Constructing Hierarchical DNA Clustering Model (HDCM) To Cryptanalyze Feedback Shift Register-Based Stream Cipher

Dr.Basim S. Yaseen
Department of Computer Sciences/Shatt Al-arab University College
Email: basimsahar@sa-uc.edu.iq

Abstract: When evaluating and analyzing encryption techniques and algorithms, cryptanalysis is a fundamental scientific field on which cybersecurity depends. In the past few decades, forensic science has been enhanced by adding DNA technology, which has brought powerful capabilities. Hierarchical DNA clustering is recursively clustering DNA points into many clusters. A genetic keystream can be represented by a binary, a triple, or a tree, depending on the number of genetic bases of clusters. Many encryption systems and algorithms rely on the shift register's physical component and the avalanche concept to ensure security. This makes it difficult for attackers to dismantle the components of the shift registers and attack them individually. Experts and stakeholders consider these components' increasing complexity and large size as a strength factor for these systems. However, a proposed model challenges the cohesion of these systems through the principle of fragmentation by attack. It works by fragmenting and attacking the shift registers and then reducing the initial values they adopt to produce the final key sequence. The principle of reduction depends on the concept of clustering, which involves creating clusters of initial key values whose contents are interrupted and reduced at each level of the binary tree created for the key sequence's genetic bases. The model involves two specific processes - a divide-and-conquer approach and a DNA binary tree clustering process - which significantly reduces the solution space and searches for the initials of LFSRs and NLFSRs. Our technique requires approximately $C(2^{\wedge n})$ to attack, where C is a constant, and n is the largest length of either LFSRs or NLFSRs. The first process includes splitting the shift registers individually and classifying their outputs based on the pre-calculated DNA sequence for the generator's outputs. The second process creates classification clusters (Test Tubes) for the initial values of the shift registers using the genetic bases that make up the final key sequence (Z_i). The proposed model aims to disrupt the cohesion of these systems by fragmenting and attacking the shift registers.

The novelty and contributions

- discovering a mathematical or logical base model that can be expressed as a theory. This theory can then counteract and weaken the coherence principle in stream cipher generators that rely on displacement registers. By doing so, we can effectively attack this family of generators such as the Global System for Mobile Communications (GSM) algorithms.
- Identifying hidden vulnerabilities that can be exploited to attack these generators will help manufacturing experts and users become aware of these weaknesses and take necessary precautions.

Keywords: DNA computing, Search Tree, Hierarchal Clustering, Feedback Shift Register, Stream cipher.

1 Literature Review

Over the past decade, numerous studies have been conducted in the field of using biotechnology to cryptanalysis and break cipher algorithms, the field of DNA-based cryptanalysis. Studies in this field are categorized into two main parts. The first part involves the use of traditional analysis techniques to analyze and attack ciphers that use DNA coding to produce ciphertext. The second part is considered to be the most important and involves analysis based on DNA. This type of analysis uses DNA calculation models and biological process simulation in the laboratory. The most significant works in this area are listed in Table 1 for reference.

TABLE.1: The most significant works in the area of DNA-based Cryptanalysis.

