Parameter Estimation of Biological Pathways Using Data Assimilation and Model Checking

Chen Li* **Keisuke Kuroyanagi**

Masao Nagasaki **Satoru Miyano**

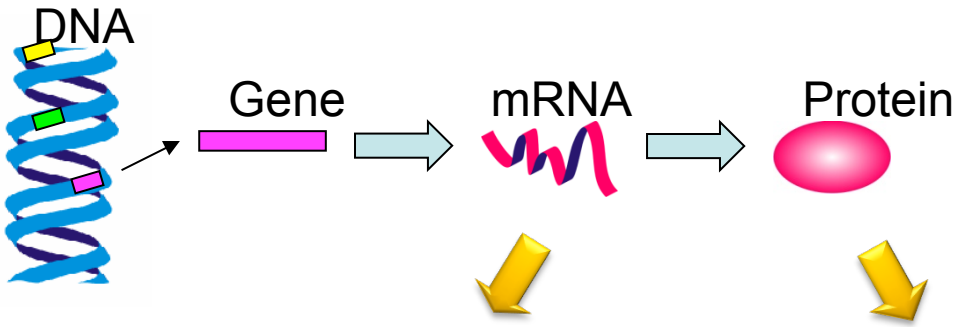
Human Genome Center
Institute of Medical Science, University of Tokyo

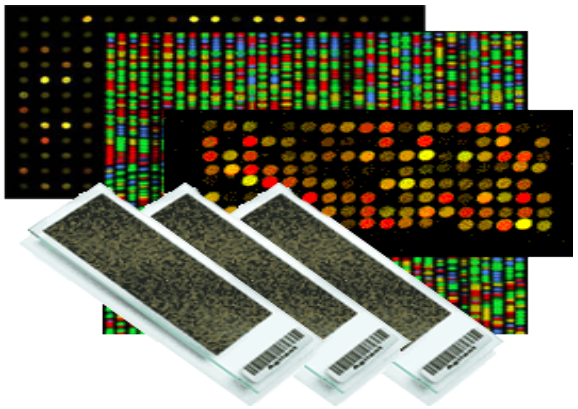
BioPPN2011, Petri Nets 2011 & ACSD 2011, 20 JUN, Newcastle Upon Tyne, UK

Outline

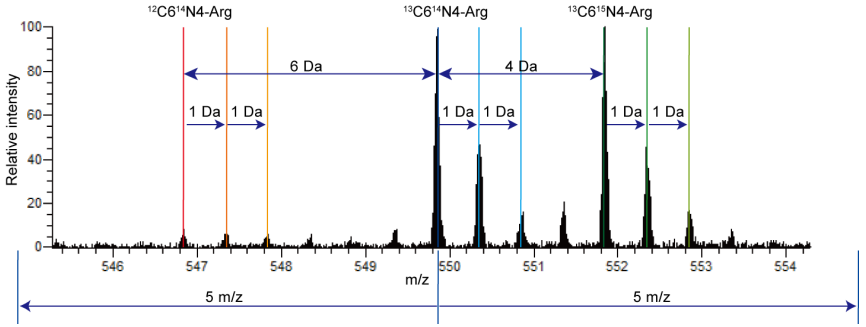
- Purpose:
  - Develop a new systematic framework to **understand** and **predict** the dynamic features of complex cellular mechanisms
  - Information resources: **Experimental Data** + **Pathway Simulation**
- Method:
  - **D**ata **A**ssimilation (**DA**) + **M**odel **C**hecking (**MC**)
    + **H**ybrid **F**unctional **P**etri **N**et with **E**xtension (**HFPNe**)
- Case study:
  - Circadian rhythm model in *mouse*
- Conclusion remarks

Information resource 1: Experimental Data





e.g. Microarray, RT- qPCR
Exon array, RNA-seq

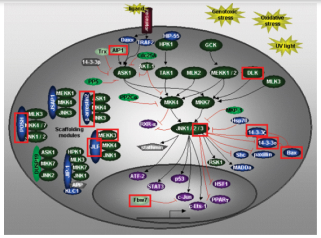


e.g. Mass Spectrometry
Western blotting

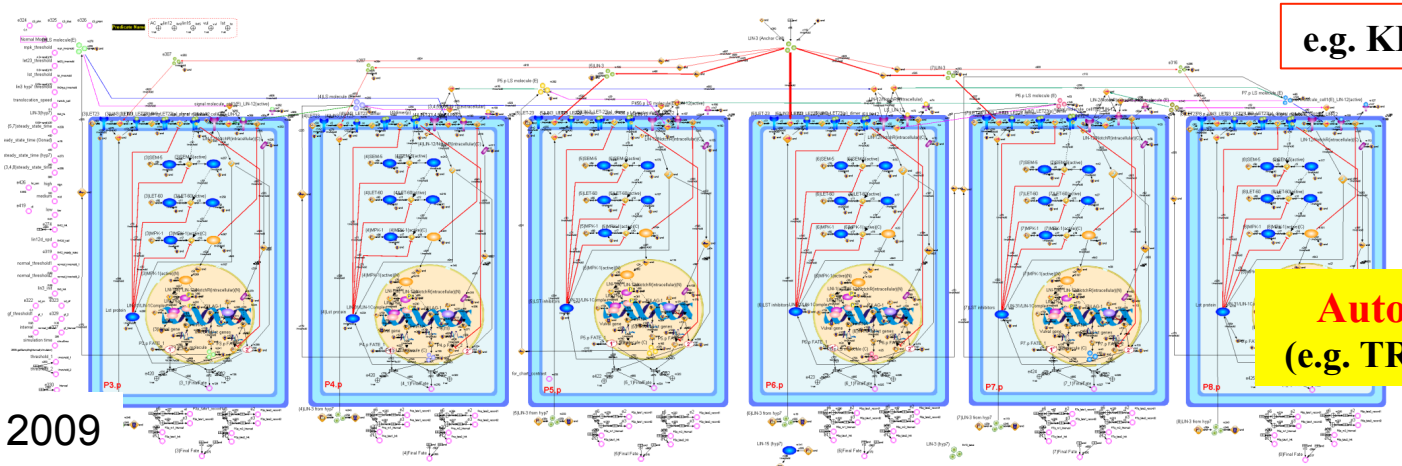
Well-formulated data

Information resource 2: Sir

Public and commercial database



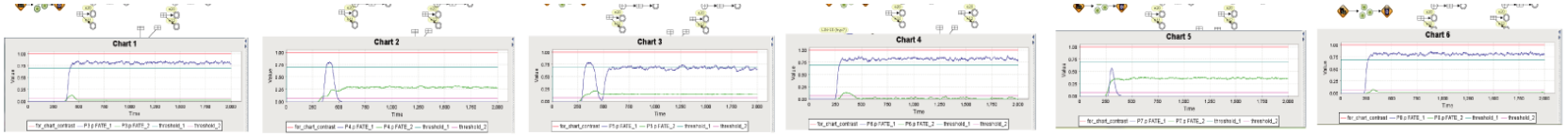
Hybrid Functional Petri Net with extension (HFPNe)



Li *et al* 2009

e.g. KEGG, TRANSPATH

Automatic convertor (e.g. TRANSPATH2CSML)



- Various simulation models using increasing biological knowledge
  - ✓ formal descriptions
  - ✓ constructing from scratch
  - ✓ converted from public and commercial database (Nagasaki *et al* 2007)

Current status of simulation models

- All parameters must be assigned in advance.

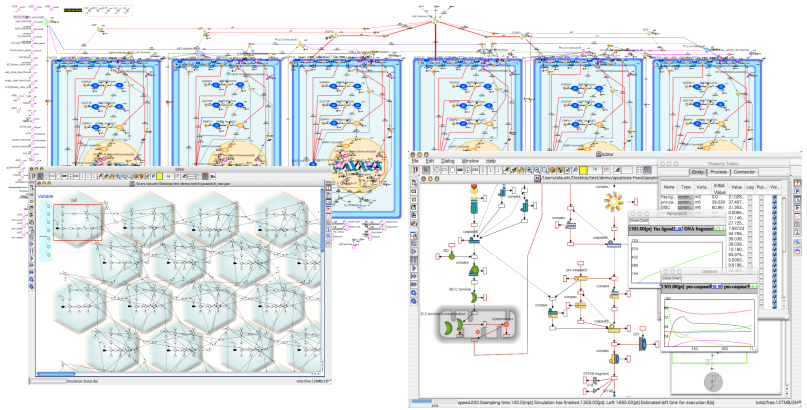
Problems

Conventional hand tuning method severely limits the **size** and **complexity** of simulation models built.

