A New Message Encryption Method based on Amino Acid Sequences and Genetic Codes

Ahmed Mahdee Abdo
Computer Science Department
Faculty of Science
University of Zakho
Zakho, Kurdistan, Iraq

Adel Sabry Essa, Abdullah A. Abdullah
Computer Science Department
Faculty of Science
University of Zakho
Duhok, Kurdistan, Iraq

***Abstract*—As the use of technology is increasing rapidly, the amount of shared, sent, and received information is also increasing in the same way. As a result, this necessitates the need for finding techniques that can save and secure the information over the net. There are many methods that have been used to protect the information such as hiding information and encryption. In this study, we propose a new encryption method making use of amino acid and DNA sequences. In addition, several criteria including data size, key size and the probability of cracking are used to evaluate the proposed method. The results show that the performance of the proposed method is better than many common encryption methods, such as RSA in terms of evaluation criteria.**

***Keywords*—Information; secure; encryption**

I. Introduction

As the digital world is growing dramatically, the need for secure and safe information is growing in the same way. There are many ways to keep the data secure such as Cryptography, Stenography or a combination of them [1]. There are two main ways to encrypt a text, either secret-key cryptography or public-key cryptography [2]. The author in [3] proposed an algorithm to encrypt a message based on DNA sequences. The algorithm uses a DNA sequence to encrypt a text using one of the complementary rules. He proposed three complementary rules based on biological and chemical features of each DNA base among each other. The algorithm also made a DNA based code to represent each DNA letter with 2 bits binary number. The final string contains English alphabetic letters that involve the hidden message. The author in [4] use another method to encrypt and hide the data in a DNA sequence. The technique uses a dictionary of codons that is used in our own method. The dictionary codon contains on 64 codons starting from ACT which is numbered 0 to GGG that is numbered 111111. The method converts the message to 8 bits ASCII binary number. Then each 6 bits of the message are represented to three DNA letter based on dictionary codon. In our new proposed algorithm, a new technique is being used in the process of cryptographic taking benefits of amino acid sequences. The rest of the paper will explain the details of the new method. Firstly, it presents a background about the information security. Then it will explore basic of molecular biology. Thereafter, the steps of the encryption and decryption with an example will be explained. Finally, the paper will assess the new method based on some criteria with existing methods.

II. Information Security

There are millions of people who get in touch to each other every day electronically via e-mail, e-commerce, e-banking machine and e-learning [2]. Among these vast amount of communications over the net, the biggest question is to which extend these communications are secure. Transferring information across the world over the Internet leads to the biggest concern related to the security of the transferred information. There are many attacks, such as Crypt analytic and Brute force, could recover the original message when the information are sensitive. Encryption is one of the methods used to secure the information. It has been developed quickly recent years to save and protect transmitted information [5].

A. Encryption

Encryption is a mathematical technique which is related to the security of information sides such as data authentication, data integrity and confidentiality. Encryption is used wildly in many fields especially in sensitive communications such as in wars, military bases and intelligent agency. Cryptography is not just meant to information security, it is rather set of techniques. There are two types of Cryptography which are used widely, public and secret key. Rivest Cipher 5 (RC5) and Advanced Encryption Standard (AES) are the most well-known algorithms based on secret key, while RivestShamirAdleman (RSA) algorithm is the most common for public key [2].

III. Basics of Molecular Biology

The life of an organism has been mapped as a very long sequence called genome. The genome is a repeated of 4 chemical bases called deoxy ribo nucleic acid (DNA). The DNA consists of 4 nucleotides (A, C, G and T). The ribo nucleic acid (RNA) sequence is one of genetic materials. The DNA sequence is convert into mRNA sequence in an operation called transaction. Then the mRNA is translated into amino acid sequence based on genetic codes. There are 20 amino acid letters which construct any protein [6]. The twenty amino acids are (E, P, A, C, G, Q, V, R, K, W, D, N, H, F, L, I, S, T, Y, M).

A. BLUSOM 50

BLOSUM (Blocks Substitution Matrix) are substitution matrices used in Bioinformatics for aligning amino acid sequences to give a score of each alignment. The intersection of

amino acids in the matrix is a spesifc score which represnts to what extend it is welling to interact with other amion acides in the matrix. The higher score in the matrix is when the amion acid interact with itself [7]. BLUSOM50 is a scoring matrix that is used by FASTA and BLAST programs for identifying distant homologous especially with fully length sequence [8].

