

Impact of environmental inputs on reverse-engineering approach to network structures

Jianhua Wu^{1,#}, James L Sinfield^{*2,#}, Vicky Buchanan-Wollaston³ and Jianfeng Feng^{*2,4}

¹Dept of Neuroscience, Columbia University, New York, NY, USA 10032

²Dept of Computer Science, Warwick University, Coventry, UK CV4 7AL

³Warwick HRI, University of Warwick, Wellesbourne, Warwick, UK CV35 9EF ⁴Centre for Computational System Biology, Fudan University, Shanghai, PR China

These authors contributed equally.

Email: Jianhua Wu - jw2663@columbia.edu; James L Sinfield - j.l.Sinfield@warwick.ac.uk; Vicky Buchanan-Wollaston - vicky.b-wollaston@warwick.ac.uk; Jianfeng Feng - Jianfeng.Feng@warwick.ac.uk;

*Corresponding author

Abstract

Background: Uncovering complex network structures from a biological system is one of the main topic in system biology. The network structures can be inferred by the dynamical Bayesian network or Granger causality, but neither techniques have seriously taken into account the impact of environmental inputs.

Results: With considerations of natural rhythmic dynamics of biological data, we propose a system biology approach to reveal the impact of environmental inputs on network structures. We first represent the environmental inputs by a harmonic oscillator and combine them with Granger causality to identify environmental inputs and then uncover the causal network structures. We also generalize it to multiple harmonic oscillators to represent various exogenous influences. This system approach is extensively tested with toy models and successfully applied to a real biological network of microarray data of the flowering genes of the model plant *Arabidopsis Thaliana*. The aim is to identify those genes that are directly affected by the

presence of the sunlight and uncover the interactive network structures associating with flowering metabolism.

Conclusions: We demonstrate that environmental inputs are crucial for correctly inferring network structures. Harmonic causal method is proved to be a powerful technique to detect environment inputs and uncover network structures, especially when the biological data exhibit periodic oscillations.

Background

One of the main topics in system biology is to uncover the complex network structures in a biological system [1,2]. In comparison with simple systems, nowadays the researchers always face larger and more complex dynamic interactive systems (e.g., neural networks and gene networks). Traditional techniques, such as the cross-correlation and partial coherence analysis [3–7], are inadequate to clearly and explicitly reveal the true network structures for such a complex system. These techniques neither take time dimension into consideration nor reveal the directional interactions, thus they cannot configure a dynamic interactive system with time. Over the past few decades several advanced techniques such as dynamic Bayesian networks [8] and Granger causality [9–13] have been developed to identify network structures in dynamic systems. Granger causality only reveals direct causality between pairwise units with linear interactions, thus conditional and partial Granger causality [14,15] and kernel Granger causality [16] have been proposed to deal with indirect causal interactions among multiple variables.

In multivariable (more than two) situations, one time series can be connected to another time series in a direct or an indirect manner, raising the important question of whether there exists a (direct) causal influence between two time series when the influence of other time series are taken into account. In such cases, repeated bivariate analysis can be misleading. For example, one time series may falsely appear to cause another if they are both influenced by a third time series but with different time delays. The conditional

Granger causality [14,15] aim to deal with the causal interactions among multiple variables. However, the applicability of the kernel Granger causality or the conditional Granger causality largely depends on the experimental ability to measure all relevant variables in the system, but it is usually not feasible in the biological recordings. Environmental inputs, including exogenous inputs from external sources and unmeasured endogenous variables, cannot be all captured by experimental techniques, but such environmental inputs can confound the accuracy of causal influences and thus degrade the credibility of the uncovered network structure. For example, in our experimental data recorded from the inferotemporal (IT) cortex of sheep, every measured neuron receives common exogenous inputs from the visual cortex and feedbacks from the prefrontal cortex [7,15]. Even with advanced multielectrode array (MEA) techniques, it is only able to record a tiny subset of interacting neurons in a single area [15,17] and there are bound to be endogenous variables. Hence controlling environmental inputs is a critical issue when applying Granger causality to experimental data. Recently, partial Granger causality [15,18] is developed to eliminate the influences of exogenous inputs and latent variables, but a full elimination is only possible if all common inputs have a more or less identical influence on all measured variables. It is generally not realistic that all measured variables can receive an almost identical and common influence in experimental recordings. In fact, the common influence degrades due to spatial increment. In such cases, it is critical to identify which measured variables received environmental inputs, and what is the impact of the environmental inputs on configuring network structures?

We take a system biology approach to answer the questions above. Most current techniques largely ignore the natural dynamical characteristics of the biological data, which usually exhibits highly rhythmic (periodic) oscillations, especially under periodic environmental influence, e.g. light-dark condition [19–23]. Such natural periodic dynamics of experimental data can provide important information in model fitting and error estimation. To overcome the limitations of current causality techniques and make full use of harmonic oscillation characteristic of experimental data, we consider a harmonic oscillator, or a set of harmonic oscillators, to represent the environmental inputs. The harmonic oscillators can be mathematically formulated by the hidden periodic model [24,25]. We extend the current

linear Granger causality model (Autoregressive model) by inclusion of the typical harmonic oscillators embedded in the experimental recordings. If the inclusion of harmonic oscillators can significantly reduce the variance of the prediction error, then the environmental causal influence can be reduced or eliminated. The mathematical representation of a harmonic oscillator model is derived in Method section and the application of the harmonic Granger causality approach is elucidated in Result section. Although the techniques of Granger causality are based on time series data, additional useful information can be revealed when the analysis is performed in the frequency domain [14, 18, 26, 27]. Investigating the causal interactions between different frequencies adds another dimension to the already complex identification of spatiotemporal and frequency-specific rhythmic oscillations. Conventional cross-frequency interactions are characterized by the synchrony of phase, recognized as 'n:m phase synchrony' [28, 29]. Phase synchrony indicates amplitude-independent phase-locking of n cycles of one oscillation to m cycles of another oscillation, however, this method largely ignores the information carried by the amplitude and the coupling effects between phase and amplitude. Importantly, phase information can be sensitive to random noise [30], while in most experimental data the true signal is heavily masked by random noise. In this study, as a by-product, we also assess whether it is reliable to use the phase information between two (oscillating) units to estimate the causality. Our simulation results clearly demonstrate that it can be very insufficient and inaccurate to use only the phase information to characterize a causal interaction, but the approach developed in the current paper works. We first apply the harmonic oscillator idea to a toy model and validate it by comparing with the conventional Granger causality. Then we investigate the effect of multiple oscillators by employing a small sparsely connected network. Finally we apply the harmonic Granger causality to a real biological network of microarray data of the flowering genes of the model plant *Arabidopsis Thaliana*. We aim to identify which genes are directly affected by the presence of the sunlight, and uncover the causal interactions among genes. Although tens of thousands of known genes within *Arabidopsis* are collected with Microarray, only those genes known to be involved in the flowering of the plant (8 genes in this case) are analyzed by our harmonic Granger causality. The method successfully

reveals the genes that most possibly receive environmental inputs. We finally compare our causal network with other candidate models in the literatures [31–33]. With this system biology approach, our causal network depicts all possible connections reported in the literatures [32, 33], and also reveals two more connections that do not exist in the known candidate models.

Results

Toy Model & Validation

In this simple toy-model example, we compare the performance of traditional Granger causality and harmonic Granger causality on four simple model configurations. We show that the traditional Granger Causality analysis is not sufficient to describe the influence of one time series upon another in the presence of an external driving force. The full simulated model under consideration is described as follows:

$$\begin{cases} X_t = 0.9X_{t-1} - 0.5X_{t-2} + C_x \cos(2\pi f_x t + \phi_x) + \epsilon_t \\ Y_t = 0.3X_{t-2} + 0.8Y_{t-1} - 0.5Y_{t-2} + C_y \cos(2\pi f_y t + \phi_y) + \eta_t \end{cases} \quad (1)$$

Within this model we consider four configurations, firstly where both node X_t and Y_t experience an environmental input, secondly where X_t experiences the environmental input and Y_t does not, thirdly where Y_t experiences the environmental input and X_t does not, and lastly where neither X_t or Y_t undergo the influence from the external driving oscillation. Despite the external driving oscillation exerted on either X or Y , there is a coupling frequency of 34Hz between X and Y for all four configurations. In each of the simulations, f_x and f_y are equal to 10Hz, the input phases, ϕ_x and ϕ_y are set to zero, and the variances of inherent noises, ϵ_t and η_t , are set to 0.1. In all the following simulations, the values of C_x and C_y are 0.07 and 0.05, respectively .

In the first case where both nodes experience the environmental inputs, Fig. 1(B_i) and C_i shows that both $X(\omega)$ and $Y(\omega)$ have components at the driving oscillation frequency of 10Hz. The causal link known to be present from X to Y should mean that there is a spike in the causality spectra at the driving frequency. Fig. 1(D_i) and (E_i) show that the normal Granger AR method can only detect the causality spectra at the intrinsic coupling frequency around 34 Hz, but the Harmonic method is capable of detecting the causal

influence at the external driving frequency at 10 Hz. The reason is that the driving force from X to Y comes from the factor $0.3X_t - 2$ (Eq. 1), which contains both intrinsic coupling frequency and external driving frequency. Harmonic method can better fit and predict the data at external driving frequency such that it can detect the external driving frequency at 10 Hz better than normal Granger AR method.

The second case is where X_t receives the driving input and Y_t does not. Fig. 1(B_{ii}) shows that the peak in the spectra can be seen in the X-channel at $10Hz$ representing the external driving oscillation, and although there is no inherent $10Hz$ oscillation in the Y_t channel, the spectra of $Y(\omega)$ (Fig. 1(C_{ii})) shows that there is an induced oscillation at this frequency, hence one would expect that the causality of $F_{X \rightarrow Y}$ will show a large component at the driving frequency. Fig. 1(D_{ii}) and E_{ii} show the causality, $F_{X \rightarrow Y}$ calculated using both AR and harmonic methods. The AR method shows little peak at the driving frequency, while the harmonic method produces a large spike in the causality at the frequency 10 Hz. In third case where Y_t receives the external input, Fig. 1(C_{iii}) shows that the peak in the spectra can be seen in the Y-channel at $10Hz$ representing the external driving oscillation. We would expect that in this case the causality shows no peak at the driving frequency as X_t contains no driving oscillation: Fig. 1(D_{iii}) and (E_{iii}) show the causality, $F_{X \rightarrow Y}$ calculated using AR (panel D_{iii}) and harmonic methods (panel E_{iii}). In this model configuration, both the AR and the harmonic methods produce similar results. However there is a drastic decrement in the spectra of the harmonic method at the driving frequency in the causality, qualitatively, this cannot be the case, so in this instance it is preferable to use the AR method to calculate the causality spectra as X_t contains no driving oscillation. The causality decrement near the driving frequency may be caused by the inclusion of the harmonic term, which may extract the intrinsic power of the signal Y at the driving frequency of 10 Hz. The decrement means that signal Y was mainly driven by the harmonic term at 10 Hz and signal X did not contribute much at this frequency. For other frequency range, harmonic term did not influence signal Y and the driving force mainly came from signal X . The final model configuration considered is where neither X_t or Y_t has the driving input. Fig. 1(D_{iv}) and (E_{iv}) show that in this case there is very little difference in the causality spectra obtained using the two methods.

Through this simple toy model we have demonstrated that the normal Granger Causality in the frequency domain is not sufficient to detect interactions at all frequencies in the presence of an environmental input. Furthermore we have demonstrated that the causal method with the additional harmonic term produces more consistent and accurate results than the traditional Granger causality method at most instances. The harmonic causal method and the traditional Granger causality method can be a good complement to each other when applying to the time series with or without oscillatory environmental influences.

