Optimized Expected Information Gain for Nonlinear Dynamical Systems

Alberto Giovanni Busetto[1,2,3] BUSETTOA@INF.ETHZ.CH
Cheng Soon Ong[1] CHENGSOON.ONG@INF.ETHZ.CH
Joachim M. Buhmann[1,2] JBUHMANN@INF.ETHZ.CH
[1]Department of Computer Science, ETH Zurich, Universitätstr. 6, 8092 Zurich, Switzerland
[2]Competence Center for Systems Physiology and Metabolic Diseases, Schafmattst. 18, 8093 Zurich, Switzerland
[3]Life Science Zurich PhD Program on Systems Biology of Complex Diseases

Abstract

This paper addresses the problem of active model selection for nonlinear dynamical systems. We propose a novel learning approach that selects the most informative subset of time-dependent variables for the purpose of Bayesian model inference. The model selection criterion maximizes the expected Kullback-Leibler divergence between the prior and the posterior probabilities over the models. The proposed strategy generalizes the standard D-optimal design, which is obtained from a uniform prior with Gaussian noise. In addition, our approach allows us to determine an information halting criterion for model identification. We illustrate the benefits of our approach by differentiating between 18 published biochemical models of the TOR signaling pathway, a model selection problem in systems biology. By generating pivotal selection experiments, our strategy outperforms the standard A-optimal, D-optimal and E-optimal sequential design techniques.

1. Introduction

Dynamical systems theory develops the mathematical approach to model and analyze complex dynamical systems, usually by employing differential or difference equations. From the origins in Newtonian mechanics, dynamical systems have broadened their range of applications to diverse fields such as physics, biology, chemistry, engineering and finance. Contemporary state-space representations enable analysis, simulation, prediction and control of dynamical processes.

When foundational assumptions exist, models can be based on first principles. In realistic cases, such models are overly complex and uncertain. Because of incomplete knowledge, conflicting theories exist and correspond to a set of alternative models. These models represent mutually exclusive hypotheses, whose degree of belief changes as evidence accumulates. In accordance with the scientific method, evidence is collected through observations of phenomena that occur in the natural world, or which are created as controlled experiments in a laboratory. In this context, the concept of surprise is fundamental in order to evaluate the expected informativeness of an experiment (Baldi & Itti, 2005). Obtaining valuable data is economically demanding and experimentally challenging and, as a consequence, quantitative measurements are usually noisy, scarce and incomplete. Therefore, designing experiments that are both feasible and maximally informative is highly desirable. The advantages of active rational design impact all the mentioned application fields but are essential for systems biology, where simple and well understood mechanisms are the exception (Kitano, 2002).

This paper describes the problem, summarizes the related work and presents our contributions in section 1. Section 2 introduces the fundamental concepts, formalizes the methods and substantiate their theoretical justification. Section 3 shows the experimental results with a real set of alternative biochemical models, demonstrating the empirical effectiveness of our approach and the obtained improvements with respect to standard techniques. The conclusion discusses the results and offers some directions for future research.

1.1. Problem Description

Given a set of models, the problem is to find the maximally informative state-subspace under given feasibil-

Appearing in *Proceedings of the 26^{th} International Conference on Machine Learning*, Montreal, Canada, 2009. Copyright 2009 by the author(s)/owner(s).

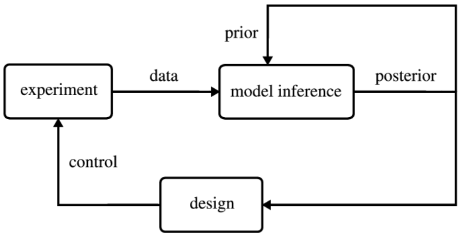

Figure 1. Schematic illustration of the relationship between experiment, model inference and design.

ity constraints. The models are defined as rules for the time evolution of a dynamical process on a state-space. The constraints are bounds on the dimension of the measurable subspace and on the number of sequential observations. In our scenario, the available models are nonlinear and can be either deterministic or stochastic. Their definition is possibly incomplete and most probably uncertain: they encode the current knowledge about the putative dynamic interaction between the variables of the system before the measurement. The task of model selection must be performed given time-series from partial and noisy experimental measurements.

In this paper, two complementary goals are considered. The first is to find the maximally informative subset of state variables for the purpose of model selection. The second is to identify their minimal number for model identification. Moreover, we want to take as much advantage as possible from prior information and domain knowledge. As often in active learning, we expect a potential feedback coupling between design and experimentation. This is schematically illustrated in Fig. 1.

