様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ６ 月 ２６ 日現在

機関番号：１２６０１
研究種目：挑戦的研究（萌芽）
研究期間：2020～2022
課題番号：２０Ｋ２１８２６

研究課題名（和文）不揮発メモリーを用いたゲノム情報学の新展開

研究課題名（英文）Genome informatics using non-volatile memory

研究代表者
笠原　雅弘（Kasahara, Masahiro）

東京大学・大学院新領域創成科学研究科・准教授

研究者番号：６０３７６６０５

交付決定額（研究期間全体）：（直接経費）　　4,800,000円

研究成果の概要（和文）：ゲノム解析の様々な場面おいて利用される配列類似性検索において、公共配列データベース上に存在する配列など検索対象データベースのサイズが非常に大きくメモリー上に検索を高速化するための索引データ構造全体を格納することができない場合には、広く一般に用いられている類似性検索ソフトウェアより高速な検索を行うためにＳＳＤなどの大容量記憶媒体を用いた索引データ構造が利用できることを見いだした。マッチ配列長さが短い場合には確率的データ構造が有効であることを見いだした。

研究成果の学術的意義や社会的意義
ＤＮＡ配列シークエンサーの発展に伴ってＤＮＡ配列解析に必要な試薬代や機器代は安くなっていくことが予想されているが、解析に必要な計算機資源を手に入れるコストはＤＮＡ配列シークエンサー関連のコスト低下ほど低下していない。
本研究により生み出された手法により、大きなデータベースに対する配列検索が従来考えられていたよりも遙かに安価なハードウェアで高速に実行できるようになった。これにより、将来的に必要となっていたゲノム解析に対する計算機資源量を減らし節約をすることができる。

研究成果の概要（英文）：In sequence similarity search, which is used in various genome analysis, when the size of the target database, such as all sequences on public genome sequence databases, is so huge that it is impossible to store the entire index data structure for speeding up the search in memory, we found that an index data structure using large-capacity storage media such as SSDs can be used for faster searching than commonly used traditional similarity search programs. We also found that probabilistic data structures are effective when the match sequence length is short.

研究分野：ゲノム情報科学

キーワード：ゲノム　不揮発メモリー　ＳＳＤ

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

ゲノム情報学の分野では、ゲノム配列を解読するゲノムアセンブリ処理や大容量データベースへの相同性解析、メタゲノム解析など、様々なゲノム解析のためにテラバイト単位の超大容量メモリーを搭載した高価なサーバーを用いてきた。例えば、国立遺伝学研究所がゲノム解析に供していたスーパーコンピューターでは１２テラバイト（１台）・３テラバイト（１０台）のメモリーを搭載した共有メモリー計算サーバーを用意していた。これらの大容量メモリー搭載サーバーは非常に高価ではあるが、数千万円以上の価格を持つ DNA シークエンサーを何台も購入し、毎年数億円の試薬代を使用する大規模ゲノム解析拠点で用いる解析インフラとしては相対的に十分に安価であり、大容量メモリーを搭載したコンピューターの使用はゲノム情報解析分野内では十分に許容されてきた。

しかし、２０１６年に本体価格１０００米ドル程度、試薬代も１ランあたり数百ドル程度となるポータブル DNA シークエンサー（Oxford Nanopore Technologies 社 MinION）が発売され、大きな予算を持たない小さい研究機関が行う小型ゲノム解析拠点の需要が突然高まった。安価な DNA シークエンサーはフィールドにおける研究や、発展途上国における感染症診断・対策への利用、宇宙空間における DNA シークエンシング、教育現場での応用など、従来のコスト感覚ではなし得なかった幅広い応用先を生み出した。とはいえ、当時の現状では大容量メモリーを必要とする解析が数多くあり、計算機側のダウンサイジングを行う需要が今後高まっていくのではないかということが予測された。

