Hindawi Publishing Corporation
EURASIP Journal on Information Security
Volume 2010, Article ID 819376, 11 pages
doi:10.1155/2010/819376

Research Article

A Simple Scheme for Constructing Fault-Tolerant Passwords from Biometric Data

Vladimir B. Balakirsky and A. J. Han Vinck

Institute for Experimental Mathematics, University of Duisburg-Essen, 45326 Essen, Germany

Correspondence should be addressed to A. J. Han Vinck, vinck@iem.uni_due.de

Received 6 April 2010; Revised 19 July 2010; Accepted 18 October 2010

Academic Editor: Bülent Sankur

We present a simple combinatorial construction for the mapping of the biometric vectors to short strings, called the passwords. A verifier has to decide whether a given vector can be considered as a corrupted version of the original biometric vector whose password is known or not. The evaluations of the compression factor, the false rejection/acceptance rates, are derived, and an illustration of a possible implementation of the verification algorithm for the DNA data is presented.

1. Introduction

Let us consider the data transmission scheme in Figure 1. The source generates a vector $\mathbf{b} \in \{0,1\}^N$ containing the outcomes of the measurements of some biometric parameters of a user. This vector is encoded as the vector $\mathrm{pw}(\mathbf{b}) \in \{0,1\}^K$, called the password of the user, which is stored in the database under the user's name. The password is read from the database upon request and given to the verifier together with the vector $\mathbf{b}' \in \{0,1\}^N$ generated by some source. The verifier has to check whether the vector $\mathbf{b}'$ can be considered as a corrupted version of the vector $\mathbf{b}$ (accept) or not (reject). The decision can be expressed as the value of a Boolean function $\varphi(\mathrm{pw}(\mathbf{b}), \mathbf{b}') \in \{\mathrm{Acc}, \mathrm{Rej}\}$, and the formal specification of the procedure is an assignment of the functions

$$\begin{aligned} \mathrm{pw}\colon \{0,1\}^N &\longrightarrow \{0,1\}^K, \\ \varphi\colon \{0,1\}^K \times \{0,1\}^N &\longrightarrow \{\mathrm{Acc}, \mathrm{Rej}\}. \end{aligned} \tag{1}$$

The scheme in Figure 1 shows a conventional biometric authentication system [1]. We apply our coding theory approaches [2–4] to find solutions for the following setup.

(1) The length of the binary representation of the password $\mathrm{pw}(\mathbf{b})$ is much less than the length of the vector $\mathbf{b}$, that is, $K \ll N$.

(2) The probability distribution over the vectors $\mathbf{b}$ is not given, and the performance is analyzed for the worst assignment of the input data.

(3) The function pw is a deterministic function. Therefore, the distribution of common randomness between the encoder and the verifier, which is a feature of randomized hashing schemes, is not relevant in our case. The probabilities of the incorrect verifier's decisions are computed over the noise ensemble.

(4) If the vector $\mathbf{b}'$ is a corrupted version of the vector $\mathbf{b}$, then the level of noise is measured by the absolute value of the difference of the Hamming weights of the vectors $\mathbf{b}$ and $\mathbf{b}'$.

Notice that many authors addressed the problem of constructing fault-tolerant passwords, and the list [5–9] is far from being complete. The main difference of the setup analyzed in our correspondence is the point that the scheme does not require randomization. As a result, our approach can essentially simplify an implementation and simultaneously cause some security problems, which are discussed below.

As pw is a deterministic function and the compression factor N/K is large; an attacker, who knows $\mathrm{pw}(\mathbf{b})$ and wants to pass through the verification stage with the acceptance

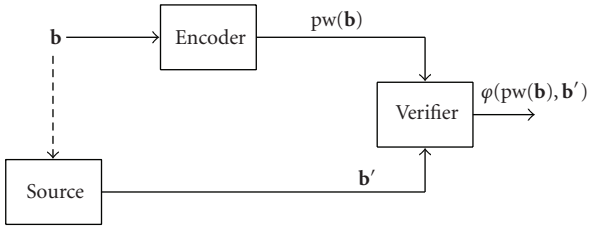

Figure 1: The data transmission scheme designed for the authentication of a user, where $\mathbf{b}, \mathbf{b}' \in \{0,1\}^N$, $\mathrm{pw}(\mathbf{b}) \in \{0,1\}^K$, and $\varphi(\mathrm{pw}(\mathbf{b}), \mathbf{b}') \in \{\mathrm{Acc}, \mathrm{Rej}\}$.

decision, can easily succeed by generating a vector $\mathbf{b}'$ such that $\mathrm{pw}(\mathbf{b}') = \mathrm{pw}(\mathbf{b})$. Therefore, the scheme is not secure in the same sense as the system, which uses the PIN codes of the users: if the PIN code is stolen and the attacker can enter it into the system, then he succeeds. Thus, one needs to encrypt passwords, and our construction can serve as a preliminary step for conventional schemes. Another kind of security is the possibility of guessing the biometric vector on the basis of its password. If the password is the weight of the vector (which is a special case of our construction), then the probability of the correct guess is very small for most of the vectors. However, the weights 0 and n uniquely determine the vector. Thus, meaning the points above, the secrecy of the scheme can be not sufficient for its separate use in practical biometric systems. However, a very large compression factor, very small probabilities of the incorrect verifier's decisions, and very small complexity of the implementation of our scheme that can be attained simultaneously make such a scheme attractive. In particular, we can recommend it for information transmission systems where the verifier has to make only the rejection decision for the vectors $\mathbf{b}'$ that definitely cannot be considered as corrupted versions of the original biometrical vector. The final decision for the vectors that passed through this test is made by some other tools in this case.

2. Model for the Noise of Observations

We will assume that

$$N = Tn, \tag{2}$$

where T, n are positive integers and n is even. Represent the vectors $\mathbf{b}$ and $\mathbf{b}'$ as concatenations of T blocks of length n and write

$$\mathbf{b} = (\mathbf{b}_1, \ldots, \mathbf{b}_T), \qquad \mathbf{b}' = (\mathbf{b}'_1, \ldots, \mathbf{b}'_T), \tag{3}$$

where $\mathbf{b}_t, \mathbf{b}'_t \in \{0,1\}^n$ for all $t = 1, \ldots, T$. The blocks will be processed in parallel, and we describe the model for the probabilistic transformation of an input block $\mathbf{b}$ to the received block $\mathbf{b}'$ having the weights

$$w = \mathrm{wt}(\mathbf{b}), \quad w' = \mathrm{wt}(\mathbf{b}'). \tag{4}$$

If the received block is generated independently of the input block, we assume that w' is the value of a random variable having the binomial probability distribution

$$(\mathrm{B}(w'),\ w' \in \{0, \ldots, n\}), \tag{5}$$

where

$$\mathrm{B}(w') = \binom{n}{w'} 2^{-n}. \tag{6}$$

If the received block is a corrupted version of the input block, we assume that w' is the value of a random variable having the given conditional probability distribution

$$(\Omega(w' \mid w),\ w' \in \{0, \ldots, n\}). \tag{7}$$

Examples. (1) Binary symmetric channel.

Suppose that the vector $\mathbf{b}'$ is the outcome of a binary symmetric channel having the crossover probability $p \in (0, 1/2)$ when the vector $\mathbf{b}$ was sent. Then,

$$\Omega(w' \mid w) = \sum_{j=0}^{w'} \binom{n-w}{j} p^j (1-p)^{n-w-j} \cdot \binom{w}{w'-j} p^{w-w'+j} (1-p)^{w'-j}. \tag{8}$$

(2) The insertion/deletion channel.

Let $\varepsilon \in (0, 1/2)$. For all $k \in \{0, \ldots, n\}$, let

$$\binom{n}{k} \varepsilon^k (1-\varepsilon)^{n-k} \tag{9}$$

be the probability that $n-k$ components of the vector $\mathbf{b}$ are noiselessly transmitted, while the remaining k positions are filled with an arbitrary vector generated with the probability $\binom{n}{k} 2^{-n}$. Then, $\Omega(w' \mid w)$ is expressed by (8) with $\varepsilon/2$ substituted for p.

