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

- 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):

\[
\begin{align*}
\frac{d[B]}{dt} &= -v_1 + v_2 \\
v_1 &= k_1[A][B] \\
v_2 &= k_2[C]
\end{align*}
\]

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

```plaintext
... ...
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

Combinatorial complexity limits the standard network modeling approach. We can understand the limitations on the conventional ODE-based modeling approaches through the simple example shown in Fig. V.2. A small number of molecules interacting in the prescribed way can generate a huge number of possible species—over 1000 in this case—which would require an equal number of ODE's to model. In practice, most modelers avoid this complexity by making additional assumptions to limit the number of possible combinations. For example, the complexity drops nearly two orders of magnitude if one assumes that only one of the adaptors (orange) can bind to the receptor (blue) at a time. Although this is currently standard practice, there is no principled way to carry out this step; doing so requires assumptions that may introduce errors.

Combination complexity arises throughout biology even in apparently simple systems involving the interaction of only a few proteins and becomes a major limiting factor in the modeling of signaling systems as diverse as the epidermal growth factor receptor (EGFR), the MAP kinase cascade, the T cell receptor, CaMKII, and the postsynaptic density. Conventional network modeling approaches based on ODEs therefore face fundamental limits on scalability and accuracy.

Rule-based Modeling (RBM) is ODE's and much more. In RBM molecular interactions, such as those shown in the contact map of Fig. V.2, are encoded as rules, which specify the properties that a particular set of reactants must possess and a function that determines the rate of interaction. The model that generated this contact map had 18 such reaction rules. BioNetGen can expand these rules to generate the full set of ODEs or simulate the model in other ways—using stochastic dynamics, PDEs, etc. (Fig. V.3).

Besides compactness, another advantage of the rule-based approach is that the coarse-grained structural features of the molecules are explicitly represented, which facilitates understanding and enables mapping to finer structural scales.

Network-free simulation provides scalable simulation of RBMs. RBM languages make it easy to encode models for which the full set of equations is too large to enumerate in advance. "Network-free" simulation methods avoid explicit generation of species and reactions by using particle-based simulation driven by the rules. These simulation methods have a computational scaling that is nearly independent of network size.
How to answer queries?

- Carry out analysis tasks
  - Perturbation
  - Sensitivity analysis
  - Bifurcation analysis
  - Model checking
  - ... ...
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)

\[
\begin{align*}
krb_{NGF} &= 0.33, \ KmAkt = 0.16, \ kpRaf1 = 0.42 \quad \ldots \\
krb_{NGF} &= 0.49, \ KmAkt = 0.08, \ kpRaf1 = 0.97 \quad \ldots \\
krb_{NGF} &= 0.88, \ KmAkt = 0.21, \ kpRaf1 = 0.05 \quad \ldots 
\end{align*}
\]
Optimization Approach

- Minimize the difference between model prediction and experimental data

\[
J(k) = \sum_{j} \| x(t_j; k) - \tilde{x}(t_j) \|^2
\]

Given data $\tilde{x}(t_j)$, find $k$ to minimize $J(k)$. $J$: objective function
Example: Steepest Decent

- Update following the direction of steepest descent on the hyper-surface of the objective function $J(k_1,k_2)$
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

---

**Figure V.2. Combinatorial complexity in a model of epidermal growth factor receptor (EGFR) signaling.**

A. Contact map showing interactions of receptor (R), its ligand (EGF), three intracellular proteins through specific components that represent sites of binding and posttranslation modification (represented by yellow circles).

