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　３ 年　 ５ 月 ２４ 日現在

機関番号：１２６０１
研究種目：基盤研究(B)（一般）
研究期間：2017～2020
課題番号：１７Ｈ０１６９３

研究課題名（和文）ビッグクエリー×ビッグデータ検索実現のためのＳＭＡＤ技術の新展開

研究課題名（英文）Development of SMAD for big query on big data

研究代表者
渋谷　哲朗（Shibuya, Tetsuo）

東京大学・医科学研究所・教授

研究者番号：６０３９６８９３

交付決定額（研究期間全体）：（直接経費）　　8,900,000円

研究成果の概要（和文）：爆発的に増加するビッグデータに対応した検索技術が求められている。これに対し、データベースの統計的挙動を活用するSMAD (Statistical Model-based Algorithm Design)技術が注目されている。本研究ではこのSMAD技術をビッグクエリー×ビッグデータ検索へ展開させる研究を行った。特に、個人ゲノムデータベース、タンパク質立体構造データベース、自然言語テキストデータベースなど様々なデータベースに対する大規模検索の基盤技術を開発することに成功したほか、プライバシー保護技術、PCMメモリにおけるメモリ分散技術、次世代シークエンサー解析などの技術開発にも成功した。

研究成果の学術的意義や社会的意義
近年のデータ爆発は、大規模ビッグデータに対する大規模なクエリーを著しく困難にしており、それに対する超効率な検索基盤技術の開発が求められている。本研究では、SMAD技術を核に、ゲノムデータベース、タンパク質立体構造データベース、自然言語テキストデータベースなど様々なデータベースに対する検索技術の開発に成功したほか、プライバシー保護、PCMメモリの活用、次世代シークエンサー解析など、様々なデータ解析の基盤技術の高度化にも貢献することに成功している。これらの技術によって、今後さらにビッグデータの利活用が高度化されることが期待できる。

研究成果の概要（英文）：A new technology for searching big data is desired. SMAD (Statistical Model-based Algorithm Design) is a candidate for improving these searching algorithms. In this research, we explored applications of SMAD technology to big query/big data searching problems. In particular, we developed new searching technologies for individual genome databases, protein 3-D databases, and natural language text databases. Moreover we succeeded in developing technologies for privacy preserving, memory distribution for PCM memories, and next generation sequencer read analysis.

研究分野：バイオインフォマティクス・アルゴリズム

キーワード：アルゴリズム理論

１．研究開始当初の背景

近年あらゆる分野でデータ爆発がおき、その活用のための情報科学技術の向上の必要性が叫ばれている。特に生物医学分野等では次世代シークエンサー等の計測技術の革命的な発展により、ムーアの法則をはるかに超える加速度で大量のデータが産出されている。一方で、ムーアの法則に関しても、物理限界による成長鈍化・停止の可能性が言及されるようになって久しい。これらのことから、増え続けるビッグデータのデータ検索や解析を行うのに計算機の高速化・並列化のみに頼っていては、近い将来必ず破綻することは火を見るより明らかである。さらにそれだけではなく、様々な分野の多くのデータベースにおいて、データベースだけでなく、検索クエリー側のデータも複雑化・大規模化しており、必要とされる高速・高精度な検索を実現するようなアルゴリズムの設計はますます困難になりつつある。これに対し、2009 年以降、本研究代表者は統計的モデルにもとづく新しいアルゴリズム設計パラダイムＳＭＡＤ(Statistical Model-based Algorithm Design)を提唱している。さらに、このパラダイムに基づくことで構造生物学における最重要データベースであるタンパク質立体構造データベースの基本検索の理論的なブレークスルーを達成するとともに、実用的にも精度の犠牲なく数千倍高速化することに成功し、構造生物学研究に対し大きなインパクトを与えた。これまでの情報科学におけるアルゴリズム設計の多くが最悪性能、単純な理想的なランダム性を仮定した平均性能を指標にしてきたのに対し、このＳＭＡＤは対象とするデータベースの精緻な統計学的モデルに基づいてアルゴリズムの設計・性能解析を行う。

