Vol.2016-CSEC-74 No.39
Vol.2016-SPT-19 No.39
2016/7/15

プライバシ保護ゲノム解析のための秘密計算フィッシャー正確検定

長谷川 聡[1,a)]　濱田 浩気[1]　千田 浩司[1]　荻島 創一[2,3]　三澤 計治[2,3]　長﨑正朗[2,3]

概要：ゲノム提供者のプライバシ保護や機密情報保護のため，最近では疾患と遺伝子の関連等を調べるゲノムワイド関連解析（GWAS）を秘密計算によって実現する研究や実装コンテストが見られるようになった．我々は関連研究として，GWAS で用いられる重要な検定手法の一つであるフィッシャー正確検定を秘密計算によって実現する手法を提案した．しかし GWAS では一般に膨大な回数の仮説検定を行うことから，フィッシャー正確検定の秘密計算を用いた GWAS はオーバーヘッドが大きい．本稿では，フィッシャー正確検定よりも軽量な演算により有意確率が有意水準を下回る候補を絞り込み，フィッシャー正確検定を秘密計算で実行する回数を削減する手法を提案する．実際のゲノムデータを用いて実験評価を行ったところ，大幅に実行回数を削減できることを確認した．

キーワード：秘密計算，フィッシャー正確検定，ゲノムワイド関連解析，プライバシー保護データマイニング

Privacy Preserving Fisher's Exact Test for GWAS

Satoshi Hasegawa[1,a)]　Koki Hamada[1]　Koji Chida[1]　Soichi Ogishima[2,3]　Kazuharu Misawa[2,3]
Masao Nagasaki[2,3]

Keywords: secure computation, Fisher's exact test, genome-wide association study, privacy preserving data mining

1. はじめに

ICT(Information and Communication Technology) の発達に伴い，多種多様な大量の情報が容易に収集できるようになり，情報の利活用による新たな価値創造への期待が高まっている．そして我々のパーソナルデータ（個人に関するデータ），例えば買物履歴，インターネットアクセス履歴，位置・移動情報，バイタル情報等も事業者によって日々蓄積され，研究やサービス向上のための分析素材として利活用され始めている．しかしパーソナルデータを扱う際はプライバシの保護に十分配慮する必要がある．このような背景から，データ提供者のプライバシを保護しつつ，提供データをデータマイニング等に利活用可能とする**プライバシ保護データマイニング**（Privacy Preserving Data Mining: PPDM）の研究が活発に進められている [1], [2][*1]．

PPDM は Lindell, Pinkas[3] および Agrawal, Srikant[4] によって 2000 年に独立に提案された．Lindell, Pinkas は，二者がそれぞれ持つデータセットを暗号化して互いに明かさぬまま，決定木学習を行う手法を提案した[*2]．Agrawal,

[1] NTT セキュアプラットフォーム研究所
NTT Secure Platform Laboratories, 3-9-11, Midori-cho, Musashino-shi, Tokyo 180–8585, Japan
[2] 東北大学東北メディカル・メガバンク機構
Tohoku Medical Megabank Organization, Tohoku University, 2-1, Seiryo-machi, Aoba-ku, Sendai 980-8573, Japan
[3] 東北大学大学院医学系研究科
Graduate School of Medicine, Tohoku, University, 2-1, Seiryo-machi, Aoba-ku, Sendai 980-8575, Japan
[a)] hasegawa.satoshi@lab.ntt.co.jp

[*1] PPDM はデータマイニングに限らず統計解析等を含むデータ分析全般を対象とすることが多い．
[*2] このような状況は，各組織が個人のデータセットを保持してお

Srikant も決定木学習を取り上げているが，データセットをランダム化して提供することでデータ提供者のプライバシを保護している．ランダム化されたデータから，ベイズ推定等を用いて誤差の少ない決定木学習を行う手法を提案した．その後の研究で，様々なデータマイニング・統計解析を対象とした PPDM が提案されている [1], [2].

PPDM の研究に先駆けて，本来の入力データを演算実行者に明かさず計算させることができる**秘密計算**[*3](Secure Computation) が古くから研究されている [5]．秘密計算は任意の関数について実行可能な手法の研究が進められてきたが，必ずしも処理効率が良くないため，関数を限定した特化型の手法も研究が行われるようになった．Lindell, Pinkas の提案手法もその一例といえる．

本研究では，疾患と遺伝子の関連等を調べるゲノムワイド関連解析 (Genome-Wide Association Study: GWAS) [6] で用いられる重要な検定手法の一つである，**フィッシャー正確検定** [7] に特化した秘密計算 (以降これを秘密計算フィッシャー正確検定と呼ぶ) に着目する．最近では GWAS を秘密計算によって実現する研究 [8], [9] や実装コンテスト [10], [11] が見られるようになったが，フィッシャー正確検定は階乗計算を繰り返し行う必要があり，従来の汎用的な秘密計算では実用が難しい．そのため我々は関連研究として秘密計算フィッシャー正確検定を提案している [12], [13]．しかし GWAS では一般に膨大な回数の仮説検定を行うことから，秘密計算フィッシャー正確検定を用いた GWAS はオーバーヘッドが大きい．

本稿では，フィッシャー正確検定よりも軽量な演算により有意確率が有意水準を下回る候補を絞り込み，秘密計算フィッシャー正確検定の実行回数を削減する手法を提案する．具体的には，フィッシャー正確検定の中間出力が有意確率以下となることを利用し，有意水準を下回る候補を絞り込む．実際のゲノムデータを用いて実験評価を行ったところ，大幅に秘密計算フィッシャー正確検定の実行回数を削減できることを確認した．