Ref.	Type of Cryptanalysis	Summarized
[1]	conventional cryptanalysis technique to cryptanalysis DNA coding	This paper presents an algorithm for cryptanalysis of an encrypting image using a 2D Hénon-Sine map and DNA coding approach. However, it has been found that the algorithm is not as secure as it was originally claimed to be. The encryption scheme uses a permutation-diffusion architecture, with DNA random coding and exclusive OR being used for image diffusion. Pixel-swapping operations are used for image scrambling.
[2]	conventional cryptanalysis technique to cryptanalysis DNA coding	The paper investigates a new image encryption scheme that uses a Feistel network and dynamic DNA encoding. The encryption scheme uses four encryption steps to encrypt plain images, which include generating chaotic sequences, Hill encryption, Feistel network, and Pixel diffusion. However, the paper identifies some analysis issues with the secret key design and encryption process of this encryption scheme. After analyzing these problems, the paper proposes necessary improvements to the encryption scheme and introduces the corresponding chosen plaintext attack algorithm.
[3]	DNA-Based Cryptanalysis	This paper explains how the logic of cipher bits, sequence, keystream, and plain text bits is transformed from propositional logic to DNA logic and executed in polynomial time.
[4]	DNA-Based Cryptanalysis	The paper suggests a sticker DNA model to cryptanalyze the cipher created by the keystream of linear and nonlinear feedback shift registers. The model is based on the principle of establishing a binary sequence as a memory strand that represents the potential plain text sequence. Subsequently, it creates all possible paths to find the correct solution by linking the stickers to the components of the solution paths that represent the key parts.
[5]	DNA-Based Cryptanalysis	The study proposes a technique that combines the GA and the DNA sticker model to perform a parallel search and attack a cryptosystem. The first step involves creating a database of all possible solution paths. Then, the technique searches for the correct path using the proposed combination of algorithms.
[6]	DNA-Based Cryptanalysis	The paper proposes a modified digital simulation of the DNA sticker model technique combined with another technique to attack linear and nonlinear feedback shift register generators.
[7]	DNA-Based Cryptanalysis	The paper introduces a new method called the DNA sticker model for cryptanalysis of a stream cipher that uses a key stream sequence generated from a linear shift register. The cipher sequence is decrypted by applying sticker operations at the binary level.
[8]	DNA-Based Cryptanalysis	The paper suggests the creation of a software computer that uses genetic operations based on a splicing DNA model. It also uses a probabilistic model of the English letter frequency in a vertical alignment. This software aims to generate genetic bases of the strands, which represent the letters of a natural language. The cohesion of these bases may depend on the frequency of occurrence of these letters in the plain text or key.

[9]	DNA-Based Cryptanalysis	The paper introduces a DNA splicing model that can cryptanalyze a stream cipher sequence generated by an unknown source. However, the cipher is known to be encoding for the plaintext. The proposed model utilizes the statistical properties of the plaintext along with the random properties of the key string segments.

2 Introduction

To date, there is no molecule more suitable for IT applications than DNA [10]. Nucleic acids possess the fundamental property of the Watson-Crick [11] rule, Figure 1, depicts the principles of the new rule, which is the basis for many DNA manipulation techniques, allowing for a unique range of potential applications for computational purposes across disciplines. These processing techniques were built on the biologically sophisticated properties of the DNA molecule and were originally developed for the life sciences but have since been gradually reused in computer applications. Late in the last century, this technology came into use to solve difficult cryptographic problems. It can be solved by traditional methods [12].

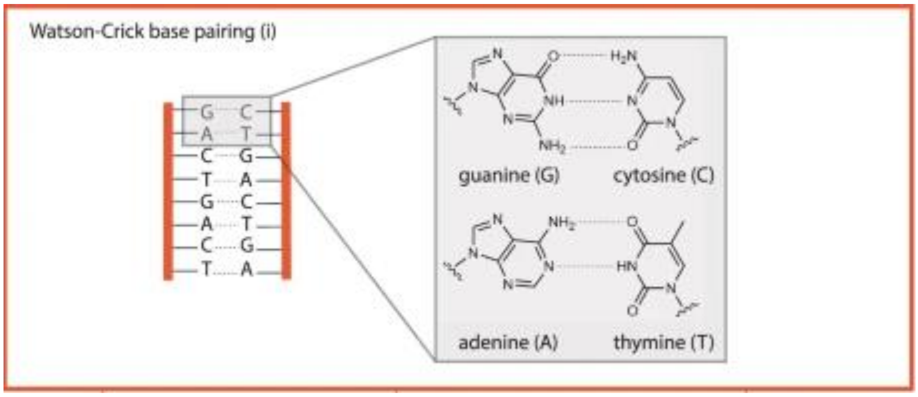
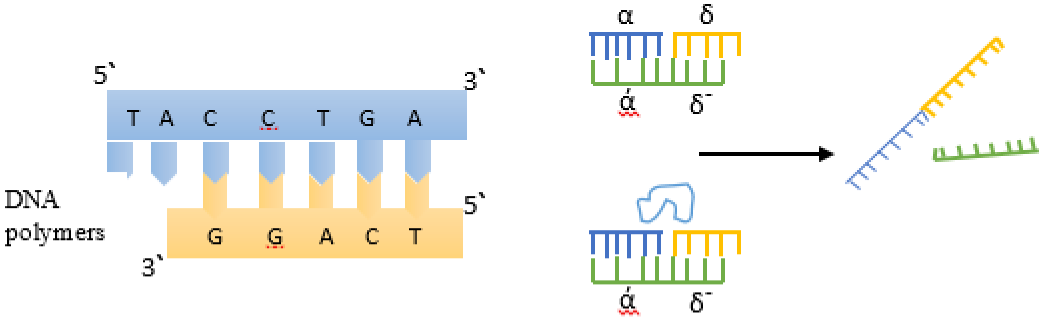


