mathrm{P}(\omega|\mathbf{D},\mathcal{M},I)$ is the *posterior* probability, $\mathrm{P}(\mathbf{D}|\omega,\mathcal{M},I)$ is the *likelihood*, $\mathrm{P}(\omega|\mathcal{M},I)$ is the *prior* probability and $\mathrm{P}(\mathbf{D}|\mathcal{M},I)$ is the *evidence*. We make use of the following shortened notation [156]: $\mathcal{P}(\omega)$ represents the posterior, $\mathcal{L}(\omega)$ the likelihood, $\pi(\omega)$ the prior and $\mathcal{Z}$ the evidence, hence Equation 1.2 becomes

$$\mathcal{P}(\omega) = \frac{\mathcal{L}(\omega)\pi(\omega)}{\mathcal{Z}}.$$

As the evidence is not a function of the parameters it does not need to be computed for parameter inference, which explains the success of MCMC methods that explore a posterior distribution proportional to the correctly normalised distribution. However calculating the evidence in crucial for Bayesian model comparison.

1.5.2 *A common likelihood function*

Maximum entropy arguments lead to the assignment of a normal distribution for the errors in the data [139], and if the $n_{\mathbf{D}}$ data points are independent the log-likelihood function resembles a least-squares residual

$$\log\mathcal{L} = -\sum_{i=1}^{n_{\mathbf{D}}}\log\left(\sigma_i\sqrt{2\pi}\right) - \sum_{i=1}^{n_{\mathbf{D}}}\frac{(d_i-y_i)^2}{2\sigma_i^2} \tag{1.3}$$

where d_i is the given data at point i, σ_i its corresponding standard deviation and y_i the value computed from the model at that point. More complex error models can be used if information is available or justified from the underlying experiment, however in this thesis Equation 1.3 is the only form of likelihood function considered. If the standard deviations σ_i are assumed to be constant throughout the data set then the first term on the right hand side is itself constant and can be ignored for the purposes of model comparison (see below).

1.5.3 *Bayesian model comparison*

Bayes' theorem not only enables us to infer parameter distributions but also provides a framework for model comparison. The posterior

probability of a model $\mathcal{M}$ is

$$\mathrm{P}(\mathcal{M}|\mathbf{D},I) = \frac{\mathrm{P}(\mathbf{D}|\mathcal{M},I)\cdot\mathrm{P}(\mathcal{M}|I)}{\mathrm{P}(\mathbf{D}|I)} \text{ or } \mathcal{P}(\mathcal{M}) = \frac{\mathcal{Z}\cdot\pi(\mathcal{M})}{\mathrm{P}(\mathbf{D}|I)}.$$

To compare models we take the *posterior odds* of two models, $\mathcal{M}_i$ and $\mathcal{M}_j$, by taking the ratio and cancelling the term $\mathrm{P}(\mathbf{D}|I)$. Thus

$$\frac{\mathrm{P}(\mathcal{M}_i|\mathbf{D},I)}{\mathrm{P}(\mathcal{M}_j|\mathbf{D},I)} = \frac{\mathrm{P}(\mathbf{D}|\mathcal{M}_i,I)\cdot\mathrm{P}(\mathcal{M}_i|I)}{\mathrm{P}(\mathbf{D}|\mathcal{M}_j,I)\cdot\mathrm{P}(\mathcal{M}_j|I)} \text{ or } \frac{\mathcal{P}(\mathcal{M}_i)}{\mathcal{P}(\mathcal{M}_j)} = \frac{\mathcal{Z}_i\cdot\pi(\mathcal{M}_i)}{\mathcal{Z}_j\cdot\pi(\mathcal{M}_j)}.$$

If we have no prior preference for either model, i.e. $\pi(\mathcal{M}_i) = \pi(\mathcal{M}_j)$, then these terms cancel out and the models are compared according to their respective evidences, which is identical to the normalisation constant in Equation 1.2. This ratio of evidences is called the *Bayes factor* [138, 162],

$$\mathcal{B}_{ij} = \frac{\mathrm{P}(\mathbf{D}|\mathcal{M}_i,I)}{\mathrm{P}(\mathbf{D}|\mathcal{M}_j,I)} = \frac{\mathcal{Z}_i}{\mathcal{Z}_j}.$$

Thus the evidence $\mathcal{Z}$ is the key quantity that can be computed by marginalising the likelihood $\mathcal{L}(\omega)$ over parameter space,

$$\mathcal{Z} = \int \mathcal{L}(\omega)\pi(\omega)\,\mathrm{d}\omega.$$

The evidence embodies the so-called Occam factor [133]. This is a measure of the extent to which the prior parameter space collapses to the posterior space after seeing the data. A model with more parameters typically has a greater volume of prior parameter space, and if the data are well described by only a small region of this space it will be penalised for this extra complexity. So a less complex model (fewer parameters) that fits well to the data for a larger region of its parameter space would be preferred by the Bayes factor calculation [133].

1.5.4 *Jeffreys' scale*

A qualitative scale for the interpretation of Bayes factors was given by Jeffreys [138] and adapted by Kass & Raftery [162]. The version used in this thesis is shown in Table 1.1 for a Bayes factor $\mathcal{B}_{ij}$. If the log-Bayes factor is negative it can trivially be reversed to provide evidence against the competing hypothesis. The interpretations are based on a natural logarithm scale and due to computational issues with underflow for the magnitude of the numbers occurring

in Bayesian inference the calculations are also on a logarithm scale. Hence the use of a log-likelihood function. For this reason if the first term on the right hand side of Equation 1.3 is constant taking the log-Bayes factor is simply a subtraction of the same term from both evidence values and so can be safely ignored, which is the case in most examples within this thesis.

$2 \ln \mathcal{B}_{ij}$	Evidence against $\mathcal{M}_j$
0–2	Hardly worth mentioning
2–6	Has some substance
6–10	Strong
> 10	Very strong

Table 1.1: Jeffreys' scale for interpreting Bayes factors. Jeffreys [138] provided a grading of decisiveness of evidence to support or reject a hypothesis, $\mathcal{M}_j$. This scale was slightly adapted by Kass & Raftery [162] in their classic paper. It should be noted that in contrast to null hypothesis significance testing (reject/fail to reject the null) the Bayes factor provides the ability to reject or accept either the null or alternative hypothesis.

NESTED SAMPLING IN SYSTEMS BIOLOGY

2

2.1 Introduction

Nested sampling [155, 156] is a technique for Bayesian inference that prioritises calculation of the evidence [133], the normalisation constant of the posterior distribution. This is an important quantity for Bayes factors, used in model comparison, but is challenging to calculate in general because it involves evaluating a multi-dimensional integral. Nested sampling focuses on calculating this integral and as a by-product of the algorithm's exploration of parameter space it can optionally produce samples from the posterior distribution. Thus it can also be used for parameter inference as is traditionally done in Bayesian computation by Markov chain Monte Carlo (MCMC) techniques. Importantly nested sampling has shown encouraging results and efficiency gains over other sampling techniques [157, 158, 163] particularly in the areas of astrophysics and cosmology. Furthermore reviewing this literature revealed that a well developed and cited implementation of the nested sampling algorithm, called MultiNest [164], existed which had been applied to astronomical data sets. Problems in physics can be of high dimension, non-linear and multimodal which is also typical of a number of problems in modelling biological processes. However systems biology, or biological models in general, had received little exposure to nested sampling when this work was initialised [159] although subsequently further articles have appeared which are relevant to the field [160, 161]. In this chapter we evaluate how well Skilling's nested sampling, and in particular MultiNest, works for system biology problems and non-linear biological models by comparing evidence values to those approximated using numerical integration and study the accuracy of parameter inferences by comparing results to those of the current workhorse of Bayesian inference, MCMC. It is demonstrated how nested sampling can be used to reverse-engineer a system's behaviour whilst accounting for the uncertainty in the results. Thereafter we present results

that employ this approach with various oscillating biological models for sparse, noisy data that is typically available to a mathematical modeller. Finally our results with nested sampling indicate that the addition of data from extra variables of a system can deliver more information for model comparison than increasing the data from one variable, thus providing a basis for experimental design.

2.1.1 *Nested sampling is a Monte Carlo technique constrained by the likelihood*

Skilling [155, 156] showed that the evidence can be calculated by a change of variables that transforms the multi-dimensional integral $\mathcal{Z} = \int \mathcal{L}(\omega)\pi(\omega)\,\mathrm{d}\omega$ over parameter space into a one-dimensional integral over likelihood space, Figure 2.1. Following Skilling [155, 156], denote the elements of prior mass as $\mathrm{d}X = \pi(\omega)\mathrm{d}\omega$ then $X(\lambda)$ is the proportion of the prior with likelihood greater than λ so that

$$X(\lambda) = \int_{\mathcal{L}(\omega)>\lambda} \pi(\omega)\mathrm{d}\omega. \tag{2.1}$$

The evidence can then be expressed as

$$\mathcal{Z} = \int_0^1 \mathcal{L}(X)\mathrm{d}X, \tag{2.2}$$

where $\mathcal{L}(X(\lambda)) \equiv \lambda$. The basic algorithm proceeds as follows:

1. Sample the prior n times to generate an active set of objects $\omega_1, \ldots, \omega_n$ and calculate each object's likelihood.
2. Sort the objects based on likelihood.
3. Withdraw the point with lowest likelihood ($\mathcal{L}^*$) from the active set, leaving $n - 1$ active samples.
4. Generate a new sample point from the prior subject to the likelihood constraint $\mathcal{L}(\omega) > \mathcal{L}^*$.
5. Add the new sample ω_{new} to the active set to return the set to n objects.
6. Repeat steps 2–5 until termination.

So by focusing on the evidence rather than the posterior distribution, a, potentially, high-dimensional integral can be replaced by a sorting problem of the likelihood [156], although high-dimensional

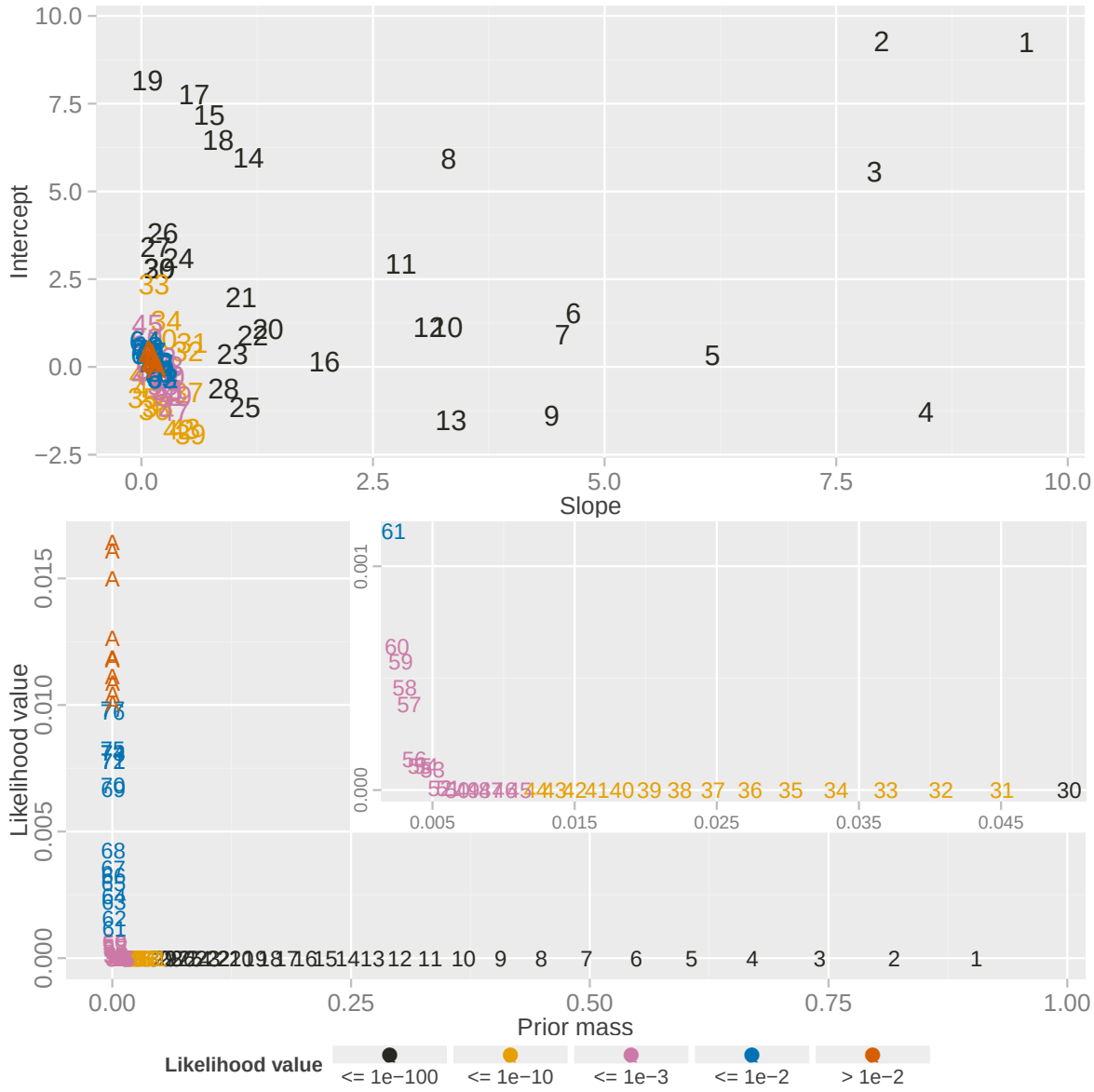


Figure 2.1: Samples from parameter space are mapped to likelihood-prior space to calculate the evidence. Numbers in the plots indicate the order in which a sample point was withdrawn from the active set of $n = 10$ objects, and are coloured by a grouping according to likelihood values. Top: Parameter space is shown for a two parameter linear model example. Bottom: The corresponding points are shown indicating the volume of the prior still remaining when that point was removed. At each rejection the remaining prior mass is multiplied by a factor $\exp(-1/n)$. The aim of nested sampling is to calculate the area under the curve. The inset zooms in on the region just as the bulk of the evidence is to be accumulated. This occurs at small values of the prior after finding the regions of highest likelihood. Points labelled 'A' are those left in the final active set at termination.

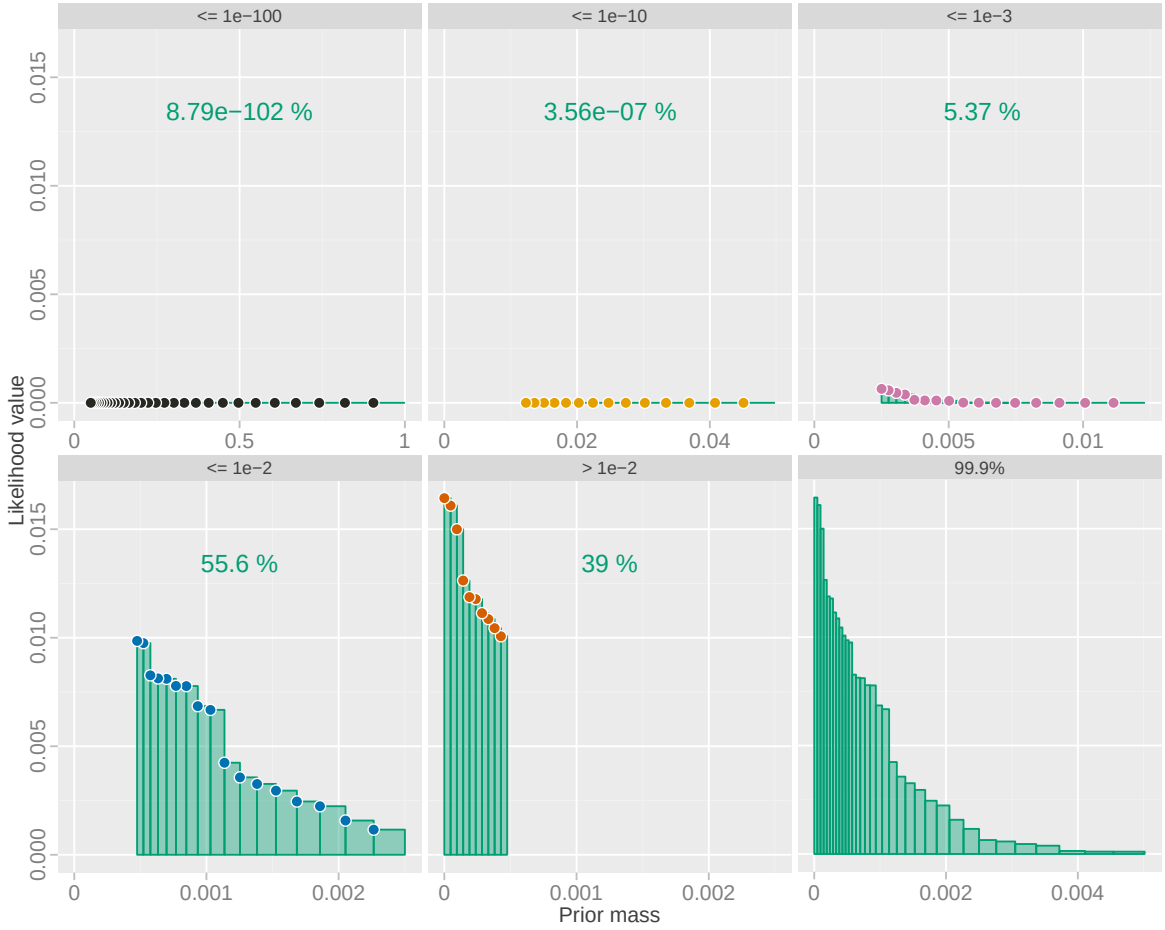


Figure 2.2: Nested sampling calculates the evidence as a sum of the likelihood weighted by the prior. The first five panels show points grouped by likelihood value and a percentage of how much each group contributes to the final evidence. The bulk of evidence is accumulated by the contribution from 17 sample points (lower left) in this example. The final panel shows that fewer than half the total samples make up 99.9% of the integral, which is found in a small fraction of the prior.

sampling around each point remains. With the generated samples, the integral (2.2) can be approximated (see Figure 2.2) using basic quadrature as

$$\mathcal{Z} \approx \sum_{k=1}^{N} h_k \mathcal{L}_k, \tag{2.3}$$

where $h_k = X_{k-1} - X_k$, $(X_0 = 1)$ is the width between successive sample points and N is the total number of samples i.e. the number of objects discarded from the active set plus those remaining in the active set at termination. Alternatively a more accurate method such as the trapezium rule could be used for the integration although the error introduced beyond this is of a higher order than that from other aspects of the algorithm [155].

The target of nested sampling is to calculate the area under the curve in likelihood-prior space as shown in the bottom of Figure 2.1. For most examples, as this one, this area is all but zero until the high likelihood regions are found. These parts of parameter space are often only found in very small domains of the prior range — notice in the top plot of Figure 2.1 how small the blue and orange regions are compared to the entire prior. As each point is removed from the active set its associated likelihood is multiplied by a prior width. This width is shrunk geometrically at each iteration because of the potentially huge range of prior-to-posterior collapse. In our examples this reduction is by 1 part in N on the log scale, corresponding therefore to each width being $\exp(1/n)$ smaller than the previous [155, 156].

2.1.2 *Posterior distribution and summary statistics*

Each accepted sample point θ_k, is assigned a weight, $h_k \mathcal{L}_k$, that corresponds to how much it contributed to the evidence. From these weights it is possible to estimate individual marginal and joint probability distributions for all parameters to examine their uncertainty, modality, correlations or other aspects. Code was written that produces a table of binned values for these estimated distributions from the posterior output of nested sampling in a suitable way for plotting. This code is freely available[6] and is made use of in figures in this thesis.

[6] *Accessible from* `https://github.com/nstjhp`.

Additionally estimating summary statistics of the posterior distribution is straightforward given the posterior output from nested sampling [155, 156]. By using the weights assigned to each point, as

above, the mean μ and standard deviation σ of a parameter θ from N samples are calculated as

$$\mu_\theta = \sum_{k=1}^{N} \frac{h_k \mathcal{L}_k}{\mathcal{Z}} \theta_k, \qquad \sigma_\theta = \left(\sum_{k=1}^{N} \frac{h_k \mathcal{L}_k}{\mathcal{Z}} \theta_k^2 - \mu_\theta^2 \right)^{1/2}.$$

When using a normal distribution with fixed standard deviation, σ, as a likelihood function, choosing a larger value of σ leads to greater evidence and larger variance of the inferred parameters in most cases.

MCMC methods produce samples from parameter space that are equally weighted and hence can be used to gain an understanding of the underlying posterior distribution. This is also possible with nested sampling. Staircase sampling can be used to generate a number of equally-weighted posterior samples [156], which is necessarily fewer than the number of nested sample points. This is implemented by default in MultiNest [164] and we make use of this later on to explore the posterior dynamics of our biological systems.

2.1.3 *MultiNest*

MultiNest is a Fortran library implementing nested sampling developed by astrophysicists in Cambridge [158, 164]. The main challenge of nested sampling is step 4 in the algorithm above — generating a new sample from the prior that must have a higher likelihood than the discarded sample. Building on pioneering work by Mukherjee et al. [157] MultiNest uses ellipsoidal rejection sampling to efficiently propose new samples. The trick is to enclose all live points in the active set by a group of ellipsoids, which are allowed to overlap. The new point is then sampled from within the volume of the enclosed ellipsoids, save for a user-chosen multiplicative factor that affects the efficiency, but also potentially a bias, in the algorithm. This (inverse) factor is chosen by the user from $(0, 1]$ with higher values reducing the time the algorithm takes but potentially missing some prior volume with likelihoods greater than the current likelihood contour. We chose the target efficiency to be 0.5 to err on the side of accurate evidence values rather than maximum efficiency. Multimodal posterior distributions can be sampled from effectively, as points falling into modes can be enclosed within their own ellipsoid. This allows for the calculation of separate "local" log-evidences for each posterior mode if required. We did not make use of this or other advanced

features, like parallelisation, available in the MultiNest software in this thesis. The MultiNest algorithm was shown to solve toy problems with multimodal or high curvature posteriors of the type that occur in cosmological problems, and additionally for less challenging examples it was shown to be highly efficient and produce similar estimates to MCMC [164]. Recent developments such as Importance Nested Sampling which uses an alternative summation of the evidence by including trial points that don't satisfy the likelihood constraint can calculate the evidence with even greater accuracy [165]. In summary, Multinest has been shown to be a fast and efficient library for nested sampling applications in cosmology and astroparticle physics (for many references read Feroz et al. [165]) which shares similar types of problems and posterior distributions with those in systems biology. Thus we used MultiNest for all nested sampling results in this thesis after a testing phase to establish the best control parameters.

2.2 Testing the accuracy of evidence calculation

An obviously important property of nested sampling to investigate is the accuracy of the method for realistic biological problems. A number of comparisons to analytic solutions are achieveable through clever choice of prior and likelihood function. These toy examples however do not reflect the use of nested sampling and Bayesian inference for complex problems. Typical scientific problems can demonstrate high dimensionality, non-linearities or other difficulties for inference methods. Difficulties in testing arise when the true solution of a realistic problem is not known. In our case we test the accuracy of nested sampling on biological data where the true evidence can be approximated using numerical integration. We took expression data of the flowering time genes *TFL1* and *FT*, determined by quantitative polymerase chain reaction (qPCR), Figure 2.3. Three different models between the antagonistic genes *TFL1* and *FT* are investigated: a linear model, a quadratic or a sigmoidal relationship. The number of parameters in these models are two, three and four respectively. The measurement errors are not known but modelled as a normal distribution with $\sigma = 0.5$ (data in arbitrary units). By keeping the dimensionality to four or below we can use brute-force integration

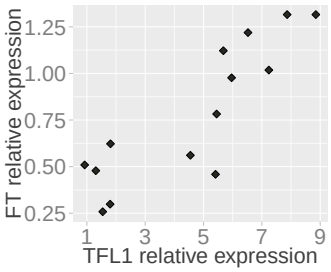


Figure 2.3: Relative expression levels of two floral genes. qPCR of the whole rosette in Arabidopsis upon the floral transition was performed by Katja Jaeger and expression of *FT* and *TFL1* quantified. This figure has been redrawn from Figure 7 published by Jaeger et al. [166].

Model	$\log \mathcal{Z}$
Linear	−12.016
Quadratic	−17.329
Sigmoidal	−8.987

Table 2.1: $\log \mathcal{Z}$ values calculated by numerical integration. Brute-force numerical integration over the prior domain using a fine grid of step-size 0.01 was used to approximate the log-evidence for three relationship models between *FT* and *TFL1* using the data in Figure 2.3.

to make a good approximation to the true value of the log-evidence. A small step-size of 0.01 was used across the uniform prior domains to calculate the numerical approximations given in Table 2.1.

We tested the accuracy of nested sampling against two main control parameters of the algorithm: the number of objects in the active set and the termination tolerance.

2.2.1 *Termination criteria*

There is no rigorous termination criterion to suggest when we have accumulated the bulk of $\mathcal{Z}$ [156]. This is because there may be a region of very high likelihood in a tiny volume of parameter space which is very hard to discover yet if found could dominate the evidence value. However whether such a region exists is impossible to know either *a priori* or *a posteriori* for many practical problems. Skilling [155] suggests three ways and importantly notes that when to stop is a matter of user judgement. The easiest way is to stop the sampling after a pre-defined number of steps. This method however could either be inefficient due to sampling far more than is required, or inaccurate due to not sampling enough. In the materials applications of Partay et al. [159] and Burkoff et al. [161] they set their convergence criteria to reflect the nature of protein folding, based on the bounded nature of the energy, whereas Aitken & Akman [160] compare log-weight ($\log h_k + \log \mathcal{L}_k$) values 50 iterations apart. The example code provided with the introduction of nested sampling [155] uses a condition that continues sampling until the number of samples significantly exceeds (in fact doubles) the number of prior objects multiplied by the current value of the information, H. A similarly plausible criterion also discussed by Skilling is implemented in the MultiNest code [164]. Termination is decided by approximating the remaining evidence that can be accumulated from the posterior. This amount can be estimated as $\Delta\mathcal{Z}_i = \mathcal{L}_{\max} X_i$, where $\mathcal{L}_{\max}$ is the maximum likelihood value of the active set and X_i is the remaining prior volume [164, 165].

We investigated how the use of different levels of tolerance affects the accuracy of log-evidence values for the three models mentioned above and a range of prior sizes, which is equivalent to the number of objects in the active set. As can be seen in Figure 2.4 the different

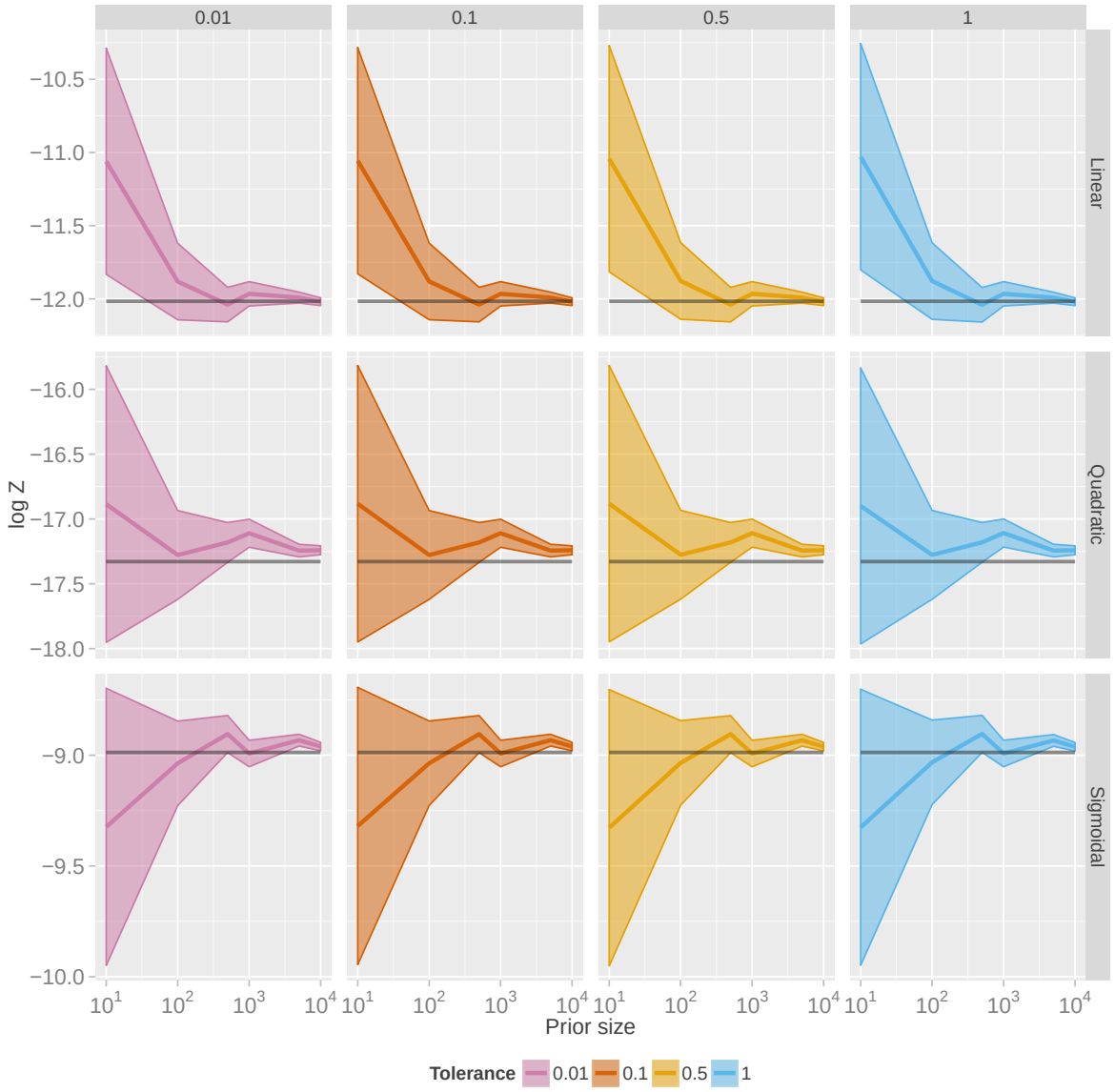


Figure 2.4: The effect of tolerance and prior size on the accuracy of nested sampling. For four different tolerance levels, 0.01, 0.1, 0.5 and 1.0 (columns), the three different models for the qPCR data (rows) and different prior sizes, the evidence (thick coloured line) and its associated numerical error (ribbon) was calculated for the same random seed. An approximation to the true evidence value was made by brute force integration over the prior domain (dark line). The chosen tolerance levels did not affect the accuracy as intra-row each ribbon looks the same. Increasing prior size increases our confidence in the evidence estimates for all models with accuracy generally very good above 100 prior objects. The rate of convergence is $\mathcal{O}(n^{-1/2})$ [156, 167], n being the total number of samples.

tolerance values have little effect, particularly when compared to the far greater effect of the prior size. Why is this? The reason that the chosen tolerance values do not affect the accuracy of the evidence value is that the bulk of evidence is already accumulated by the time the algorithm nears termination. The usual dynamics of nested sampling's progression, as noted by Skilling [155], towards the posterior is that the likelihood increases faster than the widths decrease until the decreasing width starts to dominate the likelihood values. This can be because the highest likelihood regions have been found and the increase in likelihood is not so great anymore. It is at this transitioning stage that most of the evidence integral is accrued. Thus when we come to judge when to stop the procedure we have already calculated the bulk of the posterior and the higher precision of our chosen tolerances did not affect the evidence values as shown in Figure 2.4. As Jeffreys' scale (subsection 1.5.4) suggests a differentiation between models, i.e. Bayes factor, based on half a point difference in evidence on the $\log_{10}$ scale to be safe we chose a tolerance of 0.5 in the (natural) log-evidence calculation. This value agrees with that used in example problems from the literature [164].

2.2.2 *Prior size*

As we have seen, the larger the size of the active set the more confident nested sampling is in its evidence calculation. For the nested sampling algorithm a greater sampling density from the prior distribution will increase the chances of highly probable areas being explored. In the study of protein folding [161] a set of 20000 prior objects was used to provide a wide selection of conformations. At the other end of the scale it has been shown that maintaining a set of 25 active points can produce accurate parameter mean and standard deviations that are relatively insensitive to the prior size [160]. The greater the prior size the longer the method takes to reach its stopping criterion. A robust way to measure efficiency across different active set sizes is to compare the total number of likelihood function evaluations. This negates any potential differences in computer architecture or CPU load. Additionally evaluating the likelihood function is often, in compute time, the most costly part of many inference schemes, particularly when a set of differential equations need

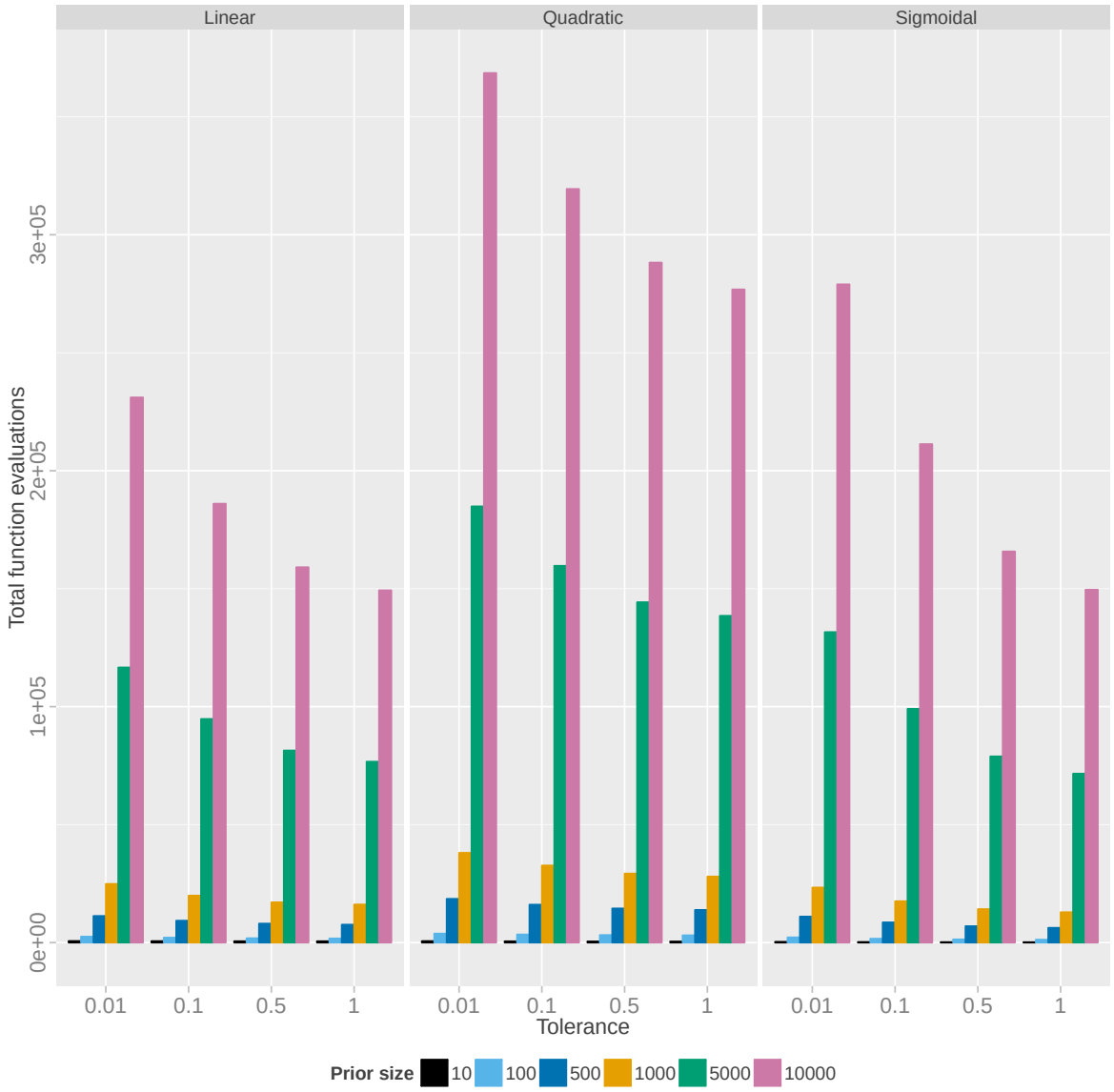


Figure 2.5: The effect of tolerance and prior size on the efficiency of nested sampling. For the four different tolerance levels and three different models for qPCR data, the total number of likelihood evaluations were compared to different prior sizes, for the same random seed. Increasing prior size increases the work done approximately linearly so that an active set of 10000 points takes roughly 10 times more function calls than 1000 points. Increasing the tolerance value reduces the number of function calls required, as expected.

Objects	Calls	Ratio
10	668	66.8
100	3844	38.44
500	18568	37.136
1000	38078	38.078
5000	184899	36.9798
10000	368564	36.8564

Table 2.2: Approximate linear increase in log-likelihood function calls with increasing number of prior samples.

to be solved or a simulation run for a given set of parameters. To choose an efficient yet accurate number of prior objects for future computations we compared six different prior sizes and the number of function calls for the three models as shown in Figure 2.5. We see an approximately linear increase in the number of evaluations with increase in prior size, see for example the results for the quadratic model with a tolerance of 0.01 in Table 2.2. As is expected increasing the tolerance serves to decrease the number of function evaluations required. We would like to choose a value of the prior size that gives accurate evidence values, is efficient, yet effective, for our future models which will be of higher dimension. Taking into account both Figure 2.4 and Figure 2.5 it was decided that 1000 active points would be a sensible trade-off to give accurate estimates for reasonable computer effort.

Table 2.3: Comparison of calculated log-evidence values. The log-evidence calculated by brute-force numerical integration over the prior domain and the nested sampling (NS) estimate using 1000 prior samples and a tolerance of 0.5 for each of the three different models for the qPCR data.

	Linear	Quadratic	Sigmoidal
$\log \mathcal{Z}$	-12.016	-17.329	-8.987
NS $\log \mathcal{Z}$	-11.97 ± 0.08	-17.11 ± 0.11	-8.99 ± 0.06

It is interesting to note that the number of likelihood calls in nested sampling does not depend solely on dimension of the problem, here the number of model parameters. Skilling states that nested sampling itself ignores dimensionality [155] which is instead a complication for the sampler within the constrained likelihood to handle. MultiNest was designed to work up to moderately high numbers of parameters [164] and thus handles these test cases without issue. The reason that the quadratic, three parameter model takes longer to run is due to the information content. The information, which is a measure of prior-to-posterior collapse, for the three models was approximated by brute-force numerical integration to be: $H_{lin} = 6.96$; $H_{quad} = 11.85$; $H_{sig} = 3.57$. After data acquisition, H is a natural logarithmic measure of the amount of information in the posterior relative to the prior [156]. The posterior, which therefore occupies approximately a fraction e^{-H} of the prior [156], will be located in a smaller domain of the prior for higher values of H, and thus is harder to find. This in turn leads to more samples being required to discover the posterior, explore it and calculate the evidence.

