DEIM Forum 2013 P2-5

データアウトソーシングにおける プライバシー保護 SNP 検索システムの実装

嶺岸　和城† 　清水　將吾†† 　權　 娟大††† 　尾崎　敬二†

† 国際基督教大学教養学部　〒 181–8585 東京都三鷹市大沢 3–10–2
†† 学習院女子大学国際文化交流学部　〒 162–8650 東京都新宿区戸山 3–20–1
††† 東京理科大学薬学部　〒 278–8510 千葉県野田市山崎 2641
E-mail: †c131257e@yamata.icu.ac.jp, ††shogo.shimizu@gakushuin.ac.jp, †††yekwon@rs.noda.tus.ac.jp, ††††keiji@icu.ac.jp

あらまし　一塩基多型（SNP）とは，標準的な塩基配列と一塩基が異なることであり，各個人差の原因とも言われていることから個人向け医療への応用が期待されている．一方，コスト削減のため，データベースの運用管理をアウトソースすることがある．このようなサービスを利用して SNP 検索システムを実現する場合，委託先の管理者からも利用者の SNP 情報を秘匿化できることが望ましい．攻撃者が背景知識を持つ場合，暗号化されたデータや問合せの頻度を用いた攻撃に対する耐性を持つ必要がある．本稿では，ハッシュに基づく索引構造と秘匿検索プロトコルを用いて，プライバシー保護 SNP 検索システムを実装し，その効率性について評価する．
キーワード　データアウトソーシング，プライバシー保護，SNP 検索

Implementation of a Privacy-Preserving SNP Search System in Data Outsourcing

Kazushiro MINEGISHI†, Shogo SHIMIZU††, Yeondae KWON†††, and Keiji OSAKI†

† College of Liberal Arts, International Christian University, 3–10–2 Osawa, Mitaka, Tokyo 181–8585, Japan
†† Faculty of Intercultural Studies, Gakushuin Women's College
3–20–1 Toyama, Shinjuku, Tokyo 162–8650, Japan
††† Faculty of Pharmaceutical Sciences, Tokyo University of Science
2641 Yamazaki, Noda, Chiba 278–8510, Japan
E-mail: †c131257e@yamata.icu.ac.jp, ††shogo.shimizu@gakushuin.ac.jp, †††yekwon@rs.noda.tus.ac.jp, ††††keiji@icu.ac.jp

Abstract　A single nucleotide polymorphism (SNP) is a DNA sequence variation occurring when a single nucleotide in the genome differs from many other DNA sequences. SNPs are considered to be the cause of personal differences and expected to be applied to personalized medicine. On the other hand, to reduce the cost of managing databases, database administration tasks are often outsourced. In outsourced environments, it is desirable that SNP contents are technically protected against database administrators. When attackers have domain knowledge about databases, protection methods should be secure from attacks using the frequency of encrypted data and queries. In this paper, we implement a privacy-preserving SNP search system using hash-based index structures and secure search protocols and evaluate its efficiency.
Key words　data outsourcing, privacy protection, SNP search

1. はじめに

Single Nucleotide Polymorphism（SNP）とは，点突然変異によって起こる局所的な DNA のバリエーションが集団内で 1% 以上を占めるタイプのことを言う．SNP は DNA 上の様々な領域で発見されるが，構造遺伝子や調整遺伝子など重要な部分の上で起きているものは正常な発現に支障をきたし，個体の疾病や身体の表現型などの特徴として現れる．このうち，個

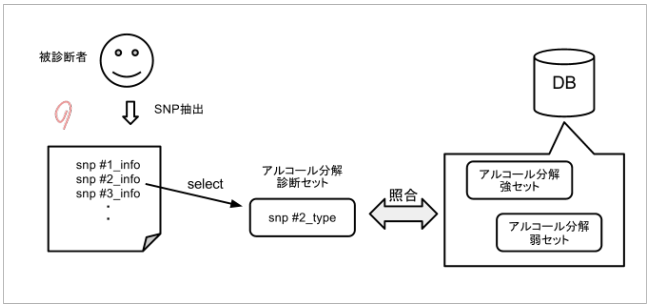

図 1　SNP 診断の概略

体の特徴とその原因 SNP がほぼ完全に解明されているものもあるが，単一或いは複数の SNP が個体の特徴と関連があることが分かっているものもある．これらの発見は Genome Wide Association Study（GWAS）と呼ばれるゲノムの統計学的な解析によって行われ，近年非常に活発な研究分野に成長している．

