15

Block Coding Schemes Designed for Biometric Authentication

Vladimir B. Balakirsky[1] and A. J. Han Vinck[2]
[1]*Data Security Association "Confident", American University of Armenia*
[2]*Institute for Experimental Mathematics*
[1]*Russia, Armenia*
[2]*Germany*

1. Introduction

We address the biometric authentication setup where the outcomes of biometric observations received at the verification stage are compared with the sample data formed at the enrollment stage. The result of comparison is either the acceptance or the rejection of the identity claim. The acceptance decision corresponds to the case when the analyzed values belong to the same person.

A possible solution to the problem, called the direct authentication, is implemented when the outcomes of biometric observations at the enrollment stage are stored in the database, and they are available to the verifier. The possible incorrect verifier's decisions are caused by the fact that these observations are noisy. The probabilities of errors are called the false rejection and the false acceptance rates. The features of the direct authentication are as follows: 1) data compression is not included at the enrollment stage; 2) the scheme does not require an additional external randomness; 3) if the stored data become available to an attacker, then he knows the outcomes of biometric observations of the person and can pass through the verification stage with the acceptance decision by presenting these data to the verifier. The considered below coding approaches to the problem require an external randomness and relax the constraint that the database has to be protected against reading. These approaches include the additive and the permutation coding schemes.

Both the direct authentication and an additive coding scheme are illustrated using a proposed mathematical model for the DNA measurements. We present the model and describe a data compression method that can be used to approach a uniform probability distribution over the obtained data for their further use in the additive scheme and other purposes. The processing of the DNA data also serves as an example of possible processing data generated by an arbitrary memoryless source.

The additive block coding scheme can be viewed as a variant of stream ciphering scheme where the data, to be hidden, are added to a key. The subtraction of the noisy version of the data creates a corrupted version of the key. If the key is a codeword of a code having certain error–correcting property, then the fact, whether the key can be reconstructed or not,

characterizes the level of the noise. In the permutation scheme, the enciphering of the input data is organized by choosing a permutation, which maps the biometric vector to a key vector. There are many permutations that can be used for this purpose, and it gives additional possibilities to the designer of the verification scheme.

The efficiency of cryptographic schemes, like the additive and the permutation schemes, is measured by the difference between the probabilities of the successful attack by an attacker, who either knows the content of the database or ignorant about these data. The additive scheme is efficient when the probability distribution over the input vectors is close to a uniform distribution. This requirement is less critical for the permutation scheme, but input vectors have to be represented by binary vectors having a fixed number of ones. We will present a simple numerical example of the implementation of the permutation scheme and describe an algorithm for the transformation of an arbitrary binary vector to a balanced vector having the same number of zeroes and ones.

There is a number of open problems in the implementation of coding schemes. One of the main problems is the representation of real biometric data in digital format, which allows one to use the memoryless assumption about the data and the Hamming distance as the measure of closeness of two observations. Another class of problems is constructing the specific codes and the decoding algorithms having a low computational complexity. We also believe that there is a request for a general theory of processing noisy data, since the known solutions in biometrics are mostly oriented to specific measurements (fingerprints, iris, palmprints, etc.) and a particular application.

The authentication problem belongs to the list of basic problems that have to be solved in the biometric direction, and it is included in the most of the books on biometrics (see Bolle et. al (2004), for example). The additive block coding scheme was suggested in Juels & Wattenberg (1999). The close relationships between the additive scheme and the wiretap channel, introduced in Wyner (1975), where the verifier receives the signals from the outputs of two parallel channels in the legitimate case and the signals from only one of channels in the case of the presence of an attacker. It implies the relevance of information and coding theory results (see Cohen & Zemor (2006), for example) to the investigation of the scheme. The permutation scheme was proposed in Dodis, et. al (2004) under the uniform probability distribution over the permutations. The algorithm for the mapping of an arbitrary binary vector to a balanced vector, which can be used in the permutation scheme, was described in Knuth (1986). The available DNA measurement data were received in the BioKey–STR project (Korte et. al (2008)).

The text of the chapter is a compressed version of the results in Balakirsky, Ghazaryan & Han Vinck (2006–2011). The general principles of constructing biometric authentication, which also include the points of rate–distortion coding, were presented in (2006a), (2006b). The described mathematical model for the DNA data was introduced in (2008a), and the data processing scheme was studied in (2009b) as an extension of the transformations for continuous random variables described in (2007). The similar analysis is relevant to the constructing passwords from biometric data, as it is indicated in (2010). The general expressions for the additive and the permutation block coding schemes for an arbitrary probability distribution over the biometric vectors are given in (2008a), (2009a). The standard technique of probability and coding theory, which is used in the chapter, can be found in Gallager (1968).

2. Notation and basic assumptions

Let $\mathcal{B} = \mathcal{B}_1 \times \cdots \times \mathcal{B}_n$, where $\mathcal{B}_t = \{0, \ldots, K_t - 1\}$ is a finite set containing K_t elements. We say that $\mathbf{b} = (b_1, \ldots, b_n) \in \mathcal{B}$ is a biometric vector and assume that the probability distribution

$$\omega = \Big(\omega(\mathbf{b}) = \Pr_{\text{bio}} \{ B = \mathbf{b} \},\ \mathbf{b} \in \mathcal{B} \Big)$$

is known. Moreover, let ω be a memoryless probability distribution, i.e.,

$$\omega(\mathbf{b}) = \prod_{t=1}^{n} \omega_t(b_t) \tag{1}$$

for all $\mathbf{b} \in \mathcal{B}$. We also write

$$\omega_t(b) = \Pr_{\text{bio}} \{ B_t = b \}$$

for all $b \in \mathcal{B}_t$. Denote the most likely biometric vector by $\mathbf{b}^* = (b_1^*, \ldots, b_n^*)$,

$$\mathbf{b}^* = \arg\max_{\mathbf{b} \in \mathcal{B}} \omega(\mathbf{b}).$$

Then, by (1),

$$b_t^* = \arg\max_{b \in \mathcal{B}_t} \omega_t(b), \quad t = 1, \ldots, n,$$

and

$$\omega(\mathbf{b}^*) = \prod_{t=1}^{n} \omega_t^*$$

where

$$\omega_t^* = \max_{b \in \mathcal{B}_t} \omega_t(b). \tag{2}$$

Furthermore, let

$$\overline{\omega}_t = \sum_{b \in \mathcal{B}_t} \omega_t^2(b) \tag{3}$$

and

$$H(\omega_t) = -\sum_{b=0}^{q_t - 1} \omega_t(b) \log \omega_t(b). \tag{4}$$

Then $\overline{\omega}_t$ is the probability that two independent runs of the t-th biometric source result in two equal symbols, and $H(\omega_t)$ is the entropy of the probability distribution ω_t, which can be understood as the number of random bits at the output of the t-th biometric source.

We will use the component–wise transformation of the vector $\mathbf{b}$ to another vector $\mathbf{z}$ and organize it in such a way that the probability distribution over the vectors $\mathbf{z}$ is close to a uniform distribution. Introduce the following notation. Let us fix $q_t \leq K_t$ as an integer power of 2 and let $\mathcal{Z}_t = \{0, \ldots, q_t - 1\}$. Let us map $b \in \mathcal{B}_t$ to $z \in \mathcal{Z}_t$ if and only if $b \in \mathcal{B}_{t,z}$, where $\mathcal{B}_{t,0}, \ldots, \mathcal{B}_{t,q_t-1}$ are pairwise disjoint sets whose union coincides with $\mathcal{B}_t$. One can see that such a specification uniquely determines z and we denote it by $z(b|q_t)$. Let

$$\mathbf{z}_{\mathbf{b}} = (z(b_1|q_1), \ldots, z(b_n|q_n)) \tag{5}$$

denote the result of the mapping $\mathcal{B} \to \mathcal{Z} = \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_n$, which is parameterized by the vector $\mathbf{q} = (q_1, \ldots, q_n)$ and the partitionings of the sets $\mathcal{B}_1, \ldots, \mathcal{B}_n$. We also denote

$$\Omega_t(z) = \sum_{b \in \mathcal{B}_{t,z}} \omega_t(b)$$

for all $z \in \mathcal{Z}_t$ and

$$\Omega(\mathbf{z}) = \prod_{t=1}^{n} \Omega_t(z_t)$$

for all $\mathbf{z} \in \mathcal{Z}$. Furthermore, let

$$\rho_t = \frac{\max_{z \in \mathcal{Z}_t} \Omega(z)}{\min_{z \in \mathcal{Z}_t} \Omega(z)}. \tag{6}$$

Let the noisy observations of the biometric vector $\mathbf{b}$ be specified by the conditional probability distributions

$$\Big(V(\mathbf{b}'|\mathbf{b}) = \Pr_{\text{err}}\{ B' = \mathbf{b}' \mid B = \mathbf{b} \},\ \mathbf{b}' \in \mathcal{B} \Big),\ \mathbf{b} \in \mathcal{B},$$

and let

$$V(\mathbf{b}'|\mathbf{b}) = \prod_{t=1}^{n} V_t(b'_t|b_t) \tag{7}$$

for all $\mathbf{b}, \mathbf{b}' \in \mathcal{B}$. We also write

$$V_t(b'|b) = \Pr_{\text{err}}\{ B'_t = b' \mid B_t = b \}$$

for all $b, b' \in \mathcal{B}_t$ and pay special attention to the conditional probability distributions such that

$$V_t(b|b) = 1 - \varepsilon, \text{ for all } b \in \mathcal{B}_t, \tag{8}$$

where $\varepsilon > 0$ is a given constant.
The transformation $\mathcal{B} \to \mathcal{Z}$ preserves the V channel in a sense that (8) implies

$$V_t(z_b|b) = \sum_{b' \in \mathcal{B}_{t,z_b}} V_t(b'|b) \geq V_t(b|b) = 1 - \varepsilon$$

for all $b \in \mathcal{B}_t$. Therefore, the V_t channel $\mathcal{B}_t \to \mathcal{B}_t$ is transformed to another V_{t,q_t} channel $\mathcal{Z}_t \to \mathcal{Z}_t$ such that

$$V_{t,q_t}(z|z) \geq 1 - \varepsilon, \text{ for all } z \in \mathcal{Z}_t. \tag{9}$$

Let

$$\text{Ham}(\mathbf{b}, \mathbf{b}') = \Big| \Big\{ t \in \{1, \ldots, n\} :\ b_t \neq b'_t \Big\} \Big|$$

denote the Hamming distance between the vectors $\mathbf{b}, \mathbf{b}' \in \mathcal{B}$ and let

$$\mathcal{D}_T(\mathbf{b}) = \Big\{ \mathbf{b}' \in \mathcal{B} :\ \text{Ham}(\mathbf{b}, \mathbf{b}') \leq T \Big\} \tag{10}$$

denote the set of biometric vectors located at distance T or less from the vector $\mathbf{b}$. The conditional probability of generating a vector belonging to the set $\mathcal{D}_T(\mathbf{b})$, given the vector

b, is defined as

$$V(\mathcal{D}_T(\mathbf{b})|\mathbf{b}) = \sum_{\mathbf{b}' \in \mathcal{D}_T(\mathbf{b})} V(\mathbf{b}'|\mathbf{b}). \tag{11}$$

Notice that if conditions (8) are satisfied, then

$$V(\mathcal{D}_T(\mathbf{b})|\mathbf{b}) = \sum_{d=0}^{T} \binom{n}{d} (1-\varepsilon)^{n-d} \varepsilon^d \tag{12}$$

for all $\mathbf{b} \in \mathcal{B}$.

