ゲノムとプライバシー

Genome and Privacy

佐久間 淳 Jun SAKUMA

アブストラクト 近年の技術的進歩によって、ゲノムシーケンシングコストが急速に低下した。このことは大規模なゲノム疫学研究の発展をもたらし、遺伝子の個人差といえる一塩基多形 (SNP) と生活習慣病を含む多様な疾患の関係が明らかになりつつある。またこれらの成果によって、個人ゲノムに基づく疾患リスク予測も可能になりつつある。個人ゲノムは予防医療を中心とした様々な個別化医療への応用の基礎となる情報であるが、「究極の個人情報」とも呼ばれ、その扱いには慎重を要する。このため、ゲノム疫学や個別化医療においてゲノム情報を安全に活用するためのプライバシー保護技術の構築が望まれている。本稿ではゲノムデータの共有や公開に伴う多様なリスクと脅威について概観を示す。また、そのプライバシー上のリスクを減ずる主要なプライバシー保護技術として、秘密計算 (セキュアマルチパーティ計算) と差分プライバシーを紹介し、そのゲノム疫学と個別化医療への適用例について解説する。

キーワード ゲノム, SNP, ゲノム疫学, 個別化医療, プライバシー, セキュリティ

Abstract The availability of genomic data resulting from rapid developments in gene sequencing technologies has accelerated the progress in genome epidemiological studies, which also contributes to the realization of genome-based disease-associated risk prediction as part of a personalized medical treatment. However, privacy concerns currently hinder the ubiquitous exploitation of personal genomic and clinical data. Accordingly, it is necessary to develop privacy-preserving methods that will allow the analysis of genomic and clinical data for genetic epidemiology and personalized medication as below.

We give a broad overview of diverse risks and threats attached to the dissemination or publication of genetic data. As major privacy-enhancing technologies to reduce privacy risks, secure multiparty computation and differential privacy are then introduced. We describe how these technologies preserve the privacy of genomic data when applied to genetic epidemiology and personalized medication.

Key words privacy, security, genome, SNP, genome wide association study, personalized medicine

1. はじめに

ヒトゲノムの塩基配列には個人間の相違が存在する。ある程度以上の頻度を持って観察される一塩基毎の多様性を一塩基多形 (SNP: Single Nucleotide Polymorphism) と呼ぶ。このようなゲノム上の“個人差”が個人の体質、ひいては特定の疾患を引き起こす原因の一つになっていると考えられている。ゲノムワイド相関解析 (GWAS: Genome-Wide Association Study) は多数の個人の遺伝情報と臨床情報から、対象疾患と関連する SNP を統計的検定によって同定することを目的とした遺伝統計学的手法である。近年の技術発展にともない、大規模 SNP 解析のコストは劇的に減少し、個人の全 SNP が 10 数万円程度のコストで解析可能になったことから (図 1, 文献 (1) から引用)、糖尿病や様々ながん

等生活習慣病をはじめとする主要な疾患と強く関連する多数の SNP が報告されている⁽²⁾⁻⁽⁵⁾。環境的要因 (例えば喫煙歴、飲酒歴、血圧) や遺伝的要因 (例えば SNP) と疾患の関連性を明らかにすることで、近い将来に個人の体質に合わせた効果的な個別化医療や予防医療が可能になると考えられている。

ゲノム疫学の課題は①有用な SNP の探索、②コストの低廉化、③プライバシーの保護、と言われている。①及び②はゲノム疫学研究の発展とシーケンシングに要するコストの低下に伴い解決されつつあるが、最後のプライバシーの問題は、技術、社会、法制度、倫理、そして人間の主観 (漏えいに対する不安感、監視に対する嫌悪感、気持ち悪さ) など、多様な問題が複雑に混ざり合い、解決が困難な問題である。

技術的には、ゲノムのプライバシーの問題は、通常の個人情報におけるプライバシーの問題とは顕著に異なる性質を持ち、これが個人ゲノムの取扱いを難しくしている。一般の個人データ解析においては、個人を識別する情報 (氏名や ID 番号、電話番号など) と、個人の特徴に関する情報 (年齢や性別、物品購買履歴や移動履歴) は明確に分離されている。個人を識別する情報は、それ自身が意味のある情報を含まないため、個人を識別する情報と

佐久間 淳 正員 筑波大学大学院システム情報工学研究科
E-mail jun@cs.tsukuba.ac.jp
Jun SAKUMA, Member (Graduate School of System and Information Engineering, University of Tsukuba, Tsukuba-shi, 305-8573 Japan).
電子情報通信学会 基礎・境界サイエンス
Fundamentals Review Vol.7 No.4 pp.348-364 2014 年 4 月
©電子情報通信学会 2014

図1 人間と同サイズの全ゲノムのシーケンシング解析コスト

個人の特徴に関する情報を切り離しても、後のデータ解析自体に影響を与えることはない。しかしゲノム解析においては、ゲノム自体が個人の特徴に関する情報（特定の病気の罹患リスクや太りやすさ、身体能力等の体質）を符号化しているとともに、ゲノム自体が個人を識別する情報として働くことから、この両者を切り離すことがそもそも不可能である。このゲノム情報の二面性が、「ゲノムは究極の個人情報である」と評されるゆえんであり、また個人ゲノム情報特有の取扱いの難しさの原因の一つとなっている。

個人ゲノムのプライバシー上のもう一つの懸念は、個人ゲノムに対する科学上の理解がまだ不完全であり、現在も新しい事実が次々と明らかになっている点にある。インターネットによる情報流通が発展した現代において、一度公開状態に置かれた情報（あるいは漏えいした情報）は二度と非公開状態に戻すことはできない一方で、ゲノム自体から判明する情報は年々増えている。このことから、現時点でリスクが低いと判断され公開されていた個人ゲノムに関連する情報も、時間がたてば個人に大きな不利益をもたらす可能性がある。

ゲノムのプライバシー保護は、個人ゲノムを取り扱う研究者やその研究の倫理上の妥当性を審査する倫理委員会等の倫理感に頼るところが大きく、技術的な解決は手つかずである。個別化医療や予防医療の普及を考えれば、今後は研究者のみならず医療関係者や患者自身が個人ゲノムに触れる機会を持つことになる。本稿では個人ゲノム解析にまつわるプライバシー上のリスクを様々な角度から議論するとともに、その技術的解決における現状の state-of-the-art を紹介する。

本稿の構成は以下のとおりである。2. ではゲノムに関する基礎的知識及び医療応用に関連する個人ゲノム解析を導入する。3. では、個人ゲノムの開示に関わるプライバシー上のリスクについて俯瞰する。4. では、個人ゲノムに関連する情報からの推論に関わるプライバシー上のリスクについて考察する。5. では、プライバシー保護における主要な要素技術である秘密計算を導入する。また、組織間で横断的に保持される個人ゲノム関連情報を用いたゲノム解析における秘密計算を解説する。6. では、もう一つの主要な要素技術である差分プライバシーを導入する。また、個人ゲ

ノム関連情報の公開におけるプライバシー上のリスクの問題と、その差分プライバシーによる解決について解説する。7. はまとめである。

2. 個人ゲノム解析と医療応用

2.1 ゲノムと一塩基多型

個々の人間を詳しく観察すれば、体質や能力、見た目や気質などの特徴は多様である。このような生物が持つ特徴を形質と呼ぶ。人間の形質には様々なものがある。最も観察しやすい特徴や目の色髪の色、肌の色やなど、人間の見た目に関わる特徴である。表立って観察できない形質には、血液型やアルコール耐性、乳糖不耐性などがある。これらは人間が生まれ持った生涯変わることがない特徴である。このように客観的に観察できる形質を表現型と呼ぶ。

親子の間で形質が似ることを遺伝現象と呼ぶ。形質に関わる情報（遺伝情報）を（全てでないにせよその多くを）伝達する情報をゲノムと呼ぶ。ゲノムは染色体と呼ばれるひも状の23対の高分子からなる（図2、文献(6)から引用）。染色体の実体はDNAと呼ばれる長い高分子が二重らせん構造を成してコンパクトに折り畳まれた分子であり、DNAに人間の形質に関わる情報のほとんどが記録され、親子間で伝達されている。

遺伝情報は、どの個体においても共通してこのひも状の染色体の特定の位置に記述されており、この部位を遺伝子座と呼ぶ。1本の染色体は1本のDNAを含み、DNAは2本の鎖状高分子一組からなる。DNAはプリン塩基であるアデニン（A）とグアニン（G）と、ピリミジン塩基であるシトシン（C）とチミン（T）の4種類が1列に結合した一組2本の鎖状高分子からなる。片方の鎖状高分子上の塩基ともう片方の鎖状高分子上の塩基は緩やかな水素結合を介して結合しており、全体として二重らせん構造をとる。また塩基同士の水素結合は相補的であり、AはTと、GはCと結合を成すことから、片方の鎖状高分子上の塩基A、T、G、Cの4種のうち、1種を決めればそれと水素結合で結ばれるもう1種も自動的に決まる。

全ての個体に共通して、ある特定の遺伝情報は、染色体上の特定の位置にある塩基の配列により符号化される。遺伝情報とこれを符号化する遺伝子座の関係を遺伝子地図と呼ぶ。DNAは多数の塩基の配列から形成されるが、その配列が符号化する情報は多様である。三つの塩基は一つのアミノ酸を符号化し、その並びが特定のアミノ酸配列からなる一つのたん白質を符号化するには、DNAはたん白質の設計図として働く。また塩基の並びが個々の遺伝子の発現制御に関連している場合もある。

遺伝子座は特定の遺伝情報を符号化する塩基配列の位置を表すが、その長さは1塩基の場合もあれば、数百万塩基に達する場合もある。個体は父母から一つずつ染色体を受け継ぎ、一つの遺伝子座に二つの塩基を持つ。遺伝子座における塩基（配列）のバリエーションを対立遺伝子と呼ぶ。対立遺伝子の構成はメンデルの法則に従う。対立遺伝子が“A”と“G”から成る場合に

図2 一塩基多形の例。DNA分子1はDNA分子2と一座位において異なる（C/T多型）。

は、対立遺伝子の組合せは“AA”、“AG”、“GG”の3種類となる。この対立遺伝子の組合せを遺伝子型と呼ぶ。

特定の座位の遺伝子型にバリエーションが見られる場合、これを遺伝的多型と呼ぶ。長さが1塩基の遺伝子座において、変異による多様性がある一定以上の頻度で観察される場合、一塩基多型（SNP）と呼ぶ。乳糖不耐性、アルコール分解酵素に関わる体質等、2型糖尿病の罹患リスクなど、単一のSNPで高い精度でその形質を予測できるものもある。一方、生活習慣病の罹患リスクなど、遺伝的要因だけでなくそれ以外の環境的要因の影響も同時に考える必要がある場合も多い。

またゲノム中には数塩基から数十塩基の特定の配列が繰り返し表れる箇所があるが、その繰り返し回数が個人によって異なることが知られている。このような多型をSTRP（short tandem repeat polymorphism）と呼ぶ。後述するように、SNPは疾患遺伝子探索などに、後者はDNA鑑定（血縁関係の鑑定等）の遺伝マーカーとして利用される。

人間の遺伝情報は約32億個の塩基で符号化されている。そのほとんどは全人類で共通であるが、約0.5%は個人ごとに異なると言われている。これが人間の形質の遺伝的多様性の元となっている。このようなSNPのうち、マイナーな対立遺伝子の割合が少なくとも数%を占めるような主要なSNP（common variant）を研究対象とする方法が主流である。ただし、現在では、発現の機序が単純ではない生活習慣病に関連するSNPの探索等において、出現割合が比較的低いrare variantにまで研究は広がっている。このようなSNPは現在約4,000万あると言われ、サンプル数や解析機器の発展とともにSNPの数は現在も増加している。

2.2 個人ゲノム

2.2.1 記法

本稿では N 人の被験者から集めた M 個のSNPの対立遺伝子を考える。あるSNP座位において、例えば A が高頻度で現れる対立遺伝子、 a が低頻度で現れる対立遺伝子とした場合、この

