Text and Image Encryption-Decryption Via Bio-Alphabets

DNA Cryptography

Neethu Manohar
M. Tech Student
Department of Computer & Information Science
SIST, Thiruvananthapuram

Ms. Renji. S
Assistant Professor
Department of Computer & Information Science
SIST, Thiruvananthapuram

***Abstract*—DNA can be used in cryptography for storing and transmitting the information as well as for computation. It's a new born cryptographic method where by DNA is the information carrier. DNA cryptography is based on computation using DNA, but not computation on DNA. The vast parallelism and extraordinary information density inherent in DNA molecules are explored for cryptographic purposes like encryption, authentication and signature. This proposed work is based on conventional cryptography. It's having three phases-key generation, encryption, and decryption. Key generation is based on One-Time-Padding. Genetic databases represent a feasible solution for OTP symmetric key generation and transmission. Transmission of a lengthy key is not required, because each sequence has a unique identification number in the database and this number itself or its combination can sent instead. Encryption is based on symmetric key cryptography. Proposed work focuses data in the form of text and image. A single algorithm is developed for both types data encryption-decryption. For key transmission, the codebook is created. Before the start of actual communication, the sender provides a copy of the codebook to the receiver. The decryption process is just the reverse of encryption. Privacy and security is of increasing concern in wireless, wired, and internet communication networks. The main goal of this work is to provide a relatively more degree of security avoiding data breaches, time complexity and space complexity.**

***Index Terms*—DNA, One-Time-Padding, DNA compression, Accession Number, codebook**

I. INTRODUCTION

Internet influences the human life to such a degree that almost every walks of life passes through this web at any time of its passage. Financial transactions, social networking, personnel data sharing, vital information sharing etcetc use this path for easy task completion. So this communication system will have to remain reliable. For this, the system has to be protected against challenging security issues like unauthorized access and hacking. From time to time cryptologists develop several protocols and standards for keeping the system reliable, but intruders succeed to the same level. This makes "Making-Cracking", a never ending task. Cryptologist has to choose the path Security-Integrity-Authenticity-Confidentiality to get around challenging security issues.

The path for secure information branches into cryptography and steganography. The former transmits the data in unintelligible form while the latter transmits the data in hidden format. Cryptography and steganography are the most widely used. A statistical report the reliability of this technique shows that that about 2 million security records were breached techniqueswhich implement the secret writing. Multiple cryptographic techniques are used for secure data transmission per day all over, that is on an average 32 records were breached per second.

An American mathematical engineer Claude Elwood Shannon estimated that human languages have redundancy. Shannon estimated the entropy of written English to be 0.6 to 1.3 bits per character based on how well people can predict successive characters in text. Cover and King concluded that human language has entropy to be 1.25 bits per character. This redundancy catalyse the action of breaking ciphers. So the internet world is searching for some new techniques which is relatively morestrong against intruding.Surely DNA cryptography can quench this search. Ongoing researches in DNA cryptography marches positively towards this target. DNA can be used in cryptography for storing and transmitting the information as well as for computation. Although in its primitive stage, DNA cryptography is shown to be very effective. DNA Cryptography is a new born cryptographic field emerged with the research of DNA Computing, in which DNA is used as an Information carrier and modern biological technology is used as implementation tool.
The remainder of this paper is divided into 9sections:- Section II describes biological background of DNA and RNA. Section III describes DNA computing. Section IV include related works in DNA cryptography and summary of literature review. Section VI describes the problem statement. Section VIIinclude the objective of the proposed work. Section VIII describes the proposed work and the working model of the proposed system. Section IX describes Design of DNA cryptosystem and module description.

II. BIOLOGICAL BACKGROUND

DNA is the genetic information carrier of cellular organisms. The polymer chains in DNA called DNA strands may be viewed as a chain of nucleotides. Nucleotides are the building molecules for DNA. Every Nucleotide carries a phosphate group, a sugar group plus a nitrogen base. The nitrogen bases are four in numbers They are named as adenine

(A), thymine (T), guanine (G) and cytosine (C), abbreviated as A, G, C and T respectively. Two separate strands of DNA bond together to form a double helix structure. A bonds with T and G bonds with C. The pairs (A, T) and (G, C) are known as Watson-Crick complementary base pairs.

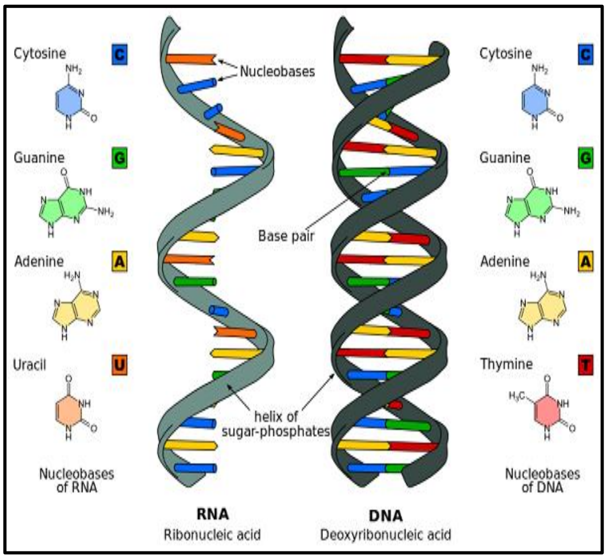


Fig.1 Structure of DNA and RNA

DNA is a polynucleotide whose monomer units are nucleotides. Nucleotide is having a 5-carbon sugar called deoxyribose, a nitrogen base attached to the sugar and a phosphate group. Four different types of nucleotides are found in DNA which differs only in the nitrogenous base. The four nucleotide bases are

- Adenine [A]
- Guanine [G]
- Cytosine [C]
- Thymine [T]

RNA (ribonucleic acid) is a polymer having one or more nucleotides. Each strand of RNA is a chain with a nucleotide at each link. Each nucleotide is made up of a base (adenine, cytosine, guanine, and uracil), a phosphate and a ribose sugar. The four bases in RNA are

- Adenine [A]
- Guanine [G]
- Cytosine [C]
- Uracil [U]

III. DNA COMPUTING

DNA computing or bio-molecular computing utilizing the combinational properties of DNA. For massively parallel computation. The idea is that with an appropriate setup and enough DNA, one can potentially solve huge mathematical problems by parallel search. Basically this means that you can attempt every solution to a given problem until you came across the right one through random calculation. Utilizing DNA for this type of computation can be much faster than utilizing a conventional computer, for which massive parallelism would require large amounts of hardware, not simply more DNA. DNA computing uses only the concept of DNA that is computation using DNA, but not computation on DNA.

DNA computing is a technique, in which DNA is used as a computation tool to solve some NP complete problem. DNA computing takes the advantage of DNA, combinational properties of DNA for massively parallel computation. DNA computing uses only the concept of DNA ie computation using DNA, but not computation on DNA. Leonard Max Adleman [1] is considered as the father of DNA computer and DNA computing. His work is based on the project in DNA steganography by Viviana Risca [2], which proposes how to hide information in a DNA microdot.

