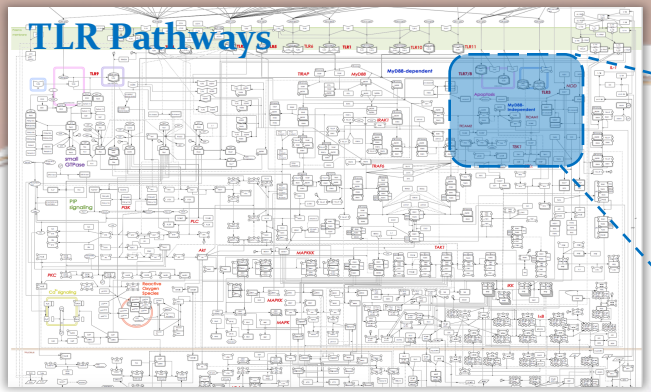
IEEE BIBM'2016, Shenzhen, China

Parameter Estimation of Rule-based Models using Statistical Model Checking

Bing Liu and James R. Faeder

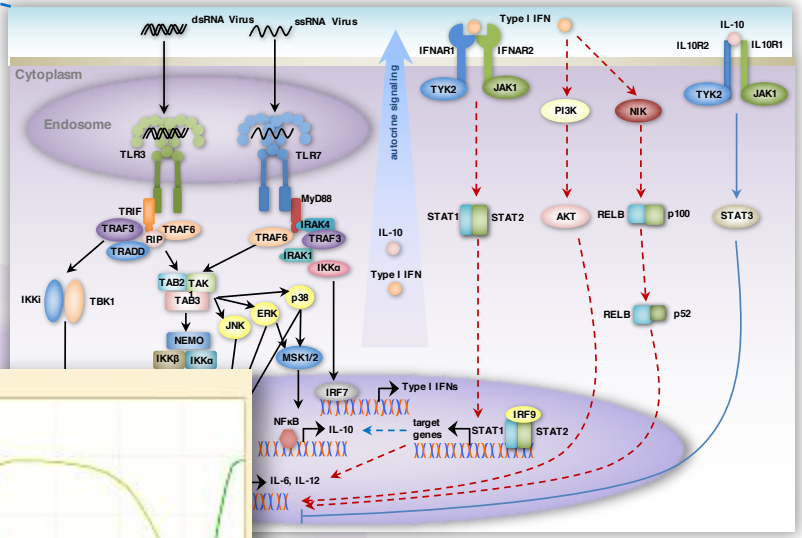
Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh

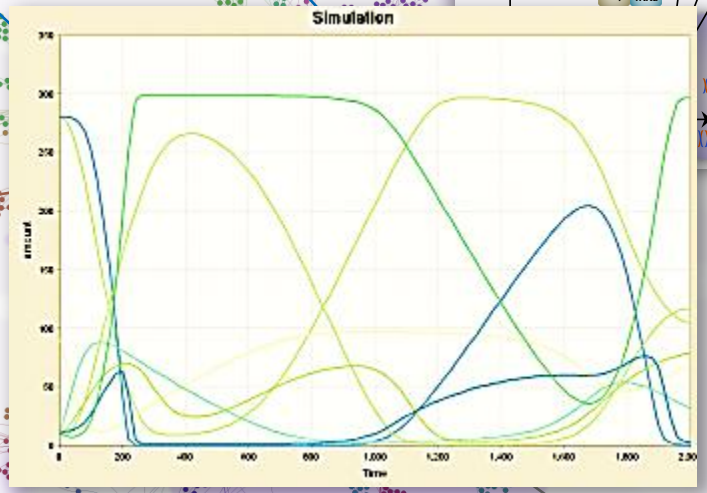
TLR Pathways



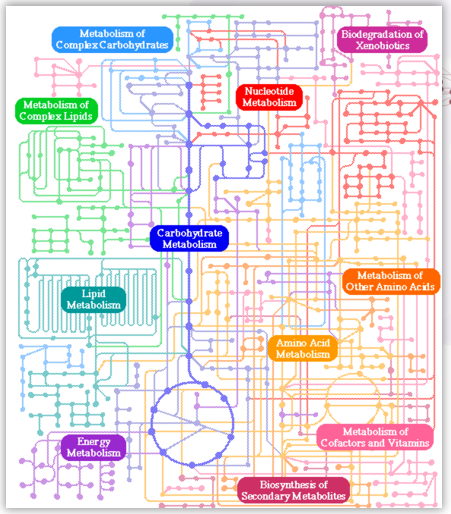
Signaling Pathways

- Cell death
- Cell differentiation
- Cell proliferation
- Cell migration
- … …

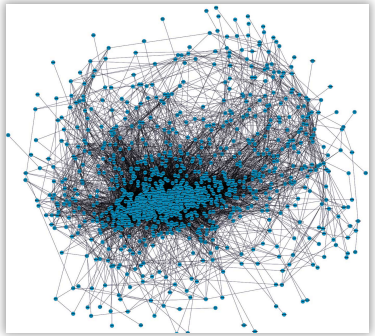




Metabolic Pathways



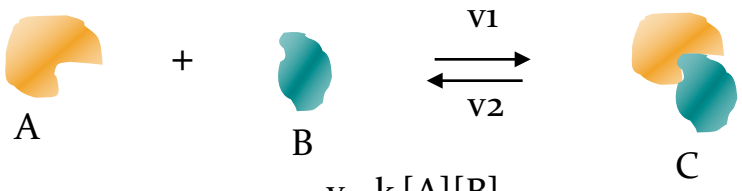
Gene Regulatory Network



Microphage

Computational Modeling

- Modeling formalisms
  - Ordinary Differential Equations
  - Petri Nets
  - Hybrid Automata
  - Markov chains
  - Rule-based models: BioNetGen, Kappa, Pathway Logic, PEPA, PRISM, …
  - …
- ODE Example (protein association):