3. Mathematical model for the DNA measurements

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, i.e., 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, e.g. genotype (10,10) is measured as (10,12); *(2) allelic drop–out*, when an allele of a heterozygous genotype is missing, e.g. genotype (7,9) is measured as (7,7); *(3) allelic shift*, when an allele is measured with a wrong repeat number, e.g. genotype (10,12) is measured as (10,13).

The points above can be formalized as follows. Suppose that there are n sources. For all $t = 1, \ldots, n$, there is a probability distribution

$$\pi_t = \Big(\pi_t(i),\, i \in \{c_t, \ldots, c_t + k_t - 1\} \Big),$$

where c_t, k_t are given positive integers. Let the probability that the t-th source generates the pair (i, j), where $i, j \in \{c_t, \ldots, c_t + k_t - 1\}$, be defined as

$$\Pr_{\mathrm{DNA}} \Big\{ (A_{t,1}, A_{t,2}) = (i, j) \Big\} = \pi_t(i)\pi_t(j).$$

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 t-th motif in the maternal and the paternal allele. We also assume that $(A_{1,1}, A_{1,2}), \ldots, (A_{n,1}, A_{n,2})$ are independent pairs of random variables, i.e.,

$$\Pr_{\mathrm{DNA}} \Big\{ (A_1, A_2) = (\mathbf{i}, \mathbf{j}) \Big\} = \prod_{t=1}^{n} \Pr_{\mathrm{DNA}} \Big\{ (A_{t,1}, A_{t,2}) = (i_t, j_t) \Big\},$$

where $A_1 = (A_{1,1}, \ldots, A_{n,1})$, $A_2 = (A_{1,2}, \ldots, A_{n,2})$ and $\mathbf{i} = (i_1, \ldots, i_n)$, $\mathbf{j} = (j_1, \ldots, j_n)$.

Let

$$S_t = \Big(\min\{A_{t,1}, A_{t,2}\}, \max\{A_{t,1}, A_{t,2}\} \Big).$$

Then

$$\Pr_{\text{DNA}} \Big\{ S_t = (i,j) \Big\} = \tilde{\pi}_t(i,j),$$

where

$$\tilde{\pi}_t(i,j) = \begin{cases} \pi_t^2(i), & \text{if } j = i, \\ 2\pi_t(i)\pi_t(j), & \text{if } j > i, \\ 0, & \text{if } j < i. \end{cases}$$

Denote $\mathcal{B}_t = \{0, \dots, K_t - 1\}$, where $K_t = k_t(k_t+1)/2$, order K_t probabilities belonging to the distribution

$$\tilde{\pi}_t = \Big(\tilde{\pi}_t(i,j), i, j \in \{c_t, \dots, c_t + k_t - 1\}, j \geq i \Big)$$

in the decreasing order, assign them indices $b = 0, \dots, K_t - 1$, and replace $\tilde{\pi}_t$ with the probability distribution

$$\omega_t = \Big(\omega_t(b), b \in \{0, \dots, K_t - 1\} \Big),$$

i.e., the probability distributions $\tilde{\pi}_t$ and ω_t contain the same entries in different order.
The transformations below are illustrated for the **TH01** allele (see Tables 2, 3), where $t = 12$, $c_t = 6$, $k_t = 4$, and

$$(\pi_t(6), \dots, \pi_t(9)) = (.234, .192, .085, .487).$$

Then

$$\Big[\pi_t(i)\pi_t(j) \Big]_{i,j=6,\dots,9} =$$

	$j=6$	$j=7$	$j=8$	$j=9$
$i=6$	.0550	.0452	.0200	.1143
$i=7$	.0452	.0371	.0165	.0939
$i=8$	.0200	.0165	.0073	.0416
$i=9$	.1143	.0939	.0416	.2376

To compute the entries of the probability distribution $\tilde{\pi}_t$, 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 the zeroes.

$$\Big[\tilde{\pi}_t(i,j) \Big]_{\substack{i,j=6,\dots,9 \\ j \geq i}} =$$

	$j=6$	$j=7$	$j=8$	$j=9$
$i=6$	.0550	.0903	.0401	.2286
$i=7$		.0371	.0329	.1878
$i=8$			.0073	.0833
$i=9$				.2376

The ordering of the non-zero entries of this matrix brings the probability distribution ω_t. Its entries and parameters ω_t^*, $\overline{\omega}_t$, defined in (2), (3), are given below.

i,j	9,9	6,9	7,9	6,7	8,9	6,6	6,8	7,7	7,8	8,8
$\tilde{\pi}_t(i,j)$	.2376	.2286	.1878	.0903	.0833	.0550	.0401	.0371	.0329	.0073
b	0	1	2	3	4	5	6	7	8	9
$\omega_t(b)$	.2376	.2286	.1878	.0903	.0833	.0550	.0401	.0371	.0329	.0073
ω_t^*	.2376									
$\overline{\omega}_t$	.2376 .2376 + … + .0073 .0073 = .0609									

b	0	9	1	8	2	5	3	4	6	7
$\omega_t(b)$	.2376	.0073	.2286	.0329	.1878	.0550	.0903	.0833	.0401	.0371
z	0		1		2		3			
$\Omega_t(z)$	.2449		.2615		.2428		.2508			
$\rho_t(z)$	.2615/.2428 = 1.08									

Table 1. Example of the mapping $\{0,\dots,9\} \to \{0,\dots,3\}$.

Let q_t be the maximum integer power of 2 such that

$$1/q_t \geq \omega_t^*,$$

where ω_t^* is defined in (2). Then one can partition the set $\mathcal{B}_t$ in q_t subsets in such a way the resulting probability distribution over these subsets is close to a uniform distribution. An example of the partitioning is given in Table 1. Notice that the entropy of the distribution ω_t is equal to 2.851 (see Table 3), while the entropy of the distribution Ω_t is less and it is close to $\log q_t$.

The available experimental data consist of probability distributions $\pi_1,\dots,\pi_{28}$, and they are given in Table 2. The computed parameters are shown in Table 3. We conclude that results of the DNA measurements can be represented by a binary vector of length 140 bits. However the probability distribution over these vectors is non–uniform and, roughly speaking, only 109 bits carry information about the measurements. The most likely vector of pairs has the probability $0.124\dots0.243 = 10^{-23}$, and the probability that the sources independently generate two equal vectors is equal to $0.013\dots0.046 = 10^{-50}$. The greedy algorithm for partitioning the sets $\mathcal{B}_1,\dots,\mathcal{B}_n$ in $q_1,\dots,q_n$ brings the vectors that can be expressed by $\log q_1 + \dots + \log q_n = 68$ bits with the property that $\rho_1\dots\rho_n \approx 16$, where $\rho_1,\dots,\rho_n$ are defined in (6). Therefore, the most likely vector of length 68 bits has the probability 2^{-64}. Notice that the spectrum of components of the vector $\mathbf{q}$ can be presented the as the sequence $(q \times N_q)$, $q = 2^1,\dots,2^6$, where N_q is the number of indices t with $q_t = q$. Namely, the constructed vector $\mathbf{q}$ has the spectrum

$$(2\times 7),(4\times 8),(8\times 9),(16\times 3),(32\times 0),(64\times 1) \tag{13}$$

and

$$\begin{aligned} 28 &= 7+8+9+3+0+1,\\ 68 &= 7\cdot\log 2 + 8\cdot\log 4 + 9\cdot\log 8 + 3\cdot\log 16 + 0\cdot\log 32 + 1\cdot\log 64. \end{aligned}$$

4. Direct authentication schemes

Let us consider the following setup. Suppose that $\mathbf{b},\mathbf{b}' \in \mathcal{B}$ are given vectors of length n. If the Hamming distance between these vectors is not greater than a fixed threshold T, then the verifier has to make the acceptance decision. Otherwise, the verifier has to make the rejection decision. Hence, the rules are as follows:

$\mathrm{R}_{\mathrm{Acc}}$: *if* $\mathbf{b}' \in \mathcal{D}_T(\mathbf{b})$, *then accept the identity claim* (Acc);

$\mathrm{R}_{\mathrm{Rej}}$: *if* $\mathbf{b}' \notin \mathcal{D}_T(\mathbf{b})$, *then reject the identity claim* (Rej).

t	Name	π_t
1	**D8S1179**	.319 .194 .173 .119 .105 .086
2	**D3S1358**	.265 .257 .218 .154 .104
3	**VWA**	.283 .202 .202 .111 .105 .095
4	**D7S820**	.248 .211 .180 .168 .155 .035
5	**ACTBP2**	.089 .080 .073 .072 .070 .064 .062 .053 .051 .049 .047 .046 .043 .039 .037 .034 .033 .028 .012 .009
6	**D7S820**	.243 .207 .177 .165 .152 .034 .018
7	**FGA**	.223 .192 .139 .139 .129 .072 .053 .026 .023
8	**D21S11**	.308 .200 .183 .160 .091 .028 .026
9	**D18S51**	.162 .142 .142 .135 .130 .129 .078 .039 .022 .016
10	**D19S433**	.382 .259 .173 .086 .082 .015
11	**D13S317**	.339 .248 .124 .112 .074 .051 .048
12	**TH01**	.487 .234 .192 .085
13	**D2S138**	.182 .146 .122 .117 .114 .093 .079 .041 .038 .033 .029
14	**D16S539**	.326 .321 .145 .112 .056 .019 .018
15	**D5S818**	.389 .365 .142 .052 .050
16	**TPOX**	.537 .244 .119 .056 .041
17	**CF1PO**	.365 .305 .219 .097 .011
18	**D8S1179**	.304 .185 .165 .114 .100 .082 .031 .011 .003
19	**VWA1**	.283 .202 .202 .111 .105 .095
20	**PentaD**	.265 .214 .189 .156 .089 .060 .014 .010
21	**PentaE**	.180 .170 .110 .105 .102 .080 .056 .051 .051 .034 .029 .010 .010 .007
22	**DYS390**	.422 .282 .164 .103 .014 .011
23	**DYS429**	.445 .325 .118 .096 .013
24	**DYS437**	.528 .317 .154
25	**DYS391**	.513 .451 .018 .016
26	**DYS385**	.551 .124 .097 .087 .059 .037 .030 .012
27	**DYS389I**	.663 .186 .150
28	**DYS389II**	.446 .272 .167 .081 .032

Table 2. The entries of the probability distributions $\pi_1, \ldots, \pi_{28}$, which are greater than 0.001, given in the decreasing order.

"The identity claim" in the description above appears because we assume that the vectors $\mathbf{b}$ and $\mathbf{b}'$ contain outcomes of measurements of some biometric parameters of two people. The verification is understood as a procedure, which checks whether the difference between the results is caused by the observation noise or by the fact that people are different.
The direct implementation of the authentication procedure includes the enrollment and the verification stages (see Figure 1).

The enrollment stage.