IV. RELATED WORKS

In 1994, Leonard Adleman [1], surprised the scientific community by using the tools of molecular biology to solve a different computational problem. His article in a 1994 issue of the journal Science outlined how to use DNA to solve a well-known mathematical problem, called the directed Hamilton Path problem, also known as the "travelling salesman" problem. The goal of the problem is to find the shortest route between number of cities, going through each city only once. He solved the instance of graph containing seven vertices by encoding it into the molecular form by using an algorithm and then computational operations were performed with the help of some standard enzymes. This was solved by brute force method.

In 1999 Viviana Risca's, Carter Bancroft, Catherine Taylor Clelland[2], which proposes how to hide information in DNA microdots. They have taken the microdot a step further and developed a DNA-based, doubly steganographic technique for sending secret messages. A DNA encoded message is first hide within the enormous complexity of human genomic DNA and then further concealed by confining this sample to a microdot.

In 1995, Lipton [3], extended the work of Adleman by solving another NP-complete problem called “satisfaction” by using DNA molecules in a test tube to encode the graph for 2 bit numbers.

In 1996, Dan Boneh et al. [4], applied the approaches of DNA computing used by Adleman and Lipton, in order to break one of the symmetric key algorithm used for cryptographic purposes known as DES (Data Encryption Standard). They performed biological operations on the DNA strands in a test tube, such as extraction, polymerization via DNA polymerase, amplification via PCR and perform operations on the DNA strands which have the encoding of binary strings. Then DES attack is planned by generating the DES-1 solution, due to which key can be easily guessed from the cipher text and further evaluate the DES circuit, lookup table and XOR gates. By using their molecular approach they broke DES in merely 4 months.

In 1997, Qi Ouyang et al. [5] applied the approaches of DNA molecular theory in order to generate the solution for maximal clique problem, which is another NP-complete problem. Thus shows the efficiency of DNA: to solve Hard-problems and vast parallelism inherent in it which makes the operations fast.

In 2009, Monica E. Borda, Olga Tornea, and Tatiana Hodorogea [6], proposed a paper presents an algorithm of secret writing by DNA hybridization, based on existing ideas. This paper investigates a variety of bioinformatics methods and proposes an algorithm for encrypting and hiding data in real or artificial DNA digital form. As in all the cryptographic methods, the DNA hybridization technique also involves the encryption and decryption processes in converting the plaintext into the cipher text and then retrieving back the original message.

In the DNA hybridization method, the original message which is referred as plain text is converted in the form of binary. This binary form of data is then compared with the randomly generated OTP key in the DNA form and the encrypted message is obtained. This obtained encrypted message is also in the form of DNA. The decryption message is carried out in reverse using the encrypted data and the OTP key and the original message is retrieved.

In the DNA hybridization technique, the plain text is converted into the binary form of the data. The OTP is generated by combining the random oligonucleotides (ssDNA) strands together with help of a short DNA fragment as template. The strands are combined using a special protein called ligase. This combining process of the oligonucleotides is performed because; the OTP key is to be generated of wider length which should be lengthier than the size of the message. That is the length of the key is 10 times longer than the plain text.

The OTP key is to be generated in the DNA form of the data. Then for each ‘1’ bit in the binary data, the key is compared with the binary digit and encrypted message produced. And if the binary digit is found to be ‘0’, no operation is performed. For this reason of the random generation of the key with huge length, it can be said that the DNA hybridization technique enables a tremendous security for the data.

In 2011, Zhang Yunpeng, Zhu Yu, Wang Zhong, Richard O.Sinnott [7], have presented a symmetric key cryptosystem based on the DNA symmetric cryptosystem using index. In this paper, a new index-based symmetric DNA encryption algorithm has been proposed. Adopting the methods of Block-Cipher and Index of string, the algorithm encrypts the DNA sequence-based plaintext. First, the algorithm encodes each character into ASCII codes. And then, according to the nucleotide sequence, the researcher should convert it to the DNA coding. Besides, the researcher selects the special DNA sequence as the encryption index, and likewise, the pre-treated plaintext will be divided into different groups.

Next, the key created by the Chaos Key Generator based on the Logistic Mapping and initialized by the number x0 and μ, will take XOR operation with the block-plaintext. The type of number x0 and μ, which is selected by the researcher, is double. Then, the result of these processes will be translated on the DNA sequence. In addition, compared to special DNA sequence, the algorithm finds the sequence which has no difference with it. Then, the algorithm will store the position as the Cipher-text. The researcher proves the validity of the algorithm through simulation and the theoretical analysis, including bio-security and math security. The algorithm has a huge key space, high sensitivity to plaintext, and an extremely great effect on encryption. Also, it has been proved that the algorithm has achieved the computing-security level in the encryption security estimating system.

In 2013 TusharMandge, Vijay Choudhary [8], author has designed a new method by integrating DNA computing in IDEA. Such conceptual works can be useful in the development of this new born technology of cryptography to fulfil the future security requirements. In this paper; a proposal is given where the concept of DNA is being used in encryption and decryption process. The theoretical analysis shows this method to be efficient in computation, storage and transmission; and it is very powerful in certain attacks. This paper also presents a secured symmetric key generation scheme which generates primary cipher and this primary cipher is then converted into final cipher using DNA sequences, so as to make it again more complicated in reading. Finally, the implementation methodology and experimental results are presented.

In 2014 Surendra Varma, K. Govinda Raju [9], proposed the DNA cryptographic using random key generation scheme. This paper analyses the different approach on DNA cryptography based on matrix manipulation and secure key generation scheme. They have presented a new DNA encryption technique based on mathematical matrix manipulation where they have used a secure generation algorithm for encryption process. The benefit of key generation scheme is, always get a new cipher text for same plaintext and same key. So it provides a good security layer which does not give any hint about plaintext.

DNA binary strands support feasibility and applicability of DNA based cryptography. The security and the performance of the DNA based cryptographic algorithms are satisfactory for multilevel security applications of today’s network. Certain DNA algorithms can resist exhaustive attack, statistical attack and differential attack. DNA computing is viable and DNA authentication methods have shown great promise in the marketplace of today and it is hoped that its applications will continue to expand. DNA cipher is the beneficial supplement to the existing mathematical cipher. If the molecular word is controlled then it may be possible to achieve vastly better performance for information storage and security.

In 2014, Bonny B Raj, Panchami V [10], presented a paper, DNA based cryptography using permutation and random key generation. Initially plaintext is converted into ASCII code, ASCII code is again converted into binary form to get the data in 0’s and 1’s. These binary values are encoded in DNA sequences to nucleotide conversion where each of the four bases is represented by combinations of 0’s and 1’s. A DNA sequence is selected as a key and grouped into the blocks in which each block is of 4 characters. Then a table is created based on the positions of each character in the key sequence. Based on table and the randomly selected DNA sequence, text gets converted into encrypted form. Finally the encrypted sequence with the key is sent to the receiver. The DNA sequence in decryption process gets decoded into binary then that binary is converted into ASCII and finally ASCII to the plaintext. The method explains how traditional cryptography differs from the emerging DNA cryptography.

