Optimising time-series experimental design for modelling of circadian rhythms: the value of transient data

Laurent Mombaerts * **Alexandre Mauroy** **
Jorge Gonçalves ***

* *Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg (e-mail: laurent.mombaerts@uni.lu)*
** *LCSB, University of Luxembourg (e-mail: alexandre.mauroy@uni.lu)*
*** *LCSB, University of Luxembourg (e-mail: jmg@uni.lu)*

Abstract: Circadian clocks consist of complex networks that coordinate the daily cycle of most organisms. In light/dark cycles, the clock is synchronized (or entrained) by the environment, which corresponds to a constant rephasing of the oscillations and leads to a steady state regime. Some circadian clocks are endogenous oscillators with rhythms of about 24-hours that persist in constant light or constant darkness. This operating mechanism with and without entrainment provides flexibility and robustness to the clock against perturbations. Most of the clock-oriented experiments are performed under constant photoperiodic regime, overlooking the transitory regime that takes place between light/dark cycles and constant light or darkness. This paper provides a comparative analysis of the informative potential of the transient time-series data with the other regimes for clock modelling.
Realistic data were simulated from 2 experimentally validated plant circadian clock models and sliced into several time windows. These windows represent the different regimes that take place before, meanwhile and after the switch to constant light. Then, a network inference tool was used over each window and its capability of retrieving the ground-truth of the network was compared for each window. The results suggest that including the transient data to the network inference technique significally improves its performance.

Keywords: Network inference; Modelling; Dynamics and control; Circadian rhythms; Systems Medicine.

1. INTRODUCTION

Most organisms have developed the capability of synchronizing their life cycle to the environment. As regards plants, the circadian clock regulatory network is responsible for controlling diverse biological processes, such as photosynthesis and flowering (Michael et al., 2008). Circadian clocks generate oscillations in gene expression and are able to synchronize to external conditions, such as temperature and light/dark (L/D) cycles (Dunlap et al., 2003). Moreover, they have the attribute of being endogenously generated and self-sustaining. When deprived of external cues, such autonomously sustained oscillations of gene expression persist with a free-running period of approximately 24 hours (Zhang and Kay, 2010).

Over the past 20 years, the circadian clock of one plant, *Arabidopsis Thaliana*, has been intensively studied. Several mathematical models have emerged, which fit the experimental data, either in light/dark cycles or constant (light (LL) or dark (DD)) condition, and elucidate the minimal regulatory structure. Conceptually, the circadian clock is composed of 3 mains components: a self-sustaining central oscillator, an input pathway that incorporates the environment conditions, and an output pathway that adjusts the plant's metabolism. The central oscillator of the clock consists in a complex network of interlocking genes activations, inhibitions and feedback loops.

The identification of the functional properties of the individual components in gene regulatory networks is challenging, due to the complexity of the interlocked network. Although several techniques have been developed for inferring genes regulatory networks from time series data (Aderhold et al., 2014), there has been no comparative study of the performance resulting from the use of different experimental time windows as an input.

During L/D cycles, the circadian central network is entrained by the light and gene expressions of similar dynamics are synchronized. Consequently, the identification of the underlying processes between the genes that compose the clock can be problematic. Indeed, the small differences in dynamical gene expression rates that could help distinguishing the subtle regulations of the genes may have faded. When switching to a constant light input, the oscillations are not forced and synchronized anymore, releasing the system to free-running condition. The system now exhibits several distinct phases where only one was observed. The transitory regime that follows the switch to the new constant condition (either constant light or constant darkness) can be short before establishing the

new regime, as described by several mathematical circadian clock models governed by deterministic differential equations. However, this time window may have the potential to shortly reveal supplementary dynamics between the genes that constitute the network.

This paper explores the performance of one network identification tool to evaluate and compare the informative potential from each time window for time-series data. We expect that other tools would lead to the same conclusion. This tool uses linear time-invariant (LTI) models to describe the causality between pairs of genes. This technique has several advantages over other inference techniques. Firstly, it is especially useful for poorly-sampled data corrupted with noise. Secondly, linear models have been proved to be a good approximation of the nonlinear biological phenomena and less prone to over-fitting than most nonlinear methods. Thirdly, this technique has been previously used for the identification of biochemical models (Dalchau et al. (2010),Herrero et al. (2012)) and circadian systems (Carignano, 2014). Finally, this tool has the capability to include hidden variables that are not part of the input/output pair, such as intermediate transcription factor, but that are necessary to describe the biological dynamics between 2 genes.

