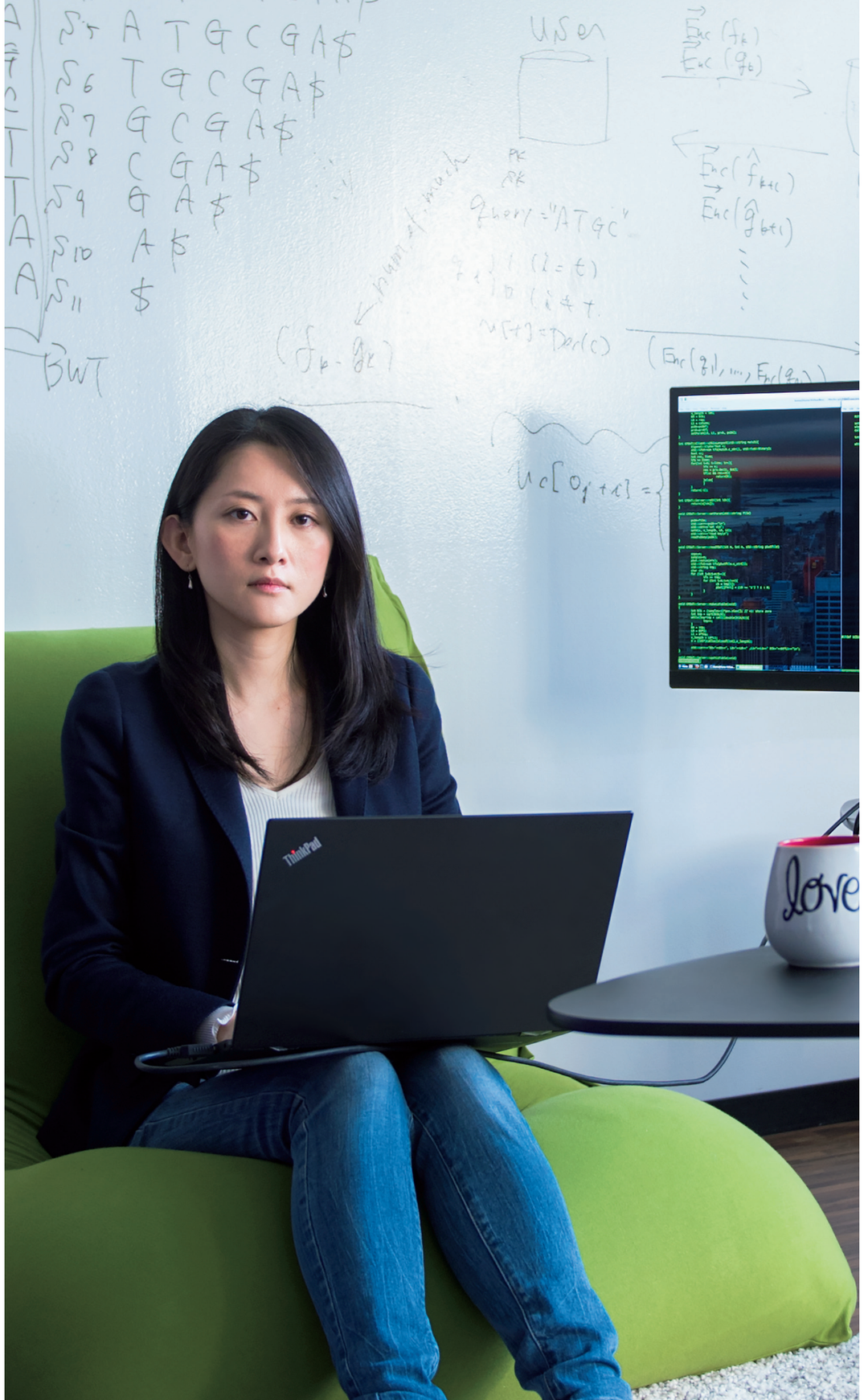
個人ゲノム情報を安全に解析する新技術「秘匿ゲノム検索」

基幹理工学部 情報理工学科
清水佳奈 准教授



ゲノム情報の効果的解析は病気の治療に生かせる。一方で、使い方を間違えると、プライバシー侵害につながる。清水研究室ではこのジレンマを解決すべく、ゲノムデータを暗号化したまま解析する技術の開発に取り組んでいる。

近年の生命科学では、計算機科学の手法を駆使した高度な情報解析が必要不可欠です。これは、計測機器の技術革新により、大量かつ多様なデータが産出されるようになったためです。最も顕著な例は、DNAの塩基配列を決定する装置の劇的な性能向上でしょう。図1はヒトの全ゲノム配列を決定するのに必要なコストを示したグラフです（アメリカ国立衛生研究所による試算）。新型装置が普及しはじめた2000年代後半から急激にコストが下落しているのが読み取れます。1990年代に実施された「ヒトゲノム計画」では、ヒトの全ゲノム配列を決定するのに約30億ドルもの予算が費やされましたが、現在はわずか1000ドルで同様の情報を取得することが可能です。