In 2014 Ashish Kumar Kaundal [11], proposed a DNA hybrid symmetric key method and algorithm which is based on DNA cryptography and feistel inspired structure. In this plaintext is converted into ascii then to binary. Reordering of binary plaintext using fiestel inspired structure is performed. SIn this paper they generate a random key sequence based on one-Time-Pad (OTP) that uses pseudo-random generator and provide the seed of 32 bytes DNA sequence as an input to it from the genetic database (GenBank) and kept the source secret. This pseudo-random generator will generate the high quality OTP sequence based on the seed and is very much secure than the other random functions that are used in C. It produces the unique result every time according to some statistical calculations.

In 2014, Ritu Gupta, Anchal Jain [12], this paper proposes a new method of image encryption based on DNA computation technology. The original image is encrypted using DNA computation and DNA complementary rule. First, a secret key is generated using a DNA sequence and modular arithmetic operations. Then each pixel value of the image undergoes the encryption process using the key and DNA computation methods.

In 2014, ToshithaKannan, M. SindhuMadhuri [13], proposed a new encryption algorithm for secret writing using DNA. In this paper, the idea of recombinant DNA technology based on use of restriction enzymes is the main principle behind the suggested crypto system. While the encryption employs the principle of restriction, the decryption involves use of primers and the concept of DNA hybridization. In the first stage of encryption they use the principle of rDNA technology and restriction enzymes. The message in DNA form is the ‘gene of interest’ and a DNA sequence from the database is the ‘vector’ which is used for ‘cloning.’ In the second stage of encryption, a DNA sequence is virtually generated as the key and the BLAST.

In 2015, AsishAich, Alosen, SatyaRanjan Dash and SatchidAnandaDehuri [14], proposed two stage encryption algorithm based on DNA sequence. In the first stage an encryption of plain text is done by generating a random key. The plain text is again encrypted to produce the cipher text in the second stage. Moreover, this encryption algorithm is based on a symmetric key cryptography system, where they provide a shared key to encrypt as well as decrypt the intended message. To encrypt the original key two stages are maintained and sending it over a separate secure channel other than the channel through which they are transferring the cipher text. A numerical study confirms that the proposed algorithm is reliable, secure, scalable, and robust for transmitting message.

In 2015, Isha Yadav, Nipun Gupta, and M.K. Beniwal [15], proposed a new DNA cryptographic approach based on one time pad. The proposed algorithm to implement data security in binary representation of DNA sequence is done using the random number generator as well as using encryption and decryption algorithm, based on the method of binary addition and binary subtraction rule. It’s having three phases key generation, encryption and decryption. This scheme uses the DNA digital coding technique, DNA synthesis and PCR amplification, Random number generation and Arithmetic operations as well as traditional cryptography. In this work, the plaintext is converted into binary form and then DNA form. Random key generation method is used for each nucleotide of DNA sequence within the range 1-99. If random number is greater than 99 then number should be subtracted by 99. The plaintext in DNA form and random key is converted into binary and perform binary addition, results sequence of binary. Then convert this sequence into DNA form using DNA encoding method.

a. SUMMARY OF LITERATURE REVIEW

Analysis of the literature survey shows that DNA cryptography merges both cryptographic and bio-molecular techniques for secure data transmission. Two approaches are there.DNA cryptography based on molecular theory and DNA cryptography based on conventional cryptography and asymmetric cryptography.

The approach based on molecular theory uses techniques like DNA micro-array, DNA fragmentation, DNA hybridization, and central dogma using symmetric as well as asymmetric key cryptography. For its implementation, high-tech lab requirements are needed.On the other hand, DNA cryptography based on conventional approach passes through key generation, encryption and decryption process. In conventional cryptography, symmetric as well as asymmetric realization can be followed. Symmetric key realization is easier than asymmetric realization.

The design issues and key generation approaches in the existing DNA cryptographic methods for text and images gives opportunities for brute force attacks. In DNA cryptography key generation is based on OTP. According to Shannon OTP is the only potentially unbreakable encryption method. In the existing methods, the OTP are usually generated using random key generators. A key is considered as OTP, if it satisfies the following constraints.

The Key must be random and generated by a non-deterministic, non-repeatable process. To achieve perfect secrecy, the key length should be greater than or equal to message length.In existing methods, key is generated using random key generator. Since random key generator is used. Truly random numbers are hard to produce and the process of key storage, key management, and transmission is somewhat difficult. The problems with existing system are

- Proportional to the size of plaintext the encryption time and decryption time varies.
- Security only depends upon the key.
- Higher security requires lengthy key, but encryption consumes more time.
- If length of DNA fragment is short, intruder can easily detect.
- Requires more memory space for storing the lengthy key and performing the operations involving it.
- Computational complexity is high based on the comparison, shifting, and the scanning processes.
- Key generation and key transmission is difficult.
- If cryptography is based on asymmetric realization, two keys are required-one for encryption and the other decryption.

- Image encryption algorithm based on DNA is complex.

V. PROBLEM STATEMENT

The fundamental target of DNA cryptography is to achieve the highest level of confidentiality, integrity and availability while sending data over a network and protect data from brute force attack. In the existing encryption-decryption techniques for text and image, time complexity, space complexity as well as computational complexity are relatively high. In existing DNA cryptographic techniques OTPs are generated using random key generator. Since random key generator is used truly random numbers are hard to produce and processing of key generation, storage and transmission is somewhat difficult. The main intent to achieve in DNA cryptography are compact storage space, relatively high computational power, generation of cryptographic keys from long sequence. For performing the encryption and decryption processes, several biological trials and tests have to be performed.

VI. OBJECTIVE

The aim of my thesis is to build a DNA cryptosystem system which satisfies the following objectives: For solving above mentioned problems, my attempt is to develop a cryptosystem based on DNA cryptography for secure data transmission. For this work, I am focusing data in the form of text and image only. For both these inputs, same encryption algorithm is being used. If data is in text form, encrypt using an encryption algorithm. If the data is in image form, two methods are used for conversion and find a time complexity of both of these algorithm. First convert image to text using suitable algorithm, then the same procedure as for the text encryption. Second method is to first convert image to binary, then the same procedure as for the text encryption.

This work is based on conventional cryptographic method. It's having three phases key generation, encryption, and decryption. Encryption is based on symmetric key technique. This proposed work is purely based on one-time-padding (OTP). The OTP is taken directly from public genetic database. There are many public databases available. I am using the database namely GenBank. GenBank is an open access genetic sequence database, a collection of all public available DNA sequences. An accession number is used for accessing DNA sequence from GenBank with the help of MATLAB Bioinformatics tool. So this accession number is kept secret and transmitted to the receiver. My plan is to use separate encryption option for transmitting key. The possibility of brute force attack is avoided since the key is lengthy. The aim of the work is to develop a system which process text and image data for secure transmission via bio alphabets.

VII. PROPOSED SYSTEM

In the proposed work a new DNA cryptographic system is introduced, which can solve the issues in conventional cryptographic method. Here a single algorithm is used for both types of data (text and images).

a. WORKING MODEL OF PROPOSED SYSTEM

The algorithm which I developed for this work is compatible for text data as well as image data. If data is in text form, encrypt using TEA. For image encryption two image pre-processing techniques are used. The first one converts image to text using suitable algorithm [16], then the same procedure as for the text encryption. Second one converts image to binary, then the same procedure as for the text encryption. In the completion of work, a comparative study between these two algorithms are included. This work is based on conventional cryptographic method. It's having three phases.

