Emerging DNA cryptography-based encryption schemes: a review

Pratyusa Mukherjee* and
Chittaranjan Pradhan

School of Computer Engineering,
KIIT (Deemed to be University),
Patia, Bhubaneshwar, Odisha, India
Email: pratyusa.mukherjee@gmail.com
Email: chitaprakash@gmail.com
*Corresponding author

Rabindra Kumar Barik

School of Computer Application,
KIIT (Deemed to be University),
Patia, Bhubaneshwar, Odisha, India
Email: rabindra.mnnit@gmail.com

Harishchandra Dubey

Center for Robust Speech Systems,
The University of Texas at Dallas,
Richardson, TX 75080, USA
Email: harishchandra.dubey@utdallas.edu

Abstract: Security has been the fundamental apprehension during information transmission and storage. Communication network is inordinately susceptible to intrusion from unpredictable adversaries thus threatening the confidentiality, integrity and authenticity of data. This is where cryptography facilitates us and encodes the original message into an incomprehensible and unintelligible form. DNA cryptography is the latest propitious field in cryptography that has transpired with the advancement of DNA computing. The immense parallelism, unrivalled energy efficiency and exceptional information density of DNA molecules is being traversed for cryptographic purpose. Currently, it is in the preliminary stage and necessitates avid scrutinisation. The foremost hindrance in the field of DNA cryptography is computational complexity and lack of sophisticated laboratories. In this paper, we discuss the existing DNA cryptographic approaches and compare their achievements and limitations to provide a better perception. In the end, a modified version of the DNA cryptography combined with soft computing is also suggested.

Keywords: security; DNA cryptography; DNA computing; bio-inspired cryptography; encryption.

Reference to this paper should be made as follows: Mukherjee, P., Pradhan, C., Barik, R.K. and Dubey, H. (2023) 'Emerging DNA cryptography-based encryption schemes: a review', *Int. J. Information and Computer Security*, Vol. 20, Nos. 1/2, pp.27–47.

Biographical notes: Pratyusa Mukherjee is a PhD Scholar of School of Computer Engineering, KIIT Deemed to be University, Bhubaneshwar, India. She received her BTech in Electronics and Communication Engineering from West Bengal University of Technology and MTech in Information Technology from Indian Institute of Engineering Science and Technology, Shibpur. Her research area is in the field of information security. It includes chaos-based encryption, lightweight block cipher, blockchain technology and DNA cryptography.

Chittaranjan Pradhan has obtained his Doctorate, Master's and Bachelor's in Computer Science and Engineering discipline. Currently, he is working as an Associate Professor at School of Computer Engineering, Kalinga Institute of Industrial Technology (KIIT) Deemed to be University, Bhubaneswar, India. He has received a total of 14 years of academic teaching experience with more than 70 publications in reputed and peer reviewed journals, edited books and conferences of national and international repute. His research area includes information security, image processing, deep learning, and multimedia systems. He has published few books published by publishers like LAP Lambert, IGI Global, and Elsevier. He is also member of various national and international professional societies in the field of engineering and research such as: IET, IACSIT, CSI, ISCA, IAENG, and ISTE.

Rabindra Kumar Barik is currently working as an Assistant Professor in School of Computer Applications, KIIT Deemed to be University, Bhubaneswar, India. He has received his both MTech and PhD from Motilal Nehru National Institute of Technology Allahabad, Prayagraj, India in 2009 and 2014, respectively. He has received best paper awards in FICTA-2020, ICSCC-2017 and ICECE-2017 conferences. His research area includes geospatial data science, geospatial big data infrastructure, geospatial database, geospatial cloud computing, fog computing and IPR. He has published more than 20 international journals and more than 30 conference papers in various top-level.

Harishchandra Dubey is interested in audio, speech and language processing, machine learning, fog and cloud computing, and social signal processing. He received his Master of Science from FAU University of Erlangen-Nuremberg, Germany in 2015 and received his Bachelor of Technology in Electronics and Communication Engineering from Motilal Nehru National Institute of Technology, Allahabad, India in 2012. Presently, he is working as a Machine Learning Expert in Microsoft. He has completed his PhD from The University of Texas at Dallas, USA.

1 Introduction

With the flourishing advancements of technology day by day, enormous amount of information is available over the internet. Several adversaries are always vigilant to intercept this crucial information and shatter its integrity. Consequently, utmost caution has to be taken in order to maintain the security and confidentiality of this information. Cryptography and steganography are commonly used techniques to protect the data. Steganography means 'covered writing' and conceals the secret message into a forged message. Cryptography signifies 'secret writing' and encrypts the original message into an unintelligible and incomprehensible form before transmission. Here, encryption and

decryption are performed with the help of a key and entire security is based on it. Only the intended authorised receiver has knowledge about these pronounced keys and thereby only he can unravel a cipher text and attain the original message.

The most frequently used technique is that of modern cryptography which is based on laborious mathematical numerical and entire security is based on the secret key. Even if the intruder is aware of the algorithm used for coding, the computational toil and no prerequisite idea about the key makes it impossible to crack the original message. Nowadays, DNA cryptography is unfolding as a new technique for encryption which depends upon rigorous biological problems. Here, DNA is used to store and transmit data and modern biological mechanisation is used as the execution tool.

In this inspection paper, in Section 2, a concise insight into important terminologies related to DNA cryptography like the biological anatomy of DNA, basics of central dogma with the flourishing advancements of technology day by day, enormous amount of information is available over the internet. Several adversaries are always vigilant to intercept this crucial information and shatter its integrity. Consequently, utmost caution has to be taken in order to maintain the security and confidentiality of this information. Cryptography and steganography are commonly used techniques to protect the data. Steganography means 'covered writing' and conceals the secret message into a forged message. Cryptography signifies 'secret writing' and encrypts the original message into an unintelligible and incomprehensible form before transmission. Here, encryption and decryption are performed with the help of a key and entire security is based on it. Only the intended authorised receiver has knowledge about these pronounced keys and thereby only he can unravel a cipher text and attain the original message.

The most frequently used technique is that of modern cryptography which is based on laborious mathematical numerical and entire security is based on the secret key. Even if the intruder is aware of the algorithm used for coding, the computational toil and no prerequisite idea about the key makes it impossible to crack the original message. Nowadays, DNA cryptography is unfolding as a new technique for encryption which depends upon rigorous biological problems. Here, DNA is used to store and transmit data and modern biological mechanisation is used as the execution tool.

In this inspection paper, in Section 2, a concise insight into important terminologies related to DNA Cryptography like the biological anatomy of DNA, basics of central dogma and genetic coding, DNA computing and bio computers have been proffered. The nitty-gritty of DNA cryptography and all its related technologies, operations and coding stratagem are also bestowed. Section 3, gives a comprehensive survey of existing ideologies in this field categorised into three broad categories and Section 4 analyses each of their achievements and loopholes. In Section 5, a rough diagram of a modified version of DNA cryptography by combining it with soft computing has been proposed. Section 6 talks about the conclusion and subsequent scope of work.

