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　２ 年　 ５ 月 ２９ 日現在

機関番号：１２６０１
研究種目：挑戦的研究（萌芽）
研究期間：2017～2019
課題番号：１７Ｋ２００２３

研究課題名（和文）家族性疾患解析のパラダイムシフトへ向けた全国民全ゲノム規模血縁推定基盤技術の開拓

研究課題名（英文）Development of kinship estimation methods on country-wide-size genome database for familial disease analyses

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）　　4,800,000円

研究成果の概要（和文）：次世代シークエンサー技術等の発達に伴い、世界各国で数十万人規模のゲノムプロジェクトが走っているが、その規模は年々大規模化の一途をたどっており、近い将来には国民全員規模のデータも得られるようになると考えられている。本研究では、そのような大規模化するデータベースにおいて網羅的家族性遺伝性疾患解析を行うための新たな技術の開拓を行った。具体的には、ゲノムワイドでの人種間のゲノム組み換えを検知する新たな技術の開発、高次元データ検索技術の開発、次世代シークエンサー高精度解析技術の開発などを行った。

研究成果の学術的意義や社会的意義
次世代シークエンサー技術の発達により、ゲノムデータベースのサイズは年々驚くべきスピードで大きくなっているが、それらを解析するための高効率な計算手法の開発はその発展スピードに追い付いているとはいえない状況にある。これを解決するために、近い将来に出現するであろう国民全員規模のゲノムデータベースを想定したアルゴリズム開発が必要となる。本研究では、そのような状況へ向けた新たな技術開発に成功するとともに、さらに今後の研究の核となる研究の基礎を固めることに成功した。

研究成果の概要（英文）：Due to the development of recent next-generation sequencers (NGS), many large-scale big genome projects are running now. As a result, genome databases becomes larger and larger, and it could be to the country-wide scale in the near future. We develop novel technologies useful for comprehensive familial genetic disease analyses, such as methods for detecting genome-wide arrangements between races, high-dimensional searching technologies, and high-accuracy NGS analysis tools.

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　人種間網羅的ゲノム機能解析　タンパク質立体構造検索　圧縮データ構造　次世代シークエンサー

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

近年登場した次世代シークエンサーは、その登場以前と較べて桁違いに高速・低コストなゲノム解読を可能とし、医学・生物学のあらゆる分野へ多大な変革をもたらしている。その産出データ量はムーアの法則をはるかに超える加速度で増えている。情報科学と医学・生物学の学際分野であるバイオインフォマティクス分野においては、それら次世代シークエンサーデータを効率的に処理するための様々なアルゴリズム研究、ツール開発等が多大なニーズに押され急ピッチで進んでいる。しかし、次世代シークエンサーのあまりに急激な進展に対応が追い付いているとは言い難い状況である。アルゴリズムの研究開発にかかる時間よりも次世代シークエンサーの開発サイクルが著しく短い現状がそのような対応の遅れにつながっている。

一方、ゲノムデータが今後もそのままムーアの法則を超える速度で指数関数的に伸びていくかというと、分野によっては上限があり得ると考えられる。例えば、国内で生存する個人の正常血液(germline)の全ゲノムデータは、その国の人口がサイズの上限である。そのような全国民規模のデータベースは、膨大な実現コストの問題と倫理上の困難を考えなければ、技術的には現状技術の延長で実現可能である。そのような実現可能な究極のデータベースに対して、現在のうちから新たな技術を開発しておけば、デファクトスタンダードとして世界のリードをとることができる可能性がある。

そして、そのような全国民規模の個人ゲノムデータベースの存在を仮定して初めて可能となる解析が、全国民規模個人ゲノムデータベースを活用した超大規模網羅的家族性疾患因子解析である。これは、データベースが全国民規模となる状況において、ほとんどどの個人について、近縁のみならず患者本人も把握できないような遠縁親族のデータが多数データベースに登録されている状況となることで、始めて考えられる解析戦略である。しかし、そのような親族データを見つけ出す方法、さらに見つけた後に解析する方法、そのいずれをとっても、現在の既存手法がそのような超大規模データベースへのスケーラビリティを持っておらず、既存技術では実現が極めて困難な状況にある。

２．研究の目的

全国民規模クラスの超大規模なゲノムデータベースが仮に構築されたならば、多くの患者について、近縁・遠縁親族のゲノムデータの多くがデータベースに含まれる可能性が高い。このことはすなわち、やはり倫理上の問題をクリアした上という前提は必要であるが、家族性疾患に関して研究・診断のパラダイムシフトが起こる可能性を示している。すなわち、家族性疾患の患者親族にゲノムデータ提供をケースバイケースでお願いする形の現在の研究から、大規模な網羅的解析研究・診断へとパラダイムシフトできる可能性がある。

