Biometric DNA and ECDLP-based Personal Authentication System: A Superior Posse of Security

Ranbir Soram [†] **Memeta Khomdram** [††]

[†] Manipur Institute of Technology, Takyelpat, Imphal -795004, India.
[††] Department of Electronics Accreditation of Computer Courses Centre, Akampat, Imphal-795008, India.

Summary

Over the past few decades there has been numerous advances in the biometric DNA proving technology, most notably among them is the development of PCR-based DNA proving methods. With the time needed to analyze DNA base sequence has dropped from weeks to hours, and with more scientific advancements, the time needed to process samples may drop to as little as few minutes; network security experts have just questioned to use this novel technology to network authentication systems. In this paper we propose a method to utilize biometric DNA information and the intractability of **E**lliptic **C**urve **D**iscrete **L**ogarithm **P**roblem (ECDLP) for personal authentication in information security systems. We also present background information on DNA and the elliptic curve discrete logarithm problem, as well as the commonly applied respective mathematics.

Key words:
Authentication, Biometric DNA, Tandem Repeat, ECDLP.

1. Introduction

Sensitive information must be protected against unauthorized access. To achieve this, computer scientists have looked for new biometrics authentication systems. Authentication is the act of establishing someone as who they claim they are. An authentication system based on biometric information offers greater security and such systems are increasingly gaining widespread use and popularity.

Biometrics are technologies used for measuring and analyzing a person's unique characteristics. There are two types of biometrics: Behavioral and Physiological. Behavioral characteristics are related to the behavior of a person. Physiological characteristics are related to the shape of the body. Some examples are fingerprint, face recognition, iris recognition, DNA matching. Behavioral biometrics are generally used for authentication while physical biometrics can be used for either identification or authentication. Identification is determining who a person is.

Public key cryptosystem supports many security mechanisms such as confidentiality, integrity, authentication, and non-repudiation. However, to successfully implement these very many security mechanisms, we must carefully plan an infrastructure to manage them. In cryptography jargon, a Public Key Infrastructure (PKI) is a system that binds public keys with respective user identities by means of a certificate authority (CA). PKI supports message encryption and digital signature that further enhances transactional security [1].

A=**A**denine
T=**T**hiamine
C=**C**ytosine
G=**G**uanine

*"Adenine in one strand only attracts thiamine in the opposite strand forming an **A-T** pair and cytosine only attracts guanine forming a **C-G** pair."*

Figure 1: Structure of DNA

Most of the biometric template data such as fingerprints or iris patterns are analog in nature and their representations are vendor specific. So, it is difficult to embed the biometric information in the PKI system directly. However, the digital nature of DNA information enhances the accuracy in authentication enabling the development of DNA based personal identifiers. DNA is same in all cells of the body [2]; regardless of age and unchangeable while the person is alive. So, DNA provides the most reliable personal identification.

Manuscript received January 5, 2010
Manuscript revised January 20, 2010

2. What is DNA

The foundation of understanding Deoxyribonucleic acid (**DNA**) analysis requires an understanding of the basic components of DNA. DNA is the genetic material found in most organisms, including humans. The main role of DNA molecules is the long-term storage of information [2]. The information in DNA is stored as a code made up of four chemical bases: adenine (A), thiamine (T), cytosine (C), and guanine (G). As shown in figure 1 above, DNA bases pair up with each other, A with T and C with G, to form units called base pairs. Each base is also attached to a sugar molecule and a phosphate molecule. The order, or sequence, of these bases make individual DNA unique and determines the information available for building and maintaining an organism, similar to the way in which letters of the alphabet appear in a certain order to form words and sentences. Together, a base, sugar, and phosphate are called a nucleotide. Nucleotides are arranged in two long strands that form a spiral called a double helix [2]. Human DNA consists of about 3 billion bases, and more than 98 percent of those bases are the same in all people. Although each individual repeating unit is very small, DNA polymers can be very large molecules containing millions of nucleotides. For instance, the largest human chromosome, chromosome number 1, is approximately 220 million base pairs long. Most DNA is located in the cell nucleus but a small amount of DNA can also be found in the mitochondria. The DNA segments that carry genetic information are called genes [2].

3. What is Chromosome

In the nucleus of each cell, the DNA molecule is packaged into thread-like structures called chromosomes. Each chromosome is made up of DNA tightly coiled many times around proteins called histones that support its structure. Chromosomes are not visible in the cell's nucleus—not even under a microscope—when the cell is not dividing [3]. However, the DNA that makes up chromosomes becomes more tightly packed during cell division and is then visible under a microscope. Most of what researchers know about chromosomes is learned by observing chromosomes during cell division. Each chromosome, as given in figure 2 below, has a constriction point called the centromere, which divides the chromosome into two sections, or "arms." The short arm of the chromosome is labeled the "p arm." The long arm of the chromosome is labeled the "q arm." The location of the centromere on each chromosome gives the chromosome its characteristic shape, and can be used to help describe the location of specific genes.

