タイトル 利用者のプライバシーを保護する医療情報共有システムの開発

清　水　將　吾　　産業技術大学院大学産業技術研究科助教

1 はじめに

DNA 情報は，通常，生物学実験を専門とする組織によって生産され，機能予測等の分析のためにデータベース化されることが望まれる.しかし,実際には組織内に情報技術の専門家を割り当てられないことがあり，データベースの効率的な構築，運用が課題となっている．そこで，インターネット経由でデータベース環境を提供する Database-as-a-Service の利用が検討される.

DNA データベースに対してよく行われる問合せの種類としてホモロジ検索がある．ホモロジ検索では，問合せとして与えられた配列と類似した配列をデータベースから検索して利用者にその一覧を返す.このとき，データベースに格納されている配列や問合せ配列が組織の外部であるデータベース管理者に漏洩することは経済的価値またはプライバシーの面から問題である．従って，データベースや問合せの内容をデータベース管理者から技術的に保護できることが望ましい.

プライバシー保護型 DNA 照合のための秘匿化通信に基づくプロトコルがこれまでにいくつか提案されている． しかし，暗号プロトコルに基づく方式は計算コストが大きく，データベース検索に適用すると多数の DNA 配列との照合計算を双方で行う必要であり，現実的ではない．これに対し，確率的手法に基づく方式は，一般に，計算結果は正確ではないが処理効率が良いという長所をもっており，データベース中の解候補のフィルタリングに適している.

本研究では，ホモロジ検索を効率的に処理する方式として q-gram 方式を採用し，フィルタリングに使われる q-gram 集合に撹乱演算子を適用することによってデータベース管理者による統計的推論攻撃を防ぐ手法を提案する.

2 問題

2-1 問題設定

二つの文字列間の類似性を編集距離により定義する．編集距離とは，文字の挿入，削除，置換のいずれかを編集操作としたとき，二つの文字列を同一の文字列にするために必要な編集操作の最小数として定義される.

データベース D を任意の長さの文字列（または，配列）の集合とする．D に対する問合せ（または，ホモロジ検索）とは，文字列 s と整数 k が与えられたとき，s との編集距離が k 以下であるようなすべての D 中の文字列を得ることである.

この問合せを効率的に処理する方式として，q-gram が知られている．q-gram では,「長さ m の二つの文字列の編集距離が d であれば，それらは少なくとも m-q(d+1)+1 個の共通の q-gram をもつ」という補題に従って，D 中の解候補のフィルタリングを行う．次に，得られた解候補に対して実際に編集距離を計算することで最終的な解を得る.

本研究では，D に対してデータの登録および検索を行う利用者は同一の一組織のみとし，利用者と異なる組織にデータベース管理業務を外部委託すると仮定する．目的は登録および検索に用いられるデータの元の配列情報を，データを閲覧できるデータベース管理者から秘匿することである．但し，データベース管理者はデータの改ざんは行わず，自身が利用者になって別のデータを登録または検索することもないと仮定する．従って，データベース管理者は正規利用者によって登録または検索時に送信されたデータの閲覧とそれらのデータからの推論のみが可能である.

データベースが大規模である場合，安全性を確保すると同時に，問合せ処理の効率性が求められる．q-gram により問合せ処理を行うためには，元の配列 s と s の q-gram 集合の組が必要であるが，このうち，s はクライアントの暗号鍵により保護することとし，q-gram 集合のみを別の手法で保護することを考える．q-gram