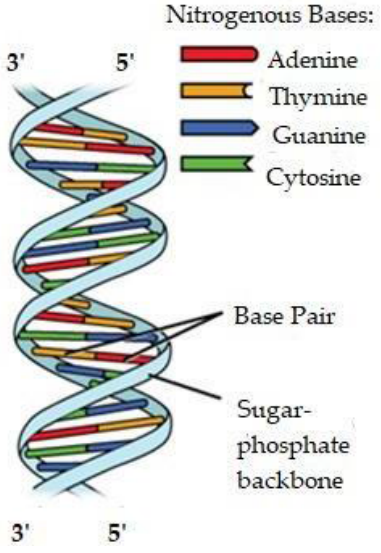
2 Foundations of DNA cryptography

2.1 Biological anatomy of DNA

DNA or Deoxyribonucleic acid is the hereditary material and information carrier of all life forms. It is composed of smaller fragments called nucleotides. The nucleotides are in turn embodied with three ingredients – nitrogenous base, five carbon sugar (deoxyribose)

and phosphate group. The nitrogenous bases are of four kinds. Adenine (A) and guanine (G) are purines, and cytosine (C) and thymine (T) are pyrimidines. The phosphate group of one nucleotide fuses covalently with the sugar molecule of the successive nucleotide. The two Nobel laureate Watson and Crick (2003) proposed that the DNA comprises of two strands that are twisted around each other, one ranging from 3' to 5', the other from 5' to 3'to appear as a right-handed helix, and hence is called a double helix. A purine always couples with a pyrimidine: namely, A bonds with T, similarly G with C. Hence, A and T, C and G are complementary pairs. Consequently, it is termed as a Watson-Crick complementary structure as in Figure 1.

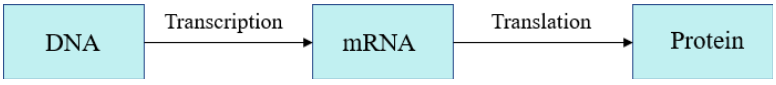
Figure 1 The double helical structure of DNA (see online version for colours)



2.2 Central dogma and genetic coding

Crick (1970) pioneered the central dogma which endows a successional clarification of the flow of familial information within a biological system. Central dogma is the comprehensive process of transforming DNA nucleotides into proteins as in Figure 2. It enables the information contained in genes to surge into the proteins which are the building blocks of life. Transcription is the procedure of synthesis of mRNA segment from a DNA segment. mRNA is similar to DNA because both consists of a long sequence of nucleotides. mRNA differs from DNA as it is single-stranded, constitutes of sugar ribose and utilises the base uracil (U) in place of thymine (T). Each DNA segment has three main parts: a promoter or starting point, introns, i.e., the non-coding regions and exons which correspond to the coding regions. During transcription, DNA segment is read from its promoter, the introns are annulled and exons are reunited. This resulting chain sequence is then transcribed into mRNA. Translation refers to the process of creating amino acid sequence of proteins from the mRNA. This analogy of the central dogma of molecular biology can be applied to the field of cryptography where we can presume the DNA to be our original message and the protein to be the corresponding cipher text.

Figure 2 The central dogma process (see online version for colours)



Crick (1968) proposed the genetic code which is a DNA code written as triplet combination of three of the bases namely A, C, T and G which are named as codons. Each codon uniquely represents an amino acid that is the building block of proteins. Three 'stop' codons and one 'start' codon denote the end and the beginning of a protein respectively. Since there are three-letters codons combinations and four distinct bases, 64 possible codons are possible which in turn cipher the 20 amino acids, thus providing redundancy to each amino acid being ciphered by multiple codons. This full set of relationship between codons and amino acids is termed as the genetic code and is summarised in Figure 3.

Figure 3 DNA to amino acid genetic code table (see online version for colours)

First Base	Second Base: U	Second Base: C	Second Base: A	Second Base: G	Third Base
U	UUU, UUC: Phe	UCU, UCC, UCA, UCG: Ser	UAU, UAC: Tyr	UGU, UGC: Cys	U
	UUA, UUG: Leu		UAA: Stop	UGA: Stop	C
			UAG: Stop	UGG: Trp	A
					G
C	CUU, CUC, CUA, CUG: Leu	CCU, CCC, CCA, CCG: Pro	CAU, CAC: His	CGU, CGC, CGA, CGG: Arg	U
			CAA, CAG: Gin		C
					A
					G
A	AUU, AUC, AUA: Ile	ACU, ACC, ACA, ACG: Thr	AAU, AAC: Asn	AGU, AGC: Ser	U
	AUG: Met or Start		AAA, AAG: Lys	AGA, AGG: Arg	C
					A
					G
G	GUU, GUC, GUA, GUG: Val	GCU, GCC, GCA, GCG: Ala	GAU, GAC: Asp	GGU, GGC, GGA, GGG: Gly	U
			GAA, GAG: Glu		C
					A
					G

2.3 DNA computing

DNA computing is a bifurcation of computing which utilises concepts of DNA, biochemistry and molecular biological hardware, in replacement of the customary silicon-based computer mechanisms. Adleman (1994) first demonstrated the use of DNA as a tool of calculation which unfolded the seven-point Hamiltonian path problem. He solved the specimen of a graph composing of seven vertices by enciphering it into the

molecular form based on an algorithm. This was followed by performing computations with the help of some standardised enzymes and finally administering brute force method. Lipton (1995) followed up on Adleman's work by showing how the satisfiability problem may also be solved using a similar approach. Boneh et al. (1995) demonstrated a break of data encryption standard (DES) by using molecular operations by recovering the key from a given arbitrary plaintext-ciphertext pair in mere four months. In 1997, an interesting demonstration of simulation of Boolean circuits using DNA computers was provided by some researchers. In 2002, researchers from Israel designed a programmable molecular computer consisting of enzymes and DNA molecules rather than silicon microchips.

2.4 DNA computer

DNA computers (Shapiro and Benenson, 2006) or bio computers utilise biotically acquired molecules like DNA, proteins and enzymes to accomplish computations involving processing, storing and retrieving data. The input, system and output of such computers are DNA molecules and they use biochemical and biological techniques hardware instead of silicon chips. Here, the DNA can be assumed to be the software and enzymes act as the hardware. When they are assembled together in a test-tube, the manner in which they undergo chemical reactions allows other basic operations to be conducted as a byproduct. Controlling the composition of DNA molecules enables to perform distinguished operations.

The main convenience of using DNA computers to expound complicated problems is that all the distinct probable solutions can be created at one go. This feature is known as parallel processing. DNA is a very cheap resource because as long as living organisms exist, there will be abundance of DNA. DNA biochips contain no toxic material and hence are environment friendly. DNA caches memory at a density of about 1 bit per cubic nanometre, whereas traditional storage mediums stores 1 bit in 10^{12} cubic nanometres, hence, DNA have minimal storage requirement. DNA has no power requirement as chemical reactions happen sans any outside power requirement. Information density of DNA is very high as 1 gram of DNA stores 10^8 TB of data. Therefore, DNA computers are very small – as small as the size of a teardrop. DNA computers also have certain limitations like accuracy as it is highly dependent on accuracy of enzymes involved. Erroneous enzymes will lead to mismatched pairs and hence DNA computers might become susceptible to errors. Each stage of parallel processing is time consuming, with substantial human or mechanical intercession amidst each step. Since a particular set of DNA strands is attuned to deal with a particular problem, a new set would have to be formulated for each new problem.