1.2. Motivation

A motivation for our work is the recent considerable interest in the design of biological experiments (Banga & Balsa-Canto, 2008). Conflicting hypotheses and scarce data severely hamper the development of predictive mechanistic models. Standard techniques are not suited to complex dynamical systems: novel techniques must incorporate prior information and take into account nonlinear effects. In general, two main challenges arise for the development of such a design. First, the measurements are time-dependent and, as a consequence, simultaneous measurements have a limited multiplicity. Second, even local nonlinearities can propagate their effects at a global scale. As a result, a successful approach must be flexible enough to consider sufficiently complex behaviors, but it should also be able to exploit the available regularities.

From a machine learning point of view, the interesting aspects are the reduced observability in the high dimensional state-space and the relation between information content and robustness of a dynamical system.

1.3. Related Work

In general, the design of experiments for the selection of mathematical models defines an important part of scientific research, i.e., how can we specify the controllable aspects of the system under consideration before we execute the measurements. This paper entertains the basic idea that model inference can be improved by appropriately selecting the values of the control variables. Since the pioneering work of Sir Ronald A. Fisher, there has been a long research tradition on the mathematical design of maximally informative experiments. A significant number of articles and surveys have been devoted to an exposition of the mathematical apparatus of optimal experimental design (Montgomery, 2004). In general, the problem can be formulated as an optimization problem for an utility function that reflects the purpose of the experiment. Standard non-Bayesian strategies include the alphabetical design methods and, between these, the most common is the D-optimal design (Boyd & Vandenberghe, 2004). A unifying theory for the formulation of Bayesian experimental design exists and is based on Shannon information criteria (Chaloner & Verdinelli, 1995). The application of these approaches is recognized as general and advantageous, but is usually severely limited to the linear-Gaussian case. The major shortcomings of the current methods, such as the lack of integration of prior knowledge and the insufficient exploration of available unmeasured data, are well documented (Chaloner & Verdinelli, 1995). Entropy-based nonlinear design approaches have been proposed before (van den Berg et al., 2003), but suffer important limitations. First, they consider static systems instead of dynamical ones. Second, by maximizing the entropy instead of the relative entropy between prior and posterior, they are unable to incorporate prior information. This prevents their consistent iterative application. Finally, they design maximally informative experiments for parameter estimation within the same model structure, instead of actively performing model selection.

In the contemporary literature, attention for experimental design is growing rapidly in the field of dynamical systems, but nonlinear dynamical systems still remain a fairly unexplored field (Heine et al., 2008). This is a significant research bottleneck, since nonlinear effects are often what differentiate alternative models in systems biology. Since linearizations are not appropriate for most of the applications, new strategies are urgently required.

1.4. Contributions

The limitations mentioned above motivate us to propose a novel approach. Our main contributions are as follows.

1. We introduce a novel optimal experimental design technique for the purpose of model selection of nonlinear dynamical systems. The key idea is to maximize the expected information gain by selecting the most appropriate measurement subspace. We provide a sequential strategy for the associated optimization problem, showing that it outperforms theoretically and empirically standard methods.
2. We propose a scheme to identify the minimal number of variables required for model identification. We provide insights into the suggested method: it seeks for subspaces of the state-space which produce informative dynamics. By combining the halting criterion with the design technique, we open the way for a better understanding of the key mechanisms that distinguish alternative dynamical behaviors.

In short, whereas other approaches are either nonlinear but static or dynamical but linear, our novel design can be applied to nonlinear dynamical systems. Moreover, being based on the maximization of the relative entropy instead of the entropy alone, it consistently incorporates prior information and enables iterative online design.

2. Methods

This section introduces the basic definitions and mathematically formalizes our methods. We start with a description of the model selection scenario for dynamical systems. This exposition is followed by the characterization of the expected information gain and by its optimization. Finally, we propose a halting criterion that finds the minimum number of variables that provide a given fraction of the experimentally accessible information.

2.1. Process and Measurement Models

A dynamical system Σ is described by two main components: the process model and the measurement model. The canonical approach in systems theory describes the process model $\mathcal{M}$ as a set of n first order coupled differential or difference equations. It is an implicit rule for the time evolution of the state x in a state-space $\mathcal{S}$. The trajectory can be deterministic or stochastic and, in the case of continuous-time continuous-valued systems, it is mathematically represented as an Ordinary Differential Equation (ODE) or as a Stochastic Differential Equation (SDE), respectively. In a time span $\mathcal{T} = [t_0, t_f]$, the Itō SDE for a process model is defined as

$$dx(t) = f(x(t), \theta, u(t), t)dt + \sigma_w(x(t))dW(t), \qquad (1)$$

where $x \in \mathbb{R}^n$ is the $n-$dimensional state vector and $\theta \in \mathbb{R}^r$ is the parameter vector. The function $u(\cdot)$ denotes the input intervention. The map $f(\cdot)$ is a nonlinear function of the current state, the possibly uncertain parameters θ and the input. It governs the deterministic component of the r.h.s. in Eq. 1 and its structure can be completely or partially known. When the stochastic component exists, it is given by the $n-$dimensional Wiener process $W(t)$, whose infinitesimal variance is σ_w^2. Given the initial conditions $x_0 = x(t_0)$, the process model is mathematically defined as an initial value problem.

