System identification of intracellular molecular cascades based on Monte Carlo sampling

吉本 潤一郎 [1,2], 井上 智文 [1], 銅谷 賢治 [1,2]

[1] （独）沖縄科学技術研究基盤整備機構　　[2] 奈良先端科学技術大学院大学

{jun-y, tf-inoue, doya}@oist.jp

概要：代謝、細胞分化や神経細胞のシナプス可塑性などの生命現象を司る分子機構の計算原理を解明しようと、数理モデルやシミュレーションによる細胞内シグナル伝達系の解析が近年進められてきている。このようなモデル研究では、分子間の反応経路の仮説立てやシミュレーションモデルのパラメータ設定が解析結果の妥当性を主張する上で重要な要因となるが、分子生物学的知見のみだけでは決定しきれないものも少なくない。このため、限られた実験データから仮説モデルや設定パラメータの合理性を保証するような計算論的手法が求められている。本研究では、このような細胞内シグナル伝達系のパラメータ同定やモデル選択問題に対してベイズ推定法とモンテカルロサンプリングに基づく解法を提案し、ベンチマーク問題に適用することによってその有用性を示す。

1　はじめに

細胞内や細胞内小器官内のシグナル伝達系は、一定体積と温度を持つ均質な溶液中で生じる I 個の分子種間の化学反応としてモデル化できる。時刻 t における全分子種の濃度の集合を系の状態 $x(t) \in \Re^I$ で表すと、その動的挙動は、化学反応速度論より以下の形を持つ微分方程式として一般的に表現することができる。

$$\dot{x}(t) = f(t, x(t), u(t); q_1) \ \text{ s.t. } \ x(0) = q_0 \qquad (1)$$

ここで、$\dot{x} \equiv dx / dt$ である。$u(t)$ は系に対する外部入力であり、生化学実験では薬品や酵素注入などの操作量に対応する。f は化学反応式から決定できるある非線形関数である。$q_0 \in \Re^I$ は系の初期状態を与えるパラメータであり、$q_1 \in \Re^M$ は反応速度係数やMichaelis係数など系の挙動を決定づける M 個のパラメータの集合である。

外部入力 $U \equiv \{u(t) \mid t \in [0,T]\}$ の下で、系を構成する分子種の部分集合 $\mathcal{O} \subset \{1,\ldots,I\}$ を N 個の時点 $t_1,\ldots,t_N \in [0,T]$ で測定する生理実験によって、観測データ $Y \equiv \{y_i(t_n) \mid n = 1,\ldots N; i \in \mathcal{O}\}$ が得られたとしよう。また、$x_i(t_n;U,q)$ を外部入力 U とあるパラメータ $q \equiv (q_0, q_1)$ の値が与えられた時に、モデル(1)式によって再現される系の軌道とする。本研究の一つの目的は、データ Y をより良く再現する、すなわち、各時点での誤差

$$\varepsilon_i(t_n) = y_i(t_n) - x_i(t_n;U,q), \ n = 1,\ldots,N; \ \ i \in \mathcal{O}$$

を小さくできるパラメータ q の空間を明確にし、シミュレーションモデルの構築に役立てることである。また、対象とするシグナル伝達系を表現できる化学反応モデルが複数存在する時に、データ Y からどのモデルが尤もらしいかを推定することが本研究のもう一つの目的である。

2　手法

この 2 つの目的を統一的に扱うために、ベイズ推定法を適用する。誤差 $\varepsilon_i(t_n)$ が注目する細胞内部の微小領域でデータ測定したために生じたと仮定すると、$\varepsilon_i(t_n)$ は近似的に平均 0、分散 $x_i^q(t_n;U) / \gamma_i$ の正規分布に従う。ここで、$\gamma \equiv \{\gamma_i; i \in \mathcal{O}\}$ の各要素は正の値をとり、一般的に未知変量となる。この結果から、全ての未知変量の集合 $\theta \equiv (q,\gamma)$ の尤度 $p(Y \mid \theta;U)$ はガウス確率密度関数の積によって明確に定義できる。また、推定前の未知変量 θ の各値に対する確信度が事前分布 $p(\theta)$ で表現されていると、データ Y を得た後の未知変量 θ の事後分布がベイズの定理を用いて以下で与えられる。

$$p(\theta \mid Y;U) = p(Y \mid \theta;U)p(\theta) \,/\, p(Y \mid U) \qquad (2)$$

この事後分布をパラメータ q の推定結果とし、各値の良し悪しが定量的に評価する。また、(2)式の分母 $p(Y \mid U) \equiv \int p(Y \mid \theta;U)p(\theta)d\theta$ は周辺尤度と呼ばれ、ベイズ推定法では各モデルの尤もらしさを与える指標となる。よって、この周辺尤度最大化基準によってモデルの良し悪しを評価する。

(2)式や周辺尤度の評価には、困難な非線形関数の積分が伴うため近似計算が必要となる。本研究では、各 γ_i の事前分布が独立なガンマ分布で与えられているものとし、未知変量 θ の要素のうち、q を適応directionサンプリング法[1]によって、γ をギブスサンプリング法によってサンプリングを繰

