A cryptographic biometric authentication system based on genetic fingerprints

Ulrike Korte [*] Michael Krawczak [†] Johannes Merkle [‡] Rainer Plaga [*]
Matthias Niesing [‡] Carsten Tiemann [§] Han Vinck [¶] Ullrich Martini [||]

Abstract: We specify a system for authentication and key derivation using genetic fingerprints which prevents the recovery of biometric information from data stored for verification. We present a detailed security analysis based on estimates of the entropy of the DNA data and formal security results. The scheme is shown to be robust and efficient by analysing the typical frequency and structure of errors in DNA measurements and selecting appropriate error correcting codes. As a result we obtain an authentication system that offers a security level equivalent to cryptographic keys with 73 bits and a FRR well below 1%.

1 Introduction

Biometric authentication systems face various risks. In [12], Jain, Nandakumar and Nagar provide a systematic and thorough analysis of the vulnerabilities due to intrinsic failures and potential attacks by adversaries. One of the most serious threats is compromising of the templates database. An attacker with access to a reference template could try to impersonate a legitimate user by reconstructing the biometric trait and by creating a physical spoof [1, 4, 9]. Therefore, compromising the database can have disastrous impact on the whole authentication system. Furthermore, the potential disclosure of digitally stored biometric data raises serious concerns about privacy and data protection [10]. However, the effectiveness of access control mechanisms is inherently limited, e.g. against internal attacks or in the presence of software vulnerabilities.

The inherent fuzziness of biometric measurements rules out a simple verification of the presented template against a reference value computed with a one-way hash function, as is common for password-based authentications. In order to overcome these limitations, several schemes have been proposed which deploy error correcting codes to provide error-tolerant biometric authentication while, at the same time, preventing the recovery of biometric information from the reference data stored for verification by cryptographic

[*] Bundesamt für Sicherheit in der Informationstechnik (BSI), Godesberger Allee 185 - 189, Bonn, Germany
[†] Institute of Medical Informatics and Statistics, University of Kiel, Kiel, Germany
[‡] secunet Security Networks AG, Kronprinzenstrae 30, Essen, Germany
[§] Labor Dr. Krone & Partner, Bad Salzuflen, Germany
[¶] Institute for Experimental Mathematics, University of Duisburg-Essen, Germany
[||] Giesecke & Devrient GmbH, Munich, Germany

techniques. These properties minimise the risk of unauthorised access to biometric templates and thereby render biometric authentication more secure and privacy friendly. Most of the schemes also use the biometric data to conceal or derive a cryptographic key[1] and are therefore also referred to as *biometric cryptosystem* or *biometric encryption*.

However, even if the stored reference data have been hashed with a one-way function, they enable an attacker to launch an exhaustive search by systematically executing the verification algorithm on a large set of candidate templates. In order to render such an approach infeasible, the biometric templates must contain sufficient entropy. Furthermore, the template must have enough entropy to ensure the secrecy and unpredictability of the concealed or derived cryptographic key.

After a detailed discussion of these aspects, Plaga [16] concluded that single conventional biometric features that are nowadays in common use do not offer the amount of entropy required for biometric cryptosystems. On the other hand, the human DNA represents a potential biometric feature which contains several megabyte of discriminatory information and hence could leverage the implementation of a secure biometric cryptosystem. Therefore, we decided to base our design on DNA data. Although genetic fingerprints do not allow real-time authentication with current technology, there are many other application scenarios where such a system could be useful. In fact, identification by genetic fingerprints is already routinely deployed for forensic and criminalistic purposes.

Our system is based on a scheme of Juels and Wattenberg [14]. By generalizing their security analysis to the case of a non-uniform distribution of the templates, we will show that the reference data stored for verification does not reveal the template. Furthermore, we will show that our system is reliable and secure against potential masquerades by estimating the false rejection rate and the false acceptance rate.

2 Previous Work

Many proposals have been made for biometric authentication in which the need to store templates for verification has been eliminated by combining error correction techniques and cryptographic methods.

In [14], Juels and Wattenberg published a *fuzzy commitment scheme* which can be used as a basis for biometric cryptosystems. Their construction was based on the Hamming metric (as shown in [6], it is even optimal for that metric) and is thus appropriate for any biometrics where the impact of measurement errors on the template is regional. Juels and Wattenberg provided a strict security analysis only for the case that the biometric templates are uniformly distributed, but in section 3.2 we will generalise their result to arbitrary distributions. Our authentication system is based on their construction.