The measurement model is continuous-valued and discrete-time. It is formalized as

$$y(t_i) = h(x(t_i), v_i), \qquad (2)$$

where the function $h(\cdot)$ maps states into observations y of dimension $m \leq n$. The experimental observations are corrupted by noise, which is modeled by the random variables v_i. The function h is usually a linear function of the states and of the noise. It can be written as

$$h(x, v) = Hx + v. \qquad (3)$$

Commonly, H can be defined as a $m \times n$ matrix, whose elements are $h_{ij} \in \{0, 1\}$. When m distinct states can be measured, it is possible to define a Variable Selection Vector (VSV) $s \in \{0, 1\}^n$ such that

$$\|s\|_1 = \sum_{i=1}^{n} s_i = m \qquad (4)$$

and $h_{ij}(s) \neq 0$ when $s_j = 1$ and $i = \sum_{k=1}^{j} s_j$. The VSV indicates which elements of the state vector can be directly observed by the measurement model.

The data set collected from an experiment whose VSV is s is defined as

$$\mathcal{D}(s) = \{(t_i, y(t_i; s))\}_{i=1}^{l}, \qquad (5)$$

where $\{t_i\}_{i=1}^{l}$ are the measurement time points.

2.2. Model Selection

Scientific inquiry determines the principles behind a series of observations as one of its fundamental tasks. Since the observer lacks complete knowledge about f in the process model $\mathcal{M}$, a set of alternative models is hypothesized $\{f_i\}_{i=1}^{q}$. They represent conflicting explanations and, therefore, are mutually exclusive hypotheses.

In the Bayesian approach to model selection, the observer makes inferences by computing the posterior probability of a model $\mathcal{M}$ given the available data set $\mathcal{D}$. The posterior probability is given by Bayes theorem as

$$p(\mathcal{M}|\mathcal{D}) = p(\mathcal{D}|\mathcal{M})p(\mathcal{M})/p(\mathcal{D}), \qquad (6)$$

where $p(\mathcal{D}|\mathcal{M})$ is the likelihood function, $p(\mathcal{M})$ is the prior and $p(\mathcal{D})$ is the normalizing constant. The likelihood function must take into account the fact that there exists an additional level of uncertainty, which originates from the fact that the parameters are unknown. Therefore, the likelihood function must be computed as

$$p(\mathcal{D}|\mathcal{M}) = \int p(\theta|\mathcal{M})p(\mathcal{D}|\mathcal{M},\theta)d\theta. \tag{7}$$

The dynamical behavior of $\mathcal{M}$ is affected by the initial conditions and possibly by stochastic effects. Consequently, between consecutive measurements, the joint probability of state and parameters is given by the solution of the Fokker-Planck equation

$$\partial p/\partial t = -\mathrm{div} fp + \Delta Dp, \tag{8}$$

where $p = p(x|\mathcal{M},\theta;t)$. We indicated the divergence operator over the continuously differentiable vector field fp as

$$\mathrm{div} fp = \sum_{i=1}^{n} \frac{\partial}{\partial x_i}[f]_i p \tag{9}$$

and the Laplacian operator over Dp as

$$\Delta Dp = \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial^2}{\partial x_i \partial x_j}[D]_{i,j}p. \tag{10}$$

Here, $\mathrm{div} fp(x,\theta;t)$ is the drift term, which is the effect obtained by the deterministic part of Eq. 1. The diffusion term $\Delta Dp(x,\theta;t)$, with the tensor D, represents the random effects introduced by the stochastic force in Eq. 1.