To produce realistic reference time-series data for system identification, we used the Millar 10 (Pokhilko et al., 2010) and Millar 12 (Pokhilko et al., 2012) models of the Arabidopsis circadian clock. These models describe the central clock through the modelling of 8 to 10 genes and the intervention of several intermediate transcription factors. They are nonlinear and mostly based on detailed dynamics including Michaelis-Menten and Hill equation dynamics.

2. METHODS

2.1 Generation of Realistic Data

Data were simulated from the Millar 10 and Millar 12 Arabidopsis circadian clock models (see Figure 1). These models have been simulated for 600 hours in 24-hour L/D cycles to remove all possible transients. Then, we switched the photoperiodic regime to constant light for another 108 hours. In order to reproduce realistic data, stochastic differential equations (SDE) were employed to account for genes intrinsic noise. SDE simulations of Millar 10 and Millar 12 models are computationally light, of the order of several seconds (MacBook Pro). This approach allows the noise to have a significant impact on the behaviour of the system (Wilkinson, Guerriero et al. (2012)). However, estimating the level of process noise is not trivial since many experiments provide only a few replicates of the dataset. Hence, a density analysis of the noise has been performed on 40 000 experimental RNA sequencing (RNA-Seq) paired data. RNA-Seq is a recently developed sequencing technology that provides a far more precise measurement of levels of transcripts than other methods. Due to the mean-variance dependency of RNA-Seq data, the noise distribution of this experimental dataset was computed by comparing the error between 2 datapoints over their mean (see Figure 2). Finally, the noise parameter of the SDE has been tuned to obtain an equivalent

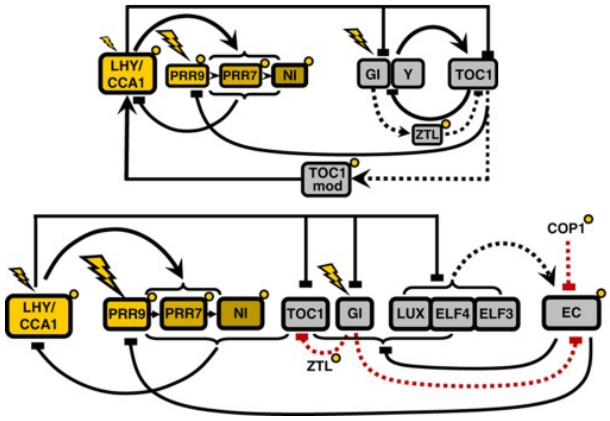

Fig. 1. Models used to obtain the simulated data. Top: Millar 10. Bottom: Millar 12 (Pokhilko et al., 2012).

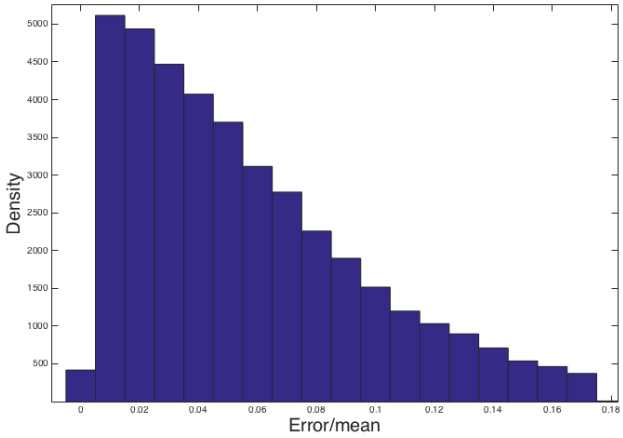

Fig. 2. Distribution of the experimental error over mean for 2 replicates over 40 000 RNAseq datapoints.

distribution of the noise for the simulated data. This setup allows to reproduce as conscientiously as possible the desynchronization phenomenon and the damping of the oscillations resulting from the switch to constant light (Guerriero et al., 2012). For each simulation, we generated 2 replicates representing 2 L/D cycles followed by 4 cycles with constant light input.In order to match realistic RNAseq experiments as closely as possible, only one data point per replicate was selected every 4 hours and only the mean of the 2 replicates was considered for the following steps. Then, this subset of the data was resampled using an interpolating cubic spline algorithm. By doing so, we could decrease the sampling period to 1 hour, which is a requirement for our inference algorithm.
We reproduced the aforementioned method by generating 50 pairs of replicates, thereby reducing the effect of noise-dependent performances of the network inference technique. Finally, each dataset was sliced into 10 equally spaced windows of 48 hours with a step size of 12 hours. Each window contains 49 (interpolated) data points. The first window starts at -48 hours and the last window starts at +60 hours (Figure 3). Constant light input starts at time 0.