– *Store the biometric vector* $\mathbf{b}$ *in the database.*

t	Name	$\log K_t$	$\lceil \log K_t \rceil$	ω_t^*	$\log q_t$	$H(\omega_t)$	$\overline{\omega}_t$
1	**D8S1179**	4.392	5	0.124	3	4.083	0.013
2	**D3S1358**	3.907	4	0.137	2	3.714	0.012
3	**VWA**	4.392	5	0.115	3	4.127	0.010
4	**D7S820**	4.392	5	0.105	3	4.074	0.008
5	**ACTBP2**	7.714	8	0.014	6	7.426	0.000
6	**D7S820**	4.807	5	0.101	3	4.241	0.008
7	**FGA**	5.492	6	0.086	3	4.916	0.005
8	**D21S11**	4.807	5	0.124	3	4.130	0.013
9	**D18S51**	5.781	6	0.046	4	5.279	0.002
10	**D19S433**	4.392	5	0.199	2	3.593	0.027
11	**D13S317**	4.807	5	0.169	2	4.151	0.018
12	**TH01**	3.322	4	0.238	2	2.851	0.061
13	**D2S138**	6.044	7	0.053	4	5.601	0.002
14	**D16S539**	4.807	5	0.210	2	3.776	0.023
15	**D5S818**	3.907	4	0.285	1	3.111	0.041
16	**TPOX**	3.907	4	0.289	1	2.909	0.087
17	**CF1PO**	3.907	4	0.223	2	3.157	0.029
18	**D8S1179**	5.492	6	0.113	3	4.487	0.011
19	**VWA1**	4.392	5	0.115	3	4.127	0.010
20	**PentaD**	5.170	6	0.114	3	4.325	0.009
21	**PentaE**	6.907	7	0.062	4	5.870	0.002
22	**DYS390**	4.392	5	0.239	2	3.238	0.039
23	**DYS429**	3.907	4	0.290	1	2.972	0.051
24	**DYS437**	2.585	3	0.335	1	2.259	0.089
25	**DYS391**	3.322	4	0.464	1	1.902	0.111
26	**DYS385**	5.170	6	0.304	1	3.607	0.093
27	**DYS389I**	2.585	3	0.440	1	2.008	0.195
28	**DYS389II**	3.907	4	0.243	2	3.145	0.046
		128.6	140	10^{-23}	68	109.1	10^{-50}

Table 3. Some characteristics of the probability distributions $\omega_1, \ldots, \omega_{28}$ that describe the DNA measurements.

The verification stage.

– *Read the biometric vector* $\mathbf{b}$ *associated with the claimed person from the database. If* $\mathbf{b}' \in \mathcal{D}_T(\mathbf{b})$, *then make the acceptance decision* (Acc). *If* $\mathbf{b}' \notin \mathcal{D}_T(\mathbf{b})$, *then make the rejection decision* (Rej).

The basic parameters of the scheme are the false rejection rate FRR, the false acceptance rate FAR, and the average false acceptance rate $\overline{\text{FAR}}$, introduced as

$$\text{FRR} = \sum_{\mathbf{b},\mathbf{b}' \in \mathcal{B}} \omega(\mathbf{b}) V(\mathbf{b}'|\mathbf{b}) \chi\{\mathbf{b}' \notin \mathcal{D}_T(\mathbf{b})\}, \tag{14}$$

$$\text{FAR} = \max_{\mathbf{b}' \in \mathcal{B}} \sum_{\mathbf{b} \in \mathcal{B}} \omega(\mathbf{b}) \chi\{\mathbf{b}' \in \mathcal{D}_T(\mathbf{b})\}, \tag{15}$$

$$\overline{\text{FAR}} = \sum_{\mathbf{b},\mathbf{b}' \in \mathcal{B}} \omega(\mathbf{b}) \omega(\mathbf{b}') \chi\{\mathbf{b}' \in \mathcal{D}_T(\mathbf{b})\}, \tag{16}$$

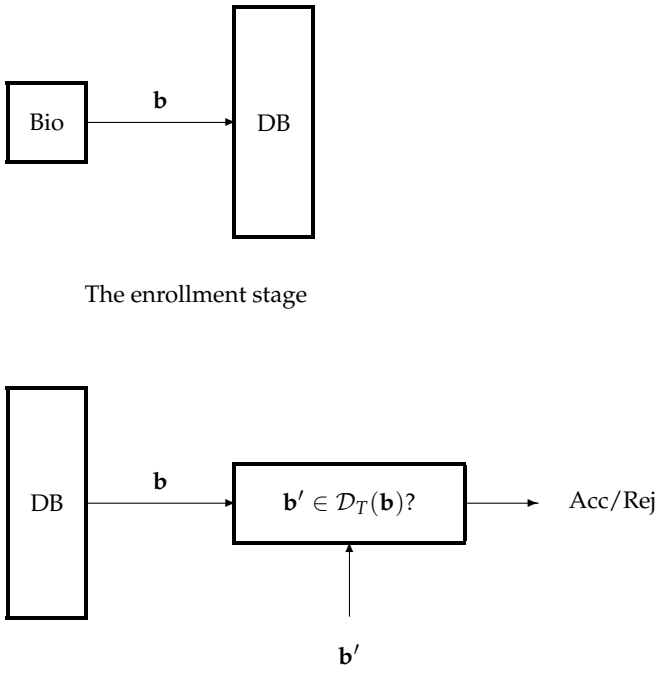


Fig. 1. The data processing in a direct authentication scheme.

where χ denotes the indicator function: $\chi\{\mathcal{S}\} = 1$ is the statement $\mathcal{S}$ is true and $\chi\{\mathcal{S}\} = 0$ otherwise. The false rejection rate is the probability of the event that the verifier makes the rejection decision when the observations belong to the same person. The false acceptance rate is the probability of the event that the verifier makes the acceptance decision when the vector $\mathbf{b}'$ is generated by an attacker. The average false acceptance rate is the probability of the event that the verifier makes the acceptance decision when the vector $\mathbf{b}'$ contains outcomes of biometric observations of a randomly chosen person.
If the V channel satisfies (8), then the false rejection rate is expressed using (12),

$$\mathrm{FRR} = \sum_{d=T+1}^{n} \binom{n}{d} (1-\varepsilon)^{n-d} \varepsilon^{d}. \tag{17}$$

To compute the false acceptance rates, we use the generating functions technique.
Let us consider the problem of computing $\overline{\mathrm{FAR}}$ and introduce the generating function

$$\overline{G}_t(z) = \overline{\omega}_t + (1-\overline{\omega}_t)z,$$

where z is a formal variable and $\overline{\omega}_t$ is defined in (3) as the probability that two independent runs of the t-th source result in two equal symbols. Furthermore, denote

$$\overline{G}(z) = \prod_{t=1}^{n} \overline{G}_t(z)$$

and represent the polynomial $\overline{G}(z)$ as

$$\overline{G}(z) = \sum_{d=0}^{n} \mathrm{Coef}_d\Big[\,\overline{G}(z)\,\Big] z^d.$$

Then the d-th term of the sum at the right-hand side is equal to the probability that two independent runs of n sources result in vectors that differ in d components. Hence,

$$\overline{\mathrm{FAR}} = \sum_{d=0}^{T} \mathrm{Coef}_d\Big[\,\overline{G}(z)\,\Big].$$

Similar manipulations bring the formula

$$\sum_{\mathbf{b}\in\mathcal{B}} \omega(\mathbf{b})\chi\{\mathbf{b}' \in \mathcal{D}_T(\mathbf{b})\} = \sum_{d=0}^{T} \mathrm{Coef}_d\Big[\,G(z|\mathbf{b}')\,\Big], \tag{18}$$

where

$$G(z|\mathbf{b}') = \prod_{t=1}^{n}\Big(\omega_t(b'_t) + (1-\omega_t(b'_t))z\Big).$$

One can easily see that the sum at the right-hand side of (18) is maximized when $\mathbf{b}' = \mathbf{b}^*$ and

$$\mathrm{FAR} = \sum_{d=0}^{T} \mathrm{Coef}_d\Big[\,G(z|\mathbf{b}^*)\,\Big],$$

where

$$G(z|\mathbf{b}^*) = \prod_{t=1}^{n}\Big(\omega_t^* + (1-\omega_t^*)z\Big)$$

and $\omega_1^*, \ldots, \omega_n^*$ are defined in (2).

Some numerical results for the DNA data are given in Table 4. We conclude that the probability of successful attack in the case when the attacker does not know the content of the database can be very small. However, the main problem with the direct authentication scheme is caused by the point that the biometric vector itself is stored in the database. If an attacker would have an access to the database, then he does not have any difficulties with the passing through the verification stage with the acceptance decision. Moreover, the biometrics, being compromized, is compromized forever and it can be also used for any other purposes. A possible solution to the hiding problem is the use of the cryptographic "one–way" hash function Hash : it is assumed that the value of the function can be easily computed for a given argument, but the value of the argument is hard to get for a given value of the function. If only $\mathrm{Hash}(\mathbf{b})$ is known to the verifier, then he can compute the values of $\mathrm{Hash}(\tilde{\mathbf{b}})$ for all vectors $\tilde{\mathbf{b}}$ located at the Hamming distance at most T from the vector $\mathbf{b}'$ and make the acceptance

	FRR				
T	$\varepsilon = 0.05$	$\varepsilon = 0.01$	FAR	$\overline{\text{FAR}}$	$\hat{\text{FAR}}$
0	$7.6 \cdot 10^{-1}$	$2.5 \cdot 10^{-1}$	$7.7 \cdot 10^{-24}$	$2.5 \cdot 10^{-50}$	$3.4 \cdot 10^{-21}$
1	$4.1 \cdot 10^{-1}$	$3.2 \cdot 10^{-2}$	$1.9 \cdot 10^{-21}$	$1.7 \cdot 10^{-46}$	$6.9 \cdot 10^{-19}$
2	$1.6 \cdot 10^{-1}$	$2.7 \cdot 10^{-3}$	$2.0 \cdot 10^{-19}$	$3.4 \cdot 10^{-43}$	$6.1 \cdot 10^{-17}$
3	$4.9 \cdot 10^{-2}$	$1.7 \cdot 10^{-4}$	$1.3 \cdot 10^{-17}$	$3.7 \cdot 10^{-40}$	$3.2 \cdot 10^{-15}$
4	$1.2 \cdot 10^{-2}$	$8.1 \cdot 10^{-6}$	$5.8 \cdot 10^{-16}$	$2.5 \cdot 10^{-37}$	$1.2 \cdot 10^{-13}$
5	$2.3 \cdot 10^{-3}$	$3.1 \cdot 10^{-7}$	$1.9 \cdot 10^{-14}$	$1.2 \cdot 10^{-34}$	$3.1 \cdot 10^{-12}$
6	$3.6 \cdot 10^{-4}$	$9.8 \cdot 10^{-9}$	$4.8 \cdot 10^{-13}$	$4.2 \cdot 10^{-32}$	$6.4 \cdot 10^{-11}$
7	$4.9 \cdot 10^{-5}$	$2.6 \cdot 10^{-10}$	$9.7 \cdot 10^{-12}$	$1.1 \cdot 10^{-29}$	$1.0 \cdot 10^{-9}$
8	$5.6 \cdot 10^{-6}$	$5.8 \cdot 10^{-12}$	$1.6 \cdot 10^{-10}$	$2.3 \cdot 10^{-27}$	$1.3 \cdot 10^{-8}$
9	$5.6 \cdot 10^{-7}$	$1.1 \cdot 10^{-13}$	$2.1 \cdot 10^{-9}$	$3.7 \cdot 10^{-25}$	$1.4 \cdot 10^{-7}$

Table 4. The false rejection and the false acceptance rates for the DNA measurements.

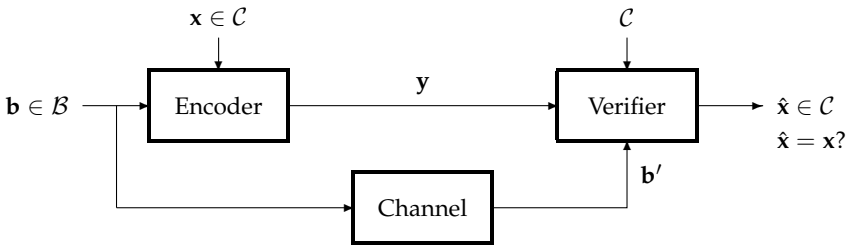


Fig. 2. General authentication scheme.

decision if one of them is equal to Hash($\mathbf{b}$). Such a scheme is secure up to the security of hashing, but requires the hash function to be defined over the set of $|\mathcal{B}|$ vectors and very large computational complexity. The block coding schemes can be viewed as solutions introduced to relax these requirements.