Chromosomes are numbered. Chromosome number 1 is the largest chromosome; chromosome number 2 a little smaller and so on. For humans, there are consistently 23 pairs of chromosomes. Among the 23 pairs of chromosomes there is a pair called the sex chromosomes. In females, the sex-chromosome pair consists of two similar size chromosomes called X chromosomes. Males have one X and one small Y chromosome.

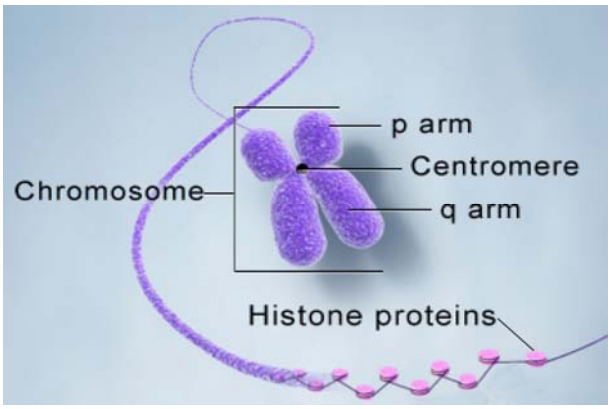

Figure 2: Structure of chromosome

4. What is Allele

An **allele(** allelomorph) is one of a series of different forms of a gene. The word was coined in the early days of genetics to describe variant forms of a gene detected as different phenotypes. Alleles may or may not result in different phenotypic traits [3]. As each chromosome has a similar chromosome partner (except for males with their X and Y chromosomes) each locus is duplicated. Loci can vary a bit. If a person has two identical versions of the locus, they are said to be homozygous. If there is a difference, they are said to be heterozygous. In any particular diploid organism, with two copies of each chromosome, the genotype for each gene comprises the pair of alleles present at that locus, which are the same in homozygotes and different in heterozygotes. A population or species of organisms typically includes multiple alleles at each locus among various individuals. Allelic variation at a locus is measurable as the number of alleles present, or the proportion of heterozygotes (heterozygosity) in the population [4].

5. Short Tandem Repeats (STR) and how are they useful

All DNA[s] have deoxyribose sugar, phosphate groups and the four bases A, T, C and G. What makes everyone different is the order and number of each base pair in their DNA. If we were to look at every single base pair in a person's DNA, we would find that no two people have exactly the same sequence. The problem is that the number of base pairs in our DNA is so huge that no DNA laboratory in the world can test your entire DNA. It would take too long to look at the entire DNA.

Instead of looking at all the DNA when a DNA fingerprint is made, all labs pin-point various 'regions' (called loci in genetics jargon), which are certain special areas of DNA. These areas are believed to be parts of the DNA that do not code for any genes, but can be very different in different people [2]. In this non-coding loci there are identical repeats of the same pattern (the motif) of 2 to 6 base pairs of DNA, which can be repeated anywhere from 1 to 50 times in a row. These repeats, as shown in figure 3 below, are next to each other on the DNA molecule and are called Tandem Repeats or **S**hort **T**andem **R**epeats (STR) [2]. An example of tandem repeat is the sequence "**A-G-C-T**-A-G-C-T-**A-G-C-T**-A-G-C-T" in which the sequence **A-G-C-T** is repeated **4** times.

The number of short tandem repeats can vary a lot from one person to another. Internet sources say that the human genome contains several 100,000 STR loci. While crime investigation and paternity verification use several locations that contain unique repeat sequences, general-purpose, absolute personal identification for information security requires the use of a larger number of loci.

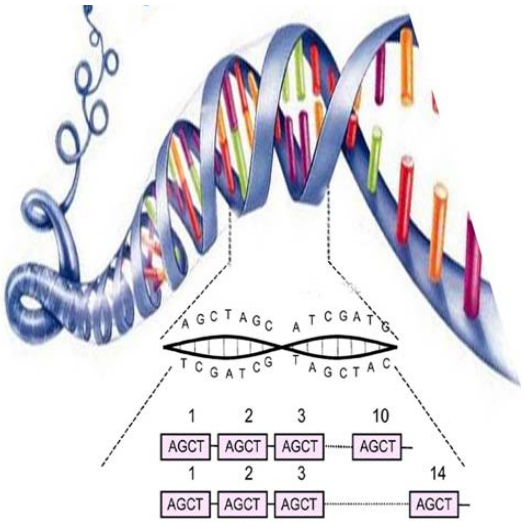

Figure 3: Tandem repeat in a locus

Today, there are more than 25 such loci that are used in DNA profiling. At each of these loci, there is a repeated sequence that is variable in length between individuals-AGCT repeated at one locus, and TTTC repeated at another. The number of repeats at each location can be measured during DNA sequencing. DNA identification is based on techniques using the non-coding tandem repetitive DNA regions [2]. Each number of repeats has statistics associated with it that can be compared to the population. The number, the pattern, and the length of these repeats are unique for each individual.

