平成23年度における研究成果

Ⅳ　戦略課題 4：大規模生命データ解析　［統括：宮野 悟（東京大学）］

特定高速電子計算機施設を中核とする HPCI に最適化した最先端・大規模シークエンスデータ解析基盤を整備した上で、生命プログラムの複雑性・多様性や進化をゲノムによって理解する研究と同時に、ゲノムを基軸とした生体分子ネットワーク解析研究を行う。それにより、薬効・副作用予測、毒性の原因の推定、オーダーメイド投薬、予後予測などへの応用に貢献することを目指す。

Ⅳ－１　宮野 悟（東京大学）
研究開発の統括

平成 23 年度の統括業務では、独立行政法人理化学研究所が実施している「HPCI 戦略プログラム分野 1　予測する生命科学・医療および創薬基盤(以下、「戦略プログラム」という)」における研究開発を推進するため、主に、戦略プログラムの一環である「課題 4 大規模生命データ解析」の研究開発を統括した。この研究開発を行う上で、関連する研究者と必要な協議等を行うとともに、学術調査を行い、本格実施に必要な研究体制の整備を行った。

以下の大学等で実施された平成 23 年度の研究課題の実施項目①～④について、大規模生命データ解析ワークショップや研究打合せを行うことにより、関係者のとりまとめを行うとともに、グループリーダー会議などにおいて理化学研究所と連携して、研究開発の統括を行った。

① 次世代シークエンサーデータ解析のための情報処理システムの開発(秋山泰・東京工業大学)
② RNA 相互作用予測技術の開発と転写物の網羅的情報解析(浅井潔・産業技術総合研究所)
③ 大規模な生体分子ネットワークの解析技術の開発(松田秀雄・大阪大学)
④ メタゲノム・比較ゲノム解析研究(五條堀孝・国立遺伝学研究所)

具体的には、以下の 1～3 を実施した。

１．大規模生命データ解析ワークショップの開催

研究の進捗を報告し、連携を密にすることを目的としてワークショップを開催した。第 1 回は、2011 年 9 月 3 日、第 2 回は 2012 年 3 月 11 日に両日とも東京国際フォーラム G504 において開催した。

２．学術調査

国際がんゲノムコンソーシアム学術ワークショップ（2012 年 3 月 21 日～22 日、フランス・カンヌ）に出席し、次世代シークエンサーによるがんゲノム解析の研究展開の状況とその大規模データ解析の手法について調査を行った。その結果、新たなシークエンス装置、データマネージメント、大規模生命データの統合的理解の方法、世界の研究のスピードなど、①及び③の研究を展開するための有用な知見を得た。また、2011 年 9 月 16 日に、兵庫県立大学・姫路書写キャンパスで開催される生物物理学会シンポジウム「高速計算機シミュレーションによる生体機能解析へのアプローチ」において、HPCI 戦略分野 1 の戦略課題 4（大規模生命データ解析）に関する講演を行い、本課題の進め方について調査を行った。

３．研究協議と調整

まず、①の研究を優先的に進めるために、「京」のリソースの調整をおこなった。また、がんゲノムの次世代シークエンサーデータを提供した。②の研究における「京」の必要性について検討す

るとともに、①の開発との相互協力の調整を行った。③の開発のために、グランドチャレンジで開発しているソフトウェア SiGN の利用を推し進め、その整合性の調整をグランドチャレンジとの間で行った。④の研究で平成 23 年度に取り組むべき内容について検討し、課題の選定を行った。また、①で開発しているソフトウェアを利用するための調整を試みた。

2012 年 12 月 3 日～4 日に理化学研究所和光研究所・鈴木梅太郎ホールで開催された研究成果報告会におけるグループの研究成果をまとめ、発表内容の調整を行った。また、2012 年 1 月 10 日に計算科学研究機構で開催された外部諮問委員会に出席し、研究の進捗と今後の計画を説明し、評価・助言を受け、その助言をグループへ反映することを検討した。毎月開催されている HPCI 戦略プログラム分野 1 運営委員会に出席し、戦略プログラム内での調整を行った。

Ⅳ－２　秋山 泰（東京工業大学）

次世代シーケンサデータ解析のための情報処理システムの開発

Ⅳ－２－１　実施計画

「大規模生命データ解析」では、ゲノムを基軸とした大規模生命データ解析により生命プログラムとその多様性を理解することを目標としている。本研究では、これを実現するために最も重要な基盤となる次世代シーケンサから産出される大量のゲノム配列情報の超高速解析を実現するための研究開発を実施する。また、「次世代シーケンサデータ解析のための情報処理システムの開発」の研究を行う上で、関連する研究者と必要な協議等を行うとともに、本格実施に必要な研究体制の整備を行う。

平成 23 年度は、ゲノム解析に必要な定型的な処理のうち、特にシーケンサからのリード配列の相同性解析、およびクラスタリング処理を対象として、自動パイプラインの開発を開始する。自動パイプラインでは、複数の処理工程から構成されるワークフローに対して、その時点で準備できるノード数およびコア数を考慮して、適切な分割数に入力データを自動分割する機能、並列分散実行した結果を再統合する機能を実現する。また、パイプラインへのジョブの投入および実行状態モニタなどを簡便に行うためのインタフェースを開発する。システムの性能は実際のゲノム解析またはメタゲノム解析の実データを使って評価する。

Ⅳ－２－２　実施内容（成果）

（１）ソフトウェアの開発・高度化の状況

１）次世代シーケンサデータ相同性解析ツール GHOST-MP の開発

平成 23 年度は、次世代シーケンサから産出される大量のゲノム配列情報の超高速解析を実現するため、得られるノード数およびコア数を考慮して処理を自動分割し、並列分散実行する自動パイプラインの開発を開始した（図 1)。自動パイプラインのワークフローの中で多くの実行時間を占めるシーケンサからのリード配列の相同性解析については、クエリ配列およびデータベース配列の双方に接尾辞配列を用いた相同配列検索ツール GHOSTX、および GHOSTX を MPI/ OpenMP ハイブリッド並列化した GHOST-MP を開発した。GHOSTX は接尾辞配列を比較し、複数の相同配列候補を同時に探索することで高速な探索を実現する。GHOST-MP は MPI のレベルでクエリおよびデータベースを分割して処理を行い、OpenMP のレベルでクエリをさらに細かい粒度で分割処理する（図 2)。

２）「京」における GHOST-MP の最適化の実施

開発した GHOST-MP を、「京」においてコードの最適化やデータベースの最適化などを行うことで、実行速度およびノード内並列性能、ノード間並列性能を改善した。

コードの最適化により、実行速度は約 1.7 倍高速化し、ノード内並列性能は 8 スレッド使用時に当初は 7.0 倍の高速化であったものを 7.6 倍の高速化へと改善した（図 3、4)。

