課題名（タイトル）:

大規模遺伝子ネットワーク推定プログラムの研究開発

利用者氏名　　:○宮野悟，玉田嘉紀
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

現在，理化学研究所が中心になって開発している次世代スーパーコンピュータ「京」を利活用するためのプログラム「次世代計算科学研究開発プログラム」において，筆者らはデータ解析融合チームの一員として研究開発課題として掲げている大規模遺伝子ネットワーク推定プログラムの研究開発を行っている．

筆者らが開発研究している大規模遺伝子ネットワーク推定プログラム SiGN は，ノンパラメトリック回帰によるベイジアンネットワーク，状態空間モデル，グラフィカルガウシアンモデル，ベクトル自己回帰モデルを遺伝子ネットワークモデルとして利用し，遺伝子発現データなどから遺伝子発現の依存関係を表す遺伝子ネットワークを推定するためのものである．それぞれのモデルのパラメータ推定及びネットワークの構造推定は非常に計算に時間がかかるため，スーパーコンピュータを利用した大規模計算が欠かせない．高精度・大規模な遺伝子ネットワークの推定が可能になることにより細胞内での遺伝子の発現の依存関係を予測することが可能になり，薬剤作用機序解明や新規薬剤標的遺伝子の同定などが可能になることが期待される．

２．　具体的な利用内容、計算方法

本研究では，RICC を利用し，申請者らが開発している大規模遺伝子ネットワークプログラムの高並列化を実現する．前述の 4 種類のネットワークモデルのうち，今年度はノンパラメトリック回帰によるベイジアンネットワークおよび状態空間モデル（state space model: SSM）による遺伝子ネットワーク推定ソフトウェアの高並列化を行った．特に SSM による遺伝子ネットワーク推定ソフトウェアの高並列化は今年度の RICC 利用の計画に挙げていなかったが，著者らの共同研究の都合上，予定を変更し今年度後半に主に取り組んだ．SSM による遺伝子ネットワーク推定ソフトウェアはこれまで SiGN とは別のソフトウェアであったが，入出力などを統一，SiGN-SSM と命名し SiGN の一部とした．SiGN-SSM はこれまで MPI 等による並列化がなされていなかったため，まずそれを RICC 上で行うことにした．また「京」での利用を見据えたコード変更，具体的には富士通製 C コンパイラへの対応および現世代機 FX1 への対応を RICC 上で行ことにした．今年度は 256 並列程度までの動作を確認し，さらなる高並列のための課題を見いだすことを目標とした．ノンパラメトリック回帰ベイジアンネットワークを用いた遺伝子ネットワーク推定ソフトウェアは新たに SiGN-BN と呼ぶことにした．SiGN-BN に関してはベイジアンネットワーク構造推定のための様々なアルゴリズムを実装しているがそのうち「発見的アルゴリズム」，および「全ゲノムサイズ遺伝子ネットワーク推定アルゴリズム」の 8000 並列以上の高並列化を RICC 上で実現し動作確認をすることにした．

次に各モデルでの計算方法を述べる．

ベイジアンネットワークによるネットワークのスコアは次のように計算される．ネットワーク G 上のノード X_i の直接の親集合を $Pa(X_i)$ と表す．このとき,マイクロアレイデータ D が得られた元でのネットワークの事後確率 $\Pr(G|D)$ に基づくスコア $S(G) = -2\log \Pr(G|D)$ に対して,分解 $S(G) = \mathrm{S}_i\, s(X_i, Pa(X_i))$ を得る．ここで、$s(X_i, Pa(X_i))$ は、ノード X_i とその直接の親集合 $Pa(X_i)$ により定義されるサブネットワークのスコアである．この $S(G)$ を最小にする非閉路有向グラフ（DAG）を求める問題がベイジアンネットワークの構造推定と呼ばれる．各局所スコアは,

$$s(X_i, Pa(X_i)) = -2\log \int \prod_n f_i(x_{nj} \mid pa(x_i)_n, \theta_i) p(\theta_i \mid \lambda_i) d\theta_i$$

の高次積分を Laplace 近似を用いることにより計算される．ここで f_i は，X_i に対するノンパラメトリック回帰に基づく確率モデル，$p(q_i \mid l_i)$ はパラメータ q_i に対する事前確率分布の密度関数である．