Profile
しみず・かな
2001年早稲田大学理工学部卒業。06年同大学院理工学研究科博士後期課程修了。博士（工学）。同年、国立研究開発法人・産業技術総合研究所に入所。13～15年米メモリアル・スローン・ケタリングがんセンター客員研究員。16年から現職。バイオインフォマティクスの研究に従事。最近は高速シークエンシングデータの解析アルゴリズムの開発、ゲノム配列の秘匿検索等に興味を持っている。

1 個人ゲノム時代の到来 プライバシー保護

ゲノム情報を安価に取得できるようになったことが後押しとなり、世界各地で大規模なゲノムコホート研究（特定の集団に属する人々からゲノム情報を収集し、その後の健康状態などを追跡することによって、ゲノムと疾患の関連性を解析する研究）が進められています。例えば、わが国最大級のゲノムバンクとして有名な東北メディカル・メガバンク機構では、15万人規模のゲノム情報を収集する計画が進められています。また、産業界においては民間企業による遺伝子検査ビジネスが拡大しつつあり、先駆けとなった米国企業の23andMeでは既に100万人分のゲノム情報を収集したとの報告もあります。

個人ゲノム情報の本格的な利用が現実となった今、医学や生物学の発展が望める一方で、「プライバシーの保護」という新たな問題が顕在化しました。個人情報という観点から眺めると、ゲノムは非常に特殊な情報です。電話番号などの一般的な個人情報は、個人を識別する性質を持ちますが、個人の特徴を示すことはありません。ところが、ゲノムの場合は、個人を識別できる情報でありながら、同時に、個人の特徴を示す情報でもあるのです。この点に関して、もう少し詳細に述べたいと思います。

DNAは4種類の塩基から構成される鎖状物質のため、四つの記号から構成される配列として表現できます。この配列は一卵性双生児などの例外を除いて個人に固有であることが知られているため、ゲノムは識別子の役割を果たすのです。また、ゲノムはいわば個人の「設計図」であるため外見上の特徴や疾患の有無などは、ゲノム配列上の特定の箇所に生じた変異によって説明できると考えられています。

例えば、BRCA遺伝子に変異があると乳がんの罹患率が高いことが知られています。女優のアンジェリーナ・ジョリーさんが遺伝子検査の結果を考慮して手術を受けたことでも有名です。つまり、特定の変異と個人の特徴の関連性が明らかになっているため、ゲノム配列の変異を見ると持ち主の特徴を言い当てられてしまうのです。さらに、ゲノム情報は遺伝しますから、個人のゲノム情報の開示は、血縁者のプライバシーにも大きく影響します。このような性質を持つ情報は他に類がなく、それゆえにゲノムは究極の個人情報と形容されるのです（図2）。

現在、ゲノム情報を利用する際には氏名などの識別子とゲノムデータを切り離す「仮名化」と呼ばれる処理を中心とした保護措置が取られています。しかしながら、ゲノムの持つ情報の

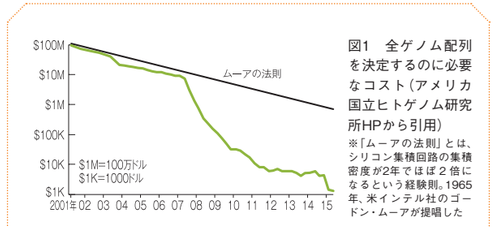


図1 全ゲノム配列を決定するのに必要なコスト（アメリカ国立ヒトゲノム研究所HPから引用）
※「ムーアの法則」とは、シリコン集積回路の集積密度が2年でほぼ2倍になるという経験則。1965年、米インテル社のゴードン・ムーアが提唱した

Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcosts [2016年12月に取得]

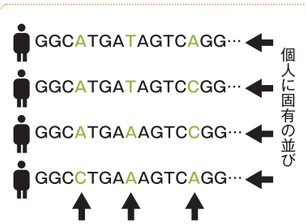


図2 ゲノム配列と個人情報

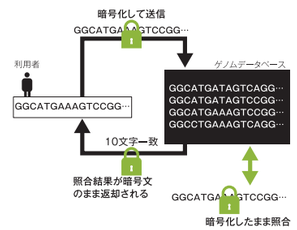


図4 秘匿ゲノム検索

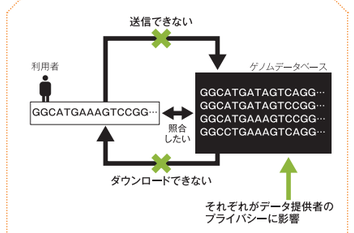


図3 ゲノム情報検索のジレンマ

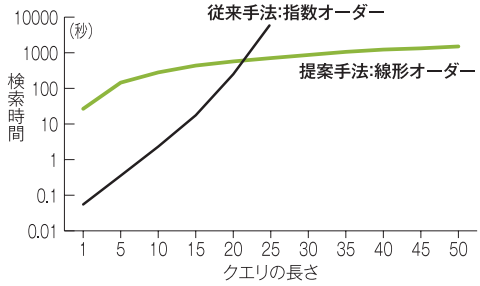


図6　秘匿ゲノム検索の性能（従来手法と提案手法の比較）。従来手法ではクエリの長さに対して指数関数的に増大する計算量が必要なのに対し、提案手法では長さに比例する計算量のみで検索できる

