International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 2, February - 2014

Data Security using DNA & Amino Acids with Four Square Cipher Decryption

Anu Rathi[1], Parmanand Astya[2]
Computer Science and Engg. Department, Mahamaya University
MIET, Meerut India

***Abstract*— In this paper, we introduce methods of decoding inspired from the DNA structure and its relation to the amino acids and also explain techniques to convert data from DNA form to binary form. This paper discusses a significant modification in techniques to enhance security by introducing DNA-based and amino acids-based structure to the core of the reverse ciphering process. Using biology in cryptography is a new approach in cryptographic research. The relevance of information security in the modern days has increased manifold as online threats are affecting millions of users. The traditional methods of cryptography are now defenseless to attacks. The idea of DNA based Cryptography has been identified as a feasible and effective methodology to create nonintrusive algorithms. In encryption procedure , all steps were implemented using DNA & Amino acids . But there are all reverse procedures are used in this paper for obtaining plaintext text .Nucleotides or codons pass through a four square cipher process based on DNA.**

Keywords*— DNA, amino acids, encryption, decryption, cryptography, security, four square ciphers*

I. INTRODUCTION

Bioinformatics is a field of science in which biology, computer science and information technology all merge into a single discipline. The ultimate goal of this science is to enable the discovery of new knowledge in biology and to create a global perspective from which unifying principles in biology can be discerned. There are three major re-search directions in bioinformatics: the development of new algorithms[5] and statistics with which they could be extracted from a large number of data elements which have common features, the analysis and interpretation of data on different types of nucleotides and amino acid sequences, protein structure, and to develop and implement tools to enable efficient access and manipulation of various types of information. A rapidly-developing technology is DNA computation or, more generally, bio molecular computation. It has emerged as a viable sub-discipline of science and engineering.

Leonard M. Adleman [3] found that the bio-computational capability of DNA can be used to solve highly complex mathematical problems. He was also able to conclude that chemistry can be used to solve un-solvable problems with the help of dedicated computers. The Hamiltonian Path Problem was solved by him in which the molecules are encoded in a sequence and bio-chemical operations are used for computations. To solve the computational problems, the data is encoded in DNA strands and molecular biological tools are used to perform operations.

In our work, we applied the conversion of character or binary form of data to the DNA form[7] and then to amino acid form. The importance of such transformation lies mainly in representing data in a biological form that can make data be able to go through biological experiments and processes, especially related to Amino Acids and DNA. It is also a way of viewing data moving through biological processes and representing it in a binary form which can be used in many computer applications. In the field of cryptography[2], the encoding techniques cannot provide security by their own as they don't include the use of a secret key. But they can be embedded into another encryption algorithm to enhance confusion and therefore enhance security. This concept is suitable for applying data integrity, digital signature and confidentiality.

II. BIOLOGICAL BACKGROUND

DNA is a very large molecule made up of a long chain of sub-units. The sub-units are called nucleotides. Each nucleotide is made up of a sugar called deoxyribose a phosphate group -PO_4 and an organic base. The sequence of bases in DNA forms the Genetic Code. A group of three bases (a triplet) controls the production of a particular amino acid in cell.The different amino acids and the order in which they are joined up determines the sort of protein being produced.

Deoxyribonucleic acid (DNA) is a nucleic acid that contains the genetic instructions used in the development and functioning of all known living organisms and some viruses. DNA is stored as a code made up of four chemical bases[10]: adenine (A), guanine (G), cytosine (C), and thymine (T).

The order, or sequence, of these bases determines the information available for building and maintaining an organism, similar to the way in which letters of the alphabet appear in a certain order to form words and sentences. The German biochemist Frederich Miescher observed DNA[8] in the late 1800s. But nearly a century passed central importance to biology.DNA binary strands support feasibility and applicability of DNA-based Cryptography[4]. The security and the performance of the DNA based

cryptographic algorithms are from that discovery until researchers unraveled the structure of the DNA molecule and realized its satisfactory for multi-level security applications of today's network. Certain DNA algorithms can resist exhaustive attack, statistical attack and differential attack.

The field of DNA computing[16] is still in its infancy and the applications for this technology have not yet been fully understood. 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[15] is the beneficial supplement to the existing mathematical cipher. If the molecular word can be controlled at will, it may be possible to achieve vastly better performance for information storage and security.

