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
理研での所属研究室名：
社会知創成事業 次世代計算科学研究開発プログラム
次世代生命体統合シミュレーション研究推進グループ　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．そのうちベイジアンネットワークを利用したものを SiGN-BN，状態空間モデルは SiGN-SSM，グラフィカルガウシアンモデル及びベクトル自己回帰モデルを利用したものを SiGN-L1 と呼んでいる．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（SiGN-BN）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．昨年までの利用において特に SiGN-SSM の MPI を用いた並列化や他の SiGN ソフトウェアとの入出力の共通化，また「京」での利用を見据えたコード変更等を富士通製 C コンパイラへの対応および現世代機 FX1 への対応通して RICC 上で行った．昨年までの範囲では 256 並列程度までの動作を確認していたが，今年度は，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．SiGN-BN に関してはスレッド並列とプロセス並列を組み合わせたハイブリッド並列化のテスト実行を行うこととした．

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 8192 並列ま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「全ゲノムサイズ探索アルゴリズム（NNSR)」を開発している．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

SiGN-BN のハイブリッド並列化はスレッド並列化に関しては SSL(BLAS/LAPACK)のスレッド並列ライブラリを利用し，プロセス並列化は現状を維持し，最小限のコストで対応することにした．RICC において SiGN-BN NNSR アルゴリズムの 32 ノード（プロセス）×8 スレッド=256 並列での動作を確認した．これ以上の並列度やパフォーマンステストは「京」試験利用が始まっているため RICC では行わなかった．

次に SiGN-SSM での高並列実行のテストを行った．結果, RICC において 8192 並列までの動作を確認した．台数効果を調べるために逐次での実行と 8 並列から 8192 並列までの実行時間を比較した(図１および表１)．使用したデータは 100 遺伝子 7 時点 3 回計測（計 21 アレイ）のデータで，4 次元から 8 次元まで EM アルゴリズムを 1000 回ずつそれぞれ 10 セットパラメータを出力する計 100,000 ジョブの設定で計算を行った．128 並列まで並列化効率(efficiency)が良くなるのは，1 プロセスをジョブの割り当て専用に使用しているため，全体で CPU の利用効率が高くなるためである．しかし，それ以降では急速に悪くなっていることが分かった．SiGN-SSM では実行時間などは使用するデータにかなり依存するため，あくまで今回使用したデータに対してのみに言えることだが，京での高並列実行には課題があることが分かった．ハイブリッド並列化はまだ適用していないため，それによりプロセス数を減らすなどによって改善が期待できるが，それ以上の工夫が必

要なことが分かった.

図 1．SiGN-SSM 台数効果

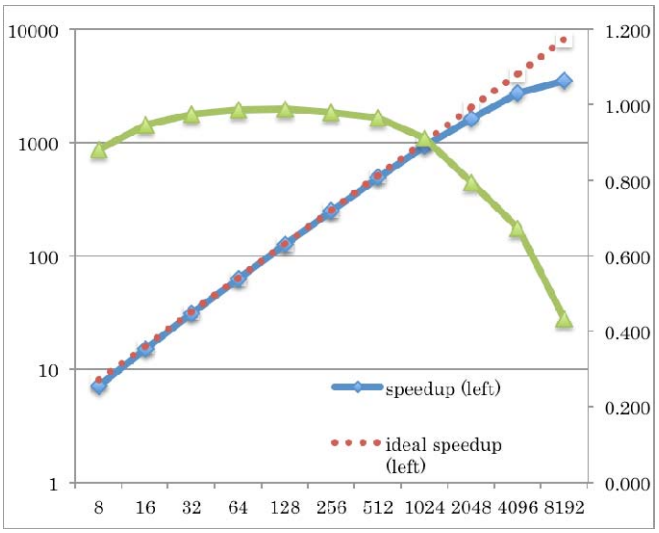

表 1．SiGN-SSM　台数効果

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
512	2161.541	494.698	0.966
1024	1146.566	932.620	0.911
2048	656.661	1628.406	0.795
4096	387.652	2758.429	0.673
8192	300.698	3556.095	0.434

4．　まとめ

「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．SiGN-SSM 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・問題点の抽出が中心であり目標としていたチューニングやアルゴリズムの改良にはあまり踏み込めなかった．来年度以降，引き続きこれらのアルゴリズムのチューニングや高効率並列化を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．またハイブリッド並列化などへも着手することができた．本プロジェクトは残り 1 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，アルゴリズムの改良などによってより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.

平成 23 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., Shimamura, T., Yamaguchi, R., Imoto, S., Nagasaki, M., and Miyano, S., SiGN: Large-scale gene network estimation environment for high performance computing, *Genome Informatics*, **25** (1), 40-52, 2011.

【国際会議、学会などでの口頭発表】

玉田 嘉紀, スーパーコンピュータによる大規模遺伝子ネットワーク推定, 情報処理学会 第 74 回全国大会 @ 名古屋工業大学御器所キャンパス（愛知県名古屋市） (Mar. 8, 2012). 口頭発表.

玉田 嘉紀, 島村 徹平, 山口 類, 新井田 厚司, 斉藤 あゆむ, 長崎 正朗, 井元 清哉, 宮野 悟, SiGN-BN: ベイジアンネットワークによる大規模遺伝子ネットワーク推定プログラム, 文部科学省「革新的ハイパフォーマンス・コンピューティング・インフラ（HPCI）の構築」・次世代ナノ統合シミュレーションソフトウェアの研究開発（ナノ）・次世代生命体統合シミュレーションソフトウェアの研究開発（ライフ）公開シンポジウム @ニチイ学館　神戸ポートアイランドセンター（兵庫県神戸市） (Mar. 5-6, 2012). ポスター発表.

玉田 嘉紀, 島村 徹平, 山口 類, 新井田 厚司, 斉藤 あゆむ, 長崎 正朗, 井元 清哉, 宮野 悟, SiGN-BN: ベイジアンネットワークによる大規模遺伝子ネットワーク推定プログラム, ISLiM 成果報告会 2011 @ 東京大学武田ホール（東京都文京区） (Feb. 21-22, 2011). ポスター発表.

玉田 嘉紀, 島村 徹平, 山口 類, 長崎 正朗, 井元 清哉, 宮野 悟,大規模遺伝子ネットワーク推定ソフトウェア SiGN, バイオスーパーコンピューティングサマースクール 2011 @ 淡路夢舞台国際会議場（兵庫県淡路市） (Sep. 26-27, 2011). ポスター掲示.