Experimental design for inference over the A. thaliana circadian clock network

Experimental design for inference over the A. *thaliana* circadian clock network*

Daniel Trejo-Banos[1], Andrew J. Millar[2,3], and Guido Sanguinetti[1,3]

[1] School of Informatics, University of Edinburgh
[2] SynthSys - Systems and Synthetic Biology, University of Edinburgh
[3] School of Biological Sciences, University of Edinburgh

Abstract. Planning experiments is a crucial step in successful investigations, which can greatly benefit from computational modeling approaches. Here we consider the problem of designing informative experiments for elucidating the dynamics of biological networks. Our approach extends previously proposed methodologies to the important case where the structure of the network is also uncertain. We demonstrate our approach on a benchmark scenario in plant biology, the circadian clock network of *Arabidopsis thaliana*, and discuss the different value of three types of commonly used experiments in terms of aiding the reconstruction of the unknown network.

1 Introduction

The execution of experiments to test a hypothesis is the essence of the scientific method. In the field of systems biology we are interested in testing and validating our hypotheses and predictions biochemical processes in living organisms, and our hypotheses are usually encoded in mathematical models which can adopt a variety of formalism. Modern biochemical experiments can be very complex and are often costly in both researcher time and other resources. For this reason, it is important to minimize the number of experiments while maximizing their information content.

Experimental design is the branch of statistics and operations research which is concerned with maximizing the information content of novel experiments. From a statistical point of view, the utility criterion for evaluating an experiment is a function of the probabilistic model chosen to represent the data-generating process. Depending on the objective of the experiment, the selection criterion can be either *m*aximize the information content of an experiment in order to estimate a set of parameters, (*estimation criterion*) or *i*mprove the prediction qualities of a fitted model (*prediction criterion*).

In this paper we use a Bayesian approach to experimental design for dynamical models of biological systems. We restrict our attention to gene regulatory

* DTB is funded by a Microsoft Research Studentship. GS acknowledges support from the European Research Council under grant MLCS30699. SynthSys was founded as a Centre for Integrative Biology by BBSRC/EPSRC award D19621 to AJM and others.

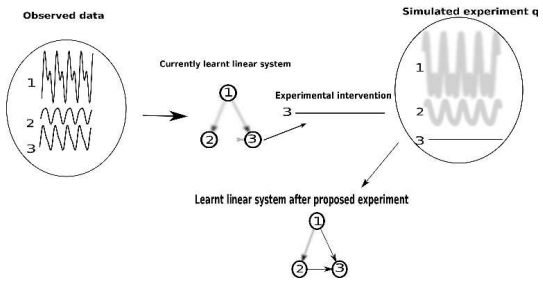

Fig. 1: Basic illustration of our experimental design approach. After a set of observations the distribution over the learnt system (blurred arrows) is used to draw samples of the experimental outcomes given an intervention (uncertainty over the outcomes is also represented by blurred functions). The aim is to choose the experiment that reduces the uncertainty over the learnt system (represented by the system with well defined arrows in the figure).

network (GRN) models, where the systems dynamics are generated by mutual interactions between genes which can modulate each others rate of expression; these models encompass a large fraction of the systems biology literature, and hence experimental design methods for this class of systems are of considerable interest. Dynamical systems such as ordinary differential equations (ODE) are widespread techniques for modeling GRNs. Previous work has considered experimental design and model selection techniques for non-linear ODE- based models of biological processes.

Liepe et al. [5] employ an approach based on mutual information which could be evaluated using Monte-Carlo simulations. This method is computationally intensive and crucially requires prior knowledge over the model components and their interactions: the structure and functional form of the equations defining the models is assumed known, and all the uncertainty is in the parametrisation. In reality, most models in systems biology are subject to considerable structural uncertainty, and clarifying the structure of interactions is the primary goal of systems biology experiments.

In this work we extend the Bayesian experimental design approach to models with structural uncertainty, formalized as hierarchical Bayesian models. We derive a *Bayesian experimental design score* for quantifying the information gain offered by different experiments. The abstract view of the method is shown on Figure 1. We start by using some preliminary data (in the form of observed oscillatory expression levels) to learn a (posterior) probability distribution over a linear approximation of the system. Experimental interventions can be simulated by constraining some components of the model to fixed values (the specific details of how we model interventions are given later), obtaining predictions of

the gene expression levels of all the other components given the experimental intervention (in the figure, the blurred lines represent uncertainty over the experimental outcomes). These enable us to quantify the information content of an intervention.