Investigating the Effects of Phase

It is a common scenario in physics to determine the driving relationship between two oscillators. The phases of the two oscillators can be interleaved throughout the time, thus phase may not be an accurate indication of the causal interactions between the two oscillators. One of the main motivations of introducing the harmonic term is to investigate the effects of phase and to determine whether it can be used as an indication of the amount of causal interactions between oscillatory signals. Consider again the case where X_t drives Y_t , Eq. 1 shows that the influence of the time series X_t upon Y_t is encapsulated in the oscillatory and noise terms. However, the amplitude of the resultant oscillation will depend on the phase difference of the two harmonic terms. This can be seen by considering the addition of the two oscillatory terms. Let $O_y = C_y \cos(2\pi f_y t + \phi_y)$ and the harmonic component from X_t be $\hat{O}_x = \hat{C}_x \cos(2\pi f_x t + \hat{\phi}_x)$, then in the case that the frequency of the oscillations are equal, $f_x = f_y = f$, the two oscillation terms can be combined as follows:

$$C_y \cos(2\pi f t + \phi_y) + \hat{C}_x \cos(2\pi f_x t + \hat{\phi}_x) = \tilde{S} \cos(2\pi f t + \tilde{\phi}) \quad (2)$$

where

$$\tilde{S} = \sqrt{C_y^2 + \hat{C}_x^2 + 2C_y \hat{C}_x \left(\cos(\phi_y) \cos(\hat{\phi}_x) + \sin(\phi_y) \sin(\hat{\phi}_x) \right)} \quad (3)$$

and

$$\tilde{\phi} = \cos^{-1} \left(\frac{\hat{C}_x \cos(\hat{\phi}_x) + C_y \cos(\phi_y)}{\tilde{S}} \right) \quad (4)$$

$$= \sin^{-1} \left(\frac{\hat{C}_x \sin(\hat{\phi}_x) + C_y \sin(\phi_y)}{\tilde{S}} \right) \quad (5)$$

It is able to analyze the effect of input phase upon the level of influence of one time series upon another. Eq. 3 reveals that the magnitude of the resultant harmonic term is indeed a function of both $\hat{\phi}_x$ and ϕ_y . The oscillating term, $(\cos(\phi_y)\cos(\hat{\phi}_x) + \sin(\phi_y)\sin(\hat{\phi}_x))$, oscillates sinusoidally and can take values in the range $[-1, 1]$.

The relative effect of this oscillation upon \hat{S} depends on the values of C_y and \hat{C}_x . The extremes of \tilde{S} are given by $\hat{S} = \sqrt{C_y^2 + \hat{C}_x^2} \pm 2C_y\hat{C}_x$, and it can be shown that the maximum effect of the phase differences happens when $C_y = \hat{C}_x$ and the minimal difference happens when either C_y or \hat{C}_x equals zero.

In order to demonstrate the effects of phase changes, the same model as Eq. 1 is used. Fig. 2 shows the enormous consequences of the magnitude of the resultant signal by simply varying the phase of the environmental input, ϕ_x and ϕ_y . Fig. 2 A shows that the configuration of the model with both nodes of the system receiving external environmental inputs and a causal link from X to Y. Fig. 2 B shows two examples in the time domain traces obtained in the absence of environmental noise. The upper panel Fig. 2B₁ shows the time domain traces when the phase difference of the external inputs is zero ($\phi_x = \phi_y$) and the lower panel Fig. 2B₂ shows the scenario when $\phi_x - \phi_y = \pi$. In each of these diagrams the blue trace, which is the node X_t , is unchanged, however the time series associated with Y_t change considerably. The difference of the maximum amplitudes of the two time series is denoted by Δ . The colourmap figure shown in Fig. 2C plotted the value of Δ against the values of the input noise and the input phase differences. It can be seen that the effects of the phase differences is lessened by increasing the noise in the system. Fig. 2D reveals the interdependence between noise and phase in this system. The upper panel Fig. 2D₁ shows the amplitude difference against the phase differences in the absence of noise. The reason why the peak does not happen at π is because C_x and C_y are different, and there is a phase difference. If $C_x = C_y$, then the peak would be exactly at π . The lower panel Fig. 2D₂ shows how a measure of the difference varies with increasing noise. The measure of the difference is defined as $\log\left(\frac{2\Delta_{max} - \Delta_{min}}{\Delta_{max} + \Delta_{min}}\right)$. As the noise increases the measure difference tends to zero, confirming that the noise levels can mask the phase effects.

We have shown that the phase of the external driving oscillation has an effect on the amplitudes of the resultant time series, then the real question is if this effect can be

detected as a difference in the level of causality. Recall that the actual level of causal interaction is not varying and the influence that X_t exerts over Y_t does not alter throughout the simulations. The investigation into the effects of phase continued with a series of calculations determining the level of causality. Fig. 3 shows the causality detected in the system obtained by four different methods. Fig. 3A₁ shows the time domain causality detected using the normal Granger method, Fig. 3B₁ shows the time domain causality detected using the harmonic method, Fig. 3A₂ and B₂ show the causality detected using the AR and harmonic methods in the frequency domain. For consistency we require that:

$$F_{X \rightarrow Y} = \int_{-\pi}^{\pi} f_{X \rightarrow Y}(\omega) d\omega \quad (6)$$

where $F_{X \rightarrow Y}$ is the causality on the time domain and $f_{X \rightarrow Y}(\omega)$ is the causality at frequency ω . Eq. 6 is the Kolmogorov condition that indicates the equivalence of the causality in time and frequency domain.

Comparing the AR and harmonic causality results, the amplitude differences influenced by the phase differences are represented in the AR method but not in the harmonic causality results, which imply that the harmonic approach must be used for a true indication of the level of causal interaction as it is robust in presence of the external driving oscillation.

Comparison of Fig. 3(A₁) and Fig. 3(A₂) shows that the Eq. 6 holds for the AR method and there is a high level of correspondence between the causality calculated in the time domain and frequency domain. Similarly, comparison of Fig. 3(B₁) and Fig. 3(B₂) shows that to a large extent the harmonic method is consistent with the Kolmogorov condition in Eq. 6. The causality calculated by harmonic method does not depend on the phase difference no matter it is in the time domain or frequency domain.

We further investigated the relationship between the phase difference of input and output signals, and the influence of noise level. The same configuration model presented in Fig. 2 A was used to demonstrate this interrelation. The phase difference between two input signals varied from 0 to 2π and the noise level (variance of the white noise) increased from 0 to 0.2. The phase difference of signal X and Y was plotted as a function against the phase difference of the input signals and the noise level. Fig. 4 A shows the colourmap of the interrelation. The color intensity represents the phase difference for the output signals. Fig. 4 B and Fig. 4 C show the averaged intensity along noise level and the input phase

difference, respectively. It is clear that the phase difference of the output signals does not depend on the phase difference of the input signals and the noise level. The results indicate that the phase information cannot be used alone to accurately determine the causal relationship between any two signals. The interpretation of the causality based on phase should be cautious as the causality may not reflect the true relationship.

Investigating Effect of Multiple Oscillators

In experimental recordings, the measured variables are usually influenced by many environmental inputs, thus multiple oscillators have to be considered. In order to reveal the power and limitations associated with the additional oscillators, a simple system was considered and the errors of seven different connection schemes was compared. The schematic plot for seven connection schemes is displayed in Fig. 5A, and the error terms corresponding to each schemes are described as follows:

- i) Σ_{NC} : AR + harmonic oscillations, no causality
- ii) Σ_{C1} : AR + harmonic oscillations, influence from causal node
- iii) Σ_{C2} : AR + harmonic oscillations, influence from non-causal nodes
- iv) Σ_{NAR} : harmonic oscillations only
- v) Σ_{ARNC} : AR only
- vi) Σ_{ARCA1} : AR, influence from causal node
- vii) Σ_{ARCA2} : AR, influence from non-causal nodes

To investigate the effect of multiple oscillators on the goodness of fit, we consider a simple network consisting of five nodes in various random configurations. The variances are estimated for each of the connection scheme in Fig. 5A, and the number of harmonic oscillators varies from 1 to 11. The results of this simulation are shown in Fig. 5B.

Inspection of Fig. 5B reveals some of the issues involved in using the harmonic oscillators to provide a full description of the time domain evolution. Firstly looking at those variance estimates with no harmonic oscillatory components (Σ_{ARNC} , Σ_{ARCA1} and Σ_{ARCA2}), as expected in Fig. 5B that these are constants for increasing numbers of oscillators (invariant as they are independent of oscillators). The estimates of Σ_{ARNC} and Σ_{ARCA2} are very similar, because the non-causal nodes do not provide useful information for fitting the target node. The causal estimate (Σ_{ARCA1}) is the best estimate when no harmonic

oscillators are considered.

Inspection of estimates in which oscillations are included reveals an important trend; one would expect that as the number of oscillators increases, the estimates become more and more accurate. Theoretically speaking this should be the case, however realistically it is shown not to be the case as in Fig. 5B. When the number of oscillators goes beyond seven, the estimated variances drastically increase. The reason for this situation is the overfitting problem. The total number of parameters needed to be estimated depends on the number of oscillators (three extra parameters per oscillator), and the known data cannot fit the model if the parameters go beyond the amount of data. As expected, the causal estimate (Σ_{C1}) provides the best estimate for small number of oscillators (minimum at oscillators = 3). This simulation reveals that the number of oscillators does not exceed certain value (seven in this case) if we do indeed obtain a good estimate far surpassing the accuracy of the estimate using only AR. We have to perform goodness of fit test to determine the number of oscillators that can help to fit the data best.

Gene Data & Network

Having shown the necessity of applying the harmonic approach to identifying causality in data sets where it is known that an external environmental oscillation is driving the time series, it is necessary to apply this system approach to real experimental data. The microarray gene data pertaining to the flowering clock cycle of the Arabidopsis is one such example where this methodology may prove enlightening. The plants (Arabidopsis) are grown in laboratory conditions, where they are subjected to 12 hours of artificial daylight followed by 12 hours of no light representing night time. Gene microarray data is collected at regular intervals (twice a day) throughout the experiment. Inspection of the time series of this gene expression data reveals that there is a clear periodic oscillation which corresponds to day/night time periods, suggesting that the expression levels of the genes depend upon the amount of sunlight present (see Fig. 6 panel A).

We consider the time domain change of eight genes involved in the flowering system of the Arabidopsis, namely CCA1, ELF4, GI, LHY, PRR5, PRR7, PRR9 and TOC1. The time domain trace of the expression of these genes is shown in Fig. 6 panel A. Each of the genes with the exception of GI exhibits highly oscillatory behaviour with period of one day. This

periodicity is attributed to the presence of incident sunlight during the day time and its absence during the night. Some of the genes are directly affected by the light and are expressed to a greater or less extent during the day. The experimental data used for this analysis is over a period of 11 days with two measurements per day, hence data for each gene consists of 22 data points and there are 16 repetitions of each time point (4 biological repetitions and 4 technical repetitions for each measurement).

The task regarding this data set is twofold, firstly we wish to identify which of the genes are driven by the external oscillation. And secondly, we wish to determine how the genes are connected to form the causal network governing flowering of the plant. The method to determine environmental input and network connectivity is as follows. There are 56 pairwise combinations possible with eight genes; for each of these 56 gene pairs the parameters of four candidate models were calculated. These models are presented below:

$$\begin{aligned}
X_g(t) &= \sum_{i=1-h}^M \sum_{j=1}^p a_{i,j} X_{i,t-j} + C_g \cos(2\pi h_g t + \phi_h) + \epsilon_1 \\
X_g(t) &= \sum_{i=1}^M \sum_{j=1}^p b_{i,j} X_{i,t-j} + C_g \cos(2\pi h_g t + \phi_h) + \epsilon_2 \\
X_g(t) &= \sum_{i=1-h}^M \sum_{j=1}^p c_{i,j} X_{i,t-j} + \epsilon_3 \\
X_g(t) &= \sum_{i=1}^M \sum_{j=1}^p d_{i,j} X_{i,t-j} + \epsilon_4
\end{aligned}$$

where M is the number of genes in the network, p is the number of lagged observations used in the model. p can be determined by a quantity called Akaike Information Criteria (AIC) [34]. The four candidate models are descriptions of an estimation of X_g used to determine the effect of X_h with or without the external driver. The variance of the errors associated with each of the models are: $\Sigma_1, \Sigma_2, \Sigma_3, \Sigma_4$ respectively.

The errors (Σ_i) - associated with each model were estimated for each gene pair. Using these errors we can infer both the presence of an external environmental driver and the possibility of a connection between the pair of genes.

Comparison of Σ_1 and Σ_3 and comparison of Σ_2 and Σ_4 reveal that whether particular gene

may have an external input as these estimates differ only by the presence of the oscillatory input. If the estimate is improved appreciably by the addition of the harmonic term, then it is possible that this gene receives an environmental input. Therefore for each gene, a measure of the likelihood of input presence is obtained as follows: $M = \log \frac{\Sigma_1}{\Sigma_3} + \sum \log \frac{\Sigma_2}{\Sigma_4}$. Fig. 6B shows the oscillation metric for each of the 8 genes in the network. This method merely states which of the genes is more likely than others to have an input, so a decision must be made as the value of the oscillation metric is the cut-off point. At a first estimate, the value of $M = 0.5$ has been selected. Figure 6B shows that selecting a cut-off value of 0.5 for the oscillation metric leads to the following genes having an external oscillatory input; ELF4, PRR7, TOC1, LFY.

Having calculated whether the genes have an external input, it is possible to obtain the causality of each pair of genes. Consider the causality between gene X and Y . If gene X has an external input, then the causal influence X exerting upon Y is: $F_{X \rightarrow Y} = \log \frac{\Sigma_1}{\Sigma_2}$.

