大規模データ解析と人工知能技術による
がんの起源と多様性の解明

文部科学省「富岳」成果創出加速プログラム

Newsletter

No.6

Unravelling origin of cancer and diversity by large-scale data analysis and artificial intelligence technology

2023年2月

「富岳」成果創出加速プログラム

大規模データ解析と人工知能技術によるがんの起源と多様性の解明

「富岳」における大規模シークエンスデータ解析

課題参加者　東京医科歯科大学M&Dデータ科学センター 助教　伊東　聰

全ゲノム解析は研究のみならず、臨床においても普及が進んでいます。次世代シーケンサの性能向上の恩恵もあり、近年の全ゲノム解析では大容量サンプルを大量に解析するようになりました。1サンプルのデータ量が増えることは高精度かつ高感度の変異検出を可能にします。

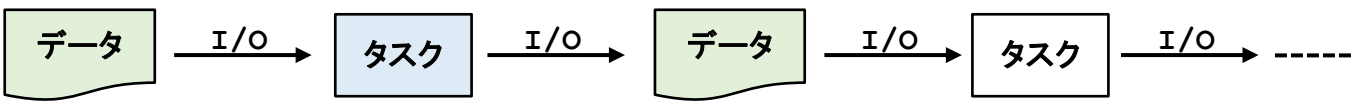
必然的に、要求される計算機能力も飛躍的に増大しており、スーパーコンピュータやクラウドサービスは全ゲノム解析を行うにあたって必須のインフラとなっています。しかし、近年はそれでも計算資源が不足する傾向にあり、富岳を始めとする超大規模スーパーコンピュータの活用を期待する声が高まっています。

ここで問題となるのが、全ゲノム解析ソフトウェアと既存のスーパーコンピュータ用ソフトウェアの特徴の違いです。スーパーコンピュータ上で運用されているソフトウェアの多くは物理や工学で用いられるシミュレーション・ソフトウェアです。その特徴をあげると、

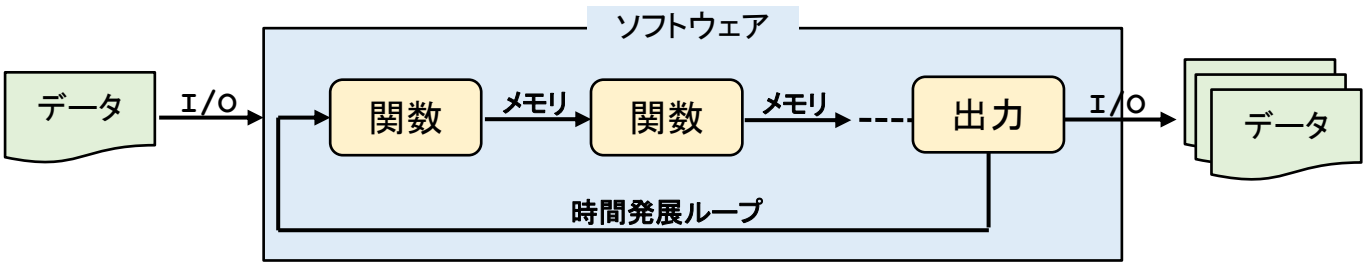
- 並列化された単一のソフトウェア
- 計算量に対して少ないI/O量
- ロードバランスをとることが出来る

などがあるでしょう。一方、全ゲノム解析ソフトウェアは、パイプライン方式（独立したソフトウェアの組み合わせ）を採用しているものが多く、計算量に対するI/O負荷が非常に大きくなります。また、並列計算時にロードバランスが著しく悪化する傾向が顕著なため、単に超大規模スーパーコンピュータを利用するだけでは、期待するほどの性能を出すことは出来ません。

我々は現在。全ゲノム解析パイプライン「Genomon」を富岳に移植、運用しています。1年に3000検体ほどの解析を実施しておりますが、安定運用に至るまでに様々な問題に直面しました。本稿では、特にI/Oとストレージシステムに関連する問題に焦点を絞り、解説をしたいと思います。