一方で、現実の状況においては、高速に大量のデータを産出する様々な機器が次々に登場しクエリーも巨大であり、大規模なデータベースに対してその巨大クエリーの検索を行う必要がある。当然ながらそのようなビッグクエリーの検索は著しく困難あるいは不可能である。それに対する新たな技術開発が必要とされている。

２．研究の目的

これまでのＳＭＡＤの適用においてはデータベース側についてのみ何らかの統計モデルを考え、それによって高速化を図っていたが、ビッグクエリー検索においては同様の統計モデルをクエリー側についても同時に考えることができる可能性がある。そこで本研究課題では、**ビッグクエリー×ビッグデータ**の検索・解析を可能とするような新たなアルゴリズム設計パラダイムとして、データベース側、クエリー側双方の適切なモデルを活用することによって高速アルゴリズムを設計する**相乗的ＳＭＡＤ**を考え新時代の検索アルゴリズムの実現を狙うとともに、そのような新たなビッグクエリーアルゴリズムの応用の開拓も狙う。

３．研究の方法

本研究では

１）相乗的ＳＭＡＤによるビッグクエリー×ビックデータ検索技術の開発

２）ビッグクエリー×ビックデータ検索を活用したビッグデータ解析技術の研究開発

の２つの研究を段階的に行った。１）のビッグクエリー×ビックデータ検索技術については、これまでＳＭＡＤ技術の適用が成功を納めてきたたんぱく質立体構造データベースやゲノムデータベースなどに対して技術開発を行った。２）では、人工知能応用やプライバシー保護など新たな技術へ展開することを行った。

４．研究成果

(1)　大規模なビッグクエリー×ビッグデータ問題として、ポピュレーションデータをビッグクエリーとして大規模個人ゲノムデータからポピュレーション間で保存されているハプロタイプブロック抽出に関する研究を行い、アフリカからヨーロッパ、アジアへとどのように人類が広がったかについて、新たな知見（図１）を得ることにも成功した。（Onuki, Yamaguchi, Shibuya, Kanehisa, Goto, 2017）

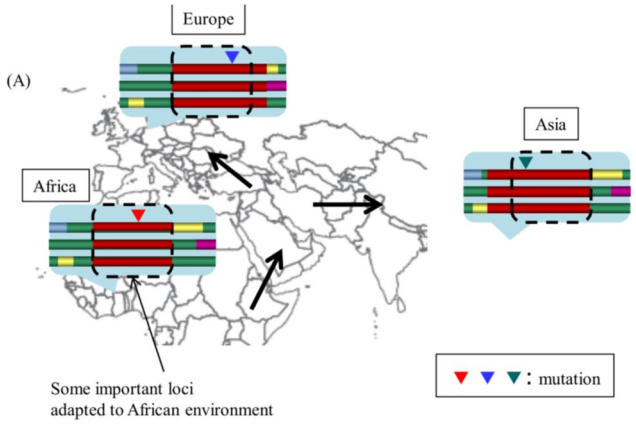

図１．民族間のハプロタイプブロック移動

(2)　多くのビッグデータが個人情報を扱っていることから、ビッグクエリー×ビッグデータ検索においてもいかにそれらを保護するかは重要である。特にビッグクエリー検索においては、データベース側を暗号化するなどだけではクエリー側のデータを保護することはできない。これに対し、本研究では、クエリー側がどこにアクセスしたかを秘匿する Oblivious RAM 技術をビッグクエリー×ビッグデータ検索においても利用可能な世界発の簡潔データ構造 Succinct Oblivious RAM（図 2）の開発に成功した。これによって従来は不可能であった規模の検索においてもアクセス秘匿を行うことが可能となった（Onodera, Shibuya, 2018）。

