Analysis of dynamic changes in gene expression under the control of the circadian clock in *Arabidopsis* (moac project 3)

Ulrich Janus
Department of Mathematics, University of Warwick, CV4 7AL, UK

September 16, 2005

Contents

Abstract

Many organisms keep track of passing time by means of a cellular mechanism called circadian clock. This clock supposedly relies on a central pacemaker that generates oscillations via a transcriptional feedback loop. The gene *LONG ELONGATED HYPOCOTYL* (*LHY*) has been identified to be part of this central feedback loop in *Arabidopsis*. Another component of this loop is *TIMING OF CAB EXPRESSION1* (TOC1), which induces an increase in *LHY* transcription. It is yet unclear how TOC1 exerts this effect on the *LHY* promoter, but it is believed that other factors like PHYTOCHROME INTERACTING FACTOR-3 (PIF3) or EARLY FLOWERING-3 (ELF3) are involved. Here the effect of overexpressing *ELF3* on the expression of *LHY* was studied by comparing *LHY* expression levels of a transgenic *elf3ox* plant line carrying a constitutively overexpressed copy of the *ELF3* gene to normal *LHY* expression of wild type plants. *elf3ox* plants were found to show a decreased response to light signals and an overall lower expression of the *LHY* gene.

The experimental time series were transformed by removing trends and normalising amplitude heights and then analysed by the fitting of two different models. An approximation model, consisting of a modified Fourier sum, was used to characterise the expression patterns. Secondly, using the Metropolis-Hastings algorithm a model based on a stochastic differential equation was fitted to the data in order to uncover the corresponding transcription rates. To overcome the problem of the low time resolution of the data (2h) an interpolation procedure based on modified Brownian bridges was implemented.

1 Introduction

The Circadian clock is a cellular mechanism by which an organism can keep track of passing time so that it is able to anticipate changes in the environment like light and temperature and to respond accordingly.

Circadian clocks have been found in organisms ranging from bacteria and fungi to animals and plants [Young 2001]. Though these clocks show a certain degree of homogeneity in the design, they seem to based on unrelated genes, which suggests that they may have involved independently several times in nature.

For the plant world the weed *Arabidopsis thaliana* has recently been established as a model organism for circadian clocks. Several studies have uncovered important clock genes [Salomé 2004].

It is believed that at the centre of any circadian clock there exists a mechanism generating stable oscillations. This *central oscillator* would then be set by environmental signals (*input pathways*) like light to local time, a process referred to as *entrainment*. The central oscillator would also control the expression of other genes via certain regulatory pathways, called *output pathways*. This three component structure consisting of the central oscillator, input and output pathways seems to be applicable to all circadian clocks studies so far.

In *Arabidopsis* a first model of the central oscillator was based on a transcriptional feedback loop between the genes *LHY*, *CCA1* and *TOC1* [Alabadi 2001]. The gene products LHY and CCA1 are transcription factors which repress the expression of *TOC1*. The protein TOC1 in turn was shown to induce the transcription of *LHY* and *CCA1*, though the mechanisms of the regulation is yet unclear. Transcription of *LHY* and *CCA1* peaks in the morning, while that of *TOC1* peaks in the evening.

Though this is a convenient working model, the picture is not that simple. For example losing the function of TOC1 does not disrupt circadian rhythms [Strayer 2000], though this would be expected in the light of the proposed model. This observation suggests that there are different levels of regulation and redundancy in the clock so that the loss of individual genes can be compensated.

A combined biology and mathematics project, as a part of the MSc program "Molecular Organisation and Assembly of Cells" (MOAC) at the University of Warwick, addressed the issue of the regulation of the *LHY* gene. Apart from the positive regulation of TOC1, whose details are not yet understood, the *LHY* expression is also subject to light induction. This project focused on the proteins PIF3 and ELF3, which are believed to play a role in the regulation of *LHY* by either mediating the impact of TOC1, being involved in light signalling, or both.

PIF3 belongs to a class of bHLH transcription factors. It is known to interact with TOC1 and PHYTOCHROME-B (PHYB). It also binds with high affinity to the G-BOX, a DNA motif present in the promoter regions of many light regulated genes like *LHY*. This make PIF3 a likely candidate in the mediating the *LHY* activation by TOC1 and the clock entrainment by light [Martinez-Garcia 2000].

Mutations in *ELF3* lead to loss of circadian rhythms in the light and a reduction in *LHY* mRNA levels Also, plants overexpressing ELF3 show a low sensitivity to light signals. [Hicks 2001].

LHY transcription activity in the *Arabidopsis* seedlings was tracked by a luminescence assay, where the promoter of the *LHY* gene was coupled with the gene luciferase (*LUC*), whose gene product emanates light when it reacts with its substrate luciferin.

Plants carrying the loss-of-function mutant *pif3-1* showed little effect on *LHY* expression activity [Janus 2005]. Here we present the results of transgenic *elf3ox* plants carrying a constitutively overexpressed copy of the *ELF3* gene.

When dealing with data in general it is desirable to be able to extract relevant information in an objective and clearly defined manner. In biology this is often complicated due to variations between experiments and also the lack of appropriate mathematical tools.

Here we applied a number of mathematical and statistical methods originally implemented by A. Morton [Morton 2004] to identify and characterise differences in the patterns of *LHY* expression, extract information about period and transcription dynamics. The analysis consisted of the following steps.

Firstly, the overall trends and the trend of the amplitudes were removed by the application of kernel smoothing. This procedure removed the data fluctuations due to degradation of luciferin over time.

Secondly, to extract information about period length and to characterise the wave form of *LHY* expression, the time series were fitted to a fourier series by ordinary least squares (OLS) regression. This step is referred to as the approximation model.

Thirdly, to simulate the dynamics of a transcription factor, the time series were fitted to a simple dynamical model defined by a stochastic differential equation (SDE). For the fitting Monte Carlo Markov Chain (MCMC) simulations using a Metropolis-Hastings algorithm were carried out.

In the MCMC simulation the sparsity of the sampled data posed a problem, as they made the estimates of the transition probabilities very inaccurate. This problem was addressed by interpolating the data by modified Brownian bridges [Durham 2002], replacing the previous simpler interpolation method used by Morton.

2 Data Analysis

2.1 Postprocessing by kernel regression

Time series data from luminescence assays typically show experimental artifacts like upwards or downward trends of the overall curve or the amplitude heights. One would like to have a means to extract these trends from the data so that the data can be adjusted accordingly.

A way to extract these trends is the application of kernel regression. Kernel regression is a way to fit a trend to a time series or two-variate data set. In contrast to parametric regression like linear regression, which fits a given function characterised by a certain number of parameters to the data, kernel regression makes no prior assumptions on the trend. Rather the trend is generated locally by making a linear or higher order parametric regression around each time point taking neighbouring data points into account according to weights distributed like the Gaussian bell curve. If the with of the Gaussian, referred to as the *bandwidth* of the kernel regression, is chosen well, then the resulting regression line will capture the overall trend but not follow the individual oscillations.

More precisely, consider a number of observations $(t_i, x_i), i = 1, \ldots, n$, which are assumed to be generated by

$$x_i = m(t_i) + \sigma\epsilon_i,$$

where ϵ_i are normally distributed independent random variables, σ^2 the variance. m is called the regression function, which we want to estimate.

This is done by performing a weighted least squares regression for each observation (t_i, x_i), where a m is taken to be linear and the weighs are normally distributed around the respective observation.