$v_1 = k_1[A][B]$

$v_2 = k_2[C]$

$d[B]/dt = - v_1 + v_2 = k_1[A][B] - k_2[C]$

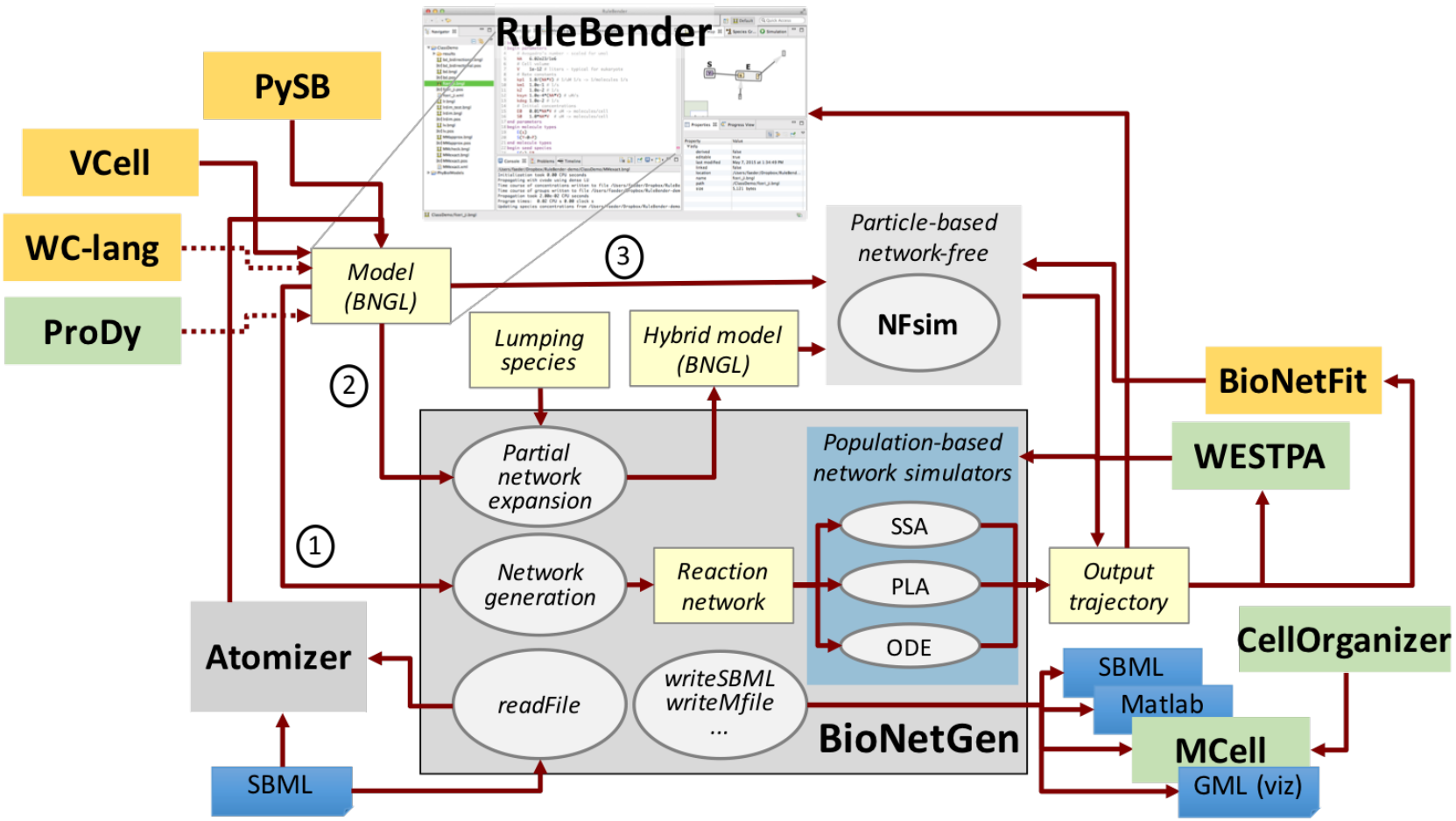
Mass action law

Rule-based Modeling

- Reactions are rules
- A compact representation of ODE and CTMC models
- Avoid the explicit enumeration of all possible molecular species or all the states of a system
- BioNetGen language

```
… …
begin molecule types
S(x~u~p,y~u~p)
E()
end molecule types
begin reaction rules
E(z) + S(y~u) <=> E(z!1).S(y~u!1)     k1, k2
E(z!1).S(y~u!1) -> E(z) + S(y~p)      k3
end reaction rules
… …
```

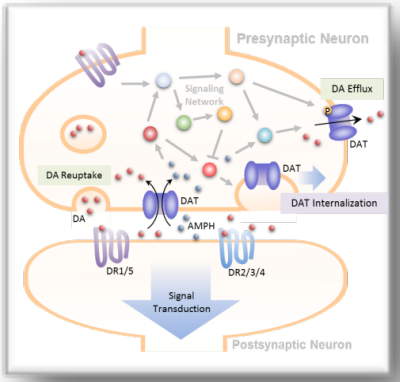
BioNetGen Software Suite

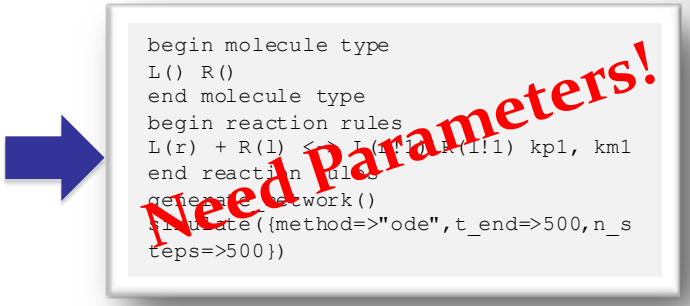


JR Faeder, unpublished

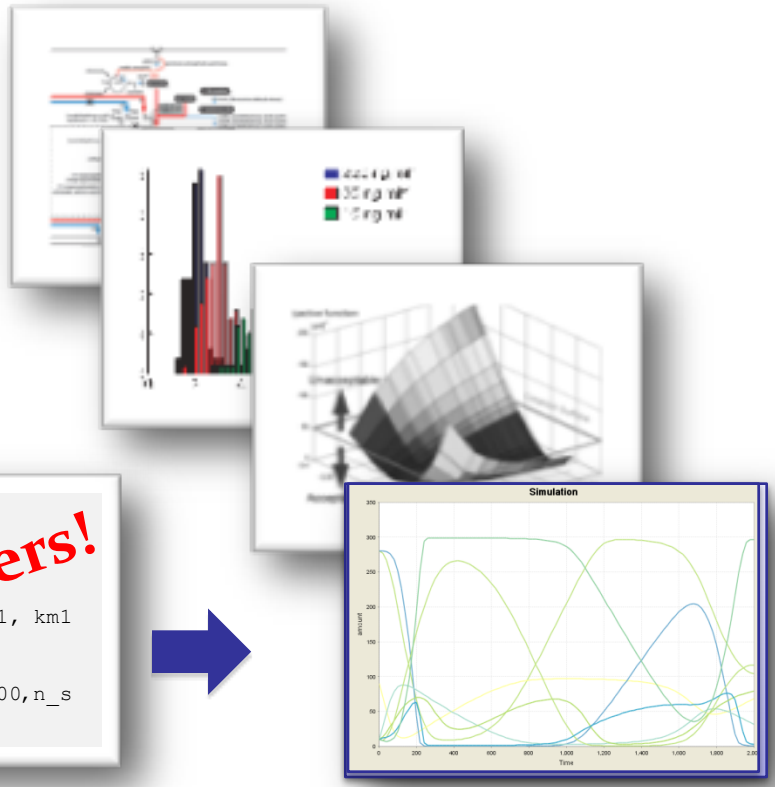
How to answer queries?

