ゲノム情報からのビッグデータの解析

大阪大学　大学院情報科学研究科
松田秀雄

本講演の流れ

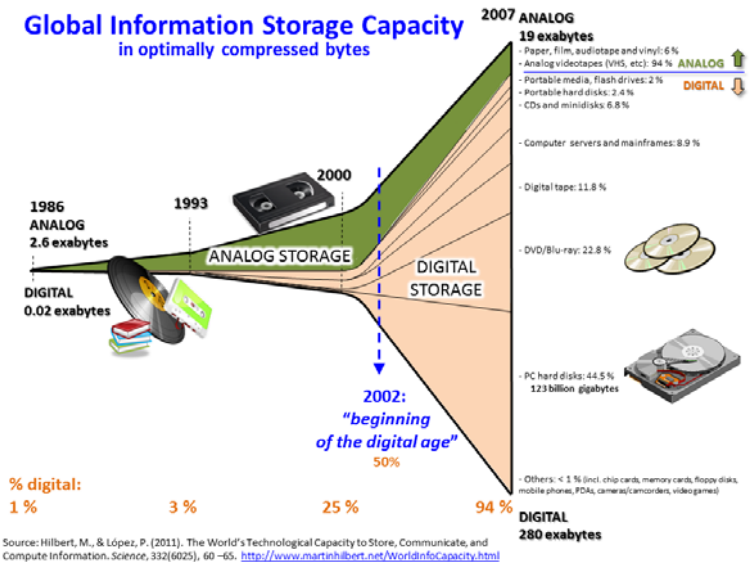
- ゲノムビッグデータについて
  - –次世代シーケンサーによるゲノムビッグデータの生成
- ゲノムビッグデータの活用事例
  - –京による大規模生命データ解析

本講演の流れ

- ゲノムビッグデータについて
  - –次世代シーケンサーによるゲノムビッグデータの生成
- ゲノムビッグデータの活用事例
  - –京による大規模生命データ解析

ビッグデータについて

- ビッグデータとは、典型的なデータベースソフトウェアが管理できる能力を超えたサイズのデータを指すとされている
- 具体的なサイズとして利用の形態に依存するが、数十テラバイトから数ペタバイトの範囲に及ぶとされている（テラ＝10^{12}、ペタ＝10^{15}）



Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

http://en.wikipedia.org/wiki/File:Hilbert_InfoGrowth.png

米国のビッグデータ・イニシアティブ

- 2012年3月29日に米国政府がビッグデータ関連の総額2億ドル以上を投じた研究開発イニシアティブの概要を発表
- このイニシアティブでは、米国の政府機関6機関が主導して、巨大なデジタルデータの組織化やそこからの知識抽出等を行うための技術やツールの開発を行うとされている

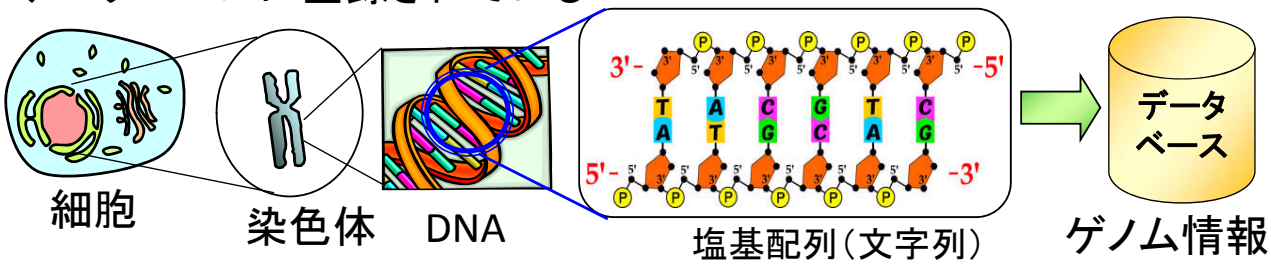
1. 巨大な量のデータの収集、保存、運用、分析、共有に必要な中核技術の進歩
2. 科学技術分野での発見速度の加速や、国家安全保障の強化、教育・学習の変化への当該技術の活用
3. ビッグデータ技術の発展・活用に必要な労働人口の拡大