Based on all the available information, the goal of Bayesian model inference for dynamical systems is the reconstruction of the model posterior. Let us denote the data set at time t_k as

$$\mathcal{D}_k = \{(t_i, y(t_i)) \in \mathcal{D} | i \leq k\}. \tag{11}$$

The Bayesian inference must be performed by integrating over the all the states and all the uncertain parameters. The probability of being in a specific state is determined by the solution of Eq. 8 and, therefore, the posterior over the models is given by

$$\begin{aligned} p(\mathcal{M};t) &= \int\int p(\mathcal{M},x,\theta;t)dxd\theta \\ &= \int\int p(\mathcal{M}|x,\theta;t)p(x,\theta;t)dxd\theta \\ &= \int\int p(\mathcal{M}|x,\theta;t)p(x|\theta;t)p(\theta;t)dxd\theta. \end{aligned} \tag{12}$$

When a new observation is performed at time t_k, extra information is available and the observer updates the current belief state about the models by recursively computing

$$p(x|\theta,\mathcal{D}_k;t) = \frac{p(y_k|x,\theta;t)p(x|\theta,\mathcal{D}_{k-1};t)}{p(y_k|\theta,\mathcal{D}_{k-1};t)} \tag{13}$$

and integrating in Eq. 12. The received information reduces the uncertainty in the model posterior. In general, the solutions of Eq. 13 cannot be computed analytically and must be approximated, usually employing Monte Carlo (MC) techniques like particle filters.

2.3. Data Collection and Information

Shannon entropy, a key notion in information theory, measures the uncertainty of a random variable. In our scenario, it measures the uncertainty about the models at certain points in time. This quantity is defined as

$$H[\mathcal{M};t] = -\sum_{i=1}^{q} p(\mathcal{M}_i;t)\log p(\mathcal{M}_i;t). \tag{14}$$

The mutual information between the models and the data is denoted by

$$\mathrm{I}(\mathcal{M},\mathcal{D}) = \sum_{i=1}^{q}\int p(\mathcal{M},\mathcal{D})\log\left(\frac{p(\mathcal{M},\mathcal{D})}{p(\mathcal{M})p(\mathcal{D})}\right)d\mathcal{D} \tag{15}$$

and represents the reduction in the uncertainty about the models as a consequence of the experimental observations. It is important to note that the mutual information is always non-negative. Therefore, from the combination of Eq. 14 and Eq. 15, we have that

$$H[\mathcal{M};t_0] - H[\mathcal{M};t_k] = \mathrm{I}(\mathcal{M},\mathcal{D}_k) \geq 0. \tag{16}$$

2.4. Optimized Information Gain

The specification of the purpose of the experiment produces various criteria for the choice of the design. In our scenario, the goal is to maximally reduce the uncertainty for the selection of the model, given all the available knowledge. Following a decision theoretic approach, a general utility function is denoted by $U(d,\mathcal{M},s,\mathcal{D})$, where d is a decision coming from a given set. For any design specified by the VSV s, the expected utility of the best decision is given by

$$\int \max_d \left\{\sum_{i=1}^{q} U(d,\mathcal{M}_i,s,\mathcal{D})p(\mathcal{M}_i|\mathcal{D},s)p(\mathcal{D}|s)\right\} d\mathcal{D}. \tag{17}$$

The Bayesian solution to the experimental design problem is provided by the design VSV s^* maximizing Eq. 17. Simultaneously, it is possible to combine the result with maximally informative initial conditions and input perturbations (Busetto & Buhmann, 2009).

By using Eq. 15, we can state our design problem as follows: find the VSV s^* such that

$$s^* = \arg\max_{s\in\{0,1\}^n}\{\mathrm{I}(\mathcal{M},\mathcal{D}(s))\}, \tag{18}$$

subject to Eq. 4. The score function can be rewritten as

$$\mathrm{I}(\mathcal{M},\mathcal{D}(s)) = \mathbb{E}_{\mathcal{D}}\left[\mathrm{KL}(p(\mathcal{M}|\mathcal{D}(s))||p(\mathcal{M}))\right], \tag{19}$$

where $\mathrm{KL}(\cdot||\cdot)$ is the Kullback-Leibler (KL) divergence, which is a non-commutative measure of the difference between two probability distributions. It is mathematically described as

$$\mathrm{KL}(p(\mathcal{M}|\mathcal{D})||p(\mathcal{M})) = \sum_{i=1}^{q} p(\mathcal{M}|\mathcal{D}) \log \frac{p(\mathcal{M}|\mathcal{D})}{p(\mathcal{M})}. \tag{20}$$

In our case, it measures the expected difference in the number of bits required to code samples from the posterior $p(\mathcal{M}|\mathcal{D}(s))$ when using a code based on $p(\mathcal{M}|\mathcal{D}(s))$ itself, and when using a code based on the prior $p(\mathcal{M})$. Eq. 19 provides an insight into the suggested optimization: we want to maximize the expected information difference between the prior and the posterior over the models.