III. DNA & AMINO ACIDS -BASED FOUR SQUARE CIPHER ALGORITHM

A. The encoding to amino acids

Any form of data can be represented in a binary form (message, image or signal). This form can be transferred to DNA form according to Table 1 .

Table 1. DNA bits representation

Bit 1	Bit 2	DNA
0	0	A
0	1	C
1	0	G
1	1	T or U

Note that each amino acid has a name, abbreviation (3-letter form), and a single character symbol (1-letter form). This character symbol is what we will use in our algorithm. The DNA form is transferred to the Amino acids form according to Table 2 which is a standard universal table of Amino acids[9] and their cordons representation in the form of DNA.

Figure1 show the flowchart which represents the logic sequence of DNA based four square cipher algorithms.

Table 2: Amino acids and their 64 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

B. Constructing the English alphabet table:

English alphabets are construct from their amino acids . There are 64 codons[15] are exists in table 2. we have only 20 amino acids in addition to 1 start and 1 stop codons. Each amino acid is abbreviated with one unique character (English letter). In order to construct a complete set of alphabetical English letters, we need 26 letters with their transformation encoding to DNA. The letters we need to fill are (B, J, O, U, X, Z). So we will make these characters share some amino acids their codons. The three stop codons have 2 of one family type (UAA,UAG) to be assigned to letter B and one of other type (UGA) to be assigned to letter J. We have 3 amino acids (L, R, S) having 6 codons. The table illustrates letters from A to Z with the associated amino acids as explained before Counting the number of codons of each letter, we will find the number varies between 1 and 4 codons per letter. We will call this number 'Ambiguity' of the character.

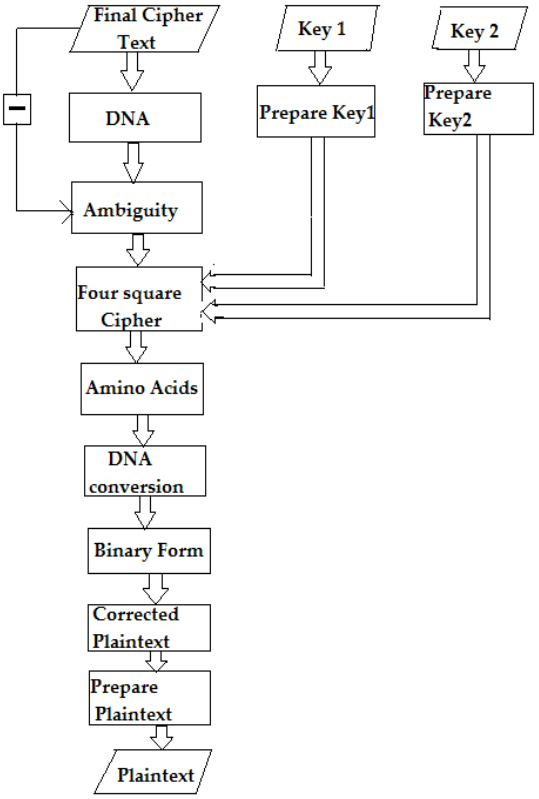

Fig1: Flow Chart of the DNA & Amino acids based four square cipher decryption algorithm

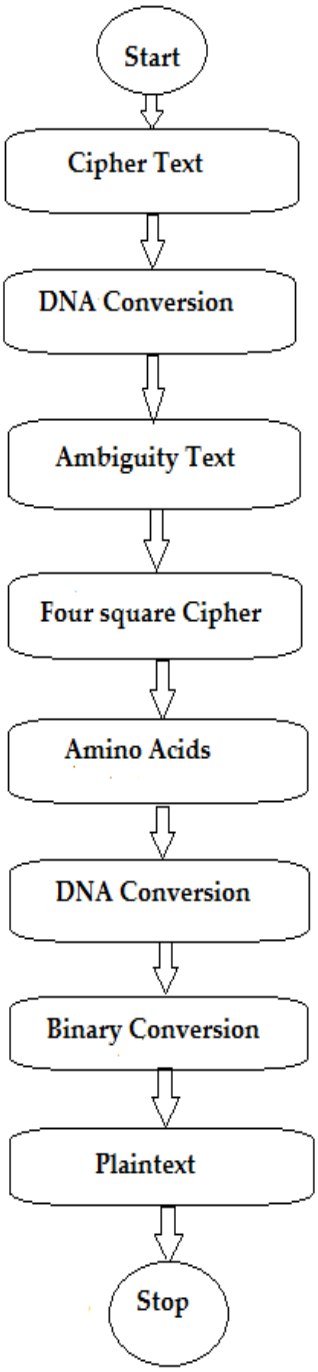

Fig2:Activity diagram of DNA & Amino acids based four square cipher Decryption algorithm

