Ministry of Higher Education
And Scientific Research
University of Diyala
College of science
Department of computer science

Hybrid RC4-DNA algorithms

Research submitted to Diyala University / College of Science /
Department of Computer Science
As one of the requirements for obtaining a Bachelor's degree
in Computer Science

By
Fahmi Sabeeh Fahmi
Ali Hatem Rasheed
Faruq Jamaal Hussain

Supervised By
Asst. Lecturer Adel Abdul-Jabbar Mohamed

2020A.D 1441A.H

جمهورية العراق
وزارة التعليم العالي والبحث العلمي
جامعة ديالى/كلية العلوم
قسم علوم الحاسوب

خوارزمياتRC4-DNA الهجينة

بحث مقدم الى جامعة ديالى/كلية العلوم/قسم علوم الحاسوب

كأحدى متطلبات لنيل شهادة البكالوريوس في اختصاص علوم الحاسوب

أعداد

فهمي صبيح فهمي

علي حاتم رشيد

فاروق جمال حسين

بأشراف

م.م. عادل عبد الجبار محمد

1441 هـ

2020م

Abstract

The cryptography algorithms play a key role in the security of information which apply complex logics and mathematical procedures to develop efficient encryption algorithms. It can also be defined as an art that allows people to hide their personal information in the world of electronics. DNA cryptography is very current state of the art and a new upcoming domain that is emerging based on DNA computing, to encrypt big message in compact volume. In this paper, a cipher algorithm is proposed using biological processes and arithmetic operations. In this research we will work on linking the DNA algorithm with the RC4 algorithm in order to increase the random output of the two algorithms, ie, we will strengthen these two algorithms through the binding process. We will rely on 100 key test to encrypt the data and generate one million bits. Random test these bits by NIST program and compare them with algorithm output The original RC4.

الخلاصة

تلعب خوارزميات التشفير دورًا رئيسيًا في أمن المعلومات التي تطبق المنطق المعقد والإجراءات الرياضية لتطوير خوارزميات التشفير الفعالة. يمكن تعريفه أيضًا على أنه فن يسمح للناس بإخفاء معلوماتهم الشخصية في عالم الإلكترونيات. تشفير DNA هو حالة حديثة للغاية ومجال جديد قادم يظهر بناءً على حوسبة DNA ، لتشفير الرسالة الكبيرة في الحجم الصغير. في هذه الورقة ، تم اقتراح خوارزمية تشفير باستخدام العمليات البيولوجية والعمليات الحسابية. سنعمل في هذا البحث على ربط خوارزمية DNA بخوارزمية RC4 من أجل زيادة الإنتاج العشوائي للخوارزميات ، أي أننا سنقوي هاتين الخوارزميتين من خلال عملية الربط. سنعتمد على 100 اختبار رئيسي لتشفير البيانات وإنشاء مليون بت. اختبار عشوائي لهذه البتات بواسطة برنامج نيست ومقارنتها مع إخراج الخوارزمية RC4 الأصلي.

بسم الله الرحمن الرحيم

((أُولَئِكَ الَّذِينَ امْتَحَنَ اللَّهُ قُلُوبَهُمْ لِلتَّقْوَى لَهُمْ مَغْفِرَةٌ وَأَجْرٌ عَظِيمٌ))

صدق الله العظيم

سورة
الحجرات
الاية (3)

الأهداء

الى اهلنا الذين وقفوا معنا في كل حياتنا بكل الحب

للإنسان الذي علمنا قيمه الموقف و المبادئ في الحياه

الى المدرسة التي علمتنا اول خطوات النجاح

الى الذين وقفوا معنا ويتأملون نجاحنا

في هذا البحث المتواضع

الشكر والعرفان

الحمد لله الذي جعل الشكر و السلام و الرحمة مفتاحا لذكرة و تقبيل عظمته و اجل و الصلاة و السلام على خير خلقه و اعز عبادة و على اله و صحبه اجمعين اما بعد:

يطيب لنا بعد التوفيق من الله عز وجل على تمام بحثنا ان نتقدم بوافر الشكر والامتنان الى المشرف العزيز الأستاذ (عادل عبد الجبار) الذي لم يدخر جهدا في إسداء النصح والتوجيه الرشيد ، ولم يتوان في تقديم المساعدة فجزاه الله عنا خير جزاء على رعايته و دعمه ومساندته طيلة فترة البحث هو الذي كان له الاثر الكبير في تخفيف الصعوبات عنا وكذلك نجد من الواجب ان نتقدم بخالص الشكر والتقديرالى الهيئة التدريسية في القسم .

واخيرا وليس اخرا يدوم الصمت طويلا و التأثر بالكلمات في ألسنتنا حيث نتوجه بالشكر الى عوائلنا وفاء لهم واكراما..

Table of Contents

LIST OF FIGURES

LIST OF TABLES

Chapter One

Introduction

1.1 Introduction

Encryption is a process or mechanism that protects information by preserving it from any threat of privacy and tampering. It calls for the need to establish encryption and data protection techniques. Encryption has existed for a long time before the advent of the computer where they were allied kings during the battle or when sending an important message, the code of Caesar and the encryption algorithm was very simple. It changed the character to be encoded with the third letter that followed. This code is now easy to break and is completely unsafe. It is believed that Julius Caesar The first to use this method and that was between 58 BC until 51 BC. There are many cipher algorithms .There are many codes and algorithms. With the emergence of many means of storage and the use of the Internet in the transfer of information and use almost exclusively in banking transactions, and the trade here is necessary to secure communication and protection, and thus devised many techniques and methods of encryption to maintain the confidentiality of information and maintain safety of importance, and these reasons are sufficient to create and find Methods and algorithms for encrypting and archiving information. With the spread of hackers is a mysterious person with unspecified trends and possesses computer skills and information security. Hacker's speech was directed at a group of clever programmers who were trying to penetrate different systems and did not necessarily commit a crime, But their success in breaking through is a success for their abilities and skills. However, the law considers them from outside who have been able to enter places where they should not be, and to do so as an optional process in which the programmer tests his capabilities without knowing his or her real name, but some use it criminally to sabotage the BI, while others use it commercially for espionage purposes to steal money. Thus, large computer companies hired some of those hackers who have high salaries to penetrate their various systems. Look for weaknesses and gaps in their systems and suggest ways to improve the system, repair damage and close gaps that may occur due to system attackers. Here we must clarify this and mentioned that there are two types of hacker I, who creates solutions to some problems and tries to innovate in his work and the second type hacker spoiler who always works from sabotage and interference for reasons not positive. The emergence of DNA coding is a new area of encryption, which effectively contributes to the security of blood information, and some of the algorithms available in DNA have limitations in that they still use standard stereotaxic coding in some of their steps or laboratory experiment as inappropriate in a digital computing environment. There are some ways that combine DNA and another coding method to close the gap.

1.2 Problem Statement

DNA it is common algorithm where it was applied encryption algorithm with many other algorithms. There were several trials of cryptanalysis to find the weaknesses in the output of the generating key. Thus, in this project we will enhance the randomness of the output of DNA.

1.3 Goals and objectives

The objectives of the study are to link the DNA algorithm with the RC4 algorithm to increase randomization in order to increase security and measure the strength of the cryptography and to identify the weaknesses.

1.4 Research Questions

The following are the research issues of the study:

1. Is there literature on simulation for factors that cause Information Security Breaches?
2. What tools and techniques were used to develop these Information Security models? And what is the impact of each factor?
3. Were the models developed subjected to end user feedback to assess ease of use and design credibility?
4. What were the findings and recommendations from previous works on Information Security behavioral modeling?

1.5 **The importance of studying:**

The study will be useful because the knowledge obtained will identify the strengths, weaknesses and subsequent work of the institutions in reducing cyber-attacks from employees. Also, the knowledge gained from this research will use the general employee management and organization process in managing cyber issues. Finally, the study will add information to the small scientific world.

1.6. Project Organization

This project was organized in the following chapters to provide a clear understanding of the proposed methodology and the result of the experience created to ascertain the proposed solution. In Chapter 2, the reader will acquire the basic knowledge necessary to understand the search reminder. The NIST will be presented in Chapter 3. In Chapter 4, the practical part of our project will be presented. Conclusion of the research in Chapter 5,
with a summary of the work done and linking the DNA algorithm with the RC4 algorithm and clarifying their strengths and weaknesses.

1-7 Previous Studies

1-Jie Chen 2003/ A DNA-based, Bimolecular Cryptography Design.

Advantages: Storing large amount of data in compact volume. Massive parallel processing capabilities of bio molecular computation.

Disadvantages: Decrypt message as given in the code book. Difficult to send message which is not in code book.

2-Tushar Mandge et al., 2008/ DNA Encryption Technique Based on Matrix Manipulation and Secure key Generation Scheme.

