完全準同型暗号によるゲノム秘匿検索の分散処理に関する一検討

山本 百合† 小口 正人†

†お茶の水女子大学

1. はじめに

バイオインフォマティクスの研究において，研究機関や各病院が所持するゲノムデータの活用が求められている．そのため，大量のゲノムデータをクラウド事業者などに委託し，利用者が問い合わせを行うことで統計処理が可能なゲノムデータ委託システムが今後広がっていくと考えられる．しかしヒトゲノムは個人の識別子となるため，プライバシ保護の観点から，暗号を適用した秘匿検索手法によるデータ活用が必要となる．またサーバ・クライアント型のゲノム秘匿検索の委託計算を行う際に適用する暗号方式として従来の共通鍵暗号を用いると，サーバ側での復号が可能となるためデータの秘匿が難しくなる [1].

先行研究 [2] では，暗号文同士の加法と乗法が成立する完全準同型暗号を秘匿検索に適用し，復号することなく統計処理などの複雑な演算が可能な手法を目指し，アルゴリズムの高速化を進めている．しかしながら，完全準同型暗号演算の計算量が大きいために，サーバ側での計算負荷が大きくなりやすい．本研究では，先行研究が用いる手法のサーバ側での演算に対してマスタ・ワーカ型の分散処理を適用することで高速化を行い，クラウドコンピューティングへの適用可能な分散処理システムの実装を目指す．

2. 先行研究

石巻ら（2015）は従来のゲノム秘匿検索手法に対して完全準同型暗号を用いることで，複雑な演算処理が可能なゲノム委託計算の整備と，複数の平文を一つの暗号文にまとめて並列計算を行う暗号文パッキングによる計算量の削減を行った [2]．石巻らの手法は，サーバとクライアントが 1:1 で問い合わせるゲノム秘匿検索システムにおいて実験が行われている．システム動作の概要は，サーバはクライアントからゲノム検索文字列を受け取り，自身が所有するゲノムデータとのマッチを計算し，クライアントに結果を返す．このときサーバとクライアント双方のデータを互いに秘匿するために，サーバがノイズを加える再帰的紛失通信を利用する．そのため，クライアントは検索文字列を 1 文字ずつ暗号化した上で送信し，返ってきた結果を利用して次の文字に対する問い合わせを作成する．また，クライアントは受け取った演算結果同士の比較によって演算結果を得ることができる [2]-[3].

3. 提案手法

本研究では，以下の完全準同型暗号を用いたゲノム秘匿検索のマスタ・ワーカ型の分散システムを提案する．提案システムの概要を図 1 に示す．

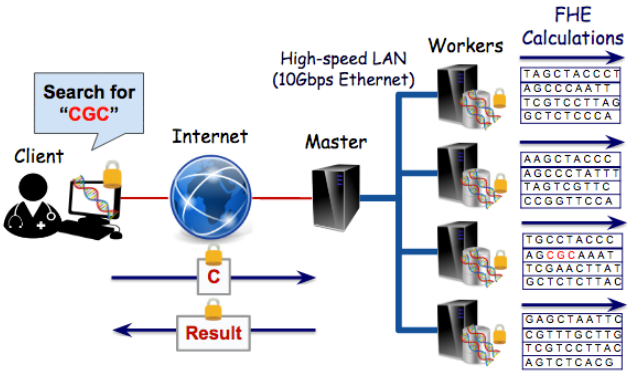

図 1: 提案手法概観

(1) クライアントはクエリの 1 文字を暗号化し，公開鍵と共にマスタへ送信する．

(2) マスタは受け取ったデータを各ワーカに転送する．

(3) ワーカは完全準同型暗号を用いた演算を行い，結果をマスタへ転送する．

(4) マスタは結果を収集し，クライアントへ結果を送信する．

(5) クライアントは復号を行い，結果を得る．また，その結果を用いてクエリの次の 1 文字を暗号化した後に再びマスタに送信する．

(6) (2)〜(5) をクエリの長さの回数分，繰り返して終了する．

以上のプロトコルを C++で実装した．また，完全準同型暗号計算には HElib[4] を，分散化における各マシンの制御のための MPI(Message Passing Interface) を利用するライブラリとして，Open MPI[5] を用いた．

4. 分散方法

今回適用するアプリケーションの分散化手法として，個体のデータごとでデータベースを分割するデータの分散，独立性の高い計算部位に適用する分散処理，また独立性の高い手順に適用する分散処理などが考えられる．先行研究の秘匿計算手法は，クエリとデータベースの要素に対して完全準同型暗号演算を行う手法を適用している．そのため，分散化した各ワーカマシンにデータベースを設置することにより，クエリとデータベースの要素間の暗号演算が可能となり，より多くのクエリとデータベース間のマッチング有無を調べることが可能となる．したがって今回の提案システムでは，もっともシンプルなデータベースの分割による分散処理を適用した．

A study of distributed system for genome secret search implemented with fully homomorphic encryption

† Yuri YAMAMOTO, † Masato OGUCHI

Ochanomizu University (†)

