Ministry Of Higher Education
and Scientific Research
University of Babylon
College of Sciences for Women
Department of Computer Sciences

Protecting Data Based on DNA Watermarking Using Docking Technique

A Thesis
Submitted to the Council of College of Sciences for Women / University of Babylon in Partial Fulfillment of the Requirements for the Degree of Master in Sciences /Computer Sciences

By

Hayder Abdul Khaleq Abdul Raheem

Supervised by

Asst. Prof. Dr. Sahar Adill K.

Prof. Dr. **Ali Hussain Al-Marzuki**

2022 A.D.

1443 A.H.

بسم الله الرحمن الرحيم

﴿قالُواْ سُبْحَانَكَ لاَ عِلْمَ لَنَا إِلاَّ مَا عَلَّمْتَنَا إِنَّكَ

أَنتَ الْعَلِيمُ الْحَكِيمُ﴾

صدَقَ آللہ العليّ آلعظيْم

سورة البقره : اية (32)

Supervisor Certification

We certify that this thesis titled **“Protecting Data Based on DNA WatermarkingUsing Docking Technique**
“ Was done by (**Hayder AbdulKhaleq Abdul Raheem**) under our supervision.

Signature:
Name : Asst. Prof. Dr. Sahar Adill Kadhum, Ph.D.
Date: / / 2022

Signature:
Name : Prof. Dr. Ali Hussain, Ph.D.
Date: / / 2022

The Head of the Department Certification

In view of the available recommendations, I forward the thesisentitled "**Protecting Data Based on DNA WatermarkingUsing Docking Technique"** for debate by the examination committee.

Signature:
Name : Asst.Prof.Dr. Saif Alalak
Date: / / 2022

Dedication

I dedicate this thesis

To my beloved country Al-IRAQTo

Souls of Martyrs of IRAQ

To everyone who educates me even a singleletter, my teachers To my friends

Hayder

2022

Acknowledgment

At first, thanks and praise be to Allah, the Lord of the world, who give me courage, patience and enabled me to achieve this work…

I would like to thank my supervisor, Dr. Sahar Adill Kadum, Dr. Ali Hussain *for his continous help, support, guidance, and knowledge he provided me throughout my research.*

I wish to express my deepest gratitude to the committee members for spending their precious time on reading my thesis.

I would like to thank all my professors and all the staff of the Department of Computer Science at the College of Sciences for women / University of Babylon for their help. I wish for them the best of luck and success.

Abstract

Digital watermarking is used to authenticate and identify the ownership of the valuable contents. In the same context, DNA watermarking is a data hiding technique that aimsto protect the copyright of DNA sequences and ensures the security of private genetic information.

In this thesis, a proposed a DNA watermarking technique that can be used for preserving the selected DNA sequences (vitamin D) used as a reference and securing of private genetic information.

The main parts the proposal consist of are: dynamic coding tables, a central Dogma encryption algorithm, message segmentation, and a docking technique

The need for coding tables to convert the plain message to DNA bases and the ciphered message to an amino acid codon. These tables are generated dynamically, i.e., each message has its coding tables. Using central dogma technique to encrypt the plain message partially (using the first step from this technique (Transcription process) producing a form of coding called (mRNA) sequence.

The mRNA sequence will be segment upon a mathematical module has been built. The number of segments depends on the length of the ciphered message length.

The mutation of codons postfix is done according to a docking technique based on two parameters (concealing key and measuring key) in embedding message segments. The embedding process for each segment has particular manipulation such that, before any embedding process, a location and an amount of distance to place the segment must be specified, and each segment is tailored by a special codon called start and end codons.

The method was tested for preserving the identity of DNA sequences and extracting the amount of similarity to other DNA sequences using BLAST software. The test results identity was a 96% with the D vitamin database. This test shows there is not identity 100% between our selected database and all others on NCBI web site which that a prove to the security and preservation to the identity of DNA sequences and biological functions, undetectable, and resistance. The cracking probability test result was very small amount less than 1e-300. Although, a cracking test has been used to present the abilityof undetectable.

List of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Full Name of Abbreviation
CDMB	Central Dogma of Biological Sciences
CK , MK	Concealing key and Measuring key
CP	Cracking Probability
DES	Data encryption standard
DNA	Deoxyribonucleic acid
mRNA	Messenger Ribonucleic acid
NCBI	National Center For Biotechnology Information
NIST	National Institute of Standards and Technology
OTP	one-time password
RNA	Ribonucleic acid
RSA	RSA Rivest Shamir and Adelman
SNP	Single Nucleotide Polymer
SSDNAS	Single Strand Deoxyribonucleic acid sequence
tRNA	Transfer Ribonucleic acid

List of Algorithms

CHAPTER ONE

GENERAL INTRODUCTION

CHAPTER ONE
General Introduction

1.1 Introduction

People's communications have altered dramatically due to the rapid development of computer and network technology. Digital multimedia material transmission over the Internet has grown more common. However, because of the openness and sharing of networks, the security of multimedia is seriously jeopardized. People must pay more attention to online data transmission security. Encryption and information hiding are two of the most powerful and well-known securing data strategies [1-3].

Information hiding is a technique to protect information from any modification, including steganography and watermarking, and each has its algorithms. [4, 5]

A watermark is a technique of inserting information such as a number, text, image, … , etc. within a digital object like image, text, audio or video to protect the rights of the owner and authentication, which is a good strategy for dealing with liability difficulties, particularly in the case of multimedia data. Using the data's great redundancy and the notion that the eye cannot distinguish little variations between data values, [5, 6]. Personalized watermarking is suitable for security systems, which are unique and difficult to copy or duplicate, Biometric characteristic can be used in copyright protection and authentication likeDNA codes [7, 8]. DNA watermarking is a method to protectthe copyright of genomic DNA and ensuring the protection of genomicsprivacy information . DNA watermarking solutions must provide security, undetectable, robustness, and biological function maintenance [9].

DNA has a number of features that make it an ideal way to hide data. These features are based on two main points: First, DNA has a tremendous storage capacity for information. Second, any desired length of DNA sequence can be generated. The various strategies for DNA-based data concealing have been presented in the previous decade. [5-9].

Although maintaining safe and completeness transmission and reception of data and information is achieved by constructing an innovative cryptographic algorithm. These algorithms are among the most effective methods for concealing information by encrypting data so only authorized individuals will access and decrypt the original data. RSA, DES, Blue Fish, OTP, ..., etc. are samples of these algorithms [7, 10].

Nowadays, introducing the modern molecular biology DNA into security aspect has opened a new direction for information protection, whether it is from the aspect of cryptography or hiding information.

A ciphered watermark is obtained by applying a well-known ciphering algorithm. In this thesis, the watermark is ciphered biologically by exploitinga special molecular biology encryption algorithm called central dogma for encrypting direct messages. Where, the biological follow of central dogma procedures explain the transfer of genetic information from a living organism's DNA molecule to the production of functional protein using two main processes (Transcription and translation). Each process has a unique contribution to the production the functional synthesis (protein). DNA cryptography uses the biological central dogma method that take the plaintext through several processes from DNA to RNA and Amino Acid coding [11,12].

Dropping the watermark at a specific location in a selected DNA database sequence, a supported mechanism is used for this purpose called the docking technique, which is used to anticipate the preferred orientation of

DNA watermarking when bound in an active site to build a stable complex according to a function and pervert the active site [13,14].

Exploiting the above techniques to innovate a new watermark scheme and its encryption method to protect the transmitted information and data while preserving the copyright privacy of the selected DNA sequence. The whole structure of the proposal thesis is based on exploiting biological specific techniques.

1.2 Related Work

DNA cryptography and DNA information concealing are two types of DNA security research that have been done since the 2000s. The purpose of this comparison is to give a knowledge to work on more efficient and accurate robust DNA cryptography and information concealing techniques in the future. The following are some of the results of these studies:

In 2022, Satir and Kendirli.[15], proposed a DNA cryptography system by cooperating DNA encoding and DNA operators into a Feistel network topology. With its biological tools, DNA is used as a transporter. To simulate biotechnical hardware, the generated DNA sequence is digitally and biologically merged. In vitro, experiments are used to demonstrate the experimental outcomes

In 2019, Basu,etal.[16], presented an encryption/decryption technique based on the Central Dogma of Biological Sciences (CDMB) that uses two processes for encryption and decryption: transcription (products mRNA) and translation (products protein). The input is in the form of 16-bit blocks. The end result is a concatenated cipher text in the form of protein derived from the blocks. The key generation was achieved by employing randomized data trained with Bidirectional Associative Memory Neural Network, by re-storing the regeneration keys on a regular basis to save memory space.

In 2017,Harry.[17], presented a method for encoding messages text into genomes and their regulatory sequences, as well as transcription, translation, protein complexes sequences. These codes represent the foundation for a cryptographic paradigm. The method is presented in a hierarchical structure that starts with encoding the message text in a DNA code (cipher-gene), then uses transcription and translation to produce a protein code (cipher-protein), resulting in a set of cryptography algorithms for securing data and providing a fully evolved form of security between multiple parties.

In 2020, Reddy, etal.[18], A DNA bio-cryptography system is suggested in this study based on encryption/ decryption, and key generation. Central Dogma of Molecular Biology is used for genetic coding, transcription, and translation. The Bidirectional Associative Memory Neural Network is used for generating and restoring keys with various number of hidden and input neurons, Whale optimization algorithm (WOA) is used to get a highest weight vector .

In 2019, RashaS, etal.[19], suggested a biomolecular computing-based modified Play fair method for encrypting/decrypting for any type of symbols ,characters and integers. To reduce time and storage space, any duplicate for characters is eliminated when producing ciphered keys based on safe lock up tables derived from other secure lock up tables. Instead of the standard method's 5*5 matrix, the suggested updated Play fair matrix has 4*16 rows. Converting Amino Acid (protein) codons to one special symbol; resulted a reduced encrypted text to 1:3, this resulting in a threefold reduction in memory storage compared to the original representation of protein codes. The suggested technique overcomes the restriction of encrypting any data type (Arab, ,Russian, English, Roman language, integers, and symbols) in a few milliseconds encryption/decryption procedure.

In 2018,Krishna,etal.[20], As an advanced crypto analysis model, Pseudo biological cryptography was presented. Through biological structure, a new cryptographic model called a central dogma was introduced to secure information. The message is transformed to DNA, MRNA, and TRNA standards. At each stage, a portion of the transformed message is sliced in an asymmetrical fashion to generate a random key. The genetic information is used in the sliced unsymmetrical key generation technique. To increase the level of security, the technique entails splicing the message and producing several random sequences of keys from various phases. The Algorithm makes use of a less of public key infrastructure.

In 2019,BMKRISHNA[21], Presented a partial configurable area biological symmetric encryption algorithm based on central dogma. DES Algorithm and DNA Cipher are used in the encryption/decryption process, followed by a symmetric key scheme and DNA cryptography. There are two stages to the suggested technique: The cipher is created in the first stage using the standard DES method. PR is used to generate the key that is used to generate the cipher, and this key is then encrypted using the tradition key. The encrypted key and cipher are then submitted to DNA computing, followed by the amino form, in which the cipher is displayed in the form of non-breakable proteins.

In 2022,Farri and Ayubi[22], proposed a method for digital video watermarking. The method based on chaotic systems, Cellular Automata (CA), and DNA sequences. The proposed approach consist from contourlet transform (CT) and single value analysis (SVD) methods with insertion requirements in low- frequency sub-bands as a practical application of copyright protection for watermarked digital videos. The DNA-based on 1D chaotic automata is built with applying many CA bases in parallel.

In 2018,Mohamed,etal.[23], proposed an algorithm for watermarking using Discrete Wavelet Transformation (DWT) and Singular Value Decomposition (SVD) to probing imperceptibility. For copyright protection and identification, biometric watermark based on DNA sequence-based image has been used, where the DNA sequence represents a unique biometric characteristics that is hard to replicate or duplicate. the watermarked image was tested at various DNA sequence-based images by applying peak signal to noise ratio (PSNR).

In 2017,Hamad,etal.[24],Introduced a new DNA watermarking approach in this research that is used to insert binary bits into DNA sequences. The proposed method alters the codon postfix based on embedded bit value. This approach was applied on a collection of DNA sequences for evaluation, and mutation resistant found from the extracted bits. The proposed DNA watermarking approach was found to be secure, invisible, resistant to biological functions, and preservative.

In 2015,Iftikhar,etal.[25], Suggested employing watermarks as proof of data ownership to allow permitted access to genetic codes. The majority of DNA-based watermarking systems alter genetic data, makingthem sensitive to data loss. Distortion-free techniques ensure that no changes are made during the watermarking process. Vulnerable to malicious attacks and therefore cannot be used to secure property (especially in the case of a threat model). A watermarking technique having the aforementioned features has been proposed in this approach.

In 2015,Agbaje,etal.[26], Introduced a digital watermarking to the cyber watermarking method, which focuses on information theft (identity &credit card theft). It is made up of an embedding algorithm for merging a picture with a sample of the owner's fingerprint. A watermarked image is created by combining them. The embedding and extract the finger print are used a same secret key. To compare the two signals, a comparator is used to scan the database. For every transaction to proceed, the decision to match or not match must be made.

In 2022,Öksüz,etal.[27],Offered a unique watermarking method for sequential genome data. To achieve this, two criteria were used: first, embedding robust watermarks such that malevolent attackers cannot modify the watermark and are easily identifiable. Second, achieving local differential privacy in all data sharing (genome data) with service providers in order to identify hostile reporters and service providers that aims to share genomic data without users' consent and unnoticed.

1.3 Problem Statement:

During the transmission process over the Internet, varieties of methods are employed to protect information and data from theft and penetration for diverse reasons. These tactics are matched by methods equal or parallel to them by force of theft and penetration, i.e, it is a stable process in finding trends and techniques that make the process of sending information safer and more reliable, one of these trends being the adoption of bio-information.

1.4 Thesis Aim

The goal of this thesis is to build a new DNA watermarking model based on two strategies: Docking and central Dogma. The proposed model includes:

a) DNA Watermarking Generation using:

- Docking technique and DNA Seq.
- Docking technique is based on a function called scoring function parameters (concealing key and measuring key).
- Hiding the encrypted message based on rRNA and tRNA Strategies.

b) Central Dogma Encryption based on:

- Using two processes technique ;Transcriptionand Translation.
- Dynamical generated encoding table to DNA
- Quadric representation based on codons and amino acids.

1.5 Thesis Layout

Additionally, to chapter one four chapters will be introduced:

 Chapter Two: Covers the theoretical background of information hiding concept and its basic techniques, cryptography bioinformatics techniques such as DNA, Central Dogma algorithm, Docking.

 Chapter Three: Covers the proposed system algorithms.

 Chapter Four: Presents the implementation and result analysis

 Chapter Five: summarizes the conclusions and future research suggestion

CHAPTER TWO

THEORTICAL BACKGROUND

Chapter Two
Theoretical Background

2.1 Introduction

The theoretical basis of the proposed work is presented in this chapter. This background will explain all the deals with theoretical background subjects that are related to our work, such as Bioinformatics, central Dogma, docking technique, cryptography, and digital watermarking disciplines.

2.2 Bioinformatics

The discipline of bioinformatics is a diverse discipline that provides strategies and technology techniques for analyzing biological information. Pauline Hogweed and Ben Hesper [68] developed the term bioinformatics to characterize "the investigation of informatics procedures in biology," It was first used when the first biology sequence data was shared . Bioinformatics hasgrown as a field breakthroughs in molecular biology and computer science, which have combined computer science, biology, mathematics, information engineering, , and statistics into a single discipline that can be usedto analyze and understand biological data. The bioinformatics integrated disciplines are depicted in Figure (2.1) [28.29].

The demand for Bioinformatics has emerged of the current expansion of available genetic material, such as the Human Genome Project, to enhance gene assessment and taxonomy and advancement to work effectively on safe and effective medication formulations and reduce the amount of time it takes to develop a pharmaceutical manually. Bioinformatics arose from the need to comprehend the DNA molecule, commonly called the code of life. Enormous DNA sequencing studies had become an actuality with the introduction of next-generation techniques, contributing to the advancement of bioinformatics science. Bioinformatics is studied in a variety of methods. Some people are working on new computational tools for improved processing and handling of biological information, both in terms of software and hardware. Others regard bioinformatics as the study of biology through informatic information services and system methods [31].

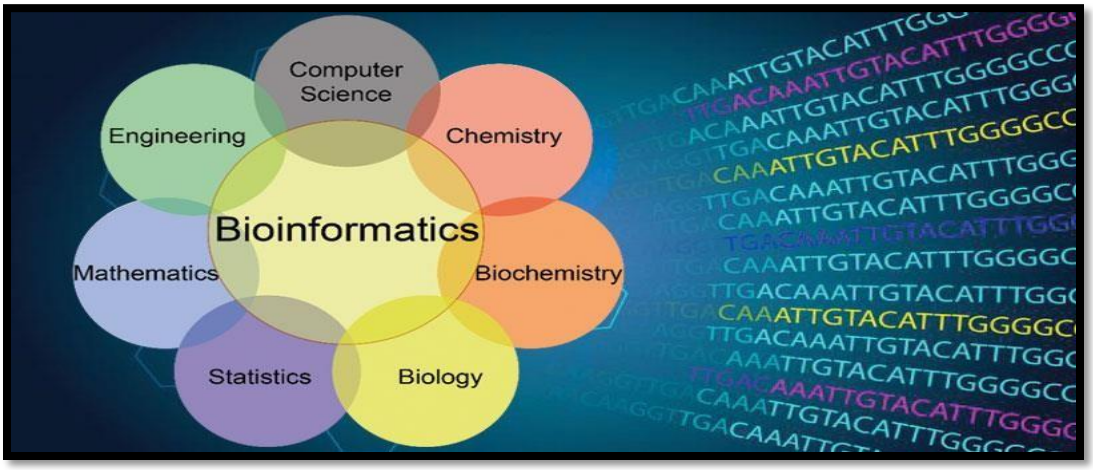


Figure (2.1): Bioinformatics disciplines [30]

2.3 The Central Dogma Technique:

The main central dogma of molecular genetics is a description of how genetic material moves through a biological system[32]. Figure (2.2) illustrates the central dogma's progression (a.the general structure of central dogma processes, b. description of each process in central dogma). In the most common living organisms, the central dogma is a framework for understanding the passage of sequence data between biopolymers that carry information in nearly all of the most common living organisms. The Dogma usually declared as "DNA forms RNA, and RNA forms protein,". This, however, indicates that once "data" has transferred into protein, it cannot be retrieved[35].

DNA → RNA → Protein.