The standard D-optimal experimental design can be obtained as a special case of the described approach, when the prior over the models is a discrete uniform distribution and the measurement noise is assumed to be Gaussian with $v \sim \mathcal{N}(0, \sigma_v^2 \mathrm{I}_m)$. This is shown by the following equivalences:

$$\begin{aligned} s^* &= \arg\max_s \left\{ \mathbb{E}_{\mathcal{D}} \left[\mathrm{KL}(p(\mathcal{M}|\mathcal{D})||p(\mathcal{M})) \right] \right\} \\ &= \arg\max_s \left\{ \int_{\mathcal{D}} p(\mathcal{D}) \sum_{\mathcal{M}} p(\mathcal{M}|\mathcal{D}) \log \frac{p(\mathcal{M}|\mathcal{D})}{p(\mathcal{M})} \right\} \\ &= \arg\max_s \left\{ - \int_{\mathcal{D}} \sum_{\mathcal{M}} p(\mathcal{D}, \mathcal{M}) \log p(\mathcal{D}) \right\} \\ &= \arg\max_s \left\{ H[\mathcal{D}] \right\} \\ &= \arg\min_s \left\{ \det\left(\mathrm{cov}[\mathcal{E}]\right) \right\}, \end{aligned} \tag{21}$$

where $H[\mathcal{D}]$ is the differential entropy of the data, that is

$$H[\mathcal{D}] = - \int p(\mathcal{D}) \log p(\mathcal{D}), \tag{22}$$

and $\mathcal{E}$ is the covariance matrix of the error $e = \widehat{x} - x$ obtained by the maximum likelihood estimation of the state. In line 3 of Eq. 21, the term $H(\mathcal{D}|\mathcal{M})$ is omitted because it does not depend on the VSV due to the zero mean and constant variance for all the state variables. As a result, the experiment minimizes the volume of the resulting confidence ellipsoid. A substantial difference from the usual static problems is that multiple simultaneous measurements can be infeasible for dynamical systems, due to temporal constraints. This distinguishes our scenario from the cases where it is possible to directly approximate the solution with the well-known relaxed experimental design problem.

Although Eq. 18 has a simple interpretation, the involved optimization is a difficult Boolean nonlinear problem. Since the number of possible VSVs grows binomially along with the dimensionality of the state space, finding the best measurable subset by exhaustive search is usually intractable. Therefore, we have to resort to feasible approximations. We formulate the optimal experimental design as a sequential optimization problem and we propose a greedy approach. Given a VSV s_k such that $\|s_k\| = k < m$, the goal of the sequential problem is the augmentation of s_k with an additional measurable element of the state, such that

$$s_{k+1}^* = \arg\max_{s_{k+1}} \left\{ \mathrm{I}(\mathcal{M}, \mathcal{D}(s_{k+1})) \right\}, \tag{23}$$

subject to

$$\|s_{k+1}\|_1 = k + 1. \tag{24}$$

This constraint maximization problem is solved by looking for the maximally discriminative state between the $n - k$ variables that remain unmeasured. Together with the simulated Bayesian inference, the evaluation of the score function requires the numerical integration of the trajectory of the dynamical system and, in the general case, requires the solution of Eq. 8. This can be done with sequential MC techniques, which achieve satisfactory approximations even for nonlinear systems (Busetto & Buhmann, 2009). Here, we propose an algorithm that iteratively augments the VSV, until the measurement output reaches the dimension m. After the solution of Eq. 8 for every model, $\mathrm{I}(\mathcal{M}, \mathcal{D}(s))$ is evaluated by integrating over

$$\mathrm{I}(\mathcal{M}, \mathcal{D}) = \int p(\mathcal{D}) \sum_{\mathcal{M}} p(\mathcal{M}|\mathcal{D}) \log \frac{p(\mathcal{M}|\mathcal{D})}{p(\mathcal{M})} d\mathcal{D}. \tag{25}$$

Algorithm 1 (Sequential Information Gain)

```
Input:  $p(x_0|\theta, \mathcal{M})$, $p(\theta|\mathcal{M})$, $p(\mathcal{M})$, $t_f$, $m$
Output: $\tilde{s}^*$
  for $i$=1 to $q$ do
    solve Eq. 8 for $\mathcal{M}_i$;
  end for
  $s_0 \leftarrow 0_{n \times 1}$
  for $i$=1 to $m$ do
    Select $s_i$ with the highest $\mathrm{I}(\mathcal{M}, \mathcal{D}(s_i))$ by augmenting $s_{i-1}$;
  end for
  $\tilde{s}^* \leftarrow s_m$;
```

This integral is simplified by using the following equivalent form

$$\sum_{\mathcal{M}} p(\mathcal{M}) \int p(\mathcal{D}|\mathcal{M}) \log \frac{p(\mathcal{D}|\mathcal{M})}{p(\mathcal{D})} d\mathcal{D}, \tag{26}$$

which is more amenable to MC computation, since it requires only the evaluation of the likelihood function and of the normalizing constant

$$p(\mathcal{D}) = \sum_{\mathcal{M}} p(\mathcal{D}|\mathcal{M}) p(\mathcal{M}). \tag{27}$$

