Comparative Study of Three DNA-based Information Hiding Methods

Nisreen Suliman Terkawi *nisreenterkawi@gmail.com*
Computer Science Department,
Imam Mohammad Ibn Saud
Islamic University
Riyadh, 11564, Saudi Arabia

Lamia Berriche *lberriche@psu.edu.sa*
Computer Science Department,
Prince Sultan University
Riyadh, 12435, Saudi Arabia

Amjad Ali Alamar *aamr@imamu.edu.sa*
Computer Science Department,
Imam Mohammad Ibn Saud
Islamic University
Riyadh, 11564, Saudi Arabia

Maimounah Abdurahman Albrahim *Maimounah6@gmail.com*
Computer Science Department,
Imam Mohammad Ibn Saud
Islamic University
Riyadh, 11564, Saudi Arabia

Wafaa Saad Alsaffar *wafaa.alsaffar@gmail.com*
Computer Science Department,
Imam Mohammad Ibn Saud
Islamic University
Riyadh, 11564, Saudi Arabia

Abstract

Cryptography is the science of protecting information by transforming data into formats that cannot be recognized by unauthorized users. Steganography is the science of hiding information using different media such as image, audio, video, text, and deoxyribonucleic acid (DNA) sequence. The DNA-based steganography is a newly discovered information security technology characterized by high capacity, high randomization, and low modification rate that leads to increased security. There are various DNA-based methods for hiding information.. In this paper, we compared three DNA-based techniques (substitution, insertion, and complementary) in terms of its capacity, cracking property, Bit Per Nucleotide (BPN), and payload. The selected algorithms combine DNA-based steganography and cryptography techniques. The results show that the substitution technique offers the best BPN for short secret messages and offers the best imperceptibility feature. We also found that both the substitution and the complementary method have a threshold BPN. On the other hand, the insertion method does not have a threshold BPN and it is more difficult to crack.

Keywords: Information Security, Steganography, Substitution, Insertion, Complementary Pair.

1. INTRODUCTION

The transmission of sensitive information through the Internet faces high risks. Therefore, exchanging messages between senders and receivers is required to be in a confidential manner to avoid attacks. Sensitive data protection from unauthorized access is provided by two major techniques; steganography and cryptography [1]. Cryptography is a technique for preventing third parties from reading a secret message by converting it to an encrypted format which is incomprehensible for intruders. Some of the methods applied in the encryption are Playfair, Rivest Shamir Adleman (RSA), and Advanced Encryption Standard (AES) [2]. Steganography is a technique of hiding a secret message inside a cover message making it unnoticeable to any illegal read. The cover media could be text [3], image [4] [5][6], audio [7], video [8], and the Deoxyribonucleic Acid (DNA) sequence [9]. A double layer of security is provided by some approaches that combine steganography with cryptography where a secret message is first encrypted then an encrypted message is hidden in a cover media. They are notable for achieving confidentiality and low cracking probabilities [10].

Data hiding based on the DNA sequence has been attracting much attention due to its potential storage capacity [11]. Several DNA steganography approaches have been proposed [12] [13] [14]. DNA-based steganography relies on three techniques; insertion technique where the secret message is inserted into the DNA sequence, a complementary technique where some DNA components are complemented based on the secret message, and the substitution technique where some DNA components are substituted by the secret message data.

In this paper, we focused on DNA-based steganographic techniques, because it has advantages such as the huge data storage capacity and the high imperceptibility [6][7][8]. We also compared three double-layered security techniques: combined DNA-based Playfair cryptography and substitution technique [15], a combined Playfair cryptography and complementary technique [16], and a combined XOR-based cryptography and insertion technique [17].

In section 2, we describe how DNA sequences are used for cryptography and steganography. In section 3, we present the three methods and their performance criteria. In section 4, we present our results and discussions.

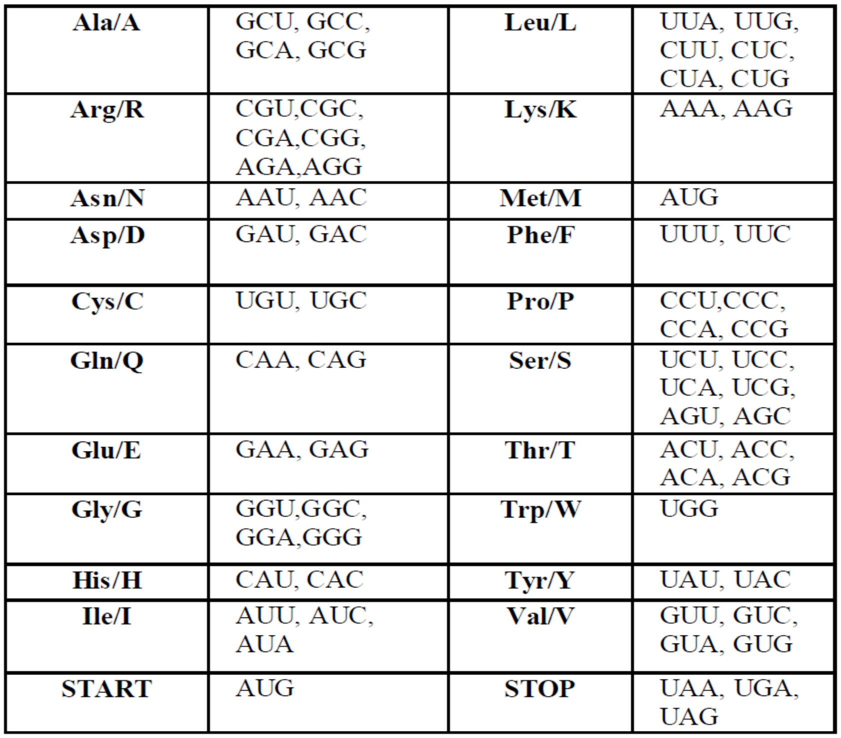
2. DNA-BASED DATA PROTECTION

This section focuses on DNA sequences, DNA-based cryptography, and DNA-based steganography.