Whereas there is no external input to gene X , then the causality will be given by: $F_{X \rightarrow Y} = \log \frac{\Sigma_3}{\Sigma_4}$. The errors (Σ_i) - and hence the causality associated with each model were found for each pair of genes and then sorted in descending order. Those with the highest level of causality deemed more likely to have a connection.

The table of errors (Σ_i) shown in Fig. 7 is used to find the most likely connection in the gene network. One such candidate network is shown in Fig. 8A.

To validate the gene network generated by harmonic Granger causality, we compare it with other candidate network in the literature of circadian clock cycle by Ueda [31, 32]. The candidate network is reproduced in Fig. 8B. The three loop model was first proposed by Locke [31] and then modified by Ueda [32, 33]. In Ueda's model two hypothesized unnamed genes are omitted in our model and the genes LHY and CCA1 are treated as one entity. Our analysis reveals that four of the genes in this network receive external inputs: PRR7, CCA1, CCA1 and ELF4. The first two of these agree with Ueda's network. The Ueda's network states that TOC1 does not receive external influence but the hypothesized gene does. It is possible that this influence has been included in the time series of TOC1. The structure of the two networks also are very closely related, both showing a high level of connectivity. Perhaps the biggest difference is that our method shows that there are

connections between PRR7/PRR9 and TOC1, while Ueda’s model does not reveal such connections.

In addition to finding the likely connections between the genes, the frequency domain analysis allows us to investigate the frequencies at which one gene drives another. Fig. 8C shows the causality spectra calculated for each of the 15 connections believed to exist in the gene network shown in Figure 8. $f_{X \rightarrow Y}(\omega)$ is shown on frequency domain between $[0, 0.5]/\text{day}^{-1}$, whether this is calculated using the harmonic or AR method depends upon whether Gene X is thought to have a external input, i.e., those with an external input are calculated with the harmonic method, and those with no external input are calculated using the AR method. It is often the case that the two methods produce very different causality spectra, so selecting the correct method is essential, in many instances the AR method predicts a causal influence at the driving frequency, yet the harmonic method does not.

Discussion

Although harmonic causal method has greatly improved the performance of causal connection detection, there are several issues that harmonic causal method cannot answer or infer at this stage. First, the application of harmonic Granger causality has a precondition that the signal is influenced by harmonic environmental inputs. Most biological data exhibits harmonic oscillatory behavior, while there may also exist other form of nonlinear exogenous input other than harmonic form. The harmonic method cannot deal with such nonlinear interactions, and its application to nonlinear exerts would depend on specific problems. Second, harmonic causal method is developed to detect the directional causal interactions between any two elements, but it has no ability to determine the influence is positive or negative. For example, one neuron can exert an excitatory or inhibitory effect on another neuron; or one gene can cause another gene’s expression level up or down. It is not possible at this stage to figure out the positive or negative effect by only determining the directional causal interactions. Third, one should take extra caution when applying multiple harmonic causal method as the overfitting problem can easily happen. Excessive number of harmonic oscillator will generate inaccurate model estimations and predictions.

Conclusion

We have presented a system biology approach to study the impact of environmental inputs on recovering network structures. The harmonic modification of the Granger causality is essential if we want to have the complete picture of causal interactions between elements in a system in the presence of a periodic environmental oscillation. The toy model example demonstrated that the conventional Granger causality was not sufficient to reveal the level of causal influence in the presence of an oscillatory driver. Furthermore, the toy model was able to validate the estimates used in the definition of the frequency domain harmonic causality. One of the motivations for the introduction of the driving oscillation was to investigate whether it is feasible to use phase differences between oscillatory signals to assess the causality. We also showed that the apparent level of influence on the conventional Granger causality was tightly related to the phase difference and noise intensity, and this artefact was enough to render the estimation of the conventional Granger causality. The harmonic Granger causality was not sensitive to these phase effects and produces more accurate estimate of the true causality. We also applied the harmonic method to detect external drivers and causal connections in a gene network. We were able to predict which genes receive an environmental input from the sun and these results are in agreement with the experimental results to a large degree. Furthermore, we were able to reproduce the network, which not only reveals known connections but also predicts new connections comparing with classical candidate models.

Our approach clearly demonstrates that by including appropriate environmental (oscillatory) inputs in a conventional reverse-engineering approach could significantly improve its accuracy. Obviously the same idea could be applied to other approaches such as the Bayesian network inferences and information theory approach.

Methods

Causality in the Time Domain

In order to infer the connections between the elements of a system constituting a network, we propose an extended Granger Causality whereby a harmonic oscillatory term is added to the normal autoregressive (AR) and error terms of the conventional Granger analysis, and such simple modification can yield surprising and useful results. To appreciate the

effect of the proposed modification and the power of the addition of the harmonic oscillation to the Granger causality analysis, we provide the conventional Granger causality in the supplementary material (Additional file 1) and proceed directly to the formulation of harmonic Granger Causality.

Consider two time series X_t and Y_t , a general form of an autoregressive model with environment inputs (sinusoidal form) has the following vector autoregressive representation:

$$\begin{cases} X_t = \sum_{j=1}^p a_{1j} X_{t-j} + C_{1x} \cos(2\pi f_{x1}t + \phi_x) + \epsilon_{1t} \\ Y_t = \sum_{j=1}^p b_{1j} Y_{t-j} + C_{1y} \cos(2\pi f_{y1}t + \phi_y) + \epsilon_{2t} \end{cases} \quad (7)$$

A joint autoregressive representation having information of past measurements of both time series X_t and Y_t can be written as

$$\begin{cases} X_t = \sum_{j=1}^p a_{2j} X_{t-j} + \sum_{j=1}^p b_{2j} Y_{t-j} + C_{2x} \cos(2\pi f_{x2}t + \phi_x) + \epsilon_{3t} \\ Y_t = \sum_{j=1}^p c_{2j} X_{t-j} + \sum_{j=1}^p d_{2j} Y_{t-j} + C_{2y} \cos(2\pi f_{y2}t + \phi_y) + \epsilon_{4t} \end{cases} \quad (8)$$

where p is the maximum number of lagged observations in the model. ϵ_{it} , $i = 1, 2, 3, 4$, are prediction errors with variance Σ_i , which are uncorrelated over time. The value of Σ_1 measures the accuracy of the autoregressive prediction of X based on its previous values and the harmonic term, whereas the value of Σ_3 represents the accuracy of predicting present value of X based on previous measurements of both X and Y and the harmonic term. According to the causality definition of Granger, if the prediction of one process is improved by incorporating past information of the second process, then the second process causes the first process. In other words, if the variance of the prediction error for the first process is reduced by the inclusion of the past histories of the second process then a causal relation from the second process to the first process exists. This causal influence is quantified by

$$F_{Y \rightarrow X} = \log \frac{\Sigma_1}{\Sigma_3} \quad (9)$$

It is clear that $F_{Y \rightarrow X} = 0$ when there is no causal influence from Y to X and $F_{Y \rightarrow X} > 0$ when there is. Similarly, define causal influence from X to Y as

$$F_{X \rightarrow Y} = \log \frac{\Sigma_2}{\Sigma_4} \quad (10)$$

Due to the natural rhythmic dynamics of the experimental recordings, the environmental inputs (denote as E) are represented by the harmonic terms. While the inclusion of the harmonic terms can exclude the periodic influence caused by the environmental inputs, thus the prediction error can be better estimated and truly reflect the interaction between two processes. We can quantify the influence of environmental inputs (E) by recalling the joint autoregressive model of X_t and Y_t .

$$\begin{cases} X_t = \sum_{j=1}^p a_{3j} X_{t-j} + \sum_{j=1}^p b_{3j} Y_{t-j} + \epsilon_{5t} \\ Y_t = \sum_{j=1}^p c_{3j} X_{t-j} + \sum_{j=1}^p b_{3j} Y_{t-j} + \epsilon_{6t} \end{cases} \quad (11)$$

By definition of Granger causality, the causal influence from E to X or Y can be defined as

$$\begin{cases} F_{E \rightarrow X} = \log \frac{\Sigma_5}{\Sigma_3} \\ F_{E \rightarrow Y} = \log \frac{\Sigma_6}{\Sigma_4} \end{cases} \quad (12)$$

Causality with multiple Oscillators

In many aspects, the addition of a single oscillator is a generalization of the Granger causality, however the application of the adaptation to the normal autoregressive approach is limited to only one external driving force. A further generalization considered here consists of adding more oscillators to the AR model. The interpretation of this approach is as follows: the first and most simple interpretation is that the additional oscillators represent more external oscillatory driving forces, thus being a mere extension of the single oscillator case. A more enlightening and more useful interpretation is that multiple oscillators represent a 'field' of unknown influences upon the network. We know from Fourier theory that any function or signal can be represented by a (possibly infinite) summation of sinusoids, therefore the addition of multiple oscillators in this fashion can, in theory, account for any incident influence upon each of the variables within the system. This interpretation has some rather useful applications. Consider a large sparsely connected network. It is a typical scenario that due to some experimental limitations, we can only record a small proportion of information in the network as a whole. Ideally we wish to reconstruct the structure of the subnetwork for which we have recorded. Hopefully by considering multiple oscillators, this Fourier-like method will provide an avenue to recover the structure of the subnetwork.

By analogy with the single oscillator case, in multiple oscillator scenario there exist a number of unknown external inputs about which we can obtain no information. These unknown inputs and their influences are to be approximated by the summation of many oscillators. In order to calculate the interaction from Y_i to X with many external oscillators and known variables, we can write the equations as

$$X(t) = \sum_{j=1-i}^M \sum_{t'=1}^p a_{t',j} Y_{t-t'}^j + \sum_{k=1}^N (A_k \cos(2\pi f_k t) + B_k \sin(2\pi f_k t)) + \epsilon_{nc} \quad (13)$$

$$X(t) = \sum_{j=1}^M \sum_{t'=1}^p a_{t',j} Y_{t-t'}^j + \sum_{k=1}^N (A_k \cos(2\pi f_k t) + B_k \sin(2\pi f_k t)) + \epsilon_c \quad (14)$$

In Eq. 14, all known variables (Y) are included in the AR terms, while in Eq. 13, the variable Y_i is excluded in the AR terms. M is the number of known variables, p is the total number of previous time steps included and N is the number of oscillators considered in the estimation. The errors associated with noncausal and causal estimations are ϵ_{nc} and ϵ_c respectively.

The level of causality from $Y_i \rightarrow X$ is quantified as:

$$F_{Y_i \rightarrow X} = \log \left(\frac{\text{var}(\epsilon_{nc})}{\text{var}(\epsilon_c)} \right) \quad (15)$$

If $F_{Y_i \rightarrow X} = 0$ then there is no causal influence from Y_i to X . If $F_{Y_i \rightarrow X} > 0$ then there is a causal influence from Y_i to X .

Causality in the Frequency Domain

The key of information extraction is to switch from temporal domain to frequency domain in which their information content can usually become more obvious. Fourier transform provides spectral power that identifies the amplitudes of sine functions of various frequencies that exist throughout the entire duration of the signal. The time domain Granger causality and the harmonic modification can be transformed into the frequency domain, whereby we can obtain the causality spectra showing the frequencies at which the influence of one variable is exerted on another. Expressing the harmonic time series approximations in matrix format leads to the following expression:

$$\begin{bmatrix} a(L) & b(L) \\ c(L) & d(L) \end{bmatrix} \begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \begin{bmatrix} E_x + O_x \\ E_y + O_y \end{bmatrix} \quad (16)$$

where the summation of over the time lags is implied such that $a(L)X = \sum_{j=1}^{\infty} a_j X_{t-j}$. L is the lag operator. And the zeroth terms of the coefficient matrix are such that $a(0) = 1$, $b(0) = 0$, $c(0) = 0$ and $d(0) = 1$. O_x and O_y are the harmonic terms. Take the Fourier transform on both sides of this matrix equation and then multiply by the inverse of the matrix, then express $X(\omega)$ and $Y(\omega)$ in terms of the error and harmonic oscillations, we can obtain the transfer function:

$$\begin{bmatrix} X(\omega) \\ Y(\omega) \end{bmatrix} = \begin{bmatrix} H_{xx}(\omega) & H_{xy}(\omega) \\ H_{yx}(\omega) & H_{yy}(\omega) \end{bmatrix} \begin{bmatrix} E_x(\omega) + O_x(\omega) \\ E_y(\omega) + O_y(\omega) \end{bmatrix} \quad (17)$$

Now the spectra of $X(\omega)$ and $Y(\omega)$ can be can be derived as

$$\begin{cases} X(\omega) = H_{xx}(\omega)(O_x(\omega) + E_x(\omega)) + H_{xy}(\omega)(O_y(\omega) + E_y(\omega)) \\ Y(\omega) = H_{yx}(\omega)(O_x(\omega) + E_x(\omega)) + H_{yy}(\omega)(O_y(\omega) + E_y(\omega)) \end{cases} \quad (18)$$