図3 ゲノムに関わる情報とその相互依存性。

SNP座位の遺伝型は AA, Aa, aa のいずれかである。

被験者のSNPによる遺伝的特徴は、 $2M$ 次元のバイナリー特徴で表現される。優性モデルにおけるバイナリー特徴では、あるSNP座位において、その対立遺伝子が一つでも A を持つ場合（つまり AA あるいは Aa の場合）を1、それ以外（つまり aa の場合）を0と符号化する。劣性モデルにおけるバイナリー特徴では、一つでも a を持つ場合（つまり Aa or aa の場合）を1、それ以外（つまり AA の場合）を0とする。前者は高頻度に出現する対立遺伝子が優性である場合に有効な特徴であり、後者は高頻度に出現する対立遺伝子が劣性である場合に有効な特徴である。以降、優性、劣性は区別せず、独立した遺伝的特徴として扱う。

i 番目の被験者の j 番目のSNPの対立遺伝子を $x_{ij} \in \{0, 1\}$ とすれば、 i 番目の被験者の個人ゲノムは $x_i = (x_{i1}, x_{i2}, \dots, x_{i2M})^T \in \{0, 1\}^{2M}$ と記述される。 i 番目の被験者に関する形質に関する情報を $y_i \in Y$ と書く。これは、特定の疾患の罹患歴や薬剤感受性など、 i 番目の被験者の遺伝的要因と関連していると信じられている情報である。これを形質情報と呼ぶ。また i 番目の被験者に関するそのほかの属性情報を $z_i \in Z$ と書く。これは、 i 番目の被験者の遺伝的要因とは必ずしも直接関連しない、年齢や飲酒歴等の環境上の要因を表す。これを非形質情報と呼ぶ。

生物学的血縁関係にある個人のSNPは独立ではないため、血縁関係を解析に用いることがある。被験者同士の血縁関係は被験者集合 $V = \{1, 2, \dots, N\}$ をノードとする木構造で表現できる。被験者 i と被験者 j の血縁関係を枝 $e_{ij} \in E$ で表せば、家系図全体はグラフ $G = (V, E)$ で表現できる。これを血縁情報と呼ぶ。個人 i に関する血縁情報を e_i と表記する。

プライバシーを考える上では、個人のアイデンティティとこれらの情報の関係を考える必要がある。アイデンティティとは、具体的には個人の名前や保険者番号、マイナンバーなど、個人と一対一で結び付く番号で表現される情報である。ここでは、 $V = \{1, 2, \dots, N\}$ でアイデンティティを表すことにする。これらを合わせ、 i 番目の被験者の情報は (i, x_i, y_i, z_i, e_i) と表記できる。これらの情報は常に全て一度に収集されるわけではなく、解析目的に応じて、必要な属性が取捨選択される。

2.2.2 ゲノムに関わる情報とその相互依存性

個人ゲノムに関連する3情報、SNPを含む個人ゲノム、形質情報、血縁情報は、互いに複雑な依存関係にある（図3）。まず個人の形質は個人のゲノムに応じて発現するため、個人ゲノムと形質

情報は非独立である．また遺伝現象のため、生物学的血縁関係にある二人の個人とその個人ゲノムは非独立である．これらより、生物学的血縁関係にある二人の個人の形質情報も非独立である．更に、複数の SNP 座位の対立遺伝子の間にランダムでない相関が見られることから、SNP 同士は互いに独立ではない（連鎖不平衡）．これに従い、個人が持つ複数の形質情報も、必ずしも独立ではない．

ゲノムを利用した解析の多くは、このゲノムと、これに関わる形質情報、血縁情報の依存関係に関する解析である．その一例を挙げれば、DNA 鑑定は、2 ゲノムの類似性から血縁関係を推定している．罹患同胞対解析は、兄弟の関係にあるゲノムと形質の依存性から、疾患遺伝子を同定する解析である．ゲノムワイド相関解析（GWAS）は、血縁関係にないゲノムと形質の依存性から、疾患遺伝子を統計的に同定する解析である．また個別化医療では、GWAS などによって構築された、特定疾患についてのリスクモデルを用いて、あるゲノムを持つ個人が、将来その疾患に罹患するリスクを評価する．これらの解析は、ゲノムに関連する情報を適切な同意と手続きの下で取得した上で、統計的な依存関係を活用して有用な未知情報を推論していると言える．

2.3 個人ゲノムの解析

2.3.1 ステークホルダー

本稿で共通して用いる用語と記法について整理する．個人ゲノム解析に関わる主要なステークホルダーは主に以下のとおりである

- ・ 被験者・患者: 被験者あるいは患者（以下被験者）は自身のゲノムサンプル（e.g., 口腔細胞など）を提供することに同意し、これを医師あるいは検査機関に提供する．また医師やそのほかの機関に、臨床情報や環境情報を提供する．
- ・ 検査機関: ゲノムシーケンスを保持している組織．被験者・患者のゲノムサンプルを解析しその塩基配列を読み取る
- ・ 医師: 患者のゲノムや環境情報に基づき疾患リスク評価を行い、個別化医療を提供する
- ・ ゲノム研究者: 患者から臨床情報やゲノムサンプルを取得し、疾患関連遺伝子を探索し、ゲノムや環境情報に基づく疾患リスクモデルを構築する

これらに加えて、ゲノムデータの保管の委託を受け、これを安全に保持するデータ収集者（クラウドストレージ）、依頼を受けてデータ解析を行うデータ解析者、データ解析者がプライバシー保護規約に従うよう規制する監督官庁などもステークホルダーとして考える場合がある．

2.3.2 個人ゲノムに基づく解析

個人ゲノムを利用した解析は以下の三つに大きく分類される．

（１）個人ゲノムの類似性等に基づく鑑定

- （a）本人鑑定は、ゲノム x と x' が同一人物のゲノム

表 1 疾患遺伝子探索に用いられる分割表の例

	対立遺伝子 AA, Aa	対立遺伝子 aa	合計
ケース	a	b	$a + b$
コントロール	c	d	$c + d$
合計	$a + c$	$b + d$	n

であるかを判定する．犯罪捜査で用いられる．

- （b）親子鑑定は、ゲノム x_i と x_j について、 $e_{ij} \in E$ か否かを判定する．

- （c）親戚関係はゲノム x_i と x_j の E 上の最短パス長から判定することができる

（２）ゲノム疫学

- （a）疾患関連遺伝子探索: サンプル $\{(x_i, z_i)\}_{i=1}^N$ について、 j 番目の SNP と特定の疾患の独立性を検定し、その疾患の発症に強く関連する SNP を発見する．

- （b）リスクモデル構築: サンプル $\{(x_i, z_i)\}_{i=1}^N$ を用いて、ゲノム x_i を持つ個人が、特定の特徴（疾患罹患リスク、薬剤感受性等） y_i を持つリスクを評価する関数 $f(x_i, z_i)$ を学習する．

（３）個別化医療/出生前診断

- （a）個人からゲノムを取得し、リスク評価関数 f に基づき、特定の疾患の罹患リスクを評価し、これに基づき予防的医療を行う．出生前診断の場合は、母親の羊水や血液に含まれる胎児のゲノムを用いる

- （b）個人からゲノムを取得し、リスク評価関数 f に基づき、特定の薬剤に関する感受性を評価し、これに基づく治療を行う

本稿では主にゲノム疫学と個別化医療に関わる問題を議論する．

2.4 ゲノム疫学と疾患関連遺伝子探索

既に述べたように、遺伝子には個人差があり、これが個人の体質、ひいては特定の疾患を引き起こす原因の一つになっていると考えられている．もし、対象疾患（例えば糖尿病）を持つ個人の集団（ケース）について、ある特定の SNP 座位における特定の対立遺伝子が、その疾患を持たない集団（コントロール）よりも高頻度で観察されるならば、その SNP 座位におけるその対立遺伝子は、その疾患に関連していることが疑われる．疾患遺伝子探索はこの観察に基づいて、統計的検定により探索される．疾患遺伝子探索の基となるデータは、対象とする疾患に罹患しているか否か（症例/非症例）、及び、リスク因子を保有するか否かを集計した分割表である．疾患遺伝子探索の場合は、リスク因子はある特定の SNP 座位の対立遺伝子となる．ここでは優性モデルにおいて、対立遺伝子 AA, Aa 及び対立遺伝子 aa を遺伝的特徴としたときの分割表の例を表 1 に示す．

統計的推論に基づく疾患遺伝子探索では、 χ^2 統計量に基づく検定がよく利用される．もし対象疾患と注目する SNP の間に何の関連もない場合には、例えば対立遺伝子 AA or Aa を持ち、かつその疾患を罹患する被験者の数の期待値は $(a + b)(a + c)/n$ となる． χ^2 統計の検定統計量

$$T = \frac{(a+b+c+d)(bc-ad)^2}{(a+b)(a+c)(b+d)(c+d)} \quad (1)$$

は、この SNP と対象疾患は独立であるという帰無仮説の下で、この期待値と実際の観測値の間に隔たりがあるか否かを定量化する。対象疾患と注目する SNP の関連性は、 p 値と呼ばれる値で定量化される。 p 値とは、この帰無仮説の下で、実際にデータから得られた統計量よりも極端な統計量が得られる確率であり、ある有意水準（例えば $\alpha = 0.05$ ）において、実測値とモデルによる推測値の間に優位な差があるか否かを評価している。 p 値が小さいほど、対象疾患と注目する SNP の関連は強いと判断できる。

GWAS における疾患関連遺伝子探索では表 1 に相当する分割表を、調査対象とした全 SNP についての個別に作成し、各 SNP と対象疾患の関連の強さを網羅的に調査するアプローチを取る。

2.5 ゲノムに基づく疾患リスクの予測

疾患リスクは、個人の遺伝的要因のみならず、環境的要因にも強く影響される。ゲノムに基づく疾患リスクは、疾患関連遺伝子同定によって発見された疾患関連遺伝子に加え、これらの環境的要因を考慮して評価される。

このような多変数特徴（この場合、遺伝的要因や環境的要因）からの、2 値の目標値に関するリスク（この場合、罹患するか否か）の予測には、生物医学分野では伝統的にロジスティック回帰が用いられる。ロジスティック回帰は以下の線形式の予測に帰着され、そのモデルパラメータは凸関数の最適化に基づき求めることができる。

$$f(x_i, z_i; \mathbf{w}) = \sum_j \overbrace{w_j x_{ij}}^{\text{遺伝的要因}} + \sum_j \overbrace{w_j z_{ij}}^{\text{環境的要因}} \quad (2)$$

ここで、 x_i は個人の遺伝的要因を表す 2 値特徴、 z_i は個人の環境的要因を表す特徴である。 w_j は遺伝的要因あるいは環境的要因に対応するリスク予測のモデルパラメータである。前節で導入したケースコントロール研究においては、ロジスティック回帰モデルの出力は対数オッズ、モデルパラメータの各要素は、その要素に対応する特徴の対数オッズ比に相当する⁽⁷⁾。

3. 個人ゲノムの漏えいに関わるリスク

個人ゲノムは「究極の個人情報」と呼ばれ、その取扱いには極めて慎重であるべきことは広く認識されている。個人ゲノムに関わる情報の一部を取得した攻撃者が、データ提供者やデータ収集者にとって意図されない目的や手段によって未知情報が推論されてしまう場合に、プライバシー上のリスクがあると言える。より正確には、攻撃者が個人ゲノムに関わる情報である $\{(i, x_i, y_i, z_i, e_i)\}$ の一部を観測したときに、未観測の情報を高い確信度で推測できるならば、個人ゲノムに関わるプライバシー上のリスクが高いと言える。

アイデンティティが取り除かれた個人ゲノムについて、大規模な個人の再識別が発生した事例として Personal Genome Project⁽⁸⁾ におけるインシデントが挙げられる。Personal Genome Project は

ボランティアでの個人ゲノムデータの提供を受け、これを集約し、医療の発展及び科学的発見に資することを目的とし、オープンデータとして公開を行っているプロジェクトである。

Personal Genome Project は 23andMe が被験者に提供した SNP データの提供を受け付けていたが、23andMe が提供する SNP データのファイル名に個人名が含まれていたこと、Personal Genome Project にファイル名を変更せずにアップロードする人が多かったこと、Personal Genome Project のオープンデータにそのファイル名が意図せず含まれてしまったことから、多数の個人について、ゲノムと個人名が結び付けられる事態となった。

現在 Personal Genome Project は、ゲノムを提供する個人について、将来において“匿名性は保証されない”ことへの同意（open consent）を前提としてデータを収集・公開している。Personal Genome Project は、遺伝学等に知識のある者が被験者のゲノムを手にした場合のリスクとして、提供された個人ゲノムが以下の目的で利用される場合があると指摘している：