2.1 DNA Sequence

The Deoxyribo Nucleic Acid (DNA) comprises six molecules; a sugar molecule called deoxyribose, a phosphate molecule, and four different nitrogenous bases (Adenine (A), Thymine (T), Cytosine (C), and Guanine (G)). These molecules are bound such that two long strands are twisted around like a ladder. Each strand is made up of units of nucleotides which consists of three basic molecules: sugar (S), a phosphate (P) group, and one of four nitrogen bases. The nitrogenous bases are Purines (A and G) and Pyrimidines (T and C). Every DNA can be viewed as a sequence of bases (AAGTCGATCGATCATCGATCATACGT). Every three adjacent bases constitute a unit known as the codon which corresponds to a specific amino acid. Exactly 61 codons of the total 64 codes for 20 amino acids. The presence of 'START' and 'STOP' codons signal the end of protein synthesis in all living organisms. Each amino acid has a name, an abbreviation, and a character symbol from the English alphabet, see TABLE 1. DNA sequences are of a huge size which allows them to provide high embedding capacity to hide the huge data [18], [19].

DNA sequences can be encoded using a binary code. A 2-bit code representation (00-01-10-11) is needed to encode the four DNA bases. Consequently, there are 4! =24 code permutations. The simplest Binary Coding Rule (BCR) to encode the 4 nucleotide bases (A, T, C, G) is: 0(00), 1(01), 2(10), 3(11) respectively [20] .



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 1: Amino Acids and their Codons.

2.2 DNA Based Cryptography

In 1999, authors of [21] proposed a cryptosystem using DNA. They developed a one-time pad encryption algorithm using DNA substitution and a bit-wise XOR scheme based on molecular computation. Going forward, [13] proposed another cryptosystem based on the manipulation of DNA binary strands. Other symmetric and asymmetric cryptosystems were proposed later [22], [23]. The Playfair algorithm is symmetric encryption based on substitution technique. The technique encrypts pairs of letters (bigrams or diagrams) by processing them in plaintext as units rather than as single letters. The Playfair algorithm uses 5 × 5 matrices of letters constructed using a keyword known at both the sender and receiver sides. The cipher replaces each pair of letters in the plaintext with another pair of letters [1]. In [24], they proposed a Playfair encryption based on amino-acids structures. The authors converted a plaintext message to a binary format and represented pairs of bits by their nucleoid symbol using the BCR. Afterward, they converted the DNA sequence to a sequence of amino acids using information presented in TABLE 2. Further, they encrypted the sequence of letters using a Playfair cipher. The advantage of this method over the non-DNA-based Playfair is its ability to encrypt messages with letters, numbers, and special characters. In [25], the authors proposed a DNA-based encryption for cloud storage.

2.3 DNA Based Steganography

DNA-based steganography uses DNA sequence from the National Center for Biotechnology Information (NCBI) [26] to hide data [12] [13] [27] [28]. DNA steganography starts by converting the DNA sequence into binary using a BCR code then the binary secret message is hidden in the DNA sequence using insertion, substitution, or complementary rule method [14]. DNA-based steganography is also categorized as blind and non-blind. In blind steganography, the secret message is extracted at the receiver side without the need for prior knowledge of the DNA reference sequence. The DNA-based steganographic approaches are the insertion, complementary rule-based, and substitution technique.

a) Insertion Technique

For this technique, the secret message is inserted in a DNA reference sequence. In [14], the secret message and the DNA reference sequence were translated into binary. The DNA sequence and the secret message were divided into equal-sized segments. This allows for easy insertion of each part of the secret message after each part of the DNA reference. The final stego-message is then converted into a DNA sequence. The main drawbacks of this technique are the increase in redundancy and the length of faked DNA which is higher than the length of the original DNA [14]. In [29], they proposed a technique composed of two phases. In the first phase, the secret data is encrypted using a DNA and Amino Acids-Based Playfair cipher. While in the second phase the encrypted data is hidden into some reference DNA sequence using an insertion technique. Their insertion method is based on dividing both the encrypted DNA sequence and the reference DNA sequence into segments of a random number of DNA nucleotides. Then, they insert each segment of encrypted DNA sequence before the segments of reference DNA sequence respectively.