Thus the spectra are given by:

$$\begin{cases} S_{xx} = X(\omega)X^*(\omega) \\ S_{yy} = Y(\omega)Y^*(\omega) \end{cases} \quad (19)$$

It is instructive to investigate the components which constitute the spectra of $X(\omega)$ and $Y(\omega)$. Expanding the expression for S_{xx} and S_{yy} yields an equation with 16 terms dependent upon the errors terms, $(E_x(\omega), E_y(\omega))$, the harmonic terms, $(O_x(\omega), O_y(\omega))$, the transfer functions, H_{xx} , H_{xy} , H_{yx} , H_{yy} , and their complex conjugates. For the X channel, these components are as follows:

$$\begin{aligned} & H_{xx}O_xO_x^*H_{xx}^* \quad , \quad H_{xx}O_xE_x^*H_{xx}^* \\ & H_{xx}O_xO_y^*H_{xy}^* \quad , \quad H_{xx}O_xE_y^*H_{xy}^* \\ & H_{xx}E_xO_x^*H_{xx}^* \quad , \quad H_{xx}E_xE_x^*H_{xx}^* \\ & H_{xx}E_xO_y^*H_{xy}^* \quad , \quad H_{xx}E_xE_y^*H_{xy}^* \\ & H_{xy}O_yO_x^*H_{xx}^* \quad , \quad H_{xy}O_yE_x^*H_{xx}^* \\ & H_{xy}O_yO_y^*H_{xy}^* \quad , \quad H_{xy}O_yE_y^*H_{xy}^* \\ & H_{xy}E_yO_x^*H_{xx}^* \quad , \quad H_{xy}E_yE_x^*H_{xx}^* \\ & H_{xy}E_yO_y^*H_{xy}^* \quad , \quad H_{xy}E_yE_y^*H_{xy}^* \end{aligned} \quad (20)$$

Each element of the spectra of S_{xx} can be thought of as either intrinsic (caused by the X_t), causal (caused by Y_t) or cross terms (caused by X_t and Y_t). Thus

$$S_{xx} = S_{intrinsic} + S_{causal} + S_{cross} \quad (21)$$

where

$$\begin{aligned} S_{intrinsic} &= H_{xx}O_xO_x^*H_{xx}^* + H_{xx}O_xE_x^*H_{xx}^* + H_{xx}E_xO_x^*H_{xx}^* + H_{xx}E_xE_x^*H_{xx}^* \\ S_{causal} &= H_{xy}O_yO_y^*H_{xy}^* + H_{xy}O_yE_y^*H_{xy}^* + H_{xy}E_yO_y^*H_{xy}^* + H_{xy}E_yE_y^*H_{xy}^* \\ S_{cross} &= H_{xx}O_xO_y^*H_{xy}^* + H_{xx}O_xE_y^*H_{xy}^* + H_{xx}E_xO_y^*H_{xy}^* + H_{xx}E_xE_y^*H_{xy}^* \\ &\quad + H_{xy}O_yO_x^*H_{xx}^* + H_{xy}O_yE_x^*H_{xx}^* + H_{xy}E_yO_x^*H_{xx}^* + H_{xy}E_yE_x^*H_{xx}^* \end{aligned}$$

In the absence of the harmonic oscillators, there are only four terms in the expression for S_{xx} and S_{yy} , and the cross term can be eliminated using the transformation proposed by Geweke [35,36]. Eliminating the cross term is essential for a consistent definition of the frequency domain causality, however the addition of the harmonic terms makes the prospect of removing the cross term rather troublesome. In the harmonic case, it is not possible, in general to eliminate the cross terms by means of transformation due to the presence of O_x and O_y terms. The oscillation terms $O_{x,y}$ are sinusoidal indicating that the Fourier transforms of these functions are delta functions, and the discontinuous nature of the delta functions makes it impossible to find a transformation eliminate all the cross terms. The method we use to eliminate the cross terms is as follows: firstly we apply the approximation of the Geweke transformation:

$$\begin{bmatrix} 1 & 0 \\ -\frac{\gamma^2}{\Sigma^2} & 1 \end{bmatrix} \begin{bmatrix} a(\omega) & b(\omega) \\ c(\omega) & d(\omega) \end{bmatrix} \begin{bmatrix} X(\omega) \\ Y(\omega) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -\frac{\gamma^2}{\Sigma^2} & 1 \end{bmatrix} \begin{bmatrix} E_x(\omega) + O_x(\omega) \\ E_y(\omega) + O_y(\omega) \end{bmatrix} \quad (22)$$

$$\approx \begin{bmatrix} 1 & 0 \\ -\frac{\gamma^2}{\Sigma^2} & 1 \end{bmatrix} \begin{bmatrix} E_x(\omega) \\ E_y(\omega) \end{bmatrix} + \begin{bmatrix} O_x(\omega) \\ O_y(\omega) \end{bmatrix} \quad (23)$$

where γ^2 is the covariance matrix between X and Y , Σ is the variance of either error E_x or E_y .

The step from Eq. 22 to Eq. 23 suggests that the transfer matrix H is an approximate to the true value. This estimate is necessary to ensure that the causality has a consistent

definition. The transfer equation is now as follows:

$$\begin{bmatrix} X(\omega) \\ Y(\omega) \end{bmatrix} \approx \begin{bmatrix} \tilde{H}_{xx} & \tilde{H}_{xy} \\ \tilde{H}_{yx} & \tilde{H}_{yy} \end{bmatrix} \begin{bmatrix} \tilde{E}_x(\omega) \\ \tilde{E}_y(\omega) \end{bmatrix} + \begin{bmatrix} O_x(\omega) \\ O_y(\omega) \end{bmatrix} \quad (24)$$

We combined Eq. 18 and Eq. 24 to define $X(\omega)$ as follows:

$$X(\omega) = \tilde{H}_{xx}(\omega)E_x(\omega) + \tilde{H}_{xx}(\omega)O_x(\omega) + \tilde{H}_{xy}(\omega)\tilde{E}_x(\omega) + \tilde{H}_{xy}(\omega)O_y(\omega) \quad (25)$$

Where $\tilde{H}_{xx} = H_{xx} + \frac{\gamma}{\Sigma}H_{xy}$, $\tilde{H}_{xy} = H_{xy}$ and $\tilde{E}_x = E_x - \frac{\gamma}{\Sigma}E_y$. This has the effect of nullifying the cross terms which contain the element of error. Then the cross terms (components of S_{cross}) are reallocated either to S_{causal} or $S_{intrinsic}$ in the following fashion:

$$\begin{aligned} \tilde{S}_{intrinsic} &= \tilde{H}_{xx}O_xO_x^*\tilde{H}_{xx}^* + \tilde{H}_{xx}O_x\tilde{E}_x^*\tilde{H}_{xx}^* + \tilde{H}_{xx}\tilde{E}_xO_x^*\tilde{H}_{xx}^* + \tilde{H}_{xx}\tilde{E}_x\tilde{E}_x^*\tilde{H}_{xx}^* + \\ &\quad \tilde{H}_{xx}O_x\tilde{E}_y^*\tilde{H}_{xy}^* + \tilde{H}_{xx}O_xO_y^*\tilde{H}_{xy} + \tilde{H}_{xy}\tilde{E}_yO_x^*\tilde{H}_{xx} \\ \tilde{S}_{causal} &= \tilde{H}_{xy}O_yO_y^*\tilde{H}_{xy}^* + \tilde{H}_{xy}O_y\tilde{E}_y^*\tilde{H}_{xy}^* + \tilde{H}_{xy}\tilde{E}_yO_y^*\tilde{H}_{xy}^* + \tilde{H}_{xy}\tilde{E}_y\tilde{E}_y^*\tilde{H}_{xy}^* + \\ &\quad \tilde{H}_{xx}\tilde{E}_xO_y^*\tilde{H}_{xy}^* + \tilde{H}_{xy}O_yO_x^*\tilde{H}_{xx}^* + \tilde{H}_{xy}O_y\tilde{E}_x^*\tilde{H}_{xx} \\ \tilde{S}_{cross} &= \tilde{H}_{xx}\tilde{E}_x\tilde{E}_y^*\tilde{H}_{xy}^* + \tilde{H}_{xy}\tilde{E}_y\tilde{E}_x^*\tilde{H}_{xx}^* \end{aligned}$$

After the reallocation of the components of the spectrum the resultant causality is an approximation rather than a precise calculation, yet it can be shown to yield convincing and consistent results. Using these two methods to approximate the spectrum, we have obtained the spectrum in such a format $\tilde{S}_{xx} = \tilde{S}_{intrinsic} + \tilde{S}_{causal}$ as the \tilde{S}_{cross} term is negligible. In the case where the harmonic term is not present, the causality is defined as:

$$f_{X \rightarrow Y} = \log \left(\frac{S_{xx}}{S_{intrinsic}} \right) \quad (26)$$

Yet, it is essential that the causality in the harmonic case is defined in terms of both $S_{intrinsic}$ and S_{causal} . Therefore, by analogy to the normal frequency domain causality (without harmonic terms), the frequency domain causality in the harmonic case is defined

as:

$$\begin{aligned}\tilde{f}_{X \rightarrow Y} &= \log \left(\frac{\tilde{S}_{xx}}{\tilde{S}_{intrinsic}} \right) \\ &= \log \left(\frac{1 + \tilde{S}_{intrinsic}}{\tilde{S}_{causal}} \right)\end{aligned}\tag{27}$$

In the Results section, we will show through examples that this approximation produces an excellent estimate of the frequency domain causality.

Authors contributions

JHW carried out model development and data analysis and wrote up the manuscript. JLS carried out the model development and data analysis. VBW provided the leaf gene data. JFF conceived of the study, participated in design, supervised the studies. All authors read and approved the final manuscript.

Acknowledgements

We thank Jing Kang to read the manuscript and give critical comments. JF was supported by an EPSRC (UK) grant CARMAN and an EU grant BION.

References

1. Alon U: **Biological networks: the tinkerer as an engineer.** *Science* 2003, **301**(5641):1866-1867.
2. Gardner TS, di Bernardo D, Lorenz D, Collins JJ: **Inferring genetic networks and identifying compound mode of action via expression profiling.** *Science* 2003, **301**(5629):102-105.
3. Lopes da Silva F, Pijn JP, Boeijinga P: **Interdependence of eeg signals: linear vs. nonlinear associations and the significance of time delays and phase shifts.** *Brain Topogr* 1989, **2**(1-2):9-18.
4. Kocsis B, Bragin A, Buzsaki G: **Interdependence of multiple theta generators in the hippocampus: a partial coherence analysis.** *J Neurosci* 1999, **19**(14):6200-6212.
5. Albo Z, Di Prisco GV, Chen YH, Rangarajan G, Truccolo W, Feng JF, Vertes RP, Ding MZ: **Is partial coherence a viable technique for identifying generators of neural oscillations?** *Biological Cybernetics* 2004, **90**(5):318-326.

6. Schelter B, Winterhalder M, Eichler M, Peifer M, Hellwig B, Guschlbauer B, Lucking CH, Dahlhaus R, Timmer J: **Testing for directed influences among neural signals using partial directed coherence.** *J Neurosci Methods* 2006, **152**(1-2):210-219.
7. Wu J, Kendrick K, Feng J: **Detecting correlation changes in electrophysiological data.** *J Neurosci Methods* 2007, **161**(1):155-165.
8. Friedman N, Linial M, Nachman I, Pe'er D: **Using Bayesian networks to analyze expression data.** *Journal of Computational Biology* 2000, **7**(3-4):601-620.
9. Granger CWJ: **Investigating Causal Relations by Econometric Models and Cross-Spectral Methods.** *Econometrica* 1969, **37**(3):414-428.
10. Granger CWJ: **Testing for Causality - A Personal Viewpoint.** *Journal of Economic Dynamics & Control* 1980, **2**(4):329-352.
11. Gourevitch B, Le Bouquin-Jeannes R, Faucon G: **Linear and nonlinear causality between signals: methods, examples and neurophysiological applications.** *Biological Cybernetics* 2006, **95**(4):349-369.
12. Ding M, Chen, Yonghong, Bressler, S.L.: **Granger Causality: Basic Theory and Application to Neuroscience.** In *Handbook of Time Series Analysis*. Edited by Schelter B. WM, Timmer J.: Wiley-VCH Verlage; 2006:451-474.
13. Seth AK: **Causal networks in simulated neural systems.** *Cogn Neurodyn* 2008, **2**(1):49-64.
14. Chen YH, Bressler SL, Ding MZ: **Frequency decomposition of conditional Granger causality and application to multivariate neural field potential data.** *J Neurosci Methods* 2006, **150**(2):228-237.
15. Guo S, Seth AK, Kendrick KM, Zhou C, Feng J: **Partial Granger causality—eliminating exogenous inputs and latent variables.** *J Neurosci Methods* 2008, **172**(1):79-93.
16. Marinazzo D, Pellicoro M, Stramaglia S: **Kernel-Granger causality and the analysis of dynamical networks.** *Physical Review E* 2008, **77** (056215):1-9.
17. Wu J, Kendrick K, Feng J: **A novel approach to detect Hot-Spots in large-scale multivariate data.** *BMC Bioinformatics*. 2007, 8:331 doi:10.1186/1471-2105-8-331.
18. Guo S, Wu J, Ding M, Feng J: **Uncovering interactions in the frequency domain.** *PLoS Comput Biol* **4**(5): e1000087.
19. Chance B, Estabrook RW, Ghosh A: **Damped sinusoidal oscillations of cytoplasmic reduced pyridine nucleotide in yeast cells.** *Proc Natl Acad Sci U S A* 1964, **51**:1244-1251.
20. Chrobak JJ, Buzsaki G: **Gamma oscillations in the entorhinal cortex of the freely behaving rat.** *J Neurosci* 1998, **18**(1):388-398.

