RESEARCH ARTICLE OPEN ACCESS

A Dna And Amino-Acids Based Implementation Of Four-Square Cipher

Sonal Namdev*, Vimal Gupta**
*(Student, Department of Computer Science & Engineering, Dr. A.P.J. Abdul Kalam Technical University, JSS Academy of Technical Education Noida. Uttar Pradesh. INDIA)
**(Assistant Professor. Department of Computer Science & Engineering, Dr. A.P.J. Abdul Kalam Technical University, JSS Academy of Technical Education Noida. Uttar Pradesh. INDIA)

ABSTRACT
The DNA cryptography is a new and very promising direction in cryptographic research. It is in the primitive stage. DNA cryptography is shown to be very effective. Currently, several DNA computing algorithms are proposed for many cryptography, cryptanalysis and steganography problems, and they are very powerful in these areas. This paper discusses a significant modification of the old approach of using DNA and Amino Acids based approach with Playfair Cipher to using the same approach with different encryption algorithm, i.e; foursquare cipher to the core of the ciphering process. In this study, a binary form of data, such as plaintext messages, or images are transformed into sequences of DNA nucleotides. Subsequently, these nucleotides pass through a Foursquare encryption process based on amino-acids structure. The fundamental idea behind using this type of encryption process is to enforce other conventional cryptographic algorithms which proved to be broken, and also to open the door for applying the DNA and Amino Acids concepts to more conventional cryptographic algorithms to enhance their security features.

I. INTRODUCTION

As some of the modern cryptography algorithms (such as DES, and more recently, MD5) are broken, the new directions of information security are to protect the data. By using the concept of DNA computing in the fields of steganography and cryptography, new powerful or even unbreakable algorithms can be designed[1]. DNA can be defined as a nucleic acid that contains genetic instructions that are used in the development and functioning of all known living organisms and some viruses[2]. The four bases found in DNA are adenine (abbreviated A), cytosine (C), guanine (G), and thymine (T). These four bases are attached to the phosphate/sugar to form the complete nucleotide. The genetic code can be expressed as either DNA codons or RNA codons. The DNA codons is read the same as the RNA codons except the nucleotide thymine (T) is found in place of Uridine. So in DNA codons we have (TCAG) and in RNA codons, we have (UCTG).

II. EXISTING FRAMEWORK

Although Playfair cipher is believed to be an old, simple and easily breakable cipher, some new modifications made it a more powerful encryption algorithm. This has been done by introducing concepts of confusion and diffusion to the core of the encryption process in addition to preserving the cipher's simplicity concept. In addition shortage in security features the plaintext message is restricted to be all upper case, without J letter, without punctuation, or even numerical values. These problems can be easily handled in any modern cipher as handled in this new algorithm[3].

The character form of a message or any form of an image can be easily transformed to the form of bits. This binary form can be transformed to DNA form through many encoding techniques implemented in previous work.

Playfair is based on the English alphabetical letters, so preserving this concept, we have used the English alphabet but from an indirect way. DNA contains four bases that can be given an abbreviation of only four letters (adenine (A), cytosine (C), guanine (G), and thymine (T)). On the other side, we have 20 amino acids with additional 3 codons to represent the Stop of coding region. Each amino acid is abbreviated by a single English character. So we are able to stretch these 20 characters to 26 characters. Amino acids are therefore represented as English alphabets. Then the DNA form is converted into amino acids form which is then passed through classical Playfair cipher. Through this conversion process, we have to keep in mind the problem of ambiguity; that most amino acids are given more than possible codon.

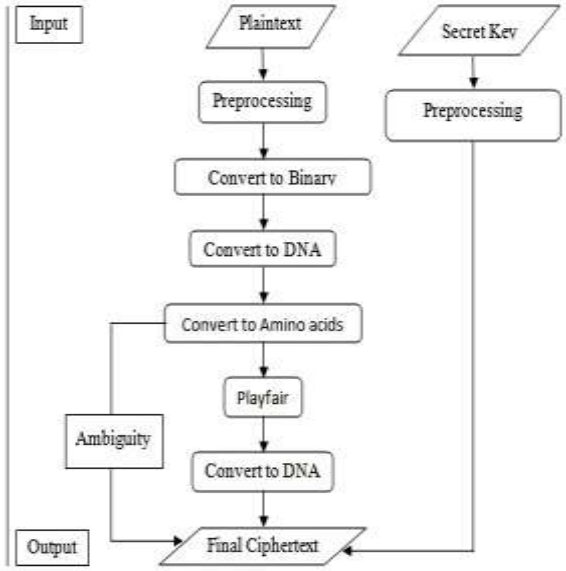

Fig: Flowchart based on PlayFair algorithm

III. PROPOSED OBJECTIVE

The main objective of this paper is to implement the DNA and Amino acid based approach of encryption and decryption on Four-square algorithm and thus to open the door for applying the DNA and Amino acids concepts[7] to more conventional cryptographic algorithms.

