SECURITY AND COMMUNICATION NETWORKS
Security Comm. Networks. 2009; **2**:1–9
Published online 14 January 2009 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sec.96

Mathematical model for constructing passwords from biometrical data

Vladimir B. Balakirsky*,†, Anahit R. Ghazaryan and A. J. Han Vinck
Institute for Experimental Mathematics, 45326 Essen, Germany

Summary

We propose a probabilistic model for constructing passwords on the basis of outcomes of biometrical measurements. An algorithm for the transformation of biometrical data to passwords is given. Performance of the authentication scheme is evaluated by the compression factor, the false acceptance/rejection rates, the probability distribution over the set of passwords, and the probability of a correct guess of the input biometrical data mapped to the known password. An application of the results to the DNA measurements is presented. Copyright © 2009 John Wiley & Sons, Ltd.

KEY WORDS: authentication; verification; hashing; decoding

1. Introduction

We consider the following authentication problem. Suppose that names and passwords of a certain group of people, called the users, are stored in the database (DB). Having received a pair (name, password), an authentication scheme checks whether the pair is valid or not. The passwords usually have length about 8 bytes, and they are stored in a protected part of the DB. The passwords "have to look as completely random sequences" to create difficulties to an attacker, who wants to enter the system. Notice that this requirement to an algorithm of assigning passwords is not easy to formalize, unless the passwords are formed by a pseudo-random number generator. Another desirable property is the possibility of correcting errors in the password presented by a legal user. Formalization of this requirement is also difficult. Moreover, the two requirements above contradict to each other.

In biometrical authentication systems [1], the password is formed as a function of the outcomes of biometrical measurements. However, the measurements can be hardly exactly repeated, and we think that finding a probabilistic description of the noise of observations is not possible. For example, in the case when locations of minutiae points of the fingerprint are measured, the errors are caused by shifts and rotations of the finger, the light, the pressure, etc. In practical biometrical systems, designers include an algorithm that tries to match two outcomes of the measurements. If n is the total number of outcomes and n' is the number of outcomes that are matched by the program, then the verifier makes the acceptance decision when $n' > n(1-\rho)$, where

*Correspondence to: Vladimir B. Balakirsky, Institute for Experimental Mathematics, 45326 Essen, Germany.
†E-mail: vinck@iem.uni-due.de

$\rho \in [0, 1]$ is a fixed threshold. We will develop a probabilistic version of this approach by assuming that the measurements of biometrical parameters of a person are obtained as results of transmission of the measurements received at the enrollment stage over a memoryless channel. The inclusion of the matching program into the channel brings the model where we only know the probability that the input symbol is not changed. Suppose that this probability is equal to $1 - \varepsilon \in (1/2, 1)$. Having received an output vector of length n, the verifier has to decide whether the vector is generated by a biometric source with the known probabilistic description ('reject') or obtained as a result of transmission of the sample vector over such a channel ('accept'). The acceptance decision is made if the number of symbols coinciding with the corresponding sample symbols is greater than $n(1 - \rho)$.

One of the main goals of biometrical authentication is generating a probability distribution (PD) over the set of passwords, which is close to the uniform PD. In this case, the scheme is highly protected against attacks when an attacker does not know the password of the person, and a possible measure of closeness is the sum of the logarithm of the probability of the most likely password and its length. A different attacker, who knows the password, is interested in guessing the biometric vector, and his power can be estimated by the average and the maximum probability of the correct guess.

We will present an approach that can be used to reach a desired trade-off between the parameters mentioned above. Section 2 is devoted to the description of a formal model for biometrical authentication. A general statement about the false acceptance and the false rejection rates for the component-wise transformation of input vectors to passwords is presented in Section 3, and the construction of the mapping for a given PD over the outcomes of biometrical measurements is given in Section 4. The use of the proposed approach to real data is illustrated for the DNA measurements in Section 5.

2. Mathematical Model for Biometrical Authentication

Let us consider the biometrical authentication scheme in Figure 1. Let $\mathcal{B}_t = \{0, \dots, Q_t - 1\}$, $\mathcal{Z}_t = \{0, \dots, q_t - 1\}$, $t = 1, \dots, n$, be given finite sets and $\mathcal{B}^{(n)} = \mathcal{B}_1 \times \cdots \times \mathcal{B}_n$, $\mathcal{Z}^{(n)} = \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_n$. Let $\mathbf{b}^* = (b_1^*, \dots, b_n^*) \in \mathcal{B}^{(n)}$ denote a fixed vector of outcomes of biometrical measurements of a person. Let $F : \mathcal{B}^{(n)} \to \mathcal{Z}^{(n)}$ be a specified mapping. The encoder maps the vector $\mathbf{b}^*$ to the vector $\mathbf{z}^* = F(\mathbf{b}^*)$, which is stored in the DB under the name of the person and given to the verifier upon request. The quantity

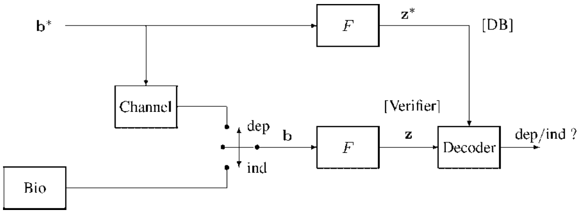

Fig. 1. The biometrical authentication scheme.