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 alleles. Without additional information, one cannot determine which allele comes from the paternal or the maternal chromosome, i.e. allele combinations (A,B) and (B,A) are indistinguishable. Therefore, we arrange the allele numbers in ascending order, i.e. as (A,B) with A≤B [5]. The range of possible genotypes differs from one STR locus to another. As shown in table 1 below, there are about 10-20 different alleles known per locus. However, these alleles are not of equal frequency in a given population. If the set of loci is large enough to allow reliable distinction between individuals, the profile is also referred to as a genetic fingerprint [4].

In the field of Computer Science, tandem repeats in DNA sequences can be efficiently detected using suffix trees or suffix arrays.

Table 1: STR loci and repeating units

STR locus	Known alleles	Commonly used repeating unit
CSF1PO	6–15	AGAT
FGA	15–30	$[TTTC]_3TTTTTTCT[CTTT]_nC$ $TCC[TTCC]_2$
TH01	3–13.3	[AATG]
TPOX	6–14	[AATG]
VWA	11–22	[AGAT]
D3S1358	9–20	[AGAT], [TCTA]
D5S818	7–15	[AGAT]
D7S820	6–14	[GATA]
D8S1179	8–18	[TATC]
D13S317	7–15	[GATA]
D16S539	5,8–15	[GATA]
D21S11	24–38	[TCTA], [TCTG]
D18S51	9–27	[GAAA]

6. How to measure STR Data

The measurement of STR data for a common set of loci is performed by using commercially available STR kits. From country to country, different DNA loci may be in use (the locus **Amelogenin** is also used for gender identification but it is not counted here). For example, the UK and some European countries use 10 core loci, Germany 8 loci, India 15 to 25 loci, the Interpol 7 loci. The US designated 13 core STR loci for the nationwide DNA database. These STR loci are given names as CSF1PO, FGA, TH01, TPOX, VWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D21S11, and D18S51. To measure the repeat count of a single locus, it is targeted with any sequence-specific primer [3] and is

amplified using Polymerase Chain Reaction (PCR). In the PCR method, a highly polymorphic regions that have short repeated sequences of DNA is used. The most common is 4 bases repeated. Developed in 1984 by Chemistry Nobel laureate Kary Mullis, PCR methods typically amplify DNA fragments up to 10 kilo base pairs (kb), although some techniques allow for amplification of fragments up to 40 kb in size. The DNA fragments that result are then separated and detected using either capillary electrophoresis (CE) or gel electrophoresis. As shown in figure 4 below, we get two counts i.e. (i,j) for a locus and each count is at most two digits. Because different unrelated people have different numbers of repeat units, these regions of DNA can be used to discriminate between unrelated individuals.

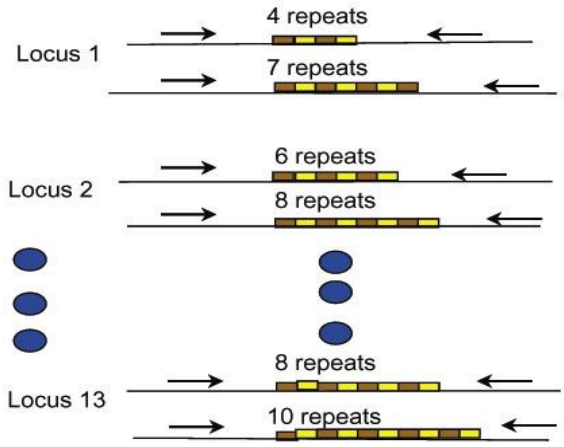

Figure 4: Repeat counts in loci

As an example we take D7S820, which is one of the 16 core genetic loci used in DNA analysis in India. This DNA is found on human chromosome 7. The DNA sequence of a representative allele of this locus is shown below. The tetrameric repeat sequence of D7S820 is "**GATA**". Different alleles of this locus have from 6 to 15 tandem repeats of the "**GATA**" sequence.

AATTTTTGTATTTTTTTTAGAGACGGGGTTTCACCA
TGTTGGTCAGGCTGACTATGGAGTTATTTTAAGGTT
AATATATATAAAGGGTATGATAGAACACTTGTCATA
GTTTAGAACGAACTAAC***GATAGATAGATAGATAGAT***
AGATAGATAGATAGATAGATAGATAGATAGACAGAT
TGATAGTTTTTTTTTATCTCACTAAATAGTCTATAGT
AAACATTTAATTACCAATATTTGGTGCAATTCTGTC
AATGAGGATAAATGTGGAATCGTTATAATTCTTAAG
AATATATATTCCCTCTGAGTTTTTGATACCTCAGATT
TTAAGGCC.

= 12 ***GATA*** repeats = ***12*** is what is reported.
= the repeat count of this allele is **12.**

In the same way, we determine the repeat count of other allele of the locus.

6.1. Errors in STR measurements

Like any measurement, DNA STR measurement is not free of errors [4,6]. Several types of error can be distinguished and some of them are as follow.

1. Allelic drop-outs:- An allele of a heterozygous genotype is missing, e.g. genotype (8,10) is measured as (8,8).
2. Allelic drop-in:- In a homozygous genotype, an additional allele is erroneously included, e.g. genotype (11, 11) is measured as (11, 13).
3. Allelic shift:- An allele is measured with a wrong repeat number, e.g. genotype (9, 12) is measured as (9, 12.2).