以降，2 節で準備としてフィッシャー正確検定，GWAS，秘密計算，および関連研究について簡単に触れる．3 節では GWAS における秘密計算フィッシャー正確検定の利用形態の考察を行う．そして 4 節で，GWAS における秘密計算フィッシャー正確検定の実行回数を削減する手法を提案する．5 節では提案手法に関する実験評価，安全性評価，および効率評価を行い，6 節で本稿をまとめる．

表 1 2×2 分割表

	Yes	No	計
カテゴリー A	a	b	X
カテゴリー B	c	d	$N-X$
計	Y	$N-Y$	N

2. 準備

2.1 フィッシャー正確検定

フィッシャー正確検定は，2 つ以上のカテゴリーの独立性の検定を行う手法である．表 1 のような 2×2 の分割表 (度数表) を考えよう．ここで $X = a+b$, $Y = a+c$, $N = a+b+c+d$. すると表 1 の分割表が得られる確率 $P(a)$ は以下の関数 $P(z)$ から得られる：

$$P(z) = \frac{X!Y!(N-X)!(N-Y)!}{N!z!(X-z)!(Y-z)!(N-X-Y+z)!} \quad (1)$$

フィッシャー正確検定 (両側検定) の有意確率は，$P(a)$ よりも極端な分割表の確率も考慮し，

$$P = P(a) + \sum_{P(i)<P(a)} P(i) \quad (2)$$

で与えられる．ただし

$$\max(0, X+Y-N) \le i \le \min(X, Y).$$

最終的に P と有意水準 α との大小関係により統計的な差の有無を得る．具体的には，$P < \alpha$ であれば帰無仮説が棄却され，「カテゴリー A とカテゴリー B には統計的な差が無いとは言えない」と帰結される．α の値は通常 0.05 や 0.01 等が用いられるが，ゲノム解析のように多重に検定を行う場合は，Bonferroni 補正法 [14] 等によって α の値が非常に小さくなることもある．

式 (1) の計算を効率化する工夫として，対数をとる方法がよく知られる．非負整数 n について $\log(n!) = \sum_{i=1}^{n} \log i$ が成り立つことから，

$$\ell_n := \sum_{i=1}^{n} \log i \quad (3)$$

として式 (1) の対数

$$\log P(n) = \ell_X + \ell_Y + \ell_{N-X} + \ell_{N-Y} - (\ell_N + \ell_n + \ell_{X-n} + \ell_{Y-n} + \ell_{N-X-Y+n}) \quad (4)$$

を計算する．特に標本数 N を固定して $\ell_1, \ell_2, \ldots, \ell_N$ を事前計算しておけば，式 (4) は事前計算値の加減算のみで簡便に計算できる．そして $\log P(n)$ から $P(n)$ を求めることで式 (2) を計算できる．

2.2 GWAS

GWAS では，ある集団のゲノム塩基配列中に見られる一塩基多型 (Single Nucleotide Polymorphism: SNP) や一

り，それらを結合してデータ分析する例が挙げられる．また個人のプライバシだけでなく，組織のプライバシ (機密情報) の保護にも有効といえる．

[*3] セキュア計算や秘匿計算とも呼ばれる．

塩基バリアント (Single Nucleotide Variants: SNVs) に基づき，疾患等との関連を統計的に調べることが行われる．SNP や SNVs は多数発見されており，例えば日本人のゲノムの解析によって 2,120 万箇所におよぶ SNVs が発見されたという報告がある [15]．すなわち数百万〜数千万種類の分割表を作成して仮説検定を行うことも珍しくない．仮説検定は，χ^2 検定による独立性の検定や，フィッシャー正確検定等が用いられる．χ^2 検定はフィッシャー正確検定の近似であり簡便に計算可能だが，分割表に小さい度数が存在する場合等に誤差の影響が大きくなることが指摘されている [16].

2.3 秘密計算

秘密計算は一般に，意中の関数の入力データを秘匿化して提供する主体と，その入力データを復元せず当該関数の演算を実行する主体の少なくとも二者が存在する．すなわちデータ提供者と演算実行者が異なっており，提供データを他者に知られたくないが，関数演算は他者の手を借りたいという動機が前提となる．秘匿化の手法として暗号化や秘密分散 [17] 等が知られる．Yao は，組合せ論理回路を実行可能な状態のまま秘匿化する主体と，秘匿化された組合せ論理回路 (Garbled Circuit) を実行する主体からなる秘密計算を提案した [18]．組合せ論理回路によりある程度汎用的な演算が可能となる．なお秘密計算において複数の主体が協調して計算する場合は**セキュアマルチパーティ計算** (Secure Multi-Party Computation: MPC) とも呼ばれる．複数の主体がそれぞれ知られたくないデータを持ち，それらを結合して計算する場合に有効といえる．

加減算や乗算の秘密計算についても多くの研究結果が知られている．Ben-Or ら [19] と Chaum ら [20] は秘密分散を用いて加減算と乗算ができる MPC を提案した．Cramer ら [21] は加法準同型暗号を用いて同様の MPC を実現した．加法準同型暗号は，暗号化したまま加算 (および減算) ができる暗号方式である．言い換えれば，$E(\cdot)$ を加法準同型暗号の暗号化関数としたとき，数値 a, b の暗号文 $E(a), E(b)$ から，それらを復号することなく $a+b$ の暗号文 $E(a+b)$ を求めることができる．加減算と乗算の組合せで回路素子の演算を構成し，前記の MPC を組合せ論理回路による汎用性の高い秘密計算とすることもできる．

