Additive Block Coding Schemes for Biometric Authentication with the DNA Data*

Vladimir B. Balakirsky, Anahit R. Ghazaryan, and A.J. Han Vinck

Institute for Experimental Mathematics, Ellernstr. 29, 45326 Essen, Germany
v_b_balakirsky@rambler.ru, a_ghazaryan@rambler.ru, vinck@iem.uni-due.de

Abstract. To implement a biometric authentication scheme, the templates of a group of people are stored in the database (DB) under the names of these people. Some person presents a name, and the scheme compares the template of this person and the template associated with the claimed person to accept or reject their identity [1]. The templates of people stored in the DB should be protected against attacks for discovery the biometrics and attacks for successful passing through the verification test. The authentication algorithm developed by Juels and Wattenberg [2] is a possible solution to the problem. However, implementations of this algorithm for practical data require generalized versions of the algorithm and their analysis. We introduce a mathematical model for DNA measurements and present such a generalization. Some numerical results illustrate the correction of errors for the DNA measurements of a legitimate user and protection of templates against attacks for successful passing the verification stage by an attacker.

1 An Additive Block Coding Scheme

An additive block coding scheme proposed in [2] can be presented as follows (see Figure 1). Let $\mathcal{C}$ be a set consisting of M different binary vectors of length n (a binary code of length n for M messages). The entries of the set $\mathcal{C}$ are called key codewords. One of the key codewords $\mathbf{x} \in \mathcal{C}$ is chosen at random with probability $1/M$. This codeword is added modulo 2 to the binary vector $\mathbf{b}$ generated by a biometrical source, and the vector $\mathbf{y} = \mathbf{x} \oplus \mathbf{b}$ is stored in the DB under the name of the person whose biometrics is expressed by the vector $\mathbf{b}$. Furthermore, the value of a one–way hash function Hash at the vector $\mathbf{x}$ (a one-to-one function whose value can be easily computed, while the inversion is a difficult problem) is also stored in the DB. Having received another binary vector $\mathbf{b}'$ and the claimed name, the verifier finds the key codeword $\hat{\mathbf{x}} \in \mathcal{C}$ located at the minimum Hamming distance from the vector $\mathbf{z} = \mathbf{y} \oplus \mathbf{b}'$. The basis for the algorithm is the observation.

$$\left.\begin{array}{r} \mathbf{y} = \mathbf{x} \oplus \mathbf{b} \\ \mathbf{b}' = \mathbf{b} \oplus \mathbf{e} \end{array}\right\} \Rightarrow \quad \mathbf{x} \oplus \mathbf{e} = \mathbf{z}.$$

* This work was partially supported by the DFG.

B. Schouten et al. (Eds.): BIOID 2008, LNCS 5372, pp. 160–169, 2008.
© Springer-Verlag Berlin Heidelberg 2008

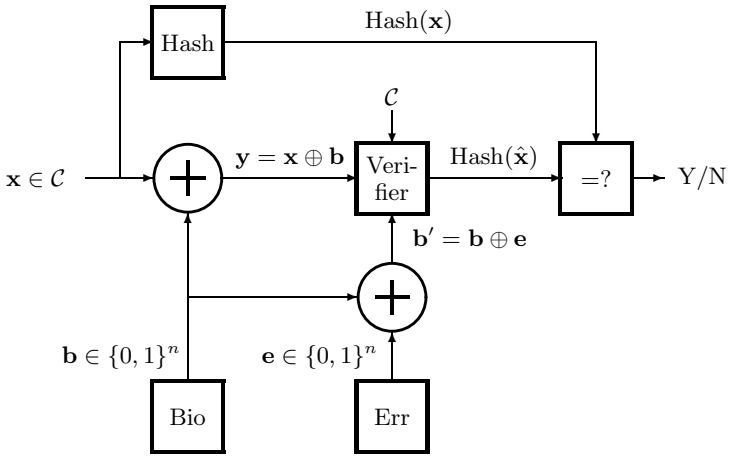

Fig. 1. Verification of a person using an additive block coding scheme with a binary code

In particular, if the number of positions where the vectors $\mathbf{b}$ and $\mathbf{b}'$ differ does not exceed $\lfloor (d_{\mathcal{C}} - 1)/2 \rfloor$, where $d_{\mathcal{C}}$ is the minimum distance of the code $\mathcal{C}$, then the key codeword used at the enrollment stage will be found. Then $\mathrm{Hash}(\hat{\mathbf{x}})$ is equal to $\mathrm{Hash}(\mathbf{x})$ and the identity claim is accepted. Otherwise, the claim is rejected.