- Parameter estimation requires **a lot of t… …nd errors**
- Small differences of parameters make large gap between simulation results and reality

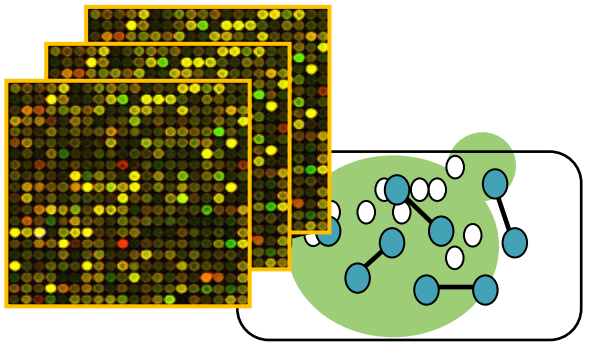
To address this problem...

Simulation model



Incomplete for the real system
Unknown parameters

Experimental data



- Time series data
- **Biological queries**

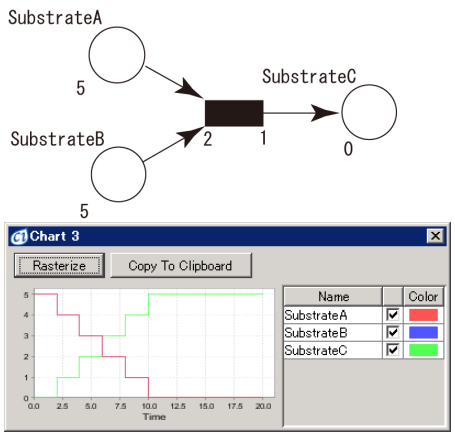
Data Assimilation + Model Checking

- **High-speed and high-accuracy** parameter estimation method
- Evaluation by analyzing *Mouse* circadian clock model

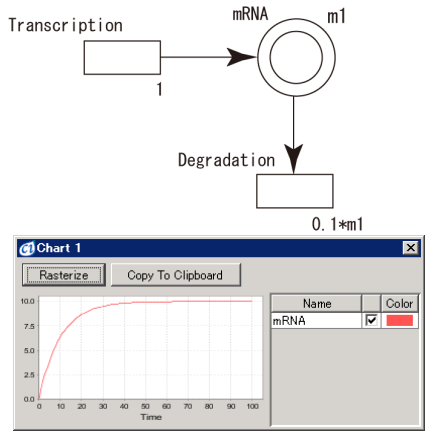
Hybrid Functional Petri Net with extension (HFPNe)

Entities	Processes	Connectors
integer Discrete entity	delay Discrete process	threshold Process connector
real number Continuous entity	rate Continuous process	threshold Association connector
any type Generic entity	any operation Generic process	threshold Inhibitory connector

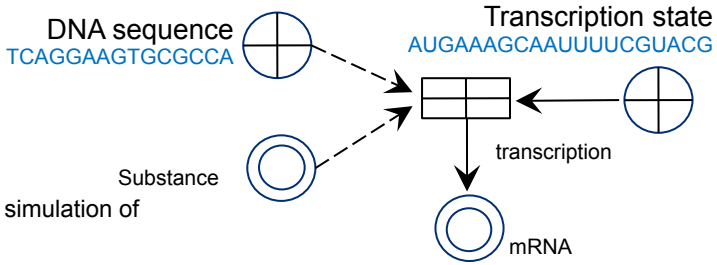
Discrete



Continuous

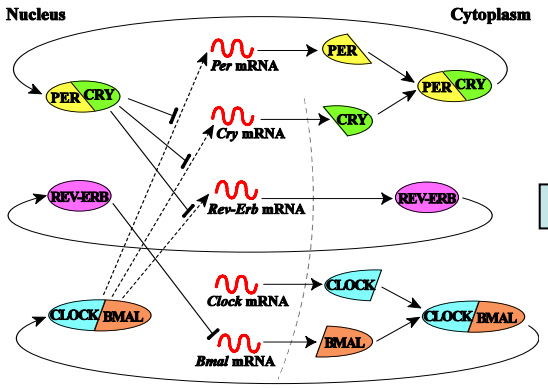


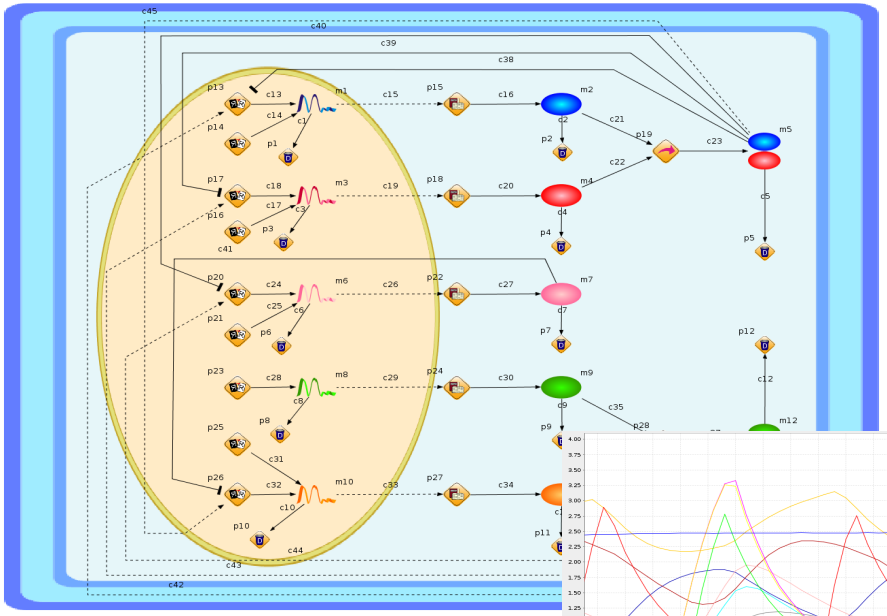
Generic

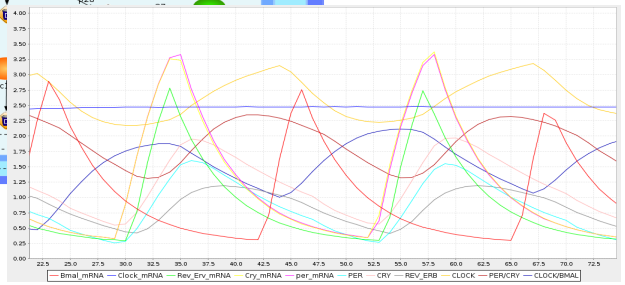


- Nagasaki, M., Doi, A., Matsuno, H., and Miyano, S., A versatile Petri net based architecture for modeling and simulation of complex biological processees, Genome Informatics, 15(1):180–197, 2004.
- https://cionline.hgc.jp

Circadian rhythm in *Mouse* by Cell Illustrator







https://cionline.hgc.jp

HFPNe model on Cell Illustrator Online 5.0

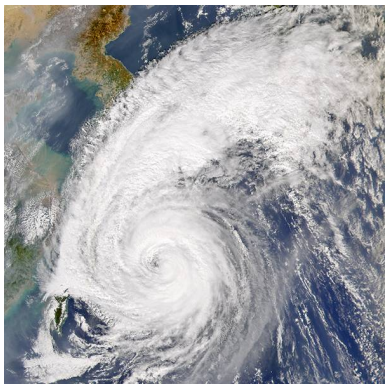
	Cell Illustrator			
	Original elements of HFPNe			Examples of biological images
Type	Discrete	Continuous	Generic	Discrete, Continuous, and Generic
Entity	○	◎	⊕	
Process	▮	▯	▯	
Connector	→ Process	- - - → Association	—⊣ Inhibitation	

Unknown Parameters (17)

$m_1(0),\ldots,m_{12}(0)$	:Initial values
k_1, k_2	:Speeds
s_1, s_2, s_3	:Thresholds

Data Assimilation

Firstly used around 2000 in Geophysics.