Advantages: Secure generation algorithm to generate a new key for encryption process. Always get new cipher data from same plaintext. It provides good security layer which does not give any hint about plaintext.

Disadvantages: It includes only basic operation. Security only depends upon key.

3-Guangzhao Cui et al. , 2008/ An Encryption Scheme Using DNA Technology.

Advantages: Prevent attack from a possible word as PCR primers. The complexity of Biological scheme and cryptography computing provide a double security safeguards for the scheme. Cost of encryption scheme is low.

Disadvantages: Security can depend only on decryption key. The encryption scheme is still far away being a perfect scheme.

4- Zhang Yunpeng et al., 2011/ Index-Based Symmetric DNA Encryption Algorithm.

Advantages: Exact position of DNA sequence is not identified. Huge key space, high sensitivity to plaintext on encryption. Proper random key sequence to improve security.

Disadvantages: Higher security, could encrypt a longer DNA Sequence takes more time. Security completely depends upon key.

Chapter Two

Background

2.1 Introduction

The research and study for codes can be described as cryptology that is the practice as well as the science of changing legible information into illegible information as well as returning it to an understandable state if it is concluded by an expected function and the methods used. Further, there are two approaches connected to cryptography as well as cryptanalysis. Symmetric cryptography is anchored on the concept of encryption. Encryption enables the sender and the recipient to share the same key that encrypts and decrypts the message. By contrast, cryptology encompasses the use of double keys. They include the public key that encrypts the message and private keys that decrypt the given messages. The main feature of the symmetric key systems is that they are simple to use. Additionally, they are faster and efficient. The symmetric cryptography applies the duality concept where a single key serves the encryption and decryption roles. The implementation of the symmetric cryptography in hardware has proved to be effective. The user does not have to witness a significant delay in time owing to the encryption and decryption that take place at the same time. The system offers a certain authentication degree since the data that is encrypted with a single symmetric key may not be decrypted using any of the existing symmetric keys. Subsequently, provided that the symmetric key is often concealed by the parties concerned in the usage of the encrypt communications. Both sides are assured of the fact that communication will take place provided that the decrypted message is sensible. With a symmetric key, the user can exchange the key with a given trustworthy participant. It helps in the production of an individual key for every particular pair of a participant in the communication. The users are guaranteed that in every message exchange that is encrypted using a given key, it can only be deciphered only by the participants. Subsequently, the message has to be kept a secret among the members. Asymmetric cryptography is the opposite of symmetric cryptography. It is sometimes referred to as the public key cryptography. It entails the usage of private and public keys to encrypt and decrypt data. In fact, the keys entail large figures which are computed in pairs but are never identical. One of the keys in the created pairs can be used by everyone in the system. Subsequently, it results in a creation of a public key. The remaining key in the pair is concealed and is often referred to as the private key. Any of the provided keys can be used in the encryption of the message while the resulting opposite key is used for decryption. In the deliverance of integrity, authenticity, and confidentiality, those who use asymmetric cryptography must be guaranteed of the authenticity of the public key. Additionally, the users must be assured that they key free from malicious damage by the third party.   It is up to the organization to select the best among the two. In a personal reflection, the symmetric key cryptography is the best alternative. It ensures utmost confidentiality of the data used in the organization's computer system. It guarantees efficiency with regards to the encryption and decryption of the data. Additionally, the users can have a meaningful competition without interruption.

2.2 Encryption

Cipher is an original word from the Arabic language which refers to as empty or zero. As a matter of fact, all the process of encrypting text means hiding its readability and meaning. Moreover, it sometimes utilizes the point that the encrypted text message uses the same method although in this case, the term cipher text is most common [3]. Encryption (encipher) can be described as the process of transforming the information into an unrecognized message. Further, the cipher can be categorized into some types as shown in the figure 2.1 below as showing in figure .

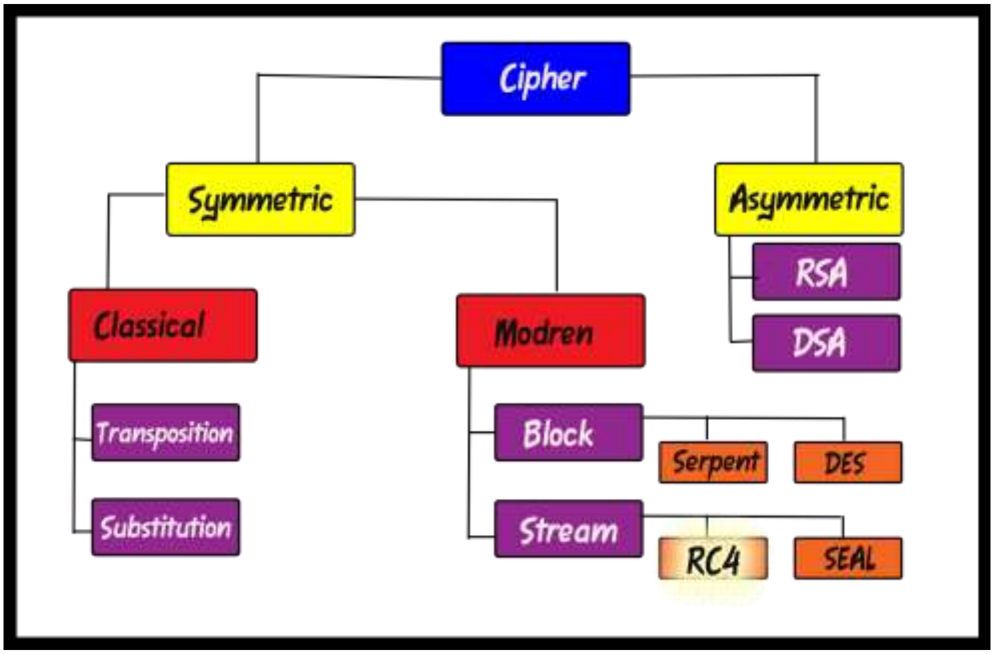


Figure2.1 The Classification of Cipher

Also**,** a cipher is regarded as part of the cryptography. It manifests as a form of an algorithm that carries the role of data encryption and decryption.

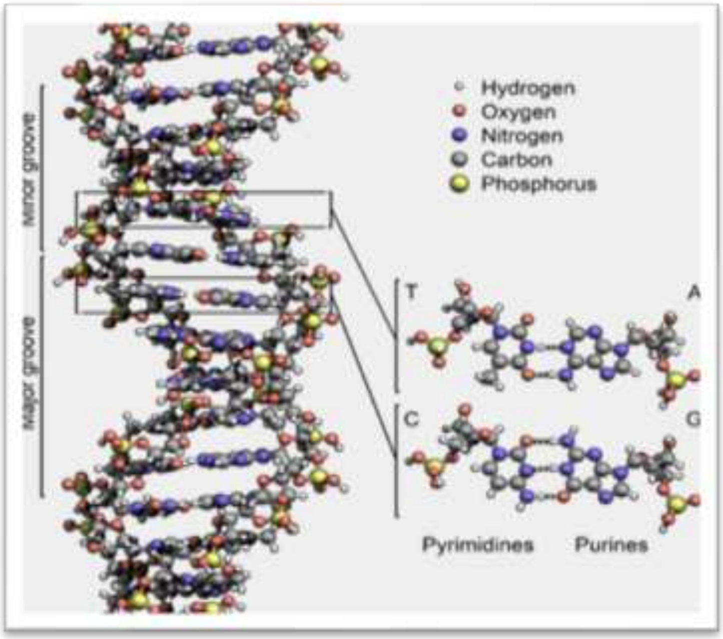
It is also considered the methods that encrypt the text message. It conceals the message by affecting its readability and meaning among the undesired users of the same text. In some cases, cipher refers to the message that is encrypted. Most ciphers operate through the realignment of the alphabet. In this case, A can be represented by L in the newly encrypted message. It manipulates the signal using a consistent pattern. Notably, most of the renowned ciphers employ a key. The key is regarded as the variable that is fused with the encrypted text using a unique manner. Additionally, ciphers use the algorithm. Algorithm entails a procedure or formula of the combinations of the provided keys with the text. There are several forms of ciphers that are often used in organizations. Block cipher involves breaking of the text into numerous chunks and combining each broken segment with a given key to create a hidden message. Block ciphers can split the text into sections that are referred to as blocks. All the generated blocks are equal in size. For example, all the blocks could have the size of 16 bytes. The algorithm pads the text with some bytes in filling out the last blocks since most of the blocks might not have the needed size. Steam cipher often applies at least a key to every bit at a given time in a bid to hide the message in the text. Most of the ciphers that are used in modern organizations are always blocked ones. Ciphers are ideal for an organization with a mass information system. The manner in which it manipulates and realigns the alphabet is its working principle. It can guarantee the discreetness of information using a newly created alphabet that only the users can decipher. People have the options between steams and block ciphers to select for an application. On a personal ground, I highly recommend the block cipher because of its increased efficiency.