- 被験者の父系やそのほかの（血縁関係に関する）情報の推定
- 被験者の雇用、保険、及び金融サービスを受ける資格に影響を与える統計的な証拠の推定
- 被験者やその親戚の犯罪性向の推定
- 被験者に相当する DNA を合成し、これを犯罪目的で（誰かに）埋め込む
- 現在において効果的な治療法が知られていない疾患の罹患リスクの開示

上記リスクのうち、2～4 番目の項目は、あくまで最悪ケースにおいてこのようなリスクがあり得ることを提示しているにすぎない。しかし Personal Genome Project による個人ゲノム開示のリスクの例示は、ゲノム科学の発展のスピードがいまだ著しく、将来的に発生し得るリスクが現時点で正確に評価できないことの傍証とも言える。特に 2 番目の遺伝的差別のリスクは法学者を中心に広く認識されており⁽⁹⁾ ⁽¹⁰⁾、医療保険市場における遺伝的差別や採用における遺伝的差別を法律により既に禁じている国もある⁽¹¹⁾。

4. 個人ゲノム関連情報からの推論とそのリスク

4.1 推論と攻撃のモデル

Personal Genome Project の事例はゲノム x_i にアイデンティティ i を付与したまま公開されたことにより発生したインシデントである。ゲノム自体にこのようなアイデンティティが付与されていない場合であっても、2.2.1 に詳述したような個人ゲノム関連情報である、個人のデモグラフィック情報、形質情報、血縁情報などが個人ゲノムに付与されている場合、その統計的な依存性を利用することで、アイデンティティや機微な情報が推定される可能性がある。ここでは前者をゲノム識別推定、後者をゲノムベース属性推定と呼ぶ。

形式的には、攻撃者がアイデンティティ以外の情報 (x_i, y_i, z_i, e_i)

表 2 攻撃者が保持する情報、攻撃者が外部から取得する情報とその情報源、及び攻撃者が推測する情報の関係

		外部情報源					
		物理的接触	公的 DB, SNS 事業者 DB	公的 DB, SNS 事業者 DB	物理的接触 SNS	研究 DB	研究 DB 論文
		取得情報					
		ID+ゲノム	ID+非形質情報	ID+血縁情報	ID+形質情報	分割表	p 値+相関係数
攻撃者情報	ゲノム	識別推定 (4.2.1)		識別推定 (4.2.3)			SNP
	ゲノム+血縁情報						
	ゲノム+形質情報				識別推定 (4.2.4)	属性推定 (4.3.1)	属性推定 (4.3.2)

の部分情報を保持するが、アイデンティティ情報は保持していないときに、 (x_i, y_i, z_i, e_i) と i を結び付けることをゲノム識別推定攻撃（あるいはレコードリンケージ、リンケージ攻撃など）と呼ぶ。一方、攻撃者がアイデンティティ以外の情報 (x_i, y_i, z_i) の一部を保持するときに、（アイデンティティとのひも付けを行う、行わないにかかわらず）未観測の属性情報を推測することをゲノムベース属性推定攻撃と呼ぶ。両者は背反ではなく、識別推定と属性推定が同時に発生する場合もある。

何らかの形で被験者のゲノム情報や機微な情報（被験者の症例など）が推測される場合、その被験者を犠牲者と呼ぶ。また何らかの手段で入手した情報を基に、犠牲者のゲノム情報や機微な情報を取得あるいは推測しようとする者を攻撃者と呼ぶ。表 2 に、本章で紹介する事例について、攻撃者が保持する情報、攻撃者が外部から取得する情報とその情報源、及び攻撃者が推測する情報の関係をまとめる。

いわゆる悪意を持って情報を不正に入手・推測しようとする者のみを攻撃者と呼ぶわけではないことに注意されたい。便宜上、攻撃の意思の有無にかかわらず、入手可能な情報を用いてあらかじめ想定された範囲を超えた情報を推測し得る立場にある者全てを攻撃者と呼ぶ。したがって、検査機関、医師、ゲノム研究者、データ収集者、データ解析者など全てのステークホルダーを攻撃者として捉えることができる。

表 2 に表される攻撃者は、個人ゲノム情報、あるいは、個人ゲノム情報とその関連情報を保持する者を指す。多くの場合、このような攻撃者として、個人ゲノムを収集した主体である検査機関あるいはゲノム研究者を想定する。ただし、研究目的で個人ゲノムを外部公開する例や、ソーシャルネットワークなどで自発的に自身の個人ゲノムを公開する例⁽¹²⁾もあり、これらの個人ゲノムデータをダウンロードした者も、攻撃者になり得ると考えるべきである。

表 2 に表される外部情報源は、個人ゲノム情報に関連する情報を提供する者を指す。多くの場合、GWAS 関連の情報（分割表や p 値、相関係数など）を研究目的で論文・Web サイト等で公開するゲノム研究者や、デモグラフィック情報、形質情報、家系情報を保持する公的データベース、事業者データベース、これらの情報を一般に公開するソーシャルネットワークなどが含まれる。一方、犠牲者は常にゲノム識別推定あるいはゲノムベース属性推定の対象となる被験者である。

以降、本章では表 2 にあるそれぞれの事例について解説してゆく。

4.2 ゲノム識別推定

この節では個人ゲノムと特定の個人を結び付けるアイデンティティは持たないが、その個人に関連する非形質情報、形質情報、家系情報を持つゲノム解析者が、これらの情報を用いて、個人ゲノムとそのアイデンティティをひも付けるゲノム識別推定について議論する。本節及び次節の議論は文献 (28) を参考しており、あわせて参照されたい。

4.2.1 ゲノム情報自体によるゲノム識別推定

ゲノムの実体は約 30 億対の塩基配列の並びである。各塩基は A,T,G,C の 4 符号語を持ち、1 塩基は 2 bit の情報を持つことから概算すれば、ゲノムは約 6,000,000,000 bit=750 MByte 程の情報量を持つ。ただし、個人ごとに異なる情報はゲノム全体の約 0.5 % と言われ、ヒトゲノムに共通して含まれる情報を取り除けば、各個人に特有のゲノムの情報量は概算で 3.75 MByte ということになる。個人ゲノム自体の情報量はさほど大きいとは言えないが、ゲノムを個人を識別する手掛かりと考えれば話は別である。

二人の人間が同一の個人ゲノムを持つ可能性は天文学的に低く、個人ゲノムはそれ自体が超強力なアイデンティティとして働く。具体的には、個人の識別に適した 45 の SNP を用いることで、極めて高い確率で個人が識別可能であると言われている⁽¹³⁾。またランダムに SNP を選んだとしても、300 程度の SNP 情報があれば個人識別には十分であるとの報告もある⁽¹⁴⁾。SNP はこれらの報告とは異なるが、犯罪捜査における DNA 本人鑑定や DNA 親子鑑定はこのゲノムの識別子としての頑健性に基づいた検査であると言える。

このため、個人ゲノムのデータベースを有する攻撃者が、犠牲者となる個人の体から物理的にサンプルを取得し、これを自身でシーケンサによって解読できる場合には、標的のゲノムを特定することは容易である（図 4）。この攻撃は、例えばゲノムデータの保管・解析を行うデータ収集者・データ解析者が、連結可能／不可能匿名化されたゲノムを特定する手段として使えるだけでなく、研究目的等で公開されたゲノムデータのレポジトリから、個人を特定する手段としても利用できる。現状ではゲノムシーケンサは極めて特殊かつ高価な機器であり、その保有者は研究者に限られているため、高いモラルの下での運用が期待できる。しかし、将来的にゲノムシーケンサが低価格化した場

図4 個人ゲノム自体によるゲノム識別推定.

合には、ゲノムシーケンサを保持する機関による解析と解析結果の外部公開には注意が必要である。具体的には、以下のよう
な措置が必要であると考えられる:

- 体毛等のサンプルは、容易に取得可能であることから、身元を偽った解析依頼が行われる可能性がある。身元不明の解析依頼者からの解析結果を受任するべきではない。
- ゲノム解析結果が本人以外に誤って開示されることがあってはならない。このため、結果の開示は何らかの本人確認を伴って行われるべきである。

4.2.2 非形質情報によるゲノム識別推定

疫学等の目的で個人ゲノムを収集した場合には、個人ゲノムとともに、年齢、居住地、生活習慣など、個人ゲノムと直接関連しない非形質情報が収集され、 (i, x_i, z_i) の形式を取る (既往歴などの個人ゲノムと強く関連する形質情報も同時に収集されることが多いが、ここでは非形質情報のみを考える。) 識別推定を防ぐためには、当然ながら個人ゲノムとアイデンティティを切り離して保管する必要がある。このため、情報からアイデンティティを取り除き、各データに新たに識別番号を付番する。これを以降仮アイデンティティと呼ぶ。このアイデンティティと仮アイデンティティの対応表を保存しない場合、これを連結不可能匿名化と呼ぶ。仮識別情報を付与した者が、この対応表を保存する場合、連結可能匿名化と呼ぶ。連結可能匿名化は、例外的な理由で匿名性を持つデータから特定個人を識別する必要性が生じる可能性がある場合に用いられる。

連結不可能匿名化された個人ゲノムとその非形質情報 (x_i, z_i) を保持する攻撃者が、アイデンティティとその個人の非形質情報 (の一部) (i, z_i) を別途入手した場合には、非形質情報 z_i を介することで再識別が起こる可能性がある (図5)。

(i, z_i) は、例えば (氏名、性別、年齢、居住地) といったレコードであり、病院等事業者であれば、業務としてこのような情報を保持していることが多い。また Facebook⁽¹⁵⁾ などのソーシャルネットワークや people finder⁽¹⁶⁾ などの情報サービスを用いて容易に収集可能である。

このような非形質情報の識別能力を表す具体的な例として、性別、生年月日及び5桁の郵便番号の組合せから、米国民のうち60%が再識別可能であると報告されている⁽¹⁷⁾。マサチューセッツ州の団体保険委員会が同時公開していた性別、生年月日及び5桁

図5 非形質情報を経由したゲノム識別推定.

の郵便番号のデモグラフィック情報を含む退院記録情報と、同じく公開情報であるマサチューセッツ州の選挙人名簿^(注1)を突き合わせ、マサチューセッツ州知事の退院日が特定された事件は有名である⁽¹⁸⁾。

このような認識の下、米国の医療情報システムにおける HIPAA セーフハーバー標準⁽¹⁹⁾では、デモグラフィック情報の記録について、89歳以上の年齢情報の削除、州より詳細な居住地情報の削除、生年月日、入院日、退院日などの日付情報からの月と日の情報の削除、などを定めている。

非形質情報による識別を防ぐための匿名性の基準としては、 k 匿名性が知られている。 k 匿名性とはデータの正確性を犠牲にして非形質情報やセンシティブ情報を改変し推測を困難にする手法である。具体的には、表形式データについては、非形質情報の属性値の組合せが同じであるデータが、少なくとも $k (> 1)$ 行存在している状態として定義される⁽¹⁸⁾。非形質情報の属性値の組合せがデータと同一でない場合には、これが同一になるように、属性全体の削除や属性値の削除、データの一般化 (“32 歳” を “ [30-39] 歳 ” にする、など) が行われる。

非形質情報はゲノム情報とは統計的な依存関係がないため、非形質情報による再識別は個人ゲノム固有の問題とは言えない。しかし、非形質情報とゲノム情報を結合させ記録しておく場合のプライバシーリスクには注意する必要がある。連結可能 / 不可能匿名化は簡便な匿名化手法であり、現在のように限定された研究者等がゲノム解析に携わる状況であれば、ゲノムのプライバシーは高いモラルの下で保護されることが期待できるが、このようなデータからの再識別 (脱匿名化) は比較的容易に起こり得る。つまり、ゲノムに氏名等のアイデンティティが結合していなくても、デモグラフィック情報などの非形質情報が結合している場合には、外部情報を用いた絞り込みによって再識別が起こる可能性が低くはないことを認識する必要がある。外部情報を用いた再識別のリスクとその対応 (匿名化手法等) については、成書⁽²⁰⁾やサーベイ⁽²¹⁾に充実した議論があるため、必要に応じて参照されたい。

4.2.3 血縁情報によるゲノム識別推定

生物学的血縁情報とゲノムは強く依存するため、罹患同胞対解析などの目的で血縁情報は個人ゲノムとともに収集される機会が多い。生物学的血縁情報は木構造で表現できるが、疫学で用いる家系図には夫婦関係を必要とする場合があり、再婚・養

(注1): 米国では選挙人名簿 (voter registration list) は公開情報である。

図6 家系情報を経由したゲノム識別推定.