更には等号判定，大小比較，ランダムシャッフル (複数のデータをランダムな順序で並び替える)，ソート等の秘密計算の研究も見られる．Schoenmakers, Tuyls[22] は，加法準同型暗号を用いて conditional gate と呼ばれる演算を MPC で構成するとともに，等号判定や大小比較への応用を示した．Damgård ら [23] と Nishide, Ohta[24] は秘密分散を用いて等号判定や大小比較の MPC を構成した．濱田らはいくつかの既存の効率的なソートアルゴリズムをランダムシャッフルを用いた MPC で実現できる手法を提案している [25], [26]．これらは何れも秘密計算の部品として利用することができる．すなわち，加減算，乗算，等号判定，大小比較，ランダムシャッフル，ソートの結果を秘匿化したまま次の計算の入力にできる．

MPC とは異なるアプローチとして，Gentry は完全準同型暗号を構成した [27]．完全準同型暗号は，加法準同型暗号の性質と乗法準同型暗号 (暗号化したまま乗算ができる暗号方式) の性質を合わせ持った暗号方式である．単一主体で計算可能なことから通信は不要だが，データサイズが大きい，計算が複雑などの欠点もあり，様々な改良研究が進められている [28].

2.4 秘密計算フィッシャー正確検定の関連研究

我々はこれまでに二種類の秘密計算フィッシャー正確検定を提案した．一つは，フィッシャー正確検定の結果を判別する決定木を事前に作成しておき，本計算では決定木の実行を秘密計算によって実現する手法である (関連研究 1)[12]．最大 1,500 の標本数について実際に決定木を作成し，十分実用的な大きさで構成できることを示している．ただし決定木の事前作成のため，標本数や有意水準は事前に与えられている必要がある．

もう一つの手法 (関連研究 2) は，前記の前提条件を不要とした．2.1 節の式 (3) を用いて，式 (4) の右辺の各項を等号判定の秘密計算によって秘匿化したまま求めている．また各項をシフト演算の秘密計算によって個別に求めるよりも効率良く計算可能とした．前記の決定木が作成されていれば関連研究 1 の方が効率的と考えられるが，決定木は標本数が大きくなるにつれ作成のオーバーヘッドが著しく増加することから，標本数が大きく決定木の作成が困難な場合は，関連研究 2 の手法を用いることが推奨される．

3. GWAS における秘密計算フィッシャー正確検定の利用形態

PPDM はデータ提供者，演算実行者，演算結果取得者，そして (データ提供者が演算実行者に対して) 秘匿すべき情報によって様々な利用形態が考えられる．また，関数の種類や秘匿方法によってアプローチが異なる場合もある．例えば Agrawal, Srikant[4] は，データ提供者は複数人のパーソナルデータセットを所持している単一組織，演算実行者および演算結果取得者は分析者を想定し (図 1)，秘匿すべき情報はパーソナルデータセットから特定または推定される機微な情報であろう．Lindell, Pinkas[3] は，データ提供者は複数の組織，演算実行者および演算結果取得者も当該複数の組織を想定し (図 2)，秘匿すべき情報は各データ提供者が所持しているデータそのものであり，パーソナルデータの保護や各組織の機密情報保護の両方の目的が考えられる．その他，データ提供者および演算結果取得者は複数人のパーソナルデータセットを所持している単一組織，

Vol.2016-CSEC-74 No.39
Vol.2016-SPT-19 No.39
2016/7/15

そして演算実行者は外部のクラウドという形態も挙げられる (図 3)．更にはパーソナルデータが逐次クラウドに集約される状況を想定し，データ提供者は複数の個人，演算実行者はクラウド，演算結果取得者は分析者 (図 4)，そしてクラウドや分析者に対して個々人のパーソナルデータを秘匿するという利用形態も考えられる．

それでは GWAS における秘密計算フィッシャー正確検定の利用形態についてはどうだろうか．先ず図 4 の形態では，秘匿化されたゲノムデータがクラウドに集約され，分析者からクラウドに分析要求があると，ゲノムデータを秘匿化したまま仮説検定を行う方法が考えられる．ただし個人が自身のゲノムデータを所有していることが前提であり，現状は想定が難しい．そこで図 2 の形態において，各組織が複数人のゲノムデータを所持し，それらを足し合わせて検定を行う方法を考える．これは昨今研究目的でゲノムデータを所持する組織が増えてきたこと，そして組織間でデータを共有や結合できれば研究の幅が広がることから現実的なモデルと考えられる．なおゲノムデータから得られる分割表が個人のプライバシの観点で問題なければ，秘密計算を用いることなく，各組織が分割表を求めて他組織に提供し，それらを足し合わせて検定を行えばよい．また，足し合わせた分割表までを秘密計算によって求め，その足し合わせた分割表を復元して仮説検定を行うことで，各組織の分割表もある程度秘匿されるかもしれない．しかし足し合わせた分割表を各組織が共有すると，自組織の分割表を差し引いて他組織の分割表を容易に特定または推定できるという問題が生じ得る．これは組織の機密情報保護の観点から望ましくない．したがって本研究では，フィッシャー正確検定の結果以外を一切漏らすことなく計算できる，GWAS に適した秘密計算フィッシャー正確検定の実現を目指している．

