Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Theoretical Biology 234 (2005) 383–393
Journal of Theoretical Biology
www.elsevier.com/locate/yjtbi

Modelling genetic networks with noisy and varied experimental data: the circadian clock in *Arabidopsis thaliana*

J.C.W. Locke[a,b,c], A.J. Millar[b,c], M.S. Turner[a,c,*]

[a]*Department of Physics, University of Warwick, Coventry CV4 7AL, UK*
[b]*Department of Biological Sciences, University of Warwick, Coventry CV4 7AL, UK*
[c]*Interdisciplinary Programme of Cell Regulation, University of Warwick, Coventry CV4 7AL, UK*

Received 8 July 2004; received in revised form 8 October 2004; accepted 23 November 2004
Available online 22 January 2005

Abstract

Circadian clocks in all organisms include feedback loops that generate rhythmic expression of key genes. We model the first such loop proposed for the clock of *Arabidopsis thaliana*, the experimental model species for circadian timing in higher plants. As for many biological systems, there are no experimental values for the parameters in our model, and the data available for parameter fitting is noisy and varied. To tackle this we constructed a cost function, which quantifies the agreement between our model and various key experimental features. We then undertook an efficient global search of parameter space, to test whether the proposed circuit can fit the experimental data. Using this approach we show that circadian clock models can function well with low cooperativity in transcriptional regulation, whereas high cooperativity has been a feature of previous (hand-fitted) clock models in other species. Our optimized solution for the *Arabidopsis* clock model fits several, but not all, of the key experimental features. We test the predicted effects of well-characterized mutations in the clock circuit and show the phases of the circadian cycle where additional components that are yet to be identified experimentally must be present to complete the circadian feedback loop.

Keywords: Circadian; Gene network; Parameter estimation; *Arabidopsis*

1. Introduction

A circadian network or "biological clock" appears to confer a competitive advantage to an organism, probably by enabling it to anticipate daily light/dark (or warm/cold) cycles in the environment. Circadian rhythms with very similar properties are found in almost all eukaryotic organisms and in some prokaryotes, controlling processes from cyanobacterial cell division to human sleep–wake cycles (Dunlap et al., 2003). There is now evidence (Dunlap, 1999) that these rhythms can be generated by a central "loop" or loops of genes (and gene products), that communicate by positive and negative feedback. Input signals from light and/or temperature alter the level of one or more components of the loops, in order to reset the phase of the rhythm. Such loops have been proposed and modelled for a variety of organisms, including the fungus *Neurospora crassa* (Leloup et al., 1999, Ruoff et al., 1999a, b, 2001), the fruit fly *Drosophila melanogaster* (Ueda et al., 2001; Tyson et al., 1999), and the mouse (Leloup and Goldbeter, 2003; Forger and Peskin, 2003). Oligonucleotide array experiments in the model plant *Arabidopsis thaliana* (Harmer et al., 2000) suggest that RNA transcripts from at least 6 percent of the genome, or over a thousand genes, are expressed rhythmically, under the control of output pathways leading from the circadian clock. Although, there is evidence that there are

*Corresponding author. Department of Physics, University of Warwick, Coventry CV4 7AL, UK. Tel.: +44 24 76 522257; fax: +44 24 76 692016.
E-mail addresses: j.c.w.locke@warwick.ac.uk (J.C.W. Locke), andrew.millar@warwick.ac.uk (A.J. Millar), m.s.turner@warwick.ac.uk (M.S. Turner).

0022-5193/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2004.11.038

independent circadian rhythms in plants and other organisms not under the central clock control, such as the light promoted circadian oscillations in nitrate reductase (NR) activity (Ramlaho et al., 1995; Christensen et al., 2004), it appears that the molecular rhythms generated from the central oscillator control a variety of the observed, macroscopic rhythms including leaf movement, flowering time, and photosynthesis.

One of the fundamental problems facing biological scientists in the post-genomic era is how to obtain, and test, models for the genetic networks that represent the regulatory "wiring diagram" of a living cell. These networks can easily involve thousands of genes and gene products (RNAs and proteins). Circadian clocks can be thought of as genetic subnetworks that are responsible for generating the circadian rhythm (Goldbeter, 2002). Their complex regulatory behaviour and relatively small number of components makes them tractable examples for joint theoretical and experimental analysis.

Much molecular data aims to identify components of the network and to define connections between them. Generally there is little or no biochemical data for the numerous parameters, such as the chemical rate constants, that control the circadian network. In all but one recent mathematical model of the circadian clock (Forger and Peskin, 2003), these parameters have been chosen "by hand". Such an approach becomes more and more time consuming, and potentially unreliable, as the number of components in the model increases. This opens up the very real possibility that apparent deficiencies in existing models may not be caused by an incomplete experimental understanding, or model, but rather by non-optimum choices of the parameters.

We present a general method for comparing noisy experimental data with model networks to derive parameter estimates, and apply this method to improve our understanding of the circadian network in *Arabidopsis*. The first multi-gene loop identified in the *Arabidopsis* circadian clock comprises a negative feedback loop, in which two partially redundant genes *LATE ELONGATED HYPOCOTYL* (*LHY*) and *CIRCADIAN CLOCK ASSOCIATED* 1 (*CCA1*) repress the expression of their activator, *TIMING OF CAB EXPRESSION* 1 (*TOC1*) (Alabadi et al., 2001). Several other clock genes have been discovered in *Arabidopsis* (reviewed in Eriksson and Millar, 2003; Hall et al., 2003; McWatters et al., 2000), but these either have accessory functions (Mas et al., 2003) or have not yet been located relative to the one-loop LHY/CCA1–TOC1 model. We therefore initiate modelling of the *Arabidopsis* clock by analysing the one-loop LHY/CCA1–TOC1 model, using a method that will be widely applicable in tackling larger genetic networks. By comparison with experimental data, the model suggests where additional component(s) of the clock network may function.

2. Model description

Even the minimal description of the *Arabidopsis* clock network as outlined in Fig. 1 required seven coupled differential equations to model the central loop, yielding a total of 29 parameters. As in previous clock models (Leloup et al., 1999; Leloup and Goldbeter, 2003; Ueda et al., 2001; Kurosawa and Iwasa, 2002; Kurosawa et al., 2002) Michealis–Menten kinetics were used to describe enzyme-mediated degradation of proteins, and Hill functions were used to describe the transcriptional activation term of the mRNA for *LHY* and *TOC1*. As *LHY* and *CCA1* are indistinguishable for our purposes, we retain only one gene, *LHY*, in our model. We took the following as our mathematical model for the central circadian network: a LHY–TOC1 feedback loop which involves the cellular concentrations $c_i^{(j)}(t)$ of the products of the ith gene ($i = L$ labels LHY, $i = T$ labels TOC1) where $j = m, c, n$ denotes that it is the corresponding mRNA, or protein in the $c$ytoplasm or $n$ucleus, respectively.

$$\frac{\mathrm{d}c_L^{(m)}}{\mathrm{d}t} = L(t) + \frac{n_1 c_T^{(n)a}}{g_1^a + c_T^{(n)a}} - \frac{m_1 c_L^{(m)}}{k_1 + c_L^{(m)}}, \tag{1}$$

$$\frac{\mathrm{d}c_L^{(c)}}{\mathrm{d}t} = p_1 c_L^{(m)} - r_1 c_L^{(c)} + r_2 c_L^{(n)} - \frac{m_2 c_L^{(c)}}{k_2 + c_L^{(c)}}, \tag{2}$$