In the following numerical illustrations, we assume that the conditional probabilities $\Omega(0 \mid w), \ldots, \Omega(n \mid w)$ are defined by (8).

Discussion over the Model. As the input vector $\mathbf{b}$ is fixed, the vector $\mathbf{w}$ is also fixed. Given an acceptance set, the probability that the verifier makes an incorrect rejection decision can be computed after the conditional probabilities $\Omega(0 \mid w), \ldots, \Omega(n \mid w)$ are specified. However, one cannot compute the probability that the verifier makes an incorrect acceptance decision for the best strategy of an attacker, unless the probability distribution over the input vectors (which determines the probability distribution over passwords) is given. We can only compute this probability for a blind attacker, who generates the vector $\mathbf{b}'$ by flipping a fair coin, which results in the binomial probability distribution over

passwords $\mathbf{w}'$. Then, computations become equivalent to the estimation of the ratios of the cardinalities of the sets of input vectors with coinciding passwords and 2^{-Tn}. Notice that this estimation is a typical problem when universal hashing schemes are studied [10]. Since our scheme is oriented to the preprocessing of the pairs of received vectors, the performance of the scheme for a blind attacker is also of interest for practical biometric applications.

3. Description of the Verification Scheme

Given the vectors $\mathbf{b} = (\mathbf{b}_1, \ldots, \mathbf{b}_T)$ and $\mathbf{b}' = (\mathbf{b}'_1, \ldots, \mathbf{b}'_T)$, let $\mathrm{pw}(\mathbf{b}) = \mathbf{w}$ and $\mathrm{pw}(\mathbf{b}') = \mathbf{w}'$, where components of the vectors $\mathbf{w}$ and $\mathbf{w}'$ are defined as $w_t = \mathrm{wt}(\mathbf{b}_t)$ and $w'_t = \mathrm{wt}(\mathbf{b}'_t)$ for all $t = 1, \ldots, T$. Thus,

$$\begin{aligned} \mathrm{pw}(\mathbf{b}) &= (\mathrm{wt}(\mathbf{b}_1), \ldots, \mathrm{wt}(\mathbf{b}_n)), \\ \mathrm{pw}(\mathbf{b}') &= (\mathrm{wt}(\mathbf{b}'_1), \ldots, \mathrm{wt}(\mathbf{b}'_n)). \end{aligned} \tag{10}$$

For all vectors $\mathbf{w} \in \{0, \ldots, n\}^T$, let $\mathcal{D}^{(T)}(\mathbf{w}) \subseteq \{0, \ldots, n\}^T$ be a subset of vectors of the length T whose components belong to the alphabet $\{0, \ldots, n\}$, which is called the acceptance set and associated with the following decoding rule:

$$\varphi(\mathbf{w}, \mathbf{b}') = \begin{cases} \mathrm{Acc}, & \text{if } \mathbf{w}' \in \mathcal{D}^{(\mathrm{T})}(\mathbf{w}), \\ \mathrm{Rej}, & \text{if } \mathbf{w}' \notin \mathcal{D}^{(\mathrm{T})}(\mathbf{w}). \end{cases} \tag{11}$$

The verification scheme is illustrated in Figure 2.

Notice that the compression factor, defined as the ratio of the length of the biometric vector and the length of the corresponding password, is equal to

$$\beta = \frac{n}{\lceil \log(n+1) \rceil}, \tag{12}$$

and it does not depend on T.

The possible verification errors are the false rejection of the identical biometric entity and the false acceptance of the different biometric entity. The probabilities of these events, called the false rejection and the false acceptance rates, can be expressed as

$$\begin{aligned} \mathrm{FRR}(\mathbf{w}) &= \sum_{\mathbf{w}' \notin \mathcal{D}^{(T)}(\mathbf{w})} \Omega(\mathbf{w}' \mid \mathbf{w}), \\ \mathrm{FAR}(\mathbf{w}) &= \sum_{\mathbf{w}' \in \mathcal{D}^{(T)}(\mathbf{w})} \mathrm{B}(\mathbf{w}'), \end{aligned} \tag{13}$$

where

$$\begin{aligned} \Omega(\mathbf{w}' \mid \mathbf{w}) &= \prod_{t=1}^{T} \Omega(w'_t \mid w_t), \\ \mathrm{B}(\mathbf{w}') &= \prod_{t=1}^{T} \mathrm{B}(w'_t). \end{aligned} \tag{14}$$

The false rejection event corresponds to the case when the blocks of the input biometric vector are transmitted over a channel in such a way that weights of these blocks are transformed to the weights of the received blocks by a memoryless channel specified by the conditional probabilities $\Omega(0 \mid w), \ldots, \Omega(n \mid w)$. The false acceptance event corresponds to the case when the blocks of the received vector are generated by a Bernoulli source having the probabilities of zeroes and ones equal to 1/2.

The goals of the designer of the system can be different. In particular, the acceptance set $\mathcal{D}^{(T)}(\mathbf{w})$ can be assigned according to the maximum likelihood decision rule. Another assignment is oriented to the minimization of the absolute value of the difference of $\mathrm{FRR}(\mathbf{w}, \mathcal{D}^{(T)}(\mathbf{w}))$ and $\mathrm{FAR}(\mathbf{w}, \mathcal{D}^{(T)}(\mathbf{w}))$. Furthermore, this set can be assigned in such a way that the false rejection/acceptance rate is fixed and the false acceptance/rejection rate is minimized. We will present the assignments of the decision sets that provide us with small decoding error probabilities of both types, which makes efficient solutions to the above problems possible.

Our main claim can be summarized as follows.

Theorem 1. *The decision sets $\mathcal{D}^{(T)}(\mathbf{w})$, $\mathbf{w} \in \{0, \ldots, n\}^T$, can be assigned in such a way that the scheme has the following features:*

(a) *the compression factor β is expressed by (12), and it tends to 0 as an almost linear function of n independently of T, and*

(b) *the false acceptance and the false rejection rates tend to 0 as exponential functions of T in such a way that*

$$\begin{aligned} \mathrm{FRR}(\mathbf{w}) &\le \exp\{-TE_{\mathrm{FRR}}\}, \\ \mathrm{FAR}(\mathbf{w}) &\le \exp\{-TE_{\mathrm{FAR}}\}, \end{aligned} \tag{15}$$

and $E_{\mathrm{FRR}}, E_{\mathrm{FAR}}$ tend to constants depending only on p, as n increases.

The (a) part of the claim directly follows from the description of the scheme. The (b) part of the claim follows from the analysis presented in Section 5. Notice that the fact that the probabilities of error exponentially vanish with T when the expected values of the corresponding random variables differ is a classical result of detection and estimation theory [11]. We will meet the situation of coinciding expected values, and such a behavior is attained due to the difference of the variances of these variables.

Let us first discuss possible approaches to constructing verification schemes for the noiseless case ($p = 0$) when the biometric vectors are mapped to passwords by a deterministic function. In this case, the verifier constructs the password for the vector $\mathbf{b}'$ and makes the acceptance decision if and only if it coincides with the password associated with the claimed user. As a result, the false rejection rate is equal to 0: if $\mathbf{b}' = \mathbf{b}$, then the passwords are identical.