また、初期に、データベース読み込みが律速となって 384 ノード以上でスケールしなかったノード間並列性能（768 ノード使用時と 384 ノード使用時を比較しストロングスケーリング 0.43）について、データベースに対してファイルストライピングを適用することで、12,288 ノードまでスケールするようになった(12,288 ノード使用時と 6,144 ノード使用時を比較し、ストロングスケール 0.85)。これらの一連の改良によって、12,288 ノード使用時に、8,000 万リード／時の処理速度を実現した。これらの性能評価は、メタゲノム解析の実データを用いて実施した。

自動パイプライン化については、パイプラインへのジョブの投入や実行状態モニタなどを簡便に

行うためのウェブユーザーインタフェースも開発した。現在は、実行時間やネットワークなどの様々な制限から「京」では利用できないが、このインタフェースはウェブブラウザを通して、ジョブの投入、キャンセル、再実行および実行状態の確認などが可能で、解析結果の要約をグラフィカルに表示する機能も有している。

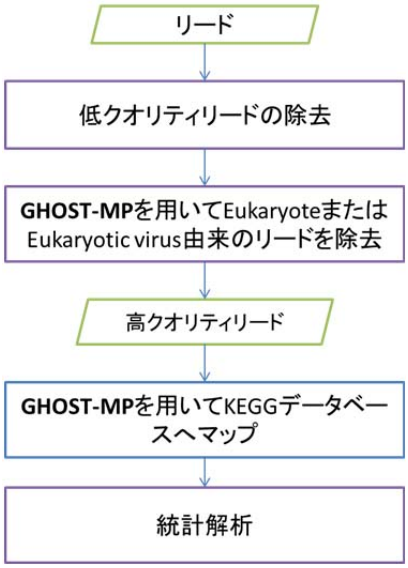

図 1　メタゲノム解析の自動パイプラインの概要

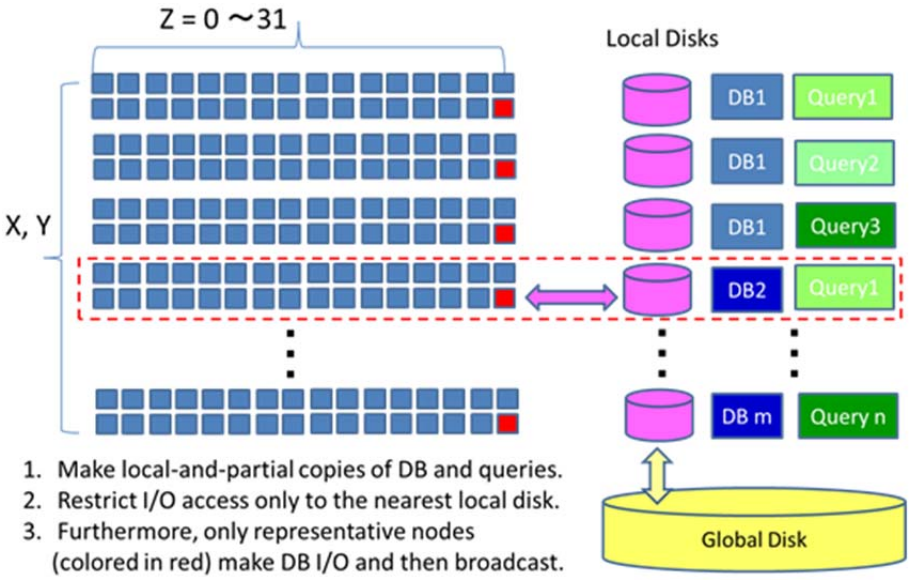

図 2　GHOST-MP の概要

GHOST-MP は、分割したデータベースとクエリを、ステージングによりローカルディスクに配置し、代表ノードが読み込んだ後に MPI の通信により各ノードへ転送する。

ノード内の計算では、分配されたクエリを、OpenMP を用いてさらに細かい粒度で並列に処理する。

（平成 23 年度では、「京」にステージング機能が未導入のため、代替としてファイルストライピングされたデータベースとクエリを代表ノードが読み込むシステムとしている。）

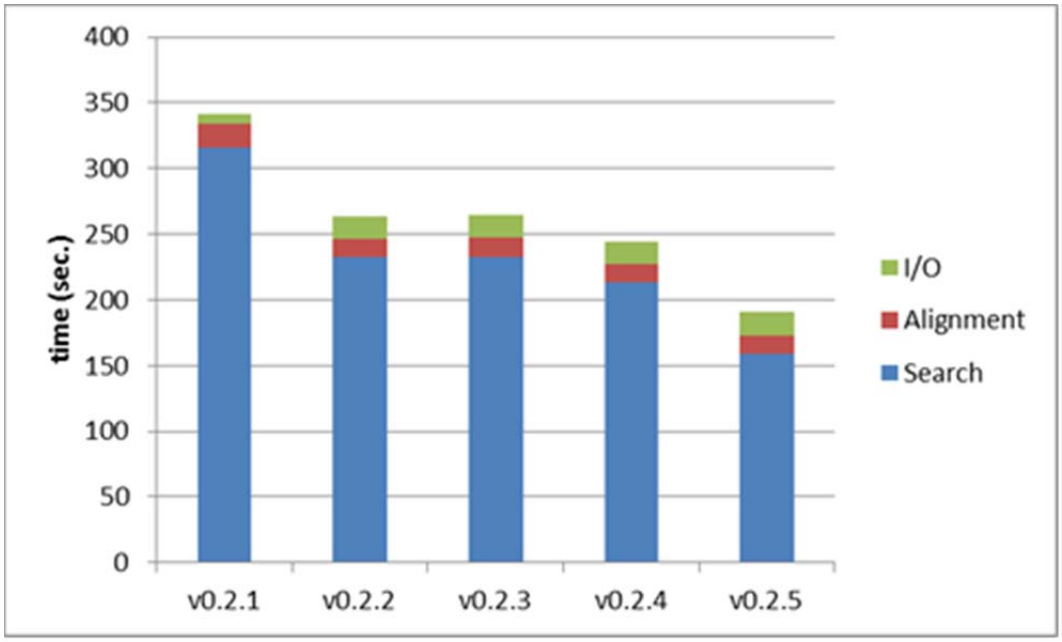

図 3　メタゲノム解析に GHOST-MP を適用した際の実行時間
横軸は GHOST-MP のバージョン、縦軸は実行時間。I/O はファイル I/O、Alignment はヒットした相同配列のアラインメント計算部分、Search は相同配列の探索部分の実行時間を表す。
v0.2.1 から v0.2.5 までで 1.7 倍高速化した。

