Ⅳ　戦略課題４：大規模生命データ解析

（統括：宮野悟・東京大学医科学研究所）

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－１－１　実施計画

本委託では、戦略プログラムの「課題４　大規模生命データ解析」研究の一環として、研究開発を統括する。

また、上記の研究を行う上で、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。

以下の大学等で実施される平成２４年度の研究課題の実施項目について、適宜、関連する研究者とワークショップや研究打合せ、及び研究調査を行うことにより、関係者のとりまとめを行うとともに、理化学研究所と連携して、研究開発の統括を行う。

①　次世代シークエンサ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　比較ゲノム解析研究(五條堀孝・国立遺伝学研究所)

平成２４年度は、ワークショップ・研究打合せを通した(a)-(d)の研究調整と(e)の研究調査を行う。

(a) ヒトゲノムシークエンスデータに対応するための①の開発調整

(b) ②の開発における京コンピュータ活用調整、及び③の開発との連携の調整

(c) ③の開発とグランドチャレンジで開発しているソフトウェアSiGNの整合性の調整

(d) ④の開発における京コンピュータ活用調整、及び解析データの調整

(e) 米国における臨床シークエンスについての研究調査

Ⅳ－１－２　実施内容（成果）

（１）研究調整

(a)-(d)の研究調整に関して以下のことを実施した。

1. 毎月第4木曜日に開催されるHPCI戦略プログラム分野１運営委員会に出席し、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。
2. ヒトゲノムシークエンスデータに対応する①の開発調整のため、東工大に秋山泰教授を訪問し、研究の進捗と京コンピュータの運用に関する意見をもらい、運営委員会に反映させた。ソフトウェアの開発が順調に行えるよう、また、メタゲノムデータの解析が順調に進むよう計算リソースの利用について調整した。
3. ②の開発における京コンピュータ活用調整のため、産総研に浅井潔教授を数回訪問し、研究の進捗の報告のプレゼンテーションを受けた。研究副統括及び分野別作業部会からの意見を研究に反映するように指示した。また、③の開発との連携の可能性について検討したが、次期が早いことが判明したため、それ以上のことは進めなかった。
4. ③の開発とグランドチャレンジで開発しているソフトウェアSiGNの整合性について、大阪大学の松田秀雄教授を東大・情報理工の玉田嘉紀助教と数回にわたり訪問し、開発のための調整

を行った。京コンピュータでの高並列化を推進できるよう調整した。

5. ④の開発における京コンピュータ活用調整、及び解析データの調整のため、国立遺伝学研究所に五條堀孝教授を数回訪ね、研究のフォーカスの仕方及び京コンピュータの利用促進のための調整を行い、阪大・松田教授との連携を指示した。
6. 2013 年２月２５日の分野別作業部会の資料作成及びプレゼンテーションを行った。
7. 2012 年 11 月 29 日－30 日に行われた分野１の全体ワークショップに出席し、グループ全体の発表及び意見交換を行った。
8. 外部諮問委員会からの意見への対応を行った。
9. 戦略統括及び副統括に研究の進捗を報告し、意見をうかがい、それに基づき、平成２５年度の研究内容の見直しを行った。

（２）研究調査

1. 米国における臨床シークエンスについての研究調査のため、米国サンフランシスコの Hotel Kabuki で開催された臨床ゲノム会議に出席し、臨床シークエンスの医療応用について、臨床応用、遺伝子情報ビッグデータ、計算資源量、データ解析、医療保険、倫理、検査技術、レギュレーションについて調査した。The Medical College of Wisconsin の Howard Jacob 教授及び Elisabeth Worthey 教授と臨床シークエンスについて意見を交わした。
2. 2013 年 1 月 15 日、戦略分野１の課題４「大規模生命データ解析」に関連して、細胞の分化とがんの研究について、iPS 細胞を使って行っている岡山大学工学部・妹尾昌治教授の研究について調査し、戦略分野１での研究戦略について意見をうかがい議論をした。
3. 2013 年 1 月 16 日に計算科学研究機構で開催された第５回 HPCI 戦略プログラム合同研究交流会に出席し、ゲノムワイド遺伝子ネットワーク解析についての現状と今後の展望について講演し、戦略５分野間で意見交換を行った。

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２４年度は、昨年度に開発したリード配列の相同性解析のための並列ソフトウェア(GHOST-MP)のコード最適化を行い「京」での実行効率をさらに高めるとともに、80,000 ノード級の超大規模実行を実施して性能の測定を行う。平成２３年度においては、最大で 12,288 ノードまでの測定を行っているが、その数倍に当たる超大規模実行を可能とするためには入出力の負荷を分散する必要があり、代表ノードのみに制限した入出力とネットワーク転送の組み合わせによるデータ配信機能を開発する。また、平成２３年度に開発した、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の機能を高度化し、将来的には「京」との遠隔接続も視野にいれた設計とする。平成２４年度内に、システム性能の実証のために、次世代シーケンサによる複数回の実験に相当する大規模なメタゲノム解析データを用いた評価実験を行う。

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 年度の研究成果を受け、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自動パイプラインの高度化を行った。自動パイプラインは、相同性解析を行うプログラムを中心として、次世代シーケンサの出力するサンプル中の遺伝子の機能解析を行う。

自動パイプラインは、一連の処理を通して次世代シーケンサから得られる各リード配列に対して､配列データベースとの比較参照を通じてアノテーションを行う(図 1)。このアノテーションによって、未知の生物種由来のリード配列に対しても遺伝子の機能等を考慮した詳細な解析が可能となる。自動パイプラインの高度化では、パイプラインの中で多くの実行時間を占めていた、シーケンサからのリード配列に対して相同性解析を行う部分に対して、アラインメント候補探索アルゴリズムの改良や、その相同性解析に用いる配列データベースの高速な分配方法の検討を行い、また、パイプラインへの機能追加などを行った。

相同性解析では、クエリ配列の類似配列を配列データベースから検索する。この目的でよく利用されている BLAST (Basic Local Alignment Search Tool) では、アラインメント候補探索の際に配列データベース全体を走査している (図 2(a))。これに対し、接尾辞配列を用いたアラインメント候補探索では、データベース全体を走査することなく候補を探すことができる (図 2(b))。接尾辞配列は、部分文字列 s[i, n] (s:元の文字列、i=1,2,...,n、n:s の長さ) の開始位置を部分文字列の辞書順に並べた配列であるため、ソート済みであることを利用して二分探索による高速な検索が可能となっている (図 2)。ただし二分探索は、探索の回数自体は減らせるものの、計算機メモリに連続的にアクセスしないため、キャッシュメモリの恩恵を得にくい部分があった。そこで、この接尾辞配列の一部に対しルックアップテーブルを追加することで、一定文字数 k までの探索を $O(k)$ で行えるよう改良した (以前の二分探索では $O(k \log n)$)。この変更により検索速度が大きく改善した (図 3)。

配列データベースの高速な分配方法の開発では、「京」の Tofu ネットワークの詳細な速度測定を行い、これに基づいて分配方法の検討を行った。配列データベースの分配には MPI ライブ