Suppose that the password is defined as a binary vector of length T where the tth bit is the parity of the tth block of the vector $\mathbf{b}$ (the tth bit of the password is equal to 1 if and only if the weight of the vector $\mathbf{b}_t$ is odd), $t = 1, \ldots, T$. Then, the compression factor is equal to $Tn/T = n$ and the false acceptance rate is equal to 2^{-T}, that is, the scheme has a similar features as our scheme. However, to attain a large

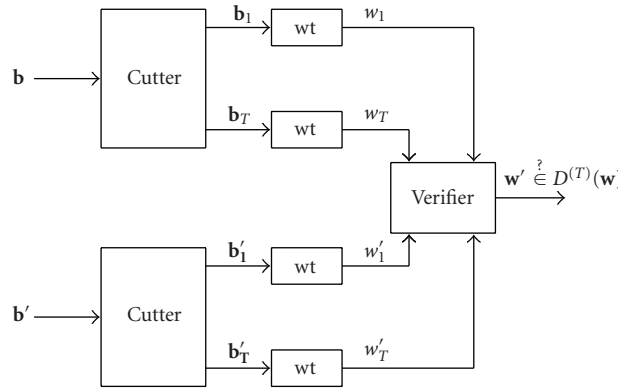

Figure 2: The structure of the verification scheme.

compression factor for $p > 0$, one needs a very large T to obtain low false rejection and false acceptance rates. Another approach to the verification for the noiseless case is based on the specification of the password as a vector consisting of weights of the blocks. Then, the compression factor is equal to β while the false acceptance rate is equal to

$$\prod_{t=1}^{T}\binom{n}{w_t}2^{-n} \le \left(\sqrt{\frac{2}{\pi n}}\right)^{T}. \tag{16}$$

It decreases with T as an exponential function and decreases with n as a polynomial function. We claim that a similar conclusion is also valid for $p \in (0, 1/2)$.

4. Processing the 1-Block Vectors

Suppose that $T = 1$, denote $\mathbf{b} = \mathbf{b}$, $\mathbf{b}' = \mathbf{b}'$, and use the notation (4). We also write $\mathcal{D}(w) = \mathcal{D}^{(1)}(w)$ and represent (11) as

$$\varphi(w, \mathbf{b}') = \begin{cases} \text{Acc}, & \text{if } w' \in \mathcal{D}(w), \\ \text{Rej}, & \text{if } w' \notin \mathcal{D}(w). \end{cases} \tag{17}$$

The maximum likelihood decision rule is implemented by using the acceptance set

$$\mathcal{D}(w) = \{w' \in \{0, \dots, n\} : \ \Omega(w' \mid w) > \mathrm{B}(w')\}. \tag{18}$$

Then, the false rejection and the false acceptance rates are expressed as

$$\begin{aligned} \mathrm{FRR}(w) &= \sum_{w' \notin \{w-\delta_0, \dots, w+\delta_1\}} \Omega(w' \mid w), \\ \mathrm{FAR}(w) &= \sum_{w' \in \{w-\delta_0, \dots, w+\delta_1\}} \mathrm{B}(w'), \end{aligned} \tag{19}$$

where δ_0 and δ_1 are the minimum integers satisfying the inequalities $\Omega(w - \delta_0 \mid w) > \mathrm{B}(w - \delta_0)$ and $\Omega(w + \delta_1 \mid w) > \mathrm{B}(w + \delta_1)$.

To check the (b) claim of the theorem, we use the Gaussian approximations

$$\Omega(w' \mid w) \longrightarrow \widetilde{\Omega}(w' \mid w), \tag{20}$$

$$\mathrm{B}(w') \longrightarrow \mathrm{B}(w'), \tag{21}$$

where

$$\begin{aligned} \widetilde{\Omega}(w' \mid w) &= \mathrm{G}(w'; (n-w)p + w(1-p), np(1-p)), \\ \widetilde{\mathrm{B}}(w') &= \mathrm{G}\left(w'; \frac{n}{2}, \frac{n}{4}\right), \\ \mathrm{G}(z \mid m, \sigma^2) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(z-m)^2}{2\sigma^2}\right\} \end{aligned} \tag{22}$$

stands for the Gaussian probability density function with the mean m and the variance σ^2. The convergence (21) is the standard Gaussian approximation for the binomial distribution. The convergence (20) follows from

$$\begin{aligned} &\binom{n-w}{j} p^j (1-p)^{n-w-j} \\ &\qquad \longrightarrow \mathrm{G}(j; (n-w)p, (n-w)p(1-p)), \\ &\binom{w}{w'-j} p^{w-w'+j} (1-p)^{w'-j} \\ &\qquad \longrightarrow \mathrm{G}(w'-j; wq, wp(1-p)) \end{aligned} \tag{23}$$

for all $j \in \{0, \dots, w'\}$. Furthermore, the replacement of the sum over j at the right-hand side of (8) with the integral over j taken over the interval $(-\infty, +\infty)$ results in (20).

In particular, $\widetilde{\Omega}(n/2)$ and $\widetilde{\mathrm{B}}$ are two Gaussian probability density functions having the same mean $n/2$ and different variances equal to $np(1-p)$ and $n/4$, respectively. The maximum likelihood decoding in this case is equivalent to the selection of one of two hypotheses about the variance of the Gaussian probability distributions having the same mean. It is well known (see, for example [12]) that the

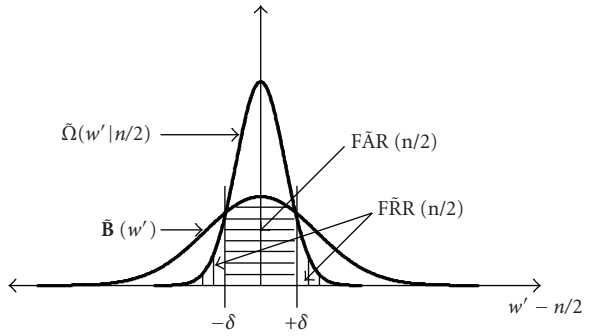

Figure 3: Example of the probability distributions $\widetilde{\Omega}(n/2)$ and $\widetilde{\mathrm{B}}$.

probabilities of the incorrect decisions are determined by the ratio of variances, which is equal to $p(1-p)/(1/4)$ and does not depend on n.

The simplest upper bound for the false acceptance and the false rejection rates can be expressed using the Bhattacharyya distance [13] between the probability density functions $\widetilde{\Omega}(w'|w)$ and $\widetilde{\mathrm{B}}(w')$. Namely, denote

$$\begin{aligned} \mathrm{F\widetilde{R}R}(w) &= \int_{\notin \widetilde{\mathcal{D}}(w)} \widetilde{\Omega}(w' \mid w)\, dw', \\ \mathrm{F\widetilde{A}R}(w) &= \int_{\in \widetilde{\mathcal{D}}(w)} \widetilde{\mathrm{B}}(w')\, dw', \end{aligned} \tag{24}$$

where

$$\widetilde{\mathcal{D}}(w) = \left\{ w' :\ \widetilde{\Omega}(w' \mid w) > B(w') \right\}. \tag{25}$$

Examples of the probability density functions $\widetilde{\Omega}(w' \mid n/2)$ and $\widetilde{\mathrm{B}}(w')$ are given in Figure 3 where we also show the false rejection and the acceptance rates for the maximum likelihood decision rule.

The values of $\mathrm{F\widetilde{R}R}(w), \mathrm{F\widetilde{A}R}(w)$ can be bounded from above as

$$\mathrm{F\widetilde{R}R}(w), \mathrm{F\widetilde{A}R}(w) \le \int_{-\infty}^{+\infty} \sqrt{\widetilde{\Omega}(w' \mid w)\widetilde{\mathrm{B}}(w')}\, dw'. \tag{26}$$

The inequalities (26) follow from the observations

$$\begin{aligned} w' \notin \widetilde{\mathcal{D}}(w) &\Longrightarrow \sqrt{\frac{\widetilde{\mathrm{B}}(w')}{\widetilde{\Omega}(w' \mid w)}} \ge 1, \\ w' \in \widetilde{\mathcal{D}}(w) &\Longrightarrow \sqrt{\frac{\widetilde{\Omega}(w' \mid w)}{\widetilde{\mathrm{B}}(w')}} \ge 1. \end{aligned} \tag{27}$$

The multiplications of the probabilities $\widetilde{\Omega}(w' \mid w)$ and $\widetilde{\mathrm{B}}(w')$ in (24) by the square roots above and extension of the integration over all possible values of w' bring the desired bounds.

The value of the integral at the right-hand side of (26) can be easily computed using the statement below.