Notice that the verification scheme in Figure 1 can be represented as transmission of the key codeword $\mathbf{x}$ over two parallel channels, because

$$\left.\begin{array}{l}\mathbf{y} = \mathbf{x} \oplus \mathbf{b} \\ \mathbf{b}' = \mathbf{b} \oplus \mathbf{e}\end{array}\right\} \Rightarrow \left\{\begin{array}{l}\mathbf{x} \oplus \mathbf{b} = \mathbf{y} \\ \mathbf{x} \oplus \mathbf{e} = \mathbf{z}.\end{array}\right.$$

Thus, we say that the verifier receives a pair of vectors $(\mathbf{x}\oplus\mathbf{b}, \mathbf{x}\oplus\mathbf{e})$ (see Figure 2), while the attacker receives only the first component and the JW decoder analyzes only the second component of that pair. The transformations $\mathbf{x} \to \mathbf{y}$ and $\mathbf{x} \to \mathbf{z}$ can be interpreted as transmissions of the key codeword over the biometric and the observation channels, respectively.

The processing of biometric data is illustrated in Table 1, where we assume that $n = 6$ and assign a binary block code $\mathcal{C}$ for $M = 8$ messages. Let 011011 be the input vector and let 011110 be the chosen key codeword. Then the vector 000101 is stored in the DB. The attacker forms the set of candidates for the biometric vectors as $000101 \oplus \mathcal{C}$ and searches for the vector having the maximum probability computed over the ensemble $\mathrm{Pr}_{\mathrm{bio}}$. If 111011 is the noisy observation of the biometric vector, then the JW decoder forms the set $000101 \oplus 111011 \oplus \mathcal{C}$, considers it as the set of possible observation noise vectors, and searches for the vector having the maximum probability computed over the ensemble $\mathrm{Pr}_{\mathrm{err}}$. The verifier analyzes the pair of these sets and searches for the pair of vectors having the maximum probability computed over the ensemble $\mathrm{Pr}_{\mathrm{bio}} \times \mathrm{Pr}_{\mathrm{err}}$.

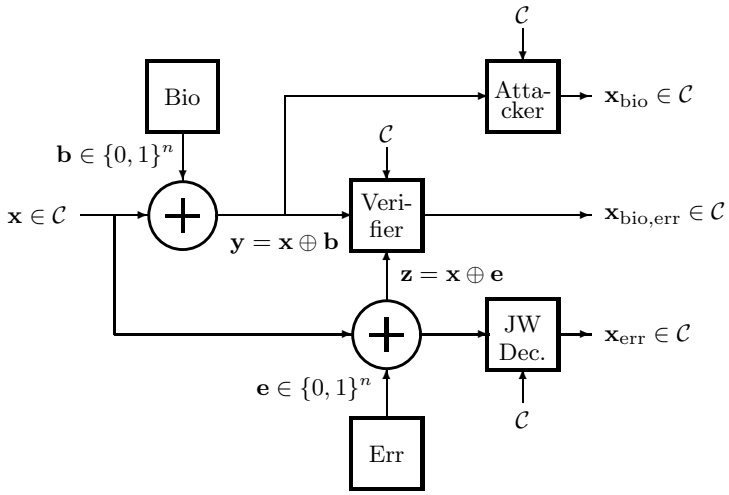

Fig. 2. Representation of the additive block coding as a scheme where a key codeword $\mathbf{x}$ is received under the biometric noise $\mathbf{b}$ and the observation noise $\mathbf{e}$

Table 1. Example of processing data with the additive block coding scheme for $n = 6$ and $M = 8$

$\mathcal{C}$	$\longrightarrow$	$\mathbf{y} \oplus \mathcal{C}$	$\mathbf{z} \oplus \mathcal{C}$
000000	$\mathbf{x} = 011110$	000101	111110
001011	$\mathbf{b} = 011011$	001110	110101
010101	$\mathbf{y} = \mathbf{x} \oplus \mathbf{b}$	010000	101011
011110	$= 000101$	011011	100000
100110	$\mathbf{b}' = 111011$	100011	011000
101101	$\mathbf{z} = \mathbf{b}' \oplus \mathbf{y}$	101011	010011
110011	$= 111110$	110110	001101
111000		111101	000110