$$C \triangleq \frac{\sum_{t=1}^{n} \lceil \log Q_t \rceil}{\sum_{t=1}^{n} \lceil \log q_t \rceil}$$

can be understood as the compression factor. Another input to the verifier is the vector $\mathbf{z} = F(\mathbf{b})$, where $\mathbf{b} = (b_1, \dots, b_n) \in \mathcal{B}^{(n)}$ is the vector containing outcomes of biometrical measurements of some person. The verifier has to decide whether the vector $\mathbf{b}$ depends on the vector $\mathbf{b}^*$ or not. In the case of dependency, components of the vector $\mathbf{b}$, called the observation vector, are obtained as a result of transmission of the vector $\mathbf{b}^*$ over a noisy channel. In the case of independence, the vector $\mathbf{b}$, called the biometric vector, is generated by a biometric source independently of the vector $\mathbf{b}^*$. In the following considerations, $B^{(n)*} = (B_1^*, \dots, B_n^*)$ and $B^{(n)} = (B_1, \dots, B_n)$ denote the vectors of random variables.

Let the biometric and the observation vectors be generated by memoryless sources,

$$\Pr_{\text{bio}} \left\{ B^{(n)} = \mathbf{b} \right\} = \prod_{t=1}^{n} \Pr_{\text{bio}} \{ B_t = b_t \}$$

and

$$\Pr_{\text{err}} \left\{ B^{(n)} = \mathbf{b} \mid B^{(n)*} = \mathbf{b}^* \right\}$$
$$= \prod_{t=1}^{n} \Pr_{\text{err}} \left\{ B_t = b_t \mid B_t^* = b_t^* \right\}$$

where for all $t = 1, \dots, n$,

$$\left(\Pr_{\text{bio}} \{ B_t = b \}, b \in \mathcal{B}_t \right)$$

DOI: 10.1002/sec

is a known PD and

$$\left(\left(\underset{\text{err}}{\Pr}\left\{B_t = b \mid B_t^* = b^*\right\}, b \in \mathcal{B}_t\right), b^* \in \mathcal{B}_t\right)$$

is a collection of the conditional PD's such that

$$\underset{\text{err}}{\Pr}\left\{B_t = b^* \mid B_t^* = b^*\right\} = 1 - \varepsilon \tag{1}$$

for all $b^* \in \mathcal{B}_t$ We also assume that

$$\max_{b \in \mathcal{B}_t} \underset{\text{bio}}{\Pr}\left\{B_t = b\right\} < 1 - \varepsilon \tag{2}$$

In general, an authentication algorithm is specified by the sets $\mathcal{A}(\mathbf{z}^*) \subseteq \mathcal{Z}^{(n)}$, $\mathbf{z}^* \in \mathcal{Z}^{(n)}$, in such a way that the decision is 'dep' (accept) if $F(\mathbf{b}) \in \mathcal{A}(\mathbf{z}^*)$ and 'ind' (reject) if $F(\mathbf{b}) \notin \mathcal{A}(\mathbf{z}^*)$. The inequality (2) and our probabilistic model for the noise assume that the verifier has to make a decision on the basis of the value of the Hamming distance between received vectors

$$\mathcal{A}(\mathbf{z}^*) = \left\{\mathbf{z} \in \mathcal{Z}^{(n)} : d(\mathbf{z}, \mathbf{z}^*) \leq n\rho\right\} \tag{3}$$

where $\rho \in [0, 1]$ is a specified parameter and

$$d(\mathbf{z}, \mathbf{z}^*) \overset{\triangle}{=} \left|\left\{t \in \{1, \ldots, n\} : z_t \neq z_t^*\right\}\right|$$

The probabilities of incorrect decisions (the false acceptance and the false rejection rates) for the vector $\mathbf{b}^*$ that is mapped to the vector $\mathbf{z}^*$ are expressed as

$$\mathrm{FAR}(\mathbf{b}^*) = \sum_{\mathbf{b}: F(\mathbf{b}) \in \mathcal{A}(\mathbf{z}^*)} \underset{\text{bio}}{\Pr}\left\{B^{(n)} = \mathbf{b}\right\} \tag{4}$$

and

$$\mathrm{FRR}(\mathbf{b}^*) = \sum_{\mathbf{b}: F(\mathbf{b}) \notin \mathcal{A}(\mathbf{z}^*)} \underset{\text{err}}{\Pr}\left\{B^{(n)} = \mathbf{b} \mid B^{(n)*} = \mathbf{b}^*\right\} \tag{5}$$

We will restrict ourselves to the case when F is a component-wise mapping determined by functions $f_t : \mathcal{B}_t \to \mathcal{Z}_t$, $t = 1, \ldots, n$, and write $F(\mathbf{b}) = (f_1(b_1), \ldots, f_n(b_n))$ for all $\mathbf{b} \in \mathcal{B}^{(n)}$.

