Bioinformatics, 38(1), 2022, 196–203
doi: 10.1093/bioinformatics/btab623
Advance Access Publication Date: 31 August 2021
Original Paper

OXFORD

Systems biology

Inferring causality in biological oscillators

Jonathan Tyler[1,2], Daniel Forger[1,3] and Jae Kyoung Kim [4,5,*]

[1]Department of Mathematics, University of Michigan, Ann Arbor, MI 48109, USA, [2]Department of Pediatrics, University of Michigan, Ann Arbor, MI 48109, USA, [3]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA, [4]Department of Mathematical Sciences, Korea Advanced Institute of Science and Technology, Daejeon 34141, Republic of Korea and [5]Biomedical Mathematics Group, Institute for Basic Science, Daejeon 34126, Republic of Korea

*To whom correspondence should be addressed.
Associate Editor: Anthony Mathelier

Received on April 28, 2021; revised on August 25, 2021; editorial decision on August 26, 2021; accepted on August 27, 2021

Abstract

Motivation: Fundamental to biological study is identifying regulatory interactions. The recent surge in time-series data collection in biology provides a unique opportunity to infer regulations computationally. However, when components oscillate, model-free inference methods, while easily implemented, struggle to distinguish periodic synchrony and causality. Alternatively, model-based methods test the reproducibility of time series given a specific model but require inefficient simulations and have limited applicability.

Results: We develop an inference method based on a general model of molecular, neuronal and ecological oscillatory systems that merges the advantages of both model-based and model-free methods, namely accuracy, broad applicability and usability. Our method successfully infers the positive and negative regulations within various oscillatory networks, e.g. the repressilator and a network of cofactors at the pS2 promoter, outperforming popular inference methods.

Availability and implementation: We provide a computational package, ION (Inferring Oscillatory Networks), that users can easily apply to noisy, oscillatory time series to uncover the mechanisms by which diverse systems generate oscillations. Accompanying MATLAB code under a BSD-style license and examples are available at https://github.com/Mathbiomed/ION. Additionally, the code is available under a CC-BY 4.0 License at https://doi.org/10.6084/m9.figshare.16431408.v1.

Contact: jaekkim@kaist.ac.kr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

A fundamental goal in biology is to uncover causal interactions. Conventional methods manipulate one or more components experimentally to investigate the effect on others in the system. However, these are time-consuming and costly, particularly as the number of components increases. On the other hand, recent technological advances (e.g. GFP, luciferase, microarray, etc.) continue to make measuring time-series data easier. Accordingly, inferring direct regulations solely given time-series data is crucial to revealing the mechanisms underlying systems in a timely and inexpensive manner (Saint-Antoine and Singh, 2020).

Various model-free methods have been widely used to infer interactions because they are easy to implement and broadly applicable (Casadiego *et al.*, 2017; Deyle and Sugihara, 2011; Deyle *et al.*, 2013, 2016; Granger, 1969; Leng *et al.*, 2020; Ma *et al.*, 2017; Pourzanjani *et al.*, 2015; Runge *et al.*, 2019; Stokes and Purdon, 2017; Sugihara *et al.*, 2012; Tani *et al.*, 2020; Tsonis *et al.*, 2015; Wang *et al.*, 2020; Ye *et al.*, 2015). A popular model-free method, Granger Causality (GC), uses predictability to infer interactions, i.e. X causes Y if X has unique information that improves the prediction of Y (Granger, 1969; Stokes and Purdon, 2017). However, GC relies heavily on the assumption that the time-series data are stationary (Lütkepohl, 2005) making it challenging to apply to oscillatory time-series data that are highly non-stationary (Abel *et al.*, 2016; Pourzanjani *et al.*, 2015; Stokes and Purdon, 2017; Yang *et al.*, 2018). To overcome this limitation, inference methods for dynamical systems, such as Convergent Cross Mapping (CCM), use a differing view of predictability to infer causality, i.e. X causes Y if historical values of X can be recovered from Y alone (Deyle *et al.*, 2013; 2016; Deyle and Sugihara, 2011; Leng *et al.*, 2020; Ma *et al.*, 2017; Runge *et al.*, 2019; Sugihara *et al.*, 2012; Tani *et al.*, 2020; Tsonis *et al.*, 2015; Wang *et al.*, 2020; Ye *et al.*, 2015). Despite the success of CCM methods, they struggle to differentiate synchrony (i.e. similar periods among components) versus causality, frequently resulting in an increase in false-positive inferences in oscillatory networks. This is problematic because biological processes frequently exhibit oscillatory behavior in time-series data, e.g. about half of the protein-coding genome is transcribed rhythmically (Mure *et al.*, 2018; Zhang *et al.*, 2014).

Alternatively, model-based methods infer causality by testing reproducibility of time-series data with mechanistic models. Although testing reproducibility requires computationally expensive model simulations and fittings (Balsa-Canto *et al.*, 2008; Firman *et al.*, 2019; Geva-Zatorsky *et al.*, 2010; Gotoh *et al.*, 2016; Lillacci and Khammash, 2010; McBride and Petzold, 2018; Mhaskar *et al.*, 2002; Pitt and Banga, 2019; Radde and Kaderali, 2009; Stražar *et al.*, 2014; Toni *et al.*, 2009; Trejo Banos *et al.*, 2015; Wang *et al.*, 2018; Wang and Enright, 2013), if the underlying model is accurate, model-based methods do not suffer from false positive predictions unlike model-free methods. However, the inference results strongly depend on the choice of model, often derived from limited information. Thus, inference methods using more general ODE forms were developed (Brunton *et al.*, 2016; Jensen *et al.*, 2009; Jo *et al.*, 2018; Kim and Forger, 2012b; Konopka, 2011; Konopka and Rooman, 2010; Mangan *et al.*, 2016; McGoff *et al.*, 2016; Pigolotti *et al.*, 2007, 2009). For example, previously, we developed a method that infers causation from X to Y by checking the reproducibility of oscillatory time-series data given a common ODE model: $\frac{dY}{dt} = f(X) - g(Y)$. Here, f and g describe the synthesis and degradation rates of Y, respectively (Kim and Forger, 2012b). Pigolotti *et al.* (2007) considered the most general possible mechanistic model between two components:

$$\frac{dY}{dt} = f(X, Y). \tag{1}$$

However, unlike Kim and Forger (2012b), this method uses only the minima and maxima of the time-series data (Pigolotti *et al.*, 2007), thus requiring the restrictive assumption that all given components are in a single negative feedback loop. Moreover, extensions of the method require that a single negative feedback loop structure drive the dynamics, limiting their applicability (Jensen *et al.*, 2009; Pigolotti *et al.*, 2009).

Here, we develop an inference method for biological oscillators described by Equation (1) that is easy to implement, broadly applicable and accurate, while also computationally efficient. Specifically, we identify a fundamental relationship between the general model (Equation 1) and its oscillatory solution. Using this relationship, we develop a functional transformation (i.e. regulation-detection function) of a pair of oscillatory time-series data that easily tests the reproducibility of the time series with the general model. This transformation enables accurate and precise inference of the (self-) regulation type (e.g. positive, negative or a mixture) between two components X *and* Y described by Equation (1). Our method infers regulations within various network structures such as a cycle, multiple cycles and a cycle with outputs from *in silico* oscillatory time-series data. Furthermore, our method successfully infers regulation types from noisy experimental data measured at the molecular and organismal levels. In particular, from time-series data of the repressilator and cofactors at the pS2 promoter, our method infers networks that match current biological knowledge while popular model-free methods incorrectly infer nearly fully connected networks. Importantly, our method predicts hidden regulations for the pS2 promoter after estradiol treatment, guiding experimental investigation. Finally, we provide a user-friendly computational package (ION: Inferring Oscillatory Networks) that implements our method to infer network structures of biological oscillators.