We illustrate our approach on a benchmark systems biology problem, the circadian clock of the *Arabidopsis thaliana* model plant [9]. We consider three classes of possible experiments: alterations to the light-dark input provided to the plant, direct measurements of regulatory links via chromatin immunoprecipitation (ChIP), and gene knock-outs. These commonly performed experiments are very different in terms of costs, and our preliminary results on their relative informativeness could be useful for practitioners.

2 Methods

Classical approaches to statistical experimental design have been primarily developed for linear regression models. Let an experiment q be given an experimental design Φ^q (usually a set of covariates and a model that accounts for the variables of the experiment) and parameters θ (which determine how each of these covariates determines the measured output of the experiment), and denote the experimental observations for experiment q as $\mathbf{y}^q$. The experimental outputs are assumed to be a linear combination of the covariates such that

$$\mathbf{y}^q = \Phi^q \theta + \epsilon \tag{1}$$

where ϵ is zero-mean Gaussian noise with variance σ^2. The probability of the observed outcomes given a set of parameters θ is known as *likelihood function* (it is a function of the parameters); we will denote it as

$$p\left(\mathbf{y}^q | \Phi^q, \theta\right) = \mathcal{N}\left(\mathbf{y}^q - \Phi^q \theta, \sigma^2\right) \tag{2}$$

The *Fisher information matrix* (FIM) quantifies how much a small change in the parameters θ is expected to affect the likelihood of the observations; mathematically, the FIM is defined as

$$\mathcal{I}_{i,j}(\theta) = E_{p(\mathbf{y}^q | \Phi^q, \theta)}\left[\frac{\partial p\left(\mathbf{y}^q | \Phi^q, \theta\right)}{\partial \theta_i} \frac{\partial p\left(\mathbf{y}^q | \Phi^q, \theta\right)}{\partial \theta_j}\right] \tag{3}$$

where E_q denotes expectation under the distribution q.

The FIM encodes interaction between the observed and the experimental covariates. The most common experimental design objective seeks to select a design Φ^q in order to attain the maximum FIM according to some ordering. For estimation purposes, the optimality criteria depends on the choice of matrix function from which to evaluate the information matrix. The most popular is the *D-optimal* criterion or maximize $\det\left(\mathcal{I}\left(\theta\right)/n\right)$. This criterion minimizes the volume of the confidence ellipsoid of the estimates [4]. A good review of D-optimal design and related criteria can be consulted in [10].

In order to accommodate further uncertainties about experimental covariates and model mis-specification, a different kind of statistical tools is needed. Bayesian methods employ a *prior distribution* over the parameters $p(\theta)$ to incorporate uncertainty in a principled way. This is incorporated with observations to compute the *posterior distribution* by applying Bayes rule which is

$$p(\theta|\mathbf{y}^q, \Phi^q) = \frac{p(\mathbf{y}^q|\theta, \Phi^q)\, p(\theta)}{p(\mathbf{y}^q)}. \tag{4}$$

The denominator in eq. 4 is computed by integrating the likelihood over the prior distribution. *Bayesian experimental design* seeks to leverage prior information about the parameter distribution by averaging over the posterior distribution of the unobserved data samples [2]. For this, we employ the concept of *Mutual information*. In this context we can view the mutual information between θ and $\mathbf{y}^q$ as the reduction in uncertainty about θ that results from observing $\mathbf{y}^q$ [7] . Then, the Bayesian counterpart to D-optimal design maximizes the Mutual information between the parameters distribution and the experimental outcomes [2].