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

TABLE 2: Alphabet-Amino Acids Correspondences.

b) Complementary Rule-Based Technique

Here, the secret data is hidden in the DNA reference sequence using a complementary rule for the nucleotides; for example, ((AC) (CG) (GT) (TA)). In a study conducted by [14], the DNA sequence was parsed for the longest complimentary substrings. A longer substring of nucleotide and its complement are further inserted into the DNA sequence before the found complementary substrings. Afterward, the secret DNA message nucleotide is inserted in special positions that depend on the position of the complementary substring. An earlier study proposed an RSA encryption followed by complementary rule-based steganography where the complemented nucleotides are selected based on a random number [29]. In a previous study [30], the authors complement the secret message initially converted to a DNA sequence. Then, they select a DNA reference sequence that should be known at the receiver side. They form a message which is a sequence number where the numbers correspond to the indexes of the appearance of the pairs of the nucleotide of the DNA complemented secret message in the selected DNA reference sequence. In [31], the authors proposed a DNA hiding technique based on complementing the codon postfix nucleotide.

c) Substitution Technique

Regarding the substitution technique, the selected positions in the reference DNA sequence are substituted by other bases depending on the binary sequence of the secret message [14]. Selected positions may be generated randomly [14], using a lookup substitution table [29], [32], or using the Least Significant Base (LSB) substitution mechanism [24],[15]. The main advantage of this technique is preserving the length of the DNA sequence after hiding the secret message.

3. METHODOLOGY

We compared three recent hybrid blind DNA encryption and data hiding techniques [15]–[17]. The three methods involves encrypting the secret message and hiding the encrypted message in a DNA reference sequence. In our deductive approach, we started by comparing the three approaches performance measures. Further, we conducted experiments with different DNA reference sequences from the NCBI dataset.

3.1 The Substitution Method [15]

In [15], authors used the DNA-based Playfair method for encryption and the substitution method for steganography.

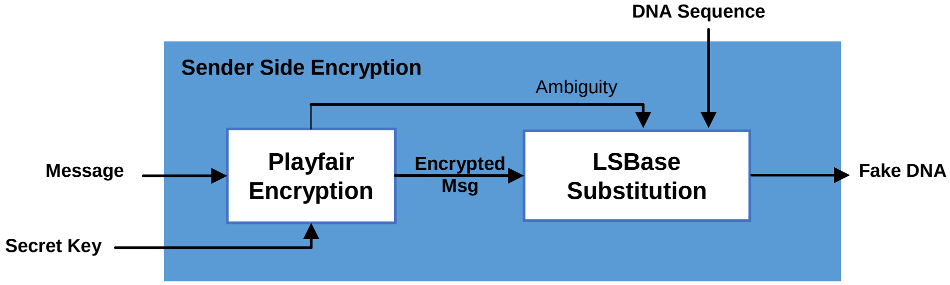


FIGURE 1: Sender side of the substitution method in [8].

Initially, the secret message is transformed into its corresponding ASCII code, then to binary using 8-bits coding. The binary secret message is then converted to a DNA nucleotide using 4-bits BCR using data presented in TABLE 3 which maps every 4 bits binary message to two 2-bits DNA nucleotides. The DNA of the secret message is then converted to amino acids. Next, Playfair with a secret key is used to encrypt the amino acid form of the secret message. Further, the cipher message is converted back to DNA by selecting a codon corresponding to each amino acid; the indexes of the codons are stored in an ambiguous message. FIGURE 1 shows the sender-side architecture.

DNA Nucleotides	Binary Representation	DNA Nucleotides	Binary Representation
AA	0000	GG	1000
AC	0001	GA	1001
AG	0010	GC	1010
AT	0011	GT	1011
CC	0100	TT	1100
CA	0101	TA	1101
CG	0110	TC	1110
CT	0111	TG	1111

TABLE 3: 4-bit BCR used in [8].

After encrypting the message, the steganography process starts by using LSBase substitution [24]. The LSBase method hides the secret message by substituting the least significant base of each codon in the reference DNA sequences. Both the ciphered DNA message and the ambiguity message are converted back to binary representation. The ciphered DNA message conversion uses 4-bits BCR. Then, they hide the binary ciphered DNA message bits and the binary ambiguity bits in the reference sequence. If LSBase is a purine base, it is substituted by (G) to encode 1 of the secret messages or (A) to encode 0. If the LSBase is a pyrimidine base, it is substituted by (C) to encode 1 of the secret messages or (U) to encode 0. The innovation idea 3:1 ratio is used to hide 3 bits of binary cipher message followed by 1 bit of binary ambiguity.

At the receiver side, first, the ciphered message and ambiguity are extracted using the LSBase method. Then Playfair decryption using the same secret key is applied, see FIGURE 2. The use of a 4-bit binary coding rule increases the algorithm security, so the likelihood of making a correct guess of the binary coding rule is decreased from $\frac{1}{4!}$ of a 2-bit BCR to $\frac{1}{16!}$. Also, the use of the 3:1 rule avoids the addition of an indicator message to separate the secret message from the ambiguous message.

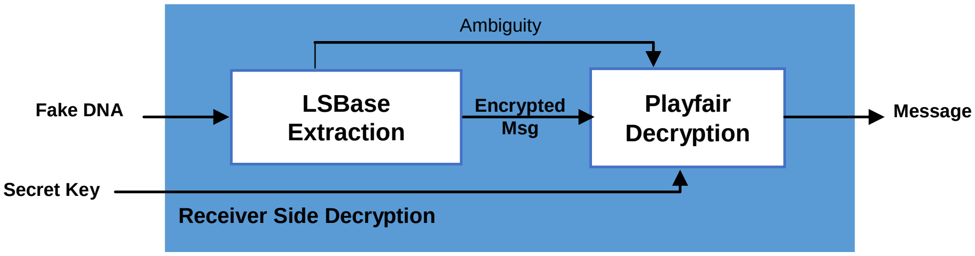


FIGURE 2: Receiver side of the substitution method in [8].