2 Results

2.1 Inferring regulation types from oscillatory time series

The reduced FitzHugh-Nagumo model (Fig. 1A) (FitzHugh, 1961) describes the interactions between the membrane potential of a neuron (V) and the accommodation and refractoriness of the membrane (W) (FitzHugh, 1961; Nagumo *et al.*, 1962). In particular, W positively regulates V while V negatively regulates W. In addition, V displays a mixture of positive and negative self-regulation while W negatively regulates itself.

How are such inter- and self-regulations reflected in the oscillatory change of V and W simulated with the model (Fig. 1B)? Notably, the changes in V and W do not directly reflect their regulatory interactions. For instance, although W positively regulates V, when W increases, V does not always increase (e.g. in the yellow region, Fig. 1B). This is because W positively regulates $\dot{V}$ rather than V (Fig. 1A). However, the relationship between the change in W and $\dot{V}$ also does not reflect the positive regulation of W on V. For example, in the yellow region (Fig. 1B), $\dot{V}$ decreases despite increasing W because the self-regulation of V on $\dot{V}$ masks the effect of W on $\dot{V}$. Thus, to infer the effect of W on $\dot{V}$ independently of V, we investigate the relationship between W and $\dot{V}$ at time points t and the *reflection time*, t_V, where $V(t) = V(t_V)$ (Fig. 1B). Since $V(t) = V(t_V)$, the difference $\dot{V}(t) - \dot{V}(t_V) = f(V(t), W(t)) - f(V(t_V), W(t_V))$ is solely determined by W. Thus, because W positively regulates V (Fig. 1A), if $W(t)$ is greater (less) than $W(t_V)$, $\dot{V}(t)$ should be greater (less) than $\dot{V}(t_V)$. Similarly, to infer the type of self-regulation of V, we must remove the variation of $\dot{V}$ due to W that masks the effect of V on $\dot{V}$. Thus, we investigate the relationship between V and $\dot{V}$ at time points t and the reflection time, t_W, where $W(t) = W(t_W)$ (Fig. 1B). To quantify such relationships between W and $\dot{V}$ and V and $\dot{V}$, we develop the *regulation-detection functions*:

$$\begin{aligned} R^{t_V}_{W\to V}(t) &:= (W(t) - W(t_V)) \cdot (\dot{V}(t) - \dot{V}(t_V)) \\ &:= W^{t_V}_d(t) \cdot \dot{V}^{t_V}_d(t), \end{aligned} \tag{2}$$

and

$$\begin{aligned} R^{t_W}_{V\to V}(t) &:= (V(t) - V(t_W)) \cdot (\dot{V}(t) - \dot{V}(t_W)) \\ &:= V^{t_W}_d(t) \cdot \dot{V}^{t_W}_d(t). \end{aligned} \tag{3}$$

As W positively regulates V, the functions $W^{t_V}_d$ and $\dot{V}^{t_V}_d$ have the same sign and thus, $R^{t_V}_{W\to V}(t) \geq 0$ throughout the cycle (Fig. 1C, left). That is, if $W^{t_V}_d = W(t) - W(t_V) \geq 0$, then $\dot{V}^{t_V}_d = \dot{V}(t) - \dot{V}(t_V) = 3(W(t) - W(t_V)) \geq 0$ (Fig. 1A). On the other hand, due to the mixed self-regulation of V, the relationship between the signs of $V^{t_W}_d(t)$ and $\dot{V}^{t_W}_d(t)$, and thus the sign of $R^{t_W}_{V\to V}(t)$, varies throughout the cycle (Fig. 1C, right).

As the profiles of the sign of the regulation-detection functions (Equations 2 and 3) reflect the regulation type, we develop a *regulation-detection score* that quantifies the variation in the sign of the regulation-detection functions. For instance, the regulation-detection score for the regulation of W on V is defined as

$$\begin{aligned} \langle R_{W\to V}\rangle &:= \frac{\int_0^\tau R^{t_V}_{W\to V}(t)dt}{\int_0^\tau |R^{t_V}_{W\to V}(t)|dt} \\ &= \frac{\text{Positive Area}}{\text{Total Area}} - \frac{\text{Negative Area}}{\text{Total Area}}, \end{aligned} \tag{4}$$

where τ is the period (e.g. $\tau = 1$ in Fig. 1C, left). The regulation-detection score $\langle R_{W\to V}\rangle = 1$ because W positively regulates V, and thus $R^V_{W\to V}(t) \geq 0$ (i.e. the negative area is zero) (Fig. 1C, left). On the other hand, because V both positively and negatively regulates itself, $R^{t_W}_{V\to V}(t)$ takes both positive and negative values, so $\langle R_{V\to V}\rangle = 0.6 - 0.4 = 0.2$ (Fig. 1C, right).

Similarly, we can obtain information about the regulation of V on W and the self regulation of W with the regulation-detection functions $R^{t_W}_{V\to W} := V^{t_W}_d \cdot \dot{W}^{t_W}_d(t)$ (Fig. 1D, left) and $R^{t_V}_{W\to W} := W^{t_V}_d \cdot \dot{W}^{t_V}_d(t)$ (Fig. 1D, right). Because V negatively regulates W, $R^{t_W}_{V\to W}(t) \leq 0$. Also, because the self-regulation of W is purely negative, $R^{t_V}_{W\to W}(t) \leq 0$. Thus, $\langle R_{V\to W}\rangle = -1$, and $\langle R_{W\to W}\rangle = -1$ (Fig. 1D). Taken together, in general, if X positively (negatively) regulates Y, then $\langle R_{X\to Y}\rangle = 1$ ($\langle R_{X\to Y}\rangle = -1$) (see Supplementary Theorem S1 in Supplementary Information).

Next, we calculated the regulation-detection scores from experimentally measured oscillatory time-series data of two bacteria: *Paramecium* (P) and *Didinium* (D) (Fig. 1E) (Veilleux, 1976). As P is a prey of the predator D (Veilleux, 1976), D is expected to negatively regulate P and P is expected to positively regulate D. Reflecting this, $\langle R_{P\to D}\rangle = 1$ and $\langle R_{D\to P}\rangle = -1$ (Fig. 1E). Furthermore, reflecting the positive (i.e. birth) and negative (i.e. death) self-regulation of both P and D, $\langle R_{D\to D}\rangle = 0.51 - 0.49 = 0.02$ and $\langle R_{P\to P}\rangle = 0.63 - 0.37 =$