We will assume that particular binary vectors $\mathbf{b}$ and $\mathbf{e}$ are chosen as the biometric and the observation noise vectors according to the probability distributions (PDs)

$$\Big(\Pr_{\rm bio}\Big\{ B = \mathbf{b} \Big\},\ \mathbf{b} \in \{0,1\}^n \Big), \quad \Big(\Pr_{\rm err}\Big\{ E = \mathbf{e} \Big\},\ \mathbf{e} \in \{0,1\}^n \Big).$$

Let $\mathbf{x}_{\rm bio}$, $\mathbf{x}_{\rm err}$, and $\mathbf{x}_{\rm bio,err}$ denote results of the decoding when the vectors $\mathbf{y}$, $\mathbf{z}$, and the pair of vectors $(\mathbf{y}, \mathbf{z})$ are available. One can easily check that the maximum probabilities of correct decoding are attained by the maximum *a posteriori* probability decoding rules, i.e., the optimum estimates of the key codeword satisfy the equalities

$$\Pr_{\rm bio}\Big\{ B = \mathbf{x}_{\rm bio} \oplus \mathbf{y} \Big\} = \max_{\mathbf{x} \in \mathcal{C}} \Pr_{\rm bio}\Big\{ B = \mathbf{x} \oplus \mathbf{y} \Big\},$$

$$\Pr_{\text{err}}\Big\{E = \mathbf{x}_{\text{err}} \oplus \mathbf{z}\Big\} = \max_{\mathbf{x}\in\mathcal{C}} \Pr_{\text{err}}\Big\{E = \mathbf{x} \oplus \mathbf{z}\Big\},$$

and

$$\Pr_{\text{bio}}\Big\{B = \mathbf{x}_{\text{bio,err}} \oplus \mathbf{y}\Big\} \Pr_{\text{err}}\Big\{E = \mathbf{x}_{\text{bio,err}} \oplus \mathbf{z}\Big\} = \max_{\mathbf{x}\in\mathcal{C}}\Big[\Pr_{\text{bio}}\Big\{B = \mathbf{x} \oplus \mathbf{y}\Big\} \Pr_{\text{err}}\Big\{E = \mathbf{x} \oplus \mathbf{z}\Big\}\Big].$$

Then the probabilities that the decoded codewords coincide with the transmitted key codewords can be expressed as

$$\Lambda_{\text{bio}} = \frac{1}{M}\sum_{\mathbf{y}} \max_{\mathbf{x}\in\mathcal{C}} \Pr_{\text{bio}}\Big\{B = \mathbf{x} \oplus \mathbf{y}\Big\},$$

$$\Lambda_{\text{err}} = \frac{1}{M}\sum_{\mathbf{z}} \max_{\mathbf{x}\in\mathcal{C}} \Pr_{\text{err}}\Big\{E = \mathbf{x} \oplus \mathbf{z}\Big\},$$

$$\Lambda_{\text{bio,err}} = \frac{1}{M}\sum_{\mathbf{y},\mathbf{z}} \max_{\mathbf{x}\in\mathcal{C}}\Big[\Pr_{\text{bio}}\Big\{B = \mathbf{x} \oplus \mathbf{y}\Big\} \Pr_{\text{err}}\Big\{E = \mathbf{x} \oplus \mathbf{z}\Big\}\Big].$$

2 Structure of the DNA Data and Mathematical Model

The most common DNA variations are Short Tandem Repeats (STR): arrays of 5 to 50 copies (repeats) of the same pattern (the motif) of 2 to 6 pairs. As the number of repeats of the motif highly varies among individuals, it can be effectively used for identification of individuals. The human genome contains several 100,000 STR loci, 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. Let the t-th source generate a pair of integers according to the PD

$$\Pr_{\text{DNA}}\Big\{\,(A_{t,1}, A_{t,2}) = (a_{t,1}, a_{t,2})\,\Big\} = \pi_t(a_{t,1})\pi_t(a_{t,2}),$$

where $a_{t,1}, a_{t,2} \in \{c_t, \ldots, c_t + k_t - 1\}$ and c_t, k_t are given positive integers. Thus, *we assume that $A_{t,1}$ and $A_{t,2}$ are independent random variables* that contain

