

DNA and Bernoulli Random Number Generator Based Security Key Generation Algorithm

Sodhi, G. K. and Gaba, G. S.*

*Discipline of Electronics and Communication Engineering, Lovely Professional University,
Jalandhar 144411, India*

ABSTRACT

Security is a major concern for the communication sector. The technique presented in this paper provides a novel security key generation mechanism. The proposed technique aims to generate a security key using the biological characteristics of the human body and the mathematically generated pseudo random sequences, thus producing different keys for different individuals. The final key is produced through the fusion of deoxyribonucleic acid (DNA) sequence of 1024 characters and Bernoulli Random Number Generator sequence of 256 bits. The performance of produced keys is evaluated using National Institute of Standards and Technology (NIST) tests and uniqueness is verified through avalanche test.

Keywords: Authentication, Bernoulli random number generator, biometrics, communication, confidentiality, DNA, integrity, security

INTRODUCTION

Security is critical in communication networks following the growth of the internet. Biometrics and integrity are two mechanisms which can improve security considerations. Biometrics ensures the identification and authentication of individuals by observing their personal unique features (Hao, Anderson, & Daugman, 2006). A practical system that integrates the iris biometrics into cryptographic applications can be found in Hao et al. 2006. A system that works using audio fingerprints has also been proposed by (Brown & Seberry, 2001; Chouakri, Bereksi, Ahmaidi, Fokapu, 2005; Covell & Baluja, 2007. Ktata, Ouni, Ellouze, 2009) studied the use of electrocardiogram (ECG) signals for enhanced security. Baluja, Covell, 2007 proposed to create personal signatures for authentication. Chen & Chandran, 2007) did a study on identification based on image processing is carried out by various researchers.

DNA has been used in many cryptographic algorithms to provide confidentiality (Chang, Kuo, Lo, & Wei, 2012). Bernoulli random

Article history:

Received: 29 December 2016

Accepted: 21 April 2017

E-mail addresses:

gurpreetsodhi123@gmail.com (Sodhi, G. K.),

er.gurjotgaba@gmail.com (Gaba, G. S.)

*Corresponding Author

number generator (BRNG) is used to make the technique more efficient and effective. The BRNG works on a secret seed value which ensures the generation of a different sequence for each seed value provided. The work reported in this paper is based on the idea of unique and random attributes of DNA and Bernoulli random number generator sequence.

The approach to generate a security key is implemented through numeric coding of the DNA sequence and the Bernoulli random sequence generated with the secret seed value. The output of key generating algorithm is tested using NIST tests of randomness as well as the strict avalanche criterion, the results of which are formulated in Table 4 and Table 5 respectively. Results point out the applicability of the proposed approach in areas where security is a concern.

The paper is organized as follows; characteristics of DNA and Bernoulli random number generator are described in Section 2. In Section 3, proposed algorithm for the 256-bit key generation is presented where the DNA data is taken from the MIT-BIH database, followed by result and analysis in Section 4. Finally, conclusions and future work are reported in Section 5.

Characteristics of DNA and BRNG

Progress in the field of biotechnology has made the DNA sequencing more effective. DNA sequencing for various organisms has been done with higher accuracy (Goldberger, Amaral, Glass, Hausdorff, Ivanov, Mark, Mietus, Moody, Peng & Stanley, 2000). Investigating the biological relationships of different species is known as analysing the DNA sequence.

The internet provides various databases of DNA sequences which can be easily accessed from the World Wide Web (Ensembl Genome Browser, NCBI databases). The DNA is composed of two polymeric strands made of monomers that include a nitrogenous base (A-adenine, C-cytosine, G-guanine, and T-thymine), deoxyribose sugar and a phosphate group. According to most of the techniques a DNA sequence is taken as symbolic data that is composed of four characters A, C, G and T corresponding to the four types of nucleic acids present. The DNA samples from a genome are sequenced using a Genome Sequencer, the signals created in the sequencing process are then analysed by the software to generate millions of sequenced bases.