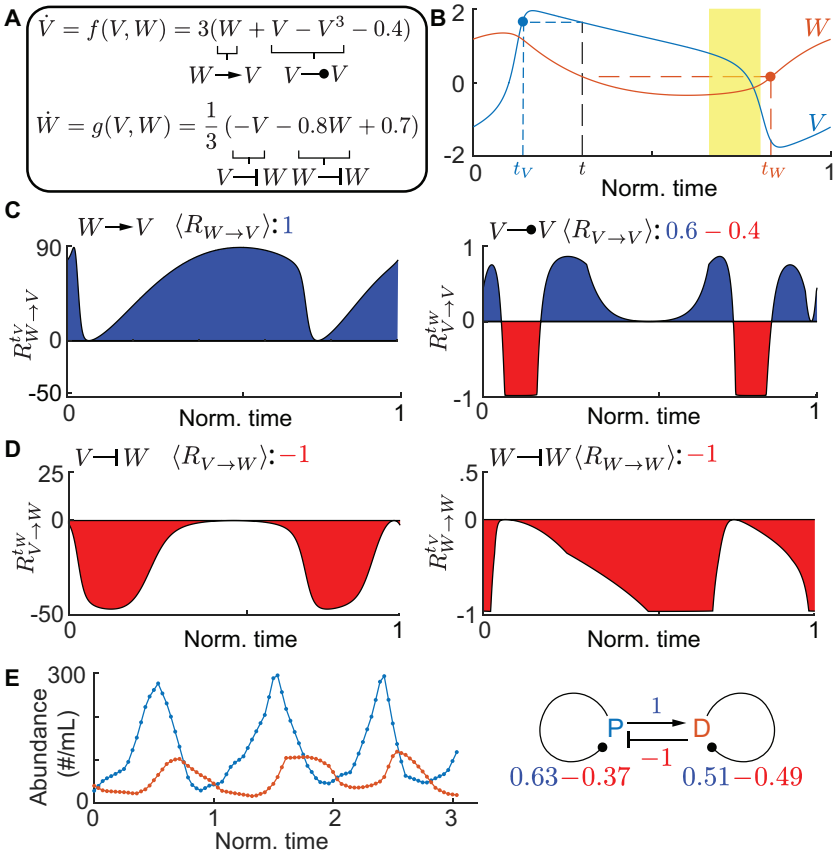

Fig. 1. Regulation-detection functions and scores reflect regulation types. (**A**) The FitzHugh-Nagumo model describes the interactions between the membrane potential of a neuron (V) and the accommodation and refractoriness of the membrane (W). W positively regulates V while V negatively regulates W. In addition, V displays a mixture of positive and negative self-regulation while W negatively regulates itself. (**B**) Time series of one cycle simulated with the FitzHugh-Nagumo model (A). Although W positively regulates V (i.e. $\dot{V}$ positively depends on W), $\dot{V}$ decreases despite increasing W (yellow region) because the self-regulation of V on $\dot{V}$ masks the effect of W on $\dot{V}$. On the other hand, for the time points t and reflection time t_V where $V(t) = V(t_V)$, if $W(t)$ is greater (less) than $W(t_V)$, then $\dot{V}(t)$ should be greater (less) than $\dot{V}(t_V)$. Similarly, as V negatively regulates W, if $V(t)$ is greater (less) than $V(t_W)$, then $\dot{W}(t)$ should be less (greater) than $\dot{W}(t_W)$ for the time points t and t_W, where $W(t) = W(t_W)$. (**C**) The regulation-detection function $R^{t_V}_{W\rightarrow V}$ (Equation 2) is positive, and thus the regulation-detection score $\langle R_{W\rightarrow V}\rangle$ (Equation 4) equals one, reflecting the positive regulation of W on V. The sign of $R^{t_W}_{V\rightarrow V}(t)$ (Equation 3) changes, and thus $-1 < \langle R_{V\rightarrow V}\rangle < 1$ (Equation 4), reflecting the mixture of positive and negative self-regulation of V. (**D**) Both $R^{t_W}_{V\rightarrow W}$ and $R^{t_V}_{W\rightarrow W}$ are negative, and thus $\langle R_{V\rightarrow W}\rangle = \langle R_{W\rightarrow W}\rangle = -1$, reflecting the negative regulation of V on W and the self-regulation of W. (**E**) Regulation-detection scores are calculated from the time-series population data of two bacteria: *Paramecium*, P (blue) and *Didinium*, D (red) (data taken from (Sugihara *et al.*, 2012)). Reflecting the known predatory interaction, $\langle R_{P\rightarrow D}\rangle = 1$ and $\langle R_{D\rightarrow P}\rangle = -1$. Furthermore, reflecting that the self-regulation of both P and D consists of positive (i.e. birth) and negative (i.e. death) regulation, $\langle R_{P\rightarrow P}\rangle = 0.26$ and $\langle R_{D\rightarrow D}\rangle = 0.02$(Color version of this figure is available at *Bioinformatics* online.)

0.26 (Fig. 1E). The regulation-detection scores appear to accurately reflect regulation types even for noisy and discrete time-series data.

2.2 Regulation inference method from oscillatory time series

If X positively (negatively) regulates Y, then the reflection score $\langle R_{X\rightarrow Y}\rangle = 1$ (resp., –1). That is, $-1 < \langle R_{X\rightarrow Y}\rangle < 1$ indicates either mixed regulation or the absence of regulation. Thus, when interactions are not mixed (i.e. monotonic), such as gene regulation by a transcription factor or predator-prey relationships, $-1 < \langle R_{X\rightarrow Y}\rangle < 1$ indicates the absence of regulation. This can be used to infer regulations from time-series data, as positive or negative regulation is present only when $\langle R_{X\rightarrow Y}\rangle = 1$ or –1, respectively. Similarly, self-regulation, which is either positive or negative, is possible only when $\langle R_{Y\rightarrow Y}\rangle = 1$ or –1. However, since depletion of a component typically increases as its own concentration increases, self-regulation can be assumed to be negative (i.e. $\langle R_{Y\rightarrow Y}\rangle = -1$). In this case, positive or negative regulation from X to Y is possible only when $\vec{R} = (\langle R_{X\rightarrow Y}\rangle, \langle R_{Y\rightarrow Y}\rangle) = (1,-1)$ or $(-1,-1)$, and thus, $\vec{R} \neq (\pm 1,-1)$ indicates the absence of regulation (**Rule 1**, Fig. 2A). Furthermore, we use $\vec{R} = (1,-1)$ or $(-1,-1)$ to infer positive or negative regulation (**Rules 2 and 3**, Fig. 2A). Note that, if positive or mixed self-regulation is possible, as in Figure 1E, Rules 2 and 3 can be relaxed to $\langle R_{X\rightarrow Y}\rangle = 1$ and $\langle R_{X\rightarrow Y}\rangle = -1$, respectively.

We illustrate how the three rules (Fig. 2A) infer regulations using as an example the Kim-Forger model (Fig. 2B), a simple model describing the mammalian circadian clock (Kim, 2016; Kim and Forger, 2012a). To infer the network structure (Fig. 2B, bottom) from the time series (Fig. 2B, top), we compute $\vec{R}$ for each possible interaction and self-regulation pair (Fig. 2B, box). Using Rule 1, three regulations are inferred as absent (Fig. 2B, box). Furthermore, Rules 2 and 3 identify the two positive regulations ($M \rightarrow P_C$ and $P_C \rightarrow P$) and the one negative regulation ($P \dashv M$), which have $\vec{R} = (1,-1)$ and $\vec{R} = (-1,-1)$, respectively (Fig. 2B, box). This successfully infers the negative feedback loop structure (Fig. 2B, bottom). Our method also successfully infers regulations in the *Frzilator* negative feedback loop (Igoshin *et al.*, 2004) (Fig. 2C and Supplementary Table S1) and a 4-state Goodwin oscillator (Goodwin, 1965) (Fig. 2D and Supplementary Table S2).

In fact, the order of peaks and nadirs of the time series in single feedback loops matches the order of regulation in the feedback loop (Fig. 2B–D). For instance, the peak of M is followed by the peaks of P_C and then P (Fig. 2B). This property has been used in previous