Let us address the cryptographic issues of the biometrical authentication. One can easily see that if an attacker wants to guess the biometric vector of a person, then the best estimate is the vector having the maximum probability. The probability that this guess is correct is equal to

$$\hat{\omega} \overset{\triangle}{=} \prod_{t=1}^{n} \hat{\omega}_t$$

where

$$\hat{\omega}_t \overset{\triangle}{=} \max_{b \in \mathcal{B}_t} \omega_t(b)$$

and

$$\omega_t(b) \overset{\triangle}{=} \underset{\text{bio}}{\Pr}\left\{B_t = b\right\}, \quad b \in \mathcal{B}_t$$

The transformation of biometric vectors to passwords and changing the attacker's task as guessing the password makes the probability of success equal to

$$\hat{\pi} \overset{\triangle}{=} \prod_{t=1}^{n} \hat{\pi}_t$$

where

$$\hat{\pi}_t \overset{\triangle}{=} \max_{z \in \mathcal{Z}_t} \pi_t(z)$$

and

$$\pi_t(z) = \sum_{b: f_t(b) = z} \underset{\text{bio}}{\Pr}\left\{B_t = b\right\}, \quad z \in \mathcal{Z}_t \tag{6}$$

is the PD over the tth component of the passwords. Furthermore, the best prediction of the input biometric vector, given the password $\mathbf{z}^*$ is the biometric vector having the maximum probability among the vectors mapped to $\mathbf{z}^*$ As it is easy to see, the probability that this prediction is correct can be computed as

$$\hat{\gamma}(\mathbf{z}^*) \overset{\triangle}{=} \prod_{t=1}^{n} \hat{\gamma}_t(z_t^*)$$

where

$$\hat{\gamma}_t(z_t^*) \overset{\triangle}{=} \max_{b \in \mathcal{B}_t} \gamma_t(b|z_t^*)$$

and

$$\left(\gamma_t(b|z_t^*), b \in \mathcal{B}_t\right)$$

DOI: 10.1002/sec

Table I. Example of the mapping $f_t : \{0, \dots, 9\} \to \{0, \dots, 3\}$.

$b =$	9	7	3	5	6	0	8	1	2	4
$\omega_t(b) =$	0.2401	0.0081	0.2254	0.0342	0.1862	0.0529	0.0882	0.0874	0.0414	0.0361
$z = f_t(b) =$	0		1		2		3			
$\pi_t(z) =$	0.2482		0.2596		0.2381		0.2531			
$\gamma_t(b\|z) =$	0.9674	0.0326	0.8682	0.1318	0.7820	0.2180	0.3485	0.3453	0.1636	0.1426
$\hat{\gamma}_t, \overline{\gamma}_t =$	0.9674, 0.7401									
$\hat{\pi}_t =$	0.2596									
$\hat{\omega}_t =$	0.2401									

where

$$\gamma_t(b|z_t^*) \triangleq \frac{1}{\pi_t(z_t^*)} \begin{cases} \omega_t(b), & \text{if } f_t(b) = z_t^*, \\ 0, & \text{if } f_t(b) \neq z_t^*, \end{cases}$$

is the z_t^* conditional PD over the tth component of the passwords.

The notation above is illustrated in Table I, where $\overline{\gamma}_t \triangleq \sum_{z^* \in \mathcal{Z}_t} \pi_t(z)\hat{\gamma}_t(z^*)$ and $\hat{\gamma} \triangleq \max_{z^* \in \mathcal{Z}_t} \hat{\gamma}_t(z^*)$ denote the average and the maximum probability of the successful guess of the input biometric vector. In the stationary case, when the biometric source is described by the presented PD for all $t = 1, \dots, n$, the probabilities of the correct guess of the biometric vector and the password are equal to $(0.2401)^n$ and $(0.2596)^n$, i.e., they are close enough. The guessing algorithm for the attacker, who has access to the password and wants to find the biometric vector, depends on the password. The average and the maximum probabilities of the success are equal to $(0.7401)^n$ and $(0.9674)^n$, respectively.

3. Performance of the Authentication Algorithm with a Component-Wise Mapping of the Input Vectors to Passwords

The formulations and proofs of the presented results are based on the generating function technique developed in probability theory [2]. Given a polynomial $G(\lambda)$ of a formal variable λ, write

$$G(\lambda) = \sum_{r \geq 0} \mathrm{Coef}_r[G(\lambda)]\lambda^r$$

For example, if

$$G(\lambda) = \left(\frac{1}{2} + \frac{1}{2}\lambda\right)\left(\frac{1}{5} + \frac{4}{5}\lambda\right)\left(\frac{1}{10} + \frac{9}{10}\lambda\right)$$

then

$$G(\lambda) = 0.01 + 0.14\lambda + 0.49\lambda^2 + 0.36\lambda^3$$

and

$$(\mathrm{Coef}_0[G(\lambda)], \dots, \mathrm{Coef}_3[G(\lambda)]) = (0.01, 0.14, 0.49, 0.36)$$

Proposition. *If the function F is a component-wise mapping $\mathcal{B}^{(n)} \to \mathcal{Z}^{(n)}$ and the decision sets are defined by (3), then, for any vector $\mathbf{b}^* \in \mathcal{B}^{(n)}$ mapped to the vector $\mathbf{z}^* \in \mathcal{Z}^{(n)}$ and any $\rho \in (0, 1)$,*

$$\mathrm{FAR}(\mathbf{b}^*) = \sum_{r=0}^{\lfloor n\rho \rfloor} \mathrm{Coef}_r \left[\prod_{t=1}^{n} \left(\pi_t(z_t^*) + (1 - \pi_t(z_t^*))\lambda\right) \right] \quad (7)$$

$$\mathrm{FRR}(\mathbf{b}^*) \leq \sum_{r=\lfloor n\rho \rfloor + 1}^{n} \binom{n}{r} (1-\varepsilon)^{n-r}\varepsilon^r \quad (8)$$

where the entries of the PD's $\pi_1, \dots, \pi_n$ are defined in Equation (6). Furthermore,

$$\max_{\mathbf{b}^* \in \mathcal{B}^{(n)}} \mathrm{FAR}(\mathbf{b}^*) = \sum_{r=0}^{\lfloor n\rho \rfloor} \mathrm{Coef}_r \left[\prod_{t=1}^{n} (\hat{\pi}_t + (1 - \hat{\pi}_t)\lambda) \right]$$

where $\hat{\pi}_t$ is the maximum probability of a symbol that can occur as the t th component of the password.