2.5 DNA cryptography

DNA cryptography (Gehani et al., 2003) is grounded on DNA computing where the original message is enciphered in the form of a DNA nucleotide using either a symmetric key or an asymmetric key. It uses modern biological technology as the implementation tool. Traditional cryptography dates back to Caesar cipher almost 2000 year ago whereas DNA cryptography is still in its infancy stage where it is being explored to utilise its advantages to the fullest. An interloper with immense computation power can break the traditional cryptosystems theoretically. The entire security of DNA cryptography relies

on biological techniques and hence have no computation involved and are immune to attacks.

DNA cryptography can be implemented by using classical or biologically inspired computation techniques or a combination of both.

2.5.1 Biologically inspired DNA techniques

Several biologically inspired techniques are used to perform DNA computation on instructive DNA nucleotides. Table 1 gives a detailed description of some of these techniques along with their merits and demerits.

Table 1 Biologically inspired DNA techniques advantages and limitations

Technique	*Advantages*	*Limitations*
Gel electrophoresis (Lewis, 2001) Method for separation and analysis of DNA fragments based on their length and charge by applying electric current on polyacrylamide or agarose gel.	• Simple • Comparatively inexpensive • Small quantity of gel is adequate	• Excessive heat is generated • Molecules must be compulsorily indistinguishable in size • External power supply is needed
Polymerase chain reaction (PCR) (Mullis and Faloona, 1987) PCR enables us to quantify a single copy of a DNA sequence into millions of more copies by loading two primers at the beginning and at the end of target DNA.	• Eases handling small amount of DNA by amplifying them • Small amount of genetic material is sufficient to work with • Highly specific procedure as only target amplified	• Highly sensitive technique because amplification occurs only primers match • Only amplifies a very limited portion of the target strand
DNA chip technology (Lemieux et al., 1998) DNA chips enable to store data as DNA sequences.	• High throughput • Many strands can be experimented with at one go	• Expensive to create. • Time consuming • DNA chips have short shelf life
DNA fragment assembly (Gibson, 2011) Technique for aligning and amalgamating fragments of DNA sequences to rebuild the aboriginal one	• Faster than gel electrophoresis • No scars remain after assembly hence fragments cannot be differentiated. • Rejects important but redundant fragments	• Difficult to be implemented
DNA splicing (Freund et al., 1999) Process of slashing of DNA fragments from one sequence and affiliating onto another. The result is recombinant DNA that incorporates features of the host sequence modified by the trait in the foreign sequence.	• Enhances the diversity of the host DNA sequence	• Might lead to deletion of important portions of the host sequence and redundant portions of foreign sequence being affiliated to it

Table 1 Biologically inspired DNA techniques advantages and limitations (continued)

Technique	*Advantages*	*Limitations*
DNA hybridisation Method of merging two complementary single stranded DNA and permitting them to form double-stranded molecules through a base pairing process.	• Enables the favourable traits to sustain for longer durations • Gives better results than simple mutation	• Specialised laboratories are needed for practical implementation

2.5.2 Computationally inspired DNA techniques

Within DNA cryptography techniques, computational operations like arithmetical, mathematical, and logical operations can also be performed. It first uses the DNA coding scheme and then deals with the DNA sequence as a series of zeros and ones. After which, any computational operation can be applied to perform the encryption.

2.6 DNA encoding

The most fundamental coding pattern to encipher the four nucleotide bases (A, T, C, G) is 0(00), 1(01), 2(10), 3(11) respectively as in Table 2. According to the Watson-Crick base pairing rules, A and T are complementary and C and G are complementary. Since four digits are involved 4! = 24 patterns are possible. But, among these 24 patterns, only 8 of them fit the complementary rule. It is recommended that the coding pattern in conformance with the sequence (0123/CTAG) is the best as in Table 3.

Table 2 Simplest DNA encoding scheme

Nucleotide base	*Decimal representation*	*Binary representation*
A	0	00
T	1	01
C	2	10
G	3	11

Table 3 DNA encoding scheme according to Watson-Crick complementary rule

	A	*T*	*C*	*G*
Rule 1	00	11	10	01
Rule 2	00	11	01	10
Rule 3	11	00	10	01
Rule 4	11	00	01	10
Rule 5	10	01	00	11
Rule 6	01	10	00	11
Rule 7	10	01	11	00
Rule 8	01	10	11	00

2.7 DNA subtraction, XOR, addition, operation

XOR, addition and subtraction operations among DNA nitrogenous bases is illustrated below in Table 4.

Table 4 DNA subtraction, XOR and addition operation

-	A	T	C	G	⊕	A	T	C	G	+	A	T	C	G
A	A	G	C	T	A	A	T	C	G	A	T	G	A	C
T	T	A	G	C	T	T	A	G	T	T	G	T	T	A
C	C	T	A	G	C	C	G	A	C	C	A	C	C	T
G	G	C	T	A	G	G	C	T	A	G	C	A	G	G

2.8 DNA coding scheme

DNA coding scheme takes an input which is usually an alphabet or a number or a special character. It then converts the input characters into their corresponding DNA codes based upon a pre-decided code set that is already fed into the DNA processing unit as in Figure 4. This coding scheme is shared amongst the sender and receiver through a secure communication channel. An illustration is in Table 5 here.

Figure 4 DNA coding scheme processing unit (see online version for colours)

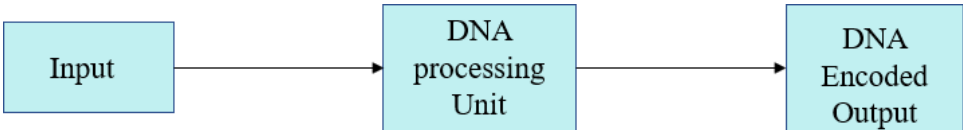


Table 5 DNA encoding scheme

A = CGA	B = CCA	C = GTT	D = TTG	E = GGC	F = GTT	G = TTT	H = CGC
I = ATG	J = AGT	K = AAG	L = TGC	M = TCC	N = TCT	O = GGA	P = GTG
Q = AAC	R = TCA	S = ACG	T = TTC	U = CTG	V = CCT	W = CCG	X = CTA
Y = AAA	Z = CTT	0 = ACT	1 = ACC	2 = TAG	3 = GCA	4 = GAG	5 = AGA
6 = TTA	7 = ACA	8 = AGG	9 = GCG	= ATA	, = TCG	. = GAT	: = GCT

3 Related work

The prime question that arises is why DNA-based encryption rather than traditional digital encryption. The explanation is that majority of the traditional cryptosystems have been partially broken, even though not fully with the assistance of modern generation super computers. The most facile and elementary method to break any encryption mechanism is 'brute force attack', where the intruder exhaustively checks all possible keys and secret parameters until the correct one is found. Binary keys comprise of only 0 and 1, hence have the exponential power to be two. On the contrary, DNA codes have exponential power of four due to four bases – A, T, C and G. This makes a single bit key eight times stronger using DNA encryption. Also, the complexity and randomness of

DNA nucleotides, high storage capacity and parallel processing, provide additional security. Thus, the concept of incorporating DNA into cryptography has been recognised as a conceivable technology that brings a new aim to enhance the robustness of algorithms.

