Computer Science and Application 计算机科学与应用, 2023, 13(9), 1641-1654
Published Online September 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.139163

Hans汉斯

一种云上的高效生物隐私保护协议

吴　铎

青岛大学计算机科学技术学院，山东　青岛

收稿日期：2023年8月3日；录用日期：2023年9月1日；发布日期：2023年9月8日

摘　要

由于生物识别的可靠性和便捷性，这项技术已经成为一种重要且可靠的识别技术，常被用于身份验证。但生物特征数据具有很高的敏感性，因此，在隐私保护生物识别协议中，安全性便成为了一大挑战。现存的大多数协议都存在效率低下或者安全级别低的问题，从而限制了他们在实践中的广泛应用。为了提高安全性和效率，本文提出了一种新的隐私保护生物识别协议。新的协议对原始的生物数据库进行了预处理操作，使得上传到云服务器上的密文操作的对象大大减少，即算法匹配操作只需要在最近邻候选者之间进行，进一步提高了效率，并且，我们的新协议在大型数据库上也有很好的表现，具有更高的现实意义。

关键词

生物隐私保护，安全外包，最近邻居

An Efficient Biometric Identification Privacy Protection Protocol on the Cloud

Duo Wu

College of Computer Science and Technology, Qingdao University, Qingdao Shandong

Received: Aug. 3rd, 2023; accepted: Sep. 1st, 2023; published: Sep. 8th, 2023

Abstract

Due to the reliability and convenience of biometrics, this technology has become an important and reliable identification technology, which is often used for identity verification. However, biometric data is highly sensitive, so security becomes a major challenge in privacy-protecting biometric protocols. Most existing protocols are inefficient or have low security levels, which limits their widespread use in practice. In order to improve security and efficiency, this paper proposes a new pri-

文章引用: 吴铎. 一种云上的高效生物隐私保护协议[J]. 计算机科学与应用, 2023, 13(9): 1641-1654.
DOI: 10.12677/csa.2023.139163

vacy protection biometric protocol. The new protocol performs pre-processing operations on the original biological database, greatly reducing the objects of ciphertext operations uploaded to the cloud server, that is, the algorithm matching operation only needs to be carried out between the nearest candidates, further improving the efficiency. In addition, our new protocol also has good performance on large databases, and has higher practical significance.

Keywords

Biological Privacy Protection, Security Outsourcing, Nearest Neighbor

1. 介绍

1.1. 研究背景及意义

生物特征识别技术是一种使用人体独特的生理特征(如虹膜、声音、脸、指纹等)或行为特征(如步态、声音、击键等)来区分、识别和验证个人身份的技术。相比于传统的身份验证方法(例如密码或智能卡)，生物特征不会被丢失、窃取或遗忘。目前，生物特征识别技术被广泛应用。该技术在电子健康[1]、工业物联网[2]、无线传感器网络[3]、辅助机器人[4]等领域得到广泛应用。

一般来说，生物特征识别包含生物特征数据库和查询请求。数据拥有者(DO)对数据库执行搜索任务来匹配查询数据并对查询做出反馈。然而，在当前的大数据时代，生物特征数据的数量正以指数级速度增长，这给 DO 端带来了沉重的存储和计算负担。因此，将存储和识别任务外包给云服务器已成为一种常见的解决方法。在云计算环境中，资源受限的用户可以将复杂的计算任务外包到云服务器，并通过按需付费的方式享受云计算平台提供的计算和存储资源。

虽然云辅助生物特征识别在实践中显示出了良好的应用前景，但它也面临着一些严重的安全挑战。由于生物数据的高敏感性，一旦这些生物数据泄露或者被滥用，将会到导致不可估量的财产损失。显然，云服务器不会是完全可信的，并且还会存在恶意用户试图突破安全屏障。

因此，一个设计良好的生物识别外包协议应保证敏感数据的机密性。同时，因为 DO 需要花费额外的成本来保护隐私信息，所以隐私保护方法必须是高效的。也就是说，相对于由 DO 端独自完成识别任务(不外包)，设计的外包协议应确保 DO 端节省计算和存储开销。

所以如何设计高效且保护隐私的生物特征识别协议成为了重要研究问题。

1.2. 相关工作

在安全的两方计算模型中，数据库所有者(DO)和用户(QU)可以共同的执行生物特征识别算法同时保证自己所拥有的信息不被泄露。在这个前提下，学者们提出了很多基于两方的安全隐私方法。Erkin [5]等提出了第一个增强隐私的双方人脸识别协议，在标准特征人脸识别算法的基础上，采用同态加密进行隐藏进行生物识别和匹配操作。然而，Sadeghi [6]等指出，[5]中的协议通信和计算效率较低，实际并不适用大规模的数据库。为了进一步提高效率，他们使用乱码电路(GC)进行阈值比较，大大降低了计算成本。另外，针对于[5]和[6]协议存在的不足，Osadchy [7]等提出了新的协议修正了他们存在的一些图像识

别问题的不准确性。[7]中的协议是基于一种特殊设计的面部识别算法，是第一个安全系统在现实生活中的人脸识别应用。为了进一步优化其构建，Huang [8]等和 Blanton [9]等随后设计了基于同态的隐私保护指纹和虹膜识别协议加密和乱码电路(GC)，与[7]相比，实现了更低的开销。然而，这些针对两方提出的隐私保护协议大都是依靠于同态加密，因此计算成本比较高，效率较低，不适用于实际应用，并且，大多数的两方协议不能直接运用到云环境下，因此不能直接应用于外包模型。

与本文主题密切相关的工作是设计高效且保护隐私的云辅助生物识别协议。在这种情况下，DO 将数据库和相关查询操作外包给资源丰富的云服务器来降低服务器的计算压力。Wong 等[10]开发了一种高效的非对称标量积保持加密(APSE)来实现安全的密文数据库查询操作。随后，Yuan 和 Yu [11]指出，这种协议是不安全的，文献[10]中的协议没有考虑到云服务器和恶意用户相互勾结的情况。因此，如果存在这两方的恶意勾结，机密的生物特征数据将会被攻破。针对这一问题，他们提出了一种新的可以抵抗已知明文攻击的外包协议。不幸的是，Wang [12]等证明了文献[10]中的协议也是不安全的，他们的协议可以通过消除随机性或者利用欧氏距离结果来破解。为了修补这一安全漏洞，Wang [12]等设计了改进的协议 CloudBI-II。不过由于该协议依靠于大矩阵乘法，所以在实践中效率低下。为了实现更高的安全性，Zhang [13]等提出了一种新的协议 PTBI-II。新的协议技术利用中心极限定理构造扰动项，并巧妙地使得扰动项的总和满足正态分布。不过，该协议需要添加足够多的扰动项，因此会有大量的冗余，效率较低。继 Wang 等之后，Zhu [14]等也进一步完善了 Yuan 和 Yu [11]提出的协议中存在的安全漏洞。另外，Hu [15]等在文献[12]中的协议的基础上，提出了一种基于部分同态加密技术的隐私保护协议。但是 Liu 等[16]指出 Zhu 等和 Hu 等的设计存在固有缺陷，无法抵抗已知明文攻击。因此，他们通过在加密算法中插入阈值，提出了一种更安全的外包协议。