2.1 Bayesian experimental design.

In his seminal work, Lindley [6] sets experimental design in a decision-theory framework. First he states that the previous knowledge over a system is encoded in the prior probability of its model parameters. The knowledge about parameters θ obtained after an experiment, given the observations y^q and experimental conditions ξ^q will be contained in the posterior distribution $p(\theta|y^q, \xi^q)$. Thus the information gained after an experiment can be expressed in terms of the expected *KL-divergence* between both distributions over the distribution of the observations

$$I(\theta; \mathbf{y}^q) = \int KL\left(p(\theta|\mathbf{y}^q) \,\|p(\theta)\right) p(\mathbf{y^q})\, d\mathbf{y^q}.$$

Thus the *utility* of an experiment q with conditions ξ^q(which we will denote by $U(\theta; \mathbf{y}^q; \xi^q)$) is obtained by solving

$$U(\theta; \mathbf{y}^q; \xi^q) = \int\int \log \frac{p(\theta|\mathbf{y}^q, \xi^q)}{p(\theta)} p(\theta, \mathbf{y}^q|\xi^q)\, d\theta dy^q. \tag{5}$$

This utility function gives rise to what is known as *Bayesian D-optimal design* [2]. In order to choose the best experimental design, the objective is to maximize the value of the utility function $U(\theta, y^q, \xi^q)$ over the set of parameters and (unobserved) responses. Unlike classic optimal design, we aim at leveraging prior information encoded in the prior distribution of the parameters.

Whereas these ideas were introduced in the linear regression case, extending to different scenarios is conceptually trivial; however, the computational simplifications afforded by linear models are then lost, giving rise to an analytically intractable problem. Liepe et al. [5] employ the same utility criteria over a set of parameters for a nonlinear system of differential equations and then proceed to

compute the utility function by Monte Carlo simulation. This requires at each step to simulate the experimental outcomes by solving the system, a procedure which may incur in severe computational overhead depending of the model size and parameters. Furthermore, the model structure is assumed fixed; introducing uncertainty in the model structure would add a further dimension to the already complex computational problem, ruling out all but the simplest problems.

In this work, we take the complementary approach of catering for structural uncertainty in the models, while simplifying the dynamics by assuming linearity and time invariance (LTI models). We approach the problem by adopting a probabilistic linear model of the frequency spectrum of the gene expression levels. In the case of oscillating networks, this linear model can offer a reasonable approximation to the system dynamics, and has been shown to be effective in capturing structural uncertainty in a network inference scenario [11]. The advantage of the LTI approximation is that sampling from the experimental outcomes "reduces" to sampling from a Multivariate Normal conditioned on a subset of variables, confining the need for Monte Carlo simulation to integrating out the structural uncertainty.

2.2 Frequency-domain model of gene expression levels.

We briefly review now the LTI approach to modelling GRN dynamics taken in [11]. We start by representing the LTI equations in frequency domain through the *Discrete Fourier Transform* (DFT). Under certain conditions the DFT is a discrete sample of the Fourier spectrum of the signal, see [8]. With this approximation we derive a matrix equation for the linearized network dynamics, this matrix equation is

$$\dot{\mathbf{X}}^q = \mathbf{X}^q \mathbf{A}^T + \mathbf{U}^q \mathbf{C}^T. \tag{6}$$

Here, matrix $\mathbf{X}^q$ is the matrix whose columns represent the DFT coefficients (spectrum) of the expression level samples of a set of N genes for an experiment q. Analogously, $\mathbf{U}^q$ will represent the DFT of the system inputs. We denote by $\dot{\mathbf{X}}^q$ the time derivative of the spectra, which can be computed by the matrix product $\mathbf{DX}$, being $\mathbf{D}$ a derivative operator. The DSS model presented in [11] proposes a Gaussian likelihood regression model for estimating coefficients $\mathbf{A}$ and $\mathbf{C}$ by the distribution of the residues $\mathbf{Q}^q = \dot{\mathbf{X}}^q - \mathbf{X}^q \mathbf{A}^{\mathrm{T}} - \mathbf{U}^q \mathbf{C}^{\mathrm{T}}$ such that

$$\mathrm{p}\left(\mathbf{Q}^q | \sigma_D\right) = \mathcal{N}\left(\dot{\mathbf{X}}^q - \left[\mathbf{X}^q \ \mathbf{U}^q\right] \begin{bmatrix} \mathbf{A}^{\mathrm{T}} \\ \mathbf{C}^{\mathrm{T}} \end{bmatrix}, \sigma_D^2\right).$$

In order to estimate the parameters $\{\mathbf{A}, \mathbf{C}\}$, a sparsity inducing prior is set over these parameters. This prior is a spike and slab distribution of the form presented in [3]: intuitively, this is a mixture distribution where parameters (LTI coefficients) can either be sampled from a distribution concentrated at zero (the spike) or a broad distribution (the slab). Thus, conditioning on data, spike and slab models carry out automatic feature selection by assigning the value zero to irrelevant features (in our case interaction coefficients between non-interacting genes).