ラリの提供する MPI_Bcast を用いるが、「京」では特定の条件の下に使用可能な高速な MPI_Bcast が用意されている。一つは、3 次元トーラスをセグメント分割した隣接通信のみで実装し、通信の衝突を避けた Tofu 専用アルゴリズム、もう一つは、全ランクの要素数が同じであることを保証した場合の高速なアルゴリズムである。測定の結果、通常の MPI_Bcast を用いた場合、Tofu 専用アルゴリズムに比べ 25-3500 倍遅く、計算ノードの増加に比例して計算時間も増加することが分かった（図 4)。一方、Tofu 専用アルゴリズムを使用した場合、12288 ノードまでの測定結果では、計算時間は使用ノード数に依らず一定であった（図 4)。この結果から、Tofu 専用アルゴリズム使用のための条件を満たすようにデータベース（およびクエリ）の分配方法を変更した。

近年の配列データベースサイズの増大は著しく、使用する配列データベースによっては、「京」の 1 ノード当たりのメモリ容量 16GB に収まらない。そのため、データベースを分割して計算ノードごとに担当データベースを割り当てる方法を用いた。このデータベースの割り当ては、従来はランク番号とデータベースの分割数の剰余にあたるデータベースを割り当てていたが､前述の測定結果を受けて､データベースの分割数だけ 3 次元直方体に分割した MPI コミュニケータを用意し、データベースの分配を行うよう変更した（図 5)。これは、Tofu 専用アルゴリズムの MPI_Bcast の適用条件の 1 つとして、コミュニケータの形状が 3 次元直方体であることを要求しているためである。開発した相同性解析のプログラムの大規模性能評価を「京」で行った。1 ノード当たりの問題規模を固定して並列性能（weak scaling）を測定したところ、「京」の全計算ノード（82944 ノード）を使用した場合でも、スケールする様子が観察され、大規模な解析を短時間で行えることが示された。

（２）研究開発の実施状況

１）ゲノム配列情報解析自動パイプラインの実データへの適用

自動パイプラインを用いた実証実験を、ヒト口腔内細菌叢のメタゲノムを対象として行った。サンプルデータとして、米国国立衛生研究所の Human Microbiome Project（HMP）でシーケンシングされた口腔内 9 部位（角化歯肉、頬粘膜、硬口蓋、口蓋扁桃、咽喉、舌背、唾液、歯肉縁上の歯垢､歯肉縁下の歯垢)､395 サンプルのリード配列計 200 億リードを対象データとして用い､KEGG genes の配列データベースに対して相同性解析を行った（図 6)。KEGG genes の各配列には、代謝経路に関するアノテーションが付与されているため、サンプル中にどのような機能を有するオーソログ遺伝子がどの程度含まれているか解析することができる。これにより、図 7 のように代謝経路上で機能するオーソログ遺伝子の多寡を部位間、あるいはサンプル間で比較することができた。また、各オーソログ遺伝子の多寡を入力として、主成分分析を行った結果、第 1、第 2 主成分で口腔前庭、（狭義の）口腔および歯垢と部位によって明確に分離されることが分かった（図 8)。本自動パイプラインを用いることで、膨大なリード配列に対し相同性解析を利用した機能アノテーションが可能になり、遺伝子の機能に着目した解析が可能となることが示された。

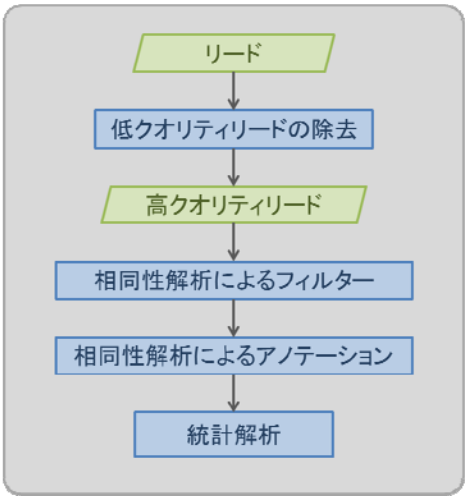



図 1　メタゲノム解析の自動パイプラインの概要

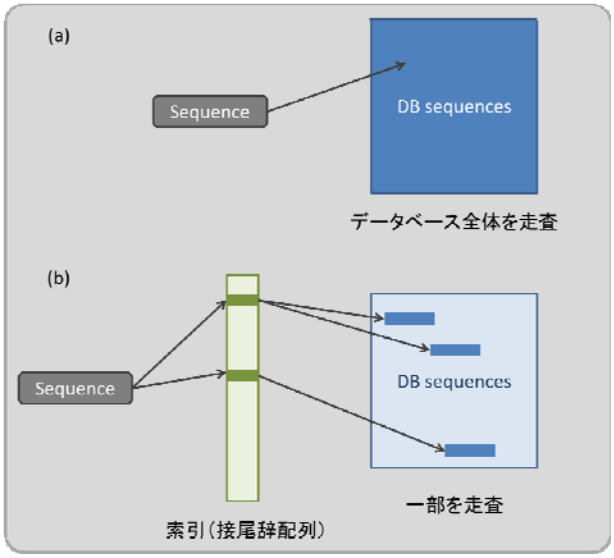



図 2　(a) BLAST 等の古典的な相同性解析ツールのアラインメント候補探索と、(b) 索引（接尾辞配列）を用いたアラインメント候補探索。後者ではあらかじめ作成された索引を用いることで、配列データベースの一部を走査するだけで全てのアラインメント候補を探索できる。

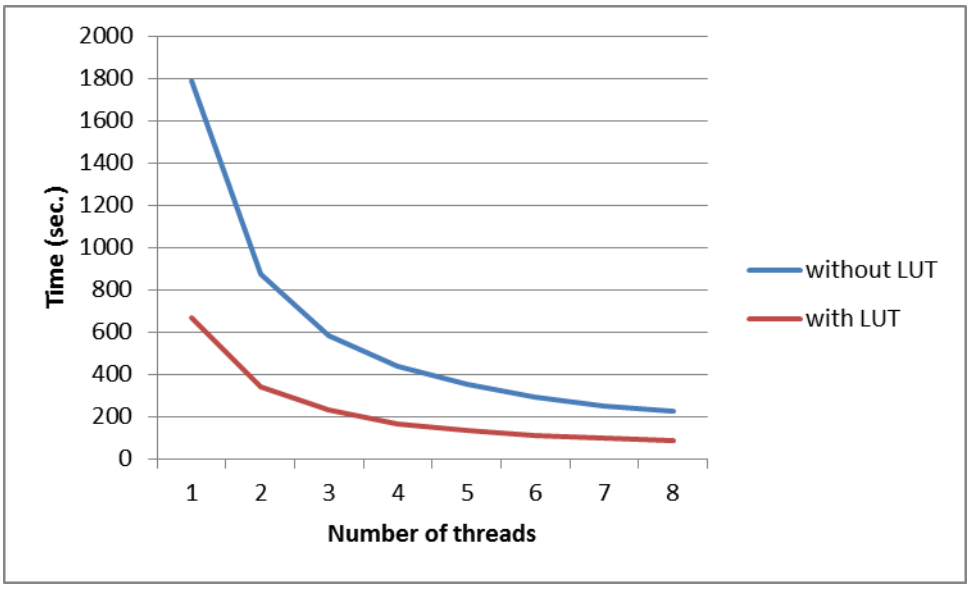



図 3　接尾辞配列の一部に対してルックアップテーブル（LUT）を導入することによる検索速度の改善。横軸は使用スレッド数、縦軸は経過時間（秒）を示す。クエリに土壌メタゲノム 40000 リード（リード長 75)、256MB の配列データベースを用いた。

