LiSDAS: Life Science Data Assimilation Systems

予測発見戦略研究センター　データ同化グループ
助教　吉田　亮

Computational Systems Biology in New Dimensions Molecular biology is driving a need for state-of-art data science due to a rapid progress of experimental technology accompanying massive information in life science. Over the past decades, a wide variety of statistical technologies has been developed in bioinformatics and computational systems biology to uncover a complex world of cellular systems made of several types of *biological circuits*. LiSDAS provides a cutting-edge statistical toolbox targeting *data-driven* model building of *in silico* biological circuits—biochemical simulator—using high-performance computing.

Data-Driven In Silico Modelling of Biological Circuits Biological circuits, such as transcription factor regulatory circuits and signal transduction pathways, can sense many different environmental signals, and respond to these extraneous stimuli by switching activation/inactivation of target proteins and also enhancing/repressing rates of gene decoding. Kinetic modelling of biochemical reactions using stochastic/non-stochastic differential equations has become a major branch of systems biology. Simulation realizes experiments *in silico* to enhance our understanding of dynamic cellular activities. It also provides a way to test plausibility of currently-obtained reaction models and to create further new hypotheses where inconsistencies arise in simulation, domain knowledge and obtained experimental data.

Despite of a growing need for simulation-based approaches, some fundamental issues limit drawing their potential in practical applications. The major problems are associated with model uncertainty. To proceed to simulations, it is an essential first step to find effective values of kinetic rates that are difficult to measure from *in vivo* experiments and also in theoretical kinetic analyses. Besides, a circuit structure modeled upon interactome database or literature is, in many applications, totally unreliable because of environmental dependency and diversity of cells. LiSDAS features basic functions to explore kinetic parameter values and also to retrieve hypothetical biological circuits from huge configuration space of potential model sets such that simulation trajectories fit experimentally-obtained profiles of biological species on diverse scales from molecules to omics.

State Space Models A biological circuit is modeled as a set of differential equations $dx_i(t)/dt = f_i(\mathrm{pa}_i(x), \theta)$ that defines rates of change in concentrations of p biological entities, $x(t) = (x_i(t))_{1\leq i\leq p}$, over continuous times $t \in \mathcal{T}$. The ith variable is regulated by the parent variables $\mathrm{pa}_i(x)$ with the rate equation f_i having a set of kinetic parameters, θ. Conduction of *in vivo* or *in vitro* time course experiments enables us to measure changes in concentrations of target molecules $y_n = (y_{in})_{1\leq i\leq p} \in \mathbb{R}^p_+$ during discrete time points $n \in \mathcal{N} \subset \mathcal{T}$. To proceed with a statistical learning, we here relate the differential equations to the experimental data using the state space model:

$$y_{in} = x_{in} + w_{in} \quad \text{with } x_{in} := x_i(t) \tag{1}$$

$$\frac{dx_i(t)}{dt} = f_i(\mathrm{pa}_i(x), \theta) + v_i(t) \tag{2}$$

where w_{in} and $v_i(t)$ denote respectively measurement and system noises independently and identically distributed. The process of generating $Y \equiv \{y_n | n \in \mathcal{N}\}$ and $X \equiv \{x(t) | t \in \mathcal{T}\}$ follows (1) and (2) with initial concentration $x(0)$ having a certain distribution $x(0) \sim p(x(0))$. Bayesian inversion analysis explores the unknown parameters in the model—initial state $x(0)$ and kinetics θ—through the posterior distributions $p(x(0), \theta | Y, G)$ and $p(G|Y)$ under which a circuit structure G—$\mathrm{pa}_i(x)$, i.e. set of reactants for each variable—is specified or unknown, and *a priori* knowledge on reaction kinetics is expressed in a prior distribution $p(\theta)$.