集合はフィルタリング時に類似性の判定に使われるため，暗号化を適用することはできない．問合せ処理方式の枠組みを図 1 に示す．

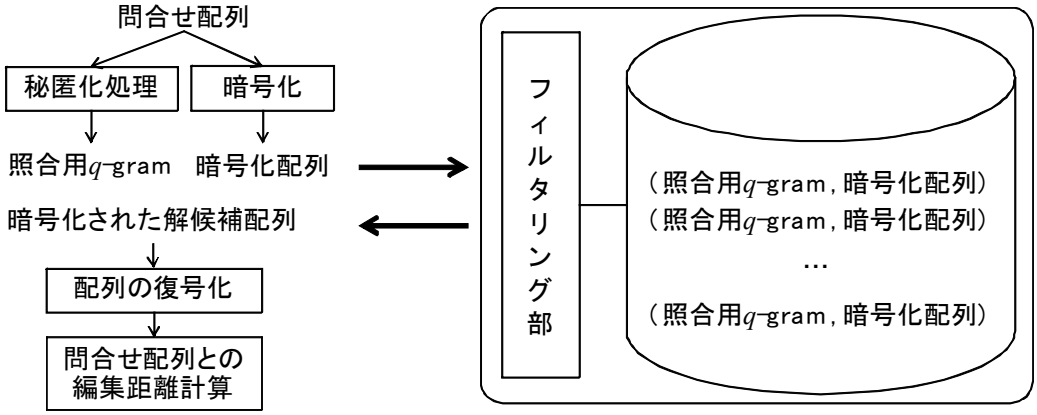

図 1 問合せ処理方式の枠組み

2-2 ハッシュに基づく手法

配列情報を秘匿する方法として，まず，q-gram のハッシュ値で照合する方法を考える．使用するハッシュ関数を h とし，配列 s から生成されるすべての q-gram からなる集合を Q(s)とする．この方法では，データベース D に配列 s を登録するときに，暗号化された s と照合用に q-gram のハッシュ値の集合 Qh(s)を送信する．検索時には，問合せ配列 t から生成される q-gram のハッシュ値の集合 Qh(t)と各 Qh(s)を照合し，共通要素の数が閾値 th 以上であるようなすべての s をクライアントに返す．ここで，th は q,s,t および編集距離 k から決定される q-gram 補題の閾値である．

この方法では，h が非可逆である限り，Qh(s)から s を得ることはできない．しかし，D に格納されている配列数が十分多い場合，D のハッシュ値の頻度分布が実際の値に近付き，事前知識を用いた統計的推論攻撃が可能になる．実際に，隠れマルコフモデルに基づくモチーフ検索では，同じ機能を持つファミリ毎にアミノ酸の出現の依存関係を数値化したスコア行列が利用されており，これを事前知識として利用できる．この情報から D に格納されているハッシュ値と元の q-gram との対応関係を推測し，照合結果から元の配列を推測できる．更に，塩基配列の場合，ある q-gram の次に出現する q-gram は，前の q-gram の先頭の塩基を取り除いた配列に ACGT のいずれかを連接した 4 通りしかなく，この依存関係も推論攻撃に利用できる．例えば，あるハッシュ値と q-gram の組が既知である場合，スコア行列を用いて次の q-gram の予測頻度を計算し，これと大きく外れた出現頻度をもつハッシュ値をその q-gram のハッシュ値候補から除外できる．

従って，ハッシュに基づく手法では配列情報を十分に保護できず，更なる洗練化が必要である．

3 確率的撹乱に基づく手法

前章で述べた統計的推測攻撃の可能性を減少させるため，データベースに格納する q-gram 集合に確率的撹乱を加え，頻度分布を変形する．

3-1 問合せ処理

まず，撹乱演算子 R を次のように定義する．以下，s を配列，|Q(s)|=m とする．Gq を長さ q のすべての q-gram からなる集合とする．

{p[i]}を{0,1,...,m}上の確率分布とし，{p'[i]}を{0,1,...,4^q}上の確率分布とする．配列 s の q-gram 集合 Q(s)が与えられたとき，R は以下の処理によって別の q-gram 集合 Q'(s)=R(Q(s))を生成する（図 2 参照）．

(1) {0,1,...,m}の中から{p[i]}に従って無作為に整数 i を選択する．
(2) Q(s)から無作為に i 個の q-gram を選択し，これらを Q'(s)に含める．
(3) {0,1,...,4^q}の中から{p'[j]}に従って無作為に整数 j を選択する．
(4) G'q を Gq の部分集合とする．G'q に含まれ，かつ Q(s)に含まれないような q-gram を j 個無作為に生成し，これらを Q'(s)に含める．