Figure 1: the principles of the Watson-Crick rule of the genetics.

The idea behind DNA-based cryptanalysis [13][14] is to use biomolecular components, the processes performed on them in the laboratory, the biocomputational models derived from them, and the test tube processes performed on them in the laboratory [15], to accomplish complex and difficult computing tasks, instead of electronic and silicon components that were unable to accomplish them. These are the tasks that we use in traditional computing. Although the computational models proposed and the processes exploited by DNA computing [16] vary, there are fundamental processes that have been used in test tubes in the biological laboratory and are used in every biocomputational technique invented to solve a problem, and these processes are annealing, Ligase, melting, is-empty, is-full, ... Figure 2 shows the processes of annealing and ligase that take place in the laboratory and that can be performed on the digital representation of the genetic bases.



Annealing of primer creates a short stretch of double-stranded DNA

Ligase Joins 5¯ phosphate to 3¯ hydroxyl

Figure 2: Annealing and Ligating operations by polymers

2 The HDCM

2.1 Methodology

For the attack to be successful, the key sequence needs to be transformed into two or more sequences, based on the subsequences of the final keystream, of genetic bases using Table 1. The attack consists of two stages. During the first stage, the stream encryption generator is divided and the shift registers are dealt with individually. A series of outputs are produced from each register, and the output length of all the initial values of the FSR is proportional to the length of the stream key sequence prepared at the start of the attack. In the second stage, the initial values are isolated based on whether they achieve the genetic base at the location that matches the initial output location. For each location, two genetic bases are present, representing the final key sequence. These are linked to two clusters of initial values of the shift registers, which are classified based on their fulfillment of each of the key sequence bases Figure 3 depicts the classification of FSR initials based on two DNA key sequences. During the second phase of the attack, an iterative process is executed where each cluster of a particular level is intersected by clusters that belong to the next level. The lowest level in the binary tree built for the clusters is considered in this case.

2.2 Processes

The proposed model focuses on attacking individual registers of the stream key generator in the first stage of the work. This approach differs from known attack methods that target registers together and surpasses the avalanche characteristic of the stream cipher. A specific set of keystream binaries must be obtained to fulfill a critical work requirement. There are different methods to obtain these binaries, such as guessing or predicting the letters of the plaintext. The number of bases needed is proportional to the number of keystream components required to transform the key stream into a series of genetic bases. These bases are essential to produce a genetic base sequence from the key stream outputs.

We are constructing the Hieratical DNA Clustering Model (HDCM) by using Test Tube operations.

Let:

- FSR_1, FSR_2,…, and FSR_n is a feedback shift that registers components of the stream key generator.
- Z_i is a final key sequence. $Z_i = FSR_{1i}$ xor FSR_{2i} ,…, xor FSR_{nj} ,where i=1,…,j.
- For simplicity, n=2, so have only FSR_1 and FSR_2 and type of the tree is binary tree.
- DNA codes

Table 2, Outlines the bases, type of tree, and corresponding codons required for two FSRs.