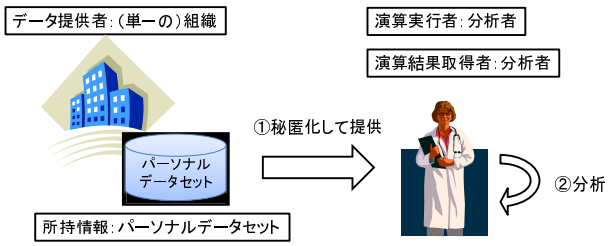

図 1 利用形態 1

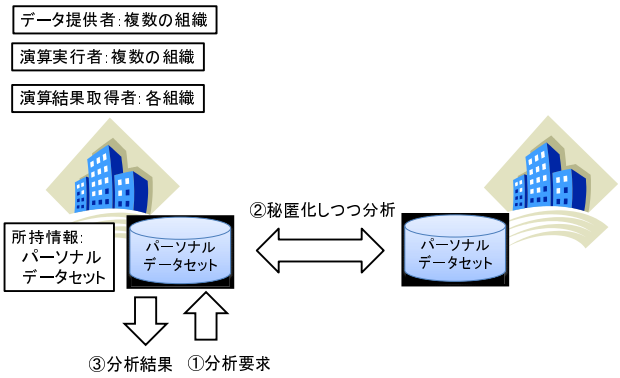

図 2 利用形態 2

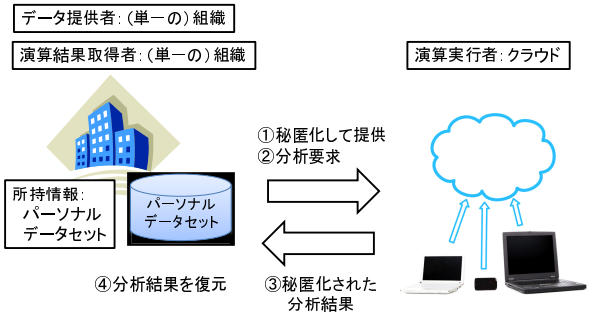

図 3 利用形態 3

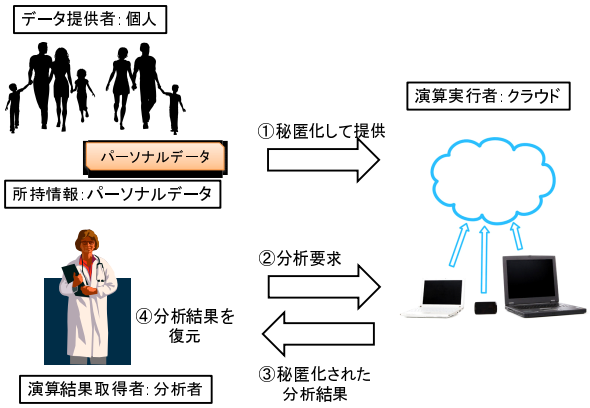

図 4 利用形態 4

4. 提案手法

4.1 基本アイデア

本研究では，前節で紹介した関連研究 1[12] や関連研究 2[13] のような単一のフィッシャー正確検定に対するアプローチではなく，GWAS のように膨大な回数のフィッシャー正確検定を一括で実行する場合の効率化について検討した．提案手法は，所定の回数 (M とする) の秘密計算フィッシャー正確検定を行うよりも軽量な演算により，有意確率が有意水準を下回る候補を絞り込む．そして絞り込まれた入力データについて，関連研究 1 または 2 等の秘密計算フィッシャー正確検定を実行する．したがってより多くの入力データを絞り込むほど効率良く計算できる．

入力データの絞り込みは，2.1 節で説明したフィッシャー正確検定の計算過程で得られる $P(a)$ を用いる．有意確率が有意水準を下回る条件 $P < \alpha$ を考えると，式 (2) より明らかに $P \geq P(a)$ となるため，$P(a) \geq \alpha$ であれば $P < \alpha$ となることはない．したがって $P(a) < \alpha$ となる入力データのみ，有意確率が有意水準を下回る候補と見なすことができる．

また M 個の $P(a)$ は，**秘密計算一括写像** [29], [30] を用

いて効率的に計算できる．秘密計算一括写像は，図 5 に示すように，ある整数 K, L について K 組の (秘匿化された) データとインデックスがあるとき，L 個の秘匿化されたインデックスを入力すると，それらのインデックスに紐づいた L 個の秘匿化されたデータを出力する．L 個の秘匿化されたインデックスを個別に入力して実行すると全体で $O(KL)$ の計算量となるのに対し，秘密計算一括写像は $O((K+L)\log(K+L))$ の計算量で実現できるため，K や L が大きい場合は特に効果的となる．

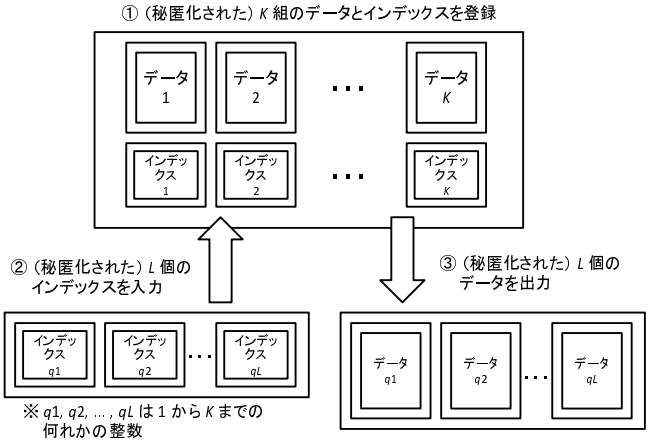

図 5 秘密計算一括写像の処理イメージ

4.2 提案プロトコル

4.1 節で提案した基本アイデアに基づくプロトコルを以下に示す．入力は M 組の分割表の度数 a_i, b_i, c_i, d_i $(i = 1, 2, \ldots, M)$ を秘匿化した値とする (秘匿化関数を $E(\cdot)$，その逆関数を $D(\cdot)$ と表記する)．すなわち 3 節で考察した利用形態において，(足し合わせた) 分割表が秘匿化された値までは秘密計算によって得られているものとする．これは例えば単純に加算の秘密計算を用いればよい．フィッシャー正確検定の関数を Fisher$(\cdot,\cdot,\cdot,\cdot,\cdot)$ と表記し，出力は

$$\mathrm{Fisher}(\alpha, a_i, b_i, c_i, d_i) = \begin{cases} 1 & \text{if } P < \alpha \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

とする．

Input: $\{E(a_i), E(b_i), E(c_i), E(d_i)\}_{i=1}^{M}$, $\{(\ell_i, i)\}_{i=0}^{\max\{N_i\}}$, α (ℓ_i は式 (3) 参照).