子縁組を含む場合の家系図はグラフ構造で表現される.

個人ゲノムから、4世代あるいは5世代前の両親を共有する親戚であるかどうかを比較的高い精度で同定できることが知られており⁽²²⁾、遺伝子検査サービスを提供する23andMeは、遺伝子検査の結果判明した(被検査者が知らない)親類縁者を紹介するサービスを行っている⁽²³⁾。

家系図情報は、規模が大きい場合、識別能力が高いことが知られている。文献⁽²⁴⁾では、6万人から成る2,500の2世代にわたる家系を調査しており、血縁関係そのものが分からなくとも、対象とする家系の各世代が含まれる男女数を知ることができれば、約30%の家系が再識別されとの報告がある⁽²⁵⁾。血縁情報による再識別におけるもう一つの注意点は、家族の中の誰か一人が再識別された場合、家族の残りのメンバーの再識別が容易である点である。

一般に多数の家系情報の入手は困難であるが、イスラエルでは2006年に住民登録に関するデータが漏えいし、多世代にわたるイスラエル市民の家系図が公になる事件があった⁽²⁶⁾。またFacebookなどのソーシャルネットワークサービスでは、個人同士の血縁関係が公開されているケースは少なくない。家系図そのものを対象とするソーシャルネットワークサービスも存在する⁽²⁷⁾。

4.2.4 形質情報によるゲノム識別推定

ゲノムは形質情報や血縁情報などと統計的な依存性を持つ。これらの統計的な性質を利用した再識別について説明する。本節の議論の多くは文献⁽²⁸⁾に充実したサーベイがあるため、参考に使いたい。

生物学的血縁関係にある父子は、Y染色体を共有する。また父系出自の社会においては、姓は父親から子へ継承される。このことから、男性のY染色体の遺伝型と姓には高い関連性があることが予想され、近年になりゲノムからの姓の推定に関する研究が進んだ⁽²⁹⁾⁽³⁰⁾。調査は中流から上流階級に属する米国白人男性を対象としており、ゲノム単体では、6%程度の誤りを認めた場合、12%程度の確率で姓が正しい姓が推定されると報告している⁽³⁰⁾。このように形質の推定を媒介したゲノム識別推定概念図を図7に示す。自身の祖先について調査できる商用のancestry dna database⁽³¹⁾を用いて、匿名の精子提供者から出生した養子が、生物学上の父親を探し当てた例は複数例報告されている⁽³²⁾⁽³³⁾。姓の特定は必ずしも個人の特定につながらないが、

(注2): この引用文献はイスラエル語の文献であり、ここに記載した情報は文献⁽²⁸⁾に基づく。

図7 形質情報によるゲノム識別推定.

ゲノムからは性別が推定可能である。また瞳の色、髪の色も比較的高い確率で推定が可能であり、これらも再識別の有力な手掛かりとなる。確実な推定とは言えないが、またゲノムから人相を再構築する試みも行われている⁽³⁴⁾。

4.3 ゲノムベース属性推定

本節では、個人ゲノムを何らかの形で利用して、個人のアイデンティティと、その個人のセンシティブな属性の属性値(疾病の罹患歴等)をひも付けるゲノムベース属性推定(攻撃)について議論する。

4.3.1 サマリー統計量からのゲノムベース属性推定

Homerらは、GWASを目的とした何らかの対象疾患をケースとした対立遺伝子の頻度が公開されており、かつ、攻撃者が犠牲者のSNPを保持しているときに、犠牲者があるグループ(つまり疾患を持つケースグループ)に属するか否かを統計的に推測可能であることを示した(図8)⁽³⁵⁾。International HapMapプロジェクト⁽³⁶⁾はreferenceとなる集団に関する対立遺伝子の頻度分布を公開している。このときに、GWASで収集された特定のケース/コントロールグループから得た対立遺伝子の頻度分布が与えられ、かつ、そのGWASにサンプルを提供した犠牲者のSNPが与えられているならば、犠牲者がケースグループとコントロールグループ、どちらの群に属するかを統計的に検定することができる。これによって、Homerらは犠牲者がその対象疾患を罹患しているかどうかを推定可能であると結論付けている。犠牲者から1万程度の独立なSNPのプロファイルを得ることができれば、高い信頼度でケースグループに属するか否かが決定できると報告している。この報告を受け、NIHは公開情報としていたGWASに関連する頻度表を非公開とすることを決定した⁽³⁷⁾⁽³⁸⁾。この報告に関する数理的解析は文献⁽³⁹⁾~⁽⁴¹⁾に詳しい。

Wangらは、Homerらとほぼ同様の設定において、ほとんどのGWAS論文で公開される検定統計量である p 値ですら、適切な保護がない場合には、無視できないのプライバシー上のリスクがあることを示した⁽⁴²⁾。Homerらの攻撃と同様、この攻撃では、攻撃者が、犠牲者のSNPプロファイルが与えられること、reference populationに関する分割表が与えられること、連鎖不平衡と呼ばれるSNP間の相関を表す r^2 値が与えられること、などの仮定を置いている。これらはやや強い仮定にも見えるが、reference populationに関する分割表や相関を表す r^2 値は、研究目的で論文や公開データベースにおいて公開されている情報であ

図 8 頻度分布からある SNP がケースグループに属するかどうかを推定。

る。Wang は、犠牲者について 200 程度の SNP プロファイルが取得できれば、犠牲者が GWAS データ中に存在するかどうか、高い信頼度において決定可能であるとの報告している。

4.3.2 連鎖不平衡を用いたゲノムベース属性推定

同一染色体上にある複数の遺伝子座における対立遺伝子は必ずしも独立ではなく相関を持つ現象を連鎖不平衡と呼ぶ。Marchini らは、完全な遺伝型の情報を持つ reference population の情報と連鎖不平衡から、欠損値を持つ遺伝情報について、その欠損値を隠れマルコフモデルなどの統計的推測により補完する手法を提案している⁽⁴³⁾。この手法はセンシティブリティの低い SNP 情報を使ってセンシティブリティの高い SNP 情報の推定を可能にすることから、ゲノムベース属性推定攻撃としても利用できる。DNA のらせん構造を解明したジェームス・ワトソン博士の遺伝子はアメリカ国立生命工学情報センター (NCBI) に登録されており閲覧可能であるが、本人の意思によりアルツハイマー病と関連の強い APOE 遺伝子の情報は公開されていない。Nyholt らは、連鎖不平衡を利用してワトソン博士の APOE 遺伝子の値を推測したことを報告している⁽⁴⁴⁾。

4.4 推論に対抗するプライバシー保護技術

これまでに個人ゲノムが持つプライバシーの問題、特に、個人ゲノムとそのゲノムの保有者を結び付けるゲノム識別推定と、個人ゲノムを介して個人とその個人のセンシティブ属性を結び付けるゲノムベース属性推定の問題を詳細に紹介してきた。

これらの推論のリスクを軽減することを目的としたプライバシー保護技術 (PETs: Privacy Enhancing Technologies) の要素技術には、データや統計量の公開を前提とする匿名性保障や差分プライバシーと、データの公開を前提としない秘密計算が知られている。

秘密計算では、そもそも秘密データを共有・公開するからゲノム識別推定やゲノムベース属性推定が発生するとの立場をとる。本来興味があるのはデータそのものではなく、最終的な解析結果のみであることを考慮し、データを外部公開せずに、結果のみを安全に解析者に与えることを目指す。具体的には、データが複数の組織に分散しており、これを互いに閲覧できない状態

図 9 秘密計算の概念図。

図 10 差分プライバシーの概念図。

にあるときに (e.g., 解析機関が個人ゲノム情報を持ち、病院が個人の臨床情報を持つが、これを互いに閲覧出来ない状況)、これらのデータを互いに共有することなく、データ解析 (e.g., ある疾患の発症に有意に影響する SNP を同定する) を実現する方法を議論する (図 9)。

一方、差分プライバシーでは、1 行が一人の個人に対応する表形式のデータに関する統計値の公開の問題 (e.g., GWAS における分割票の公開) を考える。具体的には、「このような統計値を閲覧した者が、その表にある個人のデータが含まれていたか否かを推測できない」ように公開する統計量をランダム化する方法を議論する (図 10)。

匿名性の保障では、1 行が一人の個人に対応する表形式のデータ (e.g., 性別、年齢、疾患、SNP) の公開の問題を考える。具体的には、「このような表形式のデータを閲覧した者がレコードと個人を一对一でひも付けられない」ように公開するデータを修正する。匿名性の保障は前者二つと同様に重要なプライバシー保護技術であるが、本稿ではの議論は、4.2.2 の範囲にとどめる。

秘密計算と差分プライバシーは本来互いに相反する技術ではないことに注意されたい。解析過程におけるプライバシー保護が秘密計算、解析結果の提供におけるプライバシー保護が差分プライバシーによって達成されると考えれば、両者を組み合わせて利用することも当然考えられる⁽⁴⁵⁾。次章以降では、それぞれの技術について、ゲノムデータの取扱いを例に、個別に紹介する。

5. 秘密計算による安全なゲノム解析

秘密計算とは複数のパーティーが互いに相手と共有すること

ができない秘密の入力を所持しているときに、秘密入力を互いに他者に開示することなく、これらの秘密データを入力に取る関数を評価し、その出力を入力を持ついずれかのパーティーのみ、あるいは第三者のみに報告する分散計算の総称である。個人ゲノムの主な解析用途は、ケース症例研究による疾患関連遺伝子の同定、個人ゲノムからの個別疾患の罹患リスクモデルの構築、個人ゲノムからの個別疾患の罹患リスク評価などである。

秘密計算の文脈においては、個人ゲノム、臨床情報、形質情報、デモグラフィック情報・非形質情報のプライバシー情報が、解析機関や医療機関など複数のパーティーに分散して保持されている状態において、秘密入力を互いに共有することなく、疾患関連遺伝子同定、罹患リスクモデル構築、個別疾患の罹患リスク評価等の個人ゲノムに関連した解析を実現することが目的となる。

秘密計算はこれまでに大きく分けて、秘匿回路評価、準同形性暗号系による暗号プロトコル、秘密分散に基づく秘密計算、の三種類の実現方式が知られており、それぞれの方式には計算効率性や計算モデルの自由度の観点において、長所と短所がある。本章ではまずこの三つの方式を簡単に導入する。またそれぞれの方式を用いた個人ゲノムに関する応用について、概説する。

以降では、二つのパーティー、Alice と Bob、間の秘密計算を扱う。Alice と Bob はそれぞれは秘密入力 x^A 及び x^B を持ち、これらを入力とする関数 f を評価する。その結果として、Alice と Bob はそれぞれ秘密出力 y^A 及び y^B を得るものとする。以降のこの秘密計算を $f(x^A, x^B) \rightarrow (y^A, y^B)$ と記述する。

5.1 秘密計算

秘密計算には大きく分けて三つの実現方法が知られる。以下、それぞれの実現方法を概説する。

5.1.1 秘匿回路評価による秘密計算

秘匿回路評価 (Yao's garbled circuit)⁽⁴⁶⁾ とはブーリアン回路の秘密計算を構成要素とした秘密計算の技法であり、1986 年に Yao によりその基礎が築かれた。秘匿回路評価は、紛失送信⁽⁴⁷⁾ が利用可能であることを前提として設計されている。紛失通信とは、送信者が複数項目からなるリストを保持し、受信者がそのリストのインデックスを保持しているときに、送信者に受信者が持つインデックスを与えることなく、受信者が送信者からリスト中のそのインデックスに対応する値を得るプロトコルである。

秘匿回路評価では、まず評価対象である関数 f をブーリアン回路で記述し、そのブーリアン回路の各ゲート (AND, OR, NOT) を、Alice 及び Bob 間で共有せずに評価することを一つの計算単位とし、これを回路全体に適用することで、秘密計算を実現する。1 ゲートの具体的な評価について簡単に説明する。まず 2 入力 (一つは Alice の入力、一つは Bob の入力) 1 出力のゲートの真理値表を用意する。AND ゲートの真理値表の例は以下のとおりである。

