秘密計算による化合物データベースの検索技術

― 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 ―

平成 23 年 11 月 1 日
独立行政法人 産業技術総合研究所
国 立 大 学 法 人　筑 波 大 学
国 立 大 学 法 人　東 京 大 学

■ ポイント ■

- **ユーザー側とデータベース側がお互いに情報を開示しないで、実用的な速度で検索可能**
- **全工程で暗号化されたデータだけを用いるため、情報漏えいのリスクが大幅に低下**
- **創薬のオープンイノベーションの加速に期待**

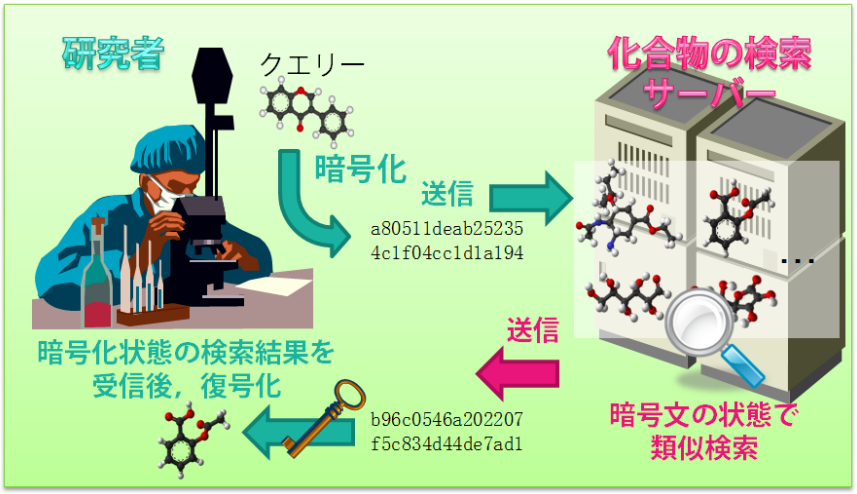
■ 概　要 ■

独立行政法人 産業技術総合研究所【理事長 野間口 有】（以下「産総研」という）生命情報工学研究センター RNA 情報工学研究チーム 清水 佳奈 研究員ら、国立大学法人 筑波大学【学長 山田 信博】（以下「筑波大」という） 荒井 ひろみ 研究員ら、国立大学法人 東京大学【総長 濱田 純一】（以下「東大」という）浅井 潔 教授らは共同で、秘密計算による化合物データベースの検索技術を開発した。

創薬などに用いられる化合物の情報は企業秘密として厳重に管理されるため、外部データベースに情報を送って類似化合物の検索を行うことが難しかった。今回開発した技術により、ユーザー側とサーバー側の双方が互いに情報を開示することなく、互いのデータを暗号化したままで比較し、検索結果だけを得ることができる。この技術では、データを加法準同型暗号により暗号化し、加算だけで検索を行えるアルゴリズムによって、大量のデータに対して実用的な速度で検索を実行できる。今回開発した技術は、企業間での安全かつ効果的な情報交換を促進し、創薬におけるオープンイノベーションに貢献すると期待される。

なお、この技術の詳細は、2011 年 11 月 8～10 日に神戸国際会議場（神戸市中央区）で開催される情報計算化学生物学会 2011 年大会/2011 年日本バイオインフォマティクス学会年会（CBI/JSBi2011 合同大会）で発表する予定である。

＿＿＿＿＿は別紙【用語の説明】参照



■ 開発の社会的背景 ■

生命情報科学の分野では、個人ゲノムや創薬関連化合物の情報など秘匿性の高いデータを用いて情報解析を行う必要があるが、プライバシー保護の方法論については深く議論されず、単純にデータをネットワーク上から切り離すといった措置がとられてきた。そのため、秘匿性の高い情報同士を組み合わせた解析が行えず、データを有効活用することができなかった。例えば、創薬ターゲットに対して活性をもつ化合物をスクリーニングする際には、既知の化合物データベースから類似化合物を検索することが有効であるが、化合物の情報は企業秘密として厳重に管理されるため、外部のデータベースを用いて検索することができなかった。また、フォーカスドライブラリーのように有料で販売されているデータベースでは、ユーザーの保持するデータと適合しているかどうかを事前に知る手段がないため、多くのユーザーが無駄な投資となることを恐れて購入をためらい、ビジネスチャンスの喪失につながっていた。

一方、情報科学の分野では、近年、プライバシー保護データマイニングが注目を集めている。プライバシー保護データマイニングでは暗号や統計などの技術を駆使し、秘匿性の高い情報そのものを開示することなくデータの解析結果だけを得る手法が提案されている。

このような背景から、生命情報科学の分野にプライバシー保護データマイニングの手法を応用し、秘匿性の高いデータを開示せずに情報解析する技術の開発に着手した。

■ 研究の経緯 ■

産総研と東大はこれまでに、生命情報科学の分野においてさまざまに分散するサーバーやデータベースなどの統合化に取り組んできた。また、この分野では個人ゲノムや創薬関連化合物など、秘匿性の高い情報を扱う必要性が今後ますます増大すると考え、プライバシー保護を考慮した情報統合の重要性について検討してきた。一方、筑波大は、プライバシー保護データマイニングの基礎研究を進めてきた。

今回、具体的なサービスに直結する応用例として、創薬に重要な役割を果たす化合物の検索問題をターゲットとして定め、3 者共同で加法準同型暗号を応用したアルゴリズムを考案した。また、筑波大からプライバシー保護データマイニングの技術に関するノウハウの提供を受けながら、産総研がプロトタイプの開発、および、テストデータでの検証実験を担当した。