GWAS が注目されている理由として，癌やアルツハイマーの新たな診断方法とする臨床応用や DNA 鑑定，先祖解析など商業的価値の高さが挙げられる．米国 National Center for Biotechnology Information（NCBI）は SNP の情報を共有するためのデータベース dbSNP を公開している．Online Mendelian Inheritance in Man（OMIM）データベースでは疾病関連 SNP の情報が蓄積されている．更に，SNPedia という wiki を用いた疾病関連 SNP のデータベースも存在する．GWAS の商業的利用では 23AndMe（`http://www.23andme.com/`）が唾液から個人の SNP を解析し，家系分析や疾病の傾向などの情報を提供するサービスを提供している．

このような個人 DNA の解析が盛んになりつつある中で，プライバシーの観点から DNA 情報は特に慎重な扱いが求められる．本稿では，個人 DNA のプライバシー保護や計算量上の問題を踏まえた上で，今日の GWAS や DNA 照合の業務モデルを想定した SNP 照合手法を提案する．以下に本稿の構成を示す．2 章では，問題設定について説明する．3 章では，2 章で述べた状況を想定し，プライバシー保護を実現するための要素技術を紹介する．4 章では，SNP 診断手順について述べる．5 章では，4 章の手順を実装し，各種のパラメータを変えながら実行時間の変化を調べ，効率化の手法も取り入れて比較する．最後に，6 章でまとめと今後の課題について述べる．

2. 問題設定

GWAS では，例えばがん患者のグループで SNP を網羅的に調べて，非がん発症者のグループと比較し，がん患者グループのみに共通する SNP を洗い出すといった解析を行う．このようにして得られた SNP は疾病関連 SNP と呼ばれ，特定の疾病と強い関連性があると考えられている．このような疾病関連 SNP や生体的特徴関連 SNP を用いた診断は，被診断者から抽出した網羅的 SNP から特定の診断を行うために選ばれた特定の SNP セットと陽性の SNP セットとの照合によって行われる．

SNP 診断では，図 1 のように，照合対象となる SNP セットが多い場合データベースに保管し，照合の際に問合せを行う．更に，データベースの運用管理コスト削減のため，DaaS（Database-as-a-Service）の利用，つまりデータを生成する組織がデータベース管理業務を専門の業者に外部委託することも実際の業務モデルとして考えられる [4]．このようなデータアウトソーシング型 SNP 診断では，外部に個人 DNA 情報の流出を防ぐための工夫が必要となる．

まず，データの格納や通信において，個人の SNP に関する情報はすべて暗号化を施す必要がある．例えば，耳あかの dry/wet を診断する SNP である rs17822931 はこれ自体は特に重要な個人情報を持つ DNA とは言えないが，同じ SNP が日本人女性の乳がんと関連があることが知られている．このように，一つの SNP が複数の 診断用 SNP セットに含まれる可能性を考えると，すべての SNP 情報は重要な情報と捉えて秘匿することが望ましい．23andMe のようなサービスにおいても，もし個人の結果が公になると各項目の診断結果から使用した SNP 診断セットと型を予想し，それらの要素を組み合わせることで，がん診断やアルコール依存診断など意図しない結果が暴露される可能性がある．

更に，暗号化されたデータの出現回数の統計的頻度から，元データの分布に関する背景知識を利用して暗号文と平文の対応関係を推測する頻度攻撃について対策する必要がある．平文が固定的であったり或いは確定的暗号であれば頻度分析が可能になる．本業務モデルでは，攻撃者となり得るサーバ管理者が暗号文をすべて収集可能であるため，データベースのレコード及びそれらへのアクセス情報も頻度分析の情報となる．