6.2. Encoding of STR value

An STR profile is composed by the genotypes of the individual loci, each of which is given by an ordered pair (a_i, b_i) of numbers representing the alleles on both chromosomes at this locus. An appropriate encoding of the STR profiles to a template should minimize the impact of measurement errors to the template. Based on the characterization of errors in DNA analysis given above, the selection of encoding function is decided solely by the following considerations:

1. Independent encoding for each locus: Each error in an STR measurement only affects the genotype of a single locus. Therefore, we encode each locus independently and concatenate the result using a fixed order of the loci.

2. Encoding of homozygous genotypes: In order to prevent bit insertions or deletions resulting from allelic drop-ins or drop-outs, we must encode homozygous genotypes by the same number of bits as heterozygous genotypes. There are two options to accomplish this:

I. The allele number is doubled in the encoding, i.e. a genotype (A, A) is encoded as α||α, where α is an encoding of A and || denotes concatenation. With this encoding, at least half of the bits in the encoding remain correct for allelic drop-ins and drop-outs.

II. The allele number is encoded and concatenated with a constant, e.g. genotype (A, A) is encoded as α||0. With this encoding, at least half of the bits in the encoding remain correct for allelic shifts.

Experimental analysis has shown that allelic shifts of homozygous genotypes are very rare compared to allelic drop-ins or drop-outs [4]. Therefore, we decided to use the first encoding method.

7. DNA personal ID generation method

For each locus the two repeat counts are encoded and arranged in ascending order. That is, $L_i=C_{i,1}C_{i,2}$ where $C_{i,1}$ and $C_{i,2}$ are the two repeat counts of locus L_i and $C_{i,1}\leq C_{i,2}$. The DNA personal ID is obtained by concatenating the repeat counts of all loci as given below [6].

DNA_{ID} = $Locus_1$|| $Locus_2$||$Locus_3$|| …….|| $Locus_{13}$
= L_1 ||L_2||L_3||……..||L_{13}
= CSF1PO || FGA || TH01 || TPOX || VWA || D3S1358 || D5S818 || D7S820 || D8S1179 || D13S317 || D16S539 || D21S11 || D18S51
= $C_{1,1}C_{1,2}$ || $C_{2,1}C_{2,2}$ || $C_{3,1}C_{3,2}$ ||….|| $C_{13,1}C_{13,2}$

For example, Alice has the following alleles at the respective loci: 04/07, 06/08, 13/15, 29/32, 15/16, 11/11, 08/08, 12/15, 23/24, 06/07, 08/11, 24/32.2, 08/10. Then, the DNA_{ID} is defined as follows.

DNA_{ID}=04070608131529321516111108081215232406070811243220810.

When the STR number of an allele has a fractional component, such as allele 32.2 in D21S11, the decimal point is removed, and all of the numbers, including those after the decimal point, are retained.

8. Genotype Probability at any STR Locus

One of the important works of DNA analysis is the creation of population databases for the STR loci studied. Probability calculations are based on knowing allele frequencies for each STR locus for a representative human population (and Hardy-Weinberg equilibrium for the population by statistical tests). Allele frequency is defined as the number of copies of the allele in a population divided by the sum of all alleles in a population [3]. For a heterozygous individual, if the two alleles have frequencies of p and q in a population [3,12], the probability (P) of an individual of having both alleles at a single locus is

$$P = 2xpxq.$$

If an individual is homozygous for an allele with a frequency of p, the probability (P) of the genotype is

$$P = p^2.$$

If Alice has the genotype 15, 18 at the locus D3S1358 and the frequency of the alleles 15 and 18 is 0.2825 and 0.1450 respectively, then the frequency of the 15, 18 genotype is therefore

$$P = 2x(0.2825)x(0.1450) =0.0819, \text{ or } 8.2\%.$$

8.1. Probability for a DNA profile of Multiple Loci

If databases of allele frequency for different loci can be shown to be independently inherited by appropriate statistical tests, the probability for the combined genotype can be determined by the multiplication (product rule) [12]. Then, the probability (P) for a DNA profile is the product of the probability (P_1, P_2, ... P_n) for each individual locus, i.e.

$$\text{Profile Probability} = (P_1)\,(P_2)\,\ldots\,(P_n)$$

The probability can be an extremely low number when all 13 STR loci are included in the DNA profile. It could be less than, say, 1.3 times 10^{-16}, or no more frequent than 1 in 7.7 quadrillion individuals (7.7 million billion), which is more than a million times the population of the planet.

9. Entropy of Loci

Information is the state of a system of interest. The most important quantities of information are entropy and mutual information. The entropy is the information in a random variable and mutual information is the amount of information in common between two random variables. The Shannon entropy [7] is a measure of the average information content one is missing when one does not know the value of the random variable. The security of the authentication depends on the entropy of DNA templates [4,6].