（a）　パイプライン型ソフトウェアのワークフロー



（b）　シミュレーション・ソフトウェアのワークフロー

図1：　ワークフローの異なるソフトウェア

1. 多数検体解析の必要性

Genomonを含む全ゲノム解析ソフトウェアのほとんどはパイプライン形式のソフトウェア（図1）です。具体的には、タスクと呼ばれる独立した処理（ソフトウェア）の実行を繰り返して最終的な結果を得ます。通常、各タスクは一つ前のタスクの結果ファイルを入力ファイルとしています。そのため、前のタスクが終わるまで、自身の処理を開始することは出来ません（依存性）。また、並列化が可能で多数のプロセスによる高速化が見込める処理もあれば、逐次的に計算しなければならない処理もあります。図2左に富岳上で全ゲノム解析1サンプルを実行した際の実行状況を示します。この例では、約400プロセス（100ノード）を使用して1サンプルの解析を行っています。全解析に4時間程度の時間が掛かっていることが分かります。ここで重要なことは、使用した400プロセスが4時間の間、ずっと処理をしているわけではないことです。この図では、プロセスが何らかの処理を担当している部分だけが色付きで表示されています。例えば、アラインメント処理（図中のmap_dna_sequence）ではほぼすべてのプロセスが稼働しています。一方、それ以外の部分では、多くのプロセスが計算処理の割り当てをされておらず（遊休状態）、待機状態にあることが確認できます。この計算負荷の不均一性（ロードインバランス）はGenomon特有の問題ではなく、広く全ゲノム解析を含むシーケンスデータ解析で一般的にみられる問題です。この問題を解決しなければ、大きな計算機を利用しても大部分の資源を無駄に消費することになります。

その解決策が多数検体の同時解析です。遊休状態が発生する原因は、ロードインバランスとタスク間の依存性が原因です。しかし、この依存性は対象のサンプルに関係するタスクにのみ影響しており、他のサンプルの解析には影響がありません。そのため、複数のサンプルを同時に解析すれば、依存性のないタスクの処理が悠久状態のプロセスに割り当て可能になるのです。図2右に10サンプルを同時に解析した際の実行状況を示します。図から明らかなように、遊休資源が無く、効率的に計算資源が利用されていることが確認できます。