2.3 Stream Cipher

A stream cipher can be defined as a symmetric key cipher in which the digits of plaintext are combined or added to the key stream of pseudorandom. Moreover, within the stream cipher, every digit of plaintext can be encrypted with another digit key stream generated, in order to obtain a digit of cipher text stream. However, since the encryption of all the digits is subjected to the current condition, it is therefore preferred to be known as state cipher. In specific, the digit, in this case, refers to a bite and hence the combination operation is an exclusive function (XOR) for instance; RC4, LILI, SOBER, and LEX.

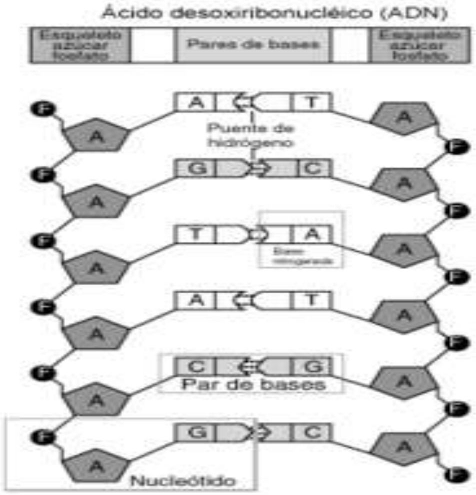
2.4 DNA

DNA cryptography is a branch of biological science, which has large data storage capacity.It stores information of living organisms. Living organisms has unique DNA information. It is defined as information storage, massive parallel processing and highly secured data transmission. DNA cryptography is based on one-time-pads scheme. Cryptography has to combine with molecular biology for more secure data transmission and data hiding. A plaintext message is encoded in DNA sequences. DNA sequences get powerful, when combined with nucleotide base A-T and C-G. DNA cryptography technology is needed in information security to protect and hide data. In traditional cryptography (like as DES, RSA), encrypted messages are detectable by attacker. DNA has capacity to store huge information rather than existing algorithm. DNA is introduced as a new technology for unbroken data. Genetic information is encoded as a sequence of nucleotides Guanine-G, Adenine-A, Thymine-T and Cytosine-C. Adenine, Thymine and Guanine, Cytosine are base pairs, which are attached to a sugar and a phosphate to maintain helical structure. DNA strands combined with hydrogen bond. A and T DNA sequences are combined with double hydrogen bond while C & G are combined with triple bond. Each nucleotide consists of the following three components, A Nitrogenous Base, A five carbon Sugar, A Phosphate Group



There are two types of DNA structure- single strand DNA and double strand DNA, which are complementary to each other. The encryption methods encrypt plaintext to cipher-text through one-time-pad scheme. The decryption methods decrypt received individual cipher-text packets to plaintext. The advantages of DNA molecular structure is its vast parallelism, exceptional energy efficiency and extra ordinary storage space. The disadvantages of DNA cryptography is it require huge computing time, high computational complexity and high tech bimolecular laboratory. Existing cryptography uses modern biological technologies. These technologies include PCR amplification and hybridization. These biological technologies are costly, complex and require a lot of time. DNA has much more storage capacity which is equal to (1gm=10^8 Tera bytes). It means small amount of DNA can stores world’s information.

DNA chain has a large scale of parallelism. Its computing speed is 1 billion times per second. DNA cryptography is a subfield of information science and emerged after the research of DNA computing in 1994. It provides a parallel processing capability with molecular level, to solve complex computational problem. DNA cryptography and information science is an effective application in design, analysis and application of DNA cryptosystem. Research on DNA cryptography is in the initial stage and required a lot of research in this field. DNA technology used to solve Hamilton path problems, combinatorial problems and extends to solve NP-complete problem by Lipton. DNA digital coding is based on binary digital coding, which is encoded by combination of 0 and 1. This paper focuses on different DNA methods of encryption process, which are powerful and secure than other traditional cryptography.



DNA and RNA are media for data storage due to very large amounts of data that can be stored in compact volume. They far exceed the storage capacities of conventional electronic, magnetic and optical media

2.5. RC4 ALGORITHM

RC4 is a stream cipher, symmetric key algorithm. The same algorithm is used for both encryption and decryption as the data stream is simply XORed with the generated key sequence. The key stream is completely independent of the plaintext used. It uses a variable length key from 1 to 256 bit to initialize a 256-bit state table. The state table is used for subsequent generation of pseudo-random bits and then to generate a pseudo-random stream which is XOR with the plaintext to give the Ciphertext.

The algorithm can be broken into two stages: initialization, and operation. In the initialization stage the 256-bit state table, **S** is populated, using the key, **K** as a seed. Once the state table is setup, it continues to be modified in a regular pattern as data is encrypted. The initialization process can be summarized by the pseudo-code**.**

Key-scheduling algorithm(KSA).

```
j = 0;
for i = 0 to 255:
    S[i] = i;
for i = 0 to 255:
    j = (j + S[i] + K[i]) mod 256;
    swap S[i] and S[j];
```

Pseudp-random generation algorithm (PRGA).

```
i = j = 0;
for (k = 0 to N-1) {
i = (i + 1) mod 256;
j = (j + S[i]) mod 256;
swap S[i] and S[j];
pr = S[ (S[i] + S[j]) mod 256]
Ciphertext>>M[k] XOR pr.
```

Table2.1 The History of RC4

Years	Happen
1987	Ron Rivest designed the RC4.
1994	RC4 was publicly established to be used on the web.
1997	RC4 became part of usually used encryption protocols like WEP.
2001	Combinatorial problem related to the number of inputs and outputs of the RC4 cipher was first posed by ItsikMantin and Adi Shamir.
2003/2004	RC4 became part of usually used encryption protocols like WPA for wireless cards.
2005	Andreas Klein presented an analysis of the RC4 stream cipher showing more correlations between the RC4 keystream and the key.
2008	Rivest has an article on RC4 in his own course notes
013	Breaking the RC4.

2014	Rivest has confirmed the history of RC4 and its code in the paper.
2015	RC4 prohibited for all versions of TLS by RFC 7465.

2.6.DNA Conversion

In this phase, output of (PRGA) operation’s binary digits are converted into DNA sequence as per the following rule:

□ When there are two consecutive:

00 = A;

01 = T;

10 = C;

11 = G.

2.7.DNA Encryption

In this phase, we first complement the DNA bases as per the DNA nitrogenous bases combined together (also called Complementary Rule) in which A always combined with T and C always combined with G. According to complementary rules we replace:

A with T;

T with A;

C with G;

G with C.

2.8. RC4 –DNA Steps

The steps for RC4-DNA encryption algorithm is as follows:
1- Get the data to be encrypted and the selected key.

2- Create two string arrays.

3- Initiate one array with numbers from 0 to 255.

4- Fill the other array with the selected key.

5- Randomize the first array depending on the array of the key.

6- Randomize the first array within itself to generate the final key stream.

8. Output operation of RC4 is represented in DNA bases format. Representation is: A=00; T=01; C=10; G=11.

9. Now complement the DNA bases as: A=T; T=A; C=G; G=C .

10 .We test a million bits in the NIST program to test random output.

Chapter Three

Methodology

&

Analysis

3.1 Introduction

The randomness of Pseudo-random generation algorithm has a vital function within the encryption process in that; at the time when the digits are increased in randomness, this feature makes the algorithm to be stronger as well as secure when it forms refraction. Moreover, the random generator digits make sure that a quality data is generated which remains to be a difficult function of the encryption. Besides, the randomness can be determined by statistic tests, where every test evaluates the value of randomness through various dimensions. Also, the randomness feature is deemed to be a probabilistic element that is the random elements of the digits can be described in the form of probability. Thus, the probable results of these tests can be applied in the real random series which is determined prior and hence can be described using the probabilistic. In addition, there are some tests that can be used to determine if a series of digits within the encryption is either random or not. Moreover, it is significant to acknowledge that there is no particular definite test that can be deemed to be complete. However, there are numerous tests to evaluate if a series of a given digits are random, not random and the value of their randomness. Furthermore, the results obtained from these tests should explain the care taken to avoid inaccurate inferences from a given set of digits generated. The test comprises of 17 tests which have an objective to critically analysis the binary series. Moreover, the test evaluates the randomness of the information by various statistic bits or blocks of bits. Further, all these tests evaluate the randomness of the whole entire bit stream. More importantly, the outcomes of these tests are always in the shape of *p*-value which symbolizes chances that are likely to be generated by random digits. Moreover, the p-value generated from the test can be centering on single features which have a clear inference of statistic and the inference of results of testing. The tests include:

3.2.NIST Statistical Test

Generally, these sixteen tests are categorized into two groups. The first group is called non-parameterized test and the second group called parameterized test are includes:

Table 3.1 NIST Statistical Test

Non-Parameter Test	Parameter Test
Cumulative Sum Test (forward and reserve)	Serial Test
Runs Test	Overlapping Template of All Ones Test
Longest Run of ones in Block Test	Non-Overlapping Template Matching Test
Discrete Fourier Transform Test	Approximate Entropy Test
Lempel-Ziv Compression Test	Block Frequency Test
Rank Test	Universal Statistical Test
Random Excursions Test	
Random Excursions Variant Test	
Frequency Test	

A brief description of each test follows [8]:

3.2.1 Frequency (Monbiot) Test

is centered on assessing the proportion of zeros and ones for the whole sequences.The purpose of the test is to limit whether the number of ones and zeros in a sequence are as the same as would be expected for a truly random sequence. The test evaluated the closeness of the part of ones to 0.5, that is the number of ones and zeroes in the sequence almost be the same.

