平成 22 年度 RICC 利用報告書

課題名（タイトル）：

生命体シミュレーションのためのデータ同化研究

利用者氏名　　：○宮野悟，吉田亮，中野慎也，長尾大道，斎藤正也
所属　　：社会知創成事業次世代計算科学研究開発プログラムデータ解析融合研究開発チーム

1．　本課題の研究の背景、目的、関係するプロジェクトとの関係

申請者らは、次世代スーパーコンピュータ（以下、ペタコン）上で動作するソフトウェア開発プロジェクト「グランドチャレンジアプリケーション」の一環として、生命体データ同化ソフトウェア「LiSDAS」（Life Science Data Assimilation System）の開発を実施している。LiSDAS では、細胞内の生化学反応を記述した微分方程式群のシミュレーションを実施し、実験で得られた mRNA の観測データとデータ同化の手法を用いて照合しながら、モデルパラメータである Michaelis-Menten 係数の事後分布を推定する。また観測データをより良く説明するために、求められた事後分布を吟味しつつモデルの再構築を実施するための構造学習機能をプラグインする。

LiSDAS は、大規模生化学反応モデルをターゲットとしており、このような大量のモデルパラメータを許容範囲の時間内で推定するためには、RICC のような超多並列計算が必須となる。

2．　具体的な利用内容、計算方法

ターゲットとしている生命体モデルの時刻 t (t = 1, …,T)における状態を x_t、観測データを y_t とし、推定すべきパラメータをまとめて、超パラメータ θ とする。θ には生命体モデルに含まれる Michaelis-Menten 係数の他、初期状態 x_0 も含まれる。この生命体モデルに関する状態空間モデルはあるノイズ項 v_t、w_t によって次のように記述される。

$$x_t = f(x_{t-1}, v_t) \quad (1), \qquad y_t = h(x_{t-1}) + w_t \quad (2)$$

(1) 式はシミュレーションモデル、(2)はシミュレーション変数と観測データとの関係を表す。LiSDAS ではまず、予め設定された超パラメータに対する事前分布 $p(\theta)$ に従う粒子 $\theta^{(i)}$を、並列計算機の各コアにつき数個程度ずつサンプリングする。次に各 $\theta^{(i)}$に対して粒子フィルタを適用することにより、尤度 $p(y_{1:T}|\ \theta^{(i)})$を計算する。ただし、$y_{1:T}$は時刻 1 から Tまでの観測データ、すなわち全ての観測データを表す。そして事後分布 $p\left(\theta^{(i)} \mid y_{1:T}\right) \propto p\left(y_{1:T} \mid \theta^{(i)}\right) p(\theta)$ に対して MCMC 法を適用することにより、粒子が事後分布からのサンプリングとなるように十分収束させた後、各ノードに搭載されているメモリが許容する個数まで粒子数を増殖させる。これにより、最終目的となる各パラメータの事後分布を表現するための十分な数のサンプリングを得ることができる

3．　結果

以下では、哺乳類概日周期変動モデルをテストベットとする実験結果を紹介する。

まず、実装したアルゴリズムの動作を確認するために、MCMC の反復によって経験分布がどのように変化するかを調べた。図 1 は、一部の Michaelis-Menten 係数に対する分布で、事前分布(赤色)から出発して、500 ステップ以降では分布が収束しており、さらに事前分布よりも分散が小さくなっていることが観察される。