IV. Proposed Approach

Based on the score among of the amino acids in BLUSOM 50, four amino acids which have less score with other amino acids have been neglected from the proposed algorithm. The amino acids that are not considered are (S, T, Y, M) while the rest sixteen amino acids (E, P, A, C, G, Q, V, R, K, W, D, N, H, F, L, I) have been used. Then each of these amino acids has been given a 4 bits binary number as it is illustrated in Table 1.

TABLE I. Representing Each Amino Acid by 4 bits Binary Number

Amino acid	Code	Amino acid	Code
A	0000	C	0100
R	0001	Q	0101
N	0010	E	0110
D	0011	G	0111
H	1000	F	1100
I	1001	P	1101
L	1010	W	1110
K	1011	V	1111

Two kinds of complementary rules have been applied. A complementary rule means which amino acid is more or less applicable to interact with other amino acids according to the scores taken from BLOSUM50. The higher the score, the higher the probability of interaction. The lower the score, the lower the interaction probability. Accordingly, the first complementary rule is based on the maximum score between a specific amino acid and the rest, and then we take that amino acid as a complementary rule for the proposed approach. The second complementary rule is based on the minimum score between a specific amino acid and the rest, and then we take that amino acid as a complementary as it is shown in Tables 2 and 3.

TABLE II. Rule 1 , The Complementary Rule According to Maximum Value in BLOSUM50

(A, G), (G, A)	(R, Q),(Q, R)	(N, D), (D, N)	(C, V), (V, C)
(E, K), (K, E)	(H, F), (F, H)	(I, L), (L, I)	(P, W), (W, P)

TABLE III. Rule 2 , The complementary Rule According to Minimum Value in BLOSUM50

(A, F), (F, A)	(R, C),(C, R)	(N, L), (L, N)	(K, P), (P, k)
(Q, I), (I, Q)	(E, G), (G, E)	(H, V), (V, H)	(D, W), (W, D)

In addition, a 6 bits binary number has been given for every codon starting from 000000 for TTT to 111111 for GGG as it is revealed in Table 3.

A. Proposed Data Encryption Algorithm: Main Steps

1) Represent each letter in the original message as a binary of 8 bits.
2) Represent each part (4 bits) as a letter based on Table 1.

TABLE IV. Dictionary of Codons, Adopted [4]

Codon	6 Bits number	Codon	6 Bits number
TTT	000000	TTC	010000
TCT	000001	TCC	010001
TAT	000010	TAC	010010
TGT	000011	TGC	010011
CTT	000100	CTC	010100
CCT	000101	CCC	010101
CAT	000110	CAC	010110
CGT	000111	CGC	010111
ATT	001000	ATC	011000
ACT	001001	ACC	011001
AAT	001010	AAC	011010
AGT	001011	AGC	011011
GTT	001100	GTC	011100
GCT	001101	GCC	011101
GAT	001110	GAC	011110
GGT	001111	GGC	011111
TTA	100000	TTG	110000
TCA	100001	TCG	110001
TAA	100010	TAG	110010
TGA	100011	TGG	110011
CTA	100100	CTG	110100
CCA	100101	CCG	110101
CAA	100110	CAG	110110
CGA	100111	CGG	110111
ATA	101000	ATG	111000
ACA	101001	ACG	111001
AAA	101010	AAG	111010
AGA	101011	AGG	111011
GTA	101100	GTG	111100
GCA	101101	GCG	111101
GAA	101110	GAG	111110
GGA	101111	GGG	111111

3) Apply one of the complementary rule (either rule one or two) on the letter) based on Table 2.
4) Take the first occurrence position of the 16 amino acids in any sequence as indexes.
5) Represent the decimal index numbers to 8 bits binary numbers.
6) Finally, represent each 6 bits binary numbers represented to DNA letters according to Table 3. Extra zeros are added to make the length of binary number 6 or multiply of 6.