3.2.2 Frequency Test within a Block

is centered on evaluating the proportion of ones within M–bit blocks. The purpose of the test is to limit whether the frequency of ones in an M-bit block is approximately M/2 It appears as expected under the random assumption. For bock size M=1, this test moves to the Frequency (Monbiot) Test.

3.2.3 Run Test

measures the total numbers of runs in the sequence, where a run is a continuous sequence of corresponding bits .the run of length k consists of completely k corresponding bits and is bounded before and after with a bit of obverse value. The purpose of this test is too limited whether the number of runs of ones and zeros of several lengths is as expected for a random sequence. Especially, the Run test limited whether the oscillation among zeros and ones is too fast or too slow.

3.2.4Test for the longest Run of Ones in a Block

evaluate the longest run of ones into M-bit blocks. The purpose of this test is to limit whether the long of the longest run of ones within the tested sequence is regular with the long of the longest run of ones that it appears as expected random sequence. A non-regularity in the expected long of the longest run of ones indicates that there is also a non-regularity in the expected long of the longest run of zeroes.

3.2.5 Binary Matrix Rank Test

Is centered on assesses the rank of disjoint sub-matrices of the whole sequence. The purpose of Rank test is to check for linear reliance among constant-length substrings of the main sequence.

3.2.6 Discrete Fourier Transform (Spectral) Test

Reveals periodic features in the tested sequence that would denote a variation from the presumption of randomness. The purpose of this test is to discover whether the number of summits surpasses the 95% threshold is very different 5%.

3.2.7 Non-Overlapping Template Matching Test

Reveals generators that produce more than events of a given non-periodic (aperiodic) pattern. For Non-overlapping test and The Overlapping Template Matching Test an m-bit window that used for search of a particular m-bit pattern. If the pattern is not found, the window slides one-bit placement but if the pattern is found, the search resumes after the window is reset to the bit after the found pattern.

3.2.8 Overlapping Template Matching Test

Counts the number of statuses of pre-specified aimed strings. Together Overlapping test and the Non-overlapping Template Matching Test use the m-bit window to search for a specific m-bit pattern. If the pattern is not found, the window slides one-bit placement. The difference among this test and Non- overlapping Template Matching Test that is resuming the search after the pattern is found the window slides only one bit.

3.2.9 Universal Statistical Test

This test appears through the counts the numbers of bits of the suitable pattern. The number is closed related to the entropy of the number generator.

3.2.10 Lempel-Ziv Compression Test

Limits how far the tested sequence can be pressed. The sequence is considered to be non-random if it can be pressed more.

3.2.11 Linear Complexity Test

Assesses the length of a linear feedback register (LFSR). The purpose of Linear Complexity Test is to limit whether the sequence is complex enough considered random or not. By Longer LFSRs can be summarized Random sequences. If the LFST is shorter that mean is non-randomness.

3.2.12 Serial Test

This test determines whether the number of overlaps of 2m m-bit nested patterns bit is approximately the same as expected for a random sequence. Every m-bit pattern has the similar prospect to show the different m-bit pattern. The Serial test is equivalent to the Frequency Test when m = 1.

3.2.13 Approximate Entropy Test

Concentrated to the frequency of all potential overlapping m-bit patterns via the whole sequence.

3.2.14 Cumulative Sums (CUSUM) Test

This test is calculated the maximal trip (from zero) of the random walk realized by the cumulative sum for (-1,+1) adjusted number in the sequence. The purpose of this test is to limit whether the cumulative sum for the partial sequences happing in the tested sequence is to great or too little relative to the expected attitude of the cumulative sum for random sequences, the trip of the random walk should be close to zero. The trip to this random walk will be large from zero in these specific types of non-random sequences. The cumulative sum can be in a forward way (mode=0) beginning from the starting of the sequence to the final of the partial sequence and the larger statistic value reference that 1s or 0s happen in the large numbers at the early stages of the sequence. And can also be in a reverse way (mode=1) beginning from the end of the sequence to the starting of the partial sequence. If it has large statistic value, 1s or 0s do happen in large numbers at the final of the sequence.

3.2.15 Random Excursion Test

Calculates exactly the number of the period having K that visit in the cumulative sum random walk. The cumulative sum random walk comes from the partial sum after the (0, 1) sequence is transferred to the suitable (-1,+1) sequence. The period of the random walk has consisted of the sequence of steps of the unit length taken at random that started at and return to the origin. This test is intended to limit if the number of visits to the designated state within a given period deviates from what one expects for a random sequence. This test is a series of eight tests and conclusion, one test and conclusion for each of these states: -4, -3, -2, -1, +1, +2, +3,+4.

3.2.16 Random Excursion Variant Test

Calculate the total number of events in which a particular event occurs in the cumulative sum random walk. The purpose of the test is to identify deviations from the potential number of events to several states in random walking. This test is a series of eighteen tests and conclusions, one test, and one conclusion for each of the events -9, -8, …, -1, +1, +2, ..., +9.

3.3.Configuration

The design of the proposed was done using Visual C++ and the tests of this PRNG were conducted using NIST STS-1.6. The probability of good or bad random number generator is represented by p-value. The testing process compared p-value to 0.01. If the p-value is more than 0.01, then the process accepts the sequence; otherwise, it rejects the sequence because the sequence is not random. Conversely, some tests accept large sizes of sequence and others fail in a small size; still other tests accept both sizes.

If the tests give a p-value asymptotically to 1, then the sequence appears to have perfect randomness. A p-value of zero indicates that the sequence appears to be completely non-random. The SUCCESS indicates the sequence is acceptable and has enough randomness, whereas the FAILURE indicates that the sequence is not acceptable due to non-randomness.

In the NIST statistical suite, there are two tests (Random Excursion and Random Excursion Variant) that do not provide results each running because these tests give a result when the number of cycles exceeds 500.

Therefore, the test results are shown as a text file in NIST. The result of the random excursion test and random excursion variant test are shown in Figure 3.1 and Figure 3.2 respectively.

Chapter Four

Implementation

&

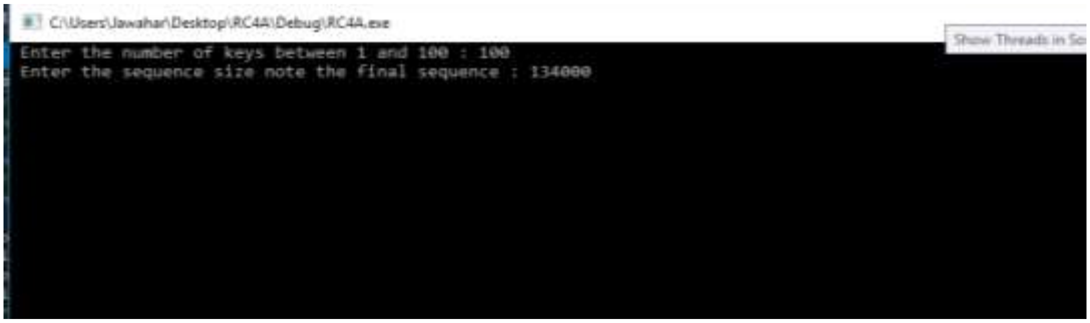
Testing

4.1 Introduction

In our program, we used a large size 134,000 bytes (1,072,000 bits) generated by each key. These sequences were tested, and we subsequently calculated the average of the *p*-values resulting from these tests. The *p*-values are acceptable when greater than 0.01, and the produced sequence can be deemed random, uniformly distributed and suitable for cryptography.

In the beginning, we will talk in detail about the RC4 encryption program. In the beginning, we used the site https://www.random.org/integer-sets to generate 100 random key each one size 16 bytes .