本研究の目的は、現状の規模のデータベースでは不可能あるいは著しく困難であるようなビッグデータ時代にふさわしい家族性遺伝性疾患解析を、現在世界で走る大規模プロジェクトと比べてもはるかに大規模な全国民規模の個人ゲノムデータベースが登場した際に可能とするための、まったく新しいアプローチからの新たな解析技術の開発である。具体的には、まずそれを可能とする第一歩として、現状技術の延長では不可能と考えられる全国民規模の超大規模個人ゲノムデータベースにもスケーラブルな高速高精度血縁推定技術の確立をめざす。これは家族性疾患解析を超大規模データベース主導で行うのに必須な技術である他、D2C(Direct to consumer)遺伝子サービスや犯罪捜査・行方不明者推定などでも必要な技術である。これまでの血縁推定技術の多くは次世代シークエンサーを用いない小規模なものがほとんどであり、次世代シークエンサーを用いればより高精度の推定が可能であると考えられる。さらに、現状の家族性疾患解析技術の多くは大規模・網羅的に用いるにはあまりに計算コストが高く、スケーラブルではない。そのため本研究計画では、上記大規模高速血縁推定技術を活用しながら、超大規模ゲノムデータベース上での網羅的な解析をめざしたスケーラブルな家族性疾患解析技術の開拓も同時にめざす。

また、これらの研究と並行して、将来の全国民規模データベースの実現時に、これらの解析技術を実行するにあたって必要となる、超大規模データベースモデルの検討、圧縮技術、秘匿・暗号化技術など、周辺技術に関しても研究・検討を行う。

３．研究の方法

本研究では、超大規模高精度血縁推定技術の研究、超大規模家族性疾患解析技術の研究の２つの柱からなり、それらの研究によって将来全国民規模の超大規模個人ゲノムデータベースが登場した際に新たな大規模網羅的家族性遺伝疾患解析スキームを実現する基盤の構築をめざすとともに、その関連技術の研究も進めた。まず、家族性疾患の性質解明につながる研究として、人種間でのゲノム保存領域の違いに関する研究を行い、それらのクラスタリングを高精度に行う手法を開発した。また、大規模高次元データおよび大規模グラフの検索技術の開発を行った。さらにこれらの研究の関連技術として、大規模個人ゲノムデータベースを安全に検索・解析するためのプライバシー保護技術の研究も同時に進めた。さらに、高精度データベース作成にあたって必要な次世代シークエンサーデータ解析技術の開発も行った。これらの開発に平行して、今後の研

究の基礎技術として、個人ゲノムデータベースの表現技術の実装を進めた。

４．研究成果

（１）大規模家族性疾患クラスタリング技術の開発

遺伝性疾患や薬剤応答といった特性がどの程度人種間で保存され、それがどのようなゲノム上の差異をもたらしているか、は極めて重要な集団遺伝学上の課題である。本研究では、過去の正の自然選択（positive selection）がどのような順序で発生したのかに着目し、多人種大規模GWAS データを HHD とよばれる隠れマルコフモデルに基づく距離計算法によってクラスタリングを行う技法を開発することに成功した（PLOS ONE 2017、図１）。さらにその結果と代謝経路データベース情報を活用して実際の遺伝子アノテーションも行った。

