



**HAL**  
open science

## Principal process analysis of biological models

Stefano Casagrande, Suzanne Touzeau, Delphine Ropers, Jean-Luc Gouzé

► **To cite this version:**

Stefano Casagrande, Suzanne Touzeau, Delphine Ropers, Jean-Luc Gouzé. Principal process analysis of biological models. *BMC Systems Biology*, 2018, 12, pp.68. 10.1186/s12918-018-0586-6. hal-01818033

**HAL Id: hal-01818033**

**<https://inria.hal.science/hal-01818033>**

Submitted on 26 May 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.




Distributed under a Creative Commons Attribution 4.0 International License

METHODOLOGY ARTICLE

Open Access



# Principal process analysis of biological models

Stefano Casagrande<sup>1\*</sup> , Suzanne Touzeau<sup>1,3</sup>, Delphine Ropers<sup>2</sup> and Jean-Luc Gouzé<sup>1</sup>

## Abstract

**Background:** Understanding the dynamical behaviour of biological systems is challenged by their large number of components and interactions. While efforts have been made in this direction to reduce model complexity, they often prove insufficient to grasp which and when model processes play a crucial role. Answering these questions is fundamental to unravel the functioning of living organisms.

**Results:** We design a method for dealing with model complexity, based on the analysis of dynamical models by means of Principal Process Analysis. We apply the method to a well-known model of circadian rhythms in mammals. The knowledge of the system trajectories allows us to decompose the system dynamics into processes that are *active* or *inactive* with respect to a certain threshold value. Process *activities* are graphically represented by *Boolean* and *Dynamical Process Maps*. We detect model processes that are *always inactive*, or *inactive* on some time interval. Eliminating these processes reduces the complex dynamics of the original model to the much simpler dynamics of the core processes, in a succession of sub-models that are easier to analyse. We quantify by means of global relative errors the extent to which the simplified models reproduce the main features of the original system dynamics and apply global sensitivity analysis to test the influence of model parameters on the errors.

**Conclusion:** The results obtained prove the robustness of the method. The analysis of the sub-model dynamics allows us to identify the source of circadian oscillations. We find that the negative feedback loop involving proteins PER, CRY, CLOCK-BMAL1 is the main oscillator, in agreement with previous modelling and experimental studies. In conclusion, Principal Process Analysis is a simple-to-use method, which constitutes an additional and useful tool for analysing the complex dynamical behaviour of biological systems.

**Keywords:** Dynamical systems, Biological networks, Process analysis, Model reduction, Parameter sensitivity analysis, Circadian clock

## Background

Mathematical modelling has been used for decades as an approach to understand the functioning of biological systems in terms of their internal processes and components. The latter form complex networks that vary in nature. For instance, biochemical networks include processes controlling the intracellular level of metabolites, RNAs and proteins, which allow cells to live and grow. A process either corresponds to a single biochemical reaction, for example protein phosphorylation, or encompasses many biochemical reactions like those involved

in general cell functions (translation of proteins, transcription of RNAs...). In ecological networks, the processes can refer to events influencing the distribution and abundance of organisms, or to fluxes of energy and matter.

Numerous kinetic models of these networks have been developed in computational biology, of increasing complexity due to advances in modelling and parameter estimation approaches (see [1, 2] for an example). Complexity arises from the high dimension of the networks, the large number of biological processes involved and their non linearity due to the complex feedback loops that regulate them.

One approach often used to tackle the problem of complexity is model reduction (see [3] for a recent review).

\*Correspondence: [stefano.casagrande01@gmail.com](mailto:stefano.casagrande01@gmail.com)

<sup>1</sup> Université Côte d'Azur, Inria, INRA, CNRS, UPMC Univ Paris 06, Biocore team, Sophia Antipolis, France

Full list of author information is available at the end of the article



The simplified models are easier to analyse, while retaining the main features of the original ones and their biological significance. Briefly, methods of model reduction shorten the list of network species or of network reactions (e.g. [4, 5]), lump state variables (e.g. [6]) or decompose the system into slow and fast dynamics (e.g. [7–9]). The often used quasi-steady-state approximation falls in the latter category (e.g. [10]). Other approaches simplify the mathematical functions describing the molecular processes. For instance, piece-wise affine differential equations approximate by step functions the sigmoidal functions used to describe the regulation of gene expression. The dynamics of the simplified system can be easily analysed by means of state transition graphs [11]. However, these simplifications are generally restricted to models of gene expression and are more difficult to apply to other types of networks [12].

Reduction approaches have proven successful to significantly reduce model complexity, but they do not provide a mean to understand how the system dynamics emerges from the cascade of biological processes and regulatory mechanisms at work. This is especially true when the reduced models remain complex, with many coupled equations sharing common processes and involving complex feedback loops. For instance, regulatory mechanisms switch on certain biological processes at some times and off at others. It is thus important for a good understanding of the system behaviour to identify which and when processes significantly influence the system dynamics. In other words, instead of analysing a single reduced model in place of the original one, valid on the whole time interval, we may want to analyse series of simplified models highlighting the important processes of the original model during the periods of time in which they are *active*.

This is how we address the problem of high dimensional model analysis in this study. We develop a mathematical and numerical approach based on the boolean concept of *activity/inactivity*. The method, called *Principal Process Analysis* (PPA), determines the contribution of each biological process to the output of the dynamical system. In models of biological networks, these processes appear in a linear additive manner in each ODE. We first identify the *inactive* processes and neglect them. In a second step, we treat processes whose *activity* varies along time: we define time windows in which these processes are either always *active* or always *inactive*. We eventually create sub-models for each time window that only contain the *active* processes. This procedure leads to the simplification of the system to its core mechanisms. The simplified system can be further studied, to understand the role of each *active* process in the system dynamics.

PPA is a general approach that can be easily applied to any biological system described by ordinary differential equations (ODEs). It shares common features with a model reduction method focusing on major model parameters rather than processes [4], in which parameters that are not required for the system behaviour are removed. Another approach dedicated to chemical reactions identifies and removes chemical species that contribute less to the model output [5]. In this case, the problem is solved using optimization approaches (see also [13]). Despite these similarities, PPA is not a model reduction approach. It provides a mean to access to and dissect the more complex dynamics of the original model through the analysis of simplified versions in given time windows. Results are easily interpretable and do not require additional and complicated computations.

Preliminary work on PPA has been described in an earlier conference paper [14], in which we applied PPA to two ODE models of biochemical networks whose simplification preserved their dynamical behaviour: the model of circadian rhythms in *Drosophila* [15] and the model of the regulation of the ERK signalling pathway [16]. Questions remained open though, concerning the scalability of the approach and its robustness: to which extent does PPA preserve model dynamics in systems of higher dimension, with many more biological processes involved and including interlocked feedback loops? And since the approach requires a priori knowledge of the parameter values, how sensitive are process *activities or inactivities* to the value of these parameters? In this study, we address these questions by studying a much more complex model of circadian rhythms in mammals, including 16 variables, 76 processes, and intertwined positive and negative feedback loops [17]. Parameter sensitivity analysis of the global relative error between the original and reduced systems allows us to assess the quality and robustness of our approach.

The paper is organized as follows. “**Methods**” section describes the principle of Principal Process Analysis as well as global sensitivity analysis. “**Model description**” section introduces the model of mammalian circadian clock. We apply our approach to this complex model in “**Principal Process Analysis of the circadian clock model**” to “**Influence of parameter values**” sections, before concluding in “**Conclusions**” section.

## Methods

We summarize below the basics of the method of Principal Process Analysis. We will use as running example the 14<sup>th</sup> variable of the mammalian circadian clock model analysed in “**Model description**” section (see also Appendix B. It describes how the concentration of the nuclear form of protein BMAL1 ( $B_N = x_{14}$ ) changes:

$$\frac{dB_N}{dt} = -V_{3B} \frac{B_N}{K_p + B_N} + V_{4B} \frac{B_{NP}}{K_{dp} + B_{NP}} + k_5 B_C - k_6 B_N - k_7 B_N PC_N + k_8 I_N - k_{dn} B_N. \tag{1}$$

**Principal Process Analysis (PPA)**

Consider the following ODE model of biological network:

$$\dot{x} = f(x, p) \tag{2}$$

where  $x = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$  is the vector of component concentrations,  $x_0 = (x_{01}, x_{02}, \dots, x_{0n}) \in \mathbb{R}^n$  the vector of their initial values and  $p \in \mathbb{R}^b$  the vector of parameters. Each equation is decomposed into a sum of biological processes:

$$\dot{x}_i = \sum_j f_{ij}(x, p) \tag{3}$$

where  $f_{ij}$  represents the  $j^{th}$  process involved in the dynamical evolution of the  $i^{th}$  variable of the system over a period of time  $[t_0, T]$ .

**Example 1** Equation (1) includes seven processes, each associated with a specific biological function. They take a positive or negative value, depending on whether they affect positively or negatively the variation of BMAL1 concentration. The equation of the protein is rewritten as:

$$\dot{x}_{14} = f_{14,1} + f_{14,2} + f_{14,3} + f_{14,4} + f_{14,5} + f_{14,6} + f_{14,7} \tag{4}$$

where  $f_{14,1} = -V_{3B} \frac{B_N}{K_p + B_N}, \dots, f_{14,7} = -k_{dn} B_N$ .

Figure 1a shows the dynamical evolution of processes  $f_{14,1}$  to  $f_{14,7}$  during a day. Nuclear import of BMAL1 is the fastest process of Eq. (1), while the basal degradation of the protein is the slowest.

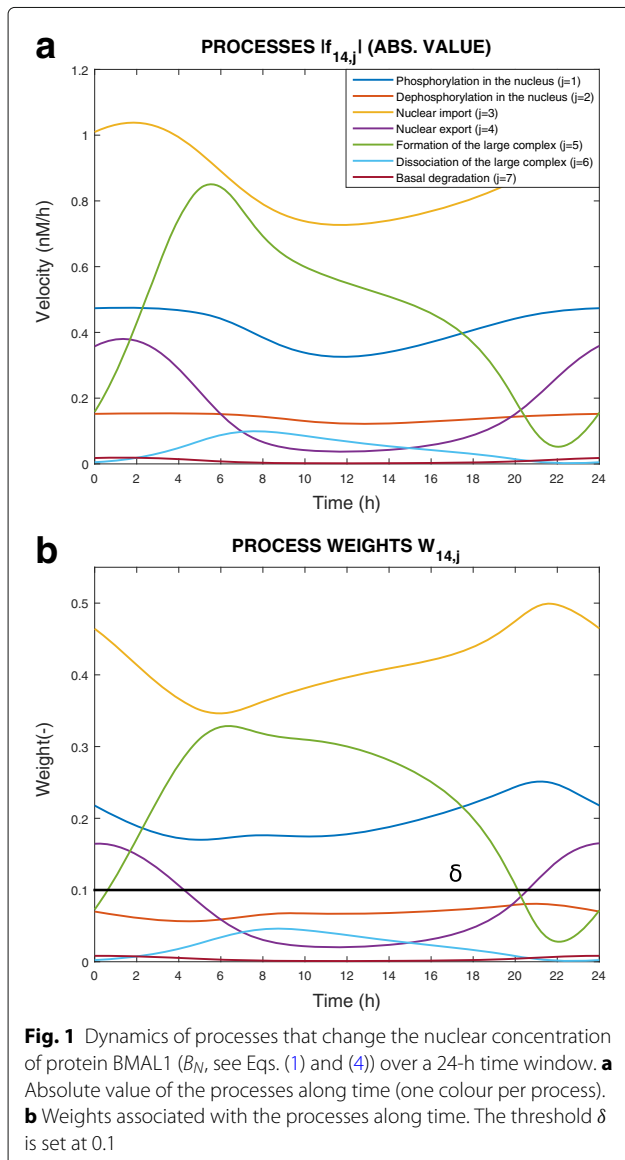
Comparison criteria are needed to weigh the influence of the different processes  $f_{ij}$  on the time evolution of each variable  $x_i$ . There are several alternatives. For instance, we can compare their absolute value  $(|f_{ij}(x, p)|)$ , scale it by the  $i^{th}$  initial condition  $(\frac{|f_{ij}(x(t), p)|}{x_{0i}})$ , or scale it by the solution of the  $i^{th}$  ODE  $(\frac{|f_{ij}(x(t), p)|}{x(t)_i})$ . In this work we associate a relative weight with each process to make it dimensionless:

$$W_{ij}(t, p) = \frac{|f_{ij}(x(t), p)|}{\sum_j |f_{ij}(x(t), p)|} \tag{5}$$

where  $0 \leq W_{ij}(t, p) \leq 1$  and  $\sum_j W_{ij}(t, p) = 1$ .

**Definition 1** Let the continuous function  $f_{ij}(x(t), p)$  be the  $j^{th}$  process of  $\dot{x}_i(t)$  in  $t \in [t_0, T]$  and let the threshold  $\delta \in [0, 1]$ . We call a process  $f_{ij}(x(t), p)$  always inactive when  $W_{ij}(t, p) < \delta \forall t \in [0, T]$ . We call a process  $f_{ij}(x(t), p)$  inactive at time  $t$  when  $W_{ij}(t, p) < \delta$ . We call a process  $f_{ij}(x(t), p)$  active at time  $t$  when  $W_{ij}(t, p) \geq \delta$ . Switching time for a process  $f_{ij}(x(t), p)$  is the time  $t_{ij}^s$  at which  $W_{ij}(t, p) = \delta$ . A process can have  $0, 1, \dots, z$  switching times. The switching time set  $S_i$  for the  $i^{th}$  variable contains all the switching times  $t_{ij}^s$  where  $j = 1, \dots, k$  and  $s = 1, \dots, z$ . The global switching time set  $S$  is the union of all  $S_i$ .

The choice of  $\delta$  is important, since it determines above which weight a process can be considered active or inactive and, as we will see it later, if the process should be kept or omitted in the simplified model. An excessively high value might lead to an oversimplified model, without many dynamical features of the original model.



Conversely a very low value might result in a model insufficiently simplified, which remains too complicated to analyse. From our experience, a convenient value is  $\delta \in [0, 0.1]$ , where the value of  $\delta$  can be adjusted to the number of processes. For instance, if an ODE contains numerous processes of similar value, each individual process weighs little. In this case,  $\delta$  should not be chosen too high to avoid omitting all these processes; it can be inversely proportional to the total number  $N$  of processes in the equation:  $\delta \propto \frac{1}{N}$ . In this paper, fine-tuning the threshold value is not justified: there are not many processes per equation and they have very different values. We will always take  $\delta = 0.1$ .

**Example 2** We apply Eq. (5) to determine the dynamical weight of the seven processes in Eq. (1). Results are shown in Fig. 1b. As expected, the nuclear import, which is the fastest process, weighs more in the dynamical evolution of BMAL1 concentration, while the basal degradation of the protein weighs little. We determine the process activities using  $\delta = 0.1$ :

- The weight of processes  $W_{14,2}$ ,  $W_{14,6}$ ,  $W_{14,7}$  is always below  $\delta$ : their related processes  $f_{14,2}$ ,  $f_{14,6}$ ,  $f_{14,7}$  are thus always inactive;
- The processes  $W_{14,1}$  and  $W_{14,3}$  are always above  $\delta$ :  $f_{14,1}$  and  $f_{14,3}$  are active during the whole system dynamics;
- The weight of processes  $W_{14,4}$  and  $W_{14,5}$  crosses the threshold twice and the switching times  $t_{14,4}^1 = 4.4h$ ,  $t_{14,4}^2 = 20.7h$ ,  $t_{14,5}^1 = 0.8h$  and  $t_{14,5}^2 = 20.3h$  are collected in the set  $S_{14}$ .

### Visualization of process activities

Graphical tools turn out to be useful to analyse the dynamical weights of complex systems such as the mammalian circadian clock model. We use three of them in PPA, which are described below.

- The Boolean Process Map summarizes qualitatively the knowledge of the process activity or inactivity along time for each variable. A black bar means that the process is active, while the white bar indicates an inactive process.

**Example:** The Boolean Process Map in Fig. 2a represents the process activities deduced from the dynamical weights in Fig. 1b. We observe that there is always an active phosphorylation of BMAL1 in the nucleus, while the basal degradation can be considered always inactive. The nuclear export is solely active in the first and last periods of time.

- The Dynamical Process Map is a network representation of the process activities. Variables (represented by boxes) are connected by processes

(arrows). Three cases arise, which depend on the activity of processes shared by several variables: black-coloured arrows represent processes that are inactive for all variables involved, while active processes are displayed as red arrows. Yellow arrows are used for processes shared by several variables that have different activities: for instance, one process is considered active in one equation, but inactive in another one. Note that the model simplification by elimination of inactive processes, as will be described in “Model simplification by elimination of always inactive processes” section, will have for effect to remove black arrows in the Dynamical Process Map.