This prior encodes the network topology through an adjacency matrix $\mathbf{H}$ within a Hierarchical Bayesian model. We call this model the DFT- Spike and slab (DSS) model of gene expression. The precise details of the model, as well as Bayesian algorithms for network inference within this framework, are provided in [11]. For the purposes of experimental design, it is sufficient to state that this framework provides us with a methodology to recast GRN dynamics in a (Bayesian) regression framework, where the (DFT projection) of the signal derivative is regressed upon the (DFT projection) of the signal. The Hierarchical Bayesian model then provides a structured prior distribution to capture the uncertainty over the underlying networks.

2.3 Experimental design for estimating parameters of a DSS model

Having specified the DSS family of models, we now discuss in detail the experimental design techniques for three classes of experiments. The starting point is a prior distribution over LTI coefficients, which in itself could be (and, generally, is) the posterior distribution from some previous experiments. The crucial problems are two, how can an experimental perturbation be encoded mathematically within the model? how can we compute the utility score for a perturbation?

The answer to these questions depends on the specific perturbation considered; here we focus on three commonly employed experiments. The first type are changes in the external input to the system, the U matrix in eq. (6). We denote this class of experiments as *photo-period experiments*, since in the case study of *A. thaliana* the input matrix represents the light inputs to the circadian clock. The second type are mutagenic experiments, where a single gene is removed from the system (*knock-out*). The third type are observation experiments, where presence/ absence of one or more edges is observed directly through experiments such as Chromatin Immunoprecipitation (ChIP) or any affinity-binding detection methods.

Notice that observation experiments are somewhat different from the other types, as they do not constitute a perturbation of the system; for this reason, in the following we describe experimental design methodologies for observation experiments separately.

Photo-period experiments and knock-out experiments In the DSS setting, we frame experimental design for photo-period and knock-out settings as choosing the best experiment q defined as $i$nterventions in matrix $[\mathbf{X^qU}^q]$ that maximizes the *information gain* over the parameters $\mathbf{B} = [\mathbf{A}, \mathbf{C}]$ of the linear dynamical model of equation 6. An $i$ntervention consists of setting a column of $\mathbf{U}^q$ or $\mathbf{X}^q$ to a known value ξ^q (zero in case of knock-out experiments or the frequency spectrum for a light signal in the case of photo-period experiments). We will denote the intervened element as column(s) $\mathbf{X}_i^q$ and the rest of the columns as $\mathbf{X}_{\setminus i}^q$.

The utility function of eq. 5 can be computed by calculating the KL-divergence between the current distribution of the LTI-coefficients (either prior distribution

or posterior distribution of a previous experiment) and the posterior distribution over said parameters after performing the desired experiment. This implies that we have to be able to compute the expected value of the next experiment's observations, in order to compute the mutual information and thus the utility of the next experiment. Explicitly this utility function is

$$U\left(\mathbf{B};\mathbf{X}^q;\xi^q\right)=\int\int p\left(\mathbf{X}^q_{\setminus i},\mathbf{B}|\mathbf{X}^q_i=\xi^q\right)\log\frac{p\left(\mathbf{B}|\mathbf{X}^q_{\setminus i},\mathbf{X}^q_i=\xi^q\right)}{p\left(\mathbf{B}\right)}d\mathbf{X^q}d\mathbf{B}$$

the prior (current knowledge) $p\left(\mathbf{B}\right)$ doesn't depend on the next, simulated experiment (we simulate using the current knowledge), as such, the selection criteria can be stated in terms of the numerator as the integral

$$\int\int\mathbf{p}\left(\mathbf{X^q_{\setminus i}},\mathbf{B}|\mathbf{X^q_i}=\xi^{\mathbf{q}}\right)\log\mathbf{p}\left(\mathbf{B}|\mathbf{X^q_{\setminus i}},\mathbf{X^q_i}=\xi^{\mathbf{q}}\right)\mathbf{dX^q dB} \tag{7}$$

