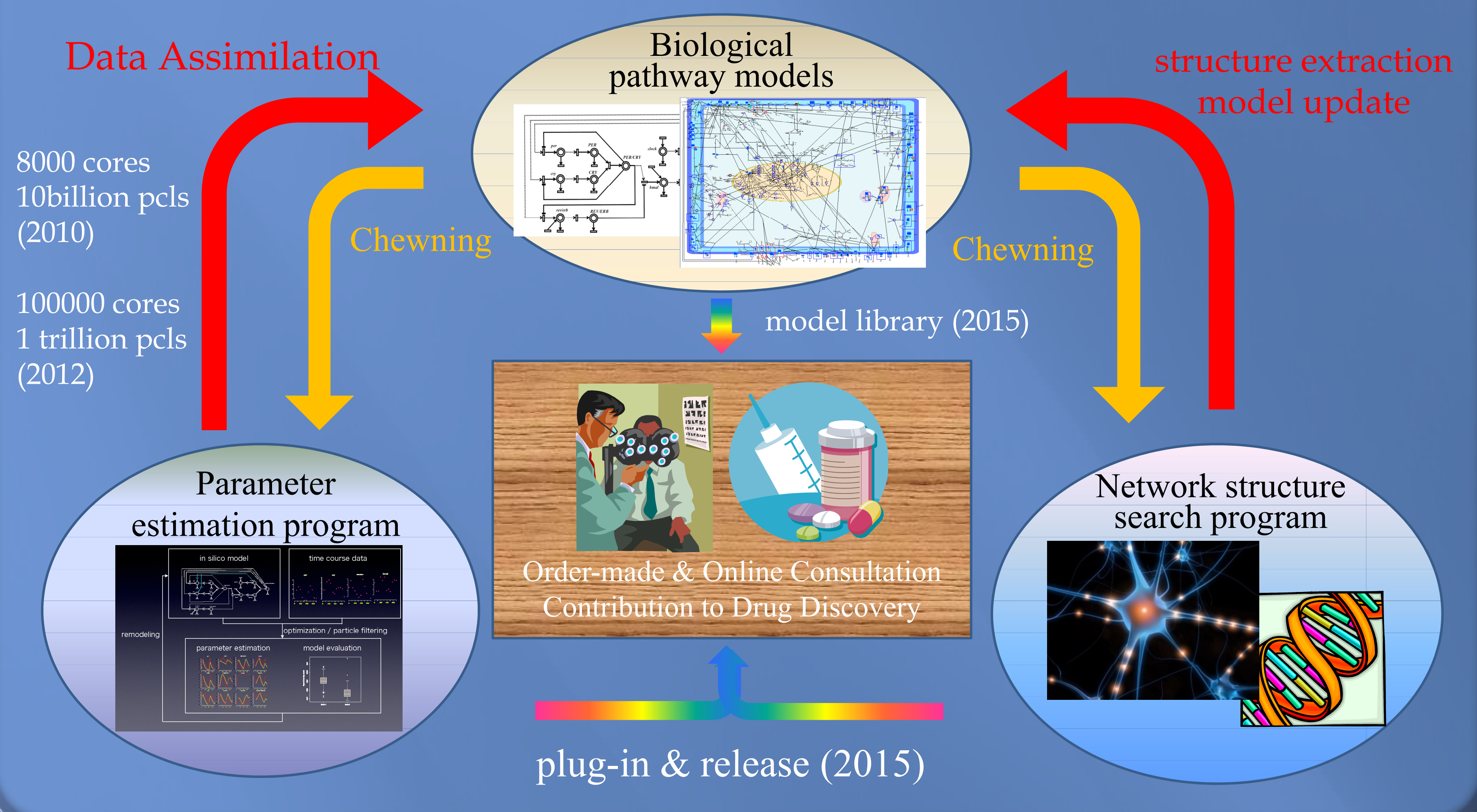
細胞内における遺伝子制御ネットワークの解明を目的とした粒子フィルタとMCMCによるハイブリッド版パラメータ推定法の開発

長尾大道　予測発見戦略研究センター 特任研究員

Abstract

We suggest a hybrid method for parallel computation that consists of particle filter (PF) and MCMC algorithms for the purpose to estimate a number of parameters evaluating their posterior distributions in large-scale biological pathway models. The PF requires, as a target model becomes larger, an exponential increase of the number of particles in order that all distribution functions are well-approximated. A combination with the MCMC enables us to obtain a better approximation for each posterior distribution with much fewer particles owing to a successive mutual evaluation among the particles.