More precisely for each t the estimator

$$\hat{m}(t, h) = \hat{\beta}_0 \tag{1}$$

of m is obtained by fitting the line

$$\beta_0 + \beta_1(t_i - t)$$

to the observations by weighted least squares regression. The weights are given by the kernel function $K_h(t_i - t)$.

$$K_h(u) = \frac{1}{h}\phi\left(\frac{u}{h}\right),$$

where $\phi(x)$ is the density of the normal distribution $N(0, 1)$ and h is called the bandwidth. This means that values close to t have higher weights than those which are further away.

The estimators $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)'$ are the minimisers of

$$\sum_{i=1}^{n} (x_i - \beta_0 - \beta_1(t_i - t))^2 K_h(t_i - t).$$

If the respective matrix is invertible, the minimiser β' is given in closed form by

$$\beta' = (T_t' W_t T_t)^{-1} T_t' W_t X,$$

Here $X = (x_1, \ldots, x_n)'$ is the vector of observations,

$$T_t = \begin{pmatrix} 1 & x_1 - x \\ \vdots & \vdots \\ 1 & x_n - x \end{pmatrix}$$

is an $n \times 2$ matrix and

$$W_t = diag\{K_h(t_1 - t), \ldots, K_h(t_n - t)\}$$

is an $n \times n$ diagonal matrix of the weights. This method is usually presented in a more general form, where m is taken as a polynomial of degree $p \geq 1$. For a more detailed description of kernel smoothing see [Wand].

Application to the collected time series data. Let

$$L(t) := \hat{m}(t, h) \tag{2}$$

denote the trend line resulting from applying kernel regression as described above (comp. (1)). The bandwidth was chosen manually to be $h = 10$. This seems to be a good value for capturing the trend while preserving the overall waveform. It would be preferable to have a means to automatically choose h, but as typical circadian clock related time series can be expected to have a similar period one should do fine with fixed bandwidth. The experimental data typically provided about 50 time series $x_i(t)$ from seedlings of the same genotype. The trend lines were estimated individually and then averaged point-wise. This is illustrated in figure 1 a), where the (point-wise) mean of the individual time series $x_i(t)$ is displayed together with the (point-wise) mean of the individual trend lines $L_i(t)$. These trend lines are then simply subtracted from the corresponding time series to yield the detrended time series denoted by

$\tilde{x}_i(t) = x_i(t) - L_i(t)$. Figure 1 b) is based on the same data as 1 a) and shows the mean of the detrened time series $\tilde{x}_i$.

Looking at 1 b) we see that though the overall trend is taken out there are still differences in amplitude. Given the experimental setup it can be assumed that most of these differences are experimental artifacts (degradation of luciferin) and carry no relevant biological information. The variations of the amplitude was eliminated as follows. Kernel smoothing was applied to the absolute values of each detrended time series $\tilde{(x)}_i(t)$, which yielded a trend for the amplitude $A_i(t)$. This is illustrated in figure 1 c) for the example data set. The figure shows that mean of the absolute values of the detrended time series together with the mean of the individual amplitude trends. Notice that the mean was taken over the absolute values of the individual time series as opposed of the absolute values of the mean, which is why the curve never actually touches the x-axis. To remove the amplitude trend $A_i(t)$, the detrended times series is divided by it to yield the transformed time series

$$\tilde{\tilde{x}}_i(t) = \frac{x_i(t) - L_i(t)}{A_i(t)}. \tag{3}$$

The mean of all the $\tilde{\tilde{x}}_i(t)$ time series of the example data set is displayed in figure (1) d). The curve in this figure was also rescaled by first multiplying it by the average over all time points of all amplitude trends $A_i(t)$ and then adding the average over all time points of all trend lines $L_i(t)$.

All following data analysis is based on the completely detrended data series $\tilde{\tilde{x}}$, which for the ease of notation will still be denoted by x.

2.2 An approximation model

This section formulates a simple approximation model. The corresponding parameters of the fitted model portray information about certain characteristics of the time series like period and the wave pattern.

After detrending the time series we expect it to be ,,close" in some sense to a periodic function with a period of $24h$ in 12h-12h light-dark conditions. This assumption motivates modelling the time series $x(t)$ as being generated by a Fourier sum $\hat{x}(t)$ while subject to a certain noise level. More precisely, the model is of the form

$$x(t) \approx \hat{x}(t) = \alpha_0 + \sum_{k=1}^{p} \left[\alpha_k \cos(\omega k t) + \beta_k \sin(\omega k t)\right], \tag{4}$$

where $\alpha_0, \alpha_k, \beta_k, \quad k = 1, \ldots, p$ are the Fourier coefficients and $\omega = \frac{2\pi}{T}$, where T is period of the periodic function $x(t)$ and $K = 2p + 1$ is the number of Fourier parameters.

As each experiment yielded a number m (≈ 50) time series with n data points, the optimal parameters were fitted by taking observation of all time series $(t_i^j, x_i^j), i = 1, \ldots, n \quad j = 1, \ldots, m$ into account. .

For fixed p, the parameter values $\alpha_k, \beta_k, \omega$ were estimated by OLS, i.e. by minimising the sum of squares,

$$SS(\theta, \omega) = \sum_{i=1}^{n} \sum_{j=1}^{m} \left[x_i^j - \left(\alpha_0 \sum_{k=1}^{p} (\alpha_k \cos(\omega k t_i) + \beta_k \sin(\omega k t_i)) \right) \right]^2,$$

with respect to $\theta = (\alpha\beta), \alpha = (\alpha_1, \ldots, \alpha_p), \beta = (\beta_1, \ldots, \beta_p)$. This was done subsequently over a certain range of ω to determine the estimate for the period. The feasibility of different numbers of Fourier parameters was estimated based on the Akaike Information Criterion (AIC), which adjusts the Log-likelihood function for a penalty proportional to the number of paramters:

$$AIC = -2l(\theta, \omega) + 2K,$$

where $l(\theta, \omega)$ is the Log-likelihood function

$$l(\theta, \omega) = -0.5nm \log \frac{SS(\theta, \omega)}{nm}.$$

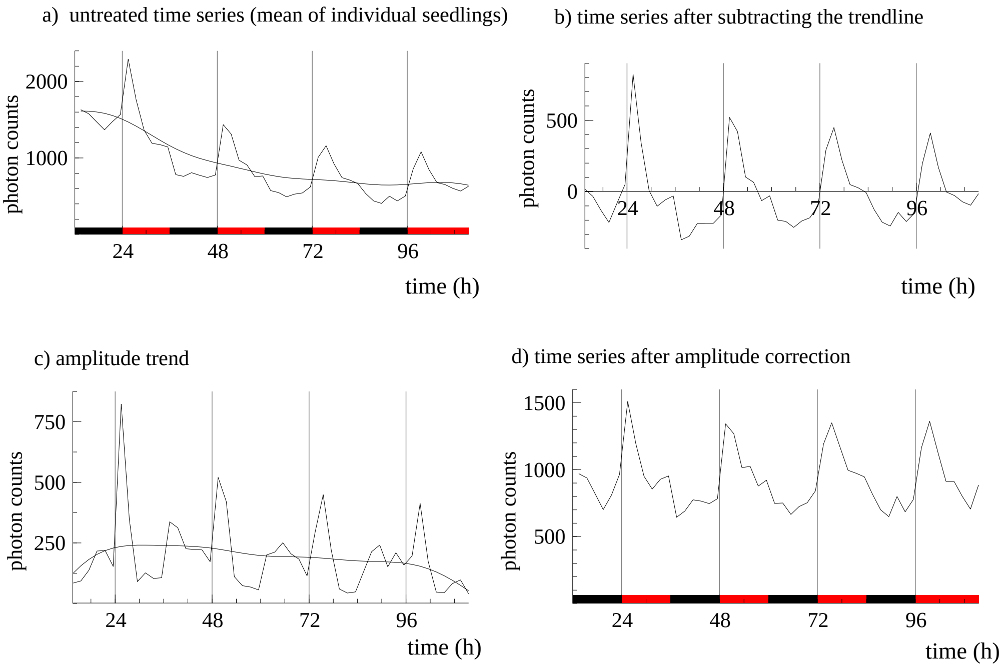


Figure 1: Example for the detrending algorithm described in section 2.1 applied to wild type data of exp. B consisting of 44 seedlings in LD (12h light, 12h darkness) conditions (comp. figure 4b, 6a). Before imaging the seedlings were entrained for 3-5 days. LD conditions were maintained for 96h. The luminescence levels were measured at a time resolution of $\delta t = 2h$. **a)** shows the curve resulting from taking the mean of the 44 individual time series at each time point. The plotted trend line is the result of calculating the trend line $L_i(t)$ (defined in (1), (2)) with a bandwidth of $h = 10$ for each individual time series $x_i(t)$ and taking the average at each time point. **b)** shows the result of subtracting each trend line $L_i(t)$ from its corresponding time series $x_i(t)$ and taking the average over all time series at each time point. **c)** shows the curve by taking the average at each time point of the absolute values of the detrended time series $|\tilde{x}_i| = |x_i(t) - L_i(t)|, i = 1, \ldots, 44$ together with the amplitude trendline $A(t)$. $A(t)$ is the time point-wise average of the individual amplitude trends $A_i(t)$ attained by applying the kernel smoothing to the $|\tilde{x}_i|$ $(h = 10)$. **d)** shows the curve resulting from taking the average at each time point of the detrended time series after amplitude correction $\tilde{\tilde{x}}_i(t)$ as defined in (3), i.e. after dividing the detrended series $\tilde{x}_i$ by the corresponding amplitude trendline $A_i(t)$. The resulting curve was further rescaled by first multiplying it with the average over all time points of all amplitude trends $A_i(t)$ and than adding the average over all time points of all trend lines $L_i(t)$.

For this project $K = 11$ was used as it had proved to be sufficient for reasonably good approximations. If the fitting algorithm yielded parameters of very low absolute values, these were set to zero.

2.3 A dynamical model

2.3.1 Synopsis

This section describes a modelling approach based based on the assumption that the observed time series is the solution of a differential equation subject to some noise. When we assume that the measured luminescence levels are proportional to the levels of LHY protein in the plant we may propose a very simple model of how they change over time: The increase of LHY levels is described by a function $\beta(t)$ representing the lumped effect of transcription of the gene and translation into the protein. The decrease of the LHY level in turn is described by degradation term $\delta x(t)$ proportional to the protein level at the given time t. The only assumption about the expression rate $\beta(t)$ is that it is a periodic function and is therefore taken to be a Fourier sum.

Unfortunately the matter of fitting the model to the data and generating estimates for the parameters is not as straight forward as for the approximation model. What we want is to find parameters which maximise the likelihood that the proposed model generated the observed data. This process is also referred to as Bayesian inference. One method to find these parameter estimates and which is applicable to our situation is the Monte Carlo Markov Chain (MCMC) algorithm of Metropolis and Hastings.

Here the idea is that initial guesses of the parameters are subsequently perturbed. Depending on how the perturbation improves the likelihood of the observed data given the model parameters the new parameter values are either accepted or rejected. This updating procedure is done many times until a distribution of the parameter values emerges. At best a given parameter will approach and oscillate around a certain value. It may also be that the parameter will traverse and a large range of values with no easily distinguishable best fit. This phenomenon could simply mean that several parameter values may equally likely have generated the data. It may also be due to noisy data or an unappropriate model.

2.3.2 The model

Going back to the model we assume that the observed time series $x(t)$ corresponds to the level of LHY and that is can be described by a (stochastic) differential equation. Changes is $x(t)$ over time are assumed to be due to an periodic expression function $\beta(t)$, representing transcription and translation, and a linear degradation rate $\delta x(t)$. More precisely we assume that the time-series is generated by the differential equation

$$\frac{dx}{dt} = \beta(t) - \delta x(t), \tag{5}$$

subject to noise. The rate of expression $\beta(t)$ was taken to be a Fourier sum of fixed order p, i.e.

$$\beta(t) = a_0 + \sum_{k=1}^{p} \left(a_k \cos \omega k t + b_k \sin \omega k t\right), \tag{6}$$

where again $\omega = \frac{2\pi}{T}$ and T is the period of $\beta(t)$. So the parameters that need to be estimated are the Fourier coefficients $a_0, a_k, b_k; k = 1, \ldots, p$, the period ω and the degradation rate δ. The order of the Fourier sum was kept fixed, though the program might be adapted so that p would also be estimated.

Data from plants under 12h in light - 12h in darkness (LD) conditions typically shows a strong increase at the dark-light transition which is difficult to capture with a $\beta(t)$ of the form (6). In order to allow a discontinuous increase in the transcription rate an addition parameter was introduced an $\beta(t)$ replaced by

$$\tilde{\beta}(t) = \beta(t) + c\chi_L, \tag{7}$$

where c is a positive parameter and χ_L is an indicator function which is 1 in the intervals corresponding to light periods and 0 during the time intervals corresponding to darkness.

2.3.3 Bayesian Inference

The process of estimating the model parameters is not as straightforward as for the approximating model. The more complex setup requires the more involved machinery of Markov Chain Monte Carlo (MCMC) methods. Therefor we quickly recall the principles of Bayesian inference and the implemented MCMC algorithm established by Metropolis and Hastings.

The Bayesian approach considers the probability distribution $p(\theta)$ of the parameters. This distribution might be uniform on the parameter space or may include some prior knowledge and is called the *prior distribution.* The model then yields a *likelihood function*

$$l(\theta) = f(x|\theta),$$

which is the probability that the observed time series x was created by the model equations with parameters θ. Bayes theorem then provides a mean to incorporate the information of the observation x into the distribution of model parameters θ:

$$\pi(\theta) := p(\theta|x) = \frac{f(x|\theta)p(\theta)}{f(x)}, \tag{8}$$

where

$$f(x) = \int f(x|\theta)p(\theta)d\theta.$$

$\pi(\theta)$ is called the *posterior distribution.* When considering only one observation, we can treat the integral $f(x)$ as a constant. So we have the proportionality

$$\pi(\theta) \propto l(\theta)p(\theta). \tag{9}$$

Assuming that the posterior distribution π is well behaved (i.e. normal-like), we can simply take the mean of π as our parameter estimate. Unfortunately we cannot sample from π directly, but need to deduce information about π by constructing a Markov chain, which has π as its limiting distribution.

2.3.4 The Metropolis-Hastings algorithm

A random process is called Markov if, given the present state, past and future states are independent. That means that any state depends on the past only through its direct predecessor. The Markov chains used in MCMC algorithms are homogeneous, i.e. it can be defined by transition probabilities $P(\theta, \phi)$ from θ to ϕ. For continuous state spaces the chain can be defined by transition densities $p(\theta, \phi)$.

