Intel SGXによるプライバシー保護 生命情報解析プラットフォーム

櫻井 碧[1,2], 岩田 大輝[1,3], 清水 佳奈[1,3]

1. 早稲田大学基幹理工学研究科 情報理工・情報通信専攻.
2. 2019年度未踏IT人材発掘・育成事業.
3. 産総研・早大 生体システムビッグデータ解析 オープンイノベーションラボラトリ.

概要: 個人のプライバシーや企業秘密を含む生命情報は，適切に保護したまま共有・分析される必要がある．データを秘匿したまま生命情報解析を行う従来研究の多くは，膨大な計算量が生じる技術に基づいており，実際の解析に必要な計算性能に遠く及ばない問題がある．そこで本研究では，ハードウェアの新技術として注目されるIntel SGXを利用して，プライバシー保護を考慮しないソフトウェアと同等の性能を有する真に実用的な解析プラットフォームを提案する．Intel SGXはCPU内の暗号エンジンによりRAM上に保護領域を形成し，データを秘匿しながらプログラムを実行する技術であり，データの秘匿に膨大な計算量を必要としない．提案プラットフォームでは，データ所有者がサーバに秘密のデータを保存し，データ利用者がサーバ上で情報解析を行って結果のみを取得することができる．具体的には，次の二つのシステムを開発した．

1. 保護領域内に仮想マシンを構築し，ユーザーが自由に解析法をプログラミングすることのできるシステム

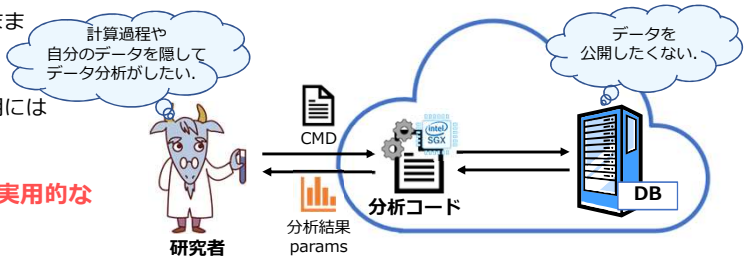
2. 解析方法を機械学習などの特別な用途に固定し，ユーザーの権限に応じて利用できるデータの範囲を柔軟に変化させることのできるシステム

実データを用いて配列検索や機械学習による解析を実施し，いずれのシステムも既存の非プライバシー保護プログラムと比較して同等の計算性能を達成したことを確認した．

1. 背景と目的

目的: 膨大な計算量が必要にならない，真に実用的な，データを秘匿したまま生命情報解析を行う手法の開発.

- 近年，生命情報を取得しやすい環境になり，これらを適切に保護したまま共有・分析をする手法の必要性が高まっている.
  - ➢ 医療機器の発達，分析手法の多様化...
- データを秘匿したまま分析を行う様々な手法が開発されているが，実用には至っていない.
  - ➢ Garbled Circuit，秘密分散，準同型暗号... といった従来手法.
- この問題を解決するため，本研究では**Intel SGXを利用した，真に実用的な新しい解析プラットフォーム**を提案する.



2. 使用技術

Intel SGX

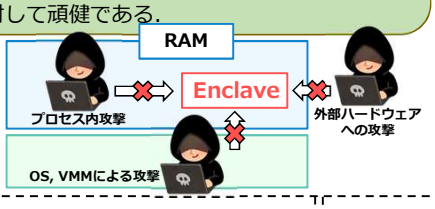
ハードウェアによるセキュリティ技術の一つ．Skylake以降のIntel製CPUで利用可能.

Enclave

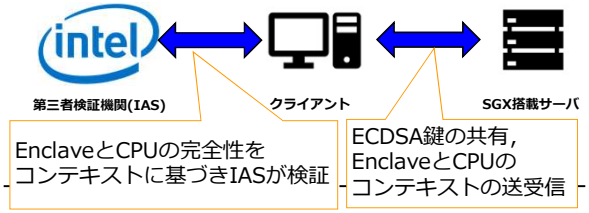
- **RAM上に構築可能な保護領域**のこと．Enclave内のデータ，プログラム実行は外部から秘匿することができる.
  - ➢ マシンの起動時にDRAM上に特別な領域が確保され，以降，CPUと当該領域の間で流通する情報は全てチップ内の暗号化回路を経由する．Enclaveはこの領域内に作られるため，仮にOSが乗っ取られたとしても情報漏洩が起きない.
- そのため，図のように様々な脅威に対して頑健である.