Corollary.

1. *If $\mathcal{Z}^{(n)} = \mathcal{B}^{(n)}$ and F is the identity mapping, then*

$$\max_{\mathbf{b}^* \in \mathcal{B}^{(n)}} \mathrm{FAR}(\mathbf{b}^*) = \sum_{r=0}^{\lfloor n\rho \rfloor} \mathrm{Coef}_r \left[\prod_{t=1}^{n} (\hat{\omega}_t + (1 - \hat{\omega}_t)\lambda) \right] \quad (9)$$

where $\hat{\omega}_t$ is the maximum probability of a symbol that can occur as the tth component of the biometric

DOI: 10.1002/sec

vector. Furthermore,

$$\mathrm{FRR}(\mathbf{b}^*) = \sum_{r=\lfloor n\rho\rfloor+1}^{n} \binom{n}{r}(1-\varepsilon)^{n-r}\varepsilon^r$$

for all $\mathbf{b}^* \in \mathcal{B}^{(n)}$.

2. *If* $1/\hat{\omega}_1, \ldots, 1/\hat{\omega}_n$ *are equal to integers* $q_1, \ldots, q_n$ *and the function F is assigned in such a way that the PD's* $\pi_1, \ldots, \pi_n$ *are uniform, then*

$$\mathrm{FAR}(\mathbf{b}^*) = \sum_{r=0}^{\lfloor n\rho\rfloor} \mathrm{Coef}_r \left[\prod_{t=1}^{n}(\hat{\omega}_t + (1-\hat{\omega}_t)\lambda)\right] \quad (10)$$

for all $\mathbf{b}^* \in \mathcal{B}^{(n)}$

The proof is given in the Appendix.

Obviously, the best performance is reached when the biometric vectors themselves are stored in the DB and such a conclusion readily follows from Corollary 1. The comparison of Equations (9) and (10) shows that, under the conditions of Corollary 2, one can get the same false acceptance rate for all vectors $\mathbf{b} \in \mathcal{B}^{(n)}$ as the maximum false acceptance rate for the identity mapping F. On the other hand, the length of the stored passwords is decreased from $\sum_{t=1}^{n}\lceil\log Q_t\rceil$ to $\sum_{t=1}^{n}\log(1/\hat{\omega}_t)$. Moreover, the PD over the set of passwords is uniform. By decreasing $q_1, \ldots, q_n$ one can keep the uniform PD, decrease the passwords lengths, and decrease the probability of the correct guess of the biometric vector for a given password. However, the false acceptance rate will also be decreased. The constraints of Corollary 2 seem to be very strong, but they can be 'almost' satisfied while processing real data, as it will be demonstrated in Section 5 where we consider three cases: $q_t \in \{Q_t, \lceil\log(1/\hat{\omega}_t)\rceil, 2\}$ for all $t = 1, \ldots, n$.

4. Constructing the Transformation *f*

In the following considerations, we omit the index $t \in \{1, \ldots, n\}$ for a formal brevity and extend the ideas presented in Reference [2] for the continuous case to the discrete case.

Let us determine the function f by a partitioning of the set $\mathcal{B}$ by q pairwise disjoint subsets $\mathcal{F}(0), \ldots, \mathcal{F}(q-1)$ in such a way that $f(b) = z$ is equivalent to $b \in \mathcal{F}(z)$. For example, the function *f* with $f(0) = f(2) = 0$, $f(1) = 1$, $f(3) = 2$ is specified by the sets $\mathcal{F}(0) = \{0, 2\}$, $\mathcal{F}(1) = \{1\}$, $\mathcal{F}(2) = \{3\}$.

'A greedy algorithm' for constructing the sets $\mathcal{F}(0), \ldots, \mathcal{F}(q-1)$ is presented below. The algorithm can be interpreted as sequential constructing $q-2$ bins. The probability of a bin is understood as the sum of probabilities of its entries and the value of the distortion as the absolute value of the difference between this probability and 2^{-q}. If there is an entry, which was not put to the bins with the smaller indices whose probability decreases the current distortion of the bin, then this entry is included. Otherwise, the bin is closed and we start to form the next bin. The $(q-1)$-st bin contains the entries that were not included in any of the previous bins.

F1: Set $z = 0$.
F2: Set $\mathcal{F}(z) = \emptyset$, $S = 0$, $\Delta = 2^{-q}$.
F3: Denote

$$\mathcal{B}_0 = \mathcal{F}(z) \bigcup \left[\bigcup_{z'=0}^{z-1} \mathcal{F}(z')\right]$$

$$\Delta_0 = \min_{b \in \mathcal{B}\setminus\mathcal{B}_0} \left|2^{-q} - (S + \omega(b))\right|$$

$$b_0 = \arg\min_{b \in \mathcal{B}\setminus\mathcal{B}_0} \left|2^{-q} - (S + \omega(b))\right|$$

If $\Delta_0 > \Delta$ then go to 5.

