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 １０ 月 ２５ 日現在

機関番号：３２６８９
研究種目：基盤研究(C)（一般）
研究期間：2019～2022
課題番号：１９Ｋ１２２０９

研究課題名（和文）プライバシ保護ゲノム情報解析技術の開発

研究課題名（英文）Development of privacy-preserving technology for genome information analysis

研究代表者
清水　佳奈（Shimizu, Kana）

早稲田大学・理工学術院・教授

研究者番号：６０３６７０５０

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：近年，爆発的に増加している個人ゲノムデータの取り扱いには高いプライバシのリスクが付随するため，有用なデータが様々な組織に囲い込まれることが懸念されている．このような状況を鑑み，本研究では，主として秘密分散法に基づき，ゲノム配列を安全に検索する手法や，ゲノム情報を含む臨床情報等をクエリとして機械学習モデルを安全に評価することのできる手法を開発した．秘密計算は秘密を保護しない方法と比較して性能が大きく劣る問題があるが，本研究では索引機能と計算性能の両立を可能にする簡潔データ構造で用いられている技術の応用や，秘密分散法に加え，TEE等を用いた技術の開発により，実用に迫る性能を達成する手法を開発した．

研究成果の学術的意義や社会的意義
ヒトゲノム情報は医学や生物学の発展，さらには医療健康を中心とした様々な産業への活用が期待されている重要なデータであるが，個人と強く結びつく情報であるため，その取り扱いには困難が伴う．本研究では，このように活用が望まれる一方で，プライバシ保護の侵害につながるデータを安全に利用することのできる技術を開発した．学術的には，秘密計算と呼ばれる，秘密の情報を開示することなく目的とする演算を行うことのできる技術に関して，部分文字列検索の高速化，決定木評価の高速化等を達成した．本研究ではゲノムを中心とした生命科学分野のデータへの応用を想定して開発を進めてきたが，分野外の様々な問題にも適応可能である．

研究成果の概要（英文）：The recent advancement in genome sequencing technology poses a new challenge to securely sharing personal genome information. To solve the problem, we developed privacy-preserving technologies such as searching personal genome sequence databases mainly by using secret sharing. To achieve the performance required for practical use, we applied an efficient algorithm used in the succinct data structure, which achieved information-theoretically lower bound of memory size while achieving high utility, and used hardware techniques such as TEE. We also developed a method that enables a user to use a machine-learning model (in our case, a decision tree) on a server while protecting the user's query and the server's machine-learning model by using secret sharing.

研究分野：生命情報科学

キーワード：ゲノム配列検索　プライバシ保護　秘密分散　暗号プロトコル　TEE　機械学習

１．研究開発当初の背景

近年，爆発的に増加している個人ゲノムデータの取り扱いには高いプライバシのリスクが付随するため，有用なデータが様々な組織に囲い込まれて孤立するサイロ化と呼ばれる現象が多発している．統計や機械学習を用いてゲノム情報を解析する際には，データの種類が豊富でサンプル数が多いほど正確な結果を得ることができるため，サイロ化したデータを安全かつ，効果的に集約し，有用な知見を発見する方法論の開発が強く望まれている．こ

２．研究の目的

上述の背景から本研究では，ゲノム情報を秘匿したまま情報解析を行う方法論の研究を行うことを目的とした．本研究では特に，(1)ゲノム配列検索と(2)ゲノムワイド関連解析の２点を中心的な課題と定め，大規模なデータ解析を安全に実施できる手法の開発を目指した．