21. Reppert SM, Weaver DR: **Molecular analysis of mammalian circadian rhythms.** *Annu Rev Physiol* 2001, **63**:647-676.
22. Buzsaki G: *Rhythms of the Brain.* Oxford: Oxford University Press; 2006.
23. Yan J, Wang HF, Liu YT, Shao CX: **Analysis of Gene Regulatory Networks in the Mammalian Circadian Rhythm.** *PLoS Comput Biol* 2008, **4**(10):13.
24. Tiao GC, Grupe MR: **Hidden Periodic Autoregressive-Moving Average Models in Time-Series Data.** *Biometrika* 1980, **67**(2):365-373.
25. He S: **Estimation of the mixed ar and hidden periodic model.** *Acta Mathematicae Applicatae Sinica* 1997, **13**(2):196-208.
26. Brovelli A, Ding M, Ledberg A, Chen Y, Nakamura R, Bressler SL: **Beta oscillations in a large-scale sensorimotor cortical network: directional influences revealed by Granger causality.** *Proc Natl Acad Sci U S A* 2004, **101**(26):9849-9854.
27. Wu J, Liu X, Feng J: **Detecting causality between different frequencies.** *J Neurosci Methods* 2008, **167**(2):367-375.
28. Tass P, Rosenblum MG, Weule J, Kurths J, Pikovsky A, Volkmann J, Schnitzler A, Freund HJ: **Detection of n : m phase locking from noisy data: Application to magnetoencephalography.** *Physical Review Letters* 1998, **81**(15):3291-3294.
29. Lachaux JP, Rodriguez E, Martinerie J, Varela FJ: **Measuring phase synchrony in brain signals.** *Human Brain Mapping* 1999, **8**(4):194-208.
30. Rubiola E: **Phase Noise and Frequency Stability in Oscillators.** In *The Cambridge RF and Microwave Engineering Series.* Cambridge University Press, 2008.
31. Locke JC, Kozma-Bognar L, Gould PD, Feher B, Kevei E, Nagy F, Turner MS, Hall A, Millar AJ: **Experimental validation of a predicted feedback loop in the multi-oscillator clock of Arabidopsis thaliana.** *Mol Syst Biol* 2006, **2**:59.
32. Ueda HR: **Systems biology flowering in the plant clock field.** *Mol Syst Biol* 2006, **2**:60.
33. Ueda HR, Chen WB, Adachi A, Wakamatsu H, Hayashi S, Takasugi T, Nagano M, Nakahama K, Suzuki Y, Sugano S et al: **A transcription factor response element for gene expression during circadian night.** *Nature* 2002, **418**(6897):534-539.
34. Akaike H: **New Look at Statistical-Model Identification.** *IEEE Transactions on Automatic Control* 1974, **19**(6):716-723.
35. Geweke J: **Measurement of linear-dependence and feedback between multiple time series.** *Journal of the American Statistical Association* 1982, **77**(378):304-313.
36. Geweke J: **Measures of conditional linear-dependence and feedback between time series.** *Journal of the American Statistical Association* 1984, **79**(388):907-915.

Figures

Figure 1

Validation of the hidden harmonic methodology. Four different model configurations are considered. In each configuration node X has a causal influence of node Y. In (i) both of the nodes have environmental inputs, in (ii) and (iii) the simulation has just node X or node Y have an external input, in (iv) neither X or Y have external inputs. (A) shows the connection configurations. In (B) and (C) the power spectra of each configuration are shown in frequency domain, node X is shown in blue, node Y is shown in green. (D) and (E) show the causality spectra from X to Y, $f_{X \rightarrow Y}$, the causality calculated using the normal Granger AR method is shown in column (D), the harmonic causal method is shown in column (E).

Figure 2

The Effects of Varying Input Phase. A): shows the model configuration where X has a causal influence upon Y and there is no feedback. Both X and Y have an external, environmental oscillatory input. B): shows approximately 1 sec of simulated time domain plots the blue trace is X_t , the green trace is Y_t . In B_1 $\phi_x = \phi_y$ and in B_2 $\delta\phi = \phi_x - \phi_y = \pi$, in both B_1 and B_2 the X trace is identical, yet the trace of Y is shown to be greatly changed simply by altering the phase difference. The difference in the magnitude is denoted by Δ . As the inherent noise will mask the effect of the phase differences, C) shows the effect on Δ by varying both input noise and $\delta\phi$. D_1 shows how this Δ changes by varying $\delta\phi$ in the absence of noise. D_2 shows how the effect of noise in the system lessens the effect of the phase differences.

Figure 3

Each colourmap shows the causality calculated whilst varying noise and input phase difference. A_1 shows the causality calculated using the AR method in the time domain, A_2 shows the causality calculated by integration of the AR method frequency domain causality. B_1 shows the causality calculated using the harmonic method in the time domain, B_2 shows the causality calculated by integration of the harmonic method frequency domain causality.

Figure 4

Investigation of the interrelation between the phase difference of input and output signals and noise level. A): a colourmap plot of the phase difference of output signal against the phase difference of input signal and noise level. The phase difference of output signal is almost uniformly distributed for varying noise level and the phase difference of the input signal. B) and C) demonstrate the averaged phase difference of the output signal against noise level and the phase difference of the input signal, respectively.

Figure 5

Schematic connection plots and estimated variance for different number of oscillators. A): Seven connection schemes for causal and non-causal influence when there are or not harmonic oscillators. B): Using a simple small sparsely connected network consisting 5 nodes. The number of oscillators was increased and various fitting algorithms are applied.

Figure 6

Time domain traces of gene expression of eight genes under scrutiny and possible environment input. A): The time domain plots, in blue is the 16 repetitions of the experimental data and in red is the parameterized fitted data (Causal harmonic fitting). With the exception of the gene GI, each of the gene exhibits high periodicity and high levels of repeatability. B): The 8 genes and the oscillation metric, those with a larger metric are more likely to contain an external oscillatory input.

Figure 7

The errors associated with each of the gene pairs for each of the four candidate models, the causality, $F_{X \rightarrow Y}$ is calculated either with $\log \frac{\Sigma_3}{\Sigma_4}$ or $\log \frac{\Sigma_1}{\Sigma_2}$ depending upon whether the gene has an external input.

Figure 8

A): Candidate gene network calculated using the both the harmonic and non-harmonic schemes. B): reproduced Ueda's candidate model [32,33]. C): Causality spectra for 15 connections inferred from the network.