以上のプライバシー保護の要件に加え，照合時の効率性も重要である．本研究では，データが暗号化されたままで，クライアントとサーバが照合できる秘匿共通集合計算プロトコル [3] を用いて効率化を図った．本プロトコルを利用した関連研究として，データベースや検索条件の情報を暗号化したまま類似化合物を検索できる方法が提案されている [6]．この手法は，化合物の情報をフィンガープリントと呼ばれる固定長のビット列として表現し，化合物の類似性をこれらフィンガープリント同士の比較により評価し，値が一致するビットの箇所が多ければ類似度が高いとしている．このフィンガープリントの情報を準同型暗号で暗号化しておき，データベースに保管し照合する．また，サーバ側がフィンガープリントの類似度を判定する計算を暗号文のまま行うので，クライアントとサーバでの処理が分散されて効率的と言える．この化合物データベースの秘匿検索は，本稿と同様データベースの利用や管理サーバへの情報の秘匿化を目的としている．一方で，化合物と異なり SNP は構造式などの特徴で表現できず，また高速で照合計算が可能なデータに変換する必要があるなどの違いがある．

DNA 配列の類似性解析では，二つの DNA 配列の照合では文字列として編集距離を計算する方法が一般的であるが，この際互いが保持する DNA を開示せずに編集距離計算を行う照合手法がいくつか提案されている [1], [2], [5]．しかしこれらはオートマトンを用いた計算を基にしており，編集距離の定義 $O(n^2)$

という計算量の観点からデータベースを用いた照合の適用が困難である．また，データベースにおいて問題となる頻度分析が考慮されていない．

3. 要素技術

3.1 Bloom フィルタ

Bloom フィルタは配列とハッシュ関数を用いる空間効率に優れたデータ構造である．主に，特定の要素が集合に含まれるかどうかを簡易的に調べる目的で利用される．

以下に，一般的な Bloom フィルタの構築方法とその性質について述べる．値の格納と検索には m ビットの論理型配列と k 個の異なるハッシュ関数を使用する．フィルタに要素を追加する際は，追加する要素に対し k 個のハッシュ関数を適用し，配列の対応する各添字の値を真にする．フィルタリング，つまり問合せの際は，調べる値に対し k 個のハッシュを適用し，得られた k 個の添字の要素がすべて真であるかどうかを調べる．但し，実際には異なる要素であってもハッシュ値の衝突が起こる可能性があるため，偽陽性が発生する．n を配列に追加されている要素数としたとき，擬陽性の確率は次式で求められる．

$$Pr = (1 - (1 - 1/m)^{kn})^k \approx (1 - e^{-kn/m})^k \qquad (1)$$

上式が示すように，配列のビット数 m が大きいほど衝突の確率が小さくなる．その一方で k の大きさには最適なもの，つまり偽陽性率を最も小さくする値が存在する．この最適値を K とすると，K は次式で表される．

$$K = m/n \ln 2 \qquad (2)$$

Bloom フィルタを単にフィルタとして使用する場合は最適化された k を用いるのが好ましいが，後述するように，本稿で提案する手法では偽陽性が比較的高くても照合結果に影響がない．

3.2 準同型暗号

準同型暗号とは二つの暗号文を用いた計算が新しい暗号文になり，それを復号すると対応する平文の計算をしたものと同じになるような性質を有する暗号であり，Enc を暗号アルゴリズム，$\otimes$ を演算としたとき，次式を満たす．

$$Enc(A) \otimes Enc(B) = Enc(A \otimes B)$$

RSA 暗号や ElGamal 暗号は積に関する同型性を有し，Paillier 暗号は和に関する同型性を有す．

3.3 秘匿共通集合プロトコル

秘匿共通集合プロトコルは，二者間で各々が持つ集合要素にいくつの共通要素が存在するかを，互いの要素を秘匿したまま計算する手法である [3]．提案方式では，ElGamal 暗号を用いて積に関する同型性を利用した秘匿共通集合計算を行う．

4. データアウトソーシングにおける SNP 診断

本稿は，SNP 診断という比較的少数の SNP を用いて個体の生理的特徴および外的特徴などを診断する分野への応用である．DNA の変形である SNP はそのものの頻度に偏りがある．この偏りを解消するために文字列である SNP 情報から Bloom フィルタを生成する．このとき，複数のハッシュ関数や特定の確率で発生する衝突によって各 SNP の頻度の偏りが撹乱される．