The backbone of each strand located on the surface of the DNA is formed by the sugar and phosphate groups, while the inside of the structure is made of bases. There exists a weak hydrogen bond between the complementary bases of each strand (i.e. between A and T and between C and G) giving rise to pairing of bases, this pairing holds the two strands together. DNA sequences are unique for every individual, including the identical twins. The pattern formed by a DNA sequence specifically represents an individual and its characteristics. Hence, there is no possibility of duplicity in the DNA sequences.

To strengthen the bond of security, a random sequence is generated by Bernoulli random number generator. This sequence is generated with a seed value that is secretly given by the user. BRNG computes n^{th} Bernoulli number for a given integer n . Bernoulli numbers are a sequence B_n of rational numbers defined by the Taylor expansion.

Bernoulli numbers have a prominent place in mathematics, for instance they appear in Taylor expansion of tangent and in Euler-Maclaurin formula. Equation (1) represents the BRNG method for generation of random sequences.

$$\sum_{n=0}^{\infty} \frac{B_n}{n!} x^n = \frac{x}{e^x - 1} \quad (1)$$

The fusion of BRNG and DNA sequence forms a very strong 256-bit key which is less susceptible to attacks and thus provides higher level of security.

The Key Generation Mechanism

The proposed key is generated by combining DNA sequence of an individual and the random sequence obtained through BRNG using a secret seed value. The method of the generation of these keys is devised into three subsequent subsections and the hierarchical structure is portrayed in Figure 1.

DNA sequence formulation

- 1) Obtaining a DNA sequence of 1024 characters from the DNA database [9]:
The DNA sequence consists of 1024 characters forming base pairs 'agct'.
- 2) Obtaining the binary sequence from DNA characters:
Each character of the DNA sequence is represented by 8 bit ASCII code. Hence, resulting in a DNA sequence of length 8192 bits.
- 3) Framing a DNA sequence of 256 bits:
 - (i) The DNA sequence is then divided into equal halves.
 - (ii) Apply exclusive-or operation on the obtained sequences.
 - (iii) The result is further divided into two equal parts and exclusive-or operation is applied again.

The step (iii) is repeated till a sequence is obtained whose length is 256 bits. The whole procedure is summarized in the flow chart (Figure 1).