この撹乱演算子は配列毎に独立に適用でき，特定の q-gram には依存しない．

Q'h(s)={h(g)｜ g は Q'(s)の要素}

とする．s をデータベースに登録するときは s の暗号化テキストと Q'h(s)の組をサーバに送信する．問合せ時は，問合せ配列 t を R で撹乱した後に，Q'h(t)をサーバに送信する．問合せを撹乱する理由は，問合せが蓄積することによって再び統計的推論攻撃が可能になるためである．例えば，基本的な機能を表す配列パターンは問合せ配列中に頻繁に含まれる可能性があり，その偏りからハッシュ値と元の q-gram との対応を推測できる．

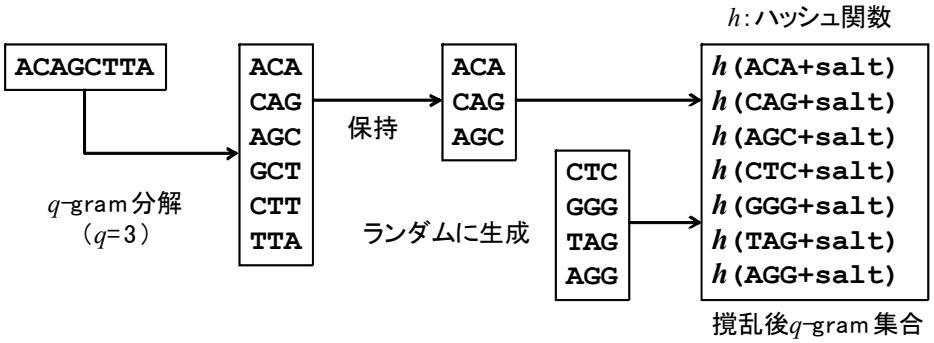

図 2 撹乱演算子 R の適用

サーバ側では各 s について，問合せ Q'h(t)と Q'h(s)との照合を行う．

|Q'h(t)∩Q'h(s)|=k'

とする．k'は s と t の実際の共通 q-gram 数よりは小さく，これをそのままフィルタリング条件に使用すると多くの偽陰性が発生する．このため，k'から|Qh(t)∩Qh(s)|の値を推定する計算を行う．この計算の結果，

p(|Qh(t)∩Qh(s)|≧m-q(d+1)+1 ｜ |Q'h(t)∩Q'h(s)|=k')≧ θ

であれば，s の暗号化データを解候補に含めることとする．ここで，θ は偽陰性の程度と効率性を調整する閾値である．

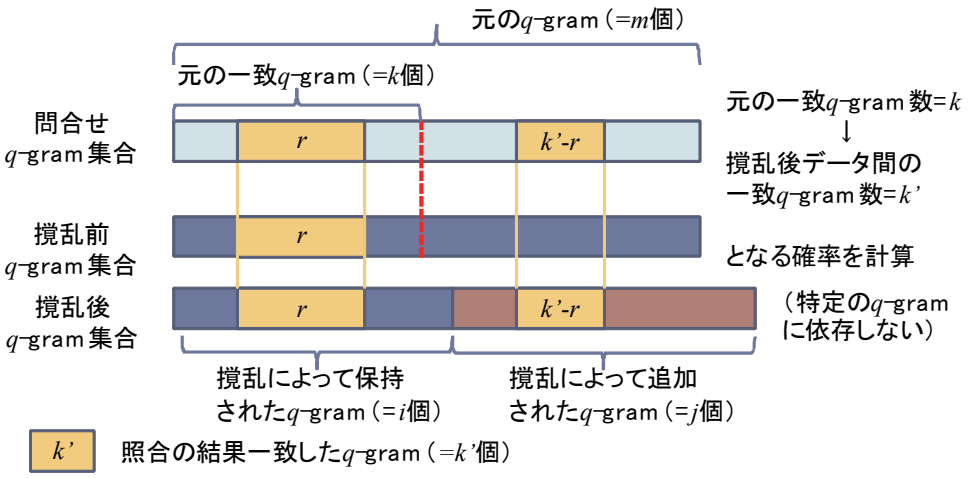

図 3 一致 q-gram 数の推定

3-2 ギャップ付き q-gram の適用

前節で述べた q-gram の連続性による頻度の偏りを利用した攻撃の可能性を減少させるために，q-gram の一般化手法であるギャップ付き q-gram の適用を検討する．