ライフサイエンス関連のビッグデータ・プロジェクト

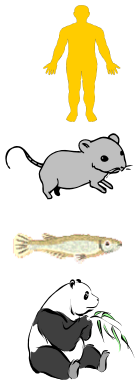
- 1000ドルゲノム
  - ヒトゲノム（約30億塩基）を読み取るコストを2013年に1000ドル、2020年には100ドル以下にすることを目標に公的資金を投入
  - 個別化医療(personalized medicine)を意図
- BD2K (Big Data to Knowledge)
  - 生物学・医学分野でのビッグデータの解析拠点、解析手法およびソフトウェア開発プロジェクト、人材育成プロジェクト等の多数の公募を、今年一斉に開始
  - NIHの直轄プロジェクトとして、Philip Bourneを Associate Directorに選定

ゲノム情報のイメージ

数千種以上の**ゲノム**（生物の持つすべてのDNA塩基配列）の情報がデータベースに登録されている



計算機でゲノム情報を比較することにより、各生物の特徴や生物間の違いを探る



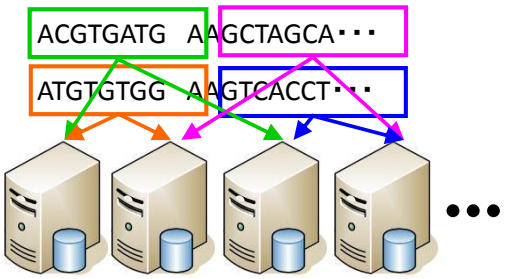
ゲノムの例（生物、発表年）	サイズ（DNA塩基数）	
ヒト (2003)	30億	ACTGA・・・
マウス(2002)	25億	AGTCT・・・
メダカ(2006)	8億	CATCT・・・
パンダ(2010)	24億	GTACA・・・

問題点

単純な比較法ではゲノムサイズの積に比例する処理時間がかかってしまう

解決方法

効率の良い文字列比較アルゴリズム
多数の計算機を使って並列に比較



sickケアからhealthケアへ 遺伝子検査サービス

- DeNAと東大医科学研究所との共同事業（COI STREAM）
- 運営会社を2014年4月に設立
- 遺伝子検査は運営会社が実施し、検査で使用する遺伝子解析ソフトウェアや、結果の開示・解説について、東大医科研が共同開発

https://mycode.jp/howitworks.html

1000ドルゲノムシーケンサーの発表

- 2014年1月に発表
- シーケンサー1台で1回1.6～1.8T塩基（ヒトゲノム3G塩基×約20人×冗長度 約30）
- シーケンサー10台をセットで販売
- 消耗品だけでなく人件費や装置の減価償却費も含めて1000ドル

http://www.illumina.com/systems/hiseq-x-sequencing-system.ilmn

次世代シーケンサーの「世代」

- 既に「次世代」ではなくて「現世代」
- 第１世代シーケンサー
  - 蛍光標識した塩基配列をサンガー法（酵素反応）で合成
  - １回当たり数本～数百本×５００塩基程度
- 第２世代シーケンサー
  - 塩基配列を合成する原理は第１世代とほぼ同じ
  - 蛍光・発光など光検出により、超並列で塩基配列を決定
- 第３世代シーケンサー
  - １分子リアルタイム・シーケンシング（以前はこれだけ第３世代と呼ばれていた）
  - 蛍光を使わないシーケンシング（第４世代シーケンサーとも呼ばれる、半導体シーケンサーなどが含まれる）

第１世代：サンガー法

- DNA複製酵素を使って塩基配列を合成（１回で数本～数百本の配列読み取り）
- 蛍光標識した塩基を混ぜておく

第２世代シーケンサーの原理

- 塩基を酵素反応で合成して、蛍光で標識して読み取るところは第１世代シーケンサーと同じ
- 一度に読み取れる配列の本数が膨大（>G本）
- 超並列シーケンサー、ギガシーケンサーとも呼ばれる

第2世代：超並列シーケンサーの例

8レーン×2カラム×50タイル×4画像×3M塩基×36サイクル=345.6 G塩基

第3世代シーケンサー（半導体シーケンサー）の原理

- 酵素反応＋蛍光スキャナーで塩基配列を読み取るのではなく、半導体チップを使って塩基配列を読み取る
- イオンセンサー方式
  - 酵素反応で塩基を取り込むところは同じ
  - 塩基を1種類ずつ順番に流し、結合時に発生する水素イオンにより生じるpH変化をイオンセンサーで検出する
  - イオンセンサーは、半導体(CMOS)チップ上のウェル（内径3.5μm）の底に配置
- ナノポア方式
  - 塩基配列を狭い穴（nanopore）に配置した電極の間を通過させ、近接した塩基のトンネル電流の違いを検出
  - 酵素反応を使わず塩基を直接読み取る