2.2.3 *Posterior samples are chosen successively in higher probability regions*

With the constraint upon the likelihood, the method moves up the likelihood gradient to regions of higher probability (even if these regions become disconnected in parameter space). This is demonstrated in Figure 2.6. As the algorithm moves through iterations there is a narrowing of the regions of higher probability as the worse samples are removed and better ones that satisfy the constraint on the likelihood survive. Thus the algorithm discovers the posterior distribution. With the points coloured by their log-likelihood value, Figure 2.7, we can see clearly how the active set of objects migrates to areas of highest likelihood. In this case all the objects left in the active set after stopping the algorithm are located in one small cigar-shaped region of parameter space. This is where the bulk of probability mass is located for this linear model and qPCR data. This region includes the maximum likelihood parameters shown by a yellow disc in the bottom right panel of Figure 2.7. If the procedure was run for even more samples, for example by reducing the tolerance level, the objects would continue to move up towards the peak of the posterior probability distribution, and cluster closer together, but as we saw in subsection 2.2.1 this will have very little effect on our final evidence.

The posterior parameter distribution allows for identification of areas where parameters can be either stiff or sloppy. Figure 2.7 conveys how in one direction the posterior distribution is wide (sloppy) whereas in the perpendicular direction it is well defined (stiff). This example demonstrates the point made by Erguler & Stumpf [134] in their Figure 1. Disperse parameter sets are commonly found in systems biology problems yet can lead to useful predictions [135, 168]. Notwithstanding the technical difficulty of visualising multiple dimensions a multimodal posterior distribution can reveal the regions of parameter space that lead to high probability yet may be disconnected above a certain probability threshold.

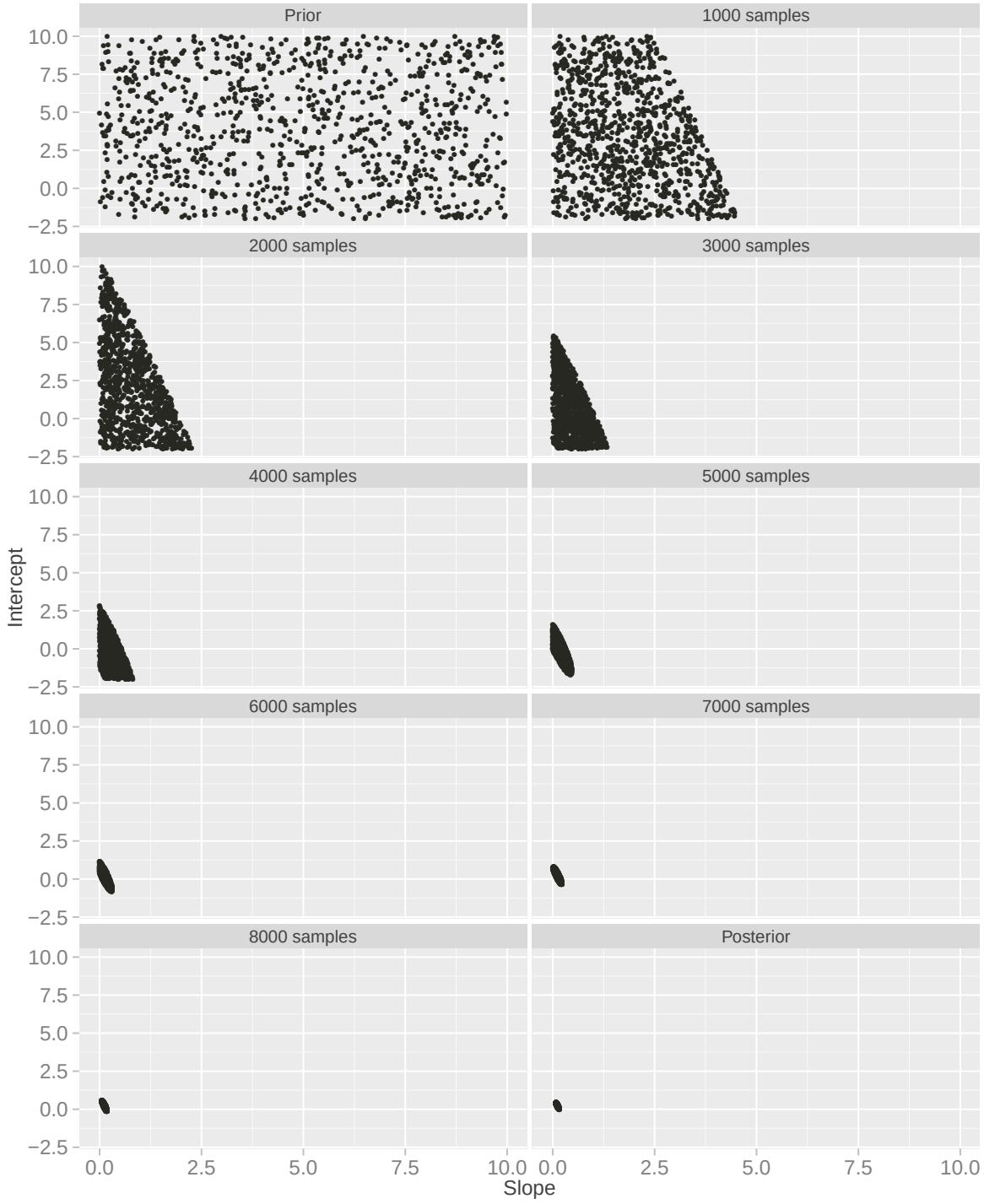


Figure 2.6: Nested sampling removes points that don't meet the likelihood constraint. From an initial uniform prior parameter distribution (Prior), nested sampling selects points that are in regions of higher likelihood. The sample sets are shown after every 1000 sampling iterations and then with the final active set (Posterior) after termination of the algorithm. In this case the sampling ends up in a single region of high probability after sample points that don't satisfy the likelihood constraint are discarded. The underlying model is the linear model for qPCR data and the samples are from the two parameters of the linear model, the slope and intercept terms.

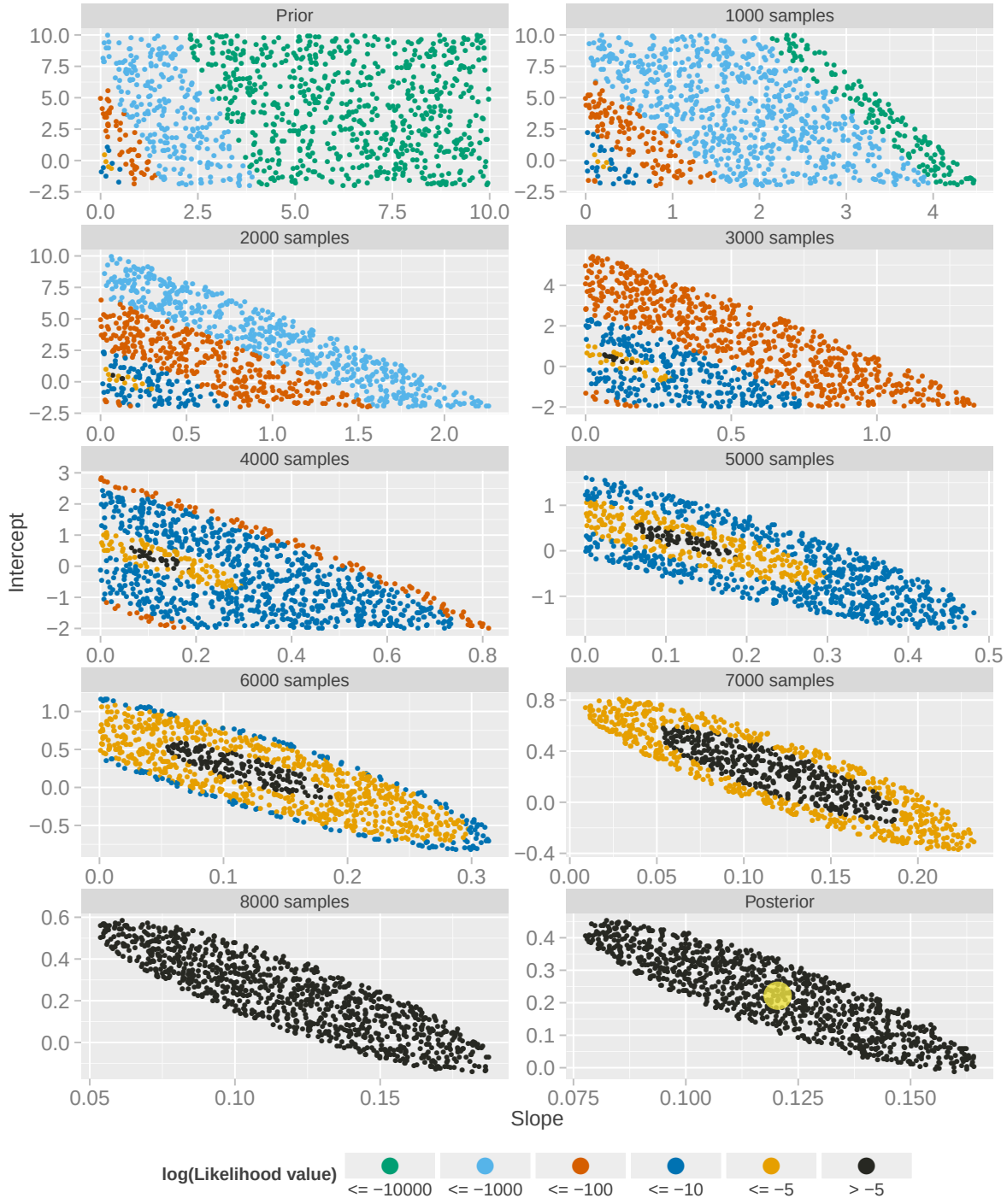


Figure 2.7: The migration of objects to higher likelihood regions. As for Figure 2.6 but now with log-likelihood values grouped into levels and coloured by these levels to show the migration to regions of high probability. We zoom in on the active set of 1000 sample points — notice the continual change of axes indicating a shrinking of the areas of highest probability. Over half the prior samples have a log-likelihood worse than -10000 yet within 2000 sampling iterations all these samples have been replaced. In the final sample set (Posterior) we indicate the location of the maximum likelihood sample point with a yellow disc. The slope parameter has a narrower range of possible values compared with the intercept parameter, which means it is a stiffer parameter. We also note the obvious correlation between the parameters.

2.3 Results

2.3.1 *Nested sampling for parameter inference in systems biology*

The current workhorse of Bayesian inference is MCMC which for a huge variety of problems will converge to the posterior distribution given enough time. The resulting posterior is only required to be proportional to the true posterior and thus calculation of the normalising constant, the evidence, can be a complicated task [133, 155]. However MCMC, or a variation, is routinely used for parameter inference as we get the full posterior distribution and thus are able to quantify our uncertainty which we are unable to do with optimisation techniques. As discussed nested sampling obtains posterior samples as a by-product of its evidence calculation and, as explained in subsection 2.1.2, from these samples we are able to perform parameter inference. Naturally it is necessary to compare the results of nested sampling to the established technique for parameter inference, MCMC. Thus output from nested sampling was compared with that of MCMC for Bayesian inference of two test problems.

In the first case, data were generated from the curve $y = 3\tanh\left(\frac{x}{2}\right)$ from $[-5, 5]$ at intervals of 0.5 to give 21 data points. Noise from a standard Gaussian, $\mathcal{N}(0, 1)$, was added to the generated data. As expected from this low dimensional problem both nested sampling and MCMC find similar solutions with identifiable parameters whose means are good summaries of their distributions given the level of noise, Figure 2.8 left.

In the second example, our data was the previously discussed qPCR expression levels of the flowering time genes *TFL1* and *FT* (Figure 2.3). As before three different models between the antagonistic genes *TFL1* and *FT* are investigated: a linear model, a quadratic or a sigmoidal relationship. The measurement errors are not known but again modelled as a normal distribution with $\sigma = 0.5$ (data in arbitrary units) which was found to be consistent with estimated noise from the data. Also a simulated annealing algorithm [120, 169] was used to optimise the parameters for a comparison with the means of our posterior parameter distributions. The fits to the data using the mean values for the three models are shown on the right in Fig-

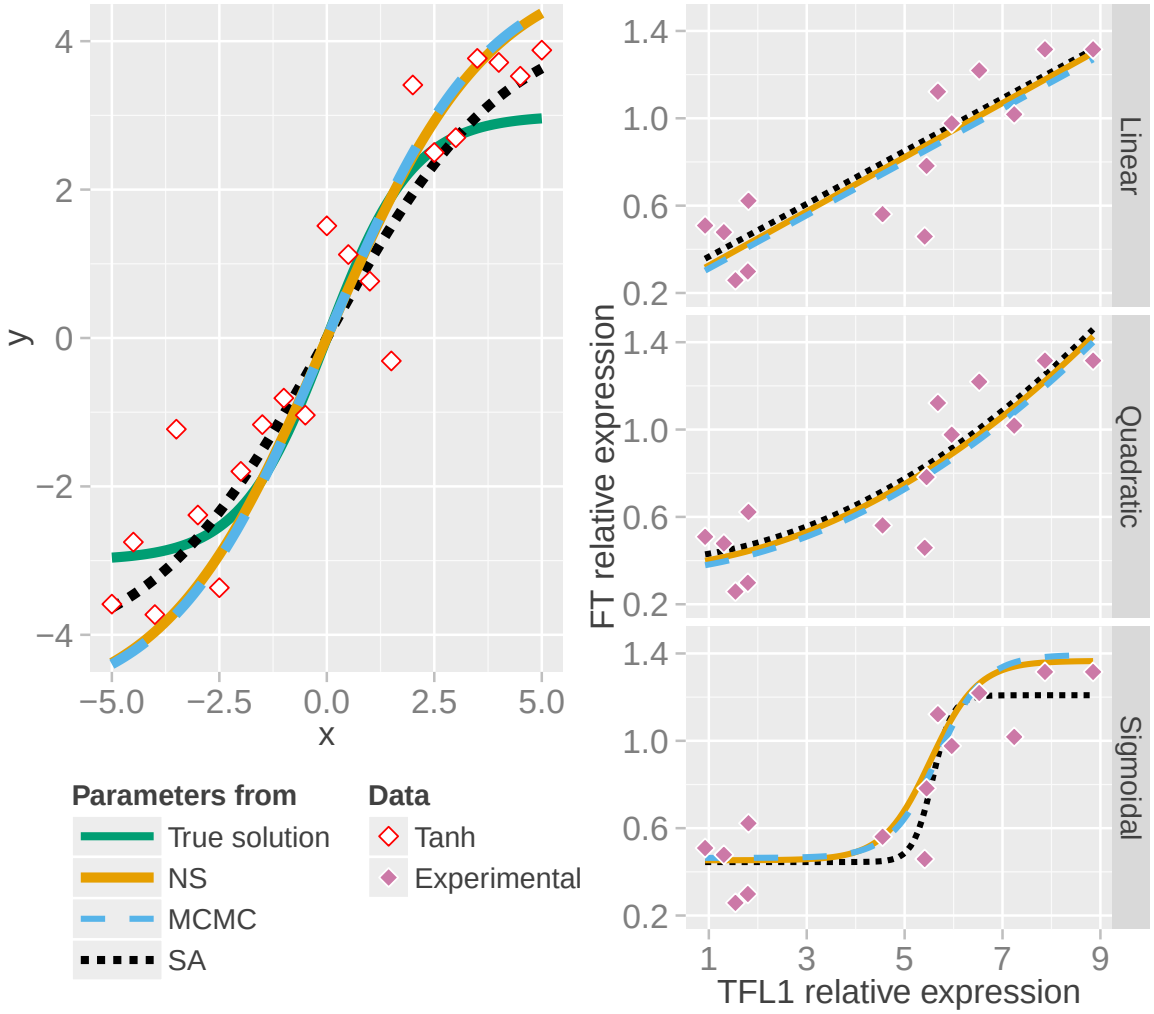


Figure 2.8: Nested sampling produces equivalent estimates to MCMC. (Left) Nested sampling (orange solid line) and MCMC (skyblue dashed) produce a similar estimate of the parameter means given noisy data (white diamonds) generated from $y = 3\tanh\left(\frac{x}{2}\right)$ (green line). The solution using an optimised point estimate of the parameters from simulated annealing is shown as a black dotted line. (Right) Using three different relationship models for flowering gene expression data, nested sampling, MCMC and simulated annealing produce near identical curves for a linear model of the experimental data (purple diamonds) and for a three parameter quadratic model, using the mean parameter values from the inference methods. Curves are offset by one line width for clarity. For a four parameter sigmoidal model MCMC and nested sampling infer comparable parameter means (given in Table 2.4). Note that some parameter sets from the posterior distributions follow a similar trajectory to the point estimate from simulated annealing.

	Hyp. tangent		Linear		Quadratic			Sigmoid			
	θ_1	θ_2	m	c	γ	β	α	k_1	k_2	k_3	k_4
NS mean	5.05	0.26	0.12	0.20	0.01	0.02	0.37	1.37	2.06	5.53	0.45
MCMC mean	5.01	0.27	0.12	0.22	0.01	0.02	0.38	1.39	2.09	5.68	0.46
SA	4.36	0.24	0.12	0.22	0.01	0.02	0.38	1.21	5.00	5.57	0.44
NS SD	1.84	0.24	0.05	0.26	0.02	0.23	0.44	0.53	1.42	2.00	0.22
MCMC SD	1.80	0.32	0.05	0.27	0.02	0.23	0.43	0.55	1.42	1.98	0.23

Table 2.4: Comparison of parameter means and standard deviations. The mean and standard deviation (SD) values of the parameters from nested sampling (NS), MCMC and the point estimates from simulated annealing (SA). The data came from $y = 3\tanh\left(\frac{x}{2}\right)$ with additional noise and from Figure 2.3 to which we fit three models: Linear $y = mx+c$; Quadratic $y = \gamma x^2+\beta x+\alpha$; Sigmoid $y = k_4+(k_1-k_4)/(1+\exp(-k_2(x-k_3)))$.

ure 2.8. All methods find a very similar solution for the linear model, and equally for the three parameter quadratic curve. For the four parameter sigmoid model $y = k_4 + (k_1 - k_4)/(1 + \exp(-k_2(x - k_3)))$ the results are also comparable. The optimisation procedure fits the data well, with a steeper gradient than the inference methods, yet this does not imply it is better despite appearances. Instead this suggests that using the mean parameters are not representative of all posterior parameter sets from MCMC and nested sampling. A number of samples from the posterior distribution are also able to follow a similarly steep trajectory as the simulated annealing result but we decided to only show one curve to avoid complicating the plot further. Furthermore the maximum likelihood values were similar between nested sampling and simulated annealing for all three models. The means and standard deviations of the parameters from nested sampling and MCMC are in good agreement, Table 2.4. The log-evidences found are in Table 2.3 which, on Jeffreys' scale, prefers the four parameter sigmoid model to explain this data set.

The remarkable similarity of the parameter moments summarised in Table 2.4 gives us confidence that nested sampling will produce parameter inferences that agree with MCMC.

2.3.2 *The repressilator*

The repressilator [170] is a frequently used system to evaluate parameter estimation developments [122–124, 171]. The repressilator is a synthetic network of transcriptional regulators comprising three genes in a feedback loop that is capable of producing oscillations. It is also the core structure of a recent circadian clock model [172]. The governing equations used are as follows

$$\begin{cases} \dfrac{\mathrm{d}m_i}{\mathrm{d}t} = -m_i + \dfrac{\alpha}{1+p_j^n} + \alpha_0 \\ \dfrac{\mathrm{d}p_i}{\mathrm{d}t} = -\beta\,(p_i - m_i) \end{cases} \tag{2.4}$$

where $i = \{lacI, tetR, cI\}$ and $j = \{cI, lacI, tetR\}$. α_0 was set to 0 and $n = 2$ so that our prior contained both stable and unstable domains [170]. Initial conditions and parameters were chosen that produce oscillations in the synthetic data, Table 2.5. To show the power of nested sampling for this example we use synthetic data from just one variable, p_{cI} (cI protein), collected at two-minute intervals for 50 minutes. The data has Gaussian noise added to it with a standard deviation of 10% of the range. It is assumed we do not know, or cannot measure, the initial conditions for the five other variables, and attempt to infer these too. Uniform priors were used for all parameters with $\alpha \sim U(0, 1000), \beta \sim U(0, 100)$ and the initial conditions are drawn from $U(0, 50)$. We choose a constant value of σ in our log-likelihood function that is equivalent to the amount of noise added. When standard deviations can be estimated from experimental data these values should be used in the error model if the noise distribution is unknown, or alternatively we could infer the standard deviation parameter. Either better quality (less noise) or greater quantity of data are both able to increase the accuracy of estimates of the parameter posterior probability distributions, as one would intuitively expect.

Using nested sampling we can produce an estimate of the means and standard deviations of the inferred parameters as explained in subsection 2.1.2. The actual values and inferred values are shown in Table 2.5. The two parameters α and β are estimated accurately and furthermore their standard deviations in Table 2.5 are much lower relative to the prior size than for the initial conditions.

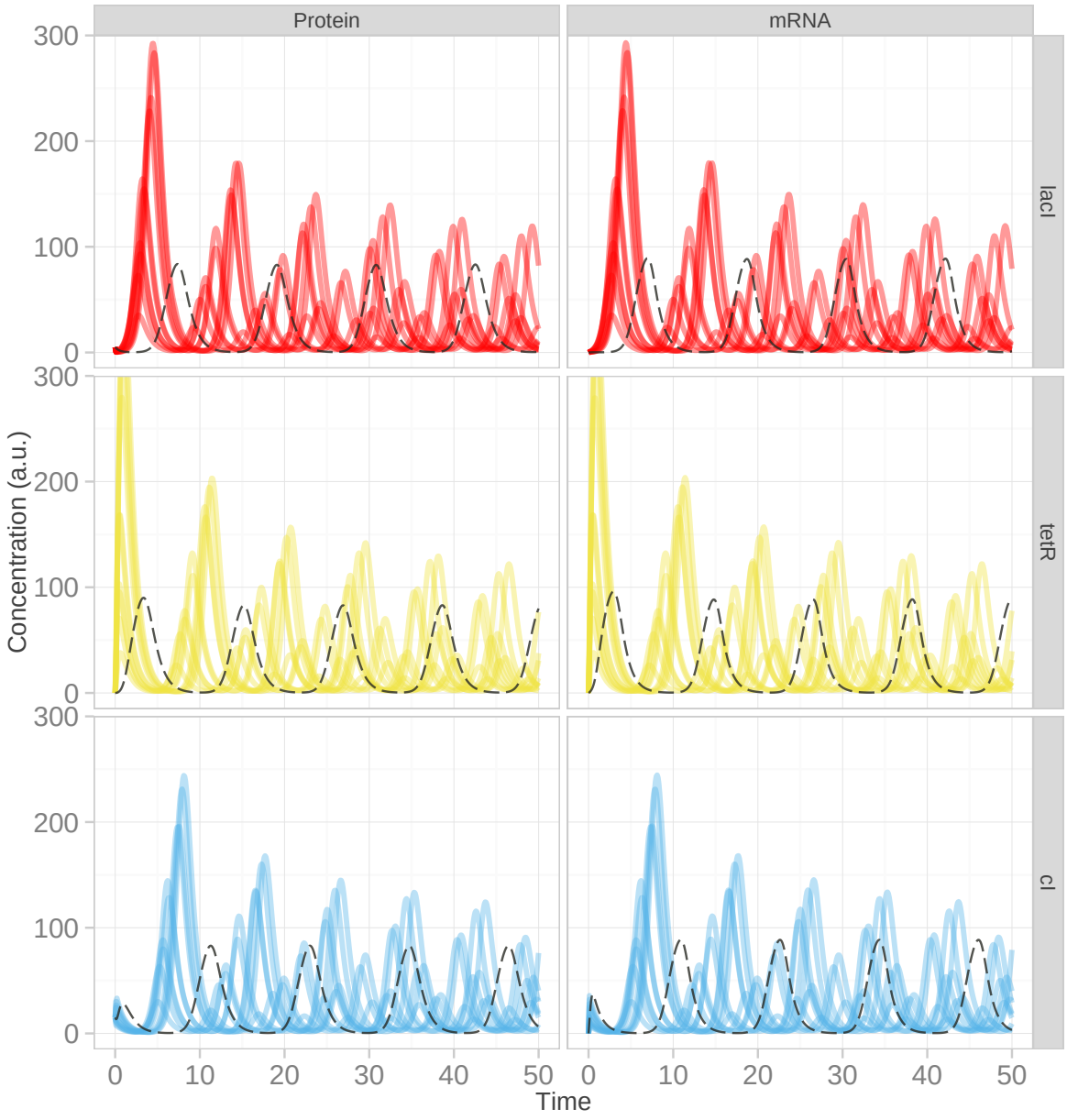


Figure 2.9: The dynamics of the repressilator with parameters sampled from the uniform prior. 10 different solutions of the system's six variables are shown with α and β chosen randomly from a uniform prior. Compared with Figure 2.10 the dynamics show a wide range of solutions. Solution with $\alpha = 125$ and $\beta = 2$, dashed black line; prior sampled dynamics, transparent coloured lines.

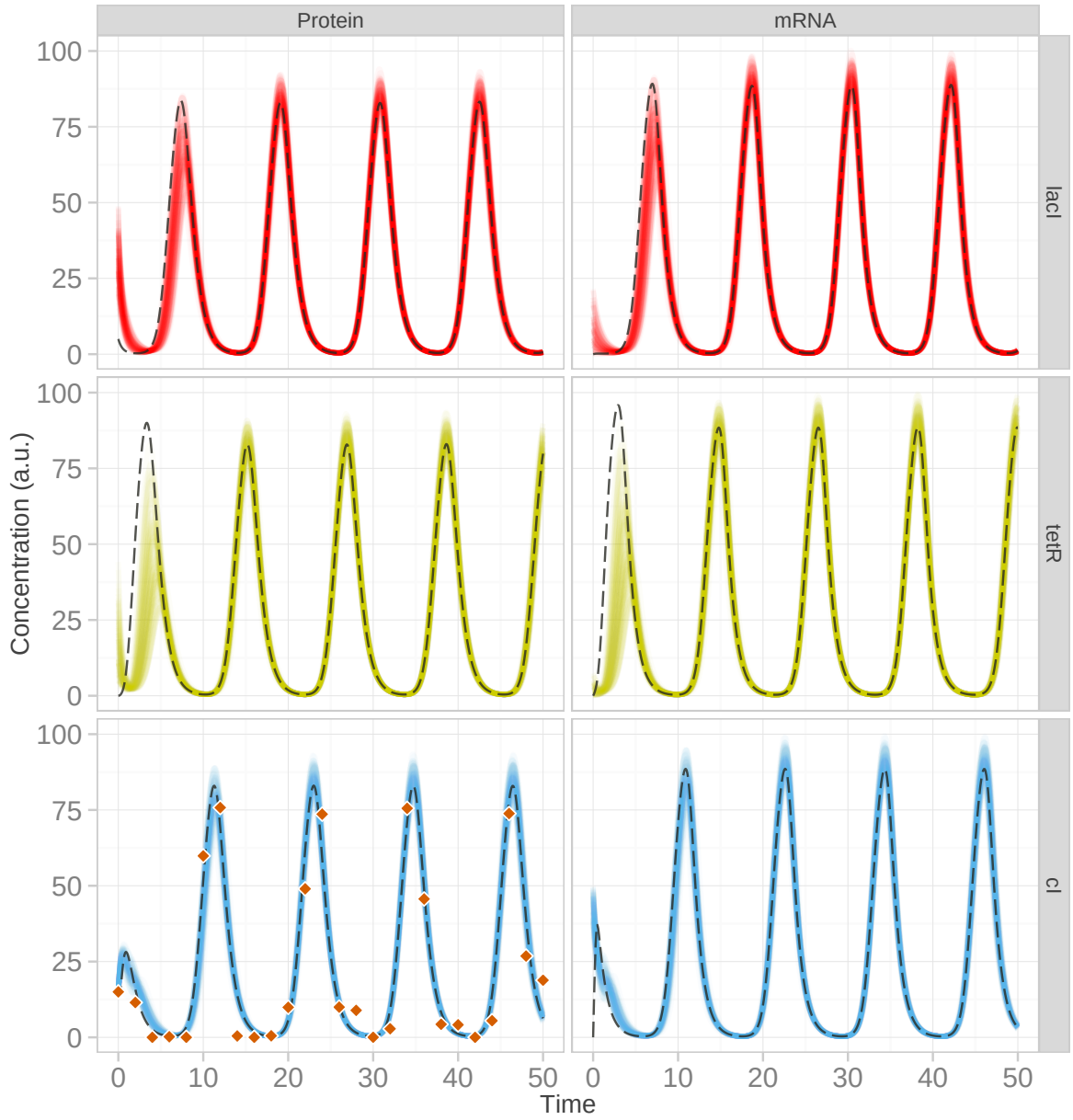


Figure 2.10: The dynamics of the repressilator with parameters sampled from the inferred posterior. 100 equally-weighted posterior samples of the system's six variables are shown. Compared with Figure 2.9 the dynamics have been significantly constrained by the data (vermilion diamonds) so that all solutions are close the true solution (dashed black line, parameters given in Table 2.5). The 26 noisy data points were informative enough to allow discovery of a posterior distribution that produces accurate dynamics for all variables. Estimated dynamics, transparent coloured lines.

	α	β	p_{lacI}	p_{tetR}	p_{cI}	m_{lacI}	m_{tetR}	m_{cI}
True	125.00	2.00	5.00	0.00	15.00	0.00	0.00	0.00
Estimated mean	128.47	2.02	33.38	15.34	-	7.21	2.66	43.21
Estimated SD	5.88	0.05	8.46	10.73	-	5.26	1.73	4.67

Table 2.5: Parameters and initial conditions of the repressilator model. The values of the parameters α, β and initial conditions of the six variables used to generate the simulated data prior to addition of Gaussian noise, and the inferred means and standard deviations (SD) from the routine. p: protein, m: mRNA. The initial amount of cI protein was assumed to be known.

If we consider the model output with 10 pairs of the parameters α and β randomly drawn from the uniform prior there is a wide range of dynamics, Figure 2.9, compared to the known solution (dashed black lines). In contrast, after the data have arrived, we can use the equally-weighted posterior samples to see how informative the data were about the parameters. Figure 2.10 shows the dynamics from 100 posterior parameter sets. The data (shown in the bottom left panel) have constrained the parameter distribution significantly such that all sets closely match the true parameters' dynamics (dashed black lines). As can be seen, despite not estimating the initial conditions well, they are not that important for capturing the qualitative dynamics of the entire system. This is because the repressilator system has a limit cycle and is therefore insensitive to most initial conditions. After the first peak the inferred oscillations match very closely to the true solution for all variables even though the algorithm only had a few, noisy data points available for one variable, cI protein. Even the first peak is fairly well estimated by the posterior distribution. The log-evidence for this model and data is -34.27 ± 0.14.

In this example, and like Figure 2.7, the data significantly reduced the probable volume of parameter space from a wide prior distribution to a narrower posterior. In spite of the fact that the data were few and noisy the simulations from the posterior distribution show us that the data were still informative enough to reconstruct the system's dynamics accurately. A lack of accuracy in parameter estimations but well captured systems dynamics is a phenomenon that has been well studied in recent years [134, 135, 168]. In this case the unknown initial conditions and a lack of parameter identifiability

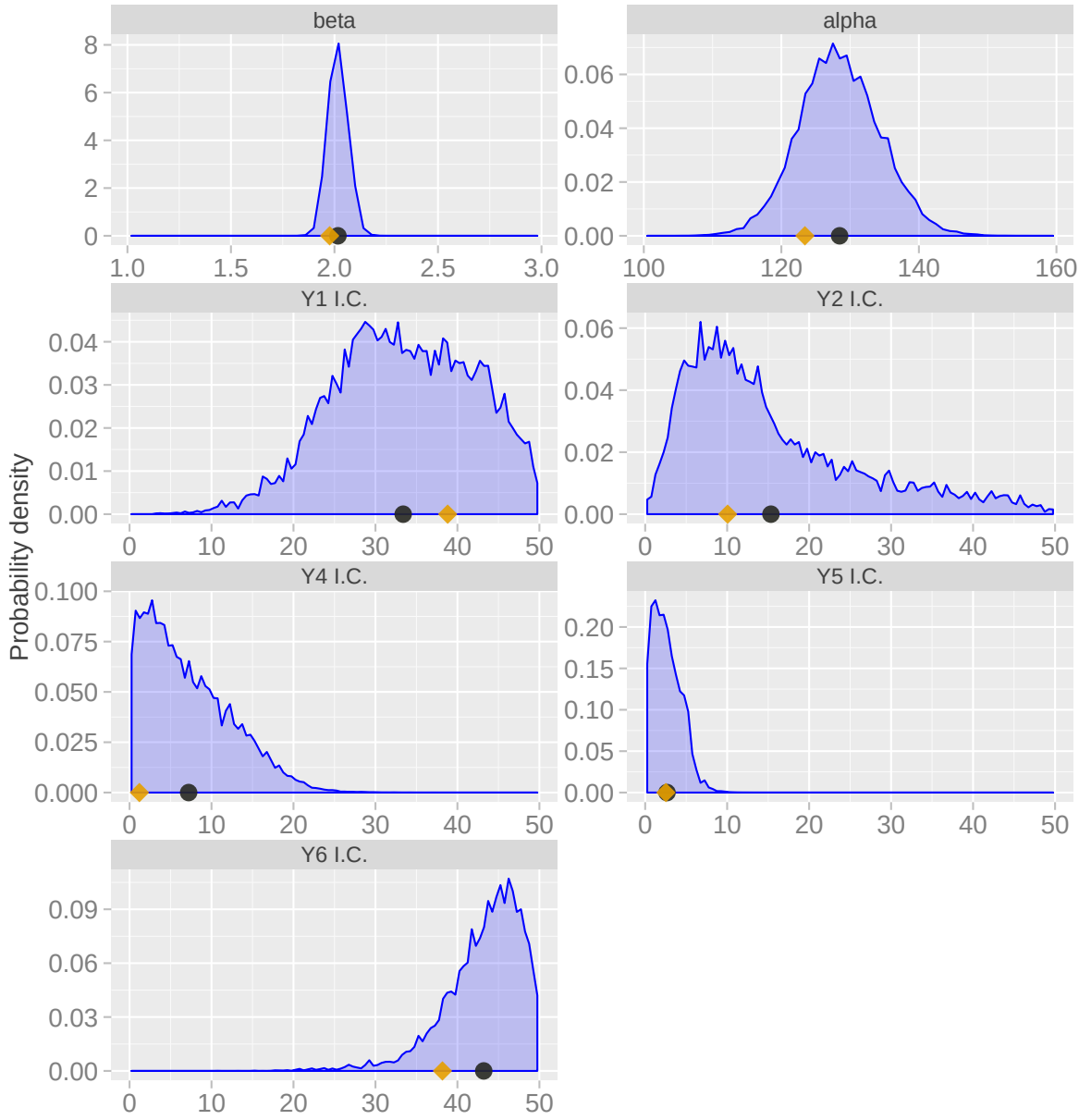


Figure 2.11: Estimated marginal distributions of the repressilator example with missing initial conditions. Using the posterior samples produced as a by-product of nested sampling we can produce marginal distributions. In this example the mean and best-fit points are close to the peak of estimated probability density. Mean parameter value, black circle; best-fit likelihood parameter value, orange diamond.

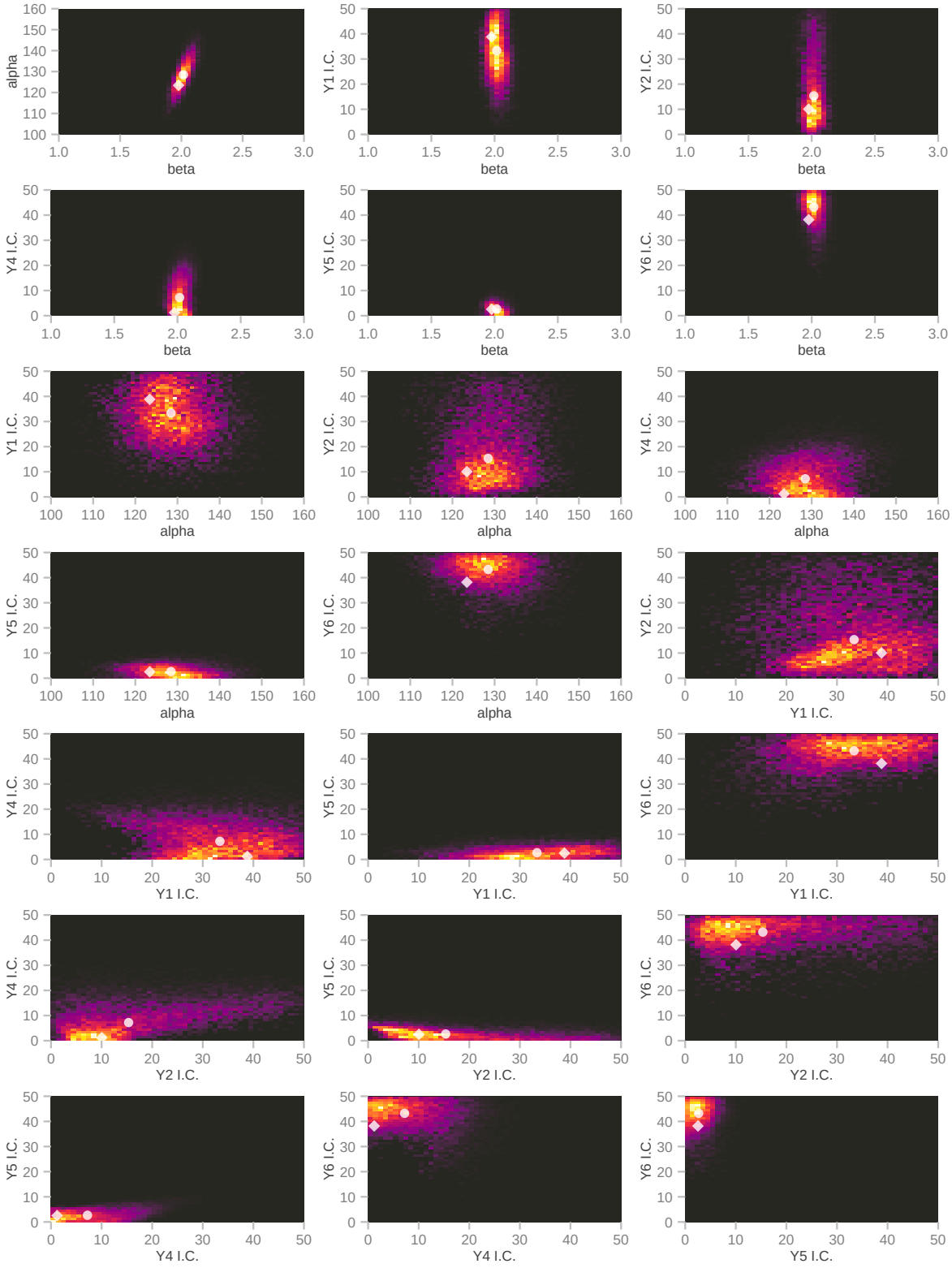


Figure 2.12: Estimated joint distributions of the repressilator example with missing initial conditions. Using the output of nested sampling we can also produce estimates of the joint distributions of pairs of parameters. The overall appearance of the posterior is roughly unimodal. Brighter colours indicate higher relative probability. Mean parameter value, white circle; best-fit likelihood parameter value, white diamond.

had little overall effect on the quality of the reproduced data. In Figure 2.11 and Figure 2.12 we show the estimated marginal and joint distributions for all parameters from this example. This enables us to see which parameters are more or less restricted and their correlations. The marginals are generally unimodal and the mean parameter values and best-fitting parameter set are similar. The joint distributions reveal that certain parameters are correlated, or somewhat disperse, but mostly they could be approximated by a Gaussian distribution. This is why the means and best-fitting parameter set are similar.