Output: Fisher$(\alpha, a_i, b_i, c_i, d_i)$= 1 となる i.

(1) 全ての i について，$n = a_i$ として式 (4) の右辺の各項のインデックスを秘匿化した値 $E(X_i)$, $E(Y_i)$, $E(N_i - X_i)$, $E(N_i - Y_i)$, $E(N_i)$, $E(a_i)$, $E(X_i - a_i)$, $E(Y_i - a_i)$, $E(N_i - X_i - Y_i + a_i)$ を計算する．

(2) 秘密計算一括写像を用いて，$9M$ 個の秘匿化されたインデックス $E(X_i)$, $E(Y_i)$, $E(N_i - X_i)$, $E(N_i - Y_i)$, $E(N_i)$, $E(a_i)$, $E(X_i - a_i)$, $E(Y_i - a_i)$, $E(N_i - X_i - Y_i + a_i)$ を入力し，$9M$ 個の秘匿化されたデータ $E(\ell_{X_i})$, $E(\ell_{Y_i})$, $E(\ell_{N_i - X_i})$, $E(\ell_{N_i - Y_i})$, $E(\ell_{N_i})$, $E(\ell_{a_i})$, $E(\ell_{X_i - a_i})$, $E(\ell_{Y_i - a_i})$, $E(\ell_{N_i - X_i - Y_i + a_i})$ を得る[*4].

(3) 全ての i について以下を行う．

(a) 加減算の秘密計算を用いて $E(\ell_{X_i})$, $E(\ell_{Y_i})$, $E(\ell_{N_i - X_i})$, $E(\ell_{N_i - Y_i})$, $E(\ell_{N_i})$, $E(\ell_{a_i})$, $E(\ell_{X_i - a_i})$, $E(\ell_{Y_i - a_i})$, $E(\ell_{N_i - X_i - Y_i + a_i})$ から $E(\log P(a_i))$ を求める[*5].

(b) 大小比較の秘密計算を用いて α と $E(\log P(a_i))$ から $\log P(a_i) < \log \alpha$ の真偽を判定する．

(c) 真であれば $E(a_i), E(b_i), E(c_i), E(d_i)$ について秘密計算フィッシャー正確検定を実行し，有意確率が有意水準を下回れば i を出力する．

4.3 秘匿性の強化

4.2 節の提案プロトコルは, 手続き (3)(b) にて $\log P(a_i) < \log \alpha$ の真偽結果を与えており，秘匿性が十分ではない．そこで，$\log P(a_i) < \log \alpha$ が真となる入力データの最大個数 $\mathcal{M}$ を設定し，常に $\mathcal{M}$ 組の入力がランダムな順序で絞り込まれるようにする．そしてその $\mathcal{M}$ 組の入力について秘密計算フィッシャー正確検定を実行し，有意確率が有意水準を下回れば i を出力する．$\mathcal{M}$ が小さいほど，秘密計算フィッシャー正確検定の実行回数が少なく済む．

以下に $\log P(a_i) < \log \alpha$ の真偽結果も秘匿するプロトコルを提案する．手続き (3)(a) までは 4.2 節のプロトコルと同様である．

Input: $\{E(a_i), E(b_i), E(c_i), E(d_i)\}_{i=1}^{M}$, $\{(\ell_i, i)\}_{i=0}^{\max\{N_i\}}$, α, $\mathcal{M}$.

Output: Fisher$(\alpha, a_i, b_i, c_i, d_i)$= 1 となる i.

(1) 全ての i について，$n = a_i$ として式 (4) の右辺の各項のインデックスを秘匿化した値 $E(X_i)$, $E(Y_i)$, $E(N_i - X_i)$, $E(N_i - Y_i)$, $E(N_i)$, $E(a_i)$, $E(X_i - a_i)$, $E(Y_i - a_i)$, $E(N_i - X_i - Y_i + a_i)$ を計算する．

(2) 秘密計算一括写像を用いて，$9M$ 個の秘匿化されたインデックス $E(X_i)$, $E(Y_i)$, $E(N_i - X_i)$, $E(N_i - Y_i)$, $E(N_i)$, $E(a_i)$, $E(X_i - a_i)$, $E(Y_i - a_i)$, $E(N_i - X_i - Y_i + a_i)$ を入力し，$9M$ 個の秘匿化されたデータ $E(\ell_{X_i})$, $E(\ell_{Y_i})$, $E(\ell_{N_i - X_i})$, $E(\ell_{N_i - Y_i})$, $E(\ell_{N_i})$, $E(\ell_{a_i})$, $E(\ell_{X_i - a_i})$, $E(\ell_{Y_i - a_i})$, $E(\ell_{N_i - X_i - Y_i + a_i})$ を得る．

[*4] $K(= \max\{N_i\} + 1)$ 組のデータとインデックス (ℓ_i, i) を登録し，秘密計算一括写像を用いて，$L(= 9M)$ 個の秘匿化されたインデックスを入力して，対応する L 個の秘匿化されたデータを出力している．

[*5] $\log P(a_i)$ は式 (4) 参照.