- Carry out analysis tasks
  - Perturbation
  - Sensitivity analysis
  - Bifurcation analysis
  - Model checking
  - … …





Need Parameters!

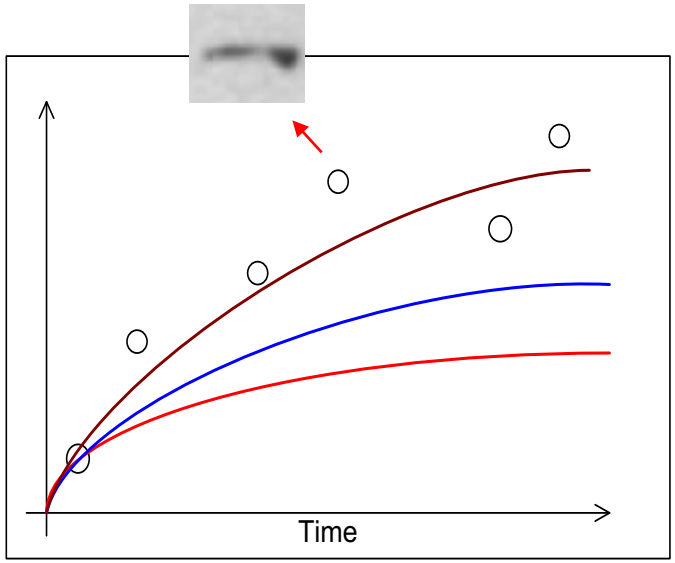


Model Parameters

- Two types of model parameters
  - Initial conditions
  - Rate constants
- Experimental measurements
  - Expensive
  - Not possible to measure all parameters
  - *In vitro* measurements may not reflect the actual physiological conditions in the cell (*Minton, J Biol Chem, 2001*)
  - Cell population-based measurements are not very accurate (*Kim & Price, Phys Rev Lett, 2010*)

Parameter Estimation

- Goal:
  - Find values of parameter so that model prediction generated by simulations using these values can match experimental data (e.g. time serials, steady state)



krbNGF = 0.33, KmAkt = 0.16, kpRaf1 = 0.42 … …

target

krbNGF = 0.49, KmAkt = 0.08, kpRaf1 = 0.97 … …

krbNGF = 0.88, KmAkt = 0.21, kpRaf1 = 0.05 … …

Optimization Approach

- Minimize the difference between model prediction and experimental data

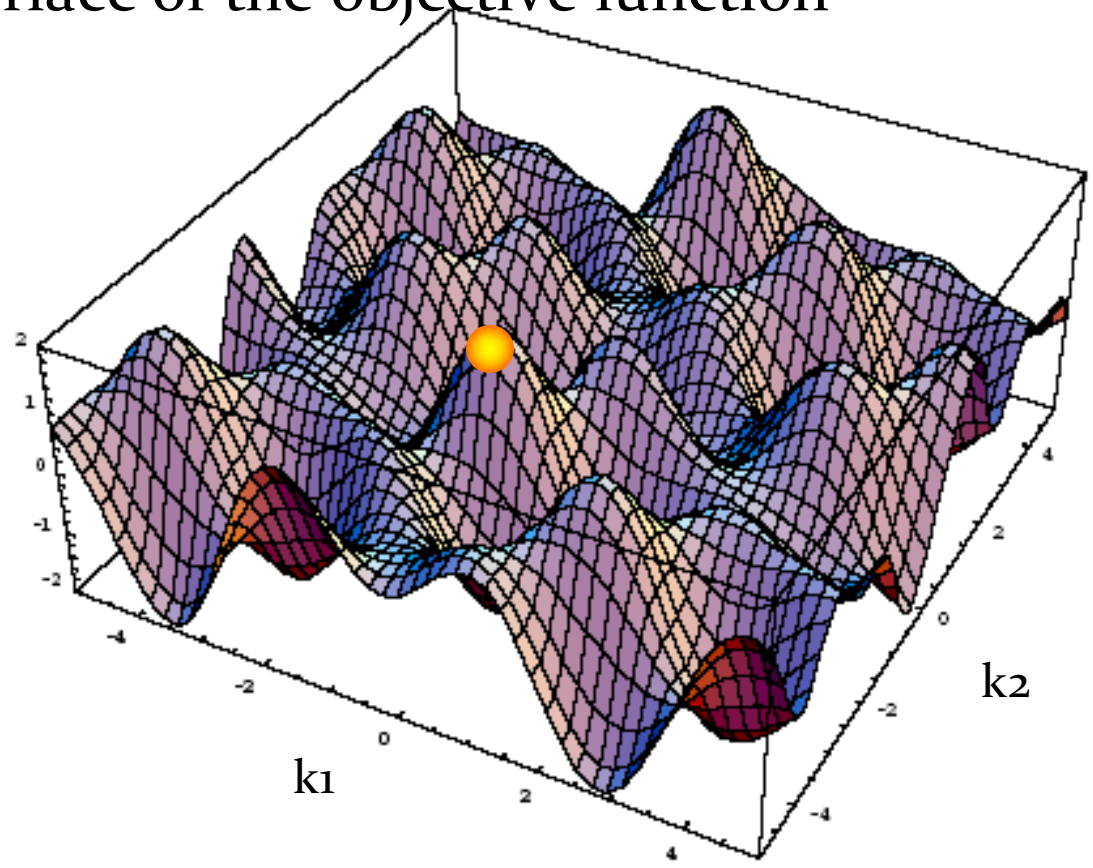
Given data $\tilde{\mathbf{x}}(t_j)$, find $\mathbf{k}$ to

minimize $J(\mathbf{k}) = \sum_j ||\mathbf{x}(t_j; \mathbf{k}) - \tilde{\mathbf{x}}(t_j)||^2$