次に、Alice はゲートの入力 x, y の値 (0 or 1) に対応する暗

input x	input y	output z
0	0	0
0	1	0
1	0	0
1	1	1

号化鍵を $k_{0x}, k_{1x}, k_{0y}, k_{1y}$ を用意する。同様に、Alice はゲートの出力 z の値 (0 or 1) の対応する暗号化鍵を k_{0z}, k_{1z} を用意する。この鍵を使い、暗号化された真理値表を生成する。

input x	input y	output z	encrypted
0	0	0	$E_{k_{0x}}(E_{k_{0y}}(k_{0z}))$
0	1	0	$E_{k_{0x}}(E_{k_{1y}}(k_{0z}))$
1	0	0	$E_{k_{1x}}(E_{k_{0y}}(k_{0z}))$
1	1	1	$E_{k_{1x}}(E_{k_{1y}}(k_{1z}))$

続いて Alice は自分の入力 $a \in \{0, 1\}$ に対応する暗号化鍵 k_{ax} と、暗号化真理値表の行をランダムシャッフルして Bob に送信する。暗号化鍵自身はランダムに生成されるため、暗号化鍵から Alice の入力知られることはないことに注意されたい。Bob は紛失送信を用いて、Alice が持つ Bob の入力に対応する二つの暗号化鍵 k_{0y}, k_{1y} のうち、Bob の実際の入力 $b \in \{0, 1\}$ に対応する暗号化鍵を k_{by} を、Alice に b を明かさずに得る。これにより、Bob は暗号化鍵 k_{ax}, k_{by} を取得する。これを用いて、Bob は暗号化真理値表に含まれる四つの暗号文のうち、 $x = a, y = b$ に対応する暗号文 $E_{k_{ax}}(E_{k_{by}}(k_{cz}))$ を復号し、 k_{cz} を得る。ここで、 $c = a \wedge b$ である。これにより、Alice の入力 a と Bob の入力 b を互いに共有することなく、出力 $c = a \wedge b$ に対応する鍵 k_{cz} を Bob が得たことになる。これを各ゲートごとに繰り返すことで、Alice と Bob は入力を共有することなく、目的とする関数 f を模倣する回路を評価することができる。

5.1.2 準同形暗号系による秘密計算

公開鍵暗号の準同形性、すなわち暗号文同士の加算あるいは乗算ができる性質を利用しても、秘密計算を実現することができる。暗号系の秘密鍵及び公開鍵のペアを (pk, sk) とする。このとき、メッセージ $m \in \mathbb{Z}_N$ の暗号化は

$$c = E_{pk}(m) \quad (3)$$

その復号は

$$m = D_{sk}(m) \quad (4)$$

とすることで得られる。ここで、公開鍵暗号系が加法的準同形性を持つ場合、演算。

$$E_{pk}(m_1) \circ E_{pk}(m_2) = E_{pk}(m_1 + m_2) \quad (5)$$

を秘密鍵に関する知識なしに実行することができる。自然数 k について、演算 $\circ$ の k 回の繰返しを $E(\cdot)^k$ と表現すれば

$$E_{pk}(m)^k = E_{pk}(km) \quad (6)$$

を得る。式 (5) は、秘密鍵の知識なしに暗号化された値同士の加

算を、式(6)は、秘密鍵の知識なしに値と暗号化された値との乗算が実現可能であることを示している。秘密計算の入力には比較的限定されたドメインの値が入力されるため、頻度攻撃^(注3)の対象となりやすい。このため、秘密計算に利用する準同形性暗号には、強秘匿性 (Semantic Security)^(注4)が要求される。この暗号の準同形性を利用した秘密計算を実現するプロトコルは、ゲノムに関する計算にかかわらず、様々なデータ解析のアルゴリズムとして利用されている⁽⁴⁸⁾⁻⁽⁵⁰⁾。

準同形暗号系を用いた具体的な秘密計算の例として、内積計算 $z = x \cdot y$ の秘密計算 $f(x, y) \rightarrow (x \cdot y, 0)$ を考える。ここで、 $x, y \in \mathbb{Z}^D$, $z \in \mathbb{Z}$ である。まず、Alice は鍵ペアを生成し、ベクトルの各要素の値を暗号化し、公開鍵 pk とともに $E_{pk}(x_i)$ を Bob に送る。Bob は受信した $(E_{pk}(x_1), E_{pk}(x_2), \dots, E_{pk}(x_D))$ 及び自身の秘密入力 y を用いて以下を計算する：

$$\begin{aligned} c &= \left(\prod_{i=1}^D E_{pk}(x_i)^{y_i} \right) = \left(\prod_{i=1}^D E_{pk}(x_i y_i) \right) \\ &= E_{pk}(\sum_{i=1}^D x_i y_i) = E_{pk}(x \cdot y). \end{aligned}$$

Bob は暗号文 c を Alice に送信し、Alice はこれを復号することで $D_{sk}(E_{pk}(x \cdot y)) = x \cdot y$ を得る。

加法的準同形性を持つ強秘匿な公開鍵暗号系は比較的限られており、よく利用されているのは Paillier 暗号⁽⁵¹⁾である。

2009 年には Gentry によって加算と乗算に関する準同形性を同時に有する暗号 (完全準同形暗号) が発表された⁽⁵²⁾。加法準同形暗号は加法についてのみ同形であり、これにより実現可能なデータ解析は限られていたが、完全準同形暗号によれば事実上任意の関数に対する 1 ラウンドの秘密計算が実現可能になる。ただし、完全準同形暗号は計算時間や空間計算量が膨大であることが知られており、その削減の努力が続いているが、ゲノムデータのような大規模データ解析に利用可能になるには、もう一段のブレークスルーが必要である。

5.1.3 秘密分散に基づく秘密計算

暗号系を用いない秘密計算手法として、秘密分散に基づく秘密計算が知られている。この方式の基礎となる秘密分散では、秘密情報を $p \geq 2$ 個のシェアと呼ばれる情報の断片に分散させる。このシェアは以下のような性質を持つ。

- $2 \leq \theta \leq p$ なる θ について、 θ 個未満のシェアを持つパーティーは、シェアから秘密情報について何の情報も得ることができない。
- しかし、 θ 個以上のシェアを持つパーティーは、それらから秘密情報を復元できる。

$p = 3, \theta = 2$ における秘密分散は、ランダムシェアと呼ばれる極めて単純な方法で実現可能である。秘密情報を $x \in \mathbb{Z}_N^{(注5)}$ とする。このとき、 $x = r_1 + r_2 + r_3 \pmod{N}$ となるように、ランダムに $r_1, r_2 \in \mathbb{Z}_N$ を決める。 $r_3 \in \mathbb{Z}_N$ は r_1, r_2 が決まれば一意に決まる。このとき、 $(r_1, r_2), (r_2, r_3)$, 及び (r_3, r_1) のそれぞれは $\mathbb{Z}_N$ 上の一様乱数と区別がつかない。つまり三つのうち二つのシェアを

持つパーティーは秘密情報 x について何の知識も得ることができない。一方、三つのシェアを得たパーティーは $x = r_1 + r_2 + r_3 \pmod{N}$ より秘密情報を復元可能である。

秘密分散に基づく秘密計算とは、複数の秘密情報をランダムシェアとして複数のパーティーに分散させたときに、ランダムシェア同士の演算がある種の代数演算を実現することに着目した手法である。具体的な実現手段や対応クラスとそのバリエーションは文献(53)や文献(54)などを参照して頂くことにし、ここでは、最も単純な、二つの秘密入力の加算の実行例を示す。

データ所有者 A が持つ秘密入力を x 、データ所有者 B が持つ秘密入力を y とし、その加算 $x + y$ の秘密計算を考える。そのそれぞれのシェアを $x = r_1 + r_2 + r_3, y = s_1 + s_2 + s_3$ とする。三つの独立したデータ収集者 C, D, E の存在を仮定し、C には r_1, s_1 を、D には r_2, s_2 を、E には r_3, s_3 を、それぞれ与える。データ収集者 C, D, 及び E はそれぞれ $t_1 = r_1 + s_1, t_2 = r_2 + s_2$, 及び $t_3 = r_3 + s_3$ を計算し、データ解析者 F に、 t_1, t_2, t_3 を与える。ここで、 t_1, t_2, t_3 それぞれは二つのランダムシェアの和にすぎないため、個々の秘密入力に関する知識を与えないが、これらの総和を計算することで望む出力 $t_1 + t_2 + t_3 = r_1 + s_1 + r_2 + s_2 + r_3 + s_3 = x + y$ を得ることが可能である。

この方式は、暗号演算や暗号プロトコルの実行を含まないため、一般に計算効率的である。ただし実現可能な関数のクラスは限定される。また、秘密分散を基礎とする方式であることから、3 以上の互いに独立なデータ収集者の存在を仮定する必要がある。このことは、秘密計算をシステムとして実装した場合には、安全性が保証された多数の計算資源の運用が要求されることを意味している。

5.2 疾患遺伝子探索におけるプライバシー保護

疾患遺伝子探索は 2.4 で紹介したとおり、多数の個人の遺伝情報とその臨床情報に基づく統計的検定により実現される。遺伝子情報や臨床情報が複数の機関に分散して蓄積されている状況において、これらを互いに共有することなく統計的検定に基づき疾患遺伝子探索を実現するシナリオは、秘密計算の問題定義と明確に合致しており、幾つかの研究が報告されている。

5.3 秘密計算と疾患遺伝子探索と疾患リスクモデル構築

Kamm らは多人数から収集した遺伝子情報と臨床情報を秘密計算によって表 1 で表現される分割表に統合し、ケースコントロールスタディにおいて、カイ二乗検定統計量を求める手法を導入している⁽⁵⁵⁾。疾患遺伝子探索のプロセスは分割表の構築と統計的検定二つの部分計算から成る。Kamm らはこれらをそれぞれ実現する秘密分散ベースの秘密計算アルゴリズムを sharemind と呼ばれる秘密計算フレームワーク⁽⁵³⁾で実現している。また、23andMe に代表されるような一つの組織がドナーから遺伝子の提供を募るモデルや、複数の機関が協調してケースコントロールスタディを行うコンソーシアムモデルなど、複数のモデルに

(注3)：暗号文の出現頻度から平文を推測する攻撃。

(注4)：暗号文から平文のどのような部分情報も得ることができないこと。

(注5)： $\mathbb{Z}_N = \{0, 1, \dots, N-1\}$ 。

おける応用シナリオも合わせて議論している。

Kamm らの方法では、分割表の各セルのカウントを、ランダムシェアとして分散保持する。ケース/コントロール及びターゲットとする SNP の遺伝型を表す指示変数が与えられることを前提とすれば、これは比較的簡単な秘密計算で準備することができる。解析対象である SNP が、目標疾患に有意に関連しているかどうかは、 p 値の評価に基づく。 p 値の評価は浮動小数点演算を含み、sharemind 上での実現には適さないため、Kamm らの実装では、有意性の統計的検定が $\mathbb{Z}_N$ 上の比較演算によって実現されるよう、評価式の変換を行っている。26 万 SNP を対象とした秘密計算が 10 分程度で終了するとの報告がある⁽⁵⁵⁾。

筆者らは、検査情報 (e.g., ピロリ菌の感染状態) と、臨床情報 (e.g., 胃がんの罹患歴) が、複数の医療機関に分散して独立に保持されている状況において、検査情報が臨床情報に与える影響を、準同形性暗号に基づくカイ二乗検定の秘密計算により評価する手法を提案している⁽⁵⁶⁾。一般に、複数機関によって独立に収集された情報には、両者を直接照合できる共通のアイデンティティが付与されていないことが多い。一般には患者の氏名や住所等、デモグラフィック情報など組合せによって一意識別性を持つ情報を用いてレコードの“名寄せ”を行うが、対象が秘密情報の場合、この名寄せ自体にプライバシー上問題があるケースがある。文献(56)では、この名寄せのプロセスと分割表の構築を秘密計算によって同時に解決する手法を提案している。シミュレーションでは 300 万レコードの名寄せに約 223 時間を要するとの結論を得ている。