5. Block coding approach to the authentication problem

The coding problem for biometric verification can be presented as designing codes for the scheme in Figure 2. Let $\mathcal{C} \subset \mathcal{B}$ be a subset whose entries are codewords assigned by the designer. The encoding is the transformation of a pair $(\mathbf{x}, \mathbf{b}) \in \mathcal{C} \times \mathcal{B}$, where the vector $\mathbf{b}$ is generated by the source and $\mathbf{x}$ is chosen according to a uniform probability distribution over the code $\mathcal{C}$, to another vector $\mathbf{y} = (y_1, \ldots, y_n)$ belonging to some finite set $\mathcal{Y}$. The mappings

$$(\mathbf{x}, \mathbf{b}) \to \mathbf{y}, \quad (\mathbf{y}, \mathbf{b}') \to \mathbf{x}$$

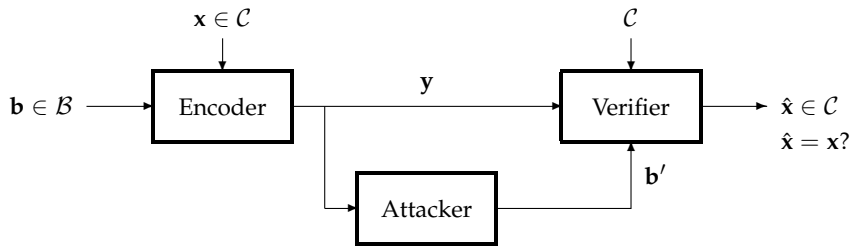


Fig. 3. General authentication scheme from the attacker's prospective.

are called the encoding and the decoding, respectively. The general requirement to the these mappings can be presented as

$$(\mathbf{x}, \mathbf{b}) \to \mathbf{y} \;\Rightarrow\; \begin{cases} \mathbf{b}' \in \mathcal{D}_T(\mathbf{b}) \Rightarrow (\mathbf{y}, \mathbf{b}') \to \mathbf{x}, \\ \mathbf{b}' \notin \mathcal{D}_T(\mathbf{b}) \Rightarrow (\mathbf{y}, \mathbf{b}') \not\to \mathbf{x}. \end{cases} \tag{19}$$

In other words, the results of the decoding for the vectors $\mathbf{b}$ and $\mathbf{b}'$ have to coincide if and only if $\mathbf{b}' \in \mathcal{D}_T(\mathbf{b})$.
Both the vector $\mathbf{y}$ and the value of $\mathrm{Hash}(\mathbf{x})$ are stored in the database under the name of the person whose biometric characteristics are expressed by the vector $\mathbf{b}$. Having received the vector $\mathbf{b}'$ and the name of the person, the decoder reads $(\mathbf{y}, \mathrm{Hash}(\mathbf{x}))$ from the database and uses the error–correcting capabilities of the code to decode "the transmitted codeword" $\mathbf{x}$ as $\hat{\mathbf{x}}$. If $\mathrm{Hash}(\hat{\mathbf{x}}) = \mathrm{Hash}(\mathbf{x})$, then the identity claim is accepted. Otherwise, the claim is rejected.
From the attacker's prospective, the authentication scheme can be viewed as the scheme in Figure 3. The attacker reads the content of the database associated with a person, presents the name of the person, and generates the vector $\mathbf{b}'$. The goal of the attacker is generating of a vector leading to the verifier's acceptance decision. The coding problem can be formulated as constructing codes that simultaneously satisfy the constraint (19) and guarantee a low probability of the attacker's success.

6. Additive block coding schemes

Given a positive integer q, let $\oplus_q$ and $\ominus_q$ denote the addition and the subtraction modulo q, respectively,

$$z \oplus_q z' = \begin{cases} z + z', & \text{if } z + z' \le q, \\ z + z' - q, & \text{if } z + z' > q \end{cases}$$

$$z \ominus_q z' = \begin{cases} z - z', & \text{if } z + z' \ge 0, \\ z - z' + q, & \text{if } z + z' < 0. \end{cases}$$

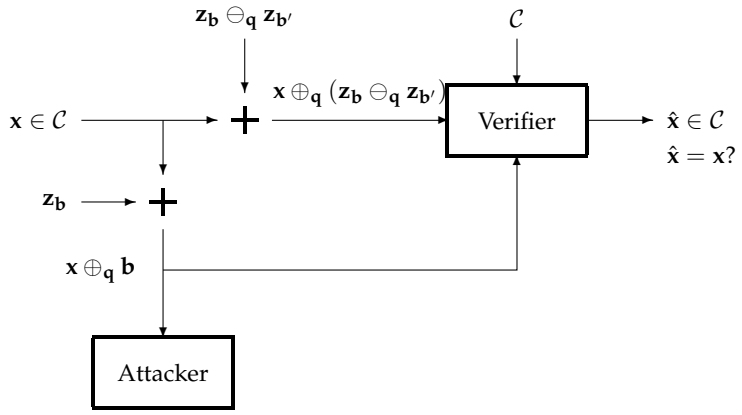


Fig. 4. Wiretap-type additive block coding scheme.

The operations $\oplus_{\mathbf{q}}$ and $\ominus_{\mathbf{q}}$, where $\mathbf{q} = (q_1, \ldots, q_n)$, being applied to the vectors of length n, are understood as component–wise addition and subtraction modulo $q_1, \ldots, q_n$, i.e.,

$$\mathbf{z} \oplus_{\mathbf{q}} \mathbf{z}' = (z_1 \oplus_{q_1} z'_1, \ldots, z_n \oplus_{q_n} z'_n),$$
$$\mathbf{z} \ominus_{\mathbf{q}} \mathbf{z}' = (z_1 \ominus_{q_1} z'_1, \ldots, z_n \ominus_{q_n} z'_n).$$

Let us consider the biometric vector $\mathbf{b}$ as an additive noise that corrupts the transmitted codeword $\mathbf{x}$ and the received vector is defined as

$$\mathbf{y} = \mathbf{x} \oplus_{\mathbf{q}} \mathbf{z_b},$$

where $\mathbf{z_b}$ is the result of the transformation of the biometric vector $\mathbf{b}$ defined in (5). The decoding is based on the observation:

$$\left.\begin{array}{l} \mathbf{y} = \mathbf{x} \oplus_{\mathbf{q}} \mathbf{z_b} \\ \mathrm{Ham}(\mathbf{z_b}, \mathbf{z_{b'}}) \leq T \end{array}\right\} \Rightarrow \mathrm{Ham}(\mathbf{y}, \mathbf{x} \oplus_{\mathbf{q}} \mathbf{z_{b'}}) \leq T.$$

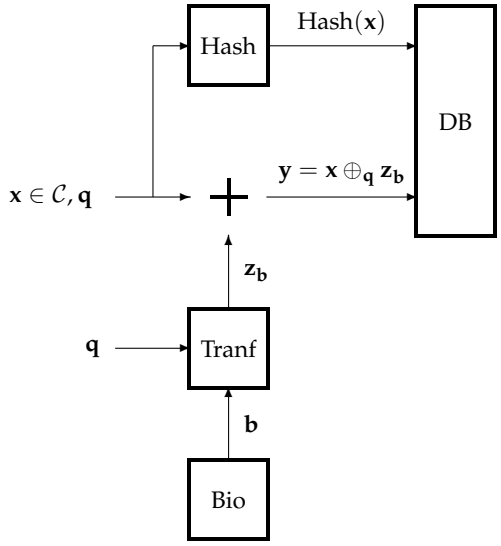
Notice also that

$$\mathbf{y} = \mathbf{x} \oplus_{\mathbf{q}} \mathbf{z_b} \Rightarrow \mathrm{Ham}(\mathbf{y}, \mathbf{x} \oplus_{\mathbf{q}} \mathbf{z_{b'}}) = \mathrm{Ham}(\mathbf{y} \ominus_{\mathbf{q}} \mathbf{z_{b'}}, \mathbf{x}) = \mathrm{Ham}(\mathbf{x} \oplus_{\mathbf{q}} (\mathbf{z_b} \ominus_{\mathbf{q}} \mathbf{z_{b'}}), \mathbf{x}).$$

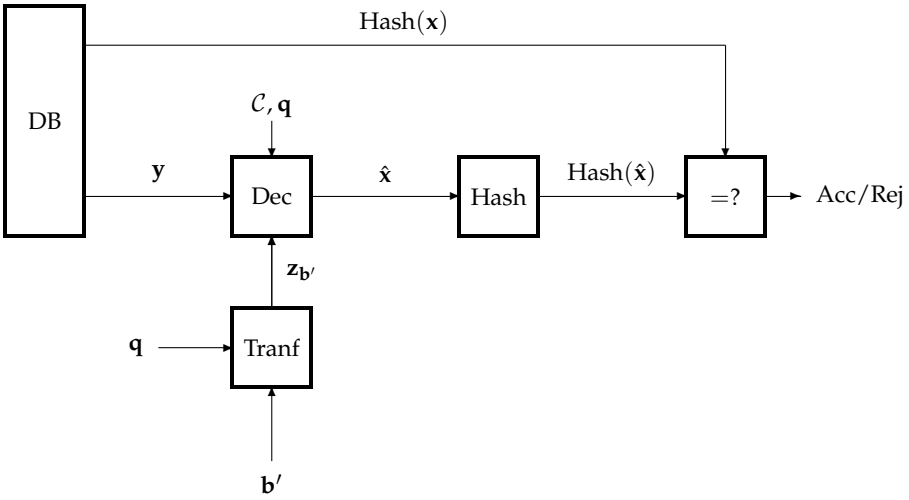
Thus, the verifier analyzes the outcomes of transmission of the codeword $\mathbf{x}$ over two parallel channels,

$$\mathbf{x} \to \mathbf{x} \oplus_{\mathbf{q}} (\mathbf{z_b} \ominus_{\mathbf{q}} \mathbf{z_{b'}}) \quad \text{(the observation channel)},$$
$$\mathbf{x} \to \mathbf{x} \oplus_{\mathbf{q}} \mathbf{z_b} \quad \text{(the biometric channel)},$$

while the attacker analyzes only the output of the biometric channel (see Figure 4).



The enrollment stage



The verification stage

Fig. 5. The data processing in an additive block coding scheme.

Processing of a given biometric vector $\mathbf{b}$ at the enrollment stage and processing of data at the verification stage when the verifier considers only the output of the observation channel is illustrated in Figure 5.

The enrollment stage.

- *Choose a key codeword* $\mathbf{x}$ *according to a uniform probability distribution over the code* $\mathcal{C}$ *and compute the value of* $\mathrm{Hash}(\mathbf{x})$.
- *Store* $(\mathrm{Hash}(\mathbf{x}), \mathbf{x} \oplus_{\mathbf{q}} \mathbf{z}_{\mathbf{b}})$ *in the database.*

The verification stage.

- *Read the data* $(\mathrm{Hash}(\mathbf{x}), \mathbf{y})$ *associated with the claimed person from the database.*
- *Decode the key codeword, given a received vector* $\mathbf{z} = \mathbf{y} \ominus_{\mathbf{q}} \mathbf{z}_{\mathbf{b}'}$, *as* $\hat{\mathbf{x}}$. *If* $\mathrm{Hash}(\hat{\mathbf{x}}) = \mathrm{Hash}(\mathbf{x})$, *then make the acceptance decision* (Acc). *If* $\mathrm{Hash}(\hat{\mathbf{x}}) \neq \mathrm{Hash}(\mathbf{x})$, *then make the rejection decision* (Rej).