次世代シーケンサーのデータ解析

- de novoシーケンス：完全に1からシーケンスして、得られた配列断片をつなぎ合わせる（アセンブル）→長い配列断片向き
- リシーケンス：得られた配列断片を、既に読まれている参照配列に張り付ける（マッピング）→短い配列断片向き
- RNA-Seq: DNAから転写されたRNAを読み取る（読み取り後にアセンブルかマッピングをする）
- ChIP-Seq: タンパク質が結合したDNAの部分を読む、クロマチンのヒストン修飾がわかる

→ データ量がとにかく大きいので、高速なアルゴリズムが求められる（de Bruijnグラフ、接尾語配列、BWT）

シーケンシング（ゲノムの読み取り）のコストの劇的低下の意味

- *遺伝子検査を情報科学の問題に還元できる!*

→ スパコンの利用による加速

  - ミスマッチを許しつつ、非常に高速な文字列照合
  - 膨大な組合せの中から健常者と患者を分離できるパターンを求めるデータマイニング

- 個別化医療に向けての期待が大きい
  - ある種の疾患はゲノムの変異による影響が大きい（例：乳がんでのアンジェリーナ・ジョリーのケース）
  - 特定の疾患に特徴的な変異が見つかれば治療に結びつく可能性がある（個人に合わせた薬剤の設計）

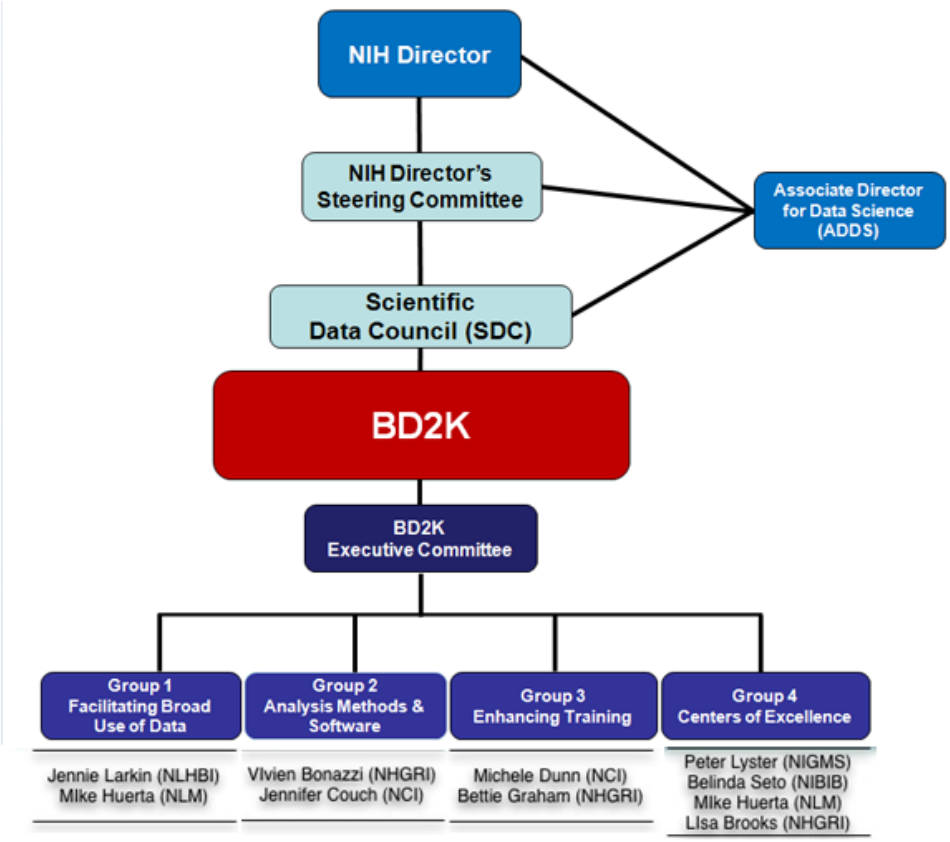
今後のゲノムのデータ量

- 次々世代シーケンサーのデータ量
  - 第3世代シーケンサー　1サンプル　1～　10TB
  - 第4世代シーケンサー　　〃　　10～100TB
- ゲノムコホート研究
  - 現状　10,000人 → 将来　＞100,000人
- 単純に掛け合わせると‥
  - 第4世代シーケンサー×次世代ゲノムコホート
    = **10EB!**