This optimized result $\tilde{s}^*$ approaches the global optimum s^* when the following assumption holds $\forall s_a, s_b$

$$\mathrm{I}(\mathcal{M}, \mathcal{D}(s_a \vee s_b)) \simeq \mathrm{I}(\mathcal{M}, \mathcal{D}(s_a)) + \mathrm{I}(\mathcal{M}, \mathcal{D}(s_b)), \quad (28)$$

where $\vee$ is the pointwise OR operator between the two VSVs. The condition holds when there exists a small subset of very informative variables which are approximately independent. As shown in the results, the assumption is justified for several biological systems. In fact, it is common to find a combination of mechanisms which evolved towards functional modularity and robustness.

2.5. Information Halting Criterion

Based on the presented design technique, we propose a halting criterion to identify the dimensionality m of the measurement vector that is sufficient to reach a certain percentage of the total experimentally accessible information. On the average, in the ideal case where $m = n$, the amount of information gained after the execution of an experiment can be easily computed by using $s_{\max} = 1_{n\times 1}$ on Eq. 26. Let us denote the expected information obtained with $s_{\max}$ as $\mathrm{I}_{\max}$. Now, for a given $\alpha \in [0, 1]$, we can formulate the halting problem as follows: find m^* such that

$$m^* = \min\{m \leq n | \mathrm{I}(\mathcal{M}, \mathcal{D}(s^{m*})) \geq \alpha \mathrm{I}_{\max}\}, \quad (29)$$

where s^{m*} is the solution of Eq. 18. The vector of dimension n which satisfies the required constraints can be computed by stopping the sequential optimization problem defined by Eq. 23 when the required information threshold is met.

This approach can be particularly useful in cases where, almost regardless of the initial conditions, a small subset of the state variables provides a significant fraction of $\mathrm{I}_{\max}$. The effect has been observed in a variety of systems, including electrical circuits, fluid dynamics and biological systems. Information concentration is due to the fact that, while some subunits are sensitive to small changes of initial conditions and parameters, other functional components of the system are robust to structural perturbation. The observed persistence of some state variables between models can be caused by decoupling between components, by incorporated redundancy in the structure or by self-stabilizing mechanisms (Wagner, 2005).

3. Results

In this section, a series of experiments are carried out to evaluate the effectiveness of the proposed methods. We test our strategies with a real set of models for the highly conserved Target-Of-Rapamycin (TOR) pathway of *Saccharomyces cerevisiæ*. The dynamical models represent mechanistically alternative hypotheses that have been proposed in the literature. We show that our design technique outperforms the standard A-optimal, D-optimal and E-optimal sequential approaches. Moreover, we illustrate the ability of the information halting criterion to identify minimal pivotal experiments that discriminate between alternative hypotheses.

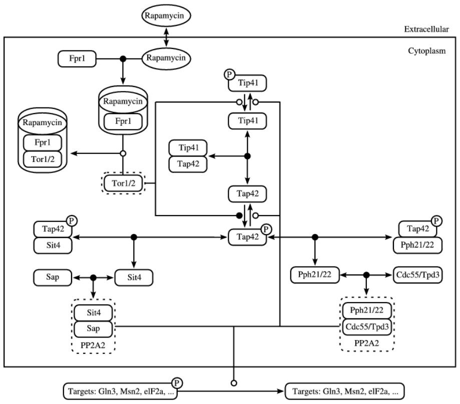

Figure 2. Representation of the molecular interactions of the core model $\mathcal{M}_0$ according to the standardized process diagram graphical notation (Kitano et al., 2005). Proteins, protein complexes and small molecules are shown as nodes, while transitions are represented by arrows. Dashed lines indicate enzymatic activity.

3.1. TOR Pathway Models

The TOR pathway of *Saccharomyces cerevisiæ* is an important biological mechanism, whose mammalian homologues are attractive drug targets. In fact, they are believed to have a role in various diseases such as cancer, autoimmunity, metabolic and cardiovascular disorders. The pathway constituents and many of the molecular interactions are known, but the overall functionality is not well understood in a quantitative sense. As a result of this uncertainty, a number of conflicting hypotheses regarding the biochemical mechanisms that control TOR signaling have been proposed (Kuepfer et al., 2007). A core model $\mathcal{M}_0$ summarizes the present available knowledge on the pathway. By incorporating new sets of chemical reactions in $\mathcal{M}_0$, a library $\{\mathcal{M}_i\}_{i=1}^{18}$ of alternative extensions has been proposed. Each of these extended models encode a different topology of the underlying reaction network. The molecular interactions of the core model are illustrated by Fig. 2.