最適な（スコアの最小な）ベイジアンネットワーク構造を探索するには, 解候補となる DAG 構造が膨大にあるため，ノード数（遺伝子数）が多い最適なネットワークを探索することは非常に困難な問題である．我々の研究グループでは探索したいネットワークのサイズに応じて様々なアルゴリズムを開発している．1000 遺伝子前後の大規模遺伝子ネットワークをベイジアンネットワークを用いて推定する場合は「発見的アルゴリズム(HC)」を用いる．この際，推定されるネットワークの信頼性を確保するために，データセットからリサンプリングを行い再構成したデータで繰り返しネットワーク推定を行い，推定されるモデルの出現頻度を計算するブートストラップ計算を行う．ブートストラップ計算はデータ並列性があるため，１回のネットワーク推定を１つのコアが担当する．この方法によるベイジアンネットワークの構造探索は１回の計算時間が不均一になるため，並列化は単純に計算を分割すると効率が悪い可能性がある．従って１つのプロセスが専属的にジョブを他のプロセスに割り当てる仕事を担当する．この方法を MPI により実装し，昨年度までに 4096 並列まで高効率で並列動作することを確認している．

ヒトの全ゲノムを含む遺伝子ネットワークを推定するためのアルゴリズムとして昨年度,「全ゲノムサイズ探索アルゴリズム（NNSR)」を開発した．これは筆者が発明した Neighbor Node Sampling 法を用いて抽出された遺伝子の部分集合に対して繰り返し HC アルゴリズムを適用することにより，超大規模並なベイジアンネットワークの構造探索の並列計算を実現した物である．このアルゴリズムの高並列化対応を，RICC を用いて行う．

SSM は時系列データ解析に用いられる統計モデルで, 時点毎に計測された遺伝子発現プロファイルに対して適用することにより動的遺伝子ネットワークモデルの予測に応用可能である．今 p 個の遺伝子からなる遺伝子ネットワークを考え，y_n を時点 n における p 個の遺伝子の発現量を表した p 次元のベクトルとする．SSM では y_n を $k(<<p)$ 次元の隠れ変数 x_n より生成されると仮定する. すなわち SSM は以下の 2 つの式により定義される.

$$x_n = Fx_{n-1} + v_n, \quad v_n \sim N(0,Q)$$
$$y_n = Hx_n + w_n, \quad w_n \sim N(0,R)$$

ここで F, H はそれぞれ状態遷移行列，観測行列といい上の式をシステムモデル，下の式を観測モデルという．v_n, w_n はそれぞれ $N(0,Q), N(0,R)$ に従うシステムノイズ，及び観測ノイズである．従って，SSM の推定は初期値 $x_0 \sim N(\mu_0, \Sigma_0)$ およびパラメータ F, H, Q, R, μ_0 を観測データより推定することが問題の本質である．これらのパラメータの推定は期待値最大化（EM）法により予測可能である．また k の大きさはベイズ型情報量規準（BIC）などの比較により決定可能である.

EM 法による推定では局所解しか得られないため, 繰り返し異なる初期値から推定を行い，最も BIC の良い推定結果を採用する．この際の繰り返し計算は並列化が可能であるため MPI を用いた並列計算アルゴリズムの開発を RICC 上で行う.

３． 結果

まずベイジアンネットワークにおける結果を述べる.

HC アルゴリズムとブートストラップ法を組み合わせた方法（HC + Bootstrap)を 8 並列から 8192 並列で実行しその実行時間を測定した．実行結果を図１及び表１に示す．8192 並列の実行は，フラット MPI による実装ではメモリを大量に消費することが分かり１プロセスあたりおよそ 300MB ほど使用量を抑えることにより高並列実行に成功した．表中の n_p，time, speedup, efficiency はそれぞれ並列数, 実行時間(秒)，高速化率および並列化効率である．使用したデータは人工的に発生した 200 遺伝子 200 サンプルのシミュレーションデータでブートストラップの繰り返し回数は 20,000 回である．またそれぞれ 8 並列との結果の比較を行っている．従って，低並列では並列化効率(efficiency)は 1.0 を超える．4096 並列では高い並列化効率（0.941)を達成したが，8192 並列では並列化効率は 0.7 まで落ちた. これは主に並列化粒度が問題であり，１回のブートストラップ計算が比較的長く，逐次アルゴリズムであるため並列化できないためである．従っ

て，問題サイズ・繰り返し回数によって大きく結果が異なる．現在のところこれ以上の高効率実行は難しく，HC + Bootstrap 法の高効率化は来年度以降の課題とする．

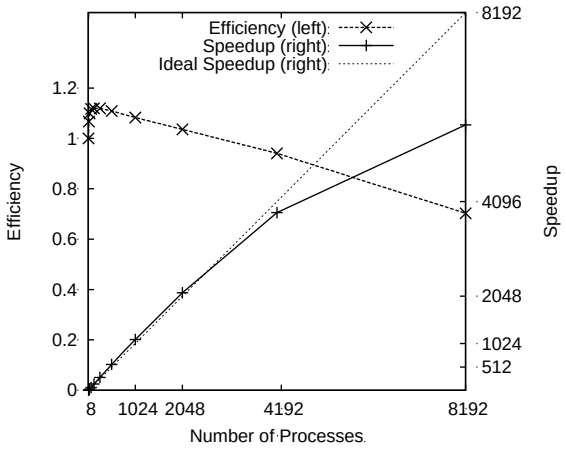

図１．HC + Bootstrap 台数効果

表１．HC + Bootstrap 台数効果

n_p	time	speedup	efficiency
8	221738.630	8.000	1.000
16	103919.283	17.070	1.067
32	50400.290	35.196	1.100
64	24811.126	71.497	1.117
128	12373.078	143.368	1.120
256	6189.733	286.589	1.119
512	3123.985	567.835	1.109
1024	1599.437	1109.083	1.083
2048	835.980	2121.951	1.036
4096	460.034	3856.040	0.941
8192	308.099	5757.588	0.703