Let us illustrate the additive block coding and the decoding algorithms that will be described in a general form by the numerical example. Let $q_1 = \cdots = q_6 = 2$, $n = 6$, and let $\mathcal{C}$ be a binary code consisting of 8 codewords,

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$	$\mathbf{x}_7$	$\mathbf{x}_8$
000000	001011	010101	011110	100110	101101	110011	111000

For example,

$$\left.\begin{array}{r} \mathbf{z}_{\mathbf{b}} = 011011 \\ \mathbf{x} = 011110 \end{array}\right\} \rightarrow \mathbf{y} = 000101,$$

and the vector $\mathbf{y}$ is stored in the database. Having received another vector $\mathbf{z}_{\mathbf{b}'}$, the verifier tries to find a codeword $\hat{\mathbf{x}}$ located at distance at most 1 from the vector $\mathbf{y} \ominus_{\mathbf{q}} \mathbf{z}_{\mathbf{b}'}$. For example,

$$\left.\begin{array}{r} \mathbf{z}_{\mathbf{b}'} = 111011 \\ \mathbf{y} = 000101 \end{array}\right\} \rightarrow \mathbf{y} \ominus_{\mathbf{q}} \mathbf{z}_{\mathbf{b}'} = 111110 \rightarrow \hat{\mathbf{x}} = 011110,$$

and the verifier makes the acceptance decision, since $\hat{\mathbf{x}} = \mathbf{x}$ implies $\mathrm{Hash}(\hat{\mathbf{x}}) = \mathrm{Hash}(\mathbf{x})$. An attacker wants to submit some vector $\mathbf{b}'$, which also leads to the acceptance. He constructs the list of candidate vectors as $\mathbf{y} \ominus_{\mathbf{q}} \mathbf{x}$, $\mathbf{x} \in \mathcal{C}$, and finds the vector $\hat{\mathbf{x}}$ such that $\Omega(\mathbf{y} \ominus_{\mathbf{q}} \mathbf{x})$ is the maximum. For example,

$\mathbf{y} \ominus_{\mathbf{q}} \mathbf{x}_1$	$\mathbf{y} \ominus_{\mathbf{q}} \mathbf{x}_2$	$\mathbf{y} \ominus_{\mathbf{q}} \mathbf{x}_3$	$\mathbf{y} \ominus_{\mathbf{q}} \mathbf{x}_4$	$\mathbf{y} \ominus_{\mathbf{q}} \mathbf{x}_5$	$\mathbf{y} \ominus_{\mathbf{q}} \mathbf{x}_6$	$\mathbf{y} \ominus_{\mathbf{q}} \mathbf{x}_7$	$\mathbf{y} \ominus_{\mathbf{q}} \mathbf{x}_8$
000101	001110	**010000**	011011	100011	101000	110110	111101

In particular, if the probabilities $\Omega(\mathbf{z})$ decrease when the weight of the vector $\mathbf{z}$ increases, then this algorithm brings the vector $\hat{\mathbf{x}} = \mathbf{x}_3$, and the attacker's vector $\mathbf{b}'$ is such that $\mathbf{z}_{\mathbf{b}'} = \mathbf{z}_{\mathbf{y} \ominus_{\mathbf{q}} \mathbf{x}_3}$. Suppose that $\mathcal{C}$ is a block code consisting of M codewords $\mathbf{x}_1, \ldots, \mathbf{x}_M \in \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_n$ and having the minimum distance greater than $2T$, i.e.,

$$\left.\begin{array}{r} \mathbf{x}, \mathbf{x}' \in \mathcal{C} \\ \mathbf{x} \neq \mathbf{x}' \end{array}\right\} \Rightarrow \mathrm{Ham}(\mathbf{x}, \mathbf{x}') \geq 2T + 1. \tag{20}$$

Then the Hamming balls of radius T centered at codewords, $\mathcal{D}_T(\mathbf{x})$, $\mathbf{x} \in \mathcal{C}$, are pairwise disjoint sets. As a result, for any $\mathbf{y}, \mathbf{z}_{\mathbf{b}'} \in \mathcal{Z}$, there is at most one codeword $\mathbf{x} \in \mathcal{C}$ such

that

$$\mathrm{Ham}(\mathbf{y}, \mathbf{x} \oplus_{\mathbf{q}} \mathbf{z}_{\mathbf{b}'}) \le T. \tag{21}$$

Let us denote this codeword by $\hat{\mathbf{x}}(\mathbf{y}, \mathbf{z}_{\mathbf{b}'})$. If the inequality (21) does not hold for all codewords, we assume that $\hat{\mathbf{x}}(\mathbf{y}, \mathbf{z}_{\mathbf{b}'})$ is a fixed vector (for example, the all–zero vector). Thus,

$$\mathrm{Ham}(\mathbf{z}_{\mathbf{b}}, \mathbf{z}_{\mathbf{b}'}) \le T \;\Rightarrow\; \mathrm{Ham}(\mathbf{x} \oplus_{\mathbf{q}} \mathbf{z}_{\mathbf{b}}, \mathbf{x} \oplus_{\mathbf{q}} \mathbf{z}_{\mathbf{b}'}) \le T \;\Rightarrow\; \hat{\mathbf{x}}(\mathbf{x} \oplus_{\mathbf{q}} \mathbf{z}_{\mathbf{b}}, \mathbf{z}_{\mathbf{b}'}) = \mathbf{x}.$$

Hence, if $\mathbf{x}$ is the codeword, which was used to encode the vector $\mathbf{z}_{\mathbf{b}}$, and the vector $\mathbf{z}_{\mathbf{b}'}$ differs from the vector $\mathbf{z}_{\mathbf{b}}$ in at most T components, then the codeword is decoded. Therefore the false rejection rate is expressed by (14),

$$\mathrm{FRR} = \sum_{\mathbf{b},\mathbf{b}' \in \mathcal{B}} \omega(\mathbf{b}) V(\mathbf{b}'|\mathbf{b}) \chi\{\mathbf{z}_{\mathbf{b}'} \notin \mathcal{D}_T(\mathbf{z}_{\mathbf{b}})\}.$$

The similar conclusion is valid for the false acceptance rate of a randomly chosen person,

$$\overline{\mathrm{FAR}} = \sum_{\mathbf{b},\mathbf{b}'} \omega(\mathbf{b}) \omega(\mathbf{b}') \chi\{\mathrm{Ham}(\mathbf{z}_{\mathbf{b}}, \mathbf{z}_{\mathbf{b}'}) \le T\}.$$

Let us analyze the situation when an attacker is present. He receives only the result of transmission of the codeword over the biometric channel and his action can be presented as the mapping

$$(\mathbf{z}_{\mathbf{b}_1} = \mathbf{y} \ominus_{\mathbf{q}} \mathbf{x}_1, \ldots, \mathbf{z}_{\mathbf{b}_M} = \mathbf{y} \ominus_{\mathbf{q}} \mathbf{x}_M) \;\to\; \mathbf{b}' = \mathbf{b}_{\hat{m}},$$

where $\hat{m} \in \{1, \ldots, M\}$ is chosen in such a way that

$$\Omega(\mathbf{z}_{\mathbf{b}_{\hat{m}}}) = \max_{1 \le m \le M} \Omega(\mathbf{z}_{\mathbf{b}_m}). \tag{22}$$

The submission of the vector $\mathbf{b}_{\hat{m}}$ to the verifier implies $\hat{\mathbf{x}} = \mathbf{x}_{\hat{m}}$, and the acceptance decision is made if and only if $\mathbf{x}_{\hat{m}}$ is the codeword that was used to encode the biometric vector at the enrollment stage. The probability of the attacker's success, given the vectors $\mathbf{z}_{\mathbf{b}_1}, \ldots, \mathbf{z}_{\mathbf{b}_M}$, is equal to

$$\frac{\Omega(\mathbf{z}_{\mathbf{b}_{\hat{m}}})}{\sum_{m=1}^{M} \Omega(\mathbf{z}_{\mathbf{b}_m})} \le \frac{\max_{1 \le m \le M} \Omega(\mathbf{z}_{\mathbf{b}_m})}{M \min_{1 \le m \le M} \Omega(\mathbf{z}_{\mathbf{b}_m})} \le \frac{\max_{\mathbf{z} \in \mathcal{Z}} \Omega(\mathbf{z})}{M \min_{\mathbf{z} \in \mathcal{Z}} \Omega(\mathbf{z})} = \frac{1}{M} \prod_{t=1}^{n} \rho_t, \tag{23}$$

where $\rho_1, \ldots, \rho_n$ are defined in (6). Since the upper bound (23) holds for any received vector $\mathbf{y}$, which determines the vectors $\mathbf{z}_{\mathbf{b}_1}, \ldots, \mathbf{z}_{\mathbf{b}_M}$,

$$\mathrm{FAR} \le \frac{1}{M} \prod_{t=1}^{n} \rho_t. \tag{24}$$

Let us evaluate the bound (24) using the standard covering arguments of coding theory. Given the vector $\mathbf{q}$, introduce the generating function

$$G(z) = \prod_{t=1}^{n} G_t(z),$$

where

$$G_t(z) = \frac{1}{q_t} + \frac{q_t - 1}{q_t} z.$$

For example, for the DNA data (see (13)),

$$G_{\mathrm{DNA}}(z) = \left(\frac{1}{2} + \frac{1}{2}z\right)^7 \left(\frac{1}{4} + \frac{3}{4}z\right)^8 \left(\frac{1}{8} + \frac{7}{8}z\right)^9 \left(\frac{1}{16} + \frac{1}{15}z\right)^3 \left(\frac{1}{64} + \frac{63}{64}z\right)^1.$$

One can easily see that the d-th coefficient of the polynomial $G(z)$ is equal to the ratio of the number of vectors $\mathbf{x}' \in \mathcal{Z}$ located at the Hamming distance d from any fixed vector $\mathbf{x} \in \mathcal{Z}$ and $q_1 \dots q_n$, i.e.,

$$\frac{1}{\prod_{t=1}^{n} q_t} \left| \left\{ \mathbf{x}' \in \mathcal{Z} : \mathrm{Ham}(\mathbf{x}, \mathbf{x}') = d \right\} \right| = \mathrm{Coef}_d[\, G(z)\,].$$

Therefore,

$$\frac{1}{\prod_{t=1}^{n} q_t} |\mathcal{D}_T(\mathbf{x})| = \sum_{d=0}^{T} \mathrm{Coef}_d[\, G(z)\,]. \tag{25}$$

Since $\mathcal{D}_T(\mathbf{x}_1), \dots, \mathcal{D}_T(\mathbf{x}_M)$ are pairwise disjoint sets,

$$\sum_{m=1}^{M} |\mathcal{D}_T(\mathbf{x}_m)| \le \prod_{t=1}^{n} q_t,$$

and (25) implies

$$\frac{1}{M} \ge \sum_{d=0}^{T} \mathrm{Coef}_d[\, G(z)\,]. \tag{26}$$

By assuming that there is a code such that (26) holds with the equality and by replacing the parameters $\rho_1, \dots, \rho_M$ with 1's, we evaluate the false acceptance rate, estimated in (24), as

$$\mathrm{FAR} \approx \hat{\mathrm{FAR}} = \sum_{d=0}^{T} \mathrm{Coef}_d[\, G(z)\,].$$

The values of $\hat{\mathrm{FAR}}$ are given in Table 4 for the DNA data. As a result, one can conclude that the additive coding scheme can give a very efficient solution to the authentication problem provided that there is a class of specific codes having the certain minimum distance and corresponding decoding algorithms that require a low computational complexity.