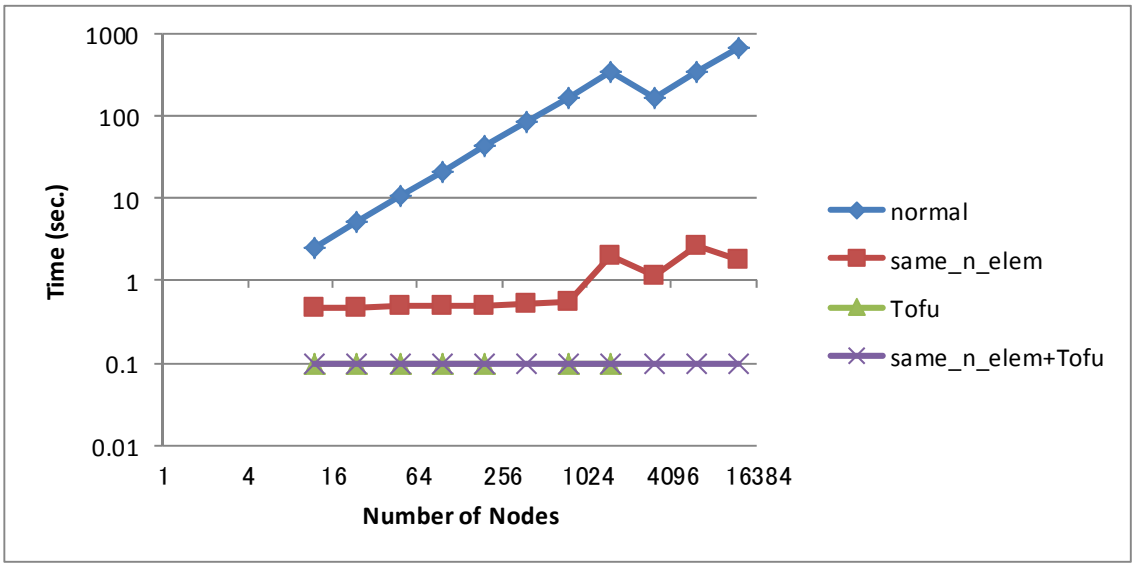



図 4　1GB のデータを送受信した際の経過時間に対して、使用ノード数をプロットした。それぞれ、normal: 通常の MPI_Bcast、same_n_elem: 要素数が全ランクで同じことを保証した場合の MPI_Bcast、Tofu: 3 次元トーラスをセグメント分割で通信の衝突を避ける Tofu 専用アルゴリズムを表している。

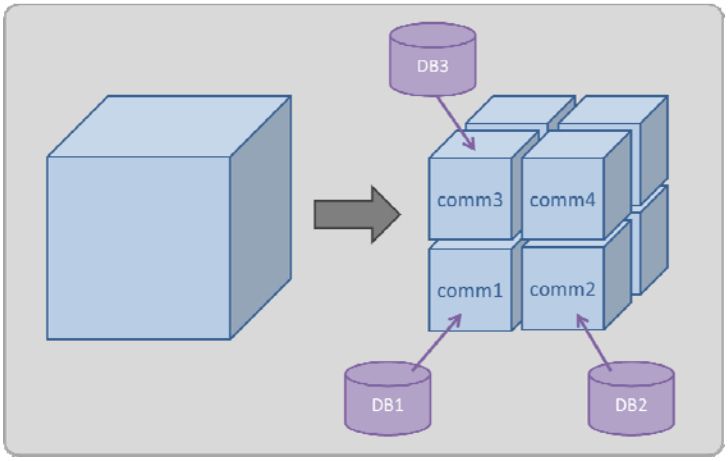



図 5　配列データベース分配時の MPI コミュニケータの分割の模式図。各コミュニケータが仮想ネットワーク座標上で 3 次元直方体になるようコミュニケータを分割し、そのコミュニケータ内の全計算ノードに同一のデータベースサブセットを割り当てる。

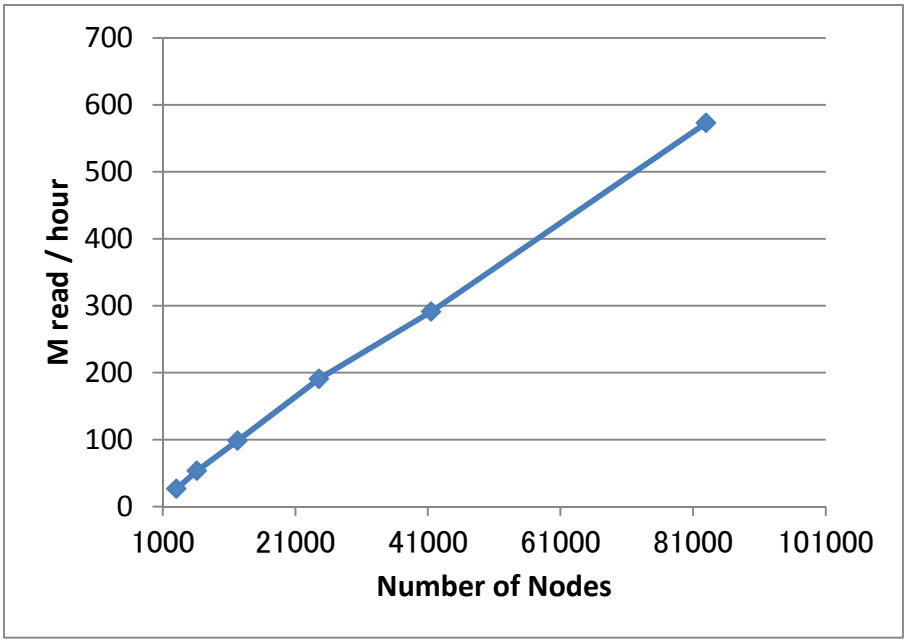



図 6　相同性解析プログラムの並列性能（ウィークスケーリング）。ヒト口腔内細菌叢の次世代シーケンサのリード配列をクエリとして、KEGG genes 配列データベースに対し検索を行った。縦軸は時間当たりの処理リード数、横軸は使用ノード数を示す。