J: objective function

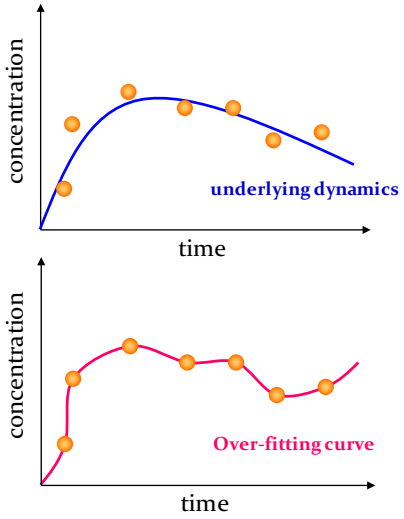
Example: Steepest Decent

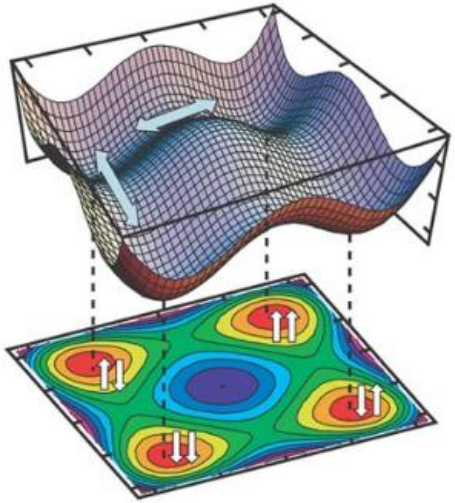
- Update following the direction of steepest descent on the hyper-surface of the objective function



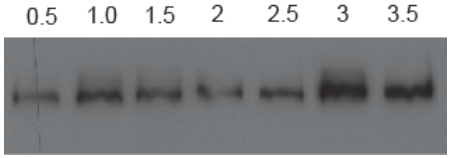
Many Challenges

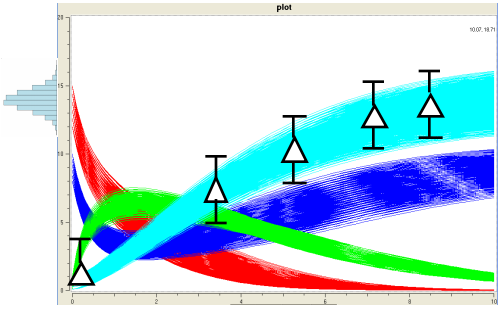
- The curse of dimensionality
- Over-fitting
- Non-identifiable models
- Inherent uncertainty of data