综上所述，设计一种安全高效的保护隐私的生物特征识别外包协议还有待进一步研究。

1.3. 动机和贡献

针对现有云辅助生物识别协议中存在的效率和安全性问题，本文提出了一种基于云辅助的生物识别技术，旨在构建云环境下更安全、更高效的生物识别协议。我们的贡献如下：

1) 我们利用基于误差加权哈希(EWH)的最近邻算法对生物特征数据集和查询数据进行预处理，生成可搜索的索引数据表和查询索引表。与线性扫描搜索相比，基于 EWH 的数据处理设计大大减少了搜索空间，从而加快了云上的查询响应速度。

2) 我们的协议对数据的加密采用的是矩阵变换等基本操作，操作简单，但是比现存的大多数云上的生物识别协议效率更高。

3) 我们的协议可以抵御已知明文攻击(KPA)，这在单服务器模型下已经是可知的云上的生物识别的协议能达到的很高的安全级别(尽管有些单服务器模型下的云上的生物识别协议声称其可以达到选择明文攻击(CPA)，但不幸的是，这些方案都被攻破了)。

4) 我们将 Liu 等的协议与 EWH 算法相结合，实现了大规模数据库上的云上的更快响应时间，具有更好的现实意义。

1.4. 文章布局

剩下的论文安排如下。第 2 章介绍了本文中的符号、系统架构图，安全模型以及我们的设计目标。在第 3 章中，我们回顾了基于误差加权哈希(EWH)的最近邻搜索算法以及介绍了其他的一些预备知识。随后，我们在第 4 节阐述了我们算法实施的主要细节。在第五章中，我们对协议的正确性、隐私性以及效率进行了分析。在第 6 章中，我们经过实验对协议的效率进一步进行了评价。最后，我们在第 7 章对本文进行了总结。

2. 系统模型及设计目标

2.1. 符号说明

在介绍设计细节之前，首先介绍一些相关的符号及其说明。具体说明如表 1 所示。

Table 1. Symbol description
表 1. 符号说明

符号	说明
D	DO 拥有的生物特征数据库 $D=\left\{\left(id_i,\boldsymbol{f}_i\right):\boldsymbol{f}_i=\left(f_{i1},\cdots f_{in}\right)\in\{0,1\}^n,1\le i\le s\right\}$
F'	生物特征数据库加密后的索引数据表
$\mathcal{C}$	最近邻候选列表 $\mathcal{C}$
$\boldsymbol{M}$	正交矩阵 $\boldsymbol{M}=\left(m_{i1},\cdots,m_{in}\right)\left(1\le i\le n\right)$
$\boldsymbol{C}_i$	$\boldsymbol{f}_i$ 由矩阵 $\boldsymbol{M}$ 加密后得到 $\boldsymbol{C}_i$
$\boldsymbol{C}$	$\boldsymbol{q}$ 由矩阵 $\boldsymbol{M}$ 加密后得到 $\boldsymbol{C}$
$\boldsymbol{R}_i$	每个最近邻候选者和 $\boldsymbol{q}$ 求内积得到 $\boldsymbol{R}_i$
$\boldsymbol{f}_i$	生物特征数据库中的第 i 个生物数据
$\boldsymbol{q}$	查询用户拥有的生物特征数据
k_i	公共参数 $pp=\{k_1,\cdots,k_\ell\}$ 中的一个随机密钥
s	中生物特征数据的个数
n	单个生物数据的维数
m	每个随机密钥 k_i 的长度
ℓ	k_i 的个数
e	误差界值
θ	相似度阈值
μ_i	赋值权重 $\{\mu_0,\cdots,\mu_e\}$ 里的一个权重值
α_i	扩维用到的随机正实数
β	扩维用到的随机正实数
r_i	安全因子，被用于数据扩维
τ	预设定的距离比较阈值

2.2. 系统架构

如图 1 所示，单服务器外包模型包括数据库拥有者(DO)、查询用户(QU)、云服务器(CS)这三个参与方：1) DO 拥有一个大规模生物特征数据库，代表权威可信任的机构，其目的是在不向 QU 和 CS 泄露真

实的数据库信息的情况下，辅助完成查询任务。2) QU 是查询用户，向 DO 发送查询请求并获得查询结果。3) CS 是据有强大计算和存储能力的云服务器，负责从密文数据库中找到与查询数据的最佳匹配数据点。但是，CS 可能会对生物特征数据库、查询数据以及查询结果感到好奇。