5. 実験

5.1 実験環境

実装したプログラムを複数のマシンにおいて実行した．各マシンの環境は，Intel®Xeon ®Processor E5-2643 v3 3.4GHz，6 コア，12 スレッド，メモリ容量 512GB，ストレージは RAID0 の SSD が 480GB，HDD が 2TB であり，同スペックのマシンを 4 台使用する．1 台をマスタの機能を持ったマシンとし，同時にワーカとして 1 スロット分の演算も行う．また他 3 台をワーカとして最大 2 スロット稼働させた．最大で 7 スロット分のワーカを稼働させてワーカ数ごとの実行時間を比較する実験を行った．実験に使用するゲノムデータは 1 サンプルあたり 10,000 文字のデータを 2,184 サンプル用意した．また，検索クエリは長さ 5 文字のものを用いた．さらに秘匿検索の秘匿性を高めるダミー検索を加えることにより，1 文字あたりデータベースの 50 箇所を始点とした文字列検索を行った．

5.2 実験結果

クエリとデータベースとのマッチングの有無の判定を行うプログラムを分散化した環境上で稼働させた．ワーカ数ごとのマスタとクライアントの実行時間のグラフをそれぞれ図 2，図 3 に示す．

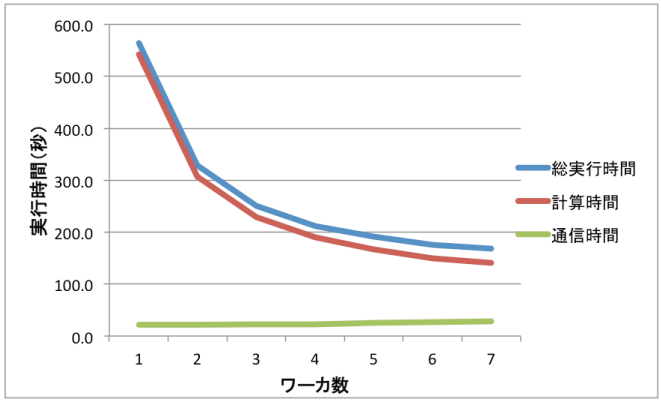

図 2: ワーカ数ごとのマスタにおける実行時間 (秒)

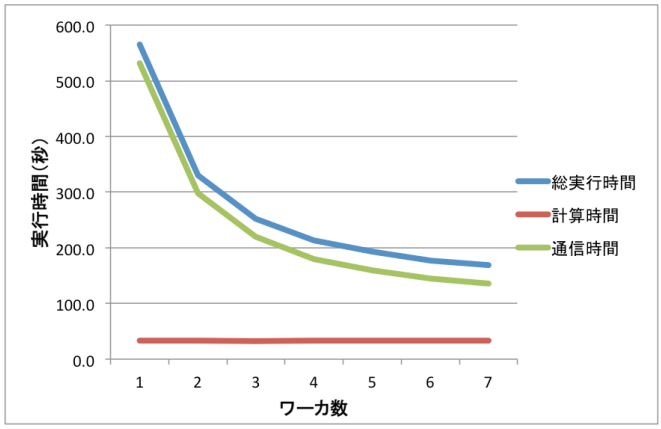

図 3: ワーカ数ごとのクライアントにおける実行時間 (秒)

マスタ側の実行時間において，ワーカ数が増加するにつれて計算時間を減少させることができた．しかし，計算手順の中にデータベースの大きさに依存しない計算が含まれることから，ワーカ数が増えるにつれて計算時間における分散化効果は徐々に横ばいになっていることがわかる．通信時間に関しては，ワーカ数の増加に対してクライアントにおける復号計算時間も含んだ通信時間はほとんど変化しないため，クライアントとの相互通信におけるオーバヘッドは小さい．

クライアント側の実行時間においては，現在マスタ側の計算が終了するまでクライアントは待機する仕様となっているため，マスタの計算時間の分だけ通信時間がかかる．クライアントにおける復号計算時間はワーカ数による大きな変化は見られなかった．

6. まとめと今後の課題

今回の実験により，完全準同型暗号を用いたゲノム秘匿検索にマスタ・ワーカ型の分散処理を適用すると，計算時間が分散台数に応じて減少した．しかし，分散台数が増えるにつれて分散化効果が横ばいになるなどの課題もある．

今後は，実装を改善し，より高速な分散処理が可能な手法や本手法が効果的に作用するアプリケーションへの適用を提案していきたい．

7. 謝辞

本研究を進めるにあたり，大変有益なアドバイスを頂いた早稲田大学山名研究室及びに工学院大学山口研究室の皆様に感謝いたします．

特に早稲田大学山名研究室所属の石巻さんからは，ゲノム秘匿検索システムのプログラムと多くの助言を賜りました．深く感謝いたします．

また本研究は，JST CREST の支援を受けております．

参考文献

[1] 草川恵太：「完全準同型暗号の概要」, 電子情報通信学会誌, Vol. 99, No. 12, pp. 1151–1158, 2016 年 12 月

[2] 石巻優, 清水佳奈, 縫田光司, 山名早人：「完全準同型暗号を用いた高速なゲノム秘匿検索」, SCIS 2016, 2A2-2, 2016 年 1 月

[3] Yu Ishimaki, Kana Shimizu, Koji Nuida, Hayato Yamana : “Poster:Privacy-Preserving String Search for Genome Sequences using Fully Homomorphic Encryption,” the 37th IEEE Symposium on Security and Privacy, 2016 年 5 月

[4] Shoup V. and Halevi S. : http://shaih.github.io/HElib/index.html, 2017 年 1 月閲覧

[5] Open MPI : https://www.open-mpi.org/, 2017 年 1 月閲覧