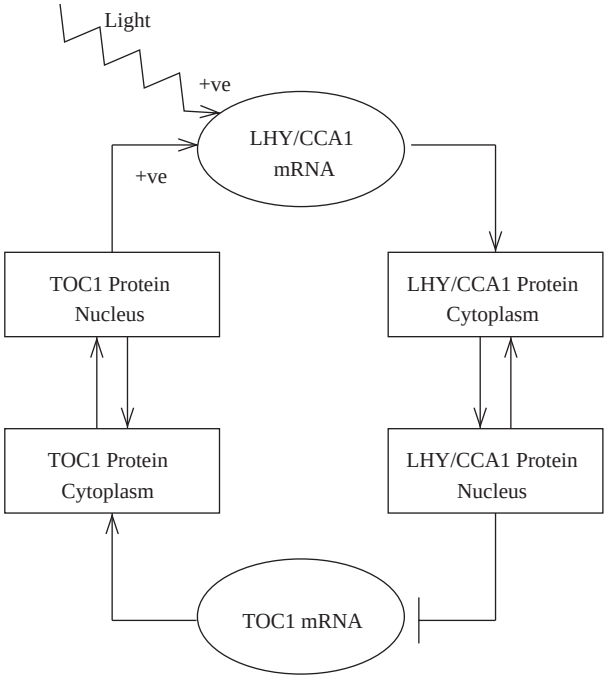

Fig. 1. Model for the central feedback loop in the *Arabidopsis* clock. TOC1 proteins in the nucleus and light, mediated by protein P (not shown) positively activates transcription of *LHY* and *CCA1* mRNA. When LHY and CCA1 proteins reach the nucleus they down regulate *TOC1* mRNA transcription.

$$\frac{dc_L^{(n)}}{dt} = r_1 c_L^{(c)} - r_2 c_L^{(n)} - \frac{m_3 c_L^{(n)}}{k_3 + c_L^{(n)}}, \quad (3)$$

$$\frac{dc_T^{(m)}}{dt} = \frac{n_2 g_2^b}{g_2^b + c_L^{(n)b}} - \frac{m_4 c_T^{(m)}}{k_4 + c_T^{(m)}}, \quad (4)$$

$$\frac{dc_T^{(c)}}{dt} = p_2 c_T^{(m)} - r_3 c_T^{(c)} + r_4 c_T^{(n)} - \frac{m_5 c_T^{(c)}}{k_5 + c_T^{(c)}}, \quad (5)$$

$$\frac{dc_T^{(n)}}{dt} = r_3 c_T^{(c)} - r_4 c_T^{(n)} - \frac{m_6 c_T^{(n)}}{k_6 + c_T^{(n)}}. \quad (6)$$

Here the various rate constants n_k, g_k etc. parameterize transcription (n_k, g_k), degradation (m_k, k_k), translation, (p_k), and the nuclear $\leftrightarrow$ cytoplasmic protein transport, (r_k). There is evidence that LHY and CCA1 proteins bind as a dimer to the promoter of *TOC1* (Daniel et al., 2004), and that there is only one active binding site on the *TOC1* promoter (Alabadi et al., 2001), so the Hill coefficient of transcription, b, was set to 2. As there is no experimental evidence for the Hill coefficient a this was set to 1. The effect of light appears through the term $L(t)$. Light is known to give an acute, transient activation response for expression of *LHY* and *CCA1* (Kim et al., 2003; Kaczorowski and Quail, 2003; Doyle et al., 2003). This was modelled through a simple mechanism involving an interaction of a protein P with the *LHY* gene promoter. P is a light-sensitive protein similar to PIF3 in stability (Bauer et al., 2004) that is present with concentration $c_P^{(n)}$.

$$L(t) = q_1 c_P^{(n)} \Theta_{light}, \quad (7)$$

where $\Theta_{light} = 1$ when light is present, 0 otherwise. We propose that P is controlled by an equation of the form

$$\frac{dc_P^{(n)}}{dt} = (1 - \Theta_{light}) p_3 - \frac{m_7 c_P^{(n)}}{k_7 + c_P^{(n)}} - q_2 \Theta_{light} c_P^{(n)}, \quad (8)$$

where the values of the four parameters that appear in this equation are chosen so as to give an acute light-activation profile which is close to that observed in experiment. In principle these could also be optimized under our scheme but since P is anyway coupled into Eqs. (1)–(6) via an arbitrary coupling constant q_1, which is varied in our optimization scheme, we consider this an adequate approach that captures the primary role of P in mediating light input. The essential features of Eq. (8) are that P is produced only when light is absent and is degraded strongly when light is present. We modelled the acute effect of light at the level of transcription but note that a similar effect on translation would result in essentially identical network behaviour (Kim et al., 2003). This left 23 parameters to be chosen by our optimization scheme.

The equations were solved using MATLAB, integrated using the inbuilt stiff equation solver ODE15s (Shampine and Reichelt, 1997). The optimization process described in the following sections was carried out by compiling the MATLAB code into C and running the code on a 'task farm' super computer consisting of 31 × 2.6 GHz Pentium4 Xeon 2-way SMP nodes (62 CPUs in total).

2.1. Scoring qualitative features of network behaviour

Various experimental data sets (Matsushika et al., 2000; Kim et al., 2003; Mizoguchi et al., 2002) give us approximate values for the phase and period of the oscillations in mRNA levels of the known central clock genes in *Arabidopsis*, see Fig. 2. *TOC1* and *LHY* mRNA expression levels have been shown to peak around dusk and dawn respectively. In contrast neither the absolute nor the relative levels of *LHY* and *TOC1* mRNA are known. Due to the lack of time points in RNA experiments, together with the level of variability (noise) currently ubiquitous in such biological data it is

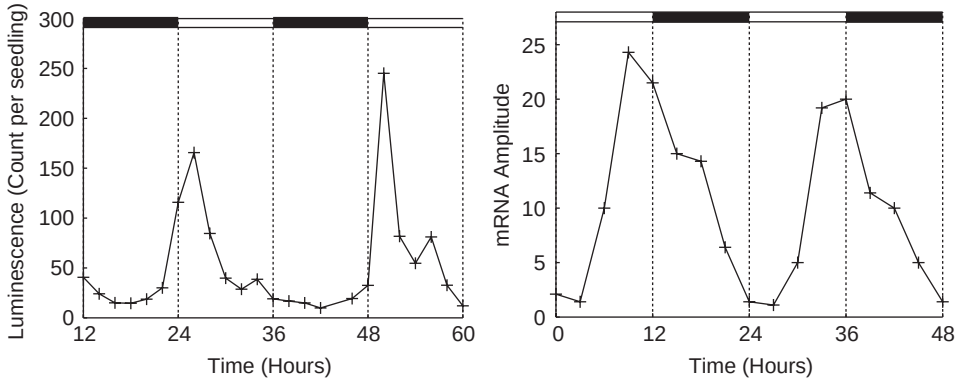

Fig. 2. Typical experimental data sets for the time-variation of mRNA levels for LHY, inferred from the luminescence levels of a Luciferase reporter gene (Kim et al., 2003) (left pane), and TOC1 (Makino et al., 2000) (right pane). Note that the phase of the peak of LHY is just after dawn while the phase of the peak in TOC1 is much later, at approximately 12 h after dawn.

inappropriate to compare a model quantitatively to any particular set of experimental mRNA traces. Although such a direct quantitative comparison has been attempted in other studies (Mendes and Kell, 1998; Forger and Peskin, 2003), as well as in cell cycle models (Zwolak et al., 2001), it is our belief that in cases with sparse experimental data this simply will not be appropriate. This motivated us to construct an empirical cost function designed to give a quantitative value for the goodness of fit of our solution to "essential" qualitative features present in the experiments.