**Example 3** Figure 2b represents the Dynamical Process Map for  $x_{14}$ , the nuclear concentration of BMAL1, in the time interval between  $t_{14,4}^1$  and  $t_{14,5}^1$ . Phosphorylation is an example of active process for the nuclear BMAL1 concentration (see the Boolean Process Map in Panel A). It is shown in red because it is also considered active at the same moment for the other variable sharing this process, the concentration of phosphorylated BMAL1.

- The 3-D Process Map represents the time-dependent evolution of the intensity of each process. Process activities are averaged per hour, which leads to a discretisation of time. Vertical bars represent process weights for each hour. Their color code represents the intensity of process weights relatively to the other weights.

**Example 4** Figure 2c describes the 3-D Process Map for the concentration of nuclear BMAL1. The phosphorylation of the protein, its nuclear import and its consumption for the formation of a large complex are the processes the most active over time.

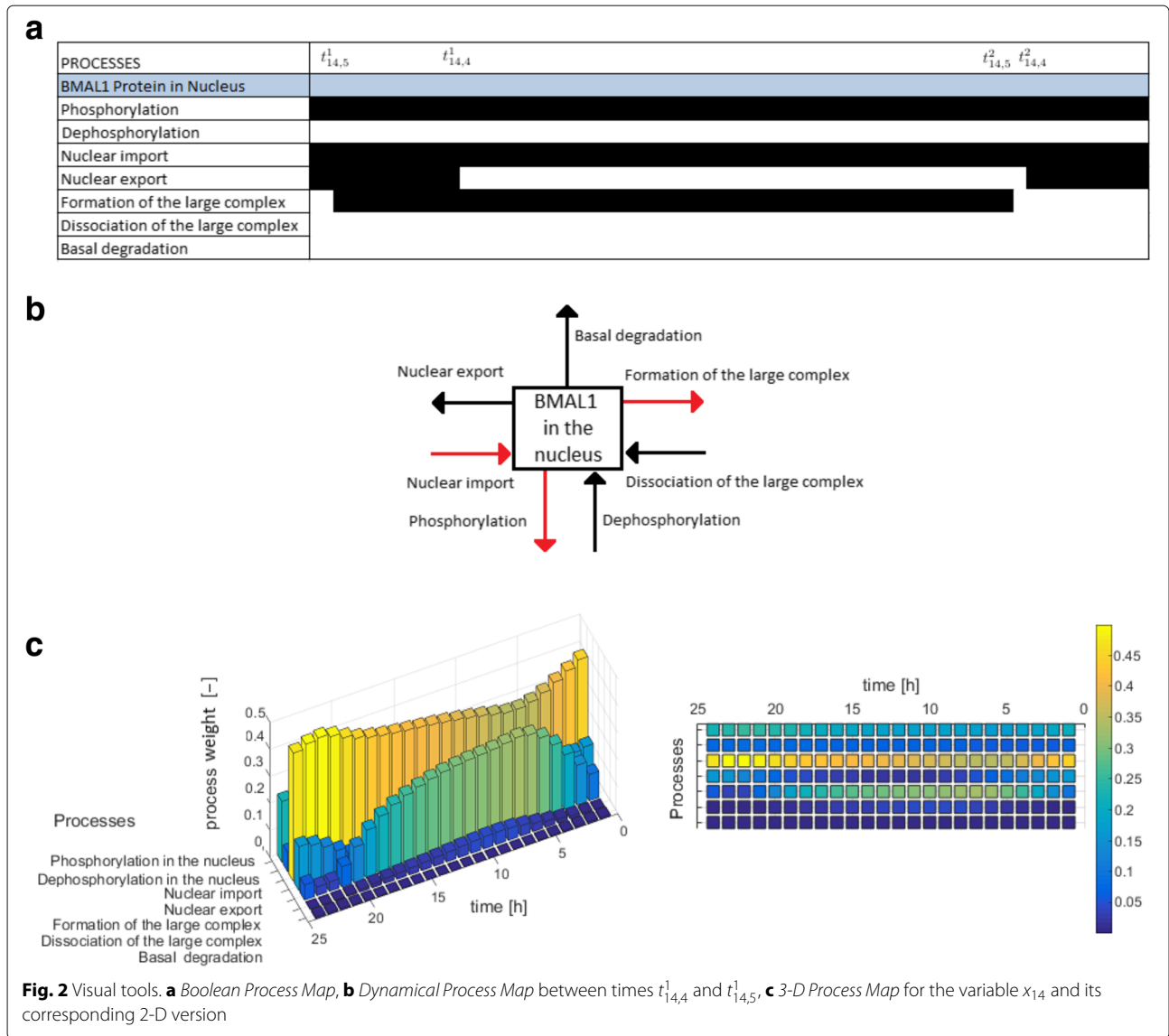
### Model simplification by elimination of always inactive processes

Eliminating processes that play a minor role in the system dynamics facilitates the analysis of large models. Since in the previous steps of PPA we have determined the process activities in system (2), we now neglect processes that are considered always inactive. This will give us  $g(x^r)$ , the function approximating  $f(x)$  in (2) with less processes.

We thus introduce the ODE system (6), which approximates system (2):

$$\dot{x}^r = g(x^r, p^r) \quad (6)$$

where  $x^r = (x_1^r, x_2^r, \dots, x_n^r) \in \mathbb{R}^n$  is the vector of component concentrations,  $x_0 = (x_{01}, x_{02}, \dots, x_{0n}) \in \mathbb{R}^n$



the vector of their initial values, and  $p^r \in \mathbb{R}^c$ , where  $c \leq b$  is the vector of parameters. The model simplification approach relies basically on the following theorem: if the vector fields of two systems are close ( $f(x) \approx g(x)$ ), then the solutions of the original and approximated systems are close during some time interval under the assumptions on the Lipschitz conditions listed in [18, p. 79, Th. 2.5].

Based on the dynamical weights determined in “Principal Process Analysis (PPA)” section and the threshold value  $\delta$ , we apply the following rule to define  $g(x^r, p^r)$ : if  $W_{ij}(x(t), p) < \delta \forall t \in [t_0, T]$  then  $g_{ij} = 0$ ; if not,  $g_{ij} \equiv f_{ij}$ .

We thus define  $x^r$  as an approximation of  $x$  and  $p^r$  as a subset of  $p$ .

**Example 5** We proceed to the simplification of processes in Eq. 1. Because  $f_{14,2}, f_{14,6}, f_{14,7}$  are always inactive,  $g_{14,2} = 0, g_{14,6} = 0, g_{14,7} = 0$  and  $g_{14,1} \equiv f_{14,1}, g_{14,3} \equiv f_{14,3}, g_{14,4} \equiv f_{14,4}, g_{14,5} \equiv f_{14,5}$ . The resulting ODE for  $x_{14}^r$  is:

$$\frac{dB_N^r}{dt} = -V_{3B} \frac{B_N^r}{K_p + B_N^r} + k_5 B_C^r - k_6 B_N^r - k_7 B_N^r PC_N^r. \quad (7)$$

Note that Principal Process Analysis is applied to each ODE separately. As a consequence, processes shared by two equations can be *active* in one equation, but *inactive* in the other. Elimination of the *inactive* processes breaks mass balance in the simplified model. For our purpose, this is not an issue: the simplification does not aim at reducing the model, but rather analysing a sub-model

of the original one, which describes the dynamics of the important phenomena.

It is interesting to quantify the extent to which the simplified system (6) preserves the behaviour of the original one. This gives a better sense of how the *active* processes kept in the simplified model are responsible for the dynamics of the original system. In addition, this helps identifying potential problems related to the model simplification, for instance involving a wrong choice of the  $\delta$  value. One can also imagine pathological cases, when the simplified system does not reproduce the main dynamical features of the original model: for instance, if it evolves towards a different basin of attraction or if the removal of a consumption term does not compensate a synthesis term any more, leading the simplified system to explode in finite time. It is non nonsensical in all these cases to analyse simplified models that behave so differently from the original ones. The  $\delta$  threshold should be adjusted to a new value and Principle Process Analysis re-run until model simplification proves satisfactory according to the criteria described below.

We present in Appendix A an a priori analysis of the error made when removing some *inactive* processes. This analysis gives a theoretical, but very conservative, bound on the error. In practice, we numerically compute the global relative error between the original and simplified models. Several forms of error are possible. We have chosen the following one, analysed over a period of time  $[t_0, T]$ , in which  $y_h$  and  $y_h^r$  are the  $h^{th}$  outputs of the original and simplified systems, respectively:

$$e_h = \frac{\int_{t_0}^T |y_h(t) - y_h^r(t)| dt}{\int_{t_0}^T |y_h(t)| dt} \tag{8}$$

How to choose the model outputs? They can correspond to all model variables or combinations of them, if the latter are involved in some biological phenomena of interest for instance. In the case of the circadian clock model, six variables were specifically studied in the original papers [17, 19], which we will use as outputs to determine the global relative error between the original and simplified models: the concentrations of *Per* mRNA ( $M_P$ ), *Cry* mRNA ( $M_C$ ), *Bmal1* mRNA ( $M_B$ ), total PER protein ( $P_{Tot}$ ), total CRY protein ( $C_{Tot}$ ) and total BMAL1 protein ( $B_{Tot}$ )<sup>1</sup>.

### Creation of sequences of sub-models

In the previous step of PPA, the models are simplified by elimination of *always inactive* processes. Here we go one step further in the simplification, by eliminating processes that are *inactive* at times. This is achieved by decomposing the period of time during which the system evolves into time intervals. To that end we use the

*switching times*  $t_b$  (with  $b = 1, \dots, d$ ) determined in “Principal Process Analysis (PPA)” section: this allows creating a succession of sub-models for each time interval, which contain the core mechanisms in that period of time.

To avoid creating large sequences of sub-models, we reduce the number of time windows by grouping proximal *switching times* with the easy-to-compute k-means clustering [20]. Hence the  $d$  *switching times* included in the *global switching time set*  $S = [t_1, t_2, \dots, t_d]$  are grouped into  $z$  ( $\leq d$ ) clusters  $C = \{C_1, C_2, \dots, C_z\}$ , so as to minimize the within-cluster sum of square (or within-cluster inertia):

$$\operatorname{argmin}_C \sum_{v=1}^z \sum_{t \in C_v} ||t - \mu_v||^2 \tag{9}$$

where  $\mu_v$  is the mean of the *switching times* in  $C_v$ . The consequence is that processes with *switching times* belonging to cluster  $C_v$  are assumed to switch together at the same time  $t_v^r = \mu_v$ , the mean *switching time* in cluster  $C_v$ .

How to define the right number of clusters? A too large number of clusters will result in a low error, but also in numerous time windows that make the simplified models still too complex to analyse. Equation (10) describes how to take into account this trade-off between the number  $z$  of clusters and the error. It is related to the difference between the maximum and the minimum number of *active* processes during the temporal evolution of the system: if this difference is low,  $z$  should be chosen low as well. We thus define  $z$ , rounded to the nearest number, as:

$$z = \frac{\max_v(n_{act}^v) - \min_v(n_{act}^v)}{2}, \tag{10}$$

where  $n_{act}^v$  denotes the number of *active* processes in the  $v^{th}$  time window.

We eventually end up with a sequence of  $z + 1$  sub-models in the time interval  $[0, T]$ , the first one being valid in  $[0, t_1^r]$  and the last one, in  $[t_z^r, T]$ .

Similarly to the global errors determined in “Model simplification by elimination of *always inactive processes*” section, we can also assess how the newly simplified models reproduce the dynamical behaviour of the original model in each time window  $[t_{v-1}^r, t_v^r]$ , by measuring the error:

$$e_h^v = \frac{\int_{t_{v-1}^r}^{t_v^r} |y_h(t) - y_h^r(t)| dt}{\int_{t_{v-1}^r}^{t_v^r} |y_h(t)| dt} \tag{11}$$

We compute the error (11) between the original model and each sub-model, with or without propagating errors: in the first case, for each time window  $[t_{v-1}^r, t_v^r]$

( $v = 1, \dots, z + 1$  with  $t_0^v = t_0$  and  $t_{z+1}^v = T$ ), the initial values of the  $h$  outputs of sub-model  $SM_v$  are equal to the final values at  $t_{v-1}^v$  of sub-model  $SM_{v-1}$ ; in the second case, they are equal to the values of the original model at  $t_{v-1}^v$ .

**Global sensitivity analysis**

Principal Process Analysis is applied to models with given parameter values and initial conditions. It may be questioned whether the uncertainty of their values influences the simplification of the model and thus, the analysis of the system dynamics. While we have shown PPA to be robust to variations of initial conditions in [21], the question remains open for parameter values.

To that aim, we perform global sensitivity analyses on the global relative errors between the original model and the reduced model (defined in Eq. (11)). Such an analysis consists in quantifying the parameter influence on the error, while varying the parameters simultaneously in given ranges. In contrast, in a local sensitivity analysis, parameters would vary one-at-a-time in the neighbourhood of their nominal value. First, we perform an analysis on each of the six errors defined for the six model outputs ( $e_{M_p}^v, e_{M_c}^v, e_{M_b}^v, e_{P_{Tot}}^v, e_{C_{Tot}}^v, e_{B_{Tot}}^v$ ). Then, in a more detailed analysis, we compute the global relative error for each state variable, according to Eq.(11) (with  $y_h = x_i, i = 1, \dots, 16$ ); sensitivity analyses are also performed on each of these 16 errors. The method used is based on factorial design [22], analysis of variance (ANOVA) and principal component analysis (PCA) [23].

We first explore the parameter space using a factorial design. We vary  $N_f = 51$  parameters of the model [17] (see “Model description” section). We choose  $N_l = 2$  levels for each parameter  $p_f$  (or factor):  $p_f^- = 0.8 p_f$  and  $p_f^+ = 1.2 p_f$ . A full factorial design, defined as all possible combinations of the parameter levels, would be necessary to estimate the main effects and interactions of all parameters. Such a full design corresponds to  $N_l^{N_f} = 2^{51}$  parameter combinations and would necessitate the same number of model simulations to compute the corresponding outputs, which are far too many. Thus we implement a fractional factorial design [24], which is a subset (fraction) of the full design of size  $N_j < N_l^{N_f}$ . The design is determined according to a given statistical model linking the error  $e_h$  to the parameters  $p_f$ , for each time window  $[t_{v-1}^v, t_v^v]$ . We choose a second order linear model, which incorporates all main effects and two-way interactions as follows:

$$e_{h,j}^v = \mu_h^v + \sum_{f=1}^{N_f} \alpha_{h,f(j)}^v + \sum_{f=1}^{N_f} \sum_{k=1, k \neq f}^{N_f} \beta_{h,f(j)k(j)}^v + \epsilon_{h,j}^v \quad (12)$$

where  $e_{h,j}^v$  is the error computed according to Eq. (8) for output (or state variable)  $h$ , time window  $v$ , and parameter combination  $j$  ( $j = 1, \dots, N_j$ ) of the fractional factorial design;  $\mu_h^v$  is the grand mean;  $\alpha_{h,f(j)}^v$  is the main effect of parameter  $p_f$  for parameter combination  $j$ ;  $\beta_{h,f(j)k(j)}^v$  is the interaction effect between parameters  $p_f$  and  $p_k$  ( $k \neq f$ ) for parameter combination  $j$ ; and  $\epsilon_{h,j}^v$  is the residual. Each main effect  $\alpha_{h,f(j)}^v$  can take two values, according to the level of parameter  $p_f$  in combination  $j$ :  $\alpha_{h,f+}^v$  or  $\alpha_{h,f-}^v$ . Similarly, each two-way interaction effect can take four values:  $\beta_{h,f+k+}^v, \beta_{h,f+k-}^v, \beta_{h,f-k+}^v, \beta_{h,f-k-}^v$ . The fractional factorial design determines the parameter combinations needed to estimate all main effects and two-way interactions. It is obtained using R package `planor`<sup>2</sup> and consists of  $N_j = 2^{12}$  parameter combinations, yielding as many simulations. According to the sparsity-of-effects principle, a system is usually dominated by main effects and low order interactions, so neglecting third-order and higher interactions can still provide good estimates.