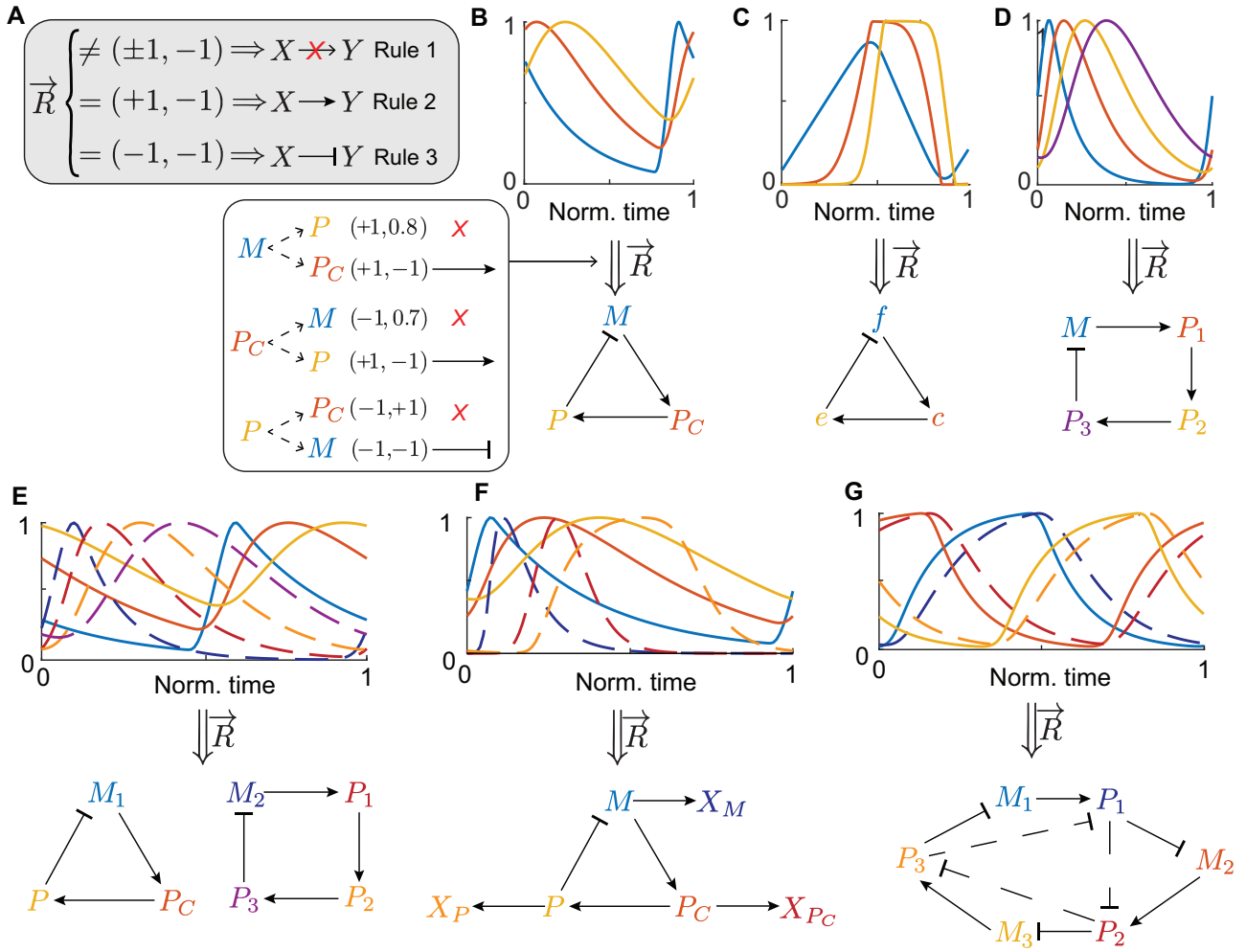

Fig. 2. The inference method successfully infers regulations within various *in silico* network structures. (**A**) The three rules for regulation inference. $\vec{R} \neq (\pm 1,-1)$ indicates the absence of regulation and $\vec{R} = (1,-1)$ or $(-1,-1)$ indicates positive or negative regulation. (**B**) The three rules successfully infer the network structure of the Kim-Forger model from simulated time-series data. According to Rule 1, the three regulations $M \rightarrow P$, $P_C \rightarrow M$ and $P \rightarrow P_C$ are inferred as absent. According to Rules 2 and 3, the two positive regulations ($M \rightarrow P_C$ and $P_C \rightarrow P$) and the one negative regulation ($P \dashv M$), which have $\vec{R} = (1,-1)$ and $\vec{R} = (-1,-1)$, are inferred. (**C**,**D**) Our inference method also successfully infers the negative feedback loop of the Frzilator (C) and a 4-state Goodwin oscillator (D). (**E**,**F**) Our inference method also successfully infers correct regulations for more challenging cases beyond the single feedback loop structure, i.e. the mixture of the Kim-Forger and Goodwin models (E) and an extended Kim-Forger model with output variables (F). (**G**) Our method also successfully infers regulations (solid arrows) of the repressilator from its three mRNA (solid lines) and three protein time-series data (dashed lines). However, our method also falsely predicts negative regulations among the proteins (dashed arrows) due to the similar time series between an mRNA and its protein (e.g. M_1 and P_1). See Supplementary Tables S1–S5 for the complete list of regulation-detection scores for (C)–(G) and Supplementary Section S4 in Supplementary Information for the equations and parameters used to simulate the data

algorithms to infer single negative feedback loop structures (Jensen *et al.*, 2009; Pigolotti *et al.*, 2007, 2009). However, since experimental datasets often contain components from more than one system, we test our method on a more challenging case when data are merged from the Kim-Forger (Fig. 2E, top; solid lines) and Goodwin (Fig. 2E, top; dashed lines) models. If only the order of peaks is used for this example, a single negative feedback loop with seven components is inferred whereas our method successfully infers the two independent underlying networks (Fig. 2E, bottom and Supplementary Table S3). Moreover, our inference method also successfully infers a cyclic network with output variables, also not adhering to the single feedback loop structure (Fig. 2F and Supplementary Table S4).

While our method successfully infers regulations within various networks, we caution that $\vec{R} = (\pm 1,-1)$ can occur even in the absence of regulation, making some correct interactions difficult to distinguish. For example, in the original repressilator model (Fig. 2G, top) (Elowitz and Leibler, 2000; Potvin-Trottier *et al.*, 2016), the mRNA and protein time series are so similar in phase (i.e. the phase difference is only 2.4% of the total period) that our method, along with inferring the actual interactions, predicts spurious interactions from one protein to the next protein. Thus, we advise users to check for nearly identical time series, which may increase false-positive inferences in our method as well as other inference methods.

2.3 Robustness of the inference method to interpolation error and noise

Experimentally measured time-series data are sampled discretely, in which case our method uses interpolation to generate continuous data (see Section 4.1.1). Accordingly, we test how sensitive our method is to interpolation error, specifically after linear interpolation, by using the *in silico* datasets in Figures 2B–F. That is, by decreasing the points measured per period from 10^2 to 10^1 (i.e. increasing the interpolation error), we quantify the accuracy of our inference method with the F_1 score, i.e. the harmonic mean of precision and recall (Fig. 3A). $F_1 = 1$ and $F_1 = 0$ indicate perfect recovery of all regulations and absence of any correct regulations, respectively. To account for interpolation error, we accept interactions based on three thresholds for $\langle R \rangle$ values: 0.99, 0.95 and 0.90. For example, a threshold of 0.99 means that we accept any interaction that

