BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 10 2013, pages 1283–1291
doi:10.1093/bioinformatics/btt130

Systems biology

Advance Access publication March 16, 2013

Sorad: a systems biology approach to predict and modulate dynamic signaling pathway response from phosphoproteome time-course measurements

Tarmo Äijö[1,2,*], Kirsi Granberg[3,4] and Harri Lähdesmäki[1,*]

[1]Department of Information and Computer Science, Aalto University, FI-00076 AALTO, Finland, [2]Division of Signaling and Gene Expression, The La Jolla Institute for Allergy and Immunology, La Jolla, CA 92037, USA, [3]Department of Signal Processing, Tampere University of Technology, FI-33101 TAMPERE, Finland and [4]Department of Pathology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA

Associate Editor: Mario Albrecht

ABSTRACT

Motivation: Signaling networks mediate responses to different stimuli using a multitude of feed-forward, feedback and cross-talk mechanisms, and malfunctions in these mechanisms have an important role in various diseases. To understand a disease and to help discover novel therapeutic approaches, we have to reveal the molecular mechanisms underlying signal transduction and use that information to design targeted perturbations.

Results: We have pursued this direction by developing an efficient computational approach, Sorad, which can estimate the structure of signal transduction networks and the associated continuous signaling dynamics from phosphoprotein time-course measurements. Further, Sorad can identify experimental conditions that modulate the signaling toward a desired response. We have analyzed comprehensive phosphoprotein time-course data from a human hepatocellular liver carcinoma cell line and demonstrate here that Sorad provides more accurate predictions of phosphoprotein responses to given stimuli than previously presented methods and, importantly, that Sorad can estimate experimental conditions to achieve a desired signaling response. Because Sorad is data driven, it has a high potential to generate novel hypotheses for further research. Our analysis of the hepatocellular liver carcinoma data predict a regulatory connection where AKT activity is dependent on IKK in TGFα stimulated cells, which is supported by the original data but not included in the original model.

Availability: An implementation of the proposed computational methods will be available at http://research.ics.aalto.fi/csb/software/.

Contact: tarmo.aijo@aalto.fi or harri.lahdesmaki@aalto.fi

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

Received on October 11, 2012; revised on February 17, 2013; accepted on March 11, 2013

1 INTRODUCTION

Binding of the ligand on the cell surface receptor initiates a signaling cascade that is propagated via several parallel pathways through phosphorylation of specific amino acid residues in signaling proteins. Building a mechanistic understanding of signaling pathway dynamics from experimental data has remained challenging for several reasons. Although mass spectrometry–based techniques have witnessed considerable progress recently, it is laborious to measure active forms of proteins in a highly quantitative and high-throughput manner. Furthermore, unraveling functional interactions between phosphoproteins demands a large number of experiments where the key signaling molecules are perturbed, or even co-perturbed, by using direct gene knockouts, gene knockdowns, specific inhibitors or other external stimuli.

A number of different computational strategies have been proposed to model signal transduction. A methodology based on Bayesian networks for describing the interconnections between signaling proteins has been used to study T cell signaling (Sachs *et al.*, 2005). Similarly, Aldridge *et al.* (2009) used an approach based on fuzzy logic, an extension to the two valued Boolean algebra, to study the downstream signaling of tumor necrosis factor, epidermal growth factor and insulin receptors. While logic-based approaches provide models that are easy to interpret and analyze computationally, they do not allow biophysically motivated mechanistic modeling approaches, which inevitably require the use of dynamic and continuous models of signaling networks (reviewed in Aldridge *et al.*, 2006; Chakraborty and Das, 2010). Ordinary differential equations (ODE), which represent the average of detailed chemical reaction models in a cell population, have thus become preferable in studies focusing on signaling dynamics. For example, Chaudhri *et al.* (2010) constructed an ODE model to describe receptor-dependent mitogen-activated protein kinase signaling based on prior biological knowledge about the network structure, and also identified a novel regulatory motif. Reverse engineering of dynamic models of molecular networks has also been extensively studied in the context of transcriptional regulation, and several ODE modeling methods have been proposed (Äijö and Lähdesmäki, 2009; Bansal *et al.*, 2006; Bonneau *et al.*, 2006; Cantone *et al.*, 2009; Gao *et al.*, 2008; Honkela *et al.*, 2010; Titsias *et al.*, 2012). Our approach differs from the previous methods by using continuous non-parametric dynamics and, importantly, by providing a way to modulate network's response as described below.

*To whom correspondence should be addressed.

A far-reaching goal is to use mathematical models and computational methods to modulate signaling pathway responses in healthy and diseased conditions. For example, Mitsos *et al.* (2009) attempted to detect drug targets from phosphoproteomic data by identifying changes in a pathway induced by a treatment relative to an untreated control. Given the complexity of signaling pathway dynamics, accurate modulation strategies inevitably require the use of a mechanistic model of network dynamics. Model-based intervention approaches have been studied in the context of probabilistic Boolean network models (Shmulevich *et al.*, 2002), but there is an interest on how the intervention strategies can be implemented for continuous and dynamical models.

Here, we present a flexible non-parametric ODE model and propose an efficient and scalable network structure learning algorithm to identify signaling pathways and their dynamics from phosphoprotein data. This modeling method, called Sorad, assumes the standard ODE formulation but alleviates the detailed parametric model specifications by using non-parametric functions. This has important consequences for the signaling pathway reconstruction as we have to estimate only the kinetic parameters α and λ and the hyperparameters of the covariance function, but no parameters directly related to the regulatory functions. When applied to previously unseen stimuli, Sorad provides quantitatively accurate signaling pathway responses. More importantly, we show how Sorad can be used to design interventions for the modulation of said pathway's response and dynamics. This novel and unique method enabling the design of optimal stimuli (e.g. receptor stimuli, drug treatment) encourages the discovery of new methodologies for experimental design, drug discovery and therapeutic applications. To demonstrate Sorad's performance, we make use of the DREAM project, an initiative to foster collaboration between experimental and computational biologists (Prill *et al.*, 2011), and validate our predictions on independent experimental data from a human carcinoma cell line (part of DREAM4) originally published in (Alexopoulos *et al.*, 2010; Saez-Rodriguez *et al.*, 2009).