B. An Example to Encrypt a Message

Let consider our message is Hi, and the amino acid sequence is MPQVKLWLSGIQICLQSNQLAPLIRELQKDSTASFHFI EGEVECGPGPGIEGIFEGP

1) Represent the message in 8bits binary: 01001000, 01101001
2) Represent each part (4 bits) as a letter based on Table 1: C, H, E, I
3) Apply one of the complementary rule (1) based on Table 2: V, F, K, L
4) The first occurrence of V F K L in the sequence is 3, 34, 4, 5 respectively.
5) Represent the decimal index numbers to 8 bits binary: 00000011, 00100010 00000100, 00000101
6) Represent each 6 bits binary numbers represented to DNA letters according to Table 3. Not: extra zero added to make all of them 6bits 000000 = TTT, 110010 =TAG, 001000 = ATT, 000100 = CTT, 000001= TCT, 010000 = TTC The faked DNA that hold the secret message is TTTTAGATTCTTTCTTTC

C. Proposed Data Decryption Algorithm: Main Steps

1) Represent each three DNA letters of the encrypted message to 6 bits binary numbers according to Table 3.
2) Take each 8bits and then convert them decimal numbers. Ignore the mod of 8.
3) Take these decimal numbers as indexes for the key and retrieve the represented value for each index in the key.
4) Apply one of the complementary rule (either rule one or two) on the letter created in previous step, so each letter will be represented to another letter of amino acids based on Table 2.
5) Represent each letter to 4bits binary number according to Table 1.
6) Represent each 8bits binary number to ASCII numbers and letters to get your original message.

D. An Example to Decrypt a Message

The faked DNA is: TTTTAGATTCTTTCTTTC The amino acid sequence: MPQVKLWLSGIQICLQSNQLAPLIRELQKD STASFHFIEGEVECGPGPGIEGIFEGP

1) Represent each three DNA letters of the faked DNA to 6 bits binary numbers according to Table 3: TTT = 000000, TAG= 110010, ATT= 001000, CTT= 000100, TCT= 000001, TTC= 010000.
2) Represent every 8bits to decimal number: 3, 34, 4, 5 Ignore the mode of 8 (here is the last 4 zeros).
3) Take these decimal numbers as indexes for the key and retrieve the represented value for each index in the key: V, F, K, L.
4) Apply complementary rule (1): C, H, E, I
5) Represent each letter to 4bits binary number according to Table 1: 0100, 1000, 0110, 1001
6) Represent each 8bits binary number to ASCII numbers and letters to get your original message: 01001000 = H 01101001 = i The message is Hi

V. Results and Comparative Analysis

The results show high performances in the proposed method comparing with other encryption techniques. There are many criteria that can be used to evaluate any encryption method or algorithm. These terms are key size, data size, security and time complexity [9]. Here, we present a comparative analysis between our proposed algorithm and well-known encryption method based on these criteria.

A. Key Size

Key size means the length of the key used in bits or letters to encrypt a message by any technique. Every method of encryption has its own key size [10]. In the asymmetric RSA cipher, the amount of key size is not limited to any number. It could be any length as it depends on the amount of information to be encrypted. The more the information, the longer the key, the more complexity time. AES is another method which has three options in terms of key size, either 256 or 192 or 128. This increase the security level of this technique as the key size has a positive correlation with security level [1]. In our proposed algorithm, the key size is composed of 16 amino acid letters. These letters are extracted from any amino acid sequence with their indexes to be used for the encryption process. The amino acid sequence could be in any length and the 16 letter could be in any position. Consequently, this increase the security level of this method to a large extend as there are hundreds of thousands of amino acid sequences.

B. Data Size

The amount of data to be encrypted which is measured by either bytes or kilobytes is called data size. The data size in some cases could be a weak point for some encryption methods. Some methods has restriction size on the amount of data to be encrypted. While others such as vigenere cipher method, the big amount of data led to repetition when the key size is small. In RSA cipher case, there is a positive relation between the key size and the data size. Each amount of data should has an appropriate key size [1] . In this new method, the data size has no limitation. In addition, it has no relation with key size. Accordingly, each of the key and the data are separated from each other.

C. Security

The security in the field of encryption means the probability of cracking. Every encryption technique has its own security level by calculating the cracking probability. The security level of the Vigenere cipher method is very weak as the key is repeated, so the cracker can guess the key length of the key using some statistical methods and ultimately decipher the text. On the other hand, the RSA cipher security level relies heavily on the key size. If the chosen key is small, the text could be cracked easily. In the AES method, the time required to crack it, depends on the key length used to cipher the text. As AES use long key size, it makes its method better than other advanced symmetric ciphers and cannot be cracked only by using the brute force cryptanalysis. In our proposed method, the key is an amino acid sequence in any length. The indexes of the first occurrence of the 16 amino acids are kept to be used later. Accordingly, the security level of the new encryption method depends on how many amino acid sequences are there?. There are many official databases that hold amino acid sequences. The Universal Protein Resource (UniProt) is one of the main amino acid and protein sequences recourse over the world. According to last release on March 2017, the database of the UniProt contains on 80204459 amino acid sequences [11]. So the security level or the probability of cracking can be calculated as it is illustrated below. The probability of cracking will be =1/280204459 . We multiplied it by 2 as we use 2 complementary rules. In addition, this is not the end, this number (80204459) is grown dramatically and this will increase the security level of the method