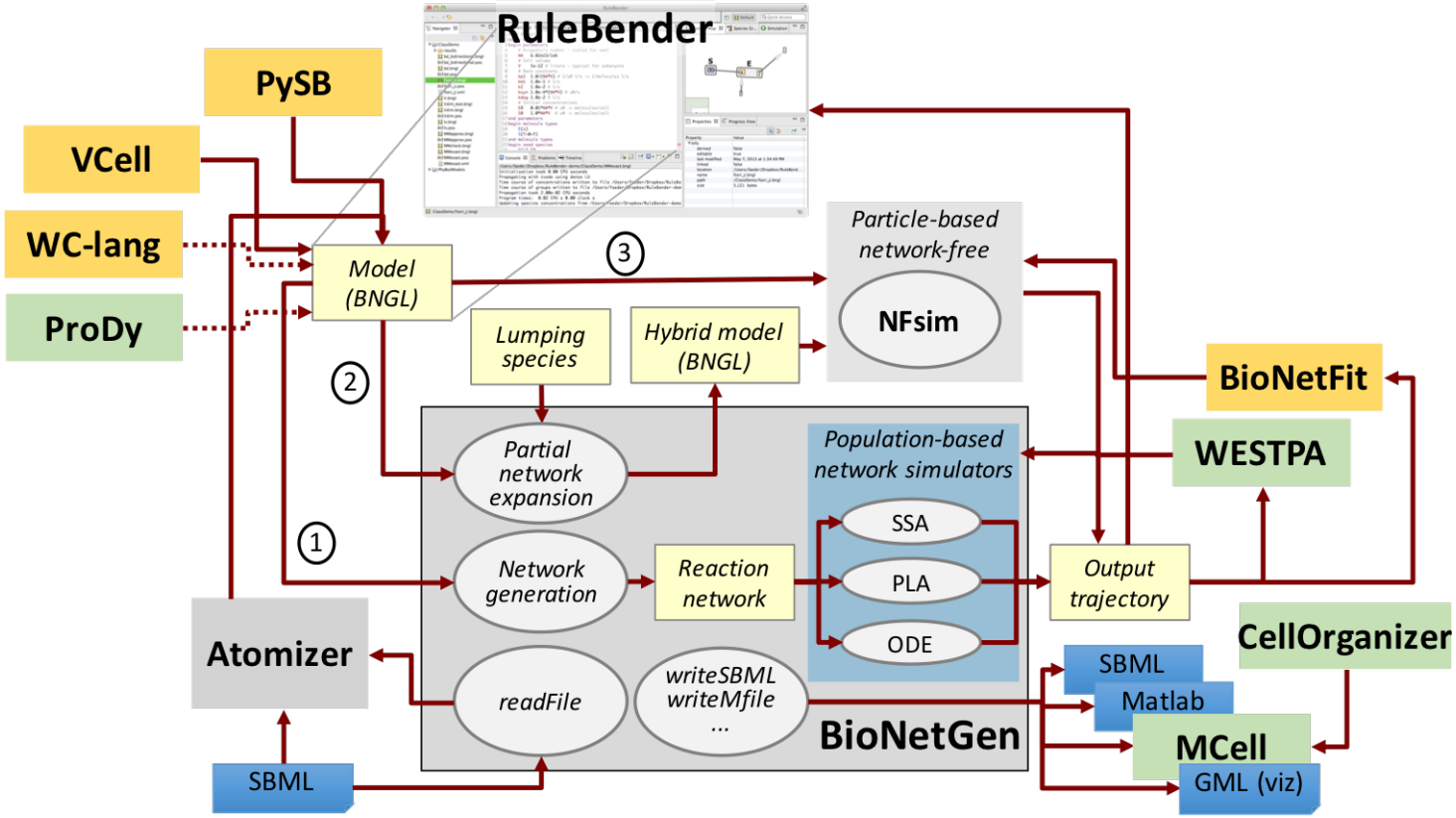
Kim et al. 2007





Parameter Estimation for BioNetGen

- Current solutions: ptempest, BioNetFit, SBML tools



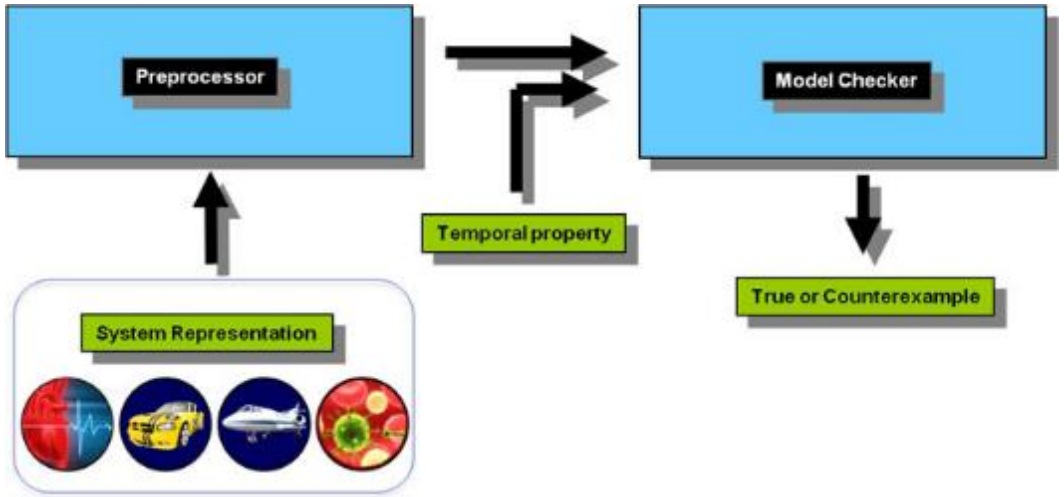
JR Faeder, unpublished

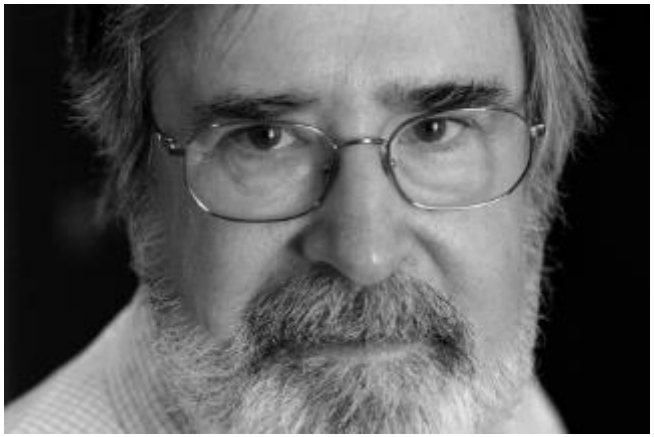
Our Solution

- A statistical model checking (SMC) based approach
  - Encode training data as a ***bounded linear temporal logic*** formula
  - Evaluate candidate parameters using SMC
  - Perform global optimization (stochastic ranking evolutionary strategy (SRES))
- Advantages
  - Utilize both ***quantitative*** and ***qualitative*** knowledge
  - Deal with uncertainty of the biological system/data
  - Good scalability due to the power of statistical testing
- Extending our previous method for ODE models with prior distribution of initial states (*Palaniappan et al, CMSB, 2013*)

Model Checking

- An automated method to formally verify a system's behavior with respect to a set of properties





Edmund M. Clarke (Turing Award 2007)

BLTL

- Atomic proposition: $(i,l,u), L_i \le l < u \le U_i$
  - the current concentration level of x_i falls in the interval $[l,u]$
- The formulas of BLTL are:
  - $\psi ::= AP \mid true \mid false \mid \psi_1 \vee \psi_2 \mid \neg \psi \mid \psi_1 \mathbf{U}^{\le t} \psi_2 \mid \psi_1 \mathbf{U}^{t} \psi_2$
  - Derived operators: $\wedge, \supset, \equiv, \mathbf{G}^{\le t}, \mathbf{G}^{t}, \mathbf{F}^{\le t}, \mathbf{F}^{t}$
- A finite set of time points $\mathbf{T} = \{0,1,\ldots,T\}$

BLTL

- Atomic proposition: $(i,l,u), L_i \le l < u \le U_i$
  - the current concentration level of x_i falls in the interval [l,u]
- The formulas of BLTL are:
  - $\psi ::= AP \mid true \mid false \mid \psi_1 \vee \psi_2 \mid \neg \psi \mid \psi_1 \mathbf{U}^{\le t} \psi_2 \mid \psi_1 \mathbf{U}^{t} \psi_2$
  - Derived operators: $\wedge, \supset, \equiv, \mathbf{G}^{\le t}, \mathbf{G}^{t}, \mathbf{F}^{\le t}, \mathbf{F}^{t}$
- A finite set of time points $\mathbf{T} = \{0,1,\ldots,T\}$

BLTL

- Semantics

The semantics of the logic is defined in terms of the relation $\sigma, t \models \varphi$ where σ is a trajectory in BEH and $t \in \mathcal{T}$.

- $\sigma, t \models (i, \ell, u)$ iff $\ell \leq \sigma(t)(i) \leq u$ where $\sigma(t)(i)$ is the i^{th} component of the n-dimensional vector $\sigma(t) \in \mathbf{V}$.
- $\neg$ and $\vee$ are interpreted in the usual way.
- $\sigma, t \models \psi \mathbf{U}^{\leq k} \psi'$ iff there exists k' such that $k' \leq k$, $t + k' \leq T$ and $\sigma, t + k' \models \psi'$. Further, $\sigma, t + k'' \models \psi$ for every $0 \leq k'' < k'$.
- $\sigma, t \models \psi \mathbf{U}^{k} \psi'$ iff $t + k \leq T$ and $\sigma, t + k \models \psi'$. Further, $\sigma, t + k' \models \psi$ for every $0 \leq k' < k$.

We can now define $models(\psi) = \{\sigma \mid \sigma, 0 \models \psi,\ \sigma \in BEH\}$.

