Cryptography using DNA Nucleotides

Mazhar Karimi
Muhammad Ali Jinnah University
Karachi, Pakistan

Waleej Haider
Sir Syed University of Engineering& technology
Karachi, Pakistan

ABSTRACT
Cryptography plays a key role in information security. Many new algorithms and techniques have been used in the same regards. Cryptography using DNA computing is very current state of the art. DNA cryptography comes with the next level of data integrity and confidentiality to protect information from intrusions, in this paper a cipher solution is proposed with a new symmetric key generation model based upon DNA strands, nucleotides, codons base pair rules, mutation and DNA to mRNA conversion. This solution emphasizes on usage of biological processes & the random changes found in DNA and simulate those processes in the key generation model.

General Terms
Algorithms, Cryptography, Information Security.DNA Cryptography, Data Integrity.

Keywords
Cryptography, DNA Cryptography, Symmetric Key Generation, Symmetric Key model, Information Security.

1. INTRODUCTION
Cryptography came into existence when human being started to realize the importance of information, and started get worried about its privacy [3] Information security is based upon 3 major elements (confidentiality, integrity, authenticity). By use of a Cryptography technique one can hide some information in a way which is not publicly readable (confidentiality). Over the time many techniques have been researched to secure the data using cryptography. Some are mathematical, some techniques involves core knowledge of physics [5].

In late 19's, Adleman resolved HPP (Hamiltonian path problem) using DNA (Deoxyribonucleic acid) Operations [1], which became the base of DNA computation.

2. BIOLOGICAL FRAMEWORK
DNA is a molecule, inside every organism. James Watson and Francis Crick formed first 3D structure of DNA based upon an X-Ray print. They found out that DNA structure is double helix/ stranded [11]like a spiral ladder. Sugar and phosphate make bond to form each strand. Each helix consists of other monomers (a molecule, that can be bonded to form a polymer) called nucleotides. Each nucleotide has sugar and phosphate groups and nitrogen base. These nitrogen bases are Adenine (A), Thymine (T), Guanine (G) and Cytosine (C).

Watson and Crick [11] observed that Adenine always make a bond with Thymine (T), and Guanine makes bond with Cytosine (C).This base pairing rule is also called complementary theory of Watson-Crick [Fig 1].Three consecutive nucleotides sequence is called Codon . Different combinations of Codon create amino acids and different combination of amino acids generates different proteins [7]. But process of creating different proteins is based upon a process called transcription which is discussed in later sections.

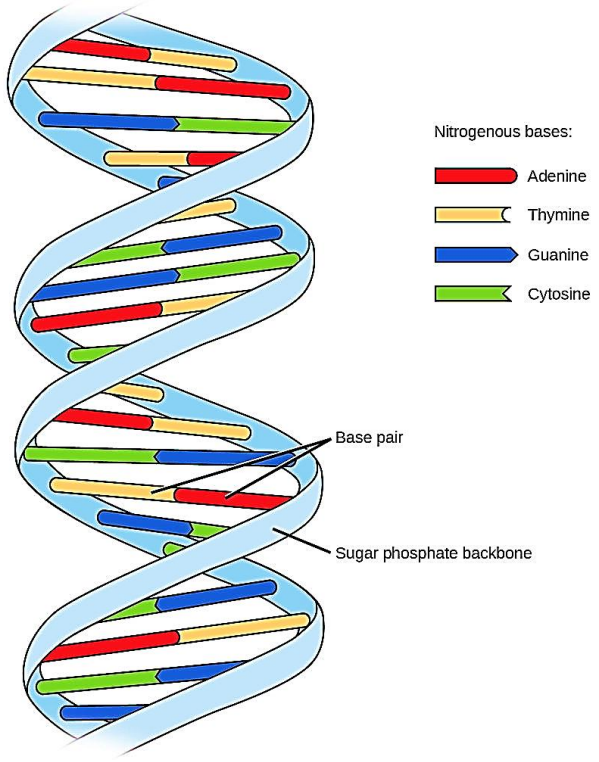

Fig1. DNA structure, its double helix form and nucleotides

2.1 Transcription
In this process Thymine in DNA gets replaced by Uracil (U). Now DNA has become RNA (Ribonucleic Acid).When transcription starts, and codons start to create protein, the generation of protein does not get stopped until a Stop Codon occurs, which is sequence of nucleotides which does not create an amino acid.

2.2 Replication
In this process, a sequence of nucleotides called primer are replicated, in the field of molecular biology, a process called polymer chain reaction (PCR) is used for replication. To start the process an enzyme is added and the whole process starts as a chain reaction. It is also called amplification. Using this technique a large quantity of DNA can be prepared from a very little quantity of DNA.