また筆者らは、秘密計算により学習したロジスティック回帰モデルから、遺伝的要因が特定の疾患に与える影響を評価し、疾患リスク評価モデルを構築する手法を開発している⁽⁵⁷⁾。疾患リスクは、遺伝的要因のみならず、年齢や性別、生活習慣など複数の要因を複合的に考慮して評価する必要があるが、カイ二乗検定は、複数の要因を考慮したリスク評価モデルの構築を意図した検定法ではない。筆者らは遺伝的要因と環境的要因の両者を考慮して学習したロジスティック回帰モデルから、遺伝的要因が疾患の発症に与えるリスクを評価することで、より精密な疾患リスク評価モデルを構築する方法を構築している。筆者らの実験では、慢性腎臓病のゲノムコホート研究のために約 5,000 人の被験者から収集した 221 の SNP データ及び臨床データ⁽⁵⁸⁾を対象として、秘密計算により疾患遺伝子探索及び疾患リスクモデル構築を行った結果、疾患に有意に関連する 27 個の SNP を同定し、その 27 個の SNP と二つの環境的要因を用いて疾患リスクモデルを構築することに成功している。

秘密分散に基づく秘密計算は、ランダム化により安全性保証するため、Yao's garbled circuit や暗号関数を利用した秘密計算よりも計算効率性が高いことが利点である。ただし、秘密情報を入力する者と、秘密計算の結果を受け取る者の他に、互いに共謀しないことが保証されている 3 つのパーティーが計算に参加することが前提となる。そのため、計算に参加するステークホルダーの構成が実際の応用シナリオと合致するか検討する必要がある。具体的には、運営母体が異なる三つのクラウド計算環境など、独立した計算主体を利用する構成が考えられる。

準同形性暗号に基づく秘密計算は、特にパーティー数等の制約はないため、応用シナリオに登場するステークホルダーを直接計算主体として計算環境を構築できる点が利点である。ただ、多数回の暗号プリミティブの評価を含むため、大規模データにおいては計算時間が障害となる。文献(56)のシミュレーション結果は必ずしも高速とは言えないが、このようなコホート研究自体が数年の期間と膨大なコストを費やして実施され、統計的検定は、その長期にわたる研究の最終プロセスにすぎないこと、また統計的検定は仮説検証プロセスにおいて一度のみ実施される計算であることを考慮すれば、疫学上の応用においては、計算時間の長さは大きな障害にはならないともいえる。

5.4 秘密計算と疾患リスク評価

Ayday らは以下の情報のフローに基づく、疾患リスクモデルに対するプライバシーを保護した疾患リスク評価の枠組みを提案している⁽⁵⁹⁾⁽⁶⁰⁾。

- 患者 (P) : CI にゲノムサンプルを提供する, MU と SPU に臨床, 環境データを提供する
- Certified institution (CI) : ゲノムサンプルを解析し, 得たゲノムデータを暗号化し SPU に提供する
- Storage and Processing Unit (SPU) : ゲノムデータと臨床・環境データを暗号化された状態で保持し, 必要に応じて MU に提供する
- Medical unit (MU) : SPU からデータを受け取り, 患者の疾患リスクを評価する

MU はいわゆる一般のクリニックであり、運用上、高いセキュリティレベルを維持することは困難であり、従業者による故意あるいは意図的な情報漏えいが脅威として想定されている。SPU はいわゆるクラウドストレージを想定すればよく、MU よりは高いセキュリティレベルを想定できるが、提供された情報が内部的に閲覧されることは避けられない。これらを考慮し、SPU は semi-honest であること、MU と SPU は互いに共謀しないことなどを仮定したプロトコル設計が行われている。

Ayday らのシステムにおいては、データは準同形性暗号により暗号化されており、疾患リスクは MU が暗号化されたゲノムデータや環境・臨床データをダウンロードした後、これらを暗号化したまま準同形性を利用しリスクを評価することで、MU が正しく、かつゲノムデータを閲覧することなく疾患リスクを評価することを可能にしている。

5.5 秘密計算によるゲノム文字列解析

ゲノム解析においては、ゲノムの類似度評価、類似検索、共通文字列のカウンティング等様々なゲノム解析を文字列の処理とみなして計算する場合がある。これらは研究上のゲノム解析のほか、実父確定検査 (paternity test) や出産前診断等の遺伝上の適合性検査 (Generic compatibilizing test)、個別疾患の罹患リスクの有無の判定など医療上の検査のためにも利用される。個人のゲノムが解析対象である場合には、プライバシー上のリスクが存在

することから、様々な文字列における計算を秘密計算で実現する手法が研究されている。

有限状態機械は正規表現や類似検索を含む多様な文字列解析をモデル化可能であることから、汎用的な文字列解析手法として利用される。秘密計算における有限状態機械のシミュレーションは紛失決定性有限オートマトン評価 (oblivious deterministic finite automaton evaluation) と呼ばれ、複数の研究例がある。

Troncoso らは、以下のクエリ評価の問題を考察している⁽⁸⁴⁾。被験者は本人のゲノム文字列を保持し、ヘルスケアサービスプロバイダがそのゲノムについてクエリ (例えばある特定の疾患に罹患するリスクを持つか否か) を発行する。ここで、被験者はプライバシー上の理由からゲノムを公開することを望まず、ヘルスケアサービスプロバイダは商業的な理由からそのクエリを秘密にしておきたいが、被験者はクエリ処理を許可したい。Troncoso らは、紛失送信を要素技術として用い、任意の決定性有限オートマトン (DFA: Definite Finite Automaton) について、2 パーティー間の紛失 DFA 評価アルゴリズムを示し、DFA の状態数について線形時間で紛失 DFA 評価が可能であることを示した。同様の問題設定で、Frikken らは Yao's garbled circuit を要素技術とし、ラウンド計算量を削減した紛失 DFA 評価アルゴリズムを提案した⁽⁶¹⁾。紛失 DFA 評価の第三者への委託計算 (outsourcing) を可能にするアルゴリズムは⁽⁶²⁾ ⁽⁶³⁾などで示されている。紛失 DFA 評価における、パーティー間の結託や悪意のある攻撃者などについては、⁽⁶⁴⁾ ⁽⁶⁵⁾にて議論されている。

これらの紛失 DFA 評価アルゴリズムは原則として DFA の評価を前提とするため、非決定性有限オートマトン (NFA, non-deterministic finite automaton) の評価においては、これを DFA に変換する必要がある。NFA から DFA への変換は指数時間の計算を必要とする場合があり、また変換によって生成された DFA は、NFA の状態数に対して指数時間の状態数を持つことがある⁽⁶⁶⁾。筆者らのグループでは、NFA を直接評価する紛失 NFA 評価アルゴリズム ONE を提案している⁽⁶⁷⁾。ONE は Troncoso らと同様の問題設定において、任意の NFA を、NFA の状態数について多項式時間・空間計算量で評価することが可能である。また ONE をベースとしたアルゴリズムは、NFA 表現された正規表現について、多項式時間・空間計算量で紛失 NFA 評価を実現している⁽⁶⁷⁾。また同様に NFA 表現された類似文字列検索について、線形時間・空間計算量で紛失 NFA 評価を実現している⁽⁶⁸⁾。

特定の文字列操作における秘密計算アルゴリズムも数多く提案されている。Smith-Waterman アルゴリズム⁽⁶⁹⁾は塩基配列やアミノ酸配列をはじめとする文字列の相同性の検出のために広く用いられているアルゴリズムである。Jha らは Yao's garbled circuit を要素技術とした文字列の編集距離及びこれを用いた Smith-Waterman アルゴリズムに基づく文字列のアラインメントを秘密計算として実現する手法を提案している⁽⁷⁰⁾。ハミング距離計算の秘密計算は文献⁽⁷¹⁾で報告されている。文字列マッチングを対象とした線形時間の秘密計算には、文献⁽⁶⁴⁾ ⁽⁷²⁾などが知られる。ワイルドカードや近似文字列マッチングの秘密計算は文献⁽⁷³⁾で実現されている。最長共通文字列発見の秘密計算は文献⁽⁷⁴⁾ ⁽⁷⁵⁾で提案されている。Baldi らは共通集合プロト

コル⁽⁷⁶⁾ ⁽⁷⁷⁾要素技術とした、実父確定検査、出産前診断等の遺伝上の適合性検査⁽⁷⁸⁾のための秘密計算を提案している。

6. 差分プライバシーによる安全なゲノム解析

ゲノム疫学においては、カイ二乗検定などの統計的検定によって、個別の SNP と疾患の関連の強さを p 値により定量化し、疾患関連遺伝子を同定する。カイ二乗検定においては、必ずしも個別のゲノム情報や臨床情報は必要せず、被験者が罹患しているか否か (症例/非症例)、及び、着目する SNP 座位においてある対立遺伝子を保有するか否かをカウントした分割表から、疾患と SNP の独立性を検定する。分割表の例は表 1 を参照されたい。

直感的には、分割表や p 値といったようなサマリー統計量から個人に関する情報は推測されないと考えがちである。しかし、GWAS では、サンプル数 (被験者数) n = 数千数万程度に対し、属性数 (SNP 数) = 数十万個を調査する。4.3.1 にて既に紹介したように、このような極めて高次元の標本については、たとえサマリー統計量からでも、個別のサンプルについての特定の情報が高い確率で推定可能な場合がある。

このような統計量から、個別のサンプルの情報に関する確率的推論が制限されることを理論的に保証するプライバシー保護の枠組みとして、差分プライバシー (differential privacy) が注目を集めている⁽⁷⁹⁾。本章では、ゲノムワイド関連解析における統計量開示における差分プライバシーの達成を議論する。

6.1 差分プライバシー

差分プライバシーとは、データベース (この場合は SNP 及び症例/非症例の組合せからなるレコードの集合) がある特定のクエリに回答し値を開示したときに (この場合は、対立遺伝子のカウント数、独立性検定に関わる統計量、 p 値、連鎖不平衡等の統計クエリへの回答)、その応答値により個別のサンプルに関する情報が推測される確率が様にバウンドされることを保証することに基づくプライバシー定義である。

ここでは厳密さを若干犠牲にしつつ、疾患遺伝子探索を例に、直観的な説明を目指す。ゲノム研究者は、ある SNP 座位の対立遺伝子 (0/1) 及び非症例 (0)/症例 (1) の組合せから成るデータを収集し、分割表 1 を得たとする。話を簡単にするために、 p 値や連鎖不平衡の開示は考えず、この分割表の開示に起因するプライバシー上のリスクについて考えることにする。

この分割表をダウンロードした攻撃者は、取得した統計情報と自身の背景知識を組合せて犠牲者 (データベース内の個別のレコード) が含む情報を推測するとしよう。攻撃者がどのような背景知識を保持しているかをあらかじめ定めることは通常困難である。差分プライバシーでは、任意の背景知識を有する攻撃者にとってさえも、データベースが保持する情報の推測がある一定の基準の下で制限されることを保証することをもってプライバシー保護が達成されるとの立場を取る。では何が達成されれば、ある一定のプライバシー保護が達成されたと言えるだろうか。

上記の例では、 f は対立遺伝子 $AA/Aa(0)$ or $aa(1)$ を持つ非症例 (0) or 症例 (1) のサンプル数を求めるカウント関数に相当することから、対応するクエリ応答を得ることで分割表を構成できる。標本集合 D に対するクエリ f の応答を $r = f(D)$ と書くことにする。ここで、プライバシーの定義のために、 n 人の個人の情報が含まれるデータベース $D = \{(i, x_i, y_i)\}$ に対して、これに類似した仮想的なデータベース $D' = \{(i', x'_i, y'_i)\}$ を考える。 D と D' は、任意の一人の個人を除いた全ての個人については同じ情報を持つデータベースである。ここで、一般性を失うことなく、 $i \neq N$ において $x_i = x'_i, y_i = y'_i$ であるとしてよい。

攻撃者が開示された統計量から情報を得ることを防ぐために、応答値に何らかの変更を施して公開することを考える。応答 $f(D)$ は実際には $M(f(D))$ として外部に公開されるとしよう。ここでプライバシーメカニズム M は、例えば統計クエリに応答を乱数でランダム化するような、プライバシー保護のための写像である。