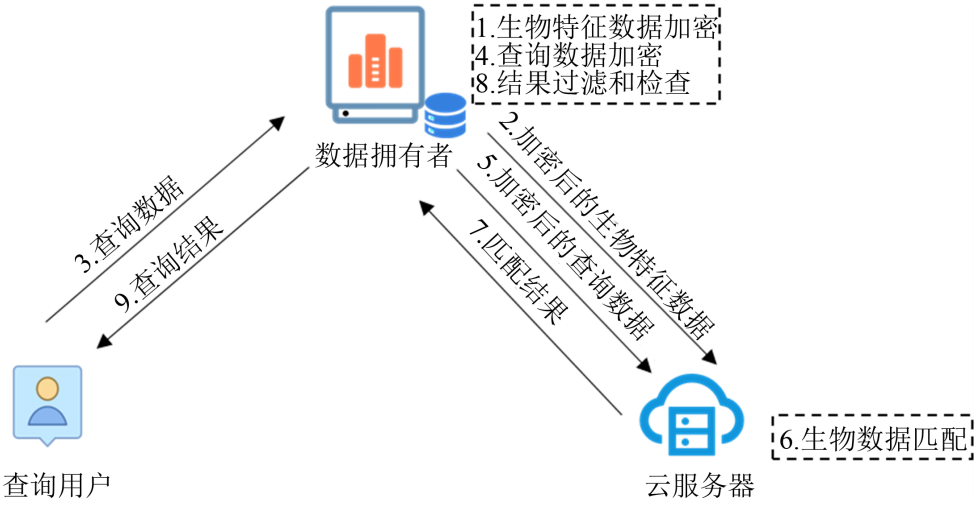

Figure 1. System architecture
图 1. 系统架构图

为了方便理解，我们对图 1 进行了详细说明：DO 拥有一个大规模生物特征数据库，该数据库在本文中表示为 $D=\left\{\left(id_i,\boldsymbol{f}_i\right):\boldsymbol{f}_i=\left(f_{i1},\cdots f_{in}\right)\in\{0,1\}^n,1\le i\le s\right\}$，其中 $\boldsymbol{f}_i$ 为索引 id_i 的生物特征数据，使用 $\boldsymbol{M}$ 表示矩阵。为了借助资源丰富的 CS 实现安全的数据库查询操作，DO 将加密后的生物特征数据 F' 给 CS (即图 1 中的步骤 1 和步骤 2)。当 QU 向 DO 发出查询请求时，DO 对查询 $\boldsymbol{q}$ 进行加密，并将其密文 $\boldsymbol{q}'$ 发送给 CS (即图 1 中的步骤 3、4、5)，CS 在 F' 中找到与 $\boldsymbol{q}'$ 的最佳匹配。系统模型并将该匹配的索引返回给 DO (即图 1 中的步骤 6 和 7)。最后，DO 查找 CS 返回的索引对应的明文数据，并使用相似度阈值检查该数据与查询数据之间的距离。根据计算结果，DO 向 QU 发送“Accept”或“Reject”(即图 1 中的步骤 8 和 9)。

2.3. 威胁和攻击模型

在云辅助的安全生物特征识别外包方案中，威胁模型有以下两种划分方式：根据云服务器的行为、云服务器的攻击能力进行划分。

根据云服务器的不同行为进行划分，威胁模型可以分为诚实但好奇模型(HBC)和恶意模型(MAL)：

1) HBC 模型：在这种威胁模型下，虽然云服务器会诚实地执行计算协议，但也会在协议执行过程从拥有的信息中推断用户的隐私数据。

2) MAL 模型：在该模型下，云服务器可能会不诚实地执行计算协议，并将错误的结果发送给查询用户。因此，在 MAL 模型下，用户应该可以验证结果的正确性，避免被恶意欺骗。

根据真实场景中不可信云服务器的不同攻击能力进行划分，本文主要考虑以下三种攻击模型：

1) 唯密文攻击(COA)模型[17]

在 COA 模型中，除了知道生物特征识别任务外，云服务器只能获得加密的数据库和加密的查询，并尝试恢复真实的生物特征数据库和真实的查询。

2) 已知样本攻击(KSA)模型[18]

攻击者不仅知道密文数据集，还可以捕获多个明文数据点。而明文数据点与密文数据点之间的对应关系是对于攻击者来说是隐藏的。这在数据库安全界被称为已知样本攻击(KSA)。

3) 已知明文攻击(KPA)模型[17]

在 KPA 模型中，云服务器知道生物特征识别任务和加密的数据库和加密的查询。并且它据有收集一部分明文数据点及其对应的加密数据点的能力。然后云服务器试图恢复真实的输入信息和真实的计算结果。

显然，KPA 模型中的云服务器是三种攻击模型中攻击能力最强的。如果一个生物特征识别方案可以抵抗 KPA 攻击，那么它也可以抵抗 COA 和 KSA 攻击。

2.4. 设计目标

根据我们的系统和威胁模型，我们的目标是构建一个 KPA 安全协议来实现安全存储并有效地识别云端的生物特征数据。确切地说，我们的设计应该满足以下要求。

1) 正确性：正确性指的是，如果云服务器诚实地执行分配的任务，协议就会对 QU 返回为“真”，当且仅当 QU 的生物特征信息与 DO 的生物特征数据库之间某个数据的距离小于某个预定义的足够小的阈值 τ。

2) 数据隐私：它涉及两个方面：生物特征数据库隐私和查询数据隐私。设计应该确保 DO 的数据库和 QU 的查询数据点都是 KPA 安全的。也就是说，即使对手可以自适应捕获 DO 的数据库或 QU 的查询数据点的一些明文－密文对，它不能区分挑战密文。

3) 高效：设计应该尽可能地减少查询－响应时间。

3. 预备知识

3.1. 基于误差加权哈希(EWH)的最近邻搜索算法

为了在汉明空间中快速搜索最近邻，Esmaeili [19]等在 2012 年提出了一种基于误差加权哈希(EWH)的高效搜索算法。与经典的局域敏感哈希(locality sensitive hash, LSH)算法[20]相比，基于 EWH 的算法考虑了误差哈希向量，并利用误差哈希向量生成一个精细化的候选列表，解决了 LSH 算法在查询的所有哈希向量碰巧都有错误并且失败概率随着最近邻距离的增加而增加的问题。此外，EWH 生成的候选列表的大小比 LSH 小得多。一般来说，基于 EWH 的算法分为两个步骤。首先，利用 EWH 对生物特征数据集进行预处理，生成索引数据表；其次，对于查询生物数据，计算误差及哈希值，并根据该哈希值从索引表中检索候选数据。表 2 中的算法 1 和表 3 中的算法 2 分别对这两个步骤进行了详细的回顾。基于 EWH 的算法的正确性可以通过以下引理来保证。

Table 2. EWH-based preprocessing algorithm
表 2. 基于 EWH 的预处理算法

算法 1：基于 EWH 的预处理算法
输入：一个生物特征数据集 $D=\left\{\left(id_i, \boldsymbol{f}_i\right): \boldsymbol{f}_i=\left(f_{i1},\cdots f_{in}\right)\in\{0,1\}^n, 1\le i\le s\right\}$，预处理过程中的密钥长度：$m$，密钥个数：$\ell$，一个哈希函数：$H:\{0,1\}^m\to\{0,1\}^\lambda$
输出：$2^\lambda\times\ell$ 的索引数据表 F'
1) 生成 ℓ 个随机密钥 $k_i=\overline{z_1z_2\cdots z_m}$，其中 $z_j\in\{1,2,\cdots,n\}$ $(1\le i\le\ell, 1\le j\le m)$

Continued

2) 将 $\boldsymbol{f}_p$ 添加到 F' 中的第 $H\left(\boldsymbol{f}_{pk_i}\right)=H\left(f_{pz_1}f_{pz_2}\cdots f_{pz_m}\right)$ 行第 i 列 $\left(1\le p\le s\right)$