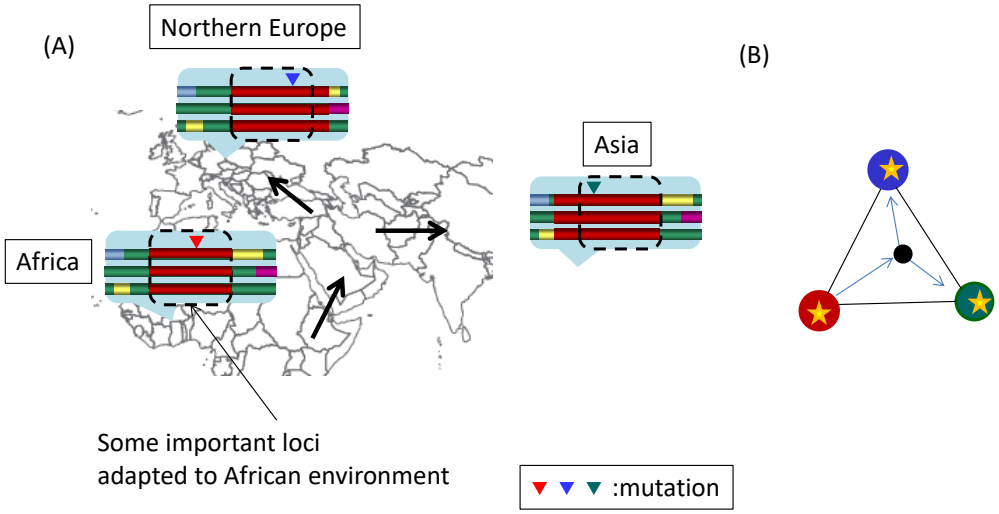

図１　ゲノム上の変異がいつ、どこで起こったか

（２）超大規模圧縮秘匿検索技術の開発

個人ゲノムデータは究極の個人情報とも呼ばれ、それらを扱う際には最大限のプライバシー保護が求められる。通常、ゲノムデータのプライバシー保護は不要時の暗号化や外部からの遮断、パスワードの設定などで行われるが、それでも様々なデータ漏洩リスクが常にある。近年の研究では、たとえデータが暗号化されていても、データベースやメモリのアクセスアドレス記録のみから様々な情報が漏洩することが指摘されている。これに対し、忘却 RAM とよばれるアクセス記録を忘却する技術が知られているが、従来の手法はゲノムデータベースのような超巨大データに対しては適用することが困難であった。それに対し、本研究では、忘却ＲＡＭに必要な付加記憶容量を漸近的に最小にすることに成功した（STACS 2018、表１）。これによって、ゲノムデータなど超大規模データのアクセス秘匿が可能となる道を開くことに成功した。

表１．世界初の劣線形サイズ忘却 RAM の開発

	アクセス時間	ストレージ容量
Square Root ORAM Goldreich. *STOC*87	$O(\sqrt{n})$ amortized	$O(\sqrt{n})$
Path ORAM Stefanov et al. *CCS*13	$O(\log^2 N)$	$>10N$
Ours (STACS 2018)	$O(\log^2 N)$	$O\left(\frac{N \log \log N}{\log^{1.4} N}\right)$

（３）高次元データベース検索技術の開発
様々な医療データベースは高次元のデータとしてあらわされる。また、タンパク質立体構造やＣＴ画像など、様々な医療関連データベースのデータが回転を考慮する必要がある。本研究では、回転を考慮した高速類似立体検索アルゴリズムの開発に成功した（IEICE Trans. Fundamentals. 2019）。

（４）次世代シークエンサーデータ解析技術の開発
次世代シークエンサーは多量のデータを輩出するが、その理由の一つがエラーが多いことである。そのために大量のデータを出力するとともに、元データとして保管を与儀なくされる。このため、全国民規模のデータベースにおいては、次世代シークエンサー出力をそのまま保管するのではなく、なるべく高精度なデータなデータのみを保管するようにすることによって、データの大規模化をとどめることが可能である。本研究では、ナノポアシークエンサーの出力を高精度化するとともに、エピゲノム解析などにも活用しやすいセグメンテーション解析を行う技術を開発した（ISMB-ECCB 2019）。

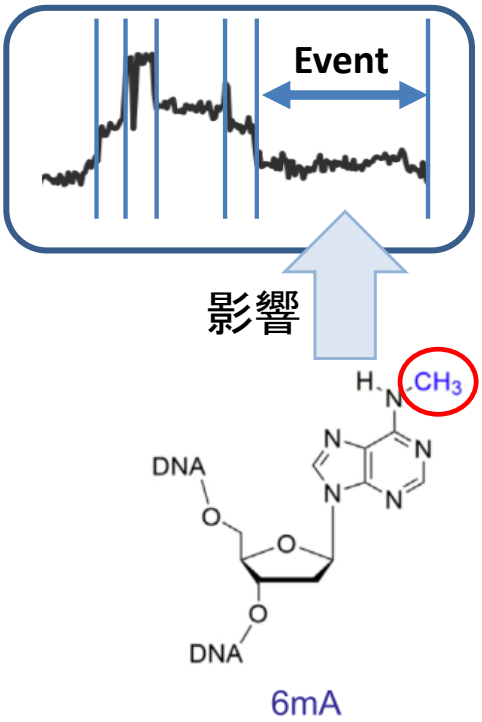

図２．セグメンテーション技術

５．主な発表論文等

〔雑誌論文〕　計5件（うち査読付論文　5件／うち国際共著　0件／うちオープンアクセス　5件）

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） 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 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 e176530～e176530
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 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 －

１．著者名 Tetsuo Shibuya	４．巻 13
２．論文標題 Application-Oriented Succinct Data Structures for Big Data	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） https://doi.org/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） https://doi.org/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 －

〔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 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

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

〔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