International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2763
Issue 11, Volume 3 (November 2016) www.ijirae.com

DNA Encryption Algorithms: Scope and Challenges in Symmetric Key Cryptography

Anurag Roy
Department of Computer Science
St Xavier's College, Kolkata

Asoke Nath
Department of Computer Science
St Xavier's College, Kolkata

Abstract *- **Data security is now a crucial issue now in our day to day life. The protection of personal identity, personal finances depend on the protection of important and irreplaceable information. Cryptography is the science of converting some readable information into something unreadable format, which are hard to decipher. In modern times, cryptography has adopted a new medium: human DNA. At a time when conventional cryptography has been losing strength to more advanced cryptanalysis, DNA cryptography has added more elements of confusion and diffusion. The use of DNA sequences to encrypt data has strengthened the existing classical encryption algorithms. Thus, DNA cryptography has added another dimension to conventional cryptography. In the present paper the authors have made a systematics study on DNA encryption algorithms and how it can be used along with standard classical encryption algorithms.***

Keywords- DNA cryptography, substitution ciphers, block ciphers, symmetric key encryption, Playfair cipher, decryption

I. INTRODUCTION

Cryptography is concerned with encryption, which is a process of converting ordinary, comprehensible messages (plaintext) into unintelligible text (ciphertext). Decryption is the inverse of this process, as it converts the ciphertext back to the original plaintext, given that the proper algorithm to invert the cipher is known to the party attempting the decryption. Another concept in the subject of "Cryptography" is cryptanalysis. This essentially defines a study of the encryption algorithm to find weaknesses or loopholes in the cipher, with an aim to extract the plaintext from the ciphertext without necessarily knowing the 'key' or the decryption algorithm.

Some of the earliest ciphers are called substitution ciphers: letters in the plaintext were replaced by other letters after undergoing a shift in the alphabet. These ciphers were easy to break, given the knowledge of the frequency of letters or the combination of letters in a language. Eventually, stronger ciphers were created which worked on either the whole of stream of plaintext at once or blocks of plaintext at a time.

Some of the modern cryptography encryption algorithms, such as the Feistel cipher, DES (influenced from the Feistel cipher), MD5, etc, are already broken. Consequently, new directions of information security are being sought to protect data. It was Adleman [1] who first demonstrated the use of DNA computing in solving real-world computation problems. His main idea was to use actual chemistry to solve intractable problems, such as the seven-point Hamiltonian path problem. This opened the door to the concept of using DNA in the field of cryptography to generate powerful, even unbreakable ciphers.

Boneh, Dunworth and Lipton [2] have also claimed that by using DNA computing the Data Encryption Standard (DES) can also be broken. This novel area in computing science has given rise to DNA steganography: the study of camouflaging a DNA encoded message within the enormous complexity of human genomic DNA and then compressing this information to a microdot [3].

Later advancements in DNA cryptography were made by Gehani, LaBean and Reif [4] with their formulation of a one-time pad with DNA strands. Kang Ning [5] used another approach wherein he did not use real DNA computing; he applied the principal ideas of molecular biology to develop an encryption method.

The advantages of DNA cryptography are the following [6]:

1. *A large degree of parallelism is supported in DNA computing, which helps in increasing computational speed.*
2. *The molecules of DNA are media with a large capacity of transmission.*
3. *Power consumption is low.*

The remainder of the paper is arranged in the following sequence: background study of DNA cryptography and cryptography, in general; a study of symmetric key cryptography or SKC (because the scope of this paper focuses primarily on SKC) and a detailed literature review.

II. BACKGROUND

A. DNA (De-oxyribo Nucleic Acid):

De-oxyribo Nucleic Acid or DNA is the genetic material in humans and almost every other organism. The same DNA is contained in nearly every cell in a person's body. The information in DNA is stored in the form of four chemical bases: adenine (A), guanine (G), cytosine (C) and thymine (T). Human DNA consists of about 3 billion bases. The order in which these bases are arranged determine the information available to build and maintain an organism. This concept is also used in many encryption algorithms which emulate the DNA structure. [7]

DNA bases pair up with each other. A is conjoined with T, and C with G. They form base pairs. Each base is also attached to a sugar molecule and a phosphate molecule. Together, this combination is called a nucleotide. Nucleotides are arranged in two long strands in a double helical structure. This construction is somewhat like a ladder, where the base pairs for the ladder's rungs.