3.2 The Insertion Method [17]

In [17], they used the XOR operation for encryption and the insertion method for steganography. First, the secret message is converted into ASCII code then into a binary sequence. The binary sequence of the secret data is split into 8-bit binary segments. An 8-bit key K1 is then XORed with the first 8 bits of the message. The resulting XOR value is again XORed with the next 8 bits of the message and so forth. All the results are finally concatenated to form the cipher message. Afterward, a binary converted DNA sequence is divided into segments using a randomly generated key K2 which should be a number less than the minimum DNA sequence length and the secret message length. The binary bits of ciphers are inserted one by one at the beginning of each segment. The resulting binary sequence is converted into DNA bases using the dictionary rule and sent as Fake DNA. At the receiver, only the two keys K1 and K2 are needed with the fake DNA message to extract the secret one. FIGURE3 shows the insertion method sender and receiver.

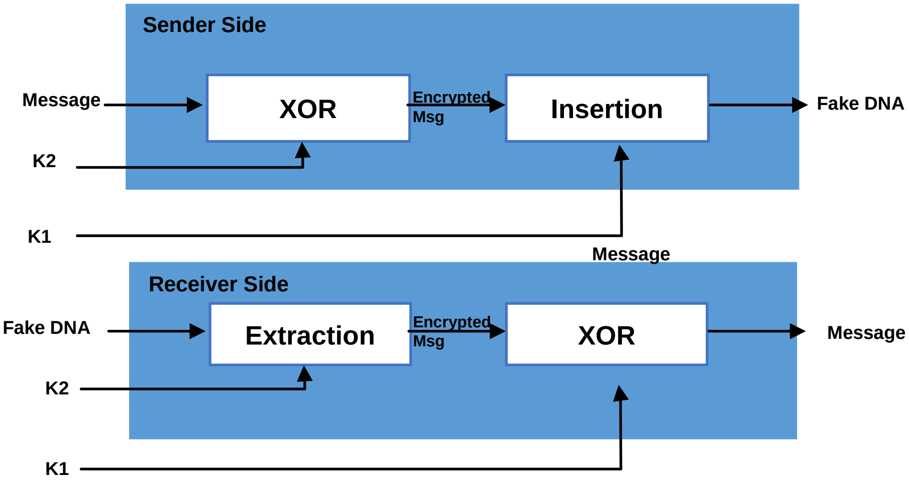


FIGURE3: Sender and receiver of the insertion method in [10].

3.3 The Complement Method [16]

In [16], the authors relied on two layers of techniques to provide higher security. Firstly, they applied a DNA-based Playfair encryption algorithm followed by a complementary substitution steganography technique. At the sender side, a DNA-based Playfair cryptography produces a cipher DNA message and an ambiguity message was adopted. Moving forward, a Generic Complementary Base Substitution (GCBS), based on TABLE 4, was employed to embed the cipher message with the ambiguity sequence into a DNA cover sequence. To indicate the end of the secret message, they embed a palindromic sequence bounded with two nucleotide T after the message. To preserve the length of the DNA sequence, the resultant DNA sequence from the previous step was truncated. Their GCBS doubles the embedding capacity compared to [14]. The final resultant DNA-sequence was inserted into the original one using the insertion method [14]. This step aims to provide the receiver with the reference sequence.

The receiver extracts the reference sequence first, locates the palindrome, and extracts the embedded message by comparison with the previously obtained reference sequence. Finally, it decrypts the message using Playfair deciphering module, see FIGURE 4.

Base	Generic Complement
A	C
C	T
G	A
T	G

TABLE 4: Complementary Rule.

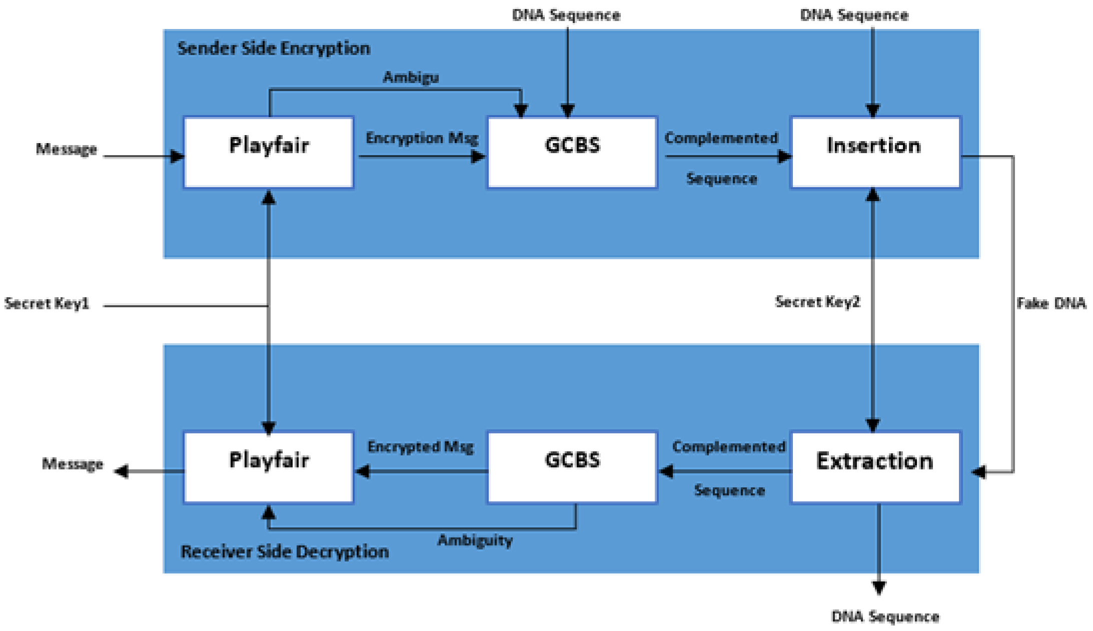


FIGURE 4: Sender and Receiver of the Complementary Technique in [9].