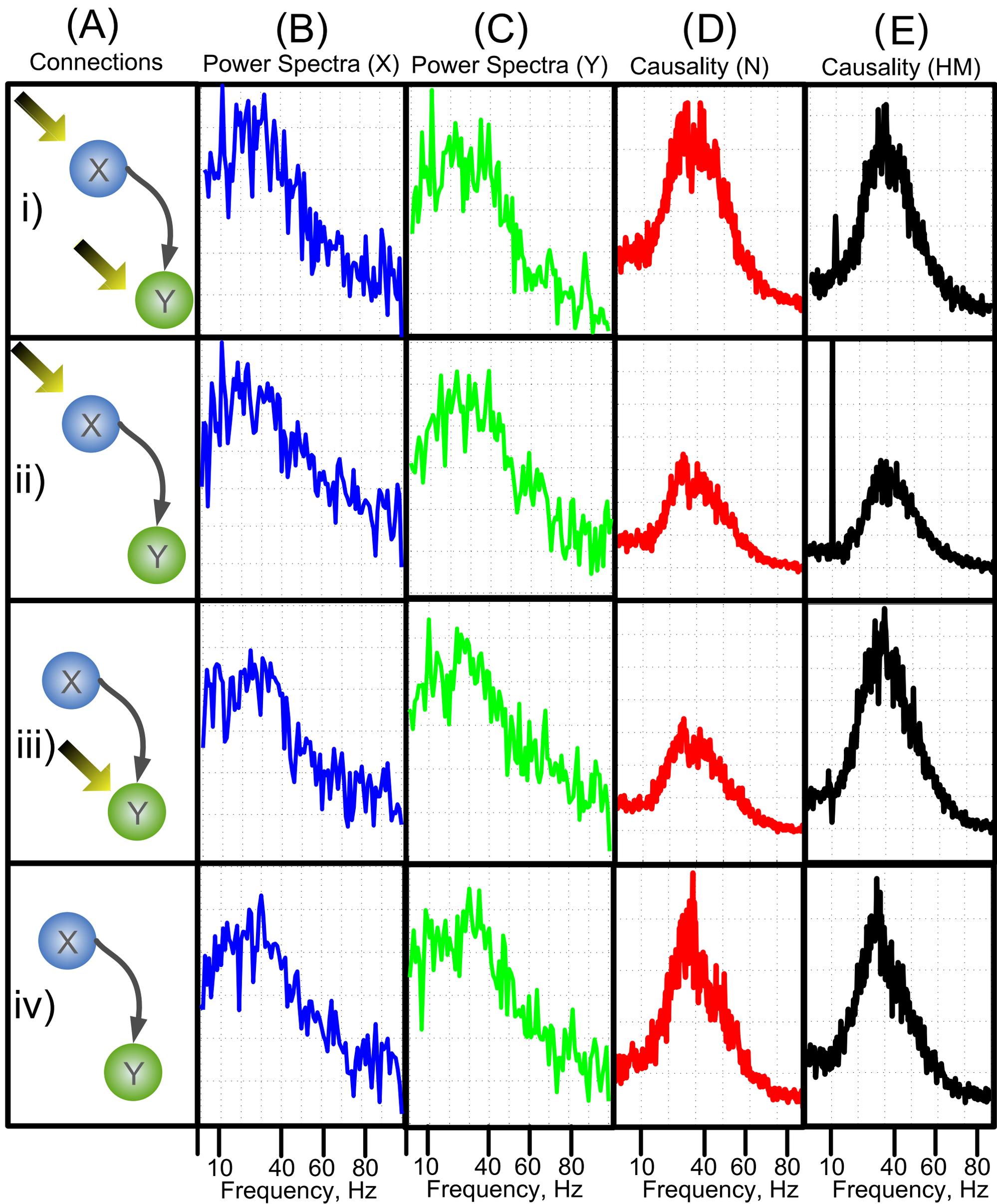


Figure 1

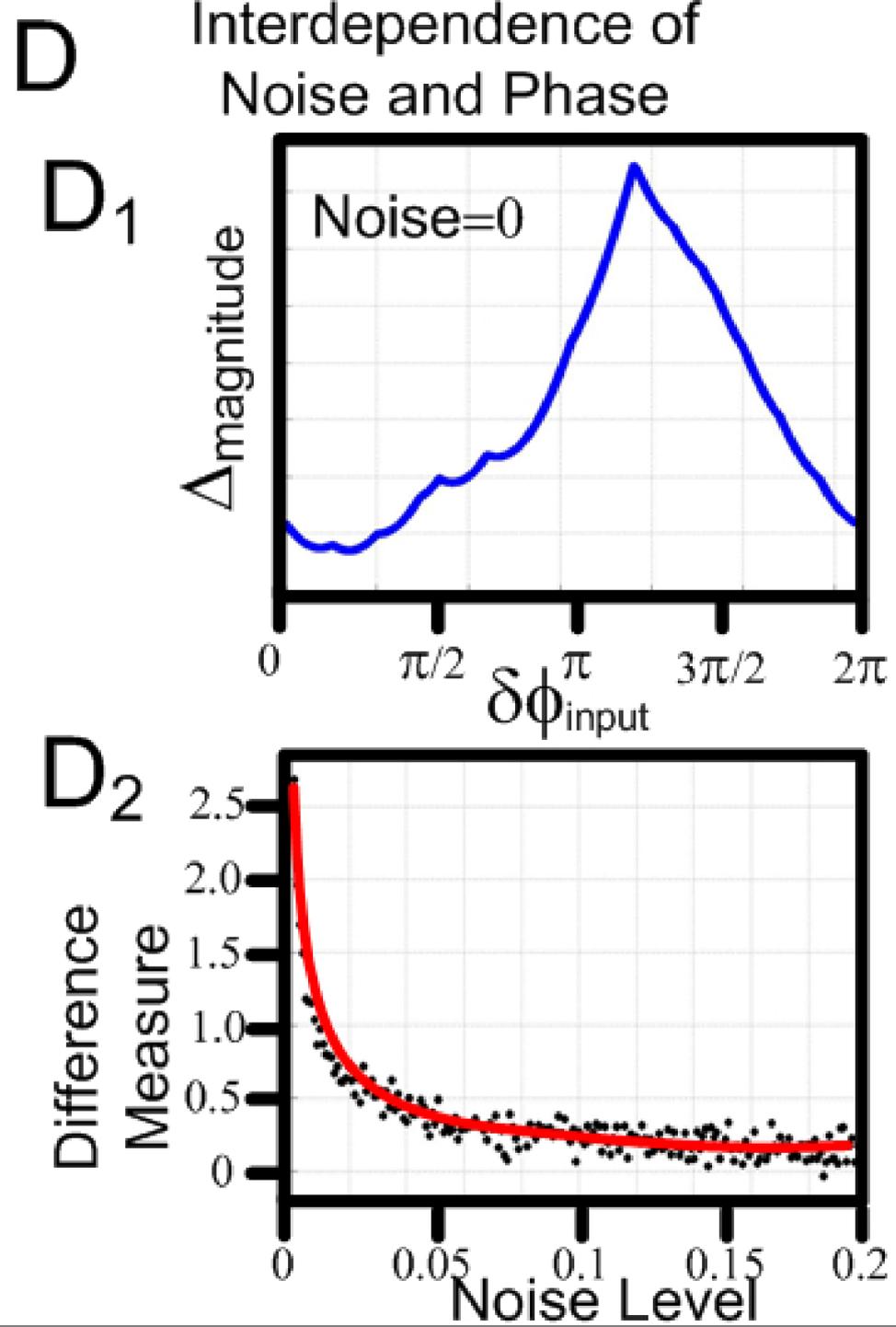
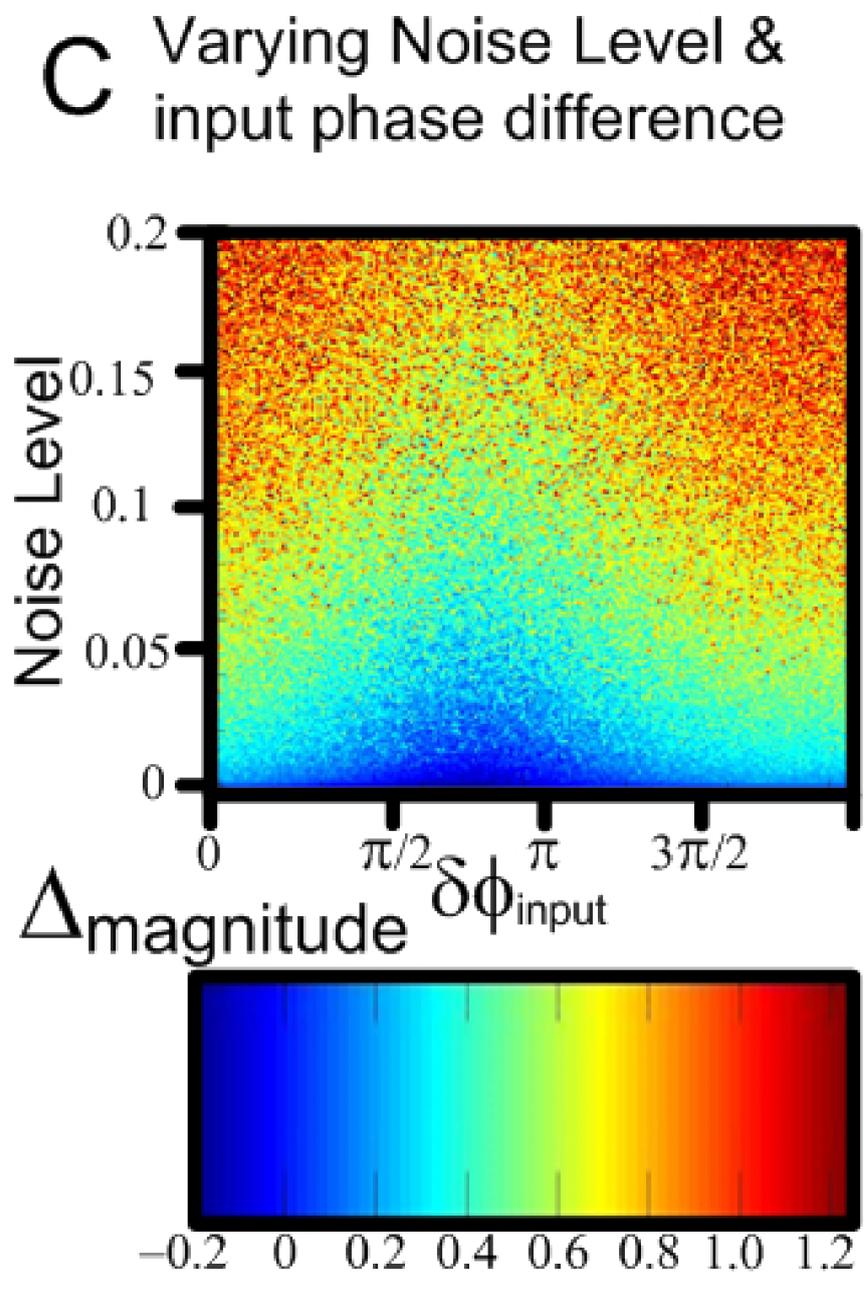
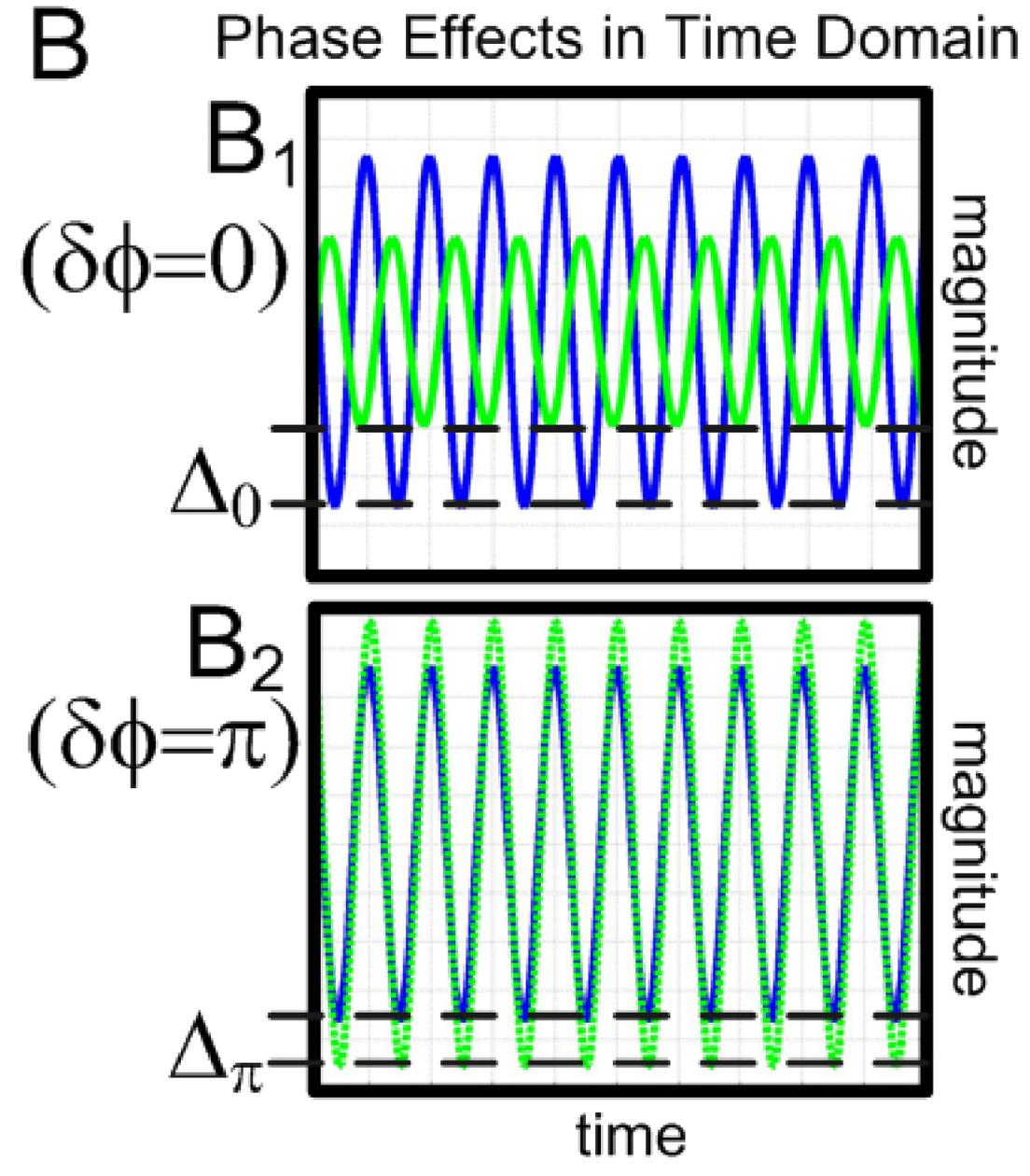
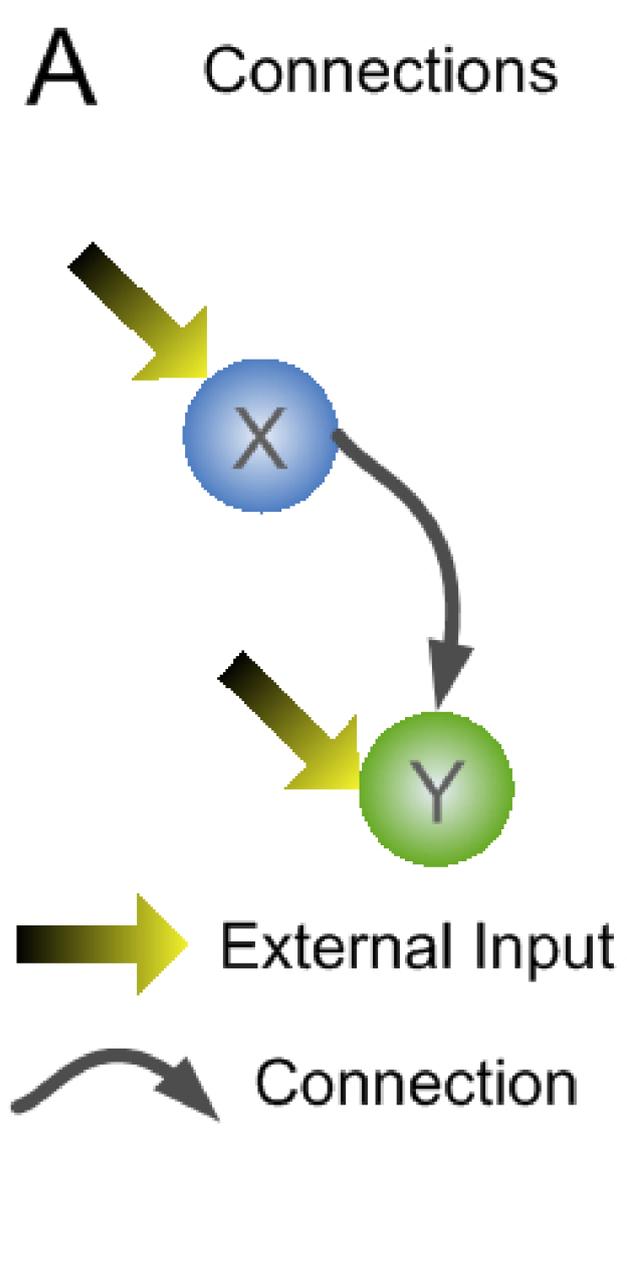