F4: Include b_0 into the set $\mathcal{F}(z)$, increase S by $\omega(b_0)$, and substitute Δ_0 for Δ.
F5: Increase z by 1. If $z \le q-2$, then go to 2.
F6: Set

$$\mathcal{F}(q-1) = \mathcal{B}\setminus \left[\bigcup_{z=0}^{q-2} \mathcal{F}(z)\right]$$

F7: Output the sets $\mathcal{F}(0), \ldots, \mathcal{F}(q-1)$ End.

The F1–F7 algorithm was used to generate data in Table I for $q_t = 4$. One can easily check that for the considered PD over the set $\{0, \ldots, 9\}$ the algorithm outputs $\mathcal{F}_0(z) = \{9, 7\}$, $\mathcal{F}_1(z) = \{3, 5\}$, $\mathcal{F}_2(z) = \{6, 0\}$, $\mathcal{F}_3(z) = \{8, 1, 2, 4\}$.

5. Application of the Authentication Algorithm to the DNA Measurements

We will use the mathematical model for the DNA measurements developed in References [3,4]. Suppose that there are n sources. Let the tth source generate a pair of integers according to the PD

$$\Pr_{\mathrm{DNA}}\left\{(R_{t,1}, R_{t,2}) = (r_{t,1}, r_{t,2})\right\} = p_t(r_{t,1})p_t(r_{t,2})$$

DOI: 10.1002/sec

where $r_{t,1}, r_{t,2} \in \mathcal{R}_t = \{c_t, \ldots, c_t + k_t - 1\}$ and integers $c_t, k_t > 0$ are given. The outcome of the t th measurement is defined as

$$R_t \overset{\triangle}{=} \big(\min\{R_{t,1}, R_{t,2}\}, \max\{R_{t,1}, R_{t,2}\}\big) \qquad (11)$$

Hence, for all $i \in \mathcal{R}_t$,

$$\Pr_{\text{DNA}} \{R_t = (i, j)\} = \begin{cases} 0, & \text{if } j \in \{0, \ldots, i-1\}, \\ p_t^2(i), & \text{if } j = i, \\ 2p_t(i)p_t(j), & \text{if } j \in \{i+1, \ldots, c_t + k_t - 1\} \end{cases}$$

We assume that $R_1, \ldots, R_n$ are mutually independent pairs of random variables, i.e.,

$$\Pr_{\text{DNA}} \left\{ R^{(n)} = \mathbf{r} \right\} = \prod_{t=1}^{n} \Pr_{\text{DNA}} \{R_t = r_t\}$$

where $R^{(n)} = (R_1, \ldots, R_n)$ and $\mathbf{r} = (r_1, \ldots, r_n)$, $r_t \in \mathcal{R}_t \times \mathcal{R}_t$. To make the notation consistent with the notation of Section 2, let us map $Q_t = k_t(k_t + 1)/2$ pairs $r_t = (i_t, j_t)$, where $j_t \geq i_t$, that can occur with positive probability to integers $b \in \mathcal{B}_t = \{0, \ldots, Q_t - 1\}$ in a lexicographic order.

The formalization above appears because the DNA measurements are usually understood as measurements of the numbers of repeats of certain motifs in the paternal and the maternal allele where the measuring device cannot distinguish between data coming from different allele. Therefore, the outcomes $r_{t,1}, r_{t,2}$ can be represented as observations of the sets $\{r_{t,1}, r_{t,2}\}$. This information can be equivalently presented as the value of the random variable R_t defined in Equation (11).

Parameters of the PDs obtained from the DNA measurements are given in Table II. One can see that the storage of biometric vectors requires 140 bits and the PD over these vectors is non–uniform, as the probability of the most likely vector is equal to $2^{-78.8}$. The encoding with parameters $q_t = \lceil \log 1/\hat{\omega}_t \rceil$ creates passwords of length 68 with the PD close to the uniform PD (the probability of the most likely vector is equal to $2^{-66.7}$). However, the maximum

Table II. The DNA measurements: some characteristics of three variants of the encoding.