2.3.3 *Nested sampling for model comparison*

Initially in this section synthetic data is used to compare four coupled ODE models:

- the Lotka-Volterra model of population dynamics [173, 174]

$$\begin{cases} \dfrac{\mathrm{d}F}{\mathrm{d}t} = \alpha F - \beta FR, \\ \dfrac{\mathrm{d}R}{\mathrm{d}t} = -\gamma R + \delta FR, \end{cases} \tag{2.5}$$

- the repressilator system in Equation (2.4),
- the Goodwin model of protein-mRNA interactions [175, 176]

$$\begin{cases} \dfrac{\mathrm{d}M}{\mathrm{d}t} = \dfrac{1}{1+E} - \alpha, \\ \dfrac{\mathrm{d}E}{\mathrm{d}t} = M - \beta, \end{cases} \tag{2.6}$$

- the trimolecular two-species Schnakenberg model [177, 178]

$$\begin{cases} \dfrac{\mathrm{d}u}{\mathrm{d}t} = \alpha - u - u^2 v, \\ \dfrac{\mathrm{d}v}{\mathrm{d}t} = \beta - u^2 v. \end{cases} \tag{2.7}$$

The synthetic data was generated from one variable of the repressilator system with known parameters before Gaussian noise was added. To ease comparison between different systems the data were scaled so that the amplitude is maximally one. All models are mechanistically different, however as all models are capable of oscillatory solutions, any of them could be used to describe the chosen data set if no further information was available. Our task is to evaluate if,

and how well, we can choose between competing models given little data.

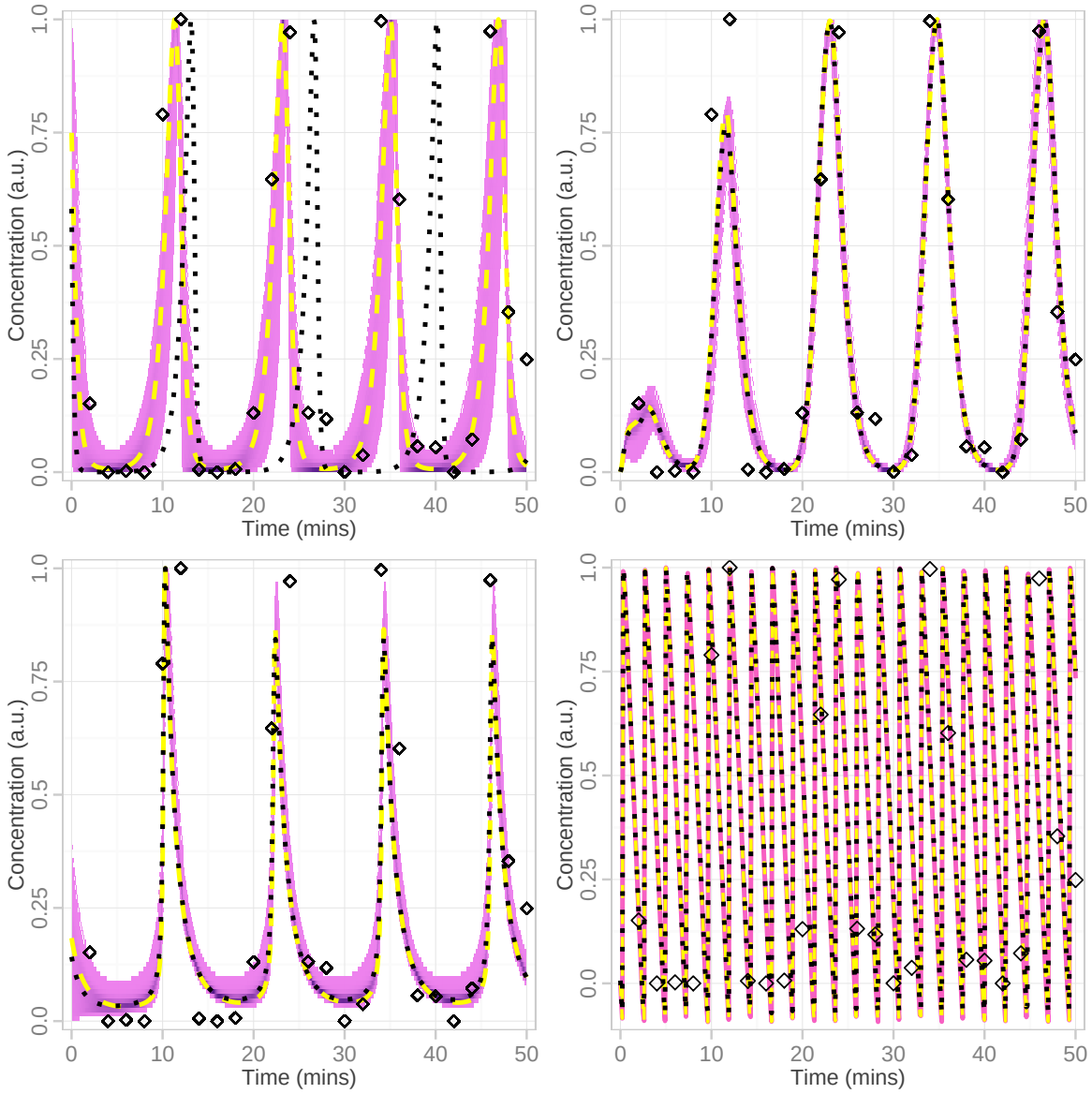


Figure 2.13: Fit to noisy data of four different oscillatory models. Clockwise from top left: Lotka-Volterra, repressilator, Goodwin and Schnakenberg models. Using the same noisy data (diamonds) 3000 equally-weighted samples (purple) were drawn from the posterior distribution of each model (except the Goodwin where we show a representative sample as all solutions were similar). The mean of the Lotka-Volterra system's posterior is not a good summary statistic for this distribution due to its non-unimodality (Figure 2.14 and Figure 2.15). The best-fit solution, dashed yellow line; solution using mean parameters, black dotted line.

Figure 2.13 shows 3000 samples from the posterior of all models (except Goodwin) along with the solution using mean parameter values and the dynamics of the best-fitting sample point from the four models. These summarising curves for the Goodwin model have much higher frequency than the others, yet can still give a good least-squares error. Note that the concentration falls below 0 for this model with these parameters, which is clearly unbiological. The other three models pick out the correct frequency in the data. The solution with the mean parameter values of the Lotka-Volterra system, Figure 2.13, is not a good summary statistic for this distribution though the best-fit likelihood line for this model in Figure 2.13 shows a good fit to the data. This indicates care should be taken when summarising distributions. However merely relying on the best fitting parameters is essentially a maximum likelihood approach, and may miss important contributions from other parts of parameter space. To visualise this, Figure 2.14 and Figure 2.15 show estimated marginal and joint distributions (with means and best-fit solution parameters indicated) for the Lotka-Volterra system which demonstrates the non-Gaussian shape of its posterior. The log-evidence values attained for the four models are shown in Table 2.6 indicating a very strong preference for the Lotka-Volterra model.

Model	$\log \mathcal{Z}$
Lotka-Volterra	-23.41 ± 0.10
Repressilator	-41.82 ± 0.13
Schnakenberg	-44.84 ± 0.14
Goodwin	-165.60 ± 0.12

Table 2.6: Log-evidence of the four models for noisy data. The log-evidence was computed by nested sampling for each model using the 25 noisy data points shown as diamonds in Figure 2.13. Using Jeffreys' scale for interpretation the data provide very strong evidence for the Lotka-Volterra model (Equation 2.5) and against the Goodwin model (Equation 2.6) compared with the other models. The repressilator (Equation 2.4) has positive evidence for it over the Schnakenberg model (Equation 2.7).

Given the nature of the sparse and noisy data it is not too surprising that a simpler model with two variables and six parameters is given preference over the model with six variables and eight parameters from which the data were actually generated. If the data are of better quality i.e. no noise and of greater density, one can see the repressilator model gaining more support in Figure 2.16 relative to the Lotka-Volterra system, but until an unreasonable amount of data is available (500 data points) the Lotka-Volterra model is preferred due to the it being the more parsimonious explanation of the data — visually both systems can fit the given data very well. Perhaps counter-intuitively, the evidence decreases with the increasing quantity of data. This is due to the log-likelihood function. As there are now more data points, unless the fit is exceptionally good, the least-squares residual increases due to summing up more errors. The evidence comprises both the Occam factor and the best fit likelihood

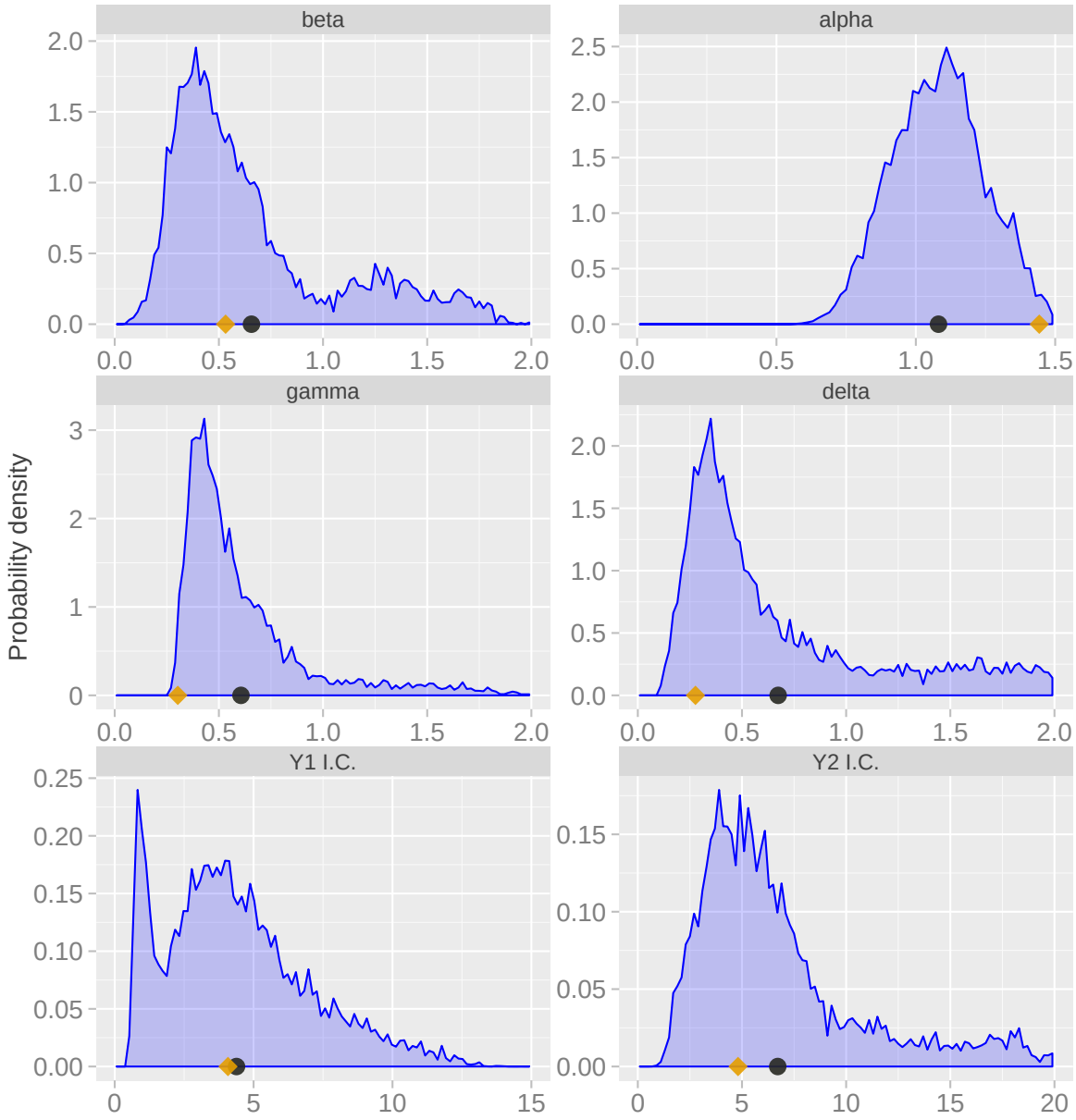


Figure 2.14: Estimated marginal distributions of the Lotka-Volterra system. 25 noisy data points generated from the repressilator system were used as the data for inference. Compared to Figure 2.11 parameters far away from the highest probability still have some probability and are accounted for in the Bayesian framework. Note the distribution for the first variable's initial condition (Y1 I.C.), which was inferred as a parameter, in asymmetrically bimodal. Mean parameter value, black circle; best-fit likelihood parameter value, orange diamond.

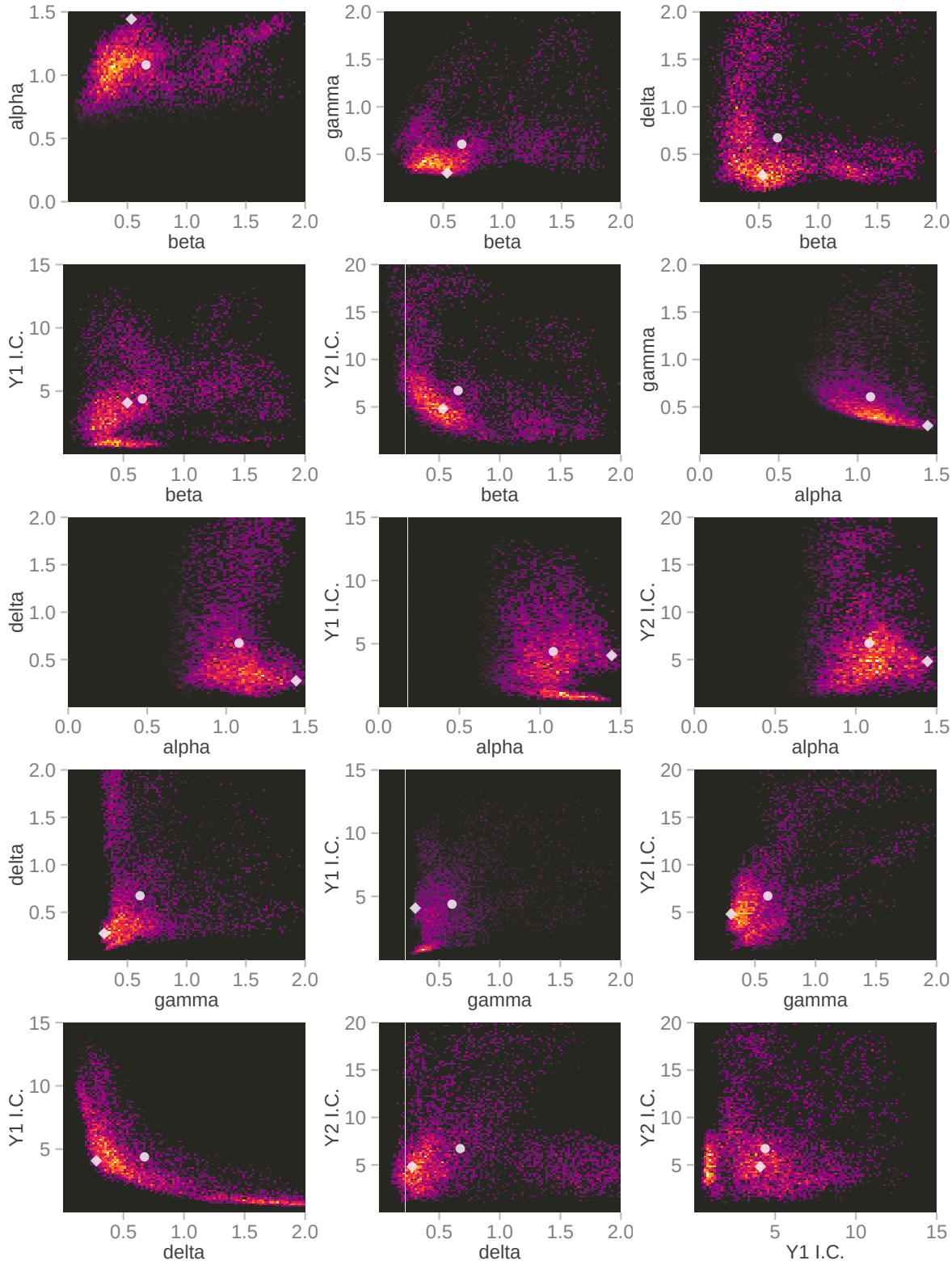


Figure 2.15: Estimated joint distributions of the Lotka-Volterra system. 25 noisy data points generated from the repressilator system were used as the data for inference. Compared to Figure 2.12 the joint probability landscapes are far less unimodal and provide an understanding of why the solution with these mean parameters shown in Figure 2.13 does not follow the same dynamics as the posterior samples. Brighter colours indicate higher relative probability. Mean parameter value, white circle; best-fit likelihood parameter value, white diamond.

(at least assuming the posterior is approximately Gaussian) [133]. Hence a worse likelihood score will similarly affect the evidence.

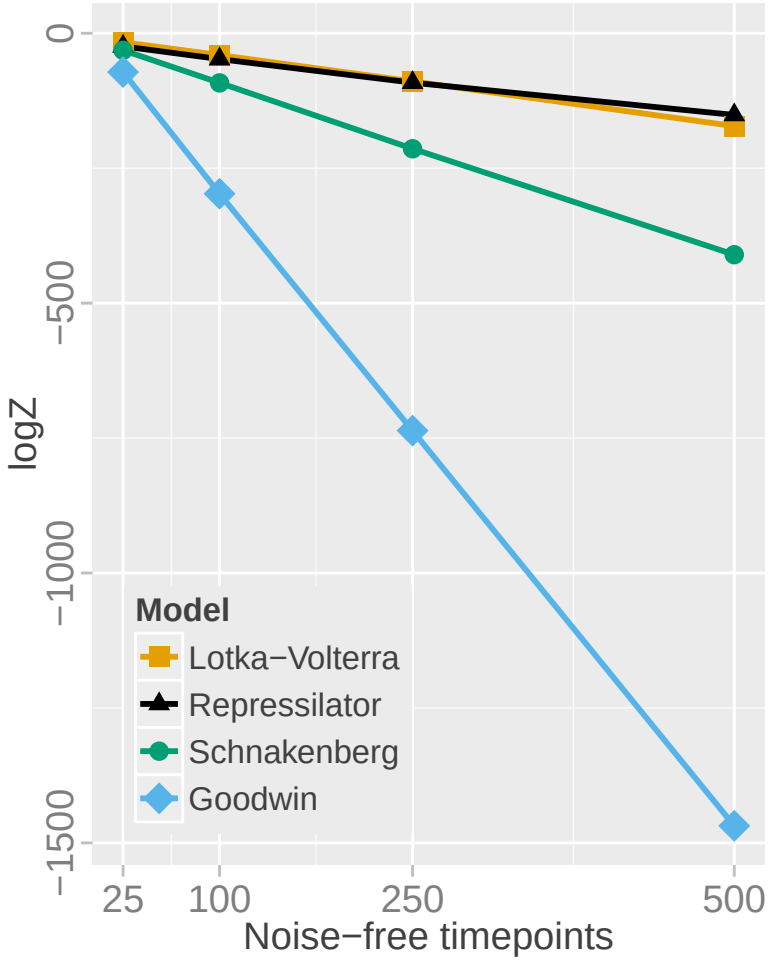


Figure 2.16: Evidence changes as a function of data quantity. As the resolution of the time-course improves the Goodwin model (skyblue, diamonds) and the Schnakenberg model (green, circles) lose support faster than the Lotka-Volterra (orange, squares) and repressilator (black, triangles) systems. The known model, the repressilator, gains preference only for a larger number of data points (500 points with a time gap of 0.1), even when using noiseless data.

During this test we normalised the amplitudes and assumed none of the initial conditions were known, whereas in practice they can be normally be measured or taken to be the first time point. With the initial condition included for the repressilator variable measured, cI protein (as in Figure 2.10), and with unnormalised amplitudes, the log-evidence improved to -34.27 ± 0.14 compared with -41.82 ± 0.13 without knowledge of the initial point.

The tests so far have all used synthetic data so that the inferences made can be compared to a reference with known parameters. Nevertheless some experimental data is available and can provide a more realistic situation in line with what typically faces a mathematical

modeller. Taking fluorescence data from the original repressilator paper [170] as a proxy for one of the variables in the system it was investigated whether this was sufficient to support the known model. The data were extracted from Figure 2C of the original work [170] and a linear increase in fluorescence equal to $(45/600) \times t$ was removed. As the data are in arbitrary units it was rescaled to be maximally one again and the algorithm was used on the four models as before. Table 2.7 shows the results which now give positive to strong evidence for the Schnakenberg model. The experimental data, solution with mean parameters and best-fit parameter's solution are plotted in Figure 2.17 which shows that although there is perhaps a fair fit in terms of residuals, in terms of the period of the data the posterior summary estimates are generally not at all close. If the frequency domain is known *a priori*, the likelihood function could be adjusted from a simple least-squares measure to take this into account. When posterior samples were plotted it was hard to gain anything visually thus for simplicity just the summary solutions are shown. Towards the second half of the experimental time series the repressilator's mean parameters' solution and best-fit solution do match the data more closely but this wasn't the case for all posterior samples.

Model	$\log \mathcal{Z}$
Schnakenberg	-101.66 ± 0.13
Repressilator	-104.85 ± 0.11
Lotka-Volterra	-124.19 ± 0.15
Goodwin	-166.70 ± 0.14

Table 2.7: Log-evidence of the four models for experimental repressilator data. The log-evidence was computed by nested sampling for each model using the 60 experimental data points given in the repressilator article [170]. The linear increase in fluorescence with time was removed and data rescaled to be maximally one. Using the interpretation on Jeffreys' scale the use of experimental data now provides positive to strong evidence for the Schnakenberg model against the repressilator and very strong evidence against the other two models.

If there was some uncertainty as to the model or its parameters, designing experiments that can maximise the information in the data is an approach that has been explored recently [179]. Experimentally it can be hard to increase the resolution of a timecourse so focusing on other genes or proteins of interest can be fruitful. With this in mind, and considering the results shown in Figure 2.16, the effect of gathering data from another variable of interest rather than trying to increase the quantity of data available from one variable was investigated. As previously the repressilator system Equation 2.4 was used to generate the timepoints, but now with two variables of 25 timepoints each and additional Gaussian noise. (The same random seed was used so as not to introduce this potential bias in generating the noise.) The four oscillatory models chosen before are used with nested sampling for model comparison. The results are presented in Table 2.8. There is now much stronger support, compared to just having data from one variable, for the repressilator model—the log-Bayes Factor has gone from 18 in favour of the Lotka-Volterra model

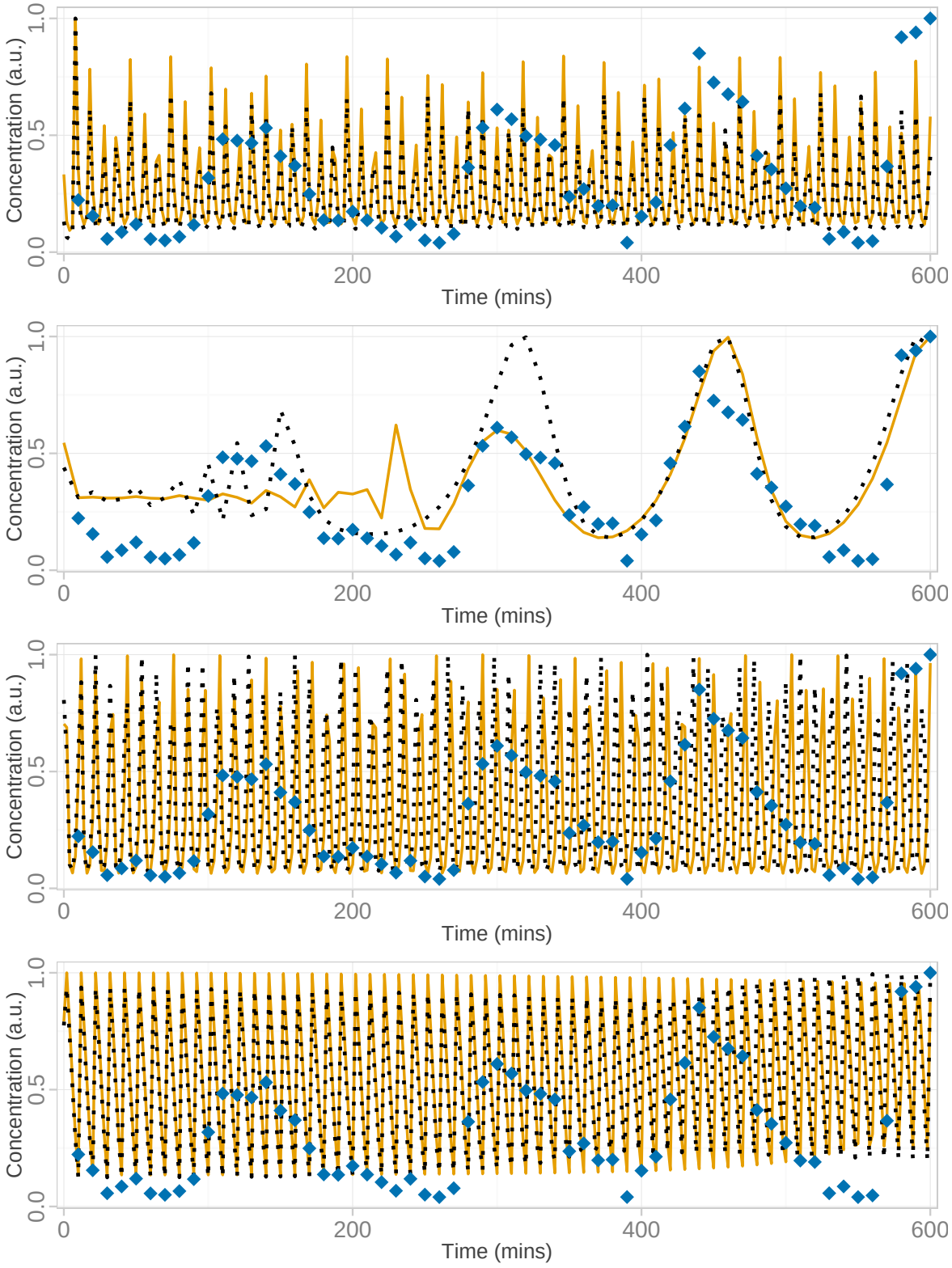


Figure 2.17: Mean and best-fit to the four models using experimental data. Experimental data points (blue diamonds) from the original repressilator paper [170] were used to compare the four models, which from top to bottom are the Schnakenberg, repressilator, Lotka-Volterra and Goodwin models. None of the models generally did well at identifying the correct period of the experimental oscillations. Solution using mean parameter values, black dotted lines. Solution using best-fit likelihood parameter value, orange solid lines.

over the other models to 72 in favour of the repressilator. This is regarded as decisive evidence for the repressilator on Jeffreys' scale. For these example models the use of data from two variables gives far more information than increasing the quantity of data from one variable and enables us to prefer the known model. We are thus able to suggest this interesting aspect should also be considered when designing experimental research, and may be very useful for Bayesian model comparison in helping to distinguish competing models of a biological process.

Model	$\log \mathcal{Z}$
Repressilator	-77.44 ± 0.14
Schnakenberg	-149.17 ± 0.14
Lotka-Volterra	-339.07 ± 0.12
Goodwin	-468.03 ± 0.12

Table 2.8: Log-evidence of the four models for noisy data from two variables. The log-evidence was computed by nested sampling for each model using 25 noisy data points from two repressilator variables. For these example models, it was found that the use of data from two variables gives more valuable information than an increase in the quantity of data from one variable. The data provide decisively strong evidence for the repressilator as judged on Jeffreys' scale.

2.4 Conclusion

Nested sampling is an effective way of calculating the evidence for a model and producing samples from the posterior distribution of the model's parameters. Nested sampling can be viewed as a Bayesian version of Monte Carlo for which initially the prior and then the likelihood are used to guide parameter space exploration. The 1D integral over the likelihood is solved by treating it as a sorting problem. As with other Bayesian approaches and in contrast to optimisation-based methods, samples are obtained from a full distribution of the parameters of interest rather than merely a point estimate for the parameter (and possibly an estimate of the variance depending on the method used). These posterior sample points can be used for further analysis such as marginal or joint distributions.

It was shown how the procedure produces samples from the posterior probability distribution of the parameters to compute the normalisation constant of the posterior. It has been demonstrated that nested sampling can produce good estimates for the parameters in systems of ordinary differential equations under typical biological scenarios of sparse, noisy data, where the data was only available from one out of six variables. Nested sampling was also shown to produce comparable parameter estimates to the established workhorse of Bayesian inference, namely MCMC, for a biological problem with experimental gene expression data.

Using Bayes' theorem for inference additionally helps reduce overfitting. In our examples the plasticity of the posterior of the Lotka-Volterra model meant that the single variable data set available was not sufficient to give preference to the repressilator model that the

data were generated from. However when data were introduced from another variable this was able to constrict the parameter space further to then convincingly give a Bayes factor in favour of the repressilator. As the mechanisms of these two models are quite different, the modeller may have background knowledge to prefer one system over another and certainly Bayes factors or any other metric for model comparison should not replace intelligent reasoning about the problem being studied.

Nevertheless a remaining problem is the reliability of model comparison in light of limited data for realistic problems where the ideal model is unknown. This difficulty could be addressed by the construction of other models, some which could, and some that could not, possibly describe the system under study accurately. Performing Bayesian model comparison assuming equal belief in all models in this case could reveal whether the target model really can provide a justifiably better fit to the data than a number of other models. If a simpler model was found that ranks higher then this would reveal the limitation in the present data set, and that adding more parameters is not justified yet until better data are available.

In our example using experimental repressilator data we could use this as a first run to clearly identify that the models are oscillating at the frequency of the data collection timepoints rather than the frequency of the fluorescence. This then could be seen as a calibration run that informs the next round of inference where we incorporate this result into our background knowledge and update our likelihood function and/or prior parameter distributions accordingly. A true Bayesian would take their prior belief for each model into account along with the evidence values for each before making a conclusion based on the posterior odds. It may be found that the evidence for one model is sufficient to overcome a strong prior belief on the model space. The data available limits the accuracy of the inferences we can make but provides the odds for hedging our bets — it does not mean that the "wrong" model can't be preferred. Alternatively the predictions from the competing models could be used as a way to distinguish between them in light of a validation data set or further data acquisition.

If we have only a small number of models we wish to evaluate, the approach of separating each model to provide an individual prediction that can be used to guide experimental validation is tractable. Bayes factors can be used to compare and select amongst models. For prediction purposes, however, the full hypothesis space is of interest to take into account parameter and model uncertainty. Model averaging is thus an important concept that provides a canopy above the layers of parameter and model inference [133, 180]. In terms of the least biased prediction, multimodel inference is therefore the approach of choice [133, 146, 180, 181]. After the new data arrive, these can be used to update the probability distributions over each model's parameter space and furthermore to then update the probabilities of the models themselves by computing the posterior distribution over model space.

Nested sampling has the advantage of calculating the evidence as its main focus, thus readily providing us with the quantity required for model comparison. For systems biologists this ability to achieve both parameter inference and model comparison with the same algorithm is clearly applicable to many current challenges in the field.

Proper statistical treatment of a biological modelling problem can be achieved with Bayesian inference and nested sampling in particular. Thus a bright future beckons for systems biologists wishing to quantify uncertainty, infer parameters and compare models.

3 MODELLING THE FLORAL TRANSITION

3.1 Introduction

The initiation of the floral transition is a key decision during plant development. Diverse species have evolved to respond to environmental cues to flower in the correct season. Despite these specific differences key properties such as irreversibility and robustness to fluctuating signals appear to be conserved in individual meristems of monocarpic annuals. In Arabidopsis many genes have been discovered and placed in regulatory networks without considering the specific temporal or spatial effect this could have on the network properties. This motivates a dynamic model of known regulators to help us gain an understanding of how genetic interactions can lead to morphological change over time. In this chapter we will use mathematical modelling to understand the essential properties of the transition. Our models will be placed on quantitative foundations with the aim of capturing some of the underlying biology in a developing plant system. The models will be trained using leaf number data from various genotypes perturbed for key genes that regulate the vegetative to floral transition and predictions made for others. The use of linear models will be explored and then a demonstration of how small regulatory networks of core components are sufficient to capture the dynamic behaviour of the floral transition. The mathematical assumptions and simplifications will be clearly stated and the models described in detail. The value of pursuing an iterative approach combining modelling with experimental work to capture key features of complex systems will be highlighted. Nested sampling, as described in chapter 2, will be used to place the models in a Bayesian context thus giving robust statistical treatment of the problems being addressed. Nested sampling will be used for parameter inference and for comparison of models of the floral transition developed in this chapter. Using the Bayesian framework allows further information to be used as it becomes available and possible avenues for taking

this work forward are discussed along with a detailed critique of the models developed.

3.2 Initial considerations

3.2.1 *Hubs*

In Arabidopsis many genetic studies have revealed major components of gene regulatory networks. From these experiments it has been found that the entire network of flowering time control is highly complex with many interacting signals and pathways [42, 43]. This complexity coupled with a lack of kinetic parameter data can lead to models that are highly underconstrained by the available biological data. Thus an appropriate way of tackling such a large system is to consider a reductionist approach. The motivation for reducing the complexity is that it is easier to work with a reduced number of variables and parameters but can still provide insights into the biology. One way of reducing such a large system is to take advantage of some genetic redundancy in the plant and approximate key genes for entire hub activities. This is not an uncommon approach in the literature when tackling redundancy (for example see van Mourik et al. [106]). Therefore an initial simplification for this chapter is to group genes with similar effects into distinct hubs or functional modules [109].

By doing this early in our analysis we will lose a direct mapping onto individual genes yet, as we have phenotypic data available for transgenic or mutant lines perturbed for the major genes controlling the hubs, we can still relate back to biological units. Simplifying the known flowering regulatory network as a set of key hubs has the advantage of making it potentially easier to identify the critical network interactions that account for the major behaviours of the system. As an example, mutations in both *FD* and its paralogue *FDP* have been shown to cause later flowering phenotypes than either individually and the double mutant significantly reduces the effect of *FT* overexpression [166] (see also Table 3.1). These two closely related genes can therefore be grouped together in to a single hub, named after its major contributor, FD. Another example would be the AP1 hub that would also include CAL, which is partially redundant with AP1 [182]. Double mutants in these genes have curious

cauliflower-like inflorescences [65, 183]. The FT hub would also include the activity of its close relative TSF [54]. As discussed in section 1.3 LFY is a master integrator that functions in multiple pathways so it makes sense that this will provide a hub. Finally we include a hub that can antagonise the floral transition and promote meristem identity. This is named after TFL1 — a key floral repressor. In total the whole flowering regulation network is simplified down into five hubs labelled after their major constituents: FT, TFL1, FD, LFY and AP1. The AP1 hub is used as the output in the modelling work presented here as its upregulation is known to be an early indicator of the floral transition in Arabidopsis [50, 184].

3.2.2 *Data*

The quantitative data used in this study were provided by Phil Wigge and Katja Jaeger who grew 16 plant genotypes and measured their rosette and cauline leaf numbers. These data are provided in Table 3.1 including various mutant and overexpressing transgenic lines under the control of the cauliflower mosaic virus 35S promoter (35S). To validate the models and predictions in terms of leaf numbers this data set was divided into a training set and a prediction set. Although the data are not independent, because a number of combinations of genes are present or absent in multiple genotypes, an appropriate way to assign to these subsets was to predict the triple mutant leaf numbers. This would also be a significant and challenging test of the models' predictive capacity because of the unknown combinatorial effects of the gene interactions.