2 METHODS

2.1 A non-parametric and probabilistic model for continuous signaling pathway dynamics

Full description and derivation of the computational methods can be found from Supplementary Material. We model signaling pathway dynamics using first-order linear ODEs where the non-linear driving function for each phosphoprotein x_i is unknown. The time-dependent phosphorylation level for each phosphoprotein is modeled using three components:

$$\dot{x}_i(t) = f_i(t) + \alpha_i - \lambda_i x_i(t) \tag{1}$$

where $x_i(t)$ is level of the ith ($i = 1, \ldots, N$) phosphoprotein at time t, $\dot{x}_i(t)$ denotes the time derivative, α_i is the basal rate, which captures the constant (non-zero) part of the phosphoprotein data, λ_i is the degradation rate, which models spontaneous decrease or degradation of phosphoprotein level, f_i is the non-parametric regulatory function, which generates the dynamic changes in the phosphoprotein level by time (a.k.a. models the actual signaling mechanism). We set a Gaussian process prior over regulatory functions f_i to prefer smooth functions. We assume that inactive form of the protein is not limiting the rate of phosphorylation. Traditional ODE modeling approaches typically encounter problems with parameter estimation and model structure comparison because neither of those have analytically tractable closed form solutions (except special cases). As is explained below, the analytical solution, which is inherited from the Gaussian process formulation, is one of our main motivations because we can thus avoid problems associated with parameter estimation steps, have an analytical model selection score and have a build-in regularization via the Bayesian analysis. Moreover, our chosen formulation also allows efficient experimental design.

2.2 Defining signaling network dynamics

For a given (fixed) signaling pathway structure, the key challenge is to define the non-parametric functions such that the signaling network dynamics agree with experimental data. Previously, a similar estimation problem for non-parametric functions f_i has been solved in the context of a discrete-time system (Penfold *et al.*, 2012) or using the first-order approximation of the derivative (Äijö and Lähdesmäki, 2009). Here, we develop a probabilistic estimation method for continuous dynamics using a two-step approach, as illustrated inside the topmost rounded rectangle in Supplementary Figure S10. The first step is to identify the driving functions, including basal rate, degradation rate and regulatory function as a function of time only, $f_i(t)$. This step is done independently for each of the measured phosphoproteins by solving the non-parametric ODE model in continuous time such that the obtained function $x_i(t)$ explain the measurements as accurately as possible, a step that only requires measurements of the corresponding phosphoprotein. Conceptually, when considered as a function of time, the non-parametric regulatory function tells for any time point the instantaneous rate of (de)phosphorylation needed to obtain the observed dynamics. The phosphorylation level of x_i can be solved to yield

$$x_i(t) = \frac{\alpha_i}{\lambda_i} + c_i e^{-\lambda_i t} + \int_0^t f_i(\tau) e^{-\lambda_i (t-\tau)} d\tau \tag{2}$$

where c_i depends on the initial condition at $t = 0$. We assume that the time-dependent driving function f_i is different for each experimental condition but phosphoprotein-specific kinetic parameters α_i and λ_i (as well as hyperparameters of the Gaussian process) are shared over different conditions. The linear integral transformation applied to the function f_i in Equation (2) preserves the Gaussian process property of f_i. Thus, the process x_i is also a Gaussian process, i.e. $x_i(t) \sim \mathcal{GP}(m_{x_i}(t), k_{x_i,x_i}(t,t'))$, where the mean is

$$m_{x_i}(t) = \frac{\alpha_i}{\lambda_i} + c_i \exp(-\lambda_i t) \tag{3}$$

and the covariance function $k_{x_i,x_i}(t,t')$ is shown in Supplementary Equation (24). The kinetic parameters α_i and λ_i and hyperparameters are estimated by optimizing the (log) marginal likelihood of the data, which can be written as

$$\begin{aligned} \log p(\mathbf{x}_i | T, \theta) = &-\frac{1}{2}(\mathbf{x}_i - \mathbf{m}_{x_i})^T (K_{x_i,x_i} + \sigma_{n,i}^2 I)^{-1} (\mathbf{x}_i - \mathbf{m}_{x_i}) \\ &-\frac{1}{2}\log |K_{x_i,x_i} + \sigma_{n,i}^2 I| - \frac{n}{2}\log 2\pi \end{aligned} \tag{4}$$

where $\mathbf{x}_i$ contains the phosphoprotein measurements for x_i, T is a vector of measurement time indices, θ represents the parameters and hyper parameters, $\mathbf{m}_{x_i}$ is a vector of values of m_{x_i} evaluated at T, K_{x_i,x_i} is the covariance matrix between x_i and x_i evaluated at T, $\sigma_{n,i}^2$ denotes the measurement noise variance for x_i, and n denotes the number of measurement time points. For partial derivatives of Equation (4), see Supplementary Equations (36–41). Finally, the driving functions f_i are estimated using the posterior means. In particular, predictive equations for f_i can be written in the standard form (omitting the phosphoprotein index i in the following)

$$\mathbf{f}_* | \mathbf{x}, T, T_* \sim \mathcal{N}(\bar{\mathbf{f}}_*, \mathrm{Cov}(\mathbf{f}_*)), \quad \text{where}$$

$$\bar{\mathbf{f}}_* = K_{f,x}(T_*, T) K_{x,x}(T, T)^{-1} \mathbf{x}, \quad \text{and}$$

$$\mathrm{Cov}(\mathbf{f}_*) = K_{f,f}(T_*, T_*) - K_{f,x}(T_*, T) K_{x,x}(T, T)^{-1} K_{x,f}(T, T_*)$$

where $K_{f,x}$ is the covariance matrix between f_i and x_i evaluated at T, $\mathbf{f}_*$ denotes the points to be predicted and $\bar{\mathbf{f}}_*$ is our final estimate (see derivations in Supplementary Material and Supplementary Equation (11) for details). As a result, we get the estimated values for the kinetic parameters α_i and λ_i and the condition-dependent functions f_i for each phosphoprotein.

The second step finds a mapping from measured phosphoprotein time-course profiles of a specific set of regulatory proteins to the inferred time-dependent regulatory function. This is done with the following regression model:

$$f_i(t) = g_i\left(\mathbf{x}_i^{\text{reg}}(t)\right) \tag{5}$$

where $\mathbf{x}_i^{\text{reg}}(t) = (x_{i_1}(t), \ldots, x_{i_k}(t))$ denotes the activities of k phosphoproteins at time t regulating the phosphoprotein x_i and g_i is an unknown non-parametric function with Gaussian process prior. Function g_i is estimated using the Gaussian process regression at measurement time points. The s.c. final g_i function will be refined during the validation step together with the construction of the final network structure. While f functions need to be estimated separately for each experiment, for the estimation of g we use all experiments together and find a mapping from regulatory phosphoproteins to the target protein that explains all the experimental (training) data. In other words, g functions represent the actual mechanisms that propagate signal through the network. A schematic diagram of the model and its inference is shown in Supplementary Figure S25.