Run screen asked the user to input the number of key and the length of the key generating as shown in Figure 4.1



```
Enter the number of keys between 1 and 100 : 100
Enter the sequence size note the final sequence : 134000
```

Figure 4.1 Run Screen of Program

Generate 50 file where each file represent the result of one key which 1072000 bit that’s mean 50 key as shown in Figure 4.2

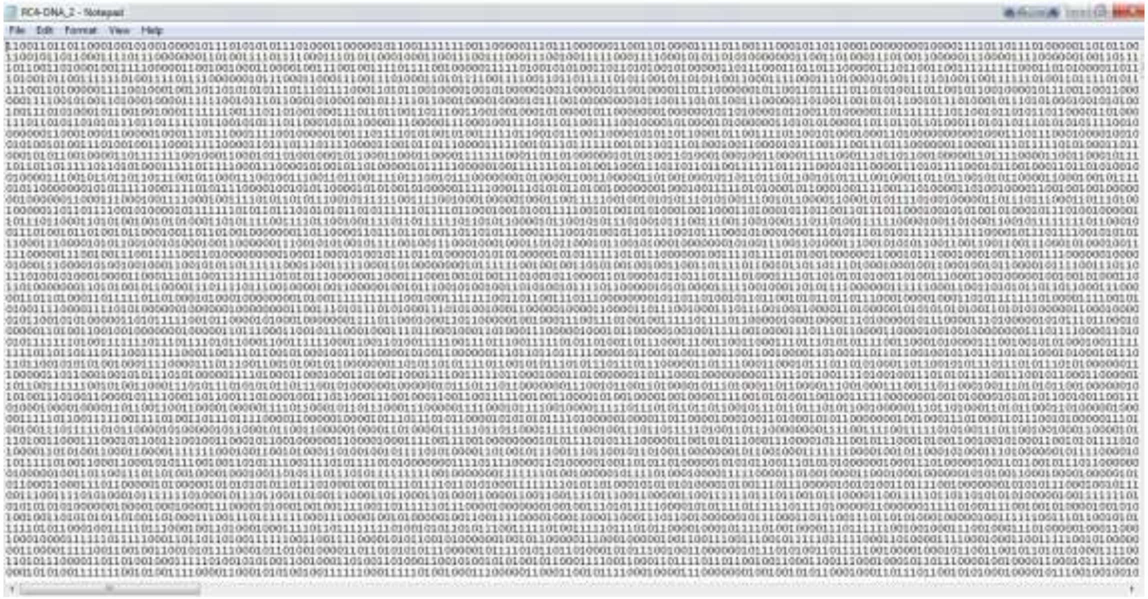


Figure 4.2 50 Key in Binary

We then modified the algorithm by using the RC4_DNA in the second phase of the algorithm (PRGA) then tested the randomness of the new algorithm by using same 50 key which we had use it in original RC4 the 50 file:

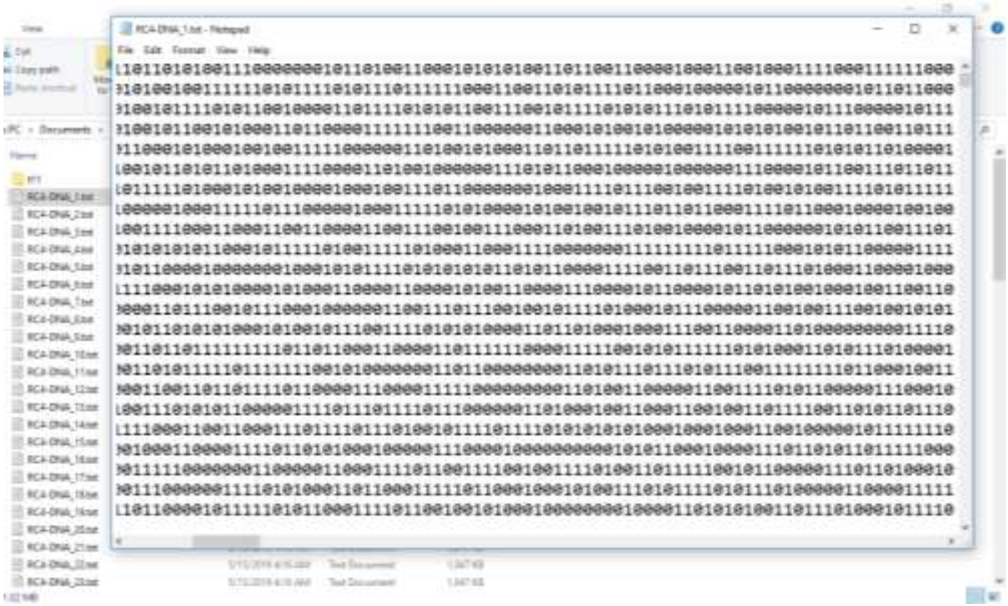
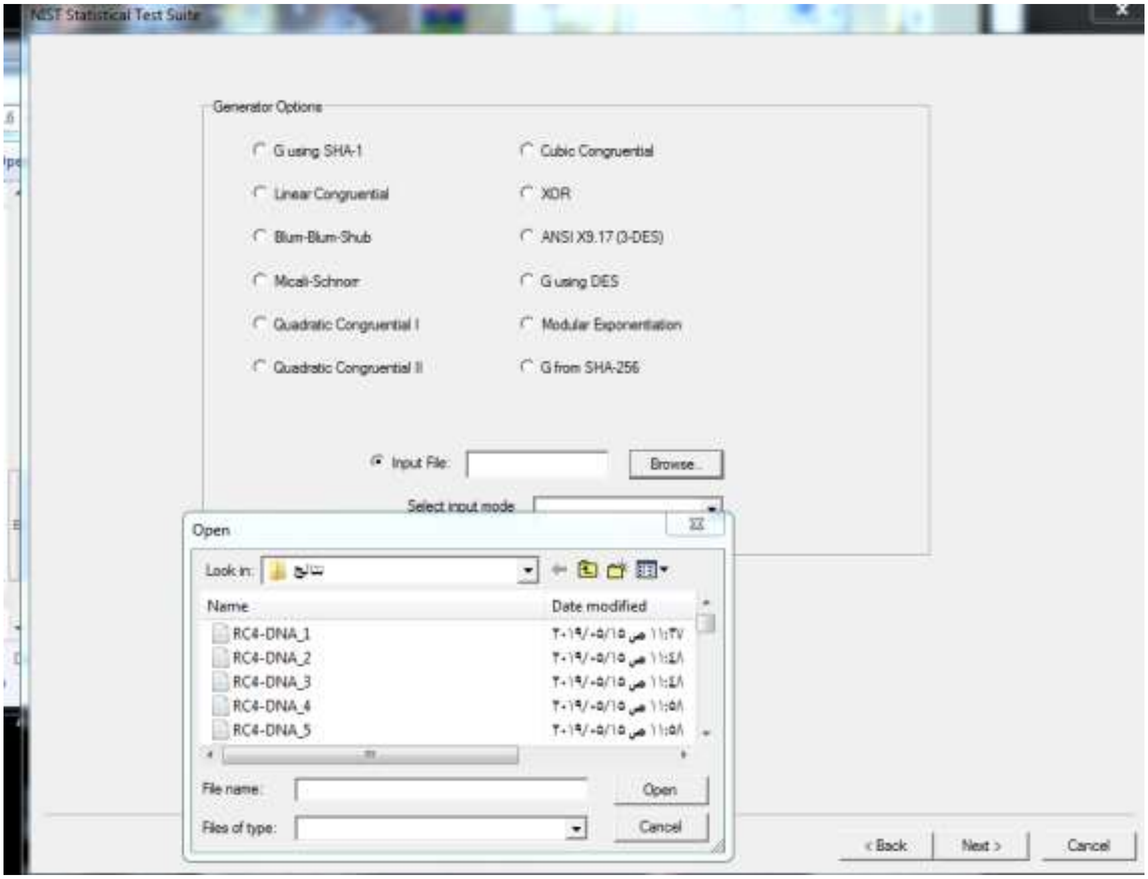


