階層ベイズモデリングと生化学反応シミュレーション

癌細胞の薬剤応答経路予測へ向けて

吉田　亮　モデリング研究系 助教

統計数理研究所 The Institute of Statistical Mathematics

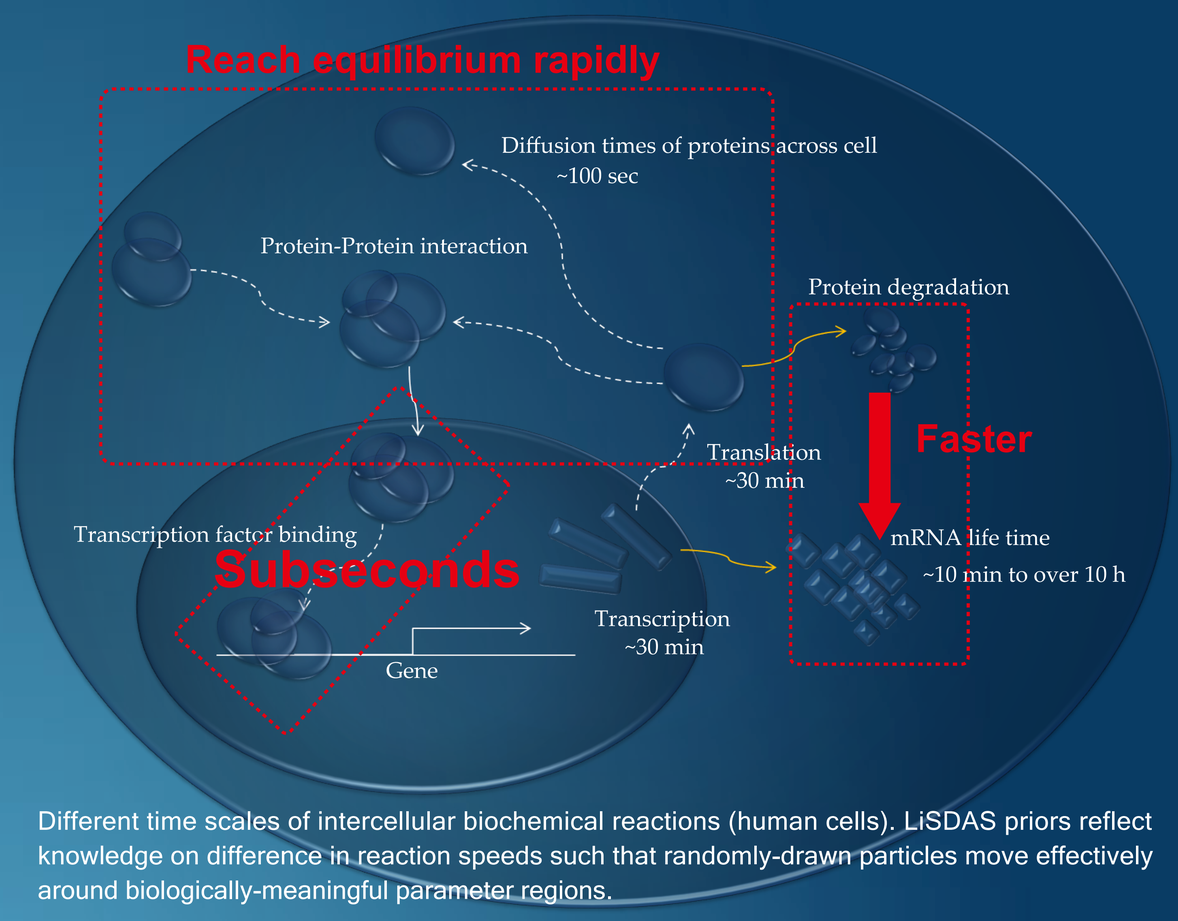
Computational Systems Biology in New Dimensions

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 respond to extraneous stimuli of cells by switching activation/inactivation of target proteins and enhancing/repressing rates of gene decoding. Kinetic modelling of biochemical reactions using differential equations has become a major branch of systems biology. Simulation realizes experiments in silico to enhance our understanding of dynamic cellular activities.

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 essential to find effective values of kinetic rates that are difficult to measure from in vivo experiments and 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.

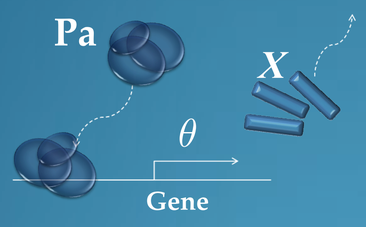


Different time scales of intercellular biochemical reactions (human cells). LiSDAS priors reflect knowledge on difference in reaction speeds such that randomly-drawn particles move effectively around biologically-meaningful parameter regions.