3.2.3 *Linear modelling*

Initially we consider a QTL-type approach for capturing the total leaf number data in terms of a number of genes involved in the floral transition. In QTL studies a single-marker analysis is the simplest conceptual method of detecting a QTL and can be conducted with statistical tests such as ANOVA or a t-test depending on how the population was crossed. Performing a linear regression is also a simple and common way of analysing traits of interest. This method was chosen for the present study because we don't have a mapping population and data set as would be considered in a traditional QTL

Genotype	No. of plants	Number of rosette leaves	Number of cauline leaves	Total leaves	SD of total	Data set
Wildtype (Col)	12	7.9	1.4	9.3	1.1	Training
35S:*FT*	10	4.4	1.0	5.4	0.7	Training
35S:*LFY*	11	3.8	1.8	5.6	0.8	Training
35S:*TFL1*	12	27.5	15.7	43.2	1.9	Training
lfy-12	9	13.0	5.3	18.3	1.2	Training
ft-10	10	36.4	9.3	45.7	1.3	Training
tfl1-1	11	7.7	0.4	8.1	0.8	Training
fd-2	12	18.5	4.63	23.13	2.47	Training
fdp-1	10	11.2	2.0	13.2	1.3	Training
fd-2 fdp-1	10	32.9	6.3	39.2	1.1	Training
35S:*TFL1 fd-2*	12	23.8	8.2	32.0	2.1	Training
tfl1-1 fd-2	12	14.4	4.6	19.0	1.2	Training
35S:*FT fd-2*	12	8.3	2.4	10.7	1.35	Training
tfl1-1 fd-2 fdp-1	12	24.83	6.67	31.5	1.38	Prediction
35S:*TFL1 fd-2 fdp-1*	10	31.33	11.0	42.33	2.89	Prediction
35S:*FT fd-2 fdp-1*	12	25.8	5.6	31.4	1.34	Prediction

Table 3.1: Experimental leaf number data. For each genotype the table lists the mean of the experimental data for rosette and cauline leaves, total leaf number (TLN) and the calculated standard deviation (SD) of the TLN. The wildtype and all single and double mutant data comprised the model training set for parameter inference. The triple mutant data are predicted using the inferred parameters from the training phase.

analysis. Thus a simple linear summation of relative estimated gene expression levels was utilised. This linear model takes a combination of the concentrations (where necessary denoted with square brackets) of FT, TFL1, FD and LFY, plus an intercept term which would represent the population mean in a QTL study. AP1 is not included as we use this as a proxy for the leaf number data which is the output of the model (discussed in more detail in subsection 3.3.1) and *AP1* mutant genotypes were not available. Thus the linear model can simply be stated as

$$\text{Total leaf numbers} = k_1 + k_2[FT] + k_3[TFL1] + k_4[FD] + k_5[LFY]$$

which is a five parameter linear regression problem with k_i being the constants to estimate. Genotypes were assigned weights for each gene depending on the contribution of mutating or overexpressing

the genes. The wildtype gene was assigned a value of 1; full knockouts were assigned 0; 35S overexpressors were assigned 10 to reflect the massive ectopic expression; and the *fd-2* and *fdp-1* mutations had levels of 0.25 and 0.8 respectively to approximately reflect their *in planta* partial knockout effect. These values were chosen because the *fd-2* mutant clearly has a stronger effect on flowering time than *fdp-1* (Table 3.1) but in combination are far more potent. Thus full FD hub mutants (in the models in this thesis represented by the data from *fd-2 fdp-1*) have a total weight of 0.05. It is not expected that these choices are particularly important due to the number of free parameters in the model that can adjust for any differences in these values.

Nested sampling was used to estimate the parameters and evidence of the linear model initially by using the known standard deviation in the total leaf numbers to place a different normal distribution on each data point. The log-evidence was worse than −11000 strongly indicating that the variation in the leaf number data could not be captured by this model. Indeed, plotting the posterior mean and standard deviations of the leaf numbers for each genotype against the true leaf numbers reveals that many genotypes are poorly estimated, as shown in Figure 3.1. If the model fitted perfectly the points would all fall on to the dark line. The small estimated standard deviations suggest the mean parameter values are the ones which maximise the likelihood function with little room for variation in those values. Note that, without further constraints, the *35S:FT* genotype is estimated to have negative leaf numbers which is obviously not biologically possible.

To try to alleviate the issues with the constraints on the data an extra parameter for the error term in the data set was added. Thus now all data points share the same standard deviation term which follows a Jeffreys-type prior (as the standard deviation is a scale parameter) [138], rather than their associated individual standard deviations. This results in a log-evidence of -69.85 ± 0.13. An equivalent plot as before is shown in Figure 3.2. This plot was constructed by taking the 1000 highest likelihood samples from the posterior set to reduce the influence of the inferred variance parameter affecting the likelihood. In other words by doing this it is expected that a good

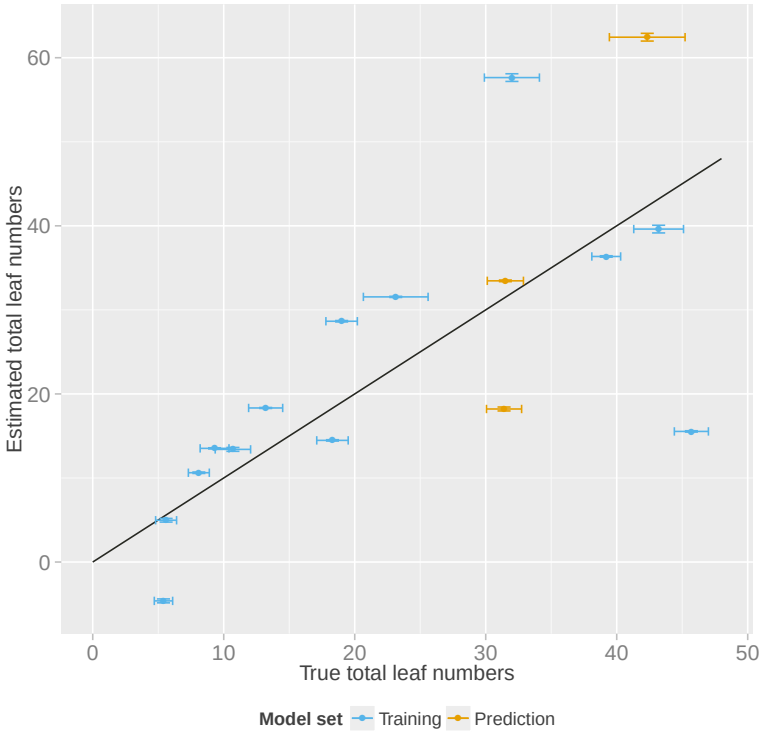


Figure 3.1: Fit of a linear model with independent errors in the data of each genotype. The general trend may be broadly correct yet this model is not particularly accurate. The tiny error bars (representing one standard deviation) of the estimated leaf numbers indicate a lack of flexibility in the parameters that maximise the likelihood function. One genotype, *35S:FT*, was estimated to have negative leaf numbers, but this could be controlled for with further constraints.

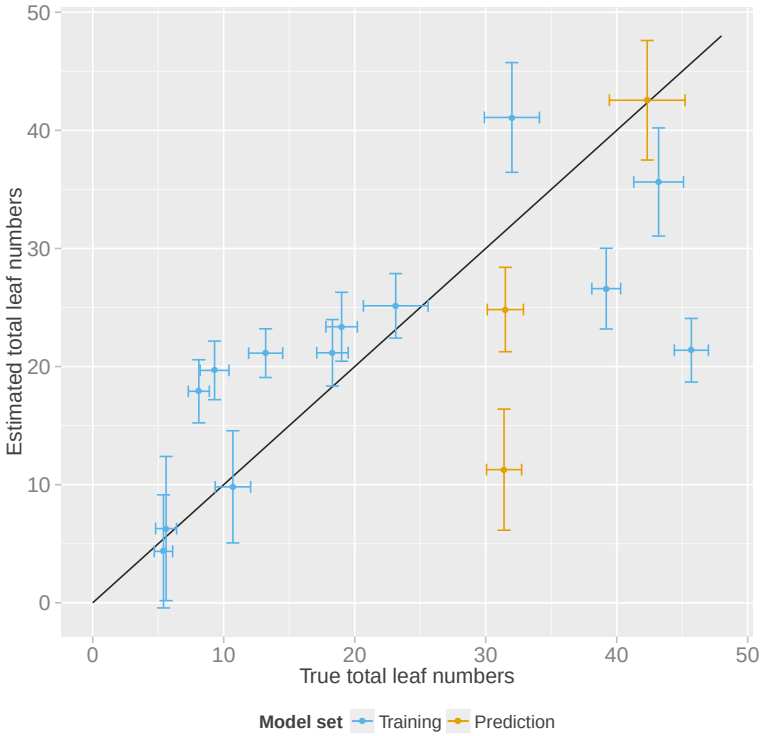


Figure 3.2: Fit of a linear model with the error model variance inferred as a parameter and individual standard deviations not considered. The estimates and predictions are slightly better than in Figure 3.1. Our error bars are much more in line with the experimental error bars than previously. The predicted mean estimate of *35S:TFL1 fd-2 fdp-1* is almost exactly correct.

likelihood fit is more likely due to the parameters of interest themselves rather than the parameter controlling the variance in the error model. The standard deviations of the resulting estimated leaf numbers are now on a similar level to the true data in a number of cases. Our predicted triple mutants are improved overall with *35S:TFL1 fd-2 fdp-1* being remarkably close to the experimental result although *35S:FT fd-2 fdp-1* is predicted to have 20 leaves fewer than in reality.

Running the linear model in this way is identical to calculating a least-squares regression. These regression statistics were confirmed using R's [185] linear model function `lm()` which produced similar estimates for the parameters. As a side note the computed Adjusted R-squared value suggests we explain 29% of the variability in the data with this linear model. The estimated against true total leaf numbers for all models considered here are shown in Table 3.2.

Taken together these results suggest that for this small system and for the given data a linear modelling approach is limited in its accuracy. Specifically it was shown that a linear model gives estimates that are too inaccurate to capture the underlying biology without further constraints. QTL-type linear modelling has a value due to its simplicity especially for genome-wide data and where there are a large number of lines available to be scored. For this data set and simplified hub model however, this approach does not appear flexible enough to be appropriate. Thus in the rest of this chapter a more detailed mechanistic approach is constructed based on an ODE model that can provide better leaf number predictions, trace the dynamics of a developing system, capture key properties of the floral transition and lead to interesting predictions that can be tested experimentally. However, this flexibility requires far more parameters and a much greater investment in computer time.

3.2.4 *Simple networks*

Introduction

For a number of species, homologues of the Arabidopsis master regulator FT are a core element of the photoperiod pathway [43, 44]. In Arabidopsis diurnal *CO* activity gives rise, in long days, to stable CO which upregulates *FT* [17, 45]. Long range signalling of FT promotes flowering time [47–49], thus oscillating input signals are interpreted

Genotype	Mean total leaves				SD of total leaves			Data set
	True	Ind. σ	Inferred σ	R `lm()`	True	Ind. σ	Inferred σ	
Wildtype (Col)	9.3	13.5	19.7	19.8	1.1	0.07	2.5	Training
35S:*FT*	5.4	-4.6	4.4	4.1	0.7	0.23	4.8	Training
35S:*LFY*	5.6	5.0	6.3	5.9	0.8	0.22	6.1	Training
35S:*TFL1*	43.2	39.6	35.6	35.7	1.9	0.46	4.6	Training
lfy-12	18.3	14.5	21.2	21.3	1.2	0.08	2.8	Training
ft-10	45.7	15.5	21.4	21.5	1.3	0.08	2.7	Training
tfl1-1	8.1	10.6	17.9	18.0	0.8	0.09	2.7	Training
fd-2	23.13	31.5	25.1	25.1	2.47	0.07	2.7	Training
fdp-1	13.2	18.3	21.1	21.2	1.3	0.06	2.1	Training
fd-2 fdp-1	39.2	36.4	26.6	26.5	1.1	0.09	3.4	Training
35S:*TFL1 fd-2*	32.0	57.6	41.1	41.0	2.1	0.46	4.6	Training
tfl1-1 fd-2	19.0	28.6	23.4	23.3	1.2	0.08	2.9	Training
35S:*FT fd-2*	10.7	13.4	9.8	9.4	1.35	0.23	4.7	Training
tfl1-1 fd-2 fdp-1	31.5	33.5	24.8	24.7	1.38	0.10	3.6	Prediction
35S:*TFL1 fd-2 fdp-1*	42.33	62.4	42.5	42.5	2.89	0.46	5.1	Prediction
35S:*FT fd-2 fdp-1*	31.4	18.2	11.3	10.8	1.34	0.23	5.1	Prediction

Table 3.2: Experimental and linear model leaf numbers. For each genotype the table lists the mean true experimental total leaf numbers and standard deviations (SD) together with those estimated (for the training set) or predicted using the linear models described in the text. R's `lm()` function does not produce an uncertainty estimate. Ind., Independent i.e. the case where each genotype's SD was used in the likelihood function.

at some level and, once in a floral primordium, the cells are committed to transition. This section presents a simple demonstration of how two of the important properties of the floral transition, namely noise-filtering and irreversibility, can be exhibited by simple three node networks in feedforward loops. The nodes consist of the complex FT with FD, and the floral transcription factors LFY and AP1. Although labelled for the Arabidopsis genes, the qualitative effects of the motifs apply equally well to other species.

A set of ODEs is used to describe the dynamic behaviour of the system and numerically solved. Binary step functions are used for the transcriptional activation of genes and AND, OR or AND/OR gating, depending on the network. The equations all follow the standard form of

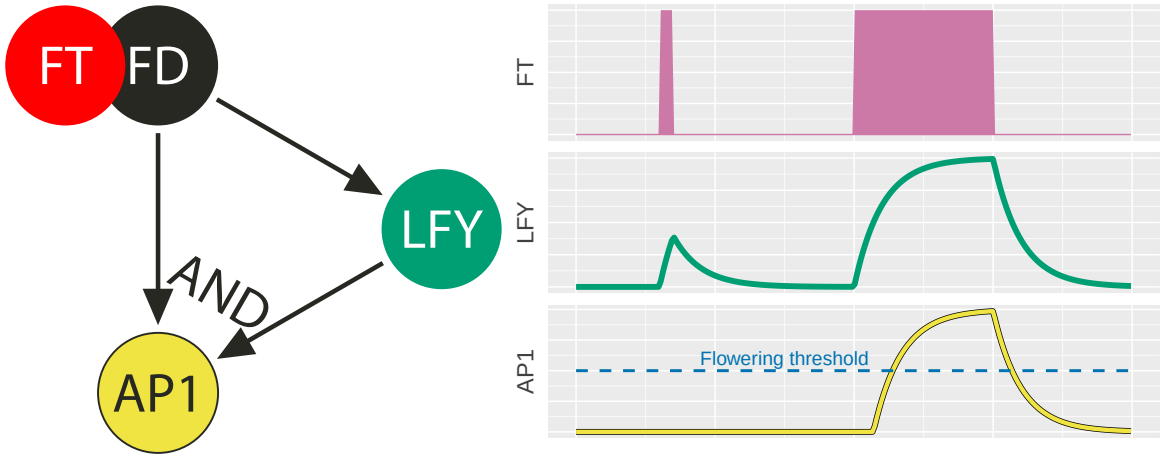
$$\frac{dx}{dt} = v - \delta x, \text{ for } x \in \{LFY, AP1\},$$

where v is the transcription term and δ the degradation rate. The FT input signal to the system in this section was modelled as a digital function, either 0 or 1. In these examples a signal of long duration and a small blip are used which might represent, for example, a one-off short exposure to sunlight. All initial conditions are set to 0 and parameters fixed.

The Coherent Feedforward Loop

The coherent feedforward loop is a network motif that is commonly found in signalling networks [109, 186]. One node regulates another, with both jointly regulating a third, Figure 3.3a. If the joint regulation is with an AND logic gate this simple network has persistence detection and thus is able to be used as a noise filter that removes blips in a signal. As the correct timing of the floral transition is crucial to reproductive success it is important that the system integrates information over time, and is not incorrectly activated by noise [187].

In the equations below we write $\theta_{FT.LFY}(FT)$ (where θ represents a Heaviside step-function) to mean that when FT is greater than (or equal to) the threshold at which it binds to the promoter site of *LFY*, FT activates *LFY* transcription. Similarly $\theta_{FT.AP1}(FT)$ means *AP1* is activated when FT is greater than or equal to the *AP1* promoter-binding threshold. The threshold for the activation of *LFY* and *AP1* is set at FT = 1. $\theta_{LFY.AP1}(LFY)$ means that when LFY reaches a



(a) This motif is very common in *Escherichia coli*, yeast and other organisms [186]. It has been referred to as a sign-sensitive delay element because it has a delayed on response but immediate off response [109].

(b) Responses of LFY (green) and AP1 (yellow) to a short and a long incoming FT signal (purple) are shown. The short pulse in filtered out and the longer signal leads to (delayed) activation of AP1. Decay in response to a fall in FT is immediate.

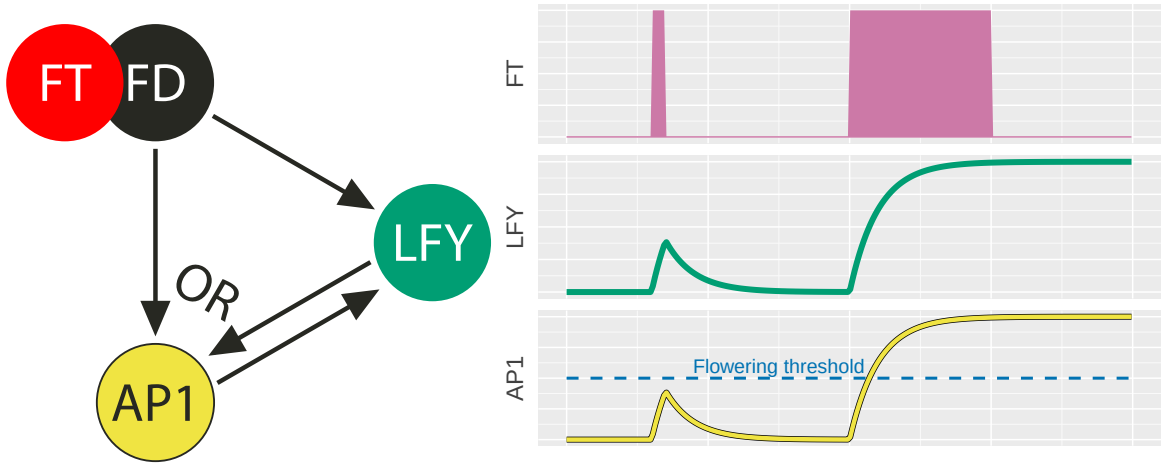
Figure 3.3: The coherent feedforward loop and its dynamics.

threshold, here 0.5, it binds the *AP1* promoter and thus activates *AP1* transcription. The activation constants, v, and degradation constants, δ, were set to 1 to scale the results to be maximally 1. The equations for this system are therefore as follows

$$\frac{\mathrm{d}LFY}{\mathrm{d}t} = v_{LFY}\theta_{FT.LFY}(FT) - \delta_{LFY}LFY,$$

$$\frac{\mathrm{d}AP1}{\mathrm{d}t} = v_{AP1}\theta_{FT.AP1}(FT)\theta_{LFY.AP1}\,(LFY) - \delta_{AP1}AP1.$$

The dynamics of this network are shown in Figure 3.3b. This network motif has been described previously and has been shown to exhibit noise filtering properties for short bursts of the incoming signal that are below the delay time through the different routes in the pathway [109, 186]. This is clearly seen in the figure where the short pulse of signal is filtered out whereas the longer signal is transferred through the network. By introducing an arbitrary threshold for flowering, seen in the lower AP1 panel of Figure 3.3b, this network shows a reversion to a non-flowering state after FT decay.



(a) This motif is often found in developmental transcription networks [186]. The feedback between the two targets of the first node can cause a memory effect as they keep active even in the absence of a signal from the initial regulator [109].

(b) Responses of LFY (green) and AP1 (yellow) to a short and a long incoming FT signal (purple) are shown. The short blip leads to a small rise in LFY and AP1 levels but not enough to be maintained. The long signal induces both targets enough so that when FT drops away the double-positive feedback loop maintains both LFY and AP1.

Figure 3.4: The regulated feedforward loop and its dynamics.

The Regulated Feedforward Loop

A similar three node network, called regulated feedforward or double-positive feedback [109], that uses an OR logic gating can exhibit irreversibility. With the same nodes, an extra activating connection between the two targets of the first complex will mean the targets are mutually activating given enough initial impulse by the first (Figure 3.4a). If these conditions are met the network can provide memory of the input signal. This is important in developmental networks because they operate on slower timescales than sensory networks. For example commitment to flowering after exposure to long days has been shown to take 1–7 days depending on plant age and seed vernalization treatment [46] whereas a sensory response in the shoot to salt stress in the root has been shown to take in the order of minutes, propagated in part by a rapid calcium wave [188].

The notation of the equations for this loop is identical to that for the coherent feedforward loop. There is an additional connection in this network as shown in Figure 3.4a. This is controlled by the term

$\theta_{AP1.LFY}(AP1)$ which means that when AP1 is greater than or equal to the threshold at which it binds to the promoter site of *LFY*, AP1 activates *LFY* transcription. Both thresholds between LFY binding *AP1* and vice versa were set at values of 0.45. This creates the double-positive feedback loop [109]. Due to using OR logic gating, rather than AND logic, the equations take the higher of whichever activator is bound, leading to

$$\frac{\mathrm{d}LFY}{\mathrm{d}t} = v_{LFY} \max\left(\theta_{FT.LFY}(FT), \theta_{AP1.LFY}(AP1)\right) - \delta_{LFY} LFY,$$

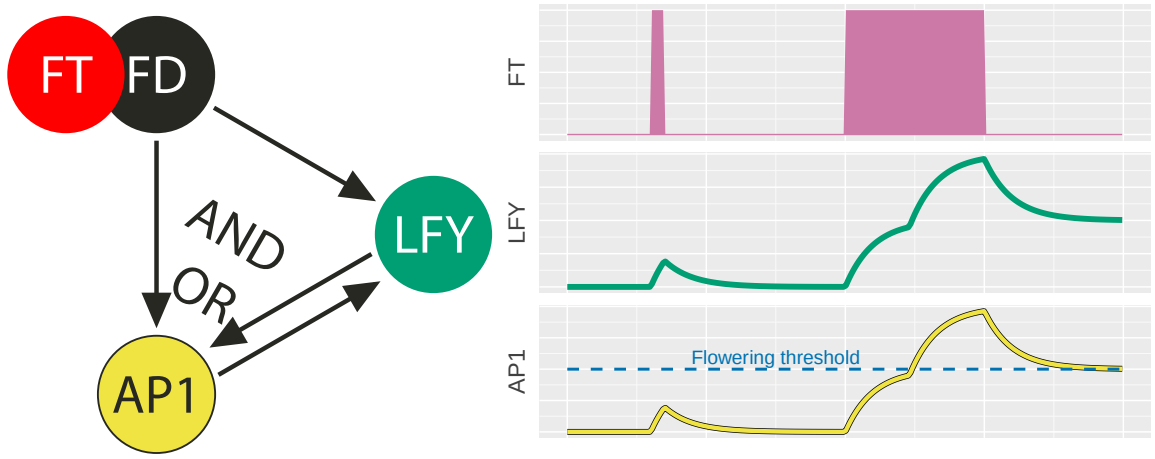
$$\frac{\mathrm{d}AP1}{\mathrm{d}t} = v_{AP1} \max\left(\theta_{FT.AP1}(FT), \theta_{LFY.AP1}(LFY)\right) - \delta_{AP1} AP1.$$

The dynamics of this motif are shown in Figure 3.4b. Once LFY reaches a concentration level that can activate *AP1*, this interaction is sufficient to maintain AP1 production even in the absence of the incoming signal FT. If the FT signal is removed before LFY has accumulated to a sufficient level then AP1 will degrade away before floral commitment. The network therefore shows a memory effect and irreversibility [109]. With the arbitrary threshold for flowering, seen in the lower AP1 panel of Figure 3.4b, the double-positive feedback shows maintenance of a flowering state even after FT decay.

A Compromise Feedforward Loop

While the previous two simple network motifs capture separate characteristics of the floral transition, in order to reproduce both noise filtering and irreversibility within the same network the logic gating rules require an extension. This can be achieved by introducing two transcription rates, a low rate that can be activated by either FT or LFY and a higher rate that requires the presence of both FT and LFY (Figure 3.5a) [7]. Hence we combine the key features of both previous network motifs by using two different levels of activation depending on the number of activators bound. The logic gating uses OR for transcriptional activation at a reduced level but requires AND for maximal activation. The higher levels, $v_{LFY,1}$ and $v_{AP1,1}$, are set to 1, and the lower levels, $v_{LFY,2}$ and $v_{AP1,2}$, are set to 0.5. These ideas lead

[7] *Espinosa-Soto et al. [104] have used a similar idea. An intermediate level of expression is determined from experimental data for a number of nodes. This, as here, allows for different activation thresholds.*



(a) This network incorporates the two previous motifs by utilising different transcription rates. A low rate requires the presence of only one signal employing an OR gate. A higher rate of transcription can be achieved by both regulators being present so using an AND gate. We get some noise filtering and a partial memory effect.

(b) Responses of LFY (green) and AP1 (yellow) to a short and a long incoming FT signal (purple) are shown. The blip in FT causes a damped response by both genes. Past a threshold during the longer signal the AND gate will be switched on enabling a higher rate of transcription. As the signal abates the transcription rate is reduced to the lower level but this is still at or above an introduced threshold for flowering.

Figure 3.5: A compromise feedforward loop and its dynamics.

to the following equations:

$$\frac{\mathrm{d}LFY}{\mathrm{d}t} = \begin{cases} v_{LFY,1} - \delta_{LFY} LFY & \text{if } \theta_{FT.LFY}(FT) = 1 \text{ and} \\ & \theta_{AP1.LFY}(AP1) = 1 \\ v_{LFY,2} \max\left(\theta_{FT.LFY}(FT), \theta_{AP1.LFY}(AP1)\right) & \\ \qquad - \delta_{LFY} LFY & \text{otherwise,} \end{cases}$$

$$\frac{\mathrm{d}AP1}{\mathrm{d}t} = \begin{cases} v_{AP1,1} - \delta_{AP1} AP1 & \text{if } \theta_{FT.AP1}(FT) = 1 \text{ and} \\ & \theta_{LFY.AP1}(LFY) = 1 \\ v_{AP1,2} \max\left(\theta_{FT.AP1}(FT), \theta_{LFY.AP1}(LFY)\right) & \\ \qquad - \delta_{AP1} AP1 & \text{otherwise.} \end{cases}$$

This gives rise to compromised characteristics for the individual properties but it is possible to capture some level of robustness to noise and a partial memory effect. By introducing a flowering threshold for AP1, depending on the threshold choice and parameters of the model, we can achieve irreversibility. Thus there is sufficient memory for the system to continue to flower as shown in Figure 3.5b.

Summary

A reduced network that represents the core structure underlying the biology can be mapped to the simple feedforward loops discussed above. As a major floral pathway integrator an active FT/FD complex was placed at the start of the transcriptional feedforward loop, upregulating another integrator, LFY, which both activate the floral initiator AP1 [63]. AP1 also mutually activates LFY in a positive feedback loop [64, 79] thus creating the important memory element which is responsible for irreversibility of a plant committed to flowering. As this network still contains the coherent feedforward loop motif it is also able to filter out some degree of noisy endogenous or exogenous input signals, which can be relevant to a plant in its natural environment.

In this section it was sought to show a simple regulatory network that can capture two major properties of the floral transition. This can be incorporated into a more complete model of the floral transition in Arabidopsis with our hubs and more realistic Hill-type functions from Michaelis-Menten kinetics and including the activity of the floral repressor *TFL1* in a repressing hub. This model can then be used to predict leaf numbers and generate hypotheses whilst having at its core a network that demonstrates key properties of the floral transition.

3.3 Methods

3.3.1 *Leaf numbers can be used to scale the network*

As mentioned earlier *AP1* levels have been shown to be a marker for floral commitment as *AP1* is detected in early floral primordia around stages 1 and 2[184, 189]. Thus in our model, the AP1 hub is chosen as the output of the floral induction pathways, and rising levels of AP1 correlate with progress through the floral transition.

After germination the first true leaves of Arabidopsis are termed rosette leaves which continue initiating during the vegetative growth phase. We map the number of rosette leaves to the initial state and low levels of our AP1 hub. Commencing the floral transition, lateral organs formed on the side of the bolting main shoot are termed cauline leaves. Once the transition is complete, flowers are made.

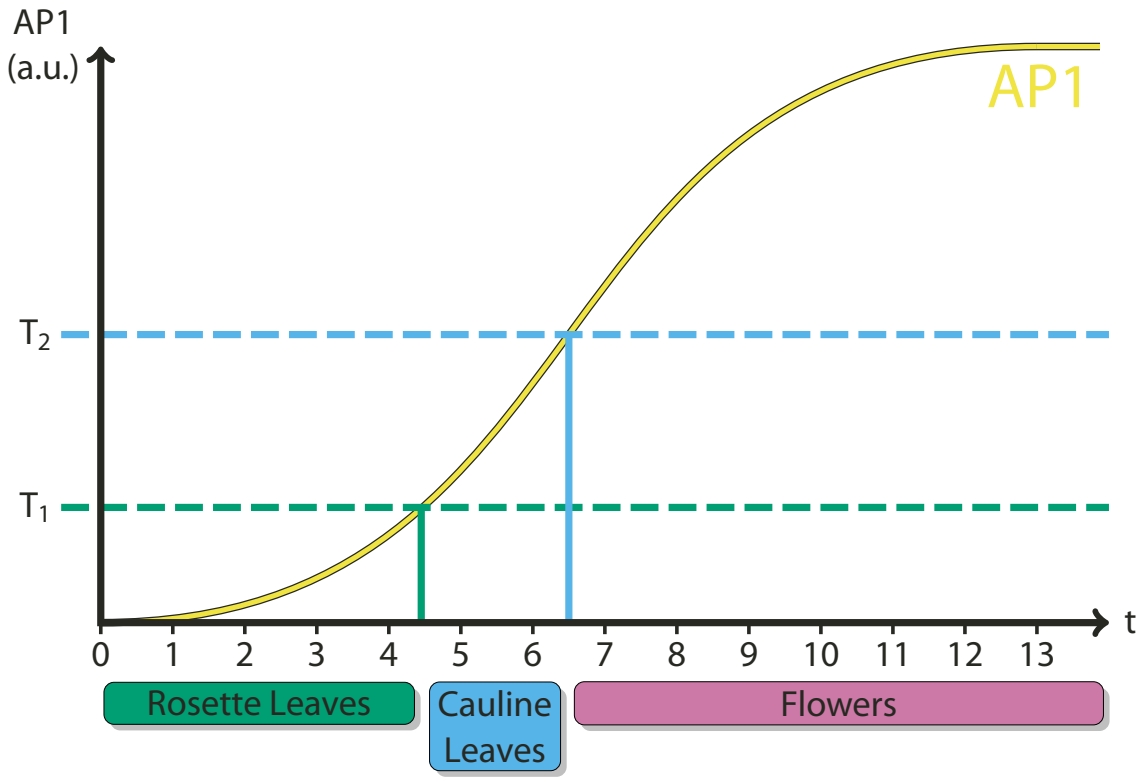


Figure 3.6: The definition of developmental decisions via AP1 hub levels. An example AP1 hub curve is shown in yellow. The developmental time taken to reach two thresholds, dashed lines, is mapped onto leaf numbers. AP1 levels below the first threshold correspond to rosette leaves. At the start of the transition, T_1, cauline leaves are produced, and on completion of the transition, T_2, flowers are made. This mapping allows for experimental leaf number data to be used to define a cost function for inference of model parameters. The values we used for T_1 and T_2 were 0.2 and 0.3. a.u., arbitrary units.

Therefore a reliable morphological indicator of time required for initiation and length of the floral transition in the Columbia accession is the number of rosette and cauline leaves made on the main stem prior to flowers and these totals are abundantly reported throughout the literature. To directly relate the simulated AP1 hub output to key events of the floral transition we defined two thresholds as pictured in Figure 3.6. The two thresholds were chosen to be at AP1 = 0.2, for the transition from rosette to cauline leaves, and at AP1 = 0.3, for the change from cauline leaves to flower production. The decision outcome takes on one of these three states: rosette leaves, cauline leaves or flowers at these defined AP1 levels. These threshold values are chosen arbitrarily as the parameters in the system are scaled relatively and will adjust to these selected values. A benefit of this approach is that it can be used to quantify information on timing of both the initiation of flowering as well as the duration of the transition. This strategy allows the developmental time to reach these states to be scaled to the leaf number data provided earlier (Table 3.1).

Having described the output of our model we now discuss the input and construction of the equations that define it. As in the simple motifs described earlier the floral pathway integrator FT comprises a major hub for many flowering pathways. Our model simplifies the effects of many biological and environmental processes leading to flowering by assuming they all feed in to the FT hub. Therefore the input function to the full network model is based on FT levels. We chose a linear time-dependent increase in FT levels until the floral transition is completed. The transition is completed in this model when the AP1 hub threshold at 0.3 is reached and flowers are produced. This linear with time increase of FT hub levels represents growth in leaf area proportional to time. Further justification for this input signal is provided by the fact that the number of rosette leaves has been shown to increase at a constant rate [190]. The equation for the synthesis of FT is therefore $v_{FT}(t) = v_{\mathrm{FT,35S}} + \eta t$, where η is the rate of FT production per unit leaf scaled to unit time (the value chosen for this parameter was 0.01) and $v_{FT,35S}$ is the transcription rate for overexpression under the control of the 35S promoter — if present this was taken as 1, therefore 100 times greater than the

base transcription rate in wildtype scenarios. Genetic networks were represented by ODEs using Hill-type gene activation and repression, assuming that protein binding is in equilibrium on the timescale of translation.

3.3.2 *Network to Equations*

A gene network can be converted to a set of ODEs following standard practice [109]. The model includes transcriptional regulation, protein-protein interactions, and protein degradation for the hub activities in the model. We introduce the following nomenclature. The concentrations of the hub activity proteins are denoted by $x_1 = [\mathrm{FT}]$, $x_2 = [\mathrm{TFL1}]$, $x_3 = [\mathrm{FD}]$, $x_4 = [\mathrm{LFY}]$ and $x_5 = [\mathrm{AP1}]$. K_{ij} is the effective binding constant between hub activity proteins i and j, $K_{i:k}$ the effective binding constant between hub protein i and a promoter site for gene k, $K_{ij:k}$ is the effective binding constant between the complex of hub activity proteins i and j and a promoter site for gene k, with $h_{i:k}$ and $h_{ij:k}$ the corresponding Hill coefficients. The equations governing the hub protein concentrations, x_i, present at any given time, t, are determined by the production rates, v_i, and the degradation rates, δ_i,

$$\frac{dx_i}{dt} = v_i - \delta_i x_i,$$

in which $i = 1, \ldots, 5$, corresponding to FT, TFL1, FD, LFY and AP1. The flowering time model can thus be described by differential equations, each describing the production and degradation rates for the five hubs. The degradation rates, δ_i, were all set to a constant value, 0.1. This can be done without loss of generality as the binding constants are allowed to change over a wide prior range to alter the effective transcription rates and adjust the effect of concentrations on downstream events. In reality the degradation rates are likely to differ between protein species and if the information became available these identified values can easily be substituted in to the model.

The model equations for the production rates are now discussed. The transcription rates for the hubs are determined by the contribution from the 35S promoter, if present, $v_{i,35S}$, and a weighted sum of the rate from nothing binding to a promoter site, $v_{i,0}$, a singly activated level, $v_{i,+}$, and a doubly activated transcription rate, $v_{i,++}$,

$$v_i = v_{i,35S} + p_{i,0} v_{i,0} + p_{i,+} v_{i,+} + p_{i,++} v_{i,++}.$$

The fractions, p_i, can be thought of as probabilities of activation and as such (including the probability of non-activation $p_{i,-}$) sum to one. These probabilities are determined from Hill type equations for gene activation and repression.

It is assumed that there is competitive binding between FT and TFL1 for FD [191] however this immediately complicates the equations and hence a full derivation is now given. We assume protein-protein binding to be in equilibrium on the timescale of protein synthesis. The total concentration of FD in the system is given by the free FD and bound FD thus $[\mathrm{FD_{total}}] = [\mathrm{FD_{free}}] + [\mathrm{FTFD}] + [\mathrm{TFL1FD}]$. The concentration of FT in the system is $[\mathrm{FT_{total}}] = [\mathrm{FT_{free}}]+[\mathrm{FTFD}]$. The aim is to calculate the amount of [FD] bound to [FT] and to [TFL1]. In the following we do not put square brackets, indicating concentration, or the subscripts, for clarity of presentation. Initially, consider the concentration of FTFD and start with the assumption that free FT and FD associate at rate k_+ and dissociate at rate k_-. Then following the law of mass-action the amount of FTFD changes over time as

$$\frac{\mathrm{dFTFD}}{\mathrm{d}t} = k_+ \cdot \mathrm{FT_{free}} \cdot \mathrm{FD_{free}} - k_- \cdot \mathrm{FTFD}.$$

At steady state $\left(\frac{\mathrm{dFTFD}}{\mathrm{d}t} = 0\right)$ and substituting in we are left with

$$\begin{aligned} k_+\mathrm{FT}\left(\mathrm{FD} - \mathrm{FTFD} - \mathrm{TFL1FD}\right) &= k_-\mathrm{FTFD}, \\ \mathrm{FTFD}\left(k_- + k_+\mathrm{FT}\right) &= k_+\mathrm{FT}\left(\mathrm{FD} - \mathrm{TFL1FD}\right), \\ \mathrm{FTFD} &= k_+\frac{\mathrm{FT}\left(\mathrm{FD} - \mathrm{TFL1FD}\right)}{k_- + k_+ \cdot \mathrm{FT}}. \end{aligned}$$

Bringing together the binding constants in to one term, k_d^{FT}, we are left with

$$\mathrm{FTFD} = \frac{\mathrm{FT}\left(\mathrm{FD} - \mathrm{TFL1FD}\right)}{k_d^{\mathrm{FT}} + \mathrm{FT}}, \tag{3.1}$$

and by similar arguments therefore also

$$\mathrm{TFL1FD} = \frac{\mathrm{TFL1}\left(\mathrm{FD} - \mathrm{FTFD}\right)}{k_d^{\mathrm{TFL1}} + \mathrm{TFL1}}. \tag{3.2}$$

Substituting (3.2) into (3.1) as

$$\mathrm{FTFD} = \frac{\mathrm{FT}}{k_d^{\mathrm{FT}} + \mathrm{FT}}\left(\mathrm{FD} - \frac{\mathrm{TFL1}\left(\mathrm{FD} - \mathrm{FTFD}\right)}{k_d^{\mathrm{TFL1}} + \mathrm{TFL1}}\right),$$

leaves us with

$$\left(k_d^{\mathrm{TFL1}} + \mathrm{TFL1}\right)\left(k_d^{\mathrm{FT}} + \mathrm{FT}\right)\mathrm{FTFD} = \mathrm{FT}\left(\mathrm{FD}\left(k_d^{\mathrm{TFL1}} + \mathrm{TFL1}\right) - \mathrm{TFL1}\left(\mathrm{FD} - \mathrm{FTFD}\right)\right),$$

and after some cancelling and rearrangement

$$\left(k_d^{\mathrm{TFL1}} + \mathrm{TFL1}\right)\left(k_d^{\mathrm{FT}} + \mathrm{FT}\right)\mathrm{FTFD} - \mathrm{FT}\cdot\mathrm{TFL1}\cdot\mathrm{FTFD} = k_d^{\mathrm{TFL1}}\cdot\mathrm{FT}\cdot\mathrm{FD}.$$

Extracting the FTFD term we are seeking and dividing across gives us

$$\mathrm{FTFD} = \frac{k_d^{\mathrm{TFL1}}\cdot\mathrm{FT}\cdot\mathrm{FD}}{\left(k_d^{\mathrm{TFL1}} + \mathrm{TFL1}\right)\left(k_d^{\mathrm{FT}} + \mathrm{FT}\right) - \mathrm{FT}\cdot\mathrm{TFL1}}.$$

Expanding the denominator and cancelling the $FT\cdot TFL1$ term simplifies our final equation for the concentration of FT-bound FD to

$$\mathrm{FTFD} = \frac{k_d^{\mathrm{TFL1}}\cdot\mathrm{FT}\cdot\mathrm{FD}}{k_d^{\mathrm{FT}}k_d^{\mathrm{TFL1}} + k_d^{\mathrm{FT}}\mathrm{TFL1} + k_d^{\mathrm{TFL1}}\mathrm{FT}}.$$

Analogously the equation for the concentration of FD bound to TFL1 is

$$\mathrm{TFL1FD} = \frac{k_d^{\mathrm{FT}}\cdot\mathrm{TFL1}\cdot\mathrm{FD}}{k_d^{\mathrm{FT}}k_d^{\mathrm{TFL1}} + k_d^{\mathrm{FT}}\mathrm{TFL1} + k_d^{\mathrm{TFL1}}\mathrm{FT}}.$$

Using the previous notation

$$x_{13} = x_1x_3 = [\mathrm{FTFD}] = \frac{K_{23}\cdot x_1\cdot x_3}{K_{13}\cdot K_{23} + K_{13}\cdot x_2 + K_{23}\cdot x_1},$$

$$x_{23} = x_2x_3 = [\mathrm{TFL1FD}] = \frac{K_{13}\cdot x_2\cdot x_3}{K_{13}\cdot K_{23} + K_{13}\cdot x_2 + K_{23}\cdot x_1}.$$

The transcription of *TFL1* hub genes consists of a single repression rate, $v_{2,+}$, if one transcription factor binds and a double repression rate, $v_{2,++}$, with the bindings of LFY and AP1 modelled as repressing Hill functions:

$$p_{5:2} = \frac{T_f^{h_{5:2}}}{T_f^{h_{5:2}} + x_5^{h_{5:2}}},$$

$$p_{4:2} = \frac{K_{4:2}^{h_{4:2}}}{K_{4:2}^{h_{4:2}} + x_4^{h_{4:2}}},$$

in which T_f is the threshold value of AP1 at which the plant enters the floral transition, 0.2. Thus the transcription of *TFL1* is $v_2 = v_{2,35S} + v_{2,+}\left(\left(1-p_{5:2}\right)p_{4:2} + \left(1-p_{4:2}\right)p_{5:2}\right) + v_{2,++}p_{5:2}p_{4:2}$.