We constructed our cost function Δ as simply a sum of terms that each quantify the agreement between our model and a qualitative experimental feature. Small values of the cost function correspond to a model (or set of parameter values) that give a good qualitative agreement with the corresponding experimental features. Wherever possible the weighting of each term in the cost function was chosen so that an $O(1)$ contribution would be given for an experimentally acceptable error. There is inevitably some arbitrariness in how we defined the level of acceptable error, but at all times we tried to base our assumptions on inferences from experimental data. In order to evaluate the terms in the cost function we solved numerically Eqs. (1)–(8) over 600, 300 h in 12 h light and 12 h dark cycles (LD), followed by 300 h in darkness (DD) (the first 200 h of each solution are discarded as transitory). In what follows we identify 1 nmol and 1 h as the typical concentration and time-scales, and measure all concentrations and rate constants in units where these are unity. We initialized our simulation at $c_i^{(j)} = 1$. The cost function is given by

$$\Delta = \delta_{\tau_{ld}} + \delta_{\tau_d} + \delta_{\phi} + \delta_{c_L} + \delta_{size}. \quad (9)$$

We now discuss the origin of these terms in turn, consigning detailed mathematical definitions to Appendix A: $\delta_{\tau_{ld}}$ measures the difference between the experimental "target" period and the mean period of the oscillation in mRNA levels of *LHY* and *TOC1* in light:dark (LD) cycles exhibited by the model. The term δ_{τ_d} gives a similar measure in constant darkness (DD). These two terms ensure that the entrained and free running clocks are near limit cycles with the experimentally observed period (stably entrained in LD cycles and with a free running period greater than 24 h (Millar et al., 1995)). The third term δ_{ϕ} measures the difference between the target phase and the average phase of the peaks of *LHY* and *TOC1* mRNA expression in LD. It also ensures that the oscillations are entrained to the LD cycles. The term δ_{c_L} contains a measure of how broad the peak of *LHY* mRNA expression is in the proposed solution in LD cycles and is small only if the trace peaks sharply, as observed experimentally. This term is also only small if the peak heights of *LHY* mRNA expression drop when going from LD to DD. Finally, δ_{size} checks that the oscillation sizes are large enough to be detectable experimentally, and quantifies the degree to which the clock in the model is damped in darkness: we require that it is not strongly damped. For details see Appendix A.

2.2. Optimum model from parameter search

We solved our system of equations for 10^6 quasi-random points, each representing a list (vector) of all parameter values. These were generated using a variant of the Sobol Algorithm (Appendix B), a scheme to distribute points (not on any lattice) so as to cover the space as uniformly as possible. We then proceeded to calculate the cost function for these 10^6 random "Sobol" points in parameter space. Encouragingly, the best cost function obtained after N steps appears to converge, see Fig. 3. Also shown is how the best 100 values from the

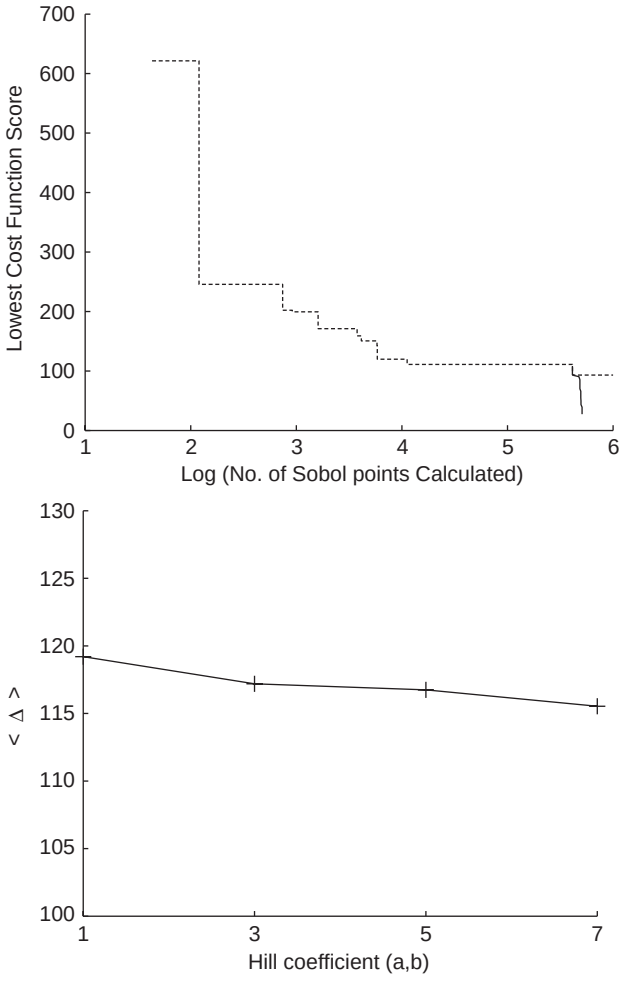

Fig. 3. Convergence of best cost function with number of Sobol points (dotted line), (Upper pane). Also shown is the trace of the annealing steps leading to the optimum solution, with $\Delta = 27$. (black line). The lower pane shows $\langle\Delta\rangle$, the average value of Δ for the best 100 solutions, for varying values of the Hill coefficients, $a = b$. The average value remains relatively constant, suggesting the value of the Hill coefficients is not crucial for the optimization of this model.

Sobol sampling vary with increasing Hill coefficients. Interestingly, we found that the high Hill coefficients adopted in many previous models are not required to obtain a near optimal solution. The 50 top solutions all have $\Delta < 122$ and were broadly distributed in parameter space (mean of parameter values range from 3.53 to 7.08, standard deviations from 2.14 to 3.19). In our scheme the 50 points with the lowest cost function score obtained from the Sobol sampling were passed to a simulated annealing routine (Appendix B). We then used the 41 annealed solutions with $\Delta < 100$ as diverse but reasonable annealed parameter sets for further analysis.

The result of this extensive parameter search is to show that it is not possible to fit the single-loop model to all the experimental data. Fig. 4 shows a typical solution obtained after the simulated annealing, with a score of $\Delta = 81$. *TOC1* mRNA expression peaks too late in the daily cycle, and *LHY* mRNA expression comes up too soon in the night. This solution oscillates on a limit cycle in DD with a period of 25 h. Fig. 5 shows the results for the optimal solution with the lowest cost function score, $\Delta = 27$. In this solution, *TOC1* and *LHY* mRNA levels are both peaking at roughly the correct time, the solution slowly damps in darkness with a period of 25 h, and the light treatment phase-response curve (PRC) has a similar shape to experiment, see Fig. 6. The phases of the peak of *LHY* mRNA and *TOC1* mRNA under LD cycles are not unduly sensitive to light levels (both an 80% reduction in light levels and a 100% increase in light levels causes no significant change in phase). Although not specified in the cost function, LHY protein levels are peaking 1–2 h after the peak of *LHY* mRNA expression, as suggested in Wang and Tobin (1998). However, *LHY* mRNA expression still over anticipates dawn. We found that 5 of the annealed solutions had low cost function scores with the pathological feature that *TOC1* mRNA expression had saturated through the night. This suggests that in future work an extra term in the cost function might be added for the profile shape of *TOC1* mRNA.