このとき、 A さんのデータがデータベース D に含まれていようが、いなかろうが、出力される統計値が大して変化しないのであれば、統計値を開示すること自体は A さんのプライバシーを侵害しない、と考えよう、というのが差分プライバシーのアイデアの直観的な説明である。逆に言えば、 A さんのデータがデータベース D に含まれているかいないかが、統計値からは高い確信度を持って推定できない程度に、統計値をプライバシーメカニズム M によって、ランダム化しよう、というのが差分プライバシーに基づくプライバシー保護措置であると言える。ここで、この「 A さんが含まれていようがいなかろうが」という考え方は、仮想的なデータベース D' を考えれば、「問合せ対象が D であろうが D' であろうが」と置き換えることができる。

厳密には、プライバシーメカニズム M が確率的アルゴリズムであるとき、 D に対するクエリ f の応答が r である確率を $Pr(r = M(f(D)))$ 、 D' に対するクエリ f の応答が r である確率を $Pr(r = M(f(D')))$ とする。このとき、任意の応答値 r 、任意の隣接するデータベースの組 (D, D') 、及び $\epsilon > 0$ について

$$\ln \left(\frac{Pr(r = M(f(D)))}{Pr(r = M(f(D')))} \right) < \epsilon \quad (7)$$

であれば、クエリ f は ϵ -差分プライバシーを満たす、と定義される⁽⁷⁹⁾。

定義では任意の隣接するデータベースについて上記のバウンズの達成を要求するため、具体的な A さん、でなく、それが誰でもあっても、その人がいようがいなかろうが統計値に影響しないことが必要である。そしてこの「大して」という概念は、パラメータ ϵ を用いて、定量化されている。

数学的な意味付けは省略するが、この差分プライバシーを達成するユニバーサルな手法として、ラプラスメカニズムが知られてる。ラプラスメカニズムは、応答値にラプラス分布から得たランダム雑音を加え、応答値をランダム化する。このメカニズムが加えるラプラス雑音のスケールは、ある個人がデータベースに含まれている場合とない場合の、応答値の最大変化量(敏感度)に比例するように定める必要がある。雑音のスケールを敏感度に応じてこのように設定することで、ラプラスメカニズムは差

分プライバシーを満足する統計量を生成することが証明されている⁽⁷⁹⁾。

カウントを応答するクエリの場合には、ある個人がデータベースに含まれている場合とない場合の、応答値の最大変化量は 1 である。このため、この場合ラプラス雑音のスケールは、評価対象となる人数に関わらず定数としてよい。このことは、対象となる人数が多い場合には、加算される雑音の分散は相対的に小さくなる。これはサンプル数が多い場合には加える雑音の分散は少なくともよいであろうという直観ともよく一致する。

6.2 GWAS 関連統計量における差分プライバシー

ラプラス雑音によるランダム化に基づき、Fienberg らは差分プライバシーを保護した χ^2 検定による疾患関連遺伝子探索の手法を示している⁽⁸⁰⁾。彼らの手法はユニバーサルなラプラスメカニズムに基づいており、マイナーな対立遺伝子のカウント、 χ^2 統計量及び p 値の敏感度を求めた上で、これらの統計量が差分プライバシーを満足するようランダム化した上で公開する手法である。実験では $M = 3$ の SNP 中に二つの疾患関連 SNP が存在することを想定し、HAP-SAMPLE シミュレーション⁽⁸¹⁾を用いて複数の標本数 N についてデータを生成している。またこの設定において、差分プライバシーが保障された検定統計量から、正しく疾患関連 SNP を発見するために必要な標本数について議論している。この仮想的な実験では、差分プライバシーのパラメータを $\epsilon = 0.4$ としたときには 10^4 オーダーのサンプル数が必要であると結論付けている。

複数の統計量を同時に開示する場合に、それらの統計量が互いに独立であれば、各統計量に独立にラプラスメカニズムを適用すればよい (Parallel composition)。Fienberg らの手法は、SNP 数を 3 と設定しているため、ナイーブなラプラスメカニズムにおいても、現実的なサンプル数で実用的な統計量を公開できると結論される。

しかし GWAS に関連する大規模な分割表の扱いはより困難である。GWAS では多くの場合数万から 10 数万の SNP を解析するが、その分割表に含まれるカウントは、連鎖不平衡のために独立ではない。このような場合には、互いに相関する統計量の数 C について、それぞれが差分プライバシーを満足するためには、雑音のスケールを C 倍する必要があることが知られているためである (Sequential composition, 文献(82))。

Johnson らは、ゲノムワイド関連解析において、このような一般の差分プライバシーのためのメカニズムを適用しようとした場合、サンプル数に比べて出力すべき統計量の数が多すぎ、統計量の有用性を維持することは困難であると主張している⁽⁸³⁾。多くの典型的なゲノムワイド関連解析研究では、SNP と疾患間の関連だけでなく、 $O(M^2)$ 個の SNP 間の相関についても調べる。この場合、ラプラスメカニズム等のユニバーサルなメカニズムを用いた場合には、 $O(M^2/N)$ のスケールの雑音を加える必要があり、典型的な GWAS ではサンプル数は $N = 10^3$ のオーダーであるのに対し SNP 数は $M = 10^5$ であることから、標準的なプライバシーメカニズムの適用はほとんど役に立たないと結論付けて

いる。

そこで Johnson らは、ゲノムワイド関連解析において興味があるのは疾患と関連の強いごく少数の SNP に関する統計量・検定量であり、それ以外の大多数の SNP については関連がないということだけを知ることができれば、それ以上の情報を得る必要がないことに着目した手法を提案している。具体的には、Johnson らの枠組みでは、疾患と有意に関連している SNP の数、最も有意に関連している SNP の位置、指定された SNP と疾患の p 値、SNP 間の相関のブロック構造のクエリ応答について差分プライバシーを保証する distance-score メカニズムを導入している。疾患遺伝子探索において、必要最小限の数のクエリ応答値から興味ある SNP を絞り込むことで、統計量・検定量に加えるべき雑音の対象 SNP 数 M に比例しない枠組みを提供している⁽⁸³⁾。

7. おわりに

この解説論文では「究極の個人情報」呼ばれる個人ゲノムを対象として、現時点で認識されている個人情報の公開が引き起こすリスクと脅威を概観した。またそのゲノム解析を対象としたプライバシー保護技術の中でも、特に秘密計算と差分プライバシーについて、疫学応用及び医療応用を中心として解説した。大規模な個人ゲノムの活用はまだ歴史が浅いが、自らのゲノムを解析し、それを自らの医療のために活用する日は近い将来に必ずやってくる。個人ゲノム利用とその保護のバランスを適切に制御するための技術の確立はまだ道半ばであるが、筆者らを含む幾つかのグループは研究開発を活性化させており、今後の技術発展が期待される。

謝辞 本研究は、JST CREST「ビッグデータ統合利活用のための次世代基盤技術の創出・体系化」領域におけるプロジェクト「自己情報コントロール機構を持つプライバシー保護データ収集・解析基盤の構築と個別化医療・ゲノム疫学への展開」の助成を受けた。本稿の執筆にあたり、重要な指摘と助言を頂いた竹内一郎氏、津田宏治氏、並びに関係する皆様に感謝致します。

文 献

- (1) National Human Genome Research Institute. DNA Sequencing Cost, <http://www.genome.gov/sequencingcosts/>
- (2) R. H. Duerr, K. D. Taylor, S. R. Brant, J. D. Rioux, M. S. Silverberg, M. J. Daly, A. H. Steinhardt, C. Abraham, M. Regueiro, A. Griffiths, et al., "A genome-wide association study identifies IL23R as an inflammatory bowel disease gene," *Science*, vol.314, no.5804, pp.1461–1463, 2006.
- (3) D. J. Hunter, P. Kraft, K. B. Jacobs, D. G. Cox, M. Yeager, S. E. Hankinson, S. Wacholder, Z. Wang, R. Welch, A. Hutchinson, et al., "A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer," *Nature genetics*, vol.39, no.7, pp.870–874, 2007.
- (4) L. J. Scott, K. L. Mohlke, L. L. Bonnycastle, C. J. Willer, Y. Li, W. L. Duren, M. R. Erdos, H. M. Stringham, P. S. Chines, A. U. Jackson, et al., "A genome-wide association study of type 2 diabetes in finns detects multiple susceptibility variants," *Science*, vol.316, no.5829, pp.1341–1345, 2007.
- (5) R. Sladek, G. Rocheleau, J. Rung, C. Dina, L. Shen, D. Serre, P. Boutin, D. Vincent, A. Belisle, S. Hadjadj, et al., "A genome-wide association study identifies novel risk loci for type 2 diabetes," *Nature*, vol.445, no.7130, pp.881–885, 2007.
- (6) D. Hall, "Single-nucleotide polymorphism," http://en.wikipedia.org/wiki/Single-nucleotide_polymorphism
- (7) D. W. Hosmer Jr and S. Lemeshow, *Applied logistic regression*, John Wiley & Sons, 2004.
- (8) Personal Genome Project, <http://www.personalgenomes.org/>
- (9) P. R. Billings, M. A. Kohn, M. D. Cuevas, J. Beckwith, J. S. Alper, and M. R. Natowicz, "Discrimination as a consequence of genetic testing," *American journal of human genetics*, vol.50, no.3, p.476, 1992.
- (10) L. N. Geller, J. S. Alper, P. R. Billings, C. I. Barash, J. Beckwith, and M. R. Natowicz, "Individual, family, and societal dimensions of genetic discrimination: a case study analysis," *Science and Engineering Ethics*, vol.2, no.1, pp.71–88, 1996.
- (11) H.R. 493 (110th): Genetic Information Nondiscrimination Act of 2008, <https://www.govtrack.us/congress/bills/110/hr493/text>
- (12) openSNP, <https://opensnp.org/>
- (13) A. J. Pakstis, W. C. Speed, R. Fang, F. C. Hyland, M. R. Furtado, J. R. Kidd, and K. K. Kidd, "SNPs for a universal individual identification panel," *Human genetics*, vol.127, no.3, pp.315–324, 2010.
- (14) Z. Lin, A. B. Owen, and R. B. Altman, "Genomic research and human subject privacy," *SCIENCE-NEW YORK THEN WASHINGTON-*, pp. 183–183, 2004.
- (15) Facebook, <https://www.facebook.com/>
- (16) PeopleFinder, <http://www.peoplefinder.com/>
- (17) L. Sweeney, "Simple demographics often identify people uniquely," *Health (San Francisco)*, pp.1–34, 2000.
- (18) L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol.10, no.5, pp.557–570, 2002.
- (19) U.S. Department of Health and Human Services, "Guidance regarding methods for de-identification of protected health information in accordance with the health insurance portability and accountability act (hipaa) privacy rule," <http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html>
- (20) C. C. Aggarwal and S. Yu. Philip, *A general survey of privacy-preserving data mining models and algorithms*, Springer, 2008.
- (21) B. Zhou, J. Pei, and W. Luk, "A brief survey on anonymization techniques for privacy preserving publishing of social network data," *ACM SIGKDD Explorations Newsletter*, vol.10, no.2, pp.12–22, 2008.
- (22) C. D. Huff, D. J. Witherspoon, T. S. Simonson, J. Xing, W. S. Watkins, Y. Zhang, T. M. Tuohy, D. W. Neklason, R. W. Burt, S. L. Guthery, et al., "Maximum-likelihood estimation of recent shared ancestry (ersa)," *Genome research*, vol.21, no.5, pp.768–774, 2011.
- (23) Discover DNA ancestry and genetic history with our DNA test - 23andMe, <https://www.23andme.com/ancestry/refinder/>
- (24) R. L. Bennett, K. A. Steinhaus, S. B. Uhrich, C. K. O'Sullivan, R. G. Resta, D. L.-Doyle, D. S. Markel, V. Vincent, and J. Hamanishi, "Recommendations for standardized human pedigree nomenclature," *Journal of Genetic Counseling*, vol.4, no.4, pp.267–279, 1995.
- (25) B. Malin, "Re-identification of familial database records," *In AMIA Annual Symposium Proceedings*, vol. 2006, p.524, American Medical Informatics Association, 2006.
- (26) Israel vs. S. B. Bilik, A. Adam, Y. Vitman, and H. Aharon, "Moche moshkowitz and meir liver (in Hebrew)," *Verdict* 2442-05-12.
- (27) Geni, <http://www.geni.com/>
- (28) Y. Erlich and A. Narayanan, Routes for breaching and protecting genetic privacy, *arXiv preprint arXiv:1310.3197*, 2013.
- (29) J. Gitschier, "Inferential genotyping of Y chromosomes in latter-day saints founders and comparison to Utah samples in the HapMap project," *The American Journal of Human Genetics*, vol.84, no.2, pp.251–258, 2009.
- (30) M. Gymrek, A. L. McGuire, D. Golan, E. Halperin, and Y. Erlich, "Identifying personal genomes by surname inference," *Science*, vol.339, no.6117, pp.321–324, 2013.
- (31) Family Tree DNA, <http://www.familytreedna.com/>
- (32) G. Naik, "Family secret: An adopted man's 26-year quest for his father," *Wall Street Journal*.
- (33) Lehman-Haupt R, "Are sperm donors really anonymous anymore?,"

Slate.