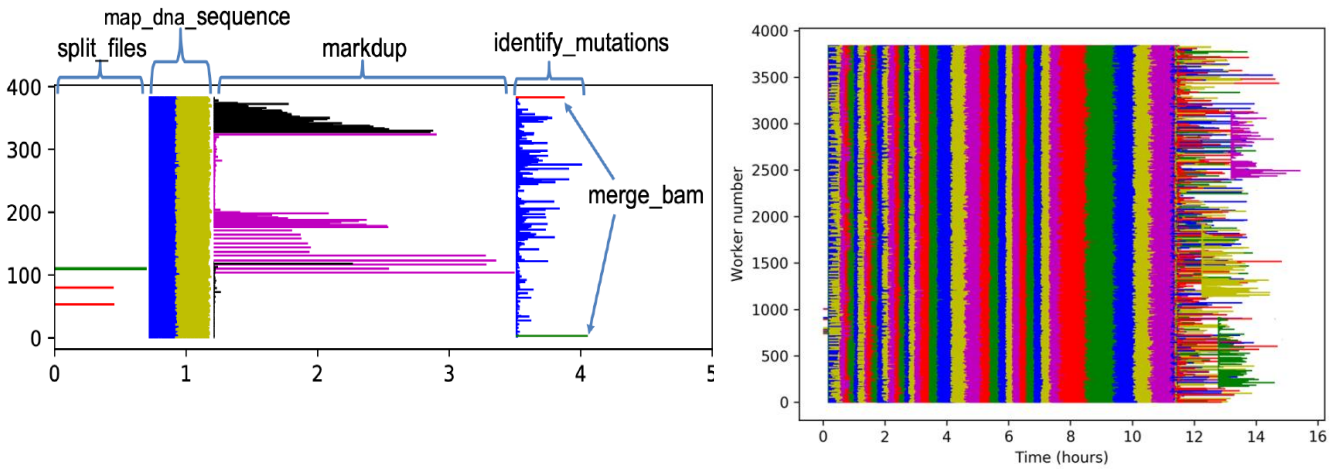


図2: 富岳上での全ゲノム解析実行の様子（左：1サンプル、右：10サンプル）
縦軸はプロセス番号、横軸は経過時間を示している。タスクを実行中のプロセスのみ、該当時間部分にカラーバーが表示されている。カラーの違いは実行中タスクの違いを示している。

2. ファイルシステムへの過負荷と対策

多数検体の同時解析は計算資源の有効活用の面では非常に有効ですが、一方で別の問題を生じます。それはファイルシステムへの負荷増大です。

初めに述べたように、スーパーコンピュータ上で運用されている多くのソフトウェアは、パイプライン形式のソフトに比べてI/O負荷がとても小さいのです。定常解析では入出力は1回ずつしか行いません。時間発展問題の場合はさらに小さく、最初の入力以外は数千～数万回の反復に1回の出力のみです。これに対し、全ゲノム解析ではタスク間で毎回入出力を行います。また、データ量も前者に比べてかなり大きく（合計数百GB～1TB）なります。このため、パイプライン形式を採用する全ゲノム解析ソフトのI/O負荷は高くなる傾向にあります。さらに、Genomonでは高速化を目的として可能な限りの並列化を実施しています。タスク間での並列数が異なる箇所ではファイルの統合・再分割などの処理を加えており、I/O負荷がさらに高くなっています。このような背景に加えて、多数検体の同時解析を実施する必要から、ファイルシステムへの負荷が非常に高くなった結果、解析時間の異常な増大という問題が発生しました。