り返すマルコフ連鎖モンテカルロアルゴリズムを実装し、これによって(2)式や周辺尤度を近似評価するものとした。

3 結果

提案手法のパラメータ推定能力を示すために、Kremling らがバイオリアクタシステムを想定して開発したベンチマーク問題[2]に適用した。この問題は、WWW 上で用意されているデータ集合 (Y,U) から反応速度係数や Michaelis 係数に対応する計 M=4個のパラメータ $q=(Y_s,k_2,k_{synmax},K_{IB})$ を推定する問題である。表 1 は、提案手法による推定結果の要約を示したものであり、合計 10^7 回のサンプリングによって近似された事後分布から求められる q の最良値（MAP 推定量）および 99%信用区間をベンチマーク問題で設定されている真の値と比較したものである。また、図 1(A)は観測データ点と MAP 推定量を用いたモデルから生成される系の軌道を比較したものであり、図 1(B)は二つのパラメータ空間 (k_{synmax},K_{IB}) 上での事後分布（各パラメータ値の良し悪し）を等高線図で示したものである。各データ点にノイズが含まれているため、MAP 推定量が真値と正確に一致することはあり得ないが信用区間を評価することによって、シミュレーションモデルとして妥当なパラメータ設定値の領域が定量化できていることが分かる。また、図 1(B)で示されるように、いくつかのパラメータ間に非線形的な相関関係があったとしてもそれを容易に見出せることが分かる。より大規模なベンチマークとして、計 M=36個の未知パラメータを推定する問題[3]を用いて、そこで最も精度が良いとされている進化アルゴリズムに基づく既存手法と比較したところ、提案手法はそれとほぼ同等か若干良い推定精度であった。

最後に、HIV proteinase のシグナル伝達経路モデル[4]を用いて提案手法のモデル比較能力を試した。この系に対する最も詳細なモデルでは、全ての化学反応が結合・解離反応として書かれ、計 9 個の分子種と 6 個の未知パラメータが存在するが、一部の伝達経路はある仮定の下で競合的抑制の酵素反応とみなせ、7 個の分子種と 5 個の未知パラメータが存在するモデルとして簡略化することもできる。この簡略化が妥当であるかどうかを各モデルごとに周辺尤度を計算し、その値と観測データの再現性を比較した。図 2 はその結果を示したものであり、点は観測データ点を、波線は MAP 推定量を用いて各モデルから生成される系の軌道である。いずれのモデルも観測データを良く再現できることが分かる。このような場合、図 2 の上部に示されるように簡略化モデルほど周辺尤度はより高い値となり、施した簡略化の妥当性に対する定量的な指標が与えられることが分かる。

表 1

パラメータ	真値	MAP	99% 信用区間	パラメータ	真値	MAP	99% 信用区間
Y_s ($\times 10^{-5}$)	7.00	7.00	[6.83,7.19]	k_2 ($\times 10^{+6}$)	6.00	6.22	[5.77,6.66]
k_{synmax} ($\times 10^{-2}$)	1.68	1.97	[1.46,3.97]	K_{IB} ($\times 10^{-2}$)	1.00	0.74	[0.28,1.28]

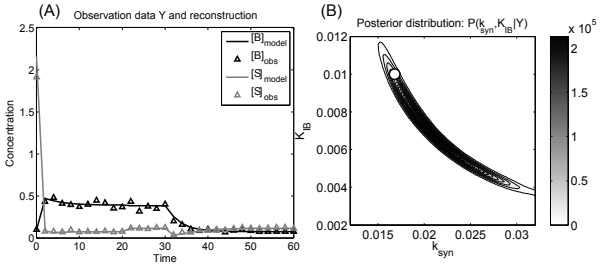

図 1

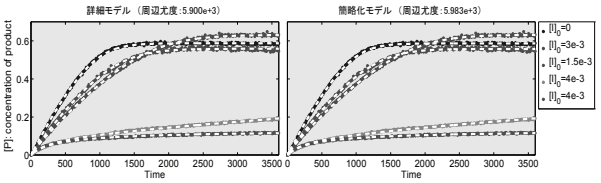

図 2

4 まとめ

本研究では、細胞内シグナル伝達系シミュレーションモデル構築のための計算論的手法を提案した。本手法の大きな特徴は、ベイズ推定法の適用により、モンテカルロサンプリングを用いてモデルパラメータの推定問題とモデル選択問題を統一的に扱うことができることである。問題点としては、事後分布や周辺尤度の近似精度を保つためには、大量のサンプリング回数が必要となり、計算時間もこの回数に比例して増大することである。しかしながら、適応 direction サンプリング法の適用により、計算ルーチンの大部分は並列化可能であるため、高性能並列計算機の性能向上に伴って計算時間の問題は将来的に取るに足りないものになると期待できる。なお、ここで提案したアルゴリズムは多くの研究者にとって容易に利用できるように GUI ソフトウェアとして WWW 上[5]で公開中である。

参考文献

[1] W.R. Gilks et al. "Adaptive direction sampling." *The Statistician* 43(1), 179–189, 1994.

[2] A. Kremling et al. "A benchmark for methods in reverse engineering and model discrimination: Problem formulation and solutions." *Genome Research* 14, 1773–1785, 2004.

[3] C.G. Moles et al. "Parameter estimation in biochemical pathways: A comparison of global optimization methods." *Genome Research* 13(11), 2467–2474, 2003.

[4] P. Kuzmic. "Program DYNAFIT for the analysis of enzyme kinetic data: application to HIV proteinase." *Analytical Biochemistry* 237(2), 260-273, 1996.

[5] http://www.nc.irp.oist.jp/software/