- (34) D. M. Lambert and L. Huynen, " Evolutionary biology: Face of the past reconstructed, " *Nature*, vol.463, no.7282, pp.739–740, 2010.
- (35) N. Homer, S. Szelling, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig, " Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays, " *PLoS genetics*, vol.4, no.8, e1000167, 2008.
- (36) International HapMap Project, <http://hapmap.ncbi.nlm.nih.gov/>
- (37) " NIH Background Fact Sheet on GWAS Policy Update, " 2008.
- (38) E. A. Zerhouni and E. G. Nabel, " Protecting aggregate genomic data, " *Science*, vol.322, no.5898, p.44a, 2008.
- (39) E. Halperin and D. A. Stephan, " SNP imputation in association studies, " *Nature biotechnology*, vol.27, no.4, pp.349–351, 2009.
- (40) K. B. Jacobs, M. Yeager, S. Wacholder, D. Craig, P. Kraft, D. J. Hunter, J. Paschal, T. A. Manolio, M. Tucker, R. N. Hoover, G. D. Thomas, S. J. Chanock and N. Chatterjee, " A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies, " *Nature genetics*, vol.41, no.11, pp.1253–1257, 2009.
- (41) P. M. Visscher and W. G. Hill, " The limits of individual identification from sample allele frequencies: Theory and statistical analysis, " *PLoS genetics*, vol.5, no.10, e1000628, 2009.
- (42) R. Wang, Y. F. Li, X. Wang, H. Tang, and X. Zhou, " Learning your Identity and disease from research papers: Information leaks in genome wide association study, " In *Proceedings of the 16th ACM conference on Computer and communications security*, pp.534–544, ACM, 2009.
- (43) J. Marchini and B. Howie, " Genotype imputation for genome-wide association studies, " *Nature Reviews Genetics*, vol.11, no.7, pp.499–511, 2010.
- (44) D. R. Nyholt, C.-E. Yu, and P. M. Visscher, " On Jim Watson's apoe status: Genetic information is hard to hide, " *European Journal of Human Genetics*, vol.17, no.2, pp.147–149, 2008.
- (45) A. McGregor, I. Mironov, T. Pitassi, O. Reingold, K. Talwar, and S. Vadhan, " The limits of two-party differential privacy, " In *51st Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pp.81–90, IEEE, 2010.
- (46) A. C.-C. Yao, " How to generate and exchange secrets, " In *27th Annual Symposium on nFoundations of Computer Science*, 1986., pp.162–167, IEEE, 1986.
- (47) M. O. Rabin, " How to exchange secrets with oblivious transfer, " *IACR Cryptology ePrint Archive*, 2005:187, 2005.
- (48) Y. Lindell and B. Pinkas, " Privacy preserving data mining, " In *Advances in Cryptology CRYPTO 2000*, pp.36–54, Springer, 2000.
- (49) J. Sakuma and S. Kobayashi, " Link analysis for private weighted graphs, " In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pp. 235–242, ACM, 2009.
- (50) J. Sakuma and S. Kobayashi, " Large-scale k-means clustering with user-centric privacy-preservation, " *Knowl. inf. syst.*, vol.25, no.2, pp.253–279, 2010.
- (51) P. Paillier, " Public-key cryptosystems based on composite degree residuosity classes, " In *Advances in cryptology EUROCRYPT '99*, pp.223–238, Springer, 1999.
- (52) C. Gentry, " A fully homomorphic encryption scheme, " *PhD thesis*, Stanford University, 2009.
- (53) D. Bogdanov, S. Laur, and J. Willemson, " Sharemind: A framework for fast privacy-preserving computations, " In *Computer Security-ESORICS 2008*, pp.192–206, Springer, 2008.
- (54) 千田浩司, " プライバシーを守った IT サービスの提供技術: 5. 安全な情報処理を目指す秘密計算技術の研究動向と実用化に向けた取り組み, " *情報処理*, vol.54, pp. 1130–1134, 2013.
- (55) L. Kamm, D. Bogdanov, S. Laur, and J. Vilo, " A new way to protect privacy in large-scale genome-wide association studies, " *Bioinformatics*, vol.29, no.7, pp.886–893, 2013.
- (56) H. Kikuchi and J. Sakuma, " Bloom filter bootstrap: Privacy-preserving estimation of the size of an intersection, " In *Data and Applications Security and Privacy XXVII*, pp. 145–163, Springer, 2013.
- (57) S. Wu, J. Kawamoto, H. Kikuchi, and J. Sakuma, " Privacy-preserving online logistic regression based on homomorphic encryption, " *信学技報 ISBISML2013-10*, vol.113, no.139, pp.67–74, 2013.
- (58) T. Yoshida, K. Kato, T. Fujimaki, K. Yokoi, M. Oguri, S. Watanabe, N. Metoki, H. Yoshida, K. Satoh, Y. Aoyagi, Y. Nishigaki, M. Tanaka, Y. Nozawa, G. Kimura, and Y. Yamada, " Association of genetic variants with chronic kidney disease in Japanese individuals, " *Clinical Journal of the American Society of Nephrology*, vol.4, no.5, pp.883–890, 2009.
- (59) J.-P. Hubaux, J. Fellay, E. Ayday, M. Laren, J. L. Raisaro, P. Jack, et al, " Privacy-preserving computation of disease risk by using genomic, clinical, and environmental data, " In *Proceedings of USENIX Security Workshop on Health Information Technologies (HealthTech'13)*, number EPFL-CONF-187118, 2013.
- (60) J.-P. Hubaux, J. Rougemont, E. Ayday, J. L. Raisaro, et al, " Protecting and evaluating genomic privacy in medical tests and personalized medicine, " In *ACM Workshop on Privacy in the Electronic Society (WPES)*, number EPFL-CONF-188284, 2013.
- (61) K. B. Frikken, " Practical private DNA string searching and matching through efficient oblivious automata evaluation, " In *Data and Applications Security XXIII*, pp.81–94, Springer, 2009.
- (62) M. Blanton and M. Aliasgari, " Secure outsourcing of DNA searching via finite automata, " *Data and Applications Security and Privacy XXIV*, pp.49–64, 2010.
- (63) L. Wei and M. K. Reiter, " Third-party private dfa evaluation on encrypted files in the cloud, " In *Computer Security-ESORICS 2012*, pp.523–540, Springer, 2012.
- (64) R. Gennaro, C. Hazay, and J. S. Sorensen, " Text search protocols with simulation based security, " In *Public Key Cryptography-PKC 2010*, pp.332–350, Springer, 2010.
- (65) P. Mohassel, S. Niksefat, S. Sadeghian, and B. Sadeghiyan, " An efficient protocol for oblivious dfa evaluation and applications, " In *Topics in Cryptology-CT-RSA 2012*, pp.398–415, Springer, 2012.
- (66) A. V. Aho, J. E. Hopcroft, and J. D. Ullman, *The design and analysis of computer algorithms*, Addison-Wesley Publishing Company, 1974.
- (67) 原田弘毅, 笹川裕人, 有村博紀, 佐久間 淳, " 多項式時間・領域秘密正規表現照合アルゴリズム, " 暗号と情報セキュリティシンポジウム予稿集 (SCIS2014), no. 4D1, 2014.
- (68) 笹川裕人, 原田弘毅, 有村博紀, 佐久間 淳, " 線形時間・領域秘密近似文字列照合アルゴリズム, " 暗号と情報セキュリティシンポジウム予稿集 (SCIS2014), 3F2, 2014.
- (69) T. F. Smith and M. S. Waterman, " Identification of common molecular subsequences, " *Journal of molecular biology*, vol.147, no.1, pp.195–197, 1981.
- (70) S. Jha, L. Kruger, and V. Shmatikov, " Towards practical privacy for genomic computation, " In *IEEE Symposium on Security and Privacy 2008 (SP 2008)* pp.216–230, IEEE, 2008.
- (71) A. J. and B. Pinkas, " Secure hamming distance based computation and its applications, " In *Applied Cryptography and Network Security*, pp.107–124, Springer, 2009.
- (72) J. Katz and L. Malka, " Secure text processing with applications to private dna matching, " In *Proceedings of the 17th ACM conference on Computer and communications security*, pp.485–492, ACM, 2010.
- (73) C. Hazay and T. Toft, " Computationally secure pattern matching in the presence of malicious adversaries, " In *Advances in Cryptology-ASIACRYPT 2010*, pp.195–212, Springer, 2010.
- (74) M. Franklin, M. Gondree, and P. Mohassel, " Communication-efficient private protocols for longest common subsequence, " In *Topics in Cryptology-CT-RSA 2009*, pp.265–278, Springer, 2009.
- (75) M. Gondree and P. Mohassel, " Longest common subsequence as private search, " In *Proceedings of the 8th ACM workshop on Privacy in the electronic society*, pp.81–90, ACM, 2009.
- (76) E. D. Cristofaro and G. Tsudik, " Practical private set intersection protocols with linear complexity, " In *Financial Cryptography and Data Security*, pp. 143–159, Springer, 2010.
- (77) M. J. Freedman, K. Nissim, and B. Pinkas, " Efficient private matching and set intersection, " In *Advances in Cryptology-EUROCRYPT 2004*, pp.1–19, Springer, 2004.
- (78) P. Baldi, R. Baronio, E. D. Cristofaro, P. Gasti, and G. Tsudik, " Countering Gattaca: gattaca: Efficient and secure testing of fully-sequenced human genomes, " In *Proceedings of the 18th ACM con-*

ference on Computer and communications security, pp. 691–702, ACM, 2011.

- (79) C. Dwork, F. McSherry, K. Nissim, and A. Smith, “ Calibrating noise to sensitivity in private data analysis, ” In Theory of Cryptography, pp.265–284, Springer, 2006.
- (80) S. E. Fienberg, A. Slavkovic, and C. Uhler, “ Privacy preserving gwas data sharing, ” In 2011 IEEE 11th International Conference on Data Mining Workshops (ICDMW), pp.628–635, IEEE, 2011.
- (81) F. A. Wright, H. Huang, X. Guan, K. Gamiel, C. Jeffries, W. T. Barry, F. P.-M. de Villena, P. F. Sullivan, K. C. Wilhelmsen, and F. Zou, “ Simulating association studies: A data-based resampling method for candidate regions or whole genome scans, ” Bioinformatics, vol.23, no.19, pp.2581–2588, 2007.
- (82) F. McSherry and I. Mironov, “ Differentially private recommender systems: Building privacy into the net, ” In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp.627–636, ACM, 2009.
- (83) A. Johnson and V. Shmatikov, “ Privacy-preserving data exploration in genome-wide association studies, ” In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pp.1079–1087, ACM, 2013.
- (84) J. R. T.-Pastoriza, S. Katzenbeisser, and M. Celik, “ Privacy preserving error resilient DNA searching through oblivious automata, ” In Proc. Comput. Commun. Security (CCS’07), pp.519–528, ACM, 2007.

(幹事団提案 , 平成 26 年 3 月 3 日受付 3 月 10 日最終受付)

佐久間 淳 (正員)

2003 東工大大学院総合理工学研究科博士後期課程了・博士 (工学) . 2003 日本 IBM に入社, 東京基礎研究所に配属 . 2004 東工大大学院総合理工学研究科助手 (助教) . 2009 筑波大学大学院システム情報工学研究科准教授 . 2009 ~ 2013JST さきがけ研究員 . 2014 から JST CREST 「ビッグデータ統合利活用のための次世代基盤技術の創出・体系化」領域におけるプロジェクト「自己情報コントロール機構を持つプライバシー保護データ

収集・解析基盤の構築と個別化医療・ゲノム疫学への展開」研究代表者.