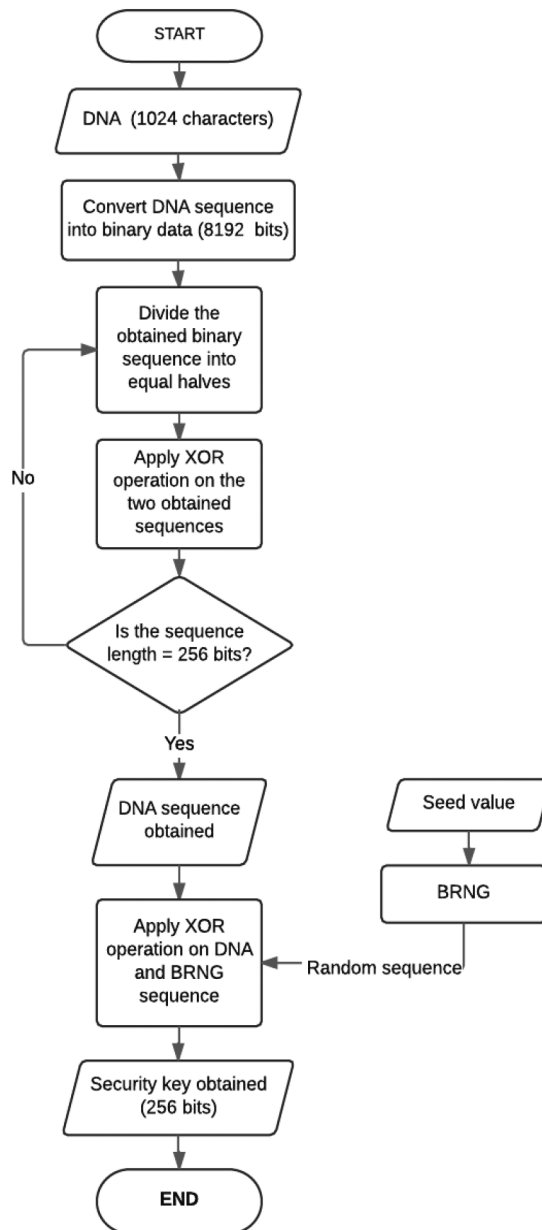


Figure 1. Key generation process

The conversion from ‘agct’ sequence of 1024 characters to 256-bit binary DNA sequence is given in Table 1.

Table 1
DNA sequence formulation

	DNA sequence (1024 characters)	DNA sequence (8192 bits)	DNA sequence (256 bits)
D ₁	gcacaatcagaagcaggcggaggagacggcg gccttcgaggaggtcatgaaggacctgacctg ac.....gcaca gaggcaaggcgtcagcagcatgccaccct gtctccgctgtcaccatcactcaggctgtagcc atg.	011001110110001101100001 011000110110000101100001 0111010001100011.....01 110100011000010110011101 100011011000110110000101 11010001100111	00000100000000000001010100 01011100000010000001100001 01010000010000000110000001 10000001100000011000010011 00000000000000000001001100 0000000000010000001100000 0000000001000000110000001 0000000001000000110000001 00000000100001001100000110 00000010000000000000010000 00000000100110000000
D ₂	ccacgcgtccggcgagaagatggcgacttcg aacaatccgcggaaatcagcagagaagatcgc gct.....ggcgtcagcc ccctgtccctgagcacagaggcaaggcgtcag caggcatcgcggccctgtcccgctgcacc at.	011000110110001101100001 011000110110011101100011 0110011101110100.....0 111010001100011011000010 110001011000110110001101 10000101110100	00000010000000100000001000 00001000000110000101110000 01 000001011100000100000000 0000000000000000000000110 00000010000001100001010100 01001100000010000000000000 00000000001000010011000001 100000000000000001000000110 00000110000000100000000000 0101110000010000000100
D ₃	agcccttaggggaagatcctgctgtgctgtg atgctccagctccagaaatccagctacctgcaac tg.....tctggagcagcagctg ccctacgctctctcaccaggcgggtcccagc agccaccgcccagcagcccccagccccgccc	011000010110011101100011 011000110110001101110100 0111010001100001.....1 000110110001101100011011 000110110011101100011011 0001101100111	00010101000000000000001000 01011100000100000101110000 00000000001000000100000001 10000000000000010000010101 00000010000001000000000000 00000000010101000000100000 00100000000000000100000000 00000101110000001000010011 00010011000001100000000000 0101110001001100010011

*D₁, D₂, D₃: DNA sequences, NCBI Database

Random sequence generation through BRNG. Three sequences are generated using three different seed values. A seed value is a secret input given by the user at the beginning of generator operation. The obtained random sequences along with the corresponding seed value are summarized as in Table 2.

Table 2
BRNG sequence

Seed value	Bernoulli random sequence (256 bits)
B ₁ 1231	101100110000001101011010011000010101001000010111011101110011101000 000011000001101010110100010101101110101000110111010101010110000001 100111011010000011011001100101111011100101000100010110010000000101 1100001111011010111011110000011011110000110010010010110
B ₂ 101355	0001011010010001010100011001011001110000001101011111001111010101110 0010010110010100001110110111001110000100110011001101101101001011100 1101110100011000001101111100111111100110100000111110011001001001100 010001000110011011100100100100101010110101101010001
B ₃ 2114	011101010000110100110000010101010000001000100001010101010101000001 10101111010101010011010011010010100000110001000101011100000001000111 0011100010000110010010111110001011100010100110101110111110101100011 00011000101110110010110100100100001111011110001101011

*B₁, B₂, B₃: Three BRNG sequences

Fusion of DNA sequence and BRNG sequence. To enhance the strength of security and randomness in the key, exclusive-or logic is applied between each DNA and BRNG sequence. The resultant key has length of 256 bits and is strong enough to survive critical attacks on networks. The process is repeated for two other DNA and BRNG sequences. Therefore, three 256 bit keys are obtained as shown in Table 3.