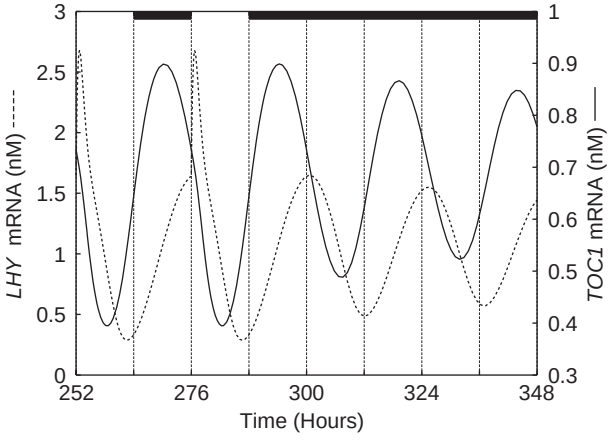

Fig. 4. mRNA Levels for a typical annealed solution, $\Delta = 81$. $q_1 = 4.5703$, $q_2 = 1.0$, $n_1 = 7.5038\,\mathrm{nM/h}$, $n_2 = 0.6801\,\mathrm{nM/h}$, $g_1 = 1.4992\,\mathrm{nM}$, $g_2 = 3.0412\,\mathrm{nM}$, $m_1 = 10.0982\,\mathrm{nM/h}$, $m_2 = 1.9685\,\mathrm{nM/h}$, $m_3 = 3.7511\,\mathrm{nM/h}$, $m_4 = 2.3422\,\mathrm{nM/h}$, $m_5 = 7.2482\,\mathrm{nM/h}$, $m_6 = 1.8981\,\mathrm{nM/h}$, $m_7 = 1.2\,\mathrm{nM/h}$, $p_1 = 2.1994\,\mathrm{h^{-1}}$, $p_2 = 9.4440\,\mathrm{h^{-1}}$, $p_3 = 0.5\,\mathrm{h^{-1}}$, $r_1 = 0.2817\,\mathrm{h^{-1}}$, $r_2 = 0.7676\,\mathrm{h^{-1}}$, $r_3 = 0.4364\,\mathrm{h^{-1}}$, $r_4 = 7.3021\,\mathrm{h^{-1}}$, $k_1 = 3.8045\,\mathrm{nM}$, $k_2 = 5.3087\,\mathrm{nM}$, $k_3 = 4.1946\,\mathrm{nM}$, $k_4 = 2.5356\,\mathrm{nM}$, $k_5 = 1.4420\,\mathrm{nM}$, $k_6 = 4.8600\,\mathrm{nM}$, $k_7 = 1.2\,\mathrm{nM}$, $a = 1$, $b = 2$.

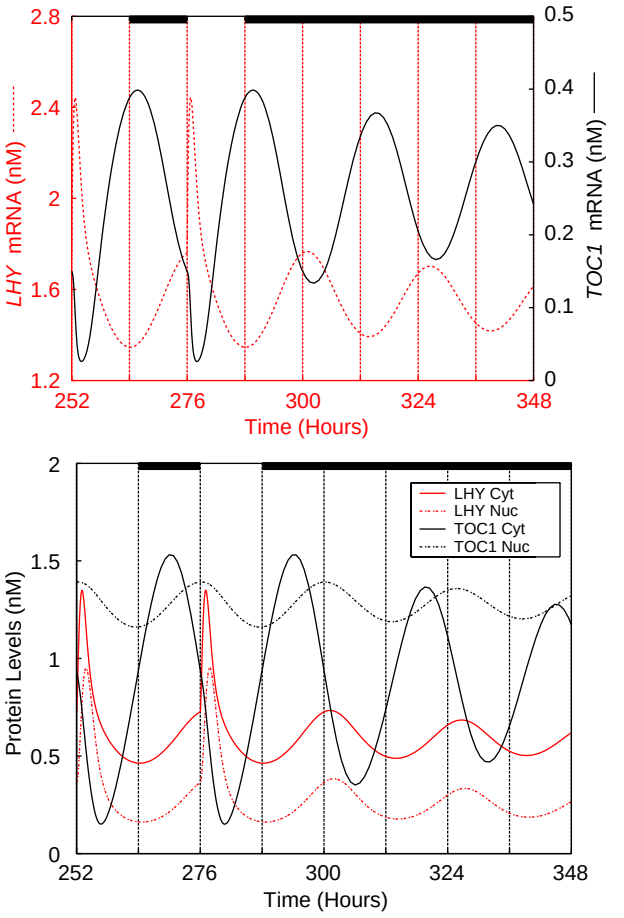

Fig. 5. *Upper pane*: mRNA Levels for optimal solution, $\Delta = 27$. $q_1 = 2.9741$, $q_2 = 1.0$, $n_1 = 16.9711\,\mathrm{nM/h}$, $n_2 = 1.3043\,\mathrm{nM/h}$, $g_1 = 3.0351\,\mathrm{nM}$, $g_2 = 0.386\,\mathrm{nM}$, $m_1 = 9.3383\,\mathrm{nM/h}$, $m_2 = 16.9058\,\mathrm{nM/h}$, $m_3 = 1.0214\,\mathrm{nM/h}$, $m_4 = 1.6859\,\mathrm{nM/h}$, $m_5 = 0.4212\,\mathrm{nM/h}$, $m_6 = 0.0484\,\mathrm{nM/h}$, $m_7 = 1.2\,\mathrm{nM/h}$, $p_1 = 4.9753\,\mathrm{h^{-1}}$, $p_2 = 1.2947\,\mathrm{h^{-1}}$, $p_3 = 0.5\,\mathrm{h^{-1}}$, $r_1 = 1.4563\,\mathrm{h^{-1}}$, $r_2 = 0.8421\,\mathrm{h^{-1}}$, $r_3 = 0.0451\,\mathrm{h^{-1}}$, $r_4 = 0.0018\,\mathrm{h^{-1}}$, $k_1 = 1.3294\,\mathrm{nM}$, $k_2 = 0.8085\,\mathrm{nM}$, $k_3 = 0.1445\,\mathrm{nM}$, $k_4 = 0.2089\,\mathrm{nM}$, $k_5 = 0.3187\,\mathrm{nM}$, $k_6 = 0.3505\,\mathrm{nM}$, $k_7 = 1.2\,\mathrm{nM}$, $a = 1$, $b = 2$. *Lower pane*: Protein levels for optimal solution.

The inadequacies in all the annealed solutions must be explained by the structure of the network, as the global search shows that no distribution of parameters gives a good fit with this network structure. Further examination of the experimental data shows that *TOC1* mRNA expression starts to fall before *LHY* mRNA expression has started to increase, meaning that experimental efforts should be focused on understanding what is missing from the network in order to bring down *TOC1* mRNA levels at night, see Fig. 7. Also, in order for

LHY mRNA to be expressed for a shorter time interval compared to *TOC1* mRNA, post-translational modification of the *TOC1* protein may be required.

Previous papers (Smolen et al., 2001) have carried out a stability analysis based on the period and amplitude of solutions after a small percentage increase and decrease for each parameter value in turn. We have also carried out such an analysis which shows that, as for previous models, the solution is extremely sensitive to small changes of a few parameters, see Fig. 8. For example, a reduction of five percent in n_2, the transcription rate of *TOC1* mRNA, causes the oscillations to damp to experimentally undetectable levels after 300 h in darkness. We also undertook a more detailed stability analysis, (data not shown), measuring the variation of cost function values with small parameter changes. In this case the results were similar to that of the stability analysis solely looking at period and amplitude change in DD, although additional information could be observed, such as if the solution becomes biphasic in LD cycles.