３．研究の方法

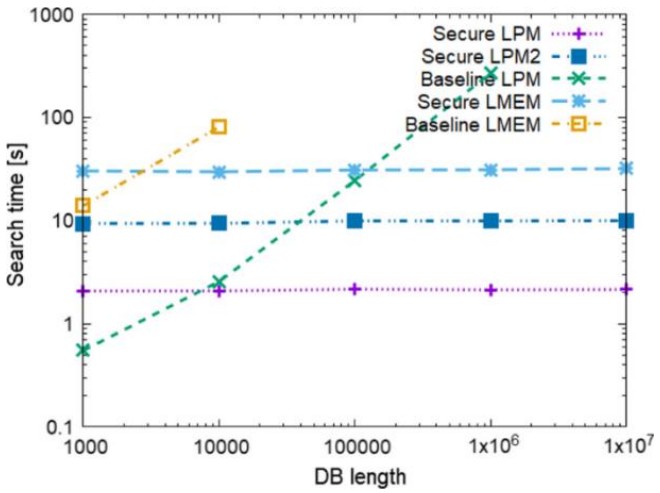
本研究では，ゲノム情報を秘匿したまま情報解析を行う方法として，個別の問題の特性を明らかにしたのち，各問題の解決に適切な技術を用いることとした．また，解決に用いる要素技術としては，秘密分散法を中心とする暗号プロトコルや Trusted Execution Environment（TEE）等のハードウェア技術を検討することとした．各要素技術の単独利用にとどまらず，要素技術を組み合わせる方法についても検討することとした．

４．研究成果

本研究では，当初の計画通り，秘密分散法に基づくゲノム配列検索法と Trusted Execution Environment（TEE）に基づきゲノムワイド関連解析を行う手法を開発した．これらに加え，近年，実用化の望まれている機械学習モデルの評価についても取り組み，秘密分散法に基づく決定木の評価手法も開発した．

４．１　秘密分散法に基づくゲノム配列検索法の開発

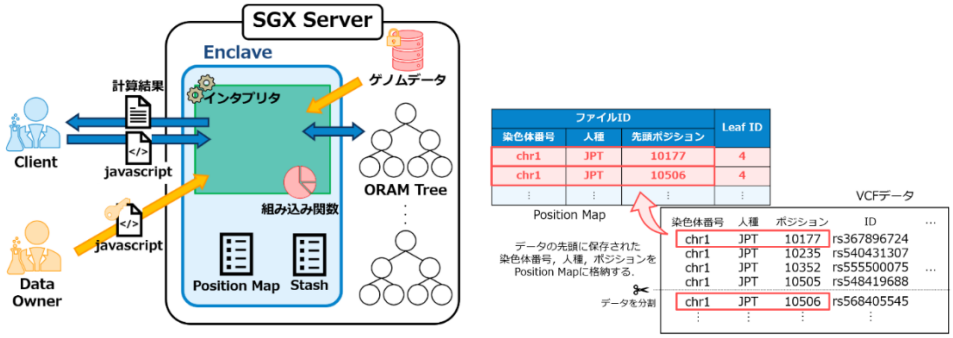
ゲノム配列や医療文書の検索は，情報分析の要となる重要なタスクである．本研究では，個人のゲノム配列等を持つユーザーが，大規模なゲノム配列等のデータベースを持つサーバー上で，文字列の一致等を検索することのできる手法を開発した．提案手法では，秘密計算技術（秘密の情報を開示することなく目的とする演算を技術）の一つである秘密分散法を用いた．秘密分散法では，計算ラウンド数や通信量が性能のボトルネックとなるが，我々は，簡潔データ構造と呼ばれる，索引機能と計算性能を両立するデータ構造で用いられている方法を応用して事前計算を工夫することにより，クエリの投入から検索結果を得るまでのオンライン計算に必要な時間計算量，通信量，ラウンド回数がデータベース長に依存せず，クエリ長のみに依存する方法を開発した．一般的な情報検索では，クエリ長はデータベース長と比較して非常に小さいため，ゲノムデータベースのような膨大な情報に対しても非常に高速に動作する．プロトタイプによる実験では，長さ一千万のゲノムデータベースへの検索が実際のインターネット環境でも 10 秒程度となることを確認した．以下のグラフでは，ベースライン法と比較して，データベース長に依存せず実用的な性能を達成していることが示されている．



上記の成果に加え，事前計算量を抑制するアルゴリズムの設計も行った．これらの成果により，2019 年度のコンピュータセキュリティシンポジウムにて研究奨励賞の受賞，国際会議での論文発表，国際論文誌での論文発表を達成した．