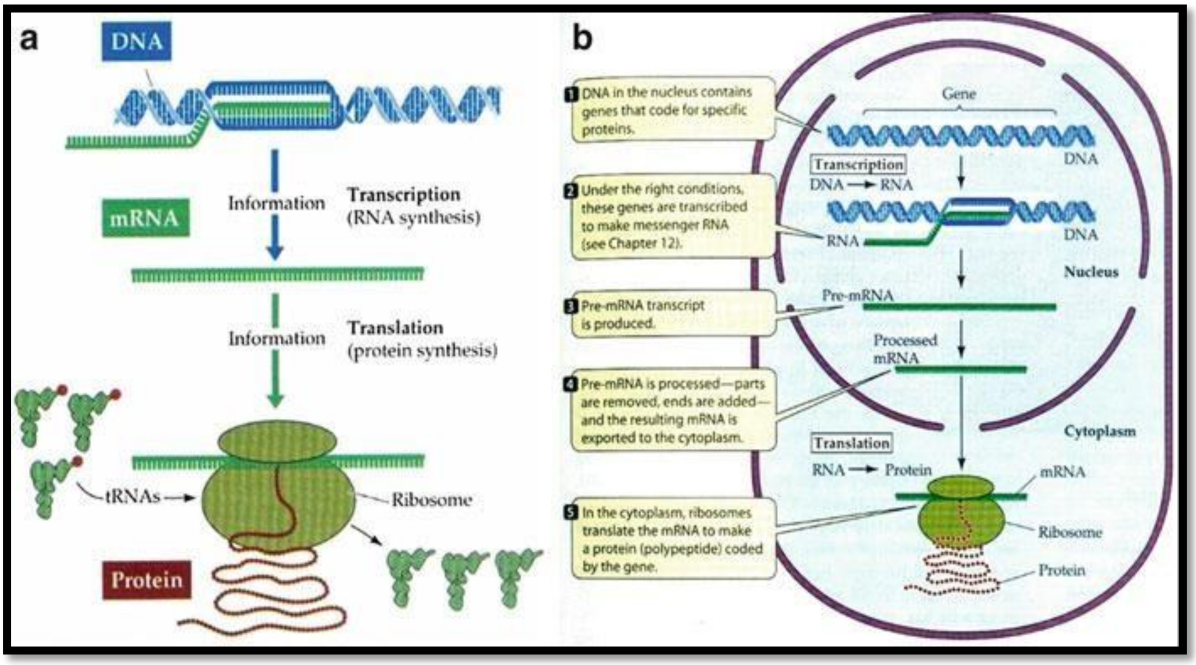

a
DNA
mRNA
Information
Transcription
(RNA synthesis)
Information
Translation
(protein synthesis)
tRNAs
Ribosome
Protein
b
1 DNA in the nucleus contains genes that code for specific proteins.
2 Under the right conditions, these genes are transcribed to make messenger RNA (see Chapter 12).
3 Pre-mRNA transcript is produced.
4 Pre-mRNA is processed—parts are removed, ends are added—and the resulting mRNA is exported to the cytoplasm.
5 In the cytoplasm, ribosomes translate the mRNA to make a protein (polypeptide) coded by the gene.
Gene
DNA
Transcription
DNA → RNA
RNA
DNA
Nucleus
Pre-mRNA
Processed mRNA
Cytoplasm
Translation
RNA → Protein
mRNA
Ribosome
Protein

Figure (2.2) Central Dogma Diagram [33]

Watson's edition varies from Crick's in that Watson depicts the basic dogma as a two-step (DNA → RNA and RNA → protein) pathway. While Crick's version of the dogma is still accrual today, Watson's version is not. Biopolymers are divided into three categories: DNA, RNA, and protein. There is a total of 3 × 3 = 9 possible fast data flows between them [34]. The dogma types organize into threecategories of 3 as shown in Table (4.1) [35].

General transfer of biological sequential information

General	Special	Unknown
DNA → DNA	RNA → DNA	Protein → DNA
DNA → RNA	RNA → RNA	Protein → RNA
DNA → Protein	DNA → Protein	Protein → Protein

Table (4.1): 3*3 Groups of transform information [35]

The three major transferring (suspected to happen across most cells), the three special transferring (known to be involved, but only under specific circumstances in the instance among some viruses and in a laboratory), and the three unexplained transports (suspect never to happen) [35].

The ordinary moving characterize the natural pass of biological data: DNA is duplicated by replication process. The DNA sequence transcript to mRNA, the proteins is produced through the co-operated between mRNA, tRNA, amino acid table by a translation. The specific transport report : RNA forms from coping of RNA (replication of RNA), RNA replication, DNA creation utilizing a template of RNA (reverse transcription) and proteins can be directly produced from template of DNA without needing of mRNA. The undiscovered moves include protein copying from another protein, The fundamental structure of the proteins can be used a template for synthesis of RNA, and creation of DNA utilizing the basic structure of a protein as a template - none of which is considered to occur naturally[36]. There are three main types of biopolymers. The three major classes of biopolymers will be discussed in the following sections.

2.3.1 Deoxyribose Nucleic Acid

Deoxyribonucleic acid is abbreviated as DNA. It is a long-structured molecule that contains the genetic code of any living creature. Just like an instruction manual includes the stages and rules for any procedure, the DNA contains the direction for a whole the proteins in any living body. This specific code stores all of a living being's features, i.e. DNA makes each individual unique, and this individuality is found on DNA and passed down the generations from parents to offspring and so to following hierarchies. Because no two people are alike, everyone will have their own DNA structure; even twins have different DNA structures. DNA is responsible for an organism's biochemistry since it contains genetic information and governs the generation of proteins. A phosphate group, a pentose sugar (ribose sugar), and a base make up every nucleotide. There are four types of bases: Adenine (A), guanine (G), cytosine (C), and thymine (T)[37]. The DNA nucleotide structure and components are depicted in Figures (2.4) and (2.5).

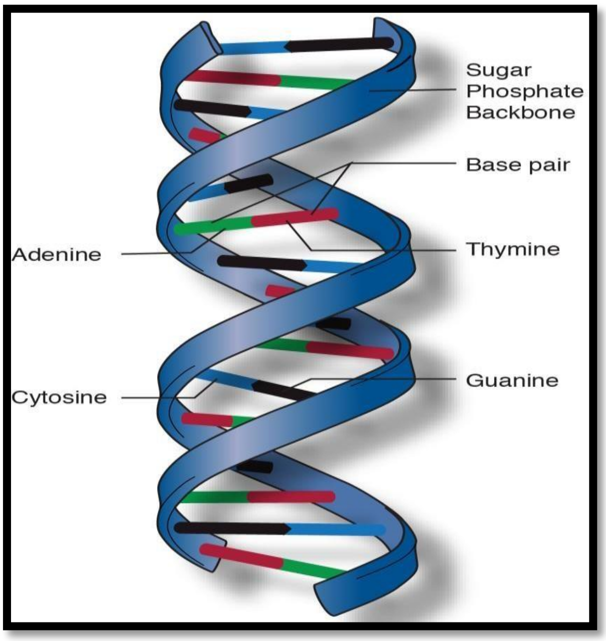


Figure (2.3): DNA Nucleotide Structure [38]

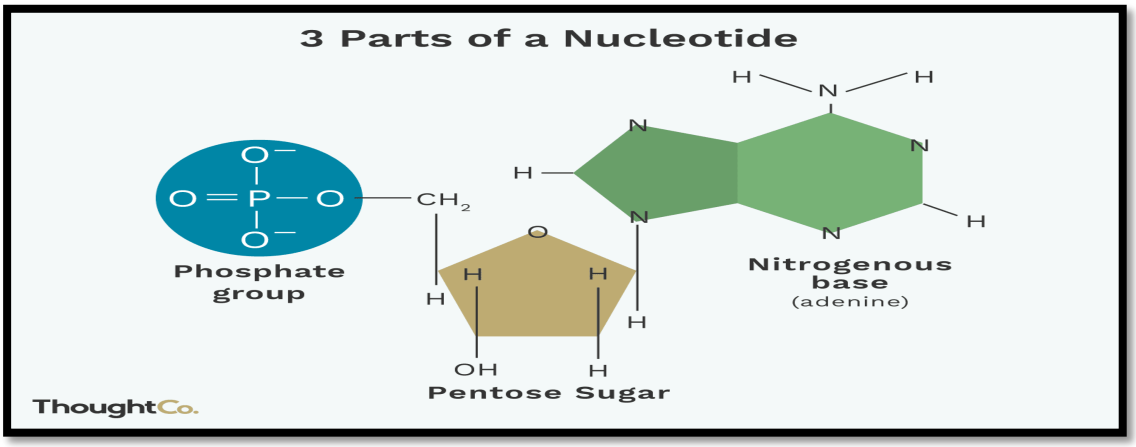


Figure (2.4): Nucleotide Parts [39]

DNA is made up of two strands that run in opposite ways. Each strand's backbone is made up of pentose and phosphate groups. Purines and pyrimidines form hydrogen bonds that bonds the two DNA strands together, forming double helix. A base A always couples with a base T on the other side by hydrogen bond, and a G with C. This binding is known as base pairing[40.41].

2.3.2 Ribonucleic Acid

The other kind of nucleotide termed ribonucleic acid (RNA). The nucleotides of ribonucleic acid (RNA) are similar to those of DNA, except the base T is replaced by a Uracil (U), as seen in figure (2.6). In most cases, RNA is a single strand. The base-pairing pattern can Become a, U-A, G-C, and C-G when an RNA strand mates with a DNA strand[40,41]. The RNA structure is shown in Figure (2.6).

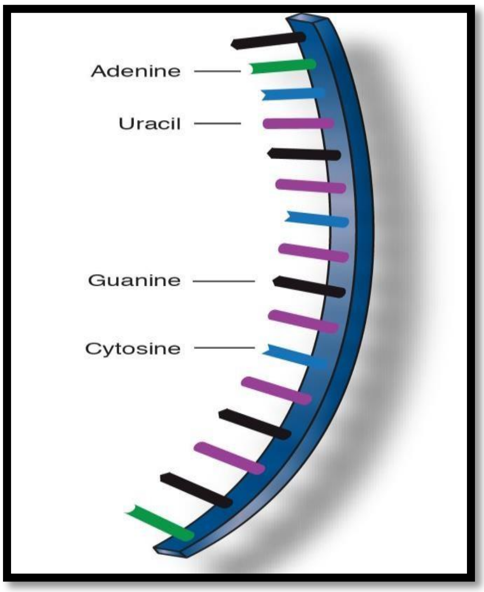


Figure (2.5): RNA structure [41]

RNA is a single-stranded molecule made up of ribonucleotides connected by phosphodiester linkages. Ribose (a pentose sugar), with four (A, U, G, C) bases, and a phosphate group make up the ribonucleotide in RNA chain. The main difference of sugar structure in DNA and RNA is: in DNA, the sugar provides additional stability to DNA, making it better for storing genetic material, but a relative instability to RNA makes it better for short-range activities[41].

The Uracil pyrimidine that is particular to RNA is utilized instead of thymine in DNA because it makes a complementary base pair with adenine. RNA has a diversity of shapes , involving ribose ribonucleic acid (rRNA),transfer ribonucleic acid (tRNA) and messenger ribonucleic acid (mRNA)[40,41].

2.3.2.1 *Messenger Ribonucleic acid (mRNA)*

The name messenger RNA was proposed by Jacob and Monod[42] for the RNA that contains data for the synthesis of proteins from the DNA (genes) to the protein production sites (ribosomes). The length and molecular weight of messenger RNA (mRNA) are highly variable [42].

2.3.2.2 *Transfer Ribonucleic Acid (tRNA)*

The transfer RNA (tRNA)is another frequent RNA found in cells (tRNA). tRNA is a tiny RNA with a molecular weight of approximately 25,000 to 30,000. tRNA is a vital constituent of the protein synthesis, that transports amino acids to mRNA through the process of protein synthesis.A certain tRNA take every amino acid. There are up to 20 amino acids. Therefore, there are at least 20 different forms of RNA. Several situations have proven that every amino acid has at least two kinds of tRNA. Its molecules outnumber amino acid types. tRNA is produced On a DNA template in the nucleus. tRNA is coded for by only 0.025 percent of DNA.

Only a small portion of the DNA molecule is used to make tRNA. As a result, there are no evident base links to DNA. A single strand coiled around itself makes up the tRNA molecule. In the cell, there are at the minimum twenty different types of trauma, each of which transports a certain amino acid to the place of translation. The short single-stranded RNA molecule folds into a specific shape to generate the L-shaped RNA[43]. Figure (2.7) shows L shape of tRNA.

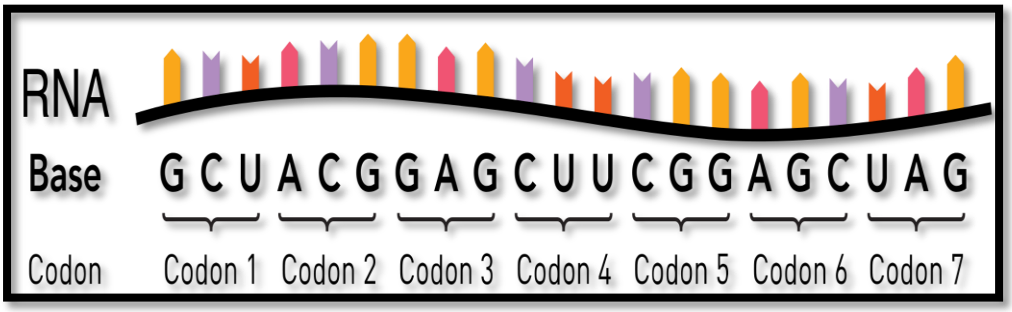
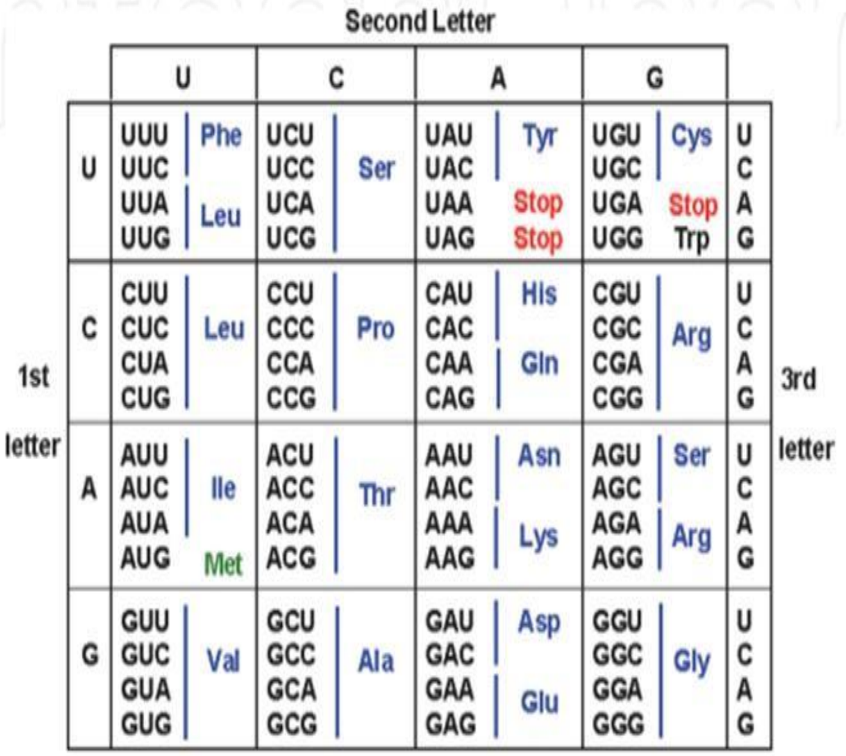


Figure (2.6): L-shaped tRNA[44]

- ***Amino Acid***

Every amino acid possesses a unique side chain and is abbreviated a three-letter as well as a one-letter symbol. The standard amino acids are 20 amino acids to synthesis proteins. The basic compositions of all 20 amino acids are shown in Figure (2.8)[45].



1st letter	Second Letter								3rd letter
	U		C		A		G		
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

Figure (2.7): Formation of Amino Acids Cracking [45]

2.3.2.3 *Ribosomal Ribonucleic Acid (rRNA)*

The ribosomal Ribonucleic Acid (rRNA) is presented in the ribosomes, as the name proposed. The ribosome composed of RNA and proteins. The percentage of RNA in the cell about 80%. The RNA sequence is generated by complementary to DNA region from which is generated. rRNA is a single strand that is twisted on itself at some places. It posses s helical areas that are joined by intermediate single-strand regions. The majority of base pairs in the helical region is the complementary and connected by hydrogen bonds. The bases have no complements in the unfolded single strand areas[47,48]. To illustrate the rRNA structure, figure (2.9) explain that.

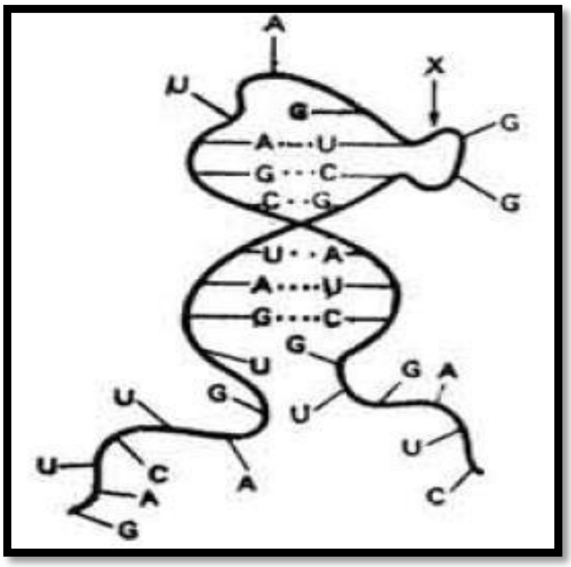


Figure (2.8): The structure of ribosomal RNA (rRNA) [48]

2.3.3 Protein

The protein is a long chain of amino acids linked together , as shown in Figure (2.10). The order in which distinct amino acids are linked together to make proteins determines their shape and function. The 20 standardamino acids are found in all proteins, regardless of their shape or function[49].

In translation process, four bases (A, U, G, and C) of mRNA read three basses known as codon to decided the direction in which amino acids are incorporated in to a protein[49,50].

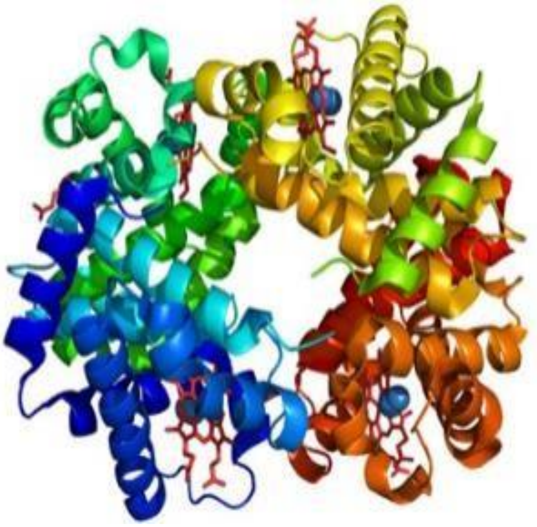


Figure (2.9): Protein structure [50]

The mechanism of protein synthesis involves two processes, as shown in Figure (2.11).