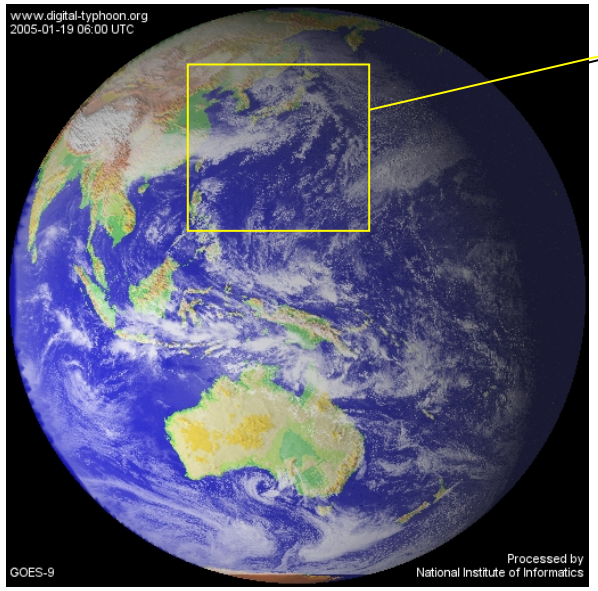
Concept:

Every simulation model is **in**complete!

If the model is incomplete, by using **observed data**, the incomplete part of the simulation model could be complemented (at each observation step) to realize real model with the mathematical background especially statistics.

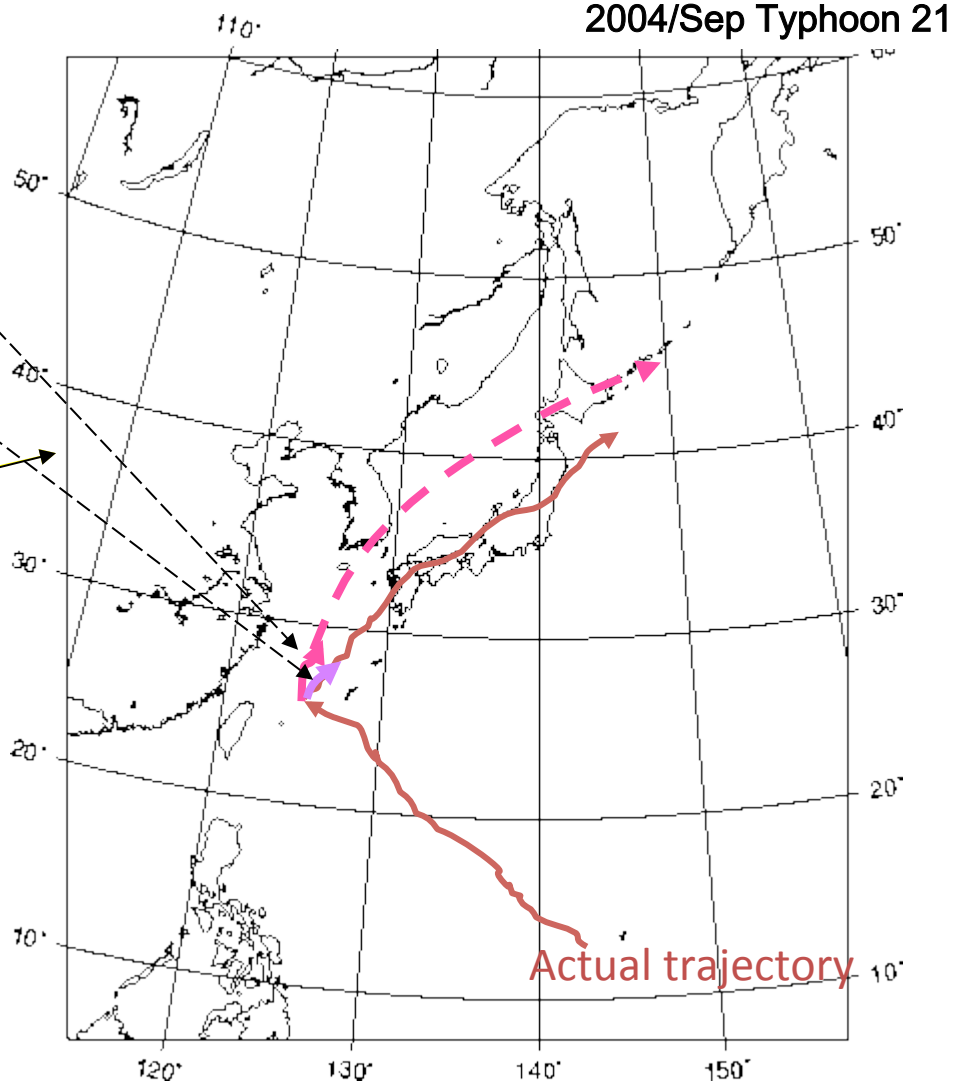
→ : Prediction only by simulation

→: Prediction using data assimilation



(Taken from NII, Japan)

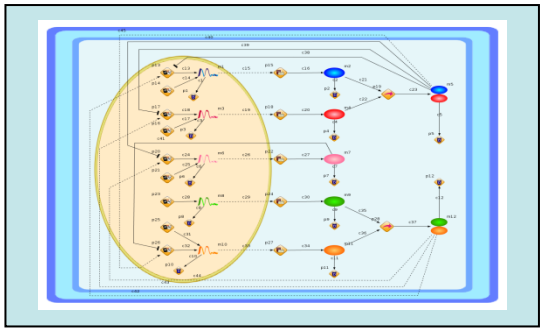
2004/Sep Typhoon 21



Technique
of
Data Assimilation

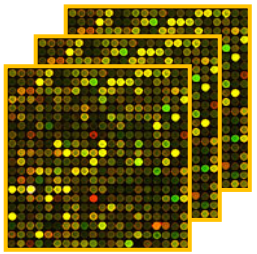
Data Assimilation

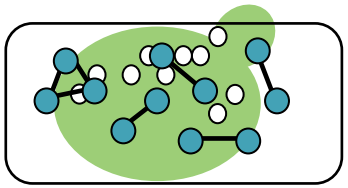
Statistical framework to integrate simulation model and data



Simulation model

+





P-P interaction

Experimental data

Formulated by the **Non-linear State Space Model**

$$m_t = f(m_{t-1}, w_t, \theta_{sys})$$ System model

$$y_t = Hm_t + \varepsilon_t$$ Observation model

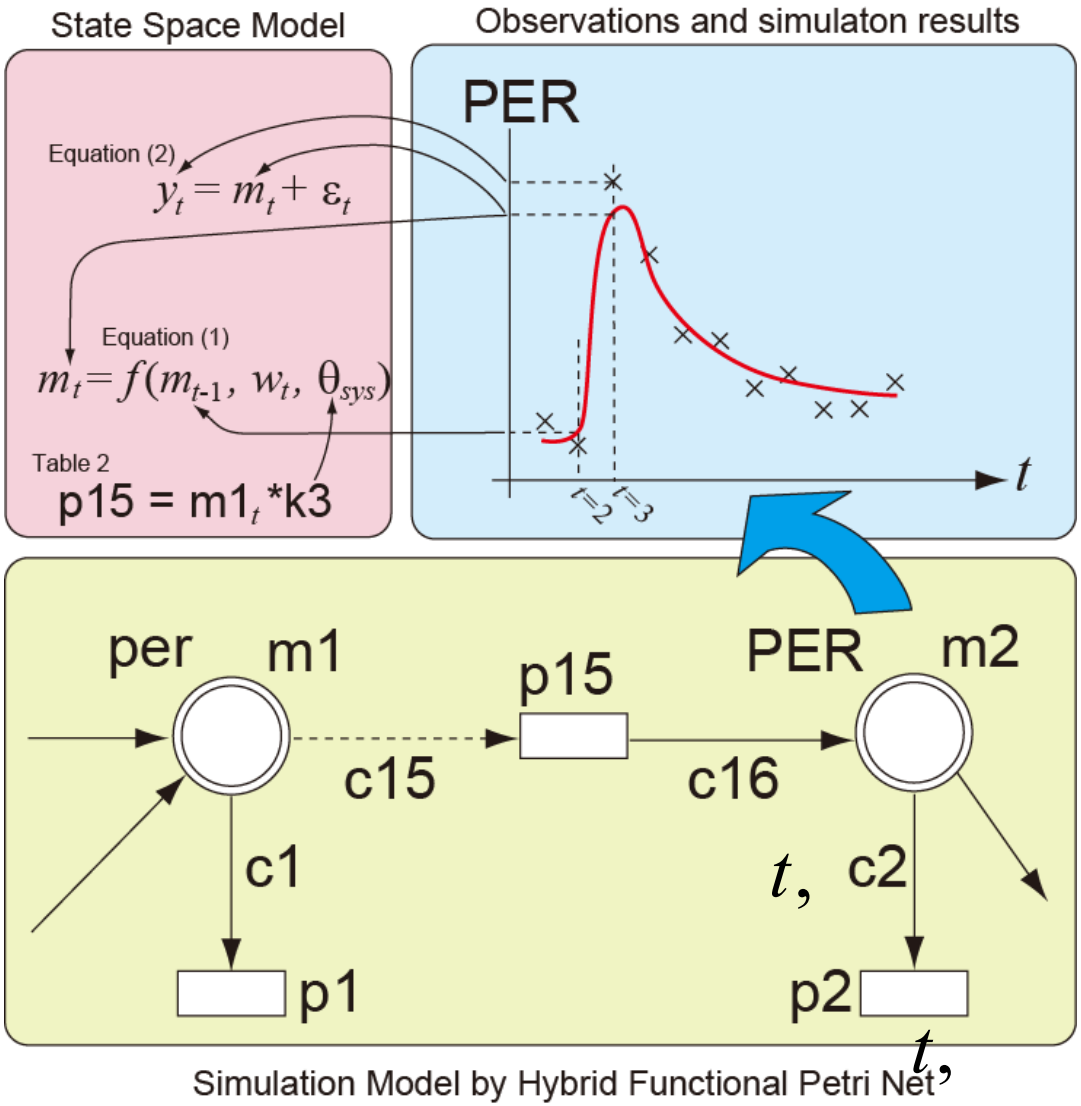
m_t : state vector at time t, f : simulation device, $t = 1, \cdots, T$