The entropy H of a discrete random variable X with possible values $\{x_1, ..., x_n\}$ is

$$H(X) = \mathrm{E}(I(X)).$$

Here E is the expected value function, and $I(X)$ is the information content or self-information of X. $I(X)$ is itself a random variable. If p denotes the probability mass function of X then the entropy can explicitly be written as [7]

$$H(X) = \sum_{i=1}^{n} p(x_i)\, I(x_i) = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i),$$

where b is the base of the logarithm used. Common values of b are 2, Euler's number e, and 10, and the unit of entropy is bit for $b = 2$, nat for $b = e$, and dit (or digit) for $b = 10$. Then, the min-entropy of the genotype G at a locus is given by H (G) = -log (max (max (p^2), max ($2xpxq$)).

The entropy of an STR profile α is the sum of the entropies of the contributing single locus genotypes G_i. Consequently, we obtain

$$H(\alpha)=\Sigma G_i.$$

10. Public Key Cryptography with Elliptic Curves

We discuss in this paper a brief introduction to Elliptic Curve Cryptography that is incorporated into our DNA based PKI. For more complete introductions to ECC, the readers are referred in, for example, [8] and [9].

Introduced in the mid 80s by Victor Miller and Neil Koblitz, ECC is based on the discrete logarithmic problem over the points on an elliptic curve. The principal

attraction of ECC compared to RSA is that it offers equal security for a far smaller key size.

An elliptic curve is the set of equations of the form

$y^2= x^3+ax+b$ (1)

or

$y^2 +xy = x^3+ax^2+b$ (2)

or

$y^2 +y = x^3+ax+b$ (3)

where **x** and **y** are variables, **a** and **b** are constants. However, for cryptography purposes these values are from a Galois field. Galois field is always a positive prime power, p^n and is denoted by $GF(p^n)$. Two special Galois fields are standard for use in Elliptic Curve cryptography. They are GF(p) when n=1 and $GF(2^n)$ when p=2. We shall discuss here only GF(p).

10.1. Elliptic curve over a GF(p)

Elliptic curves over GF(p) are of the form $E_p(a,b)$: $y^2 \pmod p = x^3+ax+b \bmod p$, where $a,b \in F_p$ and $-16(4a^3+27b^2) \bmod p \neq 0$.

All operations such as addition, subtraction, division, multiplication involves integers between **0** to **p**-1. The prime **p** is chosen such that there is finitely large number of points on the elliptic curve to make cryptosystem secure.

10.2. Operations required by ECC

The scalar multiplication or repeated addition of elliptic curve points is the main operation required by ECC schemes, although other operations such as division may also be needed. The addition of two elliptic curve points is illustrated geometrically in figure 5 given below. The line connecting the two points P and Q intercepts the curve at a point called –R. We reflect –R on the x-axis to get R. This point R is the sum of P and Q i.e. R=P+Q. To double a point P, first we draw the tangent line and find the other point of intersection -R. We reflect –R on the x-axis to get R. Now, R=P+P=2P. See figure 6 given below.

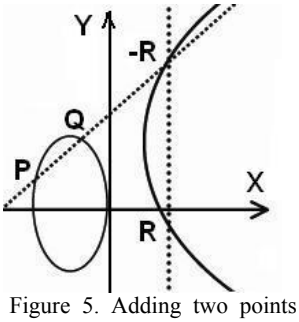

Figure 5. Adding two points

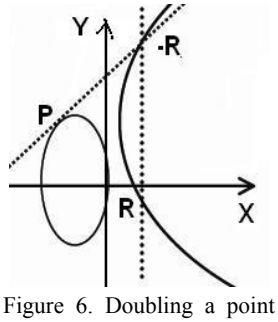

Figure 6. Doubling a point

10.3. The Group Law of Elliptic Curve

Let E be an elliptic curve defined over the field **K**. There is a *chord-and-tangent rule* for adding two points in $E_p(a,b)$ to give a third point in $E_p(a,b)$. Together with this addition operation, the set of points forms an abelian group with **0** serving as its identity. It is this group that is used in the construction of elliptic curve cryptographic systems. The group law presented here is for non-super singular elliptic curves E of the form $y^2=x^3+ax+b$.

1. *Identity.* $P +0=0+ P = P$ for all $P \in E_p(a,b)$.
2. *Negative*. If $P = (x , y) \in E_p(a,b)$, then$(x , y) + (x ,-y) =0$. The point $(x ,-y)$ is denoted by -P and is called the negative of P; note that -P is indeed a point in $E_p(a,b)$. Also, -0 = 0.
3. *Point addition*. Let $P = (x_1 , y_1) \in E_p(a,b)$ and $Q = (x_2, y_2) \in E_p(a,b)$,where $P \neq \pm Q$ .Then $P+Q =R=(x_3, y_3)$, where $x_3=\lambda^2-x_1-x_2$, $y_3=\lambda(x_1-x_3)-y_1$ and $\lambda=(y_2-y_1)/(x_2-x_1)$.
4. *Point doubling*. Let $P = (x_1 , y_1) \in E_p(a,b))$, where $P \neq -P$. Then $2P =R= (x_3 , y_3)$, where $x_3=\lambda^2-2x_1$, $y_3=\lambda(x_1-x_3)-y_1$ and $\lambda=(3x_1^2+a)/(2y_1)$.