3) 返回索引数据表 F'

Table 3. EWH-based search algorithm
表 3. 基于 EWH 的最近邻搜索算法

算法 2：基于 EWH 的最近邻搜索算法
输入：一个生物查询数据 $\boldsymbol{q}=\left(q_1,\cdots,q_n\right)$，索引数据表 F'，一组权重 $\left\{\mu_0,\cdots,\mu_e\right\}$，相似度阈值 θ
输出：查询 $\boldsymbol{q}$ 的最近邻
1) 初始化数据库中的所有数据分数为 0
2) 对于每一个 $i\left(1\le i\le \ell\right)$，计算 $\boldsymbol{q}_{k_i}:h_0^{(i)}=H\left(\boldsymbol{q}_{k_i}\right)$，并将(行，列)值等于 $\left\{h_0^{(i)},i\right\}$ 的所有数据赋值为 α_0
3) 对于误差 r 从 1 到 e，计算集合 $\left\{h_r^{(i)}:h_r^{(i)}=H\left(\boldsymbol{q}_{k_i}^{(r)}\right)\right\}$，其中 $\boldsymbol{q}_{k_i}^{(r)}$ 是与 $\boldsymbol{q}_{k_i}$ 有 r 比特误差的向量
4) 将(行，列)值等于 $\left\{h_r^{(i)},i\right\}$ 的所有数据分数增加 α_r
5) 当所有的 $\boldsymbol{f}_i$ 遍历完上述步骤后，都被赋予了一定的分数。我们选择分数大于 β 的数据作为最近邻候选者
6) 最后穷举所有的最近邻候选者，计算它们的汉明距离并返回最近邻

3.2. 正交矩阵

如果 $\boldsymbol{M}\boldsymbol{M}^{\mathrm{T}}=\boldsymbol{E}$ 或 $\boldsymbol{M}^{\mathrm{T}}\boldsymbol{M}=\boldsymbol{E}$，其中，$\boldsymbol{E}$ 为单位矩阵，$\boldsymbol{M}^{\mathrm{T}}$ 表示“矩阵 $\boldsymbol{M}$ 的转置矩阵”，则 $\boldsymbol{n}$ 阶实矩阵 $\boldsymbol{M}$ 称为正交矩阵，在本文中，我们将 $\boldsymbol{n}$ 阶正交矩阵 $\boldsymbol{M}$ 表示为：

$$\boldsymbol{M}=\begin{bmatrix} m_{11} & m_{12} & \cdots & m_{1n} \\ m_{21} & m_{22} & \cdots & m_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ m_{n1} & m_{n2} & \cdots & m_{nn} \end{bmatrix}$$

3.3. 生物特征相似度表示

对于查询向量 $\boldsymbol{q}=\left(q_1,\cdots,q_n\right)$ 以及数据库中的数据 $\boldsymbol{f}_i=\left(f_{i1},\cdots,f_{in}\right)$，若他们满足下面的式子，则可以认为 $\boldsymbol{q}$ 和 $\boldsymbol{f}_i$ 来自同一用户的两个指纹码匹配成功，查询 $\boldsymbol{q}$ 通过验证：

$$\left\|\boldsymbol{f}_i-\boldsymbol{q}\right\|=\sqrt{\sum_{j=1}^{n}\left(\boldsymbol{f}_{ij}-\boldsymbol{q}_j\right)^2}<\tau \tag{1}$$

4. 我们的外包协议设计

4.1. 概述

在本节中，我们将详细介绍我们协议的工作流程。我们的协议分为五个阶段：密钥生成阶段，数据

库加密阶段，数据库预处理及上传阶段，查询数据加密及上传阶段以及生物匹配阶段。我们协议通过这五个阶段的协同工作，实现了云上的生物识别的安全外包。首先，DO 生成一系列的密钥，密钥生成以后，DO 对数据库中的生物数据的进行扩维操作和加密操作，得到由密文组成的密文数据库 D'，之后便是对密文数据库 D' 的预处理，得到预处理后的索引数据表 F'，经过哈希函数碰撞后的数据之间相似度高的则被分到了一个哈希桶中，因此索引数据表 F' 中的每一个格中并不是只存在一个数据，而是存有很多个由哈希函数分类得到的数据之间相似度高的数据。当查询发出查询请求后，DO 根据相同的哈希函数对 $\boldsymbol{q}$ 进行处理并根据算法得到查询索引表 $F^{(q)}$。而云在接收到 F' 和 $F^{(q)}$ 后便开始进行生物匹配操作，根据查询索引表 $F^{(q)}$ 提供的行列值信息，云对 F' 对应位置中的所有的生物数据赋予不同的分数值，最后再统计每个生物数据的分数，并确定分数大于 β 的生物数据组成候选列表 $\mathcal{C}$。最后，对于候选者，云通过正交变换计算他们之间的内积 $\boldsymbol{R}_i$，若大于 0，查询 $\boldsymbol{q}$ 通过认证，云向 DO 返回“Accept”，否则返回“Reject”。

4.2. 我们协议的设计细节

4.2.1. 密钥生成阶段(DO-KeyGen)

DO 生成 s 个值为随机正实数的 α_i，s 个值为随机实数的 r_i，一个随机正实数 β，一个实数 τ，一个实数 θ，一组实数 $\{\mu_0,\cdots,\mu_e\}$，一个 $(n+5)\times(n+5)$ 维的随机正交矩阵以及 ℓ 个 k_i，其中，τ 为距离比较阈值，θ 为相似度阈值，$\{\mu_0,\cdots,\mu_e\}$ 为一组赋值权重，$k_i=\overline{z_1\ z_2\cdots z_m}$ 为公共参数 $pp=\{k_1,\cdots,k_\ell\}$ 里的一个随机数。

4.2.2. 数据库加密阶段(DO-DataEnc)

此阶段由 DO 执行。对于数据库中的任意数据 $\boldsymbol{f}_i$，我们对其的处理分为两步：

1) 扩维：对于任意的 $\boldsymbol{f}_i=(f_{i1},\cdots,f_{in})\in\{0,1\}^n,(1\le i\le s)$，DO 将其扩维到 $(n+5)$ 维，扩维后的形式为：

$$\boldsymbol{f}_i'=\left[\alpha_i f_{i1},\alpha_i f_{i2},\cdots,\alpha_i f_{in},-\frac{1}{2}\alpha_i\sum_{j=1} f_{ij}^2,\alpha_i,\frac{1}{2}\alpha_i\tau^2,r_i,0\right]$$