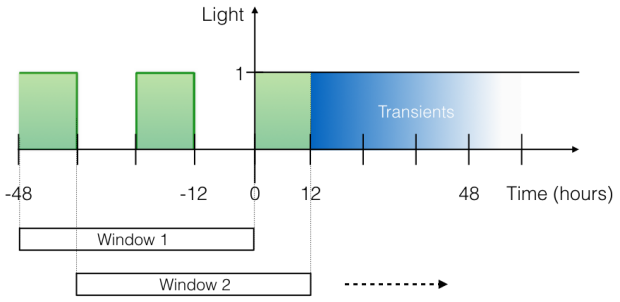

Fig. 3. **Setup for data simulation.** 48 hours of L/D data followed by 108 hours of constant light input were used for network inference. During the [-48h;0] time period, the light input is alternatively set to 1 and 0 for 12 hours. Afterwards, the light input is set to 1. The first window corresponds to the [-48h;0] time period and each of the following window is shifted by 12 hours.

2.2 Network Inference Technique

A network inference technique was applied to the previously simulated time-series data in order to capture the dynamical interactions between the genes in different photoperiodic conditions. This method was proposed in (Carignano et al., 2014) and relies on linear time-invariant (LTI) models describing the input-output interactions between each pair of genes.

We consider the LTI model

$$\begin{aligned} \frac{dx}{dt}(t) &= Ax(t) + Bu(t) + Ke(t) \\ y(t) &= Cx(t) + Du(t) + e(t) \end{aligned} \tag{1}$$

where $x \in \mathbb{R}^n$ is the vector of state variables and $e(t)$ is the white noise. In this case, the estimation problem is to identify the matrix $A \in \mathbb{R}^{n\times n}$, vectors $B, K \in \mathbb{R}^{n\times 1}$, $C \in \mathbb{R}^{1\times n}$, and scalar D. In this context, first and second order models (i.e. $n = 1$ and $n = 2$) were considered to represent the underlying dynamics between each pair of genes. Higher order models do not seem to be necessary to describe most of the biological processes.
The parameters of the LTI systems are estimated using the commonly-used prediction error method (PEM), and its implementation *pem* in Matlab (Ljung et al., 1998).For every pair of genes, first and second order LTI models were computed from the simulated dataset to describe the corresponding input-output relationship. Then, each model was characterized by a performance index that represents its capability of describing the input-output relationship. To do so, we used the fitness

$$fitness = 100 * (1 - \frac{\sum_{k=1}^{N}(y_k - \hat{y}_k)^2}{\sum_{k=1}^{N}(y_k - \bar{y})^2}) \tag{2}$$

where y is the simulated data (output), $\bar{y}$ is the average value of the simulated data, and $\hat{y}$ is the estimated output. The Matlab function *compare* can be used to compute the fitness of the model. A fitness equal to 100% corresponds

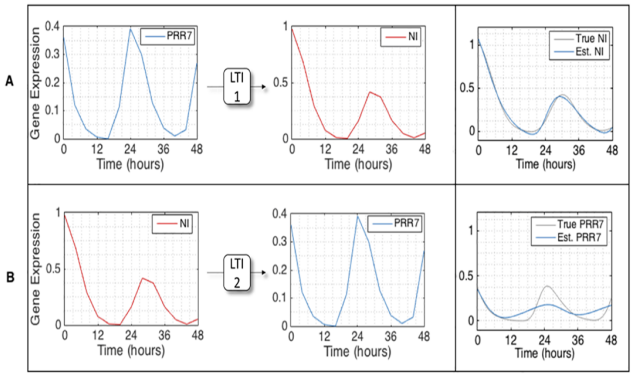

Fig. 4. **Identification of causality between one pair of genes expressions simulated with the Millar 10 model.** Transients data from the [12h;60h] window are used in this example. **(A)** and **(B)** show the identification of first order LTI systems when the input $u(t)$ is associated with PRR7 and NI, respectively. The last column in each row depicts the output $y(t)$ of the resulting LTI system (blue) and the true output (grey). The output of the estimated system in **(A)** matches NI's expression with a fitness of 90% while the other system in **(B)** fails to describe PRR7's expression accurately with an associated fitness of 32%. If the threshold of acceptance is lower than 90%, the causal relationship between PRR7 and NI will be considered as a true positive. However, NI does not regulate PRR7 and, if the acceptance threshold is below 32%, this relationships will appear as a false positive.