The multiplication over an elliptic curve group $E_p(a,b)$ is equivalent to the modular exponentiation operation in RSA. The multiplication of points by a scalar is a series of additions and doubling of points. As an example consider the point, $P=(3, 10) \in E_{23}(1, 1)$. Then $2P =(x_3,y_3)$ is equal to:

$2P = P + P =(x_1,y_1)+(x_1,y_1)$.

Since P = Q, the values of λ , x_3 and y_3 are given by:

$\lambda=((3x_1^2+a)/(2y_1)) \bmod p$

$= ((3 \times 3^2+1)/(2 \times 10)) \bmod 23=(5/20) \bmod 23$

$=4^{-1} \bmod 23=6.$

$x_3 =(\lambda^2 - x_1 - x_2) \bmod p =(6^2 - 3 - 3) \bmod 23$

$=30 \bmod 23 = 7$

$y_3 = (\lambda (x_1 - x_3) - y_1) \bmod p =(6\times (3 - 7) – 10) \bmod 23$

$= -34 \bmod 23 = 12$

Therefore, $2P =(x_3,y_3)=(7,12)$.

The value of *kP* may be computed recursively as given below:-

$$kP = \begin{cases} P \text{ for } k=1 \\ (P+P)xk/2 & \text{for k even} \\ P+(k-1)xP & \text{for k odd} \end{cases}$$

The multiplication by –1 converts **P** to **–P** by negating the **y** coordinate of **P** i,e the negative of **P**=(x,y) gives **–P**=(x,-**y**).

10.4. Elliptic Curve Discrete Logarithm Problem

Let *E* be an elliptic curve defined over a finite field and let, *P* be a point (called base point) on *E* of order *n* and k is a scalar. Calculating the point Q=*kP* from P is very easy and *Q=kP* can be computed by repeated point

additions of P. However, it is very hard to determine the value of *k* knowing the two points: *kP and P*. This leads to the definition of Elliptic Curve Discrete Logarithm Problem (ECDLP), which is defined as: "*Given a base point P and the point Q=kP, lying on the curve, find the value of scalar k, provided that such an integer exists*". The integer *k* is called the *elliptic curve discrete logarithm of Q to the base P*, denoted as $k = \log_P Q$ [8].

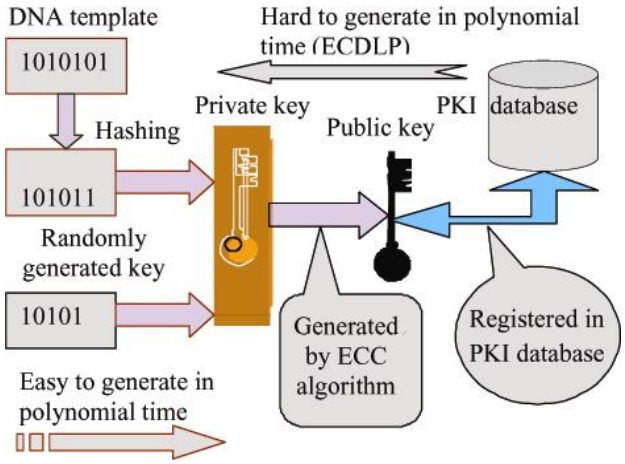

Figure 7: Keys generation and PKI

11. Generation of Private and Public Keys from DNA_{ID}

We describe in this section how to generate individual specific private key and public key from DNA information of the individual. The DNA_{ID} obtained above is a straightforward series of digits. As the DNA_{ID} contains direct individual specific information (but it may not contain any genetic information as we do not know the function of non-coding regions of DNA), it is not used in raw form as private key. Moreover, the DNA_{ID} consists of many digits. So, the DNA_{ID} is subsequently processed by a hashing function such as the secure hash algorithm-1 (SHA-1) as shown in figure 7 above. Given its general acceptance and incorporation in cryptographic standards, SHA-1 is widely used in public key cryptography where public key-based authentication mechanisms are required. SHA-**1** is a one-way hash function that takes an arbitrary length message, processes it, and returns a fixed length **1**60-bit message digest. The primary characteristic of SHA-**1** is that it provides a mechanism that makes it easy to compute a hash from some data, but difficult to determine any data from a computed hash value. Hashing protects privacy [5,10] and also reduces the data length of the DNA_{ID}. This one-way function, SHA-1, produces a private key according to the following transformation:

$$d_1 = h(DNA_{ID})$$

We also choose another private number $d_2 \in [1,p]$. Now the private key is the pair (d_1,d_2) as shown in figure 7 above.

To generate public key we take any two points, P and Q, in $E_p(a,b)$ and calculate a third point, R, in the curve such that $R=d_1P+d_2Q$, where d_1 and d_2 are the private key pair calculated above . The public key is R.