A different approach has been taken by Juels and Sudan in [13]. Their construction was based on secret sharing techniques, making it tolerant against any reordering as well as limited deletions and insertions of the substrings defining the template. This property

[1] Accordingly, in [12] the schemes are classified as *key binding* or *key generation* systems.

makes the scheme particularly interesting for biometric features like fingerprints or iris scans where such errors are commonly induced by the measurement. In [6] Dodis, Reyzin and Smith proposed an optimisation of the scheme of Juels and Sudan and provided a stringent security analysis of both the original scheme and the optimisation. Furthermore, a general model for biometric authentication with key derivation was defined that covers many of the published biometric cryptosystems.

In [5] and [8], specific systems were proposed for iris scans. The construction of [8] was based on the scheme of Juels and Wattenberg and also addresses general limitations of biometric key derivation (in particular the irrevocability and limited secrecy of biometric information) by introducing hardware tokens as a third authentication factor. Both publications estimated the security of the scheme based on estimates of the entropy of iris codes. However, the analysis of the entropy achievable with current biometric measurement techniques carried out in [16] indicated that the estimates of [5] and [8] were far too optimistic.

Bohannon et al. ([2]) addressed various cryptographic approaches to privacy of forensic DNA databases and provided a “solution framework” for this problem, strongly related to the present one. In particular, we will design and analyze the biological and cryptographic features of one of their suggested solutions in detail.

In [11], Itakura et al. discuss the applicability of DNA data for identification systems. They empirically verify that the genotypes at different loci are uncorrelated and give estimates for matching probabilites between distinct persons. Finally, they specify a cryptosystem using keys constructed from the genetic fingerprints. However, they neglect correlations between the maternal and the paternal part of the DNA (see section 4.3.2 for a discussion) and do not consider potential measurement errors and error correction.

3 General Cryptographic Scheme

This section reviews how a biometric template f is cryptographically protected in the scheme of Juels and Wattenberg (subsection 3.1) and then quantifies its security in general (subsection 3.2).

3.1 The Scheme of Juels and Wattenberg

In [14], Juels and Wattenberg proposed a very simple biometric encryption scheme based on any binary (not necessarily linear) error correcting code. In the following, we will consider the generalisation of the scheme to symbol based codes, i.e. codes over arbitrary finite fields. The elements of the finite field are referred to as *symbols*.

Let $\mathscr{C} \subset \mathbf{F}_q^n$ be an $[n,k,2t+1]$ error correcting code with encoding function $G : \mathbf{F}_q^k \leftarrow \mathscr{C}$, where $\mathbf{F}_q$ denotes the finite field with q elements. The encoding function transforms messages consisting of k symbols into n symbol code words ($n > k$), that can be retransformed into the messages even if up to t symbols of the received codeword are corrupted by errors. During enrolment, a secret key $s \in \mathbf{F}_q^k$ is randomly selected and

$$y = G(s) - f \tag{1}$$

is stored in the database. Here f is the biometric template that is obtained during enrolment. Furthermore, the secret key s is hashed with a cryptographic hash function h and $h(s)$ is stored in a database. For authentication, the template $\tilde{f}$ presented by a user is added to the value y stored in the database. The result is $G(s) + \tilde{f} - f$ and if the hamming distance between f and $\tilde{f}$ is at most t, $G(s)$ and hence s can be recovered. If the hash value of the recovered s matches the one stored in the database, the user is authenticated.

3.2 Security Analysis

For our security analysis we make use of the following notations. Let Pr(X) denote the probability of an event X and let $\mathbf{E}_{a \leftarrow A}[f(a)]$ be the expectation of the function value of a random variable A. The *min-entropy* of a random variable A is given by

$$\mathbf{H}_\infty(A) := -\log_2(\max_a(\Pr(A = a))),$$

and the *average min-entropy* of A given B is defined as

$$\widetilde{\mathbf{H}}_\infty(A \,|\, B) = -\log_2\left(\mathbf{E}_{b \leftarrow B}\left[2^{-\mathbf{H}_\infty(A \,|\, B=b)}\right]\right) = -\log_2\left(\mathbf{E}_{b \leftarrow B}\left[\max_a\left(\Pr\left(A = a \,|\, B = b\right)\right)\right]\right)$$

We use the term *B reveals u bits of A* to indicate that $u = \mathbf{H}_\infty(A) - \widetilde{\mathbf{H}}_\infty(A \,|\, B)$.