3.4 Comparison Criteria

We compared the three proposed hybrid techniques with regards to four parameters; capacity measure, payload measure, BPN measure, and cracking probability measure.

a) Capacity

Defined in [14] as the total length of increased DNA sequence after the insertion of the secret information. In line with [15], the cipher message and ambiguity bits substitute DNA reference sequence nucleotide, so the capacity is equal to the length of the DNA reference sequence $|S|$. In [16], a DNA sequence hides as much message nucleotide as its length, and the DNA sequence with the embedded secret message is inserted in itself. So, the capacity in [16] is $2 * |S|$. In [17], every two bits of the secret message are converted to a nucleotide and inserted into the DNA reference sequence. So, the capacity is equal to $|S| + \frac{|M|}{2}$ where $|S|$ is the DNA reference sequence length and $|M|$ is the secret message length in bits.

b) Payload

Defined in [14] as the length of the new sequence after the extraction of the reference DNA sequence. In [15], the DNA reference sequence length is preserved which gives a zero payload value. As per [16], a DNA sequence embeds as much nucleotide as itself and is embedded in itself which gives a payload of $|S|$. On the word of [17], every 2 bits of the secret message are inserted as a nucleotide in the DNA reference sequence which gives a payload value of $\frac{|M|}{2}$.

c) BPN measure

Defined in [14] as the number of hidden bits per nucleotide.

$$BPN = \frac{Length\ of\ secret\ message\ in\ bits}{capacity} \qquad Eq.\ 1$$

On the report of [15], they embed $|M|\ bits$ in $|S|$ nucleotide. Consequently, $bpn = \frac{|M|}{|S|}$. However, the used LSBase algorithm can hide only one bit per codon. Consequently, only $\frac{1}{3}|S|$ codons are used for the embedding process. Besides, $\frac{3}{4}$ of the codons are used for data and $\frac{1}{4}$ is used to embed the ambiguity information. So, the maximum bpn is $\frac{\frac{1}{3}|S|*\frac{3}{4}}{|S|} = \frac{1}{4}$.

In [16], they embed $|M|\,bits$ in $2|S|$ nucleotide. Consequently, $BPN = \frac{|M|}{2*|S|}$. Each DNA cipher message nucleotide is hidden in one cover DNA sequence nucleotide. So, 2-bits of data can be hidden in each nucleotide of a DNA sequence of length $|S|$ ($2|S|$ bits per $|S|$ nucleotide). The DNA cipher message includes both the encrypted secret message and the ambiguity of the DNA-based Playfair. Besides, each encrypted secret message codon corresponds to one ambiguity number. So, ¾ of the DNA sequence is used to embed the secret message whereas ¼ is used to embed the ambiguity message. So, the threshold BPN is $\frac{\frac{3}{4}*2*|S|}{2*|S|} = \frac{3}{4}$.

In [17], the total number of hidden bits is |M| and the length of the fake DNA sequence is |S|+|M|/2. So, the bpn is $\frac{|M|}{|S|+\frac{|M|}{2}}$. This method embeds unrestricted length messages into any DNA sequence.

d) Cracking Probability

The possibility for the intruder to crack the fake DNA to extract the hidden secret depends on the factors. As reported by [15], there are three pieces of information that the intruder should crack to extract the secret message namely the DNA reference sequence, binary coding rule, and LSB substituted permutations. There are 163 million DNA sequences available publicly. For the binary coding rule, there are 16 pairs (AA, AC, AG….) and each pair can be presented by a sequence of 4 bits, so there are $16!$ possibilities to represent the pairs with 4 bits. The least significant base substitution rule has two possibilities for the pyrimidine substitution and two possibilities for the purine substitution; 4 possible LSBase substitution rules. Accordingly, the cracking probability of the technique is

$$\frac{1}{1.63*10^8} * \frac{1}{16!} * \frac{1}{4} \quad Eq.\ 2$$

As stated in [16], the attacker has to find the following information to discover the secret message: The random number generator, the two seeds used in the insertion phase, the complementary rule, and the binary coding rule. The total number of possible seeds is $(2^s - 1)^2$ [16], where s is the length of the DNA sequence in bits. There are 6 possibilities of complementary rules such that $C(x) \neq CC(x) \neq CCC(x)$. Each nucleotide is encoded with two bits, so the possible number to encode the 4 nucleotides is $4!$. Accordingly, the cracking probability is