2) 数据库加密：DO 通过计算 $\boldsymbol{C}_i=\boldsymbol{f}_i'\boldsymbol{M}$ 对扩维后的数据进一步加密。

4.2.3. 数据库预处理及上传阶段(DO-DataPre)

此阶段由 DO 执行。给定一个生物特征数据集 $D=\left\{(id_i,\boldsymbol{f}_i):\boldsymbol{f}_i=(f_{i1},\cdots,f_{in})\in\{0,1\}^n,1\le i\le s\right\}$，DO 通过调用基于 EWH 的预处理算法(表 2 中的算法 1)，并使用哈希函数 $H:\{0,1\}^m\to\{0,1\}^\lambda$ 和 ℓ 个随机公共参数 $pp=\{k_1,\cdots,k_\ell\}$ 对数据库进行预处理得到如表 4 所示的加密后的规格为 $2^\lambda\times\ell$ 的索引数据表 F'。

Table 4. The hash-based index table F'
表 4. 基于哈希的索引数据表 F'

列 \ 行	k_1	…	k_1
	1	…	ℓ
0	$\left\{\left(id_{i_{(0,1)}},\boldsymbol{f}_{i_{(0,1)}}\right)\right\}$	…	$\left\{\left(id_{i_{(0,\ell)}},\boldsymbol{f}_{i_{(0,\ell)}}\right)\right\}$
⋮	⋮	⋮	⋮
$2^\lambda-1$	$\left\{\left(id_{i_{(2^\lambda-1,1)}},\boldsymbol{f}_{i_{(2^\lambda-1,1)}}\right)\right\}$	…	$\left\{\left(id_{i_{(2^\lambda-1,\ell)}},\boldsymbol{f}_{i_{(2^\lambda-1,\ell)}}\right)\right\}$

4.2.4. 查询数据加密及上传阶段(Qu-DataEnc)

此阶段由 DO 执行并且是基于查询的一次性工作。具体可以分为三部分：

1) 生成查询索引表：当 QU 处理一个查询请求时，通过公共参数 $pp=\{k_1,\cdots,k_\ell\}$，他/她首先执行表 5 中的算法 3 生成一个查询索引表 $F^{(q)}$。

2) 查询数据扩维：对于任意的 $\boldsymbol{q}=(q_1,\cdots,q_n)$，DO 将其扩维 $(n+5)$ 维，扩维后的形式为：

$$\boldsymbol{q}'=\left[\beta q_1,\beta q_2,\cdots,\beta q_n,\beta,-\frac{1}{2}\beta\sum_j^n b_j^2,\beta,0,r\right]$$

3) 查询数据加密：QU 通过计算 $\boldsymbol{C}=\boldsymbol{q}'\boldsymbol{M}$ 对扩维后的数据进一步加密。

Table 5. EWH-based query index generation algorithm
表 5. 基于 EWH 的查询索引表生成算法

算法 3：基于 EWH 的查询索引表生成算法
输入：一个查询生物数据：$\boldsymbol{q}$， $pp=\{k_1,\cdots,k_\ell\}$ 和误差界值 e
输出：查询索引表 $F^{(q)}$
1) 初始化 $F^{(q)}=0$
2) 对于每一个 $i(1\le i\le l)$，计算 $\boldsymbol{q}_{k_i}:h_0^{(i)}=H(\boldsymbol{q}_{k_i})$，并在 $F^{(q)}$ 的对应位置存储三元组 $\{(h_0^{(i)},i,0)\}$
3) 对于误差 r 从 1 到 e，计算集合 $\{h_r^{(i)}:h_r^{(i)}=H(\boldsymbol{q}_{k_i}^{(r)})\}$，其中 $\boldsymbol{q}_{k_i}^{(r)}$ 是与 $\boldsymbol{q}_{k_i}$ 有 r 比特误差的向量，并在 $F^{(q)}$ 的对应位置处存储三元组 $\{(h_r^{(i)},i,r)\}$
4) 返回 $F^{(q)}$

4.2.5. 生物匹配阶段(CS-DataMa)

此阶段由 CS 完成，该阶段分为两步：

1) 云服务器执行表 6 中的算法 4 得到查询数据 $\boldsymbol{q}$ 的最近邻候选列表。

2) 对于得到的最近邻候选列表中的数据，CS 进行按照下面的公式计算生物数据以及候选者之间的密文内积来得到距离的远近关系。

$$\boldsymbol{R}_i=\boldsymbol{C}_i\cdot\boldsymbol{C} \tag{2}$$

如果 $\boldsymbol{R}_i>0$，我们则向 DO 返回“Accept”，否则向其返回“Reject”。

Table 6. EWH-based nearest neighbor candidate list generation algorithm
表 6. 基于 EWH 的最近邻候选列表生成算法

算法 4：基于 EWH 的最近邻候选列表生成算法
输入：加密后的索引数据表 F'，查询索引表 $F^{(q)}$，一组权重 $\{\mu_0,\cdots,\mu_e\}$，一个相似度阈值 β
输出：最近邻候选列表 $\mathcal{C}$
1) 对于每个生物数据，初始化其分数为 0

Continued

2) 根据在 $F^{(q)}$ 的对应位置处存储三元组 $\left\{\left(h_r^{(i)},i,r\right)\right\}$ 的数值，我们将 F' 中(行，列)值为 $\left(h_r^{(i)},i\right)$ 的所有生物数据的分数增加 α_r

3) 赋值完成后，选择所有分数大于 β 的生物数据作为最近邻候选组成列表 $\mathcal{C}$

4) 返回最近邻候选列表 $\mathcal{C}$

5. 协议分析

5.1. 正确性分析

假设有两个向量 $\boldsymbol{a}=\left(a_1,\cdots,a_n\right)$， $\boldsymbol{b}=\left(b_1,\cdots,b_n\right)$，这两个向量内存储着生物数据。在线性代数中，我们将 $\boldsymbol{a}\to\boldsymbol{aM}$， $\boldsymbol{b}\to\boldsymbol{bM}$ 称为正交变换。因为 $\boldsymbol{M}$ 是正交矩阵，因此它满 $\boldsymbol{aM}$， $\boldsymbol{bM}$ 满足正交矩阵变换的性质，即正交矩阵变换后的两个向量的内积保持不变。

