International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-3, January 2020

VG1 Cipher – A DNA Indexing Cipher

Akhil Kaushik, Vikas Thada

Abstract*: When it comes to providing security to information systems, encryption emerges as an indispensable tool, as it has been used intensively in past few decades for securing stationary data as well as data in motion. Earlier, the security of an encryption algorithm lied in the manipulation of characters among a word or group of words, which is called the classical age of cryptography. This age ended when indigenous mathematical equations came into play and modern ciphers like DES and RSA were designed to mark the modern cryptographic era. This period witnessed two world wars and rise of machines instead of manual calculation for data secrecy. But, the time kept moving on and new advances in the information security field surfaced like Elliptical Curve Cryptography and Quantum Cryptography which added new dimensions to the secret world of confidential communication over unsecure channels. The latest addition in this count is the DNA cryptography which has combined the laws of biology with computing to form unbreakable ciphers at least theoretically. This paper introduces a new DNA cipher which is bound to provide more robust ways to safeguard vital data.

***Keywords:* DNA Cryptography,cipher,information security,encryption, decryption.**

I. INTRODUCTION

Even before the advent of technology, the human race has been profoundly in love with information. This information whether religious, personal or social has always been of great interest to rivals, gossip lovers or revolutionists. Hence, since beginning of civilization there was requirement of safeguarding this confidential information and myriad ways were adopted for it like Scytale, Heliography, Caesar cipher, Pigeon cipher, etc.**[12]** This growth increased quadruple folds with the time, especially during the world wars. Information security refers to the technique of protecting information from unauthorized access, use, disclosure, disruption and modification. Governments, military, corporations, financial institutions, hospitals, and private businesses amass a great deal of confidential information about their employees, customers, products, research, and financial status. Most of this information is now collected, processed and stored on electronic media and transmitted across networks to other computers. Encryption clearly addresses the need for confidentiality of information, in process of storage and transmission. Popular application of multimedia technology and increasingly transmission ability of network gradually leads us to acquire information directly and clearly through multimedia and hence the security of multimedia data has become inevitable.

Revised Manuscript Received on January 30, 2020.
***** Correspondence Author
Akhil Kaushik*,PhD Scholar, CSE Department, Amity University, Gurugram, Haryana, India.
Dr. Vikas Thada, CSE Department, Amity University, Gurugram, Haryana, India.

The following sections talk about the growth of cryptography with the human evolution.

II. CLASSIC CRYPTPGRAPHY

This is considered as the golden era of cryptology where the tools and tricks for encryption where in the infancy stage. It all started when a genius king designed Caesar cipher which was basically a substitution cipher i.e. replacing each character by its third character**[13]**. This was designed to be used for military operations and it instigated a large number of ciphers based on the idea of transposition and substitution. Transposition means changing the order of characters in a word while substitution means replacing a character with another one. Further development included polyalphabetic substitution instead of just mono-alphabetic i.e. if a character if repeated a number of times, it is replaced by different characters to make it more robust**[14]**. Additionally, new and fresh ideas were given for encoding like the Playfair cipher, where a matrix of English characters is formed in a 5 by 5 square (combining I and J together). Here, a particular word is used as a key which is written first and then the remaining letters were put in that matrix to form a dissimilarcipher text each time**[15]**. To put it in simple words, this was a substitution cipher which worked in a pair of words. This brilliant idea paved the way for other revolutionized ciphers like the Rail-fence cipher, Hill Cipher and Vigenere Cipher.

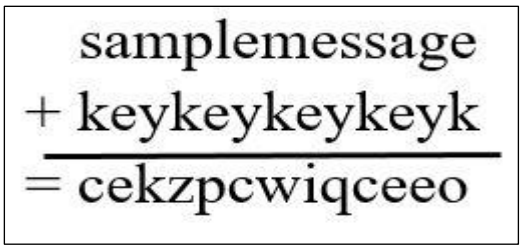

Fig 1: Example of VigenereCipher[11]

The Vigenere cipher was considered unbreakable at that time, but it also failed due to short length of the key and its repetition in accordance of the plaintext. This predicament gave birth to the idea of One-Time Pads (OTPs) which used random letters each time as the encryption key. The point here was to use a key one time only and destroy it after encryption. This sounds perfect in theory, but practically, it is hard to generate truly random characters or even numbers by humans and computers**[13]**.