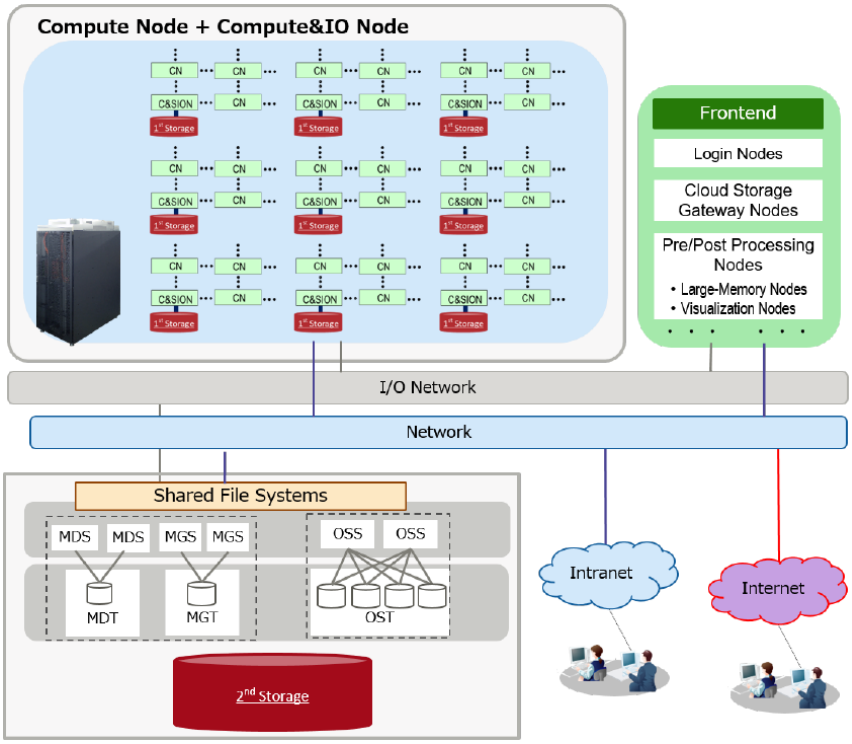


図3: 富岳の階層化ストレージ（宇野篤也, JLUG 2022より引用）

富岳のストレージは2階層構成になっている。第1階層がLLIOと呼ばれ、計算ノード（富岳を構成する最小計算機単位）に接続され高速にアクセス可能である。第2階層がいわゆるホーム領域を構成する大規模ストレージであり、ネットワークを介してアクセス可能である。第2階層は合計150PBもの大容量であるが、第1階層は1ノードあたり90GB弱しかない。

容量に制限があるが高速に読み書きできる第1階層（LLIO）と大容量であるがI/O速度に限界のある第2階層の二つの異なるファイルシステムをいかに効果的に使うかが富岳を効率的に利用するための一つのポイントである。

富岳のファイルシステムは2階層で構成されています。第1階層であるLLIOですべてのI/Oを行うことが理想的です。しかし、容量が非常に大きいこと、また、ファイルの統合・再分割等の処理の必要性から、制限の多いLLIOのみではI/O処理を完結することが出来ないため、Genomonの実行時には第2階層も利用しています。

第2階層のストレージは分散ファイルシステムの一つであるFujitsu Exabyte File System（FEFS）によって構築されています。FEFSは他の分散ファイルシステムと同様に、大量の物理ディスクを論理的に一つのストレージとしてユーザに利用できる環境を提供してくれます。複数のディスクを利用することで高速I/Oを可能にしてくれる半面、ボトルネックも存在しています。それがメタデータサーバ（MDS）です。MDSは複数の物理ディスクに分散配置されたデータ（ファイル）の状況を制御・記憶しています。このMDSは2台しかなく、富岳全ユーザのジョブで共有するため、大量のI/O要求によりファイルシステム全体のレスポンスが低下し、結果としてGenomon全体の実行が遅くなってしまいました（図4上）。そこで我々は、あえて1サンプルに対する並列数を減らし、I/O負荷を下げる調整を行いました。並列数を減らすことにより、1プロセスあたりの担当データ量が増えるため必要なCPU時間が増えますが、その増加分よりI/O負荷の減少分が大きく上回ると考えたのです。

テストの結果、当初のおよそ1/4の並列数に抑えることにより、およそ2.5倍の高速化を実現でき、現在、この最適なパラメータを用いて年間3000検体程度の全ゲノム解析を実施しています。