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

$$\Pr_{\mathrm{DNA}}\Big\{ (A_1, A_2) = (\mathbf{a}_1, \mathbf{a}_2) \Big\} = \prod_{t=1}^{n} \Pr_{\mathrm{DNA}}\Big\{ (A_{t,1}, A_{t,2}) = (a_{t,1}, a_{t,2}) \Big\},$$

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

Let us fix a $t \in \{1, \ldots, n\}$ and denote

$$\mathcal{P}_t \stackrel{\triangle}{=} \Big\{ s = (i, j) : \, i, j \in \{c_t, \ldots, c_t + k_t - 1\}, \, j \geq i \Big\}.$$

Then the PD of a pair of random variables

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

which represents the outcome of the t-th measurement, can expressed as

$$\Pr_{\mathrm{DNA}}\Big\{ S_t = (i, j) \Big\} = \omega_t(i, j),$$

where $\omega_t(i, j) \stackrel{\triangle}{=} \pi_t^2(i)$, if $j = i$, and $\omega_t(i, j) \stackrel{\triangle}{=} 2\pi_t(i)\pi_t(j)$, if $j \neq i$. Denote $\omega_t \stackrel{\triangle}{=} (\, \omega_t(i, j), \, (i, j) \in \mathcal{P}_t \,)$ and

$$\begin{aligned}
G(\omega_t) &\stackrel{\triangle}{=} -\log \max_{(i,j) \in \mathcal{P}_t} \omega_t(i, j), \\
H(\omega_t) &\stackrel{\triangle}{=} -\sum_{(i,j) \in \mathcal{P}_t} \omega_t(i, j) \log \omega_t(i, j), \\
p(\omega_t) &\stackrel{\triangle}{=} \sum_{i=c_t}^{c_t + k_t - 1} \omega_t(i, i), \\
h(\omega_t) &\stackrel{\triangle}{=} -(1 - p(\omega_t)) \log(1 - p(\omega_t)) - p(\omega_t) \log p(\omega_t).
\end{aligned}$$

One can easily see that the best guess of the output of the t-th source is a pair (i_t^*, j_t^*) such that $\omega_t(i_t^*, j_t^*) \geq \omega_t(i, j)$ for all $(i, j) \in \mathcal{P}_t$. Therefore, $2^{-G(\omega_t)}$ is the probability that the guess is correct. The value of $p(\omega_t)$ is the probability that the genotype is homozygous, $H(\omega_t)$ is the entropy of the PD ω_t, and $h(\omega_t)$ is the entropy of the PD $(1 - p(\omega_t), p(\omega_t))$.

Let us assume that $q_t \stackrel{\triangle}{=} |\, \mathcal{P}_t \,| = k_t(k_t + 1)/2$ values $\omega_t(i, j), (i, j) \in \mathcal{P}_t$, are different and introduce two transformations of a pair of measurements $(i, j) \in \mathcal{P}_t$. (a) Let $i = j$ imply $\beta(i, j) = 0$ and let $i \neq j$ imply $\beta(i, j) = 1$. (b) Given an integer $q \geq q_t$, let $\beta_q(i, j) = b$ if and only if there are b pairs $(i', j') \in \mathcal{P}_t$ such that $\omega_t(i', j') > \omega_t(i, j)$. In particular, $\beta_q(i_t^*, j_t^*) = 0$.

We will denote the vector of measurements available to the scheme at the enrollment stage by $\mathbf{s} = ((i_1, j_1), \ldots, (i_n, j_n))$. The transformations of this vector

will be denoted by $\beta(\mathbf{s}) = (\beta(i_1, j_1), \ldots, \beta(i_n, j_n))$ and $\beta_q(\mathbf{s}) = (\beta_q(i_1, j_1), \ldots, \beta_q(i_n, j_n))$. Similar notations will be used for the vector $\mathbf{s}' = ((i'_1, j'_1), \ldots, (i'_n, j'_n))$ available to the scheme at the verification stage.