(3) 全ての i について以下を行う.
  (a) 加減算の秘密計算を用いて $E(\ell_{X_i})$, $E(\ell_{Y_i})$, $E(\ell_{N_i-X_i})$, $E(\ell_{N_i-Y_i})$, $E(\ell_{N_i})$, $E(\ell_{a_i})$, $E(\ell_{X_i-a_i})$, $E(\ell_{Y_i-a_i})$, $E(\ell_{N_i-X_i-Y_i+a_i})$ から $E(\log P(a_i))$ を求める.
  (b) 大小比較の秘密計算を用いて α と $E(\log P(a_i))$ から $\log P(a_i) < \log \alpha$ の真偽 B_i (真ならば $B_i = 1$, 偽ならば $B_i = 0$) を秘匿化した値 $E(B_i)$ を求める.
(4) 全ての i について 6 つ組 $(E(a_i), E(b_i), E(c_i), E(d_i), E(i), E(B_i))$ を B_i の昇順になるようソートの秘密計算を実行する.
(5) 上位から $\mathcal{M}$ 個の 6 つ組 $(E(a_i), E(b_i), E(c_i), E(d_i), E(i), E(B_i))$ を選び, ランダムシャッフルの秘密計算を実行する.
(6) ランダムシャッフルを施した各々の $E(a_i)$, $E(b_i)$, $E(c_i)$, $E(d_i)$ について秘密計算フィッシャー正確検定を実行し, 有意確率が有意水準を下回れば $i = D(E(i))$ を計算して出力する.

上記プロトコルの手続き (4) の昇順ソートにより, $B_i = 1$(有意確率が有意水準を下回る可能性のある入力) となる 6 つ組 $(E(a_i), E(b_i), E(c_i), E(d_i), E(i), E(B_i))$ が上位に並ぶ. そして手続き (6) では上位から $\mathcal{M}$ 個の 6 つ組を選び, 秘密計算フィッシャー正確検定を実行しているが, $\mathcal{M}$ は $\log P(a_i) < \log \alpha$ が真となる, すなわち $B_i = 1$ となる 6 つ組の最大個数と仮定しているため, $B_i = 1$ となる 6 つ組は全て上位 $\mathcal{M}$ 個に含まれる. したがって, フィッシャー正確検定の有意確率が有意水準を下回る 6 つ組は必ず上位 $\mathcal{M}$ 個に含まれることが保証される.

5. 提案手法の評価

5.1 実験評価

提案手法の絞り込みの効果を NBDC ヒトデータベース [31] に登録・公開されているゲノムデータの分割表 (非制限公開データ) を利用して確認した. 具体的には表 2 に示すデータを用いた. 結果を表 3,4 に示す.

表 3 絞り込みの結果 (有意水準: $\alpha = 0.05/N$(Bonferroni 補正))

	$P < \alpha$ となる SNP 数	$P(a) < \alpha$ となる SNP 数
データ 1	8	13
データ 2	104	148
データ 3	36	51
データ 4	106	120

表 3,4 より, 条件式 $P(a) < \alpha$ を用いてかなり絞り込めることが分かった. 表 2 の例を基準にすれば, $\mathcal{M} = 200$ 程度で十分と考えられる. また $P < \alpha$ が真となる SNP 数とあまり変わらず, 絞り込みの条件式 $(P(a) < \alpha)$ が, 実際の条件式 $(P < \alpha)$ をよく近似していることが見てとれる.

表 4 絞り込みの結果 (有意水準: $\alpha = 5.0 \times 10^{-8}$)

	$P < \alpha$ となる SNP 数	$P(a) < \alpha$ となる SNP 数
データ 1	7	13
データ 2	101	133
データ 3	33	43
データ 4	91	104

5.2 安全性評価

4.3 節で与えた, 提案プロトコルの安全性 (秘匿性) について考察する. 秘匿すべき情報は分割表の度数 a_i, b_i, c_i, d_i であり, これらに関して提案プロトコルが最終出力 $\mathrm{Fisher}(\alpha, a_i, b_i, c_i, d_i) \in \{0, 1\}$ 以外の有意な情報を一切漏らさないことが目的となる.

手続き (1) の処理は, 加減算の秘密計算によって実現できる. したがって加減算の秘密計算が安全であれば, 秘匿化された値以外の一切の追加情報を与えない.

手続き (2) の処理は, 秘密計算一括写像を用いており, 出力は秘匿化されている. したがって秘密計算一括写像が安全であれば, 秘匿化された値以外の一切の追加情報を与えない.

手続き (3) の処理は, それぞれ加減算および大小比較の秘密計算を用いており, 出力は秘匿化されている. したがって加減算および大小比較の秘密計算が安全であれば, 秘匿化された値以外の一切の追加情報を与えない.

手続き (4) の処理は, ソートの秘密計算を用いており, 出力は秘匿化されている. したがってソートの秘密計算が安全であれば, 秘匿化された値以外の一切の追加情報を与えない.

手続き (5) の処理は, ランダムシャッフルの秘密計算を用いており, 出力は秘匿化されている. したがってランダムシャッフルの秘密計算が安全であれば, 秘匿化された値以外の一切の追加情報を与えない.

手続き (6) の処理は, 秘密計算フィッシャー正確検定の実行であり, 秘密計算フィッシャー正確検定が安全であれば, $\mathcal{M}$ 個の検定結果以外, 秘匿化された値以外の一切の追加情報を与えない. また $\mathcal{M}$ 個の検定結果はランダムシャッフルにより i の順序と独立に出力されるため, 検定結果の順序が有意な情報を与えることはない.