Proposition 1. *For all pairs (m_1, σ_1) and (m_2, σ_2) such that $\sigma_1, \sigma_2 > 0$,*

$$\begin{aligned} &\int_{-\infty}^{+\infty} \sqrt{G(z \mid m_1, \sigma_1) G(z \mid m_2, \sigma_2)}\, dz \\ &\quad = \left(\frac{2\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2} \right)^{1/2} \exp\left\{ -\frac{(m_1 - m_2)^2}{2(\sigma_1^2 + \sigma_2^2)} \right\}. \end{aligned} \tag{28}$$

The proof is given in the Appendix.

The use of (28) with $(m_1, \sigma_1) = ((n-w)p + w(1-p), np(1-p))$ and $(m_1, \sigma_1) = (n/2, n/4)$ shows that the worst case corresponds to $w = n/2$ and

$$\mathrm{F\widetilde{R}R}(w), \mathrm{F\widetilde{A}R}(w) \le \delta, \tag{29}$$

where

$$\delta = \left(\frac{\sqrt{p(1-p)}}{p(1-p) + 1/4} \right)^{1/2}. \tag{30}$$

The bounds (29) are very simple, but they can be useless. For example, if $p = 0.05$, then $\delta = 0.856$. If the acceptance set for the vector $\mathbf{w}$ consisting of T blocks is defined as the set of vectors $\mathbf{w}'$ such that $w_t' \in \widetilde{\mathcal{D}}(w_t)$ for at least $T/2$ indices $t \in \{1, \ldots, T\}$ and the estimate of the probability of incorrect decision for each block is greater than $1/2$, then the estimate of probability of incorrect decision for T blocks is close to 1. Nevertheless, if the acceptance set is defined differently, considerations of this section are of interest.

5. Processing the T-Block Vectors

Let us first summarize our verification scheme, which can be also called a basic scheme.

Enrollment. Represent the input vector $\mathbf{b}$ of length Tn as a result of concatenation of T blocks of length n. Compute the weights of the blocks $w_1, \ldots, w_n$ and store them in the database as the vector $\mathbf{w}$.

Verification. Having received a binary vector $\mathbf{b}'$, construct the vector of weights of its blocks and denote this vector by $\mathbf{w}'$. Compute

$$\ln \frac{\Omega(\mathbf{w}' \mid \mathbf{w})}{\mathrm{B}(\mathbf{w}')} = \sum_{t=1}^{T} \ln \frac{\Omega(w_t' \mid w_t)}{\mathrm{B}(w_t')}, \tag{31}$$

and make the acceptance decision if the obtained value is greater than a fixed threshold Λ that has to be chosen in advance depending on the requirements to the false acceptance and the false rejection rates, that is,

$$\mathcal{D}_{\Lambda}^{(T)}(\mathbf{w}) = \left\{ \mathbf{w}' :\ \sum_{t=1}^{T} \ln \frac{\Omega(w_t' \mid w_t)}{\mathrm{B}(w_t')} > T\Lambda \right\}. \tag{32}$$

We write

$$\mathrm{FRR}_{\Lambda}(\mathbf{w}) = \mathrm{FRR}(\mathbf{w}), \qquad \mathrm{FAR}_{\Lambda}(\mathbf{w}) = \mathrm{FAR}(\mathbf{w}), \tag{33}$$

Table 1: Some values of ΔT_n and ΔT.

n	β	$p = 0.01$	$p = 0.05$	$p = 0.10$
32	5.3	5.19	14.91	36.56
64	9.1	4.78	14.51	36.27
128	16.0	4.51	14.23	36.06
256	28.4	4.31	14.06	35.94
512	51.2	4.18	13.96	35.87
1024	93.1	4.10	13.90	35.82
∞	∞	4.01	13.86	35.80

when $\mathrm{FRR}(\mathbf{w}), \mathrm{FAR}(\mathbf{w})$ are defined by (13) with the set $\mathcal{D}_\Lambda^{(T)}(\mathbf{w})$ substituted for the set $\mathcal{D}^{(T)}(\mathbf{w})$. Let us also denote

$$\begin{aligned} \widetilde{\mathrm{FRR}}_\Lambda(\mathbf{w}) &= \int_{\notin \widetilde{\mathcal{D}}^{(T)}(\mathbf{w})} \widetilde{\Omega}(\mathbf{w}' \mid \mathbf{w})\, dw'_1 \dots dw'_T, \\ \widetilde{\mathrm{FAR}}_\Lambda(\mathbf{w}) &= \int_{\in \widetilde{\mathcal{D}}^{(T)}(\mathbf{w})} \widetilde{\mathrm{B}}(\mathbf{w}')\, dw'_1 \dots dw'_T, \end{aligned} \tag{34}$$

where

$$\begin{aligned} \widetilde{\Omega}(\mathbf{w}' \mid \mathbf{w}) &= \prod_{t=1}^{T} \widetilde{\Omega}(w'_t \mid w_t), \\ \widetilde{\mathrm{B}}(\mathbf{w}') &= \prod_{t=1}^{T} \widetilde{\mathrm{B}}(w'_t). \end{aligned} \tag{35}$$

The probabilities introduced above can be easily estimated for $\Lambda = 0$, which corresponds to the maximum likelihood decision rule. Namely,

$$\mathrm{FRR}_0(\mathbf{w}), \mathrm{FAR}_0(\mathbf{w}) \le \delta_n^T, \tag{36}$$

where

$$\delta_n = \sum_{w'} \sqrt{\Omega\left(w' \mid \frac{n}{2}\right) \mathrm{B}(w')}, \tag{37}$$

$$\widetilde{\mathrm{FRR}}_0(\mathbf{w}), \widetilde{\mathrm{FAR}}_0(\mathbf{w}) \le \delta^T, \tag{38}$$

where δ is defined in (30). Hence, $-\ln \delta_n$ *is a lower bound on the exponents* $E_{\mathrm{FRR}}, E_{\mathrm{FAR}}$ *in* (15).

Let us denote

$$\Delta T_n = \frac{1}{-\lg \delta_n}, \qquad \Delta T = \frac{1}{-\lg \delta}. \tag{39}$$

Then, the inequalities (36) can be represented as the following statement: if $T = k\Delta T_n$, then

$$\mathrm{FRR}_0(\mathbf{w}), \mathrm{FAR}_0(\mathbf{w}) \le 10^{-k}. \tag{40}$$

Similarly, the inequalities (38) can be represented as the following statement: if $T = k\Delta T$, then

$$\widetilde{\mathrm{FRR}}_0(\mathbf{w}), \widetilde{\mathrm{FAR}}_0(\mathbf{w}) \le 10^{-k}. \tag{41}$$

Some values of ΔT_n and ΔT are given in Table 1.

Suppose that the biometric vectors have length $N = 4$ Kbytes $= 32568$ bits. Let us partition this length in $T = 128$ blocks of length $n = 256$ bits (we will refer to the corresponding line in Table 1). In our scheme, each block is mapped to a binary vector of length $\lceil \log 257 \rceil = 9$ bits, and the length of the password is equal to $9T = 1152$ bits $= 144$ bytes. The compression factor is equal to $\beta = 256/9 = 28.4$. Suppose that $p = 0.05$. Then, the expected number of errors when the biometric vector is corrupted is equal to $32568 \cdot 0.05 = 6514$, which is 5.6 times greater than the length of the password. Nevertheless, we attain the false rejection and the false acceptance rates not greater than $10^{-128/14.06} < 10^{-9}$. Furthermore, if T is increased twice and becomes equal to 256 (the length of the vectors is equal to 8 Kbytes), then the false rejection and the false acceptance rates are not greater than $10^{-256/14.06} < (10^{-9})^2 = 10^{-18}$. Similar conclusions can be drawn for any length in a way that the increase of the length by 14 blocks reduces the false rejection and the false acceptance rates 10 times. If $p = 0.01$ or $p = 0.1$, then we have to substitute 4.31 or 35.94 for 14.06 in these considerations. Notice also that these numbers are very close to the numbers that are asymptotically attained and have a simple formal expression.