- Key generation
- Encryption
- Decryption

Key generation is based on one-time-padding (OTP). The OTP is taken directly from public genetic database. There are many public databases available like EMBL, DDBJ, and GenBank. The database used for this work is from GenBank. GenBank is an open access genetic sequence database, a collection of all publicly available DNA sequences. An accession number is used for accessing DNA sequence from GenBank with the help of MATLAB Bioinformatics toolbox

An accession number is a combination of block letters of English alphabets, numerals 0-9 and the special symbol '_' (underscore). This accession number has to be kept secret and transmitted to the receiver for decryption. In view of keeping the accession number secret, a codebook generated with the help of DNA compression algorithm (DCA). The importance of the codebook is that it has to be exchanged at least once in between the sender and receiver via publicly or privately before the actual data transmission begins.

DNA Encryption is the technique for encrypting the secret message using Bio molecular computation which makes this unique from mathematical computation. In the DNA indexing method, the plain text which is the original message is converted to the binary form and again to the DNA form. The OTP keys are generated randomly from the public database. This OTP key and the DNA form of the plain text are compared and a random index is generated, which is the encrypted data. Decryption process is carried out in the opposite order to obtain the original plain text message.

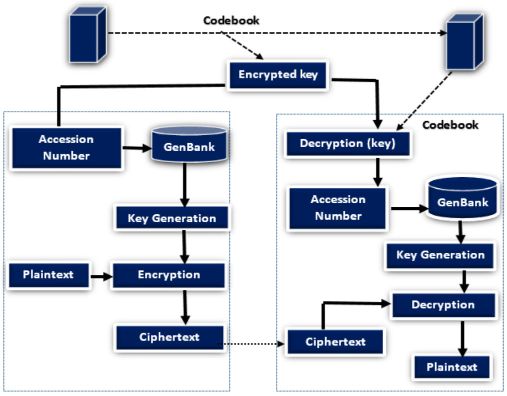


Fig 2. Proposed DNA cryptosystem

VIII. METHODOLOGY

DNA cryptographic methodology uses different ways for data encoding. DNA cryptographic methodologies like Polymerase chain reaction (PCR), Bio molecular techniques and one-time-padding (OTP) are used for secure message transmission. PCR technique is a DNA digital coding technique where messages are converted first to hexadecimal, then binary code and further to DNA sequence, which is used in DNA template. Bio molecular technique uses parallel processing capabilities of bio molecular computation. The OTP technique is used to encrypt and decrypt plain text.

The proposed system includes both text as well as image encryption.For both inputs, a single Text Encryption algorithm (TEA) works out. In addition, for image encryption two different image preprocessing techniques are checked with. Either images are converted to text using suitable algorithms or image is first converted to binary and then follow the same process as before.

A. Text Encryption Algorithm

In this algorithm first reading plain text and split the text into characters. The characters are converted to ASCII and then to base 2 binary. Binary characters are encoded into 4 character sequence using DNA encoding rule. By using accession number, key is retrieved from public data base. Compare the retrieved DNA sequence with the DNA form of the plaintext to form an index array. Randomly choose one index and write it into a file. Repeat this for entire sequence. Finally an index file is obtained. For example the key is the DNA sequence of the mitochondria, the following code is used for key retrieval and compare the retrieved DNA sequence with DNA form of the plaintext.

```
Mitochondria=getgenbank('NC_001807','SequenceOnly',true)
;
j=1
fori=1:4:length(sc)
k=strfind(Mitochondria,sc(i:i+3))
code(j)=k(randi([1 length(k)]));
j=j+1;
```

```
Step 1: Start
Step 2: Read the plaintext from file
Step 3: split the plaintext into characters
Step 4: convert each character to its corresponding ASCII value.
Step 5: Convert ASCII to binary
Step 7: Encode binary string to 4 char sequence (01=A, 10=C, 11=G, 00=T)
Step 8: Retrieve sequence from the public database (GenBank)
Step 9: compare the retrieved DNA sequence with the DNA form of the plaintext (step 7) to form an index array.
Step 10: Randomly choose one index
Step 11: Write index into file
 Step 12: Repeat step 4 to 11 until end of the string
Step 13: End
```

B. Image Encryption Algorithm (IEA)

The proposed work of image encryption is planned to be done in two techniques. Either image to text conversion or image to base2 binary. In both techniques, the same TEA algorithm is used for the purpose of encryption.

i. Image to text conversion

In this proposed work, image to text conversion is an image preprocessing technique for image encryption. Here grayscale conversion and mathematical operations are performed. For the conversion of image to text, change the image to grayscale image for processing to be carried out on a single array. The image data corresponding to each pixel is converted to ASCII characters (English alphabets) and written into text files. Each character can then be read back and converted to its corresponding pixel values through mathematical operations. The operation which is made use of is the modulus operation which gives the remainder of a division operation [16].

```
Step 1. Read an image in any of the colour spaces.
Step 2. Convert the image to grayscale image if not in gray.
Step 3. Open a text file 'F' in write mode.
Step 4. For each pixel 'x' repeat the following steps until end of image file is reached:
i.  Perform, rem=x%52 and store quotient in 'q'.
        ii. If rem>25, add 71, else add 65 to remainder to change the range to ASCII codes for English alphabets.
        iii. Convert the number to English letters.
        iv. Write the character into file 'F'
Step 5.Close the file.
```

ii. Image restoration from text

```
Step 1.  Open the corresponding text file 'F' in read mode.
Step 2.  For each pixel 'x' repeat the following steps until end of image file is reached:
i. Read the first character 'c'.
      ii. Read the second character 'q'.
iii. Convert the characters c and q to its corresponding ASCII code, say, rem1 and q1
      iv. If rem1>90, subtract remainder from 71, else subtract remainder from 65
v. Evaluate each pixel value using the formula:
          Value = (52*q1)+rem1
      vi. Store the value in an image array
Step 3.  Close the file.
Step 4.  Display the image.
```

i. Image to binary

In RGB colour model, each colour appears in its primary spectral components of red, green, and blue. The colour of a pixel is made up of three components- red, green and blue (RGB) described by their corresponding intensities. Colour components are also known as colour channels or colour planes (components).

In the RGB colour model, a colour image can be represented by the intensity function. I(RGB) =(fR ,fG ,fB) where fR(x, y) is the intensity of the pixel (x, y) in the red channel, fG(x, y) is the intensity of pixel (x, y) in the green channel and fB(x, y) is the intensity of pixel (x, y) in the blue channel.

```
r1=image(:,:,1);
g1=image(:,:,2);
b1=image(:,:,3);
```

The intensity of each colour channel is usually stored using eight bits, which indicates that the quantization level is

256 (28). That is, a pixel in a colour image requires a total storage of 24 bits (3*8). A 24 bit memory can express as 224 =256×256×256=1, 67, 77,216 distinct colours. The number of colours should adequately meet the display effect of most images. Such images may be called true colour images where information of each pixel is kept by using a 24-bit memory.

