HPCI システム利用研究課題 利用報告書
HPCI User Report

課題番号 Project Number	hp140230	
課題名	大規模生命データ解析	
Project Name	Large-scale life data analysis	
課題代表者 Project Representative	氏名	宮野 悟
	Name	Satoru Miyano
	所属機関	東京大学
	Affiliation	The University of Tokyo
	所属機関の国名	日本
	Country	Japan
キーワード [5-10 語程度]	がん、薬剤感受性、バイオマーカー、メタボリズム、関連刺激暴露、 ベージュ脂肪細胞、メタゲノム、システム生物学	
Keywords	cancer, drug-sensitivity, biomarker, metabolism, cold-shock exposure, beige adipocyte, metagenome, systems biology	
利用ソフトウェア Software	SiGN, BENIGN, Ghost-MP, Genomon, EEM	
利用枠 Project Category	HPCI 戦略プログラム（分野 1） HPCI Strategic Program (Field 1)	
実施期間 Periods of Use	2014/4/1 ~ 2015/3/31	

利用計算資源情報 Resource Information

機関名 Institutions	資源名 Computer Resources	単位 Units	割当資源量（通期） Allocated Resources	実績資源量（通期） Used Resources
理研 RIKEN AICS	京コンピュータ K computer	ノード時間 node-hours	8,940,957	8,966,119

課題番号：hp140230

大規模生命データ解析

宮野　悟　東京大学

1. 研究の背景と目的

本課題は、がん、肥満に関わる脂肪細胞、そして細菌叢という３つの方向に研究を展開しています。がんでは、最大で数万のがん検体・細胞株データをゲノムから遺伝子ネットワークまで「京」で解析し、世界最大規模でがんのシステム異常を網羅的に解析し、薬剤耐性などを獲得するがんの多様性を解明します。そして、がんの個別化医療/創薬基盤の構築を目指します。寒冷刺激で肥満白色脂肪細胞が 100 倍の熱産生能をもつアンチメタボ細胞へと機能変化することが知られています。しかし、しないものもあり、そのメカニズムは不明でした。そこで、培養細胞に加え、マウスを使い、全遺伝子からなるネットワークを時系列で解析し、脂肪細胞の熱産生・抑制の新メカニズムを in vivo レベルで解明します。この研究は、刺激トリガーで細胞に「新機能を作る」・「ブロックする」研究へ波及効果が期待されます。メタゲノムデータ解析を「京」でホモロジーの意味で世界最深度で探索できるソフトウェア GHOST-MP を開発しました。腸内細菌叢を GHOST－MP で解析し、腸内細菌叢と生体との関係を免疫学的に解明する手がかりを得ることを目標としています。これにより、メタゲノムの理解による治療・病気の予防法の創出につながると期待しています。

First, cancer. Heterogeneity of cancer makes our understanding of cancer very difficult. We need large-scale data analysis. We need to analyze the genomes and gene networks of several tens of thousands of cancer samples/cell lines. For this analysis, we developed “SiGN-Series” for gene networks, Genomon-Series” for sequence data analysis and various bioinformatics tools on K with which we can perform the world’s largest data analysis. We expected such large-scale comprehensive analysis of cancer systems disorders unravel cancer heterogeneity, drug/sensitivity, etc. Second, obesity. It is know that upon cold stimulus, normal adipocyte transforms to beige adipocyte that has adaptive thermogenesis by dissipating energy stored in adipocyte. But its molecular mechanism was unknown. For this understanding, we developed “BENIGN” that allows us to understand changing mechanisms of cells requires analyzing gene networks of ALL genes including microRNAs. As we will report later, large-scale gene network analysis, including microRNAs reveals a details of mechanisms of transformation and we validated by collaboration with wet laboratory. Third, metagenome. We live with not only bacteria but also bacteria infected by viruses that are very much related to our health and diseases. We developed Software Infrastructure for NGS Data Analysis and “GHOST-MP” realizing metagenome analysis of microbiomes by world’s deepest homology search. GHOST-MP accelerated metagenome analysis drastically.