2.3 Unraveling signaling pathway model from data

The above model inference applies for a given fixed signaling pathway structure, which is completely specified by the functions g_i and their inputs $\mathbf{x}_i^{\text{reg}}$. Fortunately, we can iteratively apply the second step to different combinations of regulatory signaling components $\mathbf{x}_i^{\text{reg}}$ to assess the confidence in each network structure. In this study, we use Bayesian and cross-validation approaches to rank different signaling network structures. In addition, we use the information included in f functions.

In each of the cross-validation cycles, the alternative models are fitted to the training data and the corresponding prediction performances are assessed using the test data. Note that we use only the original training data for model fitting. For Bayesian analysis, we need the marginal likelihoods of different models, which for Gaussian process–based regression in Equation (5) can be written analytically

$$\log p(\bar{\mathbf{f}}|T,\theta) = -\frac{1}{2}\bar{\mathbf{f}}^T K_{f,f}(T,T)^{-1}\bar{\mathbf{f}} - \frac{1}{2}\log\left|K_{f,f}(T,T)\right| - \frac{n}{2}\log 2\pi$$

where θ represents the hyperparameters, and θ is again optimized by maximizing the marginal likelihood [see Supplementary Material and Supplementary Equations (43–46) for partial derivatives]. In addition to the built-in regularization rising from the use of marginal likelihood, we noticed a need for penalizing the models, which are composed of many explanatory variables. We defined a prior distribution for the signaling networks in a similar way as in the Akaike information criterion that penalizes models based on the number of variables, i.e. $p(M) \propto \exp(-2k)$, where k is the number of directed connections in a network structure M, and $k_{\max}$ is a maximum number of explanatory variables to be considered. As a result of cross-validation and Bayesian analysis, we obtain scores for different signaling networks, which can be summarized to pair-wise relationship between phosphoproteins by summing the scores of all model structures that contain a specific directed interaction. These scores can be used to rank links between proteins.

The estimated functions f are also used to choose the signaling network structure for two reasons. First, cross-validation and Bayesian approaches depend on our choice of models of the perturbations (e.g. unknown dynamics of external cytokine and growth factor perturbations), whereas the analysis on f functions does not. Second, cross-validation and Bayesian approaches would fail to identify phosphorylation events that are active only in a single experiment. The f function represents the instantaneous rate of change of a phosphoprotein. If our method infers remarkably different dynamics for f in a single time-series experiment, that gives evidence for dependencies that are likely observable only under that particular condition. Thus, hierarchical clustering analysis is done using average linkage and the Euclidean distance over f functions to reveal dependencies between f functions and perturbations. Clustering can be interpret by looking at how different growth factor stimulations cluster separately if any of the clusters get higher or lower f values, and whether certain upstream inhibitions cause f functions to cluster into another cluster with lower/higher values than f functions in an uninhibited case under the same growth factor stimulation.

The final dynamical system we obtain by interchanging the functions f_i in Equation (1) with the functions g_i can be expressed in a vector form as

$$\dot{\mathbf{x}}(t) = \mathbf{g}(\mathbf{x}(t)) + \alpha - \text{diag}(\lambda)\mathbf{x}(t) \tag{6}$$

where $\mathbf{x}(t) = (x_1(t), x_2(t), \ldots, x_N(t))^T$, $\mathbf{g}(x(t)) = (g_1(\mathbf{x}_1^{\text{reg}}(t)),\ \mathbf{g}_2(\mathbf{x}_2^{\text{reg}}(t)), \ldots, \mathbf{g}_N(\mathbf{x}_N^{\text{reg}}(t)))^T$ and the kinetic parameters $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_N)^T$ and $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_N)^T$. This system can be solved numerically over time given the perturbations and the initial activity levels of the phosphoproteins $\mathbf{x}(0)$. Occasionally, predictions can be negative, which in this study were truncated to zero by biological reasoning.

2.4 Prediction of interventions

Note that given perturbations and initial activity levels of the phosphoproteins, it is a trivial task to simulate the behavior of the dynamic system in Equation (5) over time using, for example, Euler's method. However, a much harder problem is to estimate the perturbations given the desired behavior of a system over time. The problem can be stated as follows: Given a desired time profile of a target protein's activity level, estimate the time profiles of one or several of its regulatory proteins to achieve the desired signaling response. We propose a novel method to design interventions, which consists of two-step solution. First, given the desired behavior of the target protein x_j, estimate the regulatory function $f_j(t)$ as explained above. Second, estimate the initial values of the inputs $\mathbf{x}_j^{\text{reg}}$ to the function g_j in such a way that it approximates the ideal phosphorylation function $f_j(t)$ as well as possible. Because the analytically tractable Gaussian processes properties apply to the system state x_i as well as to the regulatory functions f_i and g_i, our modeling framework is particularly well suited to design modulation strategies. The task of predicting interventions can be seen as an inversion of a Gaussian process. For the sake of simplicity, let us consider situation of a single test point $\mathbf{t}_* \in \mathbb{R}^p$, i.e. the inputs are p-dimensional, and the training data is composed of the training inputs T (i.e. phosphorylation time-course data) and outputs $\mathbf{g}$. In that case, the predictive equation for a single test point is (Rasmussen and Williams, 2006)

$$\bar{g}_* = K_{g,g}(T,\mathbf{t}_*)K_{g,g}(T,T)^{-1}\mathbf{g} \tag{7}$$

where $K_{g,g}(T,\mathbf{t}_*) \in \mathbb{R}^{1\times N}$ holds the covariances between N training points and the test point $\mathbf{t}_*$. Now, if we turn the traditional situation the other way around, i.e. we assume that we know the desired output (denoted as $\hat{g}_*$) but we do not know the value of the test point $\mathbf{t}_*$ (input). From Equation (7) we notice that the only term that depends on $\mathbf{t}_*$ is the vector $K_{g,g}(T,\mathbf{t}_*)$. Each of the elements in the vector $K_{g,g}(T,\mathbf{t}_*)$ are given by the covariance function $k_{g,g}$, e.g. ith element is the covariance between the ith training input and the test point $\mathbf{t}_*$. Using the mean square error criterion, we can then write the optimization problem for perturbations as

$$\arg\min_{\mathbf{t}_*} ||\hat{g}_* - K_{g,g}(T,\mathbf{t}_*)K_{g,g}(T,T)^{-1}\mathbf{g}||_2 \tag{8}$$

In other words, the optimal perturbation corresponds to the initial phosphoprotein levels $\mathbf{t}_*$, which minimize Equation (8). Additionally, the search for optimal perturbations in Equation (8) can be combined

with the uncertainty of the predictions (see Supplementary Material for details).