Figure 2

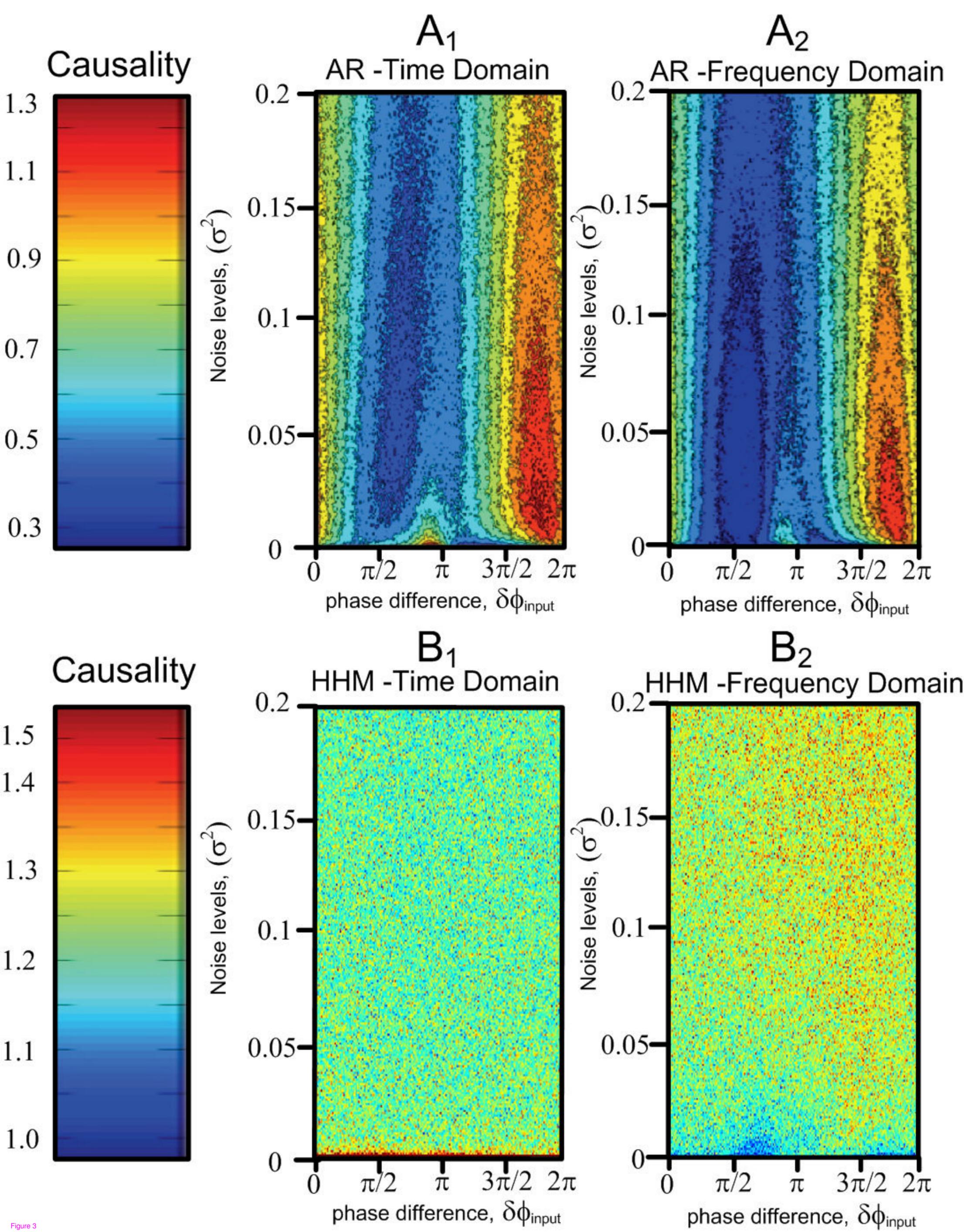


Figure 3

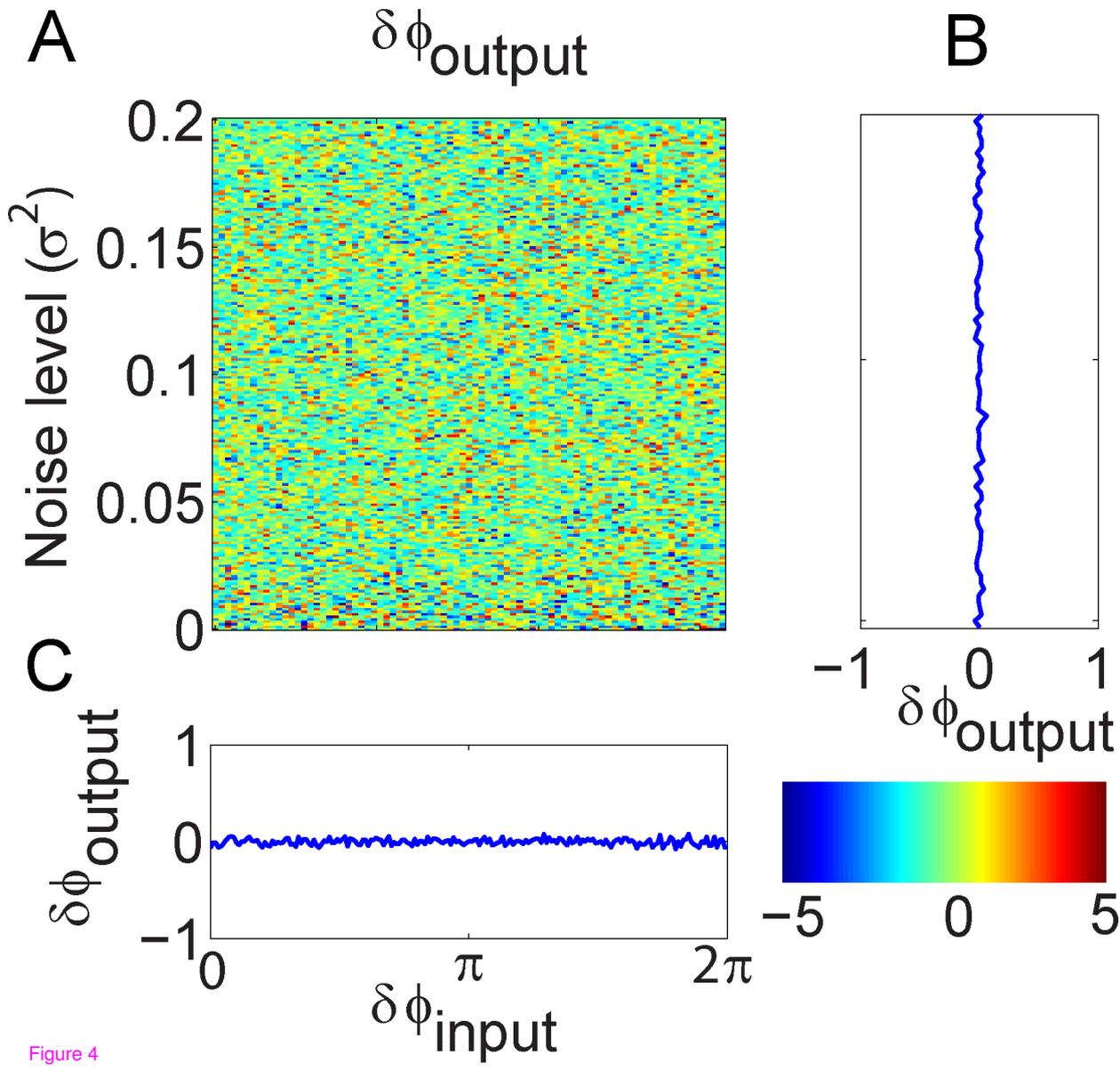
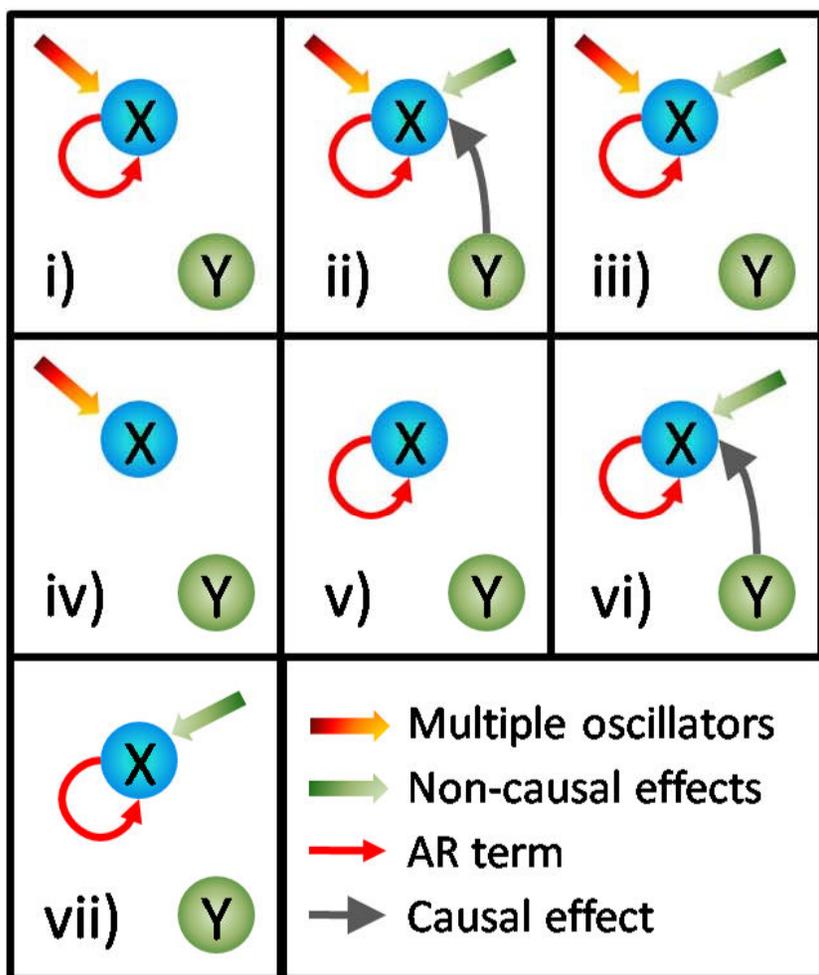
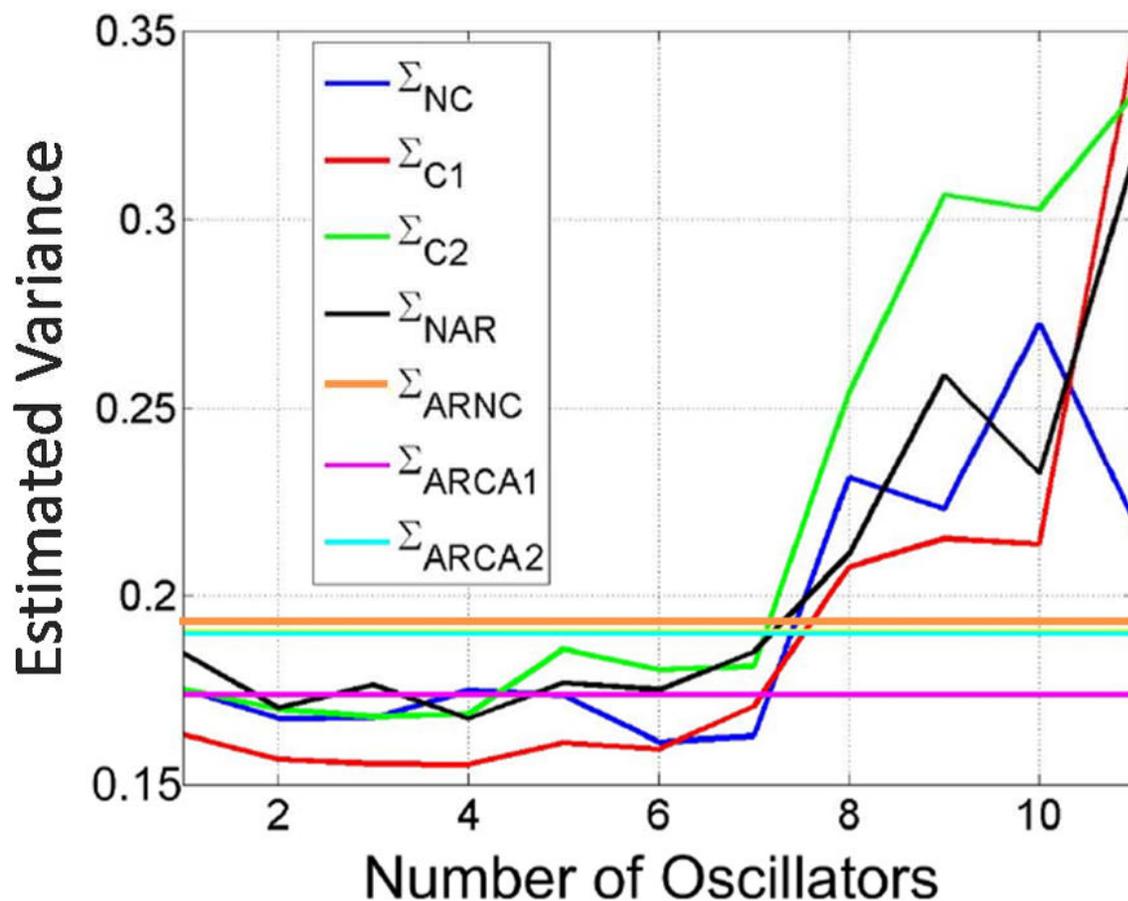