Now let S, F and Y denote the random variables for s, f and y, respectively. The distribution F of the templates refers to any fixed population and can be arbitrary, i.e. we do not assume a uniform distribution of the templates. S is uniformly distributed and the distribution of Y is induced by those of S and F.

We assume that the hash value $h(s)$ does not reveal any information about the secret key s. Consequently, we restrict our analysis to attackers who, in order to compute the template f or the secret key s of a user, take as input only the corresponding value y defined by equation (1). The following result limits the success probability of such attackers.

Theorem 1. *Any algorithm that takes as input a random output y of the scheme and tries to output the corresponding s has at most an average success probability of* $2^{-\widetilde{\mathbf{H}}_\infty(S \,|\, Y)}$. *Any algorithm that takes as input a random output y of the scheme and tries to output the corresponding f has at most an average success probability of* $2^{-\widetilde{\mathbf{H}}_\infty(F \,|\, Y)}$.

For fixed y the probability that the output of the algorithm equals the secret key is at most $\max_s(\Pr(S = s \,|\, Y = y)) = 2^{-\mathbf{H}_\infty(S \,|\, Y=y)}$. Taking expectations on both sides yields the first statement. The second statement follows analogously.

The following result shows that it is equally difficult to determine s from y as it is to determine f from y.

Lemma 2. $\widetilde{\mathbf{H}}_\infty(S\,|\,Y) = \widetilde{\mathbf{H}}_\infty(F\,|\,Y)$.

For fixed y, the key s is uniquely determined by f and vice versa. Therefore, for a given y the most likely s corresponds to a unique, and hence equally likely f, and thus $\max_s(\Pr(S = s\,|\,Y = y)) = \max_f(\Pr(F = f\,|\,Y = y))$. Taking expectations over y on both sides yields the claim.

We now turn to the security of the authentication. For biometric authentications this is usually measured by the *False Acceptance Rate (FAR)*, which can theoretically be modelled as the probability that unauthorised persons are accepted as authorised, i.e. are authenticated as a legitimate user. The probability is taken over a random choice of the enrolled users and the impostor from the considered population.

For the following result, we assume that the hash function maps all $s \in \mathbf{F}_q^k$ to distinct hash values, and consequently, that the authentication is only successful if the correct secret key is recovered.[2]

Theorem 3. *The FAR, i.e. the probability that a random impostor is accepted as one of m randomly selected users, is limited by $m2^{-\widetilde{\mathbf{H}}_\infty(F\,|\,Y)}$.*

For fixed y let $R_y(f)$ be the unique secret key s that is recovered during authentication from template f, e.g. the unique s with $|y + f - G(s)| \leq t$.[3] In case there is no such s, i.e. if the distance of $y + f$ to the next code word is greater than t, let $R_y(f) = \emptyset$.

By assumption, an impostor is authenticated as a user U_i enrolled with (f_i, s_i, y_i), if and only if his own template f satisfies $R_{y_i}(f) = s_i$. Thus, for given y_i the probability that a random impostor is accepted as user U_i is given by $Pr(S = R_{y_i}(f)\,|\,Y = y_i)$, which is at most $2^{-\mathbf{H}_\infty(S\,|\,Y=y_i)}$. Consequently, for a given set of users $U_1, \ldots, U_m$ the probability that a random impostor is accepted as one of the users is limited by $\sum_{i=1}^m 2^{-\mathbf{H}_\infty(S\,|\,Y=y_i)}$. Now, the result is obtained by taking expectations over the y_i (i.e. over the random selection and enrollment of the users) and applying Lemma 2.

Our analysis has shown that the security of the scheme can be measured by $\widetilde{\mathbf{H}}_\infty(F\,|\,Y)$, which can be determined from the entropy of the template and the number of bits of f revealed by y. In section 4.3 we will estimate the entropy of templates derived from genetic fingerprints. Subsequently, we will analyse the number of bits of f revealed by y.

