# ペタスケールコンピューティングに向けた超多数粒子フィルタの実装：遺伝子ネットワークモデルのパラメータ推定への適用

中村和幸[1]，吉田亮[1]，長崎正朗[2]，宮野悟[2]，樋口知之[1]

1 統計数理研究所, 2 東京大学医科学研究所

Contact：nakakazu@ism.ac.jp

**概要：データ同化手法の一種である粒子フィルタは並列性の高い手法である．その特性を生かし，超多数のサンプルを用いて遺伝子ネットワークモデルのパラメータ推定を行った．その結果，従来の知見よりも遥かに多くのパラメータについて，妥当な推定が可能であることを実験的に示した．同時に，アルゴリズムの検討と簡易的な時間見積もりを行い，ペタスケールコンピューティングにおいて，妥当な計算時間で結果が得られることを確認した．**

**キーワード：データ同化，粒子フィルタ，パラメータ推定**

## 1　はじめに

データ同化[1]とは，気象学・海洋学の分野において発展してきた手法であり，物理数値シミュレーションモデルに含まれる変数やパラメータを，物理モデルと観測データの両方をできるだけ満足しながら修正する手法である．数値シミュレーショモデルには，本質的に含まれる不確実性，たとえば，初期・境界条件の不確かさ，未知パラメータ，モデル化されないダイナミクスなどが存在し，必ずしも現実を反映していない．一方で，観測として得られるデータは，通常の場合，シミュレーションモデルから得られる情報よりも時空間解像度の意味ではるかに少ない情報しかもっていない．以上の数値シミュレーションモデル・観測の双方が抱える問題への対処として，観測データを用いたシミュレーション変数やパラメータの適切な推定を行うのがデータ同化である．これにより，地球科学の分野においては，予測精度の向上や物理量の推定による知識発見などがなされてきた．

システム生物学の分野においては，生化学反応ネットワークの数理モデルの構築が重要な問題となってきている．このネットワークについては，さまざまなモデリング・表現形式，アプローチが提案されてきている． Hybrid Functional Petri Net (HFPN)[2]は，これを実現する一つの形式である．ネットワークモデルの場合も，観測の情報不足と，モデルそのものの不完全性の両面が存在することから，データ同化の適用対象となり得て，従来以上の知見が得られる可能性がある．著者らのグループでは，HFPN 形式で表現されたモデルに対して，従来よりデータ同化を適用してきており，一定の成果を挙げてきた[3].

粒子フィルタ[4]は，データ同化にも適用可能なデータ解析手法である．非線形モデルを適切に扱うことができ，並列性が非常に高いという，統計理論とアルゴリズムの両方の観点から良い性質を持った手法であり，信号処理や時系列解析の分野においてその有効性が示されてきた．しかし，遺伝子ネットワークモデルのデータ同化の文脈では，近似精度の問題から適用が難しく，特に，非線形性が高い場合には，これまで高々数次元程度のパラメータ推定が限界であると見られてきた．

本発表においては，粒子フィルタの並列度の高さに注目し，粒子数を従来よりも遥かに多く使うというアプローチにより，この困難を乗り越えるアプローチをとった．以下ではその結果を示す[5].

## 2　解析対象ならびに手法

今回用いたネットワークモデルは，HFPN によって表現された circadian clock のモデル[6]である．本モデルに含まれるパラメータ数は，12 個の時変状態変数も含めて全部で 45 個である．また，時点数は 700 である．観測データは，4 つの mRNA に対応する 12 時点の状態変数に対応するデータを用いた．これらのデータは，Ueda et al. [7] による結果がもとになっている．今回，粒子フィルタについては， 500 万粒子を 1 ユニットとした粒子フィルタの結果を 20 ユニット合計する形で実装を行った．