$$\frac{1}{(2^s - 1)^2}\ X \frac{1}{6} X \frac{1}{24} \quad Eq.\ 3$$

According to [17], the attacker needs the following information to crack the secret message: The DNA sequence, the binary coding scheme, the sizes of the secret message and the prefix DNA, the keys used for the insertion phase, and the XOR combinations. The probability to predict the reference DNA sequence is $\frac{1}{1.63*10^8}$. Each of the 4 nucleotides is encoded with 2 bits. So, the total number of binary codes is $4! = 24$. For a fake DNA sequence of n bits, there are $n-1$ possibilities of secret messages of $|M|$ bits and DNA sequence of s bits such that $|M| + s = n$. The total number of guesses of segmented DNA message is $2^s - 1$ [17]. Whereas, the total number of guesses of the segmented secret message is $2^{|M|} - 1$ [17]. The total number of possible XOR operations of a key of length 8 bits and a message of length m is $2^{8|M|}$. Accordingly, the cracking probability of the insertion technique is

$$\frac{1}{1.63*\ 10\text{^}8}\ X \frac{1}{24} X \frac{1}{(n-1)} X \frac{1}{(2^{|M|}-1)} X \frac{1}{2^s-1} X \frac{1}{2^{8|M|}} \quad Eq.\ 4$$

Measures	**Substitution Method** [15]	**Complementary Substitution Method**[16]	**Insertion Method**[17]
Actual Capacity (nucleotide)	|S|	2*|S|	$\lvert S\rvert + \frac{\lvert M\rvert}{2}$
Payload (nucleotide)	0	$\lvert S\rvert$	$\frac{\lvert M\rvert}{2}$
BPN (bits per nucleotide)	$\frac{\lvert M\rvert}{\lvert S\rvert}$	$\frac{\lvert M\rvert}{2*\lvert S\rvert}$	$\frac{\lvert M\rvert}{\lvert S\rvert + \frac{\lvert M\rvert}{2}}$
Cracking Probability	$P(SG) = \frac{1}{1.63 * 10^8} * \frac{1}{16!} * \frac{1}{4}$	$\frac{1}{(2^{s} - 1)^2} \text{X} \frac{1}{6} \text{X} \frac{1}{24}$	$\frac{1}{1.63 * 10\text{^}8} \text{X} \frac{1}{24} \text{X} \frac{1}{(n-1)} \text{X} \frac{1}{(2^{\lvert M\rvert}-1)} \text{X} \frac{1}{2^{s}-1} \text{X} \frac{1}{2^{8\lvert M\rvert}}$ • n is the number of bits in the Fake DNA sequence. • $\lvert M\rvert$ is the number of bits in the secret message. • s is the number of bits in the reference DNA sequence.

TABLE 5: Performance criteria of the three methods.

TABLE 5 shows the capacity, payload, BPN, and cracking probability of the three methods. It is highlighted that the substitution method offers the lowest payload of value 0 which increases the imperceptibility of its stego-message. On the other hand, the complementary method payload is constantly equal to $|S|$ and the payload of the insertion method depends on the message length. This makes the later method more perceptible than the others. Also, we notice that the substitution method offers a higher bpn. But the later method embeds secret message bits only in the lowest significant base, this limits its usage to large DNA sequences. Especially, as only $\frac{1}{4}|S|$ is dedicated to the embedding of the secret message, the secret message length should not exceed $\frac{1}{4}|S|$. Consequently, a secret message of length $|M|$ is embedded in a DNA sequence of minimum length $4.|M|$. In the complementary method, only $\frac{3}{2}|S|$ of the DNA sequence is used for the secret message embedding which limits its use to DNA sequences of a length exceeding $\frac{2|M|}{3}$. The insertion method does not have any constraint on the length of the DNA sequence.

4. RESULTS AND DISCUSSION

We conducted the experiments of the three methods on the following input: a secret message M of size 1000 Bytes containing letters, numbers, and special characters. The Playfair secret key is 'SECURITY'. Also, we used eight different DNA reference sequences from the NCBI database to measure each capacity, payload, and BPN. The National Center for Biotechnology Information (NCBI) houses a series of databases relevant to biotechnology and biomedicine and is an important resource for bioinformatics tools and services. Major databases include GenBank for DNA sequences. All these databases are available online [26].

DNA reference	Number of nucleotides	Capacity (nucleotide)	Payload(nucleotide)	BPN
AC153526	200117	200117	0	0.0399766137
AC166252	149814	149814	0	0.0533995488
AC167221	204841	204841	0	0.0390546814

AC168901	191136	191136	0	0.0418550142
CS236146	118777	118777	0	0.0673531071
JX978467	9270	Short DNA sequence		
NC_021114	8870	Short DNA sequence		
NC_021116	8958	Short DNA sequence		

TABLE 6: Results for Substitution Technique.

DNA reference	Number of nucleotides	Capacity (nucleotide)	Payload (nucleotide)	BPN
AC153526	200117	204117	4000.0	0.0391932078
AC166252	149814	153814	4000.0	0.0520108703
AC167221	204841	208841	4000.0	0.0383066543
AC168901	191136	195136	4000.0	0.0409970482
CS236146	118777	122777	4000.0	0.0651587838
JX978467	9270	13270	4000.0	0.6028636021
NC_021114	8870	12870	4000.0	0.6216006216
NC_021116	8958	12958	4000.0	0.6173792252

TABLE 7: Results for Insertion Technique.

DNA reference	Number of nucleotides	Capacity	Payload	BPN
AC153526	200117	400234	200117	0.0199883068
AC166252	149814	299628	149814	0.0266997744
AC167221	204841	409682	204841	0.0195273407
AC168901	191136	382272	191136	0.0209275071
CS236146	118777	237554	118777	0.0336765535
JX978467	9270	18540	9270	0.4314994606
NC_021114	8870	17740	8870	0.4509582864
NC_021116	8958	17916	8958	0.4465282429

TABLE 8: Results for Complement Technique.