11.1. Public Key Binding

The public key of Alice needs to be distributed so that Bob can get it with integrity. Public-key infrastructure (PKI) has been used for this purpose. This model allows entities to accept digital certificates from outside sources and encrypt and sign the data with their own certificates [19]. This is essential for digital signature, authentication etc. The primary objective of a CA is to bind the identifying information and credentials supplied by an end entity with the public key of that end entity. The binding is declared when a trusted CA digitally signs the public key certificate with its private key.

Alice registers the tuple $\{E_p(a,b), R\}$ as his public key with the CA of PKI as shown in figure 7 above; he keeps (d_1,d_2) as his private key.

11.2. Authentication Scenario

Alice selects two random numbers r_1 and r_2 where $r_1,r_2 \in [1, n]$ and compute the following equation:-

$S=r_1P+r_2Q$. Here P and Q are the two points on the elliptic curve, $E_p(a,b)$.

Alice sends S to Bob.

Bob selects a random integer **e** such that $e \in [1, n]$ and sends it to Alice. Alice computes x_1 and x_2 from the following equations:

$x_1=r_1+ed_1$

$x_2=r_2+ed_2$

Alice sends the pair (x_1, x_2) to Bob.

Bob gets the public key R of Alice and checks whether the following equation is satisfied.

$S=x_1P+x_2Q-eR$

The authentication is successful if satisfied, otherwise unsuccessful.

11.3. Security of the System

The security of the system is governed both by the biometric DNA information and the ECDLP mentioned above. Even if Eve steals the private key d_1 derived from DNA_{ID}, he cannot find out the other private key d_2 in polynomial time due to ECDLP [8]. Given the state of today's computer technology, it is strongly believed that the ECDLP is infeasible to solve in polynomial time. At the time of writing of this paper, no one has proven that there does exist an efficient algorithm for solving the ECDLP in polynomial time. Internet sources say that the ECDLP is the hardest of the hardest problems mathematicians face today.

12. Why Choose ECC instead of RSA

Key strength is a critical factor in cryptography. An insufficiently strong key will corrupt even the best-designed public key cryptosystem. The level of cryptographic strength of any key depends on the corresponding algorithm used.

The integer factorization problem and elliptic curve discrete logarithm problem can be used as the benchmarks against which to evaluate the security of RSA and ECC. The level of effort for factoring integers and computing elliptic curve discrete logarithms may be measured in a unit called MIPS year [11]. The term MIPS year denotes the computational power of a MIPS computer utilized for one year; a **m**illion-**i**nstruction-**p**er-**s**econd processor running for one year, which is about $3x10^{13}$ instructions executed [11]. It is worthy to note that elliptic curve discrete logarithm problem of ECC appears to be relatively more difficult than that of integer factorization problem of RSA.

The following figure in table 2 shows the level of effort required for various values of **n** in bits to factor with current version of the GNFS and to compute a single elliptic curve discrete logarithm using the Pollard-rho method.

Table 2: Security comparison of RSA and ECC

RSA	ECC	MIPS years to attack
1024	160	10^{12}
2048	224	10^{24}
3072	256	10^{28}
4096	280	10^{31}
7680	384	10^{47}
21000	600	10^{81}

13. Benefits and Weaknesses of biometric DNA systems

13.1. Benefits

1 DNA is the most distinct biometric identifier available for human beings.
2 It is highly accurate; the chance of 2 individuals sharing the same DNA profile is extremely low. It is only 1 in 3 trillion.
3 DNA does not change throughout a person's life; therefore the permanence of DNA is incontestable.
4 DNA data are digital in nature.
5 DNA is very stable and resists degradation to extreme conditions.
6 We can have economical genetic system, which does not need to build up its own biological database [5].

13.2. Weaknesses

1 DNA matching is not done in real-time.
2 It is intrusive; a physical sample must be taken, while other biometric systems only use an image or a recording.
3 If sample collection is not supervised, an impostor could submit anybody's DNA.
4 The cost of DNA analysis (16 loci) in India is about Rs 11500/- for two people.
5 The main problem with DNA is that it includes sensitive information related to the genetic and medical aspects of the individuals. So, many fears that any misuse of DNA information can disclose information about genetic and medical disorders.

14. Recognization Issues with any Biometric System

There are two basic types of recognition errors: the **F**alse **A**ccept **R**ate (FAR) and the **F**alse **R**eject **R**ate (FRR). A False Accept is when a nonmatching pair of biometric data is wrongly accepted as a match by the system. A False Reject is when a matching pair of biometric data is wrongly rejected by the system. The two errors are complementary: When you try to lower one of the errors by varying the threshold, the other error rate automatically increases. There is therefore a balance to be found, with a decision threshold that can be specified to either reduce the risk of FAR, or to reduce the risk of FRR.

In a biometric authentication system, the relative false accept and false reject rates can be set by choosing a particular operating point (i.e., a detection threshold). Very low (close to zero) error rates for both errors (FAR and FRR) at the same time are not possible. By setting a high threshold, the FAR error can be close to zero, and similarly by setting a significantly low threshold, the FRR rate can be close to zero. A meaningful operating point for the threshold is decided based on the application requirements, and the FAR versus FRR error rates at that operating point may be quite different. Internet sources say that to provide high security, biometric systems operate at a low FAR instead of the commonly recommended **E**qual **E**rror **R**ate (EER) operating point where FAR = FRR.