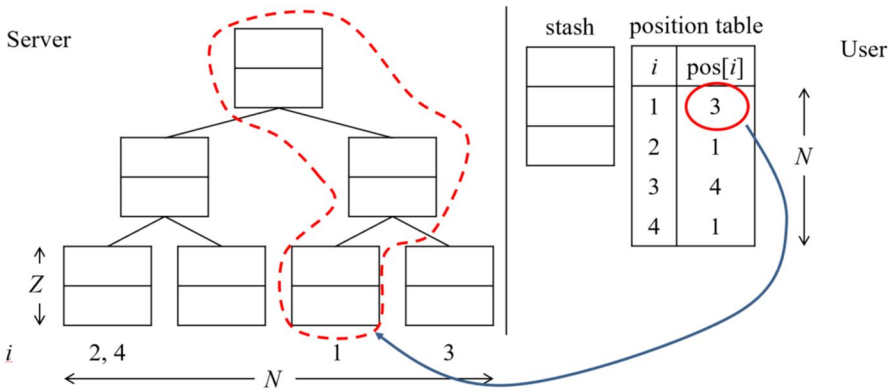

図 2. Succinct Oblivious RAM

(3) 本研究者がこれまで行ってきた SMAD 技術によるタンパク質立体構造データベースの部分構造検索について、より複雑で計算量が必要な分子の順序を考慮しない検索についても、SMAD 技術による高速化に成功した。これによって、タンパク質立体構造の中で主鎖上で連続していないような部分構造に関しても高速検索が可能となった。(Sasaki, Shibuya, Ito, Arimura, 2019)

(4) 今日の次世代シークエンサーは非常に高速に DNA を解読することができるが、一方でエラーが一定割合で含まれるためそれらをいかに高精度に補正するか、が正確なゲノムのビッグデータ解析の鍵のひとつとなっている。本研究では、ナノポアシークエンサーの出力データを深層学習を用いて高精度化するとともに、塩基修飾などさらなる付加情報の予測を行うシステムの開発に成功した。(Zhang, Akdemir, Tremmel, Imoto, Miyano, Shibuya, Yamaguchi, 2020)

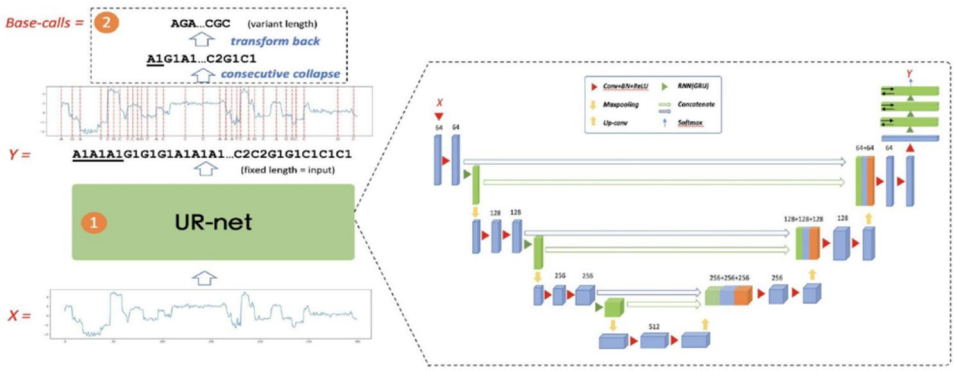

図 3. 深層学習による NGS リード解析

(5) ビッグクエリー×ビッグデータ検索を行うシステムでは、そのデータの大規模性に対処するために様々なメモリデバイスを最大限活用する必要がある。本研究では、特に phase change memory とよばれる新たなメモリーデバイスをビッグデータ解析においても最大限活用する新たなメモリ分散技術の開発に成功した。(Onodera, Shibuya 2020)

(6)ビッグクエリー×ビッグデータ検索の一つの究極の目標は多言語環境でのデータベース自然言語検索である。本研究ではそのような自然言語検索を語尾変化のきわめて激しい言語においても可能とする言語のエンベッディング技術の開発に成功した。(Arda, Shibuya, Gungor, 2021)