４．２　Trusted Execution Environment（TEE）に基づくゲノムワイド関連解析法の開発

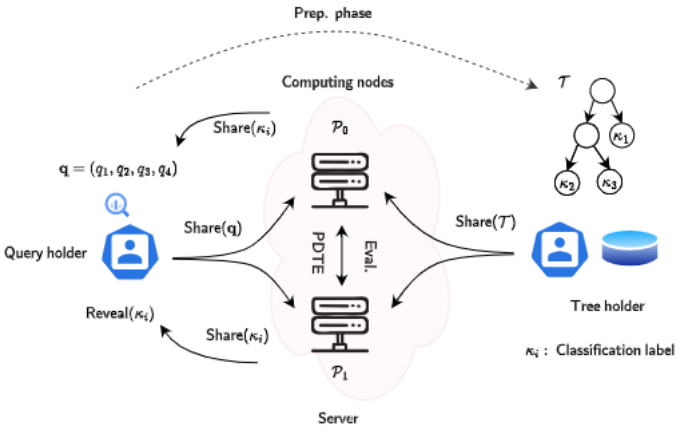
CPU の機能の拡張等により，コンピュータ上にプログラムの安全な実行環境を実現する技術を Trusted Execution Environment と呼ぶ．本研究では，TEE を実現する技術の一つである Intel 社の SGX を用いて，GWAS を行うことのできる情報分析プラットフォームを考案し，そのプロタイプ実装を行った．開発したシステムでは，全ゲノム相関解析やデータのクラスタリングを行うことができる他，データのアクセスパターンを秘匿する Oblivious RAM を用いる事により，巨大なデータにも高速にアクセスすることができる．データ分析は，ユーザーが JavaScript 等のプログラミング言語により記述し，サーバー上の Enclave 内に配備した仮想マシンがサーバー側に情報を漏らすことなく実行できる．以下にシステムの概要図を示す．200 人以上のゲノム変異データを用いた実験では，情報保護をしないソフトウェアと同等の時間で解析を行えることを確認した．結果を以下の表に示す．これらの成果により，2020 年の暗号と情報セキュリティシンポジウム（SCIS2020）において SCIS 論文賞を受賞した．



	Fisher			LR			PCA		
	SGX	C++	python	SGX	C++	python	SGX	C++	python
ファイル検索	0.613	1.64	3.78×10^{-4}	65.1	178	37.5	30.1	82.9	23.0
ゲノム解析	7.1×10^{-5}	0.0016	0.0042	0.039	0.034	0.0043	0.013	0.0087	0.0074
Total	0.614	1.64	0.799	65.1	178	37.5	30.1	82.9	23.0

４．３　秘密分散法に基づく決定木の評価手法の開発

近年，機械学習の実応用が進み，大規模なデータで学習をした機械学習モデルをエンドユーザーが利用する場面が増加している．医学，医療，健康産業の現場においては，特に，ユーザーの秘密の情報を保護しつつ，学習に用いたデータに含まれる個人情報や知的財産を保護する目的で，機械学習モデルそのものも保護する必要性が高まっている．このような状況を鑑み，本研究では，ゲノム情報を保有する小規模医療施設や個人などを想定したユーザーと，多数の臨床情報をもとに学習した機械学習モデルを保持する医療機関などを想定したサーバーが，互いに秘密情報を明かさないまま，ユーザーがサーバーの持つ機械学習モデルを用いて医療診断等の予測結果を得ることのできる手法の開発に取り組み，秘密分散法を用いて決定木を評価することのできる暗号プロトコルを開発した．提案手法では，木構造をあらかじめランダムにシャッフルにしておくことにより，秘密分散法を用いる上での性能のボトルネックとなる計算ラウンド数や通信量を削減することが可能となり，クエリ投入から計算終了までのオンライン計算時間を従来手法よりも改善することができた．



５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　3件／うち国際共著　0件／うちオープンアクセス　2件）