w_t : system noise, θ_{sys} : parameter vector,

y_t : observation vector at time t, H : observation matrix,

$\varepsilon_t \sim N(0, \sigma^2)$: observation noise

Data Assimilation



Simulation Model by Hybrid Functional Petri Net

DA to obtain

$$P(X_T, \theta_{sys} \mid Y_T)$$

$$X_T = \{m_1, \ldots, m_T\}$$

$$Y_T = \{y_1, \ldots, y_T\}$$

m_t : state vector at time

θ_{sys} : parameter vector

w_t : system noise

f : simulation device

y_t : observation vector at time

$\varepsilon_t \sim N(0, \sigma^2)$: observation noise

$t = 1, \cdots, T$

Data Assimilation

Parameter Estimation

MAP(*maximum a posteriori*)

$$(\hat{\theta}_{sys}, \hat{x}) = \arg\max_{\theta,x} p(X_T, \theta_{sys} \mid Y_T)$$

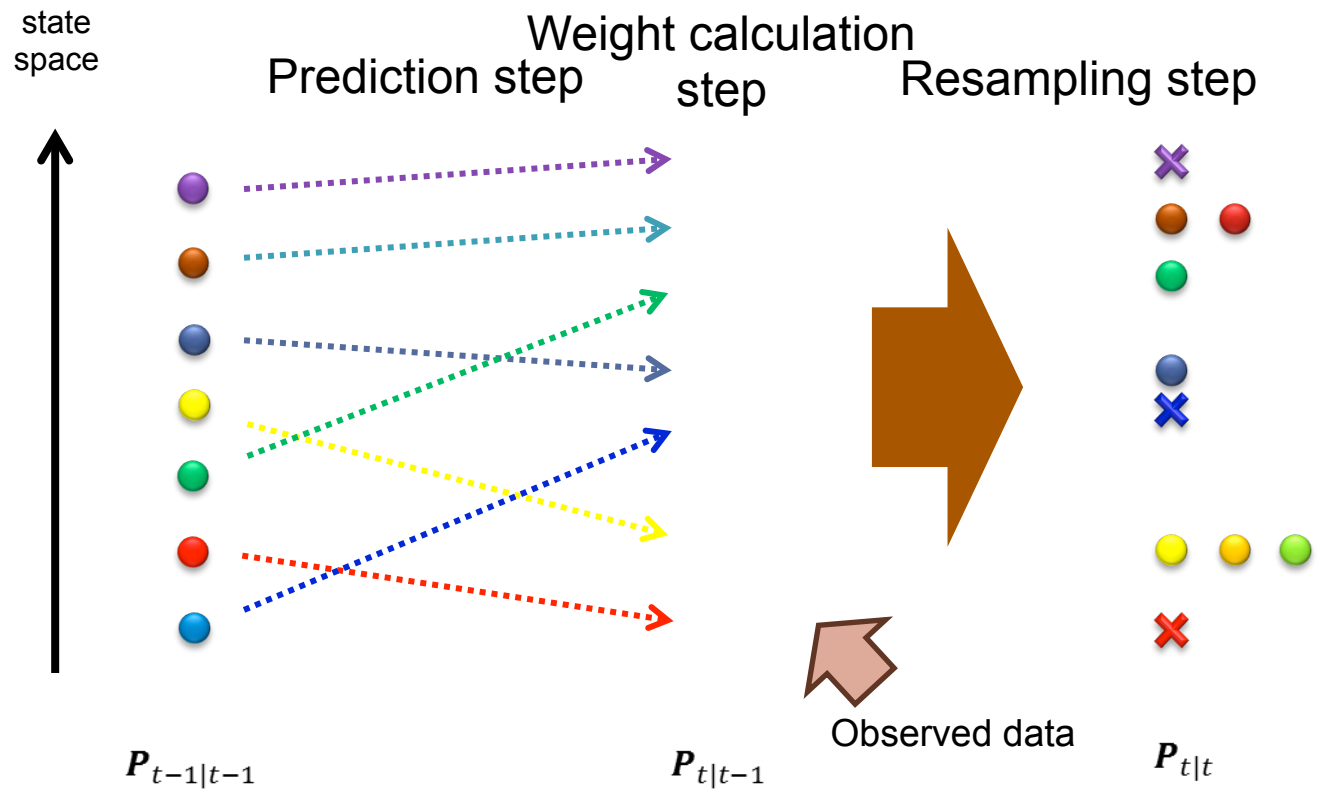
To approximate this posterior distribution. Recursive estimation algorithm **Particle Filter (PF)** is used

Other solution:
MCMC, Optimization

Monte Carlo Filter (Kitagawa, 1996)
Self-organising SSM (Kitagawa 1998)

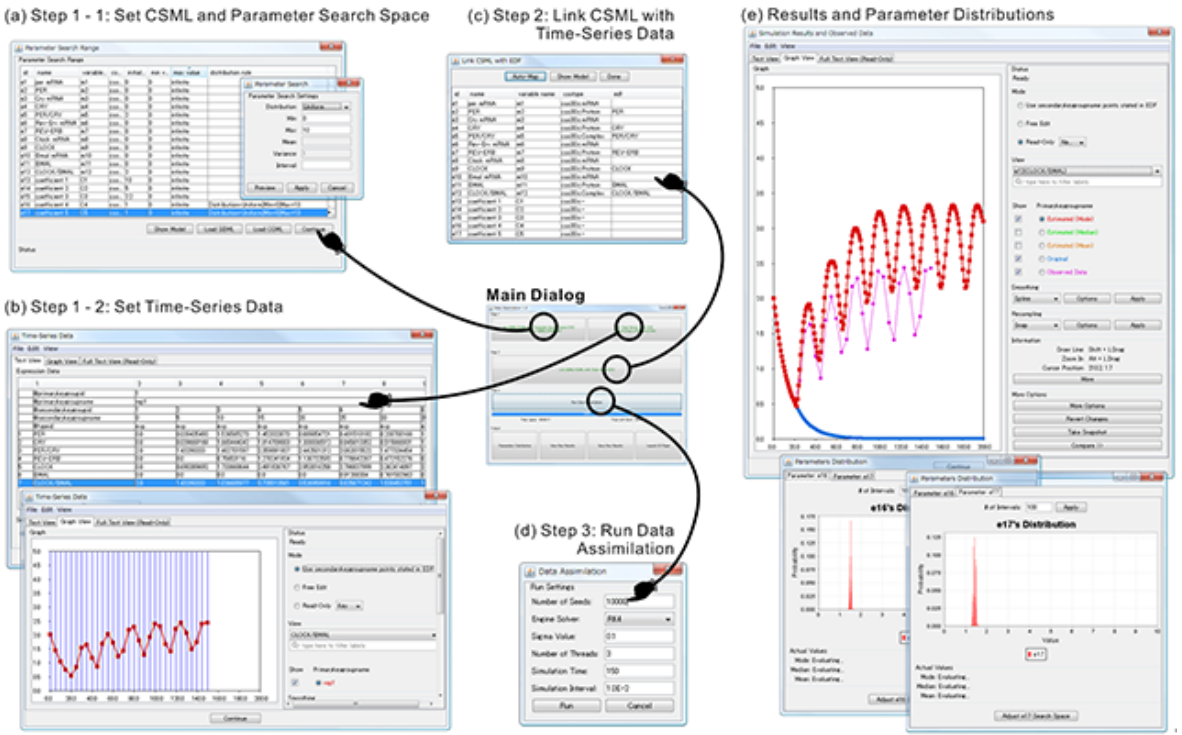
Particle Filter

- A simple and well-established statistical method
- Approximates the **joint posterior distributions** of parameters by using sequentially-generated Monte Carlo samples.