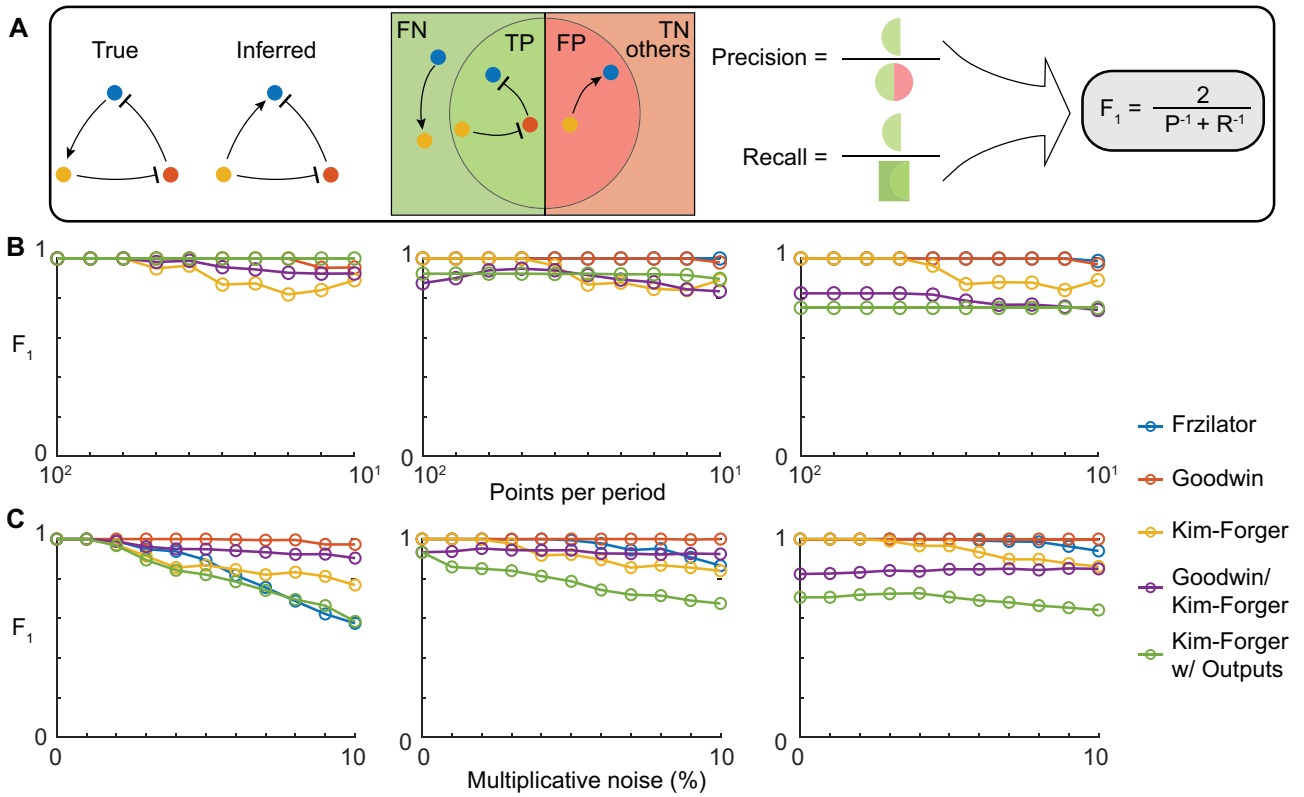

Fig. 3. Our regulation inference method is robust to interpolation error and to noise. (**A**) The illustration of F_1 calculation with a simple network (true; left, inferred; right). Precision is the number of correctly inferred regulations (TP-true positives) divided by the number of all inferred regulations (TP+FP-false positives), including those inferred incorrectly (FP). Recall is the number of correctly inferred regulations (TP) divided by the number of all true regulations (TP + FN-false negatives). Then, the F_1 score is the harmonic mean of precision and recall, which we use to measure the accuracy of our inference method. $F_1 = 1$ and 0 indicate perfect recovery of all regulations and the absence of any correct regulations, respectively. (**B**,**C**) The accuracy of our inference method when the number of points measured (B) and the level of noise (C) vary. Here, the number of points measured per period decreases from 10^2 to 10^1 (B) and the multiplicative noise increases from 0 to 10%, which is sampled from $N(0, 0.1^2)$ (C). The mean of the F_1 score for 100 different time series, which are generated with randomly chosen phases (B) and noise levels (C), is plotted (Section 4.3). Different thresholds for $\langle R \rangle$, 0.99 (left), 0.95 (middle) and 0.90 (right), are used

satisfies both $|\langle R_{X\to Y}\rangle| > 0.99$ and $\langle R_{Y\to Y}\rangle < -0.99$. We run our method beginning at 100 randomly selected times (Section 4.3). Then, we investigate how the mean of the distribution of F_1 scores changes as the sampling rate decreases (Fig. 3B). For single negative feedback loops (i.e. Frzilator, Goodwin, Kim-Forger), our method accurately recovers the network even when the number of data points measured per period is relatively low, e.g. ten per cycle. For the more complicated models (i.e. the merged Goodwin and Kim-Forger and the Kim-Forger with outputs models), slightly more data points are required for inference at high accuracy. Furthermore, our method shows similar robustness across the three thresholds, especially when the points sampled are toward the lower end.

Next, because experimental data are noisy, we increase the level of the multiplicative noise added to the dataset from 0 (no noise) to 10% multiplicative noise (sampled from $N(0, 0.1^2)$). The F_1 scores tend to decrease, but the decrease occurs more dramatically when the threshold is 0.99, indicating that the high threshold leads to higher sensitivity to noise. Moreover, this decrease in F_1 scores with the threshold of 0.99 is a result of an increase in false negatives (i.e. the exclusion of true interactions due to noise). Thus, we use a threshold of 0.9 when applying our inference method to experimental data (see below) as it leads to the most accurate results in the presence of noise (Fig. 3C). While 0.9 is recommended, depending on the weight of either avoiding false-positive or false-negative predictions, users can adjust the threshold when using our computational package, Inferring Oscillatory Networks (ION) (Fig. 4A; see Supplementary Information and Supplementary Figs S1 and S2 for a step-by-step manual). See Supplementary Information for details about how to choose the threshold.

2.4 Successful inferences from experimentally measured time series

As our inference method (ION) is quite robust to discrete data sampling and noise, we expect that our inference method can accurately infer regulations from experimentally measured time series as well. Indeed, our method successfully infers a three-gene repressilator network from experimental data of the three proteins (Potvin-Trottier *et al.*, 2016) (Fig. 4B and Supplementary Table S6). Note that our method recovers the repressilator network despite the absence of mRNA data because the shape and phase of the mRNA and protein profiles are expected to be similar, as in Figure 2G, due to the short translation time in *E.coli* relative to the period (Choi *et al.*, 2020). This indicates that our method also infers indirect regulation with a short time delay. Moreover, we compare our method with two popular model-free inference methods, PCM (Leng *et al.*, 2020) and GC (Granger, 1969) (Fig. 4B). As these methods can only infer the presence of regulation, not its type (i.e. positive and negative), unlike our method, the arrows represent inferred regulations, which could be either positive or negative. PCM recovers two correct regulations, $P_2 \to P_1$ and $P_3 \to P_2$, but fails to recover the regulation $P_1 \to P_3$ and makes two false-positive predictions, $P_1 \to P_2$ and $P_3 \to P_1$ (Fig. 4B, middle). While the GC infers all existing regulations, it makes two additional false-positive predictions, $P_1 \to P_2$ and $P_2 \to P_3$ (Fig. 4B, right). Even for this simple three-node network, the popular model-free inference methods make false-positive predictions because the network components oscillate at the same period (Cobey and Baskerville, 2016).

Next, we consider a more challenging case: the combination of two copies of the dataset in Figure 4B, one at the original phase and one with shifted phase (Fig. 4C and Supplementary Table S7). Our method successfully infers two repressilator networks, whereas PCM