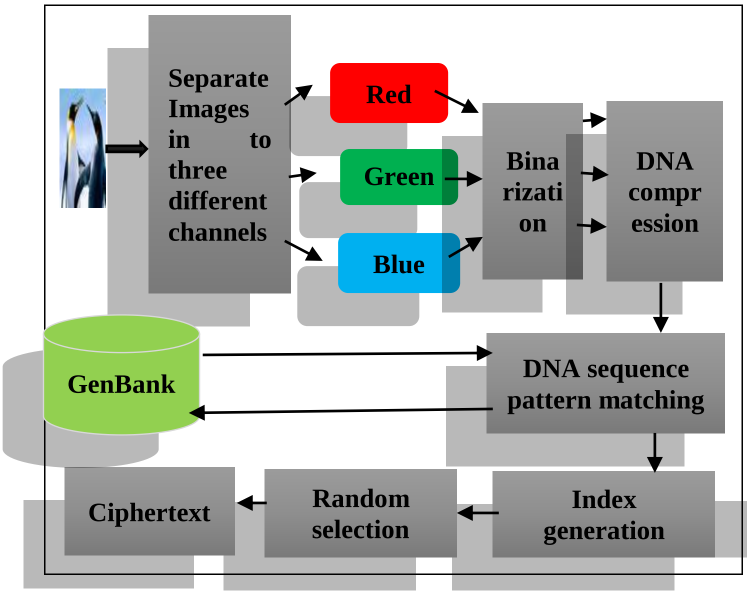


Fig 3.Block diagram for Image Encryption via image to binary conversion

Step 1. Read colour image
Step 2. Split the true colour image into three separate channel.
Step 3. Convert channel into binary.
Step 4. Repeat step 3 for each channel.
Step 5. End

C. Codebook Generation

In this work codebook is used for key exchange. This codebook is exchanged with receiver before the actual data transmission begins. For generating this codebook I am using a part of TEA. It is called DNA Compression algorithm (DCA).

Step 1. Start
Step 2. Read the key
Step 3. Split the key into characters.
Step 4. Convert character to its corresponding ASCII value.
Step 5. Convert ASCII to binary
Step 6. Encode binary string to 4 char sequence (01=A, 10=C, 11=G, 00=T)
Step 7. Repeat step 4 to 7 until end of the key.
Step 8. End

IX. DESIGN OF DNA CRYPTOSYSTEM

The proposed DNA cryptosystem decomposes into subsystem, which includes codebook generation, key generation, encryption, and decryption. For key transmission, the codebook is used. In this method, the plain text which is the original message is converted to ASCII form and then to binary form and again to the DNA form. The OTP keys are generated randomly from the public database. This OTP key and the DNA form of the plain text are compared and a random index is generated, which is the encrypted data. Decryption process is carried out in the opposite order to obtain the original plain text message. The encryption and decryption for both image and text is through the same algorithm.

In the case of image encryption two preprocessing techniques are used. By using an algorithm image is firstly converted to text then follow the same procedure as for the text encryption. The other technique converts image to base2 binary first and then follow the same procedure as for the text encryption. A comparative study in between these two method is included. This work is based on conventional cryptographic method. The phases included are key generation, encryption and decryption. Encryption is based on symmetric key cryptography.

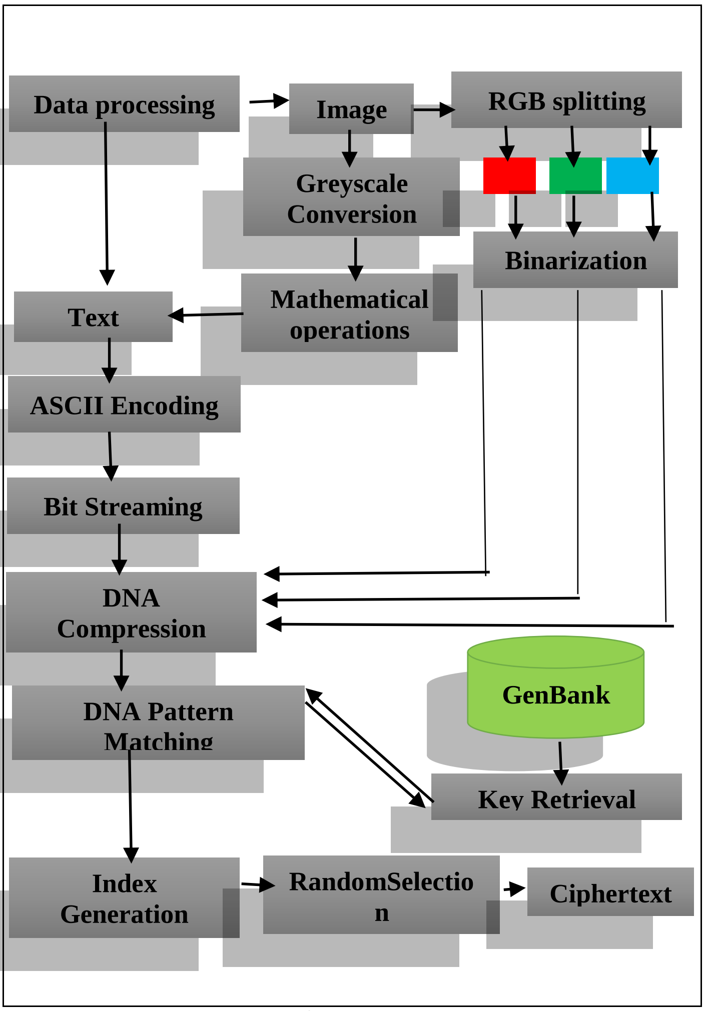


Fig 4. Block diagram for proposed DNA Cryptosystem

a. MODULE DESCRIPTION

The entire work is plan to be done in four modules.

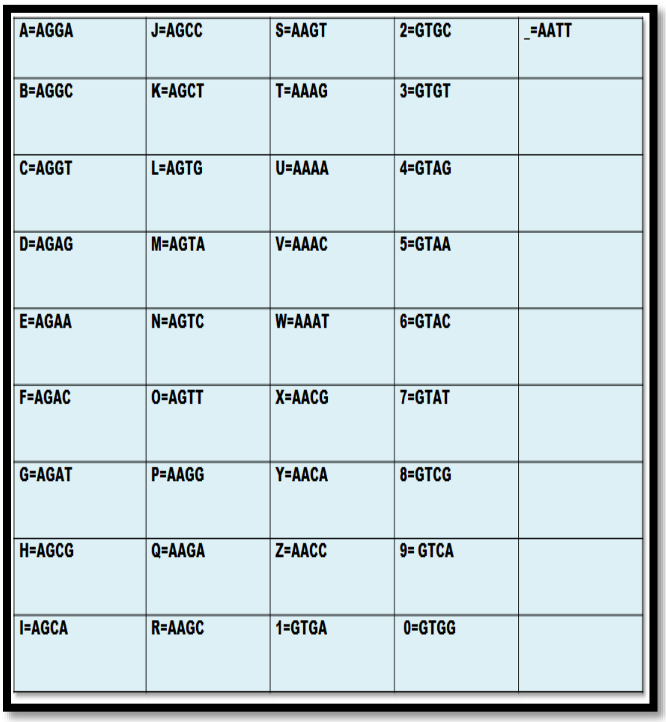
A. CODEBOOK GENERATION
B. KEY GENERATION
C. ENCRYPTION

D. DECRYPTION

A. *CODEBOOK GENERATION*

A codebook is a type of document used for gathering and storing codes. Originally codebooks were often literally books, but today codebook is a byword for the complete record of a series of codes, regardless of physical format. In cryptography, a codebook is a document used for implementing a code. A codebook contains a lookup table for coding and decoding; each word or phrase has one or more strings which replace it. To decipher messages written in code, corresponding copies of the codebook must be available at either end. The distribution and physical security of codebooks presents a special difficulty in the use of codes, compared to the secret information used in ciphers, the key, which is typically much shorter.