				$C = 140/140$			$C = 140/68$			$C = 140/28$		
t	Name	$\log Q_t$	$\log \hat{\omega}_t$	$\log q_t$	$\log \hat{\pi}_t$	$\log \hat{\gamma}_t$	$\log q_t$	$\log \hat{\pi}_t$	$\log \hat{\gamma}_t$	$\log q_t$	$\log \hat{\pi}_t$	$\log \hat{\gamma}_t$
1	**D8S1179**	4.39	−3.01	5	−3.01	0	3	−2.94	−0.00	1	−1.00	−2.01
2	**D3S1358**	3.91	−2.87	4	−2.87	0	2	−1.99	−0.87	1	−1.00	−1.87
3	**VWA**	4.39	−3.12	5	−3.12	0	3	−2.94	−0.11	1	−1.00	−2.13
4	**D7S820**	4.39	−3.25	5	−3.25	0	3	−2.92	−0.24	1	−1.00	−2.25
5	**ACTBP2**	7.71	−6.13	8	−6.13	0	6	−5.98	−0.13	1	−1.00	−5.13
6	**D7S820**	4.81	−3.31	5	−3.31	0	3	−2.92	−0.30	1	−1.00	−2.31
7	**FGA**	5.49	−3.54	6	−3.54	0	3	−2.99	−0.54	1	−1.00	−2.54
8	**D21S11**	4.81	−3.01	5	−3.01	0	3	−2.94	−0.02	1	−1.00	−2.02
9	**D18S51**	5.78	−4.43	6	−4.43	0	4	−3.85	−0.44	1	−1.00	−3.43
10	**D19S433**	4.39	−2.33	5	−2.33	0	2	−1.99	−0.33	1	−0.98	−1.35
11	**D13S317**	4.81	−2.56	5	−2.56	0	2	−1.98	−0.58	1	−0.99	−1.57
12	**TH01**	3.32	−2.07	4	−2.07	0	2	−1.93	−0.04	1	−1.00	−1.07
13	**D2S138**	6.04	−4.23	7	−4.23	0	4	−3.98	−0.22	1	−1.00	−3.23
14	**D16S539**	4.81	−2.25	5	−2.25	0	2	−1.94	−0.26	1	−1.00	−1.26
15	**D5S818**	3.91	−1.81	4	−1.81	0	1	−1.00	−0.81	1	−1.00	−0.81
16	**TPOX**	3.91	−1.79	4	−1.79	0	1	−0.86	−0.93	1	−0.86	−0.93
17	**CF1PO**	3.91	−2.16	4	−2.16	0	2	−1.90	−0.25	1	−0.95	−1.21
18	**D8S1179**	5.49	−3.15	6	−3.15	0	3	−2.99	−0.15	1	−1.00	−2.15
19	**VWA1**	4.39	−3.12	5	−3.12	0	3	−2.94	−0.11	1	−1.00	−2.13
20	**PentaD**	5.17	−3.13	6	−3.13	0	3	−2.95	−0.13	1	−1.00	−2.14
21	**PentaE**	6.91	−4.02	7	−4.02	0	4	−3.89	−0.02	1	−1.00	−3.02
22	**DYS390**	4.39	−2.06	5	−2.06	0	2	−1.95	−0.06	1	−1.00	−1.06
23	**DYS429**	3.91	−1.78	4	−1.78	0	1	−1.00	−0.79	1	−1.00	−0.79
24	**DYS437**	2.58	−1.58	3	−1.58	0	1	−1.00	−0.57	1	−1.00	−0.57
25	**DYS391**	3.32	−1.11	4	−1.11	0	1	−1.00	−0.11	1	−1.00	−0.11
26	**DYS385**	5.17	−1.72	6	−1.72	0	1	−0.98	−0.74	1	−0.98	−0.74
27	**DYS389I**	2.58	−1.18	3	−1.18	0	1	−0.99	−0.17	1	−0.99	−0.17
28	**DYS389II**	3.91	−2.04	4	−2.04	0	2	−1.99	−0.04	1	−1.00	−1.04
	$\sum$	128.6	−78.8	140	−78.8	0	68	−66.7	−9.0	28	−27.8	−49.0

DOI: 10.1002/sec

probability of the correct guess of the biometric vector is equal to $2^{-9.0}$. This probability can be decreased to $2^{-27.8}$ by assigning $q_t = 2,\ t = 1, \dots, n$, which creates passwords of length 28 bits. However, the false acceptance rate will be also increased, as it is illustrated in the end of the section.

A possible implementation of our transformations of input biometrical data is presented by the example below.

Example (the quantities below describe the **TH01** allele in Table II, $t = 12$)**.** Let $c_t = 6$, $k_t = 4$, and $(p_t(6), \dots, p_t(9)) = (0.23, 0.19, 0.09, 0.49)$. Then the entries of the matrix $\left[p_t(i)p_t(j)\right]_{i,j=6,\dots,9}$ are as follows:

	$j = 6$	$j = 7$	$j = 8$	$j = 9$
$i = 6$	0.0529	0.0437	0.0207	0.1127
$i = 7$	0.0437	0.0361	0.0171	0.0931
$i = 8$	0.0207	0.0171	0.0081	0.0441
$i = 9$	0.1127	0.0931	0.0441	0.2401

To construct the PD ω_t, we transform this matrix to the right triangular matrix below. The entries above the diagonal are doubled, and the entries below the diagonal are replaced with zeroes,

	$j = 6$	$j = 7$	$j = 8$	$j = 9$
$i = 6$	0.0529	0.0874	0.0414	0.2254
$i = 7$		0.0361	0.0342	0.1862
$i = 8$			0.0081	0.0882
$i = 9$				0.2401

Let $q_t = 4$ and let the sets $\mathcal{F}_t(0), \dots, \mathcal{F}_t(3)$ be constructed by the F1–F7 algorithm. Then $\mathcal{F}_0(z) = \{9, 7\}$, $\mathcal{F}_1(z) = \{3, 5\}$, $\mathcal{F}_2(z) = \{6, 0\}$, $\mathcal{F}_3(z) = \{8, 1, 2, 4\}$. The PDs obtained using this partitioning of the set $\{0, \dots, 9\}$ were presented in Table I.

To implement the encoding, the authentication scheme needs data of Table III. The outcome of the measurements $(i_{t,1}, j_{t,2})$ has to be found in the first row, and the symbol z in the corresponding column is sent to the output. Notice that the index of the pair $(i_{t,1}, j_{t,2})$ can be easily computed from $c_t = 6$ and $k_t = 4$. Therefore, the storage of the first and the second rows of the table is not necessary.

Suppose that there are n sources. Let the t-th source generate a pair of integers $(i_t, j_t) \in \{c_t, \dots, c_t + k_t - 1\}$, where the integers c_t and k_t are known and $i_t \leq j_t$. The PD over $Q_t = k_t(k_t + 1)/2$ pairs is known, and we map them to integers $b \in \mathcal{B}_t = \{0, \dots, Q_t - 1\}$ in a lexicographic order.