Table 3. The final distribution of Amino upon English letters with ambiguity

Letter	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
Ambiguity	4	2	2	2	2	2	4	2	3	1	2	4	1	2	2	4	2	4	4	4	2	4	1	2	1	1
A	GCU	UAA	UGU	GAU	GAA	UUU	GGU	CAU	AUU	UGA	AAA	CUU	AUG	AAU	UUA	CCU	CAA	CGU	UCU	ACU	AGA	GUU	UGG	AGU	UAU	UAC
C	GCC	UAG	UGC	GAC	GAG	UUC	GGC	CAC	AUC		AAG	CUC		AAC	UUG	CCC	CAG	CGC	UCC	ACC	AGG	GUC		AGC		
G	GCA						GGA		AUA			CUA				CCA		CGA	UCA	ACA		GUA				
U	GCG						GGG					CUG				CCG		CGG	UCG	ACG		GUG				
Family	GC	UA	UG	GA	GA	UU	GG	CA	AU	UG	AA	CU	AU	AA	UU	CC	CA	CG	UC	AC	AG	GU	UG	AG	UA	UA

IV. DNA & AMINO ACIDS -BASED FOUR SQUARE DECRYPTION ALGORITHM

Preprocessing:

1- Prepare the secret keys for decryption :

- Remove any spaces or repeated characters from [K1].
- Remove any spaces or repeated characters from [K2].
- Put the remaining characters in the UPPER case form.

[K1]→UPPER[K1].
Put the remaining characters in the UPPER case form.
[K2]→UPPER[K2].

2-Input :

DNA form of Ciphertext[DC] and Ambiguity [AMBIG] together in the suitable form→final cipher text [C].

Processing:

1- DNA [AC]=DNA form of cipher text [DC]
2- foursquare cipher [AP]= Amino acid of cipher text [AC]
3- Construct the four square cipher 5X5 matrix and add [K1] row by row, then add the rest of alphabet characters .
4- Construct the four square cipher 5X5 matrix and add [K2] row by row, then add the rest of alphabet characters.
5- AMINO [DP] (Replace each Amino acid character by 3 DNA characters ,keeping in track the ambiguity of each Amino acid [AMBIG]= Amino acids form [AP]
6- DNA [BP] (Replace DNA base representation by 2 bits) = DNA form [DP]
7- BINARY [P] (Replace binary representation-8 bits by corresponding each character) = Binary form [BP].

3- Prepare the original plaintext:

1-Select Corrected Plaintext
2-If Corrected Plaintext count = = multiple of 3(after counting spaces)
3-then Remove '#', go to 2
4-else print 'plaintext'.
5-remove '~'(which is used for represent spaces)
5-exit