6. A Variant of the Verification Scheme Based on Balancing

For all $i \in \{0, \dots, n\}$, let $1^i 0^{n-i}$ denote the vector constructed by the concatenation of i ones and $n - i$ zeroes. For example, if $n = 4$, then

$$\begin{bmatrix} 1^0 0^4 \\ 1^1 0^3 \\ 1^2 0^2 \\ 1^3 0^1 \\ 1^4 0^0 \end{bmatrix} = \begin{bmatrix} 0000 \\ 1000 \\ 1100 \\ 1110 \\ 1111 \end{bmatrix}. \tag{42}$$

The vector $\mathbf{c}$ is called *a balanced vector* if it contains equal number of zeroes and ones. Thus, the weight of a balanced vector is equal to $n/2$.

Given a vector $\mathbf{b}$, let

$$\mathcal{I}(\mathbf{b}) = \left\{ i \in \{0, \dots, n\} : \mathrm{wt}\left(\mathbf{b} \oplus 1^i 0^{n-i}\right) = \frac{n}{2} \right\} \tag{43}$$

denote the set of indices i such that the transformation

$$\mathbf{b} \longrightarrow \mathbf{b} \oplus 1^i 0^{n-i}, \tag{44}$$

which inverts the first i components of the vector $\mathbf{b}$, brings a balanced vector. For example,

$$\begin{aligned} \mathcal{I}(0000) &= \{2\}, \\ \mathcal{I}(0101) &= \{0, 2, 4\}, \\ \mathcal{I}(0100) &= \{1, 3\}. \end{aligned} \tag{45}$$

The transformation (44) is illustrated in Table 2.

Table 2: The structure of the vector $\mathbf{c} = \mathbf{b} \oplus 1^i 0^{n-i}$, where $i \in \mathcal{I}(b)$.

$\mathrm{wt}(b_1, \ldots, b_i) = j$	$\mathrm{wt}(b_{i+1}, \ldots, b_n) = w - j$
$c_1 = b_1 \oplus 1, \ldots, c_i = b_i \oplus 1$	$c_{i+1} = b_{i+1}, \ldots, c_n = b_n$
$\mathrm{wt}(c_1, \ldots, c_i) = i - j$	$\mathrm{wt}(c_{i+1}, \ldots, c_n) = w - j$
$(i - j) + (w - j) = n/2$	

It is well known [14] that

$$1 \le |\mathcal{I}(\mathbf{b})| \le n/2 + 1. \tag{46}$$

Introduce the following algorithm.

Enrollment. Represent the input vector $\mathbf{b}$ of length Tn as a result of concatenation of T blocks of length n. For each block $\mathbf{b}_t$, construct the set $\mathcal{I}(\mathbf{b})$ and choose an integer $i(\mathbf{b}_t) \in \{0, \ldots, n\}$ according to a uniform probability distribution over the set $\mathcal{I}(\mathbf{b}_t)$. Set

$$\mathrm{pw}(\mathbf{b}) = (i(\mathbf{b}_1), \ldots, i(\mathbf{b}_n)) \tag{47}$$

and store the vector $\mathrm{pw}(b)$ in the database.

Verification. Represent the input vector $\mathbf{b}'$ of length Tn as a result of concatenation of T blocks of length n. For each block $\mathbf{b}'_t$, compute

$$w'_t = \mathrm{wt}\left(\mathbf{b}'_t \oplus 1^{i(\mathbf{b}_t)} 0^{n - i(\mathbf{b}_t)}\right). \tag{48}$$

Make the acceptance decision if and only if $\mathbf{w}' \in \mathcal{D}_\Lambda^{(T)}(\mathbf{w}^*)$, where $\mathbf{w}^*$ is the vector whose components are equal to $n/2$ and the acceptance set $\mathcal{D}_\Lambda^{(T)}(\mathbf{w}^*)$ is defined in (32).

For example, if $n = 4$, then the vector 0000 is mapped to the password "2", the vector 0101 is mapped to the passwords "0", "2", "4" with the probabilities 1/3, and the vector 0100 is mapped to the passwords "1", "3" with probability 1/2.

Proposition 2. *Let a given vector* $\mathbf{b}$ *be transmitted over a binary symmetric channel having the crossover probability* p*, that is, the conditional probability of receiving the vector* $\mathbf{b}'$ *at the output of the channel is expressed as*

$$V(\mathbf{b}' \mid \mathbf{b}) = (1 - p)^{n - wt(\mathbf{b} \oplus \mathbf{b}')} p^{wt(\mathbf{b} \oplus \mathbf{b}')}. \tag{49}$$

If $i \in \{0, \ldots, n\}$ *is assigned in such a way that* $\mathbf{b} \oplus 1^i 0^{n-i}$ *is the balanced vector and*

$$V_i(w' \mid \mathbf{b}) = \sum_{\mathbf{b}'} V(\mathbf{b}' \mid \mathbf{b}) \chi\left\{wt\left(\mathbf{b}' \oplus 1^i 0^{n-i}\right) = w'\right\} \tag{50}$$

denote the probability of receiving a vector $\mathbf{b}'$ *with*

$$wt\left(\mathbf{b}' \oplus 1^i 0^{n-i}\right) = w', \tag{51}$$

then

$$V_i(w' \mid \mathbf{b}) = \Omega\left(w' \mid \frac{n}{2}\right). \tag{52}$$

The proof is given in the Appendix.

An idea of the introduction of the balanced scheme is to reduce the performance of the verifier to the worst case performance for the basic scheme when all components of the vector $\mathbf{w}$ are equal to $n/2$. Another disadvantage of the scheme is the point that an attacker passes through the verification stage with the acceptance decision by presenting an alternating vector $0101\ldots01$. On the other hand, the balancing scheme allows us to hide any biometric vector of the user in his password, contrary to the basic scheme where the password consisting of all zeroes discovers the original vector. Furthermore, in most of the cases the same biometric vector can be mapped to many different passwords, since the mapping is stochastic when the cardinality of at least one of the sets $\mathcal{I}(\mathbf{b}_1), \ldots, \mathcal{I}(\mathbf{b}_T)$ is greater than 1.

The conclusion about the secrecy of the balanced scheme, meaning the possibility of the discovery of the block given its password, is based on the considerations below. Given an $i \in \{0, \ldots, n\}$, let

$$M_i = |\{\mathbf{b} : \ i \in \mathcal{I}(\mathbf{b})\}. \tag{53}$$

Then (see Table 2),

$$\begin{aligned} M_i &= \sum_w \binom{i}{\left(w - \frac{n}{2} + i\right)/2} \binom{n-i}{\left(w + \frac{n}{2} - i\right)/2} \\ &\ge \binom{i}{\frac{i}{2}} \binom{n-i}{\frac{n-i}{2}} \\ &\ge \min_{i \in \{0, \ldots, n\}} \left[\binom{i}{i/2} \binom{n-i}{(n-i)/2} \right] \\ &= \binom{\frac{n}{2}}{\frac{n}{4}}^2 \\ &\ge \left[\frac{1}{\sqrt{2\pi(n/2)(1/4)}} 2^{n/2 - 2/(12n/4)} \right]^2 \\ &= \frac{4}{\pi n} 2^{n - 4/(3n)}, \end{aligned} \tag{54}$$

where the first inequality follows from the observation that $w = n/2$ specifies one of terms of the sum for any i. Hence, the total number of biometric vectors that are mapped to the same password is bounded from below as

$$\left(\frac{4}{\pi n}\right)^T 2^{T(n - 4/(3n))} \tag{55}$$

and the exponent asymptotically coincides with Tn.

7. Example of Using the Verification Scheme for the DNA Data

There are data received on the basis of the DNA measurements [15]. We previously used them to illustrate coding schemes in [16, 17].

The example, described in this section, is mainly introduced for the illustration, since the performance of the

verifier probably does not allow one to recommend it for practical use. Nevertheless, transformations of the outcomes of the measurements seem to be typical. Notice also that the DNA data are universal in a sense that there are 24–28 deciphered alleles where the corresponding probability distributions of the outcomes of the measurements are recognized as stable distributions, while processing fingerprints, iris, and so forth requires the description of a number of technical details.

