International Journal of Electrical and Computer Engineering (IJECE)
Vol. 4, No. 1, Feburary 2014, pp. 93~100
ISSN: 2088-8708

A Novel Implementation of an Extended 8x8 Playfair Cipher Using Interweaving on DNA-encoded Data

Safwat Hamad
Faculty of Computer and Information Sciences, Ain Shams University

Article Info

Article history:

Received Oct 27, 2013
Revised Dec 28, 2013
Accepted Jan 12, 2014

Keyword:

Brute force
Cipher
Codon
Cryptanalysis
DNA
Encryption
Iinterweaving
Playfair

ABSTRACT

Recentlty, cryptography makes extensive use of different fields including bioinformatics. The fundamental idea behind the cipher presented here is to transform any kind of binary message; such as text, sound tracks, and even images, into the form of a single-stranded DNA sequence. Subsequently, digraphs of codon triplets are encrypted using a grid of 8x8 codon matrix that is randomly constructed according to some secret key. Although the encryption/decryption rules were kept almost the same as the classical 5x5 Playfair, using the DNA encoding step makes it almost impossible for an attacker to perform a frequency analysis on that vast number of character digraphs. Furthermore, an interweaving step is added to scramble the encrypted sequence offering more randomness. When compared with other modifications of the Playfair cipher, the proposed method showed a number of advantages including the ability to cipher any type of digital media, the elimination of plain-text preprocessing step, and the applicability to be integrated into larger security systems such as DNA steganography. Furthermore, due to the very weak correlation between cipher-data and original message, the proposed method shows a strong robustness against cipher attacks.

Corresponding Author:

Safwat Hamad,
Faculty of Computer and Information Sciences,
Ain Shams University,
Khalifa Almaamoon street, Abbassia 11566, Cairo, Egypt
Email: sHamad@cis.asu.edu.eg

1. INTRODUCTION

Cryptography is known as the practice and study of techniques and protocols for secure communication in the presence of third parties. As shown in figure 1, encryption is the process of converting ordinary information (called plaintext) into incomprehensible gibberish (called ciphertext) [1]. The reverse operation is decryption, in which the received unintelligible ciphertext is converted back to plaintext. In this context, the algorithms that create the encryption/decryption pair are usually called cipher. The detailed operation of a cipher is not only controlled by the algorithms, but also by a secret parameter called the key. For a specific message exchange context with complex modern cryptographic algorithms and a key known only to the communicants, it is practically hard for an adversary to break them by any known practical means.

The main classical cipher types are transposition ciphers, which rearrange the order of letters in a message and substitution ciphers, which systematically replace letters or groups of letters with other letters or groups of letters. An example of this type of cipher is the Caesar cipher, in which each letter in the plaintext was replaced by a letter some fixed number of positions further down the alphabet.