V. SAMPLES OF THE PROGRAM STEPS AND OUTPUT

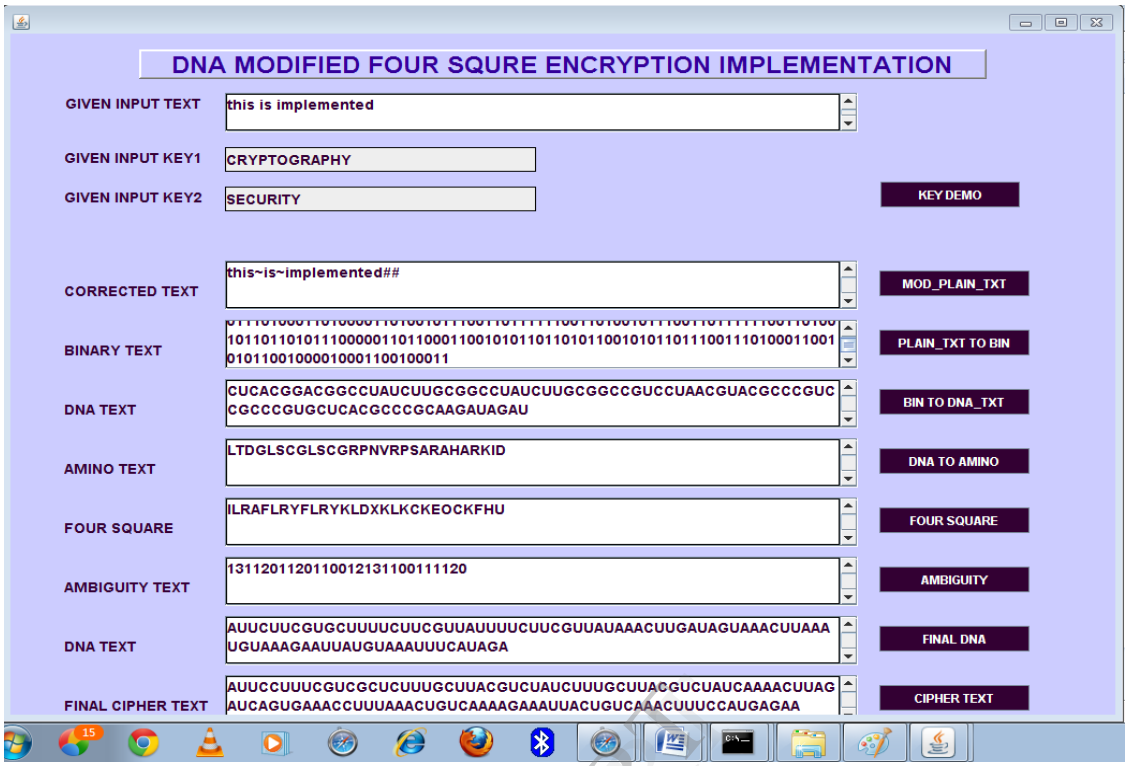

Fig 3: Encryption implemented[1]