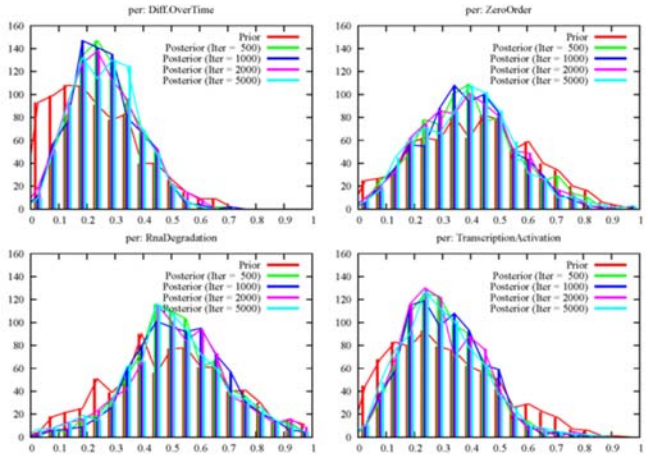

図 1 分布の収束のようす

他方、最尤推定値でのシミュレーション結果(図 2 の赤線)を見ると、観測データ(黒丸)に見られる周期性を捉えられていない。青線は通常の粒子フィルタによる最尤推定値でのシミュレーション結果である。これは粒子数 400 万の結果であるが、粒子数を 100 億まで増大してもやはり周期性を生み出すモデルパラメータは

発見できない。そのため、実装した並列 MCMC 法よりもむしろ、適切に事前分布を設計することが課題であると考えられる。

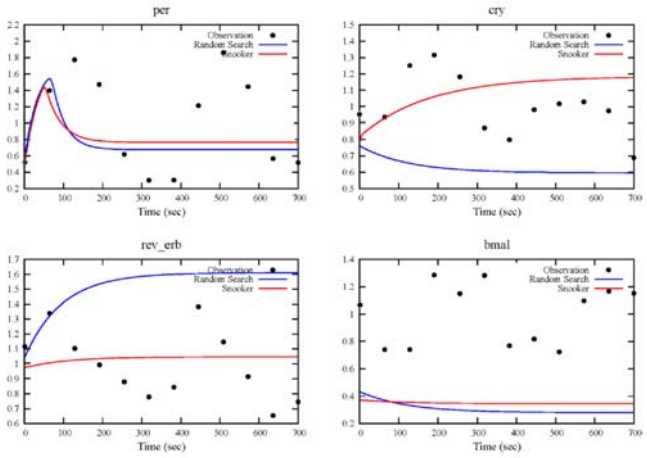

図 2 最尤推定値に対するシミュレーション

並列化性能の測定も行った。図 3 は粒子数(すなわち計算する問題のサイズ)を 1 億に固定した場合の、並列プロセス数と計算時間の関係である。プロセス数に対してほぼ線形に計算速度が上昇することが読み取れる。

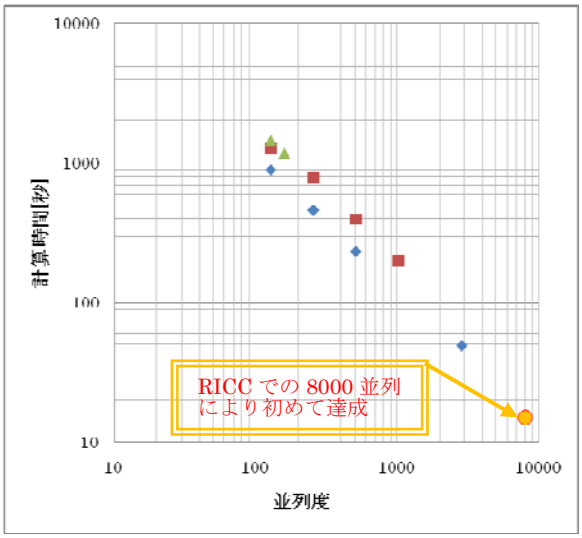

図 3.　1 億粒子を用いたパラメータ推定の計算時間

4．　まとめ

並列 MCMC 法に基づくデータアルゴリズムを実装し、性能評価を行った。事後分布への収束、並列度に線形な計算速度の向上の点では、望ましい結果を得たが、他方で標準的な事前分布ではパラメータをうまく推定できないということが明らかになった。