LiSDAS: Life Science Data Assimilation Systems



Our dream is to realize order-made/online consultations and contribute to new drug discoveries in future through rapid clarifications of large-scale biological transcriptional regulatory pathway models in cells with high-performance computing that takes personal differences into consideration. We have developed “LiSDAS”, which enables us to estimate the current/future status in a cell with a framework of the data assimilation based on the particle filter (PF) algorithm (Yoshida *et al.*, in this symposium). We consider the EGF Receptor as a target model, and plan to evaluate posterior distribution functions of model parameters with one trillion particles clarifying the biochemical pathway more precisely. We will implement the LiSDAS on the next-generation supercomputer with plugging-in a network structure search and remodelling algorithms, and release it eventually together with model libraries.

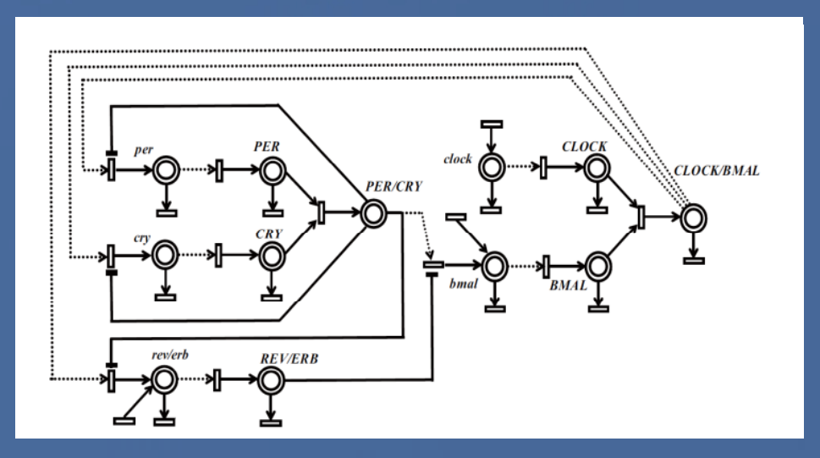
We applied the LiSDAS to the transcriptional regulatory model for circadian clock, which includes 44 unknown parameters to be determined, as a testbed with the number of 100 million particles by using a parallel computing with Intel Xeon 2,880 cores. The computation time resulted in ~1 minute, which indicates that the computation time would be >>3 hours in the case of the EGFR model with the number of one trillion particles even on the next-generation supercomputer. As Nakamura *et al.* (2009) pointed out, even such a large amount of particles are insufficient in order to approximate the posterior distribution functions of kinetic parameters, so that an intriguing improvement is necessary for the parameter estimation.

Problem in Model Parameter Estimation

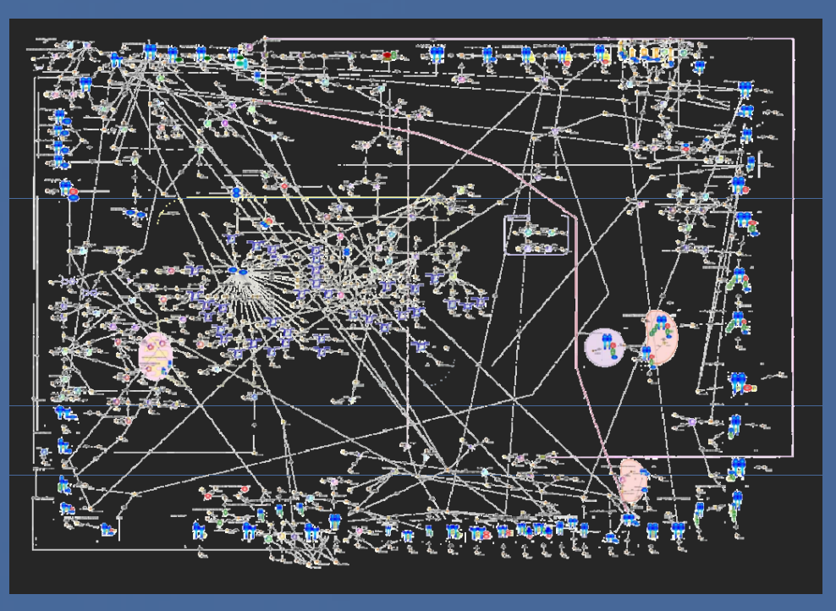
# of unknown parameters	44
# of Particles	**100 million**
Memory	**~ 40 Gbytes**
# of cores	**2,880 cores**
Computation time	**~1 minute**