shape Q を非負整数の集合とする．任意の整数 i と shape Q に対して，位置付き shape Qi は集合{i+j| jp は Q の要素}で定義される．Qi={i_1,i_2,…,iq}（但し，i=ii<i2<…<iq）とし，s=s1s2…sm を文字列とする．各 i に対して，s の位置 i での Q-gram を s[Qi]と書き，$s_{i1}s_{i2}\cdots s_{iq}$ で定義する．

s，t をハミング距離が k であるような長さ m の文字列とする．このとき，s と t の共通 Q-gram の数は少なくとも max{0,m-max{Q}-|Q|} 以上であることが示されている．

データの登録，検索時の処理は，連続した q-gram を生成する代わりにギャップ付き q-gram を生成する以外は 3 章で述べた手順と同様である．但し，shape は外部委託先には秘密にしておく．ギャップ付き q-gram の場合は特定の q-gram 間に出現の依存関係がないため，3.1 章で得た p[k→k']はより正確な値になる．

複数の shape を利用することで，Q-gram 補題および撹乱効果によるフィルタリング精度の劣化を改善することができる．これは Q-gram 索引の大きさとのトレードオフである．また，小さい k であれば，編集距離にも対応できる．

3-3 安全性

撹乱演算子 R の適用により，q-gram ハッシュ値の頻度分布が変形され，統計的推論攻撃の難しさが上がる．R のパラメータ i の値を小さくし，j の値を大きくする程安全性は高まるが，一方でフィルタリング精度が劣化するため検索効率が下がる．R で用いられる{p[i]}，{p'[j]}が一様でない限り，撹乱の効果が平準化されることはない．

但し，元の文字列中で連続して出現する q-gram のパターンに依存関係が残っており，ハッシュ値の頻度の大小関係から元の q-gram との対応を推測できる可能性がある．

ギャップ付き q-gram を使用することによって，ある Q-gram と次の Q-gram で位置が一致する文字の数が減るため，データベース中のハッシュ値の出現頻度の大小関係を Q-gram の推測に利用することは困難である．例として，q=3，Q={0,2,5}，s=ACAGCTTA の場合を考える．このとき，ACA が Q(s)に含まれることが分かれば，次の q-gram として CAA，CAC，CAG，CAT のいずれかが必ず出現するため，出現頻度の相関を用いた予測が可能になる．一方，s[Q1]=AAT であることが分かったとしても，Q2={1,3,7}であり，s[Q2]としてどのパターンもあり得るため，s[Q1]の頻度情報を用いて s[Q2]の候補を絞ることができない．従って，攻撃者は shape が分からない限り，ハッシュ値の出現頻度の相関を利用した攻撃は困難である．shape の選び方は mC|Q| 通りある．更に，shape は配列パターンを分断するため，特定のパターンが頻出するといった生物学的な事前知識を用いた推測も困難にする．

しかし，フィルタリング結果としてクライアントに返される暗号化データの頻度から，暗号化データと元の配列との対応関係を推測される可能性がある．実際，モチーフを含み，機能等がよく知られている配列は，頻繁に検索される可能性がある．これはセキュアハードウェアを使用しない限り不可避であると予想される．

3-4 実験

撹乱演算子 R の適用により安全性を高めることによってどの程度処理効率に影響があるかを実データを用いて実験する．

配列データベースとしては，糖転移酵素のアミノ酸配列 711 種を使用した．これらの配列に含まれる平均アミノ酸数は 494 である．検索処理の高速化のため，各配列は固定長（m=22）の部分配列にあらかじめ分割した．この結果，15663 個の部分配列が生成された．実際，ホモロジ検索では局所アライメントの計算を行う，問合せと一致または非常に類似した領域をまず特定することから始まる．q-gram の長さ q の値としては代表的なホモロジ検索アルゴリズムである BLAST のハッシュ表と同じ 3 を使用した．問合せの撹乱は行わず，データベース中のデータである hALG5，hb3GalT1，hb4GalT1，hFUT1 遺伝子の部分配列を使用した．類似性を示す最大編集距離 k の値は 4 とした．