As *FD* is strongly upregulated during the transition [50] when floral signals such as *FT* increase it is suspected it is under feedback control. A simple explanation is that it is auto-active in this regard. Thus the transcription of *FD* hub genes consists of a base transcription rate, $v_{3,0}$, and an enhanced rate, $v_{3,+}$, when the binding of FD leads to activation. The probability of *FD* being activated through FD is

$$p_{13:3} = \frac{x_{13:3}^{h_{13:3}}}{K_{13:3}^{h_{13:3}} + x_{13}^{h_{13:3}}},$$

where $K_{13:3}$ is the binding constant for the FTFD complex x_{13} to the promoter site of *FD* (x_3) and $h_{13:3}$ is the corresponding Hill coefficient. The modulator of the singly activated transcription rate of *FD*, $v_{3,+}$, is therefore $p_{3,+} = p_{13:3}$ and the amount of base rate transcription is modulated by $p_{3,0} = 1 - p_{3,+}$.

The transcription of *LFY* genes consists of a base transcription rate, $v_{4,0}$, and a singly activated rate, $v_{4,+}$, when the binding of FTFD or AP1 leads to activation, and a doubly activated rate, $v_{4,++}$ for when both FTFD and AP1 are bound to the promoter sites of *LFY*. FTFD and TFL1FD bind competitively to one promoter site of *LFY* with probabilities

$$p_{13:4} = \frac{K_{23:4}^{h_{23:4}} \cdot x_{13}^{h_{13:4}}}{K_{13:4}^{h_{13:4}} \cdot K_{23:4}^{h_{23:4}} + K_{23:4}^{h_{23:4}} \cdot x_{13}^{h_{13:4}} + K_{13:4}^{h_{13:4}} \cdot x_{23}^{h_{23:4}}},$$

$$p_{23:4} = \frac{K_{13:4}^{h_{13:4}} \cdot x_{23}^{h_{23:4}}}{K_{13:4}^{h_{13:4}} \cdot K_{23:4}^{h_{23:4}} + K_{23:4}^{h_{23:4}} \cdot x_{13}^{h_{13:4}} + K_{13:4}^{h_{13:4}} \cdot x_{23}^{h_{23:4}}},$$

that activate and repress the transcription, respectively. $K_{13:4}$ and $K_{23:4}$ are the binding constants for the protein complex x_{13} (FTFD) and x_{23} (TFL1FD) to the promoter site of the gene that codes for x_4 (LFY) and $h_{13:4}$ and $h_{23:4}$ are the corresponding Hill coefficients. AP1 also activates *LFY* and this was modelled as binding to a separate promoter site (not competing with FTFD or TFL1FD),

$$p_{5:4} = \frac{x_5^{h_{5:4}}}{K_{5:4}^{h_{5:4}} + x_5^{h_{5:4}}},$$

where $K_{5:4}$ is the binding constant for the protein x_5 (AP1) to the promoter site of the gene that codes for x_4 (LFY) and $h_{5:4}$ is the corresponding Hill coefficient. Thus, for the proportion of doubly activated *LFY* hub genes over time we obtain $p_{4,++} = p_{13:4}p_{5:4}$, for singly

activated $p_{4,+} = p_{13:4}(1-p_{5:4}) + (1-p_{13:4}-p_{23:4})p_{5:4}$ and for promoter sites of zero occupancy $p_{4,0} = (1 - p_{13:4} - p_{23:4})(1 - p_{5:4})$.

Similarly the transcription of *AP1* genes consists of a base transcription rate, $v_{5,0}$, and a singly activated rate, $v_{5,+}$, when the binding of FTFD or LFY leads to activation, and a doubly activated rate, $v_{5,++}$ for when both FDFT and LFY are bound to the promoter sites of *AP1*. FTFD and TFL1FD bind competitively to one promoter site of *AP1* with probabilities

$$p_{13:5} = \frac{K_{23:5}^{h_{23:5}} \cdot x_{13}^{h_{13:5}}}{K_{13:5}^{h_{13:5}} \cdot K_{23:5}^{h_{23:5}} + K_{23:5}^{h_{23:5}} \cdot x_{13}^{h_{13:5}} + K_{13:5}^{h_{13:5}} \cdot x_{23}^{h_{23:5}}},$$

$$p_{23:5} = \frac{K_{13:5}^{h_{13:5}} \cdot x_{23}^{h_{23:5}}}{K_{13:5}^{h_{13:5}} \cdot K_{23:5}^{h_{23:5}} + K_{23:5}^{h_{23:5}} \cdot x_{13}^{h_{13:5}} + K_{13:5}^{h_{13:5}} \cdot x_{23}^{h_{23:5}}}.$$

$K_{13:5}$ and $K_{23:5}$ are the binding constants for the protein complex x_{13} (FTFD) and x_{23} (TFL1FD) to the promoter site of the gene that codes for x_5 (AP1) and $h_{13:5}$ and $h_{23:5}$ are the corresponding Hill coefficients.

LFY also activates *AP1* hub genes and this was modelled as binding a separate promoter site (not competing with FTFD or TFL1FD),

$$p_{4:5} = \frac{x_4^{h_{4:5}}}{K_{4:5}^{h_{4:5}} + x_4^{h_{4:5}}}, \tag{3.3}$$

in which $K_{4:5}$ is the binding constant for the protein x_4 (LFY) to the promoter site of the gene that codes for x_5 (AP1) and $h_{4:5}$ is the corresponding Hill coefficient. From this we obtain for the proportion of doubly activated LFY genes over time $p_{5,++} = p_{13:5}p_{4:5}$, for singly activated $p_{5,+} = p_{13:5}(1-p_{4:5}) + (1-p_{13:5}-p_{23:5})p_{4:5}$, and for promoter sites of zero occupancy $p_{5,0} = (1 - p_{13:5} - p_{23:5})(1 - p_{4:5})$.

The maximal synthesis rates are set to three values, depending on whether nothing is bound, $v_{i,0}$, one type of activator is present, $v_{i,+}$, or two types of activators are working in an AND logic activation mode, $v_{i,++}$. Production and degradation rates for the AP1 hub were chosen such that the maximal concentration is unity ($AP1_{max} = 1$) in all genotypes considered. The complete set of equations are summarised in Table 3.3. With no further constraints, concentrations and binding constants are not independent so we chose to vary only the binding constants and Hill coefficients in the parameter inference. This gives a total of 19 parameters to infer.

Hub Protein Concentrations

$$\frac{dx_i}{dt} = v_i - \delta_i x_i$$

Hub Protein-Protein Binding

$$x_{13} = K_{23}x_1x_3/(K_{13}K_{23} + K_{13}x_2 + K_{23}x_1)$$
$$x_{23} = K_{13}x_2x_3/(K_{13}K_{23} + K_{13}x_2 + K_{23}x_1)$$

Hub Gene Activation

$$p_{5:2} = T_f^{h_{5:2}}/\left(T_f^{h_{5:2}} + x_5^{h_{5:2}}\right)$$
$$p_{4:2} = K_{4:2}^{h_{4:2}}/\left(K_{4:2}^{h_{4:2}} + x_4^{h_{4:2}}\right)$$
$$p_{13:3} = x_{13}^{h_{13:3}}/\left(K_{13:3}^{h_{13:3}} + x_{13}^{h_{13:3}}\right)$$
$$p_{13:4} = K_{23:4}^{h_{23:4}}x_{13}^{h_{13:4}}/\left(K_{13:4}^{h_{13:4}}K_{23:4}^{h_{23:4}} + K_{23:4}^{h_{23:4}}x_{13}^{h_{13:4}} + K_{13:4}^{h_{13:4}}x_{23}^{h_{23:4}}\right)$$
$$p_{23:4} = K_{13:4}^{h_{13:4}}x_{23}^{h_{23:4}}/\left(K_{13:4}^{h_{13:4}}K_{23:4}^{h_{23:4}} + K_{23:4}^{h_{23:4}}x_{13}^{h_{13:4}} + K_{13:4}^{h_{13:4}}x_{23}^{h_{23:4}}\right)$$
$$p_{5:4} = x_5^{h_{5:4}}/\left(K_{5:4}^{h_{5:4}} + x_5^{h_{5:4}}\right)$$
$$p_{13:5} = K_{23:5}^{h_{23:5}}x_{13}^{h_{13:5}}/\left(K_{13:5}^{h_{13:5}}K_{23:5}^{h_{23:5}} + K_{23:5}^{h_{23:5}}x_{13}^{h_{13:5}} + K_{13:5}^{h_{13:5}}x_{23}^{h_{23:5}}\right)$$
$$p_{23:5} = K_{13:5}^{h_{13:5}}x_{23}^{h_{23:5}}/\left(K_{13:5}^{h_{13:5}}K_{23:5}^{h_{23:5}} + K_{23:5}^{h_{23:5}}x_{13}^{h_{13:5}} + K_{13:5}^{h_{13:5}}x_{23}^{h_{23:5}}\right)$$
$$p_{4:5} = x_4^{h_{4:5}}/\left(K_{4:5}^{h_{4:5}} + x_4^{h_{4:5}}\right)$$

Synthesis Rates

$$v_i = v_{i,35S} + p_{i,0}v_{i,0} + p_{i,+}v_{i,+} + p_{i,++}v_{i,++}$$
$$p_{1,0} = 1, p_{1,+} = 0, p_{1,++} = 0$$
$$p_{2,0} = 1, p_{2,+} = p_{5:2}(1 - p_{4:2}) + p_{4:2}(1 - p_{5:2}), p_{2,++} = p_{5:2}p_{4:2}, T_f = 0.2$$
$$p_{3,0} = 1 - p_{3,+}, p_{3,+} = p_{13:3}, p_{3,++} = 0$$
$$p_{4,0} = (1 - p_{13:4} - p_{23:4})(1 - p_{5:4}), p_{4,+} = p_{13:4}(1 - p_{5:4}) + (1 - p_{13:4} - p_{23:4})p_{5:4}, p_{4,++} = p_{13:4}p_{5:4}$$
$$p_{5,0} = (1 - p_{13:5} - p_{23:5})(1 - p_{4:5}), p_{5,+} = p_{13:5}(1 - p_{4:5}) + (1 - p_{13:5} - p_{23:5})p_{4:5}, p_{5,++} = p_{13:5}p_{4:5}$$
$$v_{1,0} = \eta t, \eta = 0.01, v_{3,0} = v_{4,0} = 0.01, v_{2,0} = v_{5,0} = 0, v_{i,+} = 0.05, v_{i,++} = 0.1, v_{i,35S} = 1 \text{ for } i \in \{1, \ldots, 5\}$$

Initial Conditions

$$\frac{v_{i,35S} + v_{i,0}}{\delta_i} \text{ for } i \in \{1, \ldots, 5\}$$

Table 3.3: Model equations. The concentrations of the hub activity proteins are denoted by $x_1 = [\mathrm{FT}]$, $x_2 = [\mathrm{TFL1}]$, $x_3 = [\mathrm{FD}]$, $x_4 = [\mathrm{LFY}]$ and $x_5 = [\mathrm{AP1}]$. K_{ij} are effective binding constants between hub activity proteins i and j, $K_{i:k}$ the effective binding constants between hub activity protein i and a promoter site for the hub activity gene k, $K_{ij:k}$ effective binding constants between complexes of hub activity proteins i and j and a promoter site for gene k, and h_i the effective Hill coefficients. $p_{i:k}$ is the fraction of hub activity protein i bound to a promoter site of gene k, and $p_{ij:k}$ the fraction of the promoter of gene k with the complex i and j. All degradation rates were set to $\delta_i = 0.1$. Initial conditions were set at 0.1 for LFY and FD, and 0 for the other hubs in wildtype gene scenarios.

3.3.3 *Priors for the parameters*

Uninformative priors for the parameters were chosen as we have no knowledge of their likely values *in vitro* let alone *in planta*. Hill coefficients were chosen uniformly from U(1,4). These values for Hill coefficients were also chosen in a recent DREAM challenge entry [192] and are reasonable in practice [193]. Binding coeffecients have been shown to be scale parameters [194] and thus a Jeffreys-type prior was placed on those parameters with lower and upper bounds of 0.0001 and 10 respectively.

3.3.4 *Likelihood function*

As discussed earlier setting thresholds of the AP1 hub allows the developmental time to reach these thresholds to be scaled to leaf number data. It is now discussed in more detail how the introduction of fixed thresholds allows for the leaf numbers to be used constructively to explore parameter space. For all genotypes the levels of the AP1 hub are the output of our network and are mapped to the respective genotype's rosette and cauline leaf number data. The error model was chosen to be Gaussian given the data available, Table 3.1, but as leaf numbers are integer count data a Poisson error model may be more appropriate. The N=13 genotypes' leaf numbers in the training set are assumed to be independent with their own individual errors. This gives a likelihood function of the form

$$\mathcal{L} = \prod_{k=1}^{N} \frac{1}{\sigma_k \sqrt{2\pi}} \exp\left[\frac{-\left(f(\boldsymbol{\theta})_k - \mathbf{R}_k\right)^2}{2\sigma_k^2}\right] \exp\left[\frac{-\left(f(\boldsymbol{\theta})_k - \mathbf{C}_k\right)^2}{2\sigma_k^2}\right],$$

where $f(\boldsymbol{\theta})_k$ is the function of the parameters that computes the leaf numbers. $\mathbf{R}_k$ and $\mathbf{C}_k$ are, respectively, the true rosette and cauline leaf numbers for the genotype k and σ_k the experimental standard deviation in the total leaf numbers for that genotype. A more explicit way of writing this for our situation of evaluating the model when AP1 reaches 0.2 and 0.3, for rosette and cauline leaf numbers respectively, is the following log-likelihood function

$$\begin{aligned}\log \mathcal{L} = &- N \log\left(\sqrt{2\pi}\right) - \sum_{k=1}^{N} \log \sigma_k \\ &- \sum_{k=1}^{N} \frac{\left(\left(f(\boldsymbol{\theta})_k\big|_{AP1=0.2} - \mathbf{R}_k\right)^2 + \left(f(\boldsymbol{\theta})_k\big|_{AP1=0.3} - \mathbf{C}_k\right)^2\right)}{2\sigma_k^2}.\end{aligned}$$

Having defined the data, model, prior and likelihood function we now have everything in place for a proper Bayesian treatment of the problem. Hence the model parameters were inferred by nested sampling using the leaf number data from genotypes in the training set (wildtype, single and double mutants). The resulting posterior distribution of the parameters for the model was used to make predictions for the triple mutants and explore the model's dynamics, sensitivity and robustness.

3.4 Results

3.4.1 *Biological evidence contradicts statistical evidence*

Given the data available nested sampling estimates the log evidence for the derived model to be -56.17 ± 0.17. The parameters from 2000 equally-weighted posterior samples are used to give the estimated leaf numbers for all genotypes in Table 3.4 [8] . The network output is able to capture the true leaf numbers well for the training set and predict the triple mutant leaf numbers. The largest predicted leaf number deviation is for the triple mutant *tfl-1 fd-2 fdp-1* where the model predicts six more rosette leaves than recorded experimentally. These inaccuracies do not surprise us. The triple mutants were deliberately chosen as a prediction set to test the network to its fullest. The effect of adding extra mutations is clearly not simply additive. The fact that our estimated leaf numbers are in the right ball park then is encouraging because we believe we have captured some of the genetic variability in our network structure. During the undertaking of this research a paper was published [80] showing that FD is a target gene of LFY, at least at the seedling stage. This could explain the observation that FD expression is strongly upregulated during the floral transition [50]. Further experiments were suggested and analysis of the FD promoter revealed it contains two LFY binding sites [166]. It was also demonstrated that deletion of the LFY binding sites in the FD promoter abolishes FD upregulation upon the floral transition and ChIP showed that the binding was direct between LFY and the FD promoter [166]. In light of this conclusive biological evidence we removed our hypothesised auto-activation of the FD hub which accounted for the upregulation of FD during the floral transition and

[8] *Taking the parameters from the best-fit nested sample gives a maximum likelihood score of* -24.072. *The estimated leaf numbers from this set of parameters are also given in Table 3.4.*

replaced it with a term describing feedback from the LFY hub on to FD. The equation for probability of FD being activated, now through LFY, becomes

$$p_{4:3} = \frac{x_4^{h_{4:3}}}{K_{4:3}^{h_{4:3}} + x_4^{h_{4:3}}},$$

where $K_{4:3}$ is the binding constant for LFY protein, x_4, to the promoter site of the gene that codes for x_3 (FD) and $h_{4:3}$ is the corresponding Hill coefficient.

Genotype	No. of rosette leaves			No. of cauline leaves			Data set
	True	Model		True	Model		
		Best-fit	Mean ± SD		Best-fit	Mean ± SD	
Wild type (Col)	7.9	8.8	8.7 ± 0.4	1.4	1.8	1.8 ± 0.1	Training
35S:*FT*	4.4	4.3	4.1 ± 0.3	1.0	1.7	1.7 ± 0.1	Training
35S:*LFY*	3.8	4.8	4.7 ± 0.1	1.8	2.1	2.2 ± 0.1	Training
35S:*TFL1*	27.5	27.1	26.8 ± 1.8	15.7	15.5	14.4 ± 1.9	Training
lfy-12	13.0	13.4	13.6 ± 0.8	5.3	4.9	5.1 ± 0.4	Training
ft-10	36.4	36.7	37.1 ± 1.2	9.3	8.7	8.6 ± 0.9	Training
tfl1-1	7.7	8.1	8.3 ± 0.4	0.4	1.7	1.8 ± 0.1	Training
fd-2	18.5	15.6	16.1 ± 1.0	4.63	3.8	3.7 ± 0.3	Training
fdp-1	11.2	9.6	9.5 ± 0.4	2.0	1.8	2.0 ± 0.1	Training
fd-2 fdp-1	32.9	32.1	31.3 ± 1.0	6.3	7.6	7.3 ± 0.7	Training
35S:*TFL1 fd-2*	23.8	27.8	27.1 ± 1.9	8.2	5.2	5.2 ± 0.6	Training
tfl1-1 fd-2	14.4	14.2	15.1 ± 0.8	4.6	3.7	3.6 ± 0.3	Training
35S:*FT fd-2*	8.3	8.2	7.7 ± 0.9	2.4	2.8	3.0 ± 0.4	Training
tfl1-1 fd-2 fdp-1	24.83	31.0	30.8 ± 1.1	6.67	7.5	7.3 ± 0.6	Prediction
35S:*TFL1 fd-2 fdp-1*	31.33	34.2	34.0 ± 1.3	11.0	8.2	7.7 ± 0.8	Prediction
35S:*FT fd-2 fdp-1*	25.8	26.2	26.3 ± 2.2	5.6	7.4	7.2 ± 0.6	Prediction

Table 3.4: Experimental and model leaf number data for the network with FD auto-activation. For each genotype the table lists the mean experimental leaf number data and estimated (for the training set) or predicted best-fit and mean ± SD values for rosette and cauline leaves. The best-fit values use one set of parameters and thus has no possible associated error. This sample is taken from all the nested samples and is the one that maximises the likelihood function the most from the final set. Mean and SD based on 2000 posterior samples. SD, standard deviation.

Simulating this alternative network and comparing estimates of leaf number data to experimental leaf numbers reveal some differ-

ences in the training set but in sum very little. The predictions for the triple mutants slightly favour the FD auto-regulation model but of course a full Bayesian model comparison is preferred. Running the nested sampling algorithm with the slightly altered network architecture gave a worse log evidence score of -62.68 ± 0.18. Indeed on Jeffreys' scale the Bayes factor between these two models decisively favours the original assumption that FD is an auto-activator. How can we account for this difference between biology and statistics? Formally the Bayes factor is only equal to the posterior odds of two models if we assume both hypotheses to be equally true. Biologically the prior odds should favour the LFY feedback to FD model by many orders of magnitude based on the results of the experiments suggested above. In fact given that an investigator should only define models with a reasonable verisimilitude in model space, the first hypothesis could be discounted directly. Statistically the evidence incorporates an Occam factor [133], penalising more complex models if their fit is not substantially improved. Comparing a basic least squares fit to the training data for the best-fitting likelihood parameters from nested sampling reveals little difference in the sum of squared errors between models, 43.1 compared to 45.7, actually in favour of the network with LFY feedback on to FD. These numbers came from the best-fit parameter set that was found by considering the errors in the data as described by the likelihood function and then using them in a simple least-squares residual which does not consider the errors in the data. Thus we may not want to assign much weight to this particular finding.

To further build an understanding of the parameter space, a simulated annealing algorithm [120, 169] optimising the sum of squared errors was run from a number of random starting parameter sets. The procedure was cooled to near 0 from a starting temperature of 50, with the temperature reduction factor set at 0.85, and each exploration at a certain temperature involving 50 cycles with 20 subcycles. The majority (94 out of 96 runs) of results from the initial model were near to the lowest minimum value found, 27.9. In contrast for the new hypothesis a sole run (out of 83) was near its lowest minimum (32.0), with most (81/83) entering a wide local minima with a best fit more than double the optimal solution (around 82).

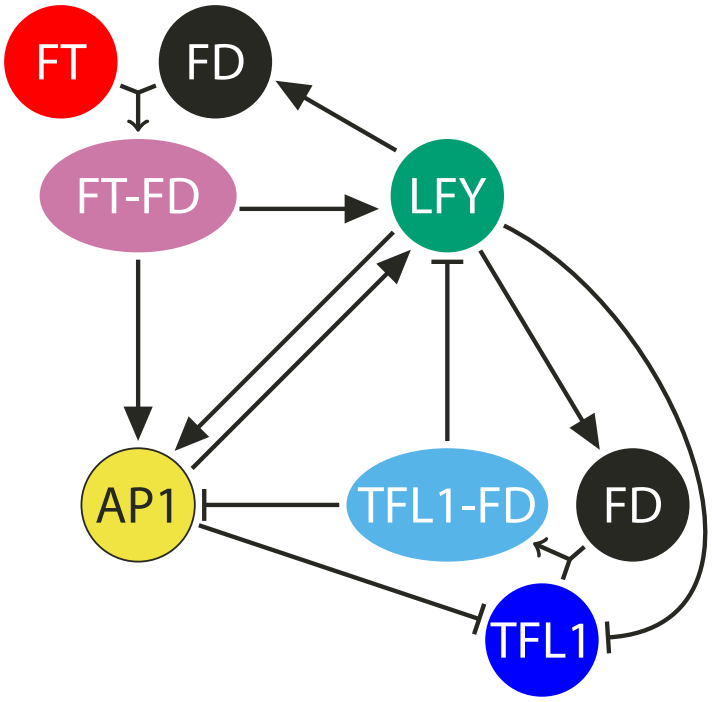


Figure 3.7: New regulatory network diagram. This network shows all connections in the network with LFY feedback on to *FD*. Both FT and TFL1 act through FD (small arrows) and these are shown as ovals regulating the other hubs (circles). FT is the input and AP1 the output of the network model. Filled arrowheads indicate activation and T-bars represent inhibition.

This suggests that a large amount of parameter space fits well with the FD auto-regulatory network, but not as much does with the alternative hypothesis. This would account for a significant difference in Occam factor and can therefore help us to explain the difference in evidence results. Ultimately the biological evidence is overwhelming for a network that has a feedback term from the LFY hub to the FD hub (depicted in Figure 3.7) and thus this is the model that was taken forward.

Posterior estimates of rosette and cauline leaf numbers from this newer network are shown in Figure 3.8. Firstly it can be observed that we can accurately infer both types of leaf numbers with a clear idea of the uncertainty attached with those estimates or predictions. Secondly the violin plots show (symmetrically) the distribution of our inferred leaf numbers is unimodal. This is important because an average prediction could be the average of two, or more, predictions that are different from the experimental value. Using a rigid methodology such as Bayesian inference allows this to be elucidated. In this case we show that due to there being no multimodality a mean and standard deviation will give a fair summary for both the estimates and the predictions.

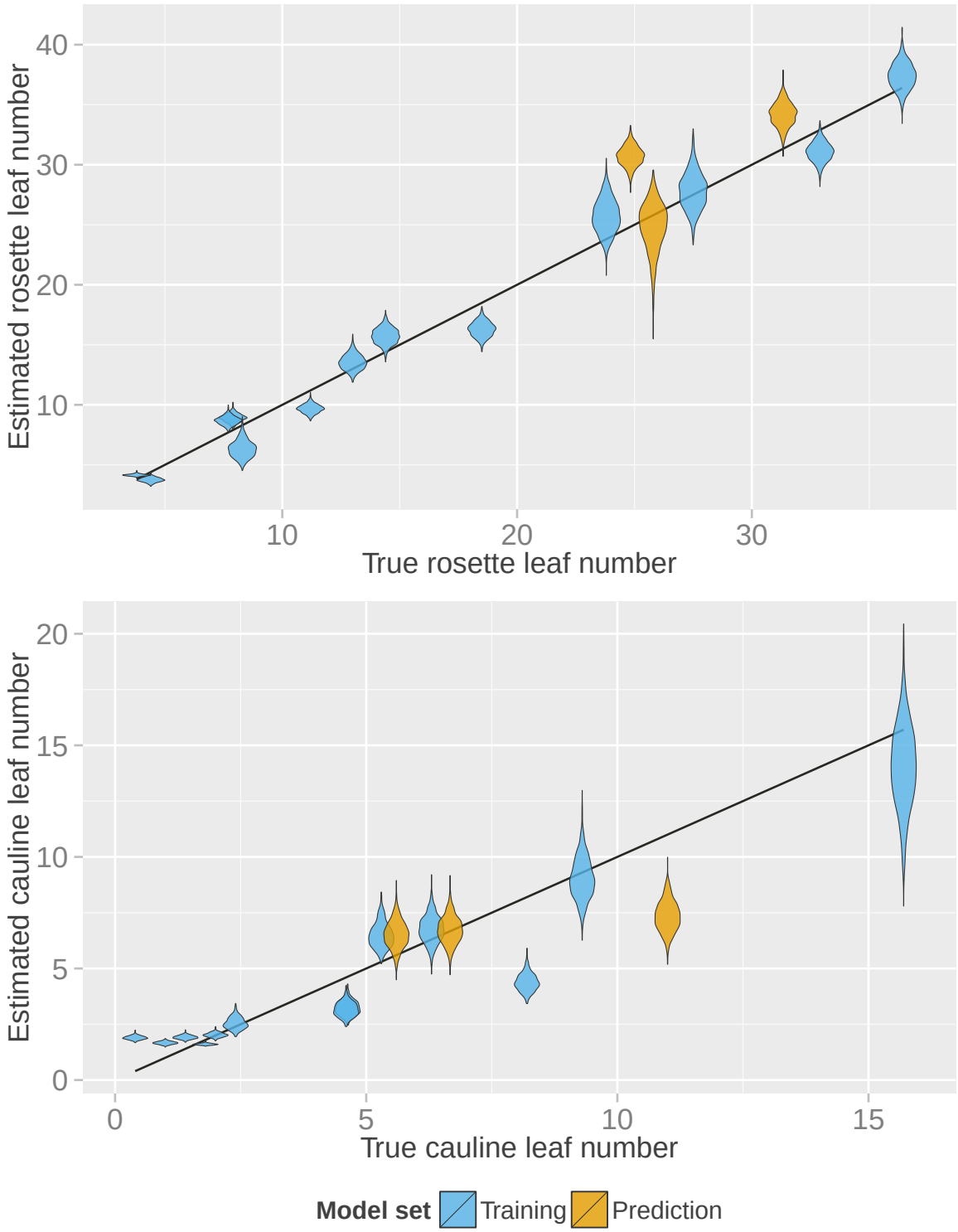


Figure 3.8: Estimated and predicted rosette and cauline leaf numbers for the model with LFY feedback onto FD. 2000 posterior samples were used for kernel density estimates of each genotype as a violin plot. The x-axis positions correspond to the mean number of experimental rosette and cauline leaves. The density estimates reveal that the predictions are unimodal. In a perfect-fit model the violins would all be on the dark line.

3.4.2 *Dynamics of the floral transition network*

The dynamics of the wildtype network's hub proteins representing concentrations in a cell on the flanks of the apex are shown in Figure 3.9 for 2000 randomly-drawn equally-weighted posterior samples. It can be seen that despite variability in all hubs the output hub, AP1, is under tight control. This is not very surprising given AP1 dynamics are what constrains the likelihood function and therefore the model fitting. FT is most variable as it approaches its steady state but in the wildtype network this has no consequence on the estimated leaf numbers. The effect on TFL1 of being repressed is clear to see and it reaches a steady state around zero. FD experiences the greatest delay in upregulation. This is because it has to wait until sufficient LFY is in the system to bind and then activate it. LFY does not accumulate immediately because there is a slight delay before the higher transcription rate is active as it needs sufficient levels of AP1 to have both promoting binding sites occupied. AP1 transcription and translation occurs very quickly in this wildtype set-up because initial levels of LFY are present to bind to the AP1 hub promoter. This is rapidly followed by rising FT levels which kick in to activate the higher rates of upregulation.

Throughout the time period of the floral transition strongly rising levels of the transcription factors in the network are observed. This is in agreement with the literature [50, 78, 79]. The behaviour of our TFL1 hub is perhaps also reflective of reality [81], at least assuming that TFL1 protein levels can be detected at a similar level to our hub levels, yet this is less clear. Because we have modelled a cell that is poised to transition to a flower on the flanks of the apex it can be thought of as initially being in floral anlagen — cells that form the foundations of, in this case floral, organs. Conti & Bradley [81] show that TFL1 protein moves without its expression domain including into anlagen cells and there *LFY* is expressed. At some point these cells experience stronger floral signals and TFL1 is restricted from floral meristems. Thus even if *TFL1* mRNA is not expressed in those primordial cells at the early stages of development that this chapter focuses on, its protein product is likely to be present before declining. This shows the early TFL1 hub dynamics in the model could be a decent description of the system.

Another point to bear in mind when critically evaluating this network's dynamics is that the input is smooth so it is not surprising that the output is smooth. Therefore the behaviour of the network to non-monotonic input signals is investigated next along with other important dynamics representative of the floral transition to reassure ourselves that the larger network still maintains qualitative properties of the network motifs.

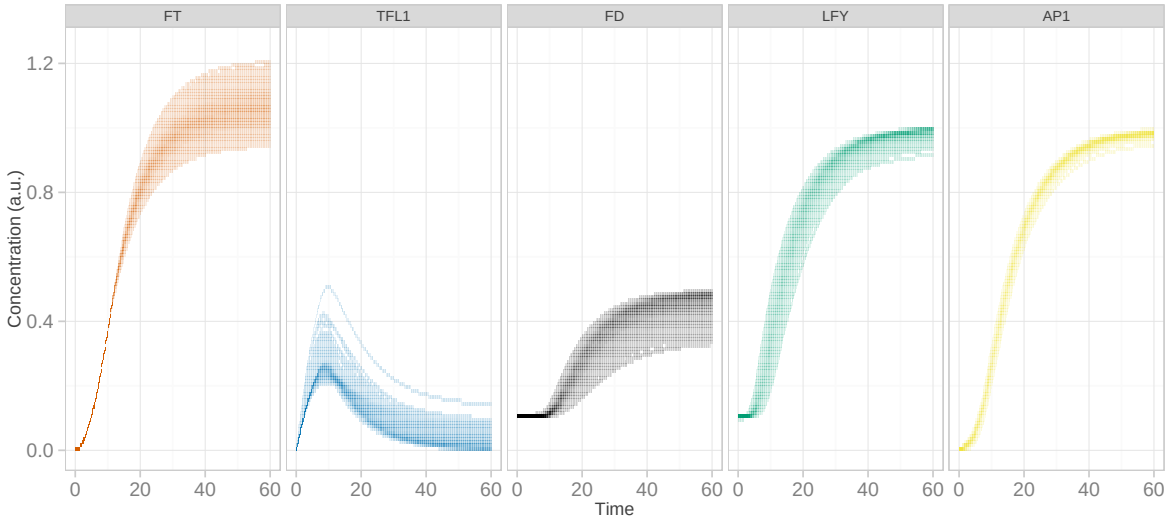


Figure 3.9: 2000 posterior samples of the wildtype network dynamics of the five variables. Darker intensity indicates more samples at that concentration for each timepoint. FT will follow the same path initially for all samples until AP1 crosses the rosette-to-cauline threshold but fans out after a while. TFL1 is repressed as LFY and AP1 become established. The predictive dynamics of the AP1 hub closely match each other despite the variation in the other hub proteins.

Irreversibility

An important characteristic of the floral transition in wildtype Arabidopsis is its irreversibility. This means that once committed to flowering the primordia can not then revert back into vegetative tissue before making floral organs. Because FT hub mutant plants flower after a long time in suitable conditions [195], by design our network can incorporate no FT hub production term and still output flowers. However what if there was initially FT production that was then withdrawn? This could represent for instance a light shift experiment from long days to darkness or a construct that can inducibly knockdown FT. This was investigated by setting the FT production

term, ν_{FT}, in the model to zero after a certain number of leaves had been observed. Simulations, using the best-fit parameters, varying the length of inductive conditions are shown in Figure 3.10. Flow-

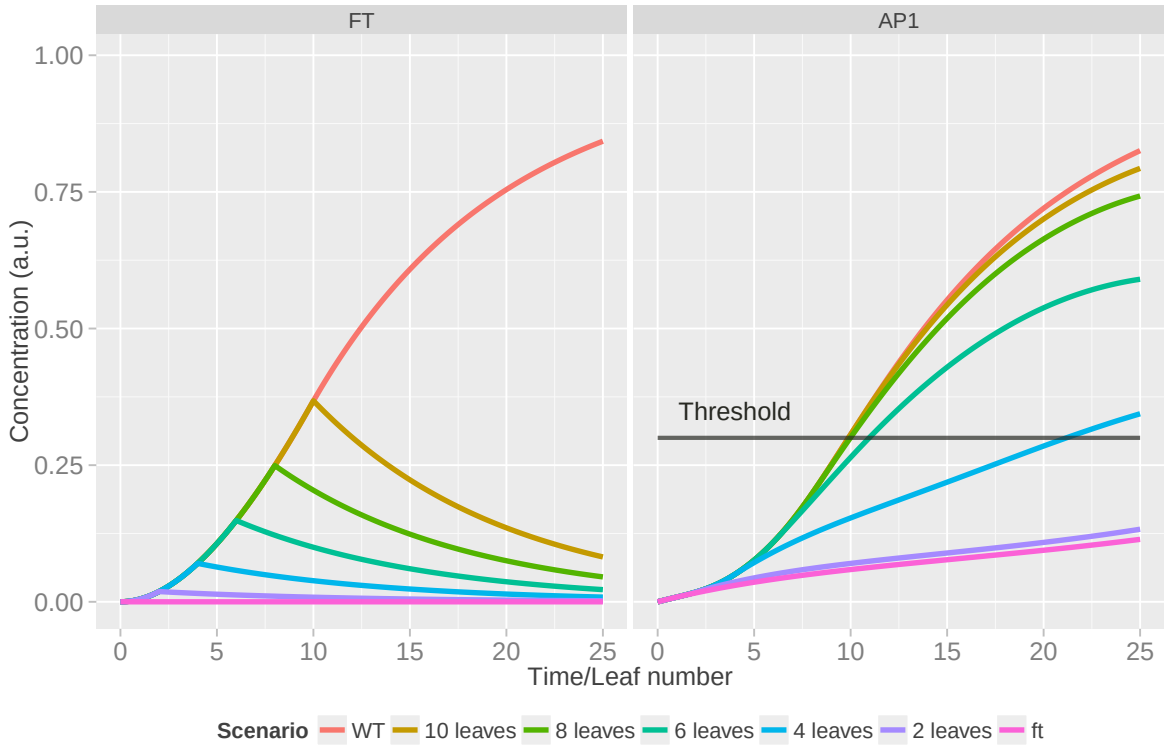


Figure 3.10: The effect of FT production withdrawal on the AP1 hub. FT is withdrawn after the indicated number of leaves and its effect on the correspondingly coloured lines for AP1 is shown. Flowering time is judged by AP1 crossing the 0.3 threshold. Longer times equate to delayed flowering. The wildtype (WT) and FT-knockout (ft) scenarios are shown for comparison. Even a small induction of FT speeds up the time to flower compared to the knockout line.

ering time is judged here by AP1 crossing the 0.3 threshold. As can be seen flowering time is identical to the wildtype if FT is stopped after a developmental time of 10 leaves, and near identical if terminated at the eight leaf stage. Ceasing FT production after the formation of six leaves causes a slightly delayed time to flower. On the other hand the duration of FT production has a strong effect on the timing of flowering if withdrawn early, at the two or four leaf stage. Compared with the simulated complete lack of the FT hub (labelled ft in Figure 3.10) flowering is still accelerated when FT production is withdrawn after the formation of only two leaves. Once AP1 has

crossed the rosette-to-cauline threshold (0.2; around the 8 leaf stage) preventing FT production has little effect on the timing of flowering. At this point the plant is committed to flowering which is consistent with a study showing that one long day can be sufficient to cause floral commitment under certain conditions [46]. Thus by introducing a flowering threshold this network can exhibit irreversibility, and from earlier we know this is due to the memory element present in the core regulatory motif.