5．　今後の計画・展望

適切な事前分布の構成が必要である。観測時系列から事前分布を構成するアルゴリズムはすでに開発しているが、「京」のような超並列計算環境で運用するには、解決すべき課題がある。

6．　RICC の継続利用を希望の場合は、これまで利用した状況（どの程度研究が進んだか、研究においてどこまで計算出来て、何が出来ていないか）や、継続して利用する際に行う具体的な内容

今回の利用では、8000 並列までの速度向上を達成しているが、プロセス並列のみを用いている。「京」のアーキテクチャではメモリ消費量の観点から、スレッド並列が推奨されているため、スレッド並列とプロセス並列の混合化への対応に取り組む予定である。

7．　一般利用で演算時間を使い切れなかった理由

ここで示した結果は、5 月のメンテナンス時の 8000 並列計算キャンペーンで得たものである。上述の課題を解決するためのソフトウェア開発に注力したため、その後の計算は、テスト計算に留まった。

平成 22 年度　RICC 利用研究成果リスト

【論文、学会報告・雑誌などの論文発表】

Yoshida, R., Saito, M.M., Nagao, H., Higuchi, T.: Bayesian experts in exploring reaction kinetics of transcription circuits. Bioinformatics 26 (18), i589-595 (2010)

【国際会議などの予稿集、proceeding】

Nagao, H., Higuchi, T.: Web application for time-series analysis based on particle filter available on cloud computing system. Proceedings of 13th International Conference on Information Fusion, Edinburgh, Scotland, July 26-29, 2010.

Hayashi. K, Saito, M.M., Yoshida R., Higuchi, T.: Implementation of sequential importance sampling in GPGPU. Proceedings of 13th International Conference on Information Fusion, Edinburgh, Scotland, July 26-29, 2010.

Nakano, S., Higuchi, T.: A dynamic grouping strategy for implementation of the particle filter on a massively parallel computer. Proceedings of 13th International Conference on Information Fusion, Edinburgh, Scotland, July 26-29, 2010.

Nakano, S.: Population-based quasi-Bayesian algorithm for high-dimensional sequential problems and hierarchization of it for distributed computing environments. 2010 IEEE World Congress on Computational Intelligence, Barcelona, Spain, Jul. 2010. (talk)

【国際会議、学会などでの口頭発表】

Nagao, H, Nakano, S., Higuchi, T.:Particle filter for time series modeling and its application to geophysical data. Heliophysics Science Division Seminar, Goddard Space Flight Center, NASA, USA, Feb. 24, 2010.

Nagao, H., Higuchi, T.: CloCK-TiME: Cloud Computing Kernel for Time-Series Modeling Engine. 3rd International Conference of the ERCIM WG on Computing & Statistics, London, UK, Dec. 10, 2010.

長尾大道，吉田亮，斎藤正也，樋口知之: Large-scale hybrid method of particle filter and MCMC for estimating parameters in biological pathway models. 2nd バイオスーパーコンピューティング・シンポジウム，東京（MY PLAZA ホール), 2010 年 3 月 18～19 日（poster）

吉田亮，長尾大道，斎藤正也，中野慎也，長崎正朗，山口類，井元清哉，山内麻衣，後藤典子，宮野悟，樋口知之: LiSDAS: Life science data assimilation systems. 2nd バイオスーパーコンピューティング・シンポジウム，東京（MY PLAZA ホール), 2010 年 3 月 18～19 日（poster）

中野慎也，樋口知之: 大規模並列計算機における高次元非線型システムの状態推定について. 2010 年度 統計関連学会連合大会，東京, 2010 年 9 月 5～8 日.

長尾大道*、吉田亮、斎藤正也、樋口知之、長崎正朗、井元清哉、山口類、宮野悟、山内麻衣、後藤典子: 遺伝子発現時系列データおよびバイオデータベース情報を基にした活性/抑制型転写因子の同定 2010 年度統計関連学会連合大会，東京（早稲田大学), 2010 年 9 月 5～8 日.