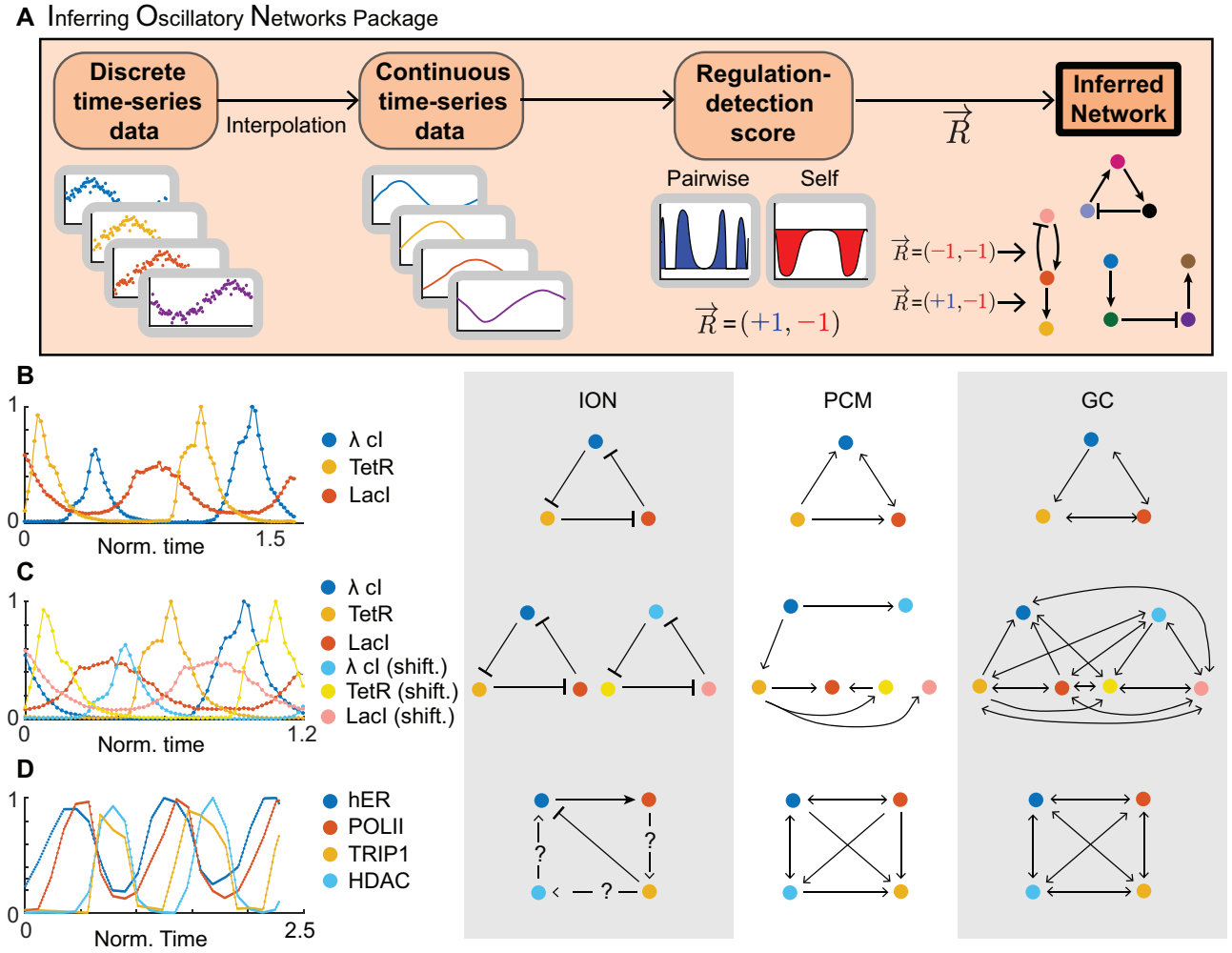

Fig. 4. The computational package ION successfully infers networks of repressilator and pS2 promoter cofactors. (**A**) ION interpolates and smooths noisy and discrete time-series data and then computes the pairwise and self regulation-detection functions and scores ($\vec{R}$) for every pair of components. If $\vec{R}$ satisfies Rules 2 or 3 (Fig. 2A), which can be relaxed depending on the threshold specified by the user, positive or negative regulation is assigned. See Supplementary Information for a comprehensive manual. (**B**) Using experimentally measured oscillatory time series from (Potvin-Trottier *et al.*, 2016), ION successfully infers a three-gene repressilator network structure (see Supplementary Fig. S3 for details). On the other hand, two popular model-free inference methods, PCM and GC, infer several false-positive regulations (e.g. λ cI regulates LacI). (**C**) ION also successfully infers two independent cycle when the experimental repressilator dataset from (A) is duplicated and the phase is shifted by about half of the period (see Supplementary Fig. S3 for details). However, PCM and GC fail to infer independent cycles due to false-positive predictions. (**D**) ION infers two direct regulations and predict three hidden regulations among cofactors at the estrogen-sensitive pS2 promoter after estradiol treatment (Métivier *et al.*, 2003) (see Supplementary Fig. S3 for details)

and GC suffer from several spurious regulations (4 and 15 false-positive interactions in Fig. 4B and C, respectively). Note that, even though we are using the same repressilator dataset, there are inconsistencies in the PCM and GC results compared with those from Figure 4B. These inconsistencies are a consequence of the shortened length of data used in Figure 4C compared with that in Figure 4B. This indicates that, in addition to the risk of false-positive inference, PCM and GC are sensitive to the amount of data, unlike ours.

For time series measuring the amount of cofactors present at the estrogen-sensitive pS2 promoter after treatment with estradiol [data from Métivier *et al.* (2003) and Lemaire *et al.* (2006)], PCM and GC infer an almost fully connected network and a fully connected network, respectively (Fig. 4D). On the other hand, our method only infers two regulations, both supported by the current biological understanding of the system. That is, estradiol triggers the binding of human ERα (hER) to the pS2 promoter to recruit RNA Polymerase II, supporting the inferred positive regulation of POLII by hER. Furthermore, TRIP1 acts as a surrogate measure for the 20S proteasome (APIS), which promotes proteasome-mediated degradation of hER (Métivier *et al.*, 2003), supporting the inferred negative regulation of hER by TRIP1. However, the inferred network (Fig. 4D, Supplementary Table S8) does not contain a negative feedback loop, which is required to generate sustained oscillations (Novák and Tyson, 2008). Thus, there may be intermediate steps between POLII and TRIP1, TRIP1 and HDAC, and also HDAC and hER that form the negative feedback loop (Fig. 4D; question marks). Altogether, this illustrates that our method can identify direct regulations while highlighting connections that require further experimental investigation.

3 Discussion

We developed a model-based method that infers regulations within networks underlying biological oscillators. The method identifies positive or negative regulation by efficiently testing the reproducibility of time-series data given Equation (1). Our method successfully inferred several networks such as single cycles (e.g. repressilator), two independent cycles and a cycle structure with outputs. Importantly, our method can distinguish direct versus indirect regulations, unlike GC and CCM (Fig. 4) (Leng *et al.*, 2020). That is, when $X \to Y \to Z$, our method typically infers $X \to Y$, not $X \to Z$ (Fig. 2B–F). However, if $Y \to Z$ is fast and thus Y and Z oscillate with nearly identical phases, our method infers $X \to Z$ as well

(Fig. 2G). Thus, if networks contain hidden steps with fast time scales (i.e. short time delays), our method may infer additional indirect regulations. Furthermore, we provide a user-friendly computational package, ION, that infers regulations within biological networks that oscillate from the molecular to the population level. When our method is coupled with evolving experimental time series, it can uncover unknown functional relationships and mechanisms that drive oscillatory behavior in biological systems.

Our method merges the advantages of model-based and model-free methods while mitigating the drawbacks of each. In particular, our model-based inference method does not suffer from the serious risk of false-positive prediction for biological oscillators or sensitivity to the amount of data, unlike the previous model-free inference methods such as GC and PCM (Fig. 4). However, as our inference method is model-based, it runs the risk that the imposed ODE model and functional relationships create false representations of the true interactions (Cobey and Baskerville, 2016). Our method minimizes this risk by using the most general form of an ODE Equation (1) to model the change in a component that is acted upon by another component and itself. In this way, we resolve the limitations of previous model-based methods that restricted the class of models, such as separable synthesis and degradation functions (Jo *et al.*, 2018; Kim and Forger, 2012b; Konopka and Rooman, 2010), specific types of functions (e.g. power or Hill functions) (Gotoh *et al.*, 2016; Konopka and Rooman, 2010) and a single feedback loop (Jensen *et al.*, 2009; Pigolotti *et al.*, 2007, 2009). Thus, we were able to uncover several varying network structures. While we considered the most general form of an ODE Equation (1) that describes the interactions between two components, an interesting future direction would be to extend our work to models that describe the interactions among multiple oscillatory components, e.g. $\frac{dY}{dt} = f(X_1, \ldots, X_n, Y)$.