Figure 4.8 50Key in Binary of New RC4

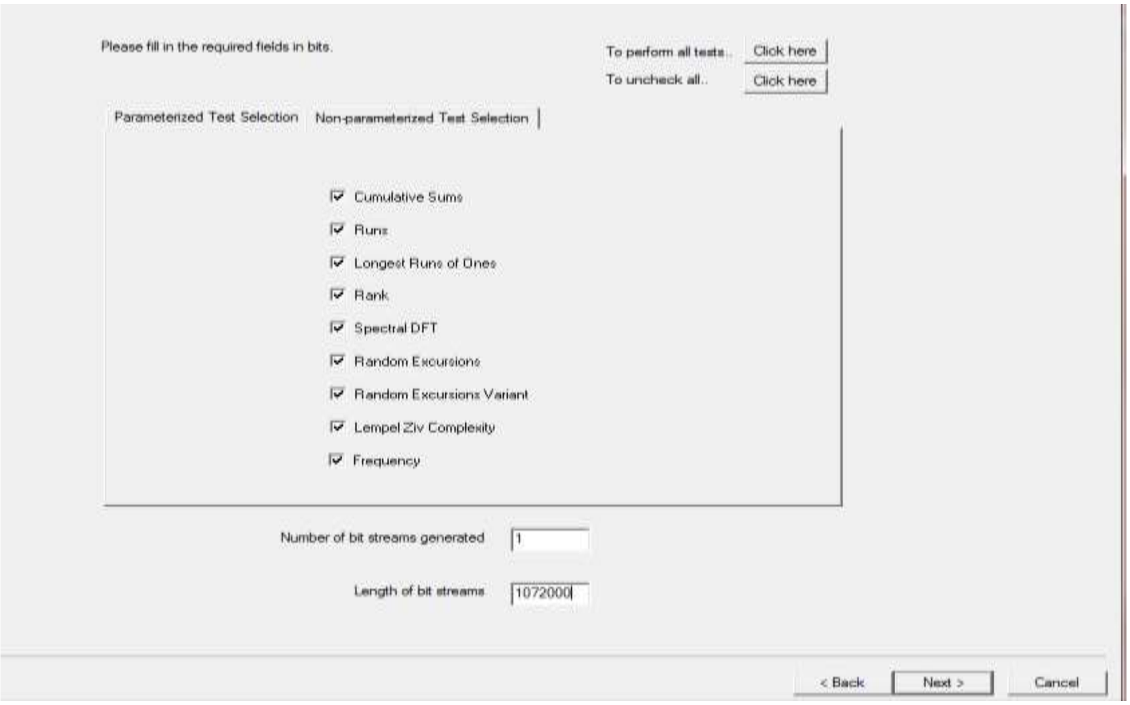
Then we introduced the entire 100 key to the new RC4 and RC4 algorithm in the test. To measure the random range of both algorithms and then save the results in the excel file, below we will show you how to use this program



Name	Date modified
RC4-DNA_1	٢٠١٩/٠٥/١٥ ص ١١:٣٧
RC4-DNA_2	٢٠١٩/٠٥/١٥ ص ١١:٤٨
RC4-DNA_3	٢٠١٩/٠٥/١٥ ص ١١:٤٨
RC4-DNA_4	٢٠١٩/٠٥/١٥ ص ١١:٥٨
RC4-DNA_5	٢٠١٩/٠٥/١٥ ص ١١:٥٨

Figure 4.10 Home Page of NIST Statistical Test

- Non-parameter Test



Please fill in the required fields in bits.

To perform all tests.. Click here

To uncheck all.. Click here

Parameterized Test Selection | Non-parameterized Test Selection |

☑ Cumulative Sums

☑ Runs

☑ Longest Runs of Ones

☑ Rank

☑ Spectral DFT

☑ Random Excursions

☑ Random Excursions Variant

☑ Lempel Ziv Complexity

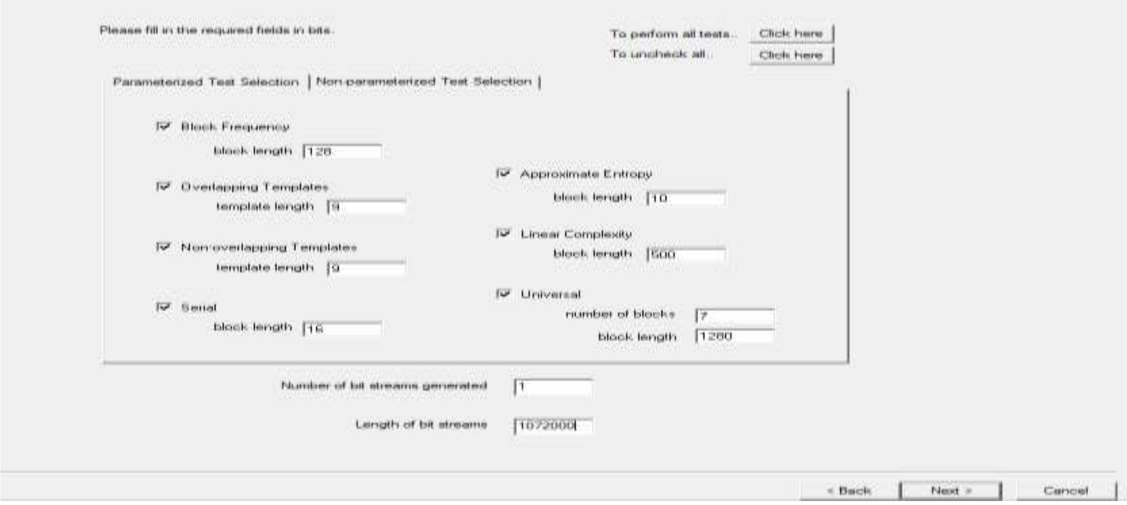
☑ Frequency

Number of bit streams generated: 1

Length of bit streams: 1072000

Figure 4.11 Non-Parameter Test

- Parameter Test



Please fill in the required fields in bits.