How do we model with LiSDAS? — Example on transcription regulatory circuit

Transcriptional Factor Activities

Controlling efficiency in transcription process is a key regulatory mechanism of cell. Typically, transcription factors can sense external stimuli of cells via signal transduction pathways. Activated transcription factors involved in the signal transduction act as either activators or repressors. A usual function that describes transcription factor activities is the Hill function which is derived from considering the equilibrium binding of the transcription factor to its target site on the promoter.

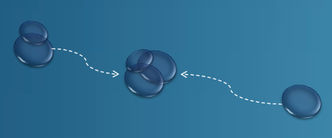


$$\frac{d[X]}{dt} = -\text{degradation} + k \prod_{i \in \text{Pa}} h_i(\text{Pa}(i)) \quad \text{with} \quad h_i(Y) = \begin{cases} \dfrac{Y^\sigma}{Y^\sigma + \omega^\sigma} & \text{(Enhancer)} \\ \dfrac{1}{1+Y^\sigma/\omega^\sigma} & \text{(Repressor)} \end{cases}$$

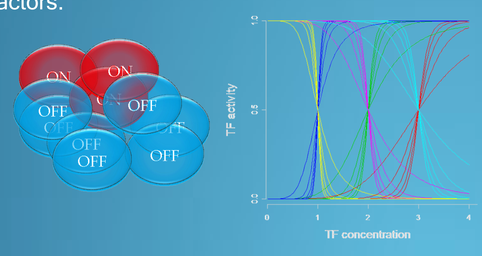
Mass action law as molecular binding process

Instantaneous rate of a reaction is proportional to concentration of each reactant. The kinetic law is known as mass action kinetics. LiSDAS describes a binding process of regulatory molecules, such as protein, with the affinity proportional to the active masses.

$$\frac{d[X]}{dt} = -\text{degradation} + k \prod_{i=1}^{n} P_i^{\sigma_i}$$



Use of sigmoid functions implicitly assumes multiple cells with the rates of complex formation between promoters and transcription factors.



Expert System in Exploring Reaction Kinetics and Circuit Structure

A key factor to 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. Transcription and translation of genes reach steady-state in many minutes. In practice, these time scales can often be inferred from experimental data. LiSDAS automatically generates prior distributions, called Expert System priors, so that successively-generated particles move around a biologically-significant subspace of kinetic parameters.

(Prior modelling using Dirichlet process mixture)

$$P(\psi|Y) \propto P(Y|\psi)P(\psi) \text{ with } \psi := \{x_0, \theta, \text{Pa}\}$$

Base measure is modelled as a mixture of Gibbs distributions with the potential functions defined by domain knowledge on biochemistry and network motives.

$$G_0(\psi) = \sum_{i=1}^{m} \alpha_i \frac{1}{Z_i} \exp\{-T^{-1}\phi_i(\psi)\}$$

Dirichlet process: Generate distributions of parameters according to DP having the base measure.

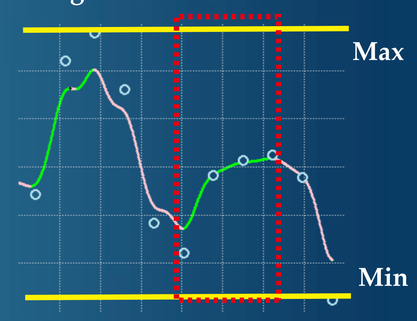
$$P(G) = DP(G|G_0, \gamma)$$

Robust analyses of system: Access the robustness of circuit structure in response to noise inclusion.

$$\psi|G \sim \int K\Big(\frac{\psi - \psi'}{h}\Big) dG(\psi')$$

(A) In DP prior modelling, LiSDAS automatically construct a set of base measures such that given kinetics of differential equations fit to reaction speeds, steady state conditions, evaluated from observed experimental data.

Nonparametric regression and gradient evaluation



Randomly-chosen focus times

Bayesian Inversion Analysis using State Space Models

A biological circuit is modeled as a set of differential equations that defines rates of change in concentrations of p biological entities, $x(t)=(x_1(t), ..., x_p(t))$, over continuous times. Each variable (i th variable) is regulated by the parent variables Pa(i) with the rate equation $f_i(\ ,\theta)$ 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 $y_n = (y_{in})$ during discrete time points. To proceed with a statistical learning, we here relate the differential equations to the experimental data using the state space model:

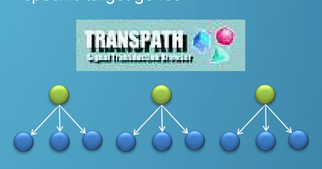
$$\text{(Simulation model)} \qquad \frac{dx_i(t)}{dt} = f_i(\text{Pa}(i), \theta)$$

$$\text{(Measurement model)} \qquad y_{in} = g(x_{in}) + w_{in} \quad \text{with} \quad x_{in} := x_i(n)$$

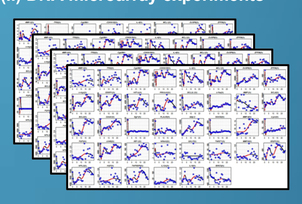
where w_{in} and $v_{in}(t)$ denote respectively measurement and system noises independently and identically distributed. Bayesian inversion analysis explores the unknown parameters in the model---initial state $x(0)$ and kinetics θ ---through the posterior distributions $P(x(0), \theta\,|Y,$ Pa) and P(Pa|Y) under which a circuit structure Pa---set of reactants for each variable---is specified or unknown, and a priori knowledge on reaction kinetics is expressed in a prior distribution.

(B) Computational strategy for data-driven automatic modeling

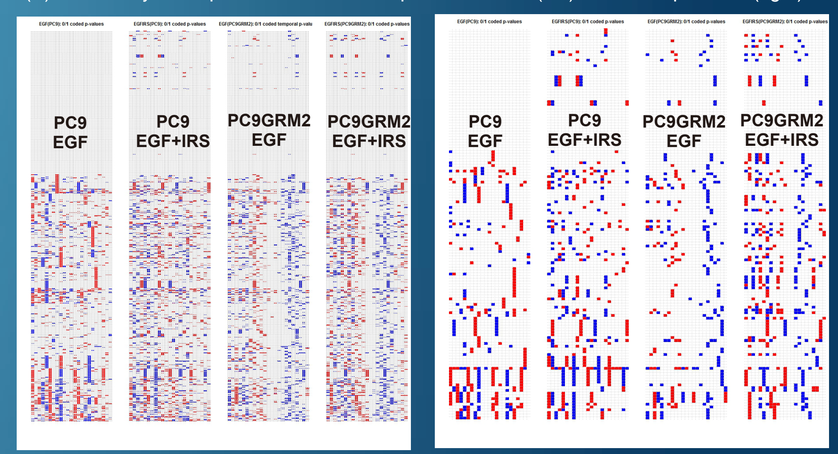
(i) TRANSPATH
Retrieve regulartory motifs of TFs and specific target genes



(ii) DNA microarray experiments



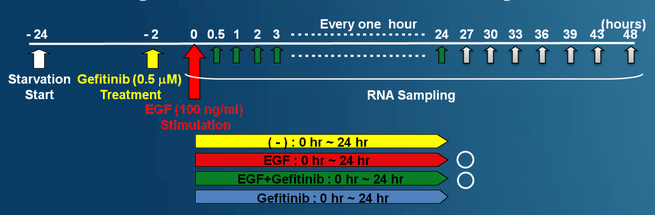
(iii) Inferred Dynamic profiles of 753 transcription activators (left) and 141 repressors (right)