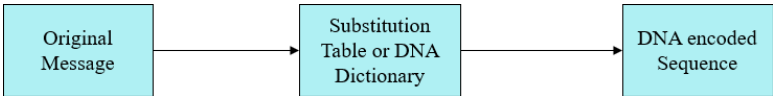
In this paper, all the existing DNA cryptographic schemes has been broadly categorised into three categories depending upon the types of algorithms used for the encryption.

3.1 Simple substitution-based DNA cryptographic schemes

These algorithms perform the entire encryption on the basis of certain pre-decided look-up table or DNA dictionary. They are the most non-complicated and simple DNA encryption techniques. Figure 5 represents the basic block diagram of these algorithms.

Sabry et al. (2012) illustrated a table where each letter from A-Z is associated with amino acids using three techniques namely discrete encoding, overlapping encoding, embedded DNA encoding. Jain and Bhatnagar (2014) suggested a method that tabulates 256 decimal numbers and their corresponding DNA sequence as a DNA dictionary. Hameed and Al-Ani (2018) proposed an encryption scheme for coloured images that disintegrates RGB image into its three components, encrypts them separately. Encrypted components are converted into DNA sequence based on a previously decided substitution table. DNA addition is performed and output is reconverted to RGB form to get final encrypted image. Nandy et al. (2018) propounded an encryption algorithm to encrypt the input image in the form of a text file using DNA cryptography by representing each pixel of the image as a combination of 'ATCG'. The content of text file is the encrypted form of the original image.

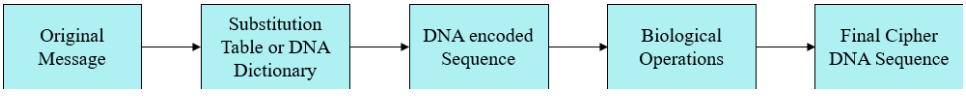
Figure 5 Block diagram of simple substitution-based DNA cryptographic schemes (see online version for colours)



3.2 Biological operations-based DNA cryptographic schemes

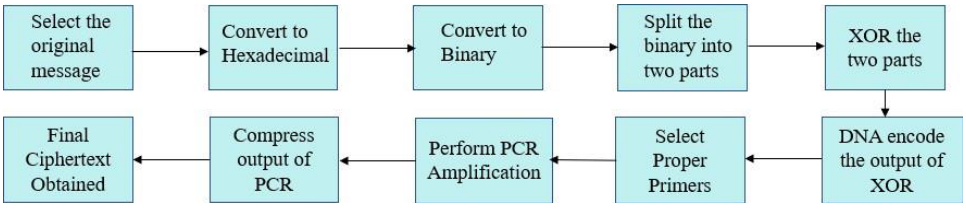
This category of algorithms uses only rigorous biological operations to perform the encryption procedure. They require minimal human intervention and hence are comparatively more secure. Figure 6 gives the block diagram of this category. Initially, using simple substitution technique, the DNA encoded form is obtained. Biological operations are then applied on them to get the final cipher DNA sequence form.

Figure 6 Block diagram of biological operations-based DNA cryptographic schemes (see online version for colours)



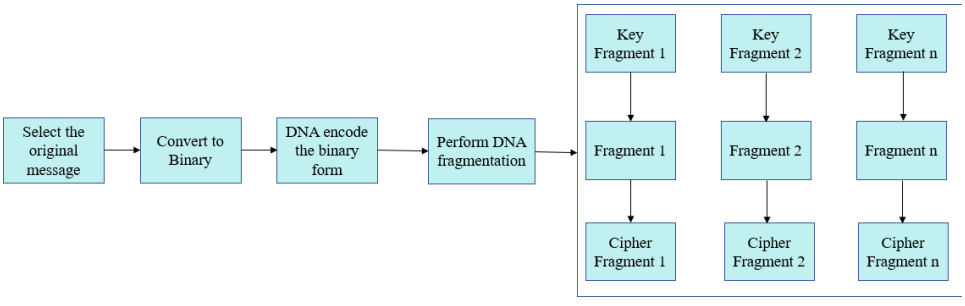
Wang and Zhang (2009) proposed an asymmetric encryption algorithm were security is obtained first through known cryptographic algorithm followed by biological operation like polymerase chain reaction (PCR) to amplify the DNA strands. Although, it furnishes a two-level enhanced security, it also leads to computational burden by aggravating the time complexity. Ning (2009) propounded an encryption scheme where data is initially converted into DNA form followed by translating it to the protein form. Since this scheme is mainly utilising biological operations, it is called pseudo-DNA cryptography. Prabhu and Adimoolam (2011) proposed a scheme where double layer security is attained through various operation performed serially as illustrated in Figure 5. Hence, it is called bi-serial encryption algorithm.

Figure 7 Block diagram of bi-serial encryption scheme suggested by Prabhu and Adimoolam (2011) (see online version for colours)



Dhawan and Saini (2012) included the concept of microdoting along with DNA digital encoding and PCR. Pramanik and Setua (2012) propounded a scheme that utilises the technique of DNA hybridisation. Zhang et al. (2012) proposed an encryption method that converts plaintext into fragmented ciphertext as depicted in Figure 6. On receiver side, fragment reassembly is done to obtain the plaintext.

Figure 8 Block diagram of encryption scheme suggested by Zhang et al. (2012) (see online version for colours)

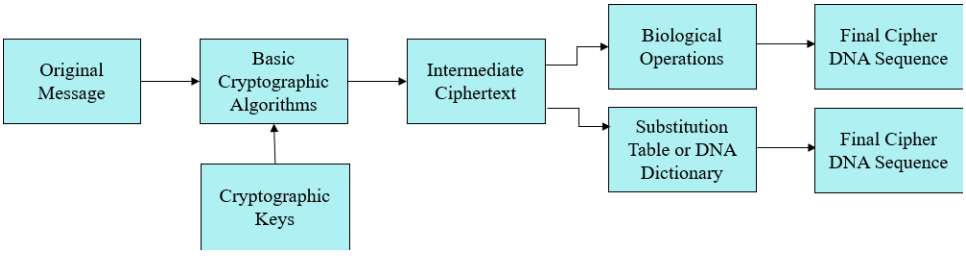


Zhang et al. (2016) proposed a new DNA Cryptography algorithm based on the bio puzzle and DNA chip technology. The DNA chips and keys are finally sent to the receiver. Wang et al. (2019) proposed a technique to hide messages in living organisms using DNA encoding as well as DNA recombinant technology. First, the secret message is DNA encoded and then along with other fake DNA sequences it is inserted into a host cell. This procedure thus makes the cracking of the message by an evildoer difficult as he has to break the double layer of security.