换句话说，对于存在的两个向量 $\boldsymbol{a}$、$\boldsymbol{b}$，我们可以将他们的内积表示为 $\boldsymbol{a}\cdot\boldsymbol{b}=\boldsymbol{a}\times\boldsymbol{b}^{\mathrm{T}}$。对于给定的正交矩阵 $\boldsymbol{M}$，有以下的性质：

$$\begin{aligned}\boldsymbol{aM}\cdot\boldsymbol{bM}&=\boldsymbol{aM}\times\left(\boldsymbol{bM}\right)^{\mathrm{T}}\\&=\boldsymbol{aMM}^{\mathrm{T}}\boldsymbol{b}^{\mathrm{T}}\\&=\boldsymbol{a}\times\boldsymbol{b}^{\mathrm{T}}\\&=\boldsymbol{a}\cdot\boldsymbol{b}\end{aligned}\tag{3}$$

根据这个性质，我们可以推导出：

$$\begin{aligned}\boldsymbol{R}_i&=\boldsymbol{C}_i\cdot\boldsymbol{C}\\&=\boldsymbol{f}_i'\boldsymbol{M}\times\boldsymbol{qM}\\&=\boldsymbol{f}_i'\times\boldsymbol{q}'\\&=\alpha_i\beta\left(\sum_{j=1}^{n}f_{ij}q_j-\frac{1}{2}\sum_{j=1}^{n}f_{ij}^2-\frac{1}{2}\sum_{j=1}^{n}q_j^2+\frac{1}{2}\tau^2\right)\\&=\frac{1}{2}\alpha_i\beta\left[\tau^2-\sum_{j=1}^{n}\left(f_{ij}-q_j\right)^2\right]\end{aligned}\tag{4}$$

如果 $\boldsymbol{R}_i>0$，则有 $\sum_{j=1}^{n}\left(f_{ij}-q_j\right)^2<\tau^2$，即在数据库中，有和查询数据 $\boldsymbol{q}$ 的距离小于 τ 的生物数据 $\boldsymbol{f}_i$，便满足了生物识别通过的条件，查询 $\boldsymbol{q}$ 便通过了验证。

5.2. 隐私性分析

在我们协议中，云服务器是诚实且好奇的，在本节中，我们将证明我们设计的协议可以抵御已知明文攻击(KPA)。下面是我们证明过程。

我们假设在攻击实验中，攻击者可以获得 d 对线性独立的明密文对 $\left\{\left\langle\boldsymbol{f}_i,\boldsymbol{C}_i\right\rangle\middle|i=1,\cdots,d\right\}$，给定某个挑战者的 $\boldsymbol{q}^*$ 的密文 $\boldsymbol{C}^*$，攻击者试图通过已知信息来恢复 $\boldsymbol{q}^*$。根据前文定义，$\boldsymbol{C}_i=\boldsymbol{f}_i'\boldsymbol{M}$。也就是说，攻击者可以得到由 $d\left(n+5\right)$ 个方程构成的方程组。

显然，在上述方程组中，$c_{ij}\left(1\le i\le d,1\le j\le n+5\right)$ 和 $f_{ij}\left(1\le i\le d,1\le j\le n+5\right)$ 是已知的系数，而 $m_{ij}\left(1\le i,j\le n+5\right)$ 是未知的，并且由于 α_i， r_i 都是随着 $\boldsymbol{f}_i$ 变化的，相当于一次加密。因此在上述方程组中，至少存在 $d\left(n+5\right)^2$ 个未知数，但是只能建立 $d\left(n+5\right)$ 个方程。并且，对于每组新的明密文对，至少

引入$(n+5)$个新方程和$d(n+5)^2$个新未知数。因为未知数的数量总比方程的数量多，故方程组至少存在指数级个不同的解向量。因此，攻击者不能得到唯一确定的解$\boldsymbol{M}$。

至于查询向量$\boldsymbol{q}$的隐私性，分析方法与上文相同，这里就不在赘述。综上可得，攻击者并没有获得任何已知之外的信息，我们的协议完全能抵御已知明文攻击。

$$\begin{cases}
\alpha_1 f_{11}m_{11}+\cdots+\alpha_1 f_{1n}m_{n1}+\left(-\frac{1}{2}\alpha_1\sum_{j=1}^{n}f_{1j}^2\right)m_{(n+1)1}+\alpha_1 m_{(n+2)1}+\frac{1}{2}\alpha_1\tau^2 m_{(n+3)1}+r_1 m_{(n+4)1}=c_{11}\\
\alpha_1 f_{11}m_{12}+\cdots+\alpha_1 f_{1n}m_{n2}+\left(-\frac{1}{2}\alpha_1\sum_{j=1}^{n}f_{1j}^2\right)m_{(n+1)2}+\alpha_1 m_{(n+2)2}+\frac{1}{2}\alpha_1\tau^2 m_{(n+3)2}+r_1 m_{(n+4)2}=c_{12}\\
\vdots\\
\alpha_1 f_{11}m_{1(n+5)}+\cdots+\alpha_1 f_{1n}m_{n(n+5)}+\left(-\frac{1}{2}\alpha_1\sum_{j=1}^{n}f_{1j}^2\right)m_{(n+1)(n+5)}+\alpha_1 m_{(n+2)(n+5)}+\frac{1}{2}\alpha_1\tau^2 m_{(n+3)(n+5)}+r_1 m_{(n+4)(n+5)}=c_{1(n+5)}\\
\alpha_2 f_{21}m_{11}+\cdots+\alpha_2 f_{2n}m_{n1}+\left(-\frac{1}{2}\alpha_2\sum_{j=1}^{n}f_{2j}^2\right)m_{(n+1)1}+\alpha_2 m_{(n+2)1}+\frac{1}{2}\alpha_2\tau^2 m_{(n+3)1}+r_2 m_{(n+4)1}=c_{21}\\
\alpha_2 f_{21}m_{12}+\cdots+\alpha_2 f_{2n}m_{n2}+\left(-\frac{1}{2}\alpha_2\sum_{j=1}^{n}f_{2j}^2\right)m_{(n+1)2}+\alpha_2 m_{(n+2)2}+\frac{1}{2}\alpha_2\tau^2 m_{(n+3)2}+r_2 m_{(n+4)2}=c_{22}\\
\vdots\\
\alpha_d f_{d1}m_{11}+\cdots+\alpha_d f_{dn}m_{n1}+\left(-\frac{1}{2}\alpha_d\sum_{j=1}^{n}f_{dj}^2\right)m_{(n+1)1}+\alpha_d m_{(n+2)1}+\frac{1}{2}\alpha_d\tau^2 m_{(n+3)1}+r_d m_{(n+4)1}=c_{d1}\\
\alpha_d f_{d1}m_{12}+\cdots+\alpha_2 f_{2n}m_{n2}+\left(-\frac{1}{2}\alpha_d\sum_{j=1}^{n}f_{dj}^2\right)m_{(n+1)2}+\alpha_d m_{(n+2)2}+\frac{1}{2}\alpha_d\tau^2 m_{(n+3)2}+r_d m_{(n+4)2}=c_{d2}\\
\vdots\\
\alpha_d f_{d1}m_{1(n+5)}+\cdots+\alpha_d f_{dn}m_{n(n+5)}+\left(-\frac{1}{2}\alpha_d\sum_{j=1}^{n}f_{dj}^2\right)m_{(n+1)(n+5)}+\alpha_d m_{(n+2)(n+5)}+\frac{1}{2}\alpha_d\tau^2 m_{(n+3)}+r_d m_{(n+4)1}=c_{d(n+5)}
\end{cases} \tag{5}$$