Example (the quantities below describe the **TH01** allele in Table 2). Let $c_t = 6$, $k_t = 4$, and $(\pi(6), \ldots, \pi(9)) = (0.23, 0.19, 0.09, 0.49)$. Then

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

	$j=6$	$j=7$	$j=8$	$j=9$
$i=6$	.0529	.0437	.0207	.1127
$i=7$	.0437	.0361	.0171	.0931
$i=8$	.0207	.0171	.0081	.0441
$i=9$	.1127	.0931	.0441	.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 the zeroes. The sum of all entries of the i-th row is equal to the probability that $\min\{A_{t,1}, A_{t,2}\} = i$ and the sum of all entries of the j-th column is equal to the probability that $\max\{A_{t,1}, A_{t,2}\} = j$ (these sums are denoted by $\omega_{t,\min}(i)$ and $\omega_{t,\max}(j)$),

$$\Big[\omega_t(i,j)\Big]_{\substack{i,j=6,\ldots,9\\ j\geq i}} =$$

	$j=6$	$j=7$	$j=8$	$j=9$	$\omega_{t,\min}(i)$
$i=6$	.0529	.0874	.0414	.2254	.4071
$i=7$		.0361	.0342	.1862	.2565
$i=8$			.0081	.0882	.0963
$i=9$				.2401	.2401
$\omega_{t,\max}(j)$	.0529	.1235	.0837	.7399	

Reading the entries of this matrix in the decreasing order of their values brings the following table,

i,j	$9,9$	$6,9$	$7,9$	$8,9$	$6,7$	$6,6$	$6,8$	$7,7$	$7,8$	$8,8$
$\beta(i,j)$	1	0	0	0	0	1	0	1	0	1
$\beta_q(i,j)$	0	1	2	3	4	5	6	7	8	9
$\omega_t(i,j)$	.2401	.2254	.1862	.0882	.0874	.0529	.0414	.0361	.0342	.0081
$G(\omega_t)$	$-\log .2401 = 2.07$									
$p(\omega_t)$	$.2401 + .0529 + .0361 + .0081 = .3372$									

Some parameters of the PDs that were under considerations in the BioKey–STR project [3] are given in Table 2. We conclude that results of the DNA measurements can be represented by a binary vector of length $\lceil \log(q_1 \ldots q_n) \rceil = 129$ bits. However the PD over these vectors is non–uniform and (roughly speaking) only 109 bits carry information about the measurements. If an attacker is supposed to guess this vector, then the best guess is the vector of pairs $\mathbf{s}^* = ((i_1^*, j_1^*), \ldots, (i_n^*, j_n^*))$. By the construction of the β_q transformation, $\boldsymbol{\beta}_q(\mathbf{s}^*)$ is the all-zero vector. The probability that the guess is correct is equal to $2^{-76.8}$. If the vector of n pairs of integers is transformed to a binary vector of length

Table 2. Some characteristics of the PDs $\omega_1, \ldots, \omega_n$ that describe the DNA measurements for $n = 28$

t	Name	$\log q_t$	$H(\omega_t)$	$G(\omega_t)$	$p(\omega_t)$	$h(\omega_t)$
1	**D8S1179**	4.39	4.08	3.01	0.20	0.73
2	**D3S1358**	3.91	3.71	2.87	0.22	0.76
3	**VWA**	4.39	4.13	3.12	0.19	0.71
4	**D7S820**	4.39	4.07	3.25	0.19	0.71
5	**ACTBP2**	7.71	7.43	6.13	0.06	0.32
6	**D7S820**	4.81	4.24	3.31	0.19	0.69
7	**FGA**	5.49	4.92	3.54	0.15	0.61
8	**D21S11**	4.81	4.13	3.01	0.20	0.73
9	**D18S51**	5.78	5.28	4.43	0.13	0.55
10	**D19S433**	4.39	3.59	2.33	0.26	0.82
11	**D13S317**	4.81	4.15	2.56	0.22	0.75
12	**TH01**	3.32	2.85	2.07	0.34	0.92
13	**D2S138**	6.04	5.60	4.23	0.12	0.52
14	**D16S539**	4.81	3.78	2.25	0.25	0.81
15	**D5S818**	3.91	3.11	1.81	0.31	0.89
16	**TPOX**	3.91	2.91	1.79	0.37	0.95
17	**CF1PO**	3.91	3.16	2.16	0.28	0.86
18	**D8S1179**	5.49	4.49	3.15	0.19	0.69
19	**VWA-1**	4.39	4.13	3.12	0.19	0.71
20	**PentaD**	5.17	4.32	3.13	0.19	0.70
21	**PentaE**	6.91	5.87	4.02	0.11	0.51
22	**DYS390**	4.39	3.24	2.06	0.30	0.88
23	**DYS429**	3.91	2.97	1.78	0.33	0.91
24	**DYS437**	2.58	2.26	1.58	0.40	0.97
25	**DYS391**	3.32	1.90	1.11	0.47	1.00
26	**DYS385**	5.17	3.61	1.72	0.34	0.93
27	**DYS389I**	2.58	2.01	1.18	0.50	1.00
28	**DYS389II**	3.91	3.14	2.04	0.31	0.89
	$\sum$	128.6	109.1	76.8	7.01	21.5