Noise Filtering

FT expression is strongly influenced by temperature [18]. Therefore plants are likely to experience large day-to-day fluctuations in *FT* levels. To simulate these conditions first uniform random noise of up to 50% of the signal v_{FT} was given as input. With this level of noise FT hub levels are only minorly perturbed, Figure 3.11 Left, with no effect on the AP1 hub. Pushing this further very high noise levels in FT production rates were simulated such that uniform random noise of up to 200% v_{FT} was given as input. These signals propagate through to the levels of FT, Figure 3.11 Right, however the network is also able to filter this out, resulting in a smooth AP1 curve. Under these conditions the model simulates the ability of this developmental system to filter noisy environmental signals and make correctly timed decisions. The buffering properties of the model result in AP1 levels that are unaffected by these perturbations because of the incorporated coherent feedforward loop.

Circadian Oscillations

Both FT hub genes *FT* and *TSF* are expressed in a circadian fashion *in planta* [37, 54]. *AP1* expression, our marker for the output of the flowering pathway, does not oscillate [196]. We wished to test how well our full regulatory network also exhibits this ability to integrate out and smooth input signals. Hence it was examined how oscillating production rates of FT influence the FT and AP1 hubs in particular. The input term, v_{FT}, was multiplied by the oscillating function $\sin(ct)^2$, where c controls the frequency, either 0.5 (Figure 3.12 Left) or 3 (Figure 3.12 Right). As shown in these figures the network is able to filter out large oscillations in FT production rate at both high

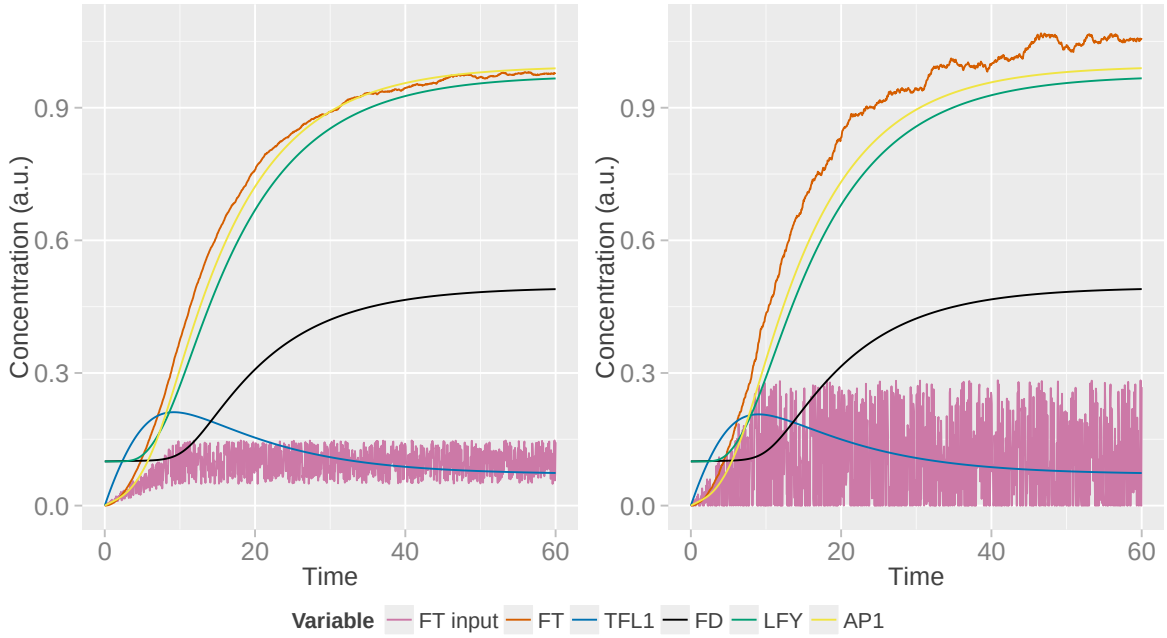


Figure 3.11: Effect of signal noise on the network hubs. 50% (left) or 200% (right) random noise was added to the signal (purple). The lower level of noise in the FT input barely filters through to the FT hub (red) thus not surprisingly the AP1 (yellow) output is smooth. The high noise levels affect the FT hub more strongly but they are also filtered out by the network so there is no effect on the AP1 hub.

and low frequencies. These modulations propagate through to FT levels yet generate an unperturbed increase in AP1.

As stated previously our proposed network contains a motif that is known to buffer noise well. However in this flowering time model the ability to buffer noise is also partly a result of the magnitude of the degradation rates compared to the steady state values. Doubling the degradation rates, such that the modulation in v_{FT} is fed through even more strongly to FT, we still find that the network filters out these perturbations, Figure 3.12 (Lower row). The full model therefore captures key properties of the floral transition in Arabidopsis, including irreversibility and the filtering of noisy and circadian signals, due to the network motifs built into its architecture.

3.4.3 *Parameter analysis*

The marginal distributions of the 19 parameters over their prior range are shown in Figure 3.13, and then after zooming in on regions of non-zero probability in Figure 3.14. These distributions show a num-

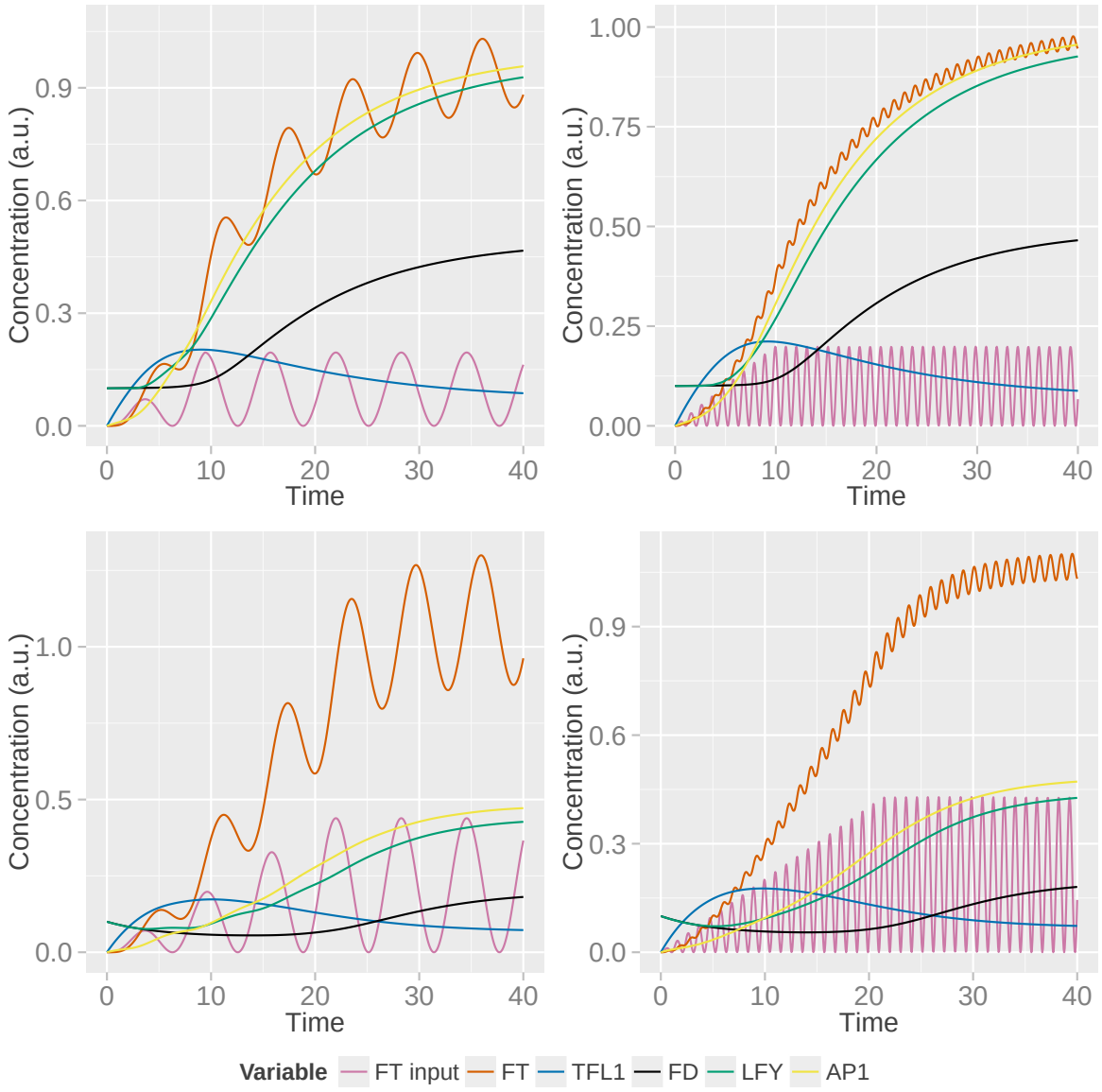


Figure 3.12: Effect of oscillating FT input signal on the network hubs. Upper row: degradation rates are set at 0.1 as used throughout this chapter. The circadian oscillations at low frequency (left) of FT activation (purple) propagate through to the FT hub (red) with no effect on the AP1 hub (yellow). High frequency (right) FT activation oscillations transmit through to the FT hub with small amplitude and thus this does not prevent a smooth rise in the AP1 hub. Lower row: the degradation rates are doubled to 0.2. Very small modulations are seen at the start of the AP1 hub curve due to the massive amplitude in the FT hub's oscillations (left). High degradation rates and fast oscillating FT input signal again damp the FT hub's dynamics without affecting AP1 (right).

ber of interesting things justifying the use of a Bayesian treatment of the problem. We note how some Hill parameters can take on most values of their prior and thus are not constrained by the data. In constrast the data constrains some parameters very strongly. As these distributions are correctly normalised to have an area equalling one the parameters with higher values are the ones that are most defined. In particular the binding coefficient of LFY on to the *TFL1* promoter, $K_{4:2}$, has a high probability in the region between 10^{-4} and 10^{-2}. The parameters controlling the binding coefficient of a number of species to *AP1* are also well defined. This perhaps does not surprise because the AP1 levels are the output of our system and must therefore be constrained to give a reasonable likelihood score. No parameters seem to show multimodality in their marginal distributions, but not all have one strong peak. This discovery might not be accounted for with classical statistical techniques. For example the mean and standard deviation of the Hill parameter controlling the binding of the TFL1FD hub complex to *AP1*, $h_{23:5}$, would completely miss the fact that more weight is given to the higher end of the parameter value, around 4.

In fact this case highlights a strength of the Bayesian approach as it can raise a concern, leading one to investigate model reliability but providing an avenue for updating our degrees of belief. It seems that high Hill coefficients are preferred for both $h_{23:5}$ and $h_{4:3}$, meaning that for a second round of inference maybe the prior range should be extended. However such large Hill terms are improbable in many realistic situations. To account for this habit of large Hill coefficients resulting from parameter inference we could either be more specific with our prior belief on the parameters by choosing a different prior distribution, or one could add a penalty to the likelihood function. A simple prior distribution to consider here could be a triangular distribution with most weight at the end nearest one and little weight at high values of the Hill parameter. If the data were not informative enough to overcome our prior belief then the parameters' posterior distribution would be dominated by the prior thus alleviating the problematic high Hill coefficients. In this flowering time model uniform priors were initially chosen for all Hill parameters as we had no knowledge of their likely values. After examining the posterior

we have revealed that firstly, if the analysis was updated then two parameters ought to have different prior distributions, and secondly that our data does not constrain the choice of all parameters to realistic values.

A selection of joint distributions between parameters is shown in Figure 3.15. No combinations of parameters appear multimodal although there are correlations in some combinations (Figure 3.16). For example, $K_{4:3}$, the parameter controlling the binding of LFY onto *FD* exhibits some correlation with other parameters. The binding constants of the two mutual activators, *AP1* and *LFY*, to each other show a surprising yet evident non-linear correlation (Figure 3.16 Bottom right). Some parameter joint distributions are so spread out that they essentially look uniformly randomly distributed as in Figure 3.15 (Bottom row). These interesting figures all together show that evaluating the posterior distribution will enable one to find more information, such as the spread or correlations, that can be missed from a point estimate of the parameters. Likewise it can be inferred that some parameters are far more sensitive to their choice than others, and thus in the future this could be an avenue for minimising the proposed model.

Additionally the effect of some parameters on the flowering time of the wildtype network was investigated qualitatively. It was found that the interplay between K_{13} and K_{23} was very important. This is intuitive as these parameters control the relative binding strength of FD to either FT or TFL1. Tighter TFL1 binding to FD leads to delayed flowering, which can be compensated for by lowering K_{13} even more, to give tighter binding between FT and FD. The competition for FD binding is therefore critical to correct flowering time for the input but what about for the output? Reducing values of $K_{23:5}$, controlling the binding of the TFL1FD complex on to the *AP1* promoter, again leads to a delay in AP1 accumulation. As previously this effect can be counterbalanced by decreasing the value of $K_{13:5}$ so that FTFD binds more tightly to the *AP1* promoter.

Satisfyingly these results support a number of experimental studies. The proposal that FT may interact with FD amongst others in a transcriptional complex more strongly than TFL1 to activate flowering genes [191] is endorsed by the qualitative analysis of our network.

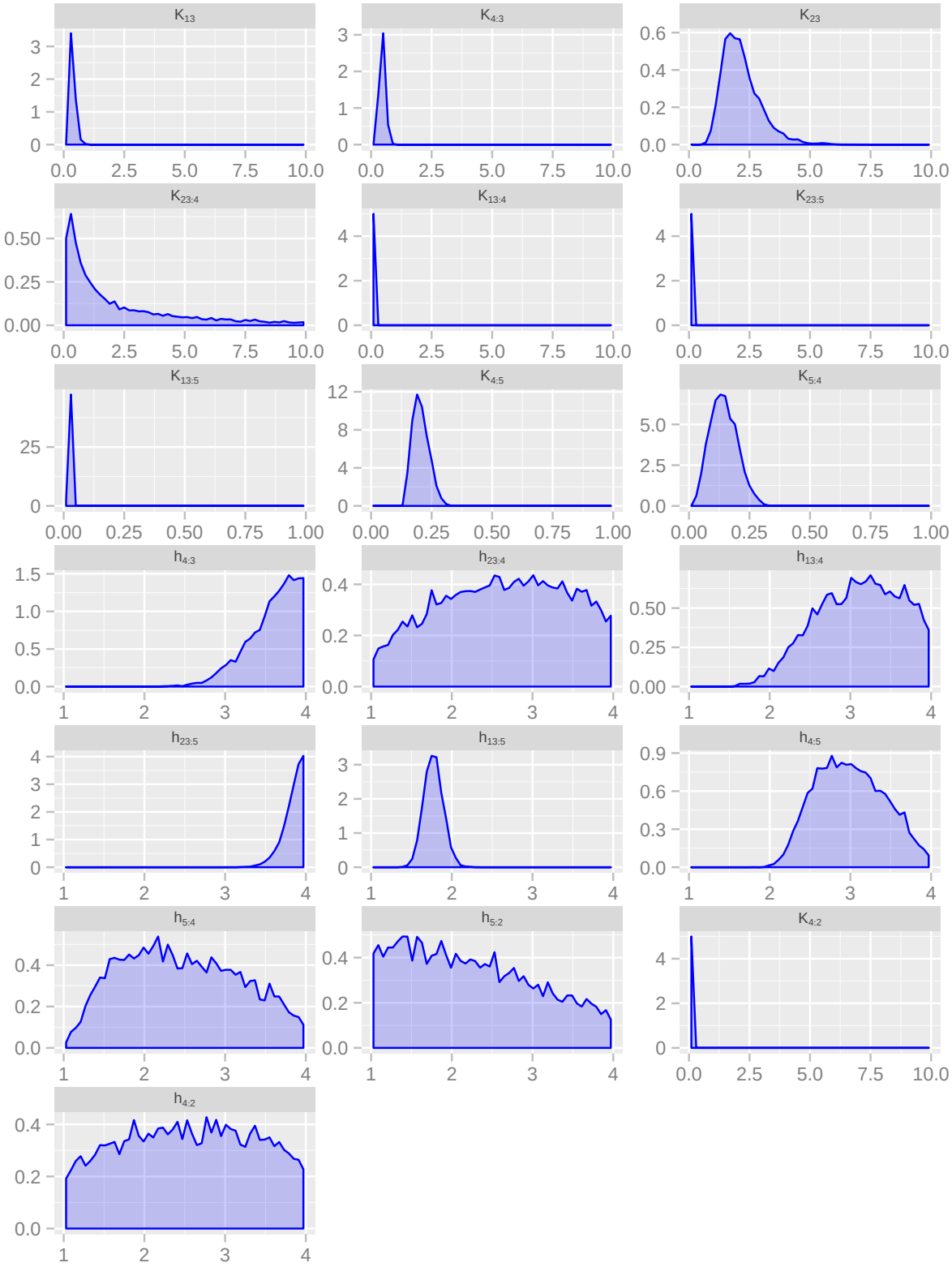


Figure 3.13: Estimated marginal posterior parameter distributions over prior range. The kinetic parameters, K_*, have a $\log_{10}$-uniform prior range of $[10^{-4}, 10]$. The Hill terms, h_*, follow the prior $U(1, 4)$. More kinetic parameters than Hill parameters have a higher probability density indicating they are the more constrained. All marginals have been properly normalised with area equal to one, thus the y-axis values represent estimated probability densities.

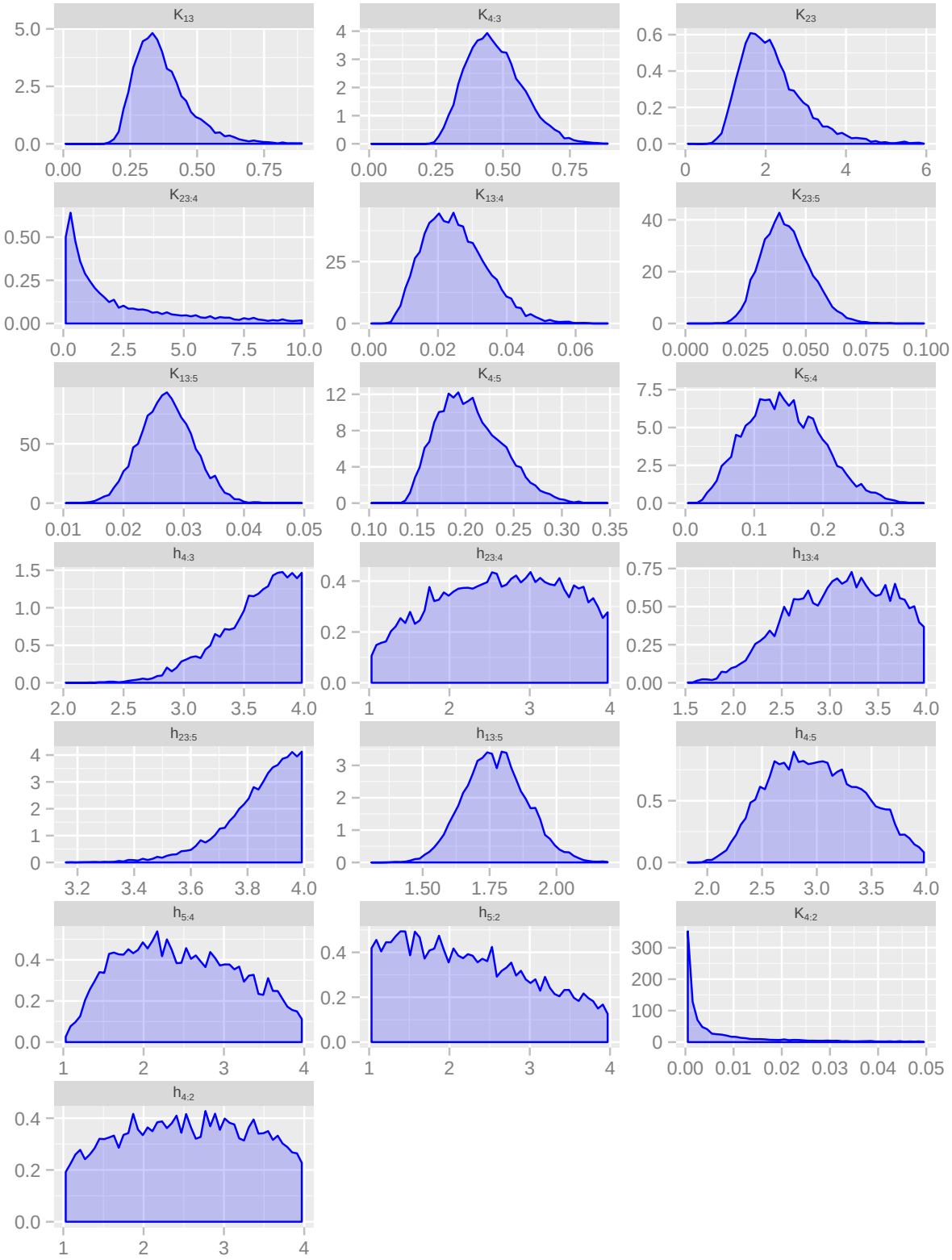


Figure 3.14: Estimated marginal posterior parameter distributions zoomed in on non-zero probability regions. No parameters are multimodal however not all have one strong mode. Some Hill parameters can take on most prior values with almost equal probability. Intriguingly this applies for $h_{4:2}$ which is the Hill term for binding of LFY on to the *TFL1* promoter whereas its corresponding kinetic parameter ($K_{4:2}$) is the mostly tightly controlled parameter. Y-axis values represent estimated probability densities with marginal distribution area equal to one.

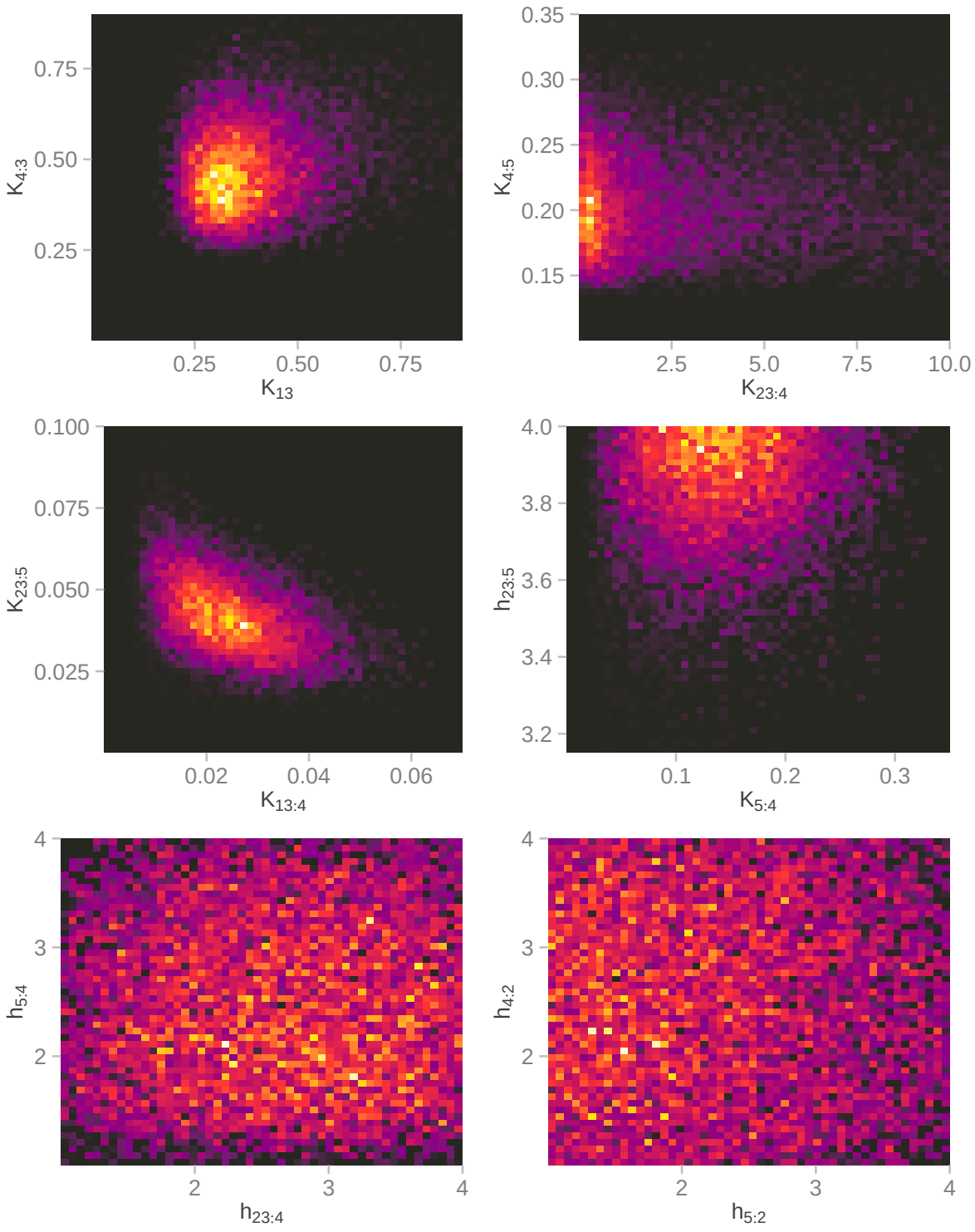


Figure 3.15: Estimated joint posterior parameter distributions for selected parameters. A sample of joint distributions of different parameter values gives a better understanding of parameter space. Brighter colours indicate higher relative probability.

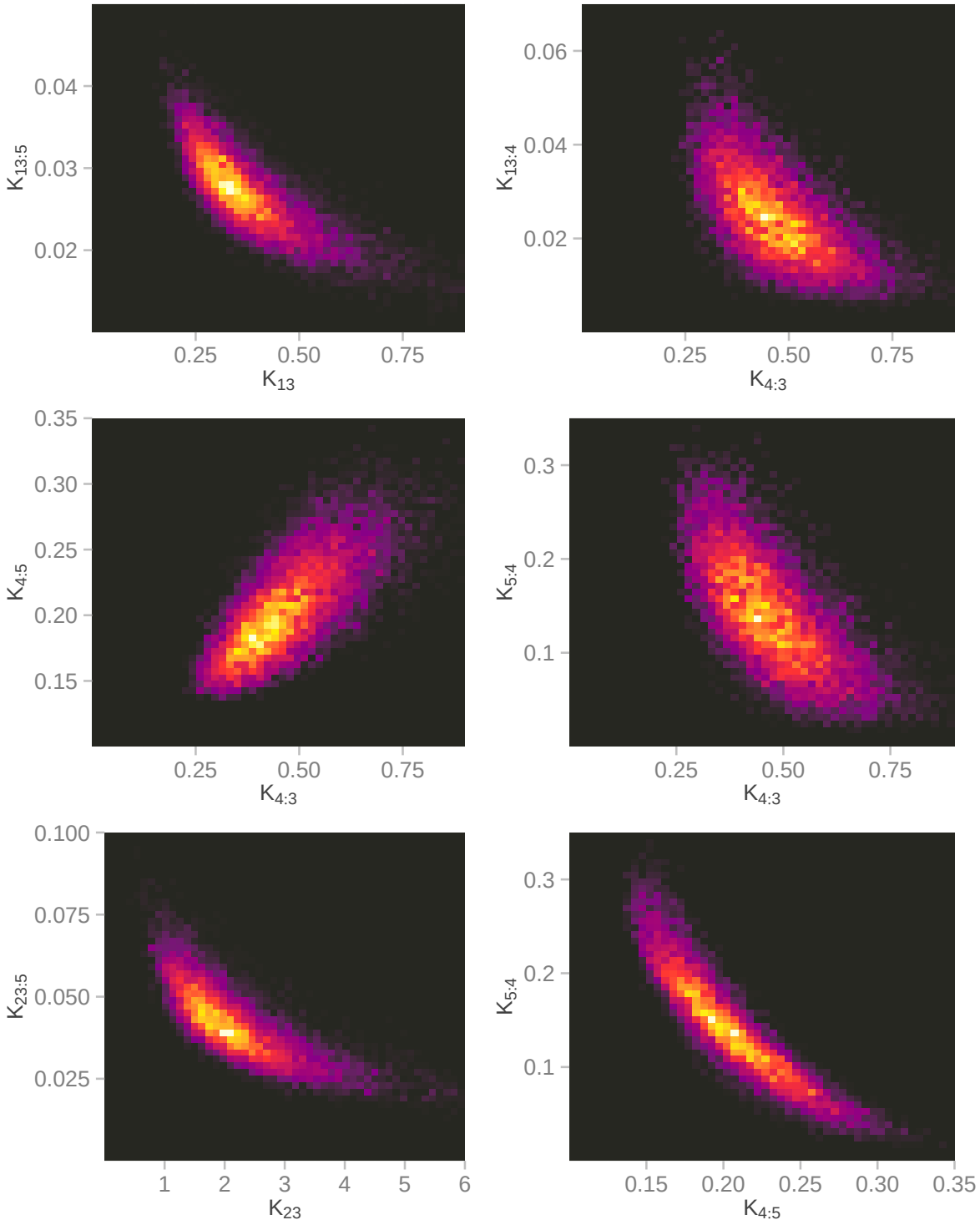


Figure 3.16: Estimated joint posterior parameter distributions showing correlations. All joint distributions were studied and the ones presented here show the most interesting correlations between parameters. Joint distributions can reveal hard to uncover information about parameter space such as correlations or multimodality. Brighter colours indicate higher relative probability.

Yeast two-hybrid assays have shown that FT interacts more strongly with FD than TFL1 in yeast [51, 191]. Looking at the marginal distributions in Figure 3.14 of the parameters controlling the same interactions, K_{13} and K_{23}, shows that the values for K_{23} are higher leading to its weaker binding to FD in our model. These results increase the belief that the model really captures some of the biology with how the network was constructed and the statistical treatment of the problem of parameter inference.

3.4.4 *A hint on spatial patterning of the SAM?*

Having established a network architecture that captures the temporal dynamics of cells undergoing the floral transition we also tentatively wondered if the model might also help us understand the spatial expression patterns of the main floral meristem regulators. This is a particularly interesting question because during development plants show sharp boundaries of gene expression in the SAM. During the floral transition initially diffuse and variable input signals, for example FT, gradually increase over time, leading to the expression of floral meristem genes such as *LFY* and *AP1* on the flanks of the shoot, while the centre of the shoot has rising *TFL1* expression and remains vegetative. It was hypothesised that low initial levels of the LFY or TFL1 hubs in the model might be sufficient to determine the stable acquisition of either a flowering (high AP1) or vegetative (high TFL1) state. Switching between these initial conditions the wildtype network has not been found to exhibit a bistable outcome between flowering or vegetative fates. Instead, a flowering state is often reached depending on the parameters (as discussed above).

This is expected for a number of reasons. As the temporal model has flowers as its output, as mentioned, it can be thought of as representing a cell on the flanks of the shoot apex poised to transition. These cells transition to a high AP1 state, but they do not experience extended upregulation of TFL1, since *TFL1* is repressed in floral meristems [79]. By contrast, within the centre of the shoot, *TFL1* is strongly upregulated upon flowering. It was found after experimental discussions that *TFL1* expression correlates across an entire Arabidopsis plant with the level of *FT* expression as shown in Figure 3.17 (exactly the same as Figure 2.3 in the previous chapter). The simplest

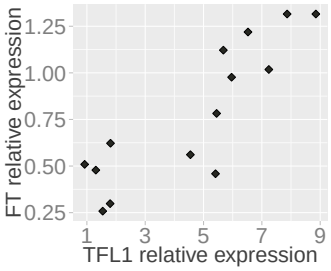


Figure 3.17: Reproduction of Figure 2.3. There is a clear positive relationship between *TFL1* and *FT* expression levels as determined by qPCR.

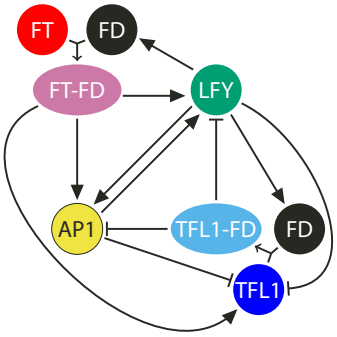


Figure 3.18: Enhanced regulatory network diagram. This network shows an additional connection between FTFD and TFL1. Filled arrowheads indicate activation and T-bars represent inhibition.

way of accounting for this behaviour in our model was to include a term for the activation of *TFL1* by FTFD in the network, as shown in Figure 3.18. Given our hub structure this need not be a direct interaction, indeed it surely involves a number of intermediaries *in planta*. In chapter 2 it was seen that for this data a sigmoidal model had a strong weight of evidence for it against a linear model. This suggests we can justifiably keep the similar structure based on Hill equations in terms of hub binding for this interaction. The addition of this connection requires four extra parameters for FTFD binding to the promoter of *TFL1* as described by the following equation

$$p_{13:2} = \frac{K_{23:2}^{h_{23:2}} x_{13}^{h_{13:2}}}{K_{13:2}^{h_{13:2}} K_{23:2}^{h_{23:2}} + K_{23:2}^{h_{23:2}} x_{13}^{h_{23:2}} + K_{13:2}^{h_{13:2}} x_{23}^{h_{23:2}}}.$$

To avoid complicating the model even further this term was simply multiplied by the doubly activated TFL1 rate which leads to a competition between activation through FTFD and repression from AP1 and LFY. Hence when AP1 and LFY are present at high concentration they will be dominantly repressive over the effect of direct FTFD-induced transcription of TFL1. The new term for TFL1 activation is then

$$\begin{aligned} v_2 &= v_{TFL1,35S} \\ &+ v_{TFL1,+}((1-p_{5:2})p_{4:2} + (1-p_{4:2})p_{5:2}) \\ &+ v_{TFL1,++} \cdot p_{5:2} \cdot p_{4:2} \cdot p_{13:2}. \end{aligned}$$

It was tested whether this network was still capable of fitting to the leaf numbers and if so, how its Bayesian evidence ranked compared to the two models previously considered. Running nested sampling on the same data set but with this extra connection in the model gave a log evidence of -59.87 ± 0.18. The previous model, with LFY feedback on to *FD*, had a log evidence of -62.68 ± 0.18 thus the Bayes factor is just under 3 — and on a normal scale is favoured 16 : 1. On Jeffreys' scale this would be strong evidence in favour of a model with this extra term despite the four additional parameters. What is the basis for this improvement in evidence? It is known (see subsection 1.5.3 and also MacKay [133]) the evidence comprises an Occam factor and a measure of goodness of fit. Here the former should penalise our extended model requiring four more parameters, and the latter should give more accurate leaf number estimates. Intriguingly much of the improvements come through superior estimates

Genotype	No. of rosette leaves			No. of cauline leaves			Data set
	True	Model		True	Model		
		Best-fit	Mean ± SD		Best-fit	Mean ± SD	
Wild type (Col)	7.9	8.9	8.7 ± 0.4	1.4	2.0	1.9 ± 0.09	Training
35S:*FT*	4.4	3.6	3.8 ± 0.2	1.0	1.6	1.7 ± 0.06	Training
35S:*LFY*	3.8	3.9	4.1 ± 0.1	1.8	1.5	1.6 ± 0.04	Training
35S:*TFL1*	27.5	26.8	28.1 ± 1.5	15.7	16.9	14.3 ± 1.9	Training
lfy-12	13.0	12.3	12.8 ± 0.7	5.3	6.2	6.6 ± 0.6	Training
ft-10	36.4	35.5	37.0 ± 1.4	9.3	8.6	8.8 ± 1.0	Training
tfl1-1	7.7	8.3	8.5 ± 0.4	0.4	1.9	1.9 ± 0.08	Training
fd-2	18.5	18.1	16.1 ± 0.8	4.63	3.9	3.4 ± 0.3	Training
fdp-1	11.2	9.8	9.4 ± 0.4	2.0	2.2	2.1 ± 0.1	Training
fd-2 fdp-1	32.9	32.8	31.6 ± 0.9	6.3	7.0	6.8 ± 0.6	Training
35S:*TFL1 fd-2*	23.8	25.3	26.0 ± 1.5	8.2	4.5	4.5 ± 0.4	Training
tfl1-1 fd-2	14.4	15.4	15.6 ± 0.7	4.6	3.4	3.3 ± 0.3	Training
35S:*FT fd-2*	8.3	8.3	7.0 ± 0.8	2.4	3.4	2.8 ± 0.3	Training
tfl1-1 fd-2 fdp-1	24.83	30.2	31.2 ± 1.0	6.67	6.6	6.7 ± 0.6	Prediction
35S:*TFL1 fd-2 fdp-1*	31.33	34.0	34.9 ± 1.1	11.0	7.2	7.3 ± 0.7	Prediction
35S:*FT fd-2 fdp-1*	25.8	28.5	26.1 ± 2.1	5.6	6.8	6.6 ± 0.6	Prediction

Table 3.5: Experimental and model leaf number data for the extended network with FTFD activating TFL1. For each genotype the table lists the mean experimental leaf number data and estimated (for the training set) or predicted best-fit and mean ± SD values for rosette and cauline leaves. The best-fit values use one set of parameters and thus have no possible associated error. This sample is taken from all the nested samples and is the one that maximises the likelihood function the most from the final set. Mean and SD based on 2000 posterior samples. SD, standard deviation.

of mutations solely affecting the FD hub Table 3.5. The genotypes *fd-2*, *fdp-1* and *fd-2 fdp-1* all have better estimated best-fits in the extended model, in particular the rosette leaves of the *fd* mutant. In the training set for all three network models the best-fit leaf number estimates of genotype *35S:TFL1 fd-2* are the source of the poorest fits. This suggests we haven't done as well capturing the suppression of *35S:TFL1* by *fd-2* as other genotypes. In terms of predicting the triple mutant leaf numbers there are no substantial improvements or regressions.