本講演の流れ

- ゲノムビッグデータについて
  - 次世代シーケンサーによるゲノムビッグデータの生成
- ゲノムビッグデータの活用事例
  - 京による大規模生命データ解析

NIHのBD2K (Big Data to Knowledge)



http://bd2k.nih.gov/about_bd2k.html

NIH BD2Kの2014年度計画

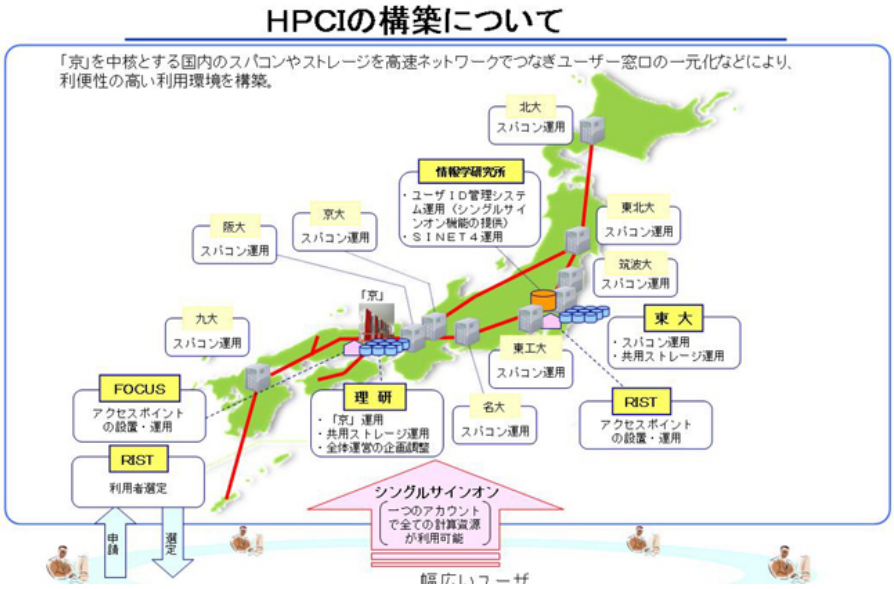
- １０月９日に、2014年度予算 総額32Mドル（35億円相当）の対象を発表

http://bd2k.nih.gov/FY14.html

- ビッグデータ・コンピューティング・COE（１１拠点）
- LINCS（ Library of Integrated Network-based Cellular Signatures） Perturbation データコーディネーションセンター
- データ探索インデックスコーディネーションコンソーシアム
- 人材育成・トレーニング

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）

- 次世代スーパーコンピュータ「京」の開発・整備
- HPCIの整備・運営
- HPCI戦略プログラム（1．生命科学・医療・創薬、2．新物質・エネルギー、3．防災・減災・地球変動予測、4．ものづくり、5．物質と宇宙の起源・構造）

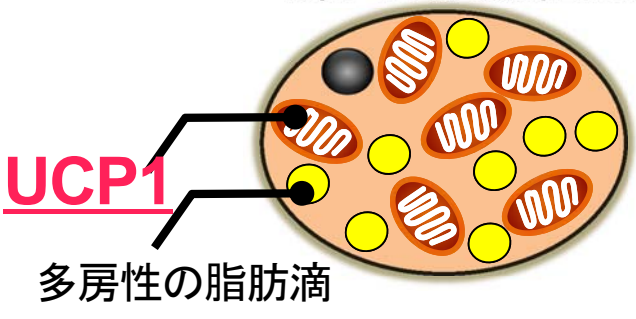


白色脂肪細胞と褐色脂肪細胞
（京大 河田教授グループとの共同研究）

白色脂肪細胞（第1の脂肪細胞）

エネルギー貯蔵・放出

褐色脂肪細胞（第2の脂肪細胞）



脂肪酸酸化・熱産生

熱産生能力は骨格筋と比較すると**100倍**高い（この熱産生能力は**UCP1**に起因することが知られている）

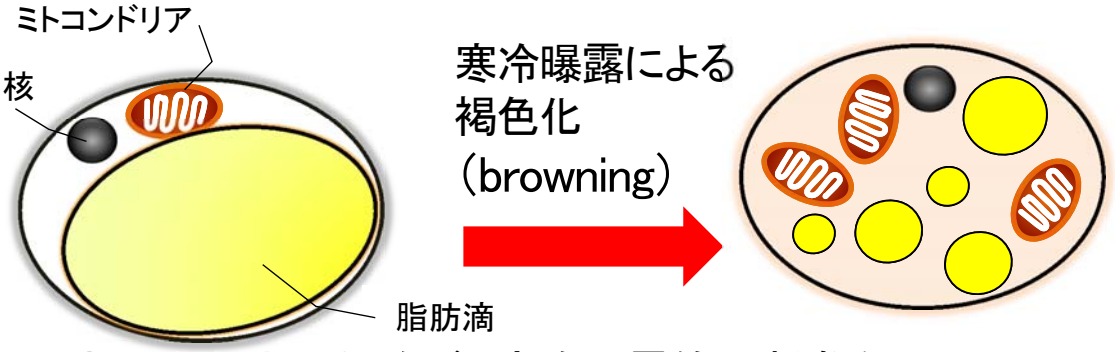
褐色脂肪細胞のヒト成人の存在は、2009年にPETによる測定で、2013年に**解剖学的にその存在が確認された**