Genetic Bases	FSR_1	FSR_2	Z_i	Type of the tree
A	0	0	0	Binary Tree
T	1	1	0	
C	0	1	1	
G	1	0	1	

- The Z_i sequence that is produced via Table 2,
  $Z_{GB(1,2)1}$, $Z_{GB(1,2)2}$, $Z_{GB(1,2)3}$, $Z_{GB(1,2)4}$, $Z_{GB(1,2)5}$, $Z_{GB(1,2)6}$.
  When Z_i =0 → Z_i is $Z_{GB(A,T)i}$
  Z_i =0 → Z_i is $ZGB_{(C,T)i}$

Algorithm Building and searching within a Hierarchical DNA Binary Tree.
Input: Solution Spaces values of FSR1 and FSR2, Genetic Bases of Keystream ($Z_{GB(\_,\_)i}$).
Output: Correct FSR initial values.
Steps:

```
(* Divide-and Conquer phase *)
 For all Solution Space Values of $FSR_1$
    For i:1 → 6
        IF initial=$Z_{GB(1,\_)i}$ → initial → $TT_{CLUSTER1i}$
        IF initial=$Z_{GB(\_,2)i}$ → initial → $TT_{CLUSTER2i}$
    Loop
Loop
For all Solution Space Values of $FSR_2$
    For i:1 → 6
        IF initial=$Z_{GB(1,\_)i}$ → initial → $TT_{cluster1i}$
        IF initial=$Z_{GB(\_,2)i}$ → initial → $TT_{cluster2i}$
     Loop
Loop
(*Phase of Building the Binary Tree and Searching within it *)
So, $TT_{cluster2i}$ is the $\overline{TT}_{cluster1i}$ (inverse)
$TT_{universal} = TT_{FSR1}. \cup TT_{FSR2}.$
$TT_{universal} = TT_{cluster11} + TT_{cluster21}$, ($TT_{universal}$ the root of the binary tree)
For i:1 → 6
For j:1 → no.of FSR initials
        $TT_{cluster1i+1} = TT_{cluster1i-1} \cap TT_{cluster1i}$ (all initials j that satisfy GB1i)
        $TT_{cluster2i+1} = TT_{cluster2i-1} \cap TT_{cluster2i}$ (all initials j that satisfy GB2i)
Loop
   For j:1→ $2^i$
         IF $TT_{cluster1i+1}$ is-empty → remove $TT_{cluster1i+1}$
         IF $TT_{cluster2i+1}$ is-empty → remove $TT_{cluster2i+1}$
   Loop
Loop
For j:1→ $2^i$
         IF $TT_{cluster1i+1}$ is not is-empty → $TT_{cluster1i+1}$ is a solution.
         IF $TT_{cluster2i+1}$ is not is-empty → $TT_{cluster2i+1}$ is a solution.
 Loop
End of the algorithm
```