An ANOVA is then performed on these simulations, for each error  $e_h$ . It consists in estimating the grand mean, main effects and interaction terms of model (12), using a least-square criterion to minimise the residuals. It is based on the following variance decomposition:

$$\underbrace{\sum_{j=1}^{N_j} (e_{h,j}^v - \bar{e}_h^v)^2}_{SS_T^{h,v}} = \underbrace{\sum_{j=1}^{N_j} (\widehat{e}_{h,j}^v - \bar{e}_h^v)^2}_{SS_M^{h,v}} + \underbrace{\sum_{j=1}^{N_j} (e_{h,j}^v - \widehat{e}_{h,j}^v)^2}_{SS_r^{h,v} = \sum_j (\epsilon_{h,j}^v)^2} \quad (13)$$

where  $\bar{e}_h^v$  is the mean error computed over all  $N_j$  simulations of the fractional factorial design; and  $\widehat{e}_{h,j}^v = \widehat{\mu}_h^v + \sum_f \widehat{\alpha}_{h,f(j)}^v + \sum_f \sum_{k \neq f} \widehat{\beta}_{h,f(j)k(j)}^v$  ( $\widehat{\phantom{x}}$  denoting an estimated value) is the error estimated from the linear model (12) for parameter combination  $j$ . The total sum of squares  $SS_T^{h,v}$  is split into the sum of squares attributed to the model  $SS_M^{h,v}$  and the residual sum of squares  $SS_r^{h,v}$ , the latter corresponding to the criterion that is minimised. In turn,  $SS_M^{h,v}$  is split into sum of squares attributed to each main effect  $\alpha_{h,f}^v$  and two-way interaction term  $\beta_{h,fk}^v$ , denoted respectively  $SS_f^{h,v}$  and  $SS_{fk}^{h,v}$ . The total sensitivity index of parameter  $p_f$  is then defined as follows:

$$tSI_f^{h,v} = \frac{SS_f^{h,v} + \sum_{k \neq f} SS_{f,k}^{h,v}}{SS_T^{h,v}} \quad (14)$$

Noting that the variance of error  $e_h^v$  computed over all  $N_j$  simulations of the fractional factorial design is



$\sigma_{e_h^v}^2 = \frac{1}{N_f-1} SS_T^{h,v}$ , the total sensitivity index  $tSI_f^{h,v}$  represents the fraction of the variance explained by parameter  $p_f$ . As an ANOVA requires a scalar variable, separate sensitivity analyses are performed for each scalar error  $e_h^v$  and separate indices are computed for each error  $e_h^v$ . To compare the parameter influence on the different errors  $e_h^v$ , we use non normalised indices, obtained by multiplying each  $tSI_f^{h,v}$  by the variance of the error:

$$tSI_f^{h,v'} = \sigma_{e_h^v}^2 tSI_f^{h,v}. \tag{15}$$

To obtain synthetic sensitivity indices that represent the influence of each parameter on the errors for all 16 state variables, we decompose the error vector ( $e_h : h = 1 \dots, 16$ ) by PCA (without normalisation). As a result, an inertia proportion  $\omega_c$  can be attributed to each component  $c$  (a component is a linear combination of the 16 errors  $e_h$ ). It represents the variability among all simulations carried by this component. Only the  $N_c$  first components whose cumulated inertia add up to 95% or more are retained. Moreover, each simulation is given a *score* on each component, a scalar representing the projection of the simulation on the component. Then, for each component retained, an ANOVA is performed on the *scores* and total sensitivity indices  $tSI_f^c$  are computed, as described in Eq. (14). Finally, a total generalised sensitivity index is calculated for each parameter  $p_f$  as the sum of the total sensitivity indices on each component, weighted by the inertia of the component:

$$tGSI_f = \sum_{c=1}^{N_c} \omega_c tSI_f^c. \tag{16}$$

We use the `multisensi` R package<sup>3</sup> for this analysis.

In what follows, we show how Principal Process Analysis can help with the analysis of complex biological models. We apply the approach to a model of the circadian clock developed in [17, 19], which we describe in the following section.

## Results

### Model description

Periodic fluctuations of the environment subject living organisms to biological rhythms. The latter are endogenous by nature, but entrained by environmental variations. For instance, circadian rhythms are generated by a molecular clock within cells, which synchronizes daily physiological variations to the day-night alternance. The model we study here describes the circadian clock in mammals [17, 19]. In this model, the clock forms a

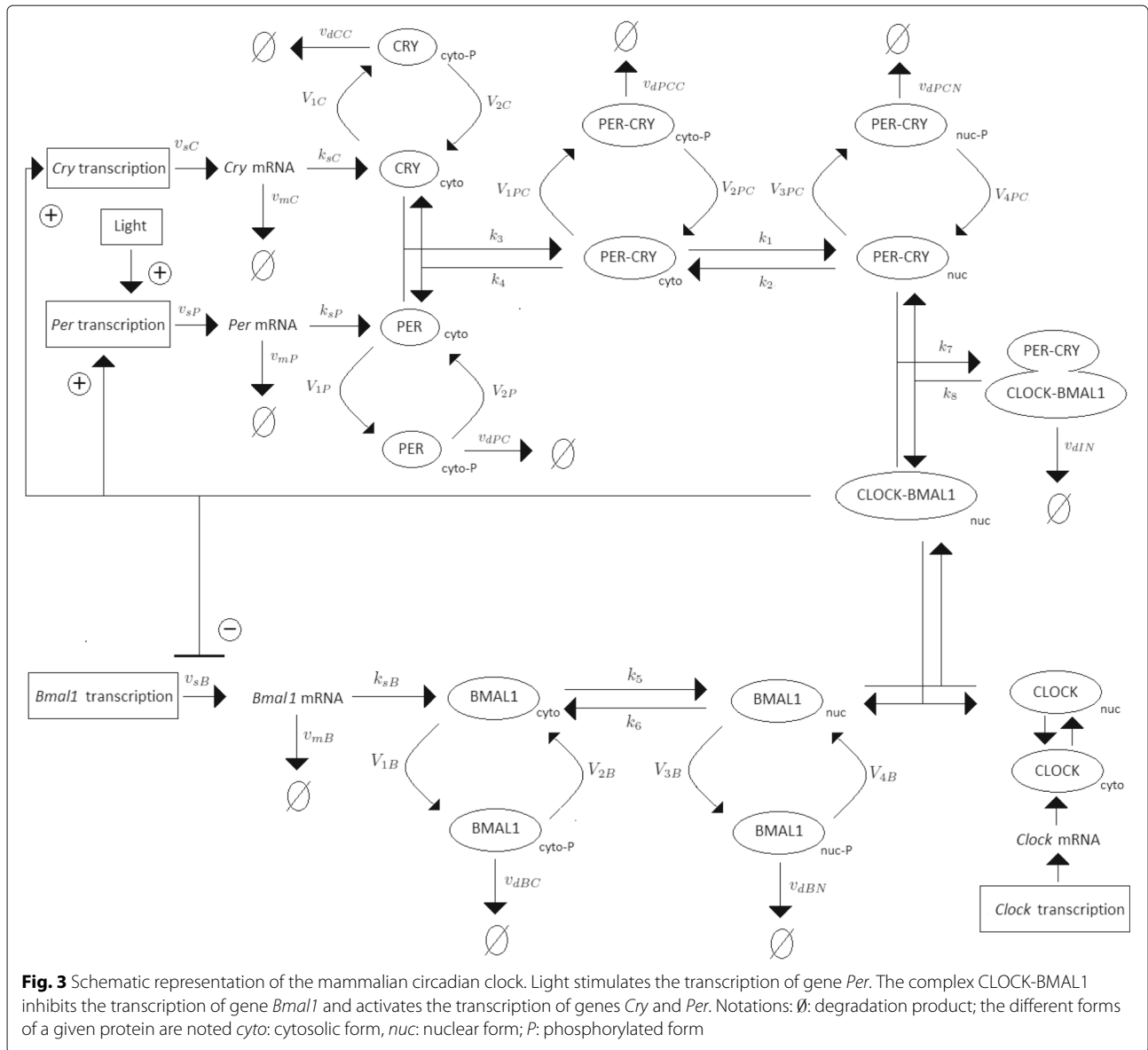
complicated network of intertwined positive and negative feedback loops involving four clock genes: *Per*, *Cry*, *Bmal1*, and *Clock*. Their mRNA and protein produce sustained oscillations with a period of 24 hours. Light affects expression of gene *Per* at the transcriptional level: the first twelve hours of day light increase its transcription rate (up to 1.8 [ $\mu\text{M}/\text{h}$ ]), while it is lowered in the next twelve hours of darkness (down to 1.5 [ $\mu\text{M}/\text{h}$ ]). The system functions as follows (for the complete schema, see Fig. 3):

- Transcription of genes *Per*, *Cry* and *Bmal1* occurs in the nucleus. The newly synthesized mRNAs are exported to the cytosol.
- In the cytosol, the mRNAs can be either degraded or translated into proteins, which ones are subsequently phosphorylated (the process is reversible). Unphosphorylated proteins PER and CRY form the complex PER-CRY, which reversibly enters the nucleus. The nuclear and cytosolic forms of the complex can be phosphorylated. Likewise, protein BMAL1 is reversibly phosphorylated and reversibly enters the nucleus, but sole its unphosphorylated form makes a complex with protein CLOCK. Phosphorylated proteins and complexes in the nucleus or the cytosol are subject to degradation.
- In the nucleus, the complex CLOCK-BMAL1 activates the transcription of *Per* and *Cry* genes. Activation is stopped by binding of the PER-CRY complex to CLOCK-BMAL1, which indirectly inhibits *Per* and *Cry* transcription.
- The concentration of CLOCK protein is not a variable in the model because it is constitutively expressed at high levels and considered to be not limiting [17].

The 16 model equations, 56 parameter and 16 initial condition values are shown in Appendix B. The model dynamics is difficult to analyse though, as the circadian clock involves numerous processes, including interlocked positive and negative feedback loops responsible for the oscillatory behaviour of the clock proteins. Reducing the original model around its core *active* processes can facilitate the model analysis, without changing significantly the original dynamics, in particular the sustained oscillations of the solutions.

### Principal Process Analysis of the circadian clock model

We apply Principal Process Analysis to identify major processes of the circadian clock model. To this end we decompose each ordinary differential equation in processes, as shown in Eq. (4) for BMAL1. Each process has a biological interpretation and corresponds to a regulatory mechanism or a biochemical reaction.

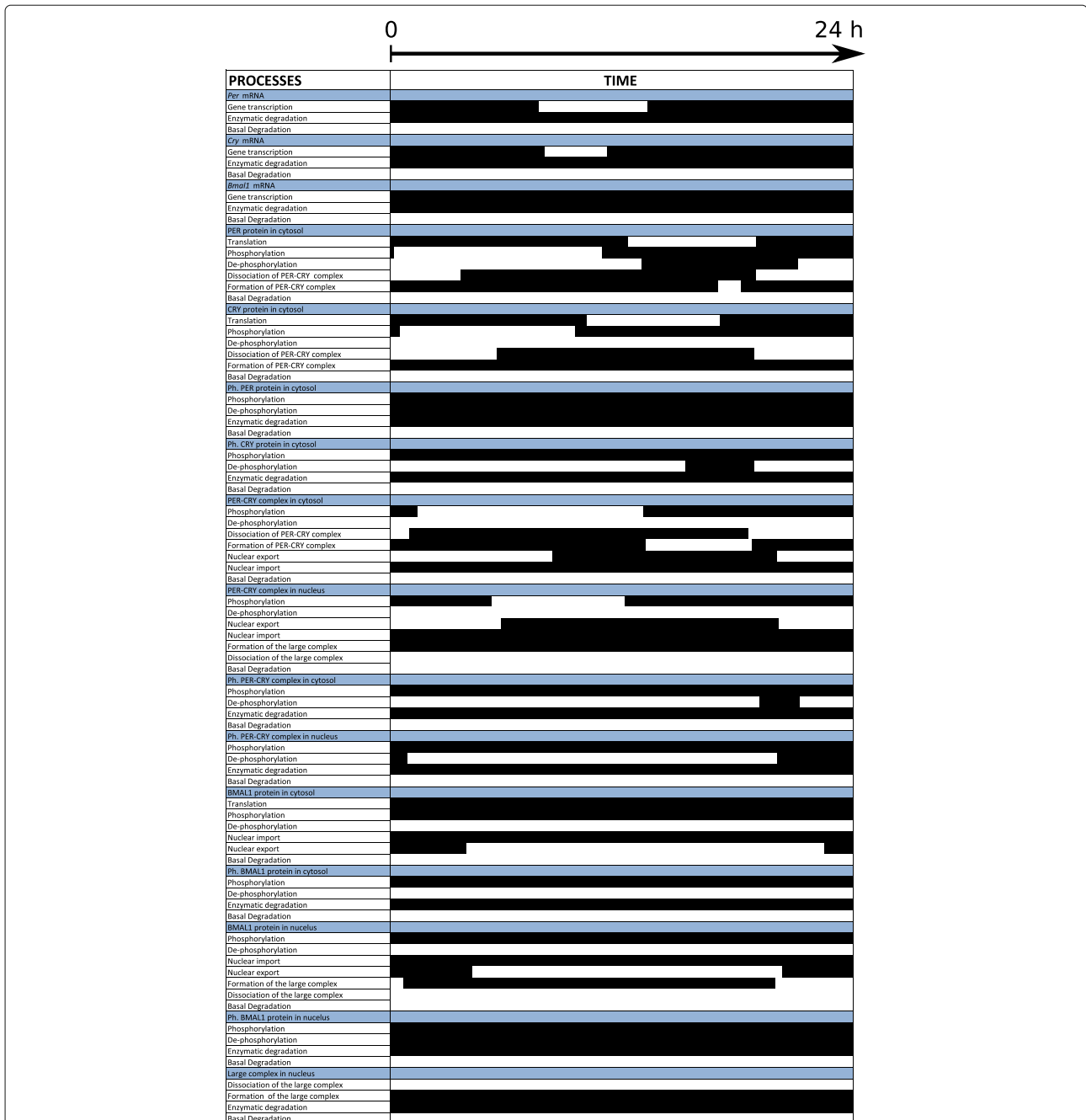


We then calculate the relative weight of each process using Eq. (5) and the threshold value  $\delta = 0.1$ .

We collect the *switching times* (values given in Appendix C) and then build a *Boolean Process Map* to visualize the *activity/inactivity* of each process, shown in Fig. 4. We simplify the model by neglecting 24 out of 76 processes, which are *always inactive* (32% of all processes). They correspond to mRNA and protein basal degradations; cytosolic dephosphorylations of CRY, BMAL1, and PER-CRY; PER-CRY-CLOCK-BMAL1 dissociation in the nucleus; and BMAL1 dephosphorylation in the nucleus. The list of neglected processes is shown in Appendix D.

We now determine the global relative errors between the original and the simplified model using Eq. (8) for

all six outputs (see Table 1). The dynamics of the two models are compared in Fig. 7a. The simplified model preserves qualitatively the trend of the original solutions, as well as their sustained oscillations. The most noticeable difference concerns the peak of the total concentration of protein PER ( $P_{Tot}$ ), which corresponds also to the highest error in Table 1 (26.48%): the peak is lower with the reduced model, which also explains the delay between the original and the simplified solutions. Nevertheless the simplified model reproduces qualitatively the oscillatory behaviour of protein PER observed in the original model. The concentrations in the original and simplified models peak at almost the same time. These global relative errors do not call for an adjustment of the threshold value  $\delta$ . In the next section, we



**Fig. 4** Activity of the 76 model processes during a 24-h period. Processes are listed in the first column (white background), ordered by variable (blue background). Their activity is depicted in the second column between 0 and 24 h: a horizontal black, resp. white, bar when the process is active, resp. inactive. Values for the switching times are given in Appendix B

proceed to the second step of the Principal Process Analysis.

**Creation of sub-models**

The simplified model obtained above can be further reduced if we also neglect processes that are some-

times inactive during the system dynamics. Based on the Boolean Process Map and the collected switching times, we identify between 38 and 45 active processes along time (Fig. 5) and a total of 46 switching times (see Fig. 6a). Clustering the switching times into 4 clusters (Fig. 6b) allows us to generate the five sub-models described

**Table 1** Global relative errors between the original and reduced models for the six outputs

Global relative error						
Output	$M_P$	$M_C$	$M_B$	$P_{Tot}$	$C_{Tot}$	$B_{Tot}$
Error	0.2499	0.2148	0.1535	0.2648	0.1326	0.2053

below. The number of clusters has been chosen according to Eq. (10).