7.1. Structure of the DNA Data and the Mathematical Model. The most common DNA variations are Short Tandem Repeats (STR), arrays of 5 to 50 copies (repeats) of the same pattern (the motif) of 2 to 6 pairs. As the number of repeats of the motif highly varies among individuals, it can be effectively used for identification of individuals. The human genome contains several 100,000 STR loci, that is, physical positions in the DNA sequence where an STR is present. An individual variant of an STR is called allele. Alleles are denoted by the number of repeats of the motif. The genotype of a locus comprises both the maternal and the paternal allele. However, without additional information, one cannot determine which allele resides on the paternal or the maternal chromosome. If the measured numbers are equal to each other, then the genotype is called homozygous. Otherwise, it is called heterozygous. The STR measurement errors are usually classified into three groups: (1) *allelic drop-in*, when in a homozygous genotype, an additional allele is erroneously included, for example, genotype (10,10) is measured as (10,12); (2) *allelic drop–out*, when an allele of a heterozygous genotype is missing, for example, genotype (7,9) is measured as (7,7); (3) *allelic shift*, when an allele is measured with a wrong repeat number, for example, genotype (10,12) is measured as (10,13).

The points above can be formalized as follows [16]. Suppose that there are N^* sources. Let the tth source generate a pair of integers according to the probability distribution

$$\Pr_{\mathrm{DNA}}\{(A_{t,1},A_{t,2})=(a_{t,1},a_{t,2})\}=\pi_t(a_{t,1})\pi_t(a_{t,2}), \tag{56}$$

where $a_{t,1},a_{t,2}\in\{c_t,\dots,c_t+k_t-1\}$ and c_t,k_t are given positive integers. Thus, *we assume that $A_{t,1}$ and $A_{t,2}$ are independent random variables* that contain information about the number of repeats of the tth motif in the maternal and the paternal allele. *We also assume that* $(A_{t,1},A_{t,2}),t=1,\dots,N^*$, *are mutually independent pairs of random variables*, that is,

$$\begin{aligned}&\Pr_{\mathrm{DNA}}\{(A_1,A_2)=(\mathbf{a}_1,\mathbf{a}_2)\}\\&\quad=\prod_{t=1}^{N^*}\Pr_{\mathrm{DNA}}\{(A_{t,1},A_{t,2})=(a_{t,1},a_{t,2})\},\end{aligned} \tag{57}$$

where $A_\ell=(A_{1,\ell},\dots,A_{n,\ell})$ and $\mathbf{a}_\ell=(a_{1,\ell},\dots,a_{n,\ell}),\ell=1,2$.

Let us fix a $t\in\{1,\dots,N^*\}$ and denote

$$\mathcal{P}_t\triangleq\{s=(i,j):\ i,j\in\{c_t,\dots,c_t+k_t-1\},\ j\geq i\}. \tag{58}$$

Then, the probability distribution of a pair of random variables

$$S_t\triangleq(\min\{A_{t,1},A_{t,2}\},\max\{A_{t,1},A_{t,2}\}), \tag{59}$$

which represents the outcome of the tth measurement, can be expressed as

$$\Pr_{\mathrm{DNA}}\{S_t=(i,j)\}=\gamma_t(i,j), \tag{60}$$

where $\gamma_t(i,j)\triangleq\pi_t^2(i)$, if $j=i$, and $\gamma_t(i,j)\triangleq 2\pi_t(i)\pi_t(j)$, if $j\neq i$. Thus, the total number of outcomes having positive probability is equal to

$$K_t=\frac{k_t(k_t+1)}{2}. \tag{61}$$

7.2. Mapping of the DNA Data to Binary Vectors and Introducing the Passwords. The outcomes of the DNA measurements bring the following results [16]: the total number of alleles is 28, one can extract 128 bits from the measurements of a person, the entropy of the probability distribution over the outcomes is equal to 109, and the maximum probability of a vector consisting of 28 outcomes is equal to 2^{-76}. In the following discussion, we will assume that $N^*=27$ (the **DYS391** allele is excluded).

Let us fix $t\in\{1,\dots,27\}$ and let $\mathcal{S}_t$ denote the set of cardinality $|\mathcal{S}_t|=K_t$ consisting of the outcomes that can be received from the t-th allele with positive probability. Associate the outcomes with the integers $1,\dots,K_t$ and let $\gamma_t^{(i)}$ denote the probability of the outcome, which is mapped to the integer i. Let us run the procedure that maps $i\in\{1,\dots,K_t\}$ to the integer $u\in\{0,\dots,7\}$: partition the set $\mathcal{S}_t$ in 8 subsets $\mathcal{S}_{t0},\dots,\mathcal{S}_{t7}$ in such a way that

$$\sum_{i\in\mathcal{S}_{tu}}\gamma_t^{(i)}\approx 2^{-3} \tag{62}$$

and set

$$i\longrightarrow u\Longleftrightarrow i\in\mathcal{S}_{tu}. \tag{63}$$

The use of this procedure for $t=1,\dots,N^*$ maps 27 outcomes to a vector $(u_1,\dots,u_{27})\in\{0,\dots,7\}^{27}$, which can be expressed by a binary vector $\mathbf{b}=(b_1,\dots,b_{81})$.

Let us apply the verification scheme described in Section 3 for $T=3$ and $n=27$. Thus, the vector $\mathbf{b}$ is mapped to the password (w_1,w_2,w_3), where $w_1,w_2,w_3\in\{0,\dots,27\}$, and we need 15 bits to express a password in binary format. Furthermore, let us postulate the following model for the noise when the DNA data of the same user are measured for the second time: with probability $1-\varepsilon'$, the outcome of the measurement at the tth allele is the same as before; with probability ε', it is equal to the integer i chosen from the set $\{1,\dots,K_t\}$ according to a uniform probability distribution. In the following formal considerations, we assume a simplified model where *the approximate equality* (62) *is replaced with the equality for all* $u\in\{0,\dots,7\}$ *and* $t\in\{1,\dots,27\}$. *One also assumes that the outcome of the measurement of the same user copies the previous value of* u

with probability $1-\varepsilon$ *and that it takes an arbitrary value belonging to the set* $\{0,\dots,7\}$ *with probability* ε, *where* ε *is less than* ε'. In a practical system, $\varepsilon' = 0.05$ [15], we set $\varepsilon = 0.02$. Notice that our assumptions do not seem to be critical: after these assumptions are relaxed, the formal analysis below has to be updated with the correction factors without essential change of the conclusions.

For $v = 0,\dots,3$, set

$$q_{v,v} = \binom{3}{v}2^{-3}\left[1-\varepsilon+\varepsilon\binom{3}{v}2^{-3}\right] \tag{64}$$

and, for $v, v' = 0,\dots,3$ and $v' \neq v$, set

$$q_{v,v'} = \binom{3}{v}2^{-3}\varepsilon\binom{3}{v'}2^{-3}. \tag{65}$$

Then, $q_{v,v'}$ is equal to the probability of the event that "the weights of the tth DNA measurements" of a randomly chosen person are equal to v and v' at the enrollment and the verification stages, respectively, $v, v' = 0,\dots,3$.

To express the conditional probabilities $\Omega(w' \mid w)$, w, $w' = 0,\dots,27$, run the following procedure.