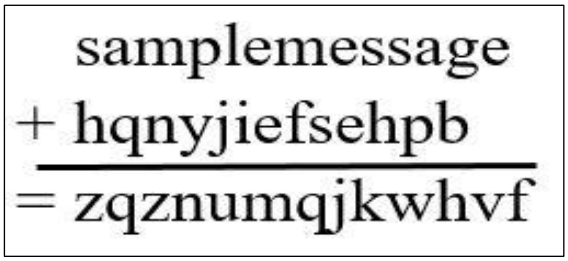

Fig 2: Example of One-Time Pad (OTP)[5]

The remainder of the paper is organized as follows. Section III gives some insight on the conventional cryptography and Section IV talks about the popular cryptographic domains i.e. Elliptic Curve Cryptography and Quantum Cryptography. The primary ideas DNA encryption are discussed in Section V and the next paragraph i.e. Section VI compares the contemporary cryptology with DNA cryptology. Section VII proposes the novel encryption algorithm and Section VIII simplifies the idea with an example. The performance analysis of the proposed cipher is done in Section IX, while conclusion and future work are proposed in Section X.

III. CONVENTIONAL CRYPTOGRAPHY

As discussed in the above section, classic cryptology was based on the characters in a human language and their manipulation. Due to the limited numbers of characters in any human language, this approach become brittle with the timeline and a new set of techniques was needed. The old era of classic cryptology was gone and new age of conventional cryptography started.

Key: FAUZANCE
01000110010000010101010101011010
00100000101001110010000110100010
Plaintext: DISASTER
01000100010010010101001101000010
10100110101010001000110101010010
Ciphertext: DISASTER
01010111101001010000010011011110
11000101011101100111000010101 1
Ciphertext: DISCSTER
11111011010101000100100100101111
11101110100001101001110101110111

Fig 3: Avalanche effect on DES[1]

This era marked the classic ciphers which used complex mathematical equations and mechanical devices like Enigma, BOMBE, etc. to produce highly efficient and perplex cipher texts, which were immune to the crypto attacks used earlier. The modern encryption ciphers use publicly known yet bamboozling mathematical algorithms and their secrecy is maintained through the secret key, which must be securely shared among the communicating parties**[16]**. Some of the popular ciphers of this age are DES, 3DES, etc. which mostly come under the symmetric category of ciphers i.e. they used one key only for encoding and decoding of data. Furthermore, the conventional ciphers are good exhibitors of the avalanche effect i.e. a small change in either the secret key or the plaintext causes a considerable change in the ciphertext. For example: changing the fourth character of the word "DISASTER" from 'A' to 'C' causes change in 35 bits which is quite significant, as shown in the figure above**[1]**.Later on in 1976, mathematical innovationby Whitfield Diffie and Martin Hellman gave birth to the idea of public key cryptography which revolutionized the security sector.The asymmetric algorithms based on the modular arithmetic uses has two different keys (per user): one for encryption and the other for decryption. This approach has two benefits: first the eavesdropper cannot decode the messages even if he/ she knows the mechanism and second the user just needs two keys for communication rather than having a lot of secret keys (one shared secret key between each pair of users).

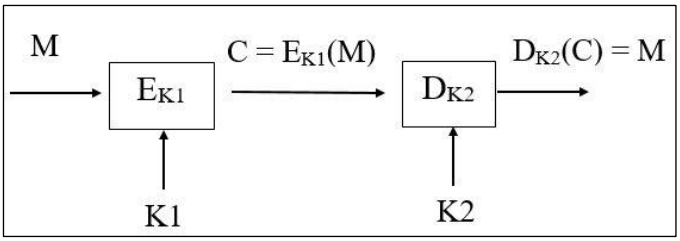

Fig 4: Block scheme of Public Key Cryptography[4]

Nevertheless the public key cryptography concept channeled the growth of other ideas like sharing secret key (Diffie-Hellman algorithms and its myriad versions), digital signatures to provide non-repudiation, digital watermarking for protecting intellectual property rights and one-way hash functions for providing authentication, which added further dimensions to the world of information security**[1]**. Additionally, the conventional cryptology can also be bifurcated on the basis of amount of data encoded each time the algorithms runs. If the encoding of data is done character by character, it is known as stream cipher and if the data is encrypted in chunks, it is called as block cipher. Obviously, the block ciphers have the upper hand over stream ciphers when it comes to speed and efficiency. Some popular block ciphers are DES, 3DES, IDEA, etc. and some well-known stream ciphers are A5, RC4, SEAL, Hughes XPD/ KPD, etc.