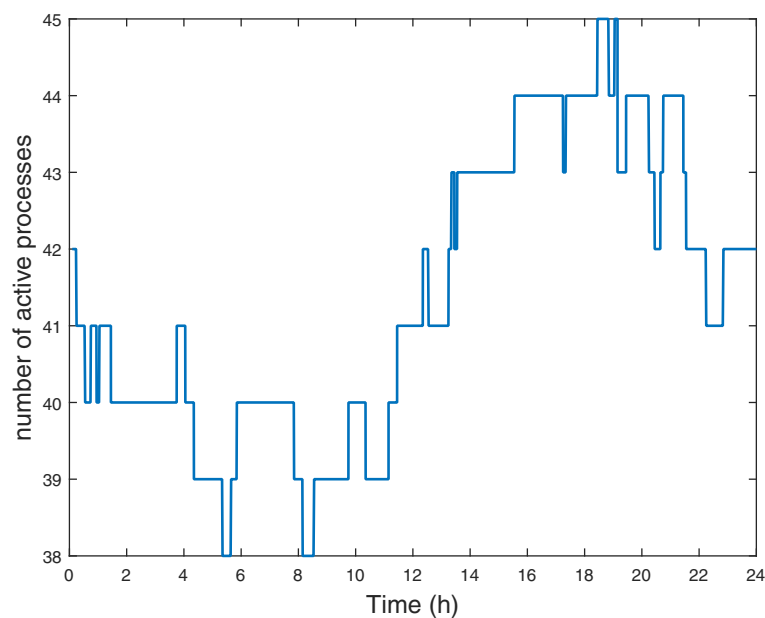
- *SM1*, valid from  $t_0^r = 0$  to  $t_1^r = 0.9$  h: neglected processes for this model are *always inactive* (32% of the total). This model corresponds to the simplified model obtained in “Principal Process Analysis of the circadian clock model” section.
- *SM2*, from  $t_1^r = 0.9$  h to  $t_2^r = 6$  h: 46% of the processes are neglected. In addition to the *always inactive* listed in “Principal Process Analysis of the circadian clock model” section, we have the following *inactive* processes in this model: cytosolic dephosphorylation of PER, CRY, and PER-CRY; cytosolic dissociation of PER-CRY; nuclear dephosphorylation of PER-CRY; PER-CRY export from the nucleus; and formation of the large complex PER-CRY-CLOCK-BMAL1.
- *SM3*, from  $t_2^r = 6$  h to  $t_3^r = 12.5$  h, in which 50% of processes are neglected. In addition to the processes listed in “Principal Process Analysis of the circadian clock model” section, *inactive* processes are in this case: transcription of *Per* and *Cry* mRNAs; cytosolic

phosphorylations and dephosphorylations of PER and CRY; cytosolic dephosphorylation of PER-CRY; nuclear phosphorylation and dephosphorylation of PER-CRY; and nuclear export of BMAL1.

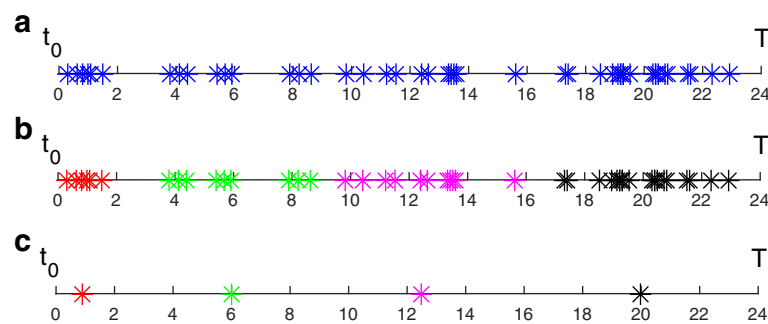
- *SM4*, from  $t_3^r = 12.5$  h to  $t_4^r = 20$  h, which neglects 42% of processes. The processes include the processes listed in “Principal Process Analysis of the circadian clock model” section, as well as: PER and CRY translation; formation of the PER-CRY complex in the cytosol; PER-CRY dephosphorylation in the cytosol and the nucleus; and export of BMAL1 from the nucleus.
- *SM5*, from  $t_4^r = 20$  h to  $t_5^r = 24$  h, in which 46% of the processes are neglected. With the processes listed in “Principal Process Analysis of the circadian clock model” section, other neglected processes are: cytosolic dephosphorylation of PER and CRY; PER-CRY dissociation in the cytosol; export of PER-CRY; PER-CRY dephosphorylation both in the cytosol and the nucleus; and PER-CRY-CLOCK-BMAL1 formation.

See also Appendix D for the list of neglected processes in each sub-model.

Table 2 gives the global relative errors (11) without propagation error, between the original model and the sub-models for the six outputs and for each time window. Figure 7b illustrates the six model outputs for the original model and the sub-models without propagation errors, while Fig. 7c compares the coupled sub-models with and without propagation error. The simplified models preserve



**Fig. 5** Evolution of the number of *active* processes as a function of time. The function increases or decreases at *switching times*, listed in Appendix B



**Fig. 6** Switching time clustering. **a** switching times  $t_b$ ,  $b = 1, \dots, 46$  (also listed in Appendix B). **b** the four switching time clusters (red, green, pink, black) obtained by the k-means method. **c** the four reduced switching times ( $t_v^*$ ,  $v = 1, \dots, 4$ ), corresponding to the mean switching time within each cluster

the oscillatory behaviour of the total concentrations of PER, CRY, and BMAL1, albeit with some discrepancies in the amplitude of the oscillations. It is in the third time window that the approximated solution differs the most from the original one (Table 2). This is visible in Fig. 7b in the third time window where the total concentrations of PER and CRY form a much higher peak in the solution of the simplified model. Recall that this error is not an issue, since our objective is primarily the qualitative analysis of the model. It is sufficient that the remaining processes in the simplified model produce a dynamical behaviour qualitatively similar and relatively close to the original model. This shows their important contribution to the system dynamics.

Applying a *Dynamical Process Map* to the third sub-model (Fig. 8; see also “Visualization of process activities” section) shows that the transcription of *Per* and *Cry* genes is *inactive* (black arrow) and that both PER and CRY phosphorylations in the cytosol and in the nucleus are not entirely *active* (they are not *active* for all the variables in which they are involved, yellow arrow). In the other time windows these processes are always entirely *active* (red arrows). This probably explains why we had an higher error in Table 2 for the variable  $M_P$ ,  $M_C$ ,  $P_{Tot}$  and  $C_{Tot}$  in SM3. The global sensitivity analysis, presented

in the next Section, will confirm the validity of this assumption.

Since the dynamics of the coupled sub-models remain close to the original one, we can further analyse the behaviour of the network simplified to its core processes. We use the *Dynamical Process Maps* for the different sub-models (Appendix E), together with the process activities in Fig. 4 and the model outputs in Fig. 7. The simplified models preserve the three main interlocked feedback loops described in the original model, one positive and two negative loops. The functioning of these loops is directly affected by changes of process activities. Among the two negative feedback loops, which one is the main oscillator? One negative feedback loop involves the inhibition of *Bmal1* transcription by the nuclear form of BMAL1 associated to the protein CLOCK. If this mechanism is the main source of oscillations, we should observe wide changes in process activities controlling BMAL1 levels. The total concentration of the protein does not vary much in amplitude (Fig. 7). It mainly decreases in SM2 and SM3, when the concentration of PER-CRY is also high and forms a complex with CLOCK-BMAL1, which is subsequently degraded. This degradation process is *active* most of the time (Fig. 4 and Appendix E), but variations of the total BMAL1 concentration do not modify strongly the transcription of *Bmal1* mRNA, which remains always *active*. As well, the other processes of translation, phosphorylation and degradation for this variable almost never switch between *inactive* and *active* states over time (Fig. 4 and Appendix E). Overall, this suggests that the negative feedback loop involving CLOCK-BMAL1 is not the main oscillator. A similar conclusion was drawn for the original model in [17].

The other negative feedback loop inhibits *Per* and *Cry* transcription through the titration of CLOCK-BMAL1 by PER-CRY to form the inhibitory complex PER-CRY-CLOCK-BMAL1. The total concentration of BMAL1 peaks before that of PER and CRY, as can be seen in

**Table 2** Global relative error between the original model and each sub-model without propagation error for the six outputs

Output	Global relative error					
	$M_P$	$M_C$	$M_B$	$P_{Tot}$	$C_{Tot}$	$B_{Tot}$
Error SM1	0.0044	0.0044	0.0044	0.0208	0.0195	0.0073
Error SM2	0.0519	0.0434	0.0453	0.0397	0.1832	0.0402
Error SM3	0.2059	0.2951	0.0360	0.1427	0.2233	0.0356
Error SM4	0.0143	0.0377	0.0389	0.0678	0.1164	0.0210
Error SM5	0.0146	0.0032	0.0230	0.1150	0.0237	0.0053

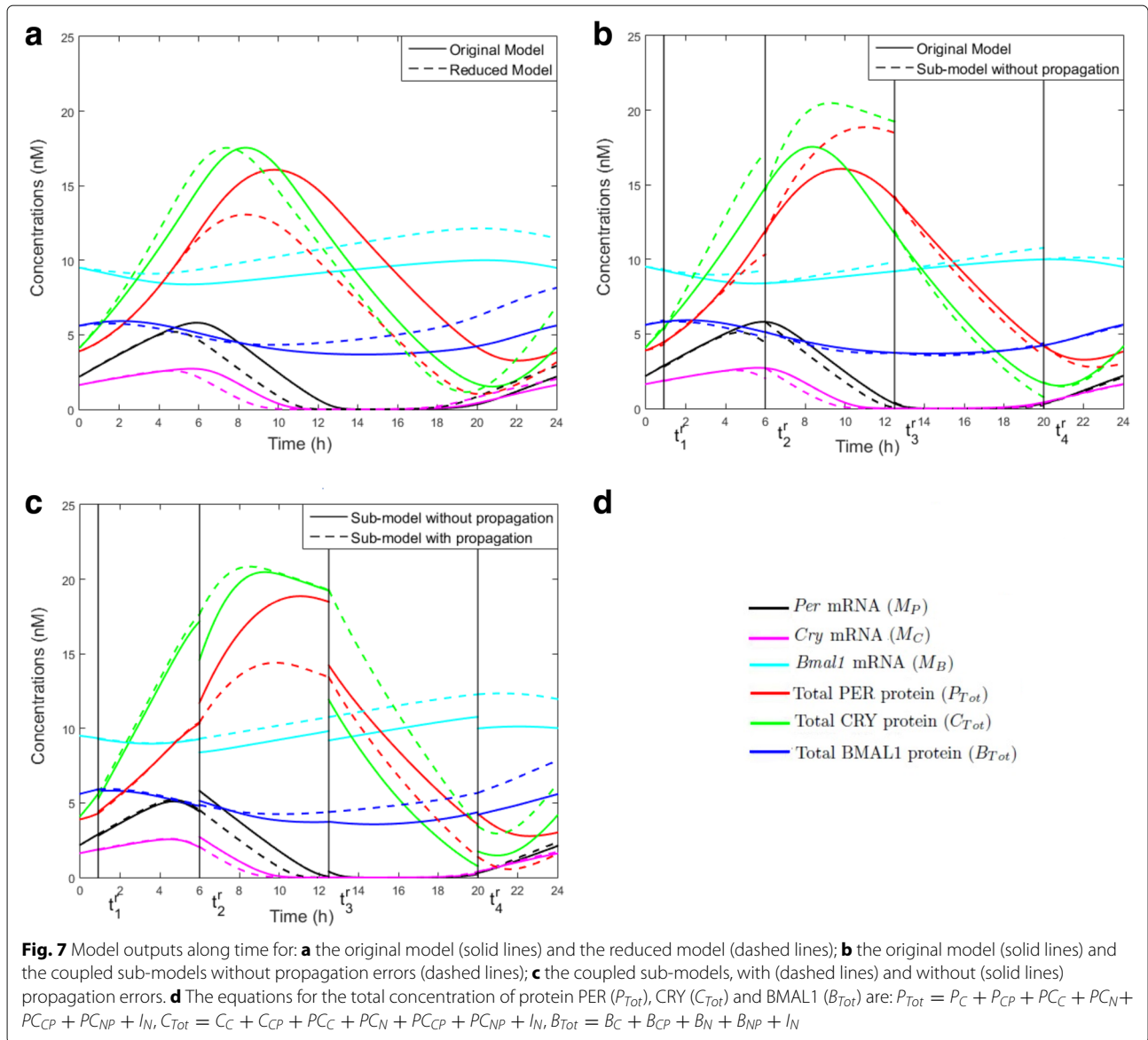
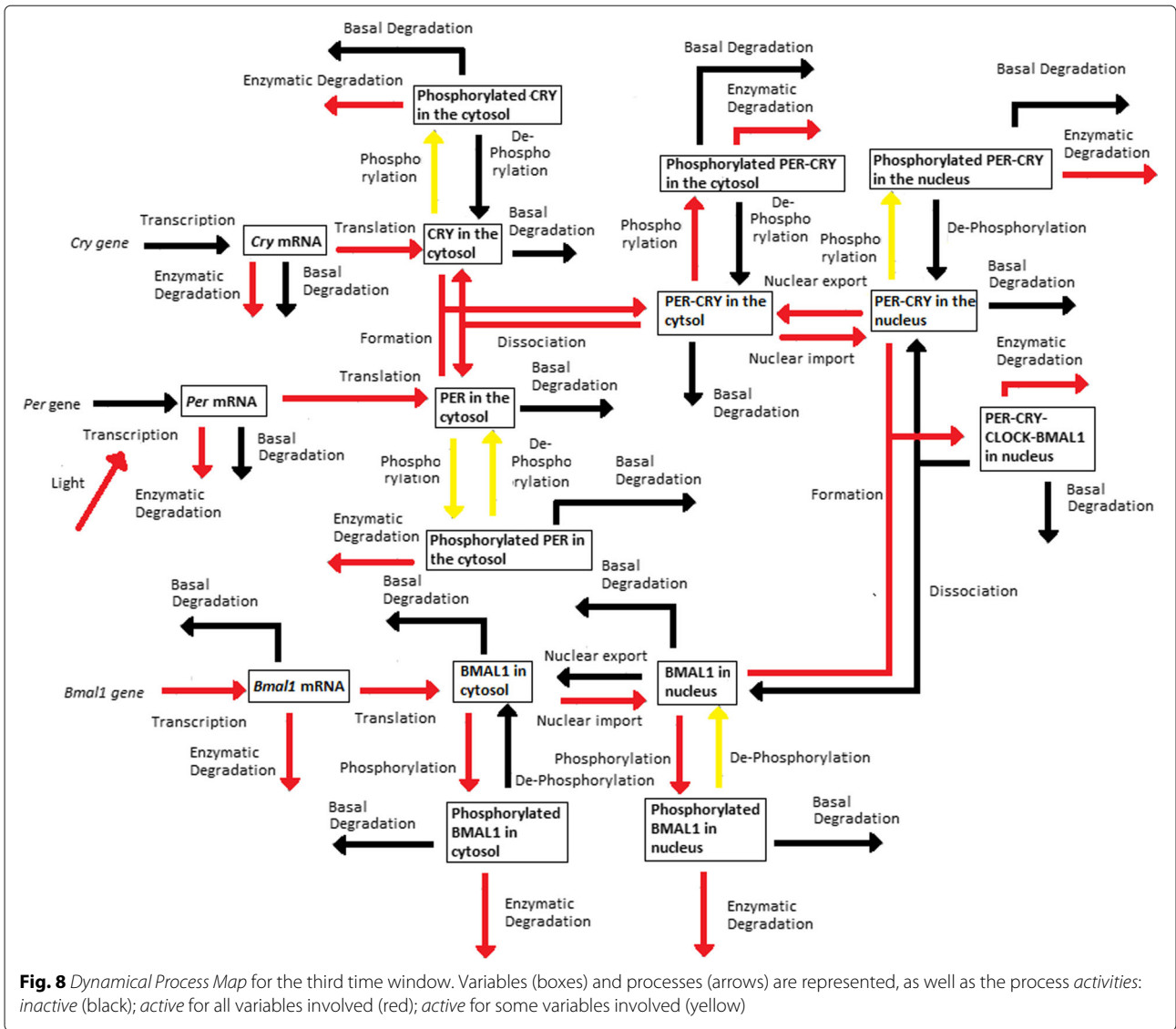


Fig. 7 for SM2 and SM3. When its concentration is maximal in SM1 and SM2, the nuclear form of the protein associated to the protein CLOCK stimulates the transcription of *Per* and *Cry* genes, in conditions where light has also a stimulatory effect on the transcription of these two genes. The processes of transcription and translation of *Per* and *Cry* are active in both models, as a result of which levels of PER and CRY raise to reach their maximal concentration in SM3. As can be seen from the process activities in Fig. 4 and the Dynamical Process Maps in Appendix E, conditions are favourable for the accumulation of high levels of complexes PER-CRY and CLOCK-BMAL1-PER-CRY in the nucleus. For instance, numerous processes decreasing PER, CRY

and PER-CRY concentrations in the cytosol and the nucleus are inactive: their phosphorylation is reduced (the process is inactive for the dephosphorylated forms but still active for the phosphorylated ones), which limits their degradation, and the nuclear import of PER-CRY is always active. During the same period of time, the formation of the large complex CLOCK-BMAL1-PER-CRY, which is active for both CLOCK-BMAL1 and PER-CRY (Fig. 4 and Appendix E), suggests that the nuclear forms of PER-CRY and CLOCK-BMAL1 bind as soon as they accumulate in the nucleus. The large complex is immediately degraded since its degradation process is always active and its dissociation, always inactive.