to a perfect identification. A high fitness suggests that most of the dynamics of the system were captured. Then, the causal relationship between these genes is validated. Hence, this technique requires to define a threshold of acceptance to select the most significant models. The identification of causality between two simulated genes expressions using LTI systems (and their respective fitness) is illustrated in Figure 4. In this example, time-series gene expression data of genes PRR7 and NI were computed through SDE simulations of the Millar 10 model. Note that these genes are both nonlinearly regulated by multiple inputs but we only consider the PRR7-NI input-output pair.

The Millar 10 and Millar 12 systems are respectively composed of 8 and 10 genes, which leads to 56 and 90 models to evaluate for each replicate, order, and time window. Although LTIs parameters estimation requires low computation time, the large number of systems to be identified makes the overall process time-consuming. Hence, 2 days were required to collect the corresponding data using the Parallel Computing Toolbox from Matlab on 12 workers. Performance of the network inference technique was evaluated by comparing the number of false positives to the number of true positives. The number of true positives corresponds to the sum of the correctly inferred causal relationships of the network while the number of false positives corresponds to the sum of non-existing links that were wrongly inferred. A causal link is identified if it has been validated either by the first order model or the second order one.

The two thresholds of acceptance $x_1, x_2 \in [0, 100]$ (based on the fitness of the 1st and 2nd order models) have been set to optimize the performance that can be obtained for each time window. Starting from 100%, thresholds were progressively decreased for each order independently. The optimal number of false positives FP^* resulting from a given number of true positive TP^* is obtained by a combination of the thresholds such that:

$$\begin{aligned} FP^* = \min_{x_1,x_2} \quad & FP(x_1, x_2) \\ \text{subject to} \quad & TP(x_1, x_2) = TP^* \end{aligned}$$

In other words, we considered the thresholds x_1, x_2 such that FP^* is minimized through the multiple combination of thresholds that return the same TP^*. Then, we computed a Receiver Operating Characteristic (ROC) by plotting the true positive rate (number of true positives divided by number of links) against the false positive rate (number of false positives divided by number of missing links) at various threshold settings. The value of the Area Under the ROC curve (AUROC) indicates the performance of the network reconstruction.

3. RESULTS

Figures 5 and 6 display the optimal ROC curves resulting from the optimization of the 2 thresholds for each time window, for the Millar 10 and 12 models, respectively. To facilitate visualization and understanding of the results, only the median value over all of the randomly replicated SDE simulations were shown here. However, the fitness of the identified systems obtained with different replicates is within a range of 5% to 10% (around the median value), so that we observed very similar overall performance.

These figures show that the results obtained with different time windows are not equivalent. Interestingly, the best result corresponds to the [0;48h] window (upper green curve) for the two models. This is the window that includes the transients data after 12 hours (see Figure 3). Data from this time window yield significantly improved results, when compared to the other ROC curves obtained with steady states data. With window [24h;72h] for the Millar 10 model, we observe in the left part of the figure a steep curve of performance showing 30% of correctly estimated links without any false positive. This curve is outperformed by the [0;48h] window curve at a true positive rates equal to 0.55. On the other hand, free-running oscillations windows (i.e. [36h;84h] to [60h;108h]) display similar shapes but show a decrease in performance, when compared to transient data. Finally, entrained data do not seem to contain enough information to retrieve accurately the ground-truth of the Millar 10 circadian network. We also observe that the different windows lead to the identification of the same links, with a few exceptions. Finally, for the Millar 10 model, the L/D windows seem to globally provide the worst results. The performances obtained from the data under constant light condition seem to lie in between the performances obtained with transient and L/D steady state conditions.