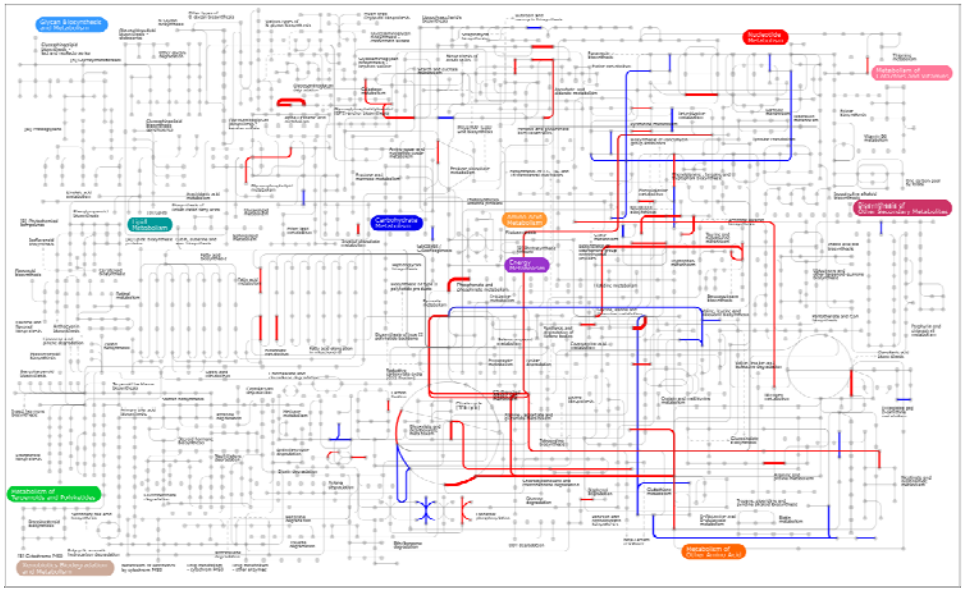


図 7　遺伝子の相対存在度の比較。歯肉縁下の歯垢と歯肉縁上の歯垢の間で相対存在度の異なる KEGG Orthology（KO）を iPATH2（http://pathways.embl.de/）によって代謝経路上に可視化した。歯肉縁下の歯垢で相対存在度の高い KO を赤、歯肉縁上の歯垢で相対存在度の高い KO を青で示した。

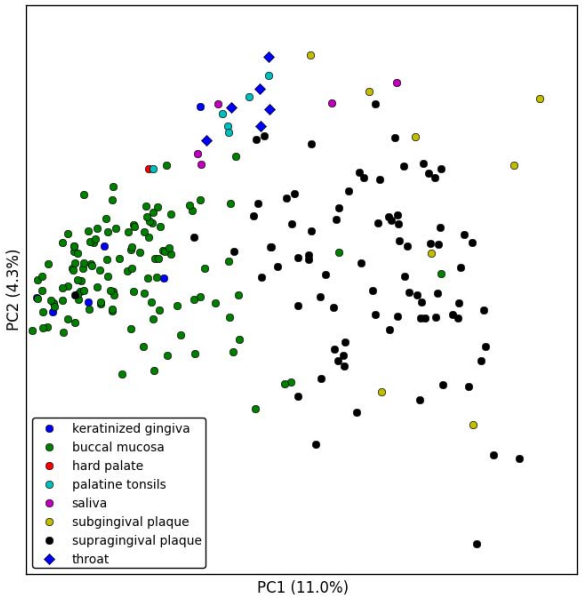



図 8　オーソログ遺伝子の多寡に基づいた主成分分析。口腔前庭（角化歯肉、頬粘膜）、狭義の口腔（硬口蓋、口蓋扁桃、咽喉、舌背、唾液）、歯垢（歯肉縁上の歯垢、歯肉縁下の歯垢）で分かれて分布している様子が見られる。

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２４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活用し、２次構造的なエネルギーを考慮した RNA 立体構造予測技術の開発と、大規模データへの適用を行う。

Ⅳ－３－２　実施内容（成果）

（１）ソフトウェアの開発・高度化の状況

１）精度、速度で BLAST を上回る配列検索・アラインメントツールである LAST と、分子内と分子間の両方の２次構造を同時に考慮した相互作用予測ツール RactIP を、京コンピュータ上に実装して稼働させた。LAST は特に塩基配列の類似性検索に有効なソフトウェアであり、次世代シークエンサーからのリード配列の相同性検索に関しても、利用価値が高いソフトウェアである。

２）京コンピュータを活用した RNA 相互作用予測パイプラインを開発した。パイプラインの部品としては、RNA の１分子内で２次構造を形成しにくく他の分子が相互作用しやすいアクセシビリティ計算ツール CapR、相同性検索ツール LAST、RNA 相互作用予測ツール RactIP を用いた。

３）２次構造的なエネルギーを考慮した RNA 立体構造予測手法としては、２次構造予測結果を考慮したフラグメントアセンブリ手法（RASCAL）を京コンピュータに実装して稼働させた。

（２）研究開発の実施状況

１）2 次構造を考慮した配列情報解析アルゴリズムを活用し、２次構造的なエネルギーを考慮した RNA 立体構造予測手法を開発した。また、2 次構造を考慮した解析手法としは、RNA 配列の塩基を置換した場合のエネルギー、エントロピー変化を計算する手法・ソフトウェア（Rchange）を開発し、誌上発表した。

２）RNA 相互作用予測手法の開発では、京コンピュータを活用したパイプラインを開発し、大規模データに適用した。各 RNA 分子の相互作用部位の候補のスクリーニングに CapR、それらの部位のうち RNA-RNA で相互作用する相補配列のスクリーニングに LAST、最終的なスクリーニングに RactIP を使用したパイプラインを開発した。作成したパイプラインを Gencode プロジェクトの long non-coding RNA transcript に対して網羅的な解析を行った。その結果、いくつかの興味深い RNA-RNA 複合体構造の候補を発見した。

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。

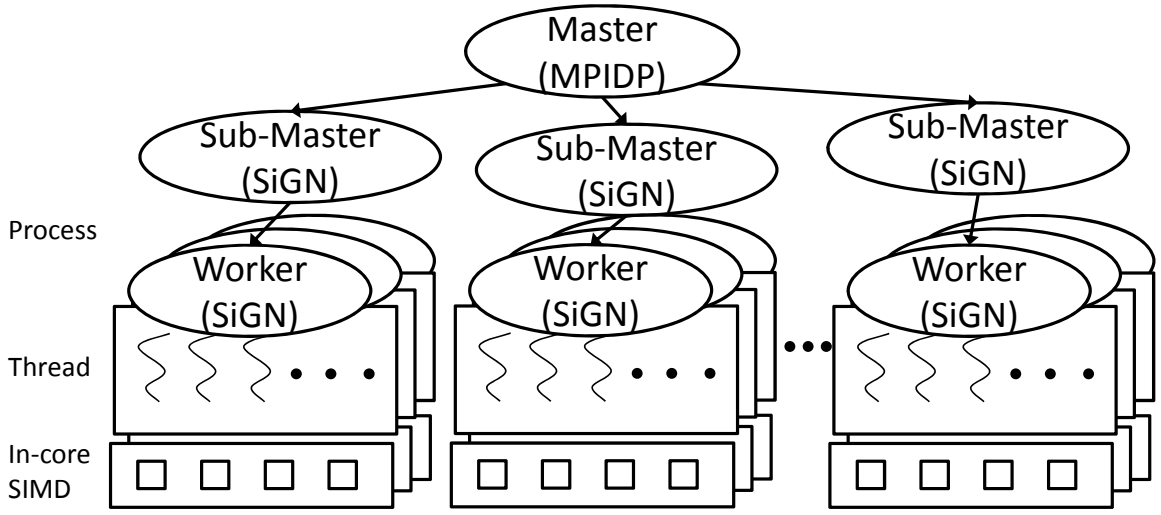
平成２４年度は、平成２３年度に開発したプロトタイプソフトウェアに負荷分散機能を加えることで、さらに大規模な HPC 環境で実行できるように拡張し、多数の条件下での多様な生体分子ネットワークの解析を一度に行うための研究開発を実施する。