1- The Transcription Process: (DNA) ⟶ (RNA)

2- The Translation Process: (RNA) ⟶ (Protein)

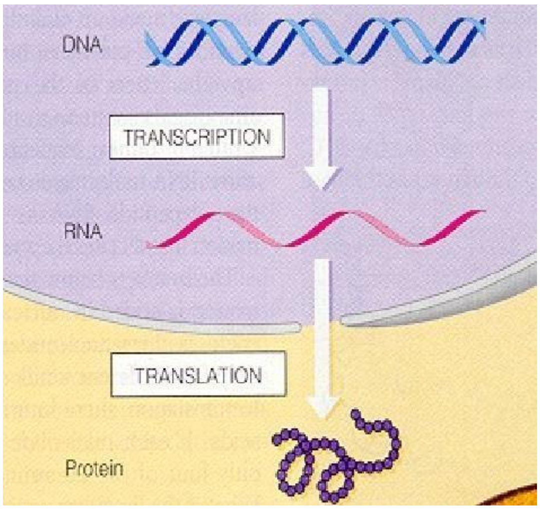


Figure (2.10): Protein Synthesizing Processes [50]

2.3.4 Central Dogma Processes

To synthesis proteins, two main processes are included they are:

2.3.4.1 *Transcription Process*

Transcription is the mechanism in which the data that found in the region of DNA is duplicated to form new assembled sections of mRNA. The new molecule of RNA is generated by inserting complementary nucleotides into the template strand. RNA varies from DNA synthesis in that just one DNA strand can use a template to form mRNA[51] as seen in Figure (2.12):

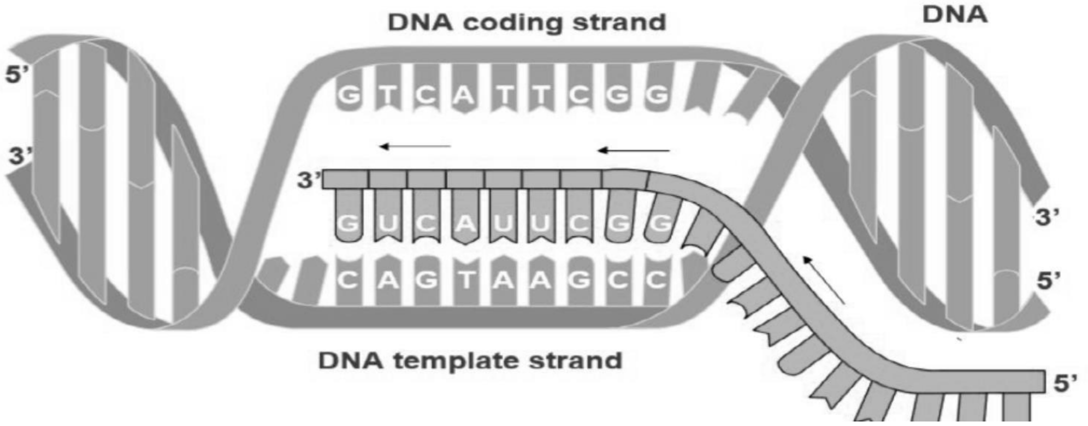


Figure (2.11): Transcription Process [52]

2.3.4.2 *Translation Process*

Transfer RNA (tRNA) is included in protein synthesis, although to messenger RNA (mRNA). The translation is a process where mRNA guides the synthesis of protein with the aid of tRNA [51.52]. As demonstrated in Figure(2.13), the procedure of translating a mRNA' sequence and producing the amino acid sequence of a protein is known as translation.

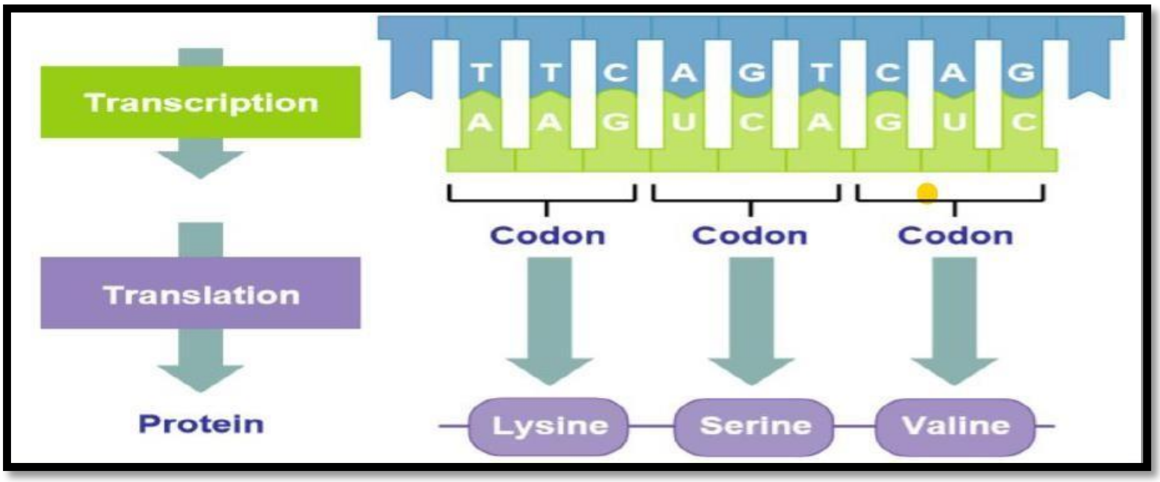


Figure (2.12): Translation Process [52]

2.4 Docking Technique

Docking is the process of attempting to identify the best possible fit between two molecules. It can also be characterized as a strategy that predicts one ligand's preferred orientation when attached in an active site to create a constant complex with the lowest total energy [53], as shown in Figure (2.14). Docking is a technique for predicting signal strength and kindness.

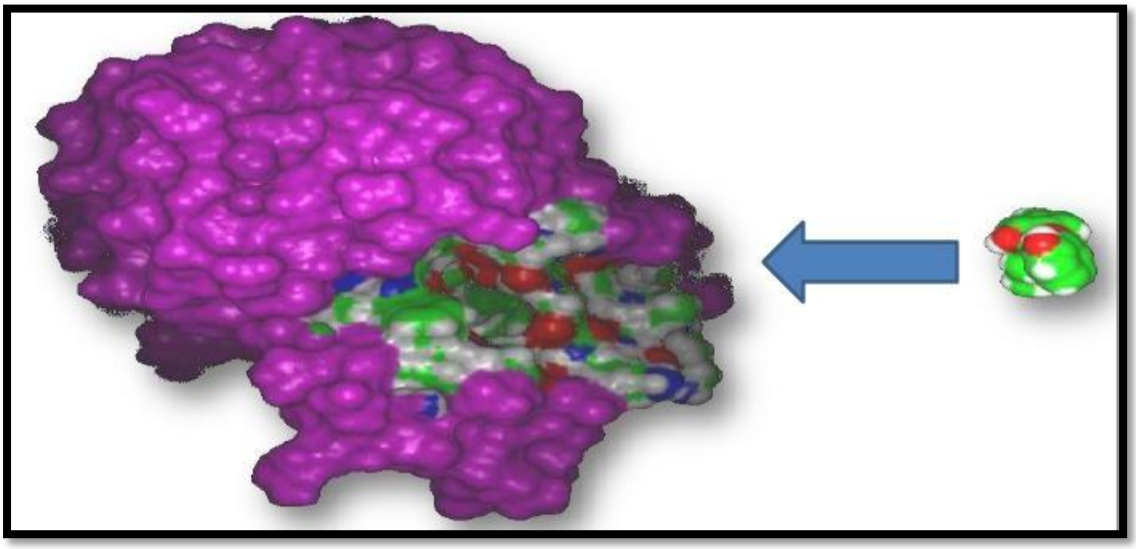


Figure (2.13) The process of Docking[53]

The goal of molecular docking is to find the best shape and relative orientation between the ligand of protein.

Scoring functions performed in docking applications form many supposition and facilitation in the assessment of modeled combination and do not completely account for a variety of physical processes that impact molecular recognition, such as entropic effects.

Scoring the assessment and classification of concluding ligand configuration is critical to the construction-depended actual checking. Even if interaction structures are accurately predicted, computations will fail if they cannot distinguish correct from wrong poses and if 'real' ligands are not identified. As a result, developing accurate scoring systems and schemes is critical [43].

2.4.1 Scoring Function

Evaluating the energy of pose can be obtained from prophesying binding affinity between two molecules and predict strength of intermolecular interactions between protein-protein, protein-DNA, and protein-drug[53].

The score function is the fractional derivative of the log-probability function, which is the standard probability function [53].

2.5 Securing Data Techniques

Because the Internet is an open and popular communications network, information security has become a significant matter. Although it is usually belief that connection is protected by information encryption, this has seldom been enough in practice. Although modern encoding techniques have started to evolve researchers have focused on techniques to hide messages rather than encrypt them. So, the study of communications security involves not only encrypts data but moreover hide their existence whose essence lies in information hiding [54].

2.5.1 Cryptography

Cryptography is known as the study of mathematical approaches connected to information security features such as integrity, confidentiality and authentication is known as cryptography. Utilizing competent methods so that the authorized person only can read this data with maintaining the characteristic of information security of data[55]. Cryptography is divided into two categories:

1. **Symmetric key Cryptography:** the key encryption and decryption methods is identical as illustrate in Figure (2.15) [55].

2. **Asymmetric Key Cryptography:** Encryption and decryption processes are performed using two distinct keys. A first key is refered to as the public key and utilized for encryption, while the second is refered to as the private key and utilized to decryption [56]. Both groups are explained in Figure (2.15 and 2.16).

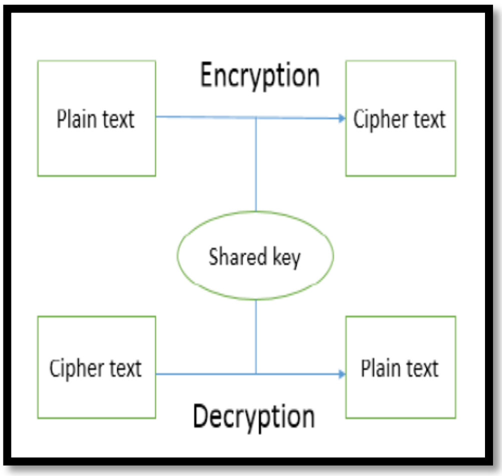


Figure (2.14): symmetric key keycryptography[55]

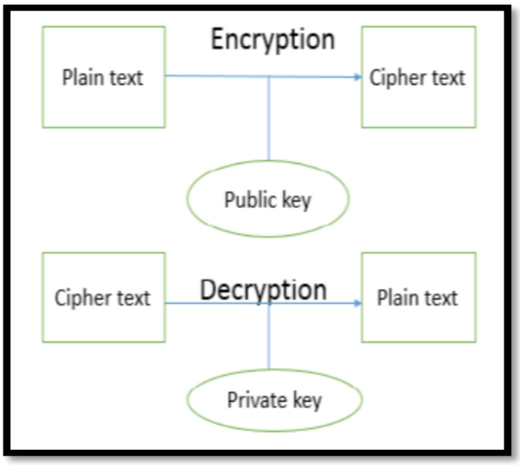


Figure (2.15:)The Asymmetric cryptography[56]

2.5.2 Information Hiding

As shown in Figure (2.17), the idea of data concealing is commonly characterized by a pair of algorithms: embedding and extracting an embedding algorithm is the process of combining two files, the secret information and carrier, plus an optionally key, to create a stego (the file containing the secret information). The extraction algorithm extracts the secret information from the stego file.

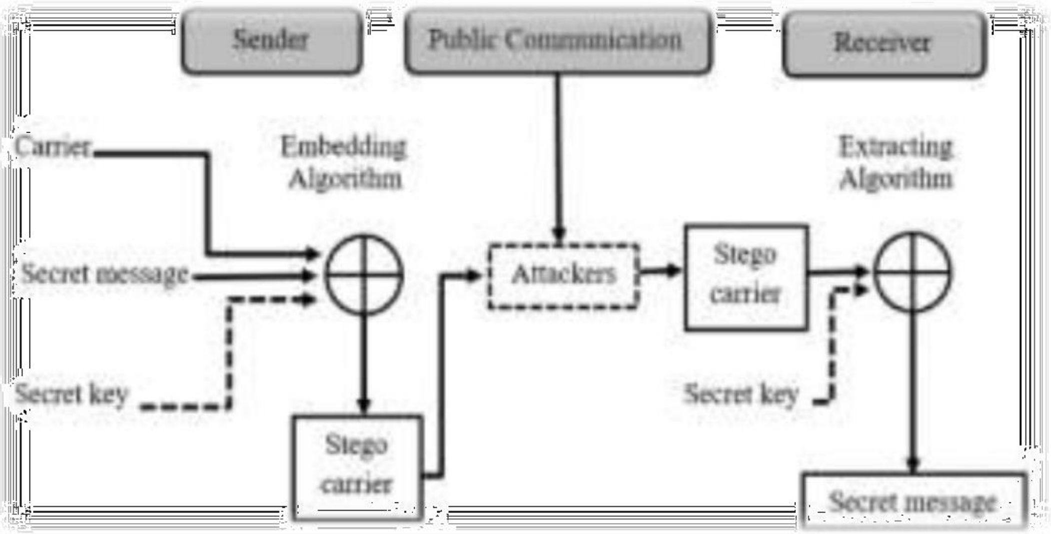


Figure (2.16) General Structure of Data Hiding System [20]

Pure data hiding is a kind of data concealing that does not require using a key; its security depends on the algorithm's privacy. As a result, it is seen as a less secure method. Another type of data hiding is secret data hiding, which uses a single key for both procedures: embedding and extraction. One of the most significant benefits of this kind is that it allows for a quick process in both processes. Unlike prior kinds, public data hiding employs two keys for both embedding and extraction: one is for embedding and the other for extracting. The major benefit of this kind is the system's resiliency; if one key is well knownby a third party, finding another key can be difficult. This kind, on the other hand, it is 100-1000 [57]. Times slower than personal data concealment. In systems of data hiding ,these applications are utilized to covering elements or carriers. Each carrier has its individual properties that aid in the data concealment process. The quantity of private documents necessary to hide data in every carrier is determined by the availability of the certain carrier's region. As a result, carriers are a crucial component of any information concealment technique. Text, video, photos, and DNA are examples of multimedia that can be used to hide data [58].

Text can be concealed by changing the design of the text by utilizing the nth characters from the text or by changing several of the parameters, like spacing, etc. The most benefit of this transmitter is that it doesn't take a lot of memory and is simple to transmit. When compared to other carriers, it possesses a tiny amount of redundant data[58].

The utilize of inaudible frequencies and a tiny alteration in the binary sequencing of an audio file can be used to hide data in audio files, the information hidden in various video frames. Because of their great level of redundancy, large capacities in images, minimal impact on visibility, and ease of manipulation, digital images have become attractive transmitters for hiding secret information. DNA is a relatively new carrier employed in the data concealing sector. DNA is the most important molecular structure in biology for encoding the information needed to build and manage all chemical components in the human body. DNA has been proposed as a potential candidate for applied in computational purposes[57.58].DNA used as a carrier in the proposal system. In general, there are two types of hiding data: steganography and digital watermarking.

2.5.2.1 *Steganography*

Steganography is the technique of writing messages that are notvisible to anybody but only the sender and the meant recipient. The concept identifies hiding sensitive information or secret messages in other media files, such as audio or video files (audio, texts, video, image, DNA) [59].

The efficiency and achievement of every steganography defined by: robustness, capability and imperceptibility. The term "robustness" refers to the length of time that the hidden information may be held, i.e., the amount of effort needed to break up the secret message without damaging the covering item. Imperceptibility refers to the ability to avoid detection, i.e., failing to realize the cover item and failing to find the hidden communication[59].

2.5.3 *Digital Watermarking*

Digital watermarking techniques are a growing topic in computer science, cryptography, signal analysis and networking. The authors of digital watermarking designed it to be a resolution to the requirement for value added security on best of information encryption and scrambling for content conservation. Digital watermarking, like most other emerging technologies, raises a variety of critical inquiry including the following[60].

2.5.3.1 *Digital DNA Watermarking*

Due to the obvious changeable of unidentified gene regulators and the insensitivity, the watermark is more successfully incorporated in the coding area than in the non-coding area. The watermarking of coding DNA sequences is discussed in this section.

2.5.3.2 *Coding DNA Watermarking*

DNA watermarked coding sequence accepts the inserted rules to replace the inserted codon with one of the codons amid similar codons. Figure (2.18) explains a sample of a watermark coding sequence in DNA. Excluding the first and end codons, this construction integral the two bits in the third base of four-fold similar codons. Likewise, in n-fold identical codons, nbits can be contained. Utilizing the DNA-Crypt algorithm, this mechanizim has been used in a DNA (steganography and watermarking).

DNA Watermark Coding converts the codons or nucleotides into a numeric value, and then uses the signaling steps to implant the watermark into the digital sequence. When analyzing a DNA signal, the four bases of the DNA sequence must be converted into integral, floating or complex numbers in order to process the signal [61].

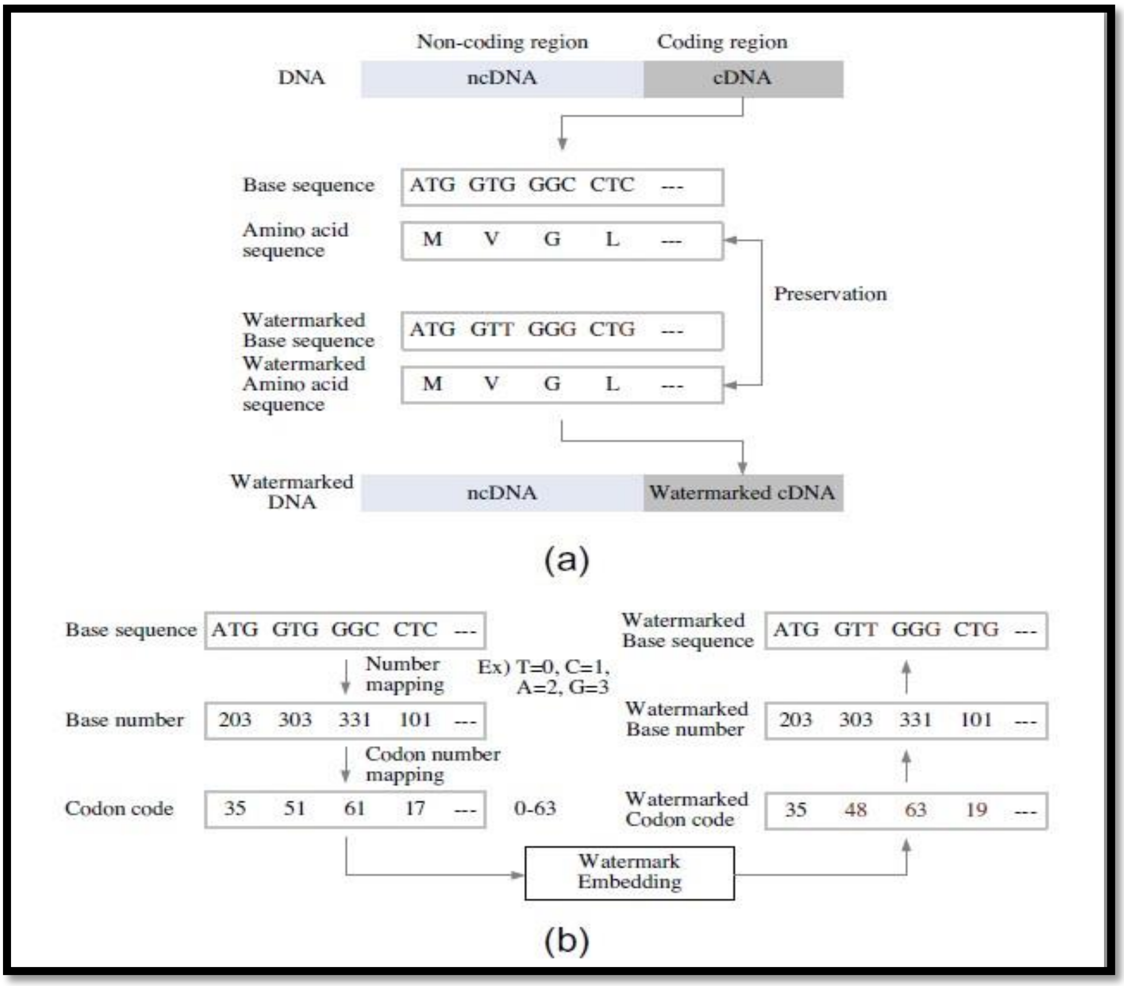


Figure (2.17: a: DNA watermark generation.b: Embedding DNA watermark) The example of constructing a DNA Watermark [57,58].