Overall, Sorad's computational complexity has two major steps. First, inference of dynamics, network structure and interventions all involve inversions of covariance matrices, which have no more than cubic asymptotic time complexity. Second, for each phosphoprotein, inference for dynamics and regulatory proteins need to be applied for all 2^N combinations of regulatory proteins. Thus, for large signaling networks, an upper bound on the number of regulatory proteins needs to be set, hence reducing the exponential term to a polynomial time complexity. Owing to the design of Sorad, however, the estimation of the f and g functions is an embarrassingly parallel problem and thus the combinatorially increasing computational load can be distributed.

3 RESULTS

3.1 Phosphoprotein time course from HepG2 cell line

In our training dataset, previously published in (Alexopoulos *et al.*, 2010), four cytokines or growth factors (IGF1, TGFα, IL1α and TNFα) were individually used to stimulate hepatocellular carcinoma cell line HepG2. In addition, inhibitors for MEK1/2, p38, PI3K and IKK proteins were used to dissect the upstream regulators for the measured proteins, resulting in altogether $(4+1)\times(4+1)=25$ different conditions. The activity levels of seven phosphoproteins, namely ERK1/2, HSP27, JNK1/2, IKB, MEK1/2, p38 and AKT, were monitored in the 25 different conditions in a time-series manner, measurements being taken at 0, 30 and 180 min. For the test dataset, we use an independent time-course dataset from (Alexopoulos *et al.*, 2010), which includes the same set of seven phosphoproteins with $5\times4=20$ different perturbations: the inhibitions were five different pairs of individual perturbations, namely p38 + MEK1/2, PI3K + MEK1/2, p38 + PI3K, p38 + IKK and PI3K + IKK, and the stimuli were IL1α, IGF1, TGFα and the pair TGFα + IGF1.

3.2 Modeling HepG2 phosphoprotein dynamics

We first estimated values of the α and λ parameters for each of the seven phosphoproteins (Supplementary Table S1). For example, based on the model fitted to the data, HSP27 has the lowest basal rate. A closer manual inspection on the data supports this finding, i.e. the level of HSP27 is low overall and is only activated when stimulated with IL1α (Supplementary Fig. S7). Similar reasoning can be done for the low basal rates of ERK1/2, IKB, JNK1/2 and p38.

The estimated regulatory functions f for AKT is visualized in Figure 1 (for other phosphoproteins see Supplementary Figs S3–S8). A separate subplot is shown for each of the individual experimental settings, including the experiment-specific measurements, f function and the continuous profile of the phosphorylated AKT level. It is evident that the estimated functions are non-linear, and moreover, that they could not be approximated with linear functions without drastic effects, emphasizing the suitability of the presented non-parametric methodology.

Recall that g is an approximation for the set of functions f, and it is a function of activity levels of regulatory phosphoproteins in contrast to functions f that were functions of time. It was found out that the behavior of f functions, e.g. for ERK1/2, are best explained by altogether three variables, TGFα, MEK1/2 and

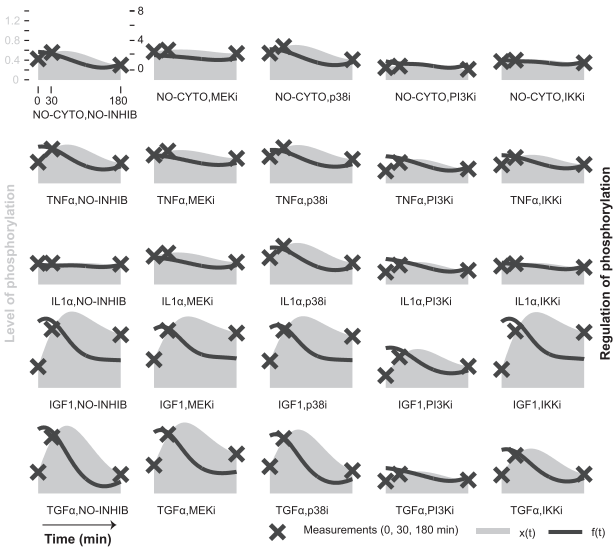

Fig. 1. The regulatory functions f of AKT over the training data. The estimated functions f are represented by the solid orange lines. The crosses are the measurements that were used to fit the model. The filled curves represent the continuous profile x that is estimated to produce the measurements. The unit of measurements is scaled/arbitrary. The left- and right-hand sided y-axes differ but remain the same between the subfigures and they cover the range from zero to the maximum value measured for AKT (the crosses and the filled curves) and from the minimum to the maximum value of the regulatory function (the solid orange lines)

ERK1/2 (see the next section). The estimated regulatory function g for ERK1/2 is shown in Supplementary Figure S9. The topmost part of the figure illustrates the estimated outputs of f and g functions across the 25 different experiments, and the lower part visualizes the values of the explanatory variables. The initial boost in the activation of ERK1/2 is explained by the stimulation of TGFα receptor, and its effect is mediated by MEK1/2.

3.3 Unraveling the cell-type–specific signaling network

Altogether three types of criteria were used to reveal the data-supported relationships between the phosphoproteins: prediction performance (cross-validation), fit of the model (marginal likelihood) and effects of individual perturbations (clustering of f functions). The data-supported confidences of the relationships for ERK1/2 are shown in Figure 2A and B. The marginal likelihood and cross-validation scores suggest that the changes in the activity level of ERK1/2 are well explained by the activity levels of ERK1/2, MEK1/2 and TGFα. It is interesting to see that ERK1/2 is the strongest hit from the cross-validation–based analysis, i.e. has the lowest negative log-likelihood score among the possible regulators. As another example, Figure 2C and D shows the data-supported confidences of the relationships for IKB. In Figure 2C, both of the analyses suggest that the activation level of IKB and IKK and the perturbation status of IL1α are good predictors of the activation of IKB. The regulatory role of TNFα is suggested by the marginal likelihood–based analysis although it was less evident according to the cross-validation results. Moreover, the roles of IL1α, TNFα and IKK are supported by the results of the hierarchical clustering of f functions as

