COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 119 (2015) 17–28

journal homepage: www.intl.elsevierhealth.com/journals/cmpb

A consensus approach for estimating the predictive accuracy of dynamic models in biology

Alejandro F. Villaverde[a], *Sophia Bongard*[b], *Klaus Mauch*[b], *Dirk Müller*[b], *Eva Balsa-Canto*[a], *Joachim Schmid*[b], *Julio R. Banga*[a,*]

[a] *Bioprocess Engineering Group, IIM-CSIC, Eduardo Cabello 6, 36208 Vigo, Spain*
[b] *Insilico Biotechnology AG, Meitnerstraße 8, 70563 Stuttgart, Germany*

ARTICLE INFO

Article history:
Received 16 July 2014
Received in revised form
19 December 2014
Accepted 2 February 2015

Keywords:
Systems biology
Metabolic engineering
Cell line engineering
Dynamic modelling
Ensemble modelling
Consensus

ABSTRACT

Mathematical models that predict the complex dynamic behaviour of cellular networks are fundamental in systems biology, and provide an important basis for biomedical and biotechnological applications. However, obtaining reliable predictions from large-scale dynamic models is commonly a challenging task due to lack of identifiability. The present work addresses this challenge by presenting a methodology for obtaining high-confidence predictions from dynamic models using time-series data. First, to preserve the complex behaviour of the network while reducing the number of estimated parameters, model parameters are combined in sets of meta-parameters, which are obtained from correlations between biochemical reaction rates and between concentrations of the chemical species. Next, an ensemble of models with different parameterizations is constructed and calibrated. Finally, the ensemble is used for assessing the reliability of model predictions by defining a measure of convergence of model outputs (consensus) that is used as an indicator of confidence. We report results of computational tests carried out on a metabolic model of Chinese Hamster Ovary (CHO) cells, which are used for recombinant protein production. Using noisy simulated data, we find that the aggregated ensemble predictions are on average more accurate than the predictions of individual ensemble models. Furthermore, ensemble predictions with high consensus are statistically more accurate than ensemble predictions with large variance. The procedure provides quantitative estimates of the confidence in model predictions and enables the analysis of sufficiently complex networks as required for practical applications.

1. Introduction

Mathematical modelling is a fundamental task in systems and computational biology [1], with important applications in biomedicine [2–5]. Among other features, models allow monitoring the state of unmeasured variables and making predictions about system behaviour for a larger number and broader variety of conditions than can be efficiently tested in experiments [6]. The construction and calibration of models

* *Corresponding author*. Tel.: +34 986231930.
E-mail addresses: afvillaverde@iim.csic.es (A.F. Villaverde), sophia.bongard@insilico-biotechnology.com (S. Bongard), klaus.mauch@insilico-biotechnology.com (K. Mauch), dirk.mueller@insilico-biotechnology.com (D. Müller), ebalsa@iim.csic.es (E. Balsa-Canto), joachim.schmid@insilico-biotechnology.com (J. Schmid), julio@iim.csic.es (J.R. Banga).

http://dx.doi.org/10.1016/j.cmpb.2015.02.001
0169-2607/

of large, complex dynamic systems is a particularly challenging task. Uncertainties appear at different stages of the process, limiting the confidence in the resulting predictions [7,8]. Shortage of experimental data can easily lead to poor identifiability. As an example, consider a well-known result from nonlinear systems theory which states that, to identify a model described by differential equations containing r parameters, $2r+1$ experimental measures may be enough [9]. This result assumes exact, noise-free measurements; however, in practice there will always be errors in the data, hence the $2r+1$ figure represents a lower bound. When the number of parameters is larger than what can be actually determined from data, the calibration procedure can sometimes—when allowed by the model structure—yield a perfect fit between model predictions and measurements. However, there is a danger of overfitting in this situation, i.e., the model is being trained to fit in detail the noise contained in the data instead of actually learning the system dynamics. This problem entails the risk that model predictions will be wrong for altered experimental conditions.