Under suitable conditions (see below), the sequence of the distributions $(\pi^n)_{n\in\mathbb{N}}$ of each step of the Markov chain converges to a limiting distribution $\pi(\phi) = \lim_{n\to\infty} P^n(\theta, \phi)$.

MCMC now exploits the fact that if a Markov chain is reversible, i.e. if for a distribution π it holds that

$$\pi(\theta)p(\theta, \phi) = \pi(\phi)p(\phi, \theta) \tag{10}$$

then π is the limiting distribution of that Markov chain.

The Metropolis-Hastings algorithm describes how to choose the transition probabilities, so that they define a Markov chain, whose limiting distribution is equal to the posterior distribution, which we need to sample from. These transition probabilities are of the form

$$p(\theta, \phi) = q(\theta, \phi)\alpha(\theta, \phi), \quad \theta \neq \phi,$$

where q is an arbitrary transition kernel and α a probability. The introduction of the acceptance probability α means that the chain has a positive probability to remain in the same state, i.e.

$$p(\theta, \theta) = 1 - \int q(\theta, \phi)\alpha(\theta, \phi)d\phi.$$

The acceptance probabilities are defined to be

$$\alpha(\theta, \phi) = \min\left\{1, \frac{\pi(\phi)q(\phi, \theta)}{\pi(\theta)q(\theta, \phi)}\right\} =^{(9)} \min\left\{1, \frac{l(\phi)q(\phi, \theta)}{l(\theta)q(\theta, \phi)}\right\} \tag{11}$$

For these transition probabilities (10) holds, which means that the chain has π as its limiting distribution.

For the implementation the algorithm can be sketched as follows:

(i) Initialize counter $j = 1$ and set initial value $\theta^{(0)}$.

(ii) Sample a proposal value ϕ from the density $q(\theta^{(j-1)}, \cdot)$.

(iii) Accept proposal $(\theta^{(j)} = \phi)$ with probability $\alpha(\theta^{(j-1)}, \phi)$. If the move is rejected, set $\theta^{(j)} = \theta^{(j-1)}$

(iv) Increase counter j and return to step 2 until a convergence criterion is met.

This algorithm was implemented using Ox as the programming language. The parameters were updated in turn. After each iteration of parameter updates there followed an update of the interpolated data points (see section 2.3.5 below). Figure 2 shows examples of processes generated by updating the model parameters.

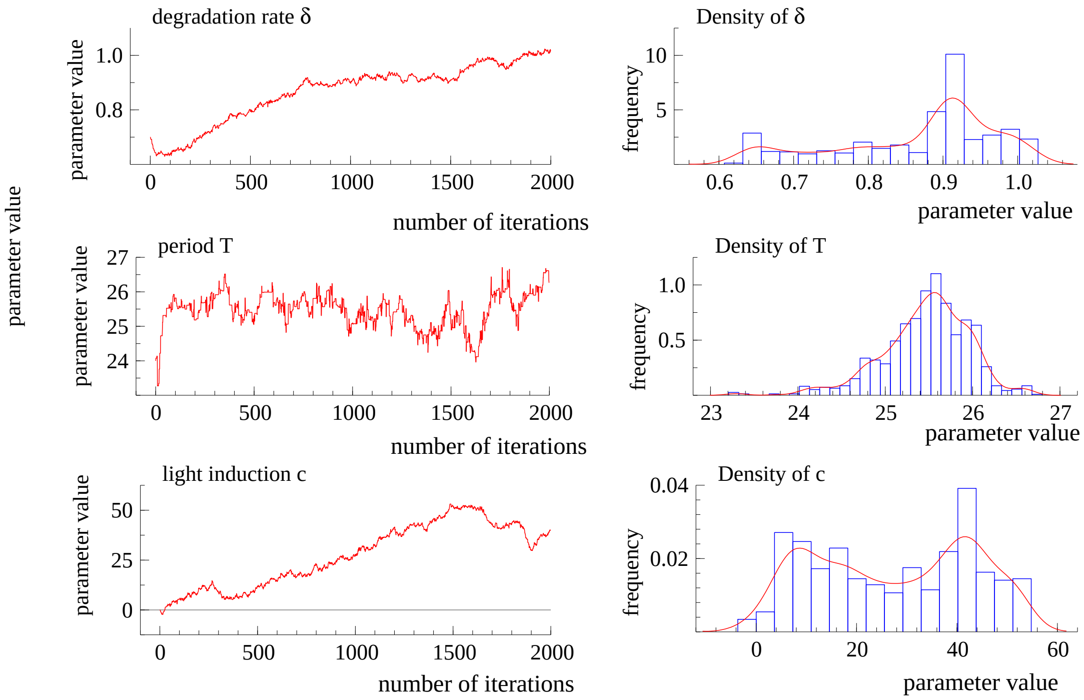


Figure 2: Example on the processes generated by updating the model parameters (see equations (5), (6) and (7)). The respective values for the parameters δ (degradation rate), T (period) and c (light induction) are plotted versus the number of iterations. These processes were the generated while fitting the dynamical model to the LD wild type data from exp B depicted in figures 1, 4b and 6a. The figure shows how the respective parameters developed over the performed iterations from their starting values (left). The densities illustrate where each parameter „spent most of its time"(right). Compare figure 6 for the model results.

2.3.5 Interpolating sparse data with Brownian bridges

When fitting the SDE model to the measured time series the problem arises, that the intervals between data points are too large to allow reliable estimates of the MCMC transition probabilities. A way to get around this problem is by introducing a number of intermediate data points. One way to do this, which tries to reflect the noise element of the data, is the Brownian bridge. The bridge between to neighbouring data points is constructed by starting a Brownian motion at the first data point which is biased towards and conditioned to terminate at the second data point. The algorithm is laid out in more detail below.

Given two data points x_s and x_t at times $s < t$. Let $s = \tau_0 < \cdots < \tau_M = t$ be a partition of the interval $[s, t]$. We then want to sample the intermediate points $(u_0, \tau_0), \ldots, (u_M, \tau_M)$, where $u_0 = x_s$ and $u_M = x_t$ are fixed.

The following approach, the modified Brownian bridge, is described in more detail in [Durham 2002]. The idea is to start a Brownian motion at u_0 conditioned to end at u_M. At each step, the next data point u_{m+1} is sampled from a Gaussian density with mean and variance depending on the position of u_m in relation to the endpoint u_M. More precisely we sample u_{m+1} from the Gaussian density

$$\phi(u_{m+1}; u_m + \tilde{\mu}_m \delta, \tilde{\sigma}_m^2 \delta),$$

where $\delta = \frac{t-s}{M}$ and

$$\tilde{\mu}_m = \frac{u_M - u_m}{t - \tau_m}, \qquad \tilde{\sigma}_m^2 = \left(\frac{M - m - 1}{M - m}\right) \bar{\sigma}^2$$

and $\bar{\sigma} = \sigma(u_m)$ is the standard deviation of noise as given by the model parameters at each step of the Markov chain. This means that at each step of the Brownian bridge is sample from a normal distribution centered around the *drift* $\tilde{\mu}_m$, which is biased towards the endpoint u_M. Further as the bridge approaches its endpoint u_M the variance decreases to minimise the probability of large jumps close to the end.

Figure 3 show an example of data interpolated with Brownian bridges.