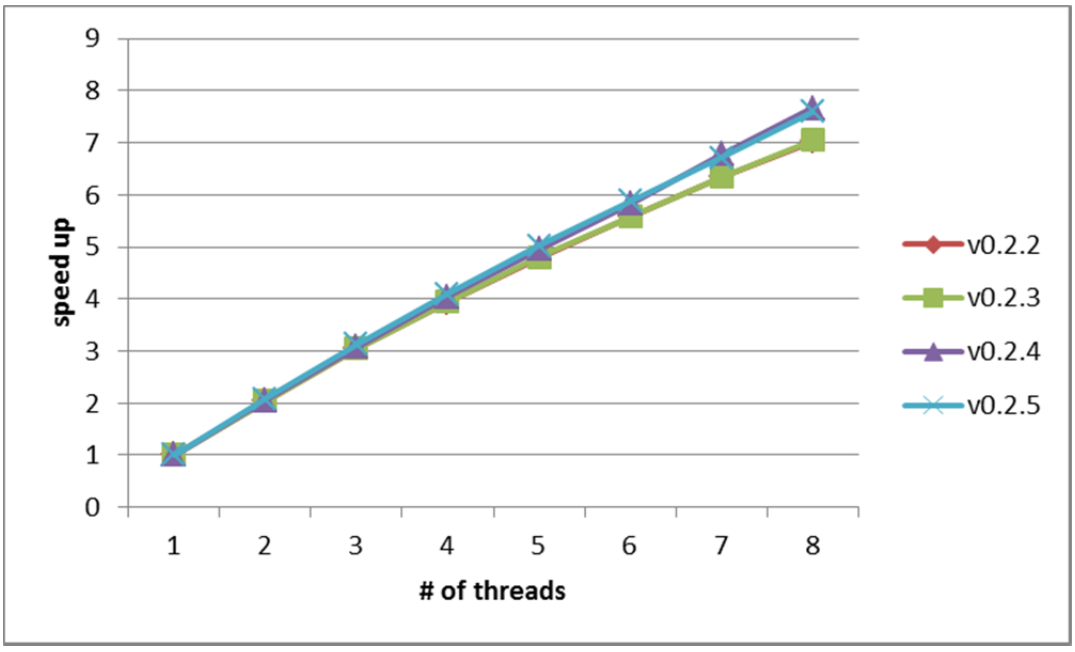

図 4　メタゲノム解析に GHOST-MP を適用した際のスレッド数に対する速度向上
横軸はスレッド数、縦軸は 1 スレッド使用時に対する速度向上を表す。
ファイル I/O を除いた結果を記載した。

＊記載した性能値については、整備中のシステムによる暫定的な数値である。

（２）研究開発の実施状況

１）接尾辞配列を用いた高速な高感度配列相同解析法の研究

通常のゲノム解析では、DNA 配列間の比較は塩基の一致・不一致のみに基づいて行う事が可能である。しかしながら、ゲノムがデータベースに登録されていない生物種もサンプル中に含まれるメタゲノム解析では、遠縁の生物との相同性を検索する必要があり、一致・不一致のみに基づいた検索では感度が不十分である。DNA 配列を 6 つの読み枠でアミノ酸配列に翻訳した後、アミノ酸残基間の類似度を評価することで、より高い感度で相同性解析を行う手法があるが、この検索には単純な一致配列の検索に比べると非常に多くの計算が必要となる。これまでは、比較的高速な相同性解析が可能な BLAST が利用されているが、その解析速度はメタゲノム解析に十分とは言えない。また、BLAST よりも高速な相同性解析法として BLAT が提案されているが、BLAT は高速であるが、相同性解析の感度が低いためメタゲノム解析への利用は適当でない。そのため、高速かつ高感度な相同性解析法が必要とされている。本研究では、クエリ配列とデータベース配列の双方に接尾辞配列（Suffix Array）を用いることで、高速かつ高感度の検索を行う手法を提案した。

提案手法を、土壌微生物のメタゲノムを Illumina 社の Genome Analyzer（Solexa)で読み取った実データ（断片長は 60-75 塩基）および MetaSim によって作成した断片長が約 500 と約 1000 の人工データ（L500、L1000）を用いて、相同性解析ツールである BLAST と BLAT と比較した。感度の評価においては、Smith-Waterman アルゴリズムによって厳密に最適なアラインメントを計算する SSEARCH プログラムの結果を正解とし、SSEARCH の結果との一致により判定を行った。提案手法を速度重視のパラメータで実行した場合、検索速度は BLAST および BLAT よりも高速であり BLAST 比で約 100 倍高速であることが分かった(図 5)。検索の感度は BLAST より低いものの、BLAT よりは高いことが分かった（図 6)。また、感度重視のパラメータで実行した場合、検索速度は BLAT よりも遅いものの BLAST 比で約 20 倍高速であり（図 5)、検索感度は BLAST と同等であることが分かった(図 6)。