In this work codebook is used for key encryption and transmission, which is accession number. An accession number is a combination of block letters of English alphabets, numerals 0-9 and the special symbol '_' (underscore).This accession number has to be kept secret and transmitted to the receiver for decryption. In view of keeping the accession number secret, a codebook is provided for accession number. The codebook is generated with the help of DNA compression algorithm. The importance of the codebook is that it has to be exchanged at least once in between the sender and receiver via publicly or privately before the actual data transmission begins.



A=AGGA	J=AGCC	S=AAGT	2=GTGC	_=AATT
B=AGGC	K=AGCT	T=AAAG	3=GTGT	
C=AGGT	L=AGTG	U=AAAA	4=GTAG	
D=AGAG	M=AGTA	V=AAAC	5=GTAA	
E=AGAA	N=AGTC	W=AAAT	6=GTAC	
F=AGAC	O=AGTT	X=AACG	7=GTAT	
G=AGAT	P=AAGG	Y=AACA	8=GTCG	
H=AGCG	Q=AAGA	Z=AACC	9= GTCA	
I=AGCA	R=AAGC	1=GTGA	0=GTGG	

Table 1. The Codebook

B. *KEY GENERATION*

Key is a piece of information or parameter that determines the functional output of a cryptographic algorithm or cipher. Without a key, the algorithm would produce no useful result. In encryption, a key specifies the particular transformation of plaintext into ciphertext, or vice versa during decryption. Keys are also used in other cryptographic algorithms such as digital signature schemes and message authentication codes. A cryptographic key is a string of bits used by a cryptographic algorithm to transform plain text into cipher text or vice versa. This key remains private and ensures secure communication.

A cryptographic key is the core part of cryptographic operations. Many cryptographic systems include pairs of operations such as encryption and decryption. A key is a part of the variable data that is provided as input to a cryptographic algorithm to execute this sort of operation. In a properly designed cryptographic scheme, the security of the scheme is dependent on the security of the keys used.

In the proposed work, key generation is based on one time padding. OTP key is directly picked from the public genetic database GenBank with the help of Matlab bioinformatics toolbox. An accession number or Accession ID is used to retrieve the key from the public database. This accession number is kept secret and transmitted to the receiver through codebook. In this proposed system symmetric key is used. In symmetric key systems, same key is used for both encryption and decryption.There is a function 'getbank' to load the DNA string from the NCBI database. The following function extracts the DNA sequence from the NCBI bank.

```
Mitochondria=getgenbank('NC_001807','SequenceOnly',true)
;
```

One-Time-Pad (OTP) is a principle of key generation applied on the stream ciphering method which offers a perfect secrecy, if all the requirements are fulfilled. It is also considered that this scheme is unbreakable in theory, but difficult to realize in practical applications. The one-time pad is a long sequence of random letters. These letters are combined with the plaintext message to produce the ciphertext.

To decipher the message, a person must have a copy of the one-time pad to reverse the process. A one-time pad should be used only once (hence the name) and then destroyed. This is the first and only encryption algorithm that has been proven to be unbreakable. Claude Shannon described in his work the principles for perfect secrecy. These characteristics for the unbreakable encryption system are the same with the OTP properties. They can be summarized as the following constrains on the encryption key:

- It must be truly random
- At least as large as the plain-text\

i. GenBank

Biological databases are huge data bases of mostly sequence data pouring in from many genome sequencing project going on all over the world. They are an important tool in assisting scientists to understand and explain a host of biological phenomena from the structure of biomolecules and their interaction to the whole metabolism of organisms to understanding the evolution of species. This knowledge helps facilitate to fight against diseases, assists in the development of medications and in discovering basic relationships amongst species in this history of life.

Information, NCBI, is a US-based organization founded in 1988 as a division of the National Library of Medicine (NLM) at the National Institutes of Health (NIH). NCBI is one of the most important public resources for DNA and protein sequence database, other life sciences-specific databases, bioinformatics tools and services. There are different types of database but for routine sequence analysis, the following are initially the most important.

1. Primary databases: Contain sequence data such as nucleic acid or protein.

Protein Databases	Nucleic Acid Databases
• SWISS-PROT	• EMBL
• TREMBL	• GenBank
• PIR	• DDBJ

Table 2. Primary databases

1. Secondary databases: These are also known as pattern databases contain results from the analysis of the sequences in the primary databases.

Secondary databases
• PROSITE
• Pfam
• BLOCKS
• PRINTS

Table 3. Secondary databases

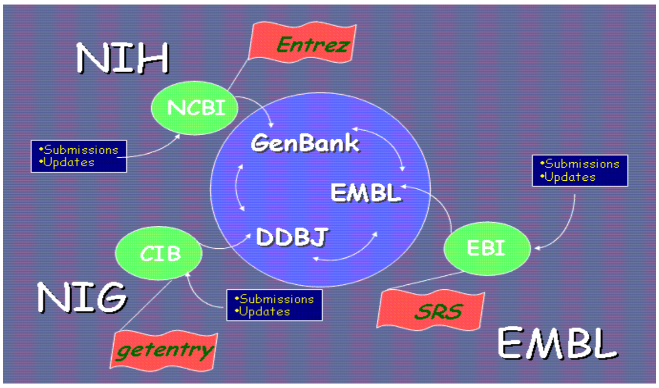


Fig 5. International collaboration of genetic database

For the implementation of work, I choose GenBank. The GenBank sequence database is an open access, annotated collection of all publicly available nucleotide sequences and their protein translations. This database is produced at National Center for Biotechnology Information (NCBI) as part of the International Nucleotide Sequence Database Collaboration, or INSDC. GenBank and its collaborators receive sequences produced in laboratories throughout the world from more than 100,000 distinct organisms. GenBank continues to grow at an exponential rate doubling every 10 months.

For this thesis work, I collected some accession numbers for key retrieval processes. The table given below indicates common name, scientific name and corresponding accession numbers of some species. For key generation Matlab bioinformatics tool box is used for retrieving key sequence from the GenBank. The data accession is through the taxonomy browser in NCBI.