For different parameter values, the vegetative centre of the SAM was simulated, where *TFL1* is initially expressed at low levels and *LFY* is absent [77–79]. In this model scenario, rising levels of FT trigger the further upregulation of *TFL1*. The negative feedback of TFL1 onto both *AP1* and *LFY* prevents their expression (Figure 3.19 Top). Under the opposite starting conditions, low levels of LFY and no initial TFL1, corresponding to the primordium prior to floral evocation, rising FT activates AP1 and LFY. Since this is a positive feedback loop, high levels of AP1 and LFY are rapidly established, and TFL1 is repressed, leading to a floral state (Figure 3.19 Bottom). Although no parameter sets have been found where this is a stable state — at very high leaf numbers (when in practice the plant may have already died) this breaks down and AP1 reaches the flowering threshold — this can be seen as a starting hypothesis that warrants future investigations. Up until late developmental times essentially the winner takes all — vegetative or flowering programs are established — depending on the initial levels of either TFL1 or LFY and their sharp rise through the transition. This simulated outcome supports the model of Ratcliffe et al. [79] who suggest that one possibility for the spatial patterning is due to the relative timing of *TFL1* and the floral genes' induction, and subsequent mutual inhibition in the centre or periphery of the apex. In addition this gives us a lead to understanding the spatial patterning of the SAM where the activators of the transition must also cause a synchronous activation of their own repression in certain domains presumably due to floral signals being perceived by upstream regulators of meristem identity. This patterning mechanism has parallels with floral induction in tomato, where the floral signal SFT upregulates a repressor of floral meristem fate in lateral meristems adjacent to floral meristems [197]. Further understanding the system of apical patterning is an exciting goal for researchers in the future.

3.5 Discussion

3.5.1 *Strengths and limitations of the model*

A challenge to modelling complex biological systems, such as the floral transition, is that many interacting components are involved

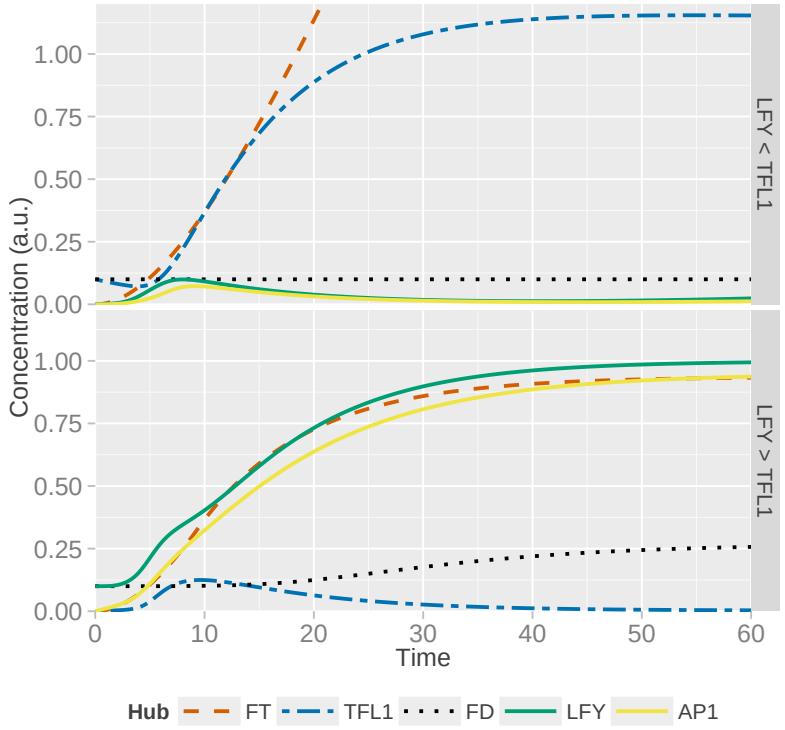


Figure 3.19: Initial conditions can determine apical cell state. For the timepoints shown a switch between initial conditions of LFY (green solid line) and TFL1 (blue dash-dot line) affect hub behaviour. Top: LFY starts at 0, TFL1 at 0.1. TFL1 is able to repress LFY and AP1 due to continued increase in FT levels. Bottom: Under opposite starting conditions where LFY is initially 0.1 and TFL1 is 0, LFY and AP1 are increased and flower normally while TFL1 is repressed.

and little is known in terms of their biophysical properties such as biochemical concentrations, binding affinities for each other or half lives within the cell. The mathematical modelling presented here thus involves considerable simplifications in terms of quantitative cell biology. We also simplified much hard work from excellent genetic studies in Arabidopsis with the reductionist approach — using knowledge of the major components and approximating key genes for entire hub activities. The list of things approximated and not modelled explicitly is vast. Therefore it is worth providing a brief list where more details are available in section 1.3. In Arabidopsis a number of pathways [5, 6] converge to stimulate flowering. This chapter focused on FT, a key floral integrator, as the input to the network. In one fell swoop this approximated the photoperiod pathway, whose output is diurnal *FT* expression; the vernalisation pathway[9]; and the autonomous pathway. The age-dependent and gibberellin pathways were accounted for by gradually rising levels of the AP1 and LFY hubs. Numerous other players in this web have been implicated: SPL transcription factors [35, 36], the floral integrator gene

[9] *Although the rapid-cycling wildtype Columbia-0 accession used as the genetic background for the genotypes in this thesis does not have this requirement.*

SOC1 [57], hormones such as auxin [28] and cytokinin [29], various microRNAs [4], ambient temperature [15, 18, 19] and the role of mechanical forces [82].

The model was motivated by known biological interactions and the idea of using the available leaf number data to make the predictions quantitative. This allowed us to train the network to the data and enabled the resulting model to suggest experiments that can be related back to biological entities. We did this by defining two thresholds at effectively arbitrary values of 0.2 and 0.3. Given the number of parameters in the model it is reasonable to believe that sensibly changing the values of these thresholds would just lead to a corresponding altering of the parameter distributions such that we still fit to the data at a similar level. The network model is also simple enough to understand some aspects of it intuitively because of the well-studied motifs it is based on. These advantages comes at a cost. First, by placing the network in an ordinary differential equation framework, we need to carry out computationally costly parameter space sampling. Although the employed nested sampling routine has been shown to perform very well for this purpose it is still true that we do not know whether or not our estimated parameter distributions are realistic or not. Second, the reduction to activity hubs means that our individual genes do not have direct *in planta* equivalents. Despite basing our equations on kinetic binding between proteins these are actually "hub proteins" and therefore an approximation of the effects of different proteins in the plant. Third, the model currently largely neglects important spatial effects. Although we can reproduce the overall behaviour of the transition, individual interactions represent spatially averaged behaviour and conclusions from this simplified network about such details must be considered carefully. For example we defined an appropriate network that represents well a single cell in the apex periphery that is capable of entering the transition. A cell elsewhere in the apex may, in fact, have a different set of connections between hubs and thus experience somewhat different behaviour.

Linear modelling of the system as presented in the early part of this chapter also suffers from these limitations, and more besides. The network approach taken here performs better than the linear

model because it helps us to understand and explain dynamic behaviour. The increase in parameters this required also gives us more flexibility in the fitting to the leaf number data. Thus we are able to be more accurate in estimating the training genotypes along with predictions that more closely match the triple mutant data — along with the individual rosette and cauline leaf data shown above compare Figure 3.20 with Figure 3.1 or Figure 3.2. It is also worth noting that all the network models had decisive evidence for them versus the linear model as judged on Jeffreys' scale.

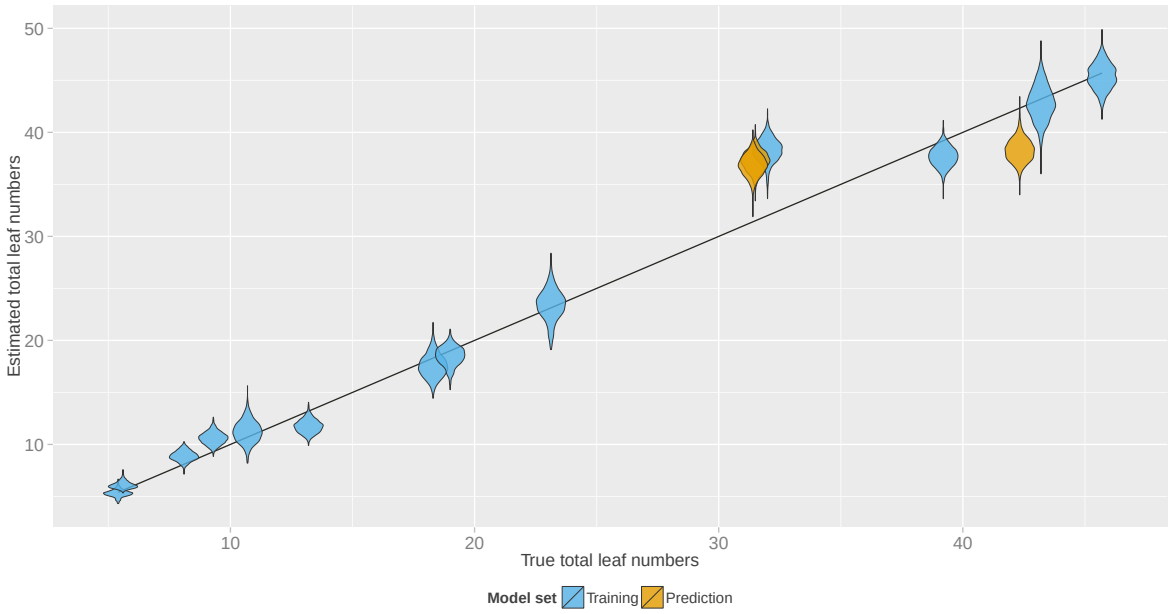


Figure 3.20: Posterior sampling of the extended model for total leaf numbers. Nested sampling was run with a likelihood function that minimised the sum of the model rosette and cauline leaves against the true total leaf number. 2000 posterior samples were taken and summary statistics calculated for each genotype. Compared with the linear models' summaries, Figure 3.1 and Figure 3.2, these estimates and predictions are far more reflective of the real data. We don't predict *35S:TFL1 fd-2 fdp-1* as accurately as in the linear model but we are now more accurate for the entire data set. The evidence for this model, -47.96 ± 0.15, is far better than the linear models' thus showing that, for this data set, the increase in parameters is quantitatively justifiable and that the network approach is considerably more powerful than linear modelling.

Our model is also extensible. Adding further hubs to this network, for example SOC1, is not too difficult and will lead to further testable hypotheses. More expansively, models for the circadian clock [110] upstream of the floral transition as well as models

for downstream processes such as MADS-box transcription factors specifying floral organs [106] could in theory all be coupled together to produce a more complete picture of floral morphogenesis at the SAM.

3.5.2 *Outlook*

Due to the clear mathematical foundations and robust statistical treatment of our modelling there are a number of directions this work could be expanded on in the future. In Arabidopsis two clear paths present a fork in the road depending on what one is trying to achieve. The main routes are either to attempt greater understanding of the detailed mechanisms involved on a simplified molecular level or to build a larger model with a greater set of features.

The first approach is the further minimisation of the presented reductionist model. When deciding upon our hubs and their structure the aim was to capture the core essentials of the flowering time network in the most basic possible model. However as touched upon when analysing the parameter distributions some parameters could potentially be removed from the system without loss of much flexibility. It would then be interesting to see the smallest network that could explain the data from the floral transition. One possibility is to remove the TFL1 hub from the network — revisiting our initial three-node motif — and define new equations to capture the data. Of course we would then have to remove the plants with altered *TFL1* expression from the data set, that is four out of 13 from the training set, and two of three from the prediction set. If this path is followed using a different data set should be strongly considered. Minimal models can help dissect complex regulation if good resolution quantitative data is available [111]. For instance a data set representing gene and protein abundance for the selected species over a number of timepoints will give far richer insights than leaf numbers alone for any flowering time modelling attempts.

The alternative path also requires a richer data set as it leads to an extended model. As mentioned our model could be coupled with others from the literature to provide a greater understanding of floral development and patterning. This would necessarily require some work in making sure all mathematical and biological assumptions

were similar between initial models and would no doubt leave a final model with a large number of parameters. A number of biological assumptions made in our simplification process could be reversed, perhaps independently of whether any models are chosen for coupling. A very interesting report from Melzer et al. [198] showed, amongst other things, that in the double-mutant *soc1-3 ful-2* the effect of *35S:FT* on flowering is largely suppressed. This suggests an important role for *SOC1* and *FUL* downstream of *FT*. How would this fit in with our hubs? Presently our model output is the AP1 hub which reflects the activities of, at least, *AP1*, *CAL* and *FUL* in Arabidopsis. A recent study [199] demonstrates the likelihood of SOC1 and FUL binding as heterodimers to the promoters of their target genes such as *LFY*. Given these data perhaps the assumption that they contribute to more than one hub could be made more explicit by adding them as extra factors. Though this of course leads to a thorny issue — how exactly is the morphology of flowers judged? At what point does a flower stop being recognised as such? In our model no architectural differences between inflorescences were accounted for. The fact that *lfy* mutant inflorescences look different to *ap1* mutant inflorescences [182] (of course not fogetting the amazing *ap1 cal ful* cauliflower-like inflorescence) may lead to headaches when specifying the outcome of such a model. A mention of architecture immediately brings one's thoughts to a question of space. The current model is temporal, and the propositions here extend this to tracking the dynamic behaviour of more proteins, and perhaps mRNAs, over time. Ultimately these genes are all interacting and possibly diffusing in the apex over time, and connections published in the literature may not be true for all cell types at all timepoints. This therefore highlights the need for future work in Arabidopsis, as the model dicot plant species, to ultimately focus on spatial cell-based models that account for the differential expression patterning of the key floral genes. Relevant experiments to inform these models would therefore include investigating the spatial and temporal dynamics of the relevant genes and proteins through techniques such as live imaging microscopy, laser microdissection and RNA sequencing throughout the initiation of the floral transition.

4 DISCUSSION & CONCLUSIONS

4.1 Modelling summary

Mathematical modelling is becoming an increasingly popular tool within the field of systems biology. Numerous modelling approaches are used in practice, ranging from machine learning of networks to partial differential equations (PDEs) on complex geometries. With a precise mathematical description of a problem, its solution can lead to unexplored research avenues or solve unexplained puzzles. In combination with experimentalists, iterative model building and biological verification or falsification can give greater knowledge about the system under study. Therefore as a function of increasing data, and by asking the right questions, scientific understanding can be increased through this interdisciplinary approach.

To this end, there has been some previous computational modelling of varying aspects of flowering time and flower development. For instance Espinosa-Soto et al. and van Mourik et al. took different approaches to build models that describe the regulation of floral organ specification [104, 106], while Prusinkiewicz et al. considered how differences in floral architecture may arise [103]. However there have been few mathematical descriptions that have focused on the dynamics of the floral transition.

In chapter 3 a reductionist approach was taken that enabled us to suggest clear experiments whilst not being overburdened by aiming for an all-encompassing model. A simplifying step of grouping genes with common or redundant function into regulatory hubs was taken. The effect of the various regulatory pathways that govern the floral transition was approximated by assuming they converge on the FT hub. Simplified models inevitably miss finer details of the biological system, yet they provide a tractable route to understanding the overall system behaviour. Though with this slight abstraction direct molecular relevance is lost, we stand to gain in terms of qualitative predictions that can be tested experimentally. To begin modelling a pathway, looking for the basic properties of simple networks that ex-

hibit the desired behaviour is a good first step. A simple three node system, as initially considered in subsection 3.2.4, can give intuitive understanding to many transcriptional or developmental networks, not just the floral transition. Starting from known components, the value of such a bottom-up approach lies in the simplicity and ease of computation, as modelling more complicated networks can require extensive computer simulations to illuminate their features. As with all simplifications, our network inevitably cannot account for the full spectrum of interacting pathways and variables seen in nature, but an experimental-modelling cycle can stimulate interesting questions that might otherwise be missed without modelling.

The models were made quantitative by scaling them to available leaf number data for a number of mutant genotypes. It was shown that a linear model is not sufficient to capture the variation in leaf numbers for the data set we had. The need for increased flexibility gave rise to a fairly simple network of core flowering time hubs that is able to capture important characteristics of the floral transition due to incorporated network motifs. Although the degree to which this behaviour manifests itself is parameter dependent, at a qualitative level this model is in agreement with many experimental observations. An intriguing feature of an extended network presented was that, for some parameter values, initial levels of LFY and TFL1 seem to control the determinacy of the cell for long developmental times. Thus this provides a hint on spatial patterning of the SAM. The type of model developed here is extensible in many directions and can provide increased power to scientists looking to develop yet deeper understanding of a crucial aspect of plant development.

4.2 Statistical summary

In this thesis Bayesian inference was used as the statistical framework of choice. Bayesian statistics allows one to place probability distributions on the elements that have some uncertainty attached to them, an obvious example being parameters. Typically, biological parameters such as degradation rates or binding constants are unknown experimentally and hence need to be inferred from the data. By being upfront about the prior assumptions our choices can be challenged by those with different degrees of belief and by sampling the poste-

rior distribution, predictions will be made that can give preference or not for a range of hypotheses.

A modern algorithm for Bayesian inference is nested sampling, which has seen success particularly in astrophysics due to the popular MultiNest implementation [164]. Nested sampling targets the important component for Bayesian model comparison by calculating the evidence — the posterior normalisation constant. In general this is a great challenge due to the need for evaluating this high-dimensional integral that arises through marginalising the likelihood over the prior. Not only can nested sampling effectively evaluate this term it also produce samples from the posterior distribution concurrently. In chapter 2 two major challenges in systems biology — parameter inference and model comparison — were addressed by the use of nested sampling. The summary statistics of parameters inferred by nested sampling were very similar to those calculated by MCMC, the go-to method for Bayesian computation. For a low dimensional example with experimental data where the evidence could be calculated by brute-force integration on a fine grid there was also good agreement between nested sampling and the numerical result. It was also shown how sampling from the posterior distribution can enable the reverse-engineering of the dynamics of the repressilator system despite little data.

Given noisy and sparse data a potential difficulty for the judicious modeller is the fair comparison of competing models. A set of four biological oscillators were compared with a limited data set from one variable. It was found that despite data being generated from the repressilator, the evidence gave preference to a different model. This also held true for published experimental data from the bacterial system, and with synthetic noiseless data up until a very high resolution timecourse was available. However when data was taken from two system variables, despite being few, the data were able to give a Bayes factor in preference of the known model. This has to be considered important for future experimental design.

The models for the floral transition in chapter 3 also benefited from the Bayesian method. In particular, despite the much better fit to the data with more parameters, the models avoided the curse of overfitting as even a model with 23 parameters was very strongly

favoured over a linear model with five parameters. For the leaf number data set available the model that best explained the data in a parsimonious way was able to fit accurately to these data and to qualitatively reproduce known properties of the floral transition. Model parameters were analysed by taking account of their marginal distribution or pairwise joint distributions. This analysis revealed a unimodal nature of the posterior parameter distribution and that there were some correlations between certain parameters. The probability density of some parameters was very high relative to others indicating that they are the most important for the model fit and thus tightly constrained by the data.

From the questions posed in this thesis it seems that nested sampling can blossom in the areas of computational and systems biology. This notwithstanding, there are a few other modern approaches for model comparison using MCMC such as annealed important sampling and thermodynamic integration [118, 154, 163, 171] that were not investigated here. These approaches for statistics examples are reviewed by Friel & Wyse [200]. For the problems herein considered nested sampling was shown to be accurate and, with the MultiNest implementation, it is an efficient algorithm, its core cost being in the computation of the log-likelihood function. It may not work as well as other techniques for certain situations but only time will tell as it gains popularity outside the physics community.

In summary, the use of Bayesian statistics through nested sampling allowed us fully quantify our uncertainty, compare models and infer parameters. Hence as more data and knowledge become available they can be used to update the models and refine our posterior inferences.

4.3 Outlook

4.3.1 *Temporal and spatial specificity*

Gene network diagrams are qualitative in nature and do not give any idea of time or space, thus missing potentially interesting dynamics. Hence it is important for further work in Arabidopsis to move away from the concept of a static GRN, whose structure does not vary with time. Although, as here, the dynamics of such networks

can be studied, it is important to recognise that the network structures themselves can be dynamic and vary over time. Simulating these dynamic networks will propose new theories and suggest new experiments that can lead to increased understanding of physiological networks.

Moreover, there should be a move towards simulating these dynamic networks in varying spatial domains. For example, the final model in chapter 3 provides a clue for future investigations into spatial patterning of the apex. However to prove or disprove the hypothesis requires a more advanced PDE approach where interacting proteins may function differently depending on their spatial localisation. Given the detailed molecular knowledge that continues to be discovered in model species it should be possible for future work in Arabidopsis to consider spatial and temporal specificity in more detail.

AP1 is a prime candidate in this regard. It was used as the output for our models so that levels of this protein could be mapped to different states, that is vegetative growth, bolting or flower production. Yet at a particular timepoint in primordium cells AP1 also represses certain flowering genes. It has been shown that *FD* is repressed in early floral primordium, around stage 2, the time when *AP1* expression is detected [50]. Thus, post-commitment, AP1 can be a repressor of meristem fate and later has other roles like activating genes required for floral morphogenesis [64]. Additionally, in the centre of the flowers A class *AP1* is repressed by C class *AGAMOUS* [183].

These results suggest manifold roles for a major floral gene that depend on developmental time and spatial localisation within the apex or establishing flower. In Arabidopsis there is detailed knowledge about these specificities, yet such details can not be captured in a single rigid representation of a genetic network. Henceforth mathematical modelling of this growing plant system should include both time and space dimensions.

4.3.2 *Outside of Arabidopsis*

In crops and other species, where there is not the detailed genetic knowledge available as in Arabidopsis, modelling GRNs can still have a great impact, thus outside of Arabidopsis the picture is rosy. The

modelling of flowering time in horticultural and agricultural crops generally takes a QTL and/or linear modelling approach. This is successful but misses much of the underlying biology. The incorporation of GRNs is little used in predictive crop scheduling or plant breeding. Yet in the era of genome-wide transcription factor binding maps and large-scale datasets, it is particularly timely to develop such approaches for other species. Orthologues of key genes considered herein such as *FT* and *TFL1* have been found in many species, for example tomato, wheat, barley, rose, apple and rice [44, 201–204].

In a number of polyploid species multiple copies of orthologous Arabidopsis genes occur. Therefore from a modelling viewpoint it may be possible to take a reductionist approach by grouping these genes into modules until the exact function of each copy is determined biologically. Simplifying the network to key hubs has the advantage of making it potentially easier to identify the critical network interactions that account for the major behaviours of a system. This can be used to further enhance the power of QTL-type approaches.

Additionally growing plants in their natural outside environment, and the effect this has on them, is far more relevant agriculturally than in stable laboratory conditions. How can this be tackled mathematically? The input to the models in chapter 3 is just FT levels. If it was known how environmental factors affected FT this could lead to the ability to predict the result in terms of leaf numbers and therefore developmental time, for many genotypes. Initially advanced growth chambers could be used that mimic outside conditions of light, temperature, rainfall and other components. Expression levels of *FT* could be recorded and correlated with these external influences and the resulting flowering time. The challenge is thus to characterise perturbations of the control variables such as temperature, CO_2 and water availability on the key inputs to the genetic networks and then to drive the change of these variables by climate models. This multi-scale approach will lead to the linking of the phenotype-based work in crops with molecular level research.

In many years' time plant breeders could benefit from knowing how predictions of a changing climate will affect the *in silico* dynamics of a modelled genotype, and thus steer their breeding programs appropriately. Furthermore combining field recording of weather

data and these ideas in important crop species could lead to live updating of computational models which will prove very helpful for farmers planning crop scheduling or harvesting at the optimal time.

All models rely on a number of parameters. Parameter determination or estimation is thus a key step towards predictions. No matter which species is chosen, being able to validate a model is vital and this includes independently evaluating the parameters in separate trials, both geographically and across seasons.

At a time when the changing climate is a hot topic the ability to build models for the regulation of developmental outcomes provides us with a means to test proposed genetic interaction networks and hence to understand which factors are affected most by environmental variability. Current crop models for predicting flowering time are highly valuable, however to fully exploit the wealth of genomic information that is becoming available these models need to bridge scales. Using gene network-based approaches should be able to calculate flowering time accurately as a function of different inputs, be they genotypic or environmental. This will lead to improvements for plant breeders and farmers as they look to feed the growing world population.

4.4 The project

4.4.1 *Evolution of the project*

Compared with sitting on a tractor cultivating fields in the relatively halcyon days pre-PhD this project, perhaps like many, has been an maelstrom of frustration, confusion and disappointment. Yet despite setbacks it is worth recording aspects that led to where we stand today and from where we can see a bright outlook on the horizon. In the early days a previous model for the floral transition developed by Richard Morris was tested. Before settling on the reduced models of chapter 3 many variants were tried, some wrong, some less wrong. For example initially it was considered that AP1 might activate *TFL1* — the reasoning being that both increase during the floral transition. Now it seems obvious that their distinct spatial expression domains can account for this observation. At the time, optimisation by simu-

lated annealing was used to estimate the parameters of our flowering models. However, we knew there was a better way.

This led to a side project that evolved into the results presented in chapter 2. Richard Morris had serendipitously coded a version of nested sampling in Fortran. The task was now to investigate parameter inference and model comparison in the Bayesian setting accounting for full uncertainty. After learning to program in Fortran and improving the original code, early tests proved very promising on simple examples and artificial data sets. It was also used successfully by Antonio Scialdone (JIC) for early investigations into a starch degradation model [205]. This gave us quite some confidence on the various applications and potential for using nested sampling more widely in systems biology modelling. It was actually after the initial submission of the paper describing our results that the MultiNest implementation [164] was tested. As it was even faster, offered more features and was easy to use given the previous fortuitous exposure to Fortran everything was redone using MultiNest, which enabled clearer results to be presented.

Nested sampling also allowed for a full re-analysis of the previous flowering time models. Hence the new results presented in chapter 3 are a culmination of over three and a half years hard work that unite two disparate ideas into a complete story.

4.4.2 *Continuation of the project*

Our success with nested sampling has also led to further use in the Morris group. Lydia Rickett (JIC) has done extensive investigations into models of bacterial growth curves with a large experimental data set. This has applications in the plant pathogen field and to food researchers where currently optimisation methods are used for parameter searching and models are compared with classical techniques. We are developing an `R` package for use by microbiologists so that they can easily compare different models of microbial growth with the posterior samples giving an idea of the uncertainty attached.

Finally Marc Jones (JIC) will hopefully enjoy a PhD on attempting to infer a flowering time network in oilseed rape (*Brassica napus*). As an allopolyploid this is a significant challenge with multiple copies of orthologous Arabidopsis genes like *FLC* and *FT*. These

different copies may exhibit sub-functionalisation and/or their effects may be different between cultivars particularly given the involvement of man for selection of the best lines in different continents. This work will hopefully build on the foundations laid by our Arabidopsis model in chapter 3 to see if it can be extended to help understand how other species effect the floral transition.

BIBLIOGRAPHY

[1] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts and P. Walter. **Molecular Biology of the Cell**. 5th ed. Garland Science, 2008 (see p. 12).

[2] Arabidopsis Genome Initiative. **Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana***. *Nature* 408.6814 (2000), 796–815 (see p. 12).

[3] F. Wellmer, M. Alves-Ferreira, A. Dubois, J. L. Riechmann and E. M. Meyerowitz. **Genome-wide analysis of gene expression during early *Arabidopsis* flower development**. *PLoS Genetics* 2.7 (2006), e117 (see p. 15).

[4] M. Schmid, N. H. Uhlenhaut, F. Godard, M. Demar, R. Bressan, D. Weigel and J. U. Lohmann. **Dissection of floral induction pathways using global expression analysis.** *Development* 130.24 (2003), 6001–12 (see pp. 15, 19, 122).

[5] G. G. Simpson and C. Dean. ***Arabidopsis*, the Rosetta stone of flowering time?** *Science* 296.5566 (2002), 285–9 (see pp. 15, 121).

[6] A. Srikanth and M. Schmid. **Regulation of flowering time: all roads lead to Rome**. *Cellular and Molecular Life Sciences* 68.12 (2011), 2013–2037 (see pp. 15, 16, 121).

[7] S. D. Michaels and R. M. Amasino. ***FLOWERING LOCUS C* encodes a novel MADS domain protein that acts as a repressor of flowering**. *The Plant Cell* 11.5 (1999), 949–956 (see p. 15).

[8] S. D. Michaels and R. M. Amasino. **Loss of *FLOWERING LOCUS C* activity eliminates the late-flowering phenotype of *FRIGIDA* and autonomous pathway mutations but not responsiveness to vernalization**. *The Plant Cell* 13.4 (2001), 935–941 (see pp. 15, 17).

[9] N. Schönrock, R. Bouveret, O. Leroy, L. Borghi, C. Köhler, W. Gruissem and L. Hennig. **Polycomb-group proteins repress the floral activator *AGL19* in the *FLC*-independent vernalization pathway**. *Genes & Development* 20.12 (2006), 1667–1678 (see p. 15).

[10] S. D. Michaels, G. Ditta, C. Gustafson-Brown, S. Pelaz, M. Yanofsky and R. M. Amasino. ***AGL24* acts as a promoter of flowering in *Arabidopsis* and is positively regulated by vernalization**. *The Plant Journal* 33.5 (2003), 867–874 (see p. 15).

[11] I. Searle, Y. He, F. Turck, C. Vincent, F. Fornara, S. Kröber, R. A. Amasino and G. Coupland. **The transcription factor FLC confers a flowering response to vernalization by repressing meristem competence and systemic signaling in *Arabidopsis***. *Genes & Development* 20.7 (2006), 898–912 (see p. 15).

[12] A. Angel, J. Song, C. Dean and M. Howard. **A Polycomb-based switch underlying quantitative epigenetic memory**. *Nature* 476.7358 (2011), 105–108 (see p. 17).

[13] D. Li, C. Liu, L. Shen, Y. Wu, H. Chen, M. Robertson, C. A. Helliwell, T. Ito, E. Meyerowitz and H. Yu. **A repressor complex governs the integration of flowering signals in *Arabidopsis***. *Developmental Cell* 15.1 (2008), 110–120 (see p. 17).

[14] J. H. Lee, S. J. Yoo, S. H. Park, I. Hwang, J. S. Lee and J. H. Ahn. **Role of *SVP* in the control of flowering time by ambient temperature in *Arabidopsis***. *Genes & Development* 21.4 (2007), 397–402 (see p. 17).

[15] M. A. Blázquez, J. H. Ahn and D. Weigel. **A thermosensory pathway controlling flowering time in *Arabidopsis thaliana***. *Nature Genetics* 33.2 (2003), 168–171 (see pp. 17, 122).

[16] R. Borner, G. Kampmann, J. Chandler, R. Gleißner, E. Wisman, K. Apel and S. Melzer. **A MADS domain gene involved in the transition to flowering in *Arabidopsis***. *The Plant Journal* 24.5 (2000), 591–599 (see pp. 17, 18, 21).

[17] A. Samach, H. Onouchi, S. E. Gold, G. S. Ditta, Z. Schwarz-Sommer, M. F. Yanofsky and G. Coupland. **Distinct roles of CONSTANS target genes in reproductive development of *Arabidopsis***. *Science* 288.5471 (2000), 1613–6 (see pp. 17, 20, 83).

[18] S. Balasubramanian, S. Sureshkumar, J. Lempe and D. Weigel. **Potent induction of *Arabidopsis thaliana* flowering by elevated growth temperature**. *PLoS Genetics* 2.7 (2006), e106 (see pp. 17, 108, 122).

[19] S. V. Kumar, D. Lucyshyn, K. E. Jaeger, E. Alós, E. Alvey, N. P. Harberd and P. A. Wigge. **Transcription factor PIF4 controls the thermosensory activation of flowering**. *Nature* 484.7393 (2012), 242–245 (see pp. 17, 122).

[20] D. Posé, L. Verhage, F. Ott, L. Yant, J. Mathieu, G. C. Angenent, R. G. Immink and M. Schmid. **Temperature-dependent regulation of flowering by antagonistic FLM variants**. *Nature* 503.7476 (2013), 414–414 (see pp. 17, 18).

[21] J. H. Lee, H.-S. Ryu, K. S. Chung, D. Posé, S. Kim, M. Schmid and J. H. Ahn. **Regulation of temperature-responsive flowering by MADS-box transcription factor repressors**. *Science* 342.6158 (2013), 628–632 (see pp. 17, 18).

[22] R. N. Wilson, J. W. Heckman and C. R. Somerville. **Gibberellin is required for flowering in *Arabidopsis thaliana* under short days**. *Plant Physiology* 100.1 (1992), 403–408 (see p. 18).

[23] F. Andrés, A. Porri, S. Torti, J. Mateos, M. Romera-Branchat, J. L. García-Martínez, F. Fornara, V. Gregis, M. M. Kater and G. Coupland. **SHORT VEGETATIVE PHASE reduces gibberellin biosynthesis at the *Arabidopsis* shoot apex to regulate the floral transition**. *Proceedings of the National Academy of Sciences of the USA* 111.26 (2014), E2760–E2769 (see p. 18).

[24] J. Moon, S.-S. Suh, H. Lee, K.-R. Choi, C. B. Hong, N.-C. Paek, S.-G. Kim and I. Lee. **The *SOC1* MADS-box gene integrates vernalization and gibberellin signals for flowering in *Arabidopsis***. *The Plant Journal* 35.5 (2003), 613–623 (see pp. 18, 21).

[25] M. A. Blázquez, R. Green, O. Nilsson, M. R. Sussman and D. Weigel. **Gibberellins promote flowering of Arabidopsis by activating the *LEAFY* promoter**. *The Plant Cell* 10.5 (1998), 791–800 (see pp. 18, 21).

[26] A. Porri, S. Torti, M. Romera-Branchat and G. Coupland. **Spatially distinct regulatory roles for gibberellins in the promotion of flowering of *Arabidopsis* under long photoperiods**. *Development* 139.12 (2012), 2198–2209 (see p. 18).

[27] N. Yamaguchi, C. M. Winter, M.-F. Wu, Y. Kanno, A. Yamaguchi, M. Seo and D. Wagner. **Gibberellin acts positively then negatively to control onset of flower formation in *Arabidopsis***. *Science* 344.6184 (2014), 638–641 (see pp. 18, 19).

[28] N. Yamaguchi, M.-F. Wu, C. M. Winter, M. C. Berns, S. Nole-Wilson, A. Yamaguchi, G. Coupland, B. A. Krizek and D. Wagner. **A molecular framework for auxin-mediated initiation of flower primordia**. *Developmental Cell* 24.3 (2013), 271–282 (see pp. 19, 122).

[29] M. D'Aloia, D. Bonhomme, F. Bouché, K. Tamseddak, S. Ormenese, S. Torti, G. Coupland and C. Périlleux. **Cytokinin promotes flowering of Arabidopsis via transcriptional activation of the *FT* paralogue *TSF***. *The Plant Journal* 65.6 (2011), 972–979 (see pp. 19, 21, 122).

[30] Z. Zhao, S. U. Andersen, K. Ljung, K. Dolezal, A. Miotk, S. J. Schultheiss and J. U. Lohmann. **Hormonal control of the shoot stem-cell niche**. *Nature* 465.7301 (2010), 1089–1092 (see p. 19).

[31] G. Bernier. **My favourite flowering image: the role of cytokinin as a flowering signal**. *Journal of Experimental Botany* 64.18 (2011), 5795–5799 (see p. 19).

[32] M. J. Aukerman and H. Sakai. **Regulation of flowering time and floral organ identity by a microRNA and its *APETALA2*-like target genes**. *The Plant Cell* 15.11 (2003), 2730–2741 (see p. 19).

[33] G. Wu and R. S. Poethig. **Temporal regulation of shoot development in *Arabidopsis thaliana* by *miR156* and its target *SPL3***. *Development* 133.18 (2006), 3539–3547 (see p. 19).

[34] G. Wu, M. Y. Park, S. R. Conway, J.-W. Wang, D. Weigel and R. S. Poethig. **The sequential action of miR156 and miR172 regulates developmental timing in *Arabidopsis***. *Cell* 138.4 (2009), 750–759 (see pp. 19, 20).

[35] J.-W. Wang, B. Czech and D. Weigel. **miR156-regulated SPL transcription factors define an endogenous flowering pathway in *Arabidopsis thaliana***. *Cell* 138.4 (2009), 738–749 (see pp. 19, 121).

[36] A. Yamaguchi, M.-F. Wu, L. Yang, G. Wu, R. S. Poethig and D. Wagner. **The microRNA-regulated SBP-box transcription factor SPL3 is a direct upstream activator of *LEAFY*, *FRUITFULL*, and *APETALA1***. *Developmental Cell* 17.2 (2009), 268–278 (see pp. 19, 121).

[37] P. Suárez-López, K. Wheatley, F. Robson, H. Onouchi, F. Valverde and G. Coupland. ***CONSTANS* mediates between the circadian clock and the control of flowering in *Arabidopsis***. *Nature* 410.6832 (2001), 1116–20 (see pp. 20, 108).

[38] R. Simon, M. Igeño and G. Coupland. **Activation of floral meristem identity genes in *Arabidopsis***. *Nature* 384.6604 (1996), 59 (see p. 20).

[39] H. An, C. Roussot, P. Suárez-López, L. Corbesier, C. Vincent, M. Piñeiro, S. Hepworth, A. Mouradov, S. Justin, C. Turnbull and G. Coupland. **CONSTANS acts in the phloem to regulate a systemic signal that induces photoperiodic flowering of *Arabidopsis***. *Development* 131.15 (2004), 3615–3626 (see p. 20).

[40] I. Kardailsky, V. K. Shukla, J. H. Ahn, N. Dagenais, S. K. Christensen, J. T. Nguyen, J. Chory, M. J. Harrison and D. Weigel. **Activation tagging of the floral inducer *FT***. *Science* 286.5446 (1999), 1962–1965 (see pp. 20, 21).

[41] Y. Kobayashi, H. Kaya, K. Goto, M. Iwabuchi and T. Araki. **A pair of related genes with antagonistic roles in mediating flowering signals**. *Science* 286.5446 (1999), 1960–1962 (see pp. 20, 23).

[42] F. Turck, F. Fornara and G. Coupland. **Regulation and identity of florigen: FLOWERING LOCUS T moves center stage.** *Annual Review of Plant Biology* 59 (2008), 573–594 (see pp. 20, 78).

[43] J. A. Higgins, P. C. Bailey and D. A. Laurie. **Comparative genomics of flowering time pathways using *Brachypodium distachyon* as a model for the temperate grasses**. *PLoS ONE* 5.4 (2010), e10065 (see pp. 20, 78, 83).

[44] S. Tamaki, S. Matsuo, H. L. Wong, S. Yokoi and K. Shimamoto. **Hd3a protein is a mobile flowering signal in rice**. *Science* 316.5827 (2007), 1033–1036 (see pp. 20, 83, 132).

[45] F. Valverde, A. Mouradov, W. Soppe, D. Ravenscroft, A. Samach and G. Coupland. **Photoreceptor regulation of CONSTANS protein in photoperiodic flowering**. *Science* 303.5660 (2004), 1003–1006 (see pp. 20, 83).