2.5.3.3 *DNA Watermarking Coding Requirements*

Amino acid preservation, Mutation resistance and privacy are the most requirements for a DNA watermark. The purposes of these necessities are comparable to that of other multimedia watermarks.

1-Mutation Resistance: The mutations are produced by introducing viral genes, radiation, transposons, and mutagenic substances and also mistakes throughout meiosis or republication of DNA [58,62]. Coding DNA watermarks must be to both purposefully acquired moreover natural and induced mutations. Mutations divided into Small-scale mutations influence one or a little nucleotide, but big mutations the structure of chromosomes. The previous involve point mutations, deletions and insertions, and the last involves many chromosomal repeats, deletions, and lack of heterozygosity. For copyright preservation or possession confirmation, DNA watermark provides a verification of the watermark in stolen DNA sequence information. During a disagreement appears across copyright or ownership of DNA sequence, Watermarked and pirated DNA is cognitively identical. The pirated DNA sequence will not be significant If the cognitively variance is large.

As a result, small-scale mutations at the single-base level can be only analyzed, as opposed to large-scale chromosomal mutations. The length of encoding DNA sections must be long enough to detect the watermark in the area surviving the large-scale mutation when large-scale mutation occurs. Chemicals or mistakes in replication of DNA can produce a point mutation, which modifies a single - base base. Point mutations involve silent mutations that express the same amino acid, nonsense mutations produce a stop codon which cut the protein and missense mutations which express a various amino acid. A transposable element (TE) frequently causes one or some of nucleotide insertions and deletions, which can change the reading frame and create a non - functional protein or produce such a fusion protein with modified, new, or no functionality. Point mutations can be reverted by either editing the code or repeating inserting the watermark. If point mutations happen at adjacent sites at the same time, however, extracting the watermark is difficult [62].

2-Amino acid preservation: The watermark in a genome is analogous to noise. Watermark alters the amino acid sequence, analogous to a missense mutation, the produced proteins may be dysfunctional. The procedure of inserting a watermark causes a meant silent mutation, Thus, the alteration span of condoms must restrict to the identical codons of every amino acid. The transparency of multimedia watermarking matches amino acids preservation [62].

3-Security: The watermark should not be taken or removed easily, and even if the watermark algorithm and biological study conditions are recognized. As a result, the watermark must always be combined into different targets or closely propagated as a Wide Spread Spectrum (WSS) in the frequency domain with the watermark encoding. Creating a specific watermark depends on different goals or positions[62].

4-Codon optimizing: When it comes to the expression of genes and genetic analysis, DNA watermarking should take the optimization of codons into account. Parameters for optimizing codons are codon utilization, codon content, GC content, DNA patterns and direct/reverse repeating[62].

5- Blind watermark detecting: In bioinformatics, a DNA watermark has remarkable benefits compared to a multimedia watermark. In general, digital watermarks can be categorized by recognition type as either blind or non-blind watermark. Non-blind watermark which needs the original information (non-watermark information) in identification methods and has greater stability. However, the multimedia watermark is rarely considered non-blind due to the leakage of the original data. DNA sequencing needs to understand the reference sequence[58.60.62].

2.6 Performance analysis

To measure the performance of the proposed system, several metrics used such as:

2.6.1 Blast Analysis

The "National Center for Biotechnology Information" (NCBI) site offers a series of biomedical and genomic information applications known as BLAST. In data base programs look for sequence similarities between proteins and nucleic acids (DNA). A number of definitional, algorithmic, and statistical enhancements have been proposed for protein comparisons. By using its site, the BLAST algorithms' execution time can be significantly reduced, while their sensitivity to low similarities is increased. BLAST has two basic terms, the firstis the input sequence of RNA (our selected Database) that is called *Query,*and the second is the reference sequence of RNA in another Database at BLAST site that matches our input query, it’s called Subject (consider the measure of similarities)[63].

2.6.2 Alignment Analysis

The technique or result of matching the genomic (DNA) or amino acid residues of two or more genomic data to reach maximum degree of identification and, in the case of amino acid sequences, conservation, in order to quantify the similarity and potential of homology. This technique is called Alignment[63].

2.6.3 Cracking Probability

There are around 163 million publicly available DNA sequences [64]. Asa result, the possibilities of an attacker making a precise estimation are

$$CP = \frac{1}{1.63 * 10^{8}}$$

Because the represented coding of A, C, G, and T yields diverse combinations of two, the likelihood of guessing the represented coding is $\frac{1}{24}$,The text and RNA are segmented using randomly generated key values, which gives the attacker with more information, the probability of predicting the text segmentation is $\frac{1}{2^{m}-1}$.

The predicting of guessing RNA segmentation (massage segments) is $\frac{1}{2^{s-1}}$

Using the experimental setup, the entire chance of finding the message concealedin the RNA sequence database is:

$$Cracking\ Probaility = \frac{1}{1.63 * 10^{8}} * \frac{1}{24} * \frac{1}{2^{m}-1} * \frac{1}{2^{s-1}} \quad \ldots\ (2.1)$$

Where:

M = codon length basis (codon representation).

2.6.4 Capacity

To measure the amount of hiding information inside the final database RNA (size of cover after embedding processes) in comparison with the size of the original message(size of plain text) use blow equation [65]:

$$Capacity = \frac{size\ of\ massage\ in\ bits}{size\ of\ cover\ in\ base} = \frac{\frac{1}{4} * |s1|}{s2} = \frac{1}{4} bpn \ldots (2.2)$$

Where:

S = length of database after hiding processes.

CHAPTER THREE
PROPOSED SYSTEM

CHAPTER THREE

The Proposed System

3.1 Introduction

This chapter presents the proposal DNA watermark technique principles clarified as algorithms, explanations, diagrams and figures. The DNA watermark is based on two biological techniques: ***Central Dogma*** and ***Docking Technique***.

3.2 Proposal system structure

The proposal aims to design a developed DNA watermark to preserve the selected DNA sequence ownership rights, provide biological data storage as large amount of data to be stored, and protect the genetic information. The proposed DNA watermark structure consists of an encryption process using the central Dogma technique and a DNA watermarkembedding process using the docking technique. Vitamin D sequence is the selected DNA sequence. With two sites: transmitter and receiver. Figure (3.1) illustrates the general proposed structure.

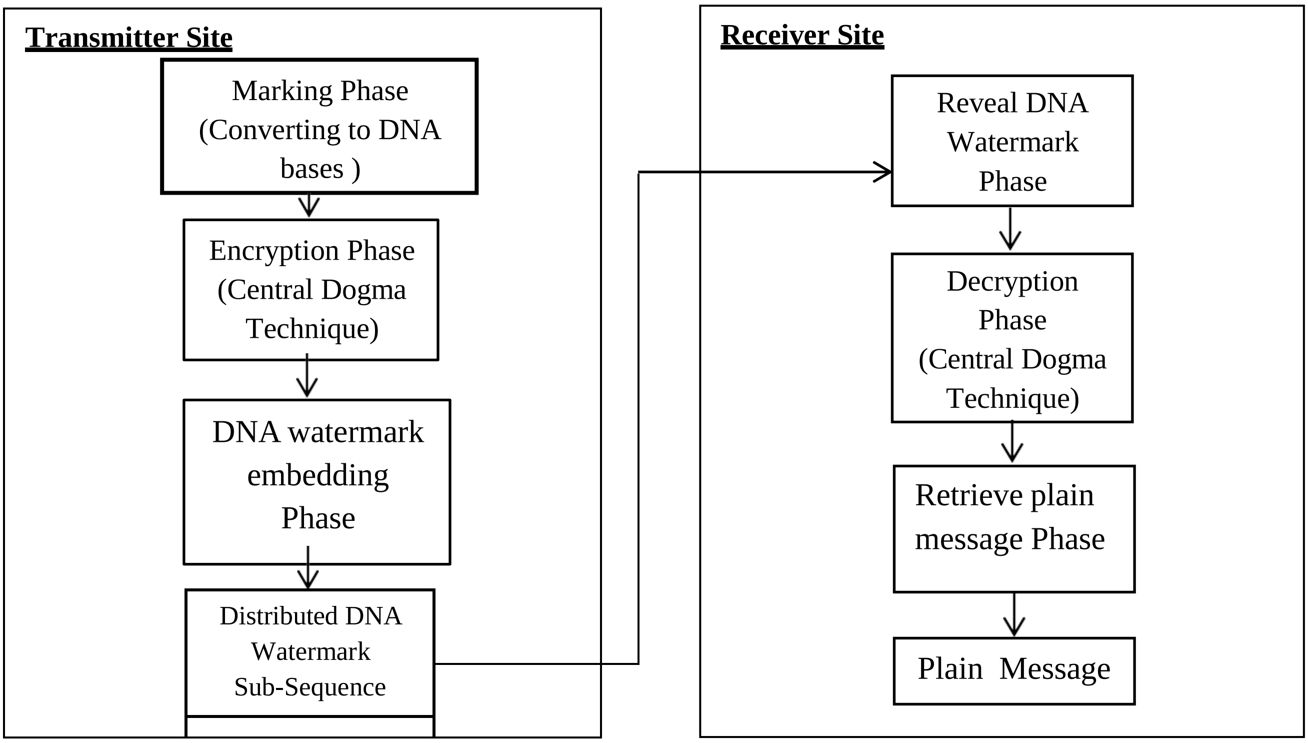

Transmitter Site
Marking Phase (Converting to DNA bases)
Encryption Phase (Central Dogma Technique)
DNA watermark embedding Phase
Distributed DNA Watermark Sub-Sequence
Receiver Site
Reveal DNA Watermark Phase
Decryption Phase (Central Dogma Technique)
Retrieve plain message Phase
Plain Message

Figure (3.1): general proposal structure.

3.2.1 Transmitter Site

On this site, several processes are executed to generate the DNA watermark.These processes will pass through three phases: (the marking phase, the encryption phase, and the embedding phase) as shown in Figure (3.2).

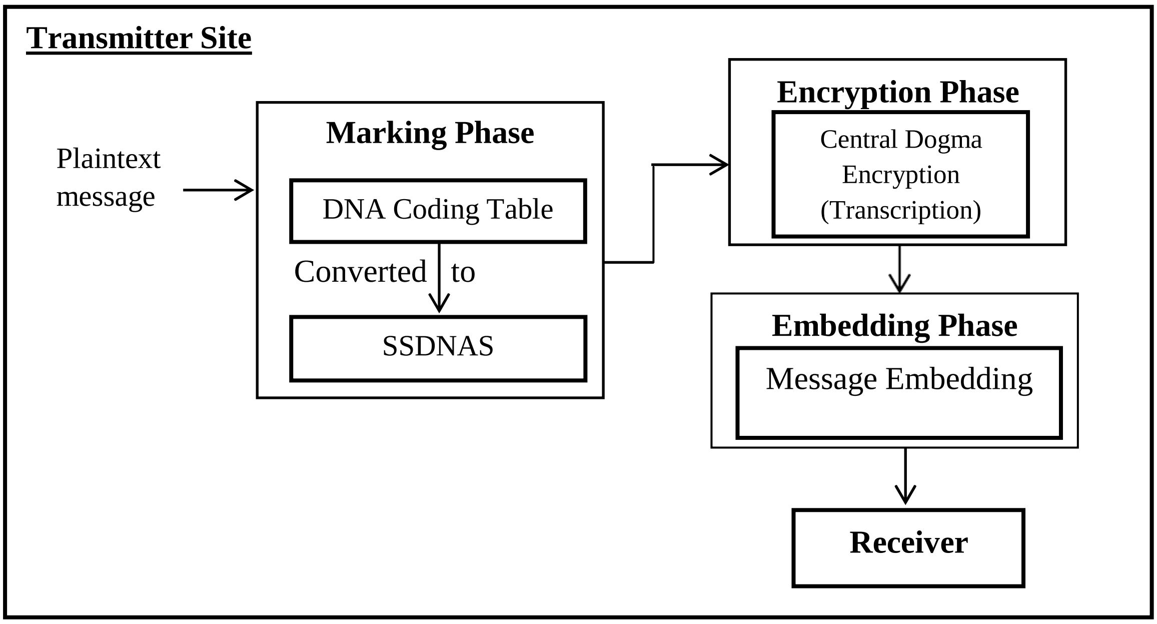


Figure (3.2): Transmitter Site Procedures

3.2.1.1 Marking phase

This phase consist of : generating (DNA coding table and DNAsingle strand). The description of these stages is illustrated bellow:

- DNA coding table Generation stage

This stage is responsible for generating a DNA watermark coding table for each plain message consist of {alphabets (A-Z, a-z), special thirty four characters (@#* %..), numbers (0-9)}. This stage represents one of the watermark data coefficients where each character in the plain message is assigned to a specific coding as a first step in generating a single strand of DNA. Algorithm (3.1) explains this process.

Algorithm (3.1): DNA codon table Generation

Input: characters (upper and lower cases, numbers and special characters), DNA bases of codons

Output: array of DNA Watermark quadruple codon table [A]

```
// Generate the DNA codon table
For the sample of codon = (ATCG)
    // Calculate the permutations of (ATCG) with repetition (for each base there are
       4 probability arrangement accordingly there are 256 probability in generation
       quadruple codons).
    For i=1 to 96
        Enter the generated codons in A[i]
    End for
    // Assign each character to a specific coding at A[i] table such that:
    Assign the first 26 coding for each characters of uppercase
    Assign the second 26 coding for each characters of lowercase
    Assign the remaining 44 coding for each special characters
End
```

The coding table is generated dynamically, i.e., the table is regenerated for each new message uniquely. For example, the character Z in some message represented by ATGC, but in other message represented by AAGG.

The natural representation of the biological codons is formed by three nucleotides (triangular base). In contrast, the codons represented by four nucleotides bases (quadruple base) form for the sake of the proposed goal. The table is up-to- date and can be developed according to the required structure.

- *Single Strand DNA Generation stage*

The DNA coding assignment process results convert the plain message to a DNA bases. These base are re-assembled to configure a single strand DNA sequence (SSDNAS).

3.2.1.2 Encryption Phase

The structure of the proposal is designed to obtain a developed DNA Watermarking method. The developing process begins with partial central dogma encryption. This partially comes from using only one of the dogma processes.

- *Partial Central Dogma Encryption Stage:*

This stage produces a ciphered message by applying central dogma technique using transcription process. Figure (3.3) illustrates the diagram of generating SSDNAS and the central dogma encryption process.

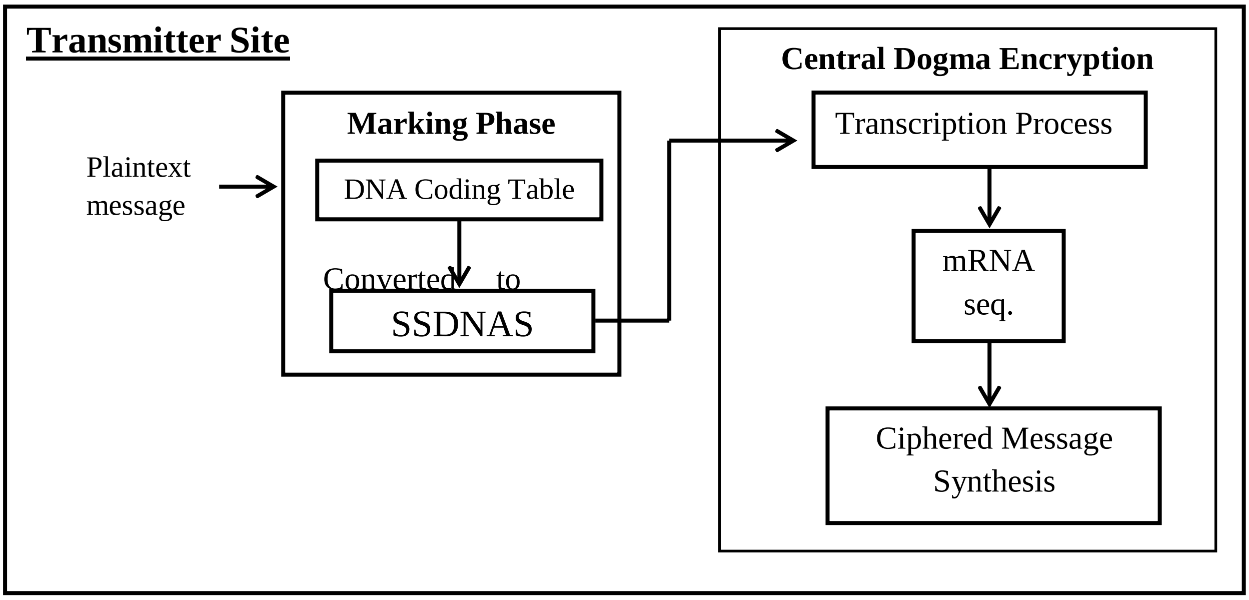


Figure (3.3): Central Dogma Encryption Stage

- *Transcription Process*

The transcription process is the first step in central Dogma encryption process. Such that, the generated SSDNAS is converted to messenger ribonucleic acid (mRNA), some conversion to the sequence is done: complement the SSDNAS and whenever a "T" base exists will convert to a "U" base. Algorithms (3.2) explains the procedures of generating SSDNAS and mRNA using transcription process.

Algorithm(3.2): SSDNAS generation and mRNA.

Input: ch (Msg. character), coding table (A), SS (Single Strand array),
Output: SSDNAS, mRNA sequence

```
//Apply Transcription process
read the length (L) of plain msg.
  for i:=1: L
     // Assign each ch (i) to specific codon according to (A)
        SS (i) = ch (i)
        SSDNAS = SS          //   SSDNAS is generated
  end // for
  complement SSDNAS
     // check SSDNAS for "T" base
    compute SSDNAS length (L1),
     for i:=1: L1
        if SSDNAS (i) ="T" then
             SSDNAS (i)= "U"
        end // if
    end //   for
mRNA sequence is generated
```

3.2.1.3 DNA Watermark Embedding Phase

In this phase the DNA Watermark embedding process is based on a docking technique, that advantage properties of docking coefficients to embed the ciphered message (mRNA) into D sequence. Figure (3.4) illustrate the block diagram of embedding process.