In [14], Juels and Wattenberg show that if the templates are uniformly distributed in ${\mathbf{F}_2}^n$, y reveals only $n - k$ bits of information about f and no information about s. However, biometric templates are usually not uniformly distributed. In section 5.3 of [14] Juels and Wattenberg argue that “a good security analysis” of the scheme for a non-uniform distribution $\mathscr{D}$ “will, in general, require detailed knowledge of $\mathscr{D}$”. Fortunately, this presumption is not true: The following theorem generalises the result of Juels and Wattenberg by giving

[2] This assumption can be justified by selecting a hash function with an output length considerably greater than the bit length of the secret keys.

[3] The uniqueness follows from the precondition that G is the encoding function of a $[n, k, 2t + 1]$ error correcting code.

an upper bound for the number of bits of s revealed by y for arbitrary distributions of the templates.[4]

We consider the generalised scheme over an arbitrary finite field $\mathbf{F}_q$. The result for the original scheme is implied by the case $q = 2$.

Theorem 4. *For any distribution F of the templates, at most $(n-k)\log_2 q$ bits of f are revealed by y, i.e. $\widetilde{\mathbf{H}}_\infty(F\,|\,Y) \geq \mathbf{H}_\infty(F) + (k-n)\log_2 q$.*

By definition, we have

$$2^{-\widetilde{\mathbf{H}}_\infty(S|Y)} = \sum_{y\in\mathbf{F}_q^n} \Pr(Y=y)\cdot\max_s\left(\Pr(S=s\,|\,Y=y)\right). \tag{2}$$

Using Bayes' theorem and equation (1) we obtain

$$\begin{aligned}\Pr(Y=y)\cdot\Pr(S=s\,|\,Y=y) &= \Pr(Y=y\,|\,S=s)\cdot\Pr(S=s)\\ &= q^{-k}\Pr(Y=y\,|\,S=s)\\ &= q^{-k}\Pr(F=G(s)-y\,|\,S=s).\end{aligned} \tag{3}$$

The distributions of F and S are statistically independent. Therefore, we can omit the condition $S=s$ on the right hand side of equation (3). Consequently, we get

$$\begin{aligned}\max_s\left(\Pr(Y=y)\cdot\Pr(S=s\,|\,Y=y)\right) &= q^{-k}\max_s\left(\Pr(F=G(s)-y)\right)\\ &\leq q^{-k}\max_z\left(\Pr(F=z)\right)\\ &= q^{-k}2^{-\mathbf{H}_\infty(F)}.\end{aligned} \tag{4}$$

Equations (2) and (4) yield $2^{-\widetilde{\mathbf{H}}_\infty(S|Y)} \leq \sum_{y\in\mathbf{F}_q^n} q^{-k}2^{-\mathbf{H}_\infty(F)} = q^{n-k}2^{-\mathbf{H}_\infty(F)}$, and with Theorem 2 we obtain the desired result.

4 DNA as Biometric Feature

In this section, we summarise some basic properties of short tandem repeats (subsection 4.1), analyse the reliability of their measurement (subsection 4.2) and estimate the information content of the resulting data (subsection 4.3).

4.1 Short Tandem Repeats

The most common DNA variations used for the identification of individuals, in particular in forensic applications, are *Short Tandem Repeats (STR)* – arrays of 5 to 50 copies

[4] In [6], a much weaker result has ben shown for a variant of the scheme.

(repeats) of the same pattern (the *motif*) of 2 to 6 base pairs. Two properties make STR particularly eligible for identification purposes:

- The number of repeats of the motif is highly variable among individuals, even in small populations.
- Forensic STR are typically located in the non-coding regions of the DNA and (at the time of writing) no biological functions of these STR loci are known.[5]

The human genome contains several 100,000 STR loci, i.e. physical positions in the DNA sequence where an STR is present. Today, approximately 20 STR loci are in practical forensic use,[6] and some more can be considered as candidates. In order to optimise their suitability for forensic applications these loci have been selected to maximise the variability of the genotype and to minimise the likelihood of any association between the genotype and a biological function, or between the genotypes of different loci.