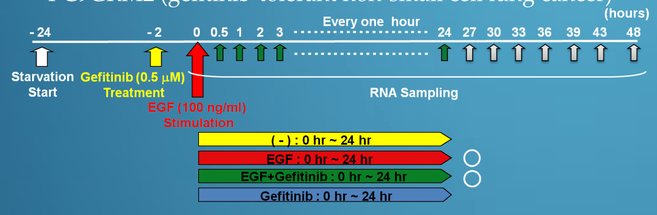
LiSDAS on Lung Cancer Systems Biology

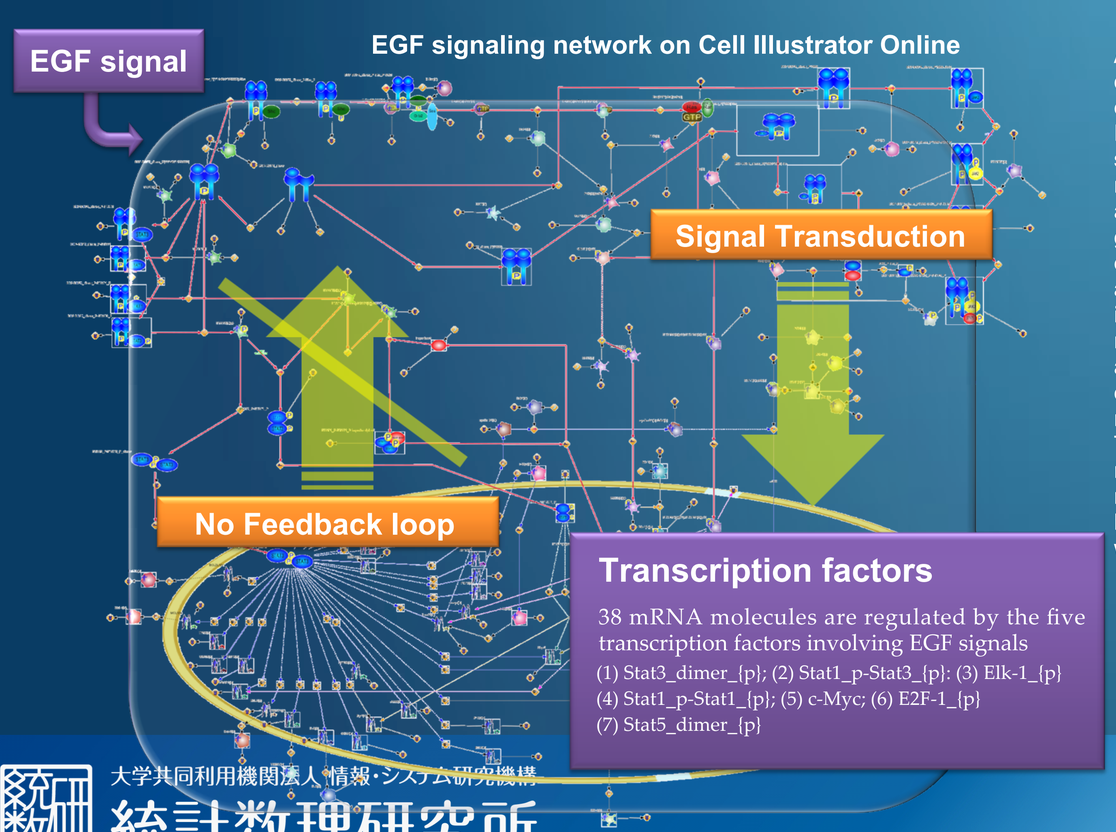
Experimental design for time course gene expression profiling

PC9 (gefitinib-sensitive non-small cell lung cancer)



PC9GRM2 (gefitinib-tolerant non-small cell lung cancer)



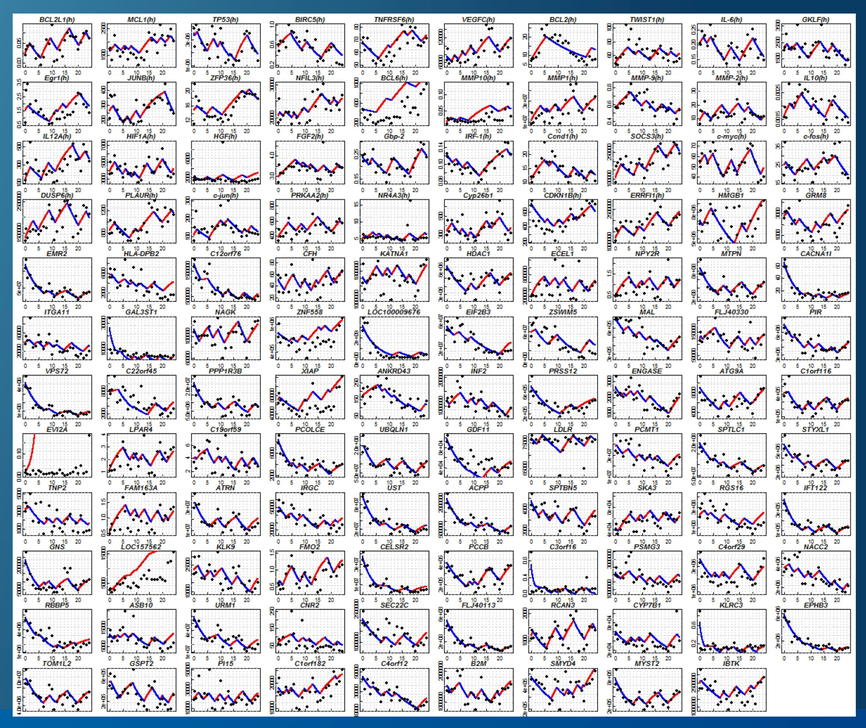


Transcription factors

38 mRNA molecules are regulated by the five transcription factors involving EGF signals

(1) Stat3_dimer_{p}; (2) Stat1_p-Stat3_{p}: (3) Elk-1_{p}
(4) Stat1_p-Stat1_{p}; (5) c-Myc; (6) E2F-1_{p}
(7) Stat5_dimer_{p}

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.



大学共同利用機関法人 情報・システム研究機構
統計数理研究所 The Institute of Statistical Mathematics