B. Symmetric Key Cryptography:

Symmetric Key Cryptography (SKC) consists of algorithms for cryptography that use the same key for both the encryption and decryption of data. The keys meant for the two functions can be identical or the key might go through a simple transformation. Symmetric key algorithms require the sender and receiver to maintain absolute secrecy of the key. Only those two parties are required to share the key with the help of a private link. This introduces drawbacks to symmetric algorithms, when compared to public-key encryption (also known as asymmetric key encryption) [8].

Symmetric key ciphers can be of two types:

- ***Stream ciphers*** *encrypt all the bytes of a plaintext message at a time.*
- ***Block ciphers*** *encrypt a number of bytes of the plaintext message at a time. Padding (increasing or decreasing the number of redundant characters) is done to make the total length of the message a multiple of the length of one block of messages.*

C. DNA Cryptography:

DNA cryptography can be thought of as hiding data in terms of DNA sequence. It is an emerging technology, which works on concepts of DNA computing [9]. DNA can be used to store and transmit data. This new concept is a hope in the direction of formulating new unbreakable algorithms.

DNA strands consist of millions of polymers of linked nucleotides. These nucleotides consist of one of four nitrogen bases (*Adenine, Guanine, Cytosine and Thymine*), a pentose sugar and a phosphate. The growing popularity of DNA cryptography can be attributed to the advantages of DNA computing over traditional computing:

- ***Speed:*** *Conventional computers have been known to perform approximately 100 million instructions per second (MIPS). Combining DNA strands has been predicted to make computations equivalent to 10^9 or better, arguably 100 times faster than the fastest computer.*
- ***Storage:*** *DNA stores memory at the rate of 1 bit per cubic nanometre, whereas conventional storage media requires 10^{12} cubic nanometre to store 1 bit.*
- ***Power Requirements:*** *DNA computing does not require power while computation is taking place. The chemical reactions that create the building blocks of DNA take place without any external power source.*

III. LITERATURE REVIEW AND DISCUSSION OF SOME RECENT DNA-BASED ENCRYPTION ALGORITHMS

Research done on DNA cryptography is extensive, though the subject is novel and a lot of growth is predicted within the scope of this emerging technology in the coming decades. As stated in an earlier section in this paper, substitution ciphers brought about the beginning of cryptography. These were the first symmetric key ciphers. However, these classical ciphers were broken over time. The advent of DNA cryptography has strengthened such obsolete encryption techniques and paved the way for further innovation. In this section, I propose to review the literature which concerns conventional techniques of encrypting data.

Given below are examples of previous research done within this scope, with a detailed discussion of novel encryption algorithms – inspired by DNA modelling - that have been introduced:

A. SECURE DATA TRANSFER THROUGH DNA CRYPTOGRAPHY USING SYMMETRIC ALGORITHM:

The paper with the above title [6] was written by Bonny B Raj, J Frank Vijay and T Mahalakshmi. The authors have proposed a simple and basic substitution cipher which is inspired by the structure and sequence of DNA. To understand this encryption technique, we have to first substitute the nucleotides for binary sequences. This paper has adopted the following values:
A – 00, C – 01, G – 10, and T – 11

After the above information is available, the following steps are performed for encryption of data:

Step 1. *The plaintext is first converted to their respective ASCII numbers (in decimal format).*
Step 2. *This stream of ASCII numbers are then grouped into blocks.*
Step 3. *The ASCII numbers are converted to binary numbers (0's and 1's).*
Step 4. *Encrypted sequences of binary numbers are broken into pairs. The possible pairs could be 00, 01, 10 and 11. The pairs are then substituted by the nucleotides, such as A for 00, T for 01, G for 10 and C for 11.*
Step 5. *This step involves the generation of a random key P_k. Now, P_k has to be a number between 1 and 256, inclusive. This random key determines the permutation of the four characters A, T, G and C. For example, when P_k is 1, the table given below is used for the conversion of ASCII code to nucleotide sequences:*