In SM2 and SM3, the degradation of the large complex is not compensated for by other mechanisms allowing BMAL1 accumulation in the nucleus: the cytosolic form of the protein is *actively* phosphorylated and then degraded, while its dephosphorylation is *inactive*, which reduces the quantity of protein to be imported in the nucleus (see Fig. 4 and the *Dynamical Process Maps* in Appendix E). In this compartment, the absence of *active* dephosphorylation, together with the *active* protein phosphorylation, also contribute to decrease pools of CLOCK-BMAL1 complexes (Fig. 4, Appendix E). This halts transcription of *Per* and *Cry* mRNAs in SM3 (the processes are *inactive* and light is also switched off towards the end of SM3). This also affects the translation of PER and CRY, which becomes *inactive* in SM4. Altogether these observations suggest that the negative feedback

loop inhibiting *Per* and *Cry* transcription via the complex CLOCK-BMAL1-PER-CRY is the main source of circadian oscillations. This is consistent with conclusions in [17], where a second oscillator based on the auto-inhibition of BMAL1 has been obtained for specific parameter values only. These results are also consistent with the observation of arrhythmic behaviours in mutant mice with double knock-out of the *Per* and *Cry* genes [25, 26].

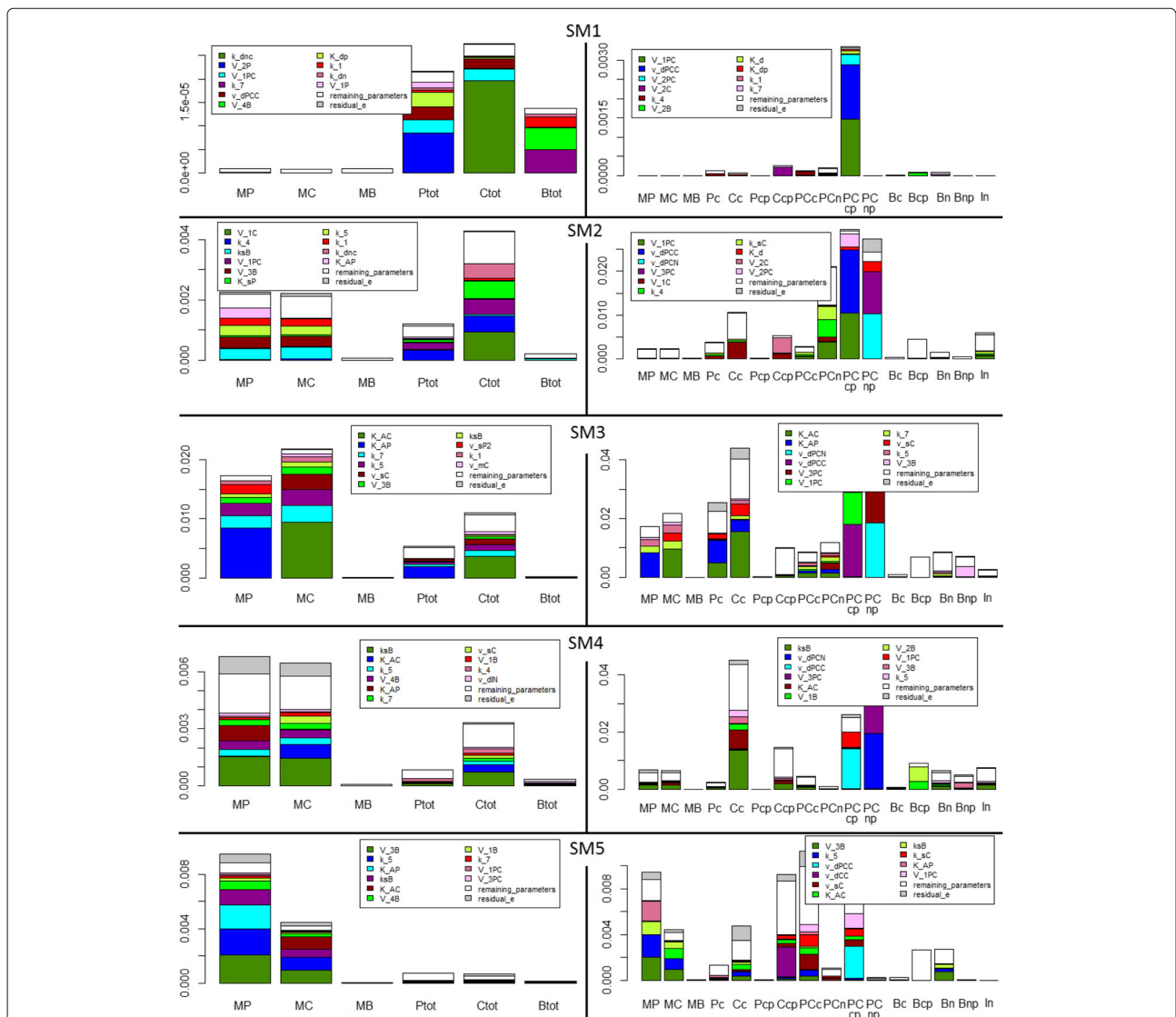
The positive feedback loop activates *Per* and *Cry* transcription through a control of protein stability mediated by the phosphorylation processes. In the model, sole the phosphorylated forms of the proteins are degraded. We observed that the reversible phosphorylation reactions are often displaced in the forward sense, as dephosphorylation processes are often found *inactive*. In particular,

they contribute to decrease the concentration of PER, CRY and PER-CRY, which also diminishes the concentration of the large complex CLOCK-BMAL1-PER-CRY and thus relieves the inhibition exerted by the complex on transcription of *Per* and *Cry* genes. Kinetic modelling of the circadian clock in *Drosophila* has shown the importance of this positive feedback loop for circadian rhythms [27].

**Influence of parameter values**

In order to check the robustness of the five sub-models, we perform a global sensitivity analysis on the output

errors for each time window ( $e_h^v$ ). We perform the analysis without propagating the errors because each sub-model is valid for a specific time window, independently from the other time windows. We vary 51 among the 56 parameters of the model: the Hill coefficients  $m$  and  $n$  are kept fixed because they represent the degree of cooperativity in gene repression/activation, while  $k_{stot}$ ,  $v_{stot}$ ,  $V_{phos}$  are function of other parameters (see Appendix B). We hence compute the non normalised total sensitivity indices for all parameters according to Eq. (15) (see Fig. 9, first column). Because the last three outputs ( $P_{Tot}$ ,  $C_{Tot}$ ,  $B_{Tot}$ ) are the sum of model variables



**Fig. 9** Global sensitivity analysis on the output (left column) or variable (right column) errors between the original model and the sub-models without propagation error for each time window (lines). Non-normalised total sensitivity indices are represented for each error (one bar per error) and for: (i) the 10 most influential parameters (color-coded); (ii) the remaining parameters (white). The residual is also represented (grey). For the biological meaning of the variables in the second column, see the equations in Appendix A



that interact, some processes have no impact on these outputs and the information on the parameter influence is lost. We also perform the global sensitivity analysis on the 16 global relative errors between the original model and the sub-model variables without propagating errors (see Fig. 9, second column). The complex PER-CRY plays an important role in every time window: its variability is due mostly to its maximal phosphorylation velocity ( $V_{1PC}$ ) and its degradation parameter ( $v_{dPCC}$ ). In the third and fourth time window the other important variation is due to the CRY protein: in SM3 the variation is mostly due to the binding constants in the transcription of *Per* and *Cry* mRNAs ( $K_{AP}$  and  $K_{AC}$ ) and in SM4, to the maximal translation rate of BMAL1 ( $k_{sB}$ ) that stimulates *Per* and *Cry* mRNA transcription. In the last time window, lots of variables contribute to the system variation: the most important parameter for the variability of the outputs is the maximal velocity of BMAL1 phosphorylation in the nucleus ( $V_{3B}$ ).

To get a more global view of the model simplification, we calculate, for each parameter combination and for each time window, the average error (averaged over the 16 variables) between the original model and the sub-model variables as follows:

$$\bar{e}^v = \frac{1}{16} \sum_{i=1}^{16} e_i^v \tag{17}$$

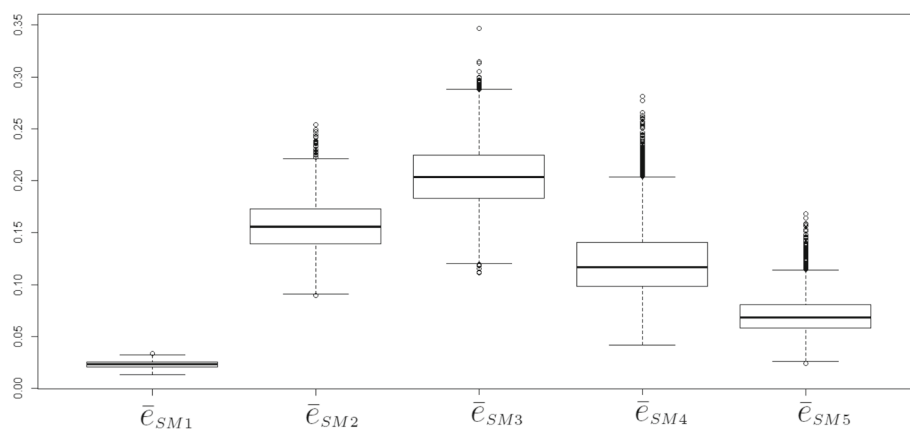
Results are shown in Fig. 10. The variability is higher in the third and four sub-model, although the difference between the lower and upper quartiles is low in all sub-models.

Then, for each time-window, we compute the total generalised sensitivity indices according to Eq. (16), which

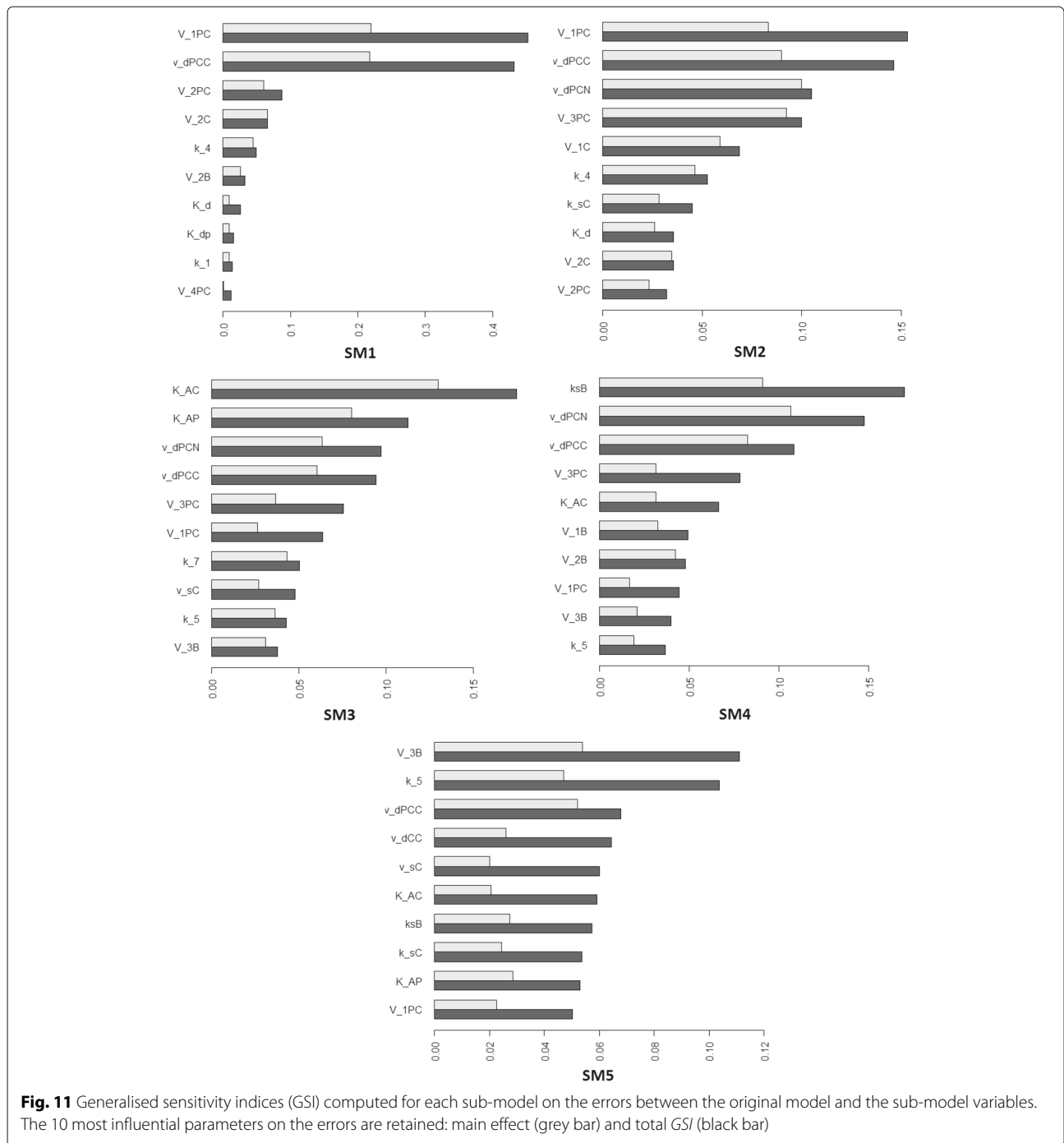
represents the fraction of error variability explained by each parameter when parameter values vary. The results are shown in Fig. 11: we obtain similar results to the ones in Fig. 9 (column 2): in SM1 and SM2, the maximal phosphorylation velocity ( $V_{1PC}$ ) and degradation ( $v_{dPCC}$ ) of PER-CRY complex play the main role; in SM3, the binding constants of *Per* and *Cry* proteins ( $K_{AP}$  and  $K_{AC}$ ); in SM4, the translation of BMAL1 protein ( $k_{sB}$ ) and in SM5, the maximal phosphorylation velocity of BMAL1 protein in the nucleus ( $V_{3B}$ ). In order to check whether the error variations between the original model and the sub-models are due to parameters appearing in neglected processes, we determine

the following ratio:  $R_h^v = \frac{\sum_{f \in \{inactive\ processes\}} tGSI_f^{h,v}}{\sum_f tGSI_f^{h,v}}$ . We only

use the 10 most informative parameters, with higher  $tGSI$ , as they explain most variability. We choose a conservative option: if a parameter is neglected in an *inactive* process but still appears in other *active* processes, we still consider that it belongs to the neglected process parameters (worst case). Results are shown in Table 3. In most time windows, the variability is mainly due to parameters still contained in the reduced sub-models, i.e. the parameters of the *active* processes. In the third time-window, however, parameters appearing in neglected processes generate more than 50% of the variability. It is consistent with Fig. 7b: the peaks of the total concentration of PER and CRY are overestimated by the sub-model and some of the most important parameters that lead to the output variability for this time window are the translation rate of PER and CRY proteins, the maximal phosphorylation velocity of PER-CRY complex in the cytosol and nucleus (as it has been shown in Fig. 11). This confirms what we



**Fig. 10** Average error between the original model and the sub-model variables calculated in each time window according to Eq. (17). Variability (box-plots) within each sub-model (or time window) is due to the various parameter combinations designed for the sensitivity analysis



have supposed when applying the *Dynamical Process Map* to SM3 (see discussion about Fig. 8 at the end of “Creation of sub-models” section).

**Discussion**

A challenging task when analysing the dynamics of biological networks is to understand the relation between the network behaviour and its numerous processes, the

**Table 3** Percentage of *tGSI* for parameters contained in *inactive* processes

	% <i>tGSI</i> inactive				
SM	SM1	SM2	SM3	SM4	SM5
<i>R<sub>h</sub></i> (%)	19.11	15.55	59.54	0	0

*activity* of which is switched on and off by regulatory mechanisms. Model reduction is one possible approach to deal with model complexity and help deciphering the design principle of these networks. However, the reduced models can be still too complex to study and this does not answer the question on the role of each individual process in the network behaviour. Ideally, one would like to identify the major processes, quantify and then understand their contribution to the system dynamics. Principal Process Analysis was developed with this objective in mind, and with the final goal of simplifying the original model in one or several sub-models around *active* processes that are responsible for the dynamics of the original system. The dynamics of the core processes is much more tractable in the sub-models than in the original one. Questions remained open though concerning the scalability and robustness of this approach.

In this paper we tested the scalability of Principal Process Analysis by applying the approach on a model of high dimension, the mammalian circadian clock model, which incorporates numerous processes and complex interlocked feedback loops responsible for oscillatory behaviours. Simplification of the original system dynamics to as much as 50% of its processes in five coupled sub-models helped us relate the dynamics of the simplified models to the system components and their *active* interactions. We hence observed that the negative feedback loop controlling *Per* and *Cry* transcription through the formation of the large complex PER-CRY-CLOCK-BMAL1 is the main oscillator, in agreement with previous experimental and modelling studies [17, 25, 26]. Principal Process Analysis has been also applied with success to diverse biogeochemical and biochemical models in our group and elsewhere, see [14, 21, 28, 29]. These case studies and the present one exemplify the applicability and scalability of the approach to models of diverse nature and complexity.