１．著者名 Nakagawa, Yoshiki ; Ohata, Satsuya ; Shimizu, Kana	４．巻 201
２．論文標題 Efficient Privacy-Preserving Variable-Length Substring Match for Genome Sequence	５．発行年 2021年
３．雑誌名 21st International Workshop on Algorithms in Bioinformatics (WABI 2021)	６．最初と最後の頁 2:1--2:23
掲載論文のDOI（デジタルオブジェクト識別子） 10.4230/LIPIcs.WABI.2021.2	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 Ahmed Mohammad Nabil、Shimizu Kana	４．巻 13849
２．論文標題 Private Evaluation of a Decision Tree Based on Secret Sharing	５．発行年 2023年
３．雑誌名 Lecture Notes in Computer Science	６．最初と最後の頁 171～194
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-031-29371-9_9	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 Nakagawa Yoshiki、Ohata Satsuya、Shimizu Kana	４．巻 17
２．論文標題 Efficient privacy-preserving variable-length substring match for genome sequence	５．発行年 2022年
３．雑誌名 Algorithms for Molecular Biology	６．最初と最後の頁 1-22
掲載論文のDOI（デジタルオブジェクト識別子） 10.1186/s13015-022-00211-1	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 －

〔学会発表〕　計11件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 Taiki Yamada, Kenichi Chiba, Keisuke Kataoka, Yuichi Shiraishi and Kana Shimizu
２．発表標題 Novel reference-free detection of somatic structural variant based on graph-aware index
３．学会等名 IIBMP 2020
４．発表年 2020年

１．発表者名
中川 佳貴　　大畑 幸矢　　清水 佳奈

２．発表標題
秘密分散に基づく秘匿全文検索

３．学会等名
コンピュータセキュリティシンポジウム2019（CSS2019）

４．発表年
2019年

１．発表者名
岩田大輝　清水佳奈

２．発表標題
Intel SGXを用いた個人ゲノム情報解析システム

３．学会等名
2020年暗号と情報セキュリティシンポジウム（SCIS2020）

４．発表年
2020年

１．発表者名
岩田大輝　清水佳奈

２．発表標題
属性ベース暗号とIntel SGXを用いた堅牢かつ柔軟なアクセス制御を実現するデータ分析プラットフォームの構築

３．学会等名
情報処理学会 第58回バイオ情報学研究会

４．発表年
2019年

１．発表者名
櫻井碧　清水佳奈

２．発表標題
BV-SGX: 生命情報解析向け仮想マシンを搭載したSGXクラウド

３．学会等名
情報処理学会 第58回バイオ情報学研究会

４．発表年
2019年

１．発表者名
清水佳奈

２．発表標題
生体情報セキュリティ

３．学会等名
第58回 日本生体医工学会大会

４．発表年
2019年

１．発表者名
Kana Shimizu

２．発表標題
Towards Privacy-preserving Biomedical Knowledge Integration

３．学会等名
JAPAN - NORDIC WORKSHOP ON DIGITAL HEALTH FOR HEALTHY LONGEVITY

４．発表年
2019年

１．発表者名
Masanobu Jimbo, Nobutaka Mitsuhashi, Shigeo Mitsunari, Shin Kawano, Toshiaki Katayama, Kiyoshi Asai, Kana Shimizu

２．発表標題
Privacy-preserving Search for Sharing Genetic Variants

３．学会等名
The 27th International Conference on Intelligent Systems for Molecular Biology and the 19 th European Conference on Computational Biology (ISMB/ECCB 2019)

４．発表年
2019年

１．発表者名
清水佳奈

２．発表標題
秘密計算による安全なゲノム配列検索

３．学会等名
2022年電子情報通信学会総合大会

４．発表年
2022年

１．発表者名
清水佳奈

２．発表標題
生命科学・医薬開発におけるプライバシ保護技術の活用

３．学会等名
第12回CSJ化学フェスタ

４．発表年
2022年

１．発表者名
清水佳奈

２．発表標題
プライバシー保護ゲノム解析技術の現状と課題

３．学会等名
日本人類遺伝学会第67回大会

４．発表年
2022年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

学会における受賞：

岩田大輝，SCIS2020暗号と情報セキュリティシンポジウム SCIS論文賞（「Intel SGXを用いた個人ゲノム情報解析システム」），2020年4月

中川佳貴，大畑幸矢，清水佳奈，コンピュータセキュリティシンポジウム2019（CSS2019）奨励賞（「秘密分散に基づく秘匿全文検索」），2019年10月

櫻井碧，第58回バイオ情報学研究会 SIGBIO優秀プレゼンテーション賞（「BV-SGX: 生命情報解析向け仮想マシンを搭載したSGXクラウド」），2019年6月

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