Concerning TABLE 6, we noticed that the substitution technique proposed in [15] is limited for use with long DNA sequences. A secret message of length $|M|$ is embedded in a DNA sequence of minimum length $4.|M|$; a message of length 8000 bits could be embedded in a DNA sequence of minimum length 32000 nucleotides.

We observed from TABLE 6, TABLE 7 and TABLE 8 that the payload of the substitution technique is the lowest (equal to zero). The payload of the insertion method depends on the secret message length; so it's identical for all the DNA sequences. On the other hand, the payload of the complementary technique is the highest as the DNA sequence is embedded into itself. When the payload is low, the stego-message becomes less perceptible. Hence, the LSBase substitution technique is the less perceptible technique.

Also, we noticed from TABLE 6, TABLE 7, and TABLE 8 that the substitution technique gives the highest BPN for the first five DNA sequences whilst the complementary techniques provide the smallest BPN. The main backdrop of the complementary technique is the fact that the DNA sequence is inserted into itself thereby doubling the fake DNA sequence size. However, one advantage of this technique is that only the secret message is hidden into the sequence compared to the two previous methods where an ambiguous message is also embedded.

The authors [33] compared different DNA-based steganography approaches, blind or non-blind with and without encryption. Findings from their study did not provide the BPN threshold. In our study, the BPN measure threshold was computed and the findings showed that the BPN measure of both the substitution and complementary techniques are upper bounded by a threshold that limits their embedding capacities.

5. CONCLUSION

Data transfer through the Internet has become necessary and important but the transmission of sensitive information through the Internet is at high risk. it is unreliable and unsafe. Therefore, exchanging messages between the sender and receiver is required to be in a confidential manner to avoid being hacked or susceptible to threats through the internet. Currently, many algorithms have been extracted in the field of steganography based on DNA to prevent unauthorized access and increase data security. For that purpose, we compared three blind hybrid steganography techniques: substitution technique, insertion technique, and complementary technique. We analyzed the performance of the three techniques. Findings from the study analysis revealed that the substitution payload was 0 which increases the imperceptibility of the message. We also noticed that for short secret messages, the substitution technique offers the best BPN. On the other hand, because of the use of the least significant bases only for the embedding of the secret message this method is restricted to DNA sequences longer than four times the length of the message. Also, this method has the highest cracking probability. Besides, we observed that both the substitution and the complementary methods have a threshold BPN. Unlike the two aforementioned techniques, the insertion method is unrestrictedly used to embed secret messages with low cracking probability enabling it to store large messages in DNA sequences. Nevertheless, the insertion approach is more perceptible than the previous ones as its payload depends on the secret message length.

6. REFERENCES

[1] W. S. and L. Brown, *Computer security: principles and practice*, Third. Pearson, 2015.

[2] A. M. Qadir and N. Varol, “A review paper on cryptography,” *7th Int. Symp. Digit. Forensics Secur. ISDFS 2019*, 2019, doi: 10.1109/ISDFS.2019.8757514.

[3] N. Alghamdi and L. Berriche, “Capacity Investigation of Markov Chain-Based Statistical Text Steganography: Arabic Language Case,” in *Proceedings of the 2019 Asia Pacific Information Technology Conference*, 2019, pp. 37–43, doi: 10.1145/3314527.3314532.

[4] N. Hamid, A. Yahya, R. B. Ahmad & Osamah, and M. M. Al-Qershi, “Image Steganography Techniques: An Overview,” *Int. J. Comput. Sci. Secur.*, no. 6, p. 168, 2012.

[5] M. E. Saleh, A. A. Aly, and F. A. Omara, “Enhancing Pixel Value Difference (PVD) Image Steganography by Using Mobile Phone Keypad (MPK) Coding,” *Int. J. Comput. Sci. Secur.*, vol. 9, no. 2, pp. 96–107, 2015.

[6] N. Pandian, “An Image Steganography Algorithm Using Huffman and Interpixel Difference Encoding,” *Nithyanandam Pandian Int. J. Comput. Sci. Secur.*, no. 8, p. 202, 2014.

[7] S. Mishra, V. K. Yadav, M. C. Trivedi, and T. Shrimali, “Audio steganography techniques: A survey,” in *Advances in Intelligent Systems and Computing*, 2018, vol. 554, pp. 581–589, doi: 10.1007/978-981-10-3773-3_56.

[8] S. Raja Ratna, J. B. Shajilin Loret, D. Merlin Gethsy, P. Ponnu Krishnan, and P. Anand Prabu, “A Review on Various Approaches in Video Steganography,” in *Lecture Notes on Data Engineering and Communications Technologies*, vol. 33, Springer, 2020, pp. 626–632.

[9] P. Johri, A. Mishra, S. Das, and A. Kumar, “Survey on steganography methods (text, image, audio, video, protocol and network steganography),” *Proc. 10th INDIACom; 2016 3rd Int. Conf. Comput. Sustain. Glob. Dev. INDIACom 2016*, 2016.

[10] M. S. Taha, M. S. Mohd Rahim, S. A. Lafta, M. M. Hashim, and H. M. Alzuabidi, “Combination of Steganography and Cryptography: A short Survey,” *IOP Conf. Ser. Mater. Sci. Eng.*, 2019, doi: 10.1088/1757-899X/518/5/052003.