また、おおむね同時期に、高価な DRAM に対して比較的安価な不揮発性メモリが発表されていた。Intel と Micron Technology が共同開発した不揮発性メモリで、3D XPoint と呼ばれのちに Intel から Optane というブランド名で発売された不揮発性の相変化メモリは、当時の Micron の見積もりでは、このメモリはダイナミックランダムアクセスメモリ（DRAM）より圧倒的に安い価格で、フラッシュメモリより圧倒的に短いレイテンシを実現していると喧伝されていた。実際に、Optane メモリーを利用した初代の SSD ドライブが約 100 KIOPS と<10μ秒のレイテンシを達成し、ランダムアクセスの低キュー深度での IOPS を大幅に向上させていた。このような新しいデバイスと、大容量メモリーを利用するゲノム解析アルゴリズムの相性を組み合わせると利点を活かすことができるのではないかと考えられた。

２．研究の目的

本研究では安価な計算機でも動作するゲノム解析アルゴリズムを開発し、安価なＤＮＡシークエンサーの登場によって爆発的に多様化しているゲノムシークエンシングの現場（発展途上国を含む病院での感染症診断、教育現場など）でのゲノム解析を可能にすることで、より多様な現場でゲノム解析を実行可能とし、ゲノム解析計算のコストを下げることでより広い分野でのゲノム解析の未来を拓くことを目的としていた。

３．研究の方法

本研究では、２０１９年にはじめて一般にも入手できるようになった低遅延の不揮発メモリー Intel Optane（以下 Optane）を用いて安価な計算機上でのゲノム解析技術構築を目指していた。Optane は Intel 社および Micron 両社が合同で開発した次世代メモリーであり、通常の DRAM と比べて数分の１の容量単価を実現すると両社は説明していた。一般に大容量ストレージとして用いられている Solid State Drive（SSD）と比べて Optane は圧倒的にランダムアクセスが低遅延であり、アクセスの単位も小さいため、簡潔データ構造や Bloom filter などのデータ構造を用いた大容量データを扱うゲノム解析アルゴリズムと極めて相性が良く、大きな性能増加が見込めると考えられた。

研究開始当時に知られていた大容量メモリーを利用する各種のゲノム解析アルゴリズムは、メモリーの容量だけではなく（不揮発記録媒体と比べたときの）圧倒的な低遅延性も利用しているため、従来のアルゴリズムを単純に Optane 向けに移植するだけでは良い性能は期待できなかった。そこで、根本的なアルゴリズムレベルから Optane の特性を考えたアルゴリズムを新開発することで、Oxford Nanopore 社の DNA シークエンサーや試薬代と釣り合う程度の価格帯のワークステーションでいくつかのゲノム解析を可能にすることを目指していた。

最初のターゲットとして、de Bruijn グラフを用いた塩基配列検索アルゴリズムを Optane 向けに設計することを目指した。DRAM スロット搭載型の Optane メモリは DRAM より高遅延・低帯域であるため、遅延の隠蔽や情報の圧縮を行う帯域節約アルゴリズムを設計することを当初は企図していた。現在知られている簡潔データ構造を用いるアルゴリズムに工夫を加えて、遅延の隠蔽やランダムアクセス回数の低減を行う手法を設計しようとした。

ただし、本研究の申請時点では Optane の価格推移を予測することが難しかった。DRAM の価格推移トレンドをもとに、Intel 社が DRAM に対して競争力のある価格を Optane メモリに対して値付けすると仮定し、そのような仮定のもと、ゲノム解析において価格競争力のあるデータ構造を見つけることを当初の目標としていたが、Optane メモリの価格が高止まりした場合には

SSD などのより高遅延な不揮発メモリーを用いて同様の目的に資するアルゴリズム設計を目指すプランも用意して研究をスタートした。

４．研究成果

大容量の不揮発メモリー(SSD)上で大きなページ単位のビットマップベースグラフを構築し、グラフ上での探索アルゴリズムを組み合わせることで、(例えば GenBank の nt のような) 大容量の DNA 配列データベースに対するデータベースインデックスを作成し、大容量の DRAM を搭載した計算機を必要とせずにデータベースサイズが極めて大きい場合でも効率的な配列検索が多数の配列に対して並列に行える新規のインデックスデータ構造を考案した。本手法を用いた配列検索アルゴリズムでは、アクセスサイズが大きく遅延が DRAM と比べて著しく大きな SSD メモリーの特性下でも性能が低下しにくく、DRAM の消費量も従来法と比べると大きく抑えることができることが分かった。