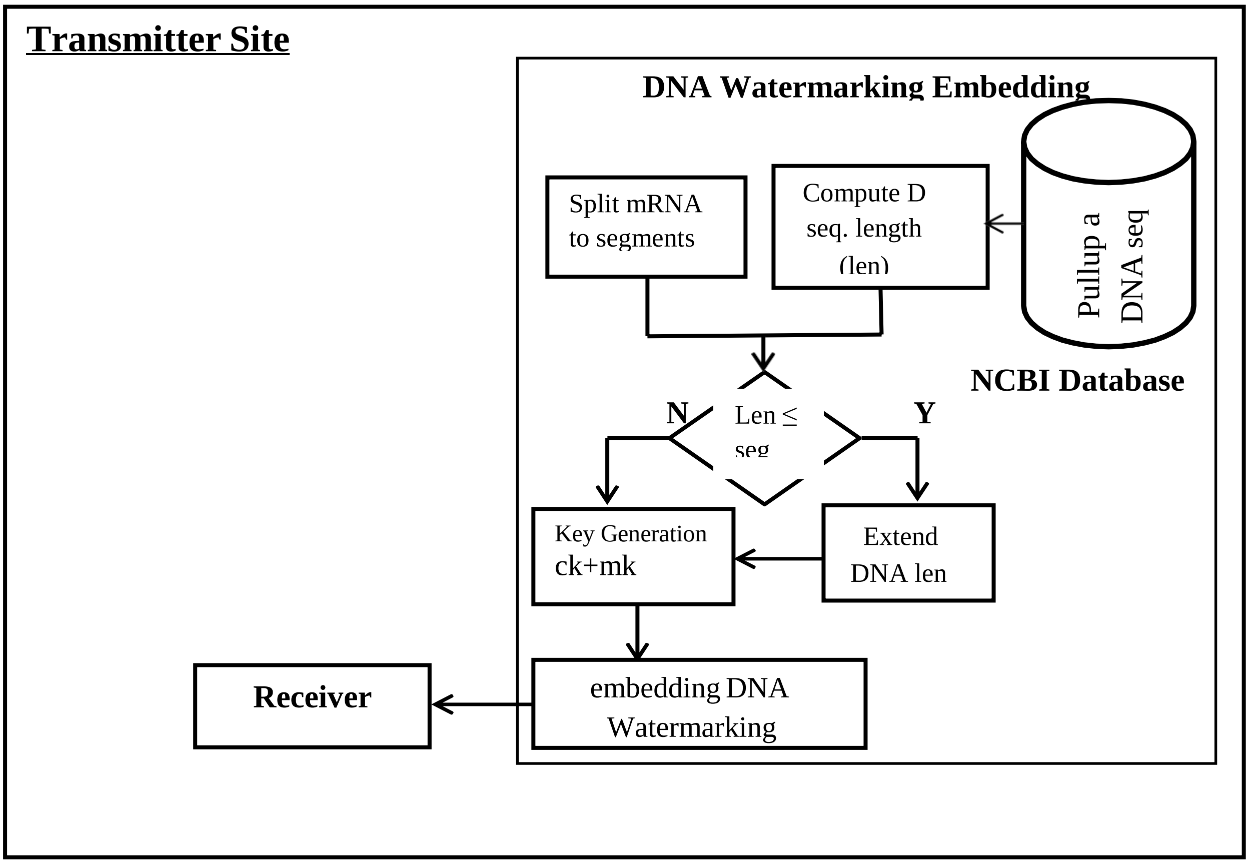


Figure (3.4): Block Diagram of DNA Watermark Embedding Process

To embed the ciphered message (mRNA) in the mark D sequence, many steps have to do such as segmentation, according to the segmentation step there will be a decision made to extend the selected sequence or not and finally: The docking step.

- *Message Segmentation step*

Processing this step results in a number of message segments being distributed randomly in the marked D sequence. To do this segmentation, a mathematical module has been built.

- Mathematical Segmentation Module

The module spilt the message to a number of segments according to the mRNA (ciphered message) length. The module will segment the message upon several length options:

1- If a message consists of 250 codons, less than 9500 characters, the number of segments will be:

$$\text{No of Section} = \left|\sqrt[2]{4 * 250}\right| = 31$$

2- If the length of the message is greater than 9500 codons or less than 95000,the number of message segments will be:

$$\text{NO of section} = |\ {}^{3}\sqrt{length\ of\ massage}\ \ |$$

3- When the length of the message is greater than 95000, the number of segments will be:

$$\text{NO of section} = |{}^{4}\sqrt{length\ of\ massage}\ |$$

4- Length of segments will be:

$$\text{Length of section} = |\frac{\boldsymbol{Length\ of\ massage}}{\boldsymbol{No\ of\ section-1}}\ | * 4$$

The procedure of executing this process illustrated in Algorithm (3.3).

Algorithm (3.3): Message Segmentation Process

Input :mRNA sequence (ciphered message)
Output :Segmented mRNA (Message), Length of Segment

```
// mRNA Segmentation  processes.
For each mRNA  do
  compute the numbers (N) of mRNA segmentation and the length of segment (len)
    N of segment =length of (mRNA)
    if $< 9500$ $then$
      N =$\sqrt[2]{Sec}$
   else
      if  $9500 > N \leq 95000$ then
         N =$\sqrt[3]{sec}$
      else
         If  N $> 95000$ $then$
            N =$\sqrt[4]{sec}$
           For all segments do
               Length of segment = | Length of massage/ No of section -1 | *4
           End for
             For each N do
                 Scoring process
                 Add the start and end code for each ck
                 Start code=AATT
               end code=GGAA
             end for
         end // if
      end //  if
  end  // if
end  //  for - Separation massage processes.
```

- Docking Stage

Docking technique locates the mRNA segments in D mark sequence. This technique has a dual process represented by selecting a proper place in D sequence for storing the mRNA segments and determining the ideal location in this place. These duties are done by a ***Scoring Function***. This function needs two parameters to be executed a concealing key and a measuring key that represents another type of watermark data coefficients.

- Concealing Key and Measuring Key Generation Process

To select a proper place in a D sequence using watermark data for hiding the mRNA segments a concealing key is needed. This key is used as an index for storing which segment in which location. For example, assume that the sum of the number of segments of the message is four, then ck will generate four random numbers in the range of number 4 as follows :{03,04,01,02} based on these numbers, such that; the third segment of the original message is hidden in the first place within D sequence while the fourth segment is in the second place, and so on until the last segment. The range of generating ck numbers within the range of 100 i.e. (0 to 99), these numbers are considered as an index.

To locate a proper location in the located place (ck) another key is generated called a measuring key . The mk ensures that the distances between segments within the D sequence are different. The steps of generating (ck) and (mk) are illustrated in Algorithm (3.4).

Algorithm (3.4): Concealing Key and Measuring Key Generation

```
Input    : len (Meg Seg ) value ,DNA coding table, integer  number
Output : ck and mk

    // ckey generation process
      let ck= len
      ck =random set of (len)
      if $0 \leq ck < 100$ $then$
        Assign each ck to a specific coding using (A) table
      else
         Ignore ck
    end if
    // Scoring process
     Insert the start and end code for each ck
    Start code= "aaccttgt"
    end code= "tcttgcaa"
    end  //  key generation process

    // mkey generation process
    for j = 0: 99
      mk = random set of (j)
    end  //measuring key generation
```

- *Masking Process*

Embedding the mRNA segments (message segments) in a D sequence starts whenever ck, mk, number of message segments, and size of DNA database already prepared. If the size of the D sequence cannot include the number of message segments, then an extended process to the size of the selected DNA sequence is required indeed. Algorithm (3.5) explain the extended process.

Algorithm (3.5): Extended Data Base Size.

Input: length of D sequence, Number of sections, length of segment.Output: selected extended Data Base.

```
    //Extended Data base processes.
    // For each new massage.
1: mRNA length =$\{No\ of\ Sec * Length\ of\ Segmant * 4 + 100 * No\ of\ sec\}$.
2: Factor of  DB size =Integer of (Massage length Require/length of D sequence).
3: Extended size of D sequence = Factor of DB size* length of D sequence.
4: End // of extended D sequence  process.
```

Masking the message segment is figured by adding additional codons at the begin and the end of the segment. These codons are called respectively start and end codons (another marking coefficients type to D sequence i.e. (DNAWatermark)). Embedding the mask segment depends on the scoring function value (ck, mk). For example, if we assume that the length of the cut-off segment is 8 and the number of codons in it is 4, the segment mask process and the embedding message in the D will be as follows:

GGGG	AUCGGCCC	CCGG	AUCGGGACUGCU	AUGC	CCGCUAACGGAUUUAC	UGUG
S codon	CK	E codon	Space of MK	S codon	Message Segment	E codon

The embedding process illustrated in Algorithm (3.6).

Algorithm (3.6): Mask Process

Input: ck, mk, D seq., mRNA segments
Output: Embedded DNA Watermark .

```
// Embedding Process .
1: start embed the segments in an order according to ck value
2: based on mk value store the message segment
3: End
```

Finally, the D sequence, codon table, ck, mk values will send on a secure channel upon agreement protocol between the transmitter and receiver.The flexibility of the proposed design can be exploited to give another direction of output, this output is about getting a ciphered message only by exploiting the complete central dogma processes.

- Fully Central Dogma Encryption:

This stage produces a ciphered message as a result of applying all central dogma processes (transcription translation) as shown in Figure (3.5).

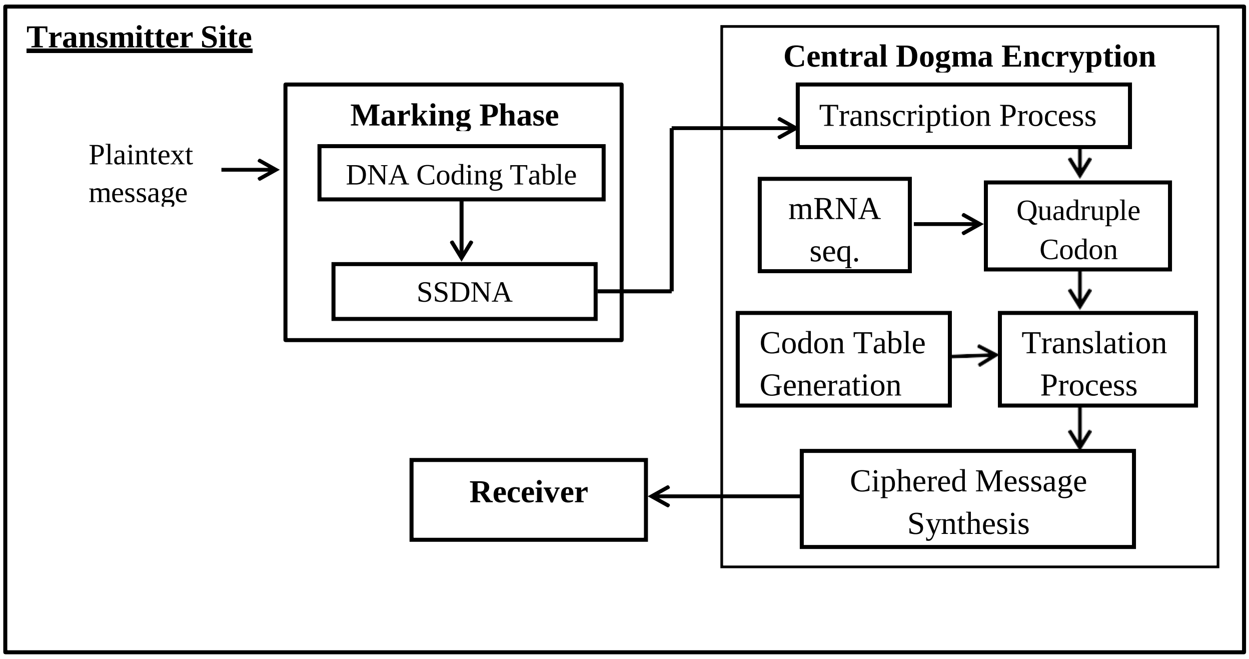


Figure (3.5): Fully Central Dogma Encryption Algorithm

The first process is transcription as explained in the above sections. The second is:

- Translation Process

The new mRNA sequence will be divided into codons of four bases. These codons will be translated to an amino acid using amino acid table generated in aspecial form that simulates the actual standard biological amino acid table. This translation is done by assigning each codon to its represented amino acid to synthesis the ciphered message by arrange these amino acid as a sequence build upon mRNA sequence . Algorithm (3.7) explains the procedure of this process.

Algorithm (3.7): Translation process

Input : mRNA sequence, amino acid table [b]

Output: protein synthesize

```
divide the mRNA sequence into subsequence of four bases called codons
 // search amino-table for a specific codons of mRNA via tRNA

For i=1 to 256
  For  j=1 to length (codons)
     If codon [j] = b[i,j]
        get the amino-acid via tRNA
     end
  end
end  // for
assemble the getting amino-acid in a sequence synthesizing a specific protein
```

The standard universal amino acid table contains 64 amino acids.However, only 20 amino acid were used. In Contrast, the new corresponding amino acids table is extended to 256 amino acids.

The actual codon representation as a biological form consist three bases, but for the sake of the proposal, the codon is represented by four bases.

3.2.2 Receiver Site

At this site, the receiver will reveal the DNA watermark using in an inverse steps.

3.2.2.1 Extract DNA Watermark

To reveal the DNA Watermark from a NCBI database called vitamin D, several processes have to take, such as searching for a DNA watermark to extract the hidden message. Figure (3.6) illustrates these processes.

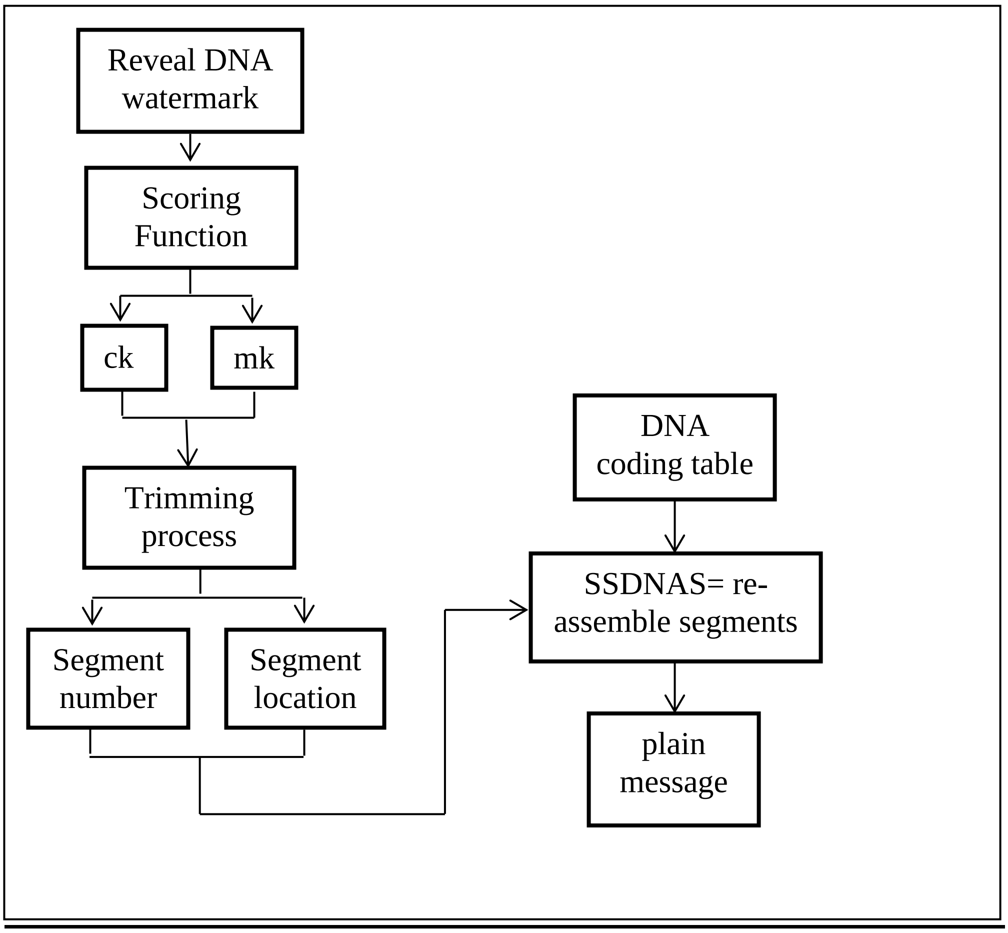


Figure (3.6): Extract DNA Watermark Phase

- DNA Watermark Revealing Process

This revealing process is based on scoring function values the (ck and mk). Algorithm (3.8) explains this process.

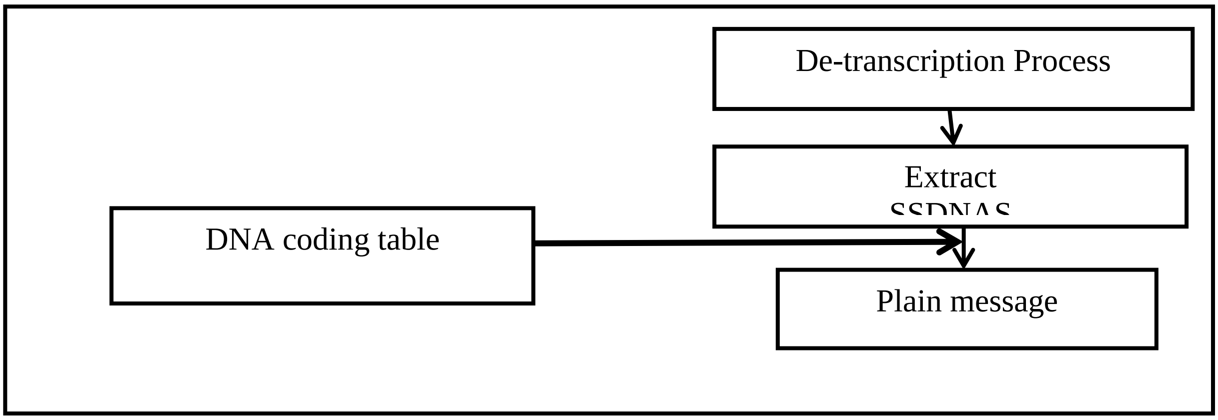
Algorithm (3.8): Secret Message Retrieving Process

Input: Vitamin D database, ck, mk, DNA coding table
Output: Pmsg (plain message)

```
// Search for Segments

for I = 1: len (D)
    get scoring function
    get ck values
     ck= trimming
      search trimming (ck) in DNA coding table
      ck = assign value from DNA coding table
      if resulted ck (value) = send ck (value)
          get mk value
          mk= trimming (mk from start & end)     // Segment trimming
         mk = assign value from DNA coding table
         if resulted mk (value) = send mk (value)
              extract the message segment (msg-s)
              p= assign (msg-s) in DNA coding table
              Pmsg = Pmsg + p
         else
                 send warning message " there are suspicious in mk value"
        end  // if
    else
             send warning message " there are suspicious in mk value"
    end // if
 end  // for
```

3.2.2.2 Partial Central Dogma Decryption

To re-synthesis the ciphered message using central dogma decryption, phase the following processes used. Figure (3.7) illustrate these processes.



Figure(3.7): Decryption Central Dogma processes

- De-transcription Process

The result [C] in the previous step will be transcribed to a form calledmRNA as explained in Algorithm (3.9).

Algorithm (3.9): De- transcription Process

Input: mRNA sequence
Output: SSDNAS

```
retrieve SSDNAS  as follows:
C=  mRNA sequence (RNA basese (A,U,C,G))

 // convert each U base to T base
for i=1: len (C)
    if  C[i] = "U"  then
        C[i] = "T"
    end // if
end   // for

 // apply complement process to each base in C

for i=1: len (C)
    S[i] = complement of C[i] according to table (3.2)
end   // for
S = SSDNAS
```

- Plain Message Retrieving

The last step for the receiver in getting the plain message is to apply Algorithm (3.10).

Algorithm (3.10): Plain Message Retrieving Process

Input: S (SSDNAS)
Output: P (plain message)

```
  // for each base in S

for i=1: len (S)
    P[i] = assign each base in S[i] according to able (3.2)
end   // for
resulted plain message (p)
```

ryption

In case of using another direction of output, i.e. a fully central dogma ciphered message only. In this case, the sequence of decrypted message begins with translation process, transcription to the retrieve the plain message; Figure (3.8) explain this direction.