[11] D. Panda, K. A. Molla, M. J. Baig, A. Swain, D. Behera, and M. Dash, “DNA as a digital information storage device: hope or hype?,” *3 Biotech*, vol. 8, no. 5. Springer Verlag, May 01, 2018, doi: 10.1007/s13205-018-1246-7.

[12] C. T. Clelland, V. Risca, and C. Bancroft, “Hiding messages in DNA microdots,” *Nature*, 1999, doi: 10.1038/21092.

[13] A. Leier, C. Richter, W. Banzhaf, and H. Rauhe, “Cryptography with DNA binary strands,” *BioSystems*, 2000, doi: 10.1016/S0303-2647(00)00083-6.

[14] H. J. Shiu, K. L. Ng, J. F. Fang, R. C. T. Lee, and C. H. Huang, “Data hiding methods based upon DNA sequences,” *Inf. Sci. (Ny).*, 2010, doi: 10.1016/j.ins.2010.01.030.

[15] G. Hamed, M. Marey, S. A. El-Sayed, and M. F. Tolba, “Hybrid technique for steganography-based on DNA with n-bits binary coding rule,” *Proceedings of the 2015 7th International Conference of Soft Computing and Pattern Recognition, SoCPaR 2015*, 2016.

[16] A. Khalifa, A. Elhadad, and S. Hamad, “Secure blind data hiding into pseudo dna sequences using playfair ciphering and generic complementary substitution,” *Appl. Math. Inf. Sci.*, 2016, doi: 10.18576/amis/100427.

[17] P. Malathi, M. Manoaj, R. Manoj, V. Raghavan, and R. E. Vinodhini, “Highly Improved DNA Based Steganography,” *Procedia Computer Science*, 2017.

[18] Q. Wang and C. Smith, “Molecular Biology of the Cell (Fifth Edition),” *Biosci. Educ.*, 2008, doi: 10.3108/beej.11.r4.

[19] “The Structure of DNA.” http://ircamera.as.arizona.edu/Astr2016/text/nucleicacid1.htm (accessed Mar. 28, 2021).

[20] D. A. Zebari, H. Haron, and S. R. M. Zeebaree, “Security issues in DNA based on data hiding: A review,” *International Journal of Applied Engineering Research*. 2017.

[21] J. H. Gehani, A., LaBean, T.H., Reif, “DNA-based Cryptography.,” *Proc. 5th DIMACS Work. DNA Based Comput.*, 1999.

[22] M. X. Lu, X. J. Lai, G. Z. Xiao, and L. Qin, “Symmetric-key cryptosystem with DNA technology,” *Sci. China, Ser. F Inf. Sci.*, 2007, doi: 10.1007/s11432-007-0025-6.

[23] X. J. Lai, M. X. Lu, L. Qin, J. S. Han, and X. W. Fang, “Asymmetric encryption and signature method with DNA technology,” *Sci. China, Ser. F Inf. Sci.*, 2010, doi: 10.1007/s11432-010-0063-3.

[24] A. Khalifa, “LSBase: A key encapsulation scheme to improve hybrid crypto-systems using DNA steganography,” *Proceedings - 2013 8th International Conference on Computer Engineering and Systems, ICCES 2013*, 2013.

[25] A. Majumdar, A. Biswas, A. Majumder, S. K. Sood, and K. L. Baishnab, “A novel DNA-inspired encryption strategy for concealing cloud storage,” *Front. Comput. Sci.*, vol. 15, no. 3, pp. 1–18, Jun. 2021, doi: 10.1007/s11704-019-9015-2.

[26] “National Center for Biotechnology Information.” https://www.ncbi.nlm.nih.gov/ (accessed Mar. 28, 2021).

[27] I. Peterson, “Hiding in DNA,” *Muse*, no. 22, 2001.

[28] B. Shimanovsky, J. Feng, and M. Potkonjak, “Hiding data in DNA,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, 2003, doi: 10.1007/3-540-36415-3_24.

[29] A. Atito, A. Khalifa, and S. Z. Rida, “DNA-Based Data Encryption and Hiding Using Playfair and Insertion Techniques,” *J. Commun. Comput. Eng.*, 2011, doi: 10.20454/jcce.2012.242.

[30] A. O. Mohammad Reza Abbasy, Pourya Nikfard and and M. R. N. Torkaman, “DNA Base Data Hiding Algorithm,” *Int. J. New Comput. Archit. Their Appl.*, vol. 2, no. 1, pp. 183–192, 2012.

[31] S. DehiaaKahdum, Ghaith AdillAl-Bawee and M. NsaifJasim, “A Proposed DNA Postfix Hiding Method,” *Int. J. Recent Technol. Eng.*, vol. 8, no. 4, pp. 12047–12051, 2019, doi: 10.35940/ijrte.C5321.118419.

[32] J. S. Taur, H. Y. Lin, H. L. Lee, and C. W. Tao, “Data hiding in DNA sequences based on table lookup substitution,” *Int. J. Innov. Comput. Inf. Control*, 2012.

[33] G. Hamed, M. Marey, S. A. El-Sayed, and M. F. Tolba, “Comparative study for various DNA based steganography techniques with the essential conclusions about the future research,” *Proc. 2016 11th Int. Conf. Comput. Eng. Syst. ICCES 2016*, pp. 220–225, Jan. 2017, doi: 10.1109/ICCES.2016.7822003.