2.3. Simulated mutant analysis

We have also further characterized the output of the 41 annealed parameter sets with $\Delta < 100$, by carrying out a simulated mutant analysis, see Fig. 9. The mutant analysis reveals further interesting information about the role of *LHY* in the network. Single null mutations in *LHY* or *CCA1* experimentally result in short periods of around 21 h (Alabadi et al., 2002). We simulated this experiment by reducing the translation rate of LHY protein to half of its original value, and for every single one of our annealed parameter sets this resulted in a

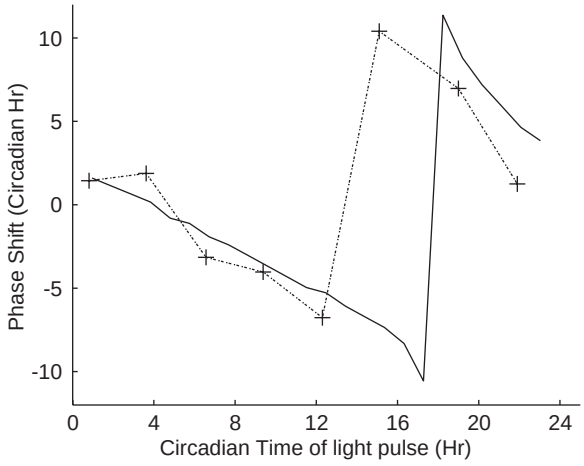

Fig. 6. Simulated phase response curve (PRC) for optimal solution (solid curve). Phase shifts in *TOC1* mRNA expression generated by 1 h light treatments are plotted against the circadian time at which the light pulses were given. Phase advances are plotted as positive values, and delays are plotted as negative values. The simulation followed the same experimental protocol as in the data extracted from a published red light PRC (Covington et al., 2001) (dashed curve).

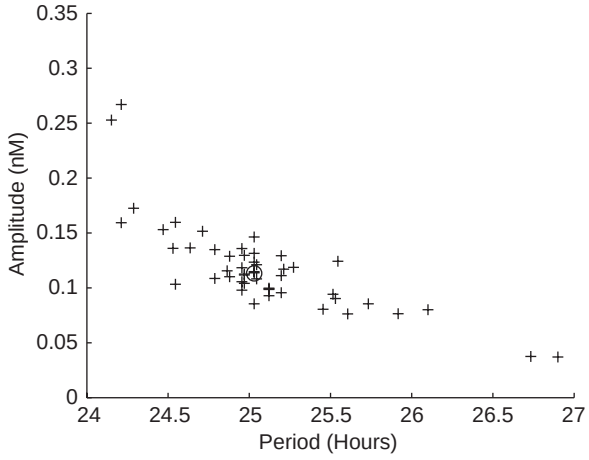

Fig. 8. Stability analysis for optimal solution. The mean period and amplitude over 300 h in DD are calculated for a small perturbation, a 5 percent increase and decrease, for each parameter value in turn. The circle represents the period and amplitude of the original parameter values.

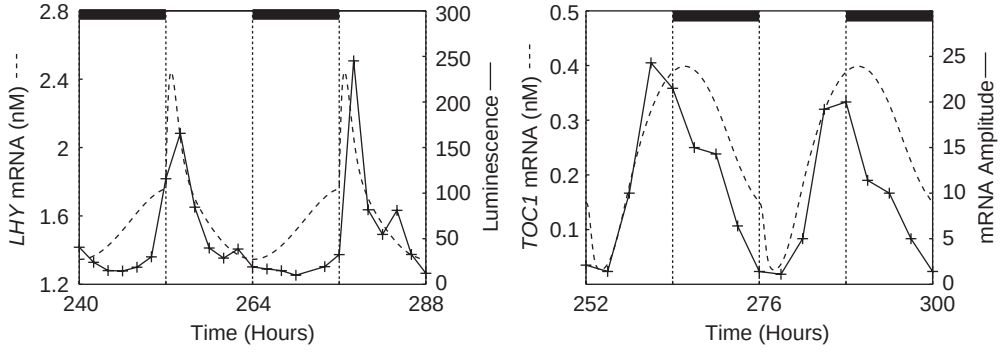

Fig. 7. Comparison between experimental concentrations (solid curves) with those obtained from our optimal model (dashed curves), as also shown in Figs. 2 and 5, respectively. Simulation models phases of *LHY* and *TOC1* mRNA correctly. *LHY* anticipates dawn in the simulation to a greater degree than experiment. Experimental traces show *TOC1* levels are falling during the night before *LHY* levels start to rise, suggesting that our model is missing some factor that would serve to bring *TOC1* down at the end of the day.

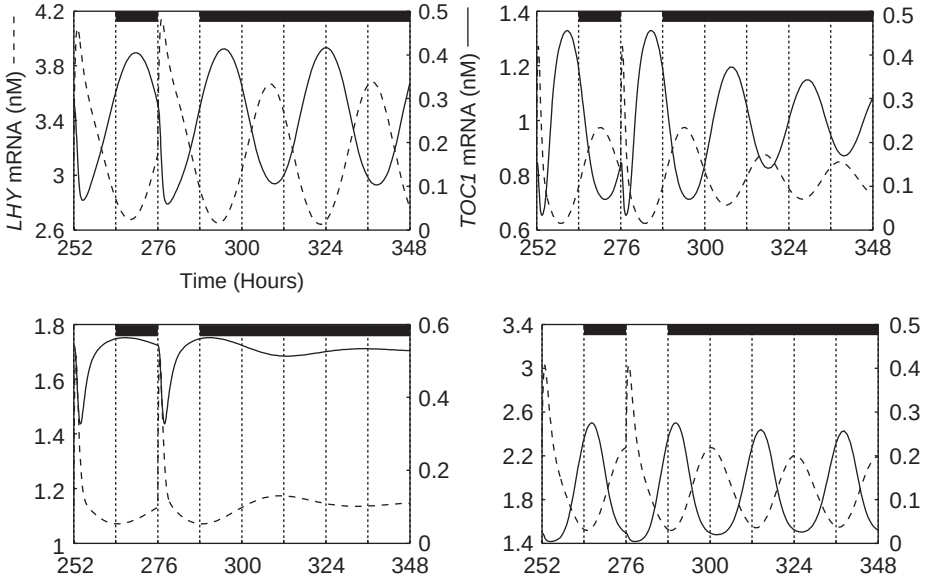

Fig. 9. Mutant analysis on optimal solution. Upper left pane, Translation rate of LHY, (p_1) halved, the average period over the 300 h in DD, $\tau_d = 29.2$ h. Upper right pane, p_1 doubled, $\tau_d = 21.3$ h. Lower left pane, translation rate of TOC1, (p_2), halved, $\tau_d = 44.7$ h. Lower right pane, p_2 doubled, $\tau_d = 23.5$ h.

long period mutant or arrhythmia. This is because in our model, a reduction in LHY level delays the repression of *TOC1*, so *TOC1* mRNA peaks later in the night, a feature that is not observed experimentally. This discrepancy suggests that the main role of LHY in the plant clock is to stop *TOC1* mRNA peaking too early in the day, rather than to bring *TOC1* levels down during the night, which it does in the model. All of the 41 solutions with $\Delta < 100$ showed similar mutant analysis, suggesting that the failure in this regard is a generic feature of the network.

3. Discussion

We have modelled the LHY/CCA1–TOC1 feedback loop of *Arabidopsis*, which was first proposed by Alabadi et al. (2001) as an important component of the plant circadian clock. We developed a cost function to score the performance of the model with a particular set of parameter values, together with an exhaustive method to explore the space of possible parameter values, in order to identify an optimal set of parameter values. The resulting model recapitulates many properties of *Arabidopsis* circadian rhythms and points to a specific phase of the circadian cycle, where further experimentation is required to identify one or more additional components of the circuit. A significant advantage of our approach is in determining that a gene network model is inconsistent with experimental data because its circuit is incorrect, not due to a poor choice of parameter values.