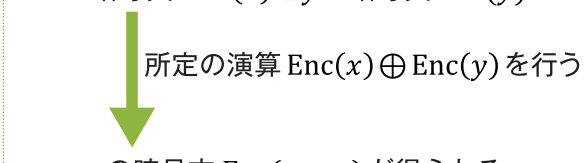


図5　準同型暗号による演算の例

本質を考慮すると、全てのケースでそのような運用が適切なのか議論の余地があるでしょう。例えば、遺伝情報と家系（名字など）を関連づけたデータベースが手に入れやすい米国の研究では、ゲノム配列と公共のデータベースの情報のみからドナーの名字を言い当て、さらにドナーの所在する州などの付加的な情報を利用した場合には、本人を含む数人の候補に到達できる危険性を指摘しています（引用：Science 18 Jan 2013: Vol. 339, Issue 6117, pp. 321-324, URL http://science.sciencemag.org/content/339/6117/321.full）。

日本国内では、2015年9月に個人情報保護法の改正が行われましたが、改正法で新たに定められた「個人識別符号」には、遺伝情報が含められることになったため、個人ゲノム情報の取り扱いに関する様々なルール作りが必要となっています。個人ゲノムデータの増大が急速に進む中、データを扱うための環境整備が追いつかず、ゲノムデータの流通は阻まれ、特定の組織内にとどまりがちです。そのため、潜在的には豊富に存在するはずの情報が十分に活用されない状況が続いており、有効な打開策が求められています。

2 秘匿ゲノム検索

それでは、個人のプライバシーを保護しつつ、データをうまく活用する手立てはないのでしょうか？ 我々はこの問題を解決するために、ゲノムデータを暗号化したまま解析する技術の開発に取り組んでいます。

どのような技術なのか、図3に示す例を用いて説明したいと思います。ここでは、医師が患者のゲノム配列をデータベースと照合して、疾患の原因を解析するケースを考えますが、患者のプライバシーを守ろうとすると、医師は検索クエリ（患者のゲノム配列）をデータベース側に送信できません。そのため、データベース全体をダウンロードして、自分の手元で照合する必要があります。ところが、データベースに保存されているゲノム配列はデータ提供者のプライバシーにかかわるため、こちらも同様に医師にダウンロードさせることができないのです。

このようなジレンマが生じるケースにおいて、我々の開発した「秘匿ゲノム検索」を用いると、医師はクエリをデータベース側に秘密にしたまま、データベースの照合結果のみを得ることが可能です。秘匿ゲノム検索では、利用者がクエリを暗号化してサーバーに送ります。そうするとサーバー側では、受け取った暗号文を復号せず、そのままデータベースと照合し、照合結果を暗号化したまま返却します（図4）。このような処理には準同型暗号と呼ばれる暗号方式を用います。この暗号方式は、単純に値を暗号化するだけでなく、暗号化したまま足し算やかけ算を行うことができる便利な性質を持っています（図5）。

実は、コンピューター上で行われる処理はすべて足し算とかけ算の組み合わせで表現できます。そのため、通常の足し算とかけ算を、準同型暗号による「安全な」足し算とかけ算に置き換えることによって、データベース検索のような複雑な処理も安全に実施できるのです。ただし、暗号化したまま行う演算には非常に重い計算が必要なため、何の工夫もせずに置き換えると計算量が膨大になってしまう問題があります。我々の手法では、簡潔データ構造と呼ばれる非常に効率の良いデータ構造と暗号技術を効果的に組み合わせることにより、従来手法よりも桁違いに高速な検索を行うことを実現しています（図6）。

3 ゲノム情報の有効活用に向けて

秘匿ゲノム検索は、収集が難しい希少疾患データの統合や、遺伝子検査結果の安全な利用など、医学研究や医療・健康産業に貢献することが期待されます。また、ゲノム情報が安全に扱われていることが世の中に伝われば、これまでプライバシーの問題を不安に思って躊躇していた人々が安心してデータを提供できるかもしれません。データが増大すれば、研究によって得られる知見も増大し、医学や生物学の向上という形で社会に還元されることになります。我々の研究室では、秘匿ゲノム検索の性能や機能を高め、ゲノム情報を有効活用する基盤技術の形成に貢献することを目指しています。

参考文献
Shimizu K, Nuida K, Rätsch G “Efficient Privacy-Preserving String Search and an Application in Genomics,” Bioinformatics, 2016 32(11): 1652-1661
清水 佳奈「生命情報科学におけるプライバシ保護検索」、日本シミュレーション学会学会誌：シミュレーション（小特集「エネルギーシミュレーションとデータ解析」について）、第35巻 第1号 p.26-31