4 Materials and methods

4.1 Inferring Oscillatory Networks (ION) computational package

We provide user-friendly MATLAB code at https://github.com/Mathbiomed/ION (Github) and https://doi.org/10.6084/m9.figshare.16431408.v1 (figshare). The ION package can be used to infer the network structure of oscillators, which are described by Equation (1), across all levels of biology. Here, we briefly describe the key steps of the ION package (see Supplementary Information for a comprehensive manual).

4.1.1 Reflection times

For each time point t_i of the given time series $X(t) = (X(t_1), X(t_2), \ldots, X(t_n))$, first, the reflection time t_{i_X} needs to be calculated (Fig. 1B). That is, we seek the time point t_{i_X} such that $X(t_i) = X(t_{i_X})$ and the signs of the slopes at $X(t_i)$ and $X(t_{i_X})$ are opposite (i.e. rising and falling phase). For this, the discrete $X(t)$ is interpolated to a continuous time series $F_X(t)$ with either the 'linear' or 'fourier' interpolation method, chosen by the user. Then, t_{i_X} is estimated by identifying the closest time point to t_i among time points t satisfying the following equation:

$$F_X(t) = X(t_i) \quad \text{and} \quad \text{sign}(F'_X(t)) \neq \text{sign}(F'_X(t_i)).$$

4.1.2 Regulation-detection function and score

Using the estimated t_{i_X}, we compute the regulation-detection function, e.g. $R^{t_{i_X}}_{Y\to X}(t_i)$, for each time point t_i as follows:

$$(Y(t_{i_X}) - Y(t_i))(\dot{X}(t_{i_X}) - \dot{X}(t_i)).$$

If the linear method is chosen, $Y(t_{i_X})$ is linearly interpolated based on the data $(Y(t_1), \ldots, Y(t_n))$, and $\dot{X}(t) = (\dot{X}(t_1), \ldots, \dot{X}(t_n))$ is estimated using a moving slope filter method. Specifically, after fitting a low-order polynomial regression model to $X(t) = (X(t_1), X(t_2), \ldots, X(t_n))$ with a sliding window (Oppenheim *et al.*, 1999), the derivative of the polynomial fit is used to estimate $\dot{X}(t)$,and then $\dot{X}(t_{i_X})$ is linearly interpolated based on the estimated $\dot{X}(t)$. The model order and the length of the sliding window parameters can be adjusted (see Supplementary Information). On the other hand, if the fourier method is chosen, both $\dot{X}(t_i)$ and $\dot{X}(t_{i_X})$ are estimated as $\dot{F}_X(t_i)$ and $\dot{F}_X(t_{i_X})$, respectively, and similarly, $Y(t_i)$ and $Y(t_{i_X})$ are estimated as $F_Y(t_i)$ and $F_Y(t_{i_X})$, respectively, where $F_Y(t)$ is the Fourier series fit to the data $Y(t)$. Finally, in both cases, the regulation-detection score (Equation 4) is estimated using the MATLAB function trapz.

4.2 Time-series data

We simulate *in silico* data using the MATLAB function ode23tb (Fig. 2). See Supplementary Information for the model equations and parameters. The experimental datasets of the repressilator (Fig. 4B) were obtained from (Potvin-Trottier *et al.*, 2016). Next, to generate the duplicated experimental repressilator dataset (Fig. 4C), we mixed two copies of the repressilator dataset from Figure 4B. We kept one copy at the original phase and, for the second copy, we shifted the phase by 115 min (almost half of the period) (Fig. 4C). Then, we removed data on the left and the right where there was only coverage of one of the two datasets. We obtained the estradiol dataset (Fig. 4D) from (Lemaire *et al.*, 2006; Métivier *et al.*, 2003) and the Paramecium/Didinium data (Fig. 1E) from (Sugihara *et al.*, 2012).

4.3 Discrete and noisy data

To generate discretely sampled data (Fig. 3B), we select a random point in the first period to begin data extraction, and then we uniformly sample two periods worth of data at a sampling rate of 100 points per period. We repeat this process 100 times—every time randomly initializing the starting point in the first period—to generate 100 distinct datasets for every model. Then, we run our algorithm and compute F_1 scores for each of the 100 datasets. Next, from each of the 100 generated datasets, we take every other data point to reduce the number of data points (e.g. $50, 33, 25, \ldots, 10$ per period).

For the multiplicative noise analysis (Fig. 3C), we begin with two periods worth of data sampled at 100 points per period. Then, we add multiplicative noise sampled randomly from a normal distribution with mean 0 and standard deviation given by the percentage. For example, at 10% multiplicative noise, we add the noise $X(t_i) \cdot \epsilon$ to $X(t_i)$, where ϵ is sampled randomly from $N(0, 0.1^2)$.

4.4 PCM and GC

We ran the Partial Cross Mapping (PCM) (Leng *et al.*, 2020), an extension of CCM, with an embedding dimension of 3, $\tau = 1$, a max delay of 3 and a threshold of 0.5684 as recommended in (Leng *et al.*, 2020). We ran the GC using the code provided in (Chandler, 2020), specifying a max delay of 3 as we did with the PCM and a significance level of 95%. We rejected the null hypothesis that Y does not Granger cause X, and thereby inferred direct regulations, if the value of the F-statistic was greater than the critical value from the F-distribution (Granger, 1969).

Acknowledgements

The authors thank Anne Shiu and Seokjoo Chae for valuable comments.

Funding

This work was supported by a National Institutes of Health Training Grant [T32 HL007622 to J.T.]; the Institute for Basic Science [IBS-R029-C3 to J.K.K.] and Samsung Science and Technology Foundation [SSTF-BA1902-01 to J.K.K.].

Conflict of Interest: none declared.

References

Abel,J.H. *et al.* (2016) Functional network inference of the suprachiasmatic nucleus. *Proc. Natl. Acad. Sci. USA*, **113**, 4512–4517.

Balsa-Canto,E. *et al.* (2008) Hybrid optimization method with general switching strategy for parameter estimation. *BMC Syst. Biol.*, **2**, 26.

Brunton,S.L. *et al.* (2016) Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc. Natl. Acad. Sci. USA*, **113**, 3932–3937.

Casadiego,J. *et al.* (2017) Model-free inference of direct network interactions from nonlinear collective dynamics. *Nat. Commun.*, **8**, 2192.

Chandler. (2020) Granger causality test (https://www.mathworks.com/matlabcentral/fileexchange/25467-granger-causality-test), MATLAB Central File Exchange. Retrieved October 2020.

Choi,B. *et al.* (2020) Bayesian inference of distributed time delay in transcriptional and translational regulation. *Bioinformatics*, **36**, 586–593.

Cobey,S. and Baskerville,E.B. (2016) Limits to causal inference with state-space reconstruction for infectious disease. *PLoS One*, **11**, e0169050.

Deyle,E.R. and Sugihara,G. (2011) Generalized theorems for nonlinear state space reconstruction. *PLoS One*, **6**, e18295.

Deyle,E.R. *et al.* (2013) Predicting climate effects on pacific sardine. *Proc. Natl. Acad. Sci. USA*, **110**, 6430–6435.

Deyle,E.R. *et al.* (2016) Global environmental drivers of influenza. *Proc. Natl. Acad. Sci. USA*, **113**, 13081–13086.

Elowitz,M.B. and Leibler,S. (2000) A synthetic oscillatory network of transcriptional regulators. *Nature*, **403**, 335–338.

Firman,T. *et al.* (2019) Maximum caliber can build and infer models of oscillation in a three-gene feedback network. *J. Phys. Chem. B*, **123**, 343–355.

FitzHugh,R. (1961) Impulses and physiological states in theoretical models of nerve membrane. *Biophys. J.*, **1**, 445– 466.