The LHY/CCA1–TOC1 loop was proposed, based upon the mRNA accumulation patterns of *Arabidopsis* plants in which these genes were mutated or misexpressed (Alabadi et al., 2001; Wang and Tobin, 1998; Schaffer et al., 1998). Our model of this loop drives rhythms with the major circadian properties observed in *Arabidopsis*, which were specified in the cost function used to select the parameter values. The model has a period close to but longer than 24 h in constant darkness, it entrains to 24 h light:dark cycles, and it shows peak expression of the *LHY*/*CCA1* and *TOC1* mRNAs close to the required phases. It also reproduces features that were not explicitly specified in the cost function, such as a PRC to light treatments with the characteristic shape, Fig. 6. For some parameter values, the model has a sustained limit cycle in constant darkness, Fig. 4, though the optimal solution is slightly damped, as is often observed in experimental data. The model supports rhythms in constant light, as it does in darkness (data not shown). We do not address any temperature effect explicitly, due to the paucity of relevant molecular data. The model is likely to be capable of temperature compensation, as much simpler models can exhibit this behaviour (Ruoff and Rensing, 1996).

The waveforms of *LHY* and *TOC1* expression during the day match the data closely, both for the optimal parameter choice, Fig. 7, and with many of the optimized parameter sets selected from a million quasi-random sets, Fig. 4, indicating that no further biochemical complexity is necessary to simulate this phase of the cycle. However, even the optimal parameter choice fails

to reproduce mRNA accumulation patterns during the night. *LHY* mRNA levels start to rise early in the night in the optimal parameter choice under LD cycles, anticipating the observed rise by several hours. Simulated mutant analysis highlights this limitation more clearly. Deletion of either the *LHY/CCA1* or the *TOC1* function leads trivially to arrhythmia, so another timing loop must be present in the plant, because complete arrhythmia is observed in the relevant mutants only in specific conditions (Mas et al., 2003), if at all (Matsushika et al., 2000; Alabadi et al., 2002). Such components might simply be paralogues of *LHY/CCA1* and *TOC1* with non-identical functions (Carre and Kim, 2002; Eriksson and Millar, 2003) which are absent from the model. More significantly, partial reduction or increase in either gene function in the model causes period changes in the opposite direction to those observed, Fig. 9. This discrepancy cannot be explained by partial gene redundancy in the plant, nor by a feature specific to the optimal parameter set in the model, as it was observed in all of the annealed solutions with $\Delta < 100$. It arises because LHY in the model represses *TOC1* expression continuously from the mid-night phase, so reduced LHY function leads to a later peak of *TOC1* expression at night and a longer period. Neither LHY nor CCA1 expression is observed until close to dawn in the plant (Wang and Tobin, 1998; Kim et al., 2003), so *LHY* and *CCA1* mutants allow an early rise in TOC1 expression in the day and have a short period. The component(s) that causes the fall in TOC1 expression during the night remains to be identified experimentally.

3.1. Model development with limited data

A cost function was used to constrain the behaviour of the model to the consistent, qualitative features of the experimental data, because only a few time-series of the main biochemical components in the model were available when we initiated modelling. This approach is effective even for noisy or underdetermined systems, when data are scarce and variable (for example, among different laboratories) so fitting to data is undesirable or impossible. It is easy to understand intuitively and can be expanded along with the available data. Terms can be added to the cost function to fit the phenotypes of mutants, for example. The terms in the cost function developed here are specifically based on the experimental data known for the *Arabidopsis* network, and each term would have to be re-examined if applied to a different network. However, a cost function containing similar terms as described here should be useful for analysing circadian networks in other organisms. The cost function can potentially combine the qualitative terms, used here, with direct fitting to data. A clock model was recently fitted to a particular set of molecular data for the first time (Forger and Peskin, 2003), though without specifying how the starting parameter values were chosen. The qualitative terms will no longer be required when sufficient, reliable time-series data are available for fitting. The need for fitting will in turn be reduced as parameter values are measured experimentally.

Neither parameter values nor the absolute levels of any component in our clock model have been measured, a situation that is almost inevitable when any gene network is first defined. Many published circadian models have found parameter values by manual fitting. This leads to a bias towards models that perform well over a broad range of parameter values (for example those with high Hill coefficients (Kurosawa et al., 2002)), because a limited exploration of parameter space is likely to find a sufficiently good solution only for these models. A model may still fail to fit the data, because the arbitrary choice of parameter values was incorrect. The more informative result is that the model fails because the network structure is incorrect.

We therefore developed a bipartite method to search parameter space exhaustively, with no prior knowledge of parameter values. A million starting parameter sets were generated quasi-randomly, leaving fewer gaps and generating fewer clusters than a truly random process, and only bounding each parameter at a maximum value. The 50 parameter sets that gave the lowest cost function scores were taken as starting points for optimization by simulated annealing. This takes predominantly small, random steps in parameter space around each starting point. Parameter changes that improve the cost function score have a high probability of being retained, so the method converges on the best parameter set near the starting parameter values. Critically, the optimization can escape local minima and approach the global minimum, because some parameter steps are large and changes that increase the cost function are sometimes retained. Our method is computationally intensive but achieves a far greater coverage of parameter space than the most extensive manual search, finding good solutions with a maximum Hill coefficient of 2. Optimization may require fewer parameter steps if these can be directed (Brown and Sethna, 2003), but our method clearly converges upon parameter values that give an excellent match to the data, (Fig. 7). The failure of the model to reproduce gene expression patterns during the night or the gross phenotypes of clock mutants, despite this global search of parameter space, indicates that the gene network in the model is incorrect. We expect that more complete models of the plant circadian system will have to incorporate additional feedback loops, not only to complete the modelling of the gene circuits we instigate here, but also to incorporate other potential oscillators, such as nitrase reductase activity (Lillo and Ruoff, 1984). We look forward to the experimental characterization of components that will complete the

model, especially those that function during the subjective night. Several genes are candidates for such components, including *EARLY-FLOWERING 3*, *EARLY-FLOWERING 4*, *GIGANTEA*, and *TIME FOR COFFEE*, though their biochemical functions are presently unknown [reviewed in Eriksson and Millar (2003)].[1]

Acknowledgements

The authors would like to express their gratitude to T. Mizoguchi and I.A. Carre for supplying *TOC1* mRNA and *LHY* mRNA data, respectively. This work was supported by a studentship from the Gatsby foundation (JCWL) and by the UK BBSRC (AJM). The computer facilities were provided by the Centre of Scientific Computing of the University of Warwick.

Appendix A

Here, we describe each term of the cost function, Eq. (9), in turn.

Firstly,

$$\delta_{\tau_{ld}} = \sum_{i=L,T} \langle (24 - \tau_i^{(m)})^2 / 0.15 \rangle_{ld}, \tag{A.1}$$

is the summed error in the period, τ, for *LHY* (L) and *TOC1* (T) mRNA levels in light:dark cycles (LD), where $\langle\rangle_{ld}$ gives the average over the cycles between $200 < t < 300$, and a "marginally acceptable" period difference of ≈ 25 min contributes $O(1)$ to the cost function.

Secondly,

$$\delta_{\tau_d} = \sum_{i=L,T} \langle (25 - \tau_i^{(m)})^2 / f \rangle_d, \tag{A.2}$$

where the average of $\langle\rangle_d$ is now over $300 < t < 600$ (DD). The biological evidence strongly indicates that the free running period of the clock is greater than 24 (Millar et al., 1995), probably about 25, but we have less confidence in assigning a precise value hence we adopt values of $f = 0.05$ if $\tau_i^{(m)} \leqslant 25$ and $f = 2$ if $\tau_i^{(m)} > 25$.