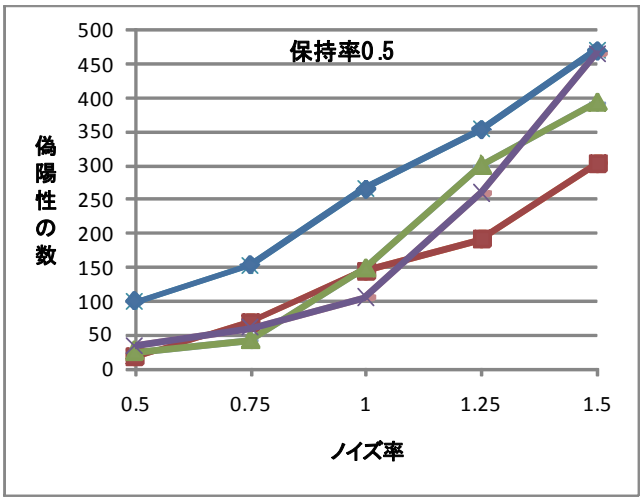

図 4 rj を変化させたときの偽陽性の数（ri=0.5）

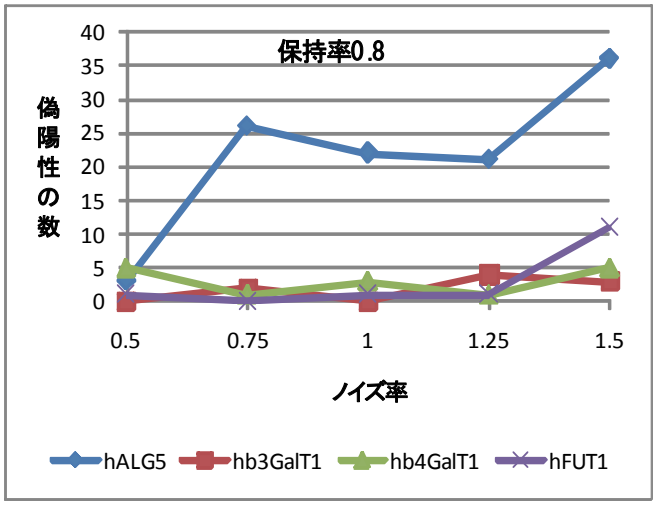

図 5　rj を変化させたときの偽陽性の数（ri=0.8）

R のパラメータである{p[i]}，{p'[j]}の値を変えることによって安全性の調整を行う．このとき，フィルタリング精度がどのように変化するかを見る．フィルタリング精度は撹乱によって新たに生じる偽陽性の数によって評価する．つまり，R を適用しない場合に q-gram フィルタリングを通過する結果を正しい解集合（偽陽性の数が 0）とみなす． θ =0.25 とした．

本実験では，{p[i]}，{p'[j]}をパラメータ ri，rj を用いて以下のように変化させる．

p[i]=1 i=m・ri のとき，そうでないとき 0

p[j]=1 j=m・rj のとき，そうでないとき 0

ここで，ri=1 であれば，R がすべての q-gram を保持することを意味する．rj=1 は元の配列の q-gram の数と同じ数の偽の q-gram を追加することを意味する．

ri=0.5 と ri=0.8 のときに，rj の比を変動させて検出された偽陽性の数を図 4，図 5 に示す．rj が大きくなる程頻度分布が歪むために安全性は高まるが，一方で偽陽性の数が増加する．また，ri=0.5 と ri=0.8 では，ri=0.8 の方が推定の精度が高まるために偽陽性の数が少ない．ri=0.5，rj=1.5 のときに最大で 470 個の偽陽性が検出される（フィルタリング率 97.0%）が，処理効率の面では実用上問題ない程度であると考える．