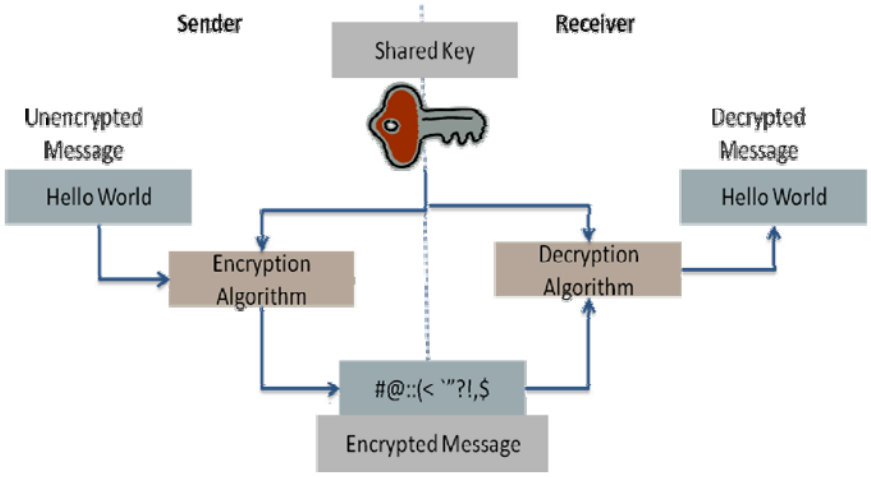


Figure 1. The encryption and decryption processes of a cipher

Modern cryptography algorithms are divided into two main categories: Symmetric-key and public-key. Symmetric-key cryptography refers to encryption methods in which both the sender and receiver share the same key. On contrast, in public-key cryptography two different but mathematically related keys are used—a public key that may be freely distributed and a private key that must remain secret [2].

The Playfair cipher is a symmetric substitution cipher which encrypts pairs of letters (digraphs). This technique uses a 5 by 5 table that is actually constructed using a key word and the ciphering process can be carried out according to few simple rules. Though the relative simplicity of the Playfair system, it is significantly hard to break with the frequency analysis used for simple monographs substitution ciphers. However, it still limits the plaintext to be formed of alphabets without punctuation, or even numerical values. In addition, the plain-text needs to be thoroughly preprocessed to remove spaces and handle any double-letter digraphs. This can make the decrypted text hard to read especially for long cipher-texts.

