

Genomic Data Assimilation for a Hybrid Functional Petri Net Model of Circadian Rhythm with Time Course Gene Expression Data

Masao Nagasaki^{*1}

masao@ims.u-tokyo.ac.jp

Seiya Imoto¹

imoto@ims.u-tokyo.ac.jp

Hiroshi Matsuno⁴

matsuno@sci.yamaguchi-u.ac.jp

Rui Yamaguchi^{*1}

ruiy@ims.u-tokyo.ac.jp

Atsushi Doi¹

doi@ims.u-tokyo.ac.jp

Satoru Miyano¹

miyano@ims.u-tokyo.ac.jp

Ryo Yoshida¹

yoshidar@ims.u-tokyo.ac.jp

Yoshinori Tamada^{2,3}

tamada@ism.ac.jp

Tomoyuki Higuchi^{2,3}

higuchi@ism.ac.jp

¹ Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo, 108-8639, Japan

² Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo, 106-8569, Japan

³ Japan Science Technology Agency, Tokyo, Japan

⁴ Faculty of Science, Yamaguchi University, 1677-1 Yoshida, Yamaguchi, 753-8512, Japan

* These authors equally contributed.

Keywords: data assimilation, state space model, particle filter, hybrid functional Petri net, circadian rhythm

1 Introduction

For simulation models of biological pathways, in most of cases, parameters to govern them are tuned by experts empirically at the present time. In this study, to estimate these parameters and to select the best model from several candidate ones automatically with observation data, we take an approach called the data assimilation (DA). The concept of DA is to incorporate information from observed data into a simulation model. We can expect to obtain more plausible results from the simulation model by this approach. From the point of view of the statistical modeling, DA can be realized by solving an inverse problem to estimate unknown state and parameters of a simulation model from observation data. To formulate the problem, we use a statistical time series model called a nonlinear state space model (SSM). Using an SSM, we can employ effective statistical methods to estimate parameters, i.e. a particle filter, which is based on Monte Carlo simulation. In order to examine the applicability of the approach for biological simulation models, the methods was applied to a model of circadian rhythm represented by a hybrid functional Petri net (HFPN) with a synthetic data [2].

2 Method

An SSM is organized according to the following two equations, e.g. [1]:

$$\mathbf{x}_t = \mathbf{f}(\mathbf{x}_{t-1}, \mathbf{w}_t, \boldsymbol{\theta}_{sys}), \quad (1)$$

$$\mathbf{y}_t = \mathbf{H}\mathbf{x}_t + \boldsymbol{\epsilon}_t. \quad (2)$$

Eq. (1) is called the system model where the $\mathbf{x}_t \in \mathcal{R}^p$ is a vector of the hidden state variables which are measured at time t . In the HFPN, the state vector contains the true signal intensities for the

concentration of the p -biological entities, i.e. mRNA transcripts, proteins, protein complex. The vector-valued function \mathbf{f} is a simulation devise for the HFPPN with the unknown parameters $\boldsymbol{\theta}_{sys}$. Note that Eq. (1), in its argument, contains the system noise \mathbf{w}_t which follows a white noise process with the density function $q(\mathbf{w}_t)$, whereas the HFPPN is a deterministic model. Eq. (2) represents the observed system of time course data $\mathbf{y}_t \in \mathcal{R}^d$, $t = 1, \dots, T$, in which the $H \in \mathcal{R}^{d \times p}$ and the $\boldsymbol{\epsilon}_t$ are the observation matrix and the observational noise, which is i.i.d. from a distribution $r(\boldsymbol{\epsilon}_t)$. Of interest is to estimate the unknown parameters $\boldsymbol{\theta}_{sys}$ and true signals for the concentrations of the p -biological entities \mathbf{x}_t from a given dataset $\mathbf{y}_1, \dots, \mathbf{y}_T$. In the most cases, a part of the expression values of the p -entities is unobserved, for example, we only monitor up to the quantities of the mRNAs when using the gene expression profiles, e.g. microarray data, for the parameter estimations. Therefore, it is a typical that $d < p$, and then, the observation matrix H is determined in the following way: $(H)_{ij}$ takes value one if the expression value of the j -th entity is measured in the i -th element of the expression vector \mathbf{y}_t , otherwise, zero. The estimation problem amounts to finding plausible values of the parameters of the HFPPN and the signal intensities of the p -entities in which the number of the estimators, $dim(\boldsymbol{\theta}_{sys}) + T \times dim(\mathbf{x}_t)$ is much greater than the number of observations $T \times dim(\mathbf{y}_t)$. One major difficulty in this problem is that such an overparameterization usually leads to overestimation in estimating the pathway model. Therefore, in this point of view, we introduce the system noise in Eq. (1) which plays a key role in controlling a trade-off between goodness of fit of the HFPPN to the data and a degree of the regularization in the parameter estimation process. $\boldsymbol{\theta}_{obs}$ denotes a parameter vector in the observation model. In this study we set an unknown parameter vector to $\boldsymbol{\theta} = [\boldsymbol{\theta}_{sys}^T, \boldsymbol{\theta}_{obs}^T]^T$. The initial seed of simulation \mathbf{x}_0 is assumed to be distributed according to $p(\mathbf{x}_0)$. The notations used here are as follows: The collection of the simulation and the observations up to time t are denoted by $\mathcal{X}_t \equiv \{\mathbf{x}_1, \dots, \mathbf{x}_t\}$ and $\mathcal{Y}_t \equiv \{\mathbf{y}_1, \dots, \mathbf{y}_t\}$.

Given this setting, the two main purposes of this analysis, i.e. the parameter estimation and the model selection, will be achieved within a framework of the Bayesian statistical inference. Suppose that we have a guess about functional form of the simulation model, denoted by \mathcal{S} , and a prior distribution of the paramter $p(\boldsymbol{\theta}|\mathcal{S})$. The parameters are estimated with the posterior distribution conditional on all T observations \mathcal{Y}_T , i.e. $p(\boldsymbol{\theta}|\mathcal{Y}_T, \mathcal{S})$. For example, we conventionally estimate the model parameters by the Bayes estimators, for example, $\hat{\boldsymbol{\theta}} = argmax_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathcal{Y}_T, \mathcal{S})$ or $E(\boldsymbol{\theta}|\mathcal{Y}_T, \mathcal{S})$. Regarding the model selection, the Bayesian approach provides us a way to knowing structure \mathcal{S} based on the posterior distributions. With a certain prior distribution $p(\mathcal{S})$, the probability of \mathcal{S} conditional on all available data are given by

$$\begin{aligned} p(\mathcal{S}|\mathcal{Y}_T) &\propto p(\mathcal{S}, \mathcal{Y}_T) \\ &= \int p(\mathcal{Y}_T|\boldsymbol{\theta}, \mathcal{S})p(\boldsymbol{\theta}|\mathcal{S})p(\mathcal{S})d\boldsymbol{\theta}. \end{aligned} \quad (3)$$

To compute the Bayes estimators, some computationally intensive methods are required. In the presentation, we will show how these estimators are obtained and the application results with the HFPPN for the circadian rhythm.

References

- [1] G. Kitagawa. Non-Gaussian state-space modeling of nonstationary time series. *Journal of the American Statistical Association*, 82(400):1032–1063, 1987. (with discussions).
- [2] M. Nagasaki, R. Yamaguchi, R. Yoshida, S. Imoto, A. Doi, Y. Tamada, H. Matsuno, S. Miyano, and T. Higuchi. Genomic data assimilation for estimating hybrid functional Petri net from time-course gene expression data. *Genome Inform*, 17(1):46–61, 2006.