Probabilistic BLTL

$P_{\geq r}(\psi), P_{\leq r'}(\psi)$, where $r \in [0,1), r' \in (0,1]$ and ψ is a BLTL formula

- The probability that a trajectory in *BEH* belong to *models*(ψ) exceeds or equal to *r*
- Based on measure theory and our assumptions, we can define *P(Models(ψ))*
- Given ODE system *S*,

$$\mathcal{S} \models P_{\geq r}\psi \text{ iff } P(Models(\psi)) \geq r$$

$$\mathcal{S} \models P_{\leq r'}\psi \text{ iff } P(Models(\psi)) \leq r'$$

SMC of PBLTL formulas

- Sequential hypothesis test between

  $H0: p \geq r + \delta$ and $H1: p \leq r - \delta$, where $p = P(Models(\psi))$

  - Generating a sequence of sample trajectories by randomly sampling *INIT*
  - Verify each trajectory and determine whether accept H0 or H1 based on Type I and Type II error bounds

- Can be an on-line method

Encoding Knowledge

- Quantitative experimental data

$$\psi_i^t = \mathbf{F}^t(i, l_i^t, u_i^t)$$

$$\psi_{\exp} = \wedge_{i \in O}(\wedge_{t \in T_i} \psi_i^t)$$

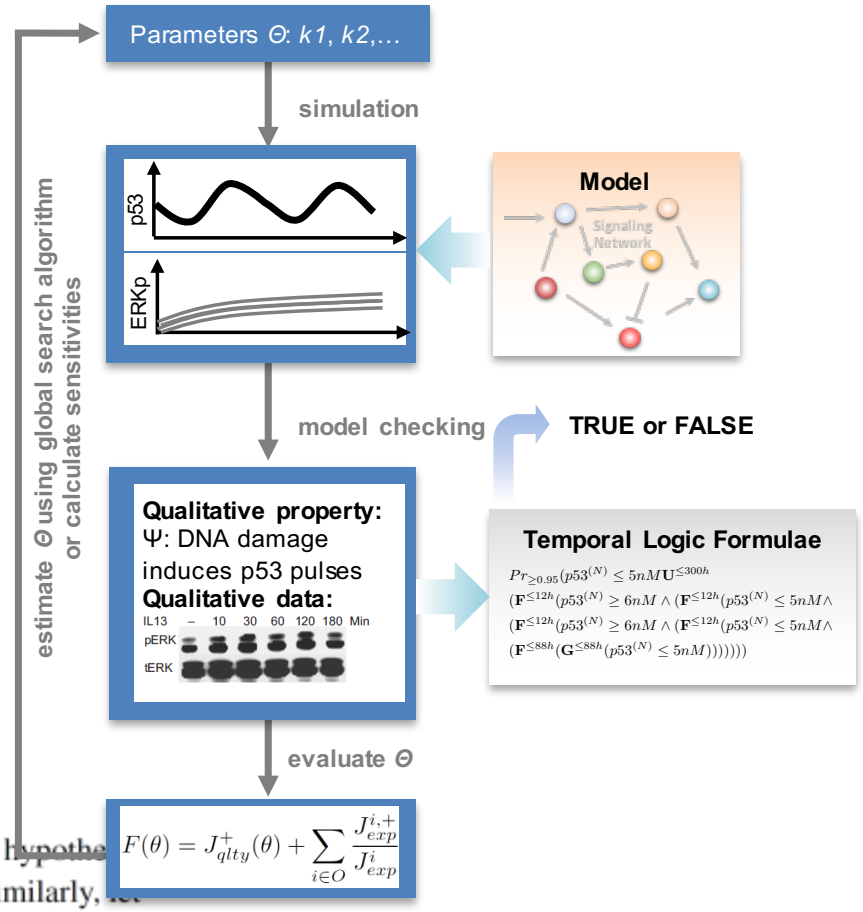
- Qualitative properties of the dynamics
  - E.g. transient/sustained activation, oscillatory behavior, bistable, …
  - '*trend*' formulas: ψ_{qlty}

- PBLTL formula: $P_{\geq r}(\psi_{\exp} \wedge \psi_{qlty})$

SMC based Parameter Estimation

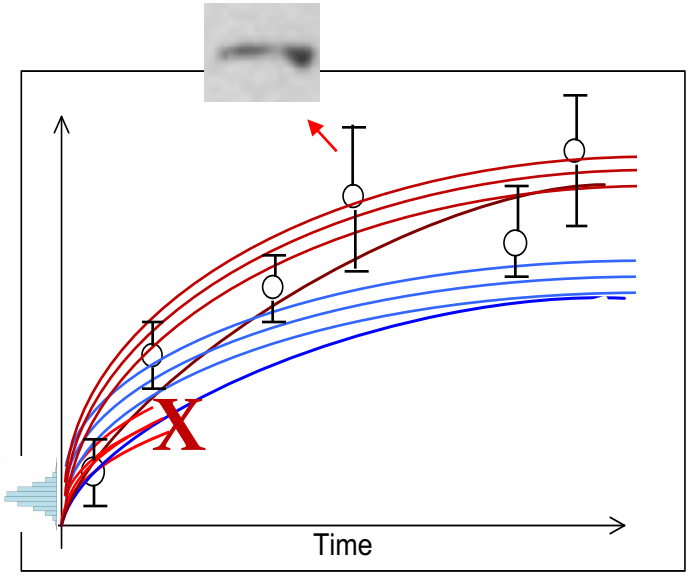
1. Guess θ_l
2. Verify $\psi_{\text{exp}} \wedge \psi_{qlty}$ with the chosen strength
3. Compute $F(\theta_l)$
4. Terminate or make a new guess (based on search strategy e.g. SRES) and repeat step 1

$$F(\theta) = J^{+}_{qlty}(\theta) + \sum_{i \in O} \frac{J^{i,+}_{exp}}{J^{i}_{exp}}$$



Let $J^{i,+}_{exp}(\theta)$ be the number of formulas of the form ψ^t_i (a conjunct in ψ^i_{exp}) such that the statistical test for $P_{\geq r}(\psi^t_i)$ accepts the null hypothe (that is, $P_{\geq r}(\psi^t_i)$ holds) with the strength $(\frac{\alpha}{J}, \beta)$, where $J = \sum_{i \in O} J^i_{exp}$. Similarly, let $J^{+}_{qlty}(\theta)$ be the number of conjuncts in ψ_{qlty} of the form $\psi_{\ell,qlty}$ that pass the statistical test $P_{\geq r}(\psi_{\ell,qlty})$ with the strength $(\frac{\alpha}{J}, \beta)$.