D. Time Complexity

Time complexity means the required time needed to encrypt a message. The time depends heavily on the steps of a specific algorithm. In our algorithm, there are 6 steps to encrypt a message. The more advanced ciphering methods such as AES and RSA shows more executing time than simpler methods such as DNA-based PCF.

We used Perl programming language to implement our method. gmtime() is used in Perl to print the executing time. It shows very fast performance to encrypt even a long message.

All the criteria above have been compared and elaborated in Table 5 which compare the proposed method with common once

TABLE V. COMPARATIVE ANALYSIS

Parameters	DNA-based PCF	RSA	Proposed method
Key size	English letters of maximum length 25	Large sizes lead to extensive computational efforts	16 amino acids letters.
Data size	Any size	Large sizes needs large key size	Any size
Time complexity	Low	The highest	Low
Security	Its cracking probability is very low	Can be cracked for small key lengths	Very low and explained in security point.

VI. CONCLUSION

To conclude, it is compulsory to keep the information safe. Using DNA and amino acids sequences as mediums to encrypt a message took much attention of the researchers. There are many algorithms and methods have been designed to encrypt a message using different kind of medium. Every cryptography technique has its own features in term of performance. Some of them are performing well in terms of key size and might fail in data size, while other present good result in time complexity meanwhile has high probability to crack the message.

Our proposed method has no restrictions in data size and the key size. In addition, its probability to crack it down is very low comparing with other methods. Moreover, this probability is decreased as long as UniProt get updated. In terms of time, as there are just six steps to encrypt any message, it shows high speed in terms of time complexity. Consequently, the new proposed algorithm has satisfied most encryption criteria to a large extend the used to evaluate and ciphering method. This method could be more improved by decreases the length of the encrypted message.

ACKNOWLEDGMENT

Thanks to everyone helped me to finish this work.

REFERENCES

[1] S. Marwan, A. Shawish, and K. Nagaty, "Dna-based cryptographic methods for data hiding in dna media," *Biosystems*, vol. 150, pp. 110–118, 2016.

[2] V. Kamalakannan and S. Tamilselvan, "Security enhancement of text message based on matrix approach using elliptical curve cryptosystem," *Procedia Materials Science*, vol. 10, pp. 489–496, 2015.

[3] K. Menaka, "Message encryption using dna sequences," in *Computing and Communication Technologies (WCCCT), 2014 World Congress on*. IEEE, 2014, pp. 182–184.

[4] R. Agrawal, M. Srivastava, and A. Sharma, "Data hiding using dictionary based substitution method in dna sequences," in *Industrial and Information Systems (ICIIS), 2014 9th International Conference on*. IEEE, 2014, pp. 1–6.

[5] A. Joshi, M. Wazid, and R. Goudar, "An efficient cryptographic scheme for text message protection against brute force and cryptanalytic attacks," *Procedia Computer Science*, vol. 48, pp. 360–366, 2015.

[6] V. Mathura and P. Kangueane, *Bioinformatics: a concept-based introduction*. Springer Science & Business Media, 2008.

[7] S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks," *Proceedings of the National Academy of Sciences*, vol. 89, no. 22, pp. 10 915–10 919, 1992.

[8] W. R. Pearson, "Selecting the right similarity-scoring matrix," *Current protocols in bioinformatics*, vol. 43, no. 1, pp. 3–5, 2013.

[9] D. A. A. G. Singh and R. Priyadharshini, "Performance analysis of data encryption algorithms for secure data transmission," *International Journal for Science and Advance Research In Technology*, vol. 2, no. 12, 2016.

[10] W. Stallings, *Cryptography and network security: principles and practice*. Pearson Upper Saddle River, NJ, 2017.

[11] U. Consortium, "Uniprot: the universal protein knowledgebase," *Nucleic acids research*, vol. 45, no. D1, pp. D158–D169, 2016.