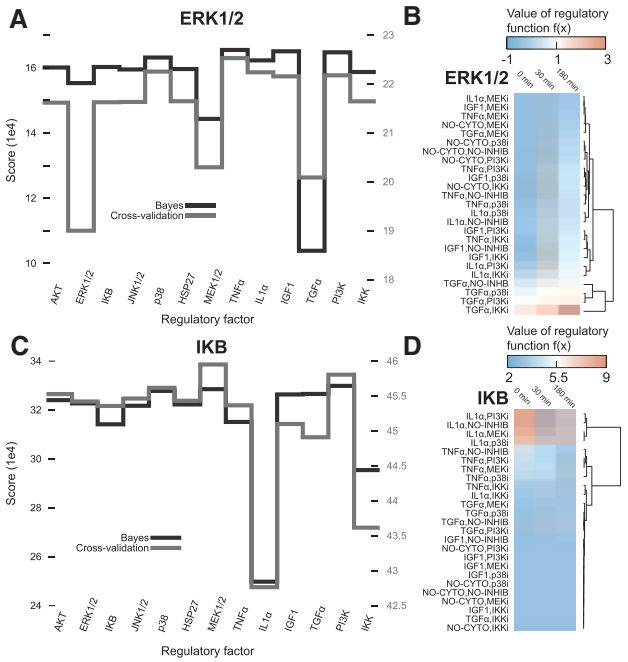

Fig. 2. The information used to infer the regulators of ERK1/2 and IKB. (**A**) Here the marginal likelihood and cross-validation scores (in negative log scale) are summarized to represent the confidences between the pairwise relationships for the target protein ERK1/2. The relationships with small scores are more reliable and are preferred. (**B**) The hierarchical clustering of f functions of ERK1/2 for all perturbations (indicated on the left) in the training data. TGFα-stimulated samples form a separate cluster with higher f function values, suggesting that TGFα activates ERK1/2. However, the *f* function under TGFα-stimulated MEK-inhibited conditions have lower values and clusters into another cluster, suggesting that MEK1/2 mediates the activating effects of TGFα. According to all the results, TGFα, MEK1/2 and ERK1/2 itself regulate ERK1/2 activity. Inhibition is denoted with the character 'i' at the end of the protein names (e.g. MEKi). (**C** and **D**) The same analysis when the target is IKB

illustrated in Figure 2D, as the stimulation of either IL1α or TNFα receptors discriminate *f* functions but this does not hold when IKK has been inhibited. Therefore, it can be concluded that IKB, IL1α, TNFα and IKK is the best set of explanatory variables that are used to predict the activity level of IKB.

Some of the dependencies between the phosphoproteins are only observable in one condition, e.g. mediating role of p38 in the activation of HSP27 when IL1α is used for stimulation (see Supplementary Fig. S7). In these cases, the cross-validation fails because the dependency is only observable in the training data or test data but not in both. However, the marginal likelihood–based approach and the clustering of *f* functions reveal these unique relationships and are hence used to supplement cross-validation. The computational method for combining the marginal likelihood–, cross-validation– and clustering-based information is not fully automatic, although such an approach could be developed. Our view on the approach is such that, owing to the diversity of interactions between signaling proteins, it is better to produce useful and orthogonal data for the user and to allow partial manual control, combined with biological knowledge, to reason about the signaling events as demonstrated above.

The data-supported network model of the signal transduction in the HepG2 cell line is shown in Figure 3 where no biological prior knowledge was used. In this study, the estimation of *g* functions cover all the models from a single explanatory variable up to the models consisting of six explanatory variables. Regulatory relationships downstream of proteins that forward the signal from the receptors to the pathway end points have been marked by dashed edges. Supplementary Figure S19 has been modified from Figure 3 by drawing the arrows coming from the cytokines or growth factors only to the mediating upstream regulators. Most of the inferred interactions are supported by the literature. For example, the mediating role of p38 in IL1α-induced activation of HSP27 was inferred correctly (Alexopoulos *et al.*, 2010). Interestingly, our model predicted donwstream signaling to be dependent on the original growth factor stimuli, whereas many models predict that signal is propagated in a similar manner in the network once it is initiated. However, it is really the case in cellular biology that initial stimuli can affect the outcome because each growth factor regulates differently the receptors and adaptor proteins that mediate the signal transduction. In principle it is possible to model this, but all the necessary information is not available at the moment. The fact that proteins can be regulated from several sites complicates their behavior even further. In addition, different signaling pathways can crosstalk together generating (in)activation effects on adjacent pathways. This all underlines the importance of modeling the signal transduction in stimulus-dependent manner.

3.4 Sorad provides accurate predictions

The ability of Sorad to predict phosphoprotein activity levels in a signaling pathway is illustrated in Supplementary Figure S10. First, the optimal model topology is inferred as described above, and its dynamics are learned from the whole training data. The model can then be initialized with different phosphoprotein levels and perturbation configurations, and the response of the system can be solved numerically. The third challenge in the DREAM4 was to unravel a signaling network based on experimental data, and the model fit was assessed by the prediction performance. The task was to predict the activity levels of the phosphoproteins at 30 min after the initiation of cytokine stimuli given the initial activity levels of the phosphoproteins and the information about the perturbations. The goal of the challenge is to find the best predictive model with the minimal number of connections in the signal transduction network.

Table 1 lists the results of Sorad and the four best performing teams from DREAM4. All the attributes listed in Table 1 are computed as in DREAM4 and are explained in 'Performance metrics' section in Supplementary Material. Sorad performed well because it had the best prediction score as well as the smallest number of edges in the network. The best performer in the DREAM4 challenge used a methodology formulated within a Boolean logic framework (Eduati *et al.*, 2010). The first step in their method statistically identifies whether a certain perturbation has an effect on the phosphorylation of a given protein (binary decision) and constructs a Boolean network model. In the second step, they link these inferred relationships with

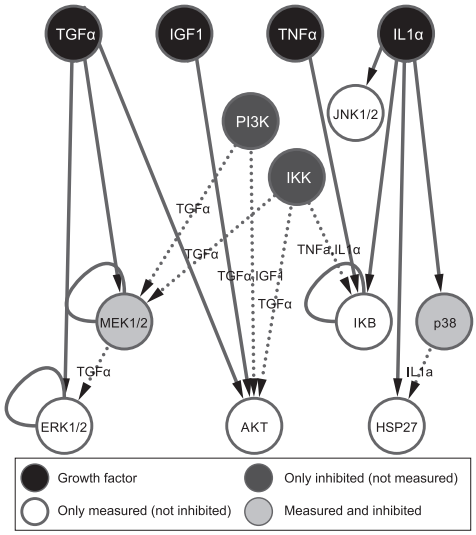

Fig. 3. The topology of the inferred network. Four cytokines or growth factors are represented by black colored nodes and other proteins either by gray or white nodes. The color of the signaling protein nodes indicates whether they have been inhibited in some of the experiments and/or measured in all the experiments. Edges in the graph show the causal interactions between the nodes. A dashed edge represents an interaction where the inhibition of the regulatory phosphoprotein dampens the activation produced by the marked growth factor stimulation