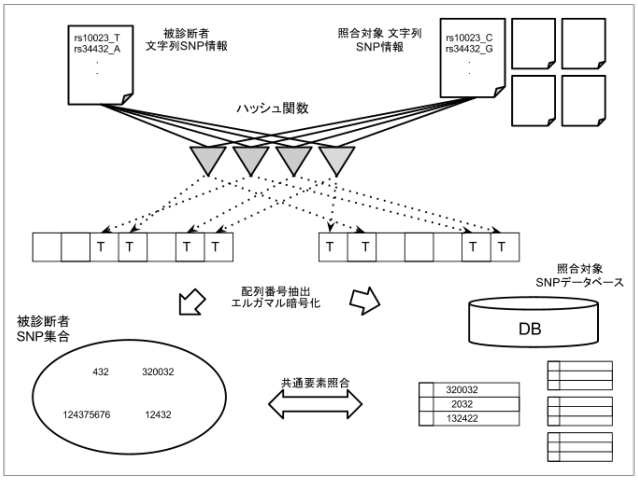

図 2 SNP 診断手順の概要

例えば，総 SNP 数を 1,000 とし，ハッシュ関数を 5 個と決める．Bloom フィルタに対応付けられる SNP のデータは一つにつき 5 個のハッシュ関数が適用され，5 個の配列添字が生成される．つまり，Bloom フィルタに対応付けられる要素数 n は合計で 5,000 となる．次に，希望の偽陽性率を決めるが，例えば 1 % 以下にしたければ，式 (1) より Bloom フィルタのビット数を 10 倍，つまり 50,000 個の配列にすれば偽陽性率は 0.94 % 程度に収まる．また式 (2) より，最適なハッシュ関数の数 k は 7 となる．更に，ElGamal 暗号の持つ準同型性を用いて秘匿共通集合計算による照合を行うことで，問合せを保護する．

図 2 に SNP 診断手順の概要を示す．以下で提案方式を，まずデータベース構築，次に照合時の手順についてクライアント側とサーバ側の処理に分けて述べる．

4.1 格納時

データベースの表に照合対象となる疾患固有の SNP セットのデータを格納する．このデータは SNP セットから Bloom フィルタを生成する際に得られた配列添字を ElGamal 暗号で暗号化したものである．データベース構築時に使用した Bloom フィルタの配列長 m とハッシュ関数は照合の際でも使用する．また，ElGamal 暗号の暗号化で使用した乱数は，後にクライアント側で復号を行う際にも使用する．

4.2 問合せ時

まず，クライアントで以下のように問合せを生成する．本照合方法の前段階として，診断する個体の DNA から診断に用いるすべての SNP を抽出した後，それらが何らかの方法で符号化されているとする．例えば，NCBI の dbSNP の識別番号を用いて {rs2980300_T, rs11260554_G,...} のような文字列に変換されているものとする．

(1) データ格納時と同様の方法で，照合元となる個人 SNP から診断のための特定の SNP セットを抽出して Bloom フィルタを生成し，対応付けられた配列添字を抽出する．

(2) 配列番号を解とする多項式を生成し，各次数の係数を

抽出する．係数を ElGamal 暗号で暗号化し，サーバに送信する．このとき，暗号化された係数を次数順に配列に入れるなどして各係数の次数情報も送信する．具体的には，SNP セットの要素 N 個を解とする N 次の多項式を生成する．

$$\Pi_{i=1}^{N}(x - c_i) = \Sigma_{i=0}^{N}(d_i x^{N-i})$$

この多項式展開した係数 $d_1, \ldots, d_N$ を暗号化し，係数の次数の情報とともにサーバに送信する．

ここで，診断用 SNP を選択するプロセスを省き，個人から抽出した全 SNP で Bloom フィルタを生成することも考えられる．この際，データベースの SNP データで生成する Bloom フィルタも全 SNP 用の Bloom フィルタで定義したものを使用する．例えば，SNP 診断に用いるため，個人から抽出する SNP が 100 万個であるとすると，要素数 n が 100 万となる．この方法がより効率的と考えられる理由は，データベースの全テーブルに対し一回照合すると，すべての診断を同時に行ったことになるからである．この方法に対して，抽出した SNP から診断するセットを選ぶ方法では，セットの選択毎にデータベース全体と照合する必要があり，無駄な処理を含むことになる．解決策として各選択 SNP セット毎に特定のテーブルのみを選ぶとしても，データベース管理者がデータベースのアクセスを監視しているので撹乱処理が必要である．