SMC based Parameter Estimation



krbNGF = 0.33, KmAkt = 0.16, kpRaf1 = 0.42 … …

target

krbNGF = 0.49, KmAkt = 0.08, kpRaf1 = 0.97 … …

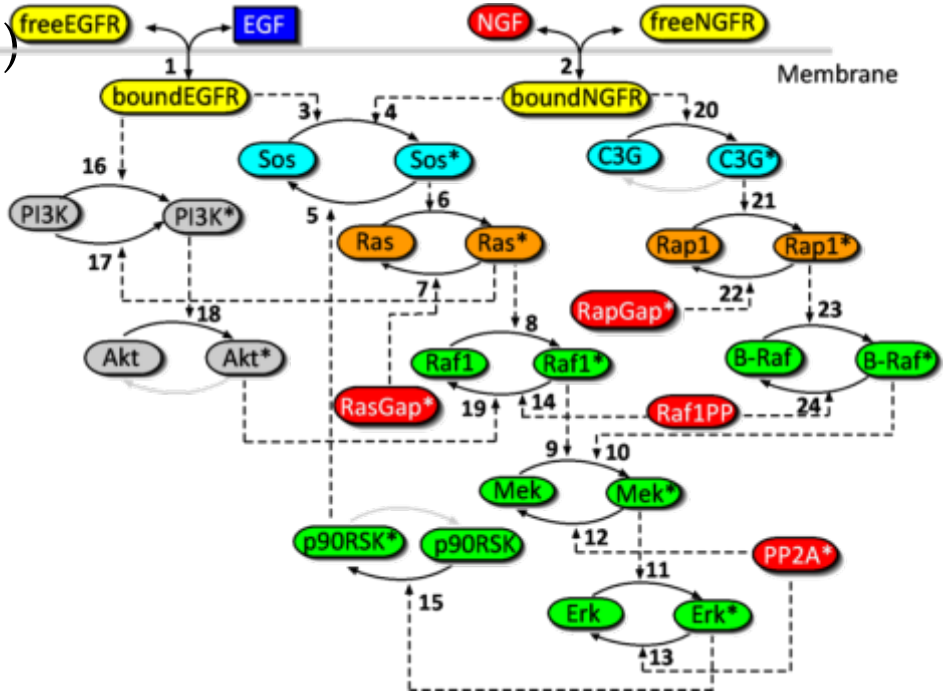
krbNGF = 0.88, KmAkt = 0.21, kpRaf1 = 0.05 … …

Case Studies

- Pathway models taken from BioModels database
- Nominal parameters
- Synthetic experimental data
- Qualitative trend

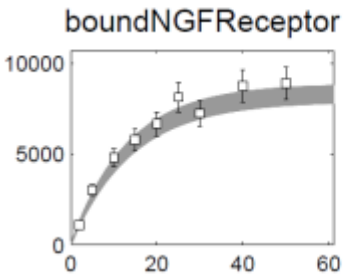
EGF-NGF Pathway

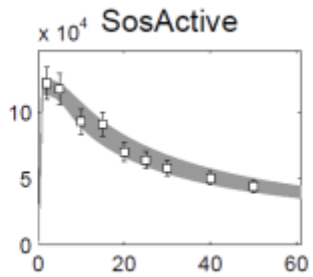
- ODE model (*Brown et al. 2004*)
  - 32 species
  - 48 parameters (20 unknown)
- Training data
  - 7 species, 9 time points
- Test data
  - 2 species, 9 time points

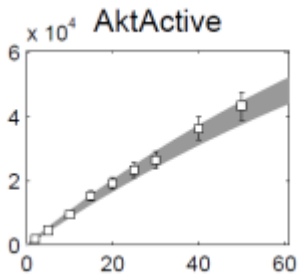


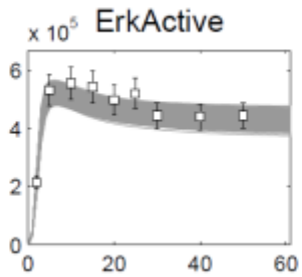
EGF-NGF Pathway

- Running time: 2.23 hours

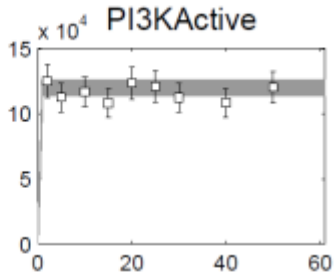








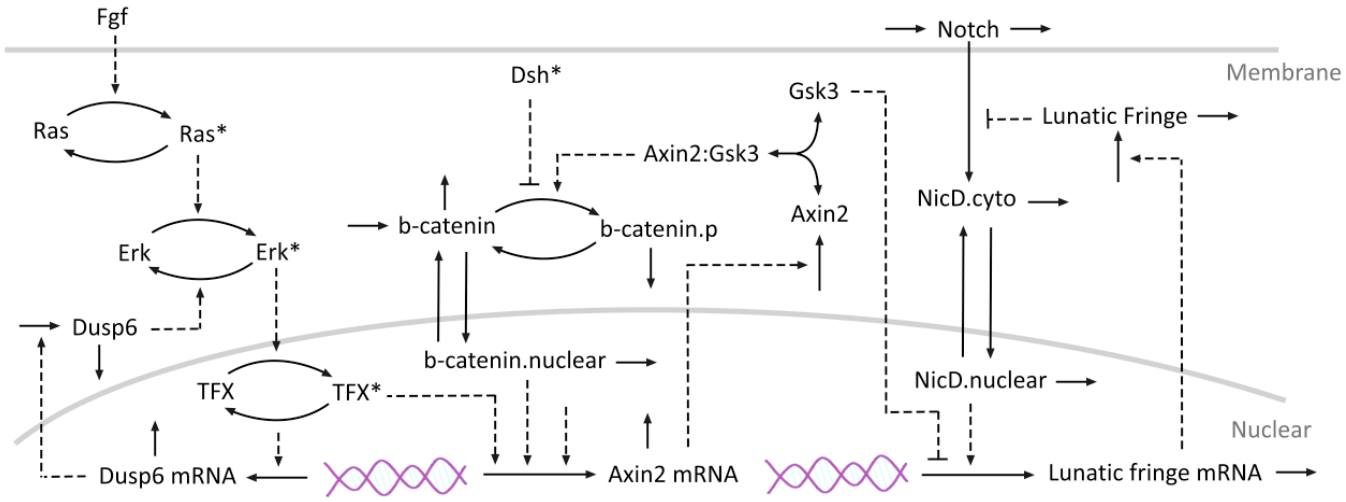
Training data



Test data

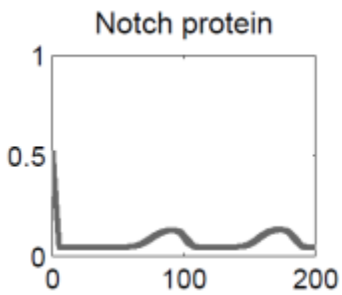
Segmentation Clock Network

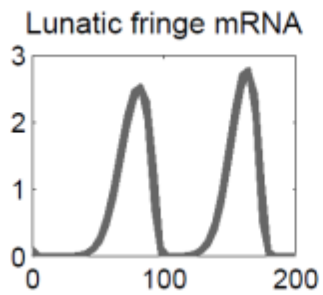
- ODE model (*Goldbeter et al. 2008*)
  - 22 species, 75 parameters (40 unknown)
- Training data
  - Time serials: Axin2 mRNA, 14 time points
  - Qualitative trend: 5 species, oscillatory behavior
    - E.g. $(([LmRNA \leq 0.4] \wedge (\mathbf{F}([LmRNA \geq 2.2] \wedge \mathbf{F}([LmRNA \leq 0.4] \wedge (\mathbf{F}([LmRNA \geq 2.2] \wedge \mathbf{F}([LmRNA \leq 0.4]))))))$
- Test data: Dusp6 protein, qualitative trend