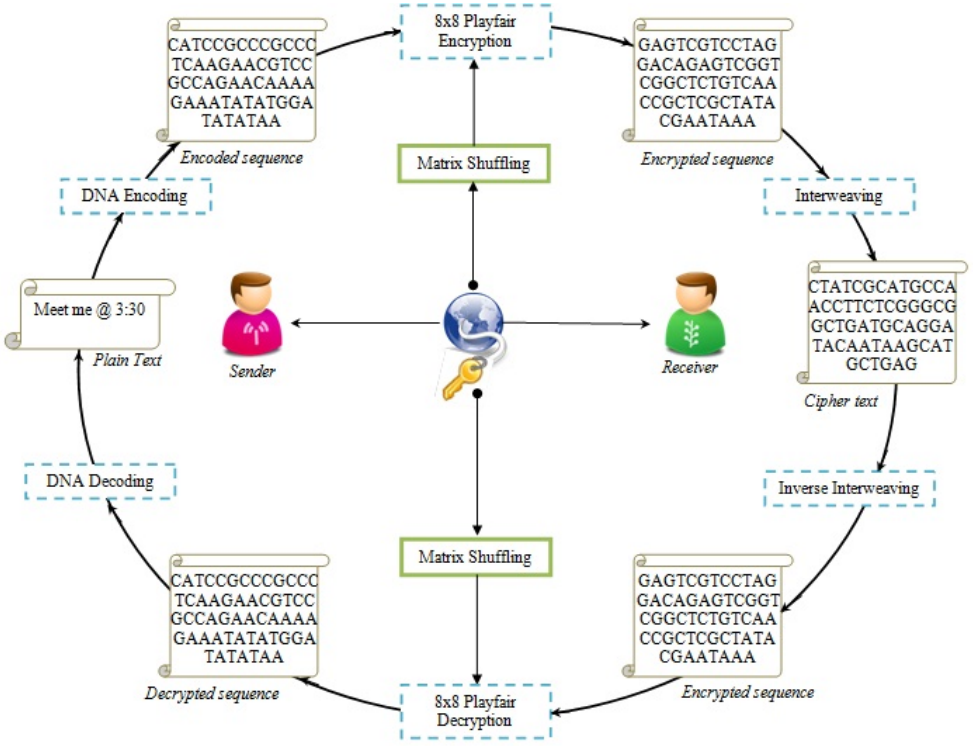


Figure 2. The general scheme of the proposed ciphering system

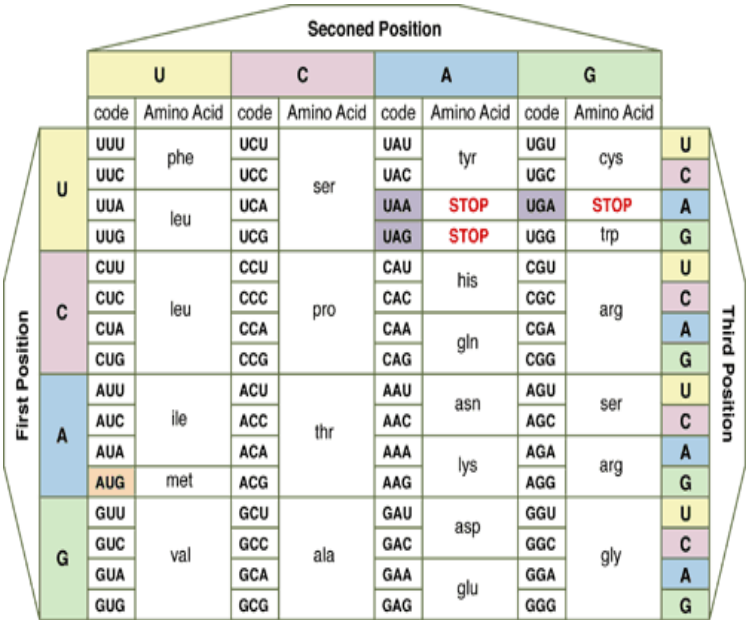
Recent research in cryptography provided some novel ways to enhance the security of the playfair cipher. An extended 8x8 playfair cipher was proposed in[3]. The proposed system was able to encrypt alphabets, numbers, and some special characters. Spaces and duplicate characters are handled using the symbols | and ^ respectively. Yet another modification appeared in [4] with a strong new cipher based on the ASCII codes of characters. However; in both cases, the proposed 64 grids were not enough to include all the keys on a standard keyboard. Another generalized 8x16 Playfair cipher was introduced in [5], which considered the 128 ASCII characters instead of the 26 characters of the English language. More research efforts are still devoted to overcome these drawbacks in the Playfair cipher. One of these trends is DNA-based cryptography. Examples of these algorithms include the "Yet Another Encryption Algorithm" (YAEA) developed by Amin et. al. [6] and an amino-acid implementation of the palyfair cipher by [7]. However, the solution suggested by the authors was complicated due to the introduction of what they called ambiguity-bits. These bits are actually added to the ciphered text in order to allow the decryption process to be carried out correctly. Furthermore, they still suffer from the plain-text preprocessing step.

In this paper, we propose a novel enhancement over the classical Playfair cipher such that any kind of binary data can be encrypted using the same rules of the 5x5 playfair cipher. The proposed technique allows cryptography to make extensive use of some concepts from the field of Bioinformatics. More specifically, both the plain- and cipher- texts will be treated as sequences of DNA. These specially encoded sequences will be encrypted and decrypted using a novel codon-based 8x8 playfair cipher. the main steps of the proposed ciphering system is depicted in figure 2.

The rest of the paper is organized as follows: Section II gives a quick overview on some basic facts and terminology that are crucial to the understanding of communing sections. In section III, we illustrate the details of the encryption/decryption processes of the proposed cipher. Section IV gives a through cryptanalysis followed by a comparative study with other similar techniques. Finally in section VI we conclude.

2. DNA PRELIMINARIES

DNA is a double stranded structure whose building blocks are called nucleotides or bases. There are only four kinds of bases: adenine (A), guanine (G), thymine (T) and cytosine (C). Hence, a sequence of DNA base pairs can be represented as a string made of these four characters i.e. <AAGTCGATCGATCATCGA>. This "genetic code" is read and eventually translated by the cellular machinery to form proteins in a long and complex process called Central Dogma [8]. The code is read and transcribed from the DNA into messenger RNA (m-RNA) three bases at a time. Each three adjacent mRNA bases (C, A, U, G) bases form a single unit known as a codon. This triplet code allows for a total of 4x4x4 or 64 different codons that are mapped to 20 different amino acids (the building blocks of proteins).