As with any other DNA polymorphism, an individual variant of an STR is called *allele*. Alleles are denoted by the number of repeats of the motif. In some cases, one or more motifs may not be complete within the STR, in which case the allele is denoted by a decimal number, where the digit after the decimal point equals the number of base pairs modulo the length of a complete repeat. For instance, if the motif is `AGT` the allele `AGTGTAGT` is denoted as 2.2.

The genotype of a locus comprises both the maternal and the paternal allele. If these two alleles are identical the genotype is called *homozygous*, and if they are different it is called *heterozygous*. However, without additional information, one cannot determine which allele resides on the paternal or the maternal chromosome, i.e. allele combinations (A,B) and (B,A) are indistinguishable. Therefore, genotypes are denoted by the allele numbers in ascending order, i.e. as (A,B) with $A \leq B$.[7] The range of possible genotypes differs from one STR locus to another. Typically, there are 10-20 different alleles known per locus. However, these alleles are not of equal frequency in a given population. In section 4.3 we give estimations of the entropy per locus.

For a given set of loci, the combined genotypes at these loci are called an *STR profile*. If the set of loci is large enough to allow reliable distinction between individuals, the profile is also referred to as a *genetic fingerprint*.

The measurement of an STR profile is conducted by specialized laboratories using commercially available STR kits. The details of this procedure are described in [3], p. 313 onward. At the time of writing, an STR profile can be determined in approximately 4 hours. Since this time is sufficient for typical forensic applications, manufacturers of STR kits have presumably made only limited efforts for further optimization.

[5] However, correlations between STR genotype and ethnical, regional or familial affiliation exist.

[6] Of these, 13 are the CODIS (Combined DNA Index System) core loci, which are the basis of the forensic system used by the FBI.

[7] A homozygous genotype is sometimes denoted by the single allele, i.e. as (A).

4.2 Reliability of STR measurements

Like any measurement, DNA analysis is not free of errors. For details on typical sources for errors in STR measurements we refer to [7].

STR measurement errors are commonly classified into three groups:

1. *Allelic drop-outs.* An allele of a heterozygous genotype is missing, e.g. genotype $(7,9)$ is measured as $(7,7)$.
2. *Allelic drop-in.* In a homozygous genotype, an additional allele is erroneously included, e.g. genotype $(10,10)$ is measured as $(10,12)$.
3. *Allelic shift.* An allele is measured with a wrong repeat number, e.g. genotype $(10,12)$ is measured as $(10,12.2)$.

The German DNA Profiling Group (GEDNAP) regularly performes ring experiments to assess the quality of laboratories performing forensic STR analysis. A central laboratory prepares several sets of identical samples. These samples are sent to all participating laboratories, who send back their results for all or a subset of samples at their discretion. The organising laboratory verifies the correctness of the results by comparison to their own, carefully conducted measurements. We refer to [17] for details.

In 2005, GEDNAP 30 and 31 were conducted with more than 175 laboratories, reporting approximately 19,000 measurements. An analysis of the (anonymised) data provided to us by GEDNAP showed that few laboratories account for most of the errors, whereas the vast majority of the laboratories have low error rates.[8] Assuming a minimum quality standard for biometric measurements, we evaluated the error rates with the results of the four worst laboratories disregarded. Table 1 summarises the results for each type of error.[9] For allelic shifts we distinguished between homozygous and heterozygous genotypes; this distinction is useful for the selection of an eligible encoding of the template (see section 5.1).

Table 1: Errors reported in GEDNAP 30 and 31

Type of Error	**Number**	**Frequency**
Allelic drop-ins	7	0.04%
Allelic drop-outs	8	0.04%
Allelic shifts (het.)	19	0.10%
Allelic shifts (hom.)	2	0.01%

The results in Table 1 show that error rates below 0.2% are achieved when a minimum quality standard for the laboratory is assumed. Furthermore, none of the good laboratories measured more than one locus incorrectly, which indicates that measurement errors occur independently for different loci.

[8] GEDNAP members explain the bad performance of some laboratories by their poor equipment and the lack of proper quality management.

[9] The classification of errors is not always unambiguous. However, the error rates based on different interpretations differ only marginally.

4.3 Entropy of STR Data

As shown in section 3.2, the security of the authentication depends on the entropy of the templates. We now estimate the entropy of an STR profile based on the 28 loci which include all loci used in the GEDNAP 30 and 31 ring experiments ([17]).