Data Assimilation Tool: DA 1.0

- Java-based software for biological pathways using DA approach*
- User-friendly interface to allow users to carry out unknown parameter estimation

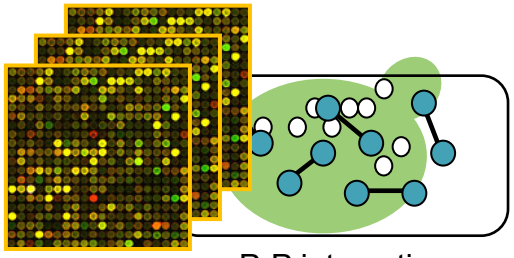


http://da.csml.org

*Koh CH, Nagasaki M, Saito A, Wong L, Miyano S, DA 1.0: Parameter Estimation of Biological Pathways using Data Assimilation approach, *Bioinformatics*. 26(14):1794-6.

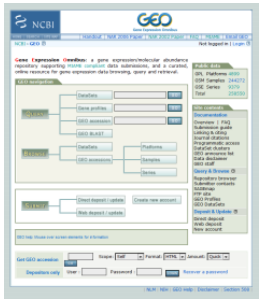
Issues with DA

Experimental data

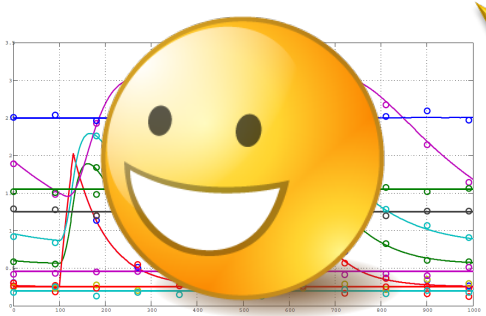


expression

P-P interaction



NCBI



Other … that cannot … in the form o… data

Well-defined time series data

- **Normal**: 10-20 time points
- **Small data**: Cost massive computational resources

DA + **Model Checking**

Prune the search space of Particle Filter

What is model checking?

- A method to verify whether models satisfy the requirement or not.
- **Pioneering works** using model checking for parameter estimation:
  - **MC + GA** (Donaldson and Gilbert 2008)
  - **MC + uniform distribution** (with HFPNe) (Li *et al* 2011)
- Temporal logic formulae often used for the biological queries
  - PLTL (*Probabilistic Linear-time Temporal Logic*)
  - Online Model checker: **MIRACH 1.0** for quantitative pathway models (Hock Koh *et al* 2011, Li *et al* 2011)

Table 1. Syntax of PLTL

ψ	::=	$\mathbf{P}_{\trianglelefteq x}(LTL) \mid \mathbf{P}_{=?}(LTL) \mid LTL$
LTL	::=	$\phi \mid \{AP\} \mid \phi$
ϕ	::=	$\mathbf{X}\,\phi \mid \mathbf{G}\,\phi \mid \mathbf{F}\,\phi \mid \phi\,\mathbf{U}\,\phi \mid \phi\,\mathbf{R}\,\phi \mid \neg\,\phi \mid \phi \,\&\&\, \phi \mid \phi \mid\mid \phi \mid \phi \Rightarrow \phi \mid AP$
AP	::=	$value\ comp\ value \mid value_{boolean}$
$value$	::=	$value\ op\ value \mid [variableName] \mid Function_{numeric} \mid Integer \mid Real$
$value_{boolean}$	::=	$true \mid false \mid Function_{boolean}$
$comp$	::=	$== \mid != \mid \geq \mid > \mid < \mid \leq$
op	::=	$+ \mid - \mid * \mid / \mid \hat{}$,

with $\trianglelefteq \in \{<, <=, >, >=\}$, $x \in [0, 1]$.

Syntax

Table 2. Semantics of temporal operators

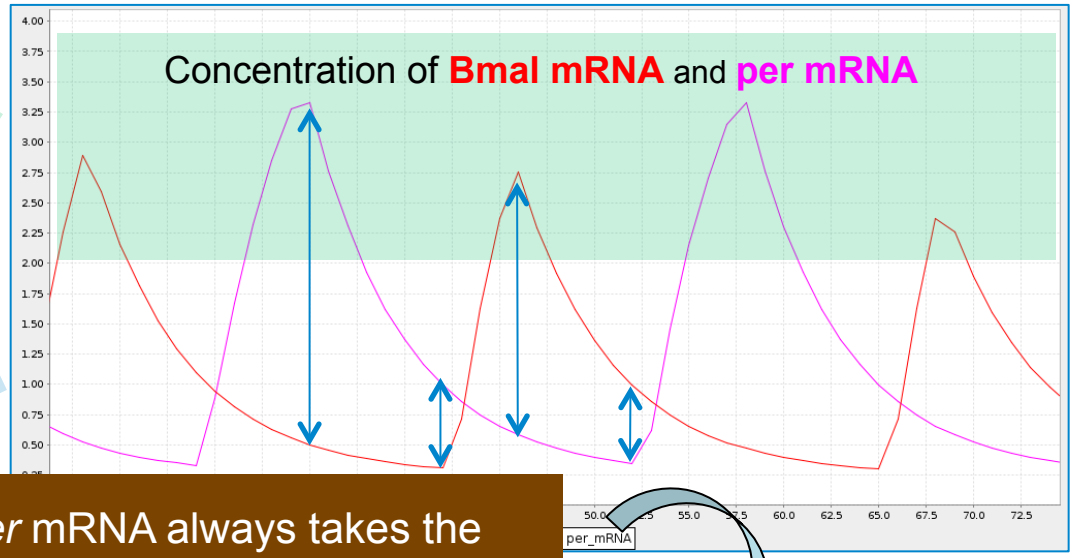
Operator	Meaning	Explanation
$\mathbf{X}\,\phi$	next time	ϕ must be true at the next time point.
$\mathbf{G}\,\phi$	globally	ϕ must always be true.
$\mathbf{F}\,\phi$	Finally	ϕ must be true at least once.
$\phi_1\,\mathbf{U}\,\phi_2$	Until	ϕ_1 must be true until ϕ_2 becomes true; ϕ_2 must become true eventually.
$\phi_1\,\mathbf{R}\,\phi_2$	Release	ϕ_2 must be true until and including the time point ϕ_1 becomes true; if ϕ_2 never true, ϕ_1 must always be true.