## 3　解析結果

**図 1** は，初期状態パラメータについての，1 億粒子の場合と 10 万粒子の場合の推定事後分布のグラフである．粒子フィルタの場合，実現値の集合で分布が表現される．そのため，横軸方向にできるだけ多くの点で，確率値が推定されているほうが望ましい．1 億粒子の場合，適切に分布していることが確認できる．これにより，適切なパーセント点や推定値などを与えることが可能となっている．一方で，10 万粒子では，分布としての推定に失敗していることが確認できる．さらに，シミュレーション精度が向上したことも確認された．以上の結果は，Opteron 2220 2.8GHz を 1 Core (5 Gflops) 用いた計算結果である．計算時間は $6.8 \times 10^5$ 秒（約 8 日）である．

ペタスケールコンピュータにおける計算時間評価のために，アルゴリズムの検討も行った．以下，粒子数を N とすると，今回用いた粒子フィルタのアルゴリズムは，フィルタリングの最終段階を除いて，O(N) の時間複雑度を持つ．また，この部分については，互いの粒子について完全に独立に計算できるため，コア数に対してスケールする．一方，フィルタリングの最終段階においては，通常 O(N log N) の計算量を必要とする．以上の結果を踏まえて計算時間を評価すると，1 億粒子の計算を今回用いたモデルで行う場合，1 petaflops の計算機において，理想的には 90 秒程度で計算できることがわかる．これは，総計のための通信時間の評価が不足しているため，より精密な評価が必要であるが，現状の見積もりの範囲では，適切な範囲内であると考えられる．

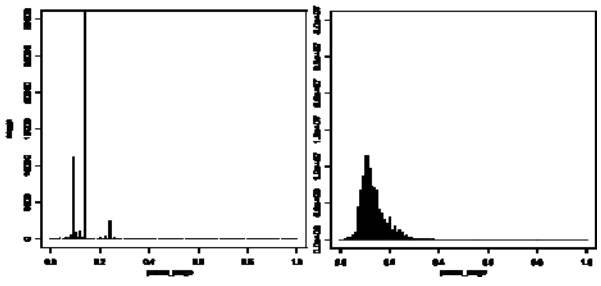

**図 1**．10 万粒子（左）と 1 億粒子（右）のそれぞれの場合の per の初期状態推定の結果．1 億粒子の方は適切に滑らかに変化しており，適切な分布の推定が得られているが，10 万粒子では滑らかに変化せず，分布としての評価に問題があることがわかる．

## 4　まとめ

粒子フィルタの並列性を生かし，1 億粒子による粒子フィルタを circadian clock model に適用した. その結果, 非線形性が強いモデルにおいて，従来は数次元程度が限界と考えられてきた推定可能パラメータが，45 次元までは適切に推定できることが示された．また，アルゴリズムの並列性の高さにより，計算時間も将来のペタスケールコンピューティングにおいては，適切な範囲であると見込まれる．今後は，大規模並列用のインプリメントを進め，同時に計算時間の正確な見積もりを進めていくことが課題となる．

**参考文献**


[1] Wunsch, C., The Ocean Circulation Inverse Problem, Cambridge University Press.

[2] Matsuno, H. et al., “A New Regulatory Interactions Suggested by Simulations for Circadian Genetic Control Mechanism in Mammals,” Journal of Bioinformatics and Computational Biology, Vol. 4,, 139-153, 2006.

[3] Nagasaki, M. et al., “Genomic Data Assimilation for Estimating Hybrid Functional Petri Net from Time-course Gene Expression Data,” Genome Informatics, Vol. 17, 46-61, 2006.

[4] Kitagawa, G., “Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State Space Model,” Journal of Computational and Graphical Statistics, Vol. 5, 1-25, 1996.

[5] Nakamura, K. et al., “Parameter Estimation of *in Silico* Biological Pathways with Particle Filtering Towards a Petascale Computing,” Proceedings of Pacific Symposium on Biocomputing 2009, to appear.

[6] Matsuno, H. et al., “A New Regulatory Interactions Suggested by Simulations for Circadian Genetic Control Mechanism in Mammals,” Journal of Bioinformatics and Computational Biology, Vol. 4, 139-153, 2006.

[7] Ueda, H.R. et al., “A Transcription Factor Response Element for Gene Expression During Circadian Night,” Nature, Vol. 481, 534-539, 2002.