L. E. Lidell et al. Nat Med 2013 May,
A. M. Cypess et al. Nat Med 2013 May

刺激により白色脂肪細胞が褐色化

白色脂肪細胞
（第1の脂肪細胞）

ベージュ脂肪細胞
（第3の脂肪細胞）

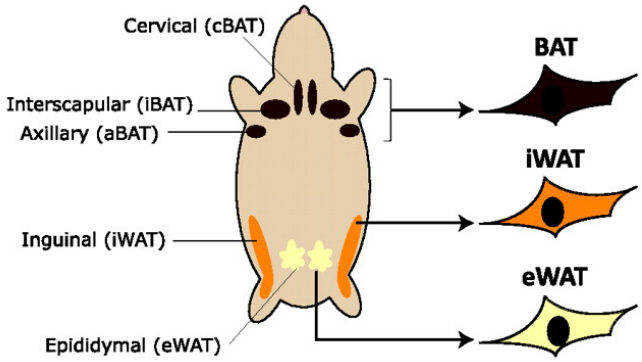


- 通常の白色脂肪細胞が、寒冷曝露等の刺激を受けると**UCP1の発現が誘導**され**褐色化（browning）**が起こり、ベージュ脂肪細胞へと変化することで、褐色脂肪細胞と同様の**熱産生**が生じる
- 褐色化をもたらす細胞の状態変化を、刺激応答での生体分子ネットワークを解析することで解明

新しい視点からの肥満是正の戦略につながる

白色脂肪細胞の褐色化（Browning）

- マウスに対して寒冷刺激（4℃の環境におく）を加えると、ある種の白色脂肪細胞（IWAT: 鼠蹊部の皮下脂肪細胞）は褐色化するが、別の白色脂肪細胞（EWAT: 精巣内の内臓脂肪細胞）は褐色化しない
- 元から褐色になっている（古典的）褐色脂肪細胞（BAT: 肩甲骨にある褐色脂肪細胞）と遺伝子発現
- プロファイルや遺伝子ネットワークを比較してこの違いが何に起因するかを明らかにする



熱産生に重要な遺伝子の発現変化

- 褐色脂肪細胞(Brown)では寒冷刺激の有無に関係なくUCP1が高発現している
- 同じ白色脂肪細胞でも、皮下脂肪細胞(Beige)では寒冷刺激とともにUCP1の発現が上昇するのに、内臓脂肪細胞(White)では発現がみられない

UCP1の発現上昇についての定説

- 寒冷刺激が脳に伝わり、交感神経を通じて血中にアドレナリンが分泌され、それを感知して発現が上がる

→細胞の種類ごとの違いを説明できない

遺伝子発現プロファイルの取得

- マウス個体の白色・褐色・ベージュの3種類の脂肪組織から、寒冷刺激前、刺激後1, 2, 4, 8, 12, 24（1日）, 48（2日）, 192（8日）, 384（16日）時間経過時のtotal RNAを取得
- マイクロアレイ(Agilent Mouse)とRNA-Seqで発現プロファイルを取得
- RNA-Seqのデータ量:

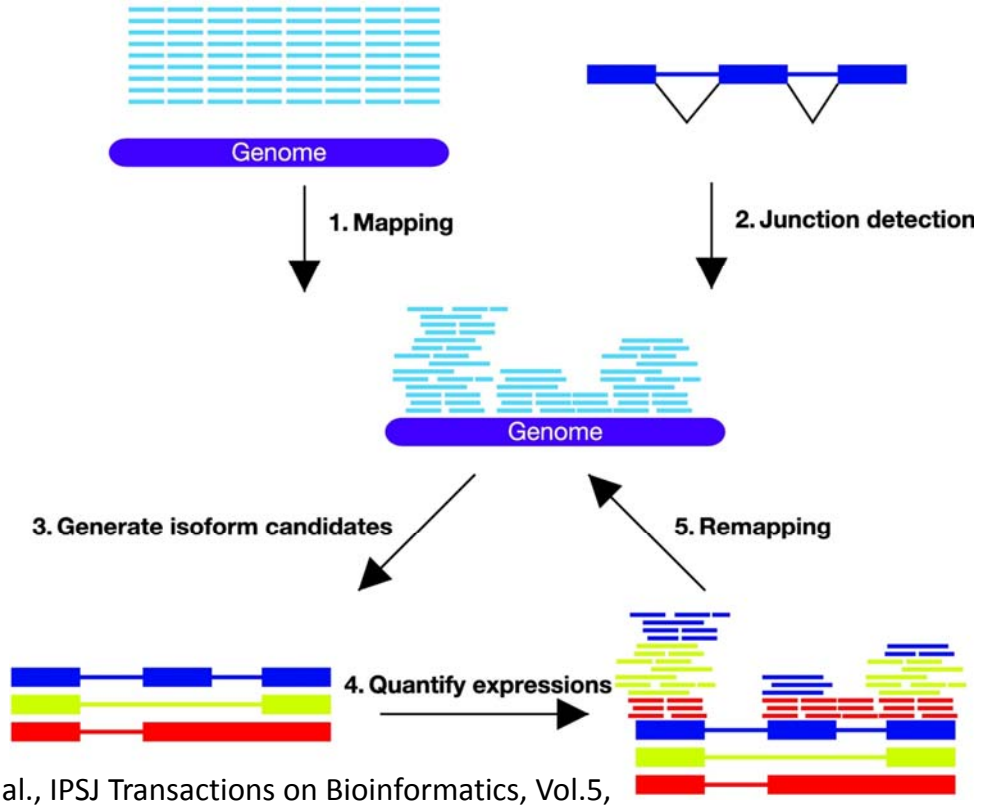
現状のRNA-Seq実験データ（1サンプルあたり）

50Mリード×100塩基×3組織×4時点（マウス個体）

マイクロアレイ　3組織×4時点×3回

（公的データベースの分も含めると**総計約300TB**）

RNAシーケンシングによる遺伝子発現量の計測



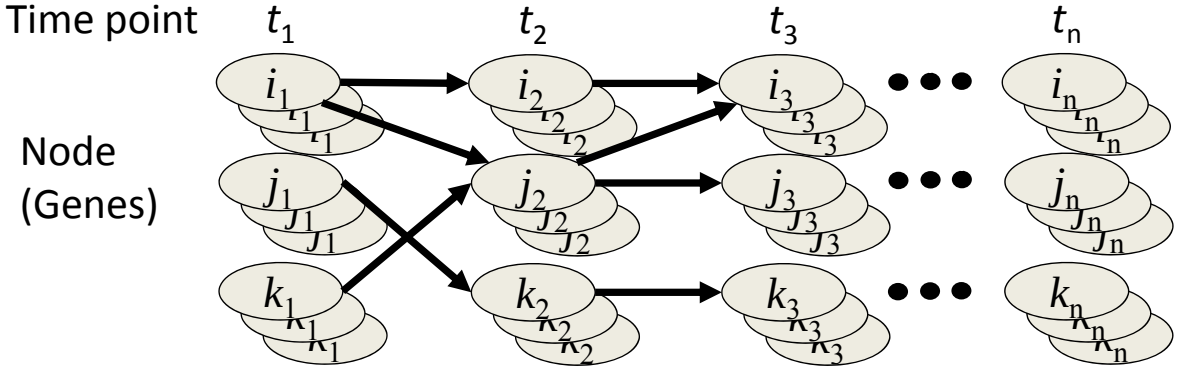
T. Ohno, et al., IPSJ Transactions on Bioinformatics, Vol.5, pp.27-33, 2012（2013年度　情報処理学会論文賞）

発現プロファイル比較とネットワーク解析

- fold-change（サンプルとコントロールの発現量比）による遺伝子抽出
  - 多数（数100～約10,000）の遺伝子が得られる
  - 抽出された遺伝子間の関連や、どれが重要かの判定が困難
- 遺伝子ネットワークを用いた解析
  - 直接的な制御関係のみを辺で結ぶ
  - ネットワークのハブ（多数の遺伝子と辺で結ばれている因子）は他に与える影響が大きい