Table III. The encoding table for the example with the **TH01** allele.

$(i, j) =$	(6,6)	(6,7)	(6,8)	(6,9)	(7,7)	(7,8)	(7,9)	(8,8)	(8,9)	(9,9)
$b_t =$	0	1	2	3	4	5	6	7	8	9
$z_t =$	3	3	2	1	2	1	3	0	3	0

Table IV. The DNA measurements: the values of the false acceptance rate and the parameter ε(FAR) such that FRR $\leq$ FAR for all $\varepsilon < \varepsilon$(FAR).

	$C = 140/140$		$C = 140/68$		$C = 140/28$	
$n\rho$	FAR	ε(FAR)	FAR	ε(FAR)	FAR	ε(FAR)
0	7.7e–24		8.3e–21		8.0e–04	0.0008
1	1.9e–21		1.6e–18		5.4e–03	0.0054
2	2.0e–19		1.4e–16		1.7e–02	0.0171
3	1.3e–17		7.0e–15		3.8e–02	0.0376
4	5.9e–16	0.0001	2.4e–13	0.0003	6.7e–02	0.0672
5	2.0e–14	0.0006	6.2e–12	0.0016	1.1e–01	0.1524
6	5.0e–13	0.0024	1.2e–10	0.0053		
7	1.0e–11	0.0066	1.9e–09	0.0129		
8	1.7e–10	0.0147	2.3e–08	0.0259		
9	2.3e–09	0.0278	2.3e–07	0.0455		
10	2.6e–08	0.0471	2.0e–06	0.0726		
11	2.5e–07	0.0735	1.4e–05	0.1077		
12	2.0e–06	0.1075	8.3e–05	0.1508		

DOI: 10.1002/sec

The formalization above appears because the DNA measurements are usually understood as measurements of the numbers of repeats of certain motifs in the paternal and the maternal allele where the measuring device cannot distinguish between data coming from different allele. Therefore, the outcomes $r_{t,1}, r_{t,2}$ can be represented as observations of the sets $\{r_{t,1}, r_{t,2}\}$. This information can be equivalently presented as the value of the random variable $(i_t = \min\{r_{t,1}, r_{t,2}\}, j_t = \max\{r_{t,1}, r_{t,2}\})$.

Some numerical results are given in Table IV where we show the values of the false acceptance rate as a function of ρ and the length of passwords.

If the outcomes of the measurements are stored in the DB one-by-one, i.e., $C = 140/140$, then the false acceptance and the false rejection rates are very small. In particular, if the verifier accepts the identity claim when the observed vector differs from the sample vector in $d \leq n\rho = 12$ positions out of $n = 28$ positions, then the false acceptance rate is not greater than 2.0e-06. The same upper bound is also valid for the false rejection rate when $\varepsilon < 0.1075$. These conclusions illustrate the point that the DNA data can be effectively used to distinguish between different persons at the verification stage. However, the PD over vectors of length 140 bits is non-uniform and the biometrical data are not protected. As an alternative, we can extract 68 almost uniformly distributed bits. The false acceptance and the false rejection rates slightly increase; in particular, the same assignment for the parameter ρ brings the false acceptance rate 8.3e-05. Another alternative is the extraction of 28 bits when we get only one bit per observation. Then data are highly protected against attacks, but the false acceptance and the false rejection rates essentially increase.

6. Concluding Remarks

We introduced a simple probabilistic model specified by the probability that the measured biometrical parameter is not changed at the verification stage as compared to the enrollment stage. The algorithm designed for this model can be easily implemented, and its robustness allows one to reach a desired trade-off between different parameters characterizing the performance of authentication schemes.

Appendix: Proof of Proposition

By the definition of the false acceptance rate in Equation (4) and the set $\mathcal{A}$ in Equation (3),

$$\mathrm{FAR}(\mathbf{b}^*) = \Pr\left\{\sum_{t=1}^{n} Y_t \leq n\rho\right\}$$

where $Y_1, \ldots, Y_n$ are independent binary random variables such that the random variable Y_t takes value 1 with probability $\pi_t(z_t^*)$ and value 0 with probability $1 - \pi_t(z_t^*)$. The probability that the sum $Y_1 + \cdots + Y_n$ is equal to $r \in \{0, \ldots, \lfloor n\rho \rfloor\}$ is equal to the sum of terms

$$\prod_{t=1}^{n} \begin{cases} \pi_t(z_t^*), & \text{if } j_t = 0, \\ 1 - \pi_t(z_t^*), & \text{if } j_t = 1 \end{cases}$$

taken over all vectors $(j_1, \ldots, j_n) \in \{0, 1\}^n$ having the Hamming weight r. By the definition of the product of polynomials,

$$\Pr\left\{\sum_{t=1}^{n} Y_t = r\right\}$$
$$= \mathrm{Coef}_r\left[\prod_{t=1}^{n} \left(\pi_t(z_t^*) + (1 - \pi_t(z_t^*))\lambda\right)\right]$$

and equality (7) follows.