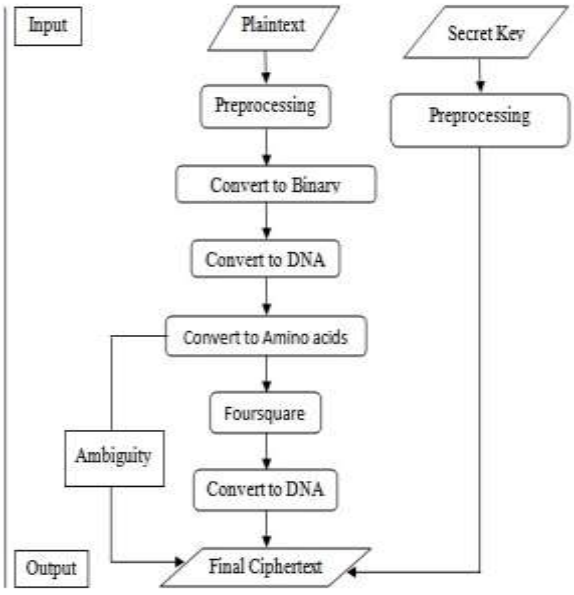

Fig: Flowchart based on Four-Square algorithm

IV. OVERVIEW OF DNA

4.1 How can we define Deoxyribonucleic acid 'DNA'?

DNA can be defined as a nucleic acid that contains genetic instructions that are used in the development and functioning of all known living organisms and some viruses. DNA molecules play main role in the long-term storage of information. DNA can be compared to a set of blueprints or a recipe, or a code, as it contains the instructions needed to construct other components of cell, such as proteins and RNA molecules[10]. Genes are the DNA segments that carry this genetic information, but other DNA sequences have structural purposes, or are involved in regulating the use of this genetic information. The DNA double helix is stabilized by hydrogen bonds between the bases attached to the two strands. The four bases found in DNA are adenine (abbreviated A), cytosine (C), guanine (G), and thymine (T). These four bases are attached to the phosphate/sugar to form the complete nucleotide[11].

4.2 The Genetic Code

The genetic code consists of 64 triplets of nucleotides. These triplets are known as codons. With three exceptions, each codon encodes for one of the 20 amino acids used in the synthesis of proteins. That produces some redundancy in the code: most of the amino acids being encoded by more than one codon[12]. The genetic code can be expressed as either DNA codons or RNA codons. RNA codons occur in messenger RNA (mRNA) and are the codons that are actually "read" during the synthesis of polypeptides (this process is known as translation). But each messenger RNA molecule acquires its sequence of nucleotides by transcription from the corresponding gene. The DNA codons is read the same as the RNA codons except the nucleotide thymine (T) is found in place of Uridine. So in DNA codons we have (TCAG) and in RNA codons, we have (UCTG)[13].

4.3 Transcription and Translation

A gene is basically defined as a sequence of DNA that contains genetic information that can even influence the phenotype of an organism. Within a gene, the sequence of bases along a DNA strand defines a messenger RNA sequence, which then defines one or more protein sequences. The relationship between the nucleotide sequences of genes and the amino-acid sequences of proteins is determined by the rules of translation, known collectively as the genetic code. The genetic code consists of three-letter 'words' called codons formed from a sequence of three nucleotides (e.g. ACT, CAG, TTT). In transcription, by RNA polymerase the codons of a gene are copied into messenger RNA. By a ribosome, copy of this RNA is then decoded that reads the RNA sequence by base-pairing the messenger RNA to transfer RNA, which carries amino acids[14]. Since there are 4 bases in 3-letter combinations, there are 64 possible codons (43 combinations). These encode the twenty standard amino acids, giving most amino acids more than one possible codon[15]. There are also three 'stop' or

'nonsense' codons signifying the end of the coding region; these are the TAA, TGA and TAG codons.

V. DNA BASED FOUR-SQUARE ALGORITHM

5.1 DNA Representation of Bits

Four-square used to be applied to English alphabet characters of plaintext. It was not able to encode any special characters or numbers which is considered a severe drawback that enforces the sender to write everything in the English letters. This problem appears while sending numerical data, equations or symbols.