サーバ側では，まず，データベースから照合対象となる SNP セットのデータが格納されている各表のデータ，つまり暗号化された添字番号を取り出し，クライアントから送られてきた各係数に対応する次数を乗じた添字番号を掛ける．この処理は多項式の各項を独立させたまま代入することに相当する．この結果得られた積の値をクライアントに返す．

クライアント側はサーバから返されてきた各次数の積の値を復号し，足し合わせる．ElGamal 暗号は積に関する同型性を有すので，データベースの照合元 SNP セットの Bloom フィルタに照合元 SNP セットの Bloom フィルタと同じ添字番号が含まれているか，または偽陽性であれば結果は 0（ElGamal 暗号で使用した素数 p を法として 0 と合同） となる．従って，照合元 SNP セットが照合対象 SNP セットと等しい場合は，表の要素すべての代入計算で 0 となる．

ここで，照合元に本来対応している照合対象はデータベースの表群の一部であるため，この計算をすべての表に対して行うことは非効率的である．例えば，照合元が耳あかの型を診断するための SNP セットを選んでいるのに，照合対象が耳あかの SNP セットだけでなく心臓病や 2 型糖尿病など本来関係ない SNP セットとも照合を試みていることになる．しかし，単純な方法で特定の表のみを選択し照合することはデータベース管理者がデータベースのアクセスやデータストリームを監視することで頻度分析可能になるので望ましくない．そこで，本稿では，無駄な表に対する復号を行わないことによって効率化を図る．

照合に用いる要素を確率的データ構造に保管するため，上記のような共通集合を用いた SNP 照合は単なる暗号による照合に比べ強い頻度耐性を持たせることが可能である．これにより SNP が持つ固有の頻度をランダム化でき，結果として頻度分析を困難にする．

5. 実　　験

本実験の主な目的の一つは提案するプロトコルを実装し，データベースのデータ量と照合元の SNP セット数を変えて実行時間を計測することである．また，これを基にして，頻度分析耐性を保持した効率化の方法についても実験する．

本来の照合方法にはクライアント・サーバモデルを想定しているが，本実験では通信環境の影響を避けるため，一つの Java アプリケーションとして実装する．データベースは MySQL を使用し，統合環境開発ソフトの eclipse のプラグインである DBViewer によりデータベースにアクセスする．

本 SNP 診断では以下のパラメータが存在する．

- a : 各診断で使用する SNP の数
- b : 診断内容の数
- c : 照合対象のテーブル数
- d : Bloom フィルタのビット数
- e : ハッシュ関数の数
- f : Bloom フィルタの要素数

次に，実験対象である照合時間に関して各パラメータの影響を考える．このため，照合処理を多項式展開，代入計算，復号処理の 3 つに分け，影響範囲を次のようにまとめる．

- 多項式展開（クライアント側）: $a \times e$, b
- 代入計算（サーバ側）: $b \times c$, $f(a, e)$
- 復号処理（クライアント側）: $b \times c$, $f(a, e)$

パラメータの影響が間接的であったり，その影響が明らかに小さい場合は除外している．多項式計算では $a \times e$ が影響するが，例えば診断で使用する SNP が rs1 と rs2 の二つであれば，これらを一つのハッシュ関数で配列にマッピングすると，衝突がなければ最終的には 2 次の多項式になるためである．つまり多項式展開の計算時間は $a \times e$ なのでどちらか一方のみを変化させれば評価可能である．代入計算と復号計算では $b \times c$ のパラメータがどちらも影響している．これは，一つの診断内容に対してテーブルすべてと照合するときに代入計算および復号処理が行われるためである．これに対し多項式展開は，診断内容一つに対しクライアントで最初に一回行われるのみであるため，b のみが影響する．また，$f(a, e)$ は各テーブル内で代入する要素数に対応しているので，代入回数とその結果を復号する回数に影響する．以上により，照合時間を計測するためのパラメータは，多項式展開に関しては $a \times e$ と b を変化させればよく，代入計算と復号処理に関しては $b \times c$ と f でよい．