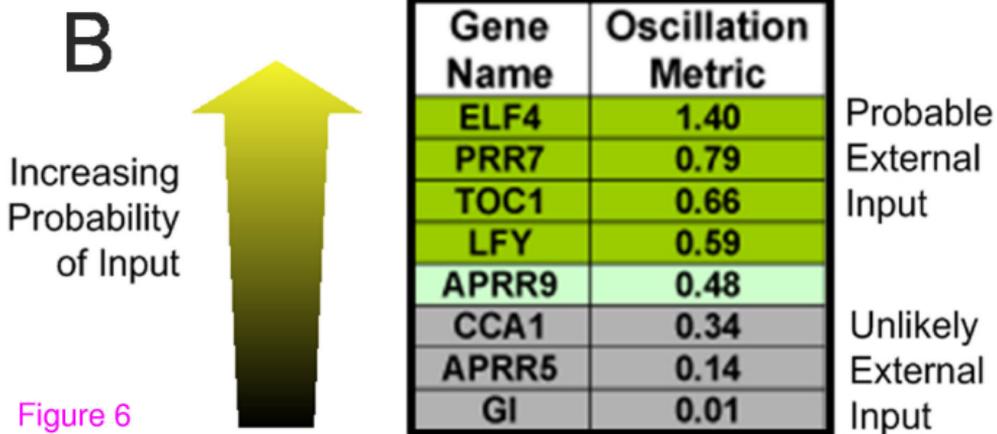
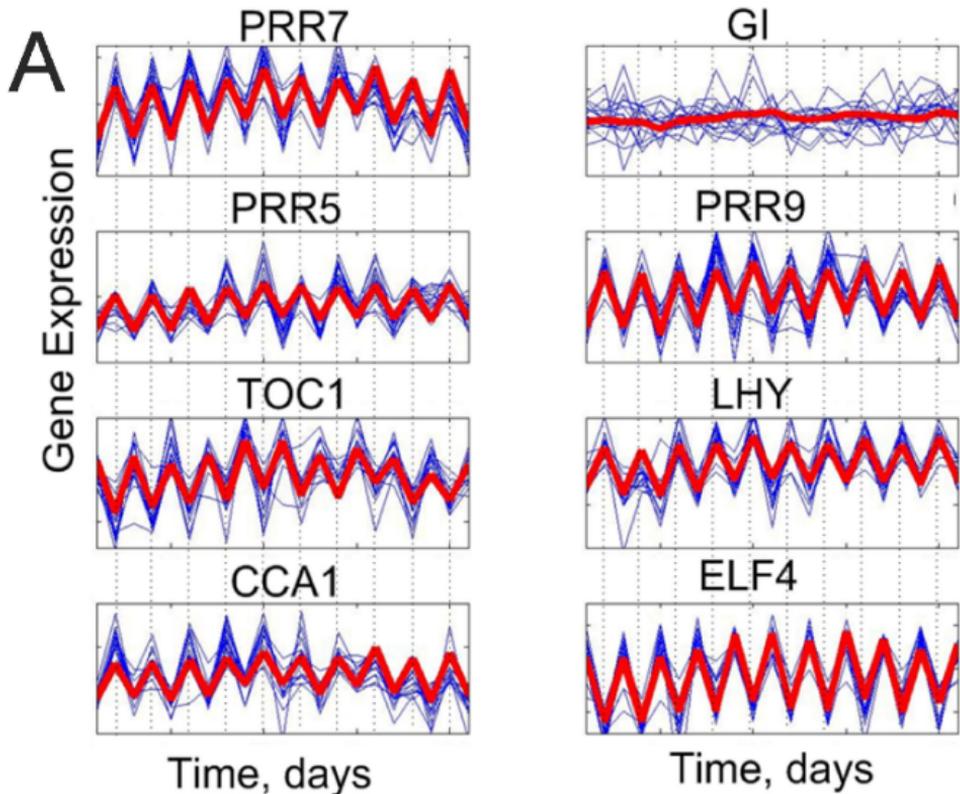
Figure 4

(A) Connection Schemes



(B) Effect of Oscillators





Gene X	Gene Y	Σ_1	Σ_2	Σ_3	Σ_4	$F_{x,y}$
LHY	TOC1	0.14	0.10	0.15	0.10	0.32
APRR9	LHY	0.18	0.14	0.19	0.14	0.28
APRR9	TOC1	0.18	0.15	0.19	0.15	0.23
PRR7	CCA1	0.13	0.11	0.17	0.11	0.23
APRR5	APRR9	0.18	0.15	0.19	0.15	0.23
LHY	GI	0.14	0.11	0.15	0.13	0.22
APRR9	CCA1	0.18	0.15	0.19	0.15	0.22
CCA1	LHY	0.14	0.11	0.15	0.12	0.22
LHY	CCA1	0.14	0.11	0.15	0.12	0.21
APRR5	CCA1	0.18	0.16	0.19	0.16	0.20
APRR5	PRR7	0.18	0.16	0.19	0.16	0.19
LHY	PRR7	0.14	0.11	0.15	0.12	0.19
CCA1	PRR7	0.14	0.12	0.15	0.12	0.18
CCA1	APRR9	0.14	0.12	0.15	0.13	0.17
APRR5	GI	0.18	0.16	0.19	0.16	0.17
APRR5	TOC1	0.18	0.16	0.19	0.16	0.16
APRR9	ELF4	0.18	0.16	0.19	0.16	0.16
PRR7	APRR9	0.13	0.12	0.17	0.13	0.15
ELF4	GI	0.11	0.09	0.13	0.11	0.14
CCA1	ELF4	0.14	0.13	0.15	0.13	0.14
LHY	APRR5	0.14	0.12	0.15	0.13	0.14
CCA1	APRR5	0.14	0.11	0.15	0.13	0.14
PRR7	TOC1	0.13	0.12	0.17	0.12	0.13
APRR5	ELF4	0.18	0.17	0.19	0.17	0.13
ELF4	TOC1	0.11	0.10	0.13	0.11	0.12
TOC1	APRR9	0.11	0.10	0.12	0.11	0.10
APRR5	LHY	0.18	0.16	0.19	0.17	0.10
FGFR7	GI	0.11	0.10	0.12	0.10	0.10

Gene X	Gene Y	Σ_1	Σ_2	Σ_3	Σ_4	$F_{x,y}$
ELF4	LHY	0.11	0.10	0.13	0.12	0.09
PRR7	LHY	0.13	0.12	0.17	0.13	0.09
CCA1	TOC1	0.14	0.14	0.15	0.14	0.09
LHY	APRR9	0.14	0.12	0.15	0.13	0.09
ELF4	PRR7	0.11	0.10	0.13	0.13	0.09
APRR9	GI	0.18	0.16	0.19	0.18	0.08
APRR9	APRR5	0.18	0.15	0.19	0.18	0.08
ELF4	APRR9	0.11	0.10	0.13	0.12	0.07
PRR7	APRR5	0.13	0.12	0.17	0.14	0.07
TOC1	LHY	0.11	0.10	0.12	0.11	0.06
TOC1	CCA1	0.11	0.10	0.12	0.11	0.06
PRR7	ELF4	0.13	0.13	0.17	0.14	0.06
TOC1	PRR7	0.11	0.10	0.12	0.11	0.06
LHY	ELF4	0.14	0.13	0.15	0.13	0.05
GI	ELF4	0.11	0.11	0.12	0.11	0.05
GI	CCA1	0.11	0.11	0.12	0.11	0.05
TOC1	APRR5	0.11	0.10	0.12	0.12	0.04
CCA1	GI	0.14	0.14	0.15	0.14	0.03
ELF4	APRR5	0.11	0.10	0.13	0.11	0.03
PRR7	GI	0.13	0.13	0.17	0.16	0.03
ELF4	CCA1	0.11	0.10	0.13	0.12	0.03
APRR9	PRR7	0.18	0.17	0.19	0.18	0.03
GI	PRR7	0.11	0.11	0.12	0.11	0.03
GI	TOC1	0.11	0.11	0.12	0.11	0.03
GI	APRR9	0.11	0.11	0.12	0.11	0.02
GI	APRR5	0.11	0.11	0.12	0.11	0.02
GI	LHY	0.11	0.11	0.12	0.11	0.01
TOC1	ELF4	0.11	0.11	0.12	0.12	0.01

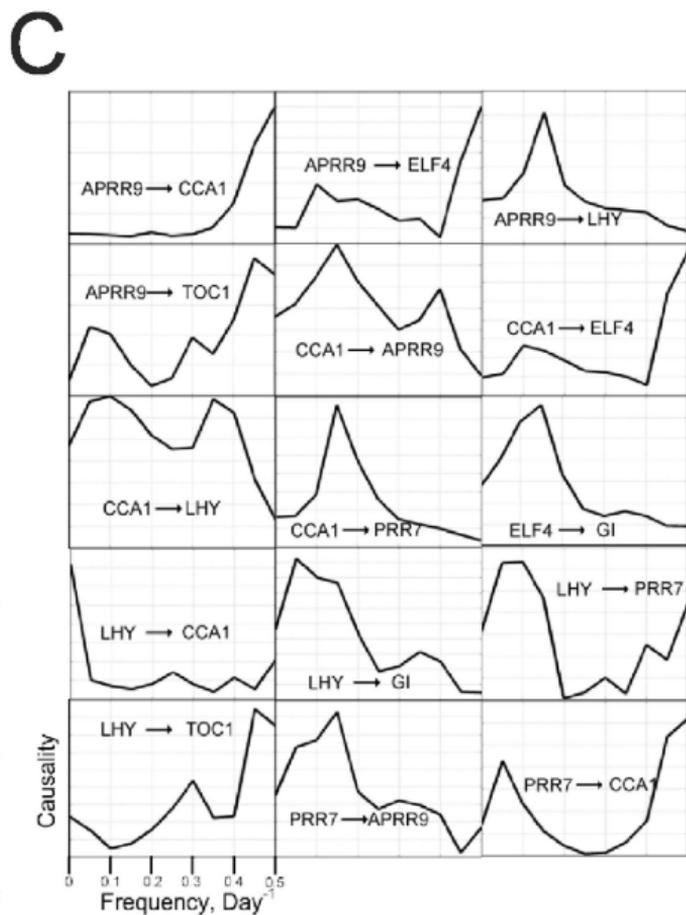
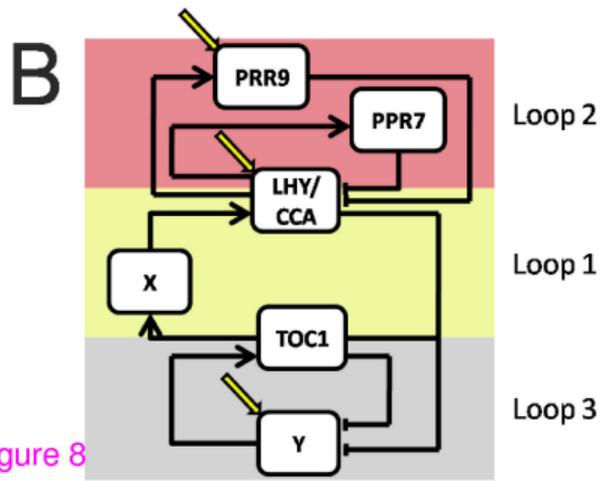
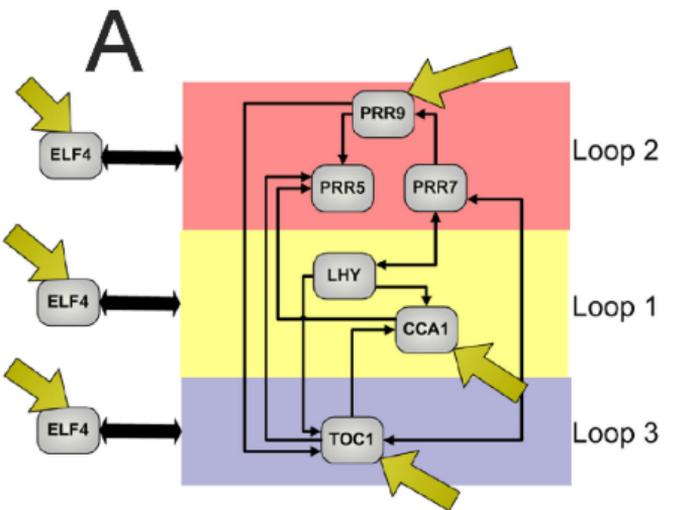


Figure 8

Additional files provided with this submission:

Additional file 1: supplementary material.pdf, 71K

<http://www.biomedcentral.com/imedia/1519807053267423/supp1.pdf>