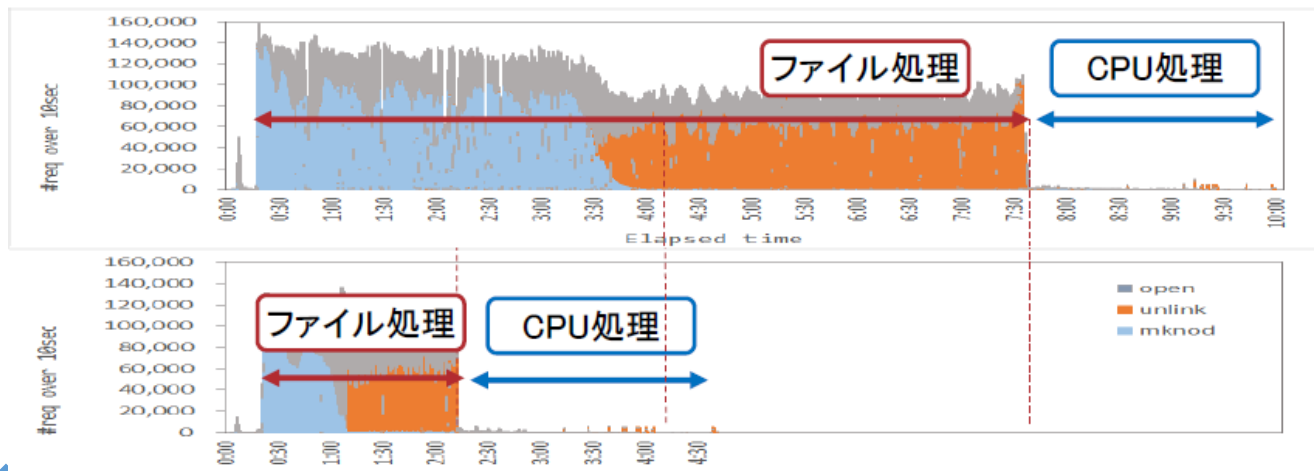


図4: Genomon実行時のI/O負荷
横軸は経過時間、縦軸がファイルオープン等のI/O処理リクエスト量。最適化前（上）ではGenomon実行初期に発生した大量のI/Oリクエスト処理に長時間を要している。最適化後はリクエストが減少し、高速化した。

編集者メモ

- 2023年2月28日：東京医科歯科大学M&Dデータ科学センターのHeewon Park教授が、ソウルの母校に移動になりました。3月からは東京医科歯科大学M&Dデータ科学センターの客員教授として継続して本プロジェクトに参加します。
- 2022年2月6-7日：The 5th R-CCS International Symposiumにおいて、宮野悟が講演及びパネルディスカッションに参加しました。https://www.r-ccs.riken.jp/R-CCS-Symposium/2023/。
- 2023年1月19日：東京医科歯科大学と東京工業大学が一つの大学に統合され、その名称が「東京科学大学(Institute of Science Tokyo)」となるそうです。東京医科歯科大学の名称の付いた最後のプロジェクトになりました。
- 2022-2023年：Heewon Park教授の研究で、がんの多様性と薬剤耐性の関係が、スパコンと独創的な数理的手法の開発で明らかになってきました。
  Park H et al. (2023) ***J Comput Biol.*** 30(2):223-243.
  Park et al. (2022) ***Int J Mol Sci***. 23(22):14398.
  Park H et al. (2022) ***BMC Bioinformatics***. 23(1):342.
  Park H et al. (2022) ***PLoS One***. 17(5): e0261630.
  Park H et al. (2022) ***J Comput Biol***. 29(3):257-275.

  また、「AI」x「富岳」でも連携機関の富士通研究所と一緒に成果をだしていますこの成果もがんの多様性に関するものです。
  Park H, Maruhashi K, et al. (2020) ***PLoS One***. 15(11):e0241508.
- 2022年12月18日：シンポジウムはNPOバイオインフォマティクス・ジャパンとの共同開催として行われました。プログラムを右につけております。100名を超える参加者がありました。
- 次回のニュースレターNo. 7が最後となりますが、課題代表者の宮野悟が、3年間の総括をする予定です。

令和4年度シンポジウム
「がんはどのようにはじまり、そしてどのように強敵にばけるのか」
日時：2022年12月18(日) 14:00-16:30
開催方法：Zoomによるオンライン開催

プログラム
14:00-14:05 開会挨拶
宮野悟（東京医科歯科大学M&Dデータ科学センター長
/NPO法人バイオインフォマティクス・ジャパン理事）
五斗進（NPO法人バイオインフォマティクス・ジャパン副理事長
/ライフサイエンス統合データベースセンター副センター長）
14:05-14:45【基調講演】がんの起源を探る
小川誠司（京都大学大学院医学研究科腫瘍生物学講座教授）
14:45-15:10【招待講演】生物の進化論で考える、がんが難治である理由（特に大腸がんに注目して）
三森功士（九州大学病院別府病院外科教授）
15:10-15:35【招待講演】「未知」を発見する人工知能
丸橋弘治（富士通(株)富士通研究所プロジェクトマネージャ）
15:35-16:00【招待講演】胃がん抗がん剤の耐性メカニズム解明を目指す遺伝子ネットワーク解析
Heewon Park（東京医科歯科大学M&Dデータ科学センター教授
/NPO法人バイオインフォマティクス・ジャパン理事）