IV. ECC AND QUANTUM CRYPTOGRAPHY

As stated above, the public-key cryptology paved the way for future developments in the safe communication and one of the elite thought is Elliptical Curve Cryptography (shortened as ECC). ECC depends on the notion of using elliptical curves to generate the pair of private key and public key instead of using product of two extreme large prime numbers for the same. Elliptical curves are basically the binary curves which are equally proportioned over the x-axis and their maximum value is kept limited by finite field theory. The two primary concepts used for producing keys in ECC are Point Addition and Point Doubling. Another chief predicaments to remember is that elliptical curves prove to be slower and inaccurate when it comes to real number's calculation**[2]**.

ECC is highly beneficialover the traditional ciphers as it produces keys of smaller length yet robust and this key feature makes it more useful for mobile applications where low battery consumption and lesser computing power is desired. Another pluses of using ECC over well-known Public key algorithms like RSA are faster key generation, quicker hardware implementation andrapid encryption and decryption processes' computation**[3]**.

Another crypto variant used now-a-days is the Quantum Cryptography, which is using laws of nature to encode the messages at the physical layer and hence the need of security at upper layers is eliminated.

Instead of using bits as the atomic information unit, the quantum computers use "qubits" which is photon (by default). The '0' and '1' state of a photon depends upon its polarity either vertical or horizontal and its primary usage is key distribution rather than data encryption.

The most vital feature of quantum key distribution is its strength against the man-in-the-middle attack i.e. if the message is intercepted and then retransmitted by someone, then the qubits settle on a solo state and hence the recipient know not to trust the message as according to the Heisenberg's uncertainty principle the eavesdropper activity must cause an irrevocable modification in the quantum states**[4]**. Although this technique is a bit expensive at the moment, but it works excellent to encode the fiber optics network. Further, this technique is yet to prove its worth over larger distances and diminishing the higher error rates. Some highlighted applications of the quantum cryptology are encoded video calls, online voting in Switzerland and smart cards.

V. DNA CRYPTOGRAPHY

As the name suggests, DNA cryptography is the combination of DNA and encryption. This is a newly emerged specific set security system based on the idea of biology. It all started when Leonard Adleman used a biological approach to unravel the illustrious traveling salesman (or Hamilton) problem in 1994 and paved the way to handle complex mathematical problems**[17]**. This renowned research took the world like a twister and several other academicians contributed their part to prove that DNA computing can prove advantageous and myriad operations can be applied like DNA slicing, DNA polymerization, PCR amplification, etc. to give answers to the unknown problematical questions.

DNA encryption was first proposed by Ashish Gehaniet. al in 2004, when their research article in Springer enlightened the use of DNA principles in symmetric encryption, DNA chip and One-Time Pad generation**[5]**. Then the exploration in the arena of DNA cryptology expanded like a bubble like the YAEA algorithm implementation by Sherif T. Amin in 2006**[6]**, Asymmetric Key algorithm, DNA digital coding, PCR amplification concepts by Guangazhao Cui in 2008**[7]**, DNA chip based encryption by LaixueJia in 2010**[8],** etcto mention the few among illustrious research work done so far. All of the above research paper suggest one common thing i.e. the parallel processing provided by DNA computing offers exponentially fast solutions to the enigmatic and big mathematical problems.

VI. MODERN CRYPTOGRAPHY VS DNA CRYPTOGRAPHY

DNA cryptography is still in its infancy stage when we compare to modern cryptography and it will take some time and deep research to grow fully as a standalone application. Some basic difference between modern cryptology and DNA encryption can be understood by the following table:

TABLE I. MODERN CRYPTOGRAPHY VS DNA CRYPTOGRAPPHY[9]

	Operates on	Time Complexity	Storage Medium	Storage Capacity	Security
Modern Cryptography	Binary numbers	In seconds	Silicon chips	In GBs	Complex Maths
DNA Cryptography	DNA strands	In Hours	DNA chips	In TBs	Biology &Maths

As clearly elucidated from the table, a scholar can figure why DNA cryptography has not yet replaced the modern cryptography. Although DNA cryptology offers several advantages like huge storage capacity and parallel processing, but two important lacunas that need to be comprehend are the time complexity and its inability to work as a standalone alternative to modern encryption. DNA encoding rather serves as a support to the modern cryptography**[10]**.