3.3 Mathematical – biological operations-based DNA cryptographic schemes

The third category of algorithms comprises of those that utilise both mathematical as well as biological operations to perform the encryption. Mathematical operations involve the usage of symmetric or asymmetric cryptographic keys and the biological operations provide an additional layer to them, thus making them the most secure DNA-based technique. Some algorithms also simply DNA encode the intermediate ciphertext on the basis of substitution techniques. The same has been illustrated in Figure 9.

Figure 9 Block diagram of mathematical – biological operations-based cryptographic schemes (see online version for colours)



Li et al. (2008) have proposed a key expansion operation using random DNA sequence selected from gene bank which enhances the security and independence of keys. Vijayakumar et al. (2011, 2013) suggested a modification to conquer this time restraint by applying the same idea with hyper elliptic curve cryptography. Sadeg et al. (2010) illustrated new concept of Bio-XOR thereby combining both biological and combinational techniques. Zhang et al. (2011) suggested an index-based symmetric DNA encryption algorithm where the original message is encrypted with an initial key then the output is DNA encoded. Finally, a special DNA sequence is selected from this DNA code. Within this sequence a random position is selected and that will be used for indexing. This index is very crucial and without its proper knowledge an intruder cannot intercept the message.

Naveen et al. (2013) implemented a technique that uses the DNA encryption to generate a DNA coded sequence which in turn is used as a key to the AES algorithm and finally ciphertext is obtained. Thus, it amends the AES algorithm with a DNA key. Mandge and Choudhary (2013) proposed a technique that applies matrix manipulation operation on the ASCII version of the original plaintext to get the final ciphertext The matrix manipulation cycle comprises of row shifting, left to right and up to down mirror operation. After this, they have applied PCR amplification methods. Chavan (2013) suggested a new asymmetric algorithm using DNA cryptography and binary one time pad (OTP) scheme.

Saranya et al. (2015) propounded an image encryption algorithm that generates a chaotic sequence from the secret key and permutes the original image using it, which is then DNA encoded. Sender generates the initial population of the mask and convert it to DNA sequence and performs XOR operation for the encryption. Next, he calculates entropy for all encrypted images and find best solution with highest entropy and lowest correlation using genetic algorithm. Barkha (2016) implemented DNA cryptography along with cloud computing and using socket programming. Here, the message is DNA

encoded and then amplified using a scheme proposed by authors. Thus, this method serves as an extension to bi-serial DNA encryption technique.

Rama Devi and Prabakaran (2016) suggested a technique involving DNA encoding, PCR and mathematical calculations involving prime numbers. Hossain et al. (2016) proposed a method that first DNA encodes the original message. This DNA sequence is transcripted into mRNA in which thymine (T) is replaced with uracil (U). mRNA is transferred to tRNA in accordance to the rule: A ⇋ U and G ⇋ C. Then, divide tRNA into two parts and create a new tRNA by interchanging them. Reverse simulation of interchanged tRNA to mRNA performed. At the end, it generates ciphertext using amino acid table that is also predefined.

Singh and Naidu (2017) suggested a novel method to secure data using DNA strands and Armstrong number. Firstly, it applies an authentication mechanism to verify the receiver. After this, plaintext is converted to ASCII and Armstrong number is added to it. Then this newly generated value is converted to cipher DNA sequence. Sukumaran and Misbahuddin (2018) proposed to use DNA cryptography for secure data storage in cloud. This technique performs DNA encoding on the data to be stored in the cloud. It is represented as D-DNA. A random DNA is selected from the gene pool and represented as R-DNA. It then indexes the R-DNA and selects the coding and non-coding regions based on indexing randomly. Finally, it inserts the D-DNA into non-coding regions of R-DNA and generate the final DNA sequence. At the end, encrypts it in binary form and stores in cloud. Popli (2018) and Pujari et al. (2018) suggested a new approach to encryption using concepts of genetic engineering and DNA cryptography.

4 Analysis and performance evaluation of existing DNA cryptography algorithms

In this section, we vigilantly dissect each of the above-mentioned encryption schemes categorically thus highlighting their demerits.

4.1 Analysis of simple substitution-based DNA cryptographic schemes

This category of algorithms does not involve any concept of key and only simple substitutions are performed, as a result of which security is compromised. They do not involve any concept of key, and hence no mathematical operations are involved. As per the DNA encoding technique since each letter, number or symbol is substituted by different combinations of any three nitrogenous bases, the size of the cipher text is too large than size of plain text. Also, repeatedly using the same look-up table or DNA dictionary for the entire procedure, same ciphertext is generated for same plaintext every time. The entire encryption is based on mere substitution and this reference table has to be shared completely with the receiver beforehand and hence this section of algorithms is highly prone to masquerading attack where any intruder can impersonate to be the receiver. They are also enormously susceptible to known plaintext attack and chosen ciphertext attacks. In these attacks, the intruder has certain plaintext-ciphertext pairs and he draws some analogy or relationship between them to intercept the entire information. An example is illustrated below.

Example 1

Consider the following plaintext-ciphertext pairs

Plaintext: MOON BOON

Ciphertext: ATCGGAGGACTG AACGGAGGACTG

In the two plaintexts, only the first alphabets are different. Hence, by comparing their corresponding ciphertexts, a relationship can be obtained. Thus, for the alphabets, M and B, their respective DNA sequence form 'ATC' and 'AAC' is retrieved. Using the same lookup table, encodes each letter into the same ciphertext. Since O occurs twice, noting the repeated sequence 'GGA', reveals the corresponding DNA encoded form of O. The left-over DNA sequence portion automatically stands out for N, i.e., 'CTG'. In this way, by analysing several plaintext-ciphertext pairs, the intruder can generate his own look-up table for further intrusion into the cryptosystem.

4.2 Analysis of biological operations-based DNA cryptographic schemes

The biological operations are beyond the control of the sender and hence desired output may not be attained sometimes. This category of algorithms does not involve any mathematical key. Information about primers, location of introns, exons, etc. comprise of the keys which is difficult to decide and implement. Obtaining the accurate plaintext even after using the correct decryption methodology is not always assured due to unrestrained operations. They are extremely difficult to be practically implemented due to highly sensitive functioning. The cost and complexity of encryption and decryption process is also highly increased as practical implementation of PCR, DNA hybridisation and DNA fragment reassembly is quite expensive.

4.3 Analysis of mathematical – biological operations-based DNA cryptographic schemes

Majority of these algorithms use simple DNA encoding as an additional layer of security to the intermediate ciphertext obtained. They perform the initial encryption using well acclaimed cryptographic techniques like AES, RSA, etc. and then simply encode the obtained ciphertext into its DNA sequence form using substitution mechanism. The number of steps is hugely increased in the encipherment procedure which adds up to the computation but does not guarantee adequate security. Involvement of cryptographic keys as well incorporating suitable DNA encoding schemes make them the most secure DNA encryption technique. However, most algorithms are restricted to text input messages. Only a few encryption algorithms have been suggested for image which is a major drawback.

4.4 Performance evaluation

Table 6 gives the performance evaluation of the three categories of the related work in terms of key involvement and type, encryption time. It also summarises their major limitations.