次に全ゲノムサイズ対応の NNSR アルゴリズムの高並列実行の確認及び実行時間の計測を行った．NNSR 法は HC + Bootstrap 法とは異なり定期的な全対全通信が必要である．図２及び表２に結果を示す．8192 並列での動作は確認したが，その際の並列化効率は 0.52 と非常に低い値になった．ハイブリッド並列化および局所的な通信のみを用いるアルゴリズムなどにより高効率化を行うことが来年度以降の課題である．

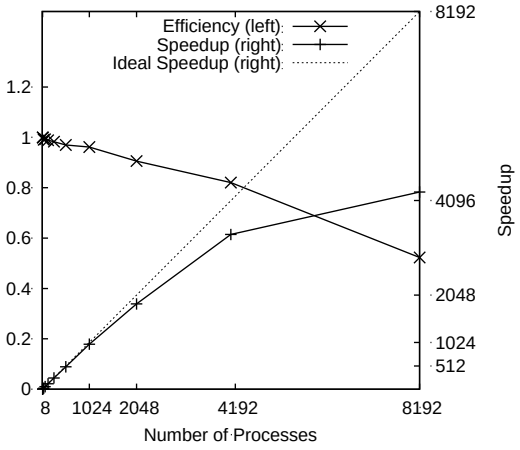

図２．NNSR 法 台数効果

表２．NNSR 法 台数効果

n_p	time	speedup	efficiency
8	166857.166	8.000	1.000
16	83485.883	15.989	0.999
32	42102.183	31.705	0.991
64	21005.062	63.549	0.993
128	10549.508	126.533	0.989
256	5305.756	251.587	0.983
512	2688.092	496.582	0.970
1024	1354.922	985.191	0.962
2048	719.525	1855.192	0.906
4096	397.033	3362.082	0.821
8192	311.721	4282.222	0.523

次に SiGN-SSM のコード化（富士通コンパイラ対応化）および並列化を行った．ベイジアンネットワークにおける HC + Bootstrap 法と同様に EM 法自体は逐次アルゴリズムであり並列化ができないが繰り返し計算を行う．従って同様に１コアがジョブ分配を行い他のコアがパラメータの推定を行うプログラムを MPI を用いて実装した．SiGN-SSM はオープンソースソフトウェアとして http://sign.hgc.jp/において現在公開中である．今年度の目標である 256 並列までの実行を確認し，台数効果を測定した．図３及び表３に結果を示す．これまでと違い，シングル版による実行結果と比較を行った．従って，MPI 版実行時，1 コア分計算を行わないため低並列時は効率が悪化する．256 並列実行時の並列化効率は 0.98 となり問題なく高効率動作することが分かった．来年度は 8000 並列以上の高並列実行を行ない高並列実行時の問題点を見いだし必要であれば改

良を加えたい．また SiGN-SSM は SPARC64 VII を搭載した，いわゆる現世代機 FX1 でのコンパイル&動作を確認している．

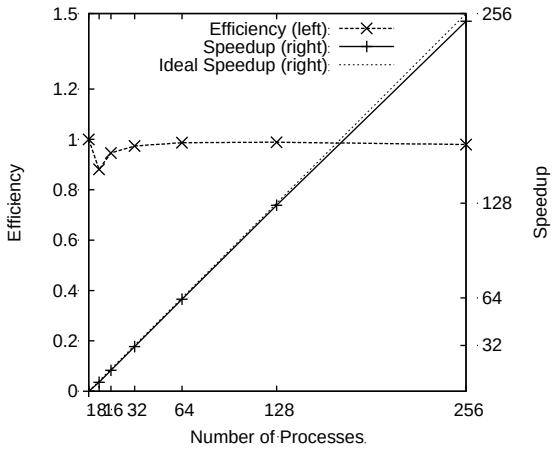

図３．SSM 台数効果

表３．SSM 台数効果

n_p	time	speedup	efficiency
1	1069310.208	1.000	1.000
8	151703.467	7.049	0.881
16	70648.527	15.136	0.946
32	34292.279	31.182	0.974
64	16930.099	63.160	0.987
128	8446.307	126.601	0.989
256	4263.709	250.793	0.980