To perform all tests.. Click here

To uncheck all.. Click here

Parameterized Test Selection | Non-parameterized Test Selection |

☑ Block Frequency
block length: 128

☑ Overlapping Templates
template length: 9

☑ Non-overlapping Templates
template length: 9

☑ Serial
block length: 16

☑ Approximate Entropy
block length: 10

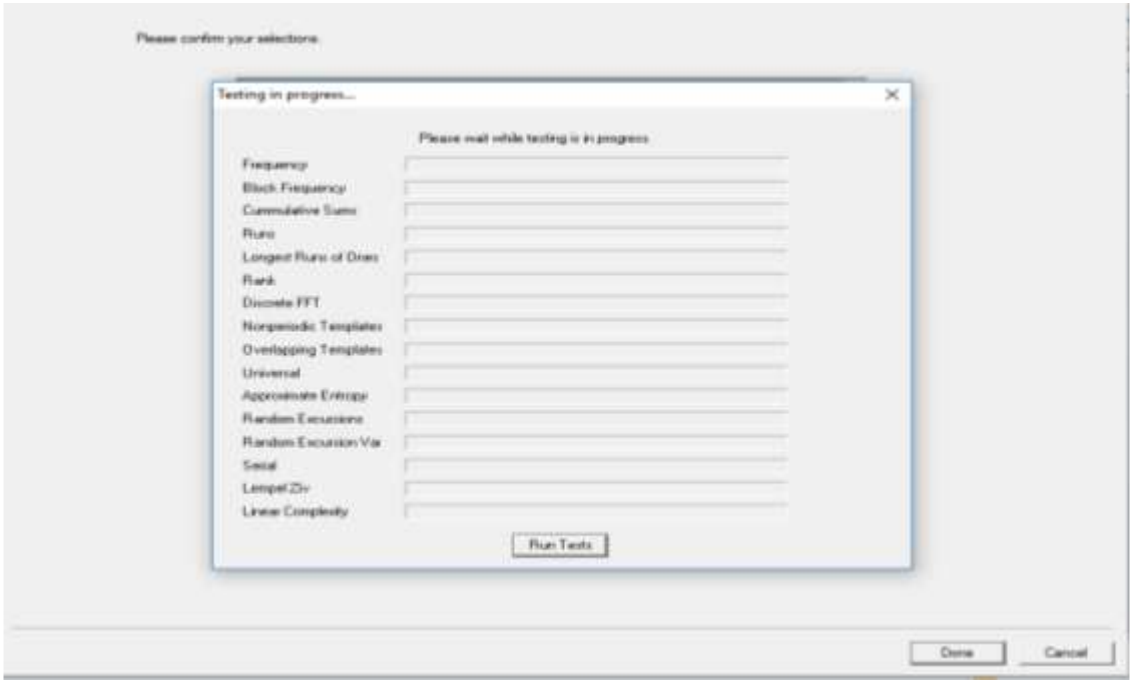
☑ Linear Complexity
block length: 500

☑ Universal
number of blocks: 7
block length: 1280

Number of bit streams generated: 1

Length of bit streams: 1072000

Figure 4.12 Parameter Test

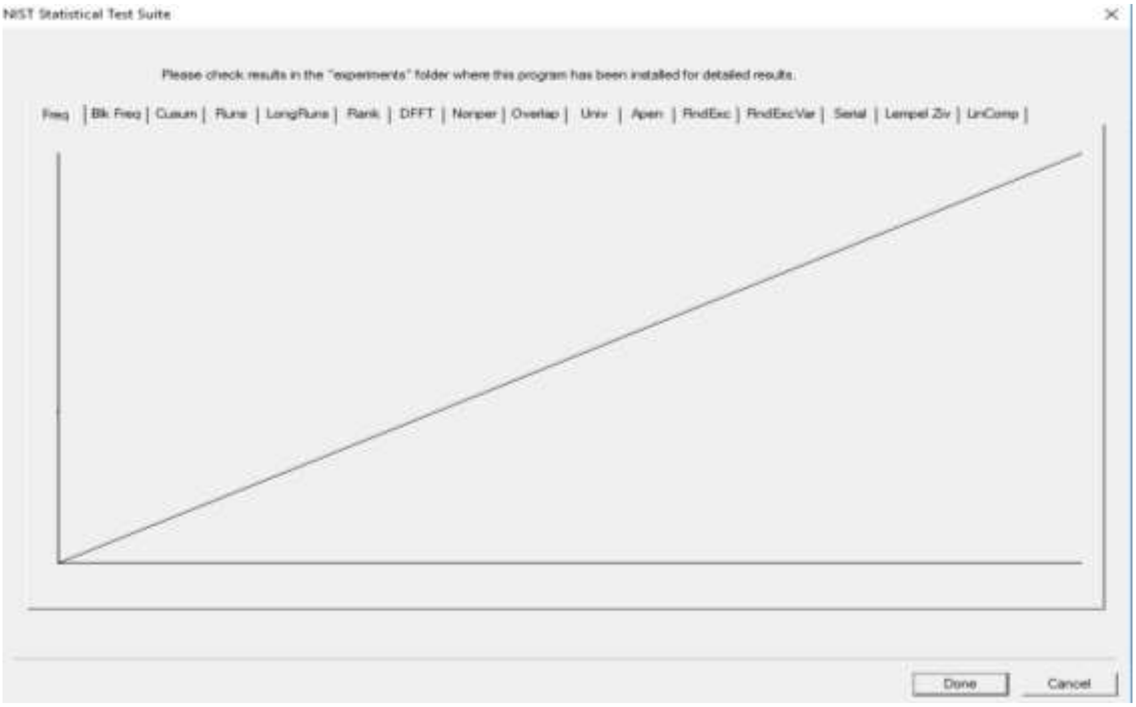


Please confirm your selections.

Testing in progress...

Please wait while testing is in progress

Figure 4.13 Run of NIST Statistical Test



NIST Statistical Test Suite

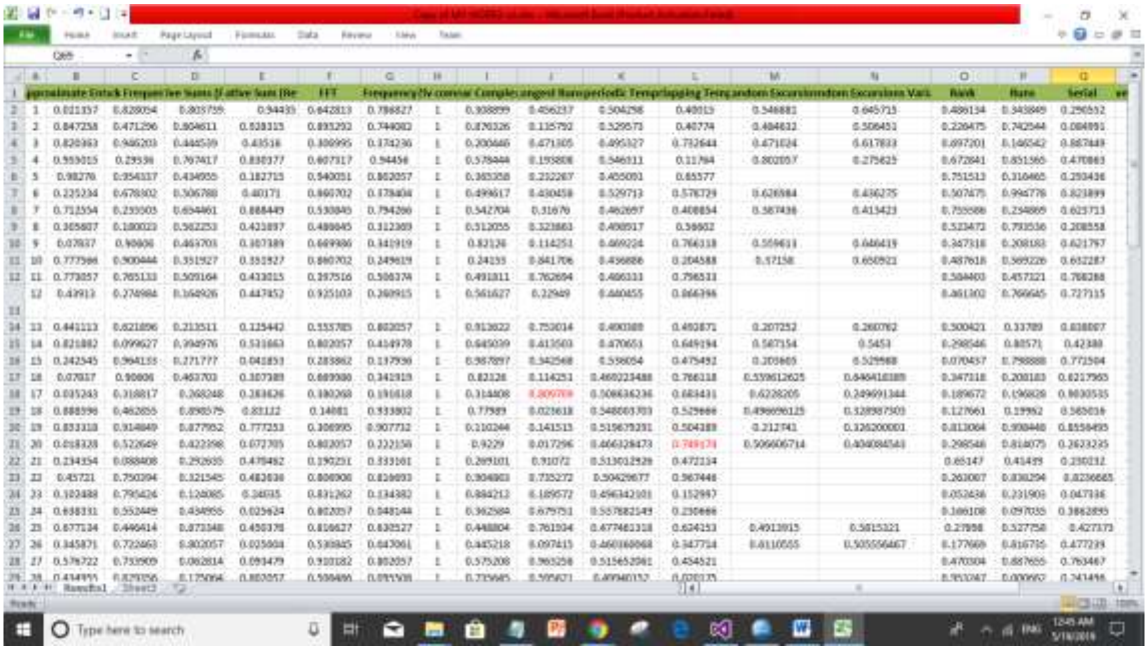
Please check results in the "experiments" folder where this program has been installed for detailed results.

Freq | Blk Freq | Cusum | Runs | LongRuns | Rank | DFFT | Nonper | Overlap | Univ | Apen | RndExc | RndExcVar | Serial | Lempel Ziv | LinComp |
Done
Cancel

Figure 4.14 End of NIST Statistical Test

Here we show you the results of Excel file after taking the rate of each test separately [5]