Table 6 Performance evaluation of existing DNA cryptographic schemes

Category	*Involve of key*	*Type of key*	*Encryption time*	*Limitations*
Simple substitution-based DNA cryptographic schemes	No keys involved	Not applicable	Least	Prone to statistical attacks as no adequate security
Biological operation-based DNA cryptographic schemes	Key are involved	Biological keys	Highest	Uncertain and beyond human control
Mathematical – biological operations-based DNA cryptographic schemes	Keys are involved	Mathematical keys	Intermediate	More predominant to text messages only

It can thus be observed that the mathematical – biological operations-based schemes are most secure and perform the best due to involvement of cryptographic keys as well as biological operations. Due to this double layer security, they must be more predominantly used.

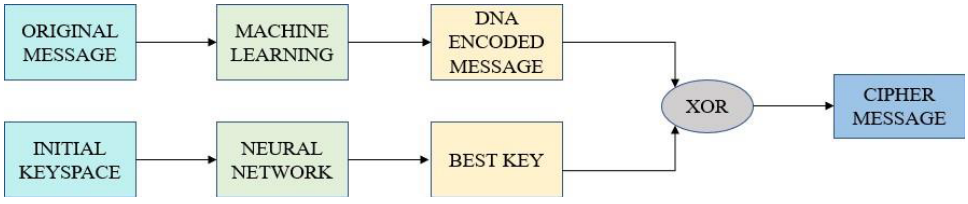
5 Proposed improvised DNA cryptography scheme combined with soft computing

The analysis of the three categories illustrates that simple substitution-based techniques are most insecure and prone to intrusion. The biological operations renders the second category quite uncertain and difficult for practical implementation. The mathematical – biological operations-based scheme is the most secure due to utilisation of cryptographic schemes as well as DNA encoding techniques. But currently, they are more predominant for text messages only. The basic block diagram of an improvised version of DNA cryptography after thoroughly scrutinising the merits and demerits of the existing literature has been proposed in this section. Soft computing (Zadeh, 1996) is capable to tolerate uncertainty and imprecision. The several soft computing methodologies include fuzzy logic, neural network, machine learning and genetic algorithms. Fuzzy logic is a modus operandi to computations based on 'degree of truth' rather than the contemporary 'true or false'. Hence, unlike Boolean expressions that involve only two values 0 (false) and 1 (true), it accepts a range of values from 0 to 1 including both. It works similar to human brain that aggregates information from a number of partial truths to combine them to form the complete truth. Artificial neural network (Schalkoff, 1997) is inspired by the biological neural network. It endeavours to understand relationship amongst several data by a process mimicking the way human brain functions. Machine learning (Michie et al., 1994) enables systems to automatically learn and improvise from their experiences without being explicitly programmed or any human intervention. Three main categories of machine learning algorithms are supervised learning, unsupervised learning and reinforcement learning. In supervised learning where the machine learns explicitly, support vector machine algorithm is very common. These three soft computing methodologies can be incorporated in DNA cryptography.

An idea to perform the entire DNA encoding using machine learning has been propounded. This will enable to accept any form of input – text, image, audio, video and represent it as corresponding DNA strands. The key can be generated by passing the

initial set of keys through several hidden layers of a neural network to obtain the best possible key. The final encrypted form of the original message will be the XORed output of the key and the DNA encoded strands. By doing so, we will probably get more optimised solutions and also reduce the amount of human interference in each step. The basic block diagram of the suggested DNA and soft computing-based cryptosystem is illustrated in Figure 10.

Figure 10 Block diagram of DNA and soft computing-based cryptosystem (see online version for colours)



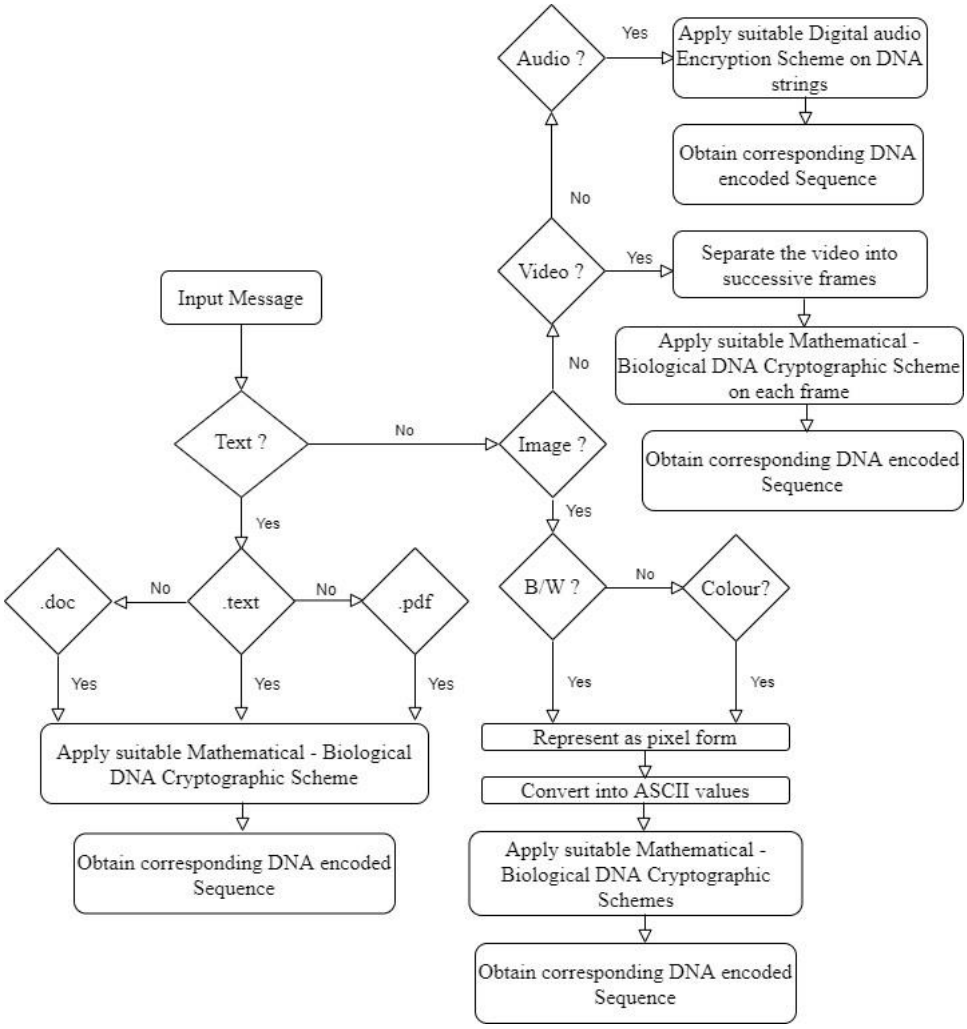
The entire methodology can thus be divided into two sections. The first involving machine learning to get the corresponding DNA encoded message from the original message. The second sections enable us to get the best possible key from the initial key space using a neural network.