B. Combinatorial complexity in the number of EGFR-containing species. Each leaf in the tree represents a different possible state of a receptor component. There are 48 possible monomeric and 1176 possible dimeric species.

---

**Figure V.3. Connectivity between BioNetGen and other selected tools for modeling and simulation.**

Green boxes indicate other MMBioS tools, orange boxes indicate tools developed by our C&SP collaborators. Current communications are file-based but would be more efficiently managed with the development of the libBNG API in Aim 3.1.
Our Solution

- A statistical model checking (SMC) based approach
  - Encode training data as a \textit{bounded linear temporal logic} formula
  - Evaluate candidate parameters using SMC
  - Perform global optimization (stochastic ranking evolutionary strategy (SRES))

- Advantages
  - Utilize both \textit{quantitative} and \textit{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 \textit{(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: \((i, l, u), L_i \leq l < u \leq U_i\)
  - the current concentration level of \(x_i\) falls in the interval \([l, u]\)
- The formulas of BLTL are:
  - \(\psi ::= AP \mid true \mid false \mid \psi_1 \lor \psi_2 \mid \neg \psi \mid \psi_1 U^{\leq t} \psi_2 \mid \psi_1 U^t \psi_2\)
  - Derived operators: \(\land, \lor, \equiv, G^{\leq t}, G^t, F^{\leq t}, F^t\)
- A finite set of time points \(T = \{0, 1, \ldots, T\}\)
BLTL

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

- **Semantics**

  The semantics of the logic is defined in terms of the relation \( \sigma, t \models \varphi \) where \( \sigma \) 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 \( i^{\text{th}} \) component of the \( n \)-dimensional vector \( \sigma(t) \in \mathbf{V} \).
  - \( \neg \) and \( \lor \) 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 \( \text{models}(\psi) = \{ \sigma \mid \sigma, 0 \models \psi, \sigma \in BEH \} \).
Probabilistic BLTL

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

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

  $S \models P_{\geq r}\psi$ iff $P(Models(\psi)) \geq r$
  $S \models P_{\leq r'}\psi$ iff $P(Models(\psi)) \leq r'$
SMC of PBLTL formulas

- Sequential hypothesis test between
  \( H_0: p \geq r + \delta \) and \( H_1: p \leq r - \delta \), where \( p = P(\text{Models}(\psi)) \)
  - Generating a sequence of sample trajectories by randomly sampling \( \text{INIT} \)
  - Verify each trajectory and determine whether accept \( H_0 \) or \( H_1 \) based on Type I and Type II error bounds

- Can be an on-line method
Encoding Knowledge

- Quantitative experimental data
  \[ \psi_i^t = F^t(i, l_i^t, u_i^t) \]
  \[ \psi_{\text{exp}} = \land_{i \in O} (\land_{t \in T_i} \psi_i^t) \]

- Qualitative properties of the dynamics
  - E.g. transient/sustained activation, oscillatory behavior, bistable, ...
  - ‘trend’ formulas: \( \psi_{qly} \)

- PBLTL formula: \( P_{\geq r} (\psi_{\text{exp}} \land \psi_{qly}) \)
SMC based Parameter Estimation

1. Guess $\theta_l$
2. Verify $\psi_{\exp} \land \psi_{\text{qlty}}$ 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

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

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

**Temporal Logic Formulae**

$P_{\text{SNAP}}(p_{\text{SNAP}}^{100}) \leq 5nM$ U $\leq \text{SNAP}$

$\left(\begin{array}{c}
P_{\text{SNAP}}(p_{\text{SNAP}}^{100}) \geq 6nM \land P_{\text{SNAP}}(p_{\text{SNAP}}^{100}) \leq 5nM \\
P_{\text{SNAP}}(p_{\text{SNAP}}^{100}) \geq 6nM \land P_{\text{SNAP}}(p_{\text{SNAP}}^{100}) \leq 5nM \land P_{\text{SNAP}}(p_{\text{SNAP}}^{100}) \leq 5nM
\end{array}\right)$
SMC based Parameter Estimation

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

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 ≤ 0.4] ∧ (F([LmRNA ≥ 2.2] ∧ F([LmRNA ≤ 0.4]) ∧ (F([LmRNA ≥ 2.2] ∧ F([LmRNA ≤ 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 analysis
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