Table 3
Security keys

KEY ₁	00010010100100010100010010000001011100100011001111011001101110110010001 0111110100010110110001001010110100110011001101101010010101110011011110001 11100011011111001001111111110100100111110111000000101100100001000010011011 100100000100101010111100001010001
------------------	---

KEY ₂	10110001000000010101100001100010101011110001110010111001110110011000010011 0000011010101101000101011010001010000101110011010000110001000000111001010 00001101100110010110101100001100011101011001000000011011101101111000010111 00111000001101010110011011001001111
------------------	--

KEY ₃	0110000000001101001100100100001000000101000001111010101010101010001100011 1101100101001101001111001000101011000000010111110000000100011100110010000 0111010010101110001011100000100110101110010001010111001110000000010001100 01110100100100100001010111101111000
------------------	--

The final key is a combination of true random source (DNA) and pseudo random source (BRNG). The keys obtained have unique and random characteristics and thus can be useful in maintaining security.

RESULTS AND DISCUSSION

A security key is said to be efficient if it is random and unique. The National Institute of Standards and Technology (NIST) tests discuss some aspects of selecting and testing random number generators. The outputs of these generators may be used in most of the security applications for the generation of security keys. The generators that are to be used for security applications should meet stronger requirements than those to be used for other applications. To be precise, their outputs need to be unpredictable for unknown inputs. These tests may be useful for determining if a generator is suitable for a particular security application. The randomness of a key is evaluated on the bases of its P-value, which must be greater than 0.01 for a random sequence.

NIST Tests

The keys obtained are evaluated on the basis of seven NIST randomness verification tests and the P-value is calculated for each test with respect to the security key. An overview of the NIST tests used for evaluating the sequences is given as:

Runs test. The purpose of applying this test is to calculate the number of runs in an entire sequence, where run specifies the number of uninterrupted sequence of identical bits (Rukhin et al., 2010). The results in Table 4 clearly depict that the proposed algorithm has higher rate of interruptions.

Frequency Test. Frequency test is used for determining the proportion of number of ones and zeros in an entire sequence. It is used to check the closeness between the number of ones and number of zeros. A sequence is said to be random if the proportion of zeros and ones are close to each other (Rukhin et al., 2010). The results in Table 4 clearly depict that the proposed algorithm produces better proximity between the count of ones and zeros as compared to the traditional techniques.

Approximate Entropy Test. This test deals with the frequency of all the overlapping bit patterns across the entire sequence. The aim of this test is to compare the frequency of overlapping blocks of two consecutive/adjacent lengths with the expected result for a random sequence.

Discrete Fourier Transform Test (DFT). The DFT test is used to find the peak heights in the Discrete Fourier Transform of a sequence. The aim of this test is to detect periodic features (i.e. repetitive patterns) in the tested sequence that indicates a deviation from the assumed randomness. The goal is to detect if the number of peaks exceeding the 95 % threshold are significantly different than 5%.

Binary Derivative Test. The test is performed using exclusive-or operation between successive bits until only one bit is left. Afterwards, the ratio of number of ones to the length of entire sequence in each case is calculated. Finally, the average of the ratio of all the sequences is observed, and where the value lies near to 0.5, then the sequence is considered as a random sequence (Rukhin et al., 2010). The results in Table 4 show that the output of proposed algorithm is random.