リモート・アテステーション (RA)

- Intel SGXが提供するセキュリティ機能のひとつ.
- RAにより遠隔計算機のEnclaveを利用することができる.
- クライアントは，Intel社が提供する検証機関(IAS)を介して，遠隔のEnclaveが信頼できるか検証できる.

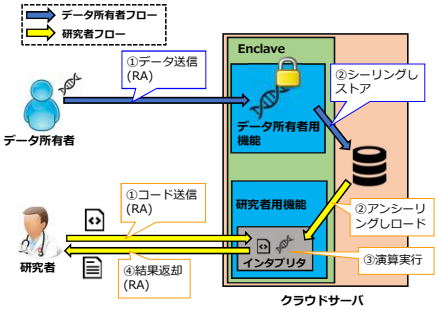


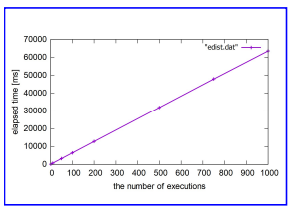
- **サイドチャネル攻撃**に対しては**別途対策**が必要.



3. 提案手法

- **Enclave上に独自の生命情報解析に特化したインタプリタを搭載する事で，高安全性・低計算負荷・低利用難易度を同時に実現.**
- **SGXが本来対応していない「コードの保護」も実現しており，SGXの弱点であるサイドチャネル攻撃も大幅に軽減可能.**

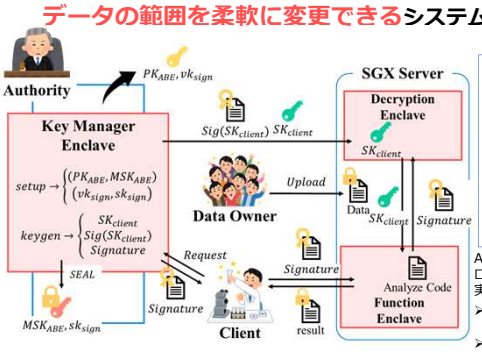


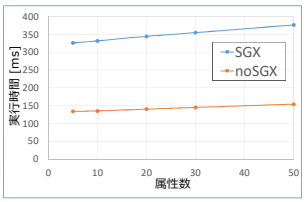


64文字同士のゲノム配列50000組の編集距離計算を1~1000ループ実行した場合の実行時間のグラフ

- クライアントに提供する分析結果から個別のデータの情報が推測できないように，インタプリタの仕様を工夫している.
- Enclaveのサイズに制限があることを鑑み，軽量なインタプリタを実装した.

- **Intel SGXを用いることで，データを秘匿したまま，機械学習等の従来の秘密計算では実現が難しい分析を可能にした.**
- **属性ベース暗号を用いることで，ユーザーの権限に応じて利用できるデータの範囲を柔軟に変更できるシステムを開発した.**





ABEで暗号化したサンプルゲノムデータを用いたロジスティック回帰分析を実行した場合の実行時間のグラフ

- ➢ 一塩基多型を含む104 項目について癌患者1579人分の2値データ.
- ➢ Iteration: 100, 学習率: 0.01, 正則化項 λ: 1.

- 二つのEnclaveを連動させ，署名による検証を実施することにより，Clientサイドでの鍵の管理を不要にした.
- 従来手法（完全準同型暗号によるシステム）と比較して，**1.54×10^6 倍高速**であった.

双方のアプローチは，同一の環境で実験を行った.
実験環境: Ubuntu 16.04.6 LTS, Intel Core i7-7700K CPU @ 4.20GHz, 16GB memory

※ この研究は，2019年度未踏IT人材発掘・育成事業の支援を受けています.

まとめ: Intel SGXを用いた二つの生命情報秘匿分析システムは，膨大な計算量を必要とする従来手法よりも，軽量で高速であることが実験により示された．ゲノムデータをはじめとする生命情報はデータサイズが膨大になる傾向があり，従来手法では分析が困難であるため，提案手法は生命情報解析の更なる発展への貢献が期待できる.