2. 結果要旨

1. 世界最大規模の網羅的がんの薬剤感受性・耐性遺伝子ネットワーク解析を行い、薬剤耐性・感受性を予測する方法を構築し、薬剤に関して、がんの 個別化医療及び創薬基盤のひとつを構築[1,2]。６００以上のサンプルと約１００の化合物に対する遺伝子発現データを解析。Garnett MJ et al. Nature 2012[3] の精度を超えた。ゲノムーシークエンス解析については、New England J Medicine, Nature Genetics などに１０近い成果をだしているが、これは、「京」の運用と I/O（バグも含む）では世界の競争に勝てないため、東大ヒトゲノム解析センターのスパコンに強烈な負荷をかけながら実施せざるをえなかった。残念なことだと考えている。

2. 肥満にとっては悪玉ともいわれる白色脂肪細胞組織が寒冷刺激によりベージュ（褐）色に変化する。このベージュ細胞組織は骨格筋の１００倍の熱産生能力をもつアンチメタボ細胞で、その全遺伝子ネットワークを「京」で解析した結果、熱産生の鍵遺伝子を惹起する新たなメカニズムが明らかになった。培養細胞では寒冷刺激でベージュ化するが、個体内では寒冷刺激をうけてもベージュ化しない細胞がある。マイクロ RNA を含む全遺伝子ネットワークを「京」で解析した結果、ベージュ化を抑制する機構が示唆された。

3. メタゲノム解析パイプラインの改良で超ディープ・データ解析では、従来法(BLASTX)の約 100 倍の高速化を実現し、さらに「京」で超並列化・高速計算する GHOST-MP を開発（遠いホモロジーを探索するツール）した[4]。腸内細菌叢のメタゲノム解析へ適用し、門レベルでは明らかな傾向は見られなかったが、属レベルでは大腸菌が IgA+群において有意に増加していた。ある種の大腸菌(*)は コレラトキシンと構造・機能的に類似し、CTB-IgA と交差反応する Heat-labile enterotoxin (LT) を産生するため、 LT 遺伝子配列を持つ大腸菌を検索し、IgA+群の中においてこの同定に成功した。コレラのワクチンはまだ確立されたものがない状況である。 また IgA 抗体は持続期間が短い。LT 産生をする病原性の大腸菌を常在化させる菌群がいると考えられ、今後これを同定することで、新たなワクチン開発につながる可能性がある（東大医科研・植松教授と共同研究を実施）。

1. One achievement is “High Precision Prediction of Anti-Cancer Drug Sensitivity/Resistance and Sensitivity-Specific Biomarkers by Large-Scale Gene Network Analysis” [1,2]. First by using gene network software SiGN-L1 and its robust version with the data “Sanger Genomics of Drug Sensitivity in Cancer”, we built a basis gene network data for understanding “how cancer acquires drug resistance”. Based on this method, the world-largest scale gene network analysis was done again for 600 cancer cell lines and 101 drugs on K computer. Then our prediction method based on this gene network data analysis exceeded the so-far best method by “Garnett MJ, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature. 2012 Mar 28; 483 (7391):570-575” [3].

2. Beige adipocyte is transformed from white (normal) adipocyte upon cold stimulus. This process is called browning. The role of beige adipocyte is adaptive thermogenesis by dissipating energy stored in white adipocyte. Thus beige adipocyte is thought to be anti-metabolic cell against obesity.

Matsuda-Kawada collaboration studied a mechanism for transforming to beige adipocyte by large-scale biomolecular network analysis including microRNAs. We found a novel mechanism to control the transformation.

3. Diarrhea by cholera or pathogenic *E. coli* is a major cause of babies and children's death in developing countries. As a fact, there exist people who have secretory IgA for cholera toxin B subunit (CTB), but have no history of cholera infection. Our metagenome analysis using GHOST-MP [4] in collaboration with Prof. Satoshi Uematsu (Institute of Medical Science, The University of Tokyo) is being unravelling “WHY?”

3. 計算モデル

ベイジアンネットワークモデル BENIGN
ベイジアンネットワークモデルによる遺伝子推定法 SiNG-BN
L1 正則化法に基づく遺伝子ネットワーク推定法 SiGN-L1
RNA シークエンス解析パイプライン Genomon-fusion
エクソーム解析パイプライン Genomon-exome
大規模な遺伝子ネットワーク解析のための方法で、遺伝子セット情報に基づいて mRNA 発現データ中で共発現している遺伝子群、発現モジュールを抽出する EEM 法（Extraction of Expression Module）
次世代シークエンサーデータ相同性解析法 GHOST-MP