提案手法を用いることで、次世代シーケンサリード配列に対し、メタゲノム解析などで必要とされる高速かつ高感度な相同性解析を行うことが可能となった。

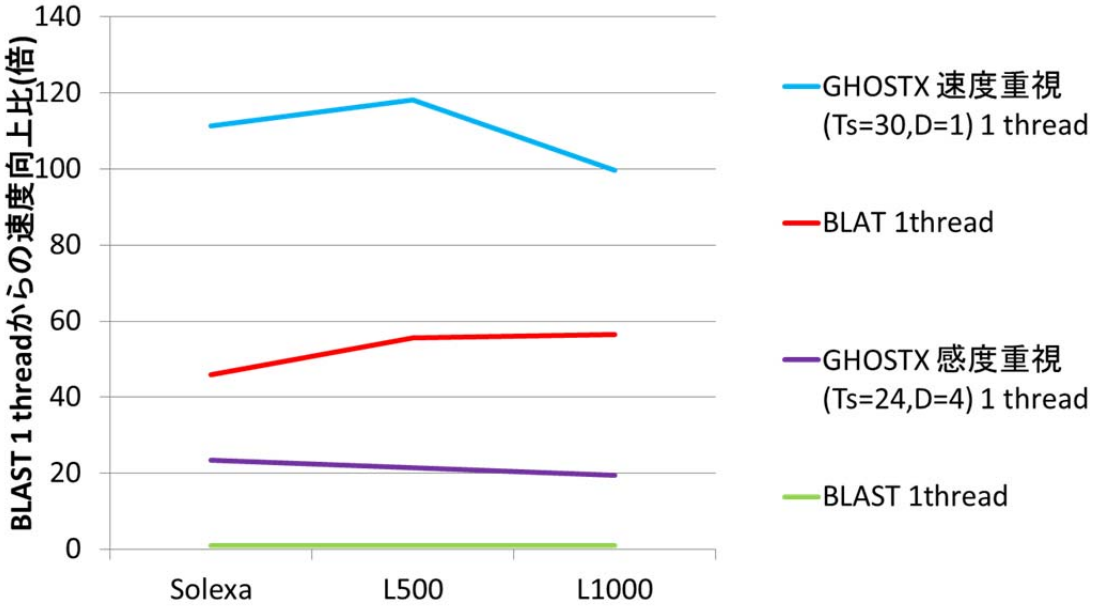

図 5　提案手法 GHOSTX、BLAST、BLAT の速度比較

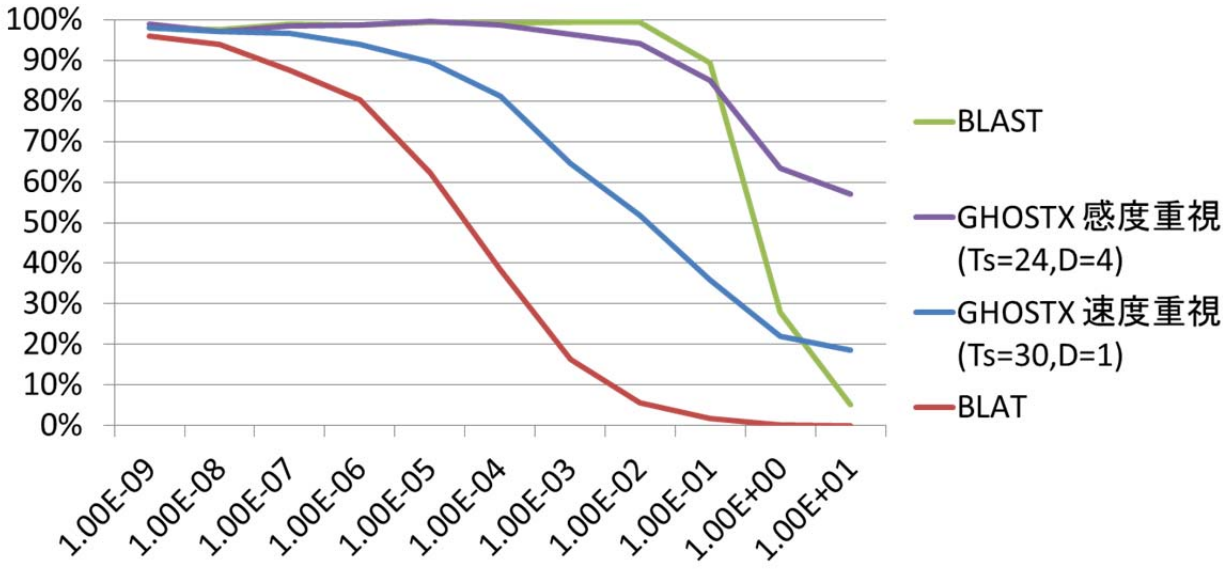

図 6　提案手法 GHOSTX、BLAST、BLAT の感度比較
各 E-value における SSEARCH の結果との一致度を示している。
縦軸は SSEARCH の結果との一致率、横軸は E-value を表す。

Ⅳ－３　浅井 潔（産業技術総合研究所）

RNA 相互作用予測技術の開発と転写物の網羅的情報解析

Ⅳ－３－１　実施計画

本試験研究では、「大規模生命データ解析」として、次世代シークエンサーによってもたらされる膨大な配列データの１次処理、得られた転写物情報の解析、ネットワーク解析、ゲノム間の比較を行う必要がある。その一環として、RNA 相互作用予測技術の開発と転写物の網羅的情報解析のための研究開発を実施する。

また、「RNA 相互作用予測技術の開発と転写物の網羅的情報解析」の研究を行う上で、関連する研究者と必要な協議等を行うとともに、本格実施に必要な研究体制の整備を行う。

平成 23 年度は、RNA 相互作用予測技術の開発の一環として、2 次構造を考慮した配列情報解析アルゴリズムを改良し、2 次構造的なエネルギーを考慮した RNA の相互作用の強度を評価する情報解析技術の開発と、転写物データによる評価を開始する。

Ⅳ－３－２　実施内容（成果）

RNA の配列情報解析では 2 次構造を考慮した解析を行わなければ正確な結果が得られないことから、転写物（RNA）の網羅的解析においては、2 次構造を考慮した網羅的解析が不可欠である。そのため、本研究課題では、2 次構造を考慮した解析手法の開発、分子シミュレーション、3 次元立体構造解析を行った。

2 次構造を考慮した解析手法の開発では、RNA 配列の任意の領域が 2 次構造的に他の分子と相互作用できる度合い（アクセサビリティ）を計算する手法・ソフトウェア（Raccess）およびシュードノットを含むRNA二次構造を期待精度最大化と整数計画法で予測する手法･ソフトウェア(IPknot)を開発し、誌上発表した。

分子シミュレーションでは、粗視化モデルを用いた RNA の分子シミュレーション手法を実装し、塩基対確率など RNA2 次構造予測にかかわるパラメータを取り込んだ手法の開発についても検討を行った。

3 次元立体構造解析では、2 次構造情報を用いた高精度のフラグメントアセンブリ手法・ソフトウェア（RASSIE[4]）を開発し、誌上発表した。

さらに､RNA 相互作用予測技術の転写物の網羅的情報解析への適用の有効性を検証するため､「RNA 相互作用評価のための長鎖 NGS シミュレーションデータ作成ツールの開発」を行った。

参考文献

[1] Hamada M et al. Prediction of RNA secondary structure using generalized centroid estimators. Bioinformatics 25(4):65-473 (2009).

[2] Kiryu H et al. A detailed investigation of accessibilities around target sites of siRNAs and miRNAs. Bioinformatics 27(13):1788-1797 (2011).

[3] Kato Y et al. RactIP: fast and accurate prediction of RNA-RNA interaction using integer programming. Bioinformatics 26(18):i460-i466 (2010).

[4] Yamasaki S et al. Prospects for tertiary structure prediction of RNA based on secondary structure information. J Chem Inf Model. 52(2):557-567 (2012).

Ⅳ－４　松田 秀雄（大阪大学）

大規模な生体分子ネットワークの解析技術の開発

Ⅳ－４－１　実施計画

本研究では、「戦略課題 4：大規模生命データ解析」の目標である、大規模生命データ解析による生命プログラム及びその多様性の理解のために必要となる、多数の生体分子間に存在する大規模な生体分子ネットワークの解析のための研究開発を実施する。

また、「戦略課題 4：大規模生命データ解析」の研究を行う上で、関連する研究者と必要な協議等を行うとともに、本格実施に必要な研究体制の整備を行う。

平成 23 年度は、シードネットワーク法を大規模 HPC 環境で実行するためのスケジューリング等の基本機能を実現するために、基本機能の設計とそれを実装したプロトタイプソフトウェアを作成し、種々の規模の生体分子ネットワークの解析に適用して性能測定を行う。

Ⅳ－４－２　実施内容（成果）

（１）ソフトウェアの開発・高度化の状況

１）生体分子ネットワーク解析技術の開発

大規模な生体分子ネットワークを、生物学的な知見を取り込む形で推定する方法として、シードネットワーク法を考案し、それに基づく分子間の制御関係のネットワーク推定を行うソフトウェアのプロトタイプを開発した。具体的には、生体分子の全セットを使って単純にネットワークの推定を行うのではなく、まずは対象とする生命現象についての生物学的な知見が既にある生体分子のみをもとにしたネットワーク（これをシードネットワークと呼ぶ）について推定を行い、その推定結果と生物学的な知見を照らし合わせることで、適切なネットワーク推定のパラメータ（連続値である時系列発現プロファイルを、何らかの数理モデルに当てはめるときのモデルパラメータ）を決定し、それをもとに全体のネットワークの推定を行うようにする。