No	Name	Scientific Name	Accession Number
1	Tiger	Panthera Tigris	JZ331708
2	Elephant	ProboqscideaElephantidae	CC935997
3	Mouse	Musmusculus	DE999383
4	Dog	CannisFamiliaris	AY345584
5	Carrot	Daucuscarota	AB027706
6	Papaya	Carica	DS981520
7	Pineapple	Ananascomosus	HM104185
8	Guava	Psidiumguajava	GU135421
9	Onion	allium cepa	AB627990
10	Cucumber	Cucumis sativas	DI183231
11	Brinjal	Solanummelongena	FJ842522
12	Tomato	Lycopersicon esculentum	AY097064
13	Orange	Citrus aurantium	EF138853
14	Potato	Solanum tubersum	L34218
15	Mango tree	Mangiferaindica	JX316911
17	Aloevera	Aloe Barbdensis miller	KJ557601
20	Hibiscus	Hibiscus	AB817499

Table 4. Accession Number

C. ENCRYPTION

The proposed work focus data in the form of text and image. If data is in text, encrypt using TEA. In the case of image as data, before encryption one of the two different image preprocessing techniques are applied and follow the procedures for TEA. A comparative study based on time complexity will also be included in this work.

I. TEXT ENCRYPTION.

In Text Encryption algorithm include following steps.

i. ASCII Encoding
ii. Bit streaming
iii. DNA Encoding
iv. DNA Encoding rule
v. DNA pattern matching
vi. Index Generation
vii. Random Extractor
viii. Ciphertext Generation

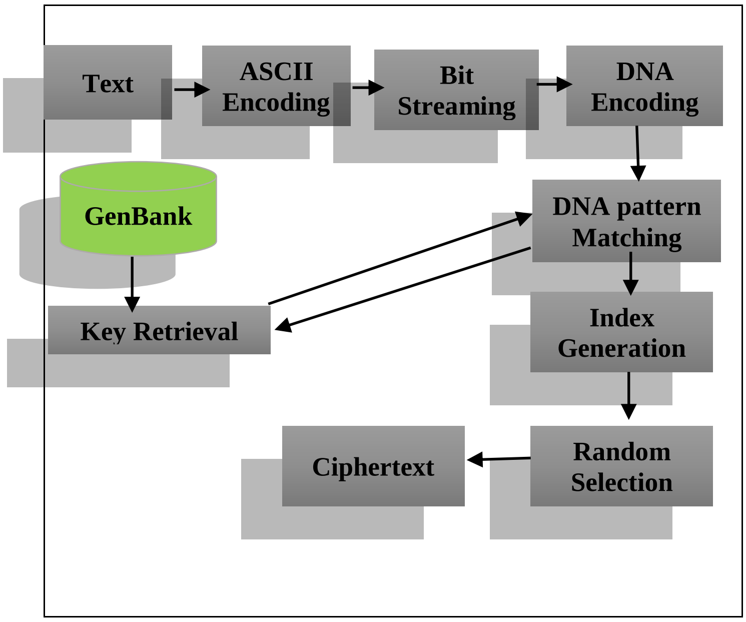


Fig 6. Block diagram for Text Encryption

i. *ASCII Encoding*

ASCII stands for American Standard Code for Information Interchange. ASCII was first introduced in 1968 as a method of encoding alphabetic and numeric data in digital format. Although ASCII code was originally developed for the teletypewriter industry, it has since found widespread use in computer and information-transfer technologies. Because ASCII code is standardized, computers and other electronic devices can use it to exchange data with each other.

This is true even for computers that use different operating systems. As originally formulated, each ASCII-encoded representation consist of a string of seven digits, where each digit was either a 0 or a 1 (i.e., binary code). This results in 128 possible ways of arranging 0s and 1s. In this representation, each alphanumeric character was uniquely assigned a number between 0 and 127, which was represented by its binary equivalent in a string of seven 0s and 1s.

ii. *Bit streaming*

A bit stream is a contiguous sequence of bits, representing a stream of data, transmitted continuously over a communications path, serially (one at a time).In this step, ASCII to binary conversion is performed. The size of the bit stream is 7 bit.

iii. *DNA compression*

DNA compression is also called DNA encoding method or DNA digital coding. The binary form of the plaintext is converted to DNA sequence. This binary sequence having eight bits is converted into four 2 bit characters using DNA encoding rule.

iv. *DNA encoding rule*

Data encoding and decoding are very important processes that are performed at the sender and receiver points of a data communication system respectively. Data encoding is also called compression or packing of data whereas decoding is also called decompression or unpacking of data. Encoding and Decoding are opposite of each other.

Data encoding or compression is frequently used when transmitting large quantities of data there by reducing the number of blocks transmitted and thus reducing the cost as well as the probability of transmission errors. For encoding and decoding different rules are used. In this proposed system use encoding and decoding rule is based on DNA. In the field of information science, the most basic encoding method is binary encoding. This is because everything can be encoded by the two states of 0 and 1. However, for DNA there are four basic units:

1. Adenine (A);
2. Thymine (T);
3. Cytosine (C);
4. Guanine (G)

Single-strand DNA sequence is composed by four bases, they are A, C, G and T, where A and T are complement to each other, so are C and G. In the modern theory of electronic computer, all information is expressed by binary system. But in DNA coding theory, information is represented by DNA sequences. So we use binary numbers to express the four bases in DNA sequence and two bits binary number to represent a base.

In the theory of binary system, 0 and 1 are complementary, so we can obtain that 00 and 11, 01 and 10 are also complementary. We can use 00, 01, 10 and 11 to express four bases and the number of coding combination kinds is 4! = 24. Due to the complementary relation between DNA bases, there are only eight kinds of coding combinations that satisfy the principle of complementary base pairing in 24 kinds of coding combinations.

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
00 – A	00 – A	00 – C	00 – C	00 – G	00 – G	00 – T	00 – T
01 – C	01 – G	01 – A	01 – T	01 – A	01 – T	01 – C	01 – G
10 – G	10 – C	10 – T	10 – A	10 – T	10 – A	10 – G	10 – C
11 – T	11 – T	11 – G	11 – G	11 – C	11 – C	11 – A	11 – A
9	**10**	**11**	**12**	**13**	**14**	**15**	**16**
00 – C	00 – C	00 – C	00 – C	00 – G	00 – G	00 – G	00 – G
01 – T	01 – T	01 – A	01 – A	01 – A	01 – A	01 – C	01 – C
10 – G	10 – A	10 – G	10 – T	10 – C	10 – T	10 – A	10 – T
11 – A	11 – G	11 – T	11 – G	11 – T	11 – C	11 – T	11 – A
17	**18**	**19**	**20**	**21**	**22**	**23**	**24**
00 – G	00 – G	00 – C	00 – T	00 – T	00 – T	00 – T	00 – T
01 – T	01 – T	01 – A	01 – A	01 – C	01 – C	01 – G	01 – G
10 – A	10 – C	10 – C	10 – G	10 – A	10 – G	10 – A	10 – C
11 – C	11 – A	11 – G	11 – C	11 – G	11 – A	11 – C	11 – A

Table 5. DNA encoding rule

Example: The binary pixel value of an image is [00111010], so the corresponding DNA sequence is [ATGG] according to the first encoding rule, similarly according to the seventh decoding rule, the decoding sequence is In the proposed algorithm, one of the 24 combination is chosen for encoding and decoding. The assigned value for bio-alphabets ACGT are 01=A, 10=C, 11=G and 00=T [11001010].

v. DNA Pattern Matching

In DNA pattern matching method, compare DNA form of the plaintext and DNA form of the key sequence. Brute force pattern matching algorithm is used for comparison.

vi. DNA indexing

As a result of DNA pattern matching steps, index is generated for each character in the plaintext.

vii. Random Selection

From the index group, one index is randomly choose for each character, which is the ciphertext.