■ 研究の内容 ■

一般的に、データベースなどでは化合物の情報はフィンガープリントと呼ばれる固定長のビット列で表現されている。各ビットはそれぞれ化合物の特徴の有無を示しており、1 の場合は対応する化合物の特徴があること、0 の場合はないことを示している。図 1 にフィンガープリントの例を示す。化合物の類似性は、これらフィンガープリント同士の比較により評価する。一般的には、値が一致するビットの箇所が多ければ類似度が高いと考えられる。

フィンガープリントの類似性を評価する指標としては、Tversky 係数が最も普及している。今回開発した技術では、Tversky 係数を変形して、新たな数式(以下「判定式」と記す)を導出した。判定式に対して、比較を行いたい 2 つのフィンガープリントを代入すると、Tversky 係数がユーザーの定めた値よりも大きい場合は計算結果が正の値となり、そうでない場合は負の値となる。

そのため､判定式の計算結果から、2 つの化合物がユーザーの定めた基準よりも類似しているか、そうでないかを判断することができる。ユーザー側の化合物とサーバー側の化合物の比較を行うには、あらかじめ、加法準同型暗号でおのおのの化合物に対応するフィンガープリントを暗号化しておく。暗号化に用いる鍵はユーザー側が発行し、暗号化を行う前にサーバー側に送信しておく。暗号文を解読するための鍵も発行するが、サーバー側には渡さないで保管する。ユーザー側は暗号化したフィンガープリントをサーバー側に送信する。サーバー側は、解読用の鍵をもっていないため、ユーザー側のデータの中身は解読できない。この状態で、サーバー側が、ユーザー側から受け取ったデータと自分がもつデータに対して、判定式の計算を行う。判定式は加算のみから成るため、 加法準同型暗号で暗号化したデータ同士の演算のみから暗号化した状態の計算結果を求めることができる。暗号化したデータ同士の演算で得られた計算結果は暗号化されたままなので、サーバー側は内容を知ることなく、ユーザー側に計算結果を送信することができる。ユーザー側は自分だけがもっている解読用の鍵を使って、受信した暗号文を解読する。解読した結果が正の値か負の値かを確認すれば、自分が送った化合物がサーバー側の化合物と類似しているかどうかを知ることができる。このようにして、お互いの化合物の内容を伏せたまま、ユーザー側が 2 つの化合物の類似性のみを知ることができる。

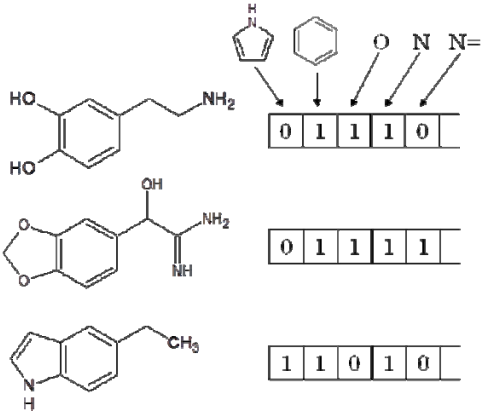


図 1 フィンガープリントの例

これらの基本技術を用いると、用途に応じてさまざまな検索を行うことができる。例えば、ユーザーのクエリー（検索情報）と類似した化合物が、データベース中に何個あるかを検索する、類似する化合物の ID だけを検索結果として返却するといったことが可能となる。

今回の技術は、プライバシー保護の手法として加法準同型暗号だけを用いるため、従来技術と比較して計算コストおよびメモリー使用量が大幅に少なく、大量データの検索にも対応可能である。また、ユーザー側とデータベース側で必要となる通信は一往復だけなので、用途に応じた通信手段を選択することができる。このように、今回開発した技術は実用性が非常に高く、企業間での安全かつ効果的な情報交換を促進し、創薬分野におけるオープンイノベーションへ貢献することが期待される。

■ 今後の予定 ■

今後は、これまでに開発したソフトウエアに関して、実データに基づく評価実験を行う。また、評価結果をもとにソフトウエアの改良を行い、企業と提携して化合物検索サービスを実施することを目指す。

【用語の説明】

◆加法準同型暗号

加法準同型暗号では、平文（暗号化する前の元のデータ）同士の加算結果と、暗号文（暗号化したデータ）同士で特定の演算（例えば、乗算など）を行った後に復号化して得られる結果が同じになる。つまり、平文を直接知らず、暗号文だけがわかっている場合でも、平文同士の加算を計算することができる。ただし、計算結果は復号化しなければわからない。

◆フォーカスドライブラリー

特定のターゲットに対して活性が高いことが期待される化合物の情報を収集したデータベース。一部の企業では有料で販売されている。

◆プライバシー保護データマイニング

データが複数地点に分散しているときに、データに含まれる秘密情報を互いに開示することなく、有用な統計情報のみを計算する技術。例えば、おのおのが誰に投票したか開示することなく、選挙の結果のみを計算する、など。

◆ビット

2 進数の 1 桁のこと。0 か 1 の状態をもつ。

◆Tversky 係数

2 つのビットベクトル間の類似性を評価する尺度。化合物の類似性を測る際には、化合物のフィンガープリント間の Tversky 係数を計算することが多い。 0～1 までの値をとり、値が 1 に近いほど類似性が高く、0 に近いほど類似性が低いことを示す。