n containing ones at positions where the genotype is homozygous, then the expected weight of the vector can be computed as $p(\omega_1) + \ldots + p(\omega_n) = 7.01$, because the weight is the sum of n independent binary random variables where the t-th variable takes value 1 with probability $p(\omega_t)$. The difference between the entropies $H(\omega_t) - h(\omega_t)$ characterizes the loss of data for the β transformation of presented measurements.

3 Verification of a Person Using the DNA Measurements

Additive block coding schemes are oriented to the correction of certain types of measurement errors with simultaneous hiding biometric data from an attacker. If only the allelic drop–in/out errors are possible, then correction of errors means the transformation of the binary vector $\beta(\mathbf{s}')$ to the binary vector $\beta(\mathbf{s})$, where

$\mathbf{s}$ and $\mathbf{s}'$ are biometric vectors presented to the scheme at the enrollment and the verification stages, respectively. This procedure can be organized using an additive block coding scheme with a binary code of length n. However, the β transformation brings an essential loss of input data, and the verifier cannot make a reliable acceptance decision.

Notice that the β_q transformation is lossless and propose the use of an additive block coding scheme with a q-ary code $\mathcal{C}_q$, where q is chosen in such a way that $q_1, \ldots, q_n \leq q$. All the vectors in Figures 1, 2 become q-ary vectors, and $\oplus$ has to be understood as the component-wise addition modulo q. To distinguish between these vectors and binary vectors, we attach the index q and introduce the following translation to parallel channels:

$$\left.\begin{array}{l}\mathbf{y}_q = \mathbf{x}_q \oplus \mathbf{b}_q \\ \mathbf{b}'_q = \mathbf{b}_q \oplus \mathbf{e}_q\end{array}\right\} \Rightarrow \left\{\begin{array}{l}\mathbf{x}_q \oplus \mathbf{b}_q = \mathbf{y}_q \\ \mathbf{x}_q \ominus \mathbf{e}_q = \mathbf{z}_q\end{array}\right.$$

where $\mathbf{z}_q = \mathbf{y}_q \ominus \mathbf{b}'_q$ and $\ominus$ denotes the component-wise difference modulo q. Our data processing algorithm is presented below.

Preprocessing. Assign a binary code $\mathcal{C}$ for M messages and a q-ary code $\mathcal{C}_q$ for M_q messages. Both codes have length n.

Enrollment (input data are specified by the vector $\mathbf{s}$).

(0) Construct the vectors $\beta(\mathbf{s})$ and $\beta_q(\mathbf{s})$.
(1) Choose a binary key codeword $\mathbf{x} \in \mathcal{C}$. Store $\mathrm{Hash}(\mathbf{x})$ and $\mathbf{y} = \mathbf{x} \oplus \beta(\mathbf{s})$ in the DB.
(2) Choose a q-ary key codeword $\mathbf{x}_q \in \mathcal{C}_q$. Store $\mathrm{Hash}(\mathbf{x}_q)$ and $\mathbf{y}_q = \mathbf{x}_q \oplus \beta_q(\mathbf{s})$ in the DB.

Verification (input data are specified by the vector $\mathbf{s}'$ and content of the DB).