時系列発現プロファイルからのシードネットワーク法によるネットワーク推定のソフトウェアを、ネットワーク推定部分はグランドチャレンジプロジェクトで宮野悟（東大）により開発された SiGN をベースに、モデルパラメータ決定支援のためのネットワーク表示部分は宮野研究室で開発されている Cell Illustrator をベースにして開発した。

マウスの脂肪細胞分化の時系列発現プロファイルを使って、シードネットワークとして 113 個の分化関連の既知の遺伝子からなる生体分子ネットワークを推定したところ、図 1 に示すようにモデルパラメータの値によって、推定されるネットワークの構造が大きく異なった（図 1 ではノードは生体分子を表し、発現時期ごとにノードを色分けし、多数のノードと接続される次数の大きなノードほどサイズが大きくなるように表示されている）。

脂肪細胞分化で代表的なマーカー遺伝子（PPARγ，C/EBPα，LPL，FABP4，ADIPOQ，GLUT4 な

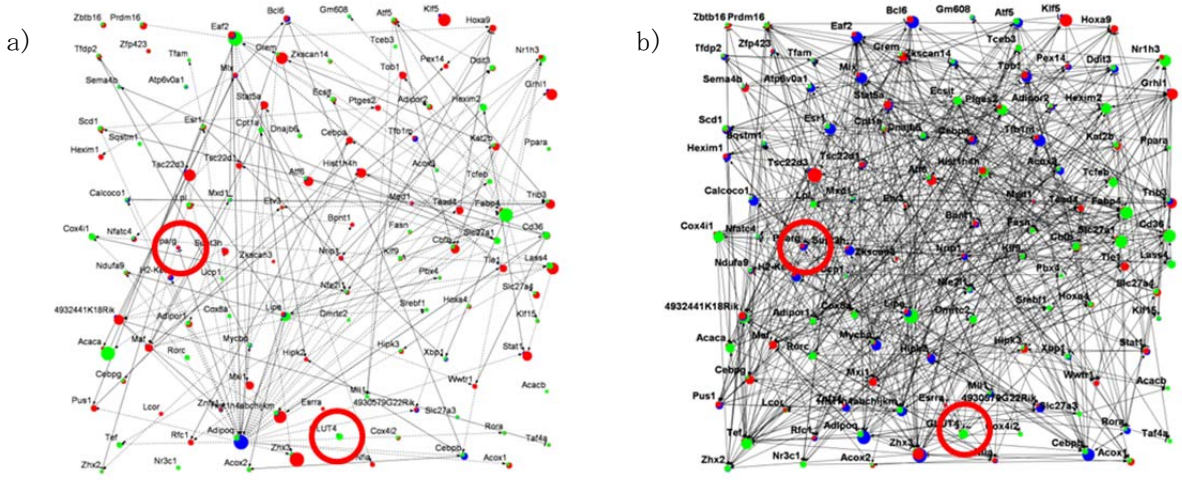

図 1 マウス脂肪細胞分化上におけるシードネットワークの推定

ど）を含む既知の分化関連因子が、その時期に大きな次数となるようなネットワークのモデルパラメータの値の組合せ(図１では赤丸で囲んだPPARγとGLUT4が図１a)だと次数が小さすぎる)を探索し、その組合せで以降のより大規模なネットワークの推定を行うようにした。

２）生体分子ネットワーク解析技術の大規模化

前節で述べたソフトウェアのプロトタイプを「京」上で大規模な並列実行が行えるように拡張し、BENIGN (Biologically Extensible Network Inference Software for Gene Expression Analysis)と名付けた。BENIGNの構成を図２に示す。

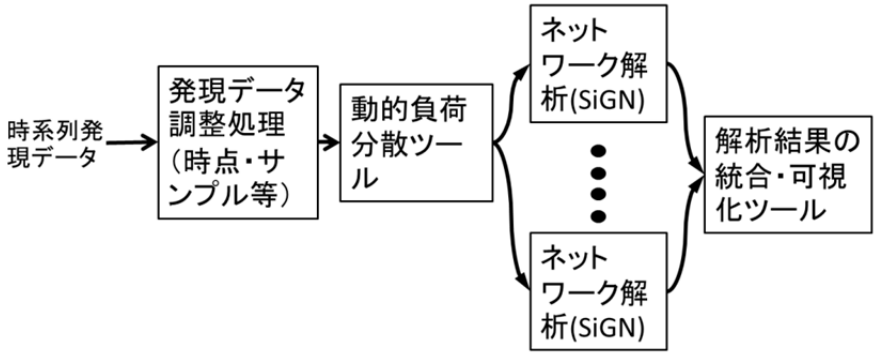

図２ソフトウェアBENIGNの構成

BENIGNでは、発現時期ごとの個別のネットワーク推定や、発現プロファイルについてのブートストラップ試行のサンプリングと、モデルパラメータの探索などのため、異なる発現データやパラメータ値でのネットワーク推定を行う。これらの組合せを調整処理し、その組合せごとにネットワーク解析プロセス（SiGNをベースにしている）を起動し、プロセスの計算ノードへの割り当てを行うことで並列処理を実現している。現状では、ブートストラップ試行のみMPI版のSiGNの負荷分散機能を使って動的な負荷分散を行っているが､それ以外の処理の並列実行は割り当て時に計算ノードを固定する静的な割り当てに留まっている。平成24年度には、これらを含めて全体を動的に負荷分散できるようにする予定である。

BENIGNの「京」での並列実行性能を表１に示す。時系列発現プロファイルとしては、マウスの間葉系幹細胞から脂肪細胞への細胞分化での発現変化を､分化誘導開始から192時間までの61時点をAffymetrixマイクロアレイMouse Genome 430 2.0 Arrayで測定したものを、前述のマーカー遺伝子などの脂肪細胞分化関連の既知遺伝子113個と、分化で発現変動がみられた転写因子を合わせた合計946個の遺伝子の発現プロファイルを使用している。

表１　BENIGNの並列実行性能

計算ノード数	実行時間（秒）	速度向上	並列化効率
6,114	2683.4797	1.00	
12,228	1386.0406	1.94	0.97

表１では、6,114ノードでの実行時間を基準として、ストロングスケーリングで速度向上と並列化効率を計測している。表からわかるように、12,228ノードでの実行では97%の並列化効率という良好な性能を達成している。

（２）研究開発の実施状況

１）脂肪細胞分化過程における生体分子ネットワーク解析の研究

BENIGNで、前述のマウスの脂肪細胞分化の時系列発現プロファイルで、113個の既知遺伝子からネットワーク推定を行った結果を図3-5に示す。埼玉医大の岡崎康司らのグループによる実行研究（Y. Tokuzawa, et al., PLoS Genetics, 6(7):e1001019, 2010）によると、脂肪細胞分化の過程は初期・中期・後期の３段階の時期に分かれるとされているので、これを参考に61時点の時系列発現プロファイルを分割し、それぞれでネットワーク推定を行っている。