5.3. 效率分析

就效率而言，我们的方案与现有的大多数方案相比，拥有更小的计算开销。在 DO 对数据库的数据预处理以及加密上传的阶段，主要的计算开销是数据库预处理过程中的哈希过程以及矩阵的运算，不过由于很多的代码语言都有集成的哈希函数库，因此在实际的操作过程中并没有占用大量时间；而我们的矩阵操作也是最常见的操作之一，又因为我们只有正交矩阵变换以及求内积操作，而且我们只求最近邻候选列表(数目较少)的内积，因此在时间上比之前的工作由极大的减少。

6. 实验评估

为了评估我们协议实际性能，本节通过实验比较了 Zhu [14]等的协议和我们的协议在各个阶段的时间成本。在一台 Intel Core i7-9700T 2.00GHz CPU 和 16.0 GB RMA 的计算机上模拟 DO 端。为 CS 端设置 10 个节点，每个节点具有与 DO 端相同的硬件配置。此外我们使用随机向量，每个生物数据的维度是 1024 维。

6.1. 生物预处理及加密阶段

该阶段由 DO 执行，代表 DO 对数据库数据处理的时间成本。如图 2 所示，我们的数据库规模从 2000

到 10,000，因为我们的安全性比 Zhu [14]等的协议安全性要高，又因为对数据有预处理阶段，因此我们在对数据库进行加密处理时间略高于 Zhu [14]等的协议。

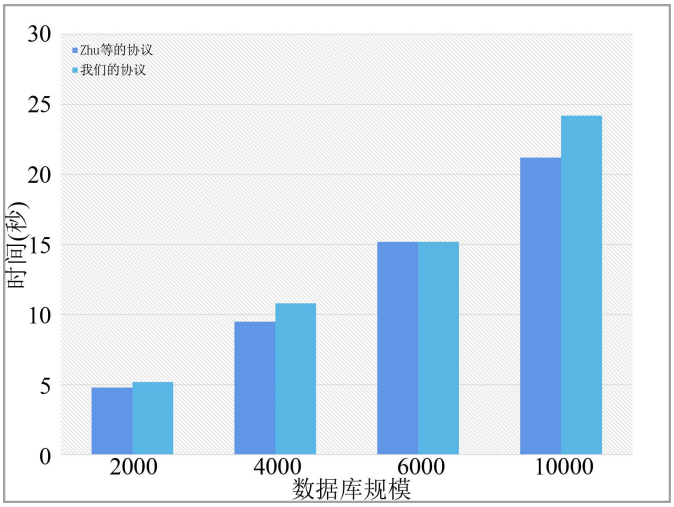

Figure 2. Comparison on the time costs of the data preprocessing and encryption stage
图 2. 生物数据预处理及加密阶段时间比较

6.2. 查询数据加密阶段

该阶段由 DO 执行。如图 3 所示，我们比较了当数据库的规模为 2000 时，在该阶段下查询数据的数目从 1 到 30 的变化范围内的查询加密所需要的时间。由于我们的协议中要根据查询数据生成查询索引表，因此时间略高于 Zhu [14]等的协议。但这个时间差是很小很小的，只有千分之几，因此在实际应用中不会造成时间上的影响。

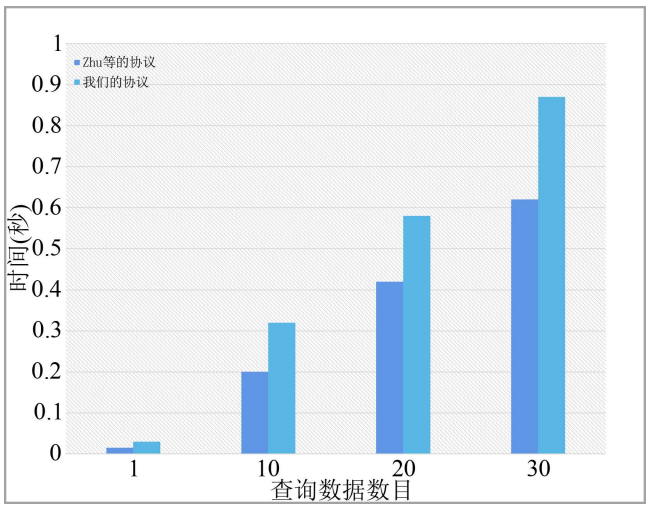

Figure 3. Comparison on the time costs of the query encryption stage
图 3. 查询数据加密阶段时间比较

6.3. 生物预处理及加密阶段

该阶段由云服务器(CS)执行，如图 4 所示，我们的数据库规模从 2000 到 10,000，显然，我们的协议

在云上的生物数据匹配阶段所需要的时间比 Zhu [14]等的协议所需要的时间小的多，应该说是了极大的提升。

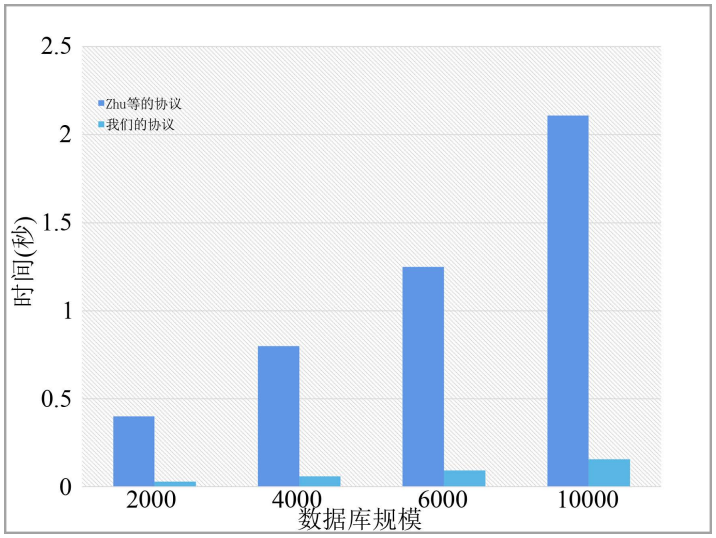

Figure 4. Comparison on the time costs of the data matching stage
图 4. 生物数据匹配阶段时间比较