(0) Construct the vectors $\beta(\mathbf{s}')$ and $\beta_q(\mathbf{s}')$.
(1) Consider $(\mathbf{y}, \mathbf{y} \oplus \beta(\mathbf{s}'))$ as the pair of received words and decode the binary key codeword as $\hat{\mathbf{x}}$. If $\mathrm{Hash}(\hat{\mathbf{x}}) \neq \mathrm{Hash}(\mathbf{x})$, then output "No" and terminate.
(2) Consider $(\mathbf{y}_q, \mathbf{y}_q \ominus \beta_q(\mathbf{s}'))$ as the pair of received words and decode the q-ary key codeword as $\hat{\mathbf{x}}_q$. If $\mathrm{Hash}(\hat{\mathbf{x}}_q) \neq \mathrm{Hash}(\mathbf{x}_q)$, then output "No". Otherwise, output "Yes".

The formal description of biometric sources for the 1-st and the 2-nd steps are as follows: for all $\mathbf{b} \in \{0,1\}^n$ and $\mathbf{b}_q \in \{0, \ldots, q-1\}^n$,

$$\Pr_{\mathrm{bio}}\Big\{ B = \mathbf{b} \Big\} = \prod_{t=1}^{n} \Pr_{\mathrm{DNA}}\Big\{ \beta(S_t) = b_t \Big\},$$

$$\Pr_{\mathrm{bio},q}\Big\{ B_q = \mathbf{b}_q \Big\} = \prod_{t=1}^{n} \Pr_{\mathrm{DNA}}\Big\{ \beta_q(S_t) = b_{t,q} \Big\}.$$

Suppose that the noise of observations is specified is such a way that, for all $\mathbf{e} \in \{0,1\}^n$ and $\mathbf{e}_q \in \{0, \ldots, q-1\}^n$,

$$\Pr_{\mathrm{err}}\Big\{ E = \mathbf{e} \Big\} = \prod_{t=1}^{n} \left\{\begin{array}{ll}1-\varepsilon, & \text{if } e_t = 0, \\ \varepsilon, & \text{if } e_t = 1,\end{array}\right.$$

$$\Pr_{\mathrm{err},q}\Big\{ E = \mathbf{e}_q \Big\} = \prod_{t=1}^{n} \begin{cases} 1-\varepsilon_q, & \text{if } e_{t,q} = 0, \\ \varepsilon_q/(q-1), & \text{if } e_{t,q} \in \{1,\ldots,q-1\}, \end{cases}$$

where ε and ε_q are given.

Let us estimate the decoding error probability at the output of the JW decoders. One can easily see that if the decoder tries to find a key codeword at distance at most $\lfloor (d_{\mathcal{C}}-1)/2 \rfloor$ from the received vector $\mathbf{y}$ and outputs an error when it is not possible, then the probability of correct decoding is expressed as

$$\hat{\Lambda}_{\mathrm{err}}(\varepsilon) = \sum_{\nu=0}^{\lfloor (d_{\mathcal{C}}-1)/2 \rfloor} \binom{n}{\nu} (1-\varepsilon)^{n-\nu} \varepsilon^{\nu}.$$

The decoding at the 2-nd step can be organized as a procedure that depends on the results of the 1-st step. Namely, the decoder can replace symbols of the vector $\mathbf{y}_q$ located at positions where the vector $\hat{\mathbf{e}} = \mathbf{y} \oplus \hat{\mathbf{x}}$ contains 1's with erasures and decode the resulting vector $\hat{\mathbf{y}}_q$. One can easily see that an estimate of the probability of correct decoding can be expressed as

$$\hat{\Lambda}^{*}_{\mathrm{err}}(\varepsilon, \varepsilon_q) = \sum_{\nu=0}^{\lfloor (d_{\mathcal{C}}-1)/2 \rfloor} \binom{n}{\nu} (1-\varepsilon)^{n-\nu} \varepsilon^{\nu} \hat{\Lambda}_{\mathrm{err},q}(\varepsilon_q | \mathrm{wt}(\hat{\mathbf{e}})),$$

where

$$\hat{\Lambda}_{\mathrm{err},q}(\varepsilon_q | \mathrm{wt}(\hat{\mathbf{e}})) \stackrel{\triangle}{=} \sum_{\tau=0}^{\lfloor (d_{\mathcal{C}_q} - \mathrm{wt}(\hat{\mathbf{e}})-1)/2 \rfloor} \binom{n-\mathrm{wt}(\hat{\mathbf{e}})}{\tau} (1-\varepsilon_q)^{n-\mathrm{wt}(\hat{\mathbf{e}})-\tau} \varepsilon_q^{\tau}$$

is the estimate of the probability of correct conditional decoding at the 2-nd step. Some numerical results are given in Table 3.