On the contrary, in the algorithm proposed in this work, any numbers, special characters, or even spaces (not preferred) can be used in the plaintext. The encryption process starts by the binary form of data (message or image) which is further represented into DNA form according to Table 1. Then the DNA form is further transferred into Amino Acids form according to Table 2 which is considered as a standard universal table of Amino Acids and their codons representation in the form of DNA.

Note that each Amino acid has a name, abbreviation, and a single character symbol. This character symbol is what we will use in our algorithm.

Table I: DNA Representation of Bits.

BIT 1	BIT 2	DNA
0	0	A
0	1	C
1	0	G
1	1	U

5.2 Construction of Alphabet Table

In table II, we have only 20 amino acids in addition to 1 start and 1 stop. While we need 25 letters to construct each of the Foursquare matrix (note that I/J are assigned to one cell).

The letters we need to fill are (B, O, U , X, Z). So we will make these characters share some amino acids their codons. The start codon is repeated with amino acid (M) so we will not use it. (B) will be assigned 3 stop codons. (L, R, S) are three amino acids having 6 codons. By observing the sequence of DNA of each, we can figure out that each has 4 codons of the same type and 2 of another type. Those 2 of the other type are shifted to the letters (O, U, X) respectively. Letter Z will take one codon from (Y), so that Y: UAU, Z: UAC. Now the new distribution of codons is illustrated in Table III. When we count the number of codons of each character, we find that number varies between 1 and 4 codons per character. We define this number to be the 'AMBIGUITY' of the character [AMBIG]. Now, the distribution of English alphabet is complete. So a message in the form of a Amino Acids can go through traditional Foursquare cipher process using the secret key.

Table II: New distribution of codons

Ala/A	GCU, GCC, GCA, GCG	**Leu/L**	UUA, UUG, CUU, CUC, CUA, CUG
Arg/R	CGU, CGC, CGA, CGG, AGA, AGG	**Lys/k**	AAA, AAG
Asn/N	AAU, AAC	**Met/M**	AUG
Asp/D	GAU, GAC	**Phe/F**	UUU, UUC
Cys/C	UGU, UGC	**Pro/P**	CCU, CCC, CCA, CCG
Gln/Q	CAA, CAG	**Ser/S**	UCU, UCC, UCA, UCG, AGU, AGC
Glu/E	GAA, GAG	**Thr/T**	ACU, ACC, ACA, ACG
Gly/G	GGU, GGC, GGA, GGG	**Trp/W**	UGG
His/H	CAU, CAC	**Tyr/Y**	UAU, UAC
Ile/I	AUU, AUC, AUA	**Val/V**	GUU, GUC, GUA, GUG
START	AUG	**STOP**	UAA, UGA, UAG