VII. THE PROPOSEDCIPHER

The proposed encryption algorithm is based on the idea of amalgamation of mathematical perplexity and DNA Indexing.Usually, the indexing mechanisms are being used for data searching and processing, however here the DNA indexing is used to make the cipher stouter and harder to guess for the eavesdropper. For this, a chromosomal sequence is downloaded from any publicly available genomic databases like GenBank, NCBI, DDBJ, etc. This chromosomal sequence is made up of copious DNA nucleotide bases (A,C,G or T) as shown in the figure below:

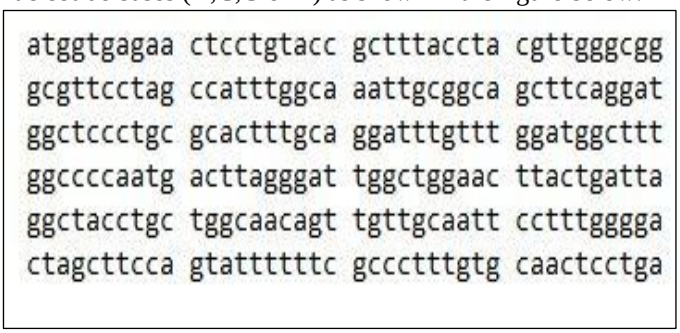

Fig 6: A fragment of DNA sequence downloaded from NCBI

According to DNA coding principle, one byte is transformed into a sequence of four DNA letters. Hence, a search is performed on every possible byte sequence and its various positions in the chromosomal sequence are stored in a table (as shown in Table II below):

TABLE II. KEY INDEXING OF GENETIC DATABASE

GGTA	58, 80, 249, 619, 645, 671, 896, 1197, 1605, 2766, 2958, 2972
AGAG	130, 161, 242, 453, 1011, 1442, 1458, 1512, 1997, 2295, 2789
AATA	27, 458, 611, 656, 924, 1059, 1332, 1518, 1521, 1539, 1584, 1647, 1695, 1698, 1734, 1767, 1770, 1885, 1933, 2166, 2225, 2365, 2401, 2625, 2700, 2754
AACT	271, 746, 1062, 1188, 1250, 1259, 1409, 1466, 1470, 1491, 1581, 1616, 1701, 1882, 1984, 2095, 2118, 2151, 2198, 2382, 2622, 2655, 2684
CTGC	10, 246, 366, 666, 1182, 1375, 1461, 1527, 1590, 1593, 1955, 2238, 2338, 2606, 2812, 2864
GGTG	521, 1754, 1877, 1992, 2422, 2531, 2618, 2675

Hence, for each byte of data, there are multiple position values which can be substituted one at a time and that too randomly. An important consideration here is to have a sufficiently long DNA sequence in order to obtain an extensive number of replacements for a byte. The encoding algorithm is primarily a stream cipher that works as follows:

1) The plaintext is read from input file.
2) The message is converted into ASCII code (or UTF-8).
3) Individual characters are arranged according to their number of occurrences and the total count of dissimiliar characters is calculated.
4) Each character is then transformed to binary '1' according to its occurrence in message and the previously recognized characters are stored as binary '0'.
5) The output of step 4 is divided bytewise and transposedto add complexity.
6) DNA coding is then applied to the data.
7) The output from previous phase is then encoded using DNA homophonic substitution cipher.
8) The enncoded data (in the form of integers)is then retransformed according to ASCII values (or UTF-8), which is the final ciphertext and is stored back in the file.

DNA Indexing ciphers are symmetric in nature that is encryption and decryption is done using same key and the process is reverse and identical. However, the receiver need to know identification number of the chromosomal sequence used as encoding key to complete the decryption process.