2.3 Anneal
In this process two single strands form their complementary strands and become double stranded DNA.

2.4 Marking
In the process of Transcription, start codon marks where to start the process, and stop codon marks where to stop the process to create a protein. In RNA stop codons are UAG (Amber), UAA (ochre) and UGA (opal).

2.5 Mutation

There are different types of mutations. Only non-sense and missense mutation are covered in the methodology.Non-sense mutation code for a stop codon, which terminates the process of protein generation. Missense mutation codes for a different codon, which generates a different protein.In Table 1, in non-sense mutation, CAG converts to UAG, CGA to UGA and UGG to UGA. And in missense mutation, GAU converts to GUU, GAG to GUG and GCC converts to ACC.

Table 1. Codon Conversion in Mutation Process.

From	To
CAG	UAG
CGA	UGA
UGG	UGA
GAU	GUU
GAG	GUG
GCC	ACC

DNA holds all genetic information required for an organism to have a form and perform functions. Genetic Code is a set of rules, which governs how information would be encoded in DNA. Genetic code is based upon codons, the order is important to create a protein. There are 64 codons, out of them 3 are stop codons, other 61 codons codes for a specific amino acid. There are many different codons which create the same amino acid. There are only 20 amino acids coded by 61 codons.

When Adleman [1] started to research in molecular biology and realized that these 4 letters (A, T, C and G) holds all the information required by an organism, and GNA processes and functions can be used for computation he successfully resolved an mathematical problem in NP complete time, which complexity problems was O(n) in silicon chips and the same problem resolved in O(1) using DNA. That was the beginning of DNA computing.

3. LITERATURE REVIEW

Hossain [4] used DNA cryptography; their method only supported ASCII characters. They used multiple rounds with substitution and transposition technique, where they generated a dynamic table of nucleotides sequence.

Ochani [9] proposed the solution to transmit a DNA image.They used modified symmetric key encryption with LSB steganography technique, but did not mention about what would be the cover image, and did not provide image compression test with regards to image size and bit quality.

Chen [2] used OTP (one time padding) and encrypted an image using DNA cryptography.

Li [6] used key expansion algorithm and diffused the user password with XOR in even number order. For key expansion they used an s-box in mars algorithm to create a random number table.

Surendra and govinda[10] used ASCII to nucleotide encoding in the first stage and in later stage they used row shifting and XOR operations.

Kang [8] used the biological methods for encryption and discussed the weakness of his method in detail.

4. METHODOLGOY

DNA cryptography is a theoretical computer science field where DNA is used for information hiding. The smallest DNA is of 30 nucleotides.DNA is an information storage, few grams of DNA can consist all the data available over the internet. DNA uses nucleotides to store the information.

The methodology is based upon method discussed in biological framework section. Number of rounds and number of keys are random, depends on the user given key.

Bits are encoded asshown in Table 2. (00) as Adenine (A), (11) as Cytosine (C), (01) as Guanine (G) and (10) as Thymine (T).

Table 2. Bits Encoding in to Nucleotides.

Bits	Nucleotide
00	A
11	C
01	G
10	T

4.1 Key Model

After binary number conversion to Nucleotides, a process of replication is applied if the length of nucleotides is less than 60, afterwards if the length of Nucleotides is not perfect divided by a codon length, the last codon is repaired by appending the last nucleotide to the sequence. Then process of anneal is started and the single helix is now bonded with its pair according to complementary rule. A full DNA is formed now. Now DNA conversion to mRNA is started, so mutation can be applied, after the mutation is applied, the separated proteins are called DNA keys. Number of DNA keys depends upon how many break codons are found and how amino acid type of Tryptophan (UGG), Glutamine (CAG), Arginine (CGA), Alanine (GCC) and Aspartic Acid (GAU) are found and converted to stop codons or other amino acids.

When all the final DNA keys are created, they are decoded to 8 bits blocks.

4.1.1 Example

Suppose text message bits are

01101101011001010111001101110011011000010110011101100101

And user password bits are

01000000001000000101111100100000001010110010000000011001000101101

When the key bits are encoded in Nucleotides, the nucleotides sequence is:

GAAAATAAGGCCATAAATTCATAAACATATCGGAAAATAAGGCCATAAATTCATAAACAT

When the processing of annealing is applied, double helix DNA is shown as:

GAAAATAAGGCCATAAATTCATAAACATATCGGAAAATAAGGCCATAAATTCATAAACATCTTTTATTCCGGTATTTAAGTATTTGTATAGCCTTTTATTCCGGTATTTAAGTATTTGTA

When process of transcription is started, the sequence becomes like following.