The conditional distribution $p\left(\mathbf{B}|\mathbf{X}^q_{\setminus i},\mathbf{X}^q_i=\xi^q\right)$ as derived in [11] is a result of a Linear regression model with Gaussian likelihood. As such the conditional over the coefficients $\mathbf{B}$ can be obtained by factorizing, and is

$$\log p\left(\mathbf{B}|\mathbf{X}^q,\xi^q\right)\propto\log\left[\det\left(\sigma_D^{-2}\boldsymbol{\Sigma}^{-1}\right)^{-1/2}\right]-\frac{1}{2\sigma_D^2}\left(-2\bar{\boldsymbol{\eta}}^{\mathrm{T}}\bar{\mathbf{B}}+\bar{\mathbf{B}}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\bar{\mathbf{B}}\right) \tag{8}$$

with the terms

$$\bar{\boldsymbol{\eta}}=vec\left(\sum_q\left[\mathbf{X}_q\ \mathbf{U}_q\right]^T\dot{\mathbf{X}}_q\right);\ \boldsymbol{\Sigma}^{-1}=\mathbf{I}\otimes\left(\sum_{\mathrm{q}}\left[\mathbf{X}_q\ \mathbf{U}_q\right]^T\left[\mathbf{X}_q\ \mathbf{U}_q\right]\right)$$

.

We evaluate equation (7) through Monte Carlo simulation by drawing a sample from the joint distribution

$$p\left(\mathbf{X}^q_{\setminus i},\mathbf{B}|\mathbf{X}^q_i=\xi^q\right)=p\left(\mathbf{X}_{\setminus\mathbf{i}}{}^q|\mathbf{B},\mathbf{X}^q_i=\xi^q\right)p\left(\mathbf{B}\right) \tag{9}$$

The Monte Carlo algorithm will consist of integrating $U_{DSS}\left(\bar{\boldsymbol{\eta}},\boldsymbol{\Sigma},\mathbf{B}\right)_{DSS}$ over both random variables

$$\frac{1}{S_1}\sum_{s_1=1}^{S_1}\left(\frac{1}{S_2}\sum_{s_2=1}^{S_2}\log p\left(\mathbf{B}^{(s_1)}|\mathbf{X}^{q(s_2)}_{\setminus i},\mathbf{X}^q_i=\xi^q\right)\right) \tag{10}$$

we draw a sample $\mathbf{B}^{(s_1)}$ from $p\left(\mathbf{B}\right)$, then we evaluate eq. 8 by drawing samples $\mathbf{X}^{q(s_2)}_{\setminus i}$ from the conditional distribution term of eq. 9. We derive the conditional distribution $p\left(\mathbf{X}_{\setminus\mathbf{i}}{}^q|\mathbf{B},\mathbf{X}^q_i=\xi^q\right)$ from the Gaussian likelihood of the regression model in [11] by using the Kronecker product and the vectorization operator. We apply the technique of completing the square[1], so we can get the distribution over the frequency spectra, from which we can draw samples as it is a Gaussian of the form

$$p\left(\mathbf{X}^q|\mathbf{B},\sigma^{\mathbf{2}}\right)\sim\mathcal{N}\left(\eta,\Lambda^{-1}\right) \tag{11}$$

with $\Lambda=\frac{1}{\sigma^2}\left(\mathbf{I}\otimes\mathbf{D}-\mathbf{A}^T\otimes\mathbf{I}\right)^T\left(\mathbf{I}\otimes\mathbf{D}-\mathbf{A}^T\otimes\mathbf{I}\right)$ and
$\eta=-\Lambda^{-1}\left(\mathbf{I}\otimes\mathbf{D}-\mathbf{A}^T\otimes\mathbf{I}\right)^T\mathbf{U}\bar{\mathbf{C}}$.

Experiments for observing interactions As a complement to the previous scores, we wished to account for an additional source of information, direct observations over DNA-protein interactions. A result of this kind of experiment can be viewed as an observation over element h_{ij} of matrix $\mathbf{H}$

Here the observed gene expression spectra are considered a fixed set $\mathbf{X}^q$. Having these observations, we aim at choosing which link h_{ij} possess the highest mutual information for learning parameters $\mathbf{B}$. This can be represented in terms of the conditional mutual information, which is a function of two conditional entropies such that $I\left(\mathbf{B};h_{ij}|\mathbf{X}^q\right)=H\left(\mathbf{B}|\mathbf{X}^q\right)-H\left(\mathbf{B}|\mathbf{X}^q,h_{ij}\right)$.