1	AAAA	33	CAAA	65	GAAA	97	TAAA	129	AGAA	161	CGAA	193	GGAA	225	TGAA
2	AAAC	34	CAAC	66	GAAC	98	TAAC	130	AGAC	162	CGAC	194	GGAC	226	TGAC
3	AAAG	35	CAAG	67	GAAG	99	TAAG	131	AGAG	163	CGAG	195	GGAG	227	TGAG
4	AAAT	36	CAAT	68	GAAT	100	TAAT	132	AGAT	164	CGAT	196	GGAT	228	TGAT
5	AACA	37	CACA	69	GACA	101	TACA	133	AGCA	165	CGCA	197	GGCA	229	TGCA
6	AACC	38	CACC	70	GACC	102	TACC	134	AGCC	166	CGCC	198	GGCC	230	TGCC
7	AACG	39	CACG	71	GACG	103	TACG	135	AGCG	167	CGCG	199	GGCG	231	TGCG
8	AACT	40	CACT	72	GACT	104	TACT	136	AGCT	168	CGCT	200	GGCT	232	TGCT
9	AAGA	41	CAGA	73	GAGA	105	TAGA	137	AGGA	169	CGGA	201	GGGA	233	TGGA
10	AAGC	42	CAGC	74	GAGC	106	TAGC	138	AGGC	170	CGGC	202	GGGC	234	TGGC
11	AAGG	43	CAGG	75	GAGG	107	TAGG	139	AGGG	171	CGGG	203	GGGG	235	TGGG
12	AAGT	44	CAGT	76	GAGT	108	TAGT	140	AGGT	172	CGGT	204	GGGT	236	TGGT
13	AATA	45	CATA	77	GATA	109	TATA	141	AGTA	173	CGTA	205	GGTA	237	TGTA
14	AATC	46	CATC	78	GATC	110	TATC	142	AGTC	174	CGTC	206	GGTC	238	TGTC
15	AATG	47	CATG	79	GATG	111	TATG	143	AGTG	175	CGTG	207	GGTG	239	TGTG
16	AATT	48	CATT	80	GATT	112	TATT	144	AGTT	176	CGTT	208	GGTT	240	TGTT
17	ACAA	49	CCAA	81	GCAA	113	TCAA	145	ATAA	177	CTAA	209	GTAA	241	TTAA
18	ACAC	50	CCAC	82	GCAC	114	TCAC	146	ATAC	178	CTAC	210	GTAC	242	TTAC
19	ACAG	51	CCAG	83	GCAG	115	TCAG	147	ATAG	179	CTAG	211	GTAG	243	TTAG
20	ACAT	52	CCAT	84	GCAT	116	TCAT	148	ATAT	180	CTAT	212	GTAT	244	TTAT
21	ACCA	53	CCCA	85	GCCA	117	TCCA	149	ATCA	181	CTCA	213	GTCA	245	TTCA
22	ACCC	54	CCCC	86	GCCC	118	TCCC	150	ATCC	182	CTCC	214	GTCC	246	TTCC
23	ACCG	55	CCCG	87	GCCG	119	TCCG	151	ATCG	183	CTCG	215	GTCG	247	TTCG
24	ACCT	56	CCCT	88	GCCT	120	TCCT	152	ATCT	184	CTCT	216	GTCT	248	TTCT
25	ACGA	57	CCGA	89	GCGA	121	TCGA	153	ATGA	185	CTGA	217	GTGA	249	TTGA
26	ACGC	58	CCGC	90	GCGC	122	TCGC	154	ATGC	186	CTGC	218	GTGC	250	TTGC
27	ACGG	59	CCGG	91	GCGG	123	TCGG	155	ATGG	187	CTGG	219	GTGG	251	TTGG
28	ACGT	60	CCGT	92	GCGT	124	TCGT	156	ATGT	188	CTGT	220	GTGT	252	TTGT
29	ACTA	61	CCTA	93	GCTA	125	TCTA	157	ATTA	189	CTTA	221	GTTA	253	TTTA
30	ACTC	62	CCTC	94	GCTC	126	TCTC	158	ATTC	190	CTTC	222	GTTC	254	TTTC
31	ACTG	63	CCTG	95	GCTG	127	TCTG	159	ATTG	191	CTTG	223	GTTG	255	TTTG
32	ACTT	64	CCTT	96	GCTT	128	TCTT	160	ATTT	192	CTTT	224	GTTT	256	TTTT

Likewise, when P_k is 2, 'AAAA' maps to the ASCII code 2, and indicates the ASCII code 3 when P_k is 3 and so on.
Step 6. An index table is generated corresponding to the value of P_k.
Step 7. The final ciphertext is obtained from the table after converting the ASCII, using the rule in Step 4. Decryption is just the reversal of the encryption process. Discussed below is an example of encryption.
Step 8. Let us take a simple plaintext, say the word **CAT**.