Semantics

Example: Biological queries of Circadian Rhythm

Q1: If the concentration of *per* mRNA always takes the maximum, its concentration is higher than 2.0

$$G(d([per\ mRNA]) \geqq 0\ \&\&\ X(d([per\ mRNA]) < 0) \rightarrow [per\ mRNA] > 2.0)$$

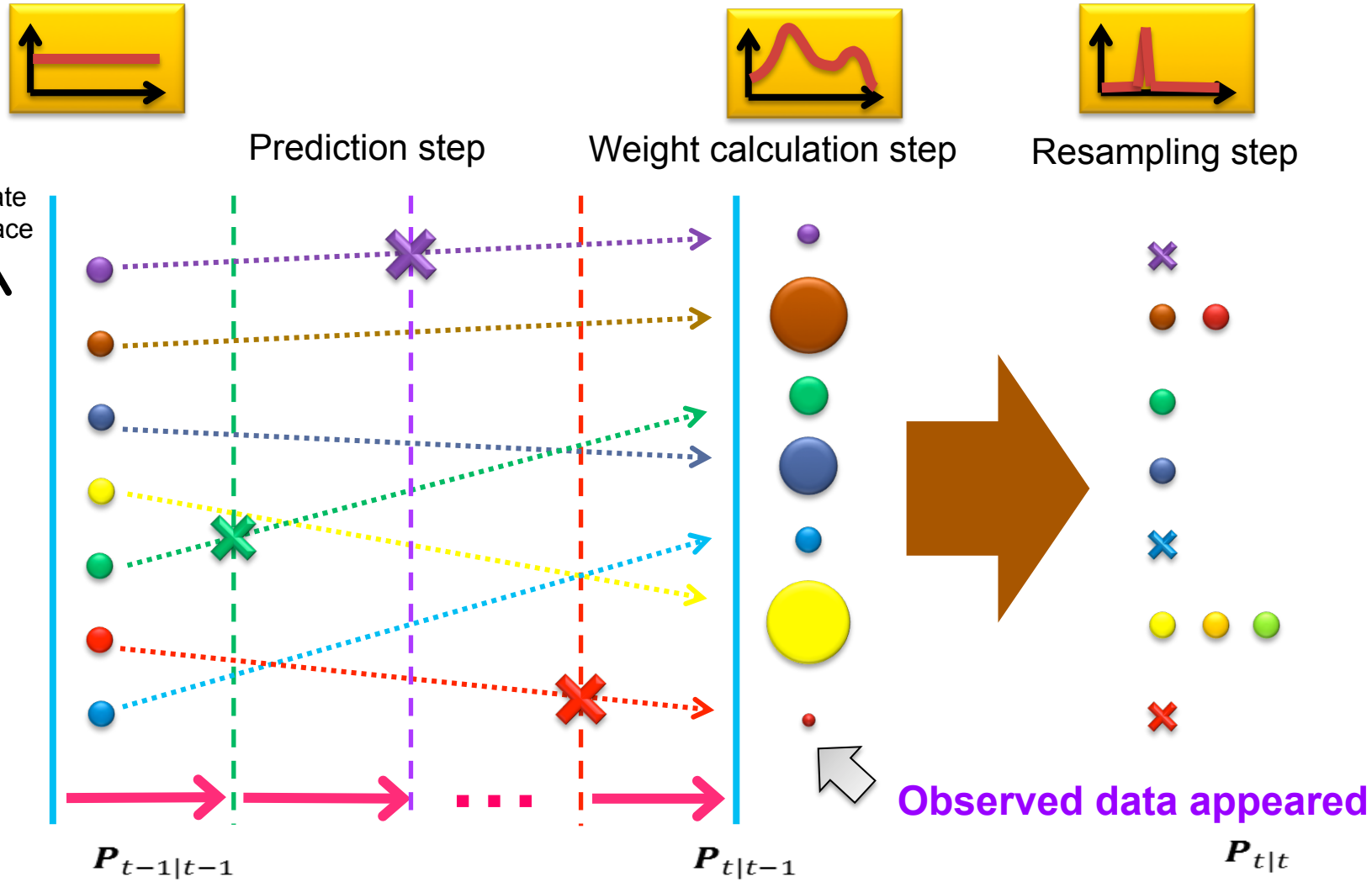


Q2: If the concentration of *per* mRNA always takes the maximum, its concentration is higher than that of *Bmal*

$$G(d([per\ mRNA]) \geqq 0\ \&\&\ X(d([per\ mRNA]) < 0) \rightarrow [per\ mRNA] > [Bmal\ mRNA])$$

We extracted and translated 23 rules with PLTL for checking

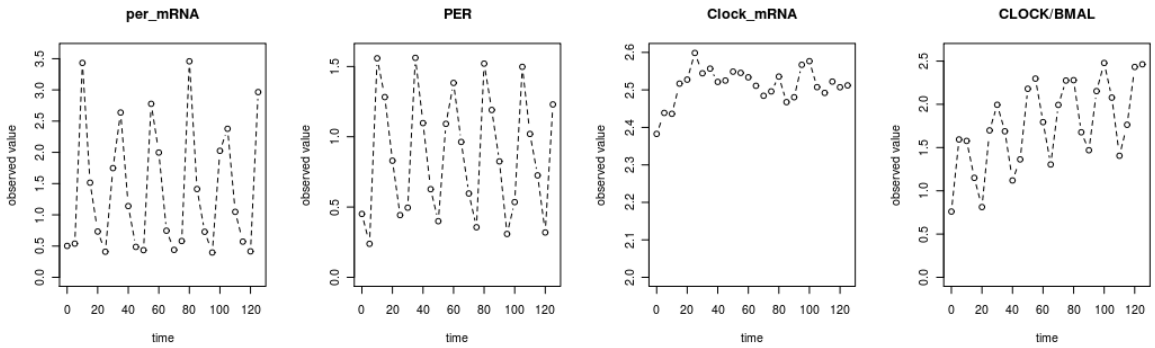
Combining DA with MC



Model checking can be executed at every simulation step until the appearance of observation data

Preparation for case study

- Purpose:
  - Evaluate the performance of the methods
    - **D**ata **A**ssimilation with **M**odel **C**hecking (**DAMC**)
    - **D**ata **A**ssimilation (**DA**)
- Target model:
  - HFPNe model of *Mouse* circadian rhythm
  - **Unknown parameters (17)**:
    12 initial values, 3 threshold values, 2 reaction rates
- Time series data generated by **simulation + Observation noise**
  - **Enough amount** of data: 26 time points for 12 entities (312 in total)
  - **Small amount**: 10 time points for 5 entities



Observed data

Preparation for case study

- Evaluation criteria

$$Score_{mean} = \sum_{t=1}^{T}\sum_{e=1}^{p} \frac{|\ SimResult(Params_{mean}, e, t) - y_{t,p}\ |^2}{|\ \boldsymbol{Y}_T\ |}$$

$$Score_{mode} = \sum_{t=1}^{T}\sum_{e=1}^{p} \frac{|\ SimResult(Params_{mode}, e, t) - y_{t,p}\ |^2}{|\ \boldsymbol{Y}_T\ |}$$

$$Score_{median} = \sum_{t=1}^{T}\sum_{e=1}^{p} \frac{|\ SimResult(Params_{median}, e, t) - y_{t,p}\ |^2}{|\ \boldsymbol{Y}_T\ |}$$

$$Score_{best} = \sum_{t=1}^{T}\sum_{e=1}^{p} \frac{|\ SimResult(Params_{best}, e, t) - y_{t,p}\ |^2}{|\ \boldsymbol{Y}_T\ |}$$

$$Score = \min\{Score_{mean}, Score_{mode}, Score_{median}, Score_{best}, Score_{current}\},$$

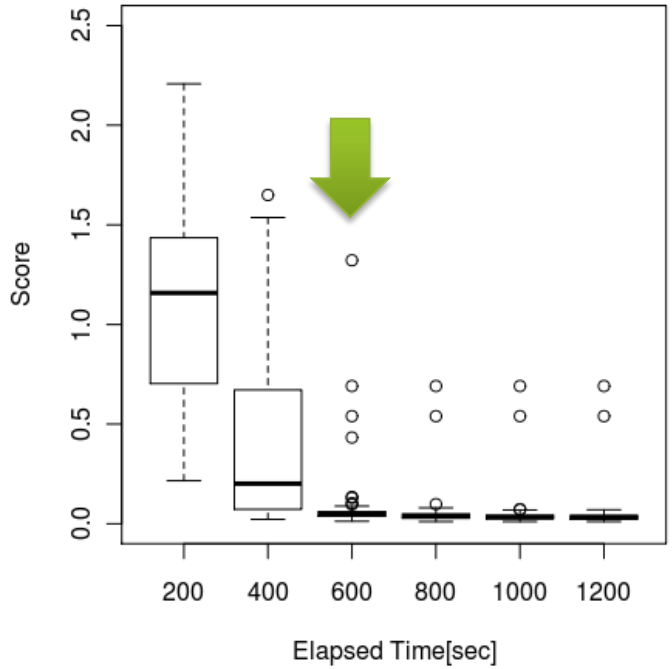
  - The average of squared difference between estimated simulation results and observed data