[46] L. Corbesier, I. Gadisseur, G. Silvestre, A. Jacqmard and G. Bernier. **Design in *Arabidopsis thaliana* of a synchronous system of floral induction by one long day**. *The Plant Journal* 9.6 (1996), 947–952 (see pp. 20, 87, 108).

[47] L. Corbesier, C. Vincent, S. Jang, F. Fornara, Q. Fan, I. Searle, A. Giakountis, S. Farrona, L. Gissot, C. Turnbull and G. Coupland. **FT protein movement contributes to long-distance signaling in floral induction of Arabidopsis**. *Science* 316.5827 (2007), 1030–1033 (see pp. 20, 83).

[48] J. Mathieu, N. Warthmann, F. Küttner and M. Schmid. **Export of FT protein from phloem companion cells is sufficient for floral induction in *Arabidopsis***. *Current Biology* 17.12 (2007), 1055–1060 (see pp. 20, 83).

[49] K. E. Jaeger and P. A. Wigge. **FT protein acts as a long-range signal in *Arabidopsis***. *Curr Biol* 17.12 (2007), 1050–1054 (see pp. 20, 83).

[50] P. A. Wigge, M. C. Kim, K. E. Jaeger, W. Busch, M. Schmid, J. U. Lohmann and D. Weigel. **Integration of spatial and temporal information during floral induction in *Arabidopsis***. *Science* 309.5737 (2005), 1056–1059 (see pp. 20, 79, 96, 100, 105, 131).

[51] M. Abe, Y. Kobayashi, S. Yamamoto, Y. Daimon, A. Yamaguchi, Y. Ikeda, H. Ichinoki, M. Notaguchi, K. Goto and T. Araki. **FD, a bZIP protein mediating signals from the floral pathway integrator FT at the shoot apex**. *Science* 309.5737 (2005), 1052–1056 (see pp. 20, 117).

[52] P. Teper-Bamnolker and A. Samach. **The flowering integrator FT regulates *SEPALLATA3* and *FRUITFULL* accumulation in *Arabidopsis* leaves**. *The Plant Cell* 17.10 (2005), 2661–2675 (see p. 20).

[53] K.-i. Taoka, I. Ohki, H. Tsuji, K. Furuita, K. Hayashi, T. Yanase, M. Yamaguchi, C. Nakashima, Y. A. Purwestri, S. Tamaki, Y. Ogaki, C. Shimada, A. Nakagawa, C. Kojima and K. Shimamoto. **14-3-3 proteins act as intracellular receptors for rice Hd3a florigen.** *Nature* 476.7360 (2011), 332–5 (see p. 20).

[54] A. Yamaguchi, Y. Kobayashi, K. Goto, M. Abe and T. Araki. ***TWIN SISTER OF FT* (*TSF*) acts as a floral pathway integrator redundantly with *FT***. *Plant & Cell Physiology* 46.8 (2005), 1175–89 (see pp. 20, 21, 79, 108).

[55] S. K. Yoo, K. S. Chung, J. Kim, J. H. Lee, S. M. Hong, S. J. Yoo, S. Y. Yoo, J. S. Lee and J. H. Ahn. ***CONSTANS* activates *SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1* through *FLOWERING LOCUS T* to promote flowering in Arabidopsis**. *Plant Physiology* 139.2 (2005), 770–8 (see p. 21).

[56] J. Lee, M. Oh, H. Park and I. Lee. **SOC1 translocated to the nucleus by interaction with AGL24 directly regulates *LEAFY***. *The Plant Journal* 55.5 (2008), 832–843 (see p. 21).

[57] C. Liu, H. Chen, H. L. Er, H. M. Soo, P. P. Kumar, J.-H. Han, Y. C. Liou and H. Yu. **Direct interaction of *AGL24* and *SOC1* integrates flowering signals in *Arabidopsis***. *Development* 135.8 (2008), 1481–1491 (see pp. 21, 122).

[58] H. Yu, T. Ito, F. Wellmer and E. M. Meyerowitz. **Repression of AGAMOUS-LIKE 24 is a crucial step in promoting flower development**. *Nature Genetics* 36.2 (2004), 157–161 (see p. 21).

[59] D. Weigel, J. Alvarez, D. R. Smyth, M. F. Yanofsky and E. M. Meyerowitz. ***LEAFY* controls floral meristem identity in Arabidopsis**. *Cell* 69.5 (1992), 843–859 (see p. 21).

[60] M. A. Blázquez, L. N. Soowal, I. Lee and D. Weigel. **LEAFY expression and flower initiation in Arabidopsis**. *Development* 124.19 (1997), 3835–3844 (see p. 21).

[61] F. Parcy, O. Nilsson, M. A. Busch, I. Lee and D. Weigel. **A genetic framework for floral patterning**. *Nature* 395.6702 (1998), 561–566 (see p. 21).

[62] S. J. Liljegren, C. Gustafson-Brown, A. Pinyopich, G. S. Ditta and M. F. Yanofsky. **Interactions among *APETALA1*, *LEAFY*, and *TERMINAL FLOWER1* specify meristem fate**. *The Plant Cell* 11.6 (1999), 1007–1018 (see p. 21).

[63] D. Wagner, R. W. Sablowski and E. M. Meyerowitz. **Transcriptional activation of APETALA1 by LEAFY**. *Science* 285.5427 (1999), 582–584 (see pp. 21, 90).

[64] K. Kaufmann, F. Wellmer, J. M. Muiño, T. Ferrier, S. E. Wuest, V. Kumar, A. Serrano-Mislata, F. Madueno, P. Krajewski, E. M. Meyerowitz, G. C. Angenent and J. L. Riechmanno. **Orchestration of floral initiation by APETALA1**. *Science* 328.5974 (2010), 85–89 (see pp. 21, 90, 131).

[65] C. Ferrándiz, Q. Gu, R. Martienssen and M. F. Yanofsky. **Redundant regulation of meristem identity and plant architecture by FRUITFULL, APETALA1 and CAULIFLOWER**. *Development* 127.4 (2000), 725–734 (see pp. 21, 79).

[66] C. Liu, J. Zhou, K. Bracha-Drori, S. Yalovsky, T. Ito and H. Yu. **Specification of *Arabidopsis* floral meristem identity by repression of flowering time genes**. *Development* 134.10 (2007), 1901–1910 (see p. 21).

[67] C. Smaczniak, R. G. Immink, J. M. Muiño, R. Blanvillain, M. Busscher, J. Busscher-Lange, Q. P. Dinh, S. Liu, A. H. Westphal, S. Boeren, et al. **Characterization of MADS-domain transcription factor complexes in *Arabidopsis* flower development**. *Proceedings of the National Academy of Sciences of the USA* 109.5 (2012), 1560–1565 (see p. 21).

[68] M. A. Busch, K. Bomblies and D. Weigel. **Activation of a floral homeotic gene in *Arabidopsis***. *Science* 285.5427 (1999), 585–587 (see p. 21).

[69] J. L. Bowman, D. R. Smyth and E. M. Meyerowitz. **Genetic interactions among floral homeotic genes of *Arabidopsis***. *Development* 112.1 (1991), 1–20 (see p. 21).

[70] E. S. Coen and E. M. Meyerowitz. **The war of the whorls: genetic interactions controlling flower development**. *Nature* 353.6339 (1991), 31–37 (see pp. 21, 30).

[71] G. N. Drews, J. L. Bowman and E. M. Meyerowitz. **Negative regulation of the Arabidopsis homeotic gene *AGAMOUS* by the *APETALA2* product**. *Cell* 65.6 (1991), 991–1002 (see p. 22).

[72] T. T. Dinh, T. Girke, X. Liu, L. Yant, M. Schmid and X. Chen. **The floral homeotic protein APETALA2 recognizes and acts through an AT-rich sequence element**. *Development* 139.11 (2012), 1978–1986 (see p. 22).

[73] S. Pelaz, G. S. Ditta, E. Baumann, E. Wisman and M. F. Yanofsky. **B and C floral organ identity functions require *SEPALLATA* MADS-box genes**. *Nature* 405.6783 (2000), 200–203 (see p. 22).

[74] G. Ditta, A. Pinyopich, P. Robles, S. Pelaz and M. F. Yanofsky. **The *SEP4* gene of *Arabidopsis thaliana* functions in floral organ and meristem identity**. *Current Biology* 14.21 (2004), 1935–1940 (see pp. 22, 31).

[75] S. Shannon and D. R. Meeks-Wagner. **A mutation in the Arabidopsis TFL1 gene affects inflorescence meristem development**. *Plant Cell* 3.9 (1991), 877–892 (see pp. 22, 23).

[76] Y. Hanzawa, T. Money and D. Bradley. **A single amino acid converts a repressor to an activator of flowering**. *Proceedings of the National Academy of Sciences of the USA* 102.21 (2005), 7748–7753 (see p. 23).

[77] D. Bradley, O. Ratcliffe, C. Vincent, R. Carpenter and E. Coen. **Inflorescence commitment and architecture in *Arabidopsis***. *Science* 275.5296 (1997), 80–83 (see pp. 23, 120).

[78] O. J. Ratcliffe, I. Amaya, C. A. Vincent, S. Rothstein, R. Carpenter, E. S. Coen and D. J. Bradley. **A common mechanism controls the life cycle and architecture of plants**. *Development* 125.9 (1998), 1609–1615 (see pp. 23, 29, 105, 120).

[79] O. J. Ratcliffe, D. J. Bradley and E. S. Coen. **Separation of shoot and floral identity in Arabidopsis.** *Development* 126.6 (1999), 1109–20 (see pp. 23, 29, 90, 105, 117, 120).

[80] C. M. Winter, R. S. Austin, S. Blanvillain-Baufumé, M. a. Reback, M. Monniaux, M.-F. Wu, Y. Sang, A. Yamaguchi, N. Yamaguchi, J. E. Parker, F. Parcy, S. T. Jensen, H. Li and D. Wagner. **LEAFY target genes reveal floral regulatory logic, *cis* motifs, and a link to biotic stimulus response**. *Developmental Cell* 20.4 (2011), 430–43 (see pp. 23, 100).

[81] L. Conti and D. Bradley. **TERMINAL FLOWER1 is a mobile signal controlling *Arabidopsis* architecture**. *Plant Cell* 19.3 (2007), 767–778 (see pp. 23, 105).

[82] O. Hamant, M. G. Heisler, H. Jönsson, P. Krupinski, M. Uyttewaal, P. Bokov, F. Corson, P. Sahlin, A. Boudaoud, E. M. Meyerowitz, Y. Couder and J. Traas. **Developmental patterning by mechanical signals in *Arabidopsis***. *Science* 322.5908 (2008), 1650–5 (see pp. 23, 122).

[83] R. S. Smith, S. Guyomarc'h, T. Mandel, D. Reinhardt, C. Kuhlemeier and P. Prusinkiewicz. **A plausible model of phyllotaxis**. *Proceedings of the National Academy of Sciences of the USA* 103.5 (2006), 1301–1306 (see p. 23).

[84] P. Barbier De Reuille, I. Bohn-Courseau, K. Ljung, H. Morin, N. Carraro, C. Godin and J. Traas. **Computer simulations reveal properties of the cell-cell signaling network at the shoot apex in *Arabidopsis***. *Proceedings of the National Academy of Sciences of the USA* 103.5 (2006), 1627–1632 (see p. 23).

[85] H. Jönsson, M. G. Heisler, B. E. Shapiro, E. M. Meyerowitz and E. Mjolsness. **An auxin-driven polarized transport model for phyllotaxis**. *Proceedings of the National Academy of Sciences of the USA* 103.5 (2006), 1633–1638 (see p. 23).

[86] G. L. Hammer, T. R. Sinclair, S. C. Chapman and E. Van Oosterom. **On systems thinking, systems biology, and the in silico plant**. *Plant Physiology* 134.3 (2004), 909–911 (see p. 24).

[87] X. Yin and P. C. Struik. **Modelling the crop: from system dynamics to systems biology**. *Journal of Experimental Botany* 61.8 (2010), 2171–2183 (see p. 24).

[88] 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* 4 (2013) (see p. 24).

[89] J. H. Thornley and I. R. Johnson. **Plant and Crop Modelling: A Mathematical Approach to Plant and Crop Physiology**. Clarendon Press, Oxford, 1990 (see p. 24).

[90] J.-F. Soussana, A.-I. Graux and F. N. Tubiello. **Improving the use of modelling for projections of climate change impacts on crops and pastures**. *Journal of Experimental Botany* 61.8 (2010), 2217–2228 (see p. 25).

[91] J. H. M. Thornley. **A model of a biochemical switch, and its application to flower initiation**. *Annals of Botany* 36.5 (1972), 861–871 (see p. 25).

[92] J. W. White, M. Herndl, L. Hunt, T. S. Payne and G. Hoogenboom. **Simulation-based analysis of effects of and loci on flowering in wheat**. *Crop Science* 48.2 (2008), 678–687 (see p. 26).

[93] J. W. Jones, G. Hoogenboom, C. H. Porter, K. J. Boote, W. D. Batchelor, L. Hunt, P. W. Wilkens, U. Singh, A. J. Gijsman and J. T. Ritchie. **The DSSAT cropping system model**. *European Journal of Agronomy* 18.3 (2003), 235–265 (see p. 26).

[94] C. Messina, J. Jones, K. Boote and C. Vallejos. **A gene-based model to simulate soybean development and yield responses to environment**. *Crop Science* 46.1 (2006), 456–466 (see p. 27).

[95] K. Boote, J. Jones, G. Hoogenboom and N. Pickering. **The CROPGRO model for grain legumes**. In: *Understanding Options for Agricultural Production*. 1998, 99–128 (see p. 27).

[96] X. Yin, P. C. Struik, J. Tang, C. Qi and T. Liu. **Model analysis of flowering phenology in recombinant inbred lines of barley**. *Journal of Experimental Botany* 56.413 (2005), 959–965 (see p. 27).

[97] X. Yin, P. C. Struik, F. A. van Eeuwijk, P. Stam and J. Tang. **QTL analysis and QTL-based prediction of flowering phenology in recombinant inbred lines of barley**. *Journal of Experimental Botany* 56.413 (2005), 967–976 (see pp. 27, 28).

[98] D. Wurr, J. R. Fellows and M. Fuller. **Simulated effects of climate change on the production pattern of winter cauliflower in the UK**. *Scientia Horticulturae* 101.4 (2004), 359–372 (see p. 27).

[99] R. Uptmoor, J. Li, T. Schrag and H. Stützel. **Prediction of flowering time in *Brassica oleracea* using a quantitative trait loci-based phenology model**. *Plant Biology* 14.1 (2012), 179–189 (see p. 27).

[100] R. Uptmoor, T. Schrag, H. Stützel and E. Esch. **Crop model based QTL analysis across environments and QTL based estimation of time to floral induction and flowering in *Brassica oleracea***. *Molecular Breeding* 21.2 (2008), 205–216 (see p. 27).

[101] R. Uptmoor, M. Osei-Kwarteng, S. Gürtler and H. Stützel. **Modeling the effects of drought stress on leaf development in a *Brassica oleracea* doubled haploid population using two-phase linear functions**. *Journal of the American Society for Horticultural Science* 134.5 (2009), 543–552 (see p. 28).

[102] S. M. Welch, J. L. Roe and Z. Dong. **A genetic neural network model of flowering time control in *Arabidopsis thaliana***. *Agronomy Journal* 95.1 (2003), 71–81 (see p. 28).

[103] P. Prusinkiewicz, Y. Erasmus, B. Lane, L. D. Harder and E. Coen. **Evolution and development of inflorescence architectures.** *Science* 316.5830 (2007), 1452–1456 (see pp. 29, 127).

[104] C. Espinosa-Soto, P. Padilla-Longoria and E. R. Alvarez-Buylla. **A gene regulatory network model for cell-fate determination during *Arabidopsis thaliana* flower development that is robust and recovers experimental gene expression profiles**. *The Plant Cell* 16.11 (2004), 2923–2939 (see pp. 30–32, 88, 127).

[105] T. Jack, L. L. Brockman and E. M. Meyerowitz. **The homeotic gene *APETALA3* of Arabidopsis thaliana encodes a MADS box and is expressed in petals and stamens**. *Cell* 68.4 (1992), 683–697 (see p. 30).

[106] S. van Mourik, A. D. J. van Dijk, M. de Gee, R. G. H. Immink, K. Kaufmann, G. C. Angenent, R. C. H. J. van Ham and J. Molenaar. **Continuous-time modeling of cell fate determination in Arabidopsis flowers.** *BMC Systems Biology* 4 (2010), 101 (see pp. 31, 33, 78, 124, 127).

[107] T. Honma and K. Goto. **Complexes of MADS-box proteins are sufficient to convert leaves into floral organs**. *Nature* 409.6819 (2001), 525–529 (see p. 31).

[108] M. Apri, M. de Gee, S. van Mourik and J. Molenaar. **Identifying optimal models to represent biochemical systems**. *PLoS ONE* 9.1 (2014), e83664 (see p. 31).

[109] U. Alon. **An Introduction to Systems Biology: Design Principles of Biological Circuits**. Taylor & Francis, 2006 (see pp. 33, 78, 85–88, 93).

[110] Y. H. Song, R. W. Smith, B. J. To, A. J. Millar and T. Imaizumi. **FKF1 conveys timing information for CONSTANS stabilization in photoperiodic flowering**. *Science* 336.6084 (2012), 1045–1049 (see pp. 33, 123).

[111] S. M. Murray, G. Panis, C. Fumeaux, P. H. Viollier and M. Howard. **Computational and genetic reduction of a cell cycle to its simplest, primordial components**. *PLoS Biology* 11.12 (2013), e1001749 (see pp. 33, 124).

[112] A. C. Hindmarsh, P. N. Brown, K. E. Grant, S. L. Lee, R. Serban, D. E. Shumaker and C. S. Woodward. **SUNDIALS: Suite of Nonlinear and Differential/Algebraic Equation Solvers**. *ACM Transactions on Mathematical Software* 31.3 (2005), 363–396 (see p. 33).

[113] P. Mendes and D. Kell. **Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation**. *Bioinformatics* 14.10 (1998), 869–883 (see pp. 33, 34).

[114] C. Moles, P. Mendes and J. Banga. **Parameter estimation in biochemical pathways: a comparison of global optimization methods**. *Genome Research* 13.11 (2003), 2467–2474 (see pp. 33, 34).

[115] M. Ashyraliyev, Y. Fomekong-Nanfack, J. Kaandorp and J. Blom. **Systems biology: parameter estimation for biochemical models**. *FEBS Journal* 276.4 (2009), 886–902 (see pp. 33, 34).

[116] J. Banga. **Optimization in computational systems biology**. *BMC Systems Biology* 2.1 (2008), 47 (see p. 34).

[117] N. Dalchau. **Understanding biological timing using mechanistic and black-box models**. *The New Phytologist* 193.4 (2012), 852–8 (see p. 34).

[118] B. Calderhead and M. Girolami. **Estimating Bayes factors via thermodynamic integration and population MCMC**. *Computational Statistics & Data Analysis* 53.12 (2009), 4028–4045 (see pp. 34, 36, 130).

[119] C. A. Floudas and C. E. Gounaris. **A review of recent advances in global optimization**. *Journal of Global Optimization* 45.1 (2009), 3–38 (see p. 34).

[120] S. Kirkpatrick, C. Gelatt and M. Vecchi. **Optimization by simmulated annealing**. *Science* 220.4598 (1983), 671–680 (see pp. 34, 56, 102).

[121] M. Schwaab, E. C. Biscaia, Jr., J. L. Monteiro and J. C. Pinto. **Nonlinear parameter estimation through particle swarm optimization**. *Chemical Engineering Science* 63.6 (2008), 1542–1552 (see p. 34).

[122] G. Lillacci and M. Khammash. **Parameter estimation and model selection in computational biology**. *PLoS Computational Biology* 6.3 (2010), e1000696 (see pp. 34, 59).

[123] M. Quach, N. Brunel and F. d'Alché-Buc. **Estimating parameters and hidden variables in non-linear state-space models based on ODEs for biological networks inference**. *Bioinformatics* 23.23 (2007), 3209–3216 (see pp. 34, 59).

[124] T. Toni, D. Welch, N. Strelkowa, A. Ipsen and M. Stumpf. **Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems**. *Journal of the Royal Society Interface* 6.31 (2009), 187–202 (see pp. 34, 36, 59).

[125] E. Granqvist, G. Oldroyd and R. Morris. **Automated Bayesian model development for frequency detection in biological time series**. *BMC Systems Biology* 5.1 (2011), 97 (see p. 34).

[126] S. Forrest. **Genetic algorithms: principles of natural selection applied to computation**. *Science* 261.5123 (1993), 872–878 (see p. 34).

[127] A. R. Mehrabian and C. Lucas. **A novel numerical optimization algorithm inspired from weed colonization**. *Ecological Informatics* 1.4 (2006), 355–366 (see p. 34).

[128] D. Slezak, C. Suárez, G. Cecchi, G. Ma rshall and G. Stolovitzky. **When the optimal is not the best: parameter estimation in complex biological models**. *PLoS ONE* 5.10 (2010), e13283 (see p. 34).

[129] D. M. Hawkins. **The problem of overfitting**. *Journal of Chemical Information and Computer Sciences* 44.1 (2004), 1–12 (see p. 34).

[130] H. Akaike. **Information theory and an extension of the maximum likelihood principle**. In: *Second International Symposium on Information Theory*. 1973, 267–281 (see p. 34).

[131] H. Akaike. **A new look at the statistical model identification**. *IEEE Transactions on Automatic Control* 19.6 (1974), 716–723 (see p. 34).

[132] P. Kirk, T. Thorne and M. P. Stumpf. **Model selection in systems and synthetic biology**. *Current Opinion in Biotechnology* 24.4 (2013), 767–774 (see p. 34).

[133] D. MacKay. **Information Theory, Inference, and Learning Algorithms**. Cambridge University Press, 2003 (see pp. 34, 36, 38, 41, 56, 70, 75, 102, 118).

[134] K. Erguler and M. P. H. Stumpf. **Practical limits for reverse engineering of dynamical systems: a statistical analysis of sensitivity and parameter inferability in systems biology models**. *Molecular BioSystems* 7 (5 2011), 1593–1602 (see pp. 34, 53, 62).

[135] 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 Computational Biology* 3.10 (2007), e189 (see pp. 34, 53, 62).

[136] M. Komorowski, M. J. Costa, D. A. Rand and M. P. H. Stumpf. **Sensitivity, robustness, and identifiability in stochastic chemical kinetics models**. *Proceedings of the National Academy of Sciences of the USA* 108.21 (2011), 8645–8650 (see p. 34).

[137] A. F. Villaverde and J. R. Banga. **Reverse engineering and identification in systems biology: strategies, perspectives and challenges**. *Journal of The Royal Society Interface* 11.91 (2014), 20130505 (see p. 35).

[138] H. Jeffreys. **Theory of Probability**. Oxford University Press, 1961 (see pp. 35, 36, 38, 39, 81).

[139] E. Jaynes and G. Bretthorst. **Probability Theory: The Logic of Science**. Cambridge University Press, 2003 (see pp. 35, 37).

[140] R. Neal. **Philosophy of Bayesian Inference**. 1998. URL: `http://www.cs.toronto.edu/~radford/res-bayes-ex.html` (visited on 08/28/2014) (see p. 35).

[141] R. T. Bayes. **An essay towards solving a problem in the doctrine of chances**. *Philosophical Transactions of the Royal Society of London* 53 (1763), 370–481 (see p. 35).

[142] D. V. Lindley. **Bayesian Statistics, A Review**. Society for Industrial and Applied Mathematics, 1972 (see p. 35).

[143] A. Blumer, A. Ehrenfeucht, D. Haussler and M. Warmuth. **Occam's razor**. *Information Processing Letters* 24.6 (1987), 377–380 (see p. 35).

[144] C. E. Rasmussen and Z. Ghahramani. **Occam's razor**. In: *Advances in Neural Information Processing Systems*. 2001, 294–300 (see p. 35).

[145] D. J. MacKay. **Bayesian interpolation**. *Neural Computation* 4 (1991), 415–447 (see p. 36).

[146] D. Posada and T. Buckley. **Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests**. *Systematic Biology* 53.5 (2004), 793–808 (see pp. 36, 75).

[147] D. Wilkinson. **Bayesian methods in bioinformatics and computational systems biology**. *Briefings in Bioinformatics* 8.2 (2007), 109–116 (see p. 36).

[148] P. Baldi and A. D. Long. **A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes**. *Bioinformatics* 17.6 (2001), 509–519 (see p. 36).

[149] T. Thorne and M. P. H. Stumpf. **Inference of temporally varying Bayesian networks**. *Bioinformatics* 28.24 (2012), 3298–305 (see p. 36).

[150] B. Calderhead and M. Girolami. **Statistical analysis of nonlinear dynamical systems using differential geometric sampling methods**. *Interface Focus* 1.6 (2011), 821–35 (see p. 36).

[151] H. Eydgahi, W. W. Chen, J. L. Muhlich, D. Vitkup, J. N. Tsitsiklis and P. K. Sorger. **Properties of cell death models calibrated and compared using Bayesian approaches**. *Molecular Systems Biology* 9.1 (2013) (see p. 36).

[152] W. Heuett, B. Miller, S. Racette, J. Holloszy, C. Chow and V. Periwal. **Bayesian functional integral method for inferring continuous data from discrete measurements**. *Biophysical Journal* 102.3 (2012), 399–406 (see p. 36).

[153] D. Schmidl, S. Hug, W. B. Li, M. B. Greiter and F. J. Theis. **Bayesian model selection validates a biokinetic model for zirconium processing in humans**. *BMC Systems Biology* 6 (2012), 95 (see p. 36).

[154] N. Lartillot and H. Philippe. **Computing Bayes factors using thermodynamic integration**. *Systematic Biology* 55.2 (2006), 195–207 (see pp. 36, 130).

[155] J. Skilling. **Nested sampling for general Bayesian computation**. *Bayesian Analysis* 1.4 (2006), 833–860 (see pp. 36, 41, 42, 45, 48, 50, 52, 56).

[156] D. Sivia and J. Skilling. **Data Analysis: A Bayesian Tutorial**. Oxford University Press, 2006 (see pp. 36, 37, 41, 42, 45, 46, 48, 49, 52).

[157] P. Mukherjee, D. Parkinson and A. R. Liddle. **A nested sampling algorithm for cosmological model selection**. *The Astrophysical Journal Letters* 638.2 (2006), L51 (see pp. 36, 41, 46).

[158] F. Feroz and M. Hobson. **Multimodal nested sampling: an efficient and robust alternative to Markov Chain Monte Carlo methods for astronomical data analyses**. *Monthly Notices of the Royal Astronomical Society* 384.2 (2008), 449–463 (see pp. 36, 41, 46).

[159] L. B. Pártay, A. P. Bartók and G. Csányi. **Efficient sampling of atomic configurational spaces**. *The Journal of Physical Chemistry B* 114.32 (2010), 10502–12 (see pp. 36, 41, 48).

[160] S. Aitken and O. Akman. **Nested sampling for parameter inference in systems biology: application to an exemplar circadian model**. *BMC Systems Biology* 7.1 (2013), 72 (see pp. 36, 41, 48, 50).

[161] N. Burkoff, C. Várnai, S. Wells and D. Wild. **Exploring the energy landscapes of protein folding simulations with Bayesian computation**. *Biophysical Journal* 102.4 (2012), 878–886 (see pp. 36, 41, 48, 50).

[162] R. Kass and A. Raftery. **Bayes factors**. *Journal of the American Statistical Association* (1995), 773–795 (see pp. 38, 39).

[163] I. Murray. **Advances in Markov chain Monte Carlo methods**. PhD thesis. Gatsby Computational Neuroscience Unit, University College London, 2007 (see pp. 41, 130).

[164] F. Feroz, M. Hobson and M. Bridges. **MULTINEST: an efficient and robust Bayesian inference tool for cosmology and particle physics**. *Monthly Notices of the Royal Astronomical Society* 398.4 (2009), 1601–1614 (see pp. 41, 46–48, 50, 52, 129, 134).

[165] F. Feroz, M. Hobson, E. Cameron and A. Pettitt. **Importance nested sampling and the MultiNest algorithm**. *arXiv:1306.2144* (2013) (see pp. 47, 48).

[166] K. E. Jaeger, N. Pullen, S. Lamzin, R. J. Morris and P. A. Wigge. **Interlocking feedback loops govern the dynamic behavior of the floral transition in *Arabidopsis***. *The Plant Cell* 25.3 (2013), 820–833 (see pp. 47, 78, 100).

[167] J. Skilling. **Nested sampling's convergence**. In: *The 29th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*. Vol. 1193. 1. 2009, 277–291 (see p. 49).

[168] M. Ashyraliyev, J. Jaeger and J. Blom. **Parameter estimation and determinability analysis applied to *Drosophila* gap gene circuits**. *BMC Systems Biology* 2.1 (2008), 83 (see pp. 53, 62).

[169] W. L. Goffe, G. D. Ferrier and J. Rogers. **Global optimization of statistical functions with simulated annealing**. *Journal of Econometrics* 60.1 (1994), 65–99 (see pp. 56, 102).

[170] M. B. Elowitz and S. Leibler. **A synthetic oscillatory network of transcriptional regulators**. *Nature* 403.6767 (2000), 335–8 (see pp. 59, 71, 72).

[171] V. Vyshemirsky and M. Girolami. **Bayesian ranking of biochemical system models**. *Bioinformatics* 24.6 (2008), 833–839 (see pp. 59, 130).

[172] A. Pokhilko, A. Fernández, K. Edwards, M. Southern, K. Halliday and A. Millar. **The clock gene circuit in *Arabidopsis* includes a repressilator with additional feedback loops**. *Molecular Systems Biology* 8.1 (2012) (see p. 59).

[173] A. Lotka. **Elements of physical biology**. Williams & Wilkins Baltimore, 1925 (see p. 65).

[174] V. Volterra. **Variazioni e fluttuazioni del numero d'individui in specie animali conviventi**. *Memorie della R. Acc. dei Lincei* 2 (1926), 31–113 (see p. 65).

[175] B. Goodwin. **Temporal organization in cells: a dynamic theory of cellular control processes**. Academic Press, London, 1963 (see p. 65).

[176] L. Edelstein-Keshet. **Mathematical Models in Biology**. Random House, 1988 (see p. 65).

[177] J. Schnakenberg. **Simple chemical reaction systems with limit cycle behaviour**. *Journal of Theoretical Biology* 81.3 (1979), 389–400 (see p. 65).

[178] J. Murray. **Mathematical Biology: I. An Introduction**. Springer, 2002 (see p. 65).

[179] J. Liepe, S. Filippi, M. Komorowski and M. P. H. Stumpf. **Maximizing the information content of experiments in systems biology**. *PLoS Computational Biology* 9.1 (2013), e1002888 (see p. 71).

[180] K. Burnham and D. Anderson. **Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach**. Springer, 2002 (see p. 75).

[181] W. Link and R. Barker. **Model weights and the foundations of multimodel inference**. *Ecology* 87.10 (2006), 2626–2635 (see p. 75).

[182] J. L. Bowman, J. Alvarez, D. Weigel, E. M. Meyerowitz and D. R. Smyth. **Control of flower development in *Arabidopsis thaliana* by *APETALA 1* and interacting genes**. *Development* 119 (1993), 721–743 (see pp. 78, 125).

[183] C. Gustafson-Brown, B. Savidge and M. F. Yanofsky. **Regulation of the Arabidopsis floral homeotic gene *APETALA1***. *Cell* 76.1 (1994), 131–143 (see pp. 79, 131).

[184] M. A. Mandel, C. Gustafson-Brown, B. Savidge and M. F. Yanofsky. **Molecular characterization of the *Arabidopsis* floral homeotic gene *APETALA1***. *Nature* 360 (1992), 273–277 (see pp. 79, 90).

[185] R Core Team. **R: A Language and Environment for Statistical Computing**. R Foundation for Statistical Computing. Vienna, Austria, 2013 (see p. 83).

[186] S. Mangan and U. Alon. **Structure and function of the feed-forward loop network motif**. *Proceedings of the National Academy of Sciences of the USA* 100.21 (2003), 11980–11985 (see pp. 85–87).

[187] J. Hasty, J. Pradines, M. Dolnik and J. J. Collins. **Noise-based switches and amplifiers for gene expression**. *Proceedings of the National Academy of Sciences of the USA* 97.5 (2000), 2075–2080 (see p. 85).

[188] W.-G. Choi, M. Toyota, S.-H. Kim, R. Hilleary and S. Gilroy. **Salt stress-induced Ca^{2+} waves are associated with rapid, long-distance root-to-shoot signaling in plants**. *Proceedings of the National Academy of Sciences of the USA* 111.17 (2014), 6497–6502 (see p. 87).

[189] F. D. Hempel, D. Weigel, M. A. Mandel, G. Ditta, P. C. Zambryski, L. J. Feldman and M. F. Yanofsky. **Floral determination and expression of floral regulatory genes in Arabidopsis**. *Development* 124.19 (1997), 3845–3853 (see p. 90).

[190] S. Pouteau and C. Albertini. **The significance of bolting and floral transitions as indicators of reproductive phase change in emphArabidopsis**. *Journal of Experimental Botany* 60.12 (2009), 3367–3377 (see p. 92).

[191] S. Hanano and K. Goto. ***Arabidopsis* TERMINAL FLOWER1 is involved in the regulation of flowering time and inflorescence development through transcriptional repression**. *The Plant Cell* 23.9 (2011), 3172–3184 (see pp. 94, 112, 117).

[192] B. Steiert, A. Raue, J. Timmer and C. Kreutz. **Experimental design for parameter estimation of gene regulatory networks**. *PLoS ONE* 7.7 (2012), e40052 (see p. 99).

[193] K. Sneppen, M. A. Micheelsen and I. B. Dodd. **Ultrasensitive gene regulation by positive feedback loops in nucleosome modification**. *Molecular Systems Biology* 4.1 (2008) (see p. 99).

[194] S. Grandison and R. J. Morris. **Biological pathway kinetic rate constants are scale-invariant**. *Bioinformatics* 24.6 (2008), 741–743 (see p. 99).

[195] W. Kim, T. Im Park, S. J. Yoo, A. R. Jun and J. H. Ahn. **Generation and analysis of a complete mutant set for the *Arabidopsis FT/TFL1* family shows specific effects on thermo-sensitive flowering regulation**. *Journal of experimental botany* 64.6 (2013), 1715–1729 (see p. 106).

[196] J. F. Sundström, N. Nakayama, K. Glimelius and V. F. Irish. **Direct regulation of the floral homeotic *APETALA1* gene by APETALA3 and PISTILLATA in Arabidopsis**. *The Plant Journal* 46.4 (2006), 593–600 (see p. 108).

[197] J. Thouet, M. Quinet, S. Lutts, J.-M. Kinet and C. Périlleux. **Repression of floral meristem fate is crucial in shaping tomato inflorescence**. *PLoS ONE* 7.2 (2012), e31096 (see p. 120).

[198] S. Melzer, F. Lens, J. Gennen, S. Vanneste, A. Rohde and T. Beeckman. **Flowering-time genes modulate meristem determinacy and growth form in *Arabidopsis thaliana***. *Nature Genetics* 40.12 (2008), 1489–92 (see p. 125).

[199] V. Balanzà, I. Martínez-Fernández and C. Ferrándiz. **Sequential action of *FRUITFULL* as a modulator of the activity of the floral regulators *SVP* and *SOC1***. *Journal of Experimental Botany* 65.4 (2014), 1193 (see p. 125).

[200] N. Friel and J. Wyse. **Estimating the evidence – a review**. *Statistica Neerlandica* 66.3 (2012), 288–308 (see p. 130).

[201] A. Shalit, A. Rozman, A. Goldshmidt, J. P. Alvarez, J. L. Bowman, Y. Eshed and E. Lifschitz. **The flowering hormone florigen functions as a general systemic regulator of growth and termination**. *Proceedings of the National Academy of Sciences of the USA* 106.20 (2009), 8392–8397 (see p. 132).

[202] L. Yan, D. Fu, C. Li, A. Blechl, G. Tranquilli, M. Bonafede, A. Sanchez, M. Valarik, S. Yasuda and J. Dubcovsky. **The wheat and barley vernalization gene *VRN3* is an orthologue of *FT***. *Proceedings of the National Academy of Sciences of the USA* 103.51 (2006), 19581–19586 (see p. 132).

[203] M. Randoux, J.-M. Davière, J. Jeauffre, T. Thouroude, S. Pierre, Y. Toualbia, J. Perrotte, J.-P. Reynoird, M.-J. Jammes, H.-S. Oyant and F. Foucher. **RoKSN, a floral repressor, forms protein complexes with RoFD and RoFT to regulate vegetative and reproductive development in rose**. *New Phytologist* 202.1 (2013), 161–173 (see p. 132).

[204] N. Mimida, N. Kotoda, T. Ueda, M. Igarashi, Y. Hatsuyama, H. Iwanami, S. Moriya and K. Abe. **Four *TFL1/CEN*-like genes on distinct linkage groups show different expression patterns to regulate vegetative and reproductive development in apple (*Malus*× *domestica* borkh.)** *Plant and Cell Physiology* 50.2 (2009), 394–412 (see p. 132).

[205] A. Scialdone, S. T. Mugford, D. Feike, A. Skeffington, P. Borrill, A. Graf, A. M. Smith and M. Howard. **Arabidopsis plants perform arithmetic division to prevent starvation at night**. *eLife* 2 (2013) (see p. 134).

COLOPHON

This thesis was typeset using LaTeX which was originally developed by Leslie Lamport and based on Donald Knuth's pioneering TeX. The body text is Minion Pro, set 11/17 pt on an almost 25 pc measure and the caption font is Myriad Pro. My favourite monospace font, Monaco, was used in the few places necessary. XeLaTeX was used as the compiler to enable easy access to these fonts with the `mathspec` package.

I tried to base as many typographical decisions on recommendations found in Bringhurst's classic text *The Elements of Typographic Style v3.2*, 2004. My use of margin figures was influenced by the books of Tufte (*The Visual Display of Quantitative Information*, 2001) and MacKay (*Information Theory, Inference, and Learning Algorithms*, 2003).

I used Ti*k*Z for diagrams and the `R` package `ggplot2` (created by Hadley Wickham) for plots. Hopefully these figures, to a great extent, follow the best guidelines for clarity of presentation and data-ink ratio. I also tried to avoid the use of true black where I could in the figures.

Finally, as far as possible I chose to use a colour-blind friendly palette for as many figures as I could. This choice was based on my strong belief that scientists can be very poor at presenting images in a clear way, especially to those with colour-blindness. My preferred reading on this matter is "*How to make figures and presentations that are friendly to colorblind people*" by Masataka Okabe and Kei Ito, 2008. Their colour-blind friendly palette can be found at `http://jfly.iam.u-tokyo.ac.jp/color/`.