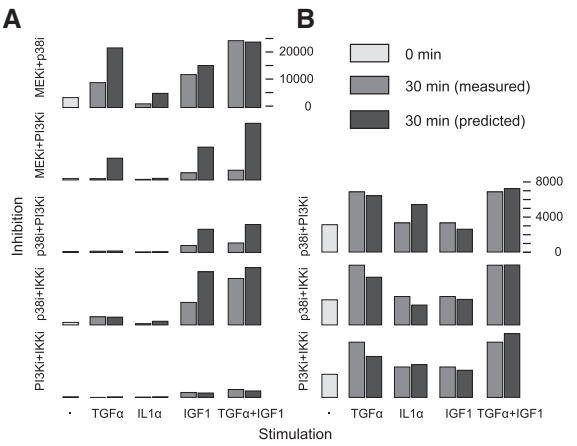

Fig. 4. Predicted behavior of (**A**) AKT and (**B**) MEK1/2 in the unseen perturbations conditions and their observed phosphorylation levels in real measurements. The leftmost bars are the provided values for 0 min time point for a given inhibition. They show the initial baseline for all the predicted and measured values at 30 min time point with the same inhibition (bars on their right side). The blue bars represent the measured phosphorylation level at 30 min and the orange bars show the predictions. Units are arbitrary (fluorescence levels), but *y*-axis is equal in all the subfigures for one predicted/measured protein covering the range from zero to the maximum value of the measurements of the corresponding protein

Table 1. Comparison of predictions on the independent test data

Attribute	Sorad	Team 441	Team 476	Team 533	Team 491
AKT	1.080e-05	4.656e-05	9.429e-04	2.085e-05	7.797e-04
ERK1/2	9.475e-013	1.568e-09	3.522e-14	1.891e-16	1.101e-08
IKB	6.140e-010	3.782e-10	5.998e-09	3.782e-10	1.292e-08
JNK1/2	1.508e-10	1.732e-10	1.310e-10	3.791e-10	5.301e-11
P38	1.375e-08	1.059e-08	1.378e-10	1.871e-06	4.821e-05
HSP27	1.782e-10	8.289e-11	5.379e-06	1.886e-06	9.276e-07
MEK1/2	1.424e-08	1.615e-07	4.014e-05	4.900e-09	1.118e-06
Edges	16	18	17	26	18
Prediction	8.783	8.167	7.730	8.430	6.505
Overall	7.460	6.678	6.324	6.279	5.016

Prediction *P*-values for different phosphoproteins, the number edges in a network, combined prediction score and the overall score for different methods.

the experimental data by combining linearly the observed effects of the perturbations. The second top best performing team also developed a two-step approach: a parametric linear ODE model is first used to infer the topology of the signaling pathway and this model is used as the starting point for a predefined non-linear ODE model that is fitted to the experimental data (The New York Academic of Sciences, 2009).

The performance metric *p*-values show that the activity levels of AKT are the most challenging to predict. To study this more closely, we have visualized the test set measurements together with our predictions for AKT in Figure 4A. The largest differences between the measurements and the predictions are due to the underestimated stimulative effect of TGFα and IGF1. As AKT lies downstream of TGFα, IGF1, PI3K and IKK (Fig. 3), it is important that cooperative effects of the perturbations are estimated correctly.

Figure 4B shows Sorad's predictions and the test set measurements for MEK1/2 across the 15 test conditions. The cooperative effect of TGFα and IGF1 receptors is challenging for modeling because the stimulation of IGF1 growth factor alone is not sufficient to activate MEK1/2. However, Sorad can capture this non-linear co-operativity well. The corresponding predictions for other five phosphoproteins are visualized in Supplementary Figures S11–S15. Taken together, we conclude that proposed non-parametric ODE model provides a predictive modeling framework that is generally quantitatively accurate.

3.5 Upstream regulators of AKT

We identified AKT as an interesting candidate protein for a more detailed analysis. We studied the effect of different perturbations on the activity level of AKT based on the training data (Supplementary Fig. S16A). Interestingly, the model suggests that TGFα or IGF1 stimulation induces approximately the same level of AKT activity. Similarly, we notice that the inhibition of PI3K has a greater negative influence on the activation of AKT than the inhibition of IKK, regardless of whether TGFα or IGF1 stimulation is used, which is in accordance with the PI3K-dependent activation of AKT (Chin and Toker, 2009). Strikingly, supporting a finding previously reported in (Eduati *et al.*, 2010), AKT activation is dependent on IKK inhibition when TGFα receptor is stimulated (Supplementary Fig. S16). High specificity of the inhibitor (Burke *et al.*, 2003) and TGFα-assiocated regulation suggests that the observed IKK-AKT interaction is not caused by unspecific side effects.

The IKK-dependent activation of AKT on TGFα stimulation is supported by both the training and the independent test data. For example, when TGFα receptor is stimulated, AKT activity decreases only on inhibition of IKK, PI3K or both, whereas IKK-dependent activation of AKT cannot be observed in IGF1-stimulated cells (Figs 1 and 4A and Supplementary Fig. S16).

3.6 Computational design of interventions to generate desired signaling pathway response

Next we study the suitability of Sorad for predicting experimental conditions to acquire a desired signaling response, for which we again use the original partitioning of the training and test data. We study two different cases where the activation of AKT is controlled: either the activity level of AKT remains constant or it is strongly increased after the perturbations (150% of the maximum activity level of AKT in training data). The estimated perturbations that are needed for the desired AKT activity levels are shown in Figure 5A–D. In each of the cases, we consider five different initially fixed perturbation conditions: none of the perturbations are fixed, TGFα stimulation is either mandatory or not permitted, or IGF1 stimulation is mandatory or not permitted. The four regulators are partioned into two groups based on the experimental setting: TGFα and IGF1 are regulators that can be used to stimulate the respective receptors, and PI3K and IKK are mediators that can be inhibited. As an example, consider the case where IGF1 receptor has been stimulated and the activity level of AKT is desired to remain constant (the top-left subfigure and the second bar chart group from top). When IGF1 receptor is stimulated, we get the estimated values for three free factors: TGFα receptor should not be stimulated and PI3K and IKK should be inhibited. From the biological point of view, this finding is arguable and, in addition, from Figure 4A (bottom row, the third panel from left), we notice that the only condition where IGF1 receptor is stimulated and AKT is not further activated is the one where TGFα receptor is not stimulated and PI3K and IKK are both inhibited, exactly as we estimated.