４． まとめ

次世代スーパーコンピュータ「京」のためのソフトウェア SiGN-BN および SiGN-SSM の高並列化対応を，RICC を用いて行った．主要なアルゴリズムにおいて 8192 並列での動作を確認し，来年度以降に行うべき改良点を見いだすことができた．

５． 今後の計画・展望

今年度はこれまで開発したアルゴリズムの 8192 並列での動作確認および新規アルゴリズムの並列化が中心であった．来年度以降，これらのアルゴリズムのチューニング期間として高効率化をハイブリッド並列化やアルゴリズムの改良などによって実現する．

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

本研究の最終的な目標は次世代スーパーコンピュータ「京」を用いて最大 640,000 コアでの超高並列・超高効率実行可能なソフトウェアの研究開発である．本年度の計画は，主要なアルゴリズムで RICC 上限である 8192 並列での実行であり，その目標は達成することができた．本プロジェクトは残り２年となり，この期間は高性能化へのチューニングの期間となる．従って，来年度以降，未着手のハイブリッド並列化やアルゴリズムの改良などによって高効率動作可能なプログラムの実現を目指す．特に全ゲノム対応ベイジアンネットワークの構造探索アルゴリズムは本プロジェクトの柱となるものであり，ハイブリッド並列化に加え全対全通信を抑えるなどのアルゴリズム上の工夫などにより高並列実行時の高効率化を目指す．

７． 一般利用で演算時間を使い切れなかった理由

他の研究プロジェクトとの関連から２年後に予定していたドキュメント化を今年度に一部先行して行った．また SSM を用いた遺伝子ネットワーク推定ソフトウェアの RICC での並列化に関しては，完全に計画外であったが，コード化およびオープンソース化とし公開するためのドキュメント化，パッケージ化および論文化の作業が発生したため，これに半年を費やした．従って，既存アルゴリズムによる 8000 並列による動作確認が中心となり，当初計画していた「全ゲノムサイズ遺伝子ネットワークプログラム」および「並列最適ネットワーク探索プログラム」の高効率化には着手しなかった．

また，8000 並列による高並列実行確認ではメモリ不足による実行失敗が何回か発生したが，それ以外では特に大きな問題なく，当初目標としていた 8000 並列を達成できたため，予備的に確保していた計算時間を消費する必要が無かった．

以上の２つの理由により申請した演算時間が余ることになった．今年度未着手の分に関しては来年度以降に取り組むことにした．

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】

Tamada, Y., Imoto, S., and Miyano, S. (2011). Parallel algorithm for learning optimal Bayesian network structure, submitted.

Tamada, Y., Yamagushi, R., Imoto, S., Hirose, O., Yoshida, R., Nagasaki, M., and Miyano, S. (2011). SiGN-SSM: open source parallel software for estimating gene networks with state space models, *Bioinformatics*, in press.

【国際会議、学会などでの口頭発表】

玉田 嘉紀, ハイパフォーマンスコンピューティングによる大規模遺伝子ネットワークの推定とその応用, 日本バイオインフォマティクス学会 第１回 応用システムバイオロジー研究会 @ 東京医科歯科大学（東京都文京区）(Feb. 28, 2011). 口頭発表.

玉田 嘉紀, 島村 徹平, 山口 類, 長崎 正朗, 井元 清哉, 宮野 悟, スーパーコンピューティングによる大規模遺伝子ネットワーク推定ソフトウェア SiGN, 第３回 バイオスーパーコンピューティングシンポジウム @ 理化学研究所計算科学研究機構および甲南大学ポートアイランドキャンパス（兵庫県神戸市）(Feb. 21-22, 2011). ポスター発表.

玉田 嘉紀, 遺伝子制御システムを解き明かす大規模遺伝子ネットワーク推定ソフトウェアの研究開発, 第３回 バイオスーパーコンピューティングシンポジウム @ 理化学研究所計算科学研究機構および甲南大学ポートアイランドキャンパス（兵庫県神戸市）(Feb. 21-22, 2011). 口頭発表.

Tamada, Y, Shimamura, T., Yamaguchi, R., Imoto, S., and Miyano, S., SiGN: Large-scale gene network estimation environment for high performance computing, The 2010 Annual Conference of the Japanese Society for Bioinformatics (JSBi2010, 2010 年日本バイオインフォマティクス学会年会) @ 九州大学医学部百年講堂（福岡県福岡市）(Dec. 13-15, 2010). ポスター＆口頭発表.

玉田 嘉紀, 非線形回帰によるベイジアンネットワークを用いた遺伝子発現データからの遺伝子ネットワーク推定とその並列アルゴリズム, 生命体統合シミュレーションサマースクール 2010 @ 湘南国際村センター（神奈川県三浦郡葉山町）(July 5, 2010). ポスター＆口頭発表.