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における負荷分散機能の実現

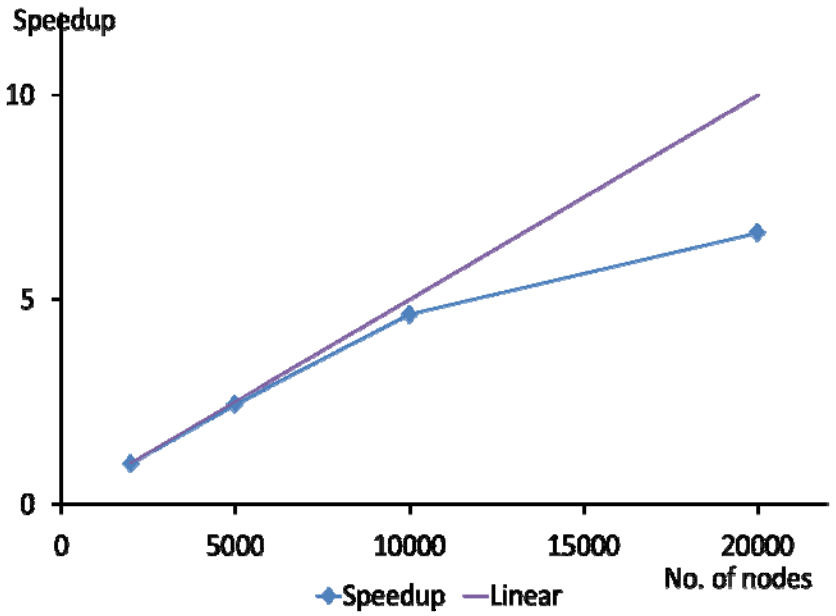
平成２３年度に開発したプロトタイプソフトウェアに対して、東京工業大学の秋山研究室で開発された負荷分散ツール MPIDP を結合することで、複数の条件下での生体分子ネットワークの解析を一度に行うための負荷分散機能を実現した（下図参照）。




２）生体分子ネットワーク解析技術の大規模化

前述の負荷分散機能を利用して、マウスの脂肪細胞組織に寒冷刺激を与えたときに発現が誘導された約１万個の遺伝子についての生体分子ネットワーク解析を「京」上で行った。このときの実行時間は、「京」の 2000 ノードでは 49659 秒かかったのに対して、5000 ノードでは 20311 秒で 2000 ノードの時間に対して約 2.4 倍の速度向上（並列化効率 97.8%）、10000 ノードでは 10715 秒で 2000 ノードの時間に対して約 4.6 倍の速度向上（並列化効率 92.7%）を達成した。さらに、20000 ノードにまで上げると実行時間は 7493 秒となり、速度向上は約 6.6 倍（並列化効率 66.3%）に留まった。以上の結果を下図に示す。2000 ノードで 14 時間近くかかっていた計算が、20000 ノードで約 2 時間にまで短縮できた。

このように、大規模な生体分子ネットワークの解析を、「京」の 10000 ノード以上を使って多数実行するソフトウェア環境が構築できたことで、本グループの平成２４年９月末の供用開始から２５年３月末までの累計実行時間は約 128 万ノード時間積となっている。




（２）研究開発の実施状況

１）マウス脂肪細胞組織の寒冷刺激誘導下での生体分子ネットワークの解析

褐色脂肪組織は通常の細胞の約 100 倍の熱を産生するが、ヒト成人においては中年期における肥満が、加齢による褐色脂肪細胞の減少と関連することが疫学的な研究により示されており、褐色脂肪細胞の増加は肥満およびそれに起因する生活習慣病の是正に向けて重要であるとされている。

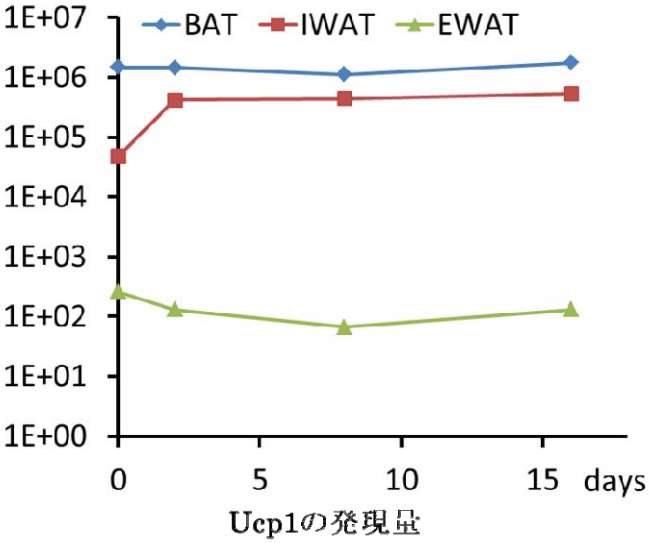
脂肪組織としては、マウス個体の、鼠蹊部の皮下白色脂肪組織（IWAT)、精巣部の内臓白色脂肪組織(EWAT)、肩甲骨間の褐色脂肪組織(BAT)の３種類の組織を対象とした。マウスの IWAT は、通常は白色であるが、低温で飼育を続ける寒冷刺激を与えると BAT のように褐色化して熱産生が増加する。このように褐色化した白色脂肪細胞を、特にベージュ細胞と呼ぶ。EWAT では寒冷刺激を与えても IWAT のような変化が見られない。最近、Harvard Medical School の Spiegelman らのグループにより、ヒト成人の BAT は、マウスの BAT よりも IWAT が褐色化したベージュ細胞に近い性質をもつことが報告(Wu et al., Cell, 2012)されており、マウスの IWAT の研究が注目を集めている。

そこで、マウスの個体に対して、寒冷刺激を与えた時の脂肪組織の遺伝子発現の変動を計測した。刺激を加える前、刺激後２日、８日、１６日の３種類の脂肪細胞組織から mRNA を採取し、マイクロアレイおよび RNA-Seq の時系列発現プロファイルを取得した。個体間のばらつきを考慮して、同一条件下でそれぞれマウス３個体から別々に mRNA を採取し発現プロファイルを取得している。