The proposed soft computing-based cryptosystem will be a dynamic algorithm overcoming the limitations of existing algorithms. It will be able to encrypt any possible type of input – text, image, audio, video rather than being restricted to only a particular type of input. To enable this, CART, i.e., classification and regression trees (Bertsimas and Dunn, 2017) will be used which is an implementation of decision trees. As an example, the following technique can be applied for different inputs:

- If the input is a text message, further it will be checked whether word, pdf or text files and appropriate mathematical – biological operations-based cryptographic schemes will be applied to get corresponding DNA encoded cipher string.
- For image inputs, they are first represented as their pixel values. For coloured images, they will be broken into their RGB components. After this, the pixel values are replaced by their corresponding ASCII values. Next suitable mathematical – biological operations-based cryptographic schemes will be applied to get corresponding DNA encoded cipher string. For audio files, digital audio encryption schemes on MP3 compressions can be applied.
- A video can be can be fragmented into its array of successive frames. Each of these frames can be encrypted into its corresponding cipher form. Ultimately, we combine the cipher frames together to achieve the cipher video.
- Affirmatively what DNA encryption algorithm will be applied to which type of input has been kept for the future scope of this work. The encryption machine will be trained to first detect the type of input and perform the appropriate encryption algorithm to get the corresponding DNA string as the intermediate cipher text. The complete detailed designing of this will be done in our later endeavours. Such a dynamic cryptosystem will reduce the human intervention in performing the

encryption and also generate faster results. Figure 11 gives the working model of our encryption machinery.

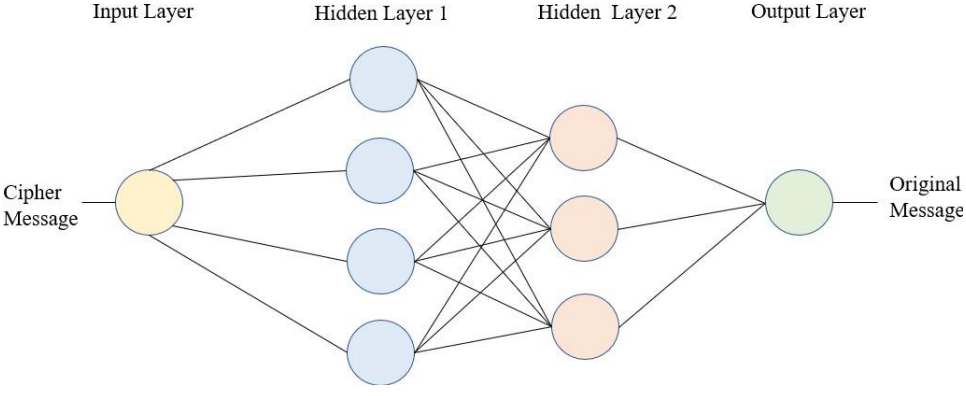
Figure 11 Proposed working model of the suggested DNA encryption Cryptosystem



A neural network can be trained by providing optimum inputs until we gain the optimum outputs. Therefore, it can be adjusted based on a collation between the desired goal and output, until the coveted goal is reached. Backpropagation algorithm (Hecht-Nielsen, 1992) is one of the most commonly used neural networks for supervised learning. The topology of the network consists of multiple layers of neurons such that each neuron of the nth layer is connected to all neurons of $(n + 1)^{th}$ layer. No connection prevails between neurons of the same layer. Synaptic weights refer to the strength of the connections between the neurons. Each of the inputs are multiplied with their initial weights and net input is calculated. This is then forward passed to the successive layers and final output is calculated. The output is then compared with the desired output and

the error is computed. To minimise this error, the weights are updated in a backward manner beginning from the end and reaching the first layer such that the coveted result is achieved and error is nullified. This ideology can be applied to an initial key space which will be passed through the several hidden layers and finally in the output layer the one having the maximum weight and no error will be selected as the best key. Figure 12 gives an illustration of the same. The exact methodology to be applied in order to train the neurons to select the best key after proper weight updating has been left for our future scope of work.

Figure 12 Basic diagram to find best key using neural network (see online version for colours)



The intermediate cipher message obtained after executing the first section will be XORed with the best key obtained in the second section to get the final cipher message that will be sent to the receiver.

6 Conclusions and future scope

DNA computation has numerous astounding features like immense parallelism, extraordinary storage capacity and minimal power requirement which make it a potential choice for cryptography. It is still in the infancy stage, with several hindrance at infrastructure level, lack of sound theoretical explanations, infeasible real-life implementations.

In this review paper, the achievements and limitations of certain encryption schemes based on DNA cryptography have been highlighted to conquer their loopholes and suitably modifying them. The existing work has been categorised into three broad categories and each has been analysed in details. It can be concluded that the mathematical – biological operations-based schemes perform the best and hence should be commonly used. Combining the new fields of DNA cryptography to the traditional encryption algorithm, and amalgamating mathematical, numerical and logical operation with biological operations will enhance the concept of confusion and diffusion. This will produce a more robust, secure and hard to decipher encryption algorithms. An improvised version based on DNA cryptography and soft computing has also been proposed. The feasibility and practicality of this scheme has to be studied and this has to be actually implemented after deciding which machine learning and neural network algorithm has to

be applied to get the best possible result. The real implementation and deciding the exact methodology and flow of actions to perform real time encryption has been kept for our future endeavours.

References

Adleman, L.M. (1994) 'Molecular computation of solutions to combinatorial problems', *Science*, Vol. 266, No. 5187, pp.1021–1024.

Barkha, P. (2016) 'Implementation of DNA cryptography in cloud computing and using socket programming', *2016 International Conference on Computer Communication and Informatics (ICCCI)*, IEEE, January, pp.1–6.

Bertsimas, D. and Dunn, J. (2017) 'Optimal classification trees', *Machine Learning*, Vol. 106, No. 7, pp.1039–1082.

Boneh, D., Dunworth, C. and Lipton, R.J. (1995) 'Breaking DES using a molecular biology', *DNA Based Computers, Proc. of DIMACS Workshop*, American Mathematics Society, pp.37–65.

Chavan, S. (2013) 'DNA cryptography based on DNA hybridization and one-time pad scheme', *International Journal of Engineering*, Vol. 2, No. 10, pp.2679–2682.

Crick, F. (1970) 'Central dogma of molecular biology', *Nature*, Vol. 227, No. 5258, p.561.

Crick, F.H. (1968) 'The origin of the genetic code', *Journal of Molecular Biology*, Vol. 38, No. 3, pp.367–379.

Dhawan, S. and Saini, A. (2012) 'Integration of DNA cryptography for complex biological interactions', *International Journal of Engineering, Business and Enterprise Application*, Vol. 2, No. 1, pp.121–127.

Freund, R., Kari, L. and Păun, G. (1999) 'DNA computing based on splicing: the existence of universal computers', *Theory of Computing Systems*, Vol. 32, No. 1, pp.69–112.

Gehani, A., LaBean, T. and Reif, J. (2003) 'DNA-based cryptography', *Aspects of Molecular Computing*, pp.167–188, Springer, Berlin, Heidelberg.

Gibson, D.G. (2011) 'Enzymatic assembly of overlapping DNA fragments', *Methods in Enzymology*, Vol. 498, pp.349–361. Academic Press.

Hameed, S.M. and Al-Ani, M. (2018) 'Image encryption using DNA encoding and RC4 algorithm', *Iraqi Journal of Science*, Vol. 59, No. 1B, pp.434–446.