II. IMAGE ENCRYPTION

In the proposed work image encryption has two options. One is image is converted into text and then the same procedure as for the TEA. Second one is image is converted into binary, then the same procedure as for the TEA.

a. Image to binary
b. Image to text

i. Image to binary

In image to binary, read the colour image and convert this image into three different channels red green, and blue. Each channel is converted into binary separately. This is called binarization.

ii. Image to text

Images can also be stored as text files by converting the corresponding pixel values to ASCII characters. This project focuses on converting pixel values to English letters (A to Z, a to z), which may be stored in a text file. The compressed image matrix stored in RLE array is written into a text file, which is a form of image to text conversion and has an added feature of hiding images as text files. The image can be reconstructed from this text file by applying the reverse process [16].

D. DECRYPTION

The proposed work is based on symmetric key encryption. So same DNA sequence is used during the decryption processes. Each integer from the ciphertext is used as pointer into the key sequence. The Receiver reads 4 letters from the indicated position and transforms them to binary representation using the same reversible DNA encoding rule. The plaintext is reconstructed when all the bytes are retrieved from the indicated positions.

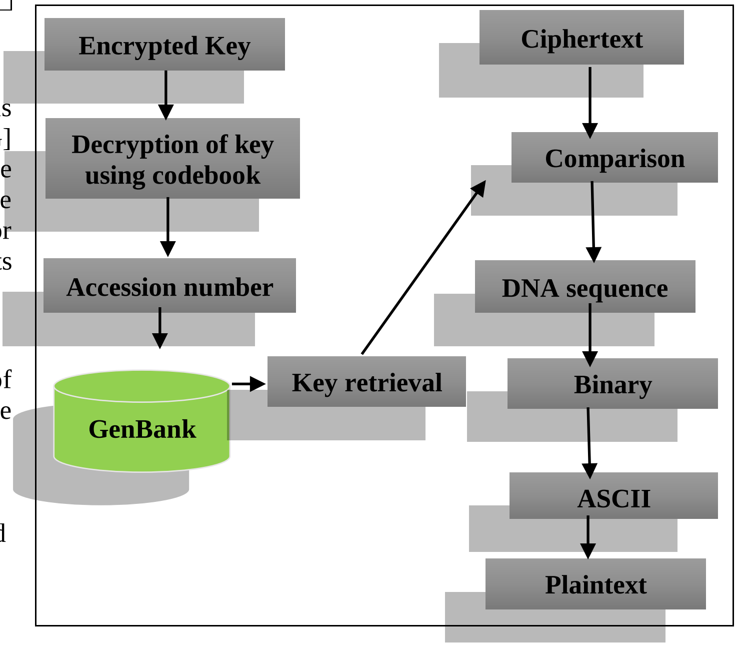


Fig 7. Block Diagram for Decryption

X. CONCLUSION

In this work, a new DNA cryptographic system is introduced, which can solve the issues in existing conventional cryptographic method. The proposed algorithm is suitable for both types of data-text and image. Key generation is based on OTP. OTP is an unbreakable encryption method used in cryptography. DNA cryptography is combine the advantage of both cryptography and bio-molecular computation. DNA cryptography uses only the concept of DNA that is computation using DNA, but not computation on DNA.

ACKNOWLEDGMENT

I would like to express my sincere gratitude to Dr. C.G. Sukumaran Nair (HOD), Associate Professor, Ms. Sudha S.K. and Ms. Renji S., Department of Computer Science and Engineering, Sarabhai Institute of Science and Technology, for their valuable guidance.

REFERENCES

[1] Adleman, Leonard M, "Molecular computation of solutions to combinatorial problems," Science-AAAS-Weekly Paper Edition 266, no. S5187, 1994.

[2] Hiding messages in DNA microdots Catherine Taylor Clelland1, Viviana Risca2 & Carter Bancroft1

[3] J. Lipton Richard, "DNA solution of hard computational problems", Science 268.5210: 542-545, 1995.

[4] Boneh Dan, Christopher Dimworth, Lipton, Richard J. "Breaking DES Using a Molecular Computer," DNA based computers 27, 37: 1996.

[5] Ouyang Qi, D. Peter Kaplan, Liu Shumao and Albert Libchaber, "DNA solution of the maximal clique problem," Science 278, 5337, 446-449, 1997

[6] M. E. Borda, O. Tornea, T. Hodorogea, "Secret Writing by DNA Hybridization", ActaTehnicaNapocensis, vol. 50, pp. 21-24, 2009.

[7] Zhang Yunpeng, Zhu Yu, Wang Zhong, Richard O.Sinnott, "Index-Based Symmetric DNA Encryption Algorithm", 2011 4th International Congress on Image and Signal Processing, IEEE.

[8] TusharMandge, Vijay Choudhary, "A DNA encryption technique based on matrix manipulation and secure key generation scheme", ICICES Journal, 2013,Print ISBN:978-1-4673-5786-9, pp.47-52.

[9] Surendra Varma, K.Govinda Raju, "Cryptography based on DNA using random key generation scheme", International Journal of Science Engineering and Advance Technology, IJSEAT ,2014, Vol 2, Issue 7, ISSN 2321-6905, pp.168-175.

[10] Bonny BRaj, Panchami, "DNA based cryptography using permutationand random key generation method, International Conference On Innovations & Advances In Science, Engineering And Technology",2014, Volume 3, Special Issue 5, ISSN (Online) : 2319 – 8753, ISSN (Print) : 2347 – 6710, pp.263-267.

[11] Ashish kumarkaundal, "Feistel Inspired structure for DNA cryptography" in June (2014).

[12] Ritu Gupta, Anchal Jain ―A New Image Encryption Algorithm based on DNA Approach‖ International Journal of Computer Applications (0975 – 8887) Volume 85 – No 18, January 2014

[13] ToshithaKannan, 2M. SindhuMadhuri "" "virtual dna based cryptography for enhanced security", International Journal of Electrical, Electronics and Computer Systems (IJEECS), ISSN (Online): 2347-2820, Volume -2, Issue-11,12 2014

[14] AsishAich, Alosen, SatyaRanjan Dash and SatchidAnandaDehuri ""A Symmetric Key Cryptosystem Using DNA sequence with OTP Key" , ISSN 2194-5357 , Springer India, Proceedings of Second International Conference INDIA 2015, Volume 2

[15] Isha Yadav, Nipun Gupta, and M.K. Beniwal "a new cipher text generation technique by digitizing the genetic dna code using random number, ijiset - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 6, June 2015.

[16] RENJI S "Application of Haar Wavelet Transform for Image Compression and Storage of Colour Information in Gray and Back" 2nd National Conference on Emerging Trends In Computing 2013 13, June 2013 - 15, June 2013