The dynamical behavior of the TOR pathway is modeled as a set of ODEs, where the state vector $c(t)$ describes the concentrations of every distinct chemical entity in the system. The change of the species concentration over time is formalized as follows

$$dc(t)/dt = N \cdot r(c(t), \theta, t) = f(c(t), \theta). \quad (30)$$

The stoichiometric matrix N describes the structural relationships between the network components and quantifies the net effect of all the involved reactions in the biochemical system. It is invariant against time, kinetics and concentrations. The vector $r(c(t), \theta, t)$

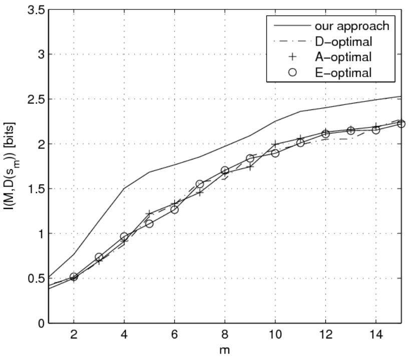

Figure 3. Average information gain as a function of the dimension of the measurement vector. Obtained from 100 design problems with 18 pathway models and randomized initial conditions. On the average, our method provides 33% more information than the standard methods for the same dimension of the measurement vector.

represents the flux distribution in the network and is a nonlinear function of the current concentrations and of several kinetic parameters. Alternative models incorporate additional control mechanisms, which are encoded in the different functions f_i. The total number of chemical species is $n = 60$ and there are 127 distinct chemical reactions in the model library, 24 of which belong to the common core. Each of the 18 extensions incorporates a distinct set of additional reactions and their complexity ranges from single reversible reactions to convoluted nonlinear feedbacks composed of 16 additional elements. Prior information about the models, their structure and parametrization comes from the integration of eleven published experimental data sets (Kuepfer et al., 2007).

3.2. Expected Information Gain

We test our algorithm against standard techniques: the A-optimal, D-optimal and E-optimal sequential designs. Respectively, they minimize the trace, the determinant and the 2-norm of the error covariance matrix $\mathcal{E}$. Their geometrical interpretations are: the minimization of the mean of the squared norm of the error and, for the resulting confidence ellipsoid, the minimization of its volume and of its diameter (Boyd & Vandenberghe, 2004).

The expected information gain is plotted in Fig. 3 as a function of m, which is the dimension of the measurement vector. The results are quantified in terms of expected Shannon information between the prior and the posterior probability over the models, as a result of the execution of the designed experiment. In this test case, we use all the 18 published models (Kuepfer et al., 2007) and we average the results obtained from 100 normalized design problems with initial concentrations sampled from the uniform distribution in $[0, 10^3]^n$ [AU]. The prior probabilities are generated at random, sampling uniformly from the parameter space of the multinomial distributions of dimension q. The measurement noise has covariance matrix $\sigma_v^2 \mathrm{I}_m = 3 \cdot 10^5$. The measurement time points are $t_1 = 2.5 \cdot 10^{-3}$ [AU] and $t_2 = 5.0 \cdot 10^{-3}$ [AU]. The numerical integration is performed by sequential importance sampling/resampling with preventive particle clustering (Busetto & Buhmann, 2009). The plot compares our strategy with the A-optimal, D-optimal and E-optimal sequential designs. For every m, our approach is more informative and, on the average, is able to provide approximately 33% more information than the standard methods. Figure 3 shows that our strategy rapidly identifies the most informative variables. Whereas standard approaches require the measurement of 4 state variables to gain 1 bit, our method yields more information with only 3 variables.

The advantage over the sequential alphabetical designs comes from the better use of the available prior information. As theoretically shown by Eq. 21, the D-optimal design assume a uniform prior and select the VSVs according to a distorted score function. More precisely, it substitutes the expected KL divergence between the prior and the posterior with the entropy of the posterior. This loss of information, which corresponds to the KL divergence between a flat prior and the "real prior" $p(\mathcal{M})$, reflects its sub-optimality in the sequential selection of s. The same considerations extend to the A-optimal and E-optimal designs: they exhibit a similar performance for these dynamical systems because $\mathcal{E}$ approximates a scalar matrix.