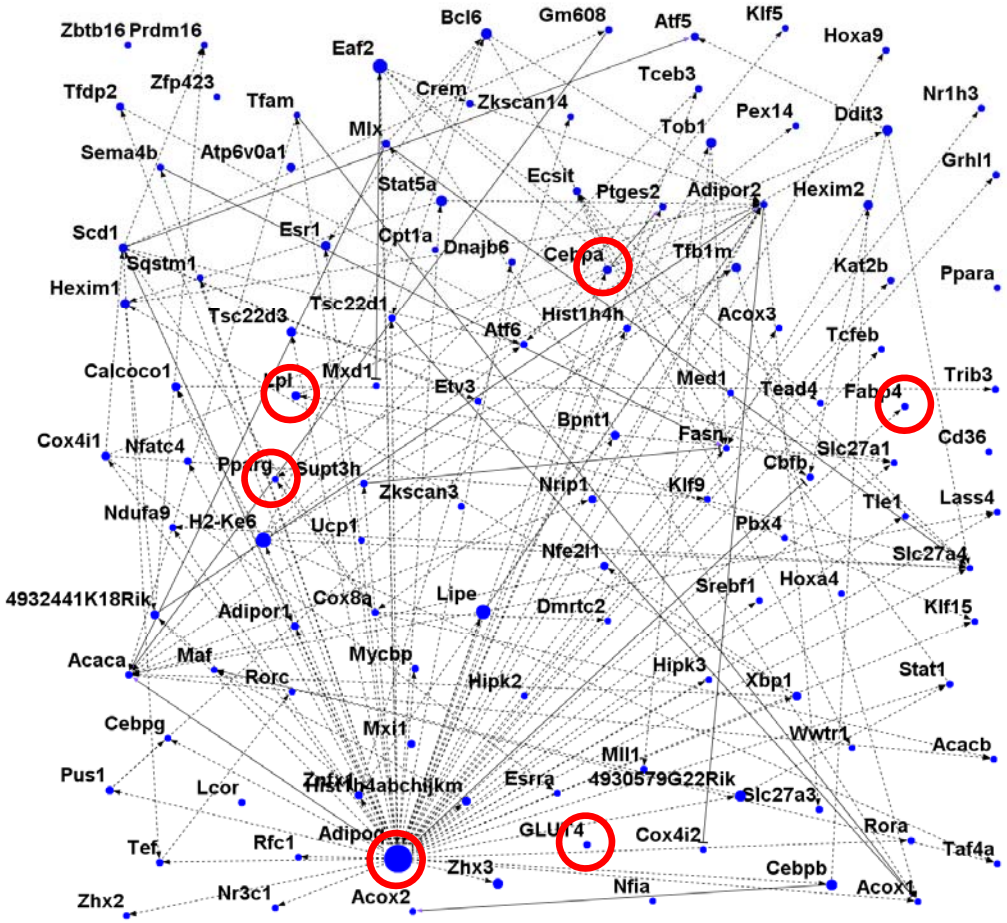

図 3 マウスの脂肪細胞分化での分化初期の生体分子ネットワーク

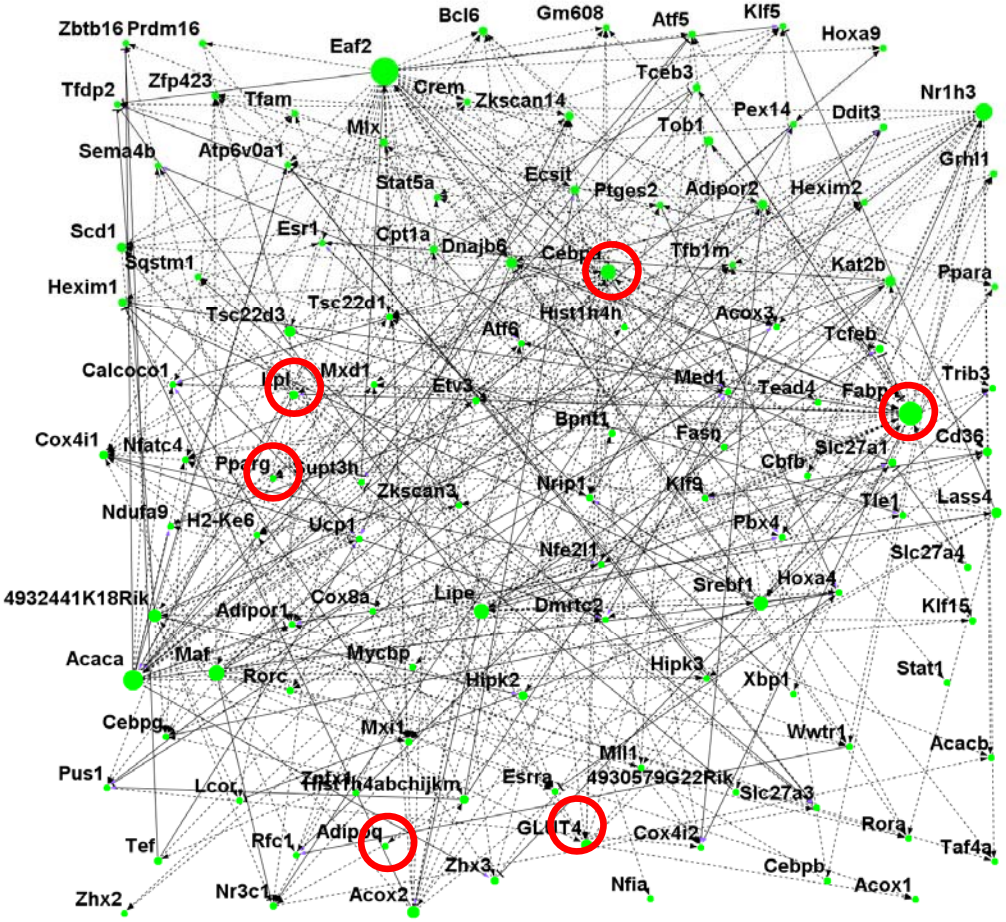

図 4 マウスの脂肪細胞分化での分化中期の生体分子ネットワーク

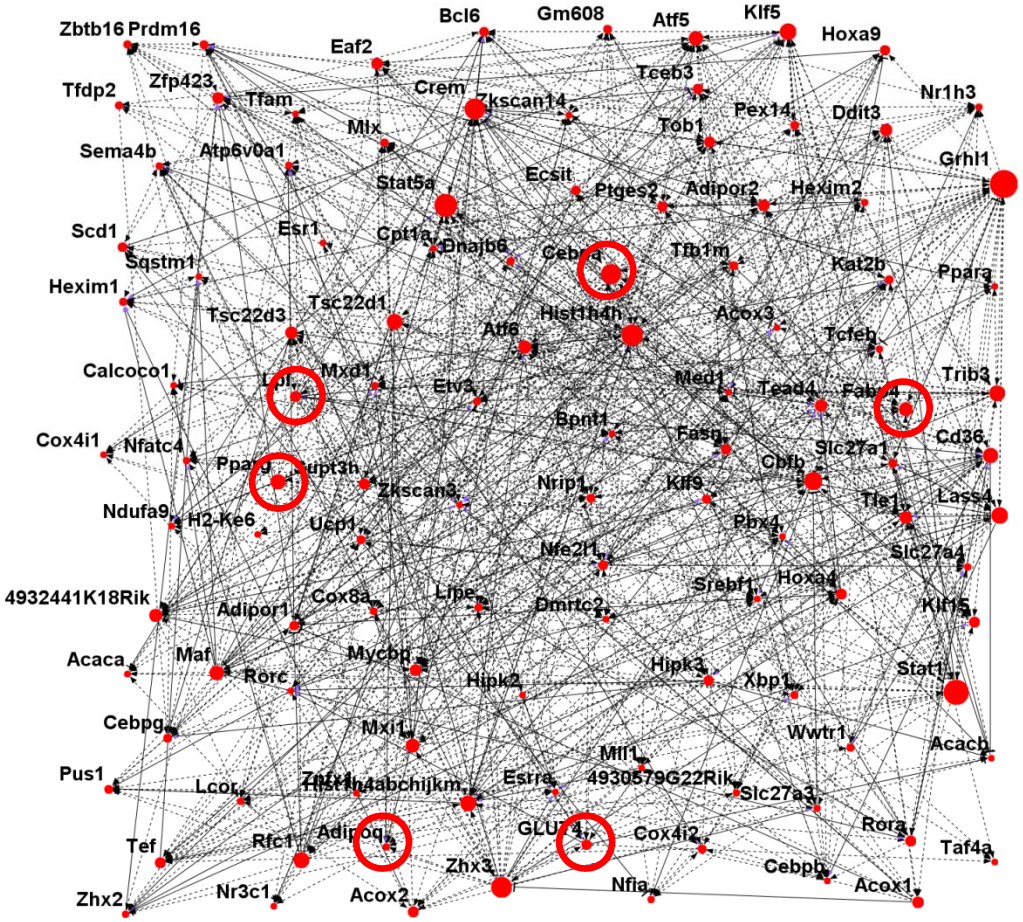

図 5 マウスの脂肪細胞分化での分化後期の生体分子ネットワーク

図 3 から図 5 に示すように、脂肪細胞分化に関連した既知のマーカー遺伝子（図中で○で囲ったもの）は、分化のいずれかの時期で何らかの制御関係に関与していることが示されており、ネットワーク推定結果は妥当であると考えられる。ただし、このネットワークでは脂肪細胞分化に関連していると考えられる既知遺伝子のみで構成されているため、未知の因子の関与は推定できていない。また、最適なネットワークのモデルパラメータが必ずしも探索しきれているとは限らない。そこで、まず、発現変動のあった転写因子も加えた 946 ノードのネットワークの推定結果を解析中である。

今後は、今回のネットワーク推定結果に基づいて、「京」の持つ膨大な計算パワーをフルに生かすべく、BENIGN の動的負荷分散機能の強化などにより、転写因子以外の遺伝子も含めた全遺伝子セットでのネットワーク推定を行う予定である。また、研究協力者の京都大学農学研究科の河田教授より、マウスの種々の脂肪細胞組織での RNA-Seq による発現データの提供を受けており、Affymetrix マイクロアレイでは得られない、miRNA などの non-coding RNA も加えた全転写単位での大規模生体分子ネットワークの推定を進める予定である。

Ⅳ－５　五條堀 孝（国立遺伝学研究所）

メタゲノム・比較ゲノム解析研究

Ⅳ－５－１　実施計画

本研究では、「大規模生命データ解析」における主目的のひとつである「地球規模ゲノム時代を先導し、生物多様性の大規模データ解析を実現する」ため、「メタゲノム・比較ゲノム解析研究」に必要となる大量情報解析のための研究開発を開始する。

また、「メタゲノム・比較ゲノム解析研究」の研究を行う上で、関連する研究者と必要な協議等を行うとともに、本格実施に必要な研究体制の整備を行う。

平成 23 年度は、具体的に、大量なメタゲノムデータの産出がなされつつある腸内細菌や海洋微生物に注目して、メタゲノムデータの品質管理をおこなうための作業データベースの構築を開始し、メタゲノム大規模データベースの構築とそれに伴うアノテーションの基盤構築の準備をおこなう。