VIII.EXAMPLE OF THE PROPOSED ALGORITHM

As displayed in the figure 6 below, the decimal coding represent here ASCII or UTF-8 coding which transform the English characters to decimal values. Then the count of dissimilar characters and the frequency of each character is displayed. The next step illustrates the binary conversion of all characters. After binary conversion, the transposition of characters is done in the following way:

Input (000 01 100)

0	0	0
0	X	1
1	0	0

Output (001 00 010)

As displayed, the input data is stored in a 3*3 matrix (keeping the center cell empty to accommodate 8 bits in 9 cells). The data is stored row-wise, but read column-wise to add another step in complexity. After operations at digital level, DNA coding is implemented. It can be done in multiple ways; one simple way is to encode using following table:

TABLE III. DNA CODING

Binary Code	DNA Chromosome
00	a
01	c
11	t
10	g

The intermediate output received after DNA encoding is then subjected to homophonic substitution, which is done according to the chosen index positions of the 4-chromosomal DNA as shown in the table II. The key sequence indicate the length of position number of the substitution done in the previous step. Finally, the cipher text is achieved after reconversion according to the decimal coding.

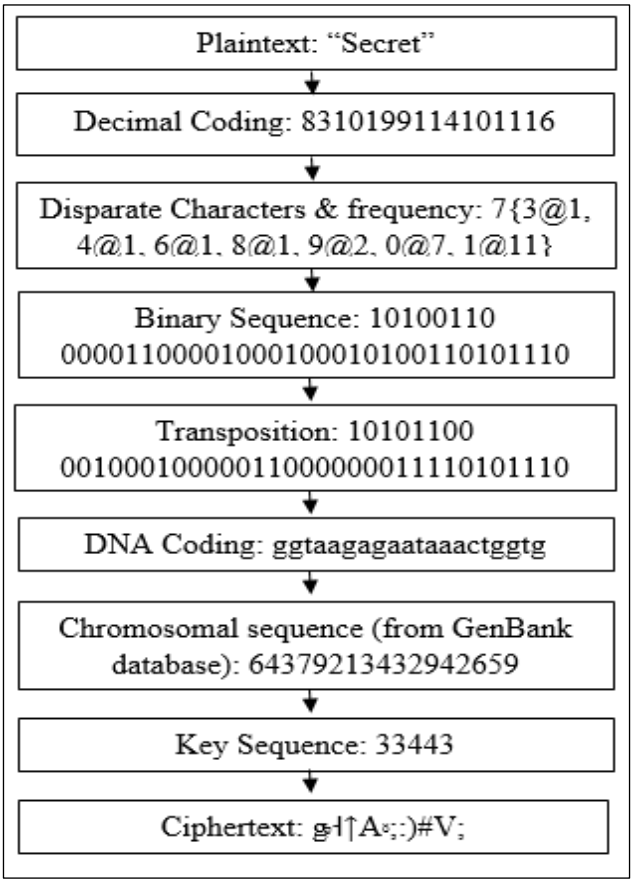

Fig. 6: Working of proposed cipher

IX. PERFORMANCE ANALYSIS

The performance of cryptosystems can be evaluated on two vital criterias: security and computational complexity. The proposed encryption algorithm is based on the theory of homophonic substitution provided by the DNA Indexing. The security provided in the proposed algorithm is three – fold. Firstly, the concept of modern cryptography (i.e. perplex mathematical function) is applied here which transmutes decimal numbers to the binary sequence depending upon its frequency. Secondly, the binary transformations are done to alter the binary sequences generated earlier. This binary operations add extra layer of refuge to the encryption system. Last but not the least, the DNA coding and chromosomal sequence from a genetic database to further mystify the encoded data. These three layers of encryption should be robust enough for safeguarding small amount of data. The homophonic substitution in VG1 cipher offers more uniform distribution of the Ciphertext and will be not show analogy to the plaintext distribution.