Hecht-Nielsen, R. (1992) 'Theory of the backpropagation neural network', in *Neural Networks for Perception*, pp.65–93, Academic Press.

Hossain, E.M.S., Alam, K.M.R., Biswas, M.R. and Morimoto, Y. (2016) 'A DNA cryptographic technique based on dynamic DNA sequence table', *2016 19th International Conference on Computer and Information Technology (ICCIT)*, IEEE, December, pp.270–275.

Jain, S. and Bhatnagar, V. (2014) 'A novel DNA sequence dictionary method for securing data in DNA using spiral approach and framework of DNA cryptography', *2014 International Conference on Advances in Engineering & Technology Research (ICAETR-2014)*, IEEE, August, pp.1–5.

Lemieux, B., Aharoni, A. and Schena, M. (1998) 'Overview of DNA chip technology', *Molecular Breeding*, Vol. 4, No. 4, pp.277–289.

Lewis, M. (2001) *Agarose Gel Electrophoresis (Basic Method). Biological Protocols*, University of Liverpool, Liverpool.

Li, X.S., Zhang, L. and Hu, Y.P. (2008) 'A novel generation key scheme based on DNA', *2008 International Conference on Computational Intelligence and Security*, IEEE, December, Vol. 1, pp.264–266.

Lipton, R.J. (1995) 'DNA solution of hard computational problems', *Science*, Vol. 268, No. 5210, pp.542–545.

Mandge, T. and Choudhary, V. (2013) 'A DNA encryption technique based on matrix manipulation and secure key generation scheme', *2013 International Conference on Information Communication and Embedded Systems ICICES)*, IEEE, February, pp.47–52.

Michie, D., Spiegelhalter, D.J. and Taylor, C.C. (1994) 'Machine learning', *Neural and Statistical Classification*, p.13.

Mullis, K.B. and Faloona, F.A. (1987) 'Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction', *Methods in Enzymology*, Vol. 155, pp.335–350, Academic Press.

Nandy, N., Banerjee, D. and Pradhan, C. (2018) 'Color image encryption using DNA based cryptography', *International Journal of Information Technology*, pp.1–8, https://doi.org/10.1007/s41870-018-0100-9.

Naveen, K.J., Karthigaikumar, P., Sivamangai, N.M., Sandhya, R. and Asok, S.B. (2013) 'Hardware implementation of DNA based cryptography', *2013 IEEE Conference on Information & Communication Technologies*, IEEE, April, pp.696–700.

Nguyen, H.T., Walker, C.L. and Walker, E.A. (2018) *A First Course in Fuzzy Logic*, CRC Press.

Ning, K. (2009) *A Pseudo DNA Cryptography Method*. arXiv preprint arXiv: 0903.2693.

Popli, M. (2018) 'DNA cryptography: a novel approach for data security using genetic algorithm', *International Journal of Advance Research in Computer Science and Management Studies*, Vol. 6.

Prabhu, D. and Adimoolam, M. (2011) *Bi-serial DNA Encryption Algorithm (BDEA)*, arXiv preprint arXiv: 1101.2577.

Pramanik, S. and Setua, S.K. (2012) 'DNA cryptography', *2012 7th International Conference on Electrical and Computer Engineering*, IEEE, pp.551–554.

Pujari, S.K., Bhattacharjee, G. and Bhoi, S. (2018) 'A hybridized model for image encryption through genetic algorithm and DNA sequence', *Procedia Computer Science*, Vol. 125, pp.165–171.

Rama Devi, K. and Prabakaran, S. (2016) 'An enhanced bilateral information security towards a conventional cryptographic system using DNA sequences', *Indian Journal of Science and Technology*, October, ISSN: 0974-5645.

Sabry, M., Hashem, M. and Nazmy, T. (2012) 'Three reversible data encoding algorithms based on dna and amino acids' structure', *International Journal of Computer Applications*, Vol. 54, No. 8, pp.24–30, https://doi.org/10.5120/8588-2339.

Sadeg, S., Gougache, M., Nabil, M. and Drias, H. (2010) 'An encryption algorithm inspired from DNA', *2010 International Conference on Machine and Web Intelligence*, IEEE, November, pp.344–349.

Saranya, M.R., Mohan, A.K. and Anusudha, K. (2015) 'Algorithm for enhanced image security using DNA and genetic algorithm', *2015 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES)*, IEEE, February, pp.1–5.

Schalkoff, R.J. (1997) *Artificial Neural Networks*, Vol. 1, McGraw-Hill, New York.

Shapiro, E. and Benenson, Y. (2006) 'Bringing DNA computers to life', *Scientific American*, Vol. 294, No. 5, pp.44–51.

Singh, M.S.P. and Naidu, M.E. (2017) 'A novel method to secure data using DNA sequence and Armstrong number', *Asian Journal for Convergence in Technology*, Founded by ISB &M School of Technology, Vol. 3, No. 3, p.3.

Sukumaran, S.C. and Misbahuddin, M. (2018) 'DNA cryptography for secure data storage in cloud', *I.J. Network Security*, Vol. 20, No. 3, pp.447–454.

Vijayakumar, P., Vijayalakshmi, V. and Zayaraz, G. (2011) 'DNA computing based elliptic curve cryptography', *International Journal of Computer Applications*, Vol. 36, No. 4, pp.18–21.

VijayaKumar, P., Vijayalakshmi, V. and Zayaraz, G. (2013) 'Enhanced level of security using DNA computing technique with hyperelliptic curve cryptography', *International Journal on Network Security*, Vol. 4, No. 1, p.1.

Wang, X. and Zhang, Q. (2009) 'DNA computing-based cryptography', *2009 Fourth International on Conference on Bio-Inspired Computing*, IEEE, October, pp.1–3.

Wang, Y., Han, Q., Cui, G. and Sun, J. (2019) 'Hiding messages based on DNA sequence and recombinant DNA technique', *IEEE Transactions on Nanotechnology*, Vol. 18, pp.299–307, https://doi.org/10.1109/tnano.2019.2904842.

Watson, J.D. and Crick, F.H. (2003) 'A structure for deoxyribose nucleic acid. 1953', *Nature*, Vol. 421, No. 6921, p.397.

Zadeh, L.A. (1996) 'Soft computing and fuzzy logic', *Fuzzy Sets, Fuzzy Logic, and Fuzzy Systems*, Selected Papers by Lotfi a Zadeh, pp.796–804.

Zhang, Y., Fu, B. and Zhang, X. (2012) 'DNA cryptography based on DNA fragment assembly', *2012 8th International Conference on Information Science and Digital Content Technology (ICIDT2012)*, IEEE, June, Vol. 1, pp.179–182.

Zhang, Y., Wang, Z., Wang, Z., Karanfil, Y.H. and Dai, W. (2016) 'A new DNA cryptography algorithm based on the biological puzzle and DNA chip techniques', *International Conference on Biomedical and Biological Engineering*, Atlantis Press, July.

Zhang, Y., Yu, Z., Zhong, W. and Sinnott, R.O. (2011) 'Index-based symmetric DNA encryption algorithm', *2011 4th International Congress on Image and Signal Processing*, IEEE, October, Vol. 5, pp.2290–2294.