Results obtained from the identification of Millar 12 time windows data show an even more distinct results between

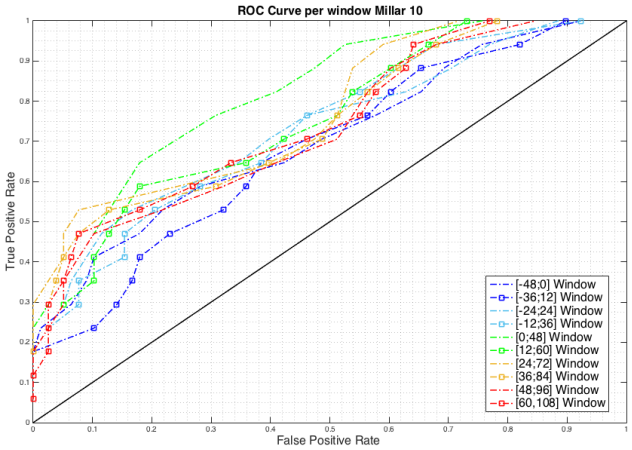

Fig. 5. **ROC Curves resulting from the network identification for each time window, for Millar 10 Model**. The dark and light blue dotted lines correspond to the performance of the system inferred from the four first windows of L/D cycles data while the red dotted curves identify the performance of the LL windows. Finally, the green and orange dotted lines stand for the transient data starting from [0;48h] to [36h;84h]

transients and steady state data. Window [12;60] provides the second best performance while the next window [24;72], however, does not yield better result than L/D data. In this case, data related to the entrained system seem to be more informative than data related to the free-running system. Furthermore, [48;96] window fails to correctly identify a large amount of the components of the network and the related ROC curve goes under the random identification curve.

Overall, a common trend appears between the 2 models: the best results are obtained when transients data are included in the window. This suggests that transient dynamics contain more information on the underlying regulation network. However, it is unclear why the window that includes the transients after 12 hours produce better results than the [12h;60h] time window in both models, which supposedly includes transient data only. Moreover, it seems that the best ratio of true positives over false positives is almost always obtained when a true light cycle is initially included in the identification process.

These results have been summarized in Table 1, which displays the AUROC value obtained with each window.

Table 1. AUROC values (IQR)

Window (hours)	AUROC Millar 10	AUROC Millar 12
-48 to 0	0.70[0.65 - 0.74]	0.64 [0.61 - 0.68]
-36 to 12	0.68 [0.66 - 0.70]	0.62 [0.60 - 0.64]
-24 to 24	0.73 [0.70 - 0.76]	0.62 [0.58 - 0.64]
-12 to 36	0.73 [0.71 - 0.75]	0.61 [0.59 - 0.63]
0 to 48	0.82 [0.78 - 0.84]	0.73 [0.70 - 0.75]
12 to 60	0.75 [0.74 - 0.76]	0.66 [0.63 - 0.68]
24 to 72	0.77 [0.75 - 0.79]	0.62 [0.58 - 0.64]
36 to 84	0.74 [0.73 - 0.77]	0.60 [0.57 - 0.62]
48 to 96	0.73 [0.70 - 0.75]	0.55 [0.52 - 0.58]
60 to 108	0.74 [0.72 - 0.77]	0.59 [0.55 - 0.63]

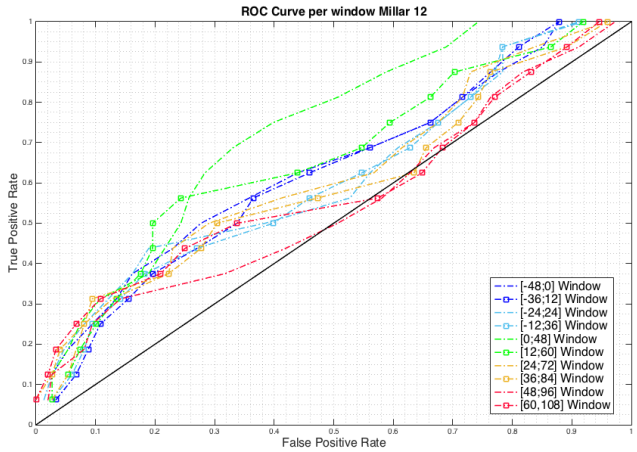

Fig. 6. **ROC Curves resulting from the network identification for each time window, for Millar 12 Model**

Table 1 shows more accurately the discrepancies between the different ROC curves for the two models. As it has been observed previously, the best results are obtained with the transient [0;48h] data. This observation is consistent between the two models. Furthermore, the time-series provided by the two first windows yield AUROC values below 0.70 for the Millar 10 model, which is relatively low. In comparison, the performances obtained with the two last windows are slightly better.

