グループ内での情報共有を目的としたプライバシー保護 DNA データベースシステム

清水將吾[†1] 權娟大[†2]
小田切和也[†1] 宮崎智[†2]

近年，e-Health などの自身の健康情報をインターネット上で管理するサービスが活用されている．類似する DNA 配列をもつ利用者同士であれば，同じ疾患にかかる可能性があり，治療や予防などの情報を共有できる．一方で，データベース管理者を含む DNA 配列を共有しない利用者に対しては情報を開示したくないという要求がある．本稿では，このような環境において，利用者の DNA 情報を保護しつつ，類似 DNA に関連付けられた情報を効率的に検索する手法を提案する．

Privacy-Preserving DNA Database System for Information Sharing within Groups

Shogo Shimizu,[†1] Yeondae Kwon,[†2] Kazuya Odagiri[†1] and Satoru Miyazaki[†2]

Recently, personal health management services on the Internet such as e-Health have prevailed. Ones that share similar DNA sequences tend to catch same diseases. Thus, exchanging their experiences on treatment and prevention through such systems is helpful. On the other hand, they are not willing to expose their DNA information to non-shared users. In this paper, we propose a database scheme that enables information sharing among groups of similar DNA sequences and provides efficient search on similar DNA sequences.

†1 産業技術大学院大学産業技術研究科
Industrial Technology Graduate Course, Advanced Institute of Industrial Technology
†2 東京理科大学薬学部
Faculty of Pharmaceutical Sciences, Tokyo University of Science

1. はじめに

近年，e-Health などの自身の健康情報をインターネット上で管理するサービスが活用されている．将来，安価で自身の DNA 配列が入手できるようになれば，類似する DNA 配列をもつサービス利用者の情報を通じて，自身がかかりやすい疾患や治療，予防などの情報を共有できるようになる．一方で，サービス提供者を含む類似 DNA を所有しない利用者に対しては，問合せやデータベースに格納されている自身の情報を開示したくないという要求がある．本稿では，利用者の DNA 情報を保護しつつ，類似 DNA に関連付けられた情報を効率的に検索できる DNA データベースシステムを提案する．

2. 提案手法

DNA 配列は ATGC の 4 種類の塩基で構成される文字列である．簡単のため，すべての DNA 配列の長さは同一であると仮定する．二つの文字列間の類似度を，二つの文字列が共通にもつ長さ q の部分文字列（以降，q-gram と呼ぶ）の数と定義する．問合せ時には DNA 配列 s が与えられ，s と類似するすべての DNA 配列に関連付けられた情報が返される．目的は，DNA 配列をデータベース管理者から秘匿すること，類似 DNA 提示者以外には関連情報を開示しないこと，類似 DNA 関連情報を効率良く抽出することである．

本方式では，DNA 配列 s の関連情報へのアクセス手段をその情報のレコード識別子 $rid(s)$ に限定する．$rid(s)$ は十分な長さをもつ乱数であり，s の類似 DNA 配列を提示できる利用者によってのみ得ることができる．これを実現するために，集合間の安全な曖昧照合手法である fuzzy vault scheme[2] を採用する．fuzzy vault scheme は二つの集合が十分類似する場合のみ，秘匿された情報を開示する仕組みである．

2.1 登録

利用者が DNA 配列 s の関連情報を登録する際の手順は次の通りである．h_1, h_2 を任意の一方向性関数とする．

(1) 任意の k 次情報多項式から n 次符号語多項式 p を生成する．

(2) 配列 s から生成される q-gram からなる集合を Q_s とする．各 $q \in Q_s$ に対して，その x 座標への写像 $x_q = h_1(q)$，および，$y_q = p(h_1(q))$ を計算する．

(3) Q_s の要素ではない m 個の任意の q-gram からなる集合を Q_c とする．ここで，m は安全性と効率を調整するパラメータであり，利用者が任意に定めてよい．各 $q' \in Q_c$ に対して，$x_{q'} = h_1(q')$ とし，$y_{q'}$ を $p(h_1(q'))$ 以外の任意の値とする．

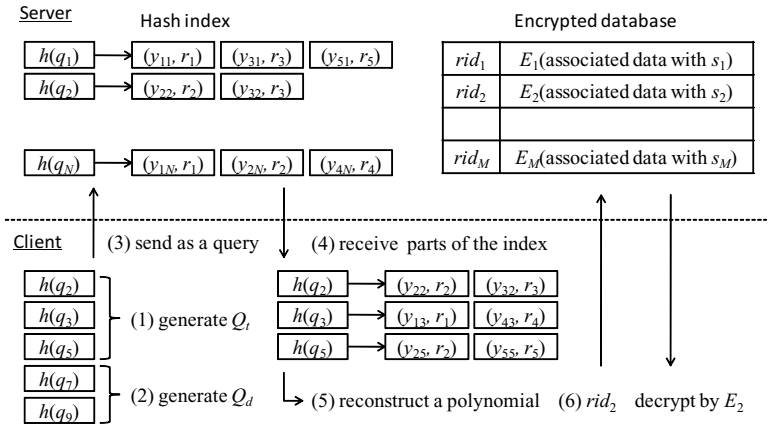

図 1 問合せ処理
Fig. 1 Query processing.