4.3.1 Assumed Probability Distribution

The entropy of a biometric template is generally limited by the size of the population from which the enrolled individuals are chosen. For instance, for the German population the entropy could not exceed $\log_2(8 \cdot 10^7) \approx 26$. In terms of security, this bound corresponds to a brute force attack that exhaustively searches through the STR profiles of all individuals within the population. For sufficiently large populations however, a comprehensive list of templates is usually not available and consequently, a brute force attack on a large scale biometric application has to be based on a much wider search space. In particular, the relevant search space and probability distribution of the templates is implied by the statistical data available on biometric sub-features. In other words, we must consider the distribution of STR profiles that can be extrapolated from statistical data under the assumption that the population is infinite.[10]

As pointed out by Bohannon et. al. [2], an analysis based on this idealised probability distribution may not be adequate in the presence of a large scale (e.g. nation wide) DNA database. Furthermore, the potential search space for attackers being related by blood to enrolled users is considerably smaller.

4.3.2 Entropy of a single locus

The probability of the occurrence of an allele in a given population is estimated from the corresponding allele frequency, as observed in scientific experiments. For many populations, allele frequencies are readily available from the literature.

In order to estimate the genotype probabilities for a given locus from individual allele probabilities, we assume that the population in question is in so-called *Hardy-Weinberg equilibrium*. This assumption is fulfilled if the population is sufficiently large and panmictic, i.e. the mating behaviour is random, and if there is no migration or selection (see [3], p. 484, or [15], p. 65–67). While the assumption of the Hardy-Weinberg equilibrium is clearly an idealisation, it is widely accepted as a method to obtain good approximations for DNA profile distributions for the large populations of industrialised countries.

If a population is in Hardy-Weinberg equilibrium, the frequency of the homozygous genotype (A,A) is ${P_A}^2$, and the frequency of the heterozygous genotype (A,B) is $2P_AP_B$, where P_A and P_B are the frequencies of alleles A and B, respectively. We refer to [3] or [15] for

[10] This approach is also used for assessing the assurance of evidences based on genetic fingerprints, e.g. in court cases.

details. Thus, the min-entropy of the genotype G at a locus is given by

$$\mathbf{H}_\infty(G) = -\log_2\left(\max\left(\max_A(P_A{}^2), \max_{A,B}(2P_AP_B)\right)\right).$$

4.3.3 Entropy of a Set of Loci

The frequencies of the compound genotypes at all loci used in a profile can be calculated from the genotype frequencies of the individual loci under the assumption that the genotype distributions of distinct loci are statistically independent. In this case, the frequency of a compound genotype is the product of the genotype frequencies at the individual loci (multiplication rule) and, consequently, the entropy of an STR profile P is the sum of the entropies of the contributing single loci genotypes G_l. Consequently, we obtain

$$\mathbf{H}_\infty(P) = \sum_l \mathbf{H}_\infty(G_l),$$

where indices l in the sums refer to the contributing loci.

The assumption of statistical independence of the genotype frequencies at different loci is an idealisation, the applicability of which depends on a sufficiently high level of homogeneity of the relevant population, and on an appropriate choice of the loci. The assumption of sufficient homogeneity is widely considered valid for populations of large western countries (see [15], pp. 78 for a discussion). The second condition is fulfilled for the loci commonly used in forensics because a main objective for their establishment was the minimisation of potential dependencies.

4.3.4 Experimental Estimation of the Entropy

In [7], we have estimated the entropy for the 18 loci used by GEDNAP 30 and 31 and 10 additional loci, for which sufficient statistical data is available. In order to match the population distribution in a potential application we have based these estimations on allele frequency data from German and Austrian populations, respectively, which have been obtained from a database of the Institute of Forensic Medicine at the University of Düsseldorf[11]. Our estimations yield a min-entropy of 85 for an STR profile based on these 28 loci.

5 “BioKey-STR”: An Authentication System Based on STR Data

We specify a biometric encryption scheme based on the scheme of Jules and Wattenberg (section 3.1) using templates obtained from Short Tandem Repeats (STR) in human DNA and analyse its properties on the basis of our previous results. Justifications for our design decisions are given in [7].