- Details of estimation environment:
  - Particles: 50,000
  - Searching range: [0, 15]

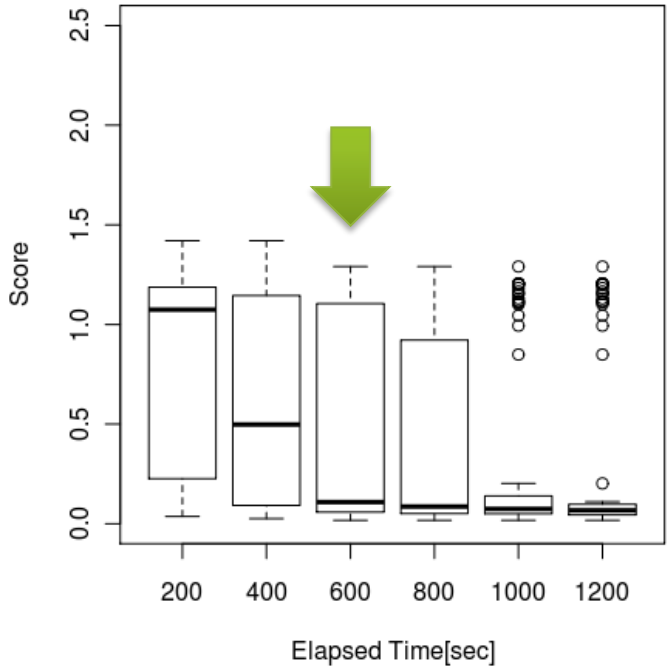
Results with **Enough Amount of Data**

- **26** time points for 12 entities (312 in total) (five cycles)

DAMC



DA



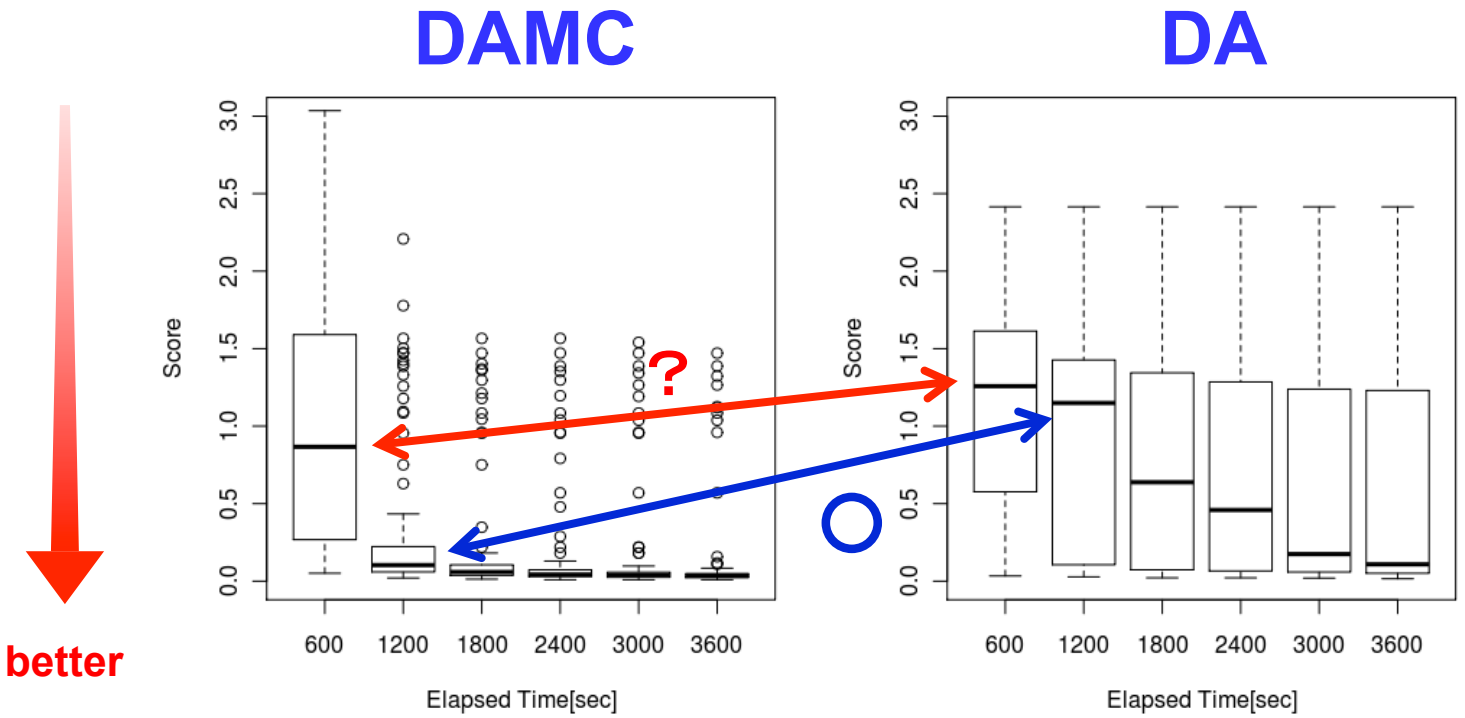
better

Welch's t-test

Elapsed time[sec]	P-value
200	1.584×10^{-6}
400	0.02169
600	1.834×10^{-10}
800	6.165×10^{-9}
1000	9.631×10^{-8}
1200	2.974×10^{-7}

Results with **Small Amount of Data**

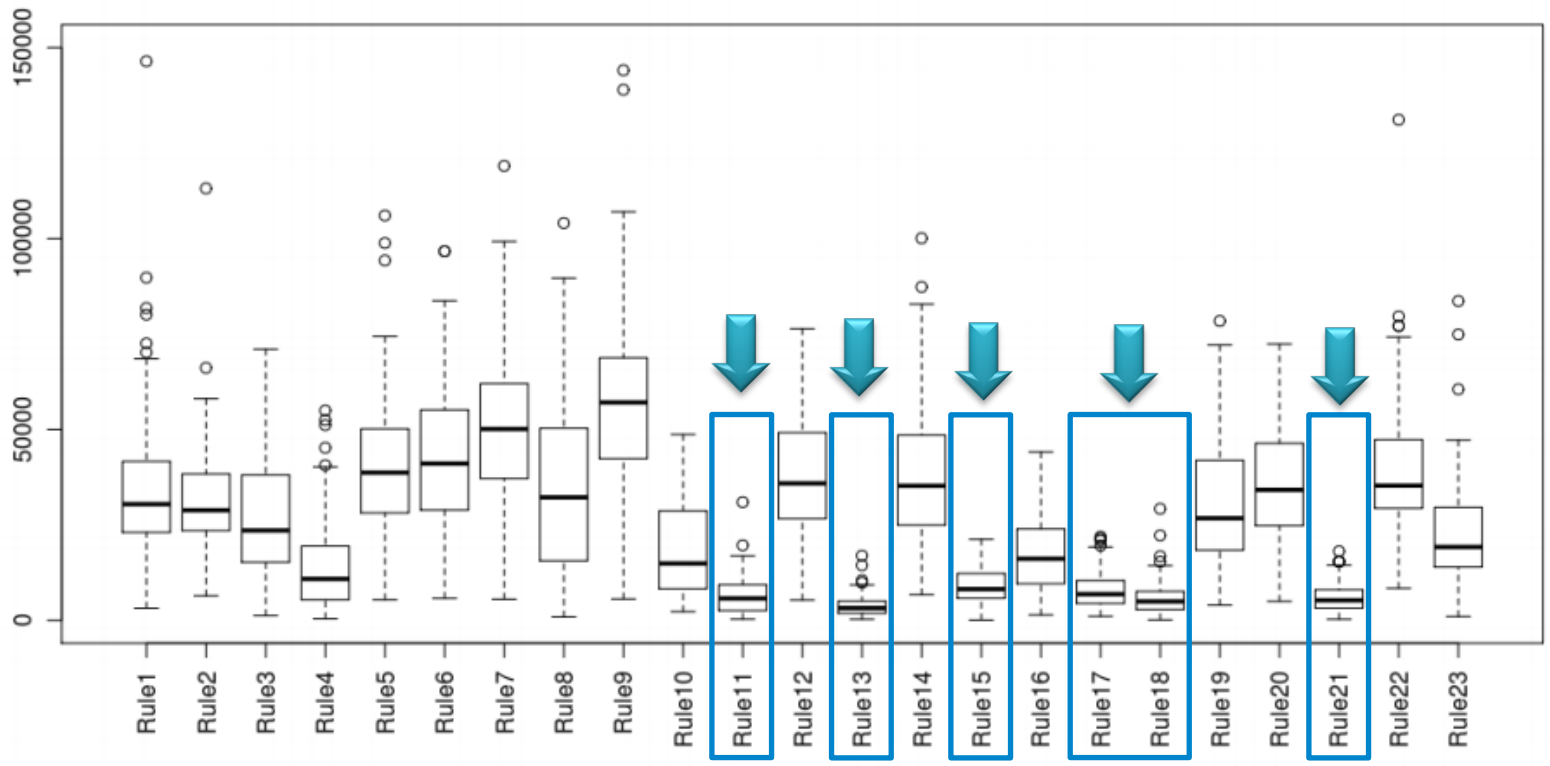
- **10** time points for five entities (about two cycles)