Maurer's "Universal Statistical" Test. The aim of this test is to detect if a sequence can be significantly compressed without any loss of information. The number of bits between matching patterns are calculated. A sequence that is significantly compressible is considered to be non-random. This test is also known as Universal test (Rukhin et al., 2010).

Random Excursion Variant Test. The test focuses on the total number of times a particular state occurs in a cumulative sum random walk. It detects the deviations from the expected number of visits. The P-value specifies if the sequence is random or not. This test considers successive sums of the binary bits as a one-dimensional random walk (Rukhin et al., 2010).

$$P_{value} = erfc \times \frac{(|\xi(x) - j|)}{\sqrt{(2 \times j \times ((4 \times |x|) - 2))}} \quad (1)$$

Where,

$erfc$: the error function

ξ : the total number of times the state x occurs

x : the state occurred

j : the total number of cycles

The P-value must be greater than 0.01 for a sequence to be random (Rukhin et al., 2010).

It is observed that the proposed technique produces better results in terms of randomness of the keys. The efficiency of the proposed technique is evaluated by comparing it with other traditional biometric techniques used for authentication and security key generation. The tests have been performed on KEY₁ and results are tabulated in Table 4.

Table 4
Comparison and analysis

S. No	Input Source of random number generator	Key Length (bits)	Runs Test		Frequency Test		Approximate DFT Test	Binary Derivative Test	Maurer's Test	Random Excursion Variant Test
			P-value	P-value	P-value	P-value	P-value	P-value		
1.	ECG	128	0.1262	0.2487	0.5468	0.0294	0.5039	0.9428	Random	
2.	Image	256	0.0809	0.8026	0.9759	0.4220	0.4887	0.9780	Random	
3.	Iris sequence	128	0.1254	0.3768	0.9409	0.3304	0.5021	0.9062	Random	
4.	Finger print	128	0.3345	0.3041	0.3345	0.7597	-	0.2757	Random	
5.	DNA & BRNG	256	0.0438	0.9005	0.9340	0.8185	0.5090	0.9920	Random	

It can be observed that the P-value obtained for the keys generated for all the seven tests is significantly greater than 0.01. Thus, it can be concluded that the keys generated are random in nature and hence fulfil the basic criteria required for security keys.

Avalanche test has also been performed on the keys obtained. The purpose of this test is to check the avalanche effect, which is a desirable property of the security keys. Wherein if the input is changed slightly the output changes significantly. It gives the percentage of bits flipped with a change in the input. It is a desirable property of security keys.

The test is performed on three sets of DNA and BRNG sequences:

Case 1: In the initial set, two security keys are generated through two DNA sequences while keeping the same BRNG sequence.

Case 2: The second set involves generation of two security keys through the same DNA sequence and two BRNG sequences.

Case 3: In the third set, two security keys are generated through two DNA and BRNG sequences.

Further, the avalanche effect is calculated for each of the three sets and results are tabulated in Table 5, Table 6 and Table 7 respectively, this is done to know the amount of randomness the proposed technique produces on changing the input.

The avalanche effect can be calculated using the formula given in equation (3).

$$Avalanche\ effect = \frac{No.of\ bits\ flipped\ in\ the\ sequence}{Total\ no.of\ bits\ in\ the\ sequence} \times 100 \tag{3}$$

The result of avalanche effect on changing the inputs is summarized in tabular form, where D₁, D₂, D₃ represent the DNA sequences as taken from Table 1 and B₁, B₂, B₃ represent the BRNG sequences characterized by the seed values, as taken from Table 2.