Another security aspect of proposed cipher is its robustness against the Ciphertext-only attacks as well as known-plaintext attacks due to the fact that each plaintext value corresponds to not just one, but numerous Ciphertext values. The brute-force attack against VG1 cipher can be possible, but a number of factors need to be considered. Firstly, the genetic sequence must be identified, which is kind of herculean task given plenty of publicly available genetic databases and plethora of DNA sequences within each (at least a million). For example: Suppose NCBI has genetic sequence which has at least 45000 nucleotide bases and each sequence is composed of A,C,G & T, then possible number of keys become 4^{45000}, which can prove to be a colossal job for hacker. Moreover, the binary operations also need to be taken care of. A vital point for deliberation here is that instead of sharing a long DNA sequence (already available on the internet), only it's ID from the genetic database needs to be shared and for the same, we can use any public-key cryptosystem. The computational complexity of an algorithm depends upon two major things: time complexity and memory requirements. The time complexity of the algorithm depends primarily on two factors: encryption and decryption time taken by the algorithm. However, importing the chromosomal sequence from file and then DNA indexing is a much bigger task and consumes the maximum amount of time.This time is nearly 10 times the time to encrypt the whole file. The total time taken to import the DNA Indexing file and to encode the input file of different file sizes is given below:

TABLE IV. TIME COMPLEXITY FOR VG1 ENCRYPTION PROCESS

Plaintext Size (in KiloBytes)	Execution Time (seconds)
256	13.27
512	28.59
768	40.23
1000	64.65

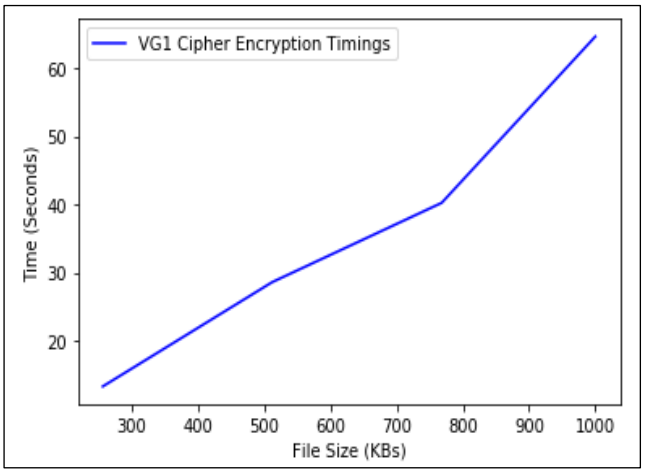

Fig. 7: Encoding Time Calculation of VG1 cipher

The total time taken to import the DNA Indexing file and to decode the received file of different file sizes is given below:

TABLE V. TIME COMPLEXITY FOR VG1 DECRYPTION PROCESS

Plaintext Size (in KiloBytes)	Execution Time (seconds)
256	15.74
512	30.11
768	41.68
1000	66.63

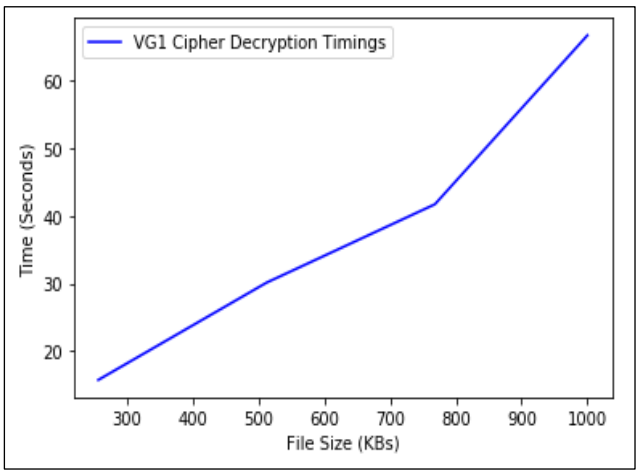

Fig. 8: Decoding Time Calculation of VG1 cipher

As evident from the above graphs, there is a positive correlation between file sizes and for encoding and decoding timings of VG1 cipher, which means the execution time upsurges with growing file sizes. Encryption and Decryption timings has been tested on a machine using Intel i5 2.2GHz processor, 4 GB RAM and Windows 10 Pro Operating System. The programming language used to implement VG1 cipher is Python version 3.7.3 64-bit using Spyder IDE.

X. SUMMARY & FUTURE SCOPE

In the world of cybercrimes and online scams, a new defense is needed to guard the vital information and once again a new form of cryptography emerges for the rescue. This new tool is called DNA Cryptology and it is derived from the roots of computing and biology. This paper give some insights on DNA encoding and designate how it can be used as a side-kick to the modern cryptography to safeguard the significant information. This paper also proposes a new encryption algorithm "VG1 Cipher" which provides a novel approach to DNA Indexing methodology which is not only fast but also produces highly secure ciphertext. The proposed algorithm is based on DNA Indexing and Modern Cryptography, which makes it more efficient and faster than its counterparts.