⬇ analogize to the case of EGFR …

# of unknown parameters	~1000
# of Particles	**1 trillion**
Memory	**400 Tbytes**
# of cores	**160,000 cores**
Computation time	**>>3 hours**

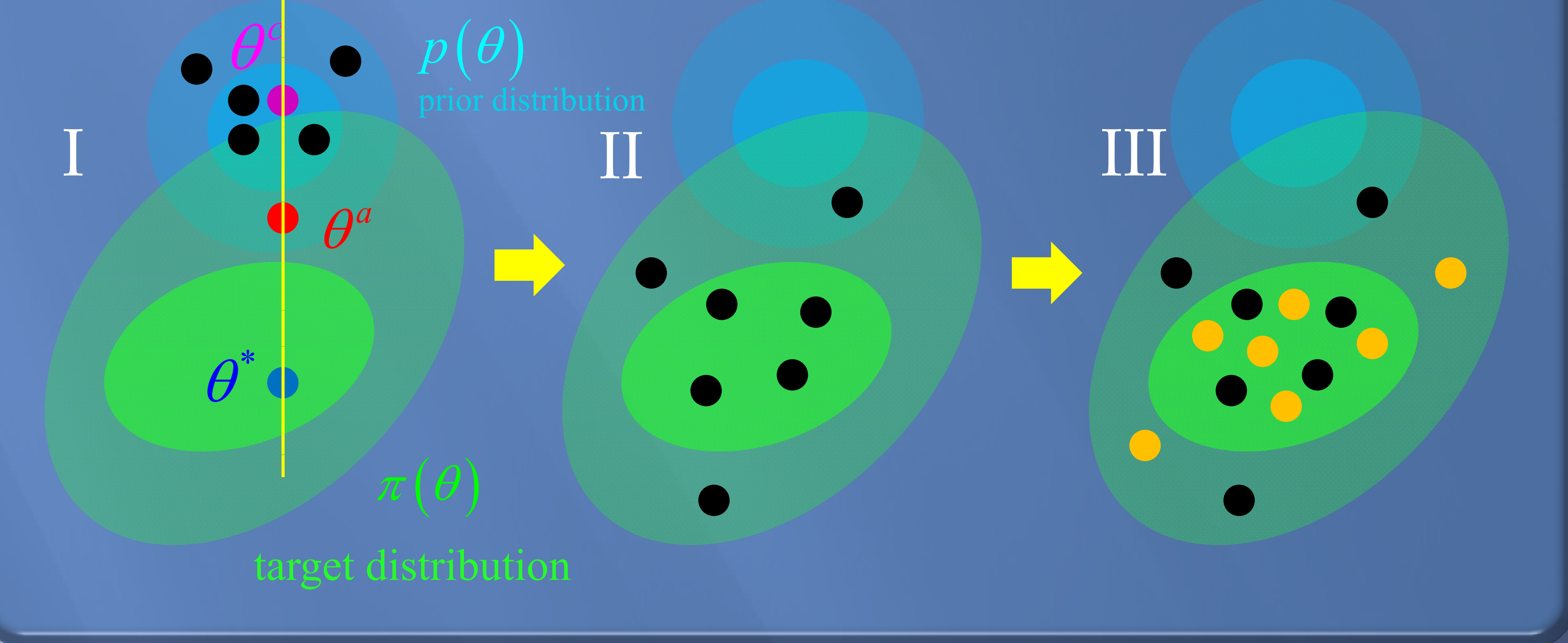


Circadian Clock



EGF Receptor

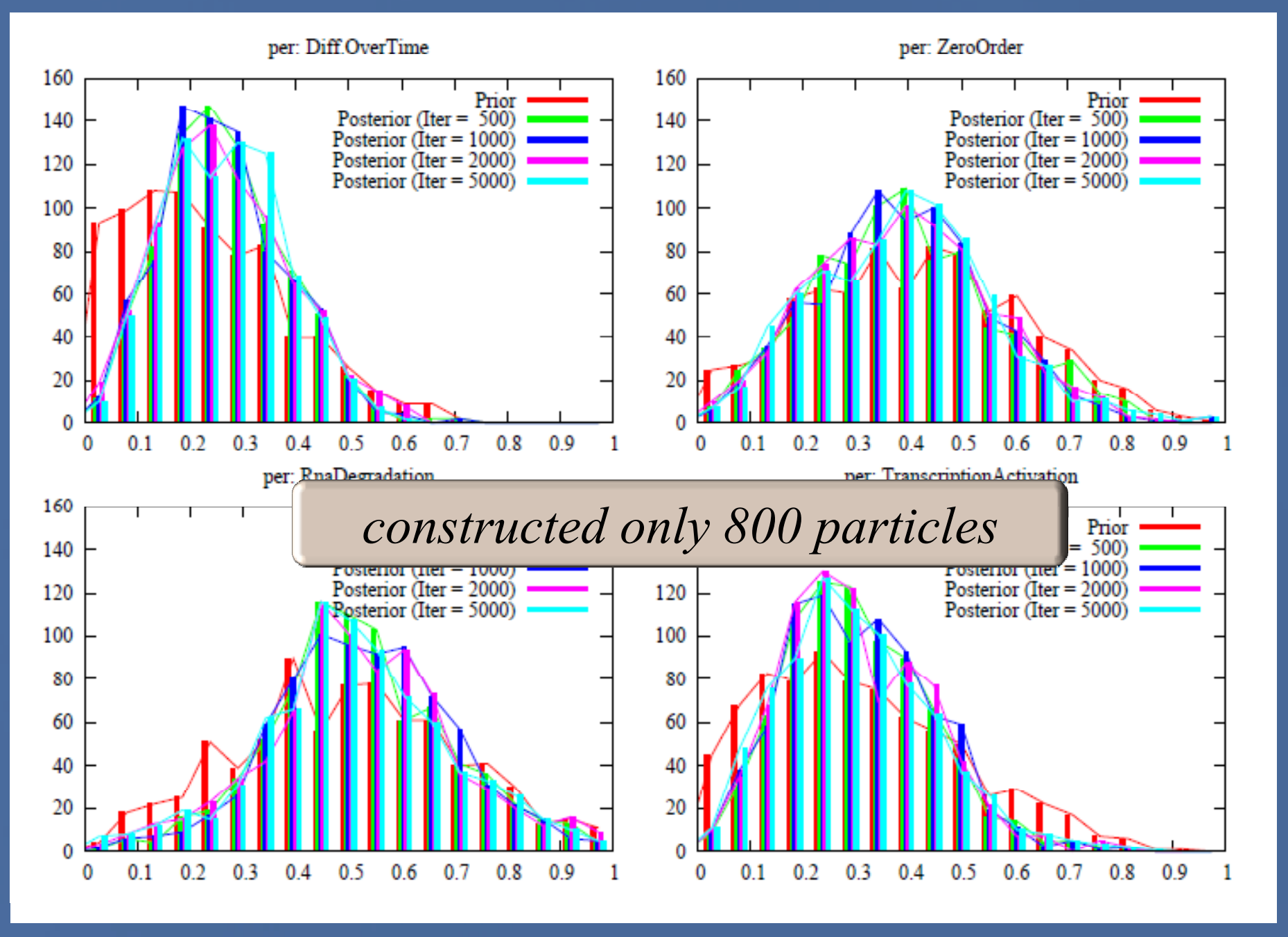
Adaptive Direction Sampling Method



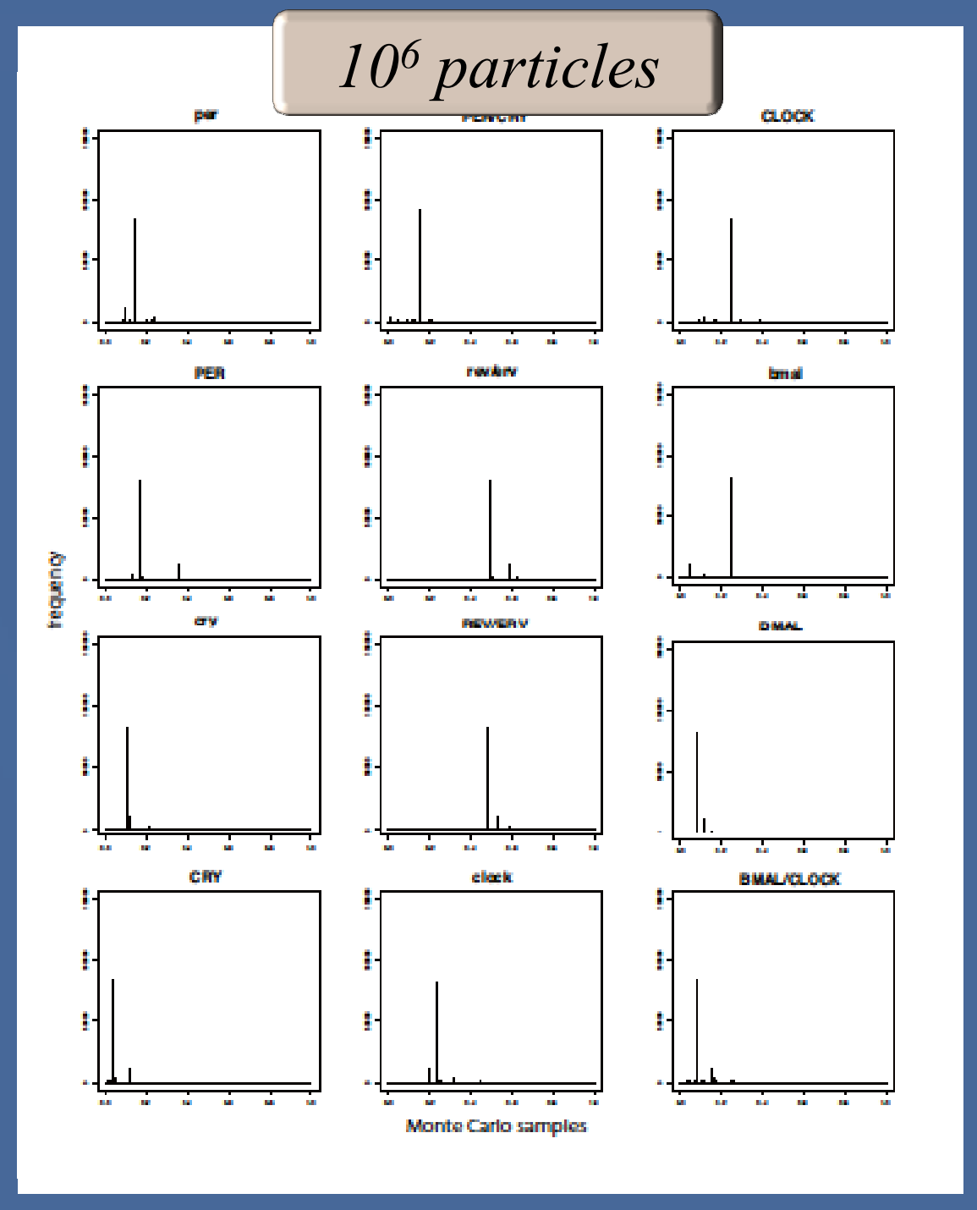
We propose a hybrid method consisting of the PF and the MCMC algorithms for the Bayesian learning of kinetic parameters in a large-scale biological pathway model. We adopt an adaptive direction sampling method for the MCMC algorithm; (I) select at random anchor point θ^a and current point θ^c from sampled particles from an appropriate prior distribution $p(\theta)$, and accept/reject a candidate point θ^* made on a line connecting θ^a and θ^c comparing with θ^c by the Metropolis-Hastings method, (II) let the particles converge “burn-in” in the target distribution $\pi(\theta)$, i.e., posterior distribution $p(\theta|Y)$ in this case, by iterating (I), and (III) proliferate particles as many as needed.

We applied our hybrid method to the transcriptional regulatory network model for circadian clock with a parallel computation using 160 cores. Five particles were given to each core, i.e., 800 particles in total, and 5,000 iterations were made until a burn-in, i.e., 4 million simulations are carried out. We successfully evaluated posterior distribution functions for all kinetic parameters with much fewer particles than before. The computation time is about twice as long as the case of a random search with the same number of particles. We have to examine the obtained results hereafter, e.g., comparison with the real observation data Y and remodelling to obtain a better biological pathway model.

Performance of the Hybrid Method



constructed only 800 particles



10^6 particles

cf. Nakamura *et al.* (2009)