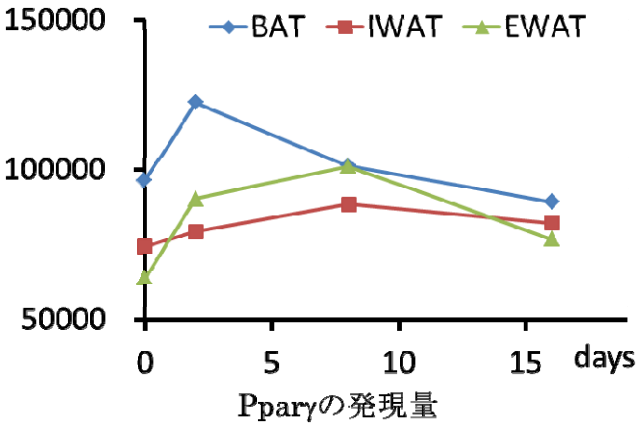
脂肪組織の熱産生で最も重要な役割を果たす遺伝子である Ucp1 は、マウスの脂肪組織においては、下図左のように EWAT では寒冷刺激を受けてもほとんど発現しないが、BAT では寒冷刺激の有無に関わらず定常的に強く発現し、IWAT では寒冷刺激により強く誘導されて発現量が急激に上昇していた。マウスの IWAT における寒冷刺激での熱産生の増加は、この Ucp1 の発現誘導がカギを握ると考えられ、そこで働く生体分子ネットワークを解明できれば、ヒト成人における熱産生の増加についての知見が得られることが期待される。

一方、脂肪細胞で最も代表的な転写因子である Pparγの発現量は、各脂肪細胞組織で寒冷刺激が加わったとき、下図右にあるように、BAT ではいったん発現が上昇するがすぐに低下し、EWAT でもやはり発現が上昇するが BAT より遅く８日目以降に発現が低下し、IWAT ではあまり発現の上昇が見られないなど、応答の違いが見られる。

このように、寒冷刺激に対する Ucp1 と Pparγの発現変化には大きな違いが見られることから、脂肪細胞で従来知られていた転写因子のネットワークとは異なるネットワークが働き、IWAT における Ucp1 の発現上昇と熱産生の増加に寄与しているのではないかと考えられる。



Ucp1の発現量

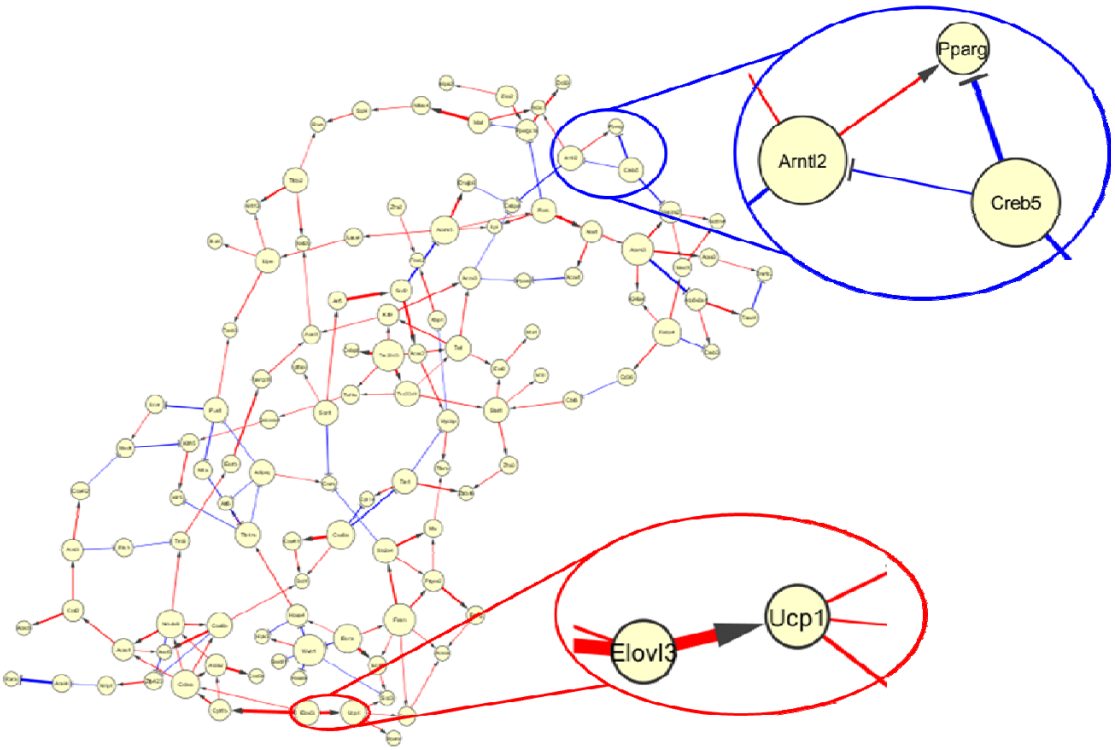


Pparγの発現量

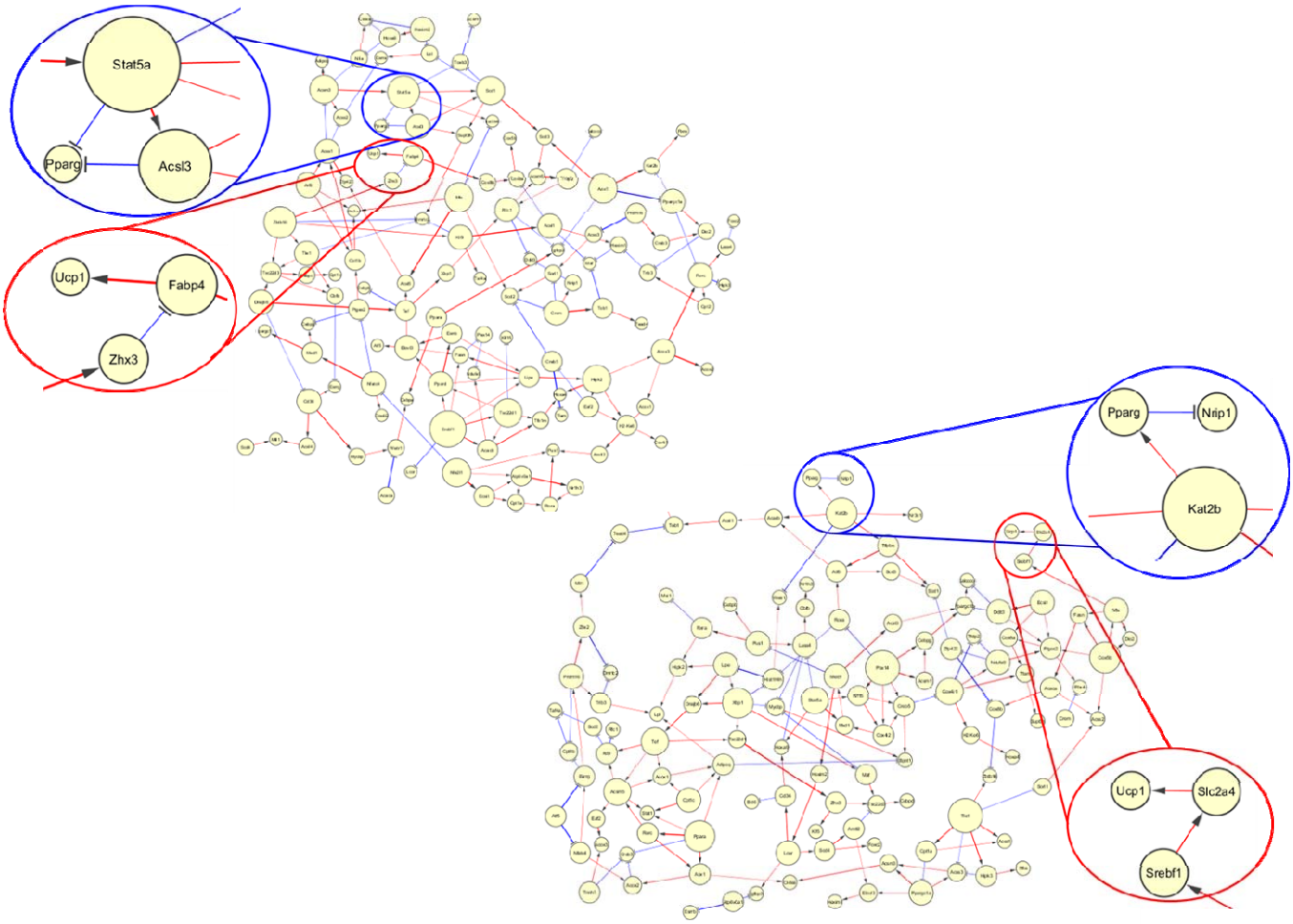
そこで、マウスの各脂肪組織における寒冷刺激下での応答を、脂肪組織で働くことが既知の130個の遺伝子と、寒冷刺激により発現変動が見られた約1万個の遺伝子について、それぞれの時系列発現プロファイルをもとに、今年度開発したソフトウェアを使ってネットワーク解析を行った。