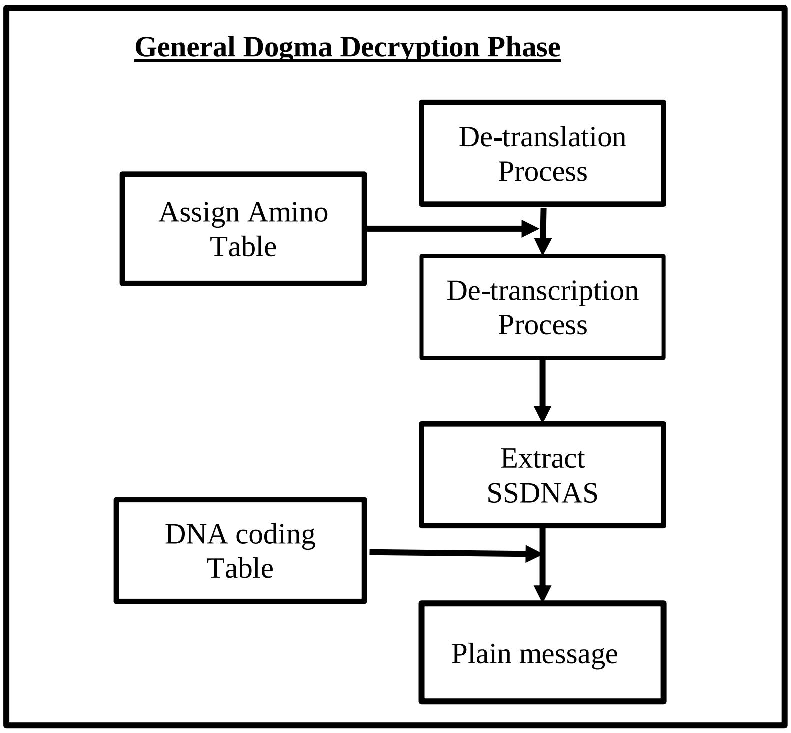


Figure (3.8): Full Central Dogma Encryption

-Translation Process

This process is a first step in retrieving the plain message as shown in Algorithm (3.11).

Algorithm (3.11): Re - translation Process

Input: C (sequence of e, Assignment table, k, As,COutput: Sequence of codons

```
retrieve the codons as follows:
k= amino acid sequence
As = assignment table
for i=1: len (k)
  for j=1: len (As)
    find k[i] in As[j]
        if  k[i] exist then
            C[i]= As[j, codon]
        end // if
    end   // for
end   // for
sequence of codons [C]
```

CHAPTER FOUR RESULTS AND DISCUSSIONS

Chapter Four

Results and Discussions

4.1 Introduction

This chapter introduces a discussion of the experimental work results of the proposed system. The system has been simulated using MATLAB program version R2016a on Windows 10 platform on Intel core i5. The experimental results were analyzed to clarify the results by some performance metrics such as cracking probability (CP), blast.

4.2 Proposal System implementation

The general follow of implementing the proposed system on both sites (senderand receiver) with their stages and the obtained results from each stage are explained below:

4.2.1 Sender Site

This site is responsible for generating DNA Watermarking securing procedures. The proposal includes two main stages: DNA Watermarking and central dogma encryption, described bellow.

4.2.1.1 DNA encryption table Generation Process.

A DNA coding table is generated to encode the plain message to DNA bases. The result of this process is illustrated in Table (4.1). The generated table contains: alphabets (upper case ,lower case), numbers, and special characters. On the other hand, each case, has its representation in a DNA coding.

Table (4.2): DNA Coding

CCAA	a	GACC	y	GTCC	W	GGCG	{
GCTA	b	GATT	z	GTTC	X	GGTC	[
CCCC	c	CACC	A	CTCG	Y	CGCG	}
CCTT	d	CATC	B	CTTC	Z	CGTA	]
TCCC	e	TACG	C	TTCG	0	TGCC	|
TCTC	f	TATC	D	TTTA	1	TGTA	\
ACCG	g	AACG	E	ATCC	2	AGCC	+
ACTC	h	AATA	F	ATTA	3	AGTT	=
GCGC	i	GAGG	G	GTGT	4	GGGA	_
GCCG	j	GACG	H	GTCG	5	GGCC	-
GCTT	k	GATA	I	GTTG	6	GGTG	)
GCAA	l	GAAG	J	GTAG	7	GGAT	(
CCGG	m	CAGA	K	CTGA	8	CGGG	*
CCCG	n	CACG	L	CTCC	9	CGCC	&
CCTA	o	CATG	M	CTTG	<	CGTT	^
CCAG	p	CAAT	N	CTAT	>	CGAA	%
TCGT	q	TAGA	O	TTGG	,	TGGC	$
TCCG	r	TACC	P	TTCC	.	TGCG	#
TCTG	s	TATG	Q	TTTT	?	TGTT	@
TCAT	t	TAAT	R	TTAA	/	TGAA	!
ACGA	u	AAGG	S	ATGC	:	AGGG	~
ACCC	v	AACC	T	ATCG	;	AGCG	`
ACTG	w	AATT	U	ATTT	“	AGTA	€
ACAT	x	AAAA	V	ATAA	‘	AGAG	£

The encoding process will be in the form of quadruple bases. While, the traditional biological representation is a triple bases.

Coding the plain message to DNA results in a single strand DNA sequence (SSDNAS). The result of this process illustrated in execution bellow:

If a sender has sent a message (**Hello Agent Sara #220**) to the receiver, the message conversion according to Table (4.1) will be:

GACG TCCC GCAA GCAA CCTA AGCG CACC ACCG TCCC CCCG TCAT AGCG ATGC GGGGTCCG GGGG AGCG TGCG ATCC ATCC TTCG.

The generation of DNA coding table and SSDNAS processes used for both central dogma encryption and DNA watermarking stages.

The next step is to apply the securing procedures for the generated SSDNAS. The first stage is applying central dogma encryption procedures as follows:

4.2.1.2 DNA watermarking stage

To conceal the plain message generated a DNA watermarking main processes take place such as.

1-Partial Central Dogma Encryption

To implement this type of encryption partially, one of two types of processes needed: it’s a transcription process.

- Transcription Process

This process converts the SSDNAS to the first version of coding of the first type of RNA, the messenger (mRNA) as shown below:

CTGCAGGGCGTTCGTTGGATTCGCGTGGTGGCAGGGGGGCAGTATCGCTA
CGCCCCAGGCCCCCTCGCACGCTAGGTAGGAAGC.

CUGCAGGGCGUUCGUUGGAUUCGCGUGGUGGCAGGGGGGCAGUAUCGC
UACGGGGGAGGCGGGGUCGCACGCUAGGUAGGAAGC .

2-Scoring Function Processes.

Select the concealing place and determine the location of concealment in the selecting place known by the Scoring function based on ck and mk values.

Example: If we take the previous example above, the notation value of (ck) 03 will be:

TTCG ATTA

- Segment Masking

The starting code and the ending code are added to each (ck) after they are preset. Suppose that the start code is: ***AAGG*** and the end code is: ***CCAA***, the final form of the ck will be in the image below:

AAGG TT CG ATTA **CCAA**
0 3

The process of adding the start code and end code to (ck) is a biological simulation of the (ck) process present in the protein synthesis process and is a supportive addition to the coding process.

- *Key Measuring Generation :*

The process of generating (mk) is by generating random numbers within a range of number 100 i.e. (0≈ 99), these numbers are considered as an indicator to set the segments to the hidden location within the database.

3-Extending Data Base Processes

The required database is pulled from the website (NCBI) as this website contains the databases for each vitamin. The vitamin D database is chosen in the design of the program to hide the cipher text inside it. The size of the database is calculated It is required to hide the cipher text inside it, after the process of calculating the total number of segments of the original message with the length of the private keys for each segment (ck+mk), applying start and end codons for each segment divided along the database.

4- Hiding Processes (Docking watermark)

To hide the massage (Hello Agent Sara #220) in a selected DNA database called vitamin D from NCBI data set. The watermarking processes give us the below results according to message segmentation process:

Length of Massage (Lm) = 21 characters

Length of vitamin’s D Database (Ldb) = 1440 bases as shown in Figure (4.1).

```
gggagcgcggaacagcuuguccacccgccggccggaccagggcuccugaaccuagcccagcuggacggagaaauaac
cuuguuuuucuauuuuucuuguuaauugguuuuuuuuuuaauuccuuaacuuguuaaguuguuuucuccuuaagu
auuuuuuugguuaacugauuaauuuuuuuuuuccuuaacuccuuuuuuaauuaacuauucuugcaaacuuaccugc
ccccugcuccuucaggcagucauugaggacauucccccaggccggcuugcuaccagacgaucauggucuacgacauu
cccucauggucacauucaggcugagaccgagccacuggcuuucacuucaaugcuaugaccugugaaggcugcaacag
ucauugcuuacgaggucuaaugccgccccucauaccgacauuccgucauggucaaccaccgucauacauggucucag
gcugacgccacugccaggccugccggcucaaacgcuguguggacaucggcaugaugaaggaguucauucugacacag
ucauuacauuccgccaggcuuuccgauuagguccaccucaugcccucccggucaaccccggucauaccgucccucag
gcugucugaggagcagcagcgcaucauugccauacugcuggacgcccaccauaagaccuacgaccccaccuacuccga
cuucugccaguuccgcagucauuucucucucacauuccgucauggucacaugcgcgcgcccggccagucauucucgg
ucucccgccgggucucgggagcgcggaacagcuuguccacccgccggccggaccagggcuccugaaccuagcccagc
uggacggagaaauaaccuuguuuuucuauuuuucuuguuaauugguuuuuuuuuuaauuccuuaacuuguuaagu
uguuuucuccuuaaguauuuuuuugguuaacugauuaauuuuuuuuuuccuuaacuccuuuuuuaauuaacuauu
cuugcaaacuuaccugccccugcuccuucaggcagucauugaggacauucccccaggccggcuugcuaccagacga
ucauggucuacgacauucccucauggucacauucaggcugagaccgagccacuggcuuucacuucaaugcuaugacc
ugugaaggcugcaacagucauugcuuacgaggucuaaugccgccccucauaccgacauuccgucauggucaaccacc
gucauacauggucucaggcugacgccacugccaggccugccggcucaaacgcuguguggacaucggcaugaugaagg
aguucauucugacacagucauuacauuccgccaggcuuuccgauuagguccaccucaugcccucccggucaaccccg
gucauaccgucccucaggcugucugaggagcagcagcgcaucauugccauacugcugcca
```

Figure (4.1): D vitamin’s Database

if

Lm $\leq$ 200 characters do

No of Sections (Ns) = $\sqrt[2]{Lm} \approx 5$

Length of Section (Ls) = $\frac{Lm}{Ns} * 4 \approx 17$

CUGCAGGGCGUUCGUUGGAUUCGCGUGGUGGCAGGGGGGCAGUAUCGC

UACGGGGGAGGCGGGGUCGCACGCUAGGUAGGAAGC

Concealing Key (cK) = Random set of numbers

Suppose a set is (4,1,5,2,3) and after convert it to mRNA form measuring Key (mK) = Random set of (100)

UUCGGUGU,UUCGUUUA,UUCGGUCG,UUCGAUCC,UUCGAUUA

We suppose a set is (2,5,12,9,99,88,60,29,72,43 …..etc.)
Start code of cK = aaccuugu
End code of cK = ucuugcaa
Variable value of Ck = uucggugu
Start code of each Section = cagucauu
End code of each Section = ucaggcug
Palin text = GGGG
Variable value of Mk= gucca….

Massage Length require (Ml_r) of each sections after add size basses of (mK+cK & Ls with its Start & End codes) = 1069 bases

Database Required = $\frac{Ml\_r}{Ldb}$* Ldb = 1440 extending the Database used.

The Database Required = Length of Database used because there is no need to extend the embedded DNA database.

As a result, the embedding massage in database of vitamin D as shown in Figure (4.2).

```
ggaaccuuguuucgguguucuugcaaguccacagucauuGGGGGAGGCGGGGU
CGCs4ucaggcuguaaccuuguuuuaaccuuguuucguuuaucuugcaauuuuuuuaa
cagucauuCUGCAGGGCGUUCGUUGs1ucaggcuguauuuuuuugguuaacu
gauuaauuuuuuuuuuccuuaacuccuuuuuuaauuaacuauucuugcaaacuuaccug
cccccugcuccuucaggcagucaaaccuuguuucggucgucuugcaauugcuaccagacga
ucauggucuacgacauucccucauggucacauucaggcugagaccgagccacuggcuuuca
cuucaaugcuaugcagucauuACGCUAGGUAGGAAGCs5ucaggcugggucua
augccgccccucauaccgacauuccgucauggucaaccaccgucauacauggucaaccuugu
uucgauccucuugcaaugccggcucaaacgcuguguggacaucggcagucauuGAUUC
GCGUGGUGGCAGs2ucaggcugacauuccgccaggcuuuccgauuagguccaccuc
augcccucccggucaaccccggucauaccgucccucagaaccuuguuucgauuaucuugcaa
cauugccauacugcuggacgcccaccauaagaccuacgaccccagucauuGGGGGCA
GUAUCGCUACs3ucaggcuguucucucucacauuccgucauggucacaugcgcgcgc
ccggccagucauucucggucucccgccgggucucgggagcgcggaacagcuuguccacccgcc
ggccggaccagggcuccugaaccuagcccagcuggacggagaaauaaccuuguuuuucuau
uuuucuuguuaauugguuuuuuuuuuaauuccuuaacuuguuaaguuguuuucuccuu
Aaguauuuuuuugguuaacugauuaauuuuuuuuuuccuuaacuccuuuuuuaauua
acuauucuugcaaacuuaccugcccccugcuccuucaggcagucauugaggacauucccca
ggccggcuugcuaccagacgaucauggucuacgacauucccucauggucacauucaggcug
agaccgagccacuggcuuucacuucaaugcuaugaccugugaaggcugcaacagucauugc
uuacgaggucuaaugccgccccucauaccgacauuccgucauggucaaccaccgucauaca
```

```
uggucucaggcugacgccacugccaggccugccggcucaaacgcuguguggacaucggcaug
augaaggaguucauucugacacagucauuacauuccgccaggcuuuccgauuagguccacc
ucaugcccucccggucaaccccggucauaccgucccucaggcugucugaggagcagcagcgca
ucauugccauacugcugcca
```

Figure (4.2): Embedded Message Configuration.

The proposed system infrastructure is a flexible structure in terms of updating the techniques and methods used or in adding new procedures. The current research path of gives a purely DNA watermark for copyright as a concealing output. Another type of output can be gained according to the proposed infrastructure by adding the reset steps of the central dogma encryption algorithm to obtain a pure ciphered message that explained in the steps bellow.

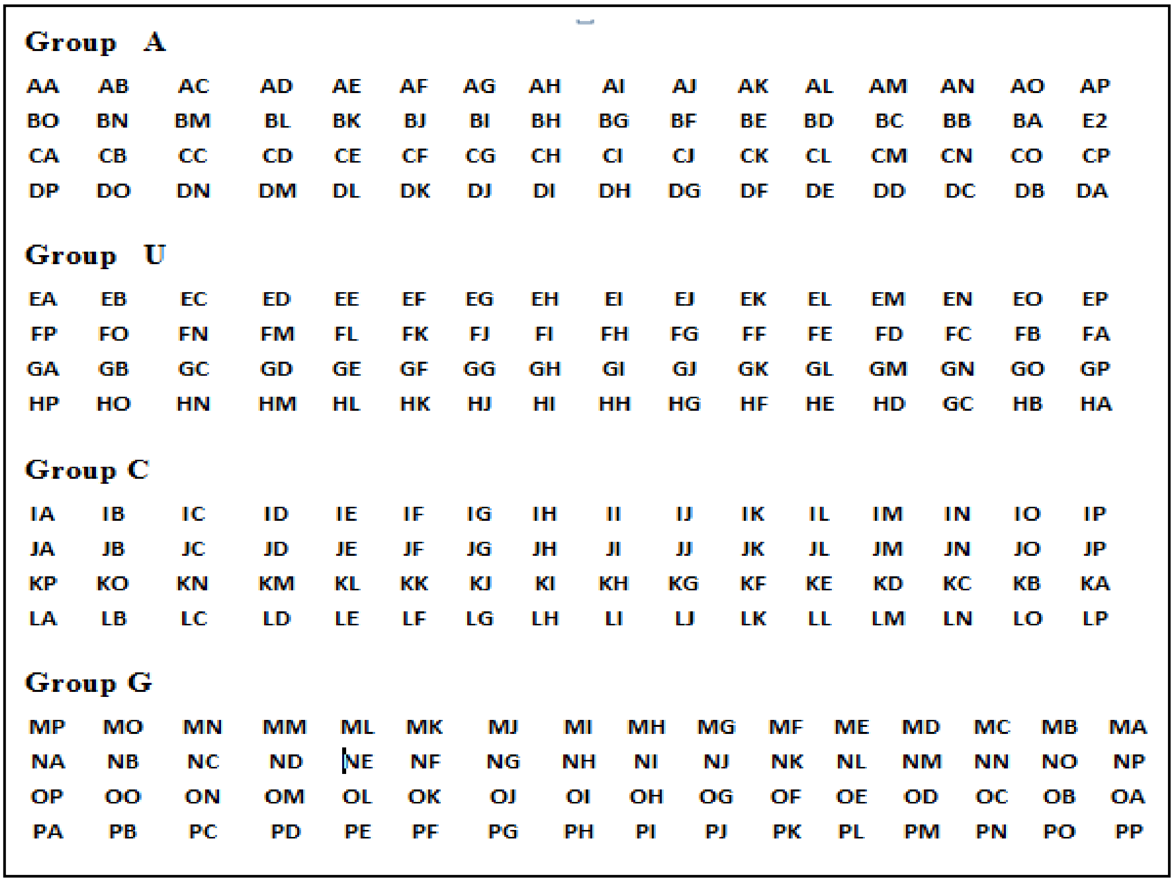
4.2.1.3 Complete Central Dogma Encryption Stage

To implement this type of encryption, two types of processes are needed: transcription and translation in addition to the three types of RNA (mRNA, tRNA, and rRNA) are needed and amino acid table. The first step in this algorithm is the transcription process mentioned in paragraph DNA watermarking stage section (4.2.1.2) in chapter (4), the next step is:

- Translation Process

In this process, the sequence of mRNA is translated to amino acids. The translation process is done by dividing the sequence into quadruple codons, during this session a table is generated called amino acid table according to the four bases (A,U,C,G) as shown in Table (4.2).

Table (4.3): Amino Acid table



Group A

AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP
BO	BN	BM	BL	BK	BJ	BI	BH	BG	BF	BE	BD	BC	BB	BA	E2
CA	CB	CC	CD	CE	CF	CG	CH	CI	CJ	CK	CL	CM	CN	CO	CP
DP	DO	DN	DM	DL	DK	DJ	DI	DH	DG	DF	DE	DD	DC	DB	DA

Group U

EA	EB	EC	ED	EE	EF	EG	EH	EI	EJ	EK	EL	EM	EN	EO	EP
FP	FO	FN	FM	FL	FK	FJ	FI	FH	FG	FF	FE	FD	FC	FB	FA
GA	GB	GC	GD	GE	GF	GG	GH	GI	GJ	GK	GL	GM	GN	GO	GP
HP	HO	HN	HM	HL	HK	HJ	HI	HH	HG	HF	HE	HD	GC	HB	HA

Group C

IA	IB	IC	ID	IE	IF	IG	IH	II	IJ	IK	IL	IM	IN	IO	IP
JA	JB	JC	JD	JE	JF	JG	JH	JI	JJ	JK	JL	JM	JN	JO	JP
KP	KO	KN	KM	KL	KK	KJ	KI	KH	KG	KF	KE	KD	KC	KB	KA
LA	LB	LC	LD	LE	LF	LG	LH	LI	LJ	LK	LL	LM	LN	LO	LP

Group G

MP	MO	MN	MM	ML	MK	MJ	MI	MH	MG	MF	ME	MD	MC	MB	MA
NA	NB	NC	ND	NE	NF	NG	NH	NI	NJ	NK	NL	NM	NN	NO	NP
OP	OO	ON	OM	OL	OK	OJ	OI	OH	OG	OF	OE	OD	OC	OB	OA
PA	PB	PC	PD	PE	PF	PG	PH	PI	PJ	PK	PL	PM	PN	PO	PP

In the Table (4.2), an appropriate amino acid for each codon in the chain is searched. These amino acids of all codons will re-assembled to configure a sequence of amino acid to synthesis the ciphered text to send it to the receiver. The table contain 256 Amino Acids each consist of 4 bases. The assignment process illustrated in Table (4.3). This table is regenerated to each new message.