7. Permutation block coding schemes

The permutation block coding scheme can be viewed as a modification of the scheme in Figure 4 where the sum modulo $\mathbf{q}$ in the link to the attacker is replaced by a stochastic mapping $f(\mathbf{x}, \mathbf{b})$, as it is shown in Figure 6. In this section, we will assume that $q = 2$. In particular, the modification of a wiretap-type block coding scheme is possible when both the vector $\mathbf{x}$ and $\mathbf{b}$ have equal weights and $f(\mathbf{x}, \mathbf{b})$ stands for the binary representation of a permutation $\boldsymbol{\pi}$ that transforms the vector $\mathbf{x}$ to the vector $\mathbf{b}$. Formally, let $\mathcal{B} = \{0,1\}_w^n$, where $\{0,1\}_w^n$ is the set consisting of binary vectors of the Hamming weight w. Thus, the biometric

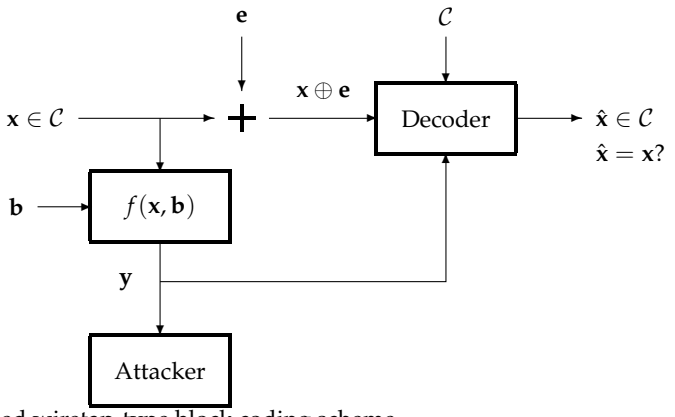


Fig. 6. Modified wiretap-type block coding scheme.

vector is a binary vector $\mathbf{b}$ of length n chosen by a combinatorial (n, w)-source, i.e.,

$$\mathrm{wt}(\mathbf{b}) \neq w \;\Rightarrow\; \Pr_{\mathrm{bio}}\{B = \mathbf{b}\} = 0. \tag{27}$$

Let $\mathcal{C}$ denote a binary code consisting of M different codewords of length n and weight w, i.e., $\mathcal{C} \subseteq \{0,1\}_w^n$ and $|\,\mathcal{C}\,| = M$.

The permutation of components of some vector $\mathbf{x} = (x_1, \ldots, x_n) \in \{0,1\}_w^n$ is determined by a vector $\boldsymbol{\pi} \in \mathcal{P}$ in such a way that $\boldsymbol{\pi}(\mathbf{x}) = (x_{\pi_1}, \ldots, x_{\pi_n})$, where $\mathcal{P}$ is the set of all possible permutations of components of the vector $(1, \ldots, n)$. Given a vector $\mathbf{b} \in \{0,1\}_w^n$ and a permutation $\boldsymbol{\pi} \in \mathcal{P}$, let $\boldsymbol{\pi}^{-1} \in \mathcal{P}$ denote the inverse permutation, i.e., $\boldsymbol{\pi}^{-1}(\mathbf{b}) = (b_{i_1(\boldsymbol{\pi})}, \ldots, b_{i_n(\boldsymbol{\pi})})$, where $i_j(\boldsymbol{\pi}) \in \{1, \ldots, n\}$ is the index determined by the equation $\pi_{i_j(\boldsymbol{\pi})} = j$.

For all vectors $\mathbf{x}, \mathbf{b} \in \{0,1\}_w^n$, let

$$\mathcal{P}(\mathbf{x} \to \mathbf{b}) = \{\, \boldsymbol{\pi} \in \mathcal{P} :\; \boldsymbol{\pi}(\mathbf{x}) = \mathbf{b} \,\} \tag{28}$$

denote the set of permutations that transform the vector $\mathbf{x}$ to the vector $\mathbf{b}$. Let us introduce the probability distribution

$$\gamma_{\mathbf{x},\mathbf{b}} = (\, \gamma(\boldsymbol{\pi}|\mathbf{x}, \mathbf{b}),\, \boldsymbol{\pi} \in \mathcal{P}\,)$$

in such a way that $\gamma(\boldsymbol{\pi}|\mathbf{x}, \mathbf{b})$ can be positive only if $\boldsymbol{\pi} \in \mathcal{P}(\mathbf{x} \to \mathbf{b})$. Let us also denote a uniform probability distribution over the set $\mathcal{P}(\mathbf{x} \to \mathbf{b})$ by

$$\overline{\gamma}_{\mathbf{x},\mathbf{b}} = (\, \overline{\gamma}(\boldsymbol{\pi}|\mathbf{x}, \mathbf{b}),\, \boldsymbol{\pi} \in \mathcal{P}\,),$$

where

$$\overline{\gamma}(\boldsymbol{\pi}|\mathbf{x}, \mathbf{b}) = \begin{cases} |\,\mathcal{P}(\mathbf{x}, \mathbf{b})|^{-1}, & \text{if } \boldsymbol{\pi} \in \mathcal{P}(\mathbf{x} \to \mathbf{b}), \\ 0, & \text{if } \boldsymbol{\pi} \notin \mathcal{P}(\mathbf{x} \to \mathbf{b}). \end{cases}$$

For example, let $n = 4, k = 2$. The set $\{0,1\}_2^4$ consists of $\binom{4}{2} = 6$ binary vectors of length 4 having the weight 2 and $\mathcal{P}$ is the set consisting of $4! = 24$ permutations of components of the vector $(1,2,3,4)$. For all $\mathbf{x}, \mathbf{b} \in \{0,1\}_2^4$, the set $\mathcal{P}(\mathbf{x} \to \mathbf{b})$ consists of $2!2! = 4$ permutations. In

particular,

$$\mathcal{P}(1100 \to 1010) = \{\, 1324, 1423, 2314, 2413 \,\}.$$

Notice that

$$\left.\begin{array}{l} \mathbf{b} \;=\; \boldsymbol{\pi}(\mathbf{x}) \\ \mathbf{b}' = \mathbf{b} \oplus \mathbf{e} \end{array}\right\} \;\Rightarrow\; \boldsymbol{\pi}^{-1}(\mathbf{b}') = \boldsymbol{\pi}^{-1}(\mathbf{b}) \oplus \boldsymbol{\pi}^{-1}(\mathbf{e}) = \mathbf{x} \oplus \boldsymbol{\pi}^{-1}(\mathbf{e}) \tag{29}$$

and

$$\mathrm{wt}(\boldsymbol{\pi}^{-1}(\mathbf{e})) = \mathrm{wt}(\mathbf{e}), \tag{30}$$

i.e., the decoder observes "the transmitted codeword" $\mathbf{x}$ as $\mathbf{x} \oplus \boldsymbol{\pi}^{-1}(\mathbf{e})$. If the source generating the noise vectors is assumed to be a memoryless source, then (30) implies that the presence of the permutation $\boldsymbol{\pi}^{-1}$ does not affect the decoding strategy, and the scheme is equivalent to the one in Figure 6.

Processing of a given biometric vector $\mathbf{b}$ at the enrollment stage and processing data at the verification stage when the verifier considers only the output of the observation channel is illustrated in Figure 7.

The enrollment stage.

- *Choose a key codeword* $\mathbf{x}$ *according to a uniform probability distribution over the code* $\mathcal{C}$ *and compute the value of* $\mathrm{Hash}(\mathbf{x})$.
- *Given a pair of vectors* $(\mathbf{x}, \mathbf{b}) \in \{0,1\}_w^n \times \{0,1\}_w^n$, *choose a permutation* $\boldsymbol{\pi} \in \mathcal{P}$ *according to the probability distribution* $\gamma_{\mathbf{x},\mathbf{b}}$.
- *Store* $(\mathrm{Hash}(\mathbf{x}), \boldsymbol{\pi})$ *in the database.*

The verification stage.

- *Read the data* $(\mathrm{Hash}(\mathbf{x}), \boldsymbol{\pi})$ *associated with the claimed person from the database.*
- *Apply the inverse permutation* $\boldsymbol{\pi}^{-1}$ *to the vector* $\mathbf{b}'$ *and decode the key codeword given a received vector* $\boldsymbol{\pi}^{-1}(\mathbf{b}')$ *as* $\hat{\mathbf{x}}$. *If* $\mathrm{Hash}(\hat{\mathbf{x}}) = \mathrm{Hash}(\mathbf{x})$, *then accept the identity claim* (Acc). *If* $\mathrm{Hash}(\hat{\mathbf{x}}) \neq \mathrm{Hash}(\mathbf{x})$, *then reject the identity claim* (Rej).

One can easily see that if the code $\mathcal{C}$ satisfies (20), then (29), (30) guarantee that the false rejection rate FRR and the false acceptance rate for a randomly chosen person $\overline{\mathrm{FAR}}$ are the same as for the additive block coding scheme. Therefore, the reasons for introducing the more advanced permutation scheme are caused by possible decrease of the false acceptance rate for an attacker. We will derive a general formula for the FAR and demonstrate the effects for a specific assignment of input data.

Let

$$\gamma = (\, \gamma_{\mathbf{x},\mathbf{b}},\ \mathbf{x}, \mathbf{b} \in \{0,1\}_w^n \,)$$

denote the list of conditional probability distributions over the set $\mathcal{P}$. In general, the attacker applies a fixed function $\psi : \mathcal{P} \to \{0,1\}^n$ to the permutation $\boldsymbol{\pi}$ stored in the DB and submits the vector $\mathbf{b}' = \psi(\boldsymbol{\pi})$ to the verifier. Let us assume that the verifier decodes the key codeword as the vector $\hat{\mathbf{x}}[\boldsymbol{\pi}^{-1}(\mathbf{b}')]$. The probability of successful attack can be expressed as

$$\mathrm{FAR} = \frac{1}{M} \sum_{\mathbf{x} \in \mathcal{C}} \sum_{\mathbf{b}} \omega(\mathbf{b}) \sum_{\boldsymbol{\pi} \in \mathcal{P}} \gamma(\boldsymbol{\pi} | \mathbf{x}, \mathbf{b}) \chi\{\, \hat{\mathbf{x}}[\boldsymbol{\pi}^{-1}(\psi(\boldsymbol{\pi}))] = \mathbf{x} \,\}, \tag{31}$$

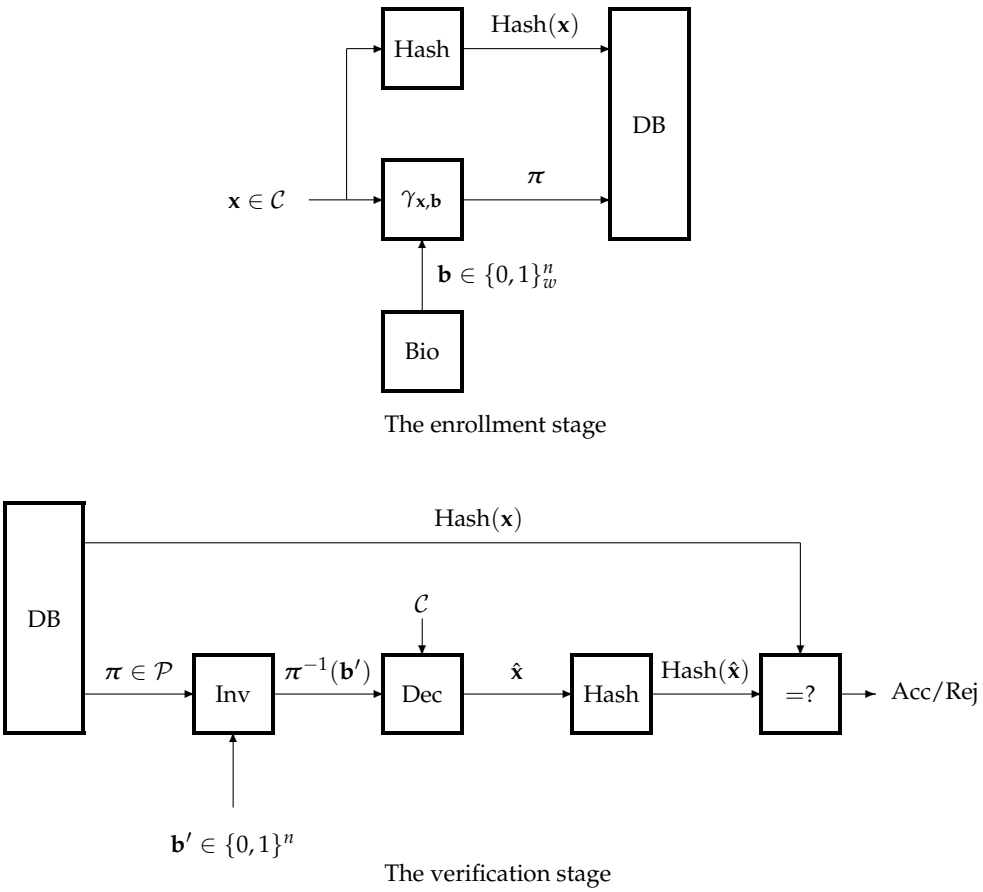


Fig. 7. The data processing in a permutation block coding scheme.