Step 1. ASCII of the characters are C = 67, A = 65 and T = 84.
C (67) → 01000011 → GAAG
A (65) → 01000001 → GAAA
T (84) → 01010100 → GCAT

Step 2. After the combination of DNA substitution, the message is **GAAG – GAAA – GCAT**.

Step 3. From the random key generation, the DNA encrypted message becomes **57**, **55**, **74**.

Step 4. Hence, the final ciphertext for this particular random key is **57, 55, 74**.

B. A DNA AND AMINO ACIDS-BASED IMPLEMENTATION OF PLAYFAIR CIPHER

This paper [10] was the result of collaboration between Mona Sabry, Mohamed Hashem, Taymoor Nazmy and Mohamed Essam Khalifa from the Ain Shams University, Cairo, Egypt. It discusses the traditional Playfair Cipher and also introduces an innovative method to implement this algorithm using principles of genetics. A drawback of the Playfair Cipher is that it can only encrypt English alphabets. It cannot encrypt numerals or special characters. This novel algorithm, however, eliminates the problem.

Firstly, it is important to understand the Playfair algorithm, which can be found here [11]. The algorithm makes use of the concept of translating DNA codes to Amino Acids to increase the confusion. Sequences of nucleotides are generated as sequences of messenger RNA (mRNA). The relationship between the nucleotide sequences and amino acid sequences are determined by a genetic code, known as codons (they are three letter words, e.g. ACT, UAG, GUU). Just like in the paper discussed previously, DNA nucleotides indicate pairs of bits; **A** for **00**, **C** for **01**, **G** for **10** and **T** for **11**. Given below are the steps of encryption using this modified algorithm:

Step 1. The plaintext is converted to bits, which is then converted to sequences of nucleotides (A, C, G and T). The DNA form is then converted to the standard amino acid form using the table given below:

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

Step 2. In the table above, we only have 20 amino acids (i.e. 20 letters), in addition to 1 START and 1 STOP codon. The Playfair Cipher would require 25 characters (since I & J are considered to be one character). B, O, U, X and Z are letters that still need to be filled. The START codon is repeated with M. Hence, that will not be used. B is assigned the 3 STOP codons. Here, we notice that 3 amino acids (L. R and S) have 6 codons each. Studying the sequence of DNA in each of them, we notice that they each have 4 codons of the same type and 2 of another type. Those 2 codons from L, R and S are shifted to the letters O, U and X respectively. Letter Z takes the codon UAC from Y.

Another aspect of this step is to determine which codon to choose to indicate each letter in the alphabet. Since each letter has codons varying from 1 to 4, a number is added which indicates the DNA sequence that would substitute the character. This number is called the "Ambiguity' of the particular character [AMBIG].

Step 3. Now that the secret key and the plaintext message is converted in the form of Amino Acids, the traditional Playfair Cipher matrix can be built using the secret key. The plaintext is then passed through this matrix to obtain the ciphertext.

Hence, if the plain text is of length L, the ciphertext would be of length 3L, because each character is converted to a 3 character-long sequence of DNA.

The Ambiguity key is helpful in increasing the "confusion" aspect of the cipher. For example, with an AMBIG value of 3, P will be converted to CCA. Hence, one character can be mapped to different sequences of DNA, which makes cryptanalysis substantially harder.