(1) For $v, v' = 0,\dots,3$, set

$$Q^{(1)}_{v,v'} = q_{v,v'}. \tag{66}$$

(2) For $k = 2,\dots,9$,

(a) for $w, w' = 0,\dots,3k$, set

$$Q^{(k)}_{w,w'} = 0; \tag{67}$$

(b) for $w, w' = 0,\dots,3(k-1)$ and $v, v' = 0,\dots,3$, increase $Q^{(k)}_{w+v,w'+v'}$ by the product $Q^{(k-1)}_{w,w'}q_{v,v'}$, that is, set

$$Q^{(k)}_{w+v,w'+v'} := Q^{(k)}_{w+v,w'+v'} + Q^{(k-1)}_{w,w'}q_{v,v'}. \tag{68}$$

(3) For $w, w' = 0,\dots,27$, set

$$\Omega(w' \mid w) = \frac{Q^{(9)}_{w,w'}}{P_w}, \tag{69}$$

where

$$P_w = \sum_{w'=0}^{27} Q^{(9)}_{w,w'}. \tag{70}$$

One can see that the same procedure, being used with $\varepsilon = 1$, gives the entries of the probabilities $\mathrm{B}(w')$, $w' = 0,\dots,27$, that describe the output probability distribution for the attacker (the value of parameter $w \in \{0,\dots,27\}$ is arbitrary in this case). The obtained probability distributions bring all necessary data for the verification algorithm of the previous section when $T = 3$ and

$$\begin{aligned}\Omega(\mathbf{w}' \mid \mathbf{w}) &= \prod_{t=1}^{3}\Omega(w'_t \mid w_t),\\ \mathrm{B}(\mathbf{w}') &= \prod_{t=1}^{3}\mathrm{B}(w'_t).\end{aligned} \tag{71}$$

Some data are presented in Table 3 where we show only the entries of the probability distributions that are greater than 0.01.

The data processing above illustrates several points that can be important for the practical implementation of the verification algorithm. In particular, notice that the conditional probability distributions $\Omega(w' \mid w)$,$w' = 0,\dots,27$, were introduced using the input probability distributions, but they are almost independent on w and their approximation, $\widehat{\Omega}(w' \mid w)$,$w' = 0,\dots,27$, can be assigned only as the function of ε,

$$\begin{aligned}&\big(\widehat{\Omega}(w-2 \mid w), \widehat{\Omega}(w-1 \mid w), \widehat{\Omega}(w \mid w),\\ &\quad \widehat{\Omega}(w+1 \mid w), \widehat{\Omega}(w+2 \mid w)\big)\\ &\qquad = (0.02, 0.04, 0.89, 0.04, 0.01),\\ &\widehat{\Omega}(w' \mid w) = 0\end{aligned} \tag{72}$$

for $w' \notin \{w-2,\dots,w+2\}$. The verification algorithm can be simplified in such a way that the acceptance decision is made if and only if $w'_t \in \{w_t - 1, w_t, w_t + 1\}$ for $t = 1,2,3$. Then, the false rejection rate is approximated as

$$1 - (0.04 + 0.89 + 0.04)^3 = 0.11 \tag{73}$$

and the false acceptance rate is approximated as

$$(0.15 + 0.15 + 0.13)^3 = 0.08. \tag{74}$$

This value has to be multiplied by a factor having the order of magnitude of $(0.15)^3 = 0.003$ if one is interested in the average false acceptance rate. Notice also that the mapping (63) gives an additional resource that decreases the false acceptance rate: if we randomize over the mapping for $t = 1,2,3$, then the same factor of the false acceptance rate is obtained for a fixed input vector consisting of pairs of outcomes of the DNA measurements.

Our example also indicates the point that the mapping of the available data to a binary string with the further computation of the weight of the vector looks as an artificial transformation, and "a more natural password" would be specified as the arithmetic average of 9 integers that form the block. However, the arithmetic average is a float, and we also meet a problem of the specification of the length of a binary string needed for its representation (it also determines the length of the password in bits). We plan to discuss this point in a future correspondence.

8. Conclusion

We presented some variants of the verification schemes oriented to practical applications where the original biometric vectors are split into blocks and converted to short strings using block-by-block transformations. The key idea is the translation of the statistical dependence between the vectors of the same user into the statistical dependence between passwords assigned to the corresponding blocks.

Table 3: Some values of the marginal and the conditional probablity distributions over the weights for the legitimate user when $\varepsilon = 0.02$ and for the attacker ($\varepsilon = 1$).

	w	P_w	$\Omega(w' \mid w)$, $w' = 8, \ldots, 19$											
$\varepsilon = 0.02$	8	0.02	0.73	0.11	0.07	0.02								
	9	0.03	0.02	0.88	0.05	0.03								
	10	0.06		0.03	0.88	0.05	0.02							
	11	0.10			0.03	0.88	0.05	0.02						
	12	0.13			0.01	0.03	0.89	0.04	0.02					
	13	0.15				0.01	0.04	0.89	0.04	0.02				
	14	0.15					0.02	0.04	0.89	0.04	0.01			
	15	0.13						0.02	0.04	0.89	0.03	0.01		
	16	0.10							0.02	0.05	0.88	0.03		
	17	0.06								0.02	0.05	0.88	0.03	
	18	0.03									0.03	0.05	0.88	0.02
	19	0.02										0.03	0.05	0.88
$\varepsilon = 1$	any		0.02	0.03	0.06	0.10	0.13	0.15	0.15	0.13	0.10	0.06	0.03	0.02

The scheme can be introduced without assumptions about a coordinate—wise dependence between the biometric vectors, which is important for many practical applications, like processing of the iris or fingerprints. In general case, "the weight of the block" is the function of the total amount of information extracted from a fixed number of outcomes of the measurements. In particular, it can be understood as the number of minutiae points belonging to a certain area while measuring the fingerprint. Different types of the observation errors, and like missing of some data, registration errors, synchronization errors, are also accumulated. To implement the verification algorithm, one is supposed to find a proper description of the conditional probability distribution Ω without specification of the errors that cause the corresponding transitions. This problem is oriented to a particular application, since we do not think that there exists a universal procedure for any biometric observations. The analysis presented in our correspondence can serve as a basis for the analysis of the verification performance depending on this probability distribution.

Notice that the verification scheme can be also effectively used when the name of a person, which is used as a pointer to a particular password stored in the database, is not given. In this case, our approach serves as a filter to make a preselection of passwords of the users whose biometric vectors can be close to the presented biometric vector. As a result, we get a typical application of hashing when the rejection decision are made with the data that are stored in a random access memory.

Notice also that there are different variants of the basic procedure. One of them, called the balancing verification scheme, was described. Another variant appears with non-uniform partitioning of the biometric vectors in blocks. In this case, the blocks of lengths $n_1, \ldots, n_T$ are created in such a way that their weights are shifted from $n_1/2, \ldots, n_T/2$ "as much as possible" to improve the performance. However, the positions of the boundaries of the blocks have to be stored, and one has to investigate the tradeoff between the performance and the required size of the memory. We did not consider this problem in the present correspondence assuming that the length of the original biometric vector and the length of the password are fixed. In this case, for the basic scheme, the values of Tn and $T\log(n+1)$ are fixed, and the values of the parameters T and n are determined.

Appendices

A. Proof of Proposition 1

We write

$$\int_{-\infty}^{+\infty} \sqrt{\mathrm{G}(z \mid m_1, \sigma_1)\mathrm{G}(z \mid m_2, \sigma_2)}\, dz = \frac{1}{\sqrt{2\pi\sigma_1\sigma_2}} \times \int_{-\infty}^{+\infty} \exp\left\{-\frac{1}{2}\left[\frac{(z-m_1)^2}{2\sigma_1^2} + \frac{(z-m_2)^2}{2\sigma_2^2}\right]\right\} dz, \tag{A.1}$$

and use the equalities

$$\begin{aligned}
&\frac{(z-m_1)^2}{2\sigma_1^2} + \frac{(z-m_2)^2}{2\sigma_2^2} \\
&= z^2\left(\frac{1}{2\sigma_1^2} + \frac{1}{2\sigma_2^2}\right) - 2z\left(\frac{m_1}{2\sigma_1^2} + \frac{m_2}{2\sigma_2^2}\right) + \left(\frac{m_1^2}{2\sigma_1^2} + \frac{m_2^2}{2\sigma_2^2}\right) \\
&= \frac{\sigma_1^2+\sigma_2^2}{2\sigma_1^2\sigma_2^2}\left[z^2 - 2z\frac{m_1\sigma_2^2 + m_2\sigma_1^2}{\sigma_1^2+\sigma_2^2} + \frac{m_1^2\sigma_2^2 + m_2^2\sigma_1^2}{\sigma_1^2+\sigma_2^2}\right] \\
&= \frac{\sigma_1^2+\sigma_2^2}{2\sigma_1^2\sigma_2^2}\left[\left(z - \frac{m_1\sigma_2^2 + m_2\sigma_1^2}{\sigma_1^2+\sigma_2^2}\right)^2 + \frac{m_1^2\sigma_2^2 + m_2^2\sigma_1^2}{\sigma_1^2+\sigma_2^2} - \frac{\left(m_1\sigma_2^2 + m_2\sigma_1^2\right)^2}{\left(\sigma_1^2+\sigma_2^2\right)^2}\right]
\end{aligned}$$