Ⅳ－５－２　実施内容（成果）

（１）ソフトウェアの開発・高度化の状況

１）最尤（ゆう）系統樹推定プログラムの開発

系統樹推定は、ゲノムから得られる塩基配列情報に多重整列(マルチプルアライメント)処理が施された複数の配列から系統樹推定を行うものである。本研究で用いる系統樹推定法は最尤法(Maximum Likelihood 法)と呼ばれ、近年の計算機性能の向上に伴って生命科学分野の研究において頻繁に利用されており、評価が安定している手法である。本手法は、与えられた配列データに対し複数の候補系統樹間で比較を行い、最も相応しい樹形を選択するものである。このため、最適な樹形を推定するためには候補系統樹を網羅的に探索する必要があるが、候補樹形の数は、配列データ数を N とすると、$O(2^N N!)$(!は階乗を示す)のオーダーで爆発的に増大するため、一定数以上の配列に対し候補樹形の全体を網羅探索することは理論上不可能である。このため、プログラムの実装上は、探索空間を狭めて探索を行うことで最尤系統樹の推定を実現している。探索の手法はプログラムによって異なるため、「京」コンピューター上でプログラムを実装するためには、プログラム毎に異なった並列最適化が必要となる。そこで本研究におけるソフトウェア開発に先立ち、最尤法を実装したプログラムのうち、「京」コンピューターの並列アーキテクチャに適合するものを調査した(表 1)。その結果、ドイツの Ludwig-Maximilians 大学の A. Stamatakis らが中心となって開発した、RAxML と呼ばれるプログラムが最も適当である事が判明したため、以降このプログラムのコードを元にソフトウェアの開発を行うこととした。前述の RAxML プログラムは、MPI(Message-Passing Interface)を使用しており、基本的に「京」コンピューター上での並列プログラミングの要求規格を充足していたが、部分的に p-Thread 等の別の並列化手法が併用されており、そのままでは

表 1：最尤系統樹推定プログラムの代表例

ML は最尤法、MP は再節約法、BI はベイズ推定法を示す。

Package name	Methods	Parallelization
fastDNAml	ML	○
MEGA	MP, ML	
MOlPHY	ML	
MrBayes	BI	○
PAML	ML, BI	
PHYLIP	MP, ML	○
RAxML	MP, ML	◎
TREE-PUZZLE	ML	○