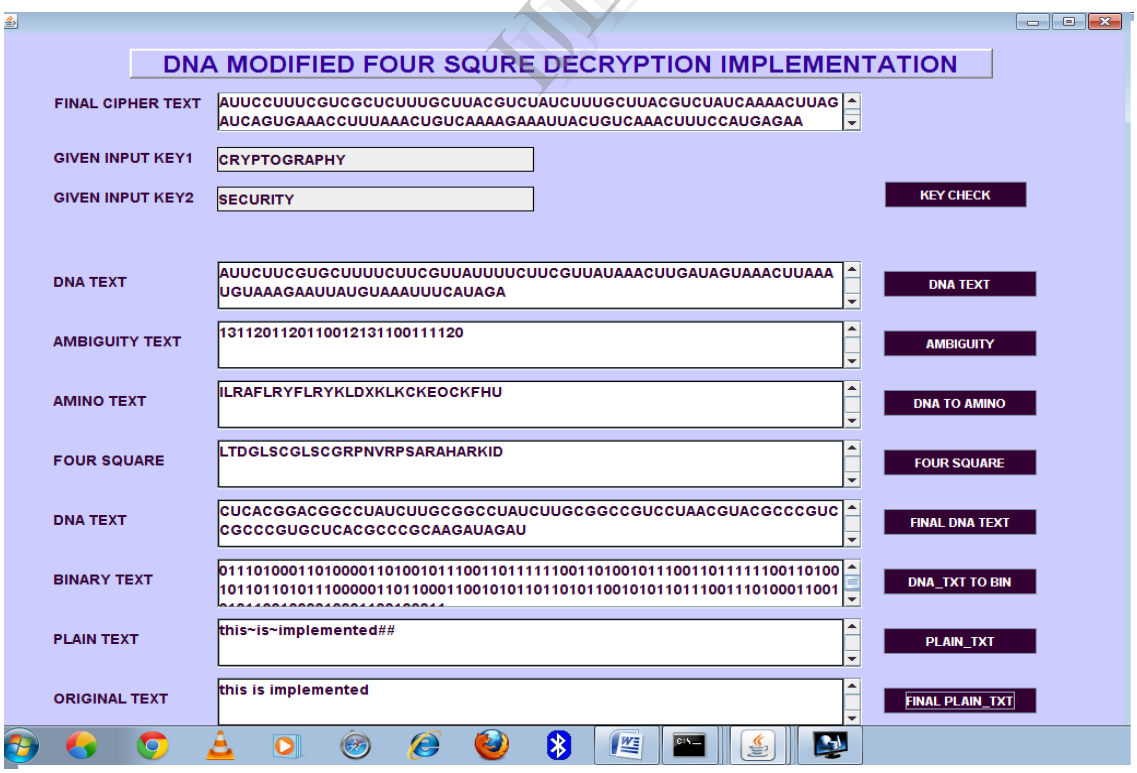

Fig 4: Decryption Process of Implemented Encryption Process

VI. ANALYSIS AND DISCUSSIONS

Data decoding to the form from Amino acids to DNA , or we can say our semi artificially built distribution of amino acids. "Artificial" because of characters we added and switching's of codons we introduced to the genetic code table and "Semi" because we kept most of the standards in the genetic code table. We have proved that the decoding algorithm is applicable on implemented encryption. The three algorithms can be implemented with one or many rounds. The idea of representing the amino acid form of data in English characters makes this form to be used as input to additional cycles for providing more security in both side.. This was implemented by calculation of the hexadecimal of each letter. Then we was converted it to the binary then DNA forms which act as input to a new round. The importance of such transformation lies mainly in representing data in a biological form that can make data be able to go through biological experiments and processes in encryption[1], especially related to Amino Acids and DNA. Above all transformation are reversed in this implemented work.

It is also a way of viewing data moving through biological processes and for getting plaintext ,the representing it in character from binary form which can be used in many computer encoding algorithms.

A. Experiment steps:

a. 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 K1 by removing spaces, repeated characters and non English letters.

3- Formatting the secret key K2 by removing spaces, repeated characters and non English letters

4- Formatting the plaintext by removing spaces between words and separating the repeated doubles by the character '~' which chosen to be a rarely used character.

b. Processing:

This includes:

1-Convert DNA form to Amino acids form after substraction of embedded ambiguity from final cipher text .

2-Do four square cipher Decryption .

3- Converting amino acids and recording ambiguity to DNA.

4- Convert DNA to Binary .

5- Convert 8 bit ASCII value to corresponding characters.

6-Find Corrected plaintext .

B. Experiment Results:

The next figure illustrates the experiments and time taken to decrypt each piece of ciphertext (each is of different data loads) in months. It shown in fig 5.

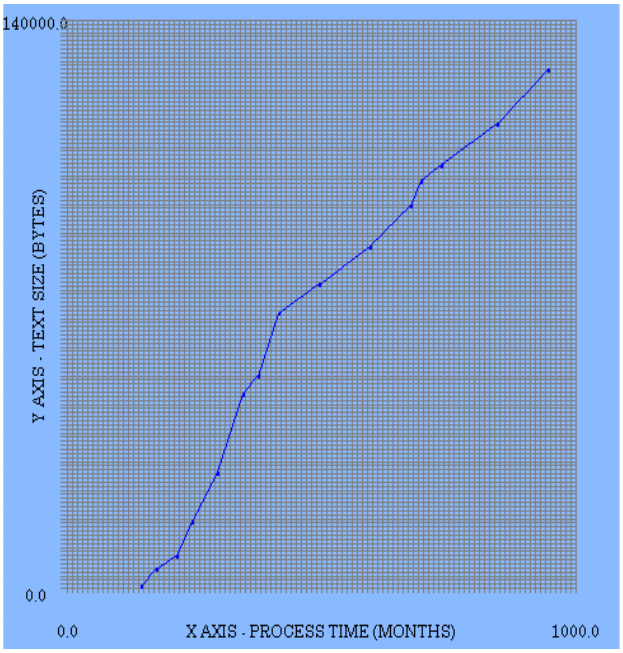

Fig 5: Reverse Complexity Graph between text size and process time in months

Process Time(Months)	Text Size(Bytes)
150	846
180	5062
220	8124
250	16599
300	28599
350	47781
380	52281
420	67543
500	74554
600	83910
680	93930
700	99940
740	103910
850	113910
950	127098

Fig6:Process Time v/s Text size for decryption

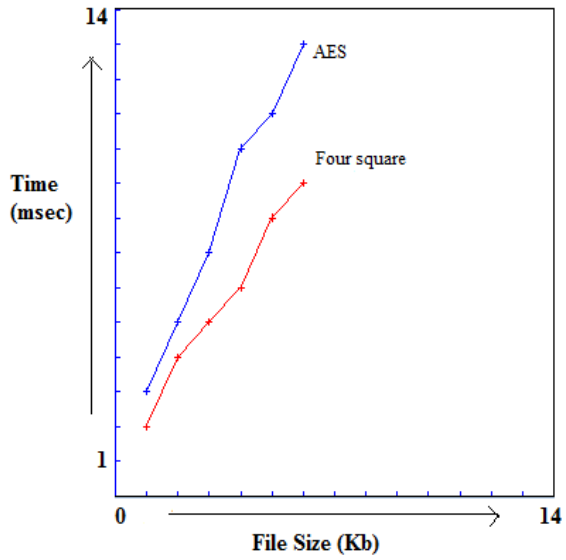

Figure 7: Comparison graph – blue line for AES & red line for Four Square

As shown in fig.7, AES provides high security in the field of security/cryptography.This implementation work is compared with AES & it provide little similar security as AES. On the basis of AES ,we can say that it provide good security for data which is used for communication .