GAAAAUAAGGCCAUAAAUUCAUAAACAUAUCGGAAAAUAAGGCCAUAAAUUCAUAAACAUCUUUUAUUCCGGUAUUUAAGUAUUUGUAUAGCCUUUUAUUCCGGUAUUUAAGUAUUUGUA

When applied the point mutation on mRNA and stop cordons are occurred.

Following proteins (keys) are created

GAAAAUAAGGCCAUAAAUUCAUAAACAUAUCGGAAAAUAAGGCCAUAAAUUCAUAAACAUCUUUUAUUCCGGUAUUUAAGUAUUUGUAUAGCCUUUUAUUCCGGUAUUUAAGUAUUUGUA

ACAUAUCGGAAAAUAAGGCCAUAAAUUCAUAAACAUCUUUUAUUCCGGUAUUUAAGUAUUUGUAUAGCCUUUUAUUCCGGUAUUUAAGUAUUUGUA

AUUCAUAAACAUCUUUUAUUCCGGUAUUUAAGUAUUUGUAUAGCCUUUUAUUCCGGUAUUUAAGUAUUUGUA

GUAUUUGUAAAA

When encoded the above mRNA keys to binary

010000000111000000100000000000110100000000111000000100000000011101010110110101000101010101001111010101101101010001010101010011110101011011010100010101010

000000110100000001110000001000000000111010101101101000101010101001111010101101101010001010101010

001000000000111010101101101010001010101010011110101011011010100010101010

010010010000

4.2 Encryption

Each DNA key creates a binary block

$$B = \{b1, b2\ldots bn\}.$$

In every binary block, each 8 bits block denoted as

$$bi = \{k1, k2 \ldots. kn\}.$$

b1= 01000000 01110000 00100000 00000011 01000000 01110000 00100000 00001110 10101101 10101000 10101010 10011110 10101101 10101000 10101010

b2=00000011 01000000 01110000 00100000 00001110 10101101 10101000 10101010 10011110 10101101 10101000 10101010

b3 = 00100000 00001110 10101101 10101000 10101010 10011110 10101101 10101000 10101010

b4 = 01001001 00000000

For encryption first 8 bits of text message is picked up, and left-shifted to 1 bit and apply XOR function with each key in block 1, this continues for all the {k1, k2 ...kn} in b1.

$$CM = (M << 1) \oplus b1kj$$

The second binary block shifts message to 2 bits and apply XOR function with each key in block 2, and this also continues for all the keys in block b2.

$$CM = (CM << 2) \oplus b2kj$$

That can be described as:

$$CM = (CM << i) \oplus bikj$$

Final cipher is:

1110100111101001111010011110100111101001111010011110100111101001

4.3 Decryption

The process of decryption is a reverse of encryption process.

5. CONCLUSION

In this approach the key size is not fixed unlike other block ciphers where key size is fixed like AES. The focus of the methodology was that the number of rounds should be random. The key size and rounds depends upon user password, which brings a great random behavior in key model, which makes cryptanalysis even harder. This can be used for those networks where processing power is not a question because user password may produce a great variety of keys.

When the user password is small and key is replicated, the partial duplicate pattern can be observed when similar nucleotides are found in great amount and key is broken in to multiple keys at time of key generation after mutation. But multiple rounds of encryption reduce this vulnerability.As number of keys and rounds are random, and time complexity is also random, small networks may suffer. But this can be avoided in future, if a limit of keys is introduced in key generation model based upon user password.

In future more biological and computation methods can be included to improve the methodology.

6. ACKNOWLEDGMENTS

Muhammad Mazhar Karimi likes to thanks his teacher Mr. Waleej Haider and Muhammad Ali Jinnah University for motivation and guidance.

7. REFERENCES

[1] Adleman, L. M. 1994. Molecular computation of solutions to combinatorial problems.

[2] Chen, J.2003. A DNA-based, bio molecular cryptography design. Circuits and Systems.

[3] Churchhouse, R. F. 2002. Codes and ciphers: Julius Caesar, the Enigma, and the Internet

[4] Hossain, E. M. 2016. A DNA cryptographic technique based on dynamic DNA sequence table.

[5] Leuenberger, M. N. 2001. Quantum computing in molecular magnets.

[6] Li, X.-s. L.-p. 2008. A novel generation key scheme based on DNA.

[7] Needleman, S. B. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins.

[8] Ning, K. A. A pseudo DNA cryptography method.

[9] Ochani, A. D. 2016. DNA image encryption using Modified Symmetric Key (MSK).

[10] Varma, P. S. 2014. Cryptography based on DNA using random key generation scheme.

[11] Watson, J. D. 1953. The structure of DNA.