Table III: New Distribution of the alphabet with the corresponding new codon

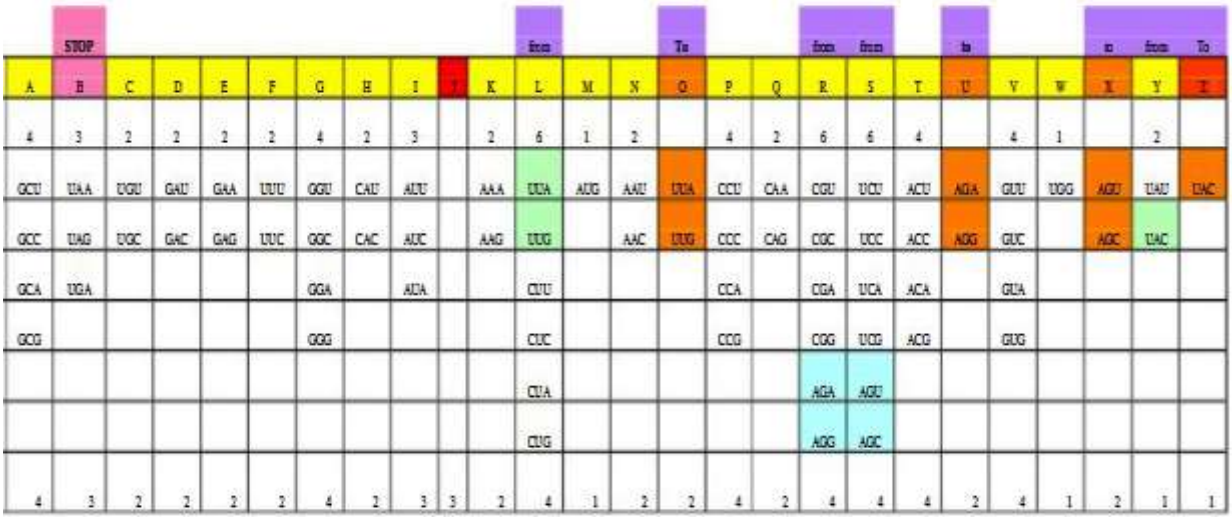

Table IV: New Distribution for codons on English Alphabet

A	GCU, GCC, GCA, GCG	E	GCU, GCC, GCA, GCG
R	CGU, CGC, CGA, CGG	K	AAA, AAG
N	AAU, AAC	M	AUG
D	GAU, GAC	F	UUU, UUC
C	UGU, UGC	P	CCU, CCC, CCA, CCG
Q	CAA, CAG	S	UCU, UCC, UCA, UCG
E	GAA, GAG	T	ACU, ACC, ACA, ACG
G	GGU, GGC, GGA, GGG	W	UGG
H	CAU, CAC	Y	UAU
I	AUU, AUC, AUA	V	GUU, GUC, GUA, GUG
B	UAA, UGA, UAG	O	UUA, UUG
U	AGA, AGG	X	AGU, AGC
Z	UAC		

DNA form of cipher text can be demonstrated also from Table IV by choosing random codons accompanied to each other. The concept that one character can have more than one DNA representation itself enhances the confusion concept that also enhances the algorithm strength. Table IV shows the new distribution of codons on the amino acids and additional alphabetical English letters according to our algorithm.

5.3 Decryption and Ambiguity Problem

The decryption process is simply the inverse/opposite of the encryption process unless that we find a problem in constructing the DNA form of plaintext from the amino acid form which is of length (L). The problem is that we are unable to choose that which codon has to be put in accordance to each amino acid character. We can relate this problem as the problem of codon-amino acid mapping problem arised with other algorithm which is based on the concept of Central Dogma like[4].

The way Nang handled this problem is to put this codon-amino acid mapping in the secret key to be sent through a secure channel[4]. This idea does not seem to be efficient as it increases the size of the key in relation to size of the plaintext.

The solution that we use for the above mentioned problem in our algorithm is located in two additional bits for each amino acid character to demonstrate which codon to choose[16]. We previously mentioned that each amino acid has 1, 2, 3 or 4 codons to represent it. This number can be put in 2 bits from 0-3.

We use Table 1 to convert these 2 bits into DNA form. This is the reason that final cipher text is both the DNA form of cipher text of length (3L) and the array carrying the ambiguity of length (L).

In decryption, the amino acid form of plaintext with the assistance of the ambiguity array can construct the correct form of plaintext in DNA form which can be transferred to binary form and then the final character form[17].

5.4 Pseudo-Code

INPUT:

[P] Plaintext (characters with spaces, numbers or any special characters).

[K] Secret Key (English characters without any number or special characters).

Algorithm Body:

Preprocessing:

1) -Prepare the secret key:

-Remove any spaces or repeated characters from [K].

-Put the remaining characters in the UPPER case form.

[K]→UPPER[K].

2) –Prepare the plaintext:

-Remove the spaces from [P] (done to avoid attacker's trace to a character which is repeated many times within the message).

Processing:

1) Binary form [BP] = Binary [P] (Replace each character by its Binary Representation-8 bits).
2) DNA form [DP] = DNA [BP] (Replace each two bits by their DNA Representation).
3) Amino Acids form [AP] = AMINO [DP] (Replace each three DNA characters by their Amino acid character keeping in track the ambiguity of each Amino acid[AMBIG]).
4) Construct the 4 Foursquare 5X5 matrices and add [K] row by row, then add the rest of alphabet characters.
5) Amino acid of cipher text [AC]=Foursquare [AP].
6) DNA form of cipher text [DC] = DNA [AC].

Output:

Add [DC] and [AMBIG] together in the suitable form→final cipher text [C].

VI. EXPERIMENT AND PERFORMANCE ANALYSIS

6.1 Experiment Inputs and Attributes

The experiment was performed on a famous novel 'Two States' cited by Chetan Bhagat.

We took paragraph from the beginning of the novel according to the estimated storage size in Kilobytes (from 1 Kb and increasing till 150 Kb).