Welch's t-test

Elapsed time[sec]	P-value
600	0.3340
1200	3.725×10^{-10}
1800	1.393×10^{-10}
2400	1.797×10^{-10}
3000	8.596×10^{-9}
3600	3.071×10^{-8}

Results: Effects on Each Rule



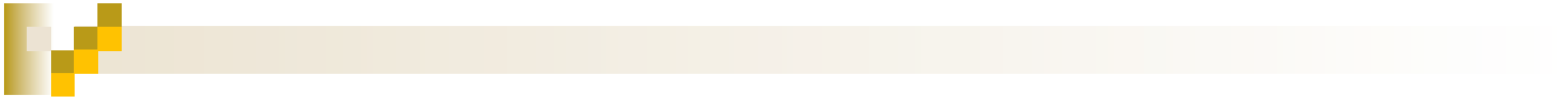
We can understand the contribution of each rule which accelerate estimation efficiency.

Concluding Remarks

- Proposed a novel parameter estimation method for biological pathways by combining MC with DA
  ⇒ Enable us to use various knowledge in addition to observed time series data
- New method and method without model checking are evaluated using *Mouse* circadian rhythm model of HFPNe
  - **Enough amounts** of observed data: our method can practically give good parameters in a short time
  - **Small amount** of observed data: new method is much **faster**
- Model checking will be a great help in improving the efficiency and accuracy of conventional parameter estimation process and eventually leading to better understanding of biological pathways

References

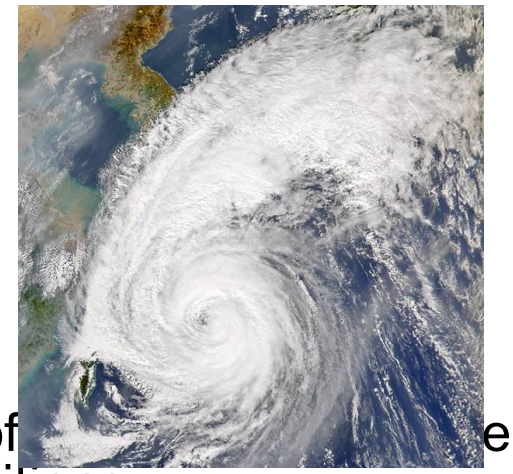
1. Yan, H., Zhang, B., Li, S., Zhao, Q.: A Formal Model for Analyzing Drug Combination Effects and its Application in TNF-alpha-induced NFkappaB Pathway. BMC Syst Biol., 4(50) (2010)
2. Antoniotti, M., Policriti, A., Ugel, N., Mishra, B.: Model Building and Model Checking for Biochemical Processes. Cell Biochem. Biophys., 38(3), 271–286 (2003)
3. Regev, A., Silverman, W., Shapiro, E.: Representation and Simulation of Biochemical Processes Using the Pi-calculus Process Algebra. Pac. Symp. Biocomput., 459–470 (2001)
4. Chaouiya, C.: Petri Net Modelling of Biological Networks. Brief Bioinform., 8(4), 210–219 (2007)
5. Koch, I., Heiner, M.: Petri nets. In Analysis of Biological Networks. Edited by Junker, B.H., Schreiber, F. A Wiley Interscience Publication, 139–180 (2008).
6. Matsuno, H., Li, C., Miyano, S.:Petri Net Based Descriptions for Systematic Understanding of Biological Pathways. IEICE Trans. Fundamentals, E89-A(11), 3166–3174 (2006)
7. Nagasaki, M., Yamaguchi, R., Yoshida, R., Seiya, I., Doi, A., Tamada, Y., Matsuno, H., Miyano, S.: Genomic Data Assimilation for Estimating Hybrid Functional Petri Net from Time-course Gene Expression Data. Genome Inform., 17(1), 46–61, 2006.
8. Tasaki, S., Nagasaki, M., Kozuka-Hata, H., Semba, K., Gotoh, N., Hattori, S., Inoue, J., Yamamoto, T., Miyano, S., Sugano, S., Oyama, M.: Phosphoproteomics-based Modeling Defines the Regulatory Mechanism Underlying Aberrant EGFR Signaling. PLoS One, 5(11), e13926 (2010)
9. B´erard, B., Bidoit, M., Finkel, A., Laroussinie, F., Petit, A., Petrucci, L., Schnoebelen, P., McKenzie, P.: Systems and Software Verification: Model-Checking Techniques and Tools, Springer (2001)
10. Donaldson, R., Gilbert, D.: A Model Checking Approach to the Parameter Estimation of Biochemical Pathways. In: Proceedings of the 6th International Conference on Computational Methods in Systems Biology (CMSB '08), pp. 269–287. SpringerVerlag, Berlin, Heidelberg (2008).
11. Li, C., Nagasaki, M., Hock Koh, C., Miyano, S.: Online model checking approach based parameter estimation to a neuronal fate decision simulation model in C. elegans with hybrid functional Petri net with extension, Mol. Biosyst., 7(5), 1576–1592 (2011)
12. Nagasaki, M., Saito, A., Jeong, E., Li, C., Kojima, K., Ikeda, E., Miyano, S.: Cell Illustrator 4.0: A Computational Platform for Systems Biology, In Silico Biol., 10, 0002 (2010)
13. Nagasaki, M., Doi, A., Matsuno, H., Miyano, S.: A Versatile Petri Net Based Architecture for Modeling and Simulation of Complex Biological Processes. Genome Inform., 15(1), 180–197 (2004)
14. Circadian rhythms in Mus musculus, http://www.csml.org/models/csml-models/circadian-rhythms-in-mouse/
15. Kitagawa, G.: Non-Gaussian State-space Modeling of Nonstationary Time Series. Journal of the American Statistical Association, 82(400), 1032–1063, 1987.
16. Kitagawa, G.: Self-organizing State Space Model. J. of the American Statistical Association, 93, 1203–1215 (1998)
17. Higuchi, T.: Self-organizing Time Series Model. In Sequential Monte Carlo Methods in Practice, Springer-Verlag New York, 429–444 (2001)
18. Kitagawa, G. and Sato, S.: Monte Carlo Smoothing and Self-Organising Statespace Model. In Sequential Monte Carlo Methods in Practice, Springer-Verlag New York, 177–195 (2001)
19. Kitagawa, G.: Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State Space Models.Journal of Computational and Graphical Statistics, 5(1), 1–25 (1996)
20. Heiner, M., Lehrack, S., Gilbert, D., Marwan, W.: Extended Stochastic Petri Nets for Model-based Design of Wetlab Experiments. Edited by Priami, C., et al., Trans. on Comput. Syst. Biol. XI, LNBI, 5750, 138–163 (2009)
21. Donaldson, R., Gilbert, D.: A Monte Carlo Model Checker for Probabilistic LTL with Numerical Constraints, Technical report, University of Glasgow, Department of Computing Science (2008)
22. H¨urzeler and Hans, M., K¨unsch, R.: Approximating and maximising the likelihood for a general state-space model. In: Sequential Monte Carlo Methods in Practice, Springer-Verlag New York, 159–175 (2001)



Thank you!

Data Assimilation in Geophysics

- Typhoon Trajectory (Cyclone, Hurricane, Willy-Willy)
- Tsunami (Seismic sea wave)



Co[illegible] part of [illegible]el, e.g. physical parameters and boundary conditions.