Let $\hat{Y}_t$ be a binary random variable chosen independently of $Y_1, \ldots, Y_n$, which takes value 1 with probability $\hat{\pi}_t$ and value 0 with probability $1 - \hat{\pi}_t$. Then

$$\Pr\left\{\sum_{t=1}^{n} \hat{Y}_t = r\right\} = \mathrm{Coef}_r\left[\prod_{t=1}^{n} (\hat{\pi}_t + (1 - \hat{\pi}_t)\lambda)\right]$$
$$= \binom{n}{r} (\hat{\pi}_t)^r (1 - \hat{\pi}_t)^{n-r}$$

and the inequality

$$\mathrm{FAR}(\mathbf{b}^*)$$
$$\leq \sum_{r=0}^{\lfloor n\rho \rfloor} \mathrm{Coef}_r\left[\prod_{t=1}^{n} (\hat{\pi}_t + (1 - \hat{\pi}_t)\lambda)\right] \quad (12)$$

DOI: 10.1002/sec

follows from

$$\Pr\left\{\sum_{t=1}^{n} Y_t \le n\rho\right\} \le \Pr\left\{\sum_{t=1}^{n} \hat{Y}_t \le n\rho\right\} \qquad (13)$$

To prove (13) we write

$$\begin{aligned}
&\Pr\left\{\sum_{t=1}^{n} Y_t \le n\rho\right\} \\
&= \sum_{r=0}^{\lfloor n\rho\rfloor-1} \Pr\left\{\sum_{t=2}^{n} Y_t = r\right\} \\
&\quad + \Pr\{Y_1 = 0\}\Pr\left\{\sum_{t=2}^{n} Y_t = \lfloor n\rho\rfloor\right\} \\
&\le \sum_{r=0}^{\lfloor n\rho\rfloor-1} \Pr\left\{\sum_{t=2}^{n} Y_t = r\right\} + \hat{\pi}_1 \Pr\left\{\sum_{t=2}^{n} Y_t = \lfloor n\rho\rfloor\right\}
\end{aligned}$$

Hence, the false acceptance rate is not decreased if the PD of the first component of the vector $(Y_1, \ldots, Y_n)$ is changed and the vector is replaced with the vector $(\hat{Y}_1, Y_2, \ldots, Y_n)$. By repeating this argument, we arrive at the vector $(\hat{Y}_1, \ldots, \hat{Y}_n)$ and complete the proof.

Let us denote

$$\Pi_t(z|b_t^*) \overset{\triangle}{=} \Pr_{\mathrm{err}}\left\{Z_t = z|B_t^* = b_t^*\right\}$$

Then we write

$$\Pi_t(z|b_t^*) = \sum_{b:\, f_t(b)=z} \Pr_{\mathrm{err}}\left\{B_t = b|B_t^* = b_t^*\right\} = 1 - \varepsilon$$

for all $z \in \mathcal{Z}_t$ and notice that (1) implies

$$\begin{aligned}
&\Pi_t(f_t(b_t^*)|b_t^*) \\
&= \sum_{b:\, f_t(b)=f_t(b_t^*)} \Pr_{\mathrm{err}}\left\{B_t = b|B_t^* = b_t^*\right\} \\
&\ge \Pr_{\mathrm{err}}\left\{B_t = b_t^*|B_t^* = b_t^*\right\} = 1 - \varepsilon \qquad (14)
\end{aligned}$$

Represent the false rejection rate defined in Equation (5) as

$$\mathrm{FRR}(\mathbf{b}^*) = \Pr\left\{\sum_{t=1}^{n} E_t > n\rho\right\} \qquad (15)$$

where $E_1, \ldots, E_n$ are independent binary random variables such that the random variable E_t takes value 1 with probability $1 - \Pi_t(f_t(b_t^*)|b_t^*)$ and value 0 with probability $\Pi_t(f_t(b_t^*)|b_t^*)$. Then, for all $r \in \{\lfloor n\rho\rfloor + 1, \ldots, n\}$,

$$\begin{aligned}
&\Pr\left\{\sum_{t=1}^{n} E_t = r\right\} \\
&= \mathrm{Coef}_r\left[\prod_{t=1}^{n}\left(\Pi_t(f_t(b_t^*)|b_t^*) + (1 - \Pi_t(f_t(b_t^*)|b_t^*))\lambda\right)\right]
\end{aligned}$$

Considerations similar to the ones used in the proof of inequality (12) and (14) bring

$$\begin{aligned}
\Pr\left\{\sum_{t=1}^{n} E_t = r\right\} &\le \mathrm{Coef}_r\left[\prod_{t=1}^{n}(1 - \varepsilon + \varepsilon\lambda)\right] \\
&= \binom{n}{r}(1-\varepsilon)^{n-r}\varepsilon^r
\end{aligned}$$

and (8) follows from (15).

If F is the identity mapping, then inequality (14) is tight, and Equation (8) holds with the equality.

Acknowledgement

This work was partially supported by the DFG.

References

1. Bolle RM, Connell JH, Pankanti S, Ratha NK, Senior AW. *Guide to Biometrics*. Springer: New York, 2004.
2. Feller WW. An introduction to probability theory and its applications. Vol 1 3rd edn. Wiley: New York, 1968.
3. Korte U, Krawczak M, Merkle J, *et al.* A cryptographic biometric authentication system based on genetic fingerprints. *Sicherheit* 2008; 263–276.
4. Balakirsky VB, Ghazaryan AR, Han Vinck AJ. Additive block coding schemes for biometric authentication with the DNA data. *Lecture Notes in Computer Science: The 1-st European Workshop on Biometrics and Identity Management*, Schouten B, *et al.* (eds). 2008; **5372**, 160–169.

DOI: 10.1002/sec