The conditional entropy is not a function of the selected link, so its computation is not necessary for discriminating between links. Then we introduce the utility function U_h equal to the negative conditional entropy of variable $\mathbf{B}$ given the gene expressions $\mathbf{X}^q$ and the observed link h_{ij}

$$U_h\left(\mathbf{B},\mathbf{X}^q,h_{ij}\right)=\sum_{\gamma\in\{0,1\}}p\left(h_{ij}=\gamma\right)\int p\left(\mathbf{B}|\mathbf{X}^q,h_{ij}=\gamma\right)\log p\left(\mathbf{B}|\mathbf{X}^q,h_{ij}=\gamma\right)d\mathbf{B}$$

where $p\left(\mathbf{B}|\mathbf{X}^q,h_{ij}=\gamma\right)$ is the posterior distribution over $\mathbf{B}$ given a fixed value for link h_{ij} (either 0 or 1).

We evaluate the integral by drawing samples from the conditional posterior $p\left(\mathbf{B}|\mathbf{X}^q,h_{ij}=\gamma\right)$, for $\gamma\in\{0,1\}$, and evaluating $\log p\left(\mathbf{B}|\mathbf{X}^q,h_{ij}=\gamma\right)$. We integrate by Monte Carlo method, with samples s_3 and s_4 drawn from the posterior distribution $p\left(\mathbf{B}|\mathbf{X}^q,h_{ij}=\gamma\right)$. As such the utility criterion is

$$U_h\left(\mathbf{B};\mathbf{X}^q;h_{ij}\right)=\frac{\sum_{s_3=1}^{S_3}\log p\left(\mathbf{B}^{(s_3)}|\mathbf{X}^q,h_{ij}=0\right)}{2S_3}+\frac{\sum_{s_3=1}^{S_4}\log p\left(\mathbf{B}^{(s_4)}|\mathbf{X}^q,h_{ij}=1\right)}{2S_4} \tag{12}$$

2.4 *A. thaliana* circadian clock model

In [9] we observe a state of the art model of the *A. thaliana* circadian clock network. It consists of the transcription factors LHY/CCA1 LHY (LATE ELONGATED HYPOCOTYL) and CCA1 (CIRCADIAN CLOCK ASSOCIATED 1), these execute an activating interaction with the transcriptional co-regulators PRR9, PRR7 and PRR5/NI (PSEUDO-RESPONSE Regulators 9, 7, 5/night inhibitor) which at the same time are interlocked in a negative feedback loop with LHY/CCA1. This feedback loop is thought to be the responsible for peak activity of day-time components.

On the other hand we have the evening loop, thought to be driven by EC (Evening complex), composed by the binding of ELF3 (EARLY FLOWERING

3), ELF4 (EARLY FLOWERING 4) and the GARP transcription LUX (LUX ARRHYTHMO) which controls LHY expression by a double negative connection [9]. A graphical representation of the model is shown in Figure 2.

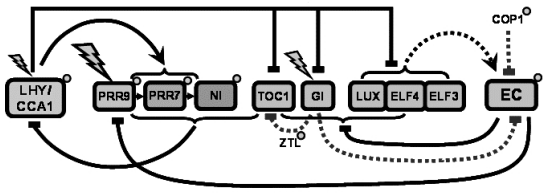

Fig. 2: Circadian clock model for *A. thaliana,* as shown in [9]. Transcriptional elements LHY, PRR579, GI, TOC1, LUX, ELF4 and ELF3 are assumed observed. While the expression levels of the Evening Complex (EC) is unobserved, along with other post-transcriptional interactions involving ZTL and COP1.

3 Results.

We simulate the *A. thaliana* circadian clock model, we selected and sub sampled the simulated data in order to get 12 samples over one light/dark cycle for a Wild Type population. We ran DSS and collected 10000 samples of the joint posterior over the model parameters. We executed DSS using standard parameters as in [11] and evaluated the mutual information criterion 10, we draw 1000 samples, thus setting parameter $S_1 = 1000$. We draw 100 samples for each gene expression level at each step, thus setting parameter $S_2 = 100$.

First, we chose photo-periods of 6/18, 8/16, 18,6 and 20/24, we computed the DFT of a {-1,1} light input (ξ^q) and added it to the spectra matrix. Thus drawing samples from the conditional distribution $p\left(\mathbf{X}^q|\mathbf{B}, \sigma^{\mathbf{2}}, \mathbf{U} = \xi^q\right)$.