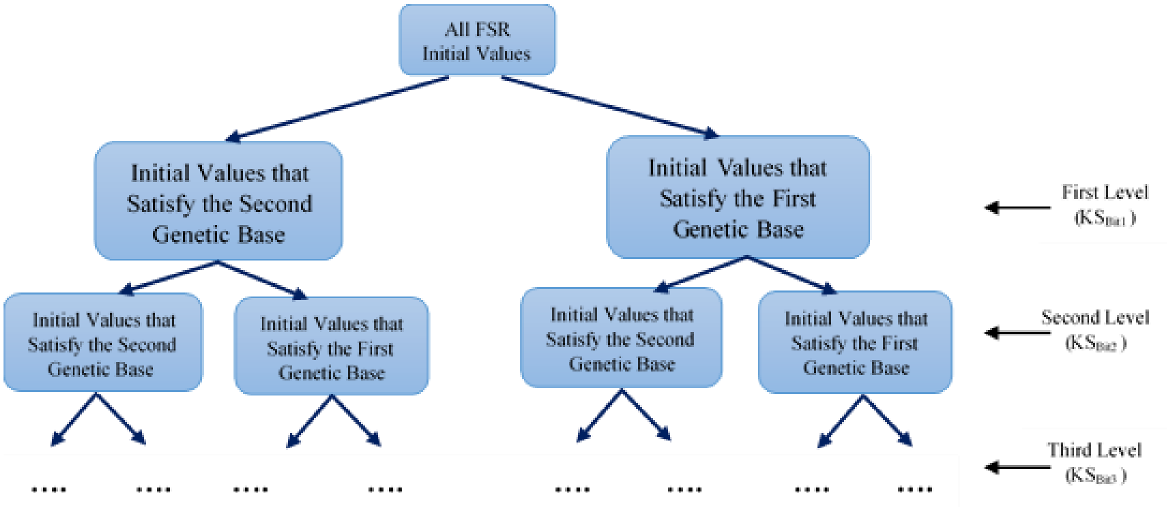

All FSR Initial Values
Initial Values that Satisfy the Second Genetic Base
Initial Values that Satisfy the First Genetic Base
Initial Values that Satisfy the Second Genetic Base
Initial Values that Satisfy the First Genetic Base
Initial Values that Satisfy the Second Genetic Base
Initial Values that Satisfy the First Genetic Base
First Level (KS_{Bit1})
Second Level (KS_{Bit2})
Third Level (KS_{Bit3})
....

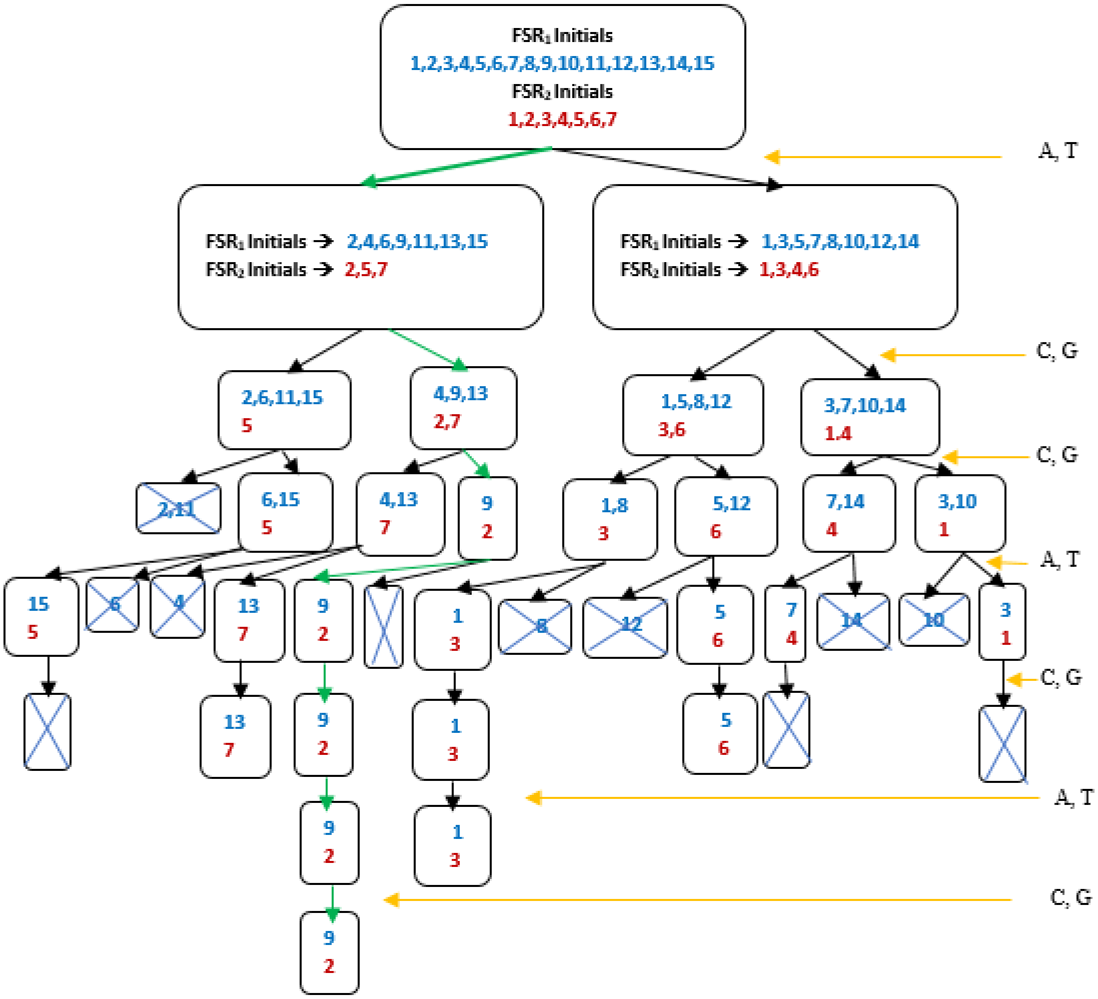
Figure 3 depicts the classification and clustering of the Hierarchical DNA Binary Tree algorithm