5.1 照合時間の実験で使用するパラメータの値と根拠

照合時間の評価については各診断で使用する SNP の数 a，診断内容の数 b，照合対象のテーブル数 c の各パラメータを変化させる．パラメータの値については a に関しては [5,15, 20]，b に関しては [100,200,300,400,500,1000] とするが，このパラメータの値を使用する根拠を以下で述べる．

まず，照合時間の評価はできる限り実際の SNP 診断で行われる処理に近い形で行うべきであると考えた．そこで 1 章で紹介した 23andMe の商業的 SNP 診断をモデルにした．23andMe

ではおよそ 100 万個の SNP が診断に用いられていると記されている．このうち，特定の SNP の組合せが特定の疾病傾向や祖先などを診断するセットとして利用されている．診断に用いる SNP セットの具体的な内容はすべては公表されていないが，SNPedia や疾病関連 SNP の文献などを見ると，セットに含まれる SNP の数は決して多くないと言える．例えば少ないもので 1 個の SNP で診断できる wet/dry の耳あかタイプがあり，多いものでも数十個が殆どである．以上の理由から診断の SNP セットは比較的小さい数で上限を 20 におさえ，診断のバリエーションに応じて増加する照合回数を計測可能な範囲まで大きくした実験を行う方がより現実的であると考えた．そこで診断数，およびテーブル数は，[100,200,300,400,500,1000] と細かく取ることにする．診断数とテーブル数のパラメータはそれぞれ独立に変化するのでなく，一つの診断に本来対応しているテーブルはただ一つと仮定して一致させることにする．

SNP データの作成方法は以下の通りである．データベース作成時，テーブルの照合対象となる SNP セットの文字列情報を 0,1,2,$\cdots$ のような数字にする．診断に使う SNP セットが 5 つならば，1 番目のテーブルには {0,1,2,3,4}，2 番目のテーブルには {5,6,7,8,9} のような数字を本来の SNP 情報である rs100212_A などの代わりに入れておく．照合時，クライアントが個人 SNP から抽出する診断用 SNP セットは，順に {0,1,2,3,4},{5,6,7,8,9},{10,11,12,13,14} のように作り Bloom フィルタを生成する．これにより，各診断 SNP セットはデータベースのテーブル中で必ず一つだけ完全に一致するものがあり，その他のテーブル内のデータと一つでも一致すればそれが偽陽性であると判別できる．

次に診断に用いる SNP の数 5,10,20 がすべての診断内容で一定にする理由について述べる．各診断で使用する SNP の数は本来診断によって変わるはずである．例えば，耳あかなら 1 個で心臓病なら 15 個などである．本実験では，診断で用いる SNP 数を各々で固有なものにすると頻度分析につながるため，ダミーとなる SNP をパディングすることを想定した．

5.2 基本的なパラメータを変化させた実験

各診断で使用する SNP 数 a が [5,10,20] の 3 通り，診断内容の数 b と照合対象のテーブル数 c が [100,200,300,400,500,1000] の 6 通りである．Bloom フィルタで使用する配列のビット数は要素数の 10 倍，ハッシュ関数の数は 1 とした．各実験でプログラムの開始から終了までの経過時間を有効数字一桁のミリ秒単位で 3 回ずつ測定し，平均値を算出した．

実験の結果を表 1 に示す．照合時間はテーブル数が 100，SNP セット数が 5 の場合を 1 として，相対値に変換している．

パラメータの変化と処理時間の関係を調べるにあたり，まずテーブル数および診断数の変化に着目する．テーブル数の変化と処理時間は n^2 よりやや少ないオーダで計算されている．一方，SNP 数の変化に着目すると，同様に，n^2 以下のオーダの計算時間となっている．SNP 数とテーブル数の変化ではどちらが計算時間に影響が大きいかを調べるにあたり，例えば，(100,5) と (400,5)，(100,5) と (100,20) が同じ 4 倍の変化になっているが，照合時間では SNP の変化の方が多くなっている．一方

表 1　基本的なパラメータの変化と照合時間

テーブル数 \ SNP セット	5	10	20
100	1.0	3.3	15. 8
200	3.5	12.0	53.9
300	7.6	26.3	116.2
400	13.2	46.4	202.6
500	20.3	72.2	313.5
1,000	79.3	288.2	1212.8