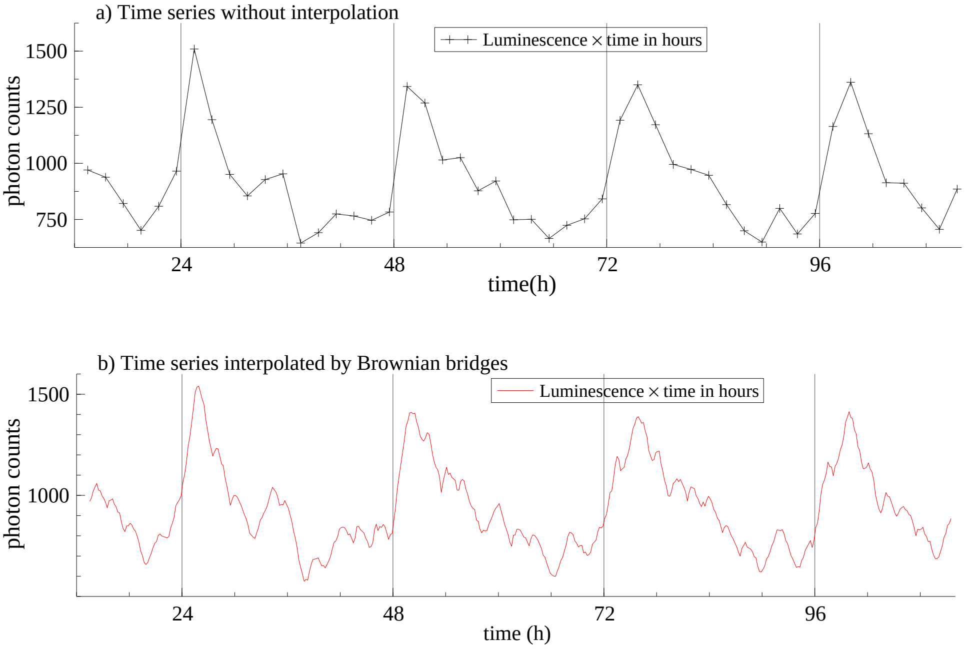


Figure 3: Example for interpolation with Brownian bridges. This interpolation was the generated while fitting the dynamical model to the detrended wild type data of exp. B in LD conditions (see figure 1). a) shows the detrended data set with the individual data points marked by ‚+'; b) shows the the data after performing 2000 iterations of the MCMC algorithm. The result of the fitting is displayed in figures 6h,j.

2.4 Outline of the work

The described algorithms except for the interpolation with Brownian bridges were implemented in Ox by Alex Morton. See [Morton 2004] for the original code and its description.

This code was changed to improve flexibility in the treatment of different data sets as well readability. Several small functions were added to automatically preprocess experimental data coming from the imaging software, for some additional analytical tools and the generation of graphical output.

The previous algorithm used for interpolation between data points based on randomly perturbing points along the straight line between neighbouring data points was replaced by the modified Brownian bridge interpolation described in section 2.3.5.

The code was applied to the data from several experiments with transgenic *Arabidopsis* seedlings. Selected results are presented below.

3 Results

3.1 Consistency of wild type data between experiments A and B

Analysing and comparing wild type data from two separate experiments yielded mixed results (figure 4). Plants in both experiments were entrained in 12h-12h light-dark (LD) conditions, which were maintained during the first 72h (exp A) respective 96h (exp B) before plants were released into constant light (LL) conditions. Panels 4 a) (exp. A) and b) (exp. B) show the detrended time series corresponding to the respective LD time intervals. The approximation (panels 4 c), d)) and the dynamical (panels 4 e), f)) were fitted to the detrended LD data. Though the fitted transcription rates $\beta(t)$ show similar behavior in the light, they diverge in the dark interval. Also the light induction parameters c were estimated differently (comp. panels 5h) for exp. A and 6j) for exp. B).

3.2 Comparison between LD and LL regimes for wild type data (exp A)

The differences of LD and LL wild type data (fig. 5 a)) of exp. A were investigated by fitting the approximation and dynamical models to each of the respective time intervals. Figure 5 b) and c) shows the wild type series after removing trends (see section 2.1). Fourier sums (see section 2.2) were fitted to the two time intervals (fig. 5 d) and e)). The results for the fitting of the dynamical model (section 2.3) are shown in figure 5 f) - i). The model fittings reflected the missing of the light induction by leading to an overall flatter curve for the LL period than for LD. Also the estimate for the period T was longer for LD than for LL (panels 5h,i). But the period estimates overall are somewhat confusing as they are distinctively larger than 24h, which did not fit with the actual for of the transcription rate $\beta(t)$.

3.3 The *elf3ox* transgenic line (exp B)

Comparing the *LHY* expression between the *elf3ox* transgenic line and the wild type in experiment B showed a number of differences (figure 6a. The light induced peaks at dawn were lower and wider for the *elf3ox* line than for wild type. This supports the notion that ELF3 is a negative regulator of the light input into the clock. Luminescence levels of the *elf3ox* were significantly lower than wild type (panels 6b and c). These observation were supported by a repeat experiment (results not shown).

The described models were fitted to the LD time intervals of wild type and the *elf3ox* line. Panels 6d and e show the detrended LD data. The corresponding fits of the approximation model (panels 6f and e) expose the „high shoulder" in *elf3ox* waveform in contrast to the lower one visible in the wild type model fit.

The fits of the transcription rates $\beta(t)$ of the dynamical model (panels 6h and i) stressed the higher light response of the wild type compared to the *elf3ox* line. This feature was also reflected in the different estimates of the light induction parameter c (panels 6j and k). The estimates for the period T seemed again to be a bit errated, though the wild type period was actually estimated to be close to 24h. Reassuringly similar estimates for the degradation rate δ were found (again panels 6j and k).