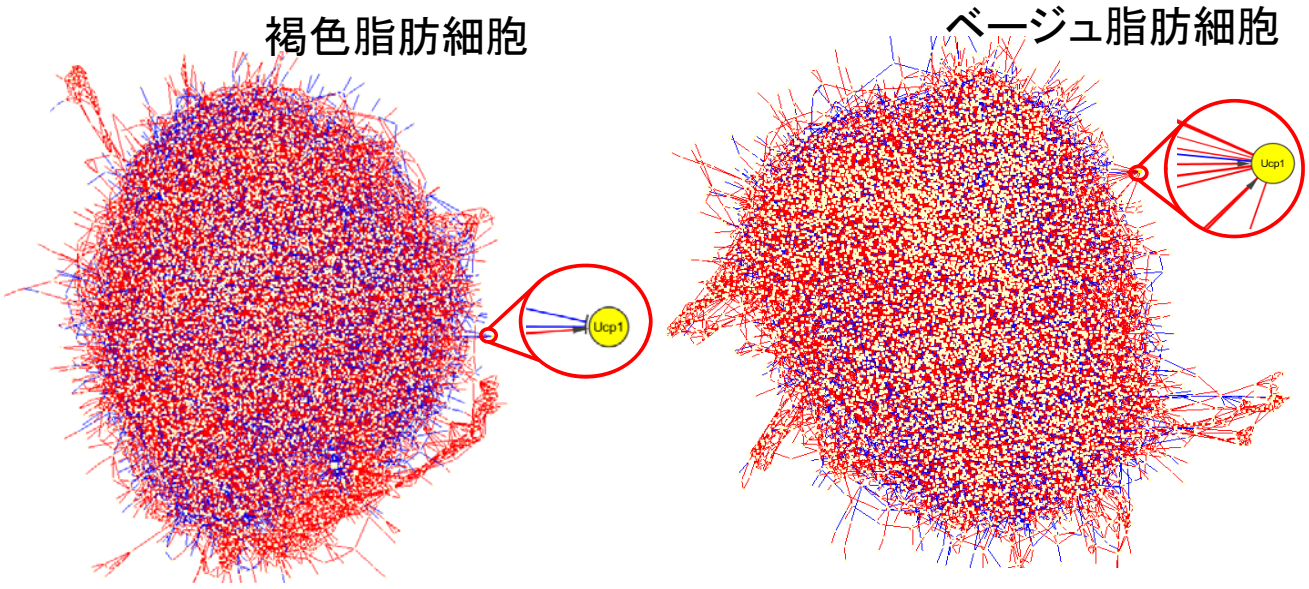
ダイナミックベイジアンネットワーク

- ベイジアンネットワークを時間的な依存関係を取り扱うように拡張
  - 連続した時点間で確率的な依存関係が存在すると仮定
- 各時点ごとに複数回の計測を行い、ブートストラップ・サンプリングでの推定結果を合成することで依存関係の信頼性（ブートストラップ確率）を計測
- 東大 宮野研で開発されたSiGN-BN（開発者 玉田嘉紀）をベース