Table (4.4): DNA Assignment Process

	CA	*AA*	*UA*	*GA*	*GU*	*UU*	*AU*	*CU*	*CC*	*AC*	*UC*	*GC*	*GG*	*UG*	*AG*	*CG*
GG	GGCA	GGAA	GGUA	GGGA	GGGU	GGUU	GGAU	GGCU	GGCC	GGAC	GGUC	GGGC	GGGG	GGUG	GGAG	GGCG
	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP
UG	UGCA	UGAA	UGUA	UGGA	UGGU	UGUU	UGAU	UGCU	UGCC	UGAC	UGUC	UGGC	UGGG	UGUG	UGAG	UGCG
	BO	BN	BM	BL	BK	BJ	BI	BH	BG	BF	BE	BD	BC	BB	BA	E2
AG	AGCA	AGAA	AGUA	AGGA	AGGU	AGUU	AGAU	AGCU	AGCC	AGAC	AGUC	AGGC	AGGG	AGUG	AGAG	AGCG
	CA	CB	CC	CD	CE	CF	CG	CH	CI	CJ	CK	CL	CM	CN	CO	CP
CG	CGCA	CGAA	CGUA	CGGA	CGGU	CGUU	CGAU	CGCU	CGCC	CGAC	CGUC	CGGC	CGGG	CGUG	CGAG	CGCG
	DP	DO	DN	DM	DL	DK	DJ	DI	DH	DG	DF	DE	DD	DC	DB	DA
CC	CCCA	CCAA	CCUA	CCGA	CCGU	CCUU	CCAU	CCCU	CCCC	CCAC	CCUC	CCGC	CCGG	CCUG	CCAG	CCCG
	EA	EB	EC	ED	EE	EF	EG	EH	EI	EJ	EK	EL	EM	EN	EO	EP
AC	ACCA	ACAA	ACUA	ACGA	ACGU	ACUU	ACAU	ACCU	ACCC	ACAC	ACUC	ACGC	ACGG	ACUG	ACAG	ACCG
	FP	FO	FN	FM	FL	FK	FJ	FI	FH	FG	FF	FE	FD	FC	FB	FA
UC	UCCA	UCAA	UCUA	UCGA	UCGU	UCUU	UCAU	UCCU	UCCC	UCAC	UCUC	UCGC	UCGG	UCUG	UCAG	UCCG
	GA	GB	GC	GD	GE	GF	GG	GH	GI	GJ	GK	GL	GM	GN	GO	GP
GC	GCCA	GCAA	GCUA	GCGA	GCGU	GCUU	GCAU	GCCU	GCCC	GCAC	GCUC	GCGC	GCGG	GCUG	GCAG	GCCG
	HP	HO	HN	HM	HL	HK	HJ	HI	HH	HG	HF	HE	HD	GC	HB	HA
GU	GUCA	GUAA	GUUA	GUGA	GUGU	GUUU	GUAU	GUCU	GUCC	GUAC	GUUC	GUGC	GUGG	GUUG	GUAG	GUCG
	IA	IB	IC	ID	IE	IF	IG	IH	II	IJ	IK	IL	IM	IN	IO	IP
UU	UUCA	UUAA	UUUA	UUGA	UUGU	UUUU	UUAU	UUCU	UUCC	UUAC	UUUC	UUGC	UUGG	UUUG	UUAG	UUCG
	JA	JB	JC	JD	JE	JF	JG	JH	JI	JJ	JK	JL	JM	JN	JO	JP
AU	AUCA	AUAA	AUUA	AUGA	AUGU	AUUU	AUAU	AUCU	AUCC	AUAC	AUUC	AUGC	AUGG	AUUG	AUAG	AUCG
	KP	KO	KN	KM	KL	KK	KJ	KI	KH	KG	KF	KE	KD	KC	KB	KA
CU	CUCA	CUAA	CUUA	CUGA	CUGU	CUUU	CUAU	CUCU	CUCC	CUAC	CUUC	CUGC	CUGG	CUUG	CUAG	CUCG
	LA	LB	LC	LD	LE	LF	LG	LH	LI	LJ	LK	LL	LM	LN	LO	LP
CA	CACA	CAAA	CAUA	CAGA	CAGU	CAUU	CAAU	CACU	CACC	CAAC	CAUC	CAGC	CAGG	CAUG	CAAG	CACG
	MP	MO	MN	MM	ML	MK	MJ	MI	MH	MG	MF	ME	MD	MC	MB	MA
AA	AACA	AAAA	AAUA	AAGA	AAGU	AAUU	AAAU	AACU	AACC	AAAC	AAUC	AAGC	AAGG	AAUG	AAAG	AACG
	NA	NB	NC	ND	NE	NF	NG	NH	NI	NJ	NK	NL	NM	NN	NO	NP
UA	UACA	UAAA	UAUA	UAGA	UAGU	UAUU	UAAU	UACU	UACC	UAAC	UAUC	UAGC	UAGG	UAUG	UAAG	UACG
	OP	OO	ON	OM	OL	OK	OJ	OI	OH	OG	OF	OE	OD	OC	OB	OA
GA	GACA	GAAA	GAUA	GAGA	GAGU	GAUU	GAAU	GACU	GACC	GAAC	GAUC	GAGC	GAGG	GAUG	GAAG	GACG
	PA	PB	PC	PD	PE	PF	PG	PH	PI	PJ	PK	PL	PM	PN	PO	PP

The result of this step is:

LLCMDKDKAGGLIMBDCMALCCGLOAAMCLAMGLFEODODNL

4.3 Experimental Results

Table (4.4), explain the coding and decoding time, while Table (4.5) explains thetotal execution time of the encoding and decoding in the proposal algorithm.

Table (4.5): Coding and Decoding time (milliseconds)

Input Size (words in count)	Encoding Time (msec)	Decoding Time(msec)
500	0.0.65	0.0.58
1000	0.1.225	0.16.405
2000	0.2.61	0.2.42
8000	0.9.8	2.11.24

Table (4.6): Total execution time

Input Size (words in count)	Total Time (msec)
500	0.1.33
1000	0.17.63
2000	0.5.142
8000	2.21.04

Table (4.7): Time Results (4 Bases Compare with 3 Bases)[67].

Input Size Massage (Plain text length)	Section Length	No. of sections	Database length (before hiding process)	Datadabe length (after hiding process)	Cracking probability	Capacity
500	96	22	4738	9476	1e-300	0.013218
1000	128	32	4738	18952	1e-300	0.01327
2000	184	45	4738	37904	1e-300	0.013257
8000	360	89	4738	66332	1e-300	0.029872

Time results (4 bases represent) quad codon compared with previous research dependon triple codon (3 bases represent) by (msec)[67]:

Table (4.8): Encoding / Decoding execution time (4 Bases Compare with 3 Bases)

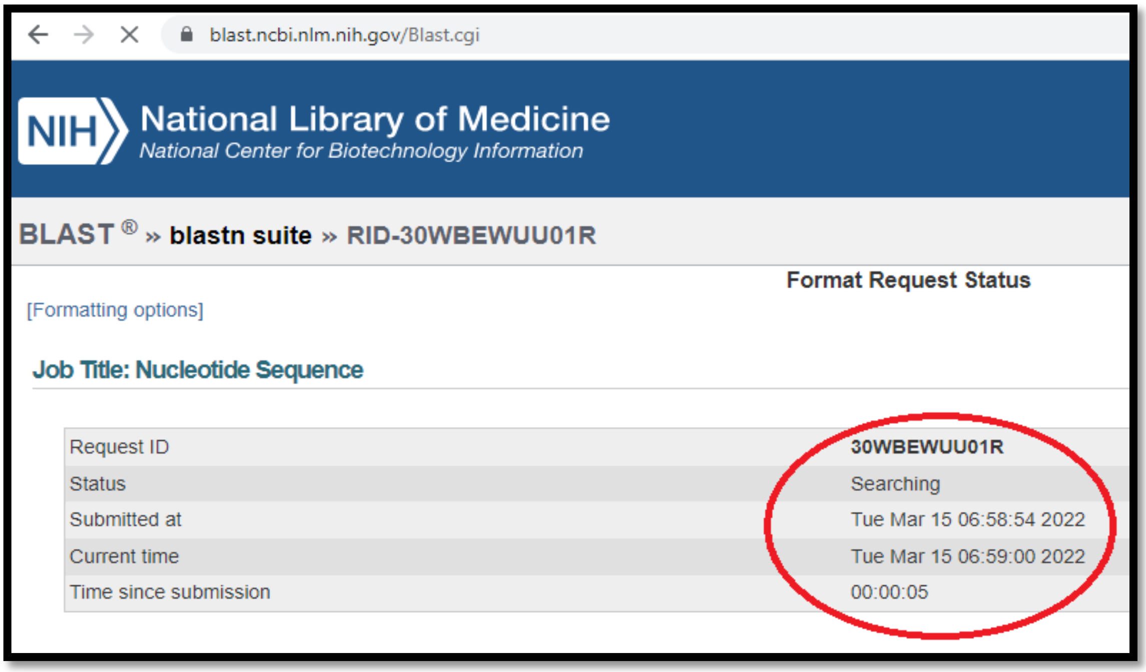
Input size	Encoding time(3 bases)	Encoding time (4 bases)	Decoding time (3 bases)	Decoding time (4 bases)	Total execution time (3base)	Total execution time (4base)
500	0.00006	0.042	0.0001	0.0007	0.0002	0.05
1000	0.0002	0.07	0.00027	0.001	0.0006	0.08
2000	0.0008	0.146	0.00092	0.003	0.001	0.1
8000	0.023	0.54	0.020	0.043	0.045	0.5

4.4 Security Analysis

To prove the copyright of DNA Watermarking and the privacy of data several measuring test used such as:

4.4.1 BLAST Software

The preservation of the selected DNA database (Vitamin D), a software called BLAST to compare and show if other databases are similar to the used database.



blast.ncbi.nlm.nih.gov/Blast.cgi

NIH National Library of Medicine
National Center for Biotechnology Information

BLAST ® » blastn suite » RID-30WBEWUU01R

Format Request Status

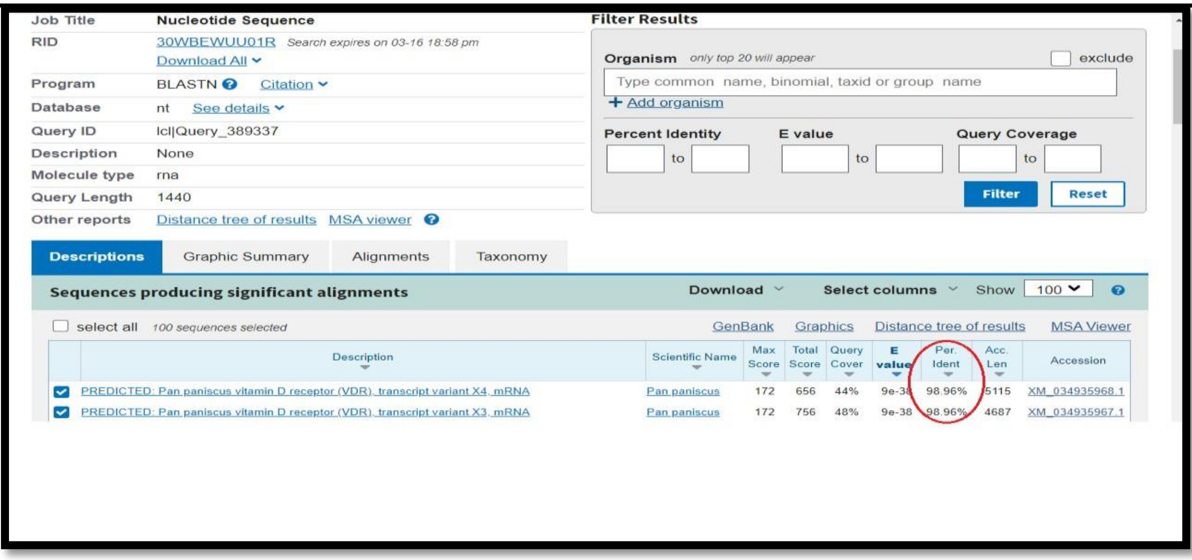
[Formatting options]

Job Title: Nucleotide Sequence

Request ID	30WBEWUU01R
Status	Searching
Submitted at	Tue Mar 15 06:58:54 2022
Current time	Tue Mar 15 06:59:00 2022
Time since submission	00:00:05

Figure (4.3): Percent Identity on NCBI sit

Date and time of database sequences test on NCBI site. Also, the identity of a selected database has been tested by BLAST as in figure (4.4).



Job Title	Nucleotide Sequence
RID	30WBEWUU01R *Search expires on 03-16 18:58 pm* Download All
Program	BLASTN Citation
Database	nt See details
Query ID	lcl|Query_389337
Description	None
Molecule type	rna
Query Length	1440
Other reports	Distance tree of results MSA viewer

Filter Results

Organism *only top 20 will appear* □ exclude

Type common name, binomial, taxid or group name

+ Add organism

Percent Identity ___ to ___ E value ___ to ___ Query Coverage ___ to ___

Sequences producing significant alignments

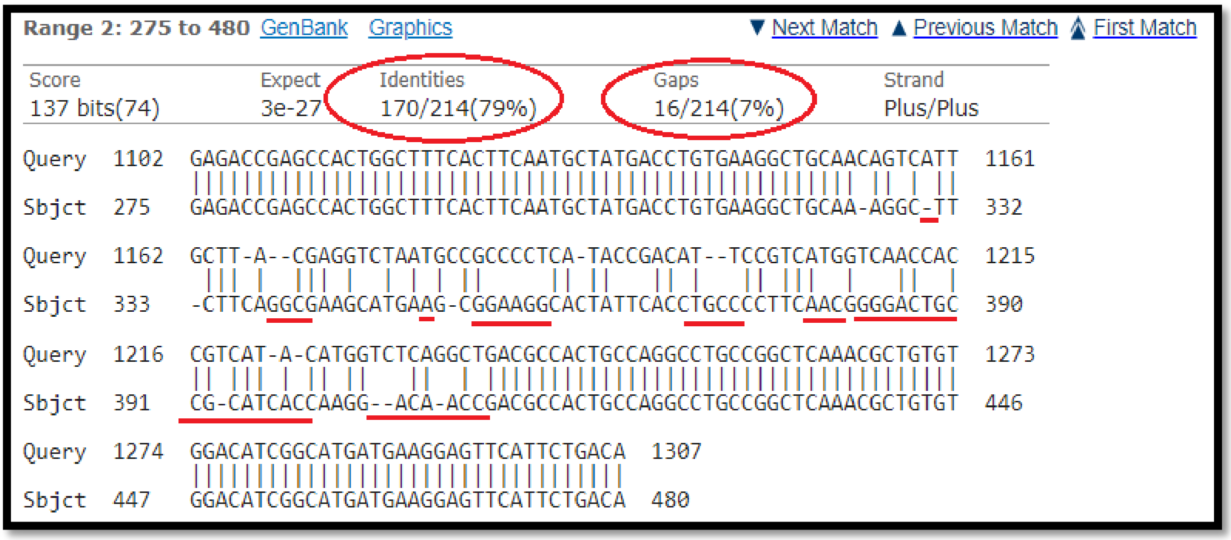
□ select all *100 sequences selected*

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	PREDICTED: Pan paniscus vitamin D receptor (VDR), transcript variant X4, mRNA	Pan paniscus	172	656	44%	9e-38	98.96%	5115	XM_034935968.1
☑	PREDICTED: Pan paniscus vitamin D receptor (VDR), transcript variant X3, mRNA	Pan paniscus	172	756	48%	9e-38	98.96%	4687	XM_034935967.1

Figure (4.4): Identity Database Test

Figure (4.4) shows no identity 100% between our selected database andall others on the NCBI site. This result gives us a proof to non-redundancy of database that used.

The alignment test is used to explains the gaps and identities of genome bases between Query and Subject sequences. Alignment resultes of identites and gapsbetween Query and Subject as in Figure (4.5).



Range 2: 275 to 480 GenBank Graphics

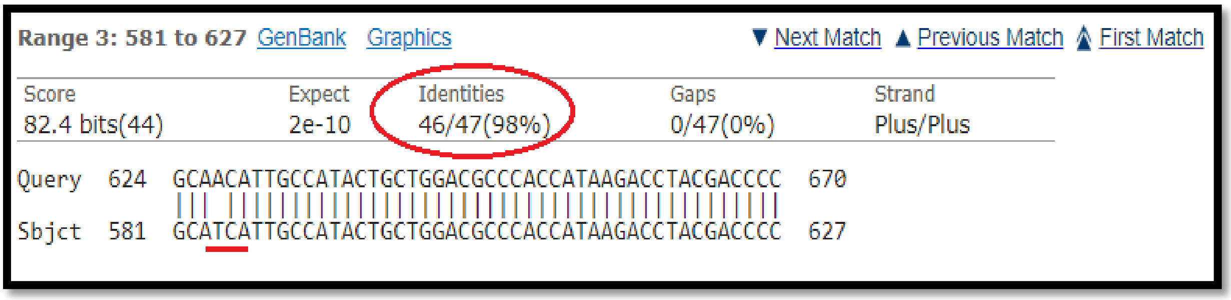
Score	Expect	Identities	Gaps	Strand
137 bits(74)	3e-27	170/214(79%)	16/214(7%)	Plus/Plus

```
Query  1102  GAGACCGAGCCACTGGCTTTCACTTCAATGCTATGACCTGTGAAGGCTGCAACAGTCATT  1161
             |||||||||||||||||||||||||||||||||||||||||||||||||||| || | ||
Sbjct  275   GAGACCGAGCCACTGGCTTTCACTTCAATGCTATGACCTGTGAAGGCTGCAA-AGGC-TT  332

Query  1162  GCTT-A--CGAGGTCTAATGCCGCCCCTCA-TACCGACAT--TCCGTCATGGTCAACCAC  1215
              ||| |  ||| |  | | | ||     || ||   || |   || |||  |   ||  |
Sbjct  333   -CTTCAGGCGAAGCATGAAG-CGGAAGGCACTATTCACCTGCCCCTTCAACGGGGACTGC  390

Query  1216  CGTCAT-A-CATGGTCTCAGGCTGACGCCACTGCCAGGCCTGCCGGCTCAAACGCTGTGT  1273
             || ||| | || ||   ||  | |||||||||||||||||||||||||||||||||||||
Sbjct  391   CG-CATCACCAAGG--ACA-ACCGACGCCACTGCCAGGCCTGCCGGCTCAAACGCTGTGT  446

Query  1274  GGACATCGGCATGATGAAGGAGTTCATTCTGACA  1307
             ||||||||||||||||||||||||||||||||||
Sbjct  447   GGACATCGGCATGATGAAGGAGTTCATTCTGACA  480
```

Figure (4.5): Dissimilarity Percentage

As shown in Figure (4.5),the dissimilarity equal 21% for one section that belong to our database after the hiding process.