In this paper, the quantification of global errors allowed us to conclude that the simplified models reproduce well the behaviour of the original ones. Even in the case of the largest errors observed on the model output, did the simplified models preserve the oscillations of the clock proteins. Since Principal Process Analysis is based on the a priori knowledge of the model parameters, it was important to assess the robustness of the approach to uncertainties on these parameter values. Through a global sensitivity analysis, we studied the impact of parameter values on the error between the original model and the simplified sub-models. Not only was the variation of the error small, but it was mostly due to parameters of the neglected processes. With this analysis, we proved the robustness of PPA to parameter uncertainty. In addition, we provided clues to identify and solve potential troubles related to the model simplification, in order to decrease

errors between the original model and the simplified sub-models.

In a recent study, we also showed by other methods the robustness of PPA to initial conditions [21]. The latter were supposed to lie in rectangles contained in a region of the variable space varying by one order of magnitude in each coordinate. Under additional assumptions on the monotonicity of biological processes within rectangles, the maximal bound of process weights was computed, which allowed identifying *active* processes in each rectangle, similarly to “Principal Process Analysis (PPA)” section, for which weight is above the threshold value  $\delta$ . Based on the behaviour of processes on the edges of the rectangles, it was then possible to determine the transitions between rectangles and deduce the evolution of process *activities* along the different transitions. The method has been applied on a small gene expression network containing a negative feedback loop [21]. The same principles could be applied to show the robustness of the model to larger variations of its parameters. In this case, the parameter space should be divided in rectangles in which the *activity/inactivity* of processes is studied. Such extension of the method is part of a future work.

In the current state of development, Principal Process Analysis is not a model reduction approach. For instance, the elimination of the *inactive* processes from the original model breaks down the mass conservation relations when eliminating a process in one equation that is considered *active* in others. As long as the purpose of PPA is to analyse the important processes in the original model, this is not an issue. Nevertheless, the approach could be extended so as to preserve mass conservation relations. In addition, simplified models with much smaller global relative errors could be obtained, so that the simplified sub-models represent more accurately the original model. We are currently studying a refinement of PPA by considering three different levels of *activities* (*inactive*, *active*, *fully active*), defined by two different thresholds in order to improve the quality of the model simplification and model analysis. Such improvements could bring PPA closer to a reduction method, since the simplified models become accurate representations of the original model.

## Conclusions

Mathematical models of biological systems have grown in complexity to include large numbers of processes. As a consequence, their contribution to the system dynamics becomes hardly tractable. The current manuscript contributes to this problem with the development of Principal Process Analysis. Provided the ODE model of the system is composed of a linear sum of terms describing each a process, the method enables the identification of the major processes contributing to the system dynamics and when they play a key role. Removing *inactive* processes

allows restricting the dynamical analysis of the system to its core processes and facilitates the understanding of the system functioning. The conclusions derived with the method are robust to fluctuations in the parameter values. As such, Principal Process Analysis can be applied to any type of ODE models with the same form.

## Endnotes

<sup>1</sup>Total protein concentrations are defined as follows:

$$P_{Tot} = P_C + P_{CP} + P_{CC} + P_{CN} + P_{CCP} + P_{CNP} + I_N,$$

$$C_{Tot} = C_C + C_{CP} + P_{CC} + P_{CN} + P_{CCP} + P_{CNP} + I_N,$$

$$B_{Tot} = B_C + B_{CP} + B_N + B_{NP} + I_N.$$

<sup>2</sup>planor: Generation of regular factorial designs <https://CRAN.R-project.org/package=planor>

<sup>3</sup>multisensi: Multivariate Sensitivity Analysis <https://CRAN.R-project.org/package=multisensi>

## Appendix A: Estimate of errors

We give in this appendix a rough estimate of the a priori error, based on bounds in the model and simple lemmas to compare the solutions of two differential equations. We refer to [18, Chapter 3], for the basic notions.

Consider the following ODE model of biological network, as given in Eq. (2):

$$\dot{x} = f(x, p) \quad (18)$$

Variable  $x$  is supposed to live in a bounded domain  $D$  of  $\mathbb{R}^n$ , and all functions in  $f$  are supposed to be smooth enough (at least  $C^2$ ) and Lipschitz on  $D$ . We denote by  $L$  the Lipschitz constant of  $f(x, p)$  on  $D$ . The decomposition into processes gives:

$$\dot{x}_i = \sum_{j=1}^{n_i} f_{ij}(x, p) \quad i = 1, \dots, n \quad (19)$$

where  $f_{ij}$  represents the  $j^{\text{th}}$  process involved in the dynamical evolution of the  $i^{\text{th}}$  variable of the system over a period of time  $[t_0, T]$ , and  $n_i$  is the number of processes for  $\dot{x}_i$ .

The weights are computed during  $[t_0, T]$ . For the sake of simplicity, we suppose that, for each variable, the weight of the first process and only this weight is lower than threshold  $\delta$ :

$$W_{i1}(t, p) = \frac{|f_{i1}(x(t), p)|}{\sum_j^{n_i} |f_{ij}(x(t), p)|} < \delta \quad i = 1, \dots, n. \quad (20)$$

It means that processes  $f_{i1}, i = 1, \dots, n$  are *inactive* during period  $[t_0, T]$ , and thus eliminated from the system, giving the new simpler system (the new variable is denoted by  $y$  for simplicity):

$$\dot{y}_i = \sum_{j=2}^{n_i} f_{ij}(y, p) \quad i = 1, \dots, n. \quad (21)$$

As variables are assumed to be bounded, vector  $f_1 = (f_{11}, \dots, f_{1n})^t$  is such that:

$$|f_{i1}(x, p)| = W_{i1} \sum_{j=1}^{n_i} |f_{ij}(x(t), p)| \leq \delta B_i \quad \forall x \in D \quad i = 1, \dots, n. \quad (22)$$

where  $B_i$  is an upper bound for  $\sum_{j=1}^{n_i} |f_{ij}(x(t), p)|$  obtained from the variable bounds on domain  $D$ . All  $B_i$  form vector  $B$ .

Therefore:

$$\|f_1(x, p)\| \leq \delta \|B\| \quad (23)$$

If the initial conditions are the same ( $x(t_0) = y(t_0)$ ), then Theorem 3.4 in [18], which is based on Gronwall's Lemma, gives a bound between the two solutions  $x$  and  $y$ :

$$\|x(t) - y(t)\| \leq \delta \frac{\|B\|}{L} (e^{L(t-t_0)} - 1) \quad \forall t \in [t_0, T]. \quad (24)$$

The same proof applies when several  $f_{ij}$  are *inactive* for some variables. One just need to sum the errors in Eq. (22).

This gives a rough bound between the two solutions; this bound is theoretical and conservative, and is not used in the practical a posteriori computation of the error in our work. Nevertheless, it shows that the error is roughly proportional to the threshold  $\delta$  used in the weight computations.

## Appendix B: Full mammalian model

### Model equations

Equations listed in [17, 19].

mRNAs of *Per* gene

$$\begin{aligned} \frac{dM_P}{dt} &= v_{sP} \frac{B_N^n}{K_{AP}^n + B_N^n} - v_{mP} \frac{M_P}{K_{mP} + M_P} - k_{dmp} M_P \\ \dot{x}_1 &= f_{1,1} + f_{1,2} + f_{1,3} \end{aligned}$$

mRNAs of *Cry* gene

$$\begin{aligned} \frac{dM_C}{dt} &= v_{sC} \frac{B_N^n}{K_{AC}^n + B_N^n} - v_{mC} \frac{M_C}{K_{mC} + M_C} - k_{dmc} M_C \\ \dot{x}_2 &= f_{2,1} + f_{2,2} + f_{2,3} \end{aligned}$$

mRNAs of *Bmal1* gene

$$\begin{aligned} \frac{dM_B}{dt} &= v_{sB} \frac{K_{IB}^n}{K_{IB}^n + B_N^n} - v_{mB} \frac{M_B}{K_{mB} + M_B} - k_{dmb} M_B \\ \dot{x}_3 &= f_{3,1} + f_{3,2} + f_{3,3} \end{aligned}$$

Non-phosphorylated PER protein in the cytosol

$$\begin{aligned} \frac{dP_C}{dt} &= k_{sP} M_P - V_{1P} \frac{P_C}{K_P + P_C} + V_{2P} \frac{P_{CP}}{K_{dP} + P_{CP}} \\ &\quad + k_4 P_{CC} - k_3 P_C C_C - k_{dn} P_C \\ \dot{x}_4 &= f_{4,1} + f_{4,2} + f_{4,3} + f_{4,4} + f_{4,5} + f_{4,6} \end{aligned}$$

Non-phosphorylated CRY protein in the cytosol

$$\begin{aligned} \frac{dC_C}{dt} &= k_{sC} M_C - V_{1C} \frac{C_C}{K_C + C_C} + V_{2C} \frac{C_{CP}}{K_{dC} + C_{CP}} \\ &\quad + k_4 P_{CC} - k_3 P_C C_C - k_{dnc} C_C \\ \dot{x}_5 &= f_{5,1} + f_{5,2} + f_{5,3} + f_{5,4} + f_{5,5} + f_{5,6} \end{aligned}$$

## Phosphorylated PER protein in the cytosol

$$\begin{aligned} \frac{dPCP}{dt} &= V_{1P} \frac{PC}{K_P + PC} - V_{2P} \frac{PCP}{K_{dP} + PCP} - v_{dPC} \frac{PCP}{K_d + PCP} - k_{dn} PCP \\ \dot{x}_6 &= f_{6,1} + f_{6,2} + f_{6,3} + f_{6,4} \end{aligned}$$

## Phosphorylated CRY protein in the cytosol

$$\begin{aligned} \frac{dCCP}{dt} &= V_{1C} \frac{CC}{K_C + CC} - V_{2C} \frac{CCP}{K_{dC} + CCP} - v_{dCC} \frac{CCP}{K_d + CCP} - k_{dn} CCP \\ \dot{x}_7 &= f_{7,1} + f_{7,2} + f_{7,3} + f_{7,4} \end{aligned}$$

## Non-phosphorylated PER-CRY complex in the cytosol

$$\begin{aligned} \frac{dPC_C}{dt} &= -V_{1PC} \frac{PC_C}{K_P + PC_C} + V_{2PC} \frac{PC_{CP}}{K_{dP} + PC_{CP}} - k_4 PC_C \\ &\quad + k_3 PC_C C_C + k_2 PC_N - k_1 PC_C - k_{dn} PC_C \\ \dot{x}_8 &= f_{8,1} + f_{8,2} + f_{8,3} + f_{8,4} + f_{8,5} + f_{8,6} + f_{8,7} \end{aligned}$$

## Non-phosphorylated PER-CRY complex in the nucleus

$$\begin{aligned} \frac{dPC_N}{dt} &= -V_{3PC} \frac{PC_N}{K_P + PC_N} + V_{4PC} \frac{PC_{NP}}{K_{dP} + PC_{NP}} - k_2 PC_N + k_1 PC_C \\ &\quad - k_7 B_N PC_N + k_8 I_N - k_{dn} PC_N \\ \dot{x}_9 &= f_{9,1} + f_{9,2} + f_{9,3} + f_{9,4} + f_{9,5} + f_{9,6} + f_{9,7} \end{aligned}$$

## Phosphorylated PER-CRY complex in the cytosol

$$\begin{aligned} \frac{dPC_{CP}}{dt} &= V_{1PC} \frac{PC_C}{K_P + PC_C} - V_{2PC} \frac{PC_{CP}}{K_{dP} + PC_{CP}} - v_{dPC} \frac{PC_{CP}}{K_d + PC_{CP}} - k_{dn} PC_{CP} \\ \dot{x}_{10} &= f_{10,1} + f_{10,2} + f_{10,3} + f_{10,4} \end{aligned}$$

## Phosphorylated PER-CRY complex in the nucleus

$$\begin{aligned} \frac{dPC_{NP}}{dt} &= V_{3PC} \frac{PC_N}{K_P + PC_N} - V_{4PC} \frac{PC_{NP}}{K_{dP} + PC_{NP}} - v_{dPC} \frac{PC_{NP}}{K_d + PC_{NP}} - k_{dn} PC_{NP} \\ \dot{x}_{11} &= f_{11,1} + f_{11,2} + f_{11,3} + f_{11,4} \end{aligned}$$

## Non-phosphorylated BMAL1 protein in the cytosol

$$\begin{aligned} \frac{dB_C}{dt} &= k_{sB} M_B - V_{1B} \frac{B_C}{K_P + B_C} + V_{2B} \frac{B_{CP}}{K_{dP} + B_{CP}} - k_5 B_C + k_6 B_N - k_{dn} B_C \\ \dot{x}_{12} &= f_{12,1} + f_{12,2} + f_{12,3} + f_{12,4} + f_{12,5} + f_{12,6} \end{aligned}$$

## Phosphorylated BMAL1 protein in the cytosol

$$\begin{aligned} \frac{dB_{CP}}{dt} &= V_{1B} \frac{B_C}{K_P + B_C} - V_{2B} \frac{B_{CP}}{K_{dP} + B_{CP}} - v_{dB} \frac{B_{CP}}{K_d + B_{CP}} - k_{dn} B_{CP} \\ \dot{x}_{13} &= f_{13,1} + f_{13,2} + f_{13,3} + f_{13,4} \end{aligned}$$

## Non-phosphorylated BMAL1 protein in the nucleus

$$\begin{aligned} \frac{dB_N}{dt} &= -V_{3B} \frac{B_N}{K_P + B_N} + V_{4B} \frac{B_{NP}}{K_{dP} + B_{NP}} + k_5 B_C - k_6 B_N - k_7 B_N PC_N \\ &\quad + k_8 I_N - k_{dn} B_N \\ \dot{x}_{14} &= f_{14,1} + f_{14,2} + f_{14,3} + f_{14,4} + f_{14,5} + f_{14,6} + f_{14,7} \end{aligned}$$

## Phosphorylated BMAL1 protein in the nucleus

$$\begin{aligned} \frac{dB_{NP}}{dt} &= V_{3B} \frac{B_N}{K_P + B_N} - V_{4B} \frac{B_{NP}}{K_{dP} + B_{NP}} - v_{dB} \frac{B_{NP}}{K_d + B_{NP}} - k_{dn} B_{NP} \\ \dot{x}_{15} &= f_{15,1} + f_{15,2} + f_{15,3} + f_{15,4} \end{aligned}$$

## Inactive complex between PER-CRY and CLOCK-BMAL1 in the nucleus

$$\begin{aligned} \frac{dI_N}{dt} &= -k_8 I_N + k_7 B_N PC_N - v_{dIN} \frac{I_N}{K_d + I_N} - k_{dn} I_N \\ \dot{x}_{16} &= f_{16,1} + f_{16,2} + f_{16,3} + f_{16,4} \end{aligned}$$

## Model parameters

Parameters listed in [17, p.546]: Set 1.

$$\begin{aligned} k_1(h^{-1}) &= 0.4, k_2(h^{-1}) = 0.2, k_3(nM^{-1}h^{-1}) = 0.4, \\ k_4(h^{-1}) &= 0.2, k_5(h^{-1}) = 0.4, k_6(h^{-1}) = 0.2, \\ k_7(nM^{-1}h^{-1}) &= 0.5, k_8(h^{-1}) = 0.1, K_{AP}(nM) = 0.7, \\ K_{AC}(nM) &= 0.6, K_{IB}(nM) = 2.2, k_{dmb}(h^{-1}) = \\ &0.01, k_{dmc}(h^{-1}) = 0.01, k_{dmp}(h^{-1}) = 0.01, k_{dnc}(h^{-1}) = \\ &0.12, k_{dn}(h^{-1}) = 0.01, K_d(nM) = 0.3, K_{dP}(nM) = 0.1, \\ K_P(nM) &= 0.1, K_{mB}(nM) = 0.4, K_{mC}(nM) = \\ &0.4, K_{mP}(nM) = 0.31, k_{stot}(h^{-1}) = 1.0, k_{sB}(h^{-1}) = \\ &0.12k_{stot}, k_{sC}(h^{-1}) = 1.6k_{stot}, k_{sP}(h^{-1}) = 0.6k_{stot}, n = 4, \\ m = 2, V_{phos}(nMh^{-1}) &= 0.4, V_{1B}(nMh^{-1}) = 0.5, \\ V_{1C}(nMh^{-1}) &= 0.6, V_{1P}(nMh^{-1}) = V_{phos}, V_{1PC}(nMh^{-1}) = \\ &V_{phos}, V_{2B}(nMh^{-1}) = 0.1, V_{2C}(nMh^{-1}) = 0.1, \\ V_{2P}(nMh^{-1}) &= 0.3, V_{2PC}(nMh^{-1}) = 0.1, V_{3B}(nMh^{-1}) = \\ &0.5, V_{3PC}(nMh^{-1}) = V_{phos}, V_{4B}(nMh^{-1}) = 0.2, \\ V_{4PC}(nMh^{-1}) &= 0.1, v_{dBC}(nMh^{-1}) = 0.5, \\ v_{dBN}(nMh^{-1}) &= 0.6v_{dCC}(nMh^{-1}) = 0.7, v_{dIN}(nMh^{-1}) = \\ &0.8, v_{dIN}(nMh^{-1}) = 0.8, v_{dPC}(nMh^{-1}) = 0.7, \\ v_{dPCC}(nMh^{-1}) &= 0.7, v_{dPCN}(nMh^{-1}) = 0.7, v_{mB}(nMh^{-1}) = \\ &0.8, v_{mC}(nMh^{-1}) = 1.0, v_{mP}(nMh^{-1}) = 1.1, \\ v_{stot}(nMh^{-1}) &= 1.0, v_{sB}(nMh^{-1}) = v_{stot}, v_{sB}(nMh^{-1}) = \\ &v_{stot}, v_{sC}(nMh^{-1}) = 1.1v_{stot}, v_{sP}(nMh^{-1}) = 1.5v_{stot} \end{aligned}$$

## Initial conditions

The unit of the initial conditions is  $nM$ .

$$\begin{aligned} M_P(0) &= 2.188M_C(0) = 1.633, M_B(0) = 9.498, P_C(0) = \\ &2.008, C_C(0) = 1.884, P_{CP}(0) = 0.129, C_{CP}(0) = 0.473, \\ P_{C_C}(0) &= 1.228, P_{C_N}(0) = 0.177, P_{C_{CP}}(0) = 0.203, \\ P_{C_{NP}}(0) &= 0.101, B_C(0) = 2.523, B_{CP}(0) = 0.929, B_N(0) = \\ &1.787, B_{NP}(0) = 0.318, I_N(0) = 0.051 \end{aligned}$$

## Appendix C: Switching times

See Table 4.