また、ビットマップベースのグラフを用いた配列検索アルゴリズムでは、検索して見つかったターゲット配列の配列メタデータを取得する計算効率／記憶容量効率の良い手法が存在しなかったが、ハッシュ関数を用いて配列そのものからメタデータへの索引を張る手法を用いることで、この問題を軽減することができることを見いだした。また、本アルゴリズムは主にＤＮＡ配列を対象として研究を進めていたが、手法をアミノ酸配列の検索に対して拡張する理論的検討を行った。

また、現在の SSD メモリーは書き込み回数の制限がハードディスクや Optane メモリと比べて非常にきつく（小さく)、本アルゴリズムによるデータベース索引作成を SSD 上で何度も繰り返すと定格の書き込み容量を超えてしまい短期間で SSD ドライブ上のデータが破損に至る可能性が高いと考えられた。もしくは、データベースセンター向けの高耐久 SSD を用いると本手法の価格競争力は完全に消えてしまうと考えられた。このため、スーパーコンピューターを用いた分散メモリー環境下において SSD を使わずに索引作成を行う手法を考案し、データ破損を心配せずとも現実的な計算機環境においても本手法が使える可能性が高いことを見いだした。

研究申請時に利用することを想定していた不揮発メモリーの一つである Intel 社 の Optane メモリーは、当初の Intel 社の宣伝に反して極めて高値で推移し、2TB クラス以上の大容量 Optane メモリーを搭載した計算機システム全体の価格は研究期間中に一度も DRAM を用いたシステムの半額以下にはならなかった。コロナ禍や戦争にともなう半導体流通の混乱もあり、特に Optane メモリーを搭載したシステム全体の価格は同容量の DRAM を用いた製品と同等かあるいはそれ以上で高止まりしていた。同じ容量のメモリーを搭載した Optane メモリーの計算機と DRAM の計算機ではもちろん後者の方が圧倒的に高い解析パフォーマンスを発揮する。このため、DRAM を用いたシステムに対するゲノム解析アルゴリズムにおけるコストパフォーマンス上優位性はゼロのまま推移していた。

悲しいことにゲノム解析アルゴリズムに留まらない一般のビジネスにおける Optane メモリーのコスト優位性も低かったことが明らかになり、2022 年 7 月に Optane メモリーを用いた現行製品の販売終了予定が Intel 社より発表された。この発表により、現状で販売されている Optane メモリーを使ってコスト競争力のあるゲノム解析アルゴリズムを構築することはできないことが確定した。eBay などのオークションサイトに流通している中古品を揃えることで Optane メモリーを使ったシステムを価格競争力のある形で構成することも検討したが、Optane メモリーと対応システム一式、および大容量メモリーに対応した L 型番の Intel 社の CPU を全て揃えることは運がないと非常に難しく、大きなブレイクスルーに繋がる 6TB の Optane メモリーを搭載したシステム一式の構築および、そのような大容量 Optane メモリーを利用したアルゴリズムの構築およびコスト競争力の証明を行うことは断念することとなった。

ただし、Optane メモリーではなく、三次元積層型の NAND 型 SSD メモリーおよびそのコントローラー開発は本研究の期間中に（市場で）大きく進展し、長いキュー深度に対しては IOPS が非常に高く容量単価の十分に小さい NVMe SSD 製品が市場に出てきたため、ターゲット配列が極めて大きい条件下における検索アルゴリズム対象に対しては有効なアルゴリズムを構築することができた。PCI gen4 さらには今後の gen5 接続の NVMe SSD 製品が一般的になるにつれ、本研究の方向性がより適切になっていくであろうと考えられる。

最後に、本研究で開発したアルゴリズムは DRAM では索引を作ることが困難であるような極めて大容量のターゲットデータベースに対する検索に適しているため、検索アルゴリズムの応用可能性としてマーカー探索などへの適用方法を模索した。特に、公開データベースに存在しているなるべく多くの種を同時に考慮することで種を高精度で検出する種特異的マーカーの探索、および品種特異的なマーカーを探索し設計するためのアルゴリズム開発において本研究で作成した索引データ構造を応用する方法を検討した。

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計0件

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関