4. 並列計算の方法と効果（性能）

大規模遺伝子ネットワーク推定ソフトウェア SiGN の開発：最適な数十遺伝子規模のネットワークからゲノムワイドネットワークまで、動的・静的・パーソナルネットワークの推定を可能にしている。

1. SiGN-BN：ベイジアンネットワークと非線形回帰を組み合わせたネットワーク推定法で、高速ブートストラップ法では 196,608 コア並列、並列化効率：0.72 (196,608 vs 98,304 コア時)を達成している。
2. SiGN-L1 及びその改良方式：L1 正則化法を用いて構造法定式モデルを推定する方法で、98,304 コア並列、実効効率 1 コアで 11.98%を達成している。これの方式をアウトライヤーにたいして頑健にするあらたな方式を開発した。パフォーマンスについては同様の状態を維持している。
3. GHOST-MP：各種データ解析パイプラインの性能評価については、現在の「京」の運用上の問題（staging プログラムのバグ、未だに利用できないプログラミング言語など「京」側の理由による）では多大な労力を必要とするため「効果（性能）」は研究の障害となり評価はしていない。

5. 研究成果

1. 世界最大規模の網羅的がんの薬剤感受性・耐性遺伝子ネットワーク解析を行い、薬剤耐性・感受性を予測する方法を構築し、薬剤に関して、がんの 個別化医療及び創薬基盤のひとつを構築。その他の成果については論文投稿中のため記載していない。
2. 肥満にとっては悪玉ともいわれる白色脂肪細胞組織が寒冷刺激によりベージュ（褐）色に変化する。このベージュ細胞組織は骨格筋の１００倍の熱産生能力をもつアンチメタボ細胞で、その全遺伝子ネットワークを「京」で解析した結果、熱産生の鍵遺伝子を惹起する新たなメカニズムが明らかになった。
3. 従来法(BLASTX)の約 100 倍の高速化を実現し、さらに「京」で超並列化・高速計算する GHOST-MP を開発（遠いホモロジーを探索するツール）し、腸内細菌叢のメタゲノム解析への適用をしている[4]。ノックアウトマウス腸内細菌叢のメタゲノム解析が進行中（植松教授と連携）。

6. まとめと今後の課題

「京」のバグ対応及び RIST の不適切な対応がなければ研究はスムーズに完了する。これまで、「京」の C コンパイラーのバグ隠しがあり（バグの入ったバージョンを消去している）、これにより我々のグループは膨大な量の再計算をよぎなくされた。その後、「京」での計算結果が正しいかどうかを常に、東大ヒトゲノム解析センターShirokane 及び東工大・TSUBAME により、全部もしくは部分的に計算結果の整合性を確認しなければ、学術的な成果として発表できないところが大問題であり、これまで運用懇談会で文章で申しいれることで対応してきた。他の課題において「京」で計算した結果が正しいかどうかの保障は全くないという感想をもっている。これを払しょくしていただきたい。

参考文献

[1] Park H, Niida A, Miyano S, Imoto S. Sparse overlapping group lasso for integrative multi-omics analysis. J Comput Biol. 2015 Feb;22(2):73-84.

[2] Park H, Shimamura T, Miyano S, Imoto S. Robust prediction of anti-cancer drug sensitivity and sensitivity-specific biomarker. PLoS One. 2014 Oct 17;9(10):e108990.

[3] Garnett MJ, Edelman EJ, Heidorn SJ, Greenman CD, Dastur A, Lau KW, Greninger P, Thompson IR, Luo X, Soares J, Liu Q, Iorio F, Surdez D, Chen L, Milano RJ, Bignell GR, Tam AT, Davies H, Stevenson JA, Barthorpe S, Lutz SR, Kogera F, Lawrence K, McLaren-Douglas A, Mitropoulos X, Mironenko T, Thi H, Richardson L, Zhou W, Jewitt F, Zhang T, O'Brien P, Boisvert JL, Price S, Hur W, Yang W, Deng X, Butler A, Choi HG, Chang JW, Baselga J, Stamenkovic I, Engelman JA, Sharma SV, Delattre O, Saez-Rodriguez J, Gray NS, Settleman J, Futreal PA, Haber DA, Stratton MR, Ramaswamy S, McDermott U, Benes CH. Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature. 2012 Mar 28;483(7391):570-575

[4] Suzuki S, Kakuta M, Ishida T, Akiyama Y. Faster sequence homology searches by clustering subsequences. Bioinformatics. 2015 Apr 15;31(8):1183-1190.