以上から, 加減算, 大小比較, ランダムシャッフル, ソートの秘密計算, および秘密計算一括写像, 秘密計算フィッシャー正確検定を用いて, a_i, b_i, c_i, d_i に関して最終出力 $\mathrm{Fisher}(\alpha, a_i, b_i, c_i, d_i)$ 以外の有意な情報を一切漏らさず計算できる.

表 2 利用したデータ

データ番号	疾患名	症例 (人)	対象群 (人)	SNP 数	アクセション番号
データ 1	心筋梗塞	1,666	3,198	455,781	hum0014.v1.freq.v1
データ 2	2 型糖尿病	9,817	6,763	552,915	hum0014.v3.T2DM-1.v1
データ 3	2 型糖尿病	5,646	19,420	479,088	hum0014.v3.T2DM-2.v1
データ 4	スティーブンス・ジョンソン症候群	117	691	449,205	hum0029.v1.freq.v1

5.3 効率評価

4.3 節で与えた提案プロトコルは，秘密計算フィッシャー正確検定の実行回数を数百万程度 M から数百程度 $\mathcal{M}$ に削減している．ただし 4.3 節で与えた提案プロトコルの手続き (1) から (5) が追加されるため，そのオーバーヘッド (計算量) を評価する．

関連研究 2 の秘密計算フィッシャー正確検定 [13] は，$O(N_i)$ 個の式 (3) を秘匿化して等号判定に用いるため，これを M 回繰り返せば $o(M\max\{N_i\})$ の計算量となる．一方，関連研究 1 の決定木を用いた秘密計算フィッシャー正確検定 [13] は，決定木の大きさ (ノード数) に依存するが，決定木の作成方法を限定していないため計算量を評価できない．しかし実際に決定木を作成した結果を見る限り，ノード数は N_i を超えており，これを M 回繰り返せば $o(M\max\{N_i\})$ の計算量と見なせるものと考えられる．

手続き (1) では $O(M)$ 回の加減算の秘密計算を実行する．手続き (2) は，既存の秘密計算一括写像 [29], [30] を用いて $O((\max\{N_i\}+M)\log(\max\{N_i\}+M))$ の計算量で実現できる．手続き (3) では $O(M)$ 回の加減算および大小比較の秘密計算を実行する．手続き (4) は，既存のソートの秘密計算 [25] を用いて $O(M)$ の計算量で実現できる．手続き (5) は，既存のランダムシャッフルの秘密計算 [25], [32] を用いて $O(M)$ の計算量で実現できる．したがって，何れも既存の秘密計算フィッシャー正確検定による $o(M\max\{N_i\})$ の計算量と比べればオーバーヘッドは十分小さいといえる．

6. まとめ

疾患と遺伝子の関連等を調べるゲノムワイド関連解析 (GWAS) におけるプライバシ保護・機密情報保護として，フィッシャー正確検定を用いた GWAS を秘密計算によって実現する手法を提案した．提案手法は，GWAS では一般に膨大な回数の仮説検定を行うことに着目し，フィッシャー正確検定よりも軽量な演算により有意確率が有意水準を下回る候補を絞り込む．実際のゲノムデータを用いて実験評価を行ったところ，大幅にフィッシャー正確検定を実現する秘密計算の実行回数を削減できることを確認した．我々は先行研究としてフィッシャー正確検定を実現する秘密計算を提案しており，それと組み合わせることで効果的にフィッシャー正確検定を用いた GWAS の秘密計算を実現できるようになる．

ゲノム解析分野において，組織の機密情報や機微な情報の保護は今後益々重要になるものと考えられる．そのため，より効率的な手法の開発や利用の幅を広げるプライバシ保護データマイニングの研究の発展が強く望まれる．

謝辞 本研究に使用したデータは，オーダーメイド医療実現化プロジェクト (代表者 理化学研究所ゲノム医科学研究センター中村祐輔センター長)，オーダーメイド医療の実現プログラム (代表者 理化学研究所統合生命医科学研究センター久保充明副センター長)，ならびに感覚器未来医療学 (代表者 京都府立医科大学医学研究科 上田真由美准教授) によって取得され，科学技術振興機構（JST）の「バイオサイエンスデータベースセンター（NBDC)」ウェブサイト（http://humandbs.biosciencedbc.jp/）を通じて提供されたものです．

計算結果の一部は，東北大学東北メディカル・メガバンク機構のスーパーコンピュータを利用して得られました．

参考文献

[1] Vaidya, J., Clifton, C.W. and Zhu, Y.M.: Privacy preserving data mining, Springer-Verlag (2005)
[2] Aggarwal, C. and Yu, P.: Privacy-preserving data mining: Models and algorithms, Springer-Verlag (2008)
[3] Lindell, Y. and Pinkas, B.: Privacy preserving data mining, CRYPTO 2000, LNCS 1880, Springer-Verlag, 36-54 (2000)
[4] Agrawal, R. and Srikant, R.: Privacy-preserving data Mining, Proc. of the ACM SIGMOD Conference on Managament of Data, 439-450 (2000)
[5] Yao, A. C.: Protocols for secure computations, Proc. of FOCS '82, 160-164 (1982)
[6] Ozaki, K., Ohnishi, Y., Iida, A., Sekine, A., Yamada, R., Tsunoda, T., Sato, H., Hori, M., Nakamura, Y. and Tanaka, T.: Functional SNPs in the lymphotoxin-a gene that are associated with susceptibility to myocardial infarction, Nature Genetics 32, 650-654 (2002)
[7] Fisher, R.A.: On the interpretation of χ^2 from contingency tables, and the calculation of P, Journal of the Royal Statistical Society 85(1), 87-94 (1922)
[8] Kamm, L., Bogdanov, D., Laur, S., Vilo, J.: A new way to protect privacy in large-scale genome-wide association studies. Bioinformatics 29(7), 886-893 (2013), http://dx.doi.org/10.1093/bioinformatics/btt066
[9] Lu, W.J., Yamada, Y. and Sakuma, J.: Privacy-preserving genome-wide association studies on cloud environment using fully homomorphic encryption, BMC medical informatics and decision making 15(Suppl 5), S1 (2015)