and one can easily see that FAR is maximized when the attacker applies the maximum *a posteriori* probability decoding, which results in

$$\psi(\boldsymbol{\pi}) = \boldsymbol{\pi}\Big(\arg\max_{\mathbf{x}\in\mathcal{C}} \gamma_{\text{bio}}(\boldsymbol{\pi}|\mathbf{x})\Big),$$

where

$$\gamma_{\text{bio}}(\boldsymbol{\pi}|\mathbf{x}) = \sum_{\mathbf{b}} \omega(\mathbf{b})\gamma(\boldsymbol{\pi}|\mathbf{x},\mathbf{b}). \tag{32}$$

Then

$$\text{FAR} = \frac{1}{M}\sum_{\boldsymbol{\pi}\in\mathcal{P}} \max_{\mathbf{x}\in\mathcal{C}} \gamma_{\text{bio}}(\boldsymbol{\pi}|\mathbf{x}).$$

Notice that ($\gamma_{\text{bio}}(\boldsymbol{\pi}|\mathbf{x})$, $\boldsymbol{\pi} \in \mathcal{P}$) is the conditional probability distribution over the set $\mathcal{P}$ and

$$\sum_{\boldsymbol{\pi}\in\mathcal{P}} \gamma_{\text{bio}}(\boldsymbol{\pi}|\mathbf{x}) = 1.$$

Notice also that the vector $\mathbf{x} \in \{0,1\}_w^n$ and the permutation $\boldsymbol{\pi} \in \mathcal{P}$ uniquely determine the vector $\mathbf{b}^0 \in \{0,1\}_w^n$ such that $\boldsymbol{\pi} \in \mathcal{P}(\mathbf{x} \to \mathbf{b}^0)$. Namely, $\mathbf{b}^0 = \boldsymbol{\pi}(\mathbf{x})$, and the sum at the right-hand side of (32) contains at most one non–zero term.
The attacker has two simple possibilities: 1) fix a codeword $\mathbf{x}' \in \mathcal{C}$ and submit the vector $\mathbf{b}' = \boldsymbol{\pi}(\mathbf{x}')$; 2) submit the most likely biometric vector. In the first case, the attacker has to know the code $\mathcal{C}$ and the stored permutation $\boldsymbol{\pi}$. In the second case, he does not know these data and equivalent to an attacker, who does not have access to the database and ignorant about the code. One can easily see that the probabilities of successful attacks are equal to $1/M$ and ω^*, respectively. Therefore the probability of successful attack under the maximum *a posteriori* probability decoding of the key codeword is bounded from below as follows:

$$\text{FAR} \geq \max\Big\{\frac{1}{M}, \omega^*\Big\}.$$

Let $n = 8$, $w = 4$, $M = 4$. Let the codewords $\mathbf{x}_1, \ldots, \mathbf{x}_4$ and the biometric vectors that can be processed at the enrollment stage be specified as

$$\begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_2\\ \mathbf{x}_3\\ \mathbf{x}_4\end{bmatrix} = \begin{bmatrix}00110011\\ 01010101\\ 10101010\\ 11001100\end{bmatrix}, \quad \begin{bmatrix}\mathbf{b}_1\\ .\\ .\\ \mathbf{b}_6\end{bmatrix} = \begin{bmatrix}00001111\\ 00110011\\ 01010101\\ 10101010\\ 11001100\\ 11110000\end{bmatrix},$$

i.e., $\mathcal{C} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\}$ and $\mathcal{B} = \{\mathbf{b}_1, \ldots, \mathbf{b}_6\}$. Then, for all pairs of vectors $(\mathbf{x}, \mathbf{b}) \in \mathcal{C} \times \mathcal{B}$,

$$|\,\mathcal{P}(\mathbf{x} \to \mathbf{b})\,| = (4!)^2 = 576 \tag{33}$$

and

$$|\,\mathcal{P}_{\mathcal{C}\to\mathcal{B}}(\mathbf{x} \to \mathbf{b})\,| = 4(2!)^4 = 64, \tag{34}$$

where $\mathcal{P}_{\mathcal{C}\to\mathcal{B}}(\mathbf{x} \to \mathbf{b})$ denotes the set of permutations $\boldsymbol{\pi} \in \mathcal{P}(\mathbf{x} \to \mathbf{b})$ such that $\boldsymbol{\pi}(\mathbf{x}') \in \mathcal{B}$ for all $\mathbf{x}' \in \mathcal{C}$.
Let us illustrate our considerations by the following examples:

$$\begin{bmatrix}\boldsymbol{\pi}'\\ \boldsymbol{\pi}'(\mathbf{x}_1)\\ \boldsymbol{\pi}'(\mathbf{x}_2)\end{bmatrix} = \begin{bmatrix}1\,2\,5\,6\,3\,4\,7\,8\\ 0\,0\,0\,0\,1\,1\,1\,1\\ 0\,1\,0\,1\,0\,1\,0\,1\end{bmatrix}, \quad \begin{bmatrix}\boldsymbol{\pi}''\\ \boldsymbol{\pi}''(\mathbf{x}_1)\\ \boldsymbol{\pi}''(\mathbf{x}_2)\end{bmatrix} = \begin{bmatrix}1\,2\,6\,5\,3\,4\,7\,8\\ 0\,0\,0\,0\,1\,1\,1\,1\\ 0\,1\,1\,0\,0\,1\,0\,1\end{bmatrix}.$$

The permutations $\boldsymbol{\pi}'$ and $\boldsymbol{\pi}''$ belong to the set $\mathcal{P}$. Furthermore, $\boldsymbol{\pi}'(\mathbf{x}_1) = \boldsymbol{\pi}''(\mathbf{x}_1) = \mathbf{b}_1$. However $\boldsymbol{\pi}'(\mathbf{x}_2) \in \mathcal{B}$, while $\boldsymbol{\pi}''(\mathbf{x}_2) \notin \mathcal{B}$. Suppose that $\boldsymbol{\pi}'$ is the permutation stored in the database. The attacker applies this permutation to all codewords of the code $\mathcal{C}$ and constructs the list $\boldsymbol{\pi}'(\mathbf{x}_1), \ldots, \boldsymbol{\pi}'(\mathbf{x}_4)$. All entries of this list are possible biometric vectors. If the permutation $\boldsymbol{\pi}''$ is stored in the database, then the list $\boldsymbol{\pi}'(\mathbf{x}_1), \ldots, \boldsymbol{\pi}'(\mathbf{x}_4)$ contains only

2 biometric vectors. The probability of successful attack is greater in the second case, and the permutation π' can be considered as "a bad" permutation.

The most of the permutations are bad permutations (see (33), (34)). This observation leads to the statement that the uniform probability distribution over the set $\mathcal{P}(\mathbf{x} \to \mathbf{b})$, where $\mathbf{x}$ is the selected codeword and $\mathbf{b}$ is the biometric vector, can bring a rather poor performance. Namely, suppose that the probability distribution over the set $\mathcal{B}$ is uniform, i.e., $\omega(\mathbf{b}) = 1/6$ for all $\mathbf{b} \in \mathcal{B}$. Let $\mathbf{x}$ be the codeword of the code $\mathcal{C}$ used at the enrollment stage. If $\gamma_{\mathbf{x},\mathbf{b}} = \overline{\gamma}_{\mathbf{x},\mathbf{b}}$, then the permutation is uniformly chosen from the set containing 576 entries. Only 64 of these permutations have the property that the set $\pi(\mathbf{x})$, $\mathbf{x} \in \mathcal{C}$ contains 4 biometric vectors, and the probability of successful attack is equal to $1/4$. For the other 512 permutations, the set $\pi(\mathbf{x})$, $\mathbf{x} \in \mathcal{C}$, contains 2 biometric vectors, and the probability of successful attack is equal to $1/2$. Thus

$$\mathrm{FAR} = \frac{64}{576}(1/4) + \frac{512}{576}(1/2) = 17/36.$$

Let us assign $\gamma_{\mathbf{x},\mathbf{b}}$ as a uniform probability distribution over the set $\mathcal{P}_{\mathcal{C}\to\mathcal{B}}(\mathbf{x} \to \mathbf{b})$ consisting of 64 entries. In all cases, the list $\pi(\mathbf{x})$, $\mathbf{x} \in \mathcal{C}$, contains 4 biometric vectors, and the probability of successful attack is equal to $1/4$. As a result, the probability of successful attack is expressed as

$$\mathrm{FAR} = \frac{64}{64}(1/4) = 1/4,$$

which is approximately twice less the value obtained with the uniform probability distribution. Moreover, we obtain that the lower bound $1/M$ on the probability FAR is attained with the equality.

Let us consider a non–uniform probability distribution over the set $\mathcal{B}$. Namely, let $a \in [1/4, 1/2]$ be a fixed parameter and let

$$\omega(\mathbf{b}) = \begin{cases} a, & \text{if } \mathbf{b} \in \{00001111, 11110000\}, \\ 1/4 - a/2, & \text{if } \mathbf{b} \in \mathcal{B}\backslash\{00001111, 11110000\}. \end{cases}$$

Notice that the set $\mathcal{P}_{\mathcal{C}\to\mathcal{B}}(\mathbf{x}_1 \to \mathbf{b}_1)$ contains 32 permutations π such that

$$\{\pi(\mathbf{x}_1), \pi(\mathbf{x}_2), \pi(\mathbf{x}_3), \pi(\mathbf{x}_4)\} = \{\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_5, \mathbf{b}_6\}$$

and 32 permutations π such that

$$\{\pi(\mathbf{x}_1), \pi(\mathbf{x}_2), \pi(\mathbf{x}_3), \pi(\mathbf{x}_4)\} = \{\mathbf{b}_1, \mathbf{b}_3, \mathbf{b}_4, \mathbf{b}_6\}.$$

Let us denote the subsets of these permutations by $\mathcal{P}'_{\mathcal{C}\to\mathcal{B}}(\mathbf{x}_1 \to \mathbf{b}_1)$ and $\mathcal{P}''_{\mathcal{C}\to\mathcal{B}}(\mathbf{x}_1 \to \mathbf{b}_1)$, respectively. Let

(a) $\gamma_{\mathbf{x}_1,\mathbf{b}_1}, \gamma_{\mathbf{x}_1,\mathbf{b}_6}$ be uniform probability distributions over the set $\mathcal{P}_{\mathcal{C}\to\mathcal{B}}(\mathbf{x}_1 \to \mathbf{b}_1)$;

(b) $\gamma_{\mathbf{x}_1,\mathbf{b}_2}, \gamma_{\mathbf{x}_1,\mathbf{b}_5}$ be uniform probability distributions over the set $\mathcal{P}'_{\mathcal{C}\to\mathcal{B}}(\mathbf{x}_1 \to \mathbf{b}_1)$;

(c) $\gamma_{\mathbf{x}_1,\mathbf{b}_3}, \gamma_{\mathbf{x}_1,\mathbf{b}_4}$ be uniform probability distributions over the set $\mathcal{P}''_{\mathcal{C}\to\mathcal{B}}(\mathbf{x}_1 \to \mathbf{b}_1)$.