一方，ri=1，rj=0 の場合に q-gram 補題を満たすデータの数は hALG5，hb3GalT1，hb4GalT1，hFUT1 についてそれぞれ 6，6，12，6 であり，ri=0.8，rj=0.5 のときでもそれぞれ 3，1，1，1 件の偽陰性が発生した．これは共通する q-gram の数が元々少ないデータが撹乱演算子の適用による破棄の影響を受けやすいためと考えられる．

4 セキュアハードウェアに基づく手法

4-1 システム構成

次に，セキュアハードウェアに基づく手法について述べる．図 6 に提案するシステムの構成を示す．サーバはデータベースとフィルタリング段階を安全に実行するセキュアプロセッサからなる．問合せ変換と洗練化段階はクライアント側で実行される．

t1, ..., tn を生文字列とする．各 i に対して，Q(ti)から問合せとの照合に用いるデータ点の集合 V(ti)を生成する.V(ti)に偽のデータから生成された点集合を含めることによってデータベース中の点集合のヒストグラムを用いた統計的推論攻撃の成功確率を下げることができる.各 ti は鍵 Ei で暗号化された形式で V(ti)の中に秘匿化されて格納される．問合せ文字列と ti が十分類似しているときのみ，Ei は V(ti)から取り出すことができる．この原理は多項式復元問題に基づいている．

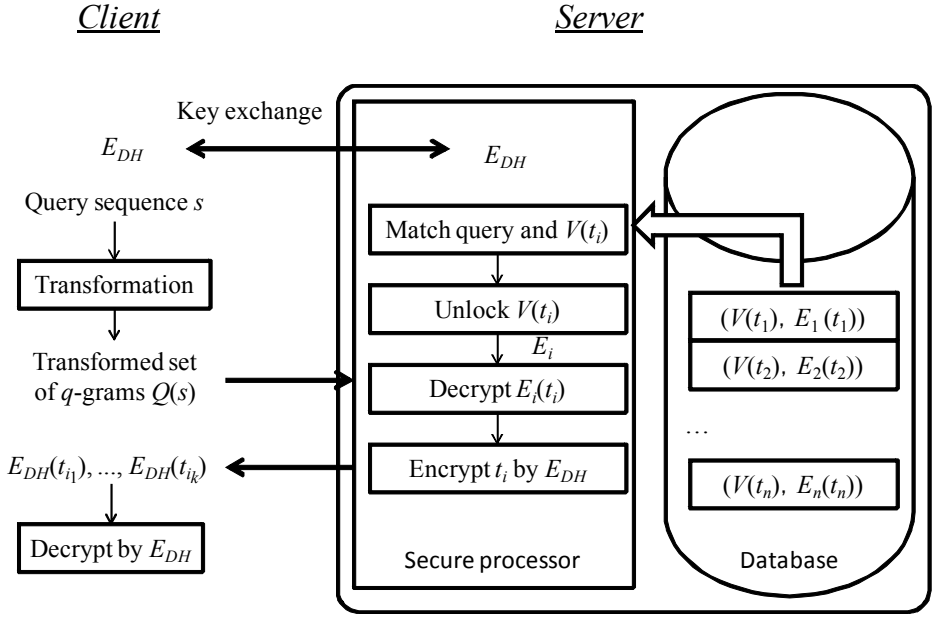

図 6 セキュアプロセッサを用いたシステム構成図

セキュアプロセッサとは物理的に保護された RAM と ROM をもつ特別なハードウェア装置である．セキュアプロセッサは鍵交換プロトコルとフィルタリング段階の実装を ROM に持ち，サーバに設置される．ここで，セキュアハードウェアに基づく解決策は，問合せ中のデータがどの V(ti)中のデータと一致したかをサーバに秘匿するために使われる．この情報は Q(ti)の変換の効果を無効化してしまう．悪意のある管理者が自身の実装を用いて直接データベースに問い合わせることを防ぐために，問合せがセキュアプロセッサから来たものであることを保証するためにデータベース側で検証を行う．このためには，例えば，セキュアプロセッサ内で秘密の置換表を用いた V(ti)中のデータ点の置換を行えばよい．このセキュアプロセッサ実装は問合せ毎に V(ti)と Ei(ti)の対を読み込めればよく，限られた大きさの RAM で実行できる．ソフトウェアのみを用いて同様のことが行えるかどうかは未知の課題である．