15. The Future of Biometric DNA System

The future of biometric DNA in terms of physical and network security will rely on experts' ability to make it a more cost efficient method of authentication and identification. Whether this means portability or mass production, development will depend on technological advances in the areas of DNA sequencing and sample comparison techniques. A professor at National University in San Diego, California is working on creating a portable DNA sequencer that will combine existing DNA biosensors with a new device called the **I**on-**S**elective

Field-**E**ffect Transistor (ISFET). This product would allow a handheld device to perform the same activities that currently must take place in a laboratory. As these kinds of advancements take place, the implementation of biometric DNA into civilian business environments for use in physical and network security will expand to a great extent. The precision and accuracy of DNA recognition will make it a much desired means of authentication, and hopefully verification, in the near future.

16. Conclusion

This paper discussed the application of DNA information and elliptic curve discrete logarithm problem in information security systems. The authentication system can be extended to the identification system when DNA proving is done in real time in near future. We also mentioned the various issues that will be encountered when using DNA information to security systems. Currently elliptic curves are believed to provide good security with smaller key sizes as compared to RSA. The uniqueness of digital DNA information supported by the security of ECC appeals us to use such authentication systems in near future.

Acknowledgment

The theme of this paper was conceived after the first author had attended a workshop held at the Department of Computer Science, Manipur University, Imphal and organized by the Machine Intelligence Unit of the Indian Statistical Institute, Kolkata in 2008. So, the first author is very thankful to the organizers of the workshop. He also wishes to thank Soram Renubala Devi for her helpful comments on special topics on genetics that ultimately clears his doubts on those topics. Lastly, he also thanks his daughter *Java Compiler Soram* and son *Chandrayan One Soram* for not complaining much when he is spending too much time with the computer.

References

[1] PKI from “http://en.wikipedia.org/”.
[2] Sanghamitra Bandyopadhyan of Indian Statistical Institute (ISI), ”Winter School on Data Mining and Computational Biology”, Manipur University, Canchipur, Imphal, Jan 28-Feb 01,2008.
[3] A.M. Campbell, L.J. Heyer, ”Discovering Genomics, Proteomics, & Bioinformatics”, Pearson Education, New Delhi, 2004.
[4] Ulrike Korte and others, “A cryptographic biometric authentication system based on genetic fingerprints”, Springer-Verlag, 2008.
[5] Yukio Itakura, Shigeo Tsujii, ”Proposal on Personal Authentication System in which Biological Information is embedded in Cryptosystem”, IACR, 2003.
[6] Masaki Hashiyada, ”Development of Biometric DNA Ink for Authentication Security”, Tohoku J. Exp. Med, 2004.
[7] C. E. Shannon, ”A Mathematical Theory of Communication”, of reprinted version. The Bell System Technical Journal, 1948.
[8] Tibor Juhas, “The Use of Elliptic Curves in Cryptography”, Master Thesis, University of Tromso, 2007.
[9] N. Koblitz, “Elliptic Curve Cryptosystems, Mathematics of computation”, 1987.
[10] Edited by Gustavus J. Simmons,” Contemporary Cryptology: The Science of Information Integrity”, IEEE press, New York.
[11] Stalling W, ”Cryptography and Network Security”, Prentice-Hall of India, New Delhi, 2001.
[12] Allele frequency and Genotype probability from “ http://www.biology.arizona.edu/”.
[13] Ms.P.G.Rajeswari, Dr.K.Thilagavathi, “An Efficient Authentication Protocol Based on Elliptic Curve Cryptography for Mobile Networks”, International Journal of Computer Science and Network Security, VOL.9 No.2, February 2009.
[14] Achintya K. Mandal and Subodh Gopal Nandi, ”Biometric Recognition: Novel Approach for Library Patron Authentication”, Visva-Bharati University, Santiniketan, India.
[15] Young-Bin Kwon, ”Biometrics in Asia ”, Chung -Ang University, Korea, 2009.
[16] Kresimir Delac, Mislav Grgic, ”A survey of biometric recognition methods”, 46th International Symposium Electronics in Marine, Croatia, 2004.
[17] Leonard M. Adleman, ”Computing with DNA”, Scientific American, 1998.
[18] D. Heider and A Barnekow, “DNA-based watermarks using the DNA-Crypt algorithm”, BMC Bioinformatics, 2007.
[19] A.S.Tanenbaum, ”Computer Networks”, 4th Edition, PHI, New Delhi, 2006.

Ranbir Soram is working as a lecturer in Computer Science and Engineering at Manipur Institute of Technology, Takyelpat, Imphal, India. His field of interest includes network security, neural network, genetic algorithm etc.

Memeta Khomdram is working at Department of Electronics Accreditation of Computer Courses Centre, Akampat, Imphal, India. Her only hobby to is give a piggyback to her small son Chandrayan One Soram while holding her small daughter Java Compiler Soram by arm.