If $\pi \in \mathcal{P}'_{\mathcal{C}\to\mathcal{B}}(\mathbf{x}_1 \to \mathbf{b}_1)$, then the *a posteriori* probabilities associated with the biometric vectors $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_5, \mathbf{b}_6$ are equal to

$$\frac{1}{32}(a/2, 1/2 - a/2, 1/2 - a/2, a/2).$$

$\tilde{\mathbf{b}}$	$\tilde{w}$	i	$\mathbf{b}$	w	$\tilde{\mathbf{b}}$	$\tilde{w}$	i	$\mathbf{b}$	w	$\tilde{\mathbf{b}}$	$\tilde{w}$	i	$\mathbf{b}$	w	$\tilde{\mathbf{b}}$	$\tilde{w}$	i	$\mathbf{b}$	w	$\tilde{\mathbf{b}}$	$\tilde{w}$	i	$\mathbf{b}$	w
0000	0	2	1100	2	0001	1	1	1001	2	0010	1	1	1010	2	0100	1	1	1100	2	1000	1	3	0110	2
1111	4	2	0011	2	1110	3	1	0110	2	1101	3	1	0101	2	1011	3	1	0011	2	0111	3	3	1001	2

Table 5. Transformation of vectors of length $n = 4$ and weights 0,1,3,4 to balanced vectors, where $\tilde{w}, w$ are the Hamming weights of the vectors $\tilde{\mathbf{b}}, \mathbf{b}$ and i is the length of the prefix of the vector $\tilde{\mathbf{b}}$, which has to be inverted to obtain the vector $\mathbf{b}$.

However $a/2 \geq 1/2 - a/2$, and the attacker outputs either the key codeword, which is mapped to the vector $\mathbf{b}_1$, or the key codeword, which is mapped to the vector $\mathbf{b}_6$. Similar considerations can be presented for the permutations belonging to the set $\mathcal{P}''_{\mathcal{C}\to\mathcal{B}}(\mathbf{x}_1 \to \mathbf{b}_1)$. As a result, we conclude that

$$\mathrm{FAR} = 64(a/64) = a,$$

i.e., the lower bound ω^* on the false acceptance rate is attained with the equality.

Let us consider the error–correcting capabilities of the verifier, who processes data of a legitimate user. Let P_w denote the probability that the vector $\mathbf{b}'$ differs from the vector $\mathbf{b}$ in w positions, $w = 0, \ldots, 8$. Then, assuming that the vectors $\mathbf{b}'$ are uniformly distributed over the set of vectors located at a fixed distance from the vector $\mathbf{b}$, we obtain that the probability of correct decoding for the code $\mathcal{C}$ and the threshold $T = 2$ is equal to

$$1 - \mathrm{FRR} = P_0 + P_1 + (16/28)P_2,$$

since the decoder makes the correct decision for all error patterns of weight at most 1 and for 16 error patterns of weight 2 (the total number of error patterns of weight 2 is equal to 28). Suppose that the processed biometric vectors are constructed as a concatenation of L vectors $\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(L)} \in \mathcal{B}$, i.e., the total length of the vector is equal to $8L$. Suppose also that the vectors $\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(L)}$ are independently generated according to a uniform probability distribution over the set $\mathcal{B}$. Let the verifier make the acceptance decision if and only if such a decision is made for all L entries. Then the probability of correct decision is equal to $(1 - \mathrm{FRR})^L$. On the other hand, the probability of successful attack, when the probability distributions $\gamma_{\mathbf{x},\mathbf{b}}$ are used is equal to $(1/4)^L$. This example illustrates the possibility of constructing the desired probability distribution over the permutations only for the subblocks of input data, and the search for good distributions is computationally feasible.

Notice that the fixed Hamming weight of the possible biometric vectors is the constraint that has to be satisfied to implement the permutation block coding scheme. It can be done if the observer takes into account only a fixed number of the most reliable biometric parameters. For example, in the case of processing fingerprints, one can put an $n_1 \times n_2$ grid on the 2-dimensional plane (in this case, $n = n_1 n_2$) and register the w most reliable minutiae points in the cells of that grid. In general case, the biometric binary vector of length n can be viewed as a vector of n features where positions of 1's index the features that are present in the outcomes of the measurements. The total number of the most reliable features taken into account by the authentication scheme can be fixed in advance.

Another useful possibility is known as balancing arbitrary binary vector by the inversion of its prefix in such a way that the obtained vector has weight $\lfloor n/2 \rfloor$. The corresponding statement is presented below, and the examples of the transformation are given in Table 5. One can see that, for any binary vector $\tilde{\mathbf{b}} \in \{0,1\}^n$, one can find an index $i \in \{0, \ldots, n\}$ in such a way that the vector $\tilde{\mathbf{b}}$ is transformed to a balanced vector by the inversion of the first i components,

i.e., $(i - \tilde{w}_i) + \tilde{w} - \tilde{w}_i = \lfloor n/2 \rfloor$, where $\tilde{w}$ and $\tilde{w}_i$ denote the Hamming weight of the vector $\tilde{\mathbf{b}}$ and the Hamming weight of the prefix of length i of the vector $\tilde{\mathbf{b}}$, respectively. The proof directly follows from the observation that the path on the plane whose coordinates are defined as $(j, \tilde{w}_j)$, $j = 0, \ldots, n$, starts at the point $(0, \mathrm{wt}(\mathbf{b}))$, ends at the point $(n, n - \mathrm{wt}(\mathbf{b}))$, and has increments ± 1. Therefore, there is at least one index i such that $\tilde{w}_i = \lfloor n/2 \rfloor$. Notice that the case $w = \lfloor n/2 \rfloor$ can be viewed as the most interesting one meaning the characteristics of the permutation block coding scheme. The claim above shows that an additional storage of the value of the parameter i used to transform an arbitrary binary vector to a vector belonging to the set $\{0,1\}^n_{\lfloor n/2 \rfloor}$ makes the implementation of such a scheme possible in general.

The mapping of the pair $(\mathbf{x}, \mathbf{b})$ to a binary string stored in the database can be viewed as the encryption of the message $\mathbf{b}$, which is parameterized by a key codeword $\mathbf{x} \in \mathcal{C}$ chosen at random. An interesting point is the possibility of decreasing the probability of successful attack, when an attacker tries to pass through the authentication stage with the acceptance decision, by using a randomized mapping, although *the values of additional random parameters are public*. In the permutation block coding scheme, a randomly chosen permutation that transforms the vector $\mathbf{x}$ to the vector $\mathbf{b}$ is used for these purposes. As the set of possible permutations has the cardinality, which is exponential in the length of the vectors, the designer has good chances to hide many of biometric vectors that differ from the most likely vector $\mathbf{b}^*$ into the information that can correspond to the vector $\mathbf{b}^*$. Thus, one can even reach exactly the same secrecy of the coded system as the secrecy of the blind guessing of the biometric vector, when the attacker does not have access to the database and ignorant about the code. In other words, one can talk about the possibility of constructing permutation block coding schemes that have *a perfect algorithmic secrecy*. This notion is different from the usual definition of perfectness, which is understood as the point that the conditional entropy of the probability distribution over the key codewords, given the content of the database, is equal to $\log M$. In our example presented in the previous subsection, the *a posteriori* probability distribution over the key codewords certainly depends on a particular permutation, and the conditional entropies of these distributions can be much less than the entropy of a uniform probability distribution. Nevertheless, an optimum attacker cannot use this fact, and his observations do not introduce changing in the decoding algorithm.

8. References

Bolle, R. M., Connell, J. H., Pankanti S., Ratha, N. K. & Senior A. W. (2004). *Guide to Biometrics*, Springer.

Cohen, G. & Zemor G. (2006). Syndrome–coding for the wiretap channel revisited, *Proceedings of IEEE Information Theory Workshop*, IEEE Press, China, pp. 33–36.

Dodis Y., Reyzin L. & Smith, A. (2004). Fuzzy extractors: How to generate strong keys from biometrics and other noisy data, *Advances in Cryptography: Lecture Notes in Computer Science*, no. 3027, Springer, pp. 523–540.

Gallager, R. (1968). *Information Theory and Reliable Communication*, Willey.

Juels, A. & Wattenberg, M. (1999). A fuzzy commitment scheme, *Proceedings of ACM Conference on Computer and Communication Security*, ACM Press, Singapore, pp. 28–36.

Knuth, D. E. (1986). Efficient balanced codes, *IEEE Transactions on Information Theory*, vol. 32, no. 1, pp. 51–53.

Korte, U., Krawczak, M., Merkle, J., Plaga, R., Niesing, M., Tiemann, C., Han Vinck, A. J., Martini, U. (2008). A cryptographic biometric authentication system based on genetic fingerprints, *Proceedings of Sicherheit*, Springer, Germany, pp. 263–276.

Wyner, A. (1975). The wiretap channel, *Bell System Technical Journal*, vol. 54, no. 8, pp. 1355–1387.

Balakirsky, V. B., Ghazaryan, A. R. & Han Vinck, A. J. (2006a). Processing fingerprints via binary codes: The BMW algorithm, *Proceedings of the 27th Symposium on Information Theory in the Benelux*, Lagendijk, R. L. & Weber, J. H. (Eds.), The Netherlands, pp. 267–274.

Balakirsky, V. B., Ghazaryan, A. R. & Han Vinck, A. J. (2006b). General principles of constructing biometric authentication schemes using block codes, *Proceedings of the International Workshop "Algorithms and Mathematical Methods in Networking"*, Han Vinck, A. J. (Ed.), Institute fur Experimentelle Mathematik Press, Germany, pp. 8–18.

Balakirsky, V. B., Ghazaryan, A. R. & Han Vinck, A. J. (2007). Testing the independence of two non–stationary random processes with applications to biometric authentication, *Proceedings of the International Symposium on Information Theory*, IEEE Press, France, pp. 2671–2675, 2007.

Balakirsky, V. B., Ghazaryan, A. R. & Han Vinck, A. J. (2008a). Additive block coding schemes for biometric authentication with the DNA data, *Lecture Notes in Computer Science*, vol. 5372, Schouten, B., et al. (Eds.), Springer, pp. 160–169.

Balakirsky, V. B., Ghazaryan, A. R. & Han Vinck, A. J. (2008b). Performance of additive block coding schemes oriented to biometric authentication, *Proceedings of the 29th Symposium on Information Theory in the Benelux*, Van de Perre, L. et. al (Eds.), Belgium, pp. 19–26.

Balakirsky, V. B., Ghazaryan, A. R. & Han Vinck, A. J. (2009a). Secrecy of permutation block coding schemes designed for biometric authentication, *Proceedings of the 30th Symposium on Information Theory in the Benelux*, Willems, F. M. J., & Tjalkens, T. J. (Eds.), The Netherlands, pp. 11–19.

Balakirsky, V. B., Ghazaryan, A. R. & Han Vinck, A. J. (2009b). Mathematical model for constructing passwords from biometrical data, *Security and Communication Networks*, vol. 2, no. 1, Wiley, pp. 1–9.

Balakirsky, V. B. & Han Vinck, A. J. (2010). A simple scheme for constructing fault–tolerant passwords from biometric data, *EURASIP Journal on Information Security*, vol. 2010, Article ID 819376, doi:10.1155/2010/819376.