[11] www.uni-duesseldorf.de/WWW/MedFak/Serology/dna.html

5.1 Encoding and Error Correction

We construct the templates from STR profiles comprising the 28 loci for which we have estimated the entropy (section 4.3.4). In order to minimise the impact of typical measurement errors in DNA fingerprinting as analysed in section 4.2, we choose the following encoding:

- Each allele is uniquely encoded with 6 bits.[12] The actual coding of alleles into integers can be done by any numbering of the known alleles.
- A genotype (A,B) is encoded by $a||b$, where a and b are the encodings of A and B, respectively, and $||$ denotes concatenation. In particular, a homozygous genotype (A,A) is encoded as $a||a$.
- The encodings of the 28 loci are concatenated using a fixed order of the loci.

With 28 loci and 12 bits per locus, we obtain templates with 336 bits.

For the error correction we decided in favour of a [56, 54, 2]-Reed-Solomon over $\mathbf{F}_{2^6}$ which encodes words with 54 symbols of 6 bits each, introduces 2 symbols (i.e. 12 bits) redundancy and can correct one symbol error.

5.2 Security

As stated in section 4.3.4, the min-entropy of our templates is approximately 85.[13] From section 3.2 we know that the scheme of Juels and Wattenberg leaks at most $(n-k)\log_2 q$ bits on the biometric template f and $n - \mathbf{H}_\infty(F)$ bits of information on the secret key s. Using Theorem 4 and $q = 64$, $n = 56$ and $k = 54$ (section 5.1) we can estimate $\widetilde{\mathbf{H}}_\infty(F\,|\,Y) \geq 73$. As has been shown in Theorem 1 and Lemma 2 this implies that an attacker who tries to determine the secret key s or the template f from the reference data y has at most a success probability of 2^{-73}. Using Theorem 3, we obtain $FAR \leq m2^{-73}$, where m is the number of enrolled individuals.

5.3 False Rejection Rate (FRR)

Our determination of the FRR is based on the error rates of STR measurements as determined empirically in section 4.2. Since this database is rather small, we can only obtain a rough estimate for the FRR.[14] As suggested by the results of the GEDNAP ring experiments, we assume that measurement errors occur independently for different loci.

[12] For the 28 loci considered in section 4.3 the cardinality of the domains of allele numbers is limited by 64.

[13] As pointed out in section 4.3.1, this estimation is based on an idealized approach which neglects blood relationships among potential users and attackers.

[14] As the FRR is not a measure of security but only of user comfort, this situation is still satisfactory.

In our setting, the BioKey system can correct up to one symbol error. With our basic encoding (see section 5.1) most measurement errors only result in one wrong symbol (*single error*); the only exception is an allelic shift in the measurement of a homozygous genotype which results in two wrong symbols (*double error*). Therefore, we can estimate $\mathrm{FRR} = p_\mathrm{d} + p_\mathrm{s}$, where p_d is the probability that at least one double error occurs in 28 independent measurements and p_s is the probability that more than one single error occurs.

From Table 1 we can infer that double errors only occurred twice in approximately 19,000 experiments and hence we estimate $p_\mathrm{d} \approx 28 \cdot 2/19{,}000 \approx 0.003$. On the other hand, $p_\mathrm{s} = 1-(1-p)^{28} - 28p(1-p)^{27}$, where p is the probability of a single error in a single locus measurement. From Table 1 we conclude that $p \approx 34/19{,}000$ for good laboratories. This yields $p_\mathrm{s} \approx 0{,}001$ and hence $\mathrm{FRR} \approx 0.4\%$.

6 Summary and Conclusion

In this paper, we have examined the feasibilty of a biometric authentication system based on genetic fingerprints that does not store genetic data in clear. The system allows the authentication of users, based on information encoded in their genome, and prevents a disclosure of the biometric information from the reference data stored in databases.

Our system is based on the biometric cryptosystem of Juels and Wattenberg. We have been able to generalize the security analysis of this scheme to the case of non-uniformly distributed templates, and to prove an upper bound of the FAR.

Our system uses a genetic fingerprint template obtained from short tandem repeats (STRs) on 28 loci in the DNA with an entropy of about 85 bits. The error rate of the measurement of the genetic information has been determined experimentally using the data of existing quality experiments for genetic labs. Error correcting codes have been applied to effectively reduce the false rejection rate to the low value of 0.4%. It has been found that approximately 70 bits of discriminatory information can be extracted using this method. This amount of information makes an exhaustive search through the template space infeasible and provides a very low FAR. By adding more suitable loci the security of the system could even be increased.