[10] IDASH PRIVACY & SECURITY WORKSHOP 2015 - SECURE GENOME ANALYSIS COMPETITION, http://www.humangenomeprivacy.org/2015/
[11] IDASH PRIVACY & SECURITY WORKSHOP 2016 - SECURE GENOME ANALYSIS COMPETITION, http://www.humangenomeprivacy.org/2016/about.html
[12] 千田浩司，長谷川聡，濱田浩気，荻島創一，三澤計治，長崎正朗: 秘密計算フィッシャー正確検定 (1) ～標本数が少ない場合, 第 74 回コンピュータセキュリティ研究会 (CSEC) 予稿集 (2016)
[13] 濱田浩気，長谷川聡，千田浩司，荻島創一，三澤計治，長崎正朗: 秘密計算フィッシャー正確検定 (2) ～標本数が多い場合, 第 74 回コンピュータセキュリティ研究会 (CSEC) 予稿集 (2016)
[14] 瀬々潤，濱田道昭: 生命情報処理における機械学習 多重検定と推定量設計, 講談社 (2014)
[15] Nagasaki, M., Yasuda, J., Katsuoka, F., Nariai, N., Kojima, K., Kawai, Y., Yamaguchi-Kabata, Y., Yokozawa, J., Danjoh, I., Saito, S., Sato, Y., Mimori, T., Tsuda, K., Saito, R., Pan, X., Nishikawa, S., Ito, S., Kuroki, Y., Tanabe, O., Fuse, N., Kuriyama, S., Kiyomoto, H., Hozawa, A., Minegishi, N., Douglas Engel, J., Kinoshita, K., Kure, S., Yaegashi, N.: Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals, Nat Commun. 6(8018), (2015)
[16] Yates, F.: Contingency tables involving small numbers and the χ^2 test, Supplement to the Journal of the Royal Statistical Society 1(2), 217-235 (1934)
[17] Shamir, A.: How to share a secret, Communications of the ACM 22(22), 612-613 (1979)
[18] Yao, A. C.: How to generate and exchange secrets, Proc. of FOCS '86, 162-167 (1986)
[19] Ben-Or, M., Goldwasser, S., and Wigderson, A.: Completeness theorems for non-cryptographic fault-tolerant distributed computation, Proc. of STOC '88, 1-10 (1988)
[20] Chaum, D., Crepeau, C., and Damgård, I.: Multiparty unconditionally secure protocols, Proc. of STOC '88, 11-19 (1988)
[21] Cramer, R., Damgård, I., and Nielsen, J.B.: Multiparty computation from threshold homomorphic encryption, EUROCRYPT 2001, LNCS 2045, Springer-Verlag, 280-300 (2001)
[22] Schoenmakers, B. and Tuyls, P.: Practical two-party computation based on the conditional gate, ASIACRYPT 2004, LNCS 3329, Springer-Verlag, 119-136 (2004)
[23] Damgård, I., Fitzi, M., Kiltz, E., Nielsen, J.B., and Toft, T.: Unconditionally secure constant-rounds multi-party computation for equality, comparison, bits and exponentiation, TCC2006, LNCS 3876, Springer-Verlag, 285-304 (2006)
[24] Nishide, T. and Ohta, K.: Multiparty computation for interval, equality, and comparison without bit-decomposition protocol, PKC 2007, LNCS 4450, Springer-Verlag, 343-360 (2007)
[25] 濱田浩気，五十嵐大，千田浩司，高橋克巳: 秘匿関数計算上の線形時間ソート, 第 28 回暗号と情報セキュリティシンポジウム (SCIS2011) 予稿集 (2011)
[26] Hamada, K., Kikuchi, R., Ikarashi, D., Chida, K. and Takahashi, K.: Practically efficient multi-party sorting protocols from comparison sort algorithms, ICISC 2012, LNCS 7839, Springer-Verlag, 202-216 (2012)
[27] Gentry, C.: Fully homomorphic encryption using ideal lattices, Proc. of STOC 2009, 169-178 (2009)
[28] Soral, A.: Achieving fully homomorphic encryption in security - A survey, SSRG-IJCSE 3(2), 22-27 (2016)
[29] 濱田浩気，五十嵐大，千田浩司: 秘密計算一括写像計算の効率化, 第 31 回暗号と情報セキュリティシンポジウム (SCIS2014) 予稿集 (2014)
[30] Laud, P.: Parallel oblivious array access for secure multiparty computation and privacy-preserving minimum spanning trees. PoPETs 2015(2), 188-205 (2015), http://www.degruyter.com/view/j/popets.2015.2015.issue-2/popets-2015-0011/popets-2015-0011.xml
[31] NBDC ヒトデータベース, http://humandbs.biosciencedbc.jp/
[32] Laur, S., Willemson, J., and Zhang, B.: Round-efficient oblivious database manipulation, ISC 2011, LNCS 7001, Springer-Verlag, 262-277 (2011)