3.3. Halting Criterion

We test the proposed information halting criterion with the 18 models of the TOR pathway. An upper bound on the amount of information that the experiment can provide is given by $\mathrm{I}_{\max}$, obtained when the entire state vector is observable. Figure 4 shows that 9 state variables provide 80% of the experimentally accessible information. The information halting criterion offers theoretical and practical advantages. First, it provides an insight about the behavior of the studied system: the few informative mechanisms are rapidly identified. Second, combining the criterion with a cost function, it defines a tradeoff between the value of the expected information gain and the expense for resources. Interestingly, our study confirms the results of computational and experimental studies (Kuepfer et al., 2007): the complex Tap42p-Tip41p plays a strongly discriminative role under a wide range of initial conditions and parameters.

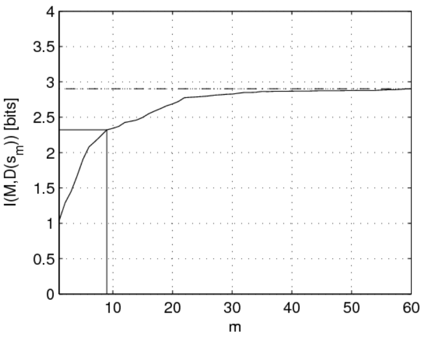

Figure 4. The information halting criterion shows that the measurement of 9 variables provides 80% of the experimentally accessible information. Key mechanisms emerge from the intersection of the subsets obtained under different initial conditions. The upper bound is given by $I_{\max}$.

4. Conclusion

We presented a novel approach to active model selection for nonlinear biochemical systems. The generality of the method permits its application to both stochastic and deterministic dynamical systems, whose initial conditions and parameters are uncertain. Our strategy overcomes some of the limitations of the current methods: they are restricted to either nonlinear but static or dynamical but linear systems and often do not directly incorporate prior information. Under the information theoretic framework, we proposed a halting criterion for the obtainment of a required amount of experimentally accessible information.

From a machine learning point of view, we introduced a new method for active learning in the context of nonlinear dynamical model selection. From the mathematical perspective, we showed that our approach is more general than the common D-optimal experimental design. The empirical results show that the method significantly outperforms the competing A-optimal, D-optimal and E-optimal sequential designs.

The advantages of the presented approach can be summarized as follows. First, it obtains a higher information gain for a given number of measurements, as shown in Fig. 3. Second, it enables the identification of a highly informative subset of measurements, as shown in Fig. 4. The obtained empirical results suggest its wide applicability to real-world problems. In particular, we expect that its ability to generate decisive experiments will prove particularly useful for structurally uncertain systems. The wide range of applications include systems biology, sensor placement, electronic circuit design, tracking in computer vision and planning of clinical trials. More generally, the relationship between dynamical robustness and learnability is a topic that deserves attention and is the subject of ongoing research.

Acknowledgments

We thank Jörg Stelling and Uwe Sauer for insightful discussions and the anonymous reviewers for helpful comments. This project was financed with a grant from the Swiss SystemsX.ch initiative, evaluated by the Swiss National Science Foundation.

References

Baldi, P., & Itti, L. (2005). Attention: bits versus wows. *Proc. IEEE Int. Conf. on Neural Networks and Brain, Beijing, China* (pp. PL56–PL61).

Banga, J. R., & Balsa-Canto, E. (2008). Parameter estimation and optimal experimental design. *Essays in Biochemistry*, *45*, 195–209.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.

Busetto, A. G., & Buhmann, J. M. (2009). Structure identification by optimized interventions. *Journal of Machine Learning Research Proceedings of the Int. Conf. on Artificial Intelligence and Statistics, Clearwater Beach, Florida USA* (pp. 49–56).

Chaloner, K., & Verdinelli, I. (1995). Bayesian experimental design: a review. *Statistical Science*, *10*, 273–304.

Heine, T., Kawohla, M., & King, R. (2008). Derivative-free optimal experimental design. *Chemical Engineering Science, Model-Based Experimental Analysis*, *63*, 4873–4880.

Kitano, H. (2002). Computational systems biology. *Nature*, *420*, 206–210.

Kitano, H., Funahashi, A., Matsuoka, Y., & Oda, K. (2005). Using process diagrams for the graphical representation of biological networks. *Nature Biotechnology*, *23*, 961–966.

Kuepfer, L., Peter, M., Sauer, U., & Stelling, J. (2007). Ensemble modeling for analysis of cell signaling dynamics. *Nature Biotechnology*, *25*, 1001–1006.

Montgomery, D. C. (2004). *Design and analysis of experiments*. Wiley.

van den Berg, J., Curtis, A., & Trampert, J. (2003). Optimal nonlinear Bayesian experimental design: an application to amplitude versus offset experiments. *Geophysical Journal Int.*, *155*, 411–421.

Wagner, A. (2005). *Robustness and evolvability in living systems (Princeton studies in complexity)*. Princeton University Press.