で (100,5) と (200,5)，(100,5) と (100,10) の 2 倍の変化の場合は逆にテーブルの変化の方が照合時間に大きい影響を与えている．その他の組合せでもどちらか一方に決まらないので，本実験では判断できず，ほぼ同じ程度の影響と考えられる．

本実験ではテーブル数と診断数が同時に変化しているので，例えばテーブル数が 10 倍 に増えると診断数も 10 倍に増え，データベース中のテーブルと照合する回数は 100 倍になる．一方で SNP 数の変化は，本プログラムのボトルネックである多項式展開の処理に影響し，また各テーブル内の要素が増えるため代入計算，復号処理にも影響する．

5.3 復号処理を省く効率化を加えた後のパラメータを変化させた実験

本稿で提案する SNP 照合方法には，ある特定の診断をするため選択した SNP セットに対して関係のない照合対象の SNP セットと照合する処理に無駄がある．しかし，診断する SNP セットに対して特定のデータベースのテーブルを固定してしまうと，データベースへのアクセスやデータストリームの監視によってその関連性を分析される恐れが生じる．頻度分析に対する耐性を下げずに効率化する方法として，計算や処理の手順の省略が考えられる．つまり本来の照合対象と出会うと，以降は擬似的なやりとりをするだけで実際の計算を省くという効率化である．照合処理はクライアント，サーバのどちらでも行うが，サーバ側に擬似的計算を行わせるにはそのタイミングや計算負荷の変化などが頻度分析の材料となるので複雑な工夫が必要になる．そのため本実験ではクライアント側の復号処理を省くことで効率化を図った．具体的には，本来対応している照合テーブル数に関するしきい値を設定した．しきい値を超えたテーブルをカウントし，そのテーブル数が照合テーブル数に達すると以降は復号処理を行わない．本実験では，本来対応している照合テーブル数は 1 である．また，しきい値は照合 SNP セット数と等しくすると，Bloom フィルタ作成時に衝突が起きた場合，得られる配列番号がセット数に満たなくなるのでここではセット数から 1 を引いた値に設定した．

以上の条件で，効率化を加えた際の a, b, c のパラメータを変化させた実験を行った．実験の結果を表 2 に示す．値はテーブル数 100，SNP セット数 5 の場合を基準とした相対値である．

パラメータの変化と処理時間の関係を調べるにあたり，テーブル数および診断数の変化に着目すると，効率化前の実験と同様，n^2 以下のオーダで計算されている．効率化前の相対値と比べると，(100,20) を除いた表の全体で共通して値が小さくなっており，復号計算の省略によって変数の変化により増える計算

表 2　効率化後の a, b, c のパラメータ変化と照合時間の相対値

テーブル数 \ SNP セット	5	10	20
100	1.0	2.9	16.0
200	3.4	10.8	52.9
300	7.3	23.7	111.3
400	12.6	41.1	189.4
500	19.2	63.6	221.9
1,000	73.5	251.9	1111.8

表 3　効率化の程度 [%]

テーブル数 \ SNP セット	5	10	20
100	165	187	163
200	169	183	168
300	172	183	172
400	173	186	176
500	174	187	233
1,000	178	188	180

時間が縮小されたと言える．

実際に全体でどの程度の効率化ができているかを調べるため，性能評価で用いる手法に従い，効率化を図った後の時間で前の時間を割った値に 100 を掛けた値 [%] を効率化の尺度とする．表 3 にその結果を示す．表 3 より，(20,500) は他のパラメータの組合せと異なり，その効率化程度が大きいので計測上の誤差と考える．これを除いた全体の平均を取ると 176.6 [%] となり，結果として，復号処理省略の工夫は効率化に有効であったと言える．