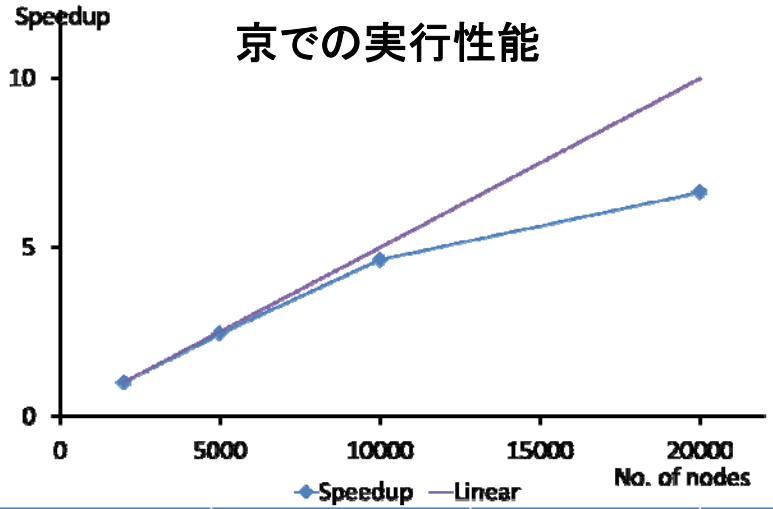


大規模生体分子ネットワークの構築

寒冷刺激で発現が誘導された約1万個の遺伝子についてネットワークを構築（赤：活性化制御辺、青：抑制制御辺）



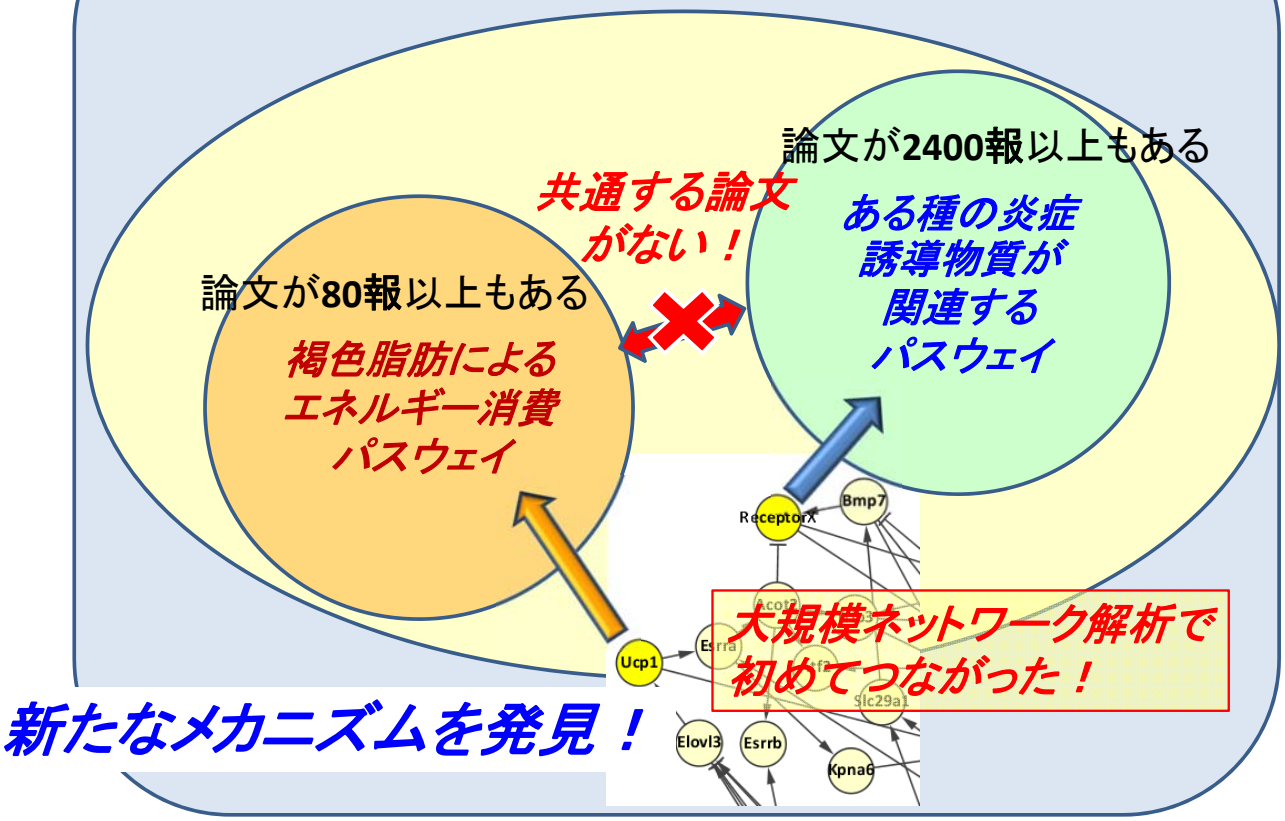
- ベージュ脂肪細胞の方が、UCP1の周りに活性化の制御辺が多く集まっており、寒冷刺激でのUCP1の発現誘導の機構を示唆している



ノード数（コア数）	実行時間（秒）	速度向上	並列化効率
2,000 (16,000)	49,659	1.0（基準）	1.0（基準）
5,000 (40,000)	20,311	2.4	0.978
10,000 (80,000)	10,715	4.6	0.927
20,000 (160,000)	7,493	6.6	0.663

種々の刺激に対する多種類の組織での細胞内の状態変化を、生体分子ネットワークにより解析するのは、「京」を使って初めてできる研究と言える

「京」によるゲノムビッグデータ解析



まとめ

- ゲノムビッグデータのデータ量はシーケンサーの劇的な性能向上による急激に増大
  - 個別化医療の実現に向けての期待が高い
- スパコン「京」と大規模ストレージがあって始めてできる規模のゲノムビッグデータ解析を実現
  - 計算量　約20万コア時間（約1万遺伝子）
  - ストレージ　約300TB
- ビッグデータ解析により、従来は関連がないと思われていた別々の生体現象がつながった（例：炎症反応と熱産生）