フィルタリング段階では,ti が問合せとの十分な類似性を示す場合にのみ復号鍵 Ei によって復元される．そうでなければ，ti を破棄し，データベース中の次のデータの処理に移る．候補解はクライアントに送信される前にセキュアプロセッサ内で共有鍵によって暗号化される．この共有鍵はセッションが開始されるときに鍵交換プロトコルによって生成され，セッションの間セキュアプロセッサ内に保管される．復号鍵 Ei を V(ti)へ埋め込むことによって ti はどのクライアントとも共有することができる点が本方式の特徴である．

4-2 問合せ処理

問合せ生成：クライアント側で問合せ文字列 t から m=|t|-q+1 個の q-gram の集合を生成し，これを登録時と同じ方法で $B=\{b_1, \cdots, b_m\}$ ($b_i \in F$)に対応付ける．B を t からの許容できる編集距離を示す整数 d とともにサーバに送信する．

照合時：サーバ側セキュアプロセッサ内でデータベース中の各 V(s)について以下の処理を行う．

(1) V(s)中の各 x 座標をセキュアプロセッサ内の置換表を用いて逆変換する．V(s)の x 座標 b_i への射影を $(x_i, y_i) \xleftarrow{(b_i,\circ)} R$ と書く．任意の y に対して，$(b_i, y) \in V(s)$ となる対(b_i, y)があれば，$(x_i, y_i)=(b_i, y)$ とする．そのような対がなければ，(x_i, y_i) には空が割り当てられる．$Q^* \leftarrow \phi$ とし，各 $i \in [1,n]$に対して，以下を行う．

$$(x_i, y_i) \xleftarrow{(b'i,\circ)} R$$

$$Q \leftarrow Q \cup (x_i, y_i)$$

(2) th=n-(d+1)q+1 とする．この値は q-gram 補題の共通 q-gram 数に関するフィルタリング条件である．もし|Q|が th より小さければ V(s)を解候補から排除し，次のデータに進む．もし|Q|が th 以上であれば，V(s)と E(s)の組を解候補集合に含め，次の処理を行う．

(3) 各解候補について，受信した符号語 Q の誤り訂正を行う．fuzzy vault scheme では，符号語が点集合からなるような Reed-Solomon 符号の一般化を使用している．復号アルゴリズムには BM アルゴリズムなどが使用できる．BM アルゴリズムの場合，訂正できる誤りの数はたかだか(n-k)/2 である．パラメータ n と k を次の式を満たすように選択する．

(n+k)/2=n-(dm+1)q+1

ここで，dm は問合せ時に指定できる編集距離の最大値である．もし前のステップで s が t と十分類似性があると判定されているならば，以下の式が成り立つため，Q は常に正しい符号語に復号される．

n-th=n-(n-(d+1)q+1)≦n-(n-(dm+1)q+1)=n-(n+k)/2=(n-k)/2

従って，s を暗号化した鍵 E は，照合時に類似文字列 t が提示され，正しい置換が V(s)の x 座標に適用されたときのみ，V(s)中に秘匿された多項式から復元できる．非類似文字列が提示された場合は，s を復元する必要はない．

(4) s を E で復号化した後に，クライアントとの鍵交換プロトコルによって得られた共有鍵によって s を再暗号化してクライアントに送信する．

洗練化時：クライアント側でデータを復号化した後に，解候補と問合せとの実際の編集距離を計算することによって最終的な解を得る．正解を見逃すことがないことが本方式の特徴である．