Table 5
Avalanche test analysis: Case 1

DNA Sequences (D _n)	Seed Value	Bernoulli Random Sequence (B _n)	Key Generated K= D _n xor B _n	Avalanche Result of Key (K)	
				No. of bits flipped	Avalanche Effect
D ₁	1231	B ₁	1011011100000011010011110111011001	58	22.65 %
			010000000100010110111010011001000		
			001111000010101010000100011001100		
			1110010001101110101010111111000001		
			100111010010000101011001100101100		
			011100011000101010110011000010001		
			111001111101111011101111000011101		
			1110000111011110010110		
D ₂	1231	B ₁	101100010000000101011000011000110	58	22.65 %
			10101000000000001111111000101000		
			000101100000110101011010001010110		
			110110100010011101100101110010001		
			010101110100100000110110011001011		
			100111101100001011101100100000000		
			001110011111010110111010110000011		
			011011110110000010010010		

* Refer Table 1 for D₁, D₂, D₃ and Table 2 for B₁, B₂, B₃

** Different DNA sequences - Same BRNG sequence

Table 6
Avalanche test analysis: Case 2

DNA Sequence (D _n)	Seed Value	Bernoulli Random Sequence (B _n)	Key Generated K= D _n xor B _n	Avalanche Result of Key (K)	
				No. of bits flipped	Avalanche effect
D ₁	1231	B ₁	101101110000001101001111011101100 101000000010001011011101001100100 000111100001010101000010001100110 01110010001101110101010111110000 011001110100100001010110011001011 000111000110001010101100110000100 01111001111101110111011110000111 011110000111011110010110	124	48.43 %
	101355	B ₂	000100101001000101000100100000010 1110010001100111110110011101110111 0010001011111010001011011000100101 0110100110011001101101010010101110 0110111110001111000110111110010011 111111101001001111101110000001011 0010000100001001101110010000010010 1010111100001010001		

* Refer Table 1 for D₁, D₂, D₃ and Table 2 for B₁, B₂, B₃

** Same DNA sequences - Different BRNG sequences

Table 7
Avalanche test analysis: Case 3

DNA Sequence (D _n)	Seed Value	Bernoulli Random Sequence (B _n)	Key Generated K= D _n xor B _n	Avalanche Result of Key (K)	
				No. of bits flipped	Avalanche Effect
D ₁	1231	B ₁	10110111000000110100111101110110 01010000000100010110111010011001 00000111100001010101000010001100 11001110010001101110101010111111 00000110011101001000010101100110 01011000111000110001010101100110 00010001111001111101111011101111 000011101111000011101111001011011	58	22.65 %
	D ₂	101355	B ₂	00010100100100110101001110010100 01110110001000101111110111111011 110011001011001010000111011011100 111110010011101100101110101111010 01111110111100011000001101111100 1101111010101001101111100110010 00001100100001001010011010100100 100100000100110111101010101	

*Refer Table 1 for D₁, D₂, D₃ and Table 2 for B₁, B₂, B₃

**Different DNA sequences – Different BRNG sequences

The avalanche test results indicate that a slight change in the inputs leads to a significant change in the output. The higher the percentage value of avalanche effect, the more is the uniqueness of the key. As the seed value may vary according to the user's choice and the DNA being unique to an individual, it can be assumed that the security system built has higher efficiency and is less susceptible to attacks.

CONCLUSION

In this work, a novel algorithm is proposed to generate biometric security keys for authentication using an individual's DNA sequence and BRNG sequence. DNA being the most unique characteristic of an individual when collaborated with the Bernoulli seed is more efficient. The proposed key finds its application in various areas where maintaining the sender's confidentiality as well as retaining the original message are important. Most of the security system researches have been carried out on fingerprint, facial, iris and voice recognition; while those focused on the DNA are rare. A system that computes a 256-bit biometric security key through the fusion of DNA and BRNG is presented and analysed using NIST Tests. Results indicate this method is superior to more traditional techniques.