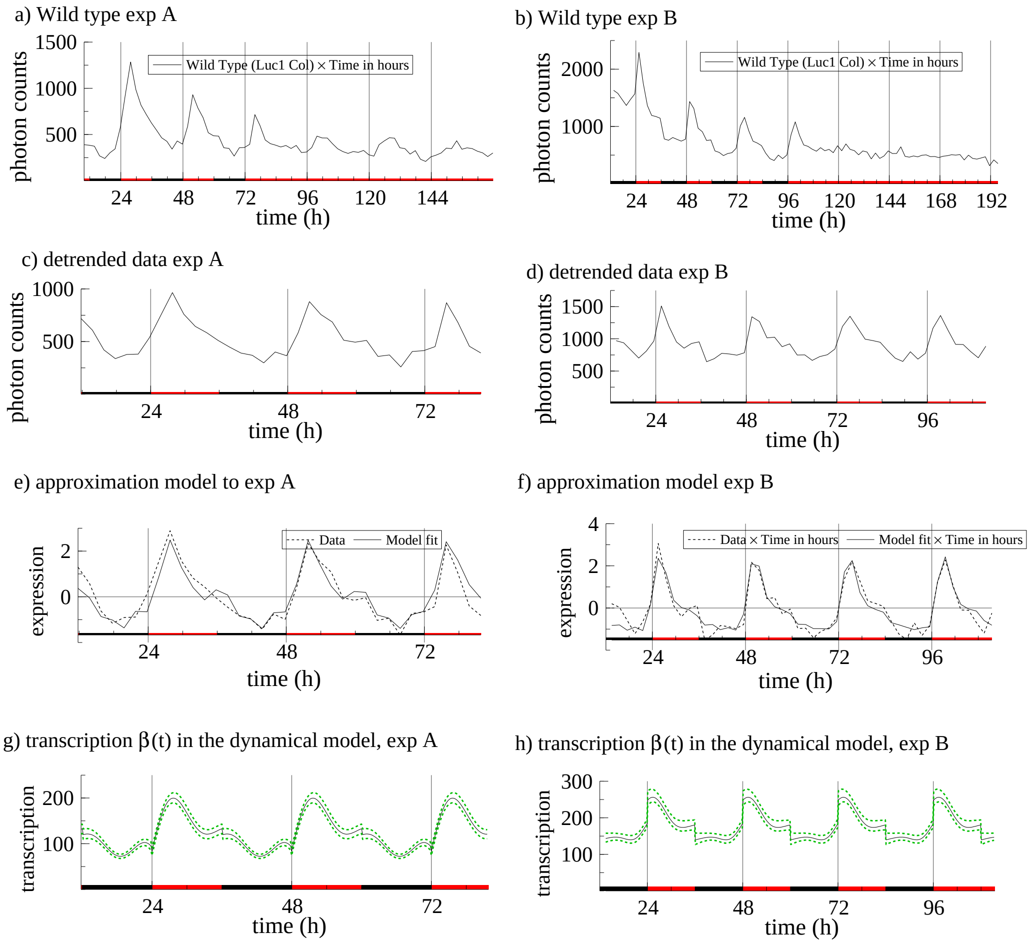


Figure 4: **Comparison of wild type data from two experiments.** For details of these experiments see figure 5 (exp. A) and figure 6 (exp. B). (a),(b) Average luminescence levels of the seedlings, (c),(d) Data of the 12h-12h light-dark (LD) interval after detrending, (e), (f) Fit of the approximation model to the detrended data, (g), (h) Fit of transcription function $\beta(t)$ to the detrended data.

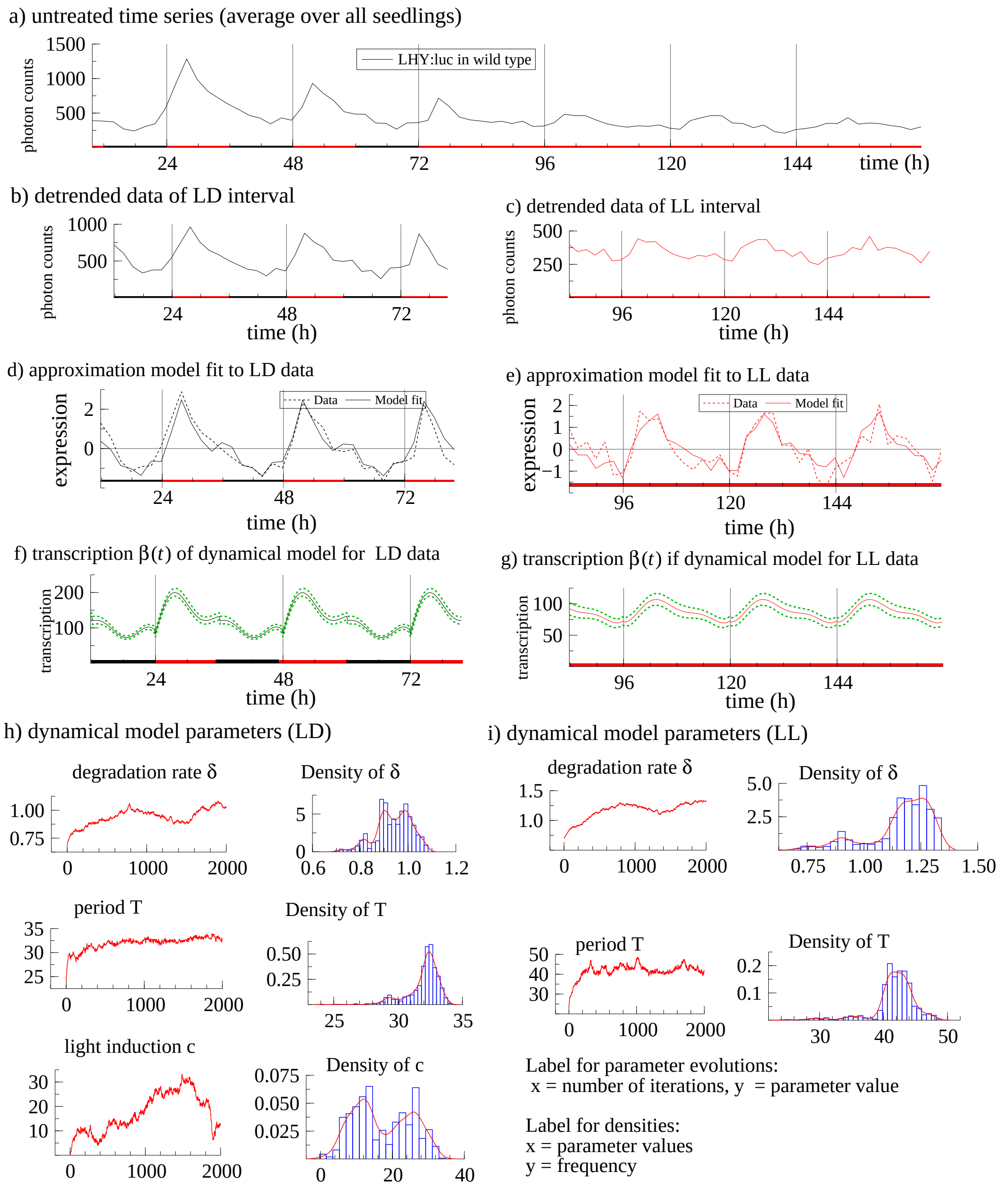
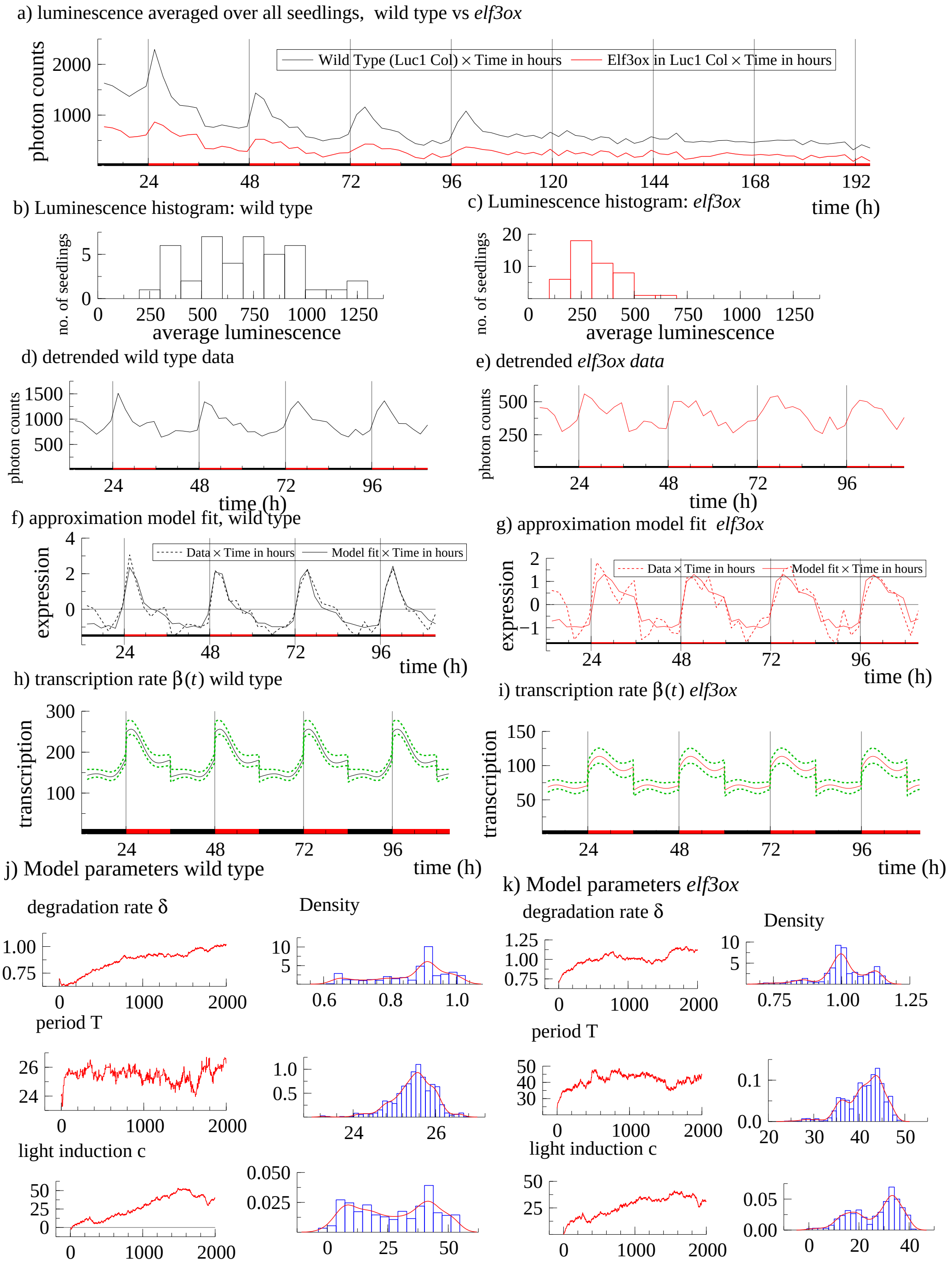