パラメータ値の制約の例を考える．符号語が 16 ビットからなるとする．従って，p=r=65,536 である．(n,k)=(128,64)とする．このとき，鍵 E のビット長は 16*48=1,048 である．各 q-gram が F の異なる要素に対応づけられるためには $4^q \leqq p$ でなければならず，q はたかだか 8 である．(n+k)/2=n-(dm+1)q+1 を満たすためには，q=3 の場合 dm=10 である．これらのパラメータ値は SNP 検索や配列の新規性をチェックするような場合に向いている．人の場合，数百から数千塩基ごとに一つの塩基置換があると言われている．例えば，Rs6313 (T102C) はヒト 5-hydroxytryptamne (serotonin) receptor 2A (HTR2A)の SNP であるが，1416 塩基対からなっている．これらのパラメータ値を使って，患者や医師は DNA 配列を問合せとして SNP データベースに安全に問合せを行うことで，将来特定の疾患にかかる可能性を判断することができる．

4-3 安全性

r 中の点の数を p と等しくすることで，すべての x 座標が各 vault にちょうど一度だけ現れる．つまり，どの x 座標の出現頻度も等しく，攻撃者に対して情報を与えることはない．さらに，y 座標の生成に異なる多項式を使うことによって，すべての vault においてどの y 座標の出現頻度もランダムに分布する．このため，データベース中のデータを用いた統計的推論攻撃は難しくなり，個別のデータの安全性は fuzzy vault scheme の安全性に依存する．例えば，$r=p=2^{16}$，n=128，k=64 であれば，一つの vault に対して 2^{127} 個の多項式が選択可能である．攻撃者は元の配列を得るために，これらの多項式のうちの一つを特定する必要がある．

セキュアプロセッサは vault 中の正しい点を秘匿化する役割を果たす．照合をセキュアプロセッサ内で行わなければ，q-gram の照合結果から容易に正しい点を推測できる．もう一つの目的は，データベースへの正規のアクセス権がない悪意のあるデータベース管理者からの辞書攻撃を防ぐことである．vault 中の x 座標をランダム化する置換表によって，データベース管理者がセキュアプロセッサを経由せずに直接データベースに問い合わせることができる場合でも，正しく q-gram 照合を行うことができず，結果として復号可能な符号語を得ることはできない．符号語が 16 ビットからなる場合，置換表を格納するためにはセキュアプロセッサには $32*2^{16}$=256KB のメモリがあればよい．置換表と vault 内に秘匿化された鍵はセキュアプロセッサ内で必要なだけ更新できる．

5 おわりに

本研究では，DNA データベースの運用管理業務を外部委託するという設定において，q-gram 集合に確率的な撹乱を加えることによって安全にホモロジ検索を実行できる手法を提案した．また，データベース管理者による統計的推論攻撃に対してより安全にデータを保護するために,ギャップ付き q-gram を利用する方式を提案した．今後は，安全性や攻撃モデルの定式化を行う予定である．

【参考文献】

[1] D. Asonov, “Querying Databases Privately: A New Approach to Private Information Retrieval,” LNCS 3128, 2004.

[2] M.J. Atallah and J. Li, “Secure outsourcing of sequence comparisons,” *Int. J. Inf. Sec.*, vol.4, no.4, pp.277- 287, 2005.

[3] K.B. Frikken, “Practical private DNA string searching and matching through efficient oblivious automata evaluation,” In *Proc. of the 23rd Annual IFIP WG 11.3 Working Conference on Data and Applications Security XXIII (DBSec 2009)*, LNCS 5645, pp.81- 94, 2009.

[4] S. Jha, L. Kruger, and V. Shmatikov, “Towards practical privacy for genomic computation,” In *IEEE Symposium on Security and Privacy (S&P2008)*, pp.216- 230, 2008.

[5] J.R. Troncoso-Pastoriza, S. Katzenbeisser, and M.U. Celik, “Privacy preserving error resilient DNA searching through oblivious automata,” In *Proc. of ACM Conference on Computer and Communications Security (CCS 2007)*, pp.519- 528, 2007.

〈発　表　資　料〉

題　　名	掲載誌・学会名等	発表年月
確率的撹乱に基づくプライバシー保護型 DNA ホモロジ検索	DEIM2010	2010 年 3 月
Secure Outsourcing of DNA Databases	iConcept Press	2010 年 6 月