Then we selected a set of knock out mutants commonly seen in experimental settings. In this way knock-out mutants ΔLHY, ΔLHY-GI, ΔLHY-TOC1 and ΔPRR7-PRR9 were simulated by conditioning the rest of the gene spectra given that the intervened genes have a constant spectrum of zero, that is $p\left(\mathbf{X}^q_{\setminus i}|\mathbf{B}, \sigma^2, \mathbf{X}^q_i = \mathbf{0}\right)$.

In figure 3 we present the results of evaluating eq. 10 for these two set of experiments. The boxes go from the 25th to the 75th percentiles and the red bar indicates the median score. It shows photo-period experiments having a median score between 220 and 225, while the knock-out mutants show less median values ranging from 210 to 217. It is of interest that the lowest information gain looks to be accredited to the ΔLHY-TOC1 double mutant, being these two genes the main drivers of circadian oscillations. This may be due to the nature of the mutual information criterion, as it accounts for the reduction in uncertainty over the estimation of parameters. It seems plausible that the disruption of these two

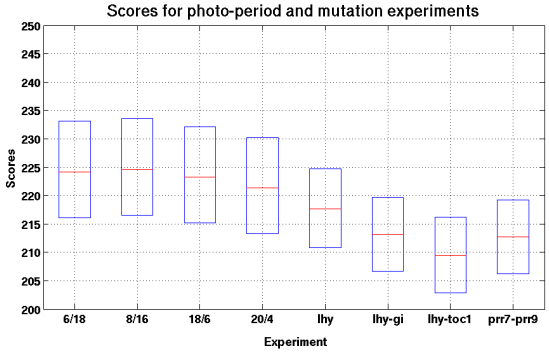

Fig. 3: Box plot for the evaluation of the DSS criterion, higher score means higher mutual information between experimental design and experimental outcomes. From left to right, photo periods of 6/18, 8/16, 18/6 and 20/4. Then knockout Mutants ΔLHY, ΔLHY-GI, ΔLHY-TOC1 and ΔPRR7-PRR9.

components alters clock behavior enough that parameter inference is less reliable, as the score suggests that the uncertainty over the model behaviour grows. This may be in fact another source of information about the importance of these two clock components.

Complementary, we computed the conditional mutual information for Chip experiments according to eq. 12. First we simulated Wild-type gene expression levels for 12 samples over a 24 hour period, using the same procedure as in the previous paragraph. Then, we selected a set of candidate links to observe, these include those known to be part of the true network, and those involving the EC components. Each one of these links was set to their possible values (one and zero), and the posterior distribution calculated for each case, this implies running DSS twice for each studied link with standard parameters as proposed in [11].

We show the resulting scores in figure 4. In this scatter plot, regulators are shown in the x axis, and the scores are presented through colored dots. Each dot is labeled according to the putative regulation tested (the regulators target is marked by a ->). Here we observe that the regulating interactions involving the elements of the EC complex (LUX, ELF4 and ELF3) as regulators show the lest information. This is not surprising as model assumptions are that the EC complex is the transcription factor involved in the evening regulation, and its effects even though essential, are not directly observable through its components. On the other hand we find that the most useful information seems to be related to the elucidation of the role of the light input over LHY and specially GI, with the highest score of 437, above of the mean value of 432.7. Another interesting interactions include that for LHY its most useful observation would be its regulation of TOC1, correspondingly, LHY would be the most informative interaction to observe for TOC1. As stated earlier, the interaction between these two components is the main driver of the morning oscillator.

Taking in account these two complimentary criteria, some decisions about the utility of the experiments can be made. In these case, it seems to points towards light-related experiments, as the expected mutual information for all the photoperiod experiments seems to be on par. This at the same time could be validated by the fact that light-input nodes of the network seem to be the most informative in first instances. Finally the LHY-TOC1 double mutant score suggest that the behaviour of the system under these circumstances is more uncertain, insight that may result useful for the researcher and thus an interesting experiment to execute.