2.3 Practice

Let us demonstrate how the proposed algorithm works by providing further clarification on its stages, and explaining how it can be used to arrive at the correct solution for attacking. Additionally, we will compare its time and storage complexity to similar methods of attacking. We will use a simple example to apply the concepts and processes that the algorithm has come up with.

Generator's description: No. of LFSR=2, Length of LFSRs are 4,3, Feedback of LFSRs = (1,4)(1,3) ,LFSRs output =feedback bits , Initials of LFSRs= 1001,010.

Keystream$_{10\ bits}$→0 1 1 0 1 0 1 0 0 1, Sequence of Genetic Bases →A, T - C, G - C, G - A, T - C, G - A, T - C, G - A, T - A, T - C, G. (Remember that, the real GB sequence is A-C-C-T-G-T-G-A-T-C). Figure 4 displays the implementation output based on the algorithm steps for both phases. The green paths are the genetic base branches that lead to the correct solution. The rectangles contain the solutions that match the genetic bases, and they disappear once they are empty. However, the rectangle with the correct solution will remain until the last step of the two genetic bases gathering. The correct solution is 9,2, with the actual GBs sequence A, C, C, T, G, T, G, A, T, C, which represents the keystream 0, 1, 1, 0, 1, 0, 1, 0, 0, 1.



The path that sets the true initial values (9,2) is A,C,C,T,G,T,G. As an example, we need 7 KS bits to determine the exactly one correct value of the initial

Figure 4: the implementation of the proposed Model

1 Discussion and Conclusion

The effectiveness of the biocomputational model relies on the number of shift registers utilized in creating the final key sequence. As the number of shift registers increases, the task of disassembling the sequence of test tubes along with the logical and mathematical operations required to be performed, becomes more complex. Additionally, the search tree also becomes more complicated. The attack on the generator is not focused on attacking the set of registers together. Instead, it involves searching within the initial values of each component. With each iteration, the solution space is reduced by a large percentage until the incorrect cases are eliminated and only the correct solution remains. The complexity of the work is directly proportional to the components of the generators. However, the proposed model offers a path that resembles an unconventional model dealing with the cohesion of stream cipher components and the resistance of its parts to disassembly. The model follows an attacking principle based on disassembling, which significantly reduces the time required to attack this type of encryption. Table 3, represents a comparison between the proposed model and some of the works summarized in paragraph number 1 by cryptanalysis stream cipher.

Table 3: a comparison between the proposed model and some of the works

Reference	The Method	Target Part of the Cryptanalysis	Evaluated Complexity
[3]	DNA logic operations	Key Stream (Z_i)	It depends on how long Z_i is processed
[4]	Sticker DNA model	Combined FSRs	Adopting a parallel search
[5]	GA with Sticker DNA	Combined FSRs	Adopting a parallel search
[6]	Sticker DNA model	Combined FSRs	=
[7]	Sticker DNA model	Combined FSRs	=
[8]	Splicing DNA model	Combined FSRs	=
[9]	Splicing DNA model	Combined FSRs	=
Proposed model	Test Tube operations	Individual FSRs	Adopting a parallel search and partitioning the generator.