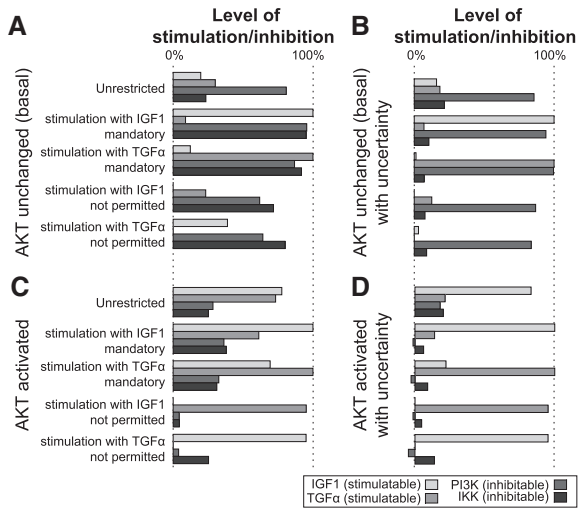

Fig. 5. Predicting the optimal perturbations to control the activation level of AKT in an unrestricted situation or under different preset conditions (stimulation with IGF1 or TGFα is mandatory or not permitted). The levels of bars in the bar charts reflect the suggested levels (concentrations) of IGF1 and TGFα growth factors as well as PI3K and IKK inhibitors, all of which can be used to acquire the desired activation level for AKT. The dashed lines mark at the levels of perturbations used in the training data. (**A**) In this prediction, it is desired that the activity level of AKT after 30 min is same as the initial activity level. (**B**) The situation is the same as in (A) but here the uncertainty of the prediction is taken into account in the optimization. (**C**) In this prediction, it is desired that AKT is fully activated after 30 min, i.e. the activity level of AKT is 150% of the maximum activity level of AKT in the training data. (**D**) Same situation as in (C), but here the uncertainty of the prediction is taken into account

A more sophisticated approach can weigh the predictions by the amount of uncertainty that is associated with it, which in principle should bias the estimated stimuli/inhibition levels from moderate to the ones used in the model construction. This is also a desired property from the experimental point of view because it would allow the use of the same experimental conditions as used already in the experiments to generate the training data. Figure 5B (corresponding to the case in Fig. 5A) and Figure 5D (corresponding to the case in Fig. 5C) show the estimated perturbations when we take into account the uncertainty of the estimations. It is clear that when the uncertainty in the estimations is taken into account, the estimated perturbations are closer to the perturbations that have already been used in the training data. For example, if it is desired to activate AKT (Fig. 5D) and it is given that either TGFα or IGF1 receptor is stimulated, then the estimated perturbations are similar to those in the experiments producing the training data. Another example includes the case where AKT is not further activated (Fig. 5A and B) under the stimulation of TGFα receptor: in this case, the approach that tries to minimize the amount of uncertainty predicts an experimental condition that requires one perturbation less than the one that is achieved by minimizing the error between the desired and predicted behavior. This can again be validated based on the data shown in Figure 4A, as we notice that it is enough to silence PI3K (inhibition p38i + PI3Ki) under the stimulation of TGFα receptor to keep AKT activity at the basal level. Because the predicted interventions to modulate the signaling response are validated by independent experimental data, we conclude that Sorad can be used to identify accurate modulation strategies.

3.7 Performance evaluation using *in silico* data

To better demonstrate Sorad's performance in predicting dynamics and perturbations, we set up an *in silico* signaling scenario similar with the one in the HepG2 cell line. We define a hypothetical ODE model (Supplementary Fig. S1A) where two proteins are responsible for the phosphorylation of a target protein and generate data from the model with additive Gaussian noise. Assuming the model structure is known, we first learn the dynamics of the model (functions f and g) together with α and λ, and then apply it to independent test data to predict the response. As shown in Supplementary Figure S1B–D, Sorad is able to learn the unknown regulatory function as well as make accurate predictions of signaling dynamics over long time intervals. We also tested Sorad on a more challenging *in silico* problem where effectively less data are available to learn the

phosphorylation dynamics *g*. Nevertheless, as shown in Supplementary Figures S2A–D, Sorad still makes relatively accurate predictions under uncertainty in estimated dynamics. To that end, we applied the estimated models (Supplementary Figs 1C and 2C) of phosphorylation dynamics to predict perturbations. First, we simulate the mathematical model and then, using the estimated dynamical model, estimated the optimal perturbation (or input) to obtain the observed model response. Results in Supplementary Figures S1E and S2E show how the estimated perturbation closely follows the (unknown) input used to generate the data, thus demonstrating Sorad's ability to predict dynamic perturbations over long time intervals.

4 DISCUSSION AND CONCLUSIONS

The presented methodology, Sorad, combines dynamic models with a data-driven non-parametric component, which leads to a flexible and probabilistic dynamical modeling framework. The first main contribution of this study is the efficient methodology, which makes the use of ODEs easy in situations where the construction of a parametric network model beforehand is challenging, as is typically the case in practice. The second main contribution is the scheme for predicting required perturbations for modulating the pathway response, which we demonstrated with a proof-of-concept example using real phosphoprotein time-course data.

To validate Sorad, we carried out a comparison that demonstrated its applicability for modeling signal propagation even without prior biological knowledge. This is supported by the performance assessment of the predictions: Sorad produced the most accurate predictions with the smallest number of relationships between the phosphoproteins. In addition to modeling signaling pathways, Sorad is also applicable for modeling other types of biological processes and it can be applied to even larger networks and datasets because the computation can be easily parallelized. However, if one is interested in analyzing larger networks consisting of hundreds or thousands of phosphoproteins or, for example, transcripts, the manual clustering step for the search of individual conditions where a target protein or gene is regulated differentially will become a bottleneck. For large networks, this step could be automated by using, for example, model-based clustering methods, which provide a quantitative and probabilistic scoring framework. Also note that the estimation of *f* functions is analytically tractable as long as it is possible to write the solution of the ODE, which only contains linear operations on *f* function.