Reverse complexity of this decryption algorithm is very high . The time taken by loading the amino acids table and preparing the secret keys is ignored because it is comparatively small to processing time. Now it is shown in fig.6 that this four square cipher reverse algorithm take large amount of time considered in months . Less bytes of data is decrypted in months by this implemented algorithm. It is shown in fig6.

VII. CONCLUSION AND FUTURE WORK

Decryption is the reverse procedure of encryption .We already has been implemented the methods of data encoding to the form of Amino acids & DNA . Encryption was implemented with one or many rounds. Encryption & Decryption of any data/information is suitable for applying data integrity, digital signature and confidentiality. As this decryption algorithm reverse of the four matrix are used to decrypt the data using both the secret keys . This is implemented by combining the traditional or biological cryptographic algorithm to create new security systems in encryption & decryption. Our future work can be done with more & more modification in this work using other algorithms i.e AES, DES, MD5,SHA1 etc. Also, Experiments should be conducted to implement the algorithm on different applications to ensure its feasibility and applicability.

REFERENCES

[1] Anu Rathi ,Parmanand Astya,Ankur Garg ," Data Security using DNA & Amino acids with Four Square Cipher Encryption", International Journal of Mobile & Adhoc Network , vol 4 ,issue 1 .Feb 2014.

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

[3] M. X. Lu, "Symmetric key cryptosystem with DNA technology, "Science in China Series F: Information Sciences, vol. 3, pp. 324–333, 2007.

[4] W.Stallings, (2005) "Cryptography and Network Security 4th Ed," Prentice Hall ,pp. 58-309

[5] W. Stallings, (1999), *"Cryptography and Network Security: Principles and Practice"*, Prentice- Hall, New Jersey, 2da.Edición.

[6] L. Kari, "DNA Computing: Arrival of Biological Mathematics," The Mathematical Tntelligencer, vol. 19, pp. 9–22, 1997.

[7] C. T. Celland, V. Risca and Bancroft C. "Hiding messages in DNA microdots," Nature, vol. 399, pp. 533–534, 1999.

[8] Mona Sabry, Mohamed Hashem, Taymoor Nazmy, Mohamed Essam Khalifa, "A DNA and Amino Acids Based Implementation of Playfair Cipher",(IJCSIS) International Journal of Computer Science and Information Security,Vol. 8, No. 3, 2010.

[9] William Stallings. "Cryptography and Network Security", Third Edition, Prentice Hall International, 2003.

[10] O. Tornea and M.E. Borda," DNA Cryptographic Algorithms", MEDITECH 2009, IFMBE Proceedings 26, pp. 223–226, 2009.

[11] A. Leier, C. Richter and W. Banzhaf, "Cryptography with DNA binary strands," Biosystems, vol. 57, pp. 13–22, 2000.

[12] A. Gehani, T. H. LaBean and J. H. Reif, "DNA based cryptography,"DNA Based Computers Providence: American Mathematical society, vol. 54, pp. 233–249, 2000.

[13] KANG Ning, "A Pseudo DNA Cryptography Method", Independent Research Study Project for CS5231, October 2004.

[14] S.V. Kartalopoulos, "DNA inspired cryptographic method in optical communications," in authentication and data mimicking Military Communications Conference, 2005, pp. 774–779.

[15] G. Z. Cui, L. M. Qin, Y. F Wang and X. C. Zhang, "Information Security Technology Based on DNA Computing," 2007 IEEE International Workshop on Anti-counterfeiting Security, Identification., 2007, pp. 288 291.

[16] Sherif T. Amin, Magdy Saeb, Salah El8Gindi, "A DNA8 based Implementation of YAEA Encryption Algorithm," IASTED International Conference on Computational Intelligence (CI 2006), San Francisco, Nov. 20, 2006.