The next step chromosomal sequence signify the homophonic substitution taken from the genetic database downloaded from either GenBank or NCBI i.e. there are several positions. The key sequence indicate the length of position number of the substitution done in previous step. Finally, the cipher text achieved at the end is shown as the final step.

REFERENCES

1. M. Kumar et. al., "Comparing Classical Encryption with Modern Techniques", S-JPSET, Vol. 1, Iss. 1, Dec. 2010, pp. 49-54.
2. J. Zargar, M. Manzoor & T. Mukhtar, "Encryption/ Decryption using using Elliptical Curve Cryptography", International Journal of Advanced Research in Computer Science, Vol. 8, Iss. 7, July 2017, pp. 48-51.

a. Ibrahim, W. Cheruiyot & M. W. Kimwele, “Data Securityin Cloud Computing withEllipticCurveCryptography”, International Journal of Computer, Vol. 26, Iss. 1, 2017, pp. 1-14
3. R. J. Hughes et. al., “Quantum Cryptography”, Contemp. Phys. 36, 1995, pp. 149-190.
4. Gehani, T. LaBean, and J. Reif, “DNA-Based Cryptography”, Lecture Notes in Computer Science 2950, Springer, pp. 167-188, 2004
5. S.T. Amin, M. Saeb, S. EI-Gindi , "A DNA-Based Implementation of YAEA Encryption Algorithm", Computational Intelligence, pp. 120-125, 2006.
6. G. Cui et. al., “An Encryption Scheme Based on DNA Microdots Technology”, Journal of Computational and Theoretical Nanoscience, Vol. 12, Iss. 7, pp. 1434-1439, 2015.
7. L. XueJia et. al., “Asymmetric encryption and signature method with DNA technology”, Science China Information Sciences, Vol. 53, No. 3, pp. 506-514, 2010.
8. S. Karthiga & E. Murugavalli, “DNA Cryptography”, International Research Journal of Engineering and Technology, Vol. 5, Iss. 3, Mar 2018, pp. 3987-3991.
9. S. Kalso, H. Kaur & V. Chang, “DNA Cryptography and Deep Learning using Genetic Algorithm with NW algorithm for Key Generation”, Springer Journal of Medical System, Vol. 42, Iss. 17, Oct. 17.
10. E. P. Dummit, “A Tour of Classical and Modern Cryptography”, University of Rochester, 2015, an online article available at https://math.la.asu.edu/~dummit/docs/talk_sums_cryptography_talk.pdf.
11. A. Kahate, “Cryptography and Network Security”, Tata McGraw Hill, New Delhi, India, 2012.
12. P.P Charles & P.L Shari, "Security in Computing: 4th edition", Prentice-Hall, lnc.,2008.
13. W.Stallings, "Cryptography and Network Security Principles and Practice," Fourth edition, Prentice hall, 2007.
14. J. Katz and Y. Lindell, Introduction to Modern Cryptography: Principles and Protocols, 1st ed. USA: Chapman & Hall/ CRC, 2007.
15. A.S. Tanenbaum, "Computer Networks", Fourth Edition, Prentice hall, 2004.
16. J. Hoffstein, J. Pipher & J.H. Silverman, An Introduction to Mathematical Cryptography, 1st ed. USA: Springer, 2010.

AUTHORS PROFILE

Akhil Kaushik has received the Master degree in Information Technology from Central Queensland University, Melbourne, Australia. Currently he is pursuing his doctorate degree in CSE Department of Amity University, Gurugram, Haryana, India. He has more than 10 years of teaching experience and nearly 6 years of research experience with contribution at International level in various proceedings like IEEE, IJCEE, ICFN, ICNIT, etc. His research interest includes network security, cryptography and machine learning.

Dr. Vikas Thada has received the doctorate degree from Dr. KNM University, Jaipur, India Currently he is employed as Associate Professor in CSE Department, Amity School of Engineering & Technology, Amity University, Gurugram, Haryana, India. He has more than 16 years of teaching experience and over 7 years of research experience. He has numerous publications in international journals and is author of number of books on programming, data structures, algorithms etc. His chief research interests are genetic algorithm, cryptography and network security, design and analysis of algorithm and web technologies.