The problem of dealing with uncertainty in cellular network modelling was reviewed in [10]. In that review, the use of ensembles—sets of models with different structures and/or parameter values—was considered as a powerful and generally applicable approach for reducing prediction errors. However, it was also acknowledged that the concept has not sufficiently matured yet. Indeed, ensemble modelling approaches have been recently applied to a variety of problems, ranging from climate prediction [11] to impact of vaccines [12]. An early example of the use of an ensemble approach in biological models was presented in [13], which was limited to ensembles of topologies of Boolean networks. Tran et al. [14] extended the approach to the dynamic case, building an ensemble of metabolic models that reached the same steady state and applying it to the central carbon metabolism of *Escherichia coli*. A related application was presented in [15]. For a review of metabolic ensemble modelling see [16].

The use of the consensus as an indication of the reliability of the predictions was explored by Bever [17], who computed time-dependent probability distributions of protein concentrations in artificial gene regulatory networks and introduced the concept of consensus sensitivity, finding that consensus among ensemble models was a good indicator of high-confidence predictions. Recently, further steps were taken with the introduction of the concept of "core prediction": a property that must be fulfilled if the model structure is to explain the data, even if the individual parameters are not accurately identified [18].

The present paper deals with the problem of evaluating and, if possible, increasing the confidence in the predictions made by kinetic metabolic models. It is assumed that the model structure—the topology of the metabolic network—is known. Actually, this assumption is not a requirement of the proposed methodology, which may be applied to ensembles of models with different topologies. However, in the present work the uncertainty in the predictions is due only to uncertainty in the parameter values. To overcome uncertainty, an ensemble of models with different parameterizations is built. As a preceding step to improve identifiability and to reduce overfitting, the initial model parameters are grouped into modules of meta-parameters, which are used during calibration. Then a measure of consensus among model outcomes is introduced, which is used to quantify the confidence in the predicted metabolite concentrations. A schematic depiction of the methodology is shown in Fig. 1.

We note that, while a consensus approach was proposed in [17], it used different measures than the ones we introduce here, and it was applied to toy models consisting of 3 or 4 genes. The present methodology includes entirely new features such as the use of meta-parameters, and it is tested on a medium-size network including 34 metabolites. We also remark that, unlike the approach presented in [18], we do not intend to characterize the model's core predictions, but instead to give estimates of the confidence in the predictions. Finally, we note that a preliminary version of this work [19] was presented at the PACBB'14 conference. This new version has been extensively rewritten, including new figures and results, which have been calculated with a new dissensus measure that enables a more sensitive discrimination of larger and smaller prediction errors.

2. Methods

2.1. *Meta-parameter approach*

The methodology aims at adapting the kinetics of interrelated reaction pathways. Highly correlated trajectories of simulated concentrations and reaction rates point at functional dynamic relations, which can be adjusted by the parameters that correspond to the correlated time courses of concentrations and fluxes. We will refer to these sets of parameters as meta-parameters and use them for improving identifiability and reducing the risk of overfitting.

Let us consider an ordinary differential equation (ODE) model with rate kinetics which follow the description in [20], where a rate of an enzyme i is defined by

$$r_i = r_i^{max} \cdot f_i(c_j, p_i^j) = r_i^{max} \cdot \left(1 + \sum_{j=1}^{npars} p_i^j \left(\frac{c_j}{c_j^0}\right)\right) \quad (1)$$

with r_i being a product of the maximal rate (r_i^{max}) and a kinetic rate expression (f_i) which is a function of the metabolite concentrations (c_j) and the parameters (p_i^j) that quantify the contribution of metabolite j in the rate i. The function f_i is following here linlog kinetics [21], but it can be any generic kinetic rate equation where parameters p_i^j are associated to specific metabolite concentrations c_j, such as K_M values in Michaelis–Menten type kinetic equations.

The model is simulated with a set of initial parameter values $p_i^{j\,0} = (p_i^{1\,0}, \ldots, p_i^{npars\,0})$, obtaining time courses of concentrations and rates. The Pearson Correlation Coefficients (PCC) are then calculated between simulated concentration time courses for all balanced species c_j, as well as between all