A closer inspection of the pathway model that was inferred using Sorad pointed out a putative regulatory role for IKK in the activation of AKT in TGFα-stimulated cells. AKT has been generally thought to regulate IKK, thereby suggesting a reverse regulatory interaction between the proteins (Manning and Cantley, 2007). Low levels of IKB phosphorylation in TGFα-stimulated cells imply that NF-κB pathway is unlikely to mediate IKK-dependent AKT phosphorylation and we could not find any publications reporting IKK-dependent activation of ILK or mTORC2, the upstream kinases for AKT. Interestingly, IRS1 has potential to regulate AKT in HepG2 cells (Khamzina *et al.*, 2005): increased IRS1 phosphorylation (Ser636/Ser639) decreases both IRS1-PI3K interaction and AKT phosphorylation after stimulation through insulin receptor. Furthermore, IKKs bind to IRS1 in a basal state in HepG2 cells (Gao *et al.*, 2002). On TNFα stimulation, IKK-IRS1 interaction is disturbed and IRS1 gets phosphorylated on Ser312, leading to decreased IRS1 activity in insulin-treated cells. IRS1 gets phosphorylated on Ser636/Ser639 residues after TGFα (but not after IGF1) stimulation in the original dataset (Alexopoulos *et al.*, 2010), which correlates with IKK-dependent AKT activation only in TGFα-treated cells. Interestingly, MEK1/2 activation is also dependent on both PI3K and IKK after TGFα stimulation, suggesting a similar type of upstream regulation. The data suggest that IKK might regulate AKT (and MEK1/2) phosphorylation through IRS1 and PI3K in TGFα-treated cells. This idea can be addressed in adjacent biological studies. What comes to the prediction itself, it demonstrates how the data-driven nature of Sorad provides high potential to generate novel hypotheses for further experimental research.

ACKNOWLEDGEMENT

The authors thank Brittany C. Parker for her careful reading and suggestions on the manuscript.

Funding: This work was supported by the Academy of Finland (Centre of Excellence in Molecular Systems Immunology and Physiology Research (2012-2017), grants 135320 and 259038, EU FP7 grant EC-FP7-SYBILLA-201106, EU ERASysBio ERA-NET and FICS graduate school.

Conflict of Interest: none declared.

REFERENCES

Äijö,T. and Lähdesmäki,H. (2009) Learning gene regulatory networks from gene expression measurements using non-parametric molecular kinetics. *Bioinformatics*, **25**, 2937–2944.

Aldridge,B.B. *et al.* (2006) Physicochemical modelling of cell signalling pathways. *Nat. Cell Biol.*, **8**, 1195–1203.

Aldridge,B.B. *et al.* (2009) Fuzzy logic analysis of kinase pathway crosstalk in TNF/EGF/Insulin-induced signaling. *PLoS Comput. Biol.*, **5**, e1000340.

Alexopoulos,L.G. *et al.* (2010) Networks inferred from biochemical data reveal profound differences in toll-like receptor and inflammatory signaling between normal and transformed hepatocytes. *Mol. Cell Proteomics*, **9**, 1849–1865.

Bansal,M. *et al.* (2006) Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinformatics*, **22**, 815–822.

Bonneau,R. *et al.* (2006) The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets *de novo*. *Genome Biol.*, **7**, R36.

Burke,J.R. *et al.* (2003) BMS-345541 is a highly selective inhibitor of I kappa B kinase that binds at an allosteric site of the enzyme and blocks NF-kappa B-dependent transcription in mice. *J. Biol. Chem.*, **278**, 1450–1456.

Cantone,I. *et al.* (2009) A yeast synthetic network for *in vivo* assessment of reverse-engineering and modeling approaches. *Cell*, **137**, 172–181.

Chakraborty,A.K. and Das,J. (2010) Pairing computation with experimentation: a powerful coupling for understanding T cell signalling. *Nat. Rev. Immunol.*, **10**, 59–71.

Chaudhri,V.K. *et al.* (2010) Integration of a phosphatase cascade with the mitogen-activated protein kinase pathway provides for a novel signal processing function. *J. Biol. Chem.*, **285**, 1296–1310.

Chin,Y.R. and Toker,A. (2009) Function of Akt/PKB signaling to cell motility, invasion and the tumor stroma in cancer. *Cell Signal.*, **21**, 470–476.

Eduati,F. *et al.* (2010) A Boolean approach to linear prediction for signaling network modeling. *PLoS One*, **5**, e12789.

Gao,Z. *et al.* (2002) Serine phosphorylation of insulin receptor substrate 1 by inhibitor kappa B kinase complex. *J. Biol. Chem.*, **277**, 48115–48121.

Gao,P. *et al.* (2008) Gaussian process modelling of latent chemical species: applications to inferring transcription factor activities. *Bioinformatics*, **24**, i70–i75.

Honkela,A. *et al.* (2010) Model-based method for transcription factor target identification with limited data. *Proc. Natl Acad. Sci. USA*, **107**, 7793–7798.

Khamzina,L. *et al.* (2005) Increased activation of the mammalian target of rapamycin pathway in liver and skeletal muscle of obese rats: possible involvement in obesity-linked insulin resistance. *Endocrinology*, **146**, 1473–1481.

Manning,B.D. and Cantley,L.C. (2007) AKT/PKB signaling: navigating downstream. *Cell*, **129**, 1261–1274.

Mitsos,A. (2009) Identifying drug effects via pathway alterations using an integer linear programming optimization formulation on phosphoproteomic data. *PLoS Comput. Biol.*, **5**, e1000591.

Penfold,C.A. *et al.* (2012) Nonparametric Bayesian inference for perturbed and orthologous gene regulatory networks. *Bioinformatics*, **28**, i233–i241.

Prill,R.J. *et al.* (2011) Crowdsourcing network inference: the DREAM predictive signaling network challenge. *Sci. Signal.*, **4**, mr7.

Rasmussen,C.E. and Williams,K.I. (2006) *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA.

Sachs,K. *et al.* (2005) Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, **308**, 523–529.

Shmulevich,I. *et al.* (2002) Gene perturbation and intervention in probabilistic Boolean networks. *Bioinformatics*, **18**, 1319–1331.

Saez-Rodriguez,J. *et al.* (2009) Discrete logic modelling as a means to link protein signalling networks with functional analysis of mammalian signal transduction. *Mol. Syst. Biol.*, **5**, 331.

The New York Academic of Sciences. (2009) Academy eBriefings. In *RECOMB regulatory Genomics/Systems Biology/DREAM conference 2009*. http://www.nyas.org/publications/ebriefings/Detail.aspx?cid=40d18ef4-6939-4deb-accc-b5e4516d78a0 (28 March 2013, date last accessed).

Titsias,M. *et al.* (2012) Identifying targets of multiple co-regulating transcription factors from expression time-series by Bayesian model comparison. *BMC Syst. Biol.*, **6**, 53.