Thirdly,

$$\delta_\phi = \sum_{i=L,T} \left[\langle \Delta\Phi_i^2 \rangle_{ld} + \left(\frac{\sigma[c_i^{(m)}(t_p)]_{ld}}{0.05 \langle c_i^{(m)}(t_p) \rangle_{ld}} \right)^2 + \left(\frac{\sigma[\Delta\Phi_i]}{5/60} \right)^2 \right] + \delta_{ent}. \tag{A.3}$$

[1]A clock modelling package including our *Arabidopsis* model with the optimal parameter choice can be found at http://www.amillar.org under Software.

The first term compares the mean difference in phase over the LD cycles, where $\Delta\Phi_i = \bar{\phi}_i - \phi_i$, ϕ_i is the phase (from dawn) of the RNA peak in the model and $\bar{\phi}_L = 1$ h, $\bar{\phi}_T = 11$ h are the target phases of the peaks in $c_L^{(m)}$ and $c_T^{(m)}$, respectively. We assume a cost that is $O(1)$ for solutions that differ by an hour. The next two terms ascribe a cost of $O(1)$ for limit cycle solutions in LD cycles whose peak heights are within 5 percent, and whose variations in peak phases are 5 min. $\sigma[]_{ld}$ is the standard deviation for the cycles in LD. The term δ_{ent} checks that the solution is truly entrained to the light/dark cycle, i.e. is not oscillating with the correct phase simply because of the initial conditions chosen, as follows: The solution is rerun for 75 h, taking the solution at 202 h and shifting it back 3 h, i.e. initializing the $t = 202$ solution as the $t = 199$ solution. The new phase of the second peak is compared to the original phase of the second peak. If the phase difference is still near 3 h, then the solution is too weakly entrained, and the solution is pathological. The LD cycles have failed to phase shift the response. Hence δ_{ent} takes the form of $\log(0.5)/\log(\delta\phi/3)$, where $\delta\phi$ is the phase difference in hours between the shifted and original solution, and $\delta\phi/3$ is therefore the fraction of the imposed 3 h phase shift remaining after 2 periods. The term $\log(0.5)$ gives the acceptable remaining phase difference of 1.5 h for the second cycle, which results in an $O(1)$ contribution to the cost function.

Next,

$$\delta_{size} = \sum_{i=L,T} \left(\frac{1}{\langle \Delta c_i^{(m)} \rangle_{ld}} \right)^2 + \left(\frac{\tau_o}{\tau_e} \right)^2. \tag{A.4}$$

The first term costs for solutions in LD cycle with oscillation sizes ($\Delta c_i^{(m)} = c_{i\max}^{(m)} - c_{i\min}^{(m)}$), less than 1 nm, and the second term checks that the oscillations do not decay too quickly when entering DD as follows: τ_o is a decay constant over the 300 h in DD, $\tau_o = -300/\log((\Delta c_{T\,ld}^{(m)} - \Delta c_{T\,d}^{(m)})/\Delta c_{T\,ld}^{(m)})$, and τ_e gives the acceptable decay constant, that the size of *TOC1* oscillations has dropped by $\frac{1}{4}$ over 300 h, $-300/\log(0.75)$.

Finally,

$$\delta_{c_L} = \sum_{i=2,-2} \left\langle \left(\frac{2/3 c_L^{(m)}(t_p)}{c_L^{(m)}(t_p) - c_L^{(m)}(t_p + i)} \right)^2 \right\rangle_{ld} + \cdots \left\langle \left(\frac{0.05(c_L^{(m)}(t_p - 2) - c_L^{(m)}(t_m))}{c_L^{(m)}(t_m) - c_L^{(m)}(t_m + i)} \right)^2 \right\rangle_{ld} + 10 \left(\frac{\langle c_L^{(m)}(t_{pd}) \rangle_{ld}}{\langle c_L^{(m)}(t_{pl}) \rangle_{ld}} \right)^4. \tag{A.5}$$

The first term checks that the *LHY* mRNA expression profile has a sharp peak in LD cycles, with an $O(1)$ contribution if *LHY*'s expression level has dropped by $\frac{2}{3}$ of its oscillation size within 2 h before and after its peak

of expression. The second term checks that *LHY* mRNA expression has a broad minimum, with an $O(1)$ contribution if 2 h before and after the minimum point *LHY*'s expression has only increased to 5 percent of the level 2 h before *LHY*'s peak. The last term checks that the peak of *LHY* mRNA expression drops from LD into DD, as it loses its light activation.

Throughout the implementation the cost function was "capped" at $\Delta_{\max} = 10^4$, such that $\Delta \to Min(10^4, \Delta)$.

Appendix B

We implemented the Antoneev–Saleev variant of the Sobol quasi-random number generator to choose parameter values (vectors) in our $d = 23$ parameter space, adapted from Press et al. (1996). The initial values as described in Joe and Kuo (2003) were used, allowing number generation in up to 1111 dimensions. We carried out the following change of parameters $\hat{g}_1 = g_1^a, \hat{g}_2 = g_2^b, \hat{n}_2 = n_2 g_2^b$, and then chose the parameter space for the 23 parameters to be optimized to be bounded $\in [0, 10]$, where our typical parameter scale is unity in nmol = hours = 1. We annealed according to a standard Metropolis algorithm (Brooks and Morgan, 1995). Temperature reduction was carried out linearly each step, from $T_{\max}$ to 0 over the 100 000 annealing steps. $T_{\max}$ was set as the average cost function value of the best 50 solutions. The step size $|\delta \mathbf{r}|$ in parameter space was set to allow the optimization routine after 100 000 annealing steps to travel a distance approximately equal to the distance between two neighbouring Sobol points. This approach yields $|\delta \mathbf{r}| = 0.0431$, and the solution was reset to the best found solution if a better solution had not been found over the previous 10 000 points (see Fig. 3).

References

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.

Alabadi, D., Yanovsky, M.J., Mas, P., Harmer, S.L., Kay, S.A., 2002. Critical role for CCA1 and LHY in maintaining circadian rhythmicity in *Arabidopsis*. Curr. Biol. 12, 757–761.

Bauer, D., Viczian, A., Kircher, S., Nobis, T., Nitschke, R., Kunkel, T., Panigrahi, K.C., Adams, E., Fejes, E., Schafer, E., Nagy, F., 2004. Constitutive photomorphogenesis 1 and multiple photoreceptors control degradation of phytochrome interacting factor 3, a transcription factor required for light signaling in *Arabidopsis*. Plant Cell 16, 1433–1445.

Brooks, S.P., Morgan, B., 1995. Optimization using simulated annealing. The Statistician 44, 241–257.

Brown, K.S., Sethna, J.P., 2003. Statistical mechanical approaches to models with many poorly known parameters. Phys. Rev. E 68.

Carre, I.A., Kim, J.Y., 2002. MYB transcription factors in the *Arabidopsis* circadian clock. J. Exp. Bot. 53, 1551–1557.

Christensen, M., Falkeid, G., Loros, J., Dunlap, J., Lillo, C., Ruoff, P., 2004. A nitrate-induced frq-less oscillator in *Neurospora crassa*. J. Biol. Rhythms 19, 280–286.

Covington, M., Panda, S., Liu, X., Strayer, C., Wagner, D., Kay, S., 2001. ELF3 modulates resetting of the circadian clock in arabidopsis. Plant Cell 13, 1305–1315.

Daniel, X., Sugano, S., Tobin, E.M., 2004. CK2 phosphorylation of CCA1 is necessary for its circadian oscillator function in *Arabidopsis*. Proc. Natl. Acad. Sci 101, 3292–3297.