本研究では、生体分子ネットワークをダイナミックベイジアンネットワークの手法を使って推定するが、まず、比較的少数の既知遺伝子で構成されるネットワーク（これをシードネットワークと呼ぶ）を、既存の生物学的知識や実験から得られている既知の制御関係を事前確率として与えて構築し、そのときの推定結果をもとにして、より大規模の遺伝子のネットワークの推定を行う。

シードネットワークの解析結果を下図に示す。下図で、赤い線は活性化、青い線は抑制の制御辺を表す。ノードの大きさは、その遺伝子が制御している遺伝子の数(Hubness)を表し、大きいほど制御する遺伝子数が多い。制御辺の太さは、3個体で別々に取得した発現プロファイルについて、ブートストラップによる統計的サンプリングでの信頼確率を反映しており、太いほど確率が高い。Ucp1とPparγの周囲の部分を、それぞれ赤と青で拡大表示している。



皮下脂肪組織(IWAT)の寒冷刺激下でのシードネットワーク

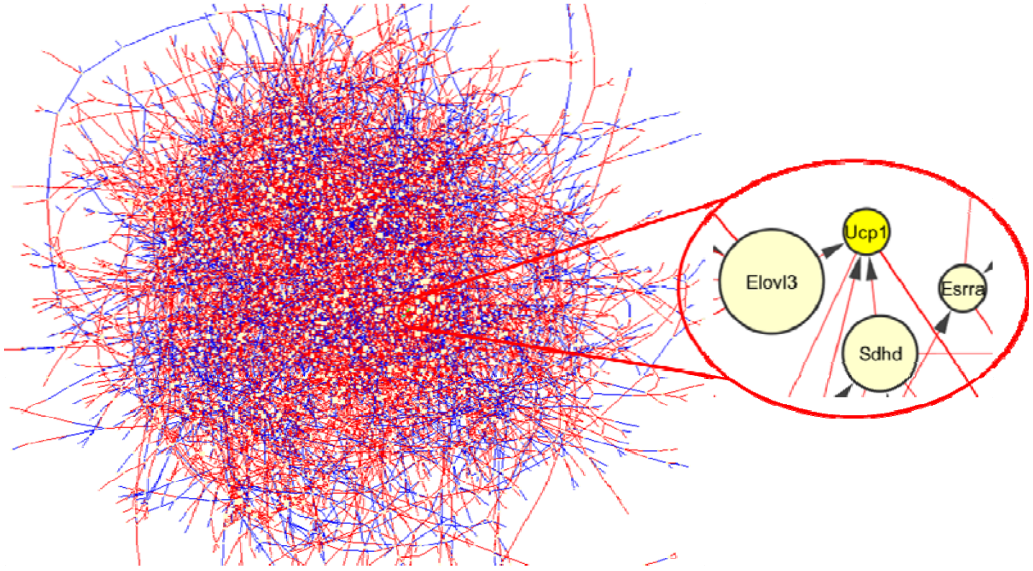



**褐色脂肪組織（BAT, 左）と内臓脂肪組織（EWAT, 右）の寒冷刺激下でのシードネットワーク**

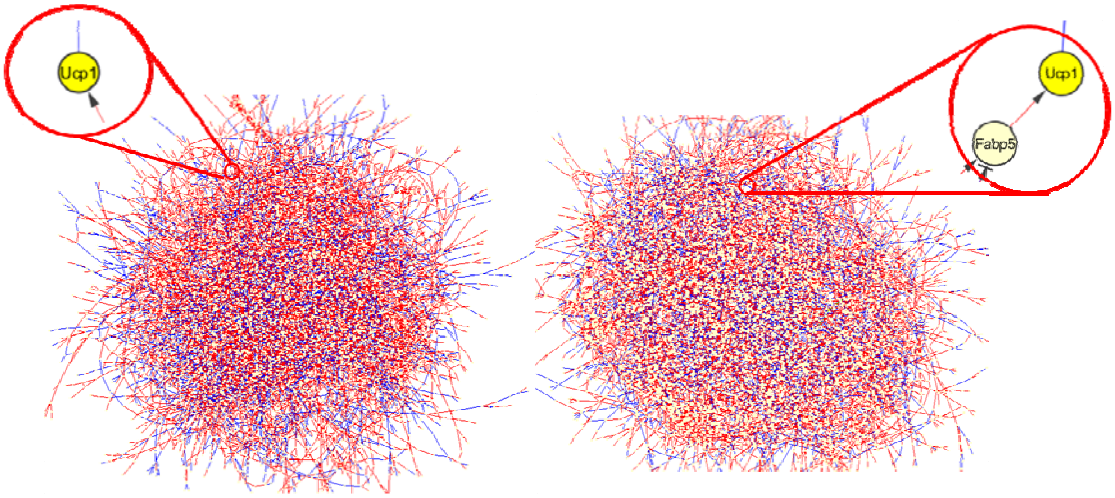
上の図で、Ucp1 の周辺では、IWAT では太い（ブートストラップ確率が 86.3%）活性化制御辺が Ucp1 につながっている。BAT でもブートストラップ確率が 67.8%の活性化制御辺が Ucp1 につながっているが、EWAT では Ucp1 につながっている活性化制御辺のブートストラップ確率は 28.5%に留まっている。Pparγの周辺では、IWAT では活性化と抑制の制御辺が、BAT では抑制の制御辺のみであり、上図の寒冷刺激の応答ネットワークでは抑制により Pparγの発現が抑えられていることが示唆される。EWAT では活性化の制御辺のみであるが、ブートストラップ確率は 23.0%と低い。