Inversion Analysis In many applications, any analytical form of the posterior distribution is unavailable, and so an efficient approximation is required to perform for solving Bayesian inversion, i.e. search for posterior means or modes. Our study started from Nagasaki et al. (2006) that provoked the use of a simple SMC (sequential Monte Carlo) technique to identify kinetic parameters in a transcription circuit of mammalian circadian clock. Since then, we have explored a range of efficient Monte Carlo search algorithms for effective

posterior learning as well as stochastic/non-stochastic search methods (Yoshida et al. (2008), Nakamura et al. (2009), Yoshida and West (2010)). A key notion of the SMC-based search lies in a common idea that the posterior distribution is evaluated approximately over a finite set of Monte Carlo samples called *particles*. Whenever following such approaches, however, a quite huge number of particles are needed to draw as raising the number of unknown parameters. The difficulty then arises from a huge space of model parameters $\{\theta, x(0)\}$ and combinatorial explosions of interacting model variables (proteins, mRNAs) involving a search of G. In a joint exploration of $\{\theta, x(0)\}$ and G, we must draw a sufficient size of particles, e.g. billion or more, corresponding to $\{\theta, x(0)\}$ for each configured network G. We are now aiming to realize such large-scale computation on the next generation of supercomputer.

Stochastic Search for Kinetic Parameters and Circuit Structures Recently, we are exploring a hybrid algorithm of SMC, deterministic/non-deterministic annealing and MCMC as well as the use of Diriclet process mixture in prior modelling. We briefly describe a conceptual view of our sampling strategy. Having a number of particles in any step of the SMC algorithm with assigned importance weights (likelihood), it is likely to be more efficient that the next draws explore more on local regions around one of the previously-obtained particles having large weights. On the contrary, some local areas enveloping particles with trivial weights should be ignored more or less. To gain efficiency, the subsequent incremental sampling shall utilize the realized weights that contain knowledge on local structure of target distribution. Another mechanism we aim to realize is an adaptive control of the degree of dispersion in sample move. At the beginning of the sequential sample increment where less information on local structure is available, we shall allocate particle points more uniformly over the entire parameter space to learn characteristics of target distribution. As raising the number of samples in which sufficient information on local structure has been aggregated, particles should be drawn intensively on more local area surrounding an important particle.

One more key factor for efficient inversion analysis lies in design of prior distributions reflecting substantial knowledge on biochemistry. Each unit of biological circuits evolves over different timescales: signal transduction pathways usually change transcription factor activities on sub-second time scales. Binding of transcription factor to its target promoter reaches equilibrium in seconds. Transcription and translation of genes take many minutes in reaching steady state. In practice, these time scales can often be inferred at some extent from experimental data. LiSDAS automatically generates prior distributions, called *steady-state priors*, so that successively-generated particles move around a biologically-significant subspace of kinetic parameters.

Lung Cancer Systems Biology As a target problem, LiSDAS project aims to discover regulatory pathways involving new, effective clinical biomarkers of lung cancers. Diversity and complexity of lung cancer dynamic systems have been obstacles in identifying key regulatory molecules applicable to prognosis and clinical treatment. We acquired time course gene expression profiles of normal lung epithelial cells treated with EGF and/or gefitinib, a specific EGF receptor tyrosine kinase (RTK) inhibitor, as well as a naive simulation model reflecting well-known molecular interactions among relevant genes. However, the currently-obtained model was unable to reproduce the experimental data in simulation. This inconsistency indicates the lack of mechanisms to be incorporated in further model revisions. Our project is going on to develop a highly-versatile, practical *in silico* circuit together with experimental biology.

References

M.Nagasaki, R.Yamaguchi, R.Yoshida, S.Imoto, A.Doi, Y.Tamada, H.Matsuno, S.Miyano, T.Higuchi (2006) Genomic data assimilation for estimating Hybrid Functional Petri Net from time-course gene expression data, *Genome Informatics*, 17(1):46-61.

R.Yoshida, M.Nagasaki, R.Yamaguchi, S.Imoto, S.Miyano, T.Higuchi (2008) Bayesian learning of biological pathways on genomic data assimilation, *Bioinformatics*, 24(22):2592-2601.

K.Nakamura, R.Yoshida, M.Nagasaki, S.Miyano, T.Higuchi (2009) Parameter estimation of in silico biological pathways with particle filtering towards a petascale computing, *Pacific Symposium on Biocomputing*, 227-238.

R.Yoshida, M.West (2010) Bayesian learning in sparse graphical factor models via annealed entropy (in revision).