C. A METHOD TO ENCRYPT INFORMATION WITH DNA-BASED CRYPTOGRAPHY

This paper was jointly authored by Mohammadreza Najaftorkaman (Griffith University Gold Coast, Australia) and Nazanin Sadat Kazazi (UTM, Malaysia) [12]. This paper, like the other two discussed previously, introduces the idea of using DNA based cryptosystem on the traditional Vigenere cipher. The discussion of the classical Vigenere Cipher can be found here [13]. The proposed algorithm follows the steps discussed below:

Step 1. **Data Preprocessing.** This is the first step. The data in binary numbers is converted yet again to a DNA string. This paper, however, proposes a unique technique to convert binary data to DNA strings. The rule is tabulated below:

Binary Data	DNA
00	AA
01	T
10	C
11	GG
0	A
1	G

Step 2. **Generation of a Key.** This process involves the generation of a key for the Vigenere substitution. In this step, the NCBI (National Center for Biotechnology Information), which is the master bank of the human genome, is used. A MATLAB software generates a very large, random stream of DNA from the NCBI database from which the key is extracted. The key (in red) can start from any position in the string and be of any length, depending on the sender's choice:
AGGTCACCGTACAGATCAGTCGTAACCAGTGACAGCATGACACGTCAGTCCAGTCGATCACGGTG
CCAAACGTGGAGACAGTCAC.....

Step 3. **Encryption Process.** In this step, an algorithm was developed to substitute the plaintext using the secret key. The Vigenere Cipher, which is a polyalphabetic cipher, follows the substitution rules of the table given below:

	A	T	C	G
A	A	T	C	G
T	T	C	G	A
C	C	G	A	T
G	G	A	T	C

Decryption can be done simply with the knowledge of the secret key. Referring to this Vigenere-DNA table, the plaintext can be obtained from the ciphertext.

IV. CONCLUSION

Data security is a major concern in the world today. With the enormous price attached to data, enterprises all over the world are putting much more effort and cost than ever to protect data. Theoretical analysis has shown that the inclusion of the biological element in computing has not only enhanced computing power, but also improved cryptosystems that were previously obsolete. Symmetric key cryptography methods which have adopted DNA technology is more powerful against attacks.

The primary inferences from this study are:

- *Binary information can be converted to DNA codes. Furthermore, binary cryptography methods can be reapplied using DNA coding technology.*
- *Preprocessing is imperative when any DNA cryptosystem is applied.*
- *DNA cryptography is stronger against attacks as it has a greater degree of intractability.*

V. REFERENCES

[1] Adleman L M, *Molecular Computation of Solutions to Combinatorial Problems*, Science, Vol 266, pp 1021-1024, November 1994.
[2] 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.
[3] TAYLOR Clelland Catherine, Viviana Risca, Carter Bancroft, 1999, "*Hiding Messages in DNA Microdots*". Nature Magazine Vol.. 399, June 10, 1999.
[4] Gehani, T. LaBean, and J. Reif, DNA-Based Cryptography, *Lecture Notes in Computer Science*, Springer, Vol 2950, pp 167-188 2004.
[5] KANG Ning, "*A Pseudo DNA Cryptography Method*", Independent Research Study Project for CS5231, October 2004.
[6] Bonny B Raj, J Frank Vijay, and T Mahalakshmi, "*Secure Data Transfer through DNA Cryptography using Symmetric Algorithm*", International Journal of Computer Applications, Volume 133- No. 2, January 2016.
[7] Watson, J. D., & Crick, F. H. C. A structure for deoxyribose nucleic acid. *Nature* **171**, 737–738 (1953).
[8] *SFWR 4C03: Computer Networks and Computer Security*, Lecturer: Kartik Krishnan, March 8-11 2004, North Carolina State University.
[9] S. T. Amin, M. Saeb, S. El-Gindi, *A DNA-based Implementation of YAEA Encryption Algorithm*, IASTED International Conference on Computational Intelligence,San Francisco, pp 120-125, 2006.
[10] Mona Sabry, Mohamed Hashem, Taymoor Nazmy, "*A DNA and Amino Acids-Based Implementation of Playfair Cipher*", International Journal of Computer Science and Information Security, Vol. 8, No. 3, 2010.
[11] Practical Cryptography, http://practicalcryptography.com/ciphers/playfair-cipher/
[12] Mohammadreza Najaftorkaman, Nazanin Sadat Kazazi, "*A Method to Encrypt Information with DNA-Based Cryptography*", International Journal of Cyber-Security and Digital Forensics (IJCSDF), The Society of Digital Information and Wireless Communications, 2015.
[13] Computer Science, Michigan Technological University: http:// www.cs.mtu.edu /~shene/NSF-4/Tutorial/VIG/Vig-Base.html