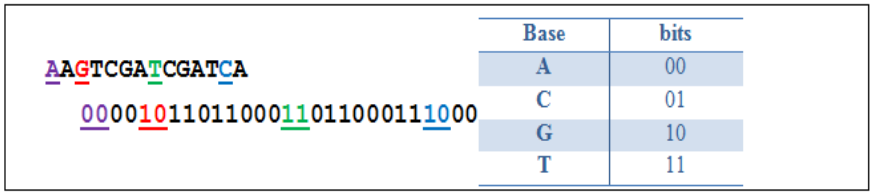


First Position	Seconed Position U code	Amino Acid	Seconed Position C code	Amino Acid	Seconed Position A code	Amino Acid	Seconed Position G code	Amino Acid	Third Position
U	UUU	phe	UCU	ser	UAU	tyr	UGU	cys	U
	UUC		UCC		UAC		UGC		C
	UUA	leu	UCA		UAA	STOP	UGA	STOP	A
	UUG		UCG		UAG	STOP	UGG	trp	G
C	CUU	leu	CCU	pro	CAU	his	CGU	arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	ile	ACU	thr	AAU	asn	AGU	ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	lys	AGA	arg	A
	AUG	met	ACG		AAG		AGG		G
G	GUU	val	GCU	ala	GAU	asp	GGU	gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	glu	GGA		A
	GUG		GCG		GAG		GGG		G

Figure 3. The Amino Acid/codons table

This means that some amino acids are coded for by more than one codon in a feature called degeneracy [8]. As shown in figure 3, the three-letter abbreviations such as “Phe” and “Leu” actually indicate the types of amino acid molecules

Therefore; from a computational point of view, DNA can be viewed as a coding medium. That is, it is convenient to adopt some coding rule to convert the DNA string of bases into binary form and vice versa. Figure 4 shows one of these rules that actually maps each base to a 2-digit binary number. That is, the bases A; C; G and T are mapped into 00, 01, 10, and 11 respectively. Similarly, the same rules can be applied on RNA sequences where the (T) nucleotide is replaced by the (U). These rules actually form the basis for the ciphering algorithm as will be illustrated shortly.



Base	bits
A	00
C	01
G	10
T	11

Figure 4. Digital coding of DNA bases

3. THE PROPOSED ALGORITHM

As we mentioned before, the proposed ciphering technique provides a novel perspective on the extended 8*8 playfair cipher based on the properties of DNA and amino acid codons. As shown in figure 2, the proposed method starts with a DNA-encoding step followed by the construction of the substitution matrix that will be used later to apply the 8x8 Playfair cipher. Furthermore, we added an interweaving step to further strengthen the security of the cipher.

Obviously, the decryption process carries out the inverse of the steps in the encryption process. That is, the binary message is encoded first into a single stranded DNA sequence. Afterwards, the inverse of the interweaving process is carried out to reveal the contents of the encrypted sequence. The nucleotides of the encrypted sequence are then grouped into codon triplets and decrypted using the 8x8 grid of codons that is constructed based on the secret key. Eventually, the contents of the original data can be disclosed by transforming the decrypted nucleotides into their binary form.

3.1. Encoding Plaintext into DNA Strands

The proposed method should start with an encoding step. In this step, the secret message M is converted for its binary form (representing any kind of media) into a single-stranded chain of DNA nucleotides based on some rule such as the one shown in Figure 3. For example, the message "Meet me @ 3:30” becomes the sequence CAT CCG CCC GCC CTC AAG AAC GTC CGC CAG AAC AAA AGA AAT ATA TGG ATA TAT AA. Of course, this encoding function is reversible and its inverse can be used to recover the original bit stream from the encoded DNA strand.