This trend is reversed for the Millar 12 model, where the AUROC value computed with the free-running data is below 0.60, which almost corresponds to random identification. We note also that the window [24h;72h] provides similar results than the entrained data, which is surprising.

4. CONCLUSION

The identification of the network based on windowed simulations of realistic data from Millar 10 and Millar 12 models allowed us to give a statistical comparison of the informative potential of the transient data. Millar 12 is a more recent circadian clock mathematical model, whose mechanism is different and more complex than Millar 10. Although the performances obtained with the two models are different, common characteristics can be retrieved from the network identification tool. These characteristics suggest that the dynamics involved in the transitory regime provide further information for system identification, in comparison to the other regimes.

Further experiments are required to understand in details the discrepancies that affect the performance obtained from different time-windows and models. In addition, the use of a combination of different windows to improve the performance of the overall network identification should also be further investigated in details. Moreover, future work will test other network inference methods to confirm that the results are not biased towards the method used in this paper. Nevertheless, this paper shows that using the transitory range of data for system identification in circadian modelling has the potential to improve the precision of the estimated network.

REFERENCES

Aderhold, A., Husmeier, D., and Grzegorczyk (2014). Statistical inference of regulatory networks for circadian regulation. *Statistical Applications in Genetics and Molecular Biology*, 13, 227–273.

Carignano, A. (2014). Genome wide analysis of differentially expressed systems: an application to circadian networks. *PhD Thesis*.

Carignano, A., Yuan, Y., Dalchau, N., Webb, A.A.R., and Gonalves, J. (2014). Understanding and predicting biological networks using linear system identification. *A Systems Theoretic Approach to Systems and Synthetic 227 Biology I: Models and System Characterizations*.

Dalchau, N., Hubbard, K.E., Robertson, F.C., Carlos T Hotta, H.M.B., Stan, G.B., Goncalves, J.M., and Webb., A.A. (2010). Correct biological timing in arabidopsis requires multiple light-signaling pathways. *Proceedings of the National Academy of Sciences of the United States of America*, 107(29), 13171?6.

Dunlap, J., L.J., and P.J., D. (2003). Chronobiology: Biological timekeeping. *Entomological Society of America*.

Guerriero, M.L., Pokhilko, A., Fernandez, A.P., Halliday, K.J., Millar, A.J., and Hillston, J. (2012). Stochastic properties of the plant circadian clock. *J.R. Soc. Interface*, 9, 744–756.

Herrero, E., Kolmos, E., Bujdoso, N., Yuan, Y., Wang, M., Berns, M.C., Uhlworm, H., Coupland, G., Saini, R., and Jaskolski, M. (2012). Early flowering4 recruitment of early flowering3 in the nucleus sustains the arabidopsis circadian clock. *The Plant Cell Online*, 428?443.

Ljung, L., Prochzka, A., Uhlil, J., Rayner, P., and Kingsbury, N. (1998). System identification. in signal analysis and prediction. *Birkhuser Boston*, 163–173.

Michael, T.P., Mockler, T.C., Breton, G., McEntee, C., Byer, A., Trout, J.D., Hazen, S.P., Shen, R., Priest, H.D., and Sullivan, C.M. (2008). Network discovery pipeline elucidates conserved time-of-day?specific cis-regulatory modules. *PLoS genetics*, 4.

Pokhilko, A., Fernandez, A.P., Edwards, K.D., Southern, M.M., Halliday, K.J., and Millar, A.J. (2012). The clock gene circuit in arabidopsis includes a repressilator with additional feedback loops. *Molecular systems biology*, 8.

Pokhilko, A., Hodge, S.K., Stratford, K., Know, K., Edwards, K.D., Thomson, A.W., Mizuno, T., and Millar, A.J. (2010). Data assimilation constrains new connections and components in a complex, eukaryotic circadian clock model. *Molecular systems biology*, 6.

Zhang, E.E. and Kay, S.A. (2010). Clocks not winding down: unravelling circadian networks. *Nature Reviews Molecular Cell Biology*, 11, 764–776.