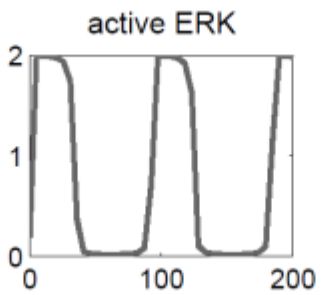


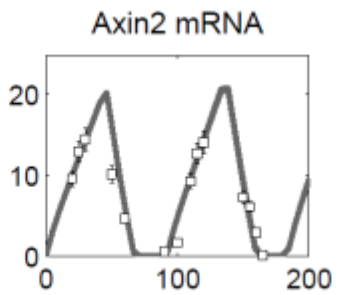
Segmentation Clock Network

- Running time: 2.2 hours

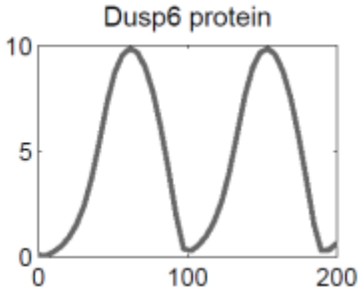








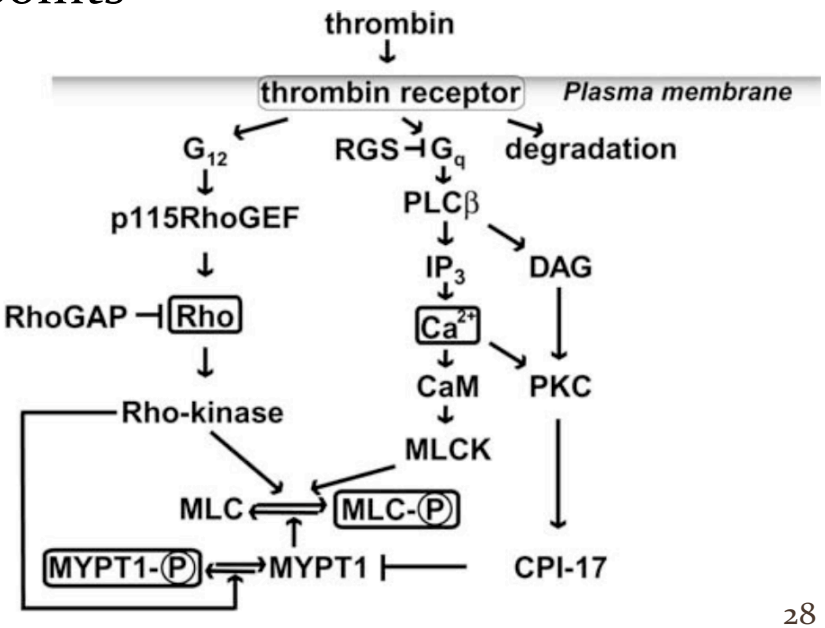
Training data



Test data

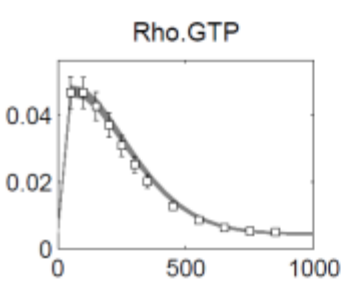
MLC Phosphorylation Pathway

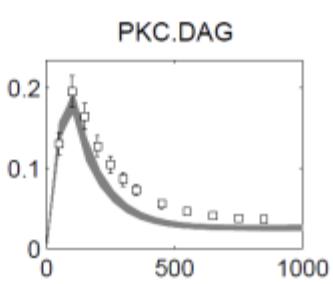
- Regulates the contraction of endothelia cells
- ODE model (*Maeda et al 2006*)
  - **105** species, 197 parameters (**100** unknown parameters)
- Training data
  - Time serials: 8 species, 12 time points
  - Qualitative trend: 2 species
- Test data
  - 2 species, 12 time points

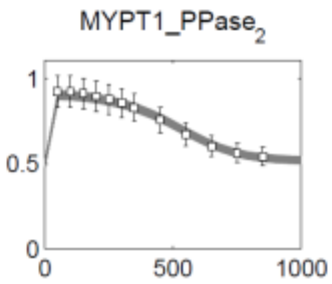


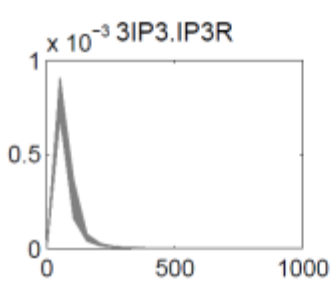
MLC Phosphorylation Pathway

- Running time: 50.67 hours

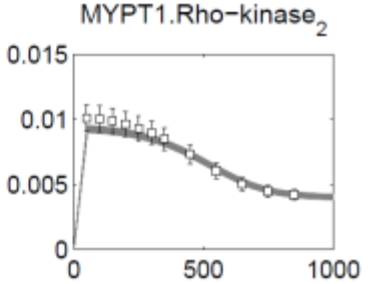








Training data



Test data

Conclusion

- A SMC based approach for the parameter estimation of bio-pathway models
- Utilize both quantitative experimental data and qualitative knowledge
- Deal with uncertainty of the initial states and the noisy cell-population data
- Employ standard search strategies
- Can be used to perform global sensitivity anlaysis

Future work

- Stochastic differential equation (SDE) based models
- Hybrid systems
- GPU acceleration

Acknowledgements

University of Pittsburgh
Ivet Bahar's Lab

Carnegie Mellon University
Edmund M. Clarke's Lab

National University of Singapore
P.S. Thiagarajan
David Hsu
Jeak Ling Ding