GGG	CCA	GAT	ACG	GTA	TGG	ATA	GCT
GCG	AAG	AAA	CCT	TGT	CGT	CGA	GAC
GGA	AGA	CGC	ACA	GCA	ACC	CAC	CTC
TTT	CCG	GAG	TAC	TTA	GTT	GGC	AAT
CTG	ACT	CAA	TGA	AAC	GTC	GGT	TTC
GTG	TCG	CAT	TCT	TTG	ATG	AGC	TGC
AGG	TCA	CGG	CTT	AGT	CTA	ATT	CAG
GAA	ATC	GCC	TAG	CCC	TCC	TAT	TAA

Figure 5. Resultant 8x8 matrix of codons when EgyRev@25Jan used as a cipher key

3.2. Random Shuffling of the Substitution Matrix:

In this step, we have to prepare the 8x8 matrix that will be used during the encryption step. However, unlike the classical playfair cipher, this matrix consists of 64 spots filled with codons. In addition, using a specific key, the matrix should be constructed in some unique pattern. One way to do this is to randomly scatter the codons across the grid using a random permutation function whose seed value depends on the cipher key. In this case, the key can be formed of any combination of characters where there is no need to apply any preprocessing on the key such as removing spaces or dropping any duplicate letters. For example, using EgyRev@25Jan as the cipher key will construct the 8x8 matrix of codons shown in figure 5:

3.3. The 8x8 Playfair Encryption:

Now, the encoded sequence is regarded as codon triplets that will be substituted using the 8x8 matrix in twos or digraphs. In fact, the rules of palyfair cipher will still be applied to each pair of codons in the DNA sequence. That is, each two codons of the digraph are considered as the opposite corners of a rectangle in the substitution matrix and are replaced according to the following rules:

1) If the codons appear on the same row of the table, replace them with the codons to their immediate right respectively. Wrap around to the left side of the row if a codon in the original pair was on the right side of the row.
2) If the codons appear on the same column of the table, replace them with the codons immediately below respectively. Wrap around to the top side of the column if a codon in the original pair was on the bottom side of the column.
3) If the codons are not on the same row or column, replace them with the codons on the same row respectively but at the other pair of corners of the rectangle defined by the original pair. Notice that the order is important, i.e. the first codon of the encrypted pair is the one that lies on the same row as the first codon of the plaintext pair.

Following the above rules, the message "Meet me @ 3:30" is encrypted into GAG TCG TCC TAG GAC AGA GTC GGT CGG CTC TGT CAA CCG CTC GCT ATA CGA ATA AA. Notice that since the final digraph is incomplete, we choose to excluded it from encryption and append it as it is to the end of the encrypted sequence.

3.4. The Interweaving Process:

We suggest this step to further randomize the encrypted sequence for the sake of better security. The interweaving process follows the same steps suggested in [5]. In fact, the process of interweaving is based on representing data in the form of a matrix. In this context, the encrypted sequence can be viewed as a matrix whose dimensions actually depend on the length of the sequence. The interweaving process starts with applying a circular rotation in the upward direction on the first column of the matrix. This is followed by a left circular rotation on the first row of the matrix. Similarly, circular rotations are applied on the second column and second row respectively. The column-row consecutive rotations continue until all columns and rows are rotated.

Here, we suggest the matrix to be constructed (in a column-wise manner) from the sequence in reverse order. That is, the above encrypted sequence will be represented as follows:

$$\begin{bmatrix} A & C & C & A & C & C & A \\ A & A & T & C & G & T & G \\ A & T & C & T & G & G & G \\ T & A & G & G & C & A & A \\ A & T & C & T & T & G & T \\ A & C & C & C & G & A & C \\ G & G & A & T & G & C & C \end{bmatrix}$$

Applying the interweaving process on the above matrix results in:

$$\begin{bmatrix} C & A & C & G & T & G & A \\ T & T & C & G & G & G & T \\ A & G & T & G & A & A & A \\ T & C & T & C & T & T & A \\ C & C & C & G & G & A & G \\ G & A & T & G & C & C & C \\ C & A & C & C & A & A & A \end{bmatrix}$$

Notice that we suggested using a square matrix; however the same procedure can still be applied on any dimensions. Hence, eventually the encrypted sequence will be:

CTATCGCATGCCAACCTTCTCGGGCGGCTGATGCAGGATACAATAAGCATGCTGAG

4. CRYPTANALYSIS

There are no absolute proofs that a cryptographic technique is secure, however, a technique is considered secure if it is practically hard for an attacker to crack its implementations. In fact, cryptanalysis of the playfair cipher is much more difficult than normal simple substitution ciphers, because digraphs (pairs of letters) are being substituted instead of monographs (single letters). If we use frequency analysis of English digraphs, we can use this information in the same way as we used the monograph frequencies; however, there are ~600 digraphs and only 26 monographs. That is, on an average, the probability of occurrence of any particular element in 5*5 Playfair matrix is 1/26=0.0384, whereas the probability of occurrence of an element in 8*8 playfair matrix is 1/64=0.0156, making the frequency analysis a tougher job.

Furthermore, there are some simple facts that make the classical Playfair cipher easy predictable by the attacker. For example, a Playfair digraph and its reverse (e.g. AB and BA) will decrypt to the same letter pattern in the plaintext. In addition, the cipher message contains an even number of letters that will never contain a double-letter digraph (e.g. EE). Matching this pattern to a list of known plaintext words is an easy way to generate possible plaintext strings with which to begin constructing the key [9].

However, the playfair cipher proposed in this paper actually broke that tight link with the English alphabets and symbols. As we discussed, any data can be eventually represented as a binary string which is then encoded into DNA codons and the encryption process can be done in the DNA domain without any direct association to the original data. For example, in the case of textual messages, each character is actually converted to the 8-bit representation of its ASCII code. And since each DNA codon consists of three nucleotides, it will be coded using only 6 bits. This means that a digraph of codons will represent only 75% of its matching character digraph. This irregular mapping makes it almost impossible to infer any statistical significance between digraphs in plaintext and ciphertext even if the length of the message is long enough. In addition, the binary coding step can be done in 4! = 24 different ways since there are 4 nucleotides. Therefore, there will be fewer clues for an attacker to positively confirm that the method of encryption is Playfair.

Two more features of the proposed cipher contribute directly in it strength against attacks. That is, the substitution process is followed by an interweaving process which will thoroughly mix the encrypted sequence at nucleotide level. Furthermore, the proposed method doesn't use the key directly to fill in the 8x8 grid. Instead, a hashed value of the key is used as the seed value for a random permutation function to construct the grid accordingly. Keeping the details of this function as a secret will definitely increase the security of the cipher. In conclusion, the time required for an attacker to guess all of these combinations is enormously large making this cipher practically difficult to crack with a brute force attack.

5. COMPARATIVE STUDY

Table 1 summarizes and compares the main features of the published different versions of the Playfair cipher. Three modifications were made on the classical 5x5 Playfair. Two of them [7] and the one presented here) are actually inspired by ideas in bioinformatics. The third one [3] basically sticks on the textual nature of the encrypted messages. According to the points listed in table 1, the proposed algorithm seems to overcome all of the problems inherent in the classical 5x5 cipher. In addition, it succeeded to provide the right solutions for all the shortcomings in the other algorithms.

Table 1. A Comparison between Different Playfair-Based Ciphering Methods.