7. 总结

在本文中，我们提出了一种新的云上的高效生物隐私保护识别协议，通过理论分析以及实验验证，我们设计的协议在生物匹配阶段只需要很少的时间，整体具有更高的效率，更适合于大型数据库，具有更好的现实意义。

参考文献

[1] Aparna, P. and Kishore, P. (2019) Biometric-Based Efficient Medical Image Watermarking in E-Healthcare Application. *IET Image Processing*, **13**, 421-428. https://doi.org/10.1049/iet-ipr.2018.5288

[2] Kraovec, A., Baldini, G. and Pejovi, V. (2021) Opposing Data Exploitation: Behaviour Biometrics for Privacy-Preserving Authentication in IoT Environments. *Proceedings of the* 16*th International Conference on Availability, Reliability and Security*, Vienna, August 2021, 1-7. https://doi.org/10.1145/3465481.3470101

[3] Das, A.K. (2017) A Secure and Effective Biometric-Based User Authentication Scheme for Wireless Sensor Networks Using Smart Card and Fuzzy Extractor. *International Journal of Communication Systems*, **30**, e2933. https://doi.org/10.1002/dac.2933

[4] Baltana, S.F., Ruiz-Sarmiento, J.R. and Gonzalez-Jimenez, J. (2020) A Face Recognition System for Assistive Robots. *APPIS* 2020: 3*rd International Conference on Applications of Intelligent Systems*, New York, January 2020, 1-6. https://doi.org/10.1145/3378184.3378225

[5] Erkin, Z., Franz, M., Guajardo, J., Katzenbeisser, S., Lagendijk, I. and Toft, T. (2009) Privacy-Preserving Face Recognition. In: Goldberg, I. and Atallah, M.J., Eds., *Privacy Enhancing Technologies*. *PETS* 2009. *Lecture Notes in Computer Science*, Springer, Berlin. https://doi.org/10.1007/978-3-642-03168-7_14

[6] Sadeghi, A.R., Schneider, T. and Wehrenberg, I. (2009) Efficient Privacy-Preserving Face Recognition. *Proceedings of the* 12*th International Conference on Information Security and Cryptology*, Berlin, 2-4 December 2009, 229-244. https://doi.org/10.1007/978-3-642-14423-3_16

[7] Osadchy, M., Pinkas, B., Jarrous, A., *et al*. (2010) SCiFI—A System for Secure Face Identification. *Proceedings of the* 2010 *IEEE Symposium on Security and Privacy*, Oakland, 16-19 May 2010, 239-254. https://doi.org/10.1109/SP.2010.39

[8] Huang, Y., Malka, L., Evans, D., *et al*. (2011) Efficient Privacy-Preserving Biometric Identification. *Proceedings of the Network and Distributed System Security Symposium*, California, 14-19 April 2013, 319-323.

[9] Blanton, M. and Gasti, P. (2011) Secure and Efficient Protocols for Iris and Fingerprint Identification. In: Atluri, V. and Diaz, C. Eds., *Computer Security—ESORICS* 2011, Springer, Berlin, 190-209. https://doi.org/10.1007/978-3-642-23822-2_11

[10] Wong, W., Cheung, D., Kao, B., *et al*. (2009) Secure kNN Computation on Encrypted Databases. *Proceedings of the* 2009 *ACM SIGMOD International Conference on Management of Data*, New York, June 2009, 139-152. https://doi.org/10.1145/1559845.1559862

[11] Yuan, J. and Yu, S. (2013) Efficient Privacy-Preserving Biometric Identification in Cloud Computing. 2013 *Proceedings IEEE INFOCOM*, Turin, 14-19 April 2013, 2652-2660. https://doi.org/10.1109/INFCOM.2013.6567073

[12] Wang, Q., Hu, S., Ren, K., *et al*. (2015) CloudBI: Practical Privacy-Preserving Outsourcing of Biometric Identification in the Cloud. *Proceedings of the* 20*th European Symposium on Research in Computer Security*, Vienna, 18 November 2015, 186-205. https://doi.org/10.1007/978-3-319-24177-7_10

[13] Zhang, C., Zhu, L. and Chang, X. (2017) PTBI: An Efficient Privacy-Preserving Biometric Identification Based on Perturbed Term in the Cloud. *Information Sciences*, **409-410**, 56-67. https://doi.org/10.1016/j.ins.2017.05.006

[14] Zhu, L., Zhang, C., Xu, C., *et al*. (2018) An Efficient and Privacy-Preserving Biometric Identification Scheme in Cloud Computing. *IEEE Access*, **6**, 19025-19033. https://doi.org/10.1109/ACCESS.2018.2819166

[15] Hu, S., Li, M.H., *et al*. (2018) Outsourced Biometric Identification with Privacy. *IEEE Transactions on Information Forensics and Security*, **13**, 2448-2463. https://doi.org/10.1109/TIFS.2018.2819128

[16] Liu, C., Hu, X., Zhang, Q., *et al*. (2019) An Efficient Biometric Identification in Cloud Computing with Enhanced Privacy Security. *IEEE Access*, **7**, 105363-105375. https://doi.org/10.1109/ACCESS.2019.2931881

[17] Delfs, H. and Knebl, H. (2007) Introduction to Cryptography. 2nd Edition, Springer, Berlin. https://doi.org/10.1007/3-540-49244-5

[18] Liu, K., Giannella, C. and Kargupta, H. (2006) An Attacker's View of Distance Preserving Maps for Privacy Preserving Data Mining. *European Conference on Principles of Data Mining and Knowledge Discovery*, Berlin, 18-22 September 2006, 297-308. https://doi.org/10.1007/11871637_30

[19] Esmaeili, M.M., Ward, R.K. and Fatourechi, M. (2012) A Fast Approximate Nearest Neighbor Search Algorithm in the Hamming Space. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **34**, 2481-2488. https://doi.org/10.1109/TPAMI.2012.170

[20] Gionis, A., Indyk, P., Motwani, R., *et al*. (1999) Similarity Search in High Dimensions via Hashing. *Proceedings of the* 25*th International Conference on Very Large Data Bases*, San Francisco, 7 September 1999, 518-529.