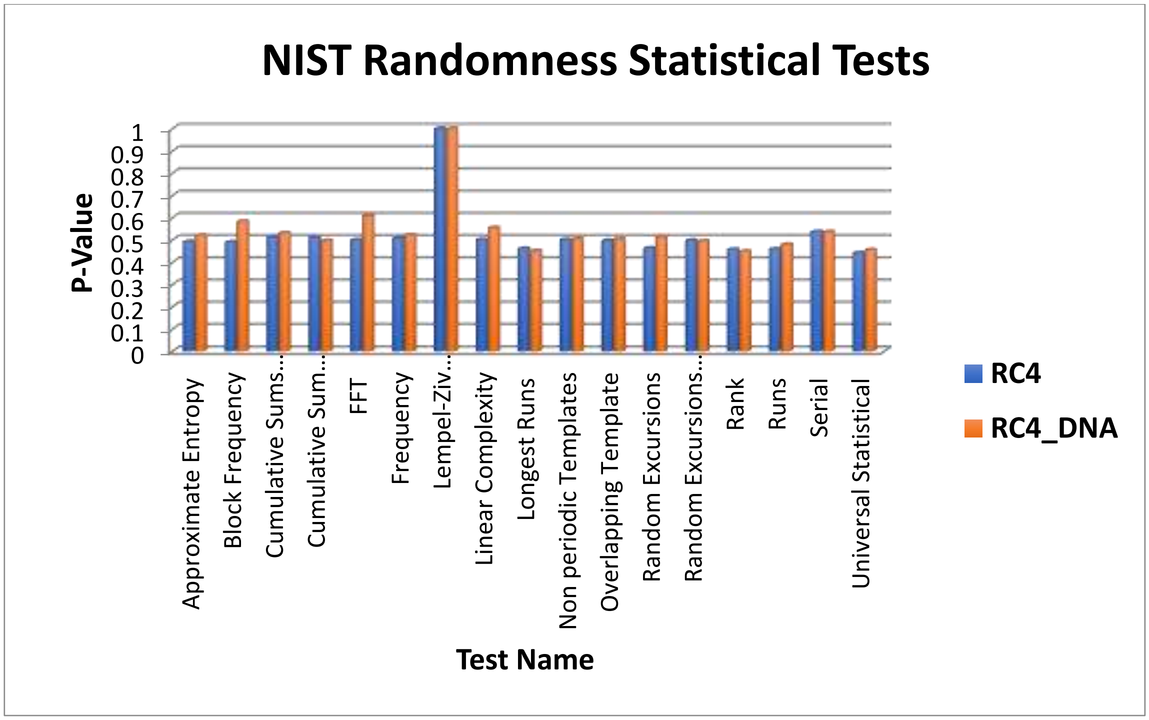
RC4_DNA Excel sheet



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	pproximate	Ent	ck Frequen	ive Sums (F	ative Sum (Re	FFT	Frequency	lv cimmar	Complex	ongest Run	periodic Temp	lapping Temp	andom Excursion	dom Excursions Vari	Rank	Runs	Serial
2	1	0.021357	0.828054	0.803739	0.94435	0.642813	0.786827	1	0.308899	0.456237	0.504298	0.40013	0.346881	0.645715	0.486134	0.343849	0.290552
3	2	0.847258	0.471296	0.804611	0.828315	0.893292	0.744082	1	0.876326	0.135792	0.529573	0.40774	0.484632	0.506451	0.226475	0.742544	0.084891
4	3	0.820363	0.946203	0.444539	0.43516	0.300995	0.374236	1	0.200446	0.471305	0.495327	0.732644	0.471024	0.617833	0.697201	0.146542	0.887449
5	4	0.553015	0.29536	0.767417	0.830377	0.607317	0.94456	1	0.578444	0.193808	0.546111	0.11784	0.802057	0.275825	0.672841	0.851565	0.470863
6	5	0.98276	0.954337	0.434955	0.182715	0.540051	0.802057	1	0.383358	0.232287	0.455091	0.85577			0.751513	0.316465	0.293436
7	6	0.225234	0.678302	0.306788	0.40171	0.860702	0.178404	1	0.499617	0.430458	0.529713	0.576729	0.626984	0.436275	0.507475	0.994778	0.823899
8	7	0.712554	0.233303	0.654461	0.888449	0.530845	0.794286	1	0.542704	0.31676	0.462697	0.408854	0.387438	0.413423	0.755586	0.234869	0.623713
9	8	0.305807	0.180023	0.582253	0.421897	0.488845	0.312369	1	0.512055	0.323863	0.498917	0.38602			0.523472	0.793536	0.208558
10	9	0.07037	0.90806	0.463703	0.307389	0.669986	0.341919	1	0.82126	0.114251	0.469224	0.766318	0.559613	0.646419	0.347318	0.208183	0.621797
11	10	0.777586	0.900444	0.351927	0.351927	0.860702	0.249819	1	0.24155	0.841706	0.456886	0.204588	0.57158	0.650921	0.487618	0.569226	0.652287
12	11	0.773057	0.765133	0.509164	0.433015	0.397516	0.508374	1	0.491811	0.762694	0.486331	0.798533			0.504403	0.457321	0.768288
	12	0.43913	0.274984	0.164926	0.447452	0.925103	0.260915	1	0.561627	0.22949	0.440455	0.866396			0.461302	0.766645	0.727115
13																	
14	13	0.441113	0.621896	0.213511	0.125442	0.553785	0.802057	1	0.913622	0.753014	0.490388	0.492871	0.207252	0.260762	0.300421	0.33789	0.838007
15	14	0.821882	0.099627	0.394976	0.531663	0.802057	0.414978	1	0.645039	0.413503	0.470651	0.649194	0.587154	0.5453	0.298546	0.80571	0.42388
16	15	0.242545	0.964133	0.271777	0.041853	0.283862	0.137936	1	0.987897	0.342568	0.558054	0.475492	0.203605	0.529988	0.070437	0.798888	0.771504
17	18	0.07037	0.90806	0.463703	0.307389	0.669986	0.341919	1	0.82126	0.114251	0.469223488	0.766318	0.559612625	0.646416189	0.347318	0.208183	0.6217965
18	17	0.035243	0.318817	0.268248	0.283626	0.380268	0.191618	1	0.314808	0.809769	0.508636236	0.683431	0.6228205	0.249691344	0.189672	0.196828	0.8030535
19	18	0.888596	0.462855	0.898579	0.83122	0.14081	0.933802	1	0.77989	0.023618	0.548003703	0.529666	0.496696125	0.328987503	0.127661	0.19962	0.585036
20	19	0.833318	0.914849	0.877952	0.777253	0.306995	0.907732	1	0.110244	0.141513	0.519675291	0.504189	0.212741	0.326200001	0.813064	0.930448	0.8558495
21	20	0.018328	0.522649	0.422398	0.072705	0.802057	0.222156	1	0.9229	0.017296	0.466328473	0.749174	0.506606714	0.404084543	0.298546	0.814075	0.2623235
22	21	0.234354	0.088408	0.292635	0.479462	0.190251	0.333161	1	0.289101	0.91072	0.513012528	0.472114			0.65147	0.41499	0.230232
23	22	0.45721	0.750394	0.321545	0.482636	0.806906	0.826693	1	0.304903	0.735272	0.50429677	0.967446			0.263007	0.838294	0.8236665
24	23	0.102488	0.795426	0.128085	0.24035	0.831262	0.134382	1	0.884212	0.189572	0.496342101	0.152997			0.052636	0.231903	0.047336
25	24	0.638331	0.552449	0.434955	0.025624	0.802057	0.048144	1	0.362584	0.679751	0.557882549	0.230666			0.166108	0.097035	0.3862895
26	25	0.677134	0.446414	0.873348	0.450378	0.816627	0.830527	1	0.448804	0.761934	0.477461318	0.624153	0.4913915	0.5815321	0.27898	0.527758	0.427373
27	26	0.345871	0.722463	0.802057	0.025604	0.538845	0.047061	1	0.445218	0.097415	0.460160068	0.347714	0.6110555	0.505556467	0.177669	0.816716	0.477239
28	27	0.576722	0.733909	0.062814	0.093479	0.910182	0.802057	1	0.575208	0.963258	0.515652061	0.454521			0.470304	0.887655	0.763467
29	28	0.434935	0.829356	0.175064	0.802057	0.508486	0.095508	1	0.735685	0.505621	0.40940152	0.020135			0.353247	0.000662	0.241498

Figure 4.19. The Result of 50 Key of RC4_DNA

Then we use visual basic 2010 to design interface that allow to user to cipher any text they want.



Test No.	Statistical Test Name	RC4	RC4_DNA
1	**Approximate Entropy**	0.49167119	0.52098782
2	**Block Frequency**	0.489457076	0.58242956
3	**Cumulative Sums (Forward)**	0.511759633	0.53093118
4	**Cumulative Sum (Reverse)**	0.509934392	0.49620828
5	**FFT**	0.498774468	0.607476354
6	**Frequency**	0.506324013	0.5217638
7	**Lempel-Ziv compression**	1	1
8	**Linear Complexity**	0.500117418	0.553840625
9	**Longest Runs**	0.460326747	0.450299569
10	**Non periodic Templates**	0.500500051	0.503571882
11	**Overlapping Template**	0.49631219	0.503581778
12	**Random Excursions**	0.461785035	0.51303963
13	**Random Excursions Variant**	0.496975747	0.494119102

14	**Rank**	0.456413038	0.448222673
15	**Runs**	0.457660949	0.478905
16	**Serial**	0.536270703	0.53626585
17	**Universal Statistical**	0.440574848	0.45459232

Table 4.1:statistical test

4.2.Conclusion

Rapid technological advancement has revolutionized the dynamics of the market and given rise to newer calamity, cybercrime which threatens the survival of organization. In the niche where only the fit survives it is important that every organization formulate and adopt strategic plans which are integrated with cyber strategies and policies. Organizations that happen to ignore the call of adopting cyber security strategies are most constant prey to all knowing cyber criminals who happen to be more dangerous than it is thought [9]. Many organizations have been blackmailed by cyber criminals and have suffered in silence for fear of spoiling the organizations image. These criminals often go scot free in many cases so instead of waiting to use reactive defense it is important to blend these cyber security strategies with organization long term and short term strategically plans so as to give this malady of cyber-crime a proactive approach. Thus, to increase the security and reduce potential attacks should be used strong encryption algorithms. In the work we improve the RC4_DNA algorithm to increase the randomness. The analysis and the results show that the new algorithm is passed the NIST statistical tests

References

[1] Mousa, A., & Hamad, A. (2006). Evaluation of the RC4 Algorithm for Data Encryption. IJCSA, 3(2), 44-56.

[2] Bhatt, (2002).Management strategies for individual knowledge and organizational knowledge. Journal of knowledge management, 6(1), pp.31-39.

[3] Guan, Z. H., Huang, F., & Guan, W. (2005). Chaos-based image encryption algorithm. Physics Letters A, 346(1), 153-157.

[4] Fluhrer, S., Mantin, I., Shamir, A.: Weaknesses in the key Scheduling Algorithm of RC4. Proceedings of Annual Workshop on Selected Areas in Cryptography, vol. 2259, Toronto, Canada, pp. 1-24, Springer, 2001.

[5] William Sharp, 2010. The past, present, and future of cybersecurity. J. Nat'l Sec. L. &Pol'y, 4, p.13.

[6] Paul, S., Preneel, B.: A New Weakness in the RC4 Keystream Generator and an Approach to Improve the Security of the Cipher. In: Fast Software Encrypt, FSE, LNCS 3017. New York: Springer-Verlag, pp. 245–259, 2004.

[7] G.Carter, E.Dawson, and K. Wong. An Analysis of the RC4 Family of Stream Ciphers against Algebraic Attacks. Proc. 8th Australian Information Security Conference(AISC 2010), 2010.

[8] Rukhin, A., Soto, J., Nechvatal, J., Smid, M., Barker, E., Leigh, S., Levenson, M., Vangel, M., Banks, D., Heckert, A., Dray, J., Vo, S.: A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications. NIST special publication 800-22, National

Institute of Standards and Technology (NIST), Gaithersburg, MD, USA (2001); See http://csrc.nist.gov/rng/ .

[9] Choo Nelson, (2011). The cyber threat landscape: Challenges and future research directions. Computers & Security, 30(8), pp.719-731.