Further research activities, as for example regarding performance aspects and the unpredictability of allele combinations, are necessary and actually performed by the forensic community. Furthermore, an analysis of potential threats arising from correlations of templates among relatives could be useful.

Comparing our results with previous publications concerning schemes based on other biometric information like iris images, fingerprints, etc., our proposed system shows a significant improvement of the security due to the higher entropy of the DNA data compared to the limited information content of other biometric features.

Acknowledgements

We are grateful to Carsten Hohoff for his consultancy in interpreting the results of the GEDNAP ring experiments which has been crucial for the success of this work.

References

[1] Adler, A.: Sample images can be independently restored from face recognition templates, Canadian Conference on Electrical and Computer Engineering (CCECE), Montral, Canada, 2003. pp. 1163-1166, 2003.

[2] Bohannon, P., Jakobson, M. and Srikwan, S.: Cryptographic Approaches to Privacy in Forensic DNA Databases. In Imai, H., Zheng, Y (eds.) Public Key Cryptography 2000, Melbourne, Australia, pages 373-390, LNCS 1751, Springer-Verlag, 2000.

[3] Butler, J.M.: Forensic DNA typing. Elsevier, 2005.

[4] Cappelli, R., Lumini, A., Maio, D. and Maltoni, D.: Fingerprint Image Reconstruction from Standard Templates. Transactions on Pattern Analysis and Machine Intelligence, Volume 29, Issue 9, pp. 1489-1503, 2007.

[5] Davida, G.I., Frankel, Y. and Matt, B.J.: On enabling secure applications through offline biometric identification. 1998 IEEE Symposium on Security and Privacy, Oakland, California, USA, pp. 148–157, 1998.

[6] Dodis, Y., Reyzin, L. and Smith, A.: Fuzzy Extractors: How to Generate Strong Keys from Biometrics and Other Noisy Data. Cachin, C., Camenisch, J. (eds) Advances in Cryptology - EUROCRYPT 2004, Interlaken, Switzerland, pp. 523–540, LNCS 3027, Springer Verlag, 2004.

[7] Korte, U, Krawczak, M., Merkle, J., et al.: A cryptographic biometric authentication system based on genetic fingerprints (Extended Version). Available at http://www.bsi.de/biokeys.pdf, 2008.

[8] Hao, F., Anderson, R. and Daugman, J.: Combining cryptography with biometrics effectively. Technical Report UCAM-CL-TR-640, University of Cambridge, 2005.

[9] Hill, C.J.: Risk of Masquerade Arising from the Storage of Biometrics, B.S. Thesis, Australian National University, 2001.

[10] Hornung, M.: Biometric Identity Cards: Technical, Legal, and Policy Issues. Paulus, S., Pohlmann, N., Reimer, H. (eds) ISSE 2004: Securing Electronic Business Processes, pp. 47-57, 2004.

[11] Itakura, Y., Hayashida, M., Nagashima, T. and Tsujii, S.: Proposal on Personal Identifiers Generated from the STR Information of DNA, International Journal of Information Security, Vol. 1, No. 3, pp. 149-160, 2002

[12] Jain, A. K., Nandakumar and K., Nagar, A.: Biometric Template Security, EURASIP Journal on Advances in Signal Processing, 2008.

[13] Juels, A. and Sudan, M.: A Fuzzy Vault Scheme. IEEE International Symposium on Information Theory, p. 408, 2002.

[14] Juels, A. and Wattenberg, M.: A Fuzzy Commitment Scheme. Sixth ACM Conference on Computer and Communication Security, pp. 28–36, 1999.

[15] Krawczak, M. and Schmidtke, J.: DNA Fingerprinting (second edition), BIOS Scientific Publishers Ltd, 1998.

[16] Plaga, R.. Biometric keys: suitable uses and achievable information content. Submitted to International Journal of Information Security, 2006.

[17] Rand, S., Schrenkamp, M., Brinkmann, B. and Hohoff, C.: The GEDNAP blind trial concept part II. Trends and developments. International Journal of Legal Medicine 118, pp. 83-89, 2004.