6.2 System Parameters

The experiments were conducted using Intel (R) Core (TM) 2 CPU T5300, 1.73 GHz, 32 bit processor with 1 GB of RAM. The simulation program was compiled using the default settings in .NET 2005 visual studio for C# Windows applications under WINDOWS XP as the operating system. The experiments was performed several times to assure that the results are consistent and valid.

6.3 Experiment Factors

The chosen factor here to determine the performance is the algorithm's speed to encrypt data blocks of various sizes. Suppose we will use the original sequence of English alphabet and embed the ambiguity inside the message and not after it. The secret key used is "MATTER CHANCE".

6.4 Experiment Steps

Experiment Preprocessing:

1- Loading the table of the 64 amino acids with their DNA encodings and number of ambiguous encodings.
2- Formatting the secret key by removing spaces, repeated characters and non English letters.
3- Formatting the plaintext by removing spaces between words and separating the repeated doubles by the character '~' which is a rarely used character.

Processing:

This includes:

1- Characters are converted into binary form.
2- The binary form is converted into DNA form.
3- DNA form is converted into amino acid acid form and their ambiguity is recorded.
4- Four-square Encryption is performed.
5- Amino acid form of cipher text is converted into DNA form and the ambiguity in DNA format is also embedded to it.

6.5 Experiment Results and Comparison

The table V and VI given on next page shows the experiment results performed on Existing and Proposed Frameworks. They give the time taken to encrypt each piece of plaintext. The time is shown in milliseconds. The time taken by loading the amino acids table and preparing the secret key is ignored as it comparatively small to processing time.

Table V: Experiment Results obtained from DNA and Amino-acids implementation of PlayFair Cipher

Input size of plaintext in KB)	Plaintext after preprocessing	Processing plaintext	From Binary to Amino acids form	Playfair	Prepare ciphertext	Total processing time	Bytes/Second
1	846B	0	0	0	15.625	15.625	65.408
10	8124B	62.500	15.625	0	125.000	203.125	48.034
20	16599B	203.125	15.625	0	171.875	390.625	51.259
50	41781B	1062.500	46.875	15.625	437.500	1562.500	32.276
100	83910B	4687.500	78.125	31.25	859.375	5656.250	17.162
150	127098B	11390.625	140.625	31.25	1343.750	12906.25	11.887

Table VI: Experiment Results obtained from DNA and Amino-acids implementation of Four-Square Ciphe

Input size of plaintext in KB)	Plaintext after preprocessing	Processing plaintext	From Binary to Amino acids form	Foursquare	Prepare ciphertext	Total processing time	Bytes/Second
1	846B	0	0	0	10.75	10.75	70.135
10	8124B	62.500	15.625	0	95.000	173.125	55.237
20	16599B	203.125	15.625	0	70.154	288.904	61.167
50	41781B	1062.500	46.875	8.625	369.234	1487.234	40.866
100	83910B	4687.500	78.125	15.75	575.736	5357.111	22.370
150	127098B	11390.625	140.625	15.75	743.134	12290.134	15.836

VII. ADDITIONAL SECURITY FEATURES

Some of the given features can enhance the security and strength of the DNA b ased Four-Square Algorithm.

1. The Key: Strength of the Key is directly proportional to the strength of the algorithm. Therefore, to increase the security and strength of algorithm, the key should be more complex.
2. Use Amino acids alphabet sequence instead of English alphabet sequence: After adding the secret key in all four 5X5 matrices, the remaining spaces can be filled by standard table of Amino acids which has a special sequence defined in the matrix [4X4] (UCAG) X (UCAG).

3. The total resulting message can be combined with DNA that can be inserted into a Microdot (Steganography): This algorithm allows the cipher text to be written in any form. It can be written in DNA form, character form, or even in Amino acid form. This idea leads to more confudion in whole process. Further the DNA form can be embedded into several steganographic techniques.
4. Ambiguity provides the feature of confusion: Ambiguity bits can be used in different ways to add the confusion in the algorithm.

VIII. CONCLUSION AND FUTURE WORK

Our future work is dedicated to implementing this encoding on other known algorithms and measuring its performance and security, Also, Experiments should be conducted to implement the algorithm on different applications to ensure its feasibility and applicability.

The project is very versatile as many amendments will be possible at any time of computing because of the support for a number of intermediary processes during encryption. The project has good current market value as it is important from the view of research and has an extremely bright future as regards the importance of DNA Cryptography in an era where the DES and MD5 have been broken.