The importance of the proposed model lies not only in what it has achieved but also in its ability to cryptanalyze a single FSR of a large length and modern block cipher cryptanalysis. Our future work aims to develop a model to address these aspects.

REFERENCES

[1] Junxin, Lei Chen, Yicong Zhou" Cryptanalysis of a DNA-based image encryption scheme", Elsevier, Information Sciences, Vol.520, Pp:130-141, May 2020.

[2] Wei Feng, Zhentao Qin, Musheer Ahmad" Cryptanalysis and improvement of the image encryption scheme based on Feistel network and dynamic DNA encoding ", IEEE-2021.

[3] A. S. Polenov: The Computing of NP-Complete Problems in Polynomial Time Using DNA - Logic World Applied Sciences Journal, IDOSI Publications 30(9), Pp:1188-1192, (2014).

[4] S. B. Sadkhan, B. S. Yaseen," A DNA-Sticker Algorithm for Cryptanalysis LFSRs and NLFSRs Based Stream Cipher", International Conference on Advanced Science and Engineering, October 9-11 2018, (IEEE-ICOASE2018), University of Zakho - Duhok Polytechnic 12University, and Submitted to the IEEE Xplore Digital Library, (2018).

[5] S. B. Sadkhan, B. S. Yaseen," DB Based DNA Computer to Attack Stream Cipher", International Conference, IEEE-ICECCPCE2019, In Mosul and Erbil,13-14 February, (2019).

[6] S. B. Sadkhan, B. S. Yaseen," Hybrid Method to Implement a Parallel Search of the Cryptosystem Keys", International Conference on Advanced Science and Engineering, April 2-3 2019, (IEEE, Springer-ICOASE2019), University of Zakho - Duhok Polytechnic University, and Submitted to the IEEE Xplore Digital Library, (2019).

[7] Noora Amir Abdulmehdi, Sahar Adel Kadum,' Cryptanalysis Using DNA-Sticker Algorithm', Iraqi Academics Syndicate International Conference for Pure and Applied Sciences: Conference Series 1818 (2021) 012088, Babylon-Iraq.

[8] B.S. Yaseen,’ Cryptanalysis of OTP Cipher Using Probabilistic DNA Computer’, Design Engineering (Toronto), Issue:8, PP:10739-10748,2021.

[9] B.S.Y,’ Splicing DNA Model for Unknown Stream Cipher Cryptanalysis’, IEEE Conferences,2021 2nd Information Technology to Enhance E-learning and Other Application (IT-ELA),2022.

[10] Andrew Travers and Georgi Muskhelishivli” DNA structure and function” the FEBS Journal, doi:10.1111/febs.13307, 2015.

[11] Leslie A. Pray” Discovery of DNA structure and function: Watson and Crick”, Nature Education 1(1):100,2018.

[12] Ashish Kumar Kendhe, Hamani Agranwel” A survey Report and various cryptanalysis techniques”, International Journal of soft computing and engineering (IJSCE), Vol.3, ISSUE-2, May 2013.

[13] Gambhir Singh, Rakesh Kumar Yadar” DNA based cryptography techniques with applications and limitations”, International Journal of Engineering and Advanced Technology (IJEAT), online, Vol.-8, ISSUE-6, Aug. 2019.

[14] Sattar B. Sadkhan, Bassim S. Yaseen” DNA-based cryptanalysis: challenges, and future trends”2019 2nd Scientific Conference of Computer Science (SCCS), IEEE-explorer,2019.

[15] Stove Minchin, Julia Lodge” Understanding biochemistry: structure and function of nucleic acids”, Essays Biochem, online 63(4):433-456, Oct.2019.

[16] G. Rozenberg, A. Salomaa” DNA computing: new Ideas and paradigms”, International Colloquium, computer science, doi: 10.1007/3-540-48523-6_9.corpus, ID:20317046, July 1999.