**Table 4** Switching times (s.t.), their values (v.) in [h] and associate reduced (cluster) switching times ( $t_1^r, t_2^r, t_3^r, t_4^r$ ) (s.t.c.):  $t_1^r$  is associated to the cluster of  $t_1 - t_6, t_2^r$  to  $t_7 - t_{15}, t_3^r$  to  $t_{16} - t_{26}$ , and  $t_4^r$  to  $t_{27} - t_{46}$

s.t.	v.	s.t.c.	s.t.	v.	s.t.c.	s.t.	v.	s.t.c.	s.t.	v.	s.t.c.
$t_0$	0		$t_{12}$	5.9	6	$t_{24}$	13.5		$t_{36}$	19.5	20
$t_1$	0.3		$t_{13}$	7.9		$t_{25}$	13.6		$t_{37}$	20.3	
$t_2$	0.6		$t_{14}$	8.2		$t_{26}$	15.6		$t_{38}$	20.4	
$t_3$	0.8	0.9	$t_{15}$	8.6		$t_{27}$	17.3		$t_{39}$	20.45	
$t_4$	1		$t_{16}$	9.8		$t_{28}$	17.4		$t_{40}$	20.5	
$t_5$	1.1		$t_{17}$	10.4		$t_{29}$	18.5		$t_{41}$	20.7	
$t_6$	1.5		$t_{18}$	11.2		$t_{30}$	18.9		$t_{42}$	20.8	
$t_7$	3.8		$t_{19}$	11.5		$t_{31}$	19.1		$t_{43}$	21.5	
$t_8$	4.1		$t_{20}$	12.4	12.5	$t_{32}$	19.2		$t_{44}$	21.6	
$t_9$	4.4		$t_{21}$	12.6		$t_{33}$	19.25		$t_{45}$	22.3	
$t_{10}$	5.4		$t_{22}$	13.3		$t_{34}$	19.3		$t_{46}$	22.9	
$t_{11}$	5.7		$t_{23}$	13.4		$t_{35}$	19.35		$T$	24	

**Appendix D: Neglected processes**

**First reduced model**

Neglected processes are:  $f_{1,3}, f_{2,3}, f_{3,3}, f_{4,6}, f_{5,3}, f_{5,6}, f_{6,4}, f_{7,4}, f_{8,2}, f_{8,7}, f_{9,6}, f_{9,7}, f_{10,4}, f_{11,4}, f_{12,3}, f_{12,6}, f_{13,2}, f_{13,4}, f_{14,2}, f_{14,6}, f_{14,7}, f_{15,4}, f_{16,1}, f_{16,4}$ .

**Second reduced model: sub-models**

Neglected processes in SM1 are:  $f_{1,3}, f_{2,3}, f_{3,3}, f_{4,6}, f_{5,3}, f_{5,6}, f_{6,4}, f_{7,4}, f_{8,2}, f_{8,7}, f_{9,6}, f_{9,7}, f_{10,4}, f_{11,4}, f_{12,3}, f_{12,6}, f_{13,2}, f_{13,4}, f_{14,2}, f_{14,6}, f_{14,7}, f_{15,4}, f_{16,1}, f_{16,4}$ .

In SM2, we supposed that processes switching state from  $t_1 = 0.33$  until  $t_6 = 1.5$  change simultaneously at time  $t'_1 = 0.9$ . Deleted processes are common to those removed in SM1, as well as:  $f_{4,3}, f_{4,4}, f_{5,4}, f_{7,2}, f_{8,3}, f_{8,5}, f_{9,2}, f_{9,3}, f_{10,2}, f_{14,5}$ .

In SM3, we supposed that processes switching state from  $t_7 = 3.8$  until  $t_6 = 1.5$  change simultaneously at time

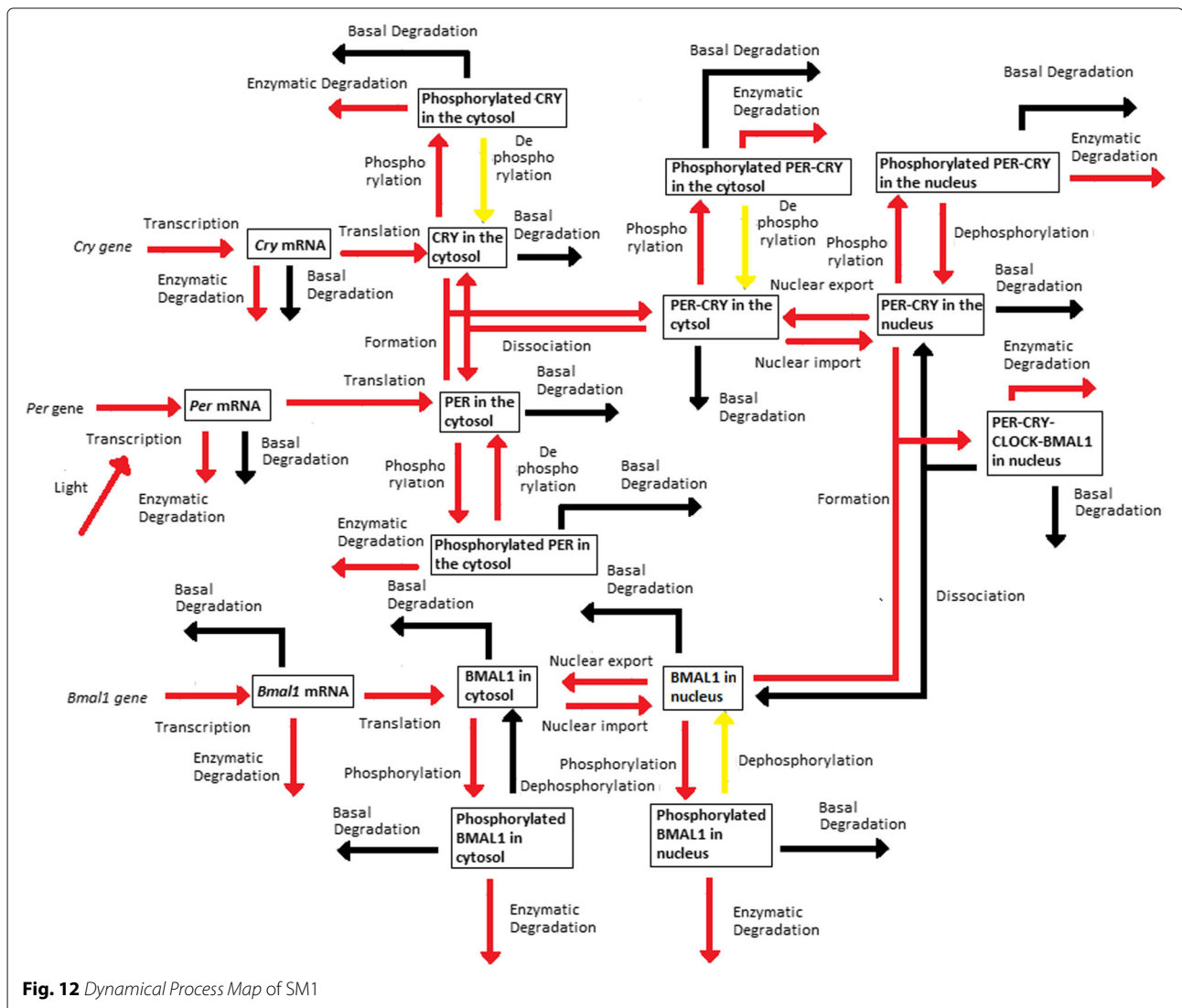
$t'_2 = 6$ . Deleted processes are common to those removed in SM1, as well as:  $f_{1,1}, f_{2,1}, f_{4,2}, f_{4,3}, f_{5,2}, f_{7,2}, f_{8,1}, f_{9,1}, f_{9,2}, f_{10,2}, f_{11,2}, f_{12,5}, f_{14,4}$ .

In SM4, we supposed that processes switching state from  $t_{16} = 9.8$  until  $t_{26} = 15.6$  change simultaneously at time  $t'_3 = 12.5$ . Deleted processes are common to those removed in SM1, as well as:  $f_{4,1}, f_{5,1}, f_{8,4}, f_{9,2}, f_{10,2}, f_{11,2}, f_{12,5}, f_{14,4}$ .

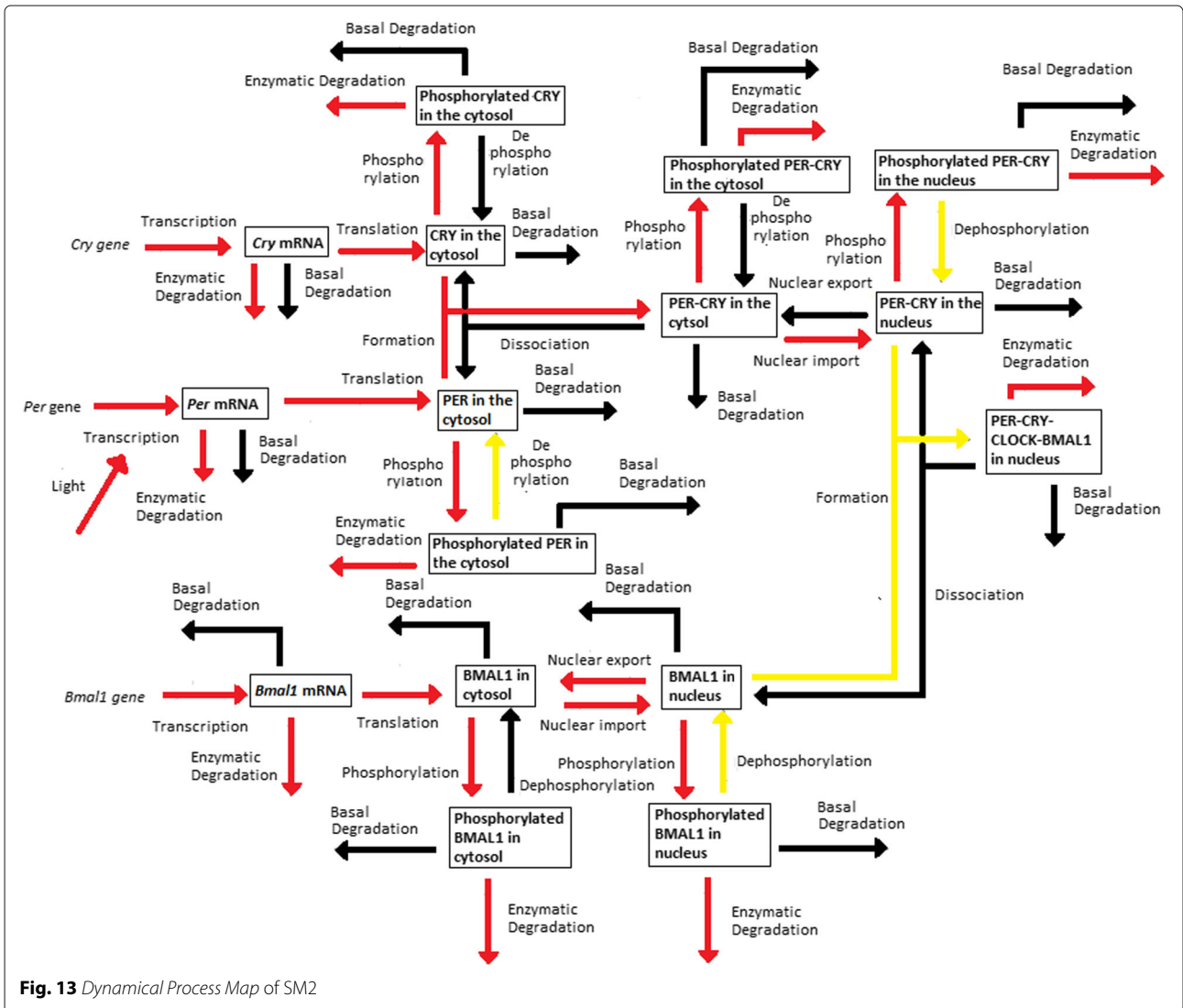
In SM5, we supposed that processes switching state from  $t_{27} = 17.3$  until  $t_{46} = 22.9$  change simultaneously at time  $t'_4 = 20.0$ . Deleted processes are common to those removed in SM1, as well as:  $f_{4,3}, f_{4,4}, f_{5,4}, f_{7,2}, f_{8,3}, f_{8,5}, f_{9,2}, f_{9,3}, f_{10,2}, f_{14,5}$ .

**Appendix E: Dynamical Process Maps**

See Figs. 12, 13, 14, 15 and 16.