５．主な発表論文等

〔雑誌論文〕　計8件（うち査読付論文　8件／うち国際共著　1件／うちオープンアクセス　5件）

１．著者名 Arda Akdemir, Tetsuo Shibuya	４．巻 2696(66)
２．論文標題 Transfer Learning for Biomedical Question Answering	５．発行年 2020年
３．雑誌名 Proc. CLEF 2020, CEUR Workshop Proceedings	６．最初と最後の頁 1-15
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Taku Onodera, Tetsuo Shibuya	４．巻 181(65)
２．論文標題 Wear Leveling Revisited	５．発行年 2020年
３．雑誌名 Leibniz International Proceedings in Informatics (LIPIcs)	６．最初と最後の頁 1-17
掲載論文のDOI（デジタルオブジェクト識別子） 10.4230/LIPIcs.ISAAC.2020.65	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

１．著者名 Arda Akdemir, Tetsuo Shibuya, and Tunga Gungor	４．巻 1
２．論文標題 Subword Contextual Embeddings for Languages with Rich Morphology	５．発行年 2021年
３．雑誌名 Proc. ICMLA 2020	６．最初と最後の頁 994-1001
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ICMLA51294.2020.00161	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Yao-zhong Zhang, Arda Akdemir, Georg Tremmel, Seiya Imoto, Satoru Miyano, Tetsuo Shibuya, and Rui Yamaguchi	４．巻 21
２．論文標題 Nanopore basecalling from a perspective of instance segmentation	５．発行年 2020年
３．雑誌名 BMC Bioinformatics	６．最初と最後の頁 1-9
掲載論文のDOI（デジタルオブジェクト識別子） 10.1186/s12859-020-3459-0	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Tetsuo Shibuya	4．巻 13
2．論文標題 Application-Oriented Succinct Data Structures for Big Data, The Review of Socionetwork Strategies	5．発行年 2019年
3．雑誌名 The Review of Socionetwork Strategies	6．最初と最後の頁 227-236
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s12626-019-00045-1	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Yoichi Sasaki, Tetsuo Shibuya, Kimihito Ito, and Hiroki Arimura	4．巻 E102.A(9)
2．論文標題 Efficient Approximate 3-Dimensional Point Set Matching Using Root-Mean-Square Deviation Score	5．発行年 2019年
3．雑誌名 IEICE Transactions on Fundamentals	6．最初と最後の頁 1159-1170
掲載論文のDOI（デジタルオブジェクト識別子） 10.1587/transfun.E102.A.986	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Onuki Ritsuko、Yamaguchi Rui、Shibuya Tetsuo、Kanehisa Minoru、Goto Susumu	4．巻 12
2．論文標題 Revealing phenotype-associated functional differences by genome-wide scan of ancient haplotype blocks	5．発行年 2017年
3．雑誌名 PLOS ONE	6．最初と最後の頁 0176530-0176530
掲載論文のDOI（デジタルオブジェクト識別子） 10.1371/journal.pone.0176530	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Taku Onodera, Tetsuo Shibuya	4．巻 96
2．論文標題 Succinct Oblivious RAM	5．発行年 2018年
3．雑誌名 Proc. STACS	6．最初と最後の頁 52:1-52:16
掲載論文のDOI（デジタルオブジェクト識別子） 10.4230/LIPIcs.STACS.2018.52	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計3件（うち招待講演　0件／うち国際学会　1件）

1．発表者名 Yao-zhong Zhang, Arda Akdemir, Georg Tremmel, Seiya Imoto, Satoru Miyano, Tetsuo Shibuya, Rui Yamaguchi
2．発表標題 Nanopore base-calling from a perspective of instance segmentation
3．学会等名 ISMB-ECCB 2019（国際学会）
4．発表年 2019年

1．発表者名 山岸大騎，髙木拓也，渋谷哲朗，有村博紀.
2．発表標題 重み付き有向非巡回グラフに対する効率良いテキスト索引の構築アルゴリズム
3．学会等名 第17回情報科学技術フォーラム
4．発表年 2018年

1．発表者名 小野寺拓，渋谷哲朗
2．発表標題 簡潔Oblivious RAM
3．学会等名 電子情報通信学会　情報セキュリティ研究会
4．発表年 2018年

〔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