「京」コンピューター上では利用出来ない事が判明したため、「京」上で利用可能な並列化手法を用いる事によりプログラムを実装した。

２）初期ベンチマークテストの実施

ミュンヘン工科大学の ARB データベース(http://www.arb-home.de)上で公開されている、1,000 生物種のリボソーム DNA の配列データから部分データ(50, 100, 150, 200, 250, 500 種) を作成した。次に、これらのデータを「京」コンピューターと比較対象である、PC クラスター(SGE)上で別々に実行し、異なる並列度(2, 4, 8, 12, 24, 36, 72 ノード)で実行時間を計測した。

$$\text{CPU occupancy}\ (\%) = \frac{user + sys}{elapse \cdot p}$$ (*user*: 総ユーザー時間, *sys*: 総システム時間, *elapse*: 経過時間 *p*: プロセッサ数)

$$\text{Speedup}\ S(n,p) = \frac{T_\sigma(n)}{T_n(n,p)}$$ ($T_\sigma(n)$: 逐次処理実行時間, $T_\pi(n,p)$: p プロセッサによる並列実行時間)

$$\text{Efficiency}\ E(n,p) = \frac{S(n,p)}{p}$$

実験の結果、図１右パネルに示すように、系統樹の本数が増えるに従って、計算時間が指数的に増大した。最尤法においては、与えられた枝数から生成される全樹形にから正しい樹形候補を網羅的に探索するが、樹形数は枝数を N とすると、$2^N N!$ (!は階乗を表す)のオーダーで増大することが知られており、実際に配列データを解析するために、「京」コンピュータの並列アーキテクチャを最大限利用したプログラムの更なるチューニングが必要である。

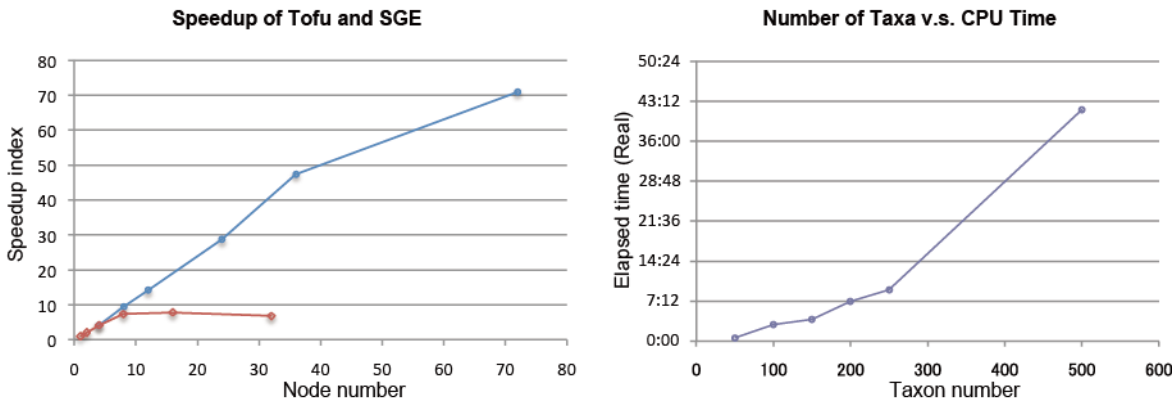

図 1: テストデータを用いた並列化効率の変化

「京」のアーキテクチャ(図中では Tofu と呼ぶ)と PC クラスターとの並列化効率の比較では(図 1 左パネル)、赤線で示した PC クラスターが 8 並列度程度で並列化効率が頭打ちとなるのに対し、青線で示した「京」のアーキテクチャではより高い並列度においても並列化効率の向上が維持されていることが示された。これらの結果は、本実施研究において採用した RAxML プログラムが、「京」コンピューターのアーキテクチャに基本的に適合していることを示す結果であり、本件のソフトウェア開発が順等に進捗していることを示している。今後予定されている実際の配列データ解析においては、今回のテストより大規模なデータの解析が予想されることから、今後は、ソフトウェアの改善を行いつつ実際の解析を行う予定である。

（２）研究開発の実施状況

１）最尤系統樹推定を用いた比較ゲノム解析の研究

生命プログラムの複雑性・多様性や進化は、原始生命から連綿と継承されてきた遺伝情報担体である「ゲノム」に刻まれている。本研究開発課題は、我が国に先導的な研究蓄積が豊富な、分子進化学の技術を基盤に、これまでに確立されてきた分子進化解析の技術を大規模ゲノム解析に適用することで、生命科学にゲノムの視点から新たな光を当てようとするものである。本研究開発課題において、研究開始時点においては、メタゲノム解析の中心課題として大規模データベースの構築に主眼に置かれていたが、外部諮問委員会における指摘を受け、研究の方向性として、ゲノム配列データの実解析を中心とすることが提案されたため、この指摘を反映するように研究開発の指向を検討した。その結果、本研究の中心テーマである、最尤法による系統樹推定の活用という軸を堅持しつつ、分かり易い成果目標として、①感染症ウイルスの進化的解析、②日本人の人類集団における進化的位置、および、③がん細胞の系統的関係の解析の３つの目標を設定した(図 2)。

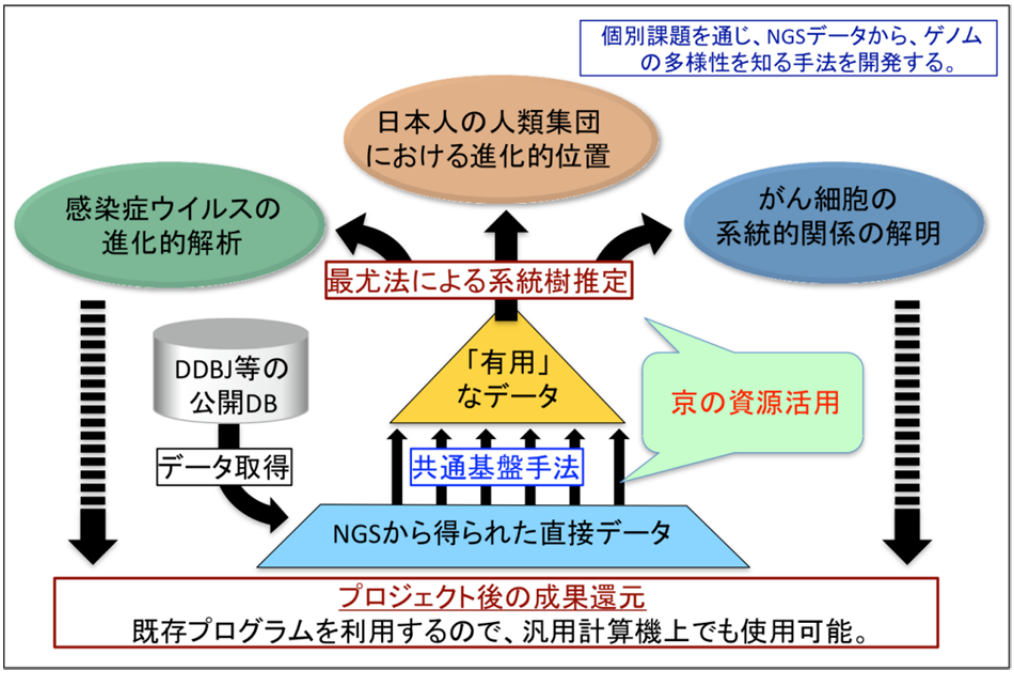

図 2：本課題プロジェクトの俯瞰図