Doyle, M.R., Davis, S.J., Bastow, R.M., McWatters, H.G., Kozma-Bognar, L., Nagy, F., Millar, A.J., Amasino, R., 2003. The ELF4 gene controls circadian rhythms and flowering time in *Arabidopsis thaliana*. Nature 419, 74–77.

Dunlap, J., 1999. Molecular bases of circadian clocks. Cell 96, 271–290.

Dunlap, J.C., Loros, J.L., DeCoursey, P.J., 2003. Chronobiology: Biological TimeKeeping. Sinauer, Sunderland.

Eriksson, M.E., Millar, A.J., 2003. The circadian clock. A plant's best friend in a spinning world. Plant Physiol. 132, 732–738.

Forger, D.B., Peskin, C.S., 2003. A detailed predictive model of the mammalian clock. Proc. Natl. Acad. Sci. 100, 14806–14811.

Goldbeter, A., 2002. Computational approaches to cellular rhythms. Nature 420, 238–245.

Hall, A., Bastow, R.M., Davis, S.J., Hanano, S., Mcwatters, H.G., Hibberd, V., Doyle, M.R., Sung, S., Amasino, K.J.H.R.M., Millar, A.J., 2003. The TIME FOR COFFEE gene maintains the amplitude and timing of *Arabidopsis* circadian clocks. Plant Cell 15, 2719–2729.

Harmer, S.L., Hogenesch, J.B., Straume, M., Chang, H.S., Han, B., Zhu, T., Wang, X., Kreps, J.A., Kay, S.A., 2000. Orchestrated transcription of key pathways in *Arabidopsis* by the circadian clock. Science 290, 2110–2113.

Joe, S., Kuo, F.Y., 2003. Remark on algorithm 659: implementing sobol's quasirandom sequence generator. ACM Trans. Math. Software (TOMS) Arch. 29, 49–57.

Kaczorowski, K.A., Quail, P.H., 2003. *Arabidopsis* PSEUDO-RESPONSE REGULATOR7 is a signaling intermediate in phytochrome-regulated seedling deetiolation and phasing of the circadian clock. Plant Cell 15, 2654–2665.

Kim, J.Y., Song, H.R., Taylor, B.L., Carre, I.A., 2003. Light-regulated translation mediates gated induction of the *Arabidopsis* clock protein LHY. EMBO J. 22, 935–944.

Kurosawa, G., Iwasa, Y., 2002. Saturation of enzyme kinetics in circadian clock models. J. Biol. Rhythms 17, 568–577.

Kurosawa, G., Mochizuki, A., Iwasa, Y., 2002. Comparative study of circadian clock models, in search of processes promoting oscillation. J. Theor. Biol. 216, 193–208.

Leloup, J.C., Goldbeter, A., 2003. Toward a detailed computational model for the mammalian circadian clock. Proc. Natl. Acad. Sci 100, 7051–7056.

Leloup, J.C., Gonze, D., Goldbeter, A., 1999. Limit cycle models for circadian rhythms based on transcriptional regulation in *Neurospora* and *Drosophila*. J. Biol. Rhythms 14, 433–448.

Lillo, C., Ruoff, P., 1984. A minimal model of light-induced circadian rhythms of nitrate reductase activity in leaves of barley. Plant Physiol. 62, 589–592.

Makino, S., Kiba, T., Imamura, A., Hanaki, N., Nakamura, A., Suzuki, T., Taniguchi, M., Ueguchi, C., Sugiyama, T., Mizuno, T., 2000. Genes encoding pseudo-response regulators: insight into His-to-Asp phosphorelay and circadian rhythm in Arabidopsis thaliana. Plant Cell Physiol. 41, 791–803.

Mas, P., Kim, W.Y., Somers, D.E., Kay, S.A., 2003. Targeted degradation of TOC1 by ZTL modulates circadian function in *Arabidopsis thaliana*. Nature 426, 567–570.

Matsushika, A., Makino, S., Kojima, M., Mizuno, T., 2000. Circadian waves of expression of the APRR1/TOC1 family of

pseudo-response regulators in *Arabidopsis thaliana*: an insite into the plant circadian clock. Plant Cell Physiol. 41, 1002–1012.

McWatters, H.G., Bastow, R.M., Hall, A., Millar, A.J., 2000. The ELF3 zeitnehmer regulates light signalling to the circadian clock. Nature 6813, 716–720.

Mendes, P., Kell, D.B., 1998. Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. Bioinformatics 14, 869–883.

Millar, A.J., Straume, M., Chory, J., Chua, N.H., Kay, S.A., 1995. The regulation of circadian period by phototransduction pathways in *Arabidopsis*. Science 267, 1163–1166.

Mizoguchi, T., Wheatley, K., Hanzawa, Y., Wright, L., Mizoguchi, M., Song, H.R., Carre, I.A., Coupland, G., 2002. LHY and CCA1 are partially redundant genes required to maintain circadian rhythms in *Arabidopsis*. Dev. Cell 2, 629–641.

Press, W.H., Teukolsky, S.A., Vettering, W.T., Flannery, B.P., 1996. Numerical Recipes in C: The Art of Scientific Computing. Cambridge University Press, Cambridge.

Ramlaho, C., Hastings, J., Colepicolo, P., 1995. Circadian oscillation of nitrate reductase activity in *Gonyaulax polyedra* is due to changes in cellular protein levels. Plant Physiol. 107, 225–231.

Ruoff, P., Rensing, L., 1996. The temperature-compensated Goodwin model simulates many circadian clock properties. J. Theor. Biol. 179, 275–285.

Ruoff, P., M Vindjevik, C.M., Rensing, L., 1999a. The Goodwin oscillator: on the importance of degradation reactions in the circadian clock. J. Biol. Rhythms 14, 469–479.

Ruoff, P., Vindjevik, M., Mohsenzadeh, S., Rensing, L., 1999b. The Goodwin model: simulating the effect of cycloheximide and heat shock on the sporulation rhythm of *Neurospora crassa*. J. Theor. Biol. 196, 483–494.

Ruoff, P., M Vindjevik, C.M., Rensing, L., 2001. The Goodwin model: simulating the effect of light pulses on the circadian sporulation rhythm of *Neurospora crassa*. J. Theor. Biol. 209, 29–42.

Schaffer, R., Ramsay, N., Samach, A., Corden, S., Putterill, J., Carre, I.A., Coupland, G., 1998. The late elongated hypocotyl mutation of *Arabidopsis* disrupts circadian rhythms and the photoperiodic control of flowering. Cell 93, 1219–1229.

Shampine, L.F., Reichelt, M., 1997. The MATLAB ODE suite. SIAM J. Sci. Comput. 18, 1–22.

Smolen, P., Baxter, D., Byrne, J.H., 2001. Modeling circadian oscillations with interlocking positive and negative feedback loops. J. Neurosci. 21, 6644–6656.

Tyson, J., Hong, C., Thron, C., Novak, B., 1999. A simple model of circadian rhythms based on dimerization and proteolysis of PER and TIM. J. Biophys. 77, 2411–2417.

Ueda, H.R., Hagiwara, M., Kitano, H., 2001. Robust oscillations within the interlocked feedback model of drosophila circadian rhythm. J. Theor. Biol. 210, 401–406.

Wang, Z.Y., Tobin, E.M., 1998. Constitutive expression of the circadian clock associated 1 (CCA1) gene disrupts circadian rhythms and suppresses its own expression. Cell 93, 1207–1217.

Zwolak, J.W., Tyson, J.J., Watson, L.T., 2001. Estimating rate constants in cell cycle models. Proceedings of the High Performance Computing Symposium 2001, San Diego, CA, pp. 53–57.