1. 代表機関

国立大学法人東京医科歯科大学
研究課題開発責任者
M&Dﾃﾞｰﾀ科学ｾﾝﾀｰ 特任教授 宮野悟

2. 協力機関

国立大学法人京都大学
協力機関代表者
大学院医学研究科 教授 小川誠司

愛知県がんセンター
協力機関代表者
研究所ｼｽﾃﾑ解析分野 分野長 山口類

3. 連携機関

(株)富士通研究所
連携機関連絡担当者
富士通株式会社 フェロー 岡本青史

国立大学法人東京大学
連携機関連絡担当者
医科学研究所 准教授 片山琴絵

Karolinska Institutet
連携機関連絡担当者
Department of Medicine/Center for Hematology and Regenerative Medicine
Visiting Professor 小川誠司

4. 事業参加者

事業参加者（代表機関）
東京医科歯科大学M&Dﾃﾞｰﾀ科学ｾﾝﾀｰ
統合解析分野 特任教授 宮野悟
准教授 長谷川嵩矩
助教 伊東聰
特任助教 角田将典
特任助教 小川弥穂
特任研究員 田中洋子
AI技術開発分野 教授 Heewon Park

事業協力者（協力機関）
京都大学大学院医学研究科
腫瘍生物学講座 教授 小川誠司
特定教授 南谷泰仁
助教 越智陽太郎
他11名

愛知県がんセンター研究所
ｼｽﾃﾑ解析分野 分野長 山口類

連携参加者（連携機関）
富士通株式会社 フェロー 岡本青史
富士通株式会社 富士通研究所
プロジェクトマネージャ 丸橋弘治
東京大学医科学研究所
ﾋﾄｹﾞﾉﾑ解析ｾﾝﾀｰ 准教授 片山琴絵
先端医療研究ｾﾝﾀｰ 教授 南谷泰仁
（2023年2月28日時点）

スーパーコンピュータ「富岳」成果創出加速プログラムについて（理化学研究所計算科学研究センターより抜粋）
https://www.r-ccs.riken.jp/jp/fugaku/promoting-researches

スーパーコンピュータ「富岳」成果創出加速プログラムは、「富岳」を用いた成果を早期に創出することを目的として文部科学省が設置しました。①人類の普遍的課題への挑戦と未来開拓、②国民の生命・財産を守る取組の強化、③産業競争力の強化、④研究基盤の4領域について課題の選定が行われ、19課題が選定されています。選定された課題は、「富岳」の計算資源を優先的に無償で使用し、速やかな成果創出を目指します（2020～2022年度）。

課題名： 大規模データ解析と人工知能技術によるがんの起源と多様性の解明
課題番号：JPMXP1020200102

課題ID
2022年度：hp220163
2021年度：hp210167
2020年度：hp200138

Information

TOKYO MEDICAL AND DENTAL UNIVERSITY
M&D Data Science Center

文部科学省「富岳」成果創出加速プログラム
課題名：大規模データ解析と人工知能技術によるがんの起源と多様性の解明
ニュースレター No. 6
発行日★2023年2月28日
課題代表者★宮野 悟
- 東京医科歯科大学M&Dデータ科学センター　統合解析分野
- 郵便物宛先：　〒113-8510　東京都文京区湯島1-5-45
- 宅配便宛先：　〒101-0062　東京都千代田区神田駿河台2-3-10
  駿河台キャンパス22号館5階
- TEL: 03-5280-8620 FAX: 03-5280-8632
- E-mail: mdsc.dsc@tmd.ac.jp
- HP: https://odcla.mddsc.jp