Table 3. Estimates of the decoding error probability for $n = 28$ and $d_{\mathcal{C}_q} = 5$

	$1-\hat{\Lambda}_{\mathrm{err}}(\varepsilon)$			$1-\hat{\Lambda}^{*}_{\mathrm{err}}(\varepsilon, \varepsilon_q = .001)$		
ε	$d_{\mathcal{C}}=5$	$d_{\mathcal{C}}=7$	$d_{\mathcal{C}}=9$	$d_{\mathcal{C}}=5$	$d_{\mathcal{C}}=7$	$d_{\mathcal{C}}=9$
.001	3.2e-06	2.0e-08	9.6e-11	1.6e-05	1.3e-05	1.3e-05
.002	2.5e-05	3.2e-07	3.0e-09	4.7e-05	2.3e-05	2.2e-05
.003	**8.4e-05**	**1.6e-06**	**2.3e-08**	1.1e-04	3.4e-05	3.3e-05
.004	1.9e-04	4.9e-06	9.3e-08	2.3e-04	4.9e-05	4.4e-05
.005	**3.7e-04**	**1.2e-05**	**2.8e-07**	4.2e-04	6.8e-05	5.7e-05

Considerations presented in [4] show that the performance of the verifier, who analyzes transmitted key codeword both under the biometric and the observation noise, corresponds to the performance of the JW decoder for the channel having crossover probability $\varepsilon' = 2\varepsilon/3$, i.e., $\hat{\Lambda}_{\mathrm{bio,err}}(\mathbf{p}, \varepsilon) = \hat{\Lambda}_{\mathrm{err}}(2\varepsilon/3)$. The value of parameter ε that can be of interest for practical systems is $\varepsilon = 0.005$, and the corresponding values of the decoding error probabilities are given in Table 3 in bold font.

We can also prove the following upper bound on the probability of correct decoding by the attacker,

$$\hat{\Lambda}_{\text{bio}}(\mathbf{p}) \leq \frac{2^n}{M} \cdot \frac{q^n}{M_q} \max_{\mathbf{s}} \Pr_{\text{DNA}} \Big\{ S = \mathbf{s} \Big\}.$$

In particular, if $\mathcal{C}$ is the code for $M = 2^{18}$ messages having the minimum disance 5 and $\mathcal{C}_8$ is the Reed–Solomon code over $GF(2^8)$ for $M_8 = (2^8)^{24}$ messages having the minimum distance 5, then $\hat{\Lambda}_{\text{bio}}(\mathbf{p})$ is equal to $2^{-18}2^{-8(28-24)}2^{-76.8} = 2^{-34.8}$.

A more detailed discussion of the implementation issues will be presented in another paper.

4 Conclusion

Additive block coding schemes can bring efficient solutions to biometric problems when the length of the auxiliary key codewords is the same as the length of biometric vectors and there is an external randomness measured by the number of possible key codewords. This approach is especially effective for correcting the drop–in/out errors in the DNA measurements.

References

[1] Bolle, R.M., Connell, J.H., Pankanti, S., Ratha, N.K., Senior, A.W.: Guide to Biometrics. Springer, NY (2004)

[2] Juels, A., Wattenberg, M.: A fuzzy commitment scheme. In: Proc. 6th ACM Conf. on Computer and Communication Security, pp. 28–36 (1999)

[3] Korte, U., Krawczak, M., Merkle, J., Plaga, R., Niesing, M., Tiemann, C., Han Vinck, A.J., Martini, U.: A cryptographic biometric authentication system based on genetic fingerprints. In: Proc. Sicherheit 2008, Saarbrucken, Germany, pp. 263–276 (2008)

[4] Balakirsky, V.B., Ghazaryan, A.R., Han Vinck, A.J.: Performance of additive block coding schemes oriented to biometric authentication. In: Proc. 29th Symp. on Information Theory in the Benelux, Leuven, Belgium (2008)