**Fig. 12** Dynamical Process Map of SM1



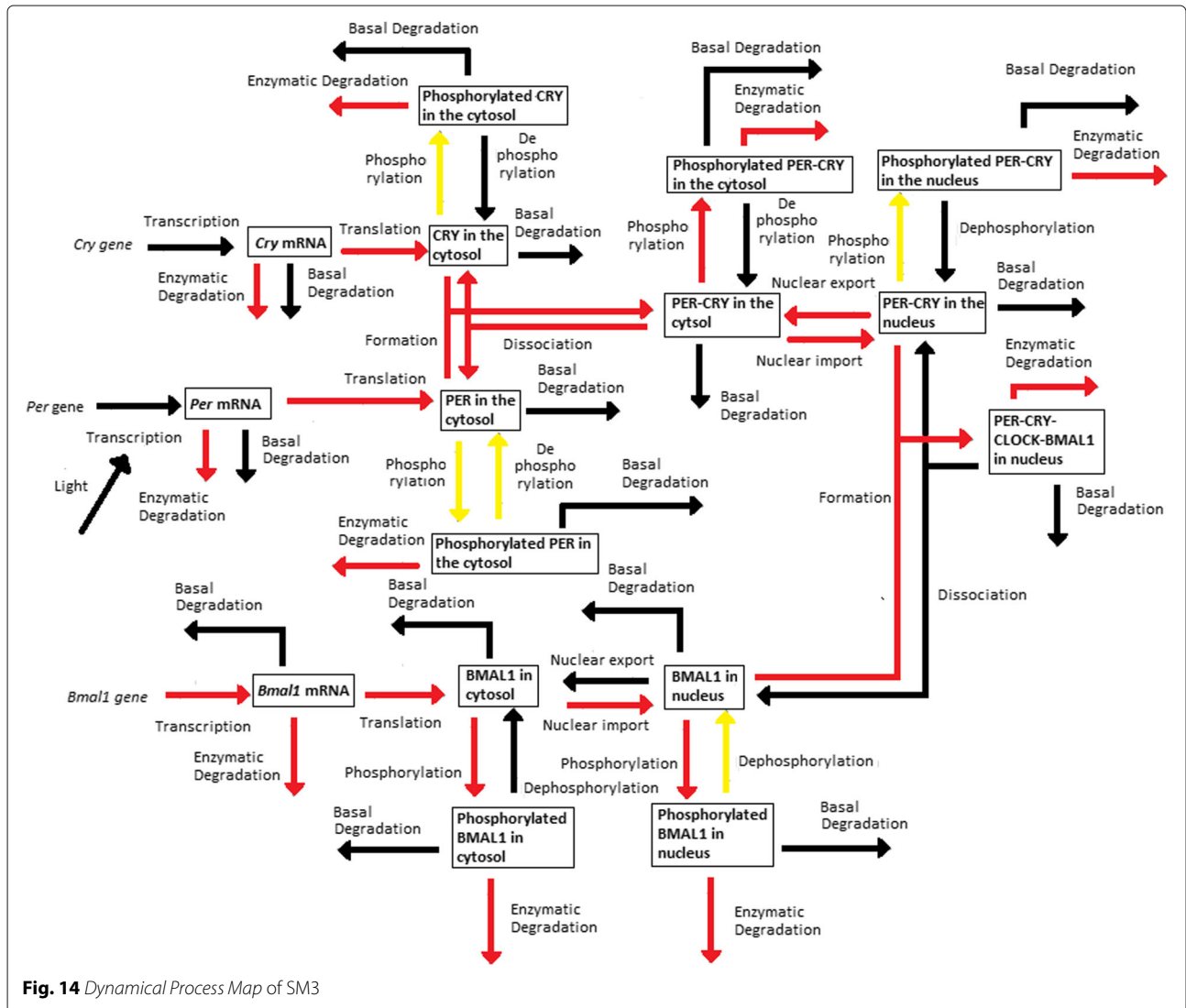
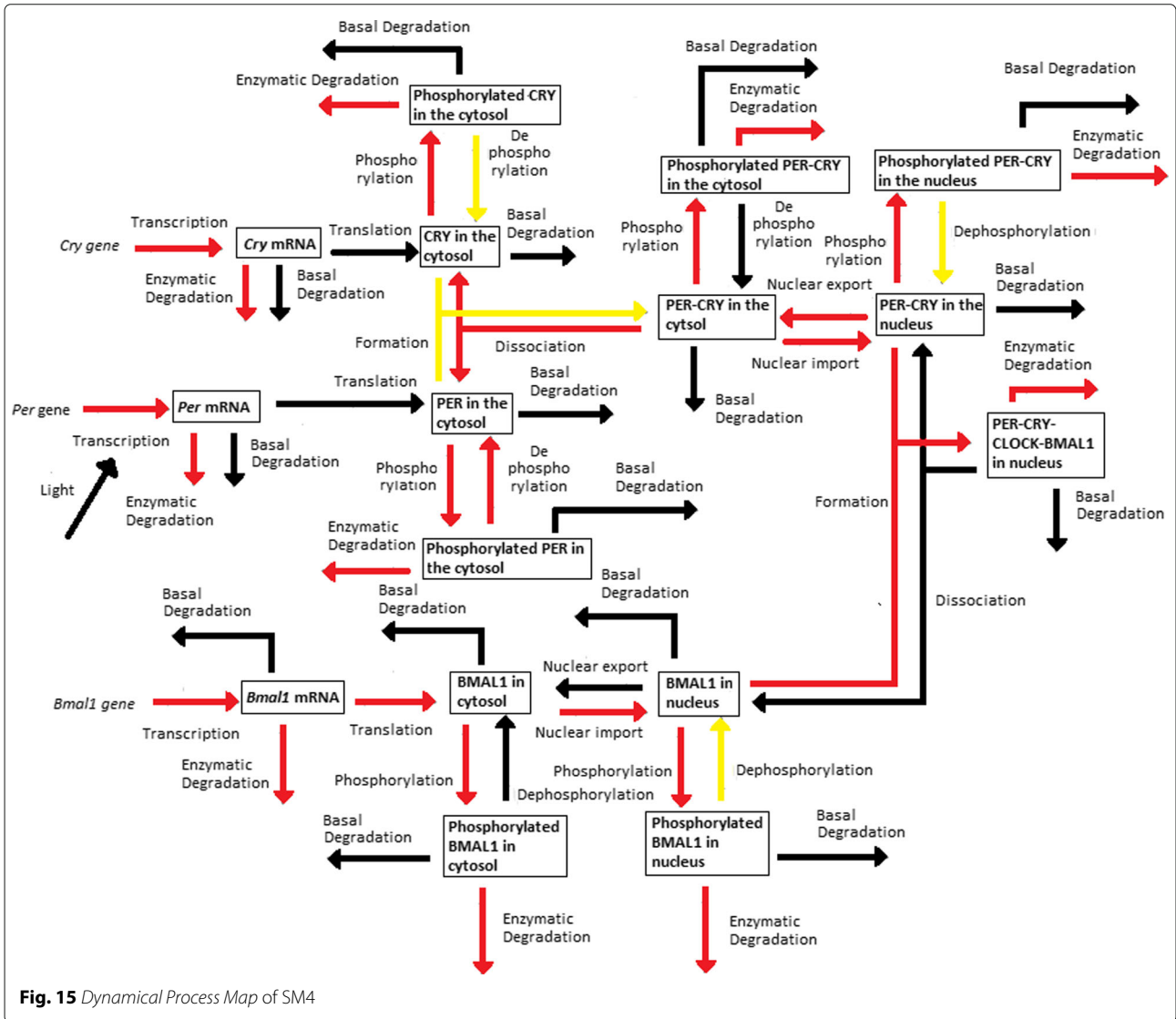


Fig. 14 Dynamical Process Map of SM3





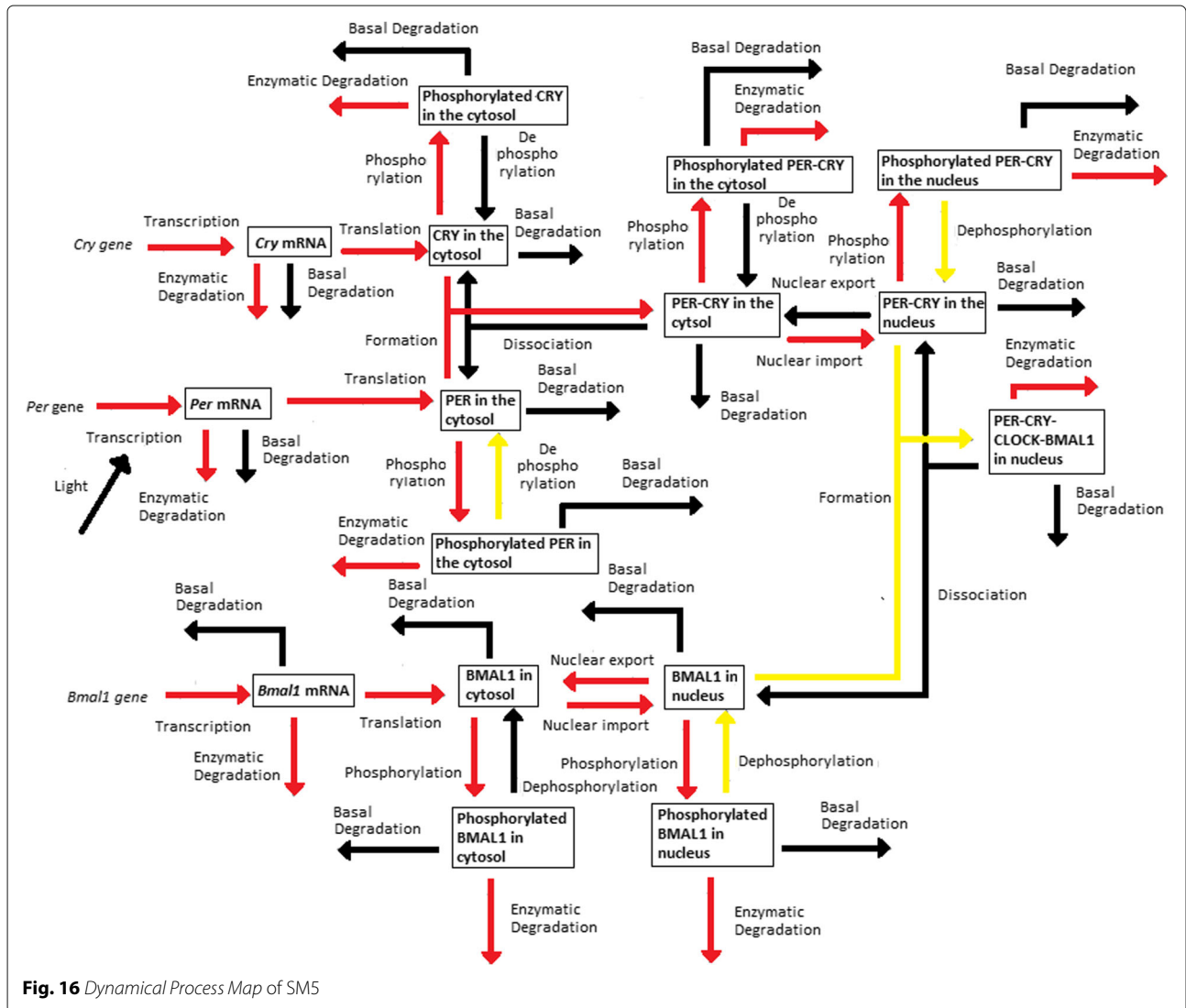


Fig. 16 Dynamical Process Map of SM5

### Acknowledgements

The authors would thank the research program LABEX SIGNALIFE (ANR-11-LABX-0028-01) and the anonymous reviewers for helpful comments.

### Funding

We acknowledge the *Conseil Régional PACA* and the *Investissements d'Avenir Bio-informatique programme* under project RESET (ANR-11-BINF-0005) for funding the PhD thesis of S. Casagrande.

### Availability of data and materials

All data generated or analysed during this study are included in the published article and the Appendix. Matlab scripts to run Principal Process Analysis on the circadian clock model are available at the following link: <http://www.sop.inria.fr/members/Jean-Luc.Gouze/BMBcode/code.zip>.

### Authors' contributions

SC, DR and JLG designed the study and developed the methodology. SC performed the analysis. ST contributed to parameter sensitivity analysis. All authors discussed the results and contributed to the final manuscript. All authors read and approved the final manuscript.

### Competing interests

The authors declare that they have no competing interests.

### Ethics approval and consent to participate

Not applicable.

### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

### Author details

<sup>1</sup>Université Côte d'Azur, Inria, INRA, CNRS, UPMC Univ Paris 06, Biocore team, Sophia Antipolis, France. <sup>2</sup>Univ. Grenoble Alpes, Inria, 38000 Grenoble, France. <sup>3</sup>Université Côte d'Azur, INRA, CNRS, ISA, Sophia Antipolis, France.

Received: 4 December 2017 Accepted: 15 May 2018

Published online: 14 June 2018

### References

- Bettenbrock K, Fischer S, Kremling A, Jahreis K, Sauter T, Gilles E-D. A quantitative approach to catabolite repression in *Escherichia coli*. *J Biol Chem*. 2006;281(5):2578–84.
- Kuepfer L, Peter M, Sauer U, Stelling J. Ensemble modeling for analysis of cell signaling dynamics. *Nat Biotechnol*. 2007;25(9):1001–6.
- Snowden TJ, van der Graaf PH, Tindall MJ. Methods of model reduction for large-scale biological systems: a survey of current methods and trends. *Bull Math Biol*. 2017;79(7):1449–86.
- Apri M, de Gee M, Molenaar J. Complexity reduction preserving dynamical behavior of biochemical networks. *J Theor Biol*. 2012;304:16–26.
- Petzold L, Zhu W. Model reduction for chemical kinetics: An optimization approach. *AIChE J*. 1999;45(4):869–86.
- Sunnåker M, Cedersund G, Jirstrand M. A method for zooming of nonlinear models of biochemical systems. *BMC Syst Biol*. 2011;5(1):140.
- Gorban AN, Karlin IV. Method of invariant manifold for chemical kinetics. *Chem Eng Sci*. 2003;58(21):4751–68.
- Anderson J, Chang Y-C, Papachristodoulou A. Model decomposition and reduction tools for large-scale networks in systems biology. *Automatica*. 2011;47(6):1165–74.
- Hangos KM, Gábor A, Szederkényi G. Model reduction in bio-chemical reaction networks with Michaelis-Menten kinetics. In: *Control Conference (ECC), 2013 European*. Zürich: IEEE; 2013. p. 4478–4483.
- Segel LA, Slemrod M. The quasi-steady-state assumption: a case study in perturbation. *SIAM Rev*. 1989;31(3):446–77.
- de Jong H, Gouzé J-L, Hernandez C, Page M, Sari T, Geiselman J. Qualitative simulation of genetic regulatory networks using piecewise-linear models. *Bull Math Biol*. 2004;66(2):301–40.
- Baldazzi V, Ropers D, Markowicz Y, Kahn D, Geiselman J, de Jong H. The carbon assimilation network in *Escherichia coli* is densely connected and largely sign-determined by directions of metabolic fluxes. *PLoS Comput Biol*. 2010;6(6):1000812.
- Bhattacharjee B, Schwer DA, Barton PI, Green WH. Optimally-reduced kinetic models: reaction elimination in large-scale kinetic mechanisms. *Combust Flame*. 2003;135(3):191–208.
- Casagrande S, Ropers D, Gouzé J-L. Model reduction and process analysis of biological models. In: *2015 23rd Mediterranean Conference on Control and Automation (MED)*. Torremolinos: IEEE; 2015. p. 1132–9.
- Leloup J-C, Goldbeter A. A model for circadian rhythms in *Drosophila* incorporating the formation of a complex between the PER and TIM proteins. *J Biol Rhythm*. 1998;13(1):70–87.
- Kwang-Hyun C, Sung-Young S, Hyun-Woo K, Wolkenhauer O, McFerran B, Kolch W. Mathematical modeling of the influence of RZIP on the ERK signaling pathway. In: Priami C, editor. *Computational Methods in Systems Biology*. Rovereto: Springer; 2003. p. 127–41.
- Leloup J-C, Goldbeter A. Modeling the mammalian circadian clock: sensitivity analysis and multiplicity of oscillatory mechanisms. *J Theor Biol*. 2004;230(4):541–62.
- Khalil HK. *Nonlinear Systems*, Second edn. New Jersey: Prentice Hall; 1996.
- Leloup J-C, Goldbeter A. Toward a detailed computational model for the mammalian circadian clock. *Proc Natl Acad Sci*. 2003;100(12):7051–6.
- Kanungo T, Mount DM, Netanyahu NS, Piatko CD, Silverman R, Wu AY. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Trans Pattern Anal Mach Intell*. 2002;24(7):881–92.
- Casagrande S, Gouzé J-L. Principal Process Analysis and reduction of biological models with order of magnitude. *IFAC-PapersOnLine*. 2017;50(1):12661–6.
- Kobilinsky A, Monod H, Bailey RA. Automatic generation of generalised regular factorial designs. *Comput Stat Data Anal*. 2017;113:311–29.
- Lamboni M, Monod H, Makowski D. Multivariate sensitivity analysis to measure global contribution of input factors in dynamic models. *Reliab Eng Syst Saf*. 2011;96(4):450–9.
- Box GE, Hunter JS. The 2 k-p fractional factorial designs. *Technometrics*. 1961;3(3):311–51.
- Zheng B, Albrecht U, Kaasik K, Sage M, Lu W, Vaishnav S, Li Q, Sun ZS, Eichele G, Bradley A, et al. Nonredundant roles of the *mPer1* and *mPer2* genes in the mammalian circadian clock. *Cell*. 2001;105(5):683–94.
- Van Der Horst GT, Muijtjens M, Kobayashi K, Takano R, Kanno S-i, Takao M, de Wit J, Verkerk A, Eker AP, van Leenen D, et al. Mammalian Cry1 and Cry2 are essential for maintenance of circadian rhythms. *Nature*. 1999;398(6728):627–630.
- Tyson JJ, Hong CI, Thron CD, Novak B. A simple model of circadian rhythms based on dimerization and proteolysis of PER and TIM. *Biophys J*. 1999;77(5):2411–7.
- Pagel H, Poll C, Ingwersen J, Kandeler E, Streck T. Modeling coupled pesticide degradation and organic matter turnover: From gene abundance to process rates. *Soil Biol Biochem*. 2016;103:349–64.
- Robles-Rodriguez C, Bideaux C, Guillolet S, Gorret N, Roux G, Molina-Jouve C, Aceves-Lara C. Multi-objective particle swarm optimization (MOPSO) of lipid accumulation in fed-batch cultures. In: *2016 24th Mediterranean Conference on Control and Automation (MED)*. Athens: IEEE; 2016. p. 979–984.