Criterion	The classical	Amino Acid based [4]	The extended [3]	Interwining-based [5]	The proposed
Secret data	Letters A-Z (except J)	• Letters (upper or lower case) • numerals • all special characters	• Letters (A-Z) • Numerals (0-9) • Some special characters	128 ASCII characters	Any form of binary data (*text, sound, image, etc*)
Grid size	5x5	5x5	8x8	8x16	8x8
Plaintext preprocessing	✓ Remove all J's, spaces, punctuation, or other characters ✓ Use an x to stuff between double letters ✓ (if necessary) Append an 'x' to the to make it even		✓ Replace spaces with | ✓ Use ^ for stuffing between double alphabets ✓ (if necessary) Use ^ at the end to get the last alphabet in pair	Not specified	Not required
Key	English Letters without any number or special characters			Any combination of characters	
Length of ciphertext	Even, same as processed plaintext	Even, 33% longer than plaintext because of ambiguity bits	Even, same as processed plaintext		Not necessarily even
Form of ciphertext	Textual with no spaces, no double letters, nor special characters	Can be represented as text, binary form, DNA string, or Amino Acid characters	Textual or binary	Textual	same as original data , binary or as a DNA sequence
Advantages	✓ Simple ✓ Easy to trace manually using paper and pencil	✓ Innovatively used amino acids as a means of representation ✓ Can be integrated in a biological DNA encryption process using one time pad or substitution.	✓ very strong cipher ✓ very effective for areas with low bandwidth or very less memory storage ✓ increased randomness using LFSR ✓ An extended allowed character set.	✓ Strong due to iteration, interwining, and interweaving. ✓ exhibits strong avalanche effect ✓ An extended allowed character set	✓ Robust against cryptanalytic attacks ✓ Same rules as traditional playfair, but different outcome with much less correlation between plain- and ciphered-data. ✓ The same technique can be applied on all kinds of binary data. ✓ Can be combined with DNA steganography techniques (A. Atito, A. Khalifa and S. Reda, 2011)
Disadvantages	✓ Requires a lot of preprocessing ✓ sometimes decrypted text is very hard to read due to limited letters allowed.	✓ Requires a lot of preprocessing ✓ Added ambiguity bits ✓ Inaccurate decrypted text	✓ Suitable only for textual data ✓ Still requires preprocessing	✓ Suitable only for textual data	

6. CONCLUSION

This paper proposed a novel implementation of the Playfair cipher using an 8x8 matrix based on the 64 DNA codon table. This enhancement was introduced to overcome the problems inherent in the classical 5x5 Playfair cipher. One great advantage over the classical Playfair; which works only on English letters, the proposed algorithm is able to cipher any kind of digital messages such as text, sound, images, etc. In addition, it doesn't require any pre-processing for the plain-text resulting in a complete and precise decrypted message. Furthermore, the cipher-text can be represented either in a variety of formats including binary and DNA. In this way, it can be considered a step in a longer and more complicated process such as Information hiding in DNA sequences.

A through cryptanalysis showed that the proposed cipher is practically difficult to crack with a brute force attack. That is, encoding binary messages into DNA codons provides irregular mapping between corresponding character digraphs. Furthermore, applying an interweaving step on the substituted sequence increases the security of the technique. Therefore, carrying a frequency analysis by an attacker would be nearly impossible especially with fewer clues leading to the Playfair cipher.

REFERENCES

[1] David K. *The Codebreakers – The Story of Secret Writing*. 1967 New York: Macmillan.
[2] 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.
[3] 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.
[4] Srivastava SS and N Gupta. A Novel Approach to Security using Extended Playfair Cipher. *International Journal of Computer Applications*. 2011; 20(6): 0975 – 8887.
[5] 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.
[6] Amin ST, M Saeb and S El-gindi. A DNA-based implementation of YAEA encryption algorithm. *Computational Intelligence*. 2006: 120-125.
[7] Sabry M, et al., A DNA and Amino Acids-Based Implementation of Playfair Cipher. *International Journal of Computer Science and Information Security*. 2010; 8(3): 129-136.
[8] Crick F, Central dogma of molecular biology. *Nature*. 1970; 227: 561–563.
[9] Department of the Army. *Basic Cryptanalysis, FM 34-40-2, FIELD MANUAL*, 1990: Washington

BIOGRAPHY OF AUTHOR

Dr. Safwat Hamad. Currently, He is working as an assistant professor of Scientific Computing at Faculty of Computers & Information Sciences, Ain Shams University, Egypt. He graduated in 2000 and worked as a teaching assistant for a number of undergraduate courses till 2004 at the same Faculty. Meanwhile, He got his MSc degree in the field of Modeling Simulation and Visualization. In 2005 He was granted a 2 years research scholarship in University of Connecticut, USA. He earned his PhD degree in 2008 in the area of High performance Computing. His main research interests are Steganography, computational biology, parallel computing, encryption and Security.