ここで，本実験におけるクライアントの復号処理の効率化の工夫は各診断において照合判定が陽性のテーブルと出会うと，その後の復号処理は行わないというものであった．更に実験で使用した擬似 SNP データについて考えると，各診断で一つだけ照合判定が陽性のテーブルが存在するように作成した．つまり，診断を全体で見ると半分のテーブル数との照合でのみ復号処理が行われており，結果として，各診断でランダムにテーブルと照合していくようなモデルと同様の実験であった．一方，先ほどの性能評価の手法と逆に，照合時間が短縮された比率を求めるため，効率化後の照合時間を前の照合時間で割ったものの外れ値を除いた平均は 0.546 となった．このことを効率化の工夫後の復号処理のテーブル数と合わせて考えると，復号処理が行われるテーブル数が半分になると，照合時間は約 0.546 倍に短縮される．つまり，厳密な評価ではないが，本稿で実験用に書かれたプログラムでは復号処理は処理全体の 45% 程度に相当すると考えられる．

最後に，効率化前と後の照合時間を折れ線グラフで比較したものを図 3 に示す．グラフ中の 6 本の折れ線は，診断で使用する SNP 数を固定した際のテーブル数の変化に対する照合時間の変化を表す．また，数字にプライムが付いた方が効率化後の照合時間である．

6. おわりに

本稿では，外部に委託したデータベースを用い，個人情報で

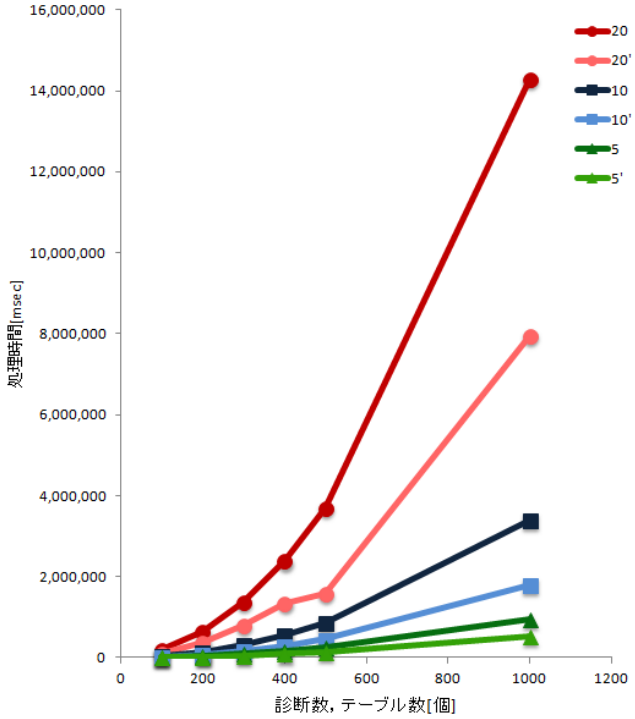

図 3　効率化前後での照合時間の比較

ある SNP を利用し疾病傾向や生体的特徴を診断する際にプライバシー保護機能を備える手法を提案した．また簡易モデルを実装し，データの可変要素を操作することで照合に掛かる時間を測定し，更に効率化のための工夫を行い比較した．

今後は，まず多項式展開の高速アルゴリズムを取り入れ，理論上より効率的な方法と現在の方法の処理速度を比較する．次に，より有望な手順でサーバクライアントモデルで実行できるプログラムを作成し，実際の SNP データで照合を行うことを予定している．

文　献

[1] M. Blanton and M. Aliasgari, Secure outsourcing of DNA searching via finite automata, Proc. of the Data and Applications Security and Privacy XXIV (LNCS 6166), pp. 49–64, 2010.

[2] K. B. Frikken, Practical private DNA string searching and matching through efficient oblivious automata evaluation, Proc. of the Data and Applications Security and Privacy XXIII (LNCS 5645), pp. 81–94, 2009.

[3] M. J. Freedman, K. Nissim, and B. Pinkas, Efficient private matching and set intersection, EUROCRYPT, pp. 1–19, 2004.

[4] H. Hacigümüs, B. Iyer, C. Li, and S. Mehrotra, Executing SQL over encrypted data in the database-service-provider model, Proc. of the 2002 ACM SIGMOD Int'l Conf. on Management of Data, pp. 216–227, 2002.

[5] S. Jha, L. Kruger, and V. Shmatikov, Towards practical privacy for genomic computation, Proc. of the IEEE Symposium on Security and Privacy, pp. 216–230, 2008.

[6] 産業技術総合研究所, 秘密計算による化合物データベースの検索技術, http://www.aist.go.jp/aist_j/press_release/pr2011/pr20111101/pr20111101.html.