次に、下図に 3 種類の脂肪組織について、寒冷刺激により発現変動のあった約 1 万個の遺伝子についてネットワーク解析を行った結果を示す。Ucp1 の周辺の部分を赤で囲んで拡大している。IWAT では、多くの活性化制御辺が Ucp1 とつながっているのに対して、BAT や EWAT ではどちらも 1 本の活性化制御辺のみであり、IWAT では寒冷刺激により Ucp1 の大きな発現上昇が見られたが、BAT や EWAT ではほとんど発現に変動がなかったことと一致している。

今後は、IWAT でのネットワークを詳細に解析することにより、寒冷刺激により IWAT がベージュ細胞に転換していく機構について関与する生体分子を明らかにしていくことで、脂肪組織での熱産生の増加に向けての知見を得る予定である。




皮下脂肪組織(IWAT)の寒冷刺激下での大規模ネットワーク




褐色脂肪組織（BAT, 左）と内臓脂肪組織（EWAT, 右）の寒冷刺激下での大規模ネットワーク

Ⅳ－４　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－４－１　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行った。

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った。

平成 24 年度の研究では、本研究における成果目標のうち、京コンピューターの計算資源を活用し、がん転移のプロセスに関するモデルを公開済みの SRA 配列を用い集団遺伝学的に解析することを目標とした。これらのデータを用い、ゲノムワイドに特定された変異サイトから細胞塊ごとの系統樹を作成することによって、使用データの症例において、ガン細胞集団において集団遺伝学的効果が示唆された。上記結果について国際学会にて報告を行った(N.V. Sasaki *et al*, the 3rd AICS International Symposium, Kobe)。

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）最尤(ゆう)系統樹推定プログラムのチューニング(RAxML)

昨年度までのチューニング作業によって得られた性能が維持されていることを確認し、1,000 ノードのスケールでの計算が実行可能であることを確認した。また、ステージングへの対応を行った。

（２）研究開発の実施状況

１）系統樹解析のがん転移モデル検証への応用

がん細胞の浸潤転移は、幾つかの過程によって構成されていると考えられているが、これらの過程を経ることにより、潜在的な転移ガン細胞に対し、大規模な減耗が起きることが知られており、この減耗過程のメカニズムを解明することは、がん研究の今日的な課題である。このプロセスを説明するためにいくつかのモデルが提唱されているが、がんの種類によって、どのモデルが優勢であるかが異なる可能性があり、対象とするがんがどのモデルに適合するかは、治療計画の設計の上で医療の見地からも重要な意味があるといえる。上記モデルの違いは、標的遺伝子等への突然変異が蓄積される時期の違いとボトルネック作用の起きる時期の違いによって、集団遺伝学的なアプローチが可能であり、その部分において系統樹解析の活用の機会が考えられた。そこで、公開済みの SRA 配列からゲノムワイドで変異サイトを探索し、細胞塊ごとの系統樹を作成し、細胞系譜から推定される原発巣と転移巣の多様性指標の違い等から、対象とした症例がどの上記モデルに適合するかを推定した。

がん細胞の浸潤転移の過程は、大別すると (1)ECM (細胞外基質)および間質層への浸潤、(2)血管系への侵入と循環中での細胞の生存維持、(3)離れた組織実質への侵入と生存維持、(4)増殖プログラムの再始動、以上 4 つの過程を経て腫瘍性新生物が生成すると考えられている。これらの転移過程において注目すべき点は、これらの過程によって、潜在的な転移能を持ったがん細胞に対し最終的に 99.3%もの減耗が起きることである。より詳細には、(2)までは 20%の減耗であるが、(3)の段階までに 96%もの大規模な減耗が起きる事が報告されている[1]。このことは、転

移先の微小環境への適応が腫瘍新生の律速段階であることを示している。現在これらの過程を説明するために提唱されているモデルのなかで有力なものは、(A)Cell-of-Origin Model と(B)Stochastic Model と呼ばれるものがある。これらのモデルの大きな違いは、血管系に侵入するがん細胞が転移能を獲得する時期の違いと集団遺伝学的なボトルネック効果の起きる時期の違いであるが、近年まではこれらの差異をゲノムワイドに実証的に検証する方法がなかった。

そこで本研究では、目的となる症例が上記モデルのどれに適合するかを推定するために、これまでに「京」コンピューター上に実装した系統樹推定プログラムである RAxML を利用し、公開済みの SRA(Short-Read Archive)配列から、ゲノムワイドで変異サイトを探索し、細胞塊ごとの系統樹を作成することによって、細胞系譜から推定される原発巣と転移巣の多様性指標の違い等を検出することを目標とし、研究を行った。

その結果、肺がんから転移を起こした患者から得られたがん細胞の SRA データから、対照区となる血中細胞と比較したとき転移巣のがん細胞において、塩基多様度が低くなっていることが示された(図 1)。このことは、前述の転移における減耗の集団遺伝学的効果を間接的に示唆する結果であり、この解析によって得られた知見を国際学会において報告した。(N.V. Sasaki *et al*, the 3rd AICS International Symposium, Kobe)

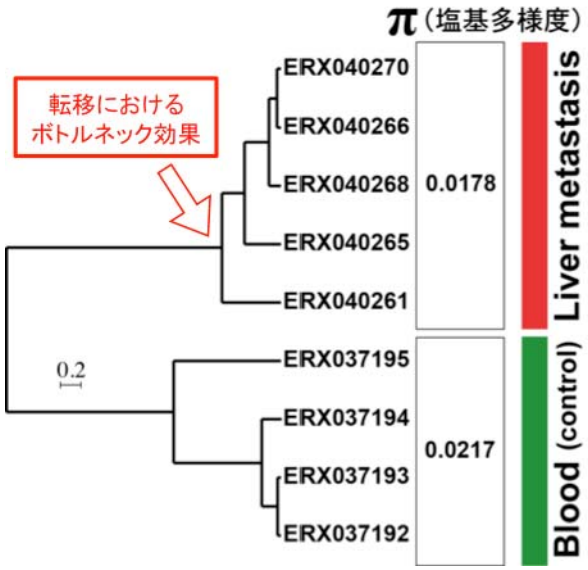



図 1：塩基多様度と最尤系統樹の関係

本研究で明らかになった大規模生命データを利用する際の課題として、今回開発に利用した SRA 等の大規模な配列データの解析を行うために必要な、プリ・ポスト処理の問題があげられる。今後の「京」コンピューターを含めた大規模計算機を利用する際には、メインとなる計算プログラムの解析対象となる入力および出力データのプリ・ポスト処理に必要な計算機資源を、どのように確保するかが重要であると考えられる。

(参考文献)

[1] Scott Valastyan and Robert A. Weinberg, “Tumor Metastasis: Molecular Insights and Evolving Paradigms,”, *Cell*, 147:275 2011