$$= \frac{\sigma_1^2+\sigma_2^2}{2\sigma_1^2\sigma_2^2}\left(z-\frac{m_1\sigma_2^2+m_2\sigma_1^2}{\sigma_1^2+\sigma_2^2}\right)^2 + \frac{\left(m_1^2\sigma_2^2+m_2^2\sigma_1^2\right)\left(\sigma_1^2+\sigma_2^2\right)^2-\left(m_1\sigma_2^2+m_2\sigma_1^2\right)^2}{2\sigma_1^2\sigma_2^2\left(\sigma_1^2+\sigma_2^2\right)}$$

$$= \frac{\sigma_1^2+\sigma_2^2}{2\sigma_1^2\sigma_2^2}\left(z-\frac{m_1\sigma_2^2+m_2\sigma_1^2}{\sigma_1^2+\sigma_2^2}\right)^2 + \frac{m_1^2-2m_1m_2+m_2^2}{2\left(\sigma_1^2+\sigma_2^2\right)}$$

$$= \frac{\sigma_1^2+\sigma_2^2}{2\sigma_1^2\sigma_2^2}\left(z-\frac{m_1\sigma_2^2+m_2\sigma_1^2}{\sigma_1^2+\sigma_2^2}\right)^2 + \frac{(m_1-m_2)^2}{2\left(\sigma_1^2+\sigma_2^2\right)}. \tag{A.2}$$

Therefore,

$$\int_{-\infty}^{+\infty}\sqrt{\mathrm{G}(z\mid m_1,\sigma_1)\mathrm{G}(z\mid m_2,\sigma_2)}\,dz$$

$$= \frac{1}{\sqrt{2\pi\sigma_1\sigma_2}}\exp\left\{-\frac{(m_1-m_2)^2}{2\left(\sigma_1^2+\sigma_2^2\right)}\right\} \cdot \int_{-\infty}^{+\infty}\exp\left\{-\frac{1}{2}\cdot\frac{\sigma_1^2+\sigma_2^2}{2\sigma_1^2\sigma_2^2}\left(z-\frac{m_1\sigma_2^2+m_2\sigma_1^2}{\sigma_1^2+\sigma_2^2}\right)^2\right\}dz$$

$$= \frac{\sqrt{2\sigma_1^2\sigma_2^2}}{\sqrt{\sigma_1\sigma_2}\sqrt{\sigma_1^2+\sigma_2^2}}\exp\left\{-\frac{(m_1-m_2)^2}{2\left(\sigma_1^2+\sigma_2^2\right)}\right\}. \tag{A.3}$$

B. Proof of Proposition 2

We write

$$\begin{aligned} V_i(w'\mid \mathbf{b}) &= \sum_{\mathbf{b}'} V(\mathbf{b}'\mid \mathbf{b})\chi\left\{\mathrm{wt}\left(\mathbf{b}'\oplus 1^i0^{n-i}\right)=w'\right\} \\ &= \sum_{\mathbf{b}''} V\left(\mathbf{b}''\oplus 1^i0^{n-i}\mid \mathbf{b}\right)\chi\left\{\mathrm{wt}(\mathbf{b}'')=w'\right\} \\ &= \sum_{\mathbf{b}''} V\left(\mathbf{b}''\mid \mathbf{b}\oplus 1^i0^{n-i}\right)\chi\left\{\mathrm{wt}(\mathbf{b}'')=w'\right\} \\ &= \sum_{\mathbf{b}''} V\left(\mathbf{b}''\mid \widehat{\mathbf{b}}\right)\chi\left\{\mathrm{wt}(\mathbf{b}'')=w'\right\} \\ &= \Omega\left(w'\mid \frac{n}{2}\right), \end{aligned} \tag{B.1}$$

where $\mathbf{b}''=\mathbf{b}'\oplus 1^i0^{n-i}$, $\widehat{\mathbf{b}}=\mathbf{b}\oplus 1^i0^{n-i}$, and (52) follows.

Acknowledgment

This work was partially supported by the DFG.

References

[1] R. M. Bolle, J. H. Connell, S. Pankanti, N. K. Ratha, and A. W. Senior, *Guide to Biometrics*, Springer, New York, NY, USA, 2004.

[2] V. B. Balakirsky, "Hashing of databases with the use of metric properties of the hamming space," *Computer Journal*, vol. 48, no. 1, pp. 4–16, 2005.

[3] V. B. Balakirsky, A. R. Ghazaryan, and A. J. Han Vinck, "Estimating the Hamming distance between binary vectors via rate distortion source coding," in *Proceedings of the 29th Symposium on Information Theory in the Benelux*, pp. 3–10, Leuven, Belgium, 2008.

[4] V. B. Balakirsky, A. R. Ghazaryan, and A. J. Han Vinck, "Combinatorial data reduction algorithm and its applications to biometric verification," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT '09)*, pp. 2246–2251, Seoul, Korea, 2009.

[5] U. Uludag, S. Pankanti, S. Prabhakar, and A. K. Jain, "Biometric cryptosystems: Issues and challenges," *Proceedings of the IEEE*, vol. 92, no. 6, pp. 948–60, 2004.

[6] N. Ratha, S. Chikkerur, J. Connell, and R. Bolle, *Security with Noisy Data*, Springer, New York, NY, USA, 2007.

[7] A. Juels and M. Wattenberg, "Fuzzy commitment scheme," in *Proceedings of the 6th ACM Conference on Computer and Communications Securit*, pp. 28–36, November 1999.

[8] Y. Dodis, L. Reyzin, and A. Smith, "Fuzzy extractors: how to generate strong keys from biometrics and other noisy data," *Lecture Notes in Computer Science*, vol. 3027, pp. 523–540, 2004.

[9] N. Frykholm and A. Juels, "Error-tolerant password recovery," in *Proceedings of the 8th ACM Conference on Computer and Communications Security*, pp. 1–9, Philadelphia, Pa, USA, 2001.

[10] D. R. Stinson, "Universal hashing and authentication codes," *Designs, Codes and Cryptography*, vol. 4, no. 3, pp. 369–380, 1994.

[11] H. L. Van Trees, *Detection, Estimation and Modulation Theory*, John Wiley & Sons, New York, NY, USA, 2002.

[12] A. Papoulis, *Papoulis, Probability, Random Variables and Stochastic Processes*, McGraw-Hill, New York, NY, USA, 1984.

[13] R. Gallager, *Information Theory and Reliable Communication*, John Wiley & Sons, New York, NY, USA, 1986.

[14] D. E. Knuth, "Efficient balanced codes," *IEEE Transactions on Information Theory*, vol. 32, no. 1, pp. 51–53, 1986.

[15] U. Korte, M. Krawczak, J. Merkle et al., "A cryptographic biometric authentication system based on genetic fingerprints," in *Proceedings of the Sicherheit*, pp. 263–276, Saarbrucken, Germany, 2008.

[16] V. B. Balakirsky, A. R. Ghazaryan, and A. J. Han Vinck, "Additive block coding schemes for biometric authentication with the DNA data," in *Proceedings of the 1st European Workshop on Biometrics and Identity Management*, B. Schouten et al., Ed., vol. 5372 of *Lecture Notes in Computer Science*, pp. 160–169, 2008.

[17] V. B. Balakirsky and A. J. Han Vinck, "Mathematical model for constructing passwords from biometrical data," *Security and Communication Networks*, vol. 2, no. 1, pp. 1–9, 2009.