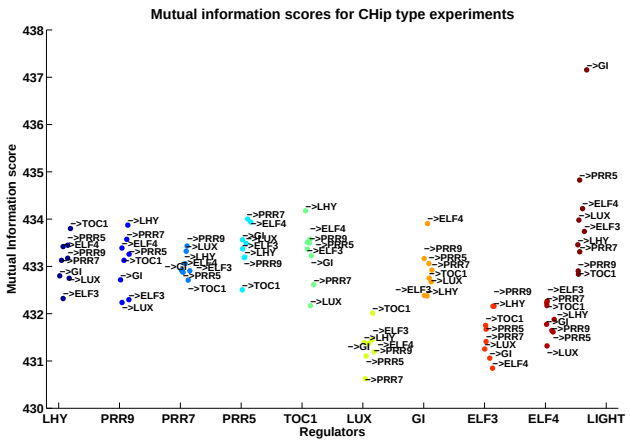

Fig. 4: Scatter plot of the conditional mutual information scores for observations over some edges. Each score is labeled with the represented interaction. The regulating interactions are symbolized by a "->" as "->targets", with the regulator being the label on the x axis tick. From left to right we have regulators LHY, PRR9, PRR7, PRR5, TOC1, LUX, GI, ELF3, ELF4 and photo-regulation in case of light inputs.

4 Conclusions

We have presented a methodology for Bayesian experimental design in biological dynamical systems with structural uncertainty. Experimental design is a branch of classical computational statistics which is gaining increasing attention in systems biology, due to inherent complexity and uncertainty of biological systems. Adapting classical methods to modern systems biology is problematic, as sources of uncertainty are ubiquitous in systems biology data, leading to computationally intractable problems and/ or predictions with large associated uncertainty. In general, handling both parametric and structural uncertainty in nonlinear systems is highly problematic. Earlier work such as [5] chose to focus on non-linear

systems without structural uncertainty. However, in many biological systems, such as oscillatory systems, it may be preferable to approximate the system dynamics to gain computational savings which will enable structural uncertainty to be considered in experimental design. Our results on the *A. thaliana* clock model show that this approach can be fruitful, highlighting potentially large differences in information content for different classes of experiments, and for different individual experiments in each class. These results are potentially precious for practitioners, whose prime preoccupation is often the prioritisation of experiments in the face of technical and resource limitations.

There are several directions along which the approach could be further developed. A simple, but potentially useful, extension would be to modify the utility function by explicitly accounting for the different costs of different experiments. It would also be of interest to develop strategies for planning multiple experiments, as the information gain is generally a non-linear function on the space of possible experiments. While the same approach can be easily deployed for small sets of experiments, the general issue of multiple experimental design yields a very challenging discrete optimisation problem. We envisage that ideas from reinforcement learning could be effective in this scenario.

References

1. Cristopher M. Bishop Pattern recognition and Machine learning. *Springer Verlag*, 2001
2. Kathryn Chaloner and Isabella Verdinelli. Bayesian Experimental Design: A Review. *Statist. Sci.*, (3):273–304, August 1995.
3. Hemant Ishwaran and J. Sunil Rao. Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics*, 33(2):730–773, April 2005.
4. Clemens Kreutz and Jens Timmer. Systems biology: experimental design, *FEBS Journal*, 276(4):923–942, 2009.
5. Juliane Liepe, Sarah Filippi, Michal Komorowski, and Michael P. H. Stumpf. Maximizing the Information Content of Experiments in Systems Biology. *PLoS Computational Biology*, 9(1):e1002888, January 2013.
6. D. V. Lindley. On a Measure of the Information Provided by an Experiment. *Ann. Math. Statist.*, (4):986–1005, December 1956.
7. David JC MacKay. *Information theory, inference, and learning algorithms*, volume 7. Cambridge University Press, Cambridge, UK, 2003.
8. Pintelon R, Schoukens J. *System Identification: A Frequency Domain Approach*,2nd ed., 2012.
9. Alexandra Pokhilko, Aurora Pinas Fernandez, Kieron D Edwards, Megan M Southern, Karen J Halliday, and Andrew J Millar. The clock gene circuit in Arabidopsis includes a repressilator with additional feedback loops. *Molecular Systems Biology*, 8, March 2012.
10. F. Pukelsheim. *Optimal Design of Experiments*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, January 2006.
11. D. Trejo-Banos, Guido Sanguinetti, and Andrew J. Millar A Bayesian approach for structure learning in oscillating regulatory networks `arXiv:1504.06553 [stat.ML]`