(4) p の係数をもとにビット列 E_s を生成し，$c_p = h_2(E_s)$ とする．

(5) ハッシュ索引の構成要素として $I_s = \{(x_q, y_q) \mid x_q \in Q_s\} \cup \{(x_{q'}, y_{q'}) \mid x_{q'} \in Q_c\}$ を，共有データとして $c_p (= rid(s))$ と s の関連情報を E_s で暗号化したデータの対をサーバに送信する．

サーバは登録データを受け取ると，登録データに対する一意の識別子 r を生成する．各 $(x_q, y_q) \in I_s$ に対して，x_q をキー，(y_q, r) を値としてハッシュ索引に追加する．

2.2 問 合 せ

利用者が DNA 配列 t を問合せとして与えるときの処理の流れを図 1 に示す．

(1) 配列 t から生成されるすべての q-gram からなる集合を Q_t とする．

(2) Q_t の要素ではない m' 個の任意の q-gram からなる集合を Q_d とする．ここで，m' は安全性と効率を調整するパラメータであり，利用者が任意に定めてよい．

(3) $I_t = \{h_1(q) \mid q \in Q_t \cup Q_d\}$ を求め，サーバに送信する．

サーバは I_t の各要素をキーとするハッシュ索引の値をすべて返す．クライアントは第二段階として次の処理を行う．sim を類似度の閾値とする．但し，DNA 配列の長さを l としたとき，$sim \geq \frac{k+(l-q+1)}{2}$ とする．この制約は rid の再構築を保証するために必要である．

(4) サーバから受け取ったハッシュ索引から，Q_d 中の値がキーである要素を削除する．残りのハッシュ索引の中で sim 回以上出現する識別子 r を求め，(y_q, r) が値であるようなすべてのキー x_q と値 y_q の組を抽出する．

(5) (x_q, y_q) の集合から多項式 p を再構築する．再構築は正しく行えることが保証される．

(6) p から登録時と同様に得た c_p をキーとして問合せを行い，結果を E_s で復号する．

3. 考 察

3.1 安 全 性

まず，類似 DNA 非保有者が偽の問合せで得られたハッシュ索引から rid を推測しようとする攻撃について考える．この強度は fuzzy vault scheme の安全性に基づき，体の大きさを e としたとき，小さい実数 μ に対して $\frac{\mu}{3} e^{k-l} (\frac{m}{l})^l$ 通りの多項式が存在する[2]．

管理者による頻度分布の偏りを用いた攻撃に対しても，Q_c によって防ぐことができる．Q_c を大きくすれば，次節で述べるように処理効率は下がるが，q-gram の頻度分布が十分変形されるために背景知識を用いた推測攻撃が困難になる．また，登録したデータは個別に再登録可能であり，利用者が定期的に更新を行うことで安全性を保つことができる．

3.2 処 理 効 率

多項式構築は，例えば BM 法などの復号アルゴリズムで効率良く行える．Q_d の大きさによって安全性と処理効率が調整される．Q_d を大きくすると sim 回以上出現する識別子が含まれる可能性が高くなり，クライアントで余分な処理を行う可能性が高まる．一方，Q_d が空であれば，すべての復号を正しく行えるが，管理者に対する Q_c の効果をなくしてしまう．

4. おわりに

本稿では，データベース外部委託[1] において，利用者が保有する DNA 情報をアクセス権として利用することによって，グループ間で安全に関連情報を共有する仕組みを提案した．今後は情報漏洩度の定量化とモニタリングシステムの開発を行う予定である．

謝 辞

本研究の成果の一部は，(財) 電気通信普及財団の助成を受けたものである．

参 考 文 献

1) H. Hacigümüs, B. R. Iyer, C. Li, and S. Mehrotra: Executing SQL over encrypted data in the database-service-provider model, In *Proc. of the 2002 ACM SIGMOD Int'l. Conf. on Management of Data (SIGMOD '02)*, pp.216–227, 2002.

2) A. Juels and M. Sudan: A fuzzy vault scheme, In *Designs, Codes and Cryptography*, vol.38, no.2, pp.237–257, 2006.