Another test for the segment distribution has been done as in Figure (4.6).



Range 3: 581 to 627

Score	Expect	Identities	Gaps	Strand
82.4 bits(44)	2e-10	46/47(98%)	0/47(0%)	Plus/Plus

```
Query  624  GCAACATTGCCATACTGCTGGACGCCCACCATAAGACCTACGACCCC  670
            ||| |||||||||||||||||||||||||||||||||||||||||||
Sbjct  581  GCATCATTGCCATACTGCTGGACGCCCACCATAAGACCTACGACCCC  627
```

Figure (4.6): Dissimilarity Percentage

Figure (4.6) shows the Alignment test of another section that belong to the same database. The Differences in results between sections due to the difference in the random distribution of the locations of the syllables at database.

4.4.2 Cracking Probability

By using the equations of cracking probability (CP) and capacity (C) for the last example, the result will be:

$$\text{CP} = \frac{1}{1.63*10^{8}} * \frac{1}{24} * \frac{1}{2^{4}-1} * \frac{1}{2^{1440-1}} \ .$$

CP = Very small amount less than 1e-300.

4.4.3 Capacity

Capacity result very small because adopting the substitution method that does not make any change on the embedding database.

$$c = \frac{\frac{1}{4} * 84}{1440} = 0.014$$

Where (1440) : length of database selected

CHAPTER FIVE CONCLUSIONS AND FUTURE WORKS

Chapter Five
Conclusion and Future Work

5.1 Conclusion

Through the implementation of the proposed system and the obtained results, several indicators have been reached that explains the strength of the system and the security level. These indicators:

1- The new representation of codons and re-generation process for DNA encoding and amino acid tables for each input message give the encryption process more complication and hard work to the hacker in predicating the ciphering message.

2- Randomizing the spaces between segments and their dependent keys(ck,mk) ensures the security of embedding and anonymity within the database.

3- The number of message segments is unpredictable because they produced up-to- time using mathematical model.

4- Measuring the strength of DNA Watermarking proved by BAST SOFTWARE test results.

5.2 Future Work

In this section, some future directions are suggested as an extension of the proposed system in this thesis:

1- Symmetric and Asymmetric ciphering algorithms can be applied instead of central dogma ciphering algorithm for coding plain messages. As example as the medical fields that use DNA in its work.

2- The proposal can be adopted in cloudy environment by oriented the used techniques according to cloudy requirements.

3- Applying Single Nucleotide Polymer (SNP) as another technique for implementing DNA watermarking

REFRENCESS

References

[1] Adithya, B., and G. Santhi. "DNA computing using cryptographic and steganographic strategies." Data Integrity and Quality (2021). DOI: 10.5772/intechopen.97620.

[2] Khadam, Umair, Muhammad Munwar Iqbal, Meshrif Alruily, Mohammed A. Al Ghamdi, Muhammad Ramzan, and Sultan H. Almotiri. "Text data security and privacy in the internet of things: threats, challenges, and future directions." Wireless Communications and Mobile Computing 2020 (2020).

[3] Al-Qudsy, and Zainab N.,"Information Hiding",March 2019.

[4] Kuribayashi, Minoru, Takuya Fukushima, and Nobuo Funabiki. "Data hiding for text document in PDF file." In International Conference on Intelligent Information Hiding and Multimedia Signal Processing, pp. 390-398. Springer, Cham, 2017.

[5] Nagaraju, G., P. Pardhasaradhi, V. S. Ghali, and G. R. K. Prasad. "Secure hybrid watermarking technique in medical imaging." Eur. J. Mol. Clin. Med 7, no. 05 (2020): 160-167.

[6] Ayday, Erman, Emre Yilmaz, and Arif Yilmaz. "Robust {Optimization-Based} Watermarking Scheme for Sequential Data." In 22nd International Symposium on Research in Attacks, Intrusions and Defenses (RAID 2019), pp. 323-336. 2019.

[7] Ahgue, Augustin Ousmanou, Jean Dieu De Nkapkop, Joseph Yves Effa, Samuel Franz, Pierre Adelis, and Monica Borda. "A DNA-based chaos algorithm for an efficient image encryption application." In 2018 International Symposium on Electronics and Telecommunications (ISETC), pp. 1-4. IEEE, 2018.

[8] Boonekamp, Francine J., Sofia Dashko, Donna Duiker, Thies Gehrmann, Marcel van den Broek, Maxime den Ridder, Martin Pabst et al. "Design and experimental evaluation of a minimal, innocuous watermarking strategy to distinguish Near-Identical DNA and RNA sequences." ACS synthetic

biology 9, no. 6 (2020): 1361-1375.

[9] Ambadekar, Sarita P., Jayshree Jain, and Jayshree Khanapuri. "Digital image watermarking through encryption and DWT for copyright protection." In Recent trends in signal and image processing, pp. 187-195. Springer, Singapore, 2019.

[10] Reddy, M. Indrasena, AP Siva Kumar, and K. Subba Reddy. "A secured cryptographic system based on DNA and a hybrid key generation approach." Biosystems 197 (2020): 104207.

[11] Kalsi, Shruti, Harleen Kaur, and Victor Chang. "DNA cryptography and deep learning using genetic algorithm with NW algorithm for key generation." Journal of medical systems 42, no. 1 (2018): 1-12.

[12] Singh, Gambhir, and Dr. Rakesh Kumar Yadav. "Improvement of Performance Metrics and Security of AODV Routing Protocol Using Central Dogma of Molecular Biology Based DNA Cryptography." International Journal of Innovative Technology and Exploring Engineering. Blue Eyes Intelligence Engineering and Sciences Engineering and Sciences Publication - BEIESP, February 28, 2020.

[13] de Oliveira, Tiago Alves, Lucas Rolim Medaglia, Eduardo Habib Bechelane Maia, Letícia Cristina Assis, Paulo Batista de Carvalho, Alisson Marques da Silva, and Alex Gutterres Taranto. "Evaluation of Docking Machine Learning and Molecular Dynamics Methodologies for DNA-Ligand Systems." Pharmaceuticals 15, no. 2 (2022): 132.

[14] Alomari, Fatimah Y., Abeer A. Sharfalddin, Magda H. Abdellattif, Doaa Domyati, Amal S. Basaleh, and Mostafa A. Hussien. "QSAR Modeling, Molecular Docking and Cytotoxic Evaluation for Novel Oxidovanadium (IV) Complexes as Colon Anticancer Agents." Molecules 27, no. 3 (2022): 649.

[15] Şatir, Esra, and Oğuzhan Kendirli. "A symmetric DNA encryption process with a biotechnical hardware." Journal of King Saud University-Science 34, no. 3 (2022): 101838.

[16] Basu, Sayantani, Marimuthu Karuppiah, Mita Nasipuri, Anup Kumar Halder, and Niranchana Radhakrishnan. "Bio-inspired cryptosystem with DNA cryptography and neural networks." Journal of Systems Architecture 94 (2019): 24-31.

[17] Shaw, Harry. "A cryptographic system based upon the principles of gene expression." Cryptography 1, no. 3 (2017): 21.

[18] Reddy, M. Indrasena, AP Siva Kumar, and K. Subba Reddy. "A secured cryptographic system based on DNA and a hybrid key generation approach." Biosystems 197 (2020): 104207.

[19] Ali, Rasha S., Rajaa Kadhom Hassoun, Inas Fadhil Jaleel, and Noor Subhi Ali. "Proposal for encryption by using modified play fair algorithm and bioinformatics techniques." In Proceedings of the International Conference on Information and Communication Technology, pp. 120-126. 2019.

[20] Krishna, B. Murali, Habibulla Khan, and G. Madhumati. "Reconfigurable pseudo biotic key encryption mechanism for cryptography applications." International Journal of Engineering & Technology 7, no. 1.5 (2018): 62-70.

[21] Krishna, B. Murali, Chella Santhosh, Shruti Suman, and SK Sadhiya Shireen. "Evolvable hardware-based data security system using image steganography through dynamic partial reconfiguration." Journal of Circuits, Systems and Computers 31, no. 01 (2022): 2250014.

[22] Farri, Elhameh, and Peyman Ayubi. "A robust digital video watermarking based on CT-SVD domain and chaotic DNA sequences for copyright protection." Journal of Ambient Intelligence and Humanized Computing (2022): 1-25.

[23] H Mohamed, Marghny, Saad Z Rida, and Ahmed A. Hafez. "Secure Watermarking Algorithm based on DNA Sequence Using DWT-SVD." Information Sciences Letters 7, no. 1 (2018): 1.

[24] Hamad, Safwat, Ahmed Elhadad, and Amal Khalifa. "DNA watermarking using Codon Postfix technique." IEEE/ACM Transactions on Computational Biology and Bioinformatics 15, no. 5 (2017): 1605-1610.

[25] Iftikhar, Saman, Sharifullah Khan, Zahid Anwar, and Muhammad Kamran. "GenInfoGuard—a robust and distortion-free watermarking technique for genetic data." PloS one 10, no. 2 (2015): e0117717.

[26] Agbaje, Micheal, Oludele Awodele, and Chibueze Ogbonna. "Applications of digital watermarking to cyber security (cyber watermarking)." In Proceedings of Informing Science & IT Education Conference (InSITE), pp. 1-11. 2015.

[27] Öksüz, Abdullah Çağlar, Erman Ayday, and Uğur Güdükbay. "Privacy-preserving and robust watermarking on sequential genome data using belief propagation and local differential privacy." Bioinformatics 37, no. 17 (2021): 2668-2674.

[28] James, Tra-Vaughn MC. "A Literature Review of Bioinformatic Workflow Management Tools and Languages." methods 18, no. 10 (2021): 1161-1168.

[29] Diniz, Wellison Jarles da Silva, and Fernanda Canduri. "Bioinformatics: an overview and its applications." Genet Mol Res 16, no. 1 (2017): 10-4238.

[30] Srinivasa, K. G., G. M. Siddesh, and S. R. Manisekhar, eds. Statistical Modelling and Machine Learning Principles for Bioinformatics Techniques, Tools, and Applications. Springer Nature, 2020.

[31] Hepsyba, S. Gladis Helen, and C. R. Hemalatha. "Basic Bioinformatics MJP Publ." Chennai. UNIT-IV 4 (2009).

[32] Cafferty, Patrick. "The Central Dogma of Molecular Biology." POGIL Activity Clearinghouse 3, no. 1 (2022).

[33] Rastogi, S. C., Parag Rastogi, and Namita Mendiratta. Bioinformatics: Methods and Applications-Genomics, Proteomics and Drug Discovery. PHI Learning Pvt. Ltd., 2022.

[34] C.F.A.Bryce,D.Pacin."The Structure and Function of Nucleic Acids", Napier University Edinburgh and Bolton,1998.

[35] Akash, Shopnil, and Mst Fateha Arefine. "A SHORT REVIEW ON CENTRAL DOGMA OF MOLECULAR BIOLOGY.", IJARR, 6(10), 2021.

[36] Camacho, M. Polo. "Beyond descriptive accuracy: The central dogma of molecular biology in scientific practice." Studies in History and Philosophy of Science Part A 86 (2021): 20-26.

[37] Das, Anupam, Shikhar Kumar Sarma, and Shrutimala Deka. "Data security with DNA cryptography." In Transactions on Engineering Technologies, pp. 159-173. Springer, Singapore, 2021.

[38] "National Human Genome Research Institute- NHGRI‖Double Helix", available at: https://www.genome.gov/genetics-glossary/Double-Helix.

[39] Helmenstine, Anne Marie, Ph.D. "What Are the 3 Parts of a Nucleotide? How Are They Connected?" ThoughtCo.October 23, 2022.

[40] Branco, Iuliia, and Altino Choupina. "Bioinformatics: new tools and applications in life science and personalized medicine." Applied microbiology and biotechnology 105, no. 3 (2021).

[41] Gray, Michael W. , and Ann L. Beyer. “Ribonucleic acid (RNA).” AccessScience, McGraw Hill, Aug. 2020.

[42] Oeffinger, Marlene, and Daniel Zenklusen, eds. The biology of mRNA: structure and function. Springer, 2019.

[43] Berg, Matthew D., and Christopher J. Brandl. "Transfer RNAs: diversity in form and function." RNA biology 18, no. 3 (2021): 316-339.

[44] Matsumoto, S., Cavadini, S., Bunker, R.D. et al. DNA damage detection in nucleosomes involves DNA register shifting. Nature 571, 79–84 (2019).

[45] Lopez MJ and Mohiuddin SS. Biochemistry, Essential Amino Acids. In: StatPearls. StatPearls Publishing, Treasure Island (FL); 2022. PMID: 32496725.

[46] Schimmel, Paul, and Karla L. Ewalt. “Genetic code.” AccessScience, McGraw Hill, June 2019.

[47] Sponer, Jiri, Giovanni Bussi, Miroslav Krepl, Pavel Banáš, Sandro Bottaro, Richard A. Cunha, Alejandro Gil-Ley et al. "RNA structural dynamics as captured by molecular simulations: a comprehensive overview." Chemical reviews 118, no. 8 (2018): 4177-4338.

[48] Bailey, Regina. "An Introduction to DNA Transcription." ThoughtCo. accessed October 23, 2022.

[49] Cianci, Michele. "INTRODUCTION TO PROTEINS, Introduction to Proteins: Structure, Function, and Motion: Amit Kessel, Nir Ben-Tal, Abingdon, UK, CRC press, Taylor & Francis Group, 2020, 932 pp., GBP 63.99 (hardback), ISBN 978-1-4987-4717-2." (2021): 47-50.

[50] Andrew-Peter-Leon, M. T., Ramchander Selvaraj, K. K. Kumar, Mehanathan Muthamilarasan, Jeshima Khan Yasin, and M. Arumugam Pillai. "Loss of function of OsFBX267 and OsGA20ox2 in rice promotes early maturing and semi-dwarfism in γ-irradiated IWP and genome-edited Pusa Basmati-1." Frontiers in plant science (2021): 1968.

[51] Mercadante, Anthony A., Manjari Dimri, and Shamim S. Mohiuddin. "Biochemistry, replication and transcription." (2019).

[52] Lisa Bartee and Jack Brook, " Biology for Health Professions" , Creative Commons Attribution 4.0 International License, May, 2016.

[53] Cox, Brian, and David R. Witty, eds. Progress in medicinal chemistry. Elsevier, 2021.

[54] Tornea, Olga, and Monica E. Borda. "Security and complexity of a DNA-based cipher." In 2013 11th RoEduNet International Conference, pp. 1-5. IEEE, 2013.

[55] Cui, Guangzhao, Limin Qin, Yanfeng Wang, and Xuncai Zhang. "An encryption scheme using DNA technology." In 2008 3rd International Conference on Bio-Inspired Computing: Theories and Applications, pp. 37-42. IEEE, 2008.

[56] Lai, XueJia, MingXin Lu, Lei Qin, JunSong Han, and XiWen Fang. "Asymmetric encryption and signature method with DNA technology." Science China Information Sciences 53, no. 3 (2010): 506-514.

[57] Najaftorkaman, Mohammadreza, and Nazanin Sadat Kazazi. "A method to encrypt information with DNA-based cryptography." International Journal of Cyber-Security and Digital Forensics (IJCSDF) 4, no. 3 (2015): 417-426.

[58] Anwar, Tausif, Abhishek Kumar, and Sanchita Paul. "DNA cryptography based on symmetric key exchange." International Journal of Engineering and Technology 7, no. 3 (2015): 938-950.

[59] Hariri, Mehdi, Ronak Karimi, and Masoud Nosrati. "An introduction to steganography methods." World Applied Programming 1, no. 3 (2011): 191-195.

[60] Begum, Mahbuba, and Mohammad Shorif Uddin. "Digital image watermarking techniques: a review." Information 11, no. 2 (2020): 110.

[61] Lee, Suk-Hwan, Seong-Geun Kwon, Eung-Joo Lee, and Ki-Ryong Kwon. "Reversible DNA Watermarking Technique Using Histogram Shifting for Bio-Security." Journal of Korea Multimedia Society 20, no. 2 (2017): 244-253.

[62] Sreeja, C. S., Mohammed Misbahuddin, and NP Mohammed Hashim. "DNA for information security: A Survey on DNA computing and a pseudo DNA method based on central dogma of molecular biology." In International Conference on Computing and Communication Technologies, pp. 1-6. IEEE, 2014.

[63] The National Center for Biotechnology Information (NCBI) on November 4, 1988/ https://blast.ncbi.nlm.nih.gov/Blast.cgi.

[64] Malathi, Pa, Ma Manoaj, Ra Manoj, Vaikunth Raghavan, and R. E. Vinodhini. "Highly improved DNA based steganography." Procedia Computer Science 115 (2017): 651-659.

[65] Khalifa, Amal, and Safwat Hamad. "Hiding secret information in dna sequences using silent mutations." British Journal of Mathematics & Computer Science 11, no. 5 (2015): 1.

[66] Hamad, Safwat, Ahmed Elhadad, and Amal Khalifa. "DNA watermarking using Codon Postfix technique." IEEE/ACM Transactions on Computational Biology and Bioinformatics 15, no. 5 (2017): 1605-1610.

[67] UbaidurRahman, Noorul Hussain, Chithralekha Balamurugan, and Rajapandian Mariappan. "A novel DNA computing based encryption and decryption algorithm." Procedia Computer Science 46 (2015): 463-475.

[68] Hesper, Ben, and Paulien Hogeweg. "Bioinformatica: een werkconcept." Kameleon 1, no. 6 (1970): 28-29.

الخــلاصــــــة

بسبب التطور المستمر لتكنولوجيا المعلومات وانتشار تبادل الوسائط الرقمية كالنصوص والفيديو والصور بين المستخدمين، أصبحت حمايتها من الضروريات. وفقًا لذلك، لابد من وجود تقنيات حماية. إحدى هذه التقنيات هي العلامة المائية الرقمية لحماية الوسائط الرقمية من العبث.

تُستخدم العلامات المائية الرقمية للمصادقة على ملكية المحتويات الإلكترونية القيمة وتحديدها ، وعلى نفس السياق، فإن العلامة المائية للحمض النووي هي تقنية لاخفاء البيانات التي تهدف إلى حماية حقوق التأليف والنشر لتسلسل الحمض النووي وتضمن أمن المعلومات الجينية الخاصة.

في هذه الأطروحة، اقترحنا تقنية العلامة المائية للحمض النووي التي يمكن استخدامها للحفاظ على تسلسل الحمض النووي المحدد (فيتامين د) المستخدمة كمرجع وتأمين المعلومات الجينية الخاصة.

الأجزاء الرئيسية التي يتكون منها الاقتراح هي: جداول الترميز الديناميكية، وخوارزمية التشفير Central Dogma ، وتجزئة الرسائل، وتقنية الإرساء.

يتم بناء جداول الترميز لتحويل الرسالة النصية الى قواعد امينية والتي تؤدي الى تحويل كامل القواعد الامينية (الرسالة النصية) الى سلسلة من الاحماض الامينية (كودونات). هذه الجداول يتم توليدها بشكل ديناميكي، أي ان كل رسالة نصية تمتلك جداول ترميز خاصة بها تختلف عن باقي جداول الترميز للرسائل الأخرى. باستخدام الجزء الأول من خوارزمية التشفيرCentral