REFERENCES

- Baluja, S., & Covell, M. (2007, April). Audio fingerprinting: Combining computer vision & data stream processing. *IEEE International Conference on Acoustics, Speech and Signal Processing, 2007, ICASSP 2007* (Vol. 2, pp. II-213). IEEE.
- Brown, L., & Seberry, J. (1990). On the design of permutation P in DES type cryptosystems. In G. Brassard (Ed.), *Advances in Cryptology—EUROCRYPT'89* (pp. 696-705). Germany: Springer Berlin/Heidelberg.
- Chang, H. T., Kuo, C. J., Lo, N. W., & Lv, W. Z. (2012). DNA sequence representation and comparison based on quaternion number system. *DNA Sequence*, 3(11), 39-46.
- Chen, B., & Chandran, V. (2007, December). Biometric based cryptographic key generation from faces. In *9th Biennial Conference of the Australian Pattern Recognition Society on Digital Image Computing Techniques and Applications* (pp. 394-401). IEEE.
- Chouakri, S. A., Bereksi-Reguig, F., Ahmaldi, S., & Fokapu, O. (2005, September). Wavelet denoising of the electrocardiogram signal based on the corrupted noise estimation. *Computers in Cardiology, 2005* (pp. 1021-1024). IEEE.
- Covell, M., & Baluja, S. (2007, April). Known-audio detection using waveprint: spectrogram fingerprinting by wavelet hashing. *IEEE International Conference on Acoustics, Speech and Signal Processing, 2007, ICASSP 2007* (Vol. 1, pp. I-237). IEEE.
- Ensembl. (n.d.). *Ensembl Genome Browser*. Retrieved from <http://www.ensembl.org/index.html>
- Garcia-Baleon, H. A., Alarcon-Aquino, V., & Starostenko, O. (2009, August). A wavelet-based 128-bit key generator using electrocardiogram signals. *52nd IEEE International Midwest Symposium on Circuits and Systems, 2009, MWSCAS'09* (pp. 644-647). IEEE.

- Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., ... & Stanley, H. E. (2000). Physio bank, physio toolkit, and physio net. *Circulation*, *101*(23), e215-e220.
- Hao, F., Anderson, R., & Daugman, J. (2006). Combining crypto with biometrics effectively. *IEEE transactions on computers*, *55*(9), 1081-1088.
- Hedayatpour, S., & Chuprat, S. (2011, September). Hash functions-based random number generator with image data source. *IEEE Conference on Open Systems (ICOS), 2011* (pp. 69-73). IEEE.
- Khokher, R., & Singh, R. C. (2015, May). Generation of security key using ECG signal. *International Conference on Computing, Communication and Automation (ICCCA), 2015* (pp. 895-900). IEEE.
- Ktata, S., Ouni, K., & Ellouze, N. (2009). A novel compression algorithm for electrocardiogram signals based on wavelet transform and SPIHT. *International Journal of Signal Processing*, *5*(4), 32-37.
- NCBI. (n.d.). *National Center for Biotechnology Information*. Retrieved from <http://www.ncbi.nlm.nih.gov/Entrez>
- NCBI. (n.d.). *National Center for Biotechnology Information*. Retrieved from <http://www.ncbi.nlm.nih.gov/nuccore/3327045?report=genbank>
- NCBI. (n.d.). *National Center for Biotechnology Information*. Retrieved From <http://www.ncbi.nlm.nih.gov/nuccore/20380066?report=genbank>
- NCBI. (n.d.). *National Center for Biotechnology Information*. Retrieved from <http://www.ncbi.nlm.nih.gov/nuccore/33874586?report=genbank>
- Rukhin, A., Soto, J., Nechvatal, J., Barker, E., Leigh, S., Levenson, M., ... & Smid, M. (2010). *Statistical test suite for random and pseudorandom number generators for cryptographic applications*. Gaithersburg: NIST Special Publication.
- Wei, W., & Jun, Z. (2013, November). Image encryption algorithm based on the key extracted from iris characteristics. In *IEEE 14th International Symposium on Computational Intelligence and Informatics (CINTI), 2013* (pp. 169-172). IEEE.
- Ying, L., Shu, W., Jing, Y., & Xiao, L. (2010, December). Design of a random number generator from fingerprint. In *International Conference on Computational and Information Sciences (ICCIS), 2010* (pp. 278-280). IEEE.