Here it is fascinating to note that several extensions of the research work carried out in this project are possible.

REFERENCES:

[1]. David K. *The Codebreakers – The Story of Secret Writing*. 1967 New York: Macmillan.
[2]. Crick F, Central dogma of molecular biology. *Nature*. 1970; 227: 561–563.
[3]. Whitfield D and EH Martin. *Multiuser cryptographic techniques*, in *Proceedings of the June 7-10, 1976, national computer conference and exposition*. 1976, ACM: New York, New York.
[4]. Department of the Army. *Basic Cryptanalysis, FM 34-40-2, FIELD MANUAL*, 1990: Washington.
[5]. Leonard Adleman. "Molecular Computation of Solutions to Combinatorial Problems". Science, 266:1021-1024, November 1994.
[6]. Dan Boneh, Cristopher Dunworth, and Richard Lipton. "Breaking DES Using a Molecular Computer". Technical Report CS-TR-489-95, Department of Computer Science, Princeton University, USA, 1995.
[7]. L. Kari, "DNA Computing: Arrival of Biological Mathematics," The Mathematical Tntelligencer, vol. 19, pp. 9–22, 1997.
[8]. TAYLOR Clelland Catherine, Viviana Risca, Carter Bancroft, 1999, "Hiding Messages in DNA Microdots". Nature Magazine Vol.. 399, June 10, 1999.
[9]. W. Stallings, (1999), *"Cryptography and Network Security: Principles and Practice"*, Prentice- Hall, New Jersey, 2da.Edición.
[10]. Ashish Gehani, Thomas LaBean and John Reif. DNA-Based Cryptography. DIMACS DNA Based Computers V, American Mathematical Society, 2000.
[11]. William Stallings."Cryptography and Network Security", Third Edition, Prentice Hall International, 2003.
[12]. G.Z. Xiao, M.Q. Lu, and L. Qin, New field of cryptography: DNA cryptography, *Chinese Science Bulletin*, (2006), 51(10), pp. 1139-1144.
[13]. Amin ST, M Saeb and S El-gindi. A DNA-based implementation of YAEA encryption algorithm. *Computational Intelligence*. 2006: 120-125.
[14]. G.Z. Cui, Y.L. Liu, and X.C. Zhang, New Direction of Data Storage: DNA Molecular Storage Technology, *Computer Engineering and Applications*, (2006), 42(26), pp. 29-32.
[15]. Mona Sabry et al.,"A DNA and Amino Acids-Based Implementation of Playfair Cipher", International Journal of Computer Science and Information Security (IJCSIS), Vol. 8 No. 3, 2010.
[16]. Sastry VUK, NR Shankar and SB Durga. A Generalized Playfair Cipher involving Intertwining, Interweaving and Iteration. *International Journal of Network and Mobile Technologies*. 2010; 1(2): 45-53.
[17]. Souhila Sadeg, Mohamed Gougache, Nabil Mansouri and Habiba Drias, "An encryption algorithm inspired from DNA", IEEE. 2010.
[18]. Srivastava SS, N Gupta and R Jaiswal. Modified Version of Playfair Cipher by using 8x8 Matrix and Random Number Generation. in *IEEE 3rd International Conference on Computer Modeling and Simulatio*. 2011. Mumbai.
[19]. Mona Sabry et al.," Three Reversible Data Encoding Algorithms based on DNA and Amino Acids' Structure", International Journal of Computer Applications (0975 – 8887), Vol. 54– No.8, September 2012.
[20]. Amal Khalifa and Ahmed Atito,"High-Capacity DNA-based Steganography", The 8th International Conference on INFOrmatics and Systems (INFOS2012) - 14-16 May. Bio-inspired Optimization Algonthms and Their Applications Track.
[21]. Meetu Skariya et al.," Enhanced Double Layer Security using RSA over DNA based

Data Encryption System", International Journal of Computer Science & Engineering Technology (IJCSET), Vol. 4 No. 06 Jun 2013.

[22]. Safwat Hamad, "A Novel Implementation of an Extended 8x8 Playfair Cipher Using Interweaving on DNA-encoded Data", International Journal of Electrical and Computer Engineering (IJECE), Vol. 4, No. 1, February 2014.

[23]. Fatma E.Ibrahim et al.,"A Symmetric Encryption Algorithm Based on DNA Computing", International Journal of Computer Applications(0975-8887), Vol. 97 No. 16, July 2014.

[24]. Diffie, W., and Hellman, M. *"Newdirections in cryptography"* IEEE Trans