Geva-Zatorsky,N. *et al.* (2010) Fourier analysis and systems identification of the p53 feedback loop. *Proc. Natl. Acad. Sci. USA*, **107**, 13550–13555.

Goodwin,B.C. (1965) Oscillatory behavior in enzymatic control processes. *Adv. Enzyme Regul.*, **3**, 425–438.

Gotoh,T. *et al.* (2016) Model-driven experimental approach reveals the complex regulatory distribution of p53 by the circadian factor Period 2. *Proc. Natl. Acad. Sci. USA*, **113**, 13516–13521.

Granger,C.W.J. (1969) Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, **37**, 424–438.

Igoshin,O.A. *et al.* (2004) A biochemical oscillator explains several aspects of *Myxococcus xanthus* behavior during development. *Proc. Natl. Acad. Sci. USA*, **101**, 15760–15765.

Jensen,M.H. *et al.* (2009) Genetic oscillation patterns. *Eur. Phys. J. Special Top.*, **178**, 45–56.

Jo,H.-H. *et al.* (2018) Waveforms of molecular oscillations reveal circadian timekeeping mechanisms. *Commun. Biol.*, **1**, 207.

Kim,J.K. (2016) Protein sequestration versus Hill-type repression in circadian clock models. *IET Syst. Biol.*, **10**, 125–135.

Kim,J.K. and Forger,D.B. (2012a) A mechanism for robust circadian timekeeping via stoichiometric balance. *Mol. Syst. Biol.*, **8**, 630.

Kim,J.K. and Forger,D.B. (2012b) On the existence and uniqueness of biological clock models matching experimental data. *SIAM J. Appl. Math.*, **72**, 1842–1855.

Konopka,T. (2011) Automated analysis of biological oscillator models using mode decomposition. *Bioinformatics*, **27**, 961–967.

Konopka,T. and Rooman,M. (2010) Gene expression model (in)validation by Fourier analysis. *BMC Syst. Biol.*, **4**, 123.

Lemaire,V. *et al.* (2006) Sequential recruitment and combinatorial assembling of multiprotein complexes in transcriptional activation. *Phys. Rev. Lett.*, **96**, 198102.

Leng,S. *et al.* (2020) Partial cross mapping eliminates indirect causal influences. *Nat. Commun.*, **11**, 2632.

Lillacci,G. and Khammash,M. (2010) Parameter estimation and model selection in computational biology. *PLoS Comput. Biol.*, **6**, e1000696.

Lütkepohl,H. (2005) *New Introduction to Multiple Time Series Analysis*. Springer, Berlin.

Ma,H. *et al.* (2017) Detection of time delays and directional interactions based on time series from complex dynamical systems. *Phys. Rev. E*, **96**, 012221.

Mangan,N.M. *et al.* (2016) Inferring biological networks by sparse identification of nonlinear dynamics. *IEEE Trans. Mol. Biol. Multi-Scale Commun.*, **2**, 52–63.

McBride,D. and Petzold,L. (2018) Model-based inference of a directed network of circadian neurons. *J. Biol. Rhythms*, **33**, 515–522.

McGoff,K.A. *et al.* (2016) The local edge machine: inference of dynamic models of gene regulation. *Genome Biol.*, **17**, 214.

Métivier,R. *et al.* (2003) Estrogen receptor-α directs ordered, cyclical, and combinatorial recruitment of cofactors on a natural target promoter. *Cell*, **115**, 751– 763.

Mhaskar,P. *et al.* (2002) Cell population modeling and parameter estimation for continuous cultures of Saccharomyces cerevisiae. *Biotechnol. Progress*, **18**, 1010–1026.

Mure,L.S. *et al.* (2018) Diurnal transcriptome atlas of a primate across major neural and peripheral tissues. *Science*, **359**, eaaao0318.

Nagumo,J. *et al.* (1962) An active pulse transmission line simulating nerve axon. *Proc. IRE*, **50**, 2061–2070.

Novák,B. and Tyson,J.J. (2008) Design principles of biochemical oscillators. *Nat. Rev. Mol. Cell Biol.*, **9**, 981–991.

Oppenheim,A.V. *et al.* (1999) *Discrete-Time Signal Processing*. Technology and Engineering. Prentice Hall, Upper Saddle River, NJ.

Pigolotti,S. *et al.* (2007) Oscillation patterns in negative feedback loops. *Proc. Natl. Acad. Sci. USA*, **104**, 6533–6537.

Pigolotti,S. *et al.* (2009) Symbolic dynamics of biological feedback networks. *Phys. Rev. Lett.*, **102**, 088701.

Pitt,J.A. and Banga,J.R. (2019) Parameter estimation in models of biological oscillators: an automated regularised estimation approach. *BMC Bioinf.*, **20**, 82.

Potvin-Trottier,L. *et al.* (2016) Synchronous long-term oscillations in a synthetic gene circuit. *Nature*, **538**, 514–517.

Pourzanjani,A. *et al.* (2015) On the inference of functional circadian networks using granger causality. *PLoS One*, **10**, e0137540.

Radde,N. and Kaderali,L. (2009) Inference of an oscillating model for the yeast cell cycle. *Discrete Appl. Math.*, **157**, 2285–2295.

Runge,J. *et al.* (2019) Inferring causation from time series in earth system sciences. *Nat. Commun.*, **10**, 2553.

Saint-Antoine,M.M. and Singh,A. (2020) Network inference in systems biology: recent developments, challenges, and applications. *Curr. Opin. Biotechnol.*, **63**, 89–98.

Stokes,P.A. and Purdon,P.L. (2017) A study of problems encountered in granger causality analysis from a neuroscience perspective. *Proc. Natl. Acad. Sci. USA*, **114**, E7063–E7072.

Stražar,M. *et al.* (2014) An adaptive genetic algorithm for parameter estimation of biological oscillator models to achieve target quantitative system response. *Nat. Comput.*, **13**, 119–127.

Sugihara,G. *et al.* (2012) Detecting causality in complex ecosystems. *Science*, **338**, 496–500.

Tani,N. *et al.* (2020) Small temperature variations are a key regulator of reproductive growth and assimilate storage in oil palm (*Elaeis guineensis*). *Sci. Rep.*, **10**, 650.

Toni,T. *et al.* (2009) Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface*, **6**, 187–202.

Trejo Banos,D. *et al.* (2015) A Bayesian approach for structure learning in oscillating regulatory networks. *Bioinformatics*, **31**, 3617–3624.

Tsonis,A.A. *et al.* (2015) Dynamical evidence for causality between galactic cosmic rays and interannual variation in global temperature. *Proc. Natl. Acad. Sci. USA*, **112**, 3253–3256.

Veilleux,B. (1976) The Analysis of a Predatory Interaction between Didinium and Paramesium. Master's Thesis, University of Alberta, Edmonton, Alberta.

Wang,B. and Enright,W. (2013) Parameter estimation for ODEs using a cross-entropy approach. *SIAM J. Sci. Comput.*, **35**, A2718–A2737.

Wang,J.-Y. *et al.* (2020) Causal effects of population dynamics and environmental changes on spatial variability of marine fishes. *Nat. Commun.*, **11**, 2635.

Wang,S. *et al.* (2018) Inferring dynamic topology for decoding spatiotemporal structures in complex heterogeneous networks. *Proc. Natl. Acad. Sci. USA*, **115**, 9300–9305.

Yang,A.C. *et al.* (2018) Causal decomposition in the mutual causation system. *Nat. Commun.*, **9**, 3378.

Ye,H. *et al.* (2015) Distinguishing time-delayed causal interactions using convergent cross mapping. *Sci. Rep.*, **5**, 14750.

Zhang,R. *et al.* (2014) A circadian gene expression atlas in mammals: implications for biology and medicine. *Proc. Natl. Acad. Sci. USA*, **111**, 16219–16224.