Figure 5: **Comparison of wild type LD and LL data (exp. A)**. Time series of luminescence of a set 109 wild type seedlings. All seedlings carried the -929 LHY :LUC promoter con- struct. Red and black bars at the bottom indicate the periods of red light and darkness respectively. Images were taken with a time resolution of 2h. Plants were entrained in white light for 7 days and then transferred to red light for imaging. Normal LD conditions were maintained for the first 72 hours. After that plants were kept in constant light. The signal strength for the two different sets of seedlings were averaged and the respective background signal subtracted (a). Detrended data for LD and LL are compared in panel (b) and (c) respectively. Fit of the approximation model are shown in panel (d) for LD data and (e) for LL data. Transcription rates $\beta(t)$ of the fit of the dynamical model are shown in panels (f) for LD data and (g) for LL data. The corresponding developments of the model parameters δ (degradation rate), T (period) and c (light induction) are shown in panels (h) for LD and (i) for LL.


a) luminescence averaged over all seedlings, wild type vs *elf3ox*

b) Luminescence histogram: wild type

c) Luminescence histogram: *elf3ox*

d) detrended wild type data

e) detrended *elf3ox data*

f) approximation model fit, wild type

g) approximation model fit *elf3ox*

h) transcription rate β(*t*) wild type

i) transcription rate β(*t*) *elf3ox*

j) Model parameters wild type

k) Model parameters *elf3ox*

Figure 6: **The *elf3ox* transgenic line vs. wild type (exp. B**. 45 seedling of the *elf3ox* transgenic line and 44 wild type seedlings were imaged after entrainment in LD conditions. The LD conditions were maintained for the first 96 hours of imaging. Then plants were released in constant light (LL). The red and black bars indicate the light conditions. (a) Averaged luminescence levels of the transgenic and wild type seedlings after noise adjustment. (b),(c) Average luminescence over the whole imaging time (0-194h) were calculated for each seedling. The frequencies of different average levels are displayed in histograms for wild type and *elf3ox* line respectively. (d),(e) LD data with trends removed. (f),(g) Fitting of the approximation model. (h),(i) Transcription rate $\beta(t)$ from fitting of the dynamical model. (j),(k) Parameter evolution from starting values generated while fitting the dynamical model. (Axis label are the same as figure 5)

4 Discussion and Conclusion

4.1 Discussion

The results presented in this paper are two-fold in nature. On the one hand side there is the experimental result of the effect of overexpressing *ELF3* on the expression of the *LHY* gene, on the other hand we have the application of mathematically based analytical methods, whose performance was tested on the experimental data.

The application of the methods on different data sets originating from wild type plants revealed that despite the recovering of common features the models results also showed some significant variations. Reasons for this may lie in the experimental data as well as in the analysis. For one the data were difficult to treat because of the low time resolution, a relatively high noise level and variability between experiments. One the other hand problems with the model implementation and the fine tuning of the model fit may also play a role. So what needs to be performed on in the future is a extended and careful analysis of a number of reliable wild type assays to build a sound basis which further studies can then be set on.

The detrending and the approximation model algorithms worked reasonably well. The model estimated the period T close to $24h$ as expected. The algorithm for the dynamical model was mathematically and computationally more involved and also produced some difficult to interpret results. As seen in the comparative studies of the two wild type experiments, the estimate of the light induction showed significant variations. In terms of the data the reason for this might lie with problems of the plant entrainment or the noise in the data collection. It may also be that the model needs to be improved. Instead of the Fourier series in combination with a step function one might also just model the LD and LL time intervals independently by polynomials, or implement some biological knowledge about the effect of light induction. Other issues of the dynamical model were that many parameter densities had medium to high variances, and that the period T was often estimated far too large. This behaviour needs to be analysed to decide whether the problem lies with the data, the model or the implementation of the algorithm. The problem of the sparsity of the data was successfully addressed by implementation of the modified Brownian bridge. The acceptance rate was reasonably high (about 20 %).

The data from the assay of the transgenic *elf3ox* line revealed a lower response to light signals and an overall lower degree of *LHY* expression than in wild type plants. These results were confirmed in a repeat experiment (data not shown). This finding supports the idea of ELF3 to act as a repressor of light signalling to the clock. Despite the issues of the dynamical model described above the comparative analysis revealed interesting differences between the expression patterns of *LHY* under LD conditions of the wild type and *elf3ox* line, which would have to be confirmed by repeat experiments. Generally these expression patterns are hard to interpret biologically, as a detailed model of the underlying transcriptional network is still missing. But even though the analytical results should be taken with some care, they demonstrate that the presented modelling approach can be useful in accessing hidden layers of information in the data.

4.2 Conclusion

The *elf3ox* assay was performed as a part of a bigger project focusing on the regulation of the *LHY* gene. Recent analysis of the *LHY* promoter has revealed two likely binding sites for regulatory factors (Spensley 2005, unpublished results). Future work needs to be performed in order to reveal if the induction of transcription of *LHY* by TOC1 is transmitted via any of those binding sites. For one, the effects of mutations of other likely regulatory factors on *LHY* transcription need to be studied. Assays with plants carrying *pif3ox* (overexpression of *PIF3*) or TOC1RNAi (here the *TOC1* gene is ,silenced' by a complementary mRNA strand) were performed, but failed to yield useful results and need to be repeated. The data of these experiments could then be compared to the *LHY* transcription in plants carrying transformed promoter constructs, where one or both of the identified binding sites are missing or non-functional. In this comparative analysis the presented methods may prove very useful.

All in all it is still a long way to a thorough understanding of the regulation of the central oscillator in *Arabidopsis* and that of *LHY* particular. But should it be reached it will be an important step forward in understanding the workings of the *Arabidopsis* clock and probably other plant clocks as well.

5 Experimental Methods

5.1 Plant materials and growth condidtions

The wild type ecotype Columbia (Col) of *Arabidopsis* was obtained from The *Arabidopsis* Stock Centre. The promoter regions of the *LHY* gene were fused to the luciferase (luc) reporter gene and to terminator sequences from the nopaline synthase (nos) gene, referred to as -929 lhy:luc construct. The construct was made by Jac-Yean Kim in the Carré laboratory. The promoter construct was introduced into the Columbia ecotype of *Arabidopsis*. These wild type plants were crossed with plants containing the *elf3ox* mutation. This was done by Mark Spensley in the Carré laboratory.

Plants were grown on a 1:1 mixture of compost (B & Q plc, UK) and vermiculite (Silvaperl, UK). The soil was soaked with water containing 2g/litre of insecticide (Intercept, Scotts,UK) before the seedlings were transferred to soil. Plants were grown in the greenhouse under 16 hour photoperiods.

Seeds were sterilized with 50 % (w/v) bleach (Fisher Scientific) and 0.01% (v/v) Tween 20 (Aldrich Chemical Co., UK) for 10 minutes and then rinsed with sterile distilled water four times. Seeds were sown on MS-agar medium [4.2g/litre of Murashige and Skoog powder (Sigma), 1% (w/v) agar, pH adjusted to 5.3-5.7 with 1M KOH, 3% (w/v) sucrose]. Seeds were stratified at $4°C$ for 4-5 days to synchronize germination.

5.2 Luminescence assay with photon-counting cameras

Two weeks before imaging, seeds were sterilized, sowed on MS medium containing 3% (w/v) sucrose and stratified as described above (see section. Plants were then grown for 7 days under 80 $\mu molm^{-2}s^{-1}$ of white light in temperature-controlled incubators (Sanyo electronic Co., Japan). The photoperiod was 12h of light and 12h of darkness. Plants were kept at a constant temperature of $22°C$. For imaging plants were transferred to red light conditions. During imaging light-dark cycles were maintained for 72h, then plants were kept in constant light. One day before, and then on the day when imaging started plants were pre-sprayed with the luciferase substrate luciferin (BIOSYNTH AG, Switzerland) at a 5mM concentration in 0.01% Triton X-100, to remove luciferase activity accumulated prior to the first luciferin treatment. Luminescence was imaged for 10 or 25 minutes every 2 hours of for 5-6 days using a photon-counting camera (Hamamatsu). The luminescence levels were quantified from the images using Metamorph$^{\mathrm{TM}}$ software (Universal Imaging).

Acknowledgements

I thank my supervisors Mark Spensley, Isabelle Carré (Department of Biology), Eleisa Heron, Barbel Finkenstadt (Department of Statistics) and David Rand (Department of Mathematics) for their support and advice. The work was supported by the Engineering and Physical Sciences Research Council (EPSRC).

References

[Alabadi 2001] Alabadi D., Oyama T., Yanovsky M.J., Harmon F.G., Mas P., Kay S.A. (2001). Reciprocal regulation between *TOC1* and *LHY/CCA1* within the *Arabidopsis* circadian clock. Science 293, 880-883

[Brooks 1998] Brooks, S P, 1998. Markov Chain Monte Carlo Method and its Application. The Statistician. 47. p69-100

[Carré 2002] Carré, I.A. and Kim, J.-Y. (2002). MYB transcription factors in the Arabidopsis circadian clock. J. of Exp. Botany, Vol. 53, No. 374, pp. 1551-1557

[Durham 2002] Durham, Garland B & Gallant, A Ronald, 2002 Numerical Techniques for Maximum Likelihood Estimation of Continuous-Time Diffusion Processes, Journal of Business & Economic Statistics, American Statistical Association, vol. 20(3), pages 297-316

[Elerian 2001] Elerian, O., Chib, S. and Shephard, N, 2001. Likelihood inference for discretely observed nonlinear diffusions, Econometrica 69: 959993

[Fujimori 2004] Fujimori T, Yamashino T, Kato T, Mizuno T. Circadian-controlled basic/helix-loop-helix factor, PIL6, implicated in light-signal transduction in Arabidopsis thaliana. 2004. Plant Cell Physiol. 45(8):1078-86

[Gamerman] Dani Gamerman 1997,Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference , Chapman & Hall . New York , ISBN: 0412818205

[Hicks 2001] Hicks, K. A., Albertson, T.M. & Wagner, D.R. (2001). *Early flowering3* encodes a novel protein that regulates circadian clock function and flowering in *Arabidopsis*. Plant Cell 13, 1281-1292

[Hall 2002] Hall A, Kozma-Bognar L, Bastow RM, Nagy F, Millar AJ. (2002). Distinct regulation of CAB and PHYB gene expression by similar circadian clocks. Plant J. Nov;32(4):529-37

[Janus 2005] Janus, U. 2005. Analysis of dynamic changes in gene expression under the control of the circadian clock in Arabidopsis. Master Thesis.

[Lazebnik 2002] Lazebnik, Y. 2002. Can a Biologist Fix a Radio?. Cancer Cell, 2, 179-182)

[Martinez-Garcia 2000] Martinez-Garcia, J.F., Huq, E. and Quail, P.H. (2000). Direct targeting of light signals to a promoter element-bound transcription factor. Science 288 (5467): 859-63

[Morton 2004] Morton, A. 2004. unpublished notes

[Øksendal] Bernt Øksendal 2003, Stochastic Differential Equations, Springer. Heidelberg, ISBN: 3540637206

[Park 1999] Park DH, Somers DE, Kim YS, Choy YH, Lim HK, Soh MS, Kim HJ, Kay SA, Nam HG. Control of circadian rhythms and photoperiodic flowering by the Arabidopsis GIGANTEA gene. (1999). Science. Sep 3;285(5433):1579-82.

[Park 2004] Park E., Kim J., Lee Y., Shin J., Oh E., Chung W., Liu J.R., G. Choi, (2004). Degradation of Phytochrome Interacting Factor 3 in Phytochrome-Mediated Light Signaling. Plant Cell Physiol. 136(8): 968975

[Salomé 2004] Salomé, P.A. and McClung, C. R. (2004). The Arabidopsis thaliana Clock, J. of Biol. Rhythms, Vol. 19 No. 5, 425-435

[Scully 2000] Scully, A. L., Kay, S. A. (2000). Time flies for Drosophila. Cell 100, 297-300.

[Strayer 2000] Strayer, C. et al. (2000). Cloning of the Arabidopsis clock gene TOC1, an autoregulatory response regulator homolog. Science 289, 768-771

[Toledo-Ortiz 2003] Toledo-Ortiz G, Huq E, Quail PH. (2003). The Arabidopsis basic/helix-loop-helix transcription factor family. Plant Cell. Aug;15(8):1749-70

[Wand] Wand M.P., Jones M.C. 1995, Kernel Smoothing, Chapman & Hall. London, ISBN: 0412552701

[Yamashino 2003] Yamashino T, Matsushika A, Fujimori T, Sato S, Kato T, Tabata S, Mizuno T. 2003. A Link between circadian-controlled bHLH factors and the APRR1/TOC1 quintet in Arabidopsis thaliana. Plant Cell Physiol. 44(6):619-29

[Young 2001] Young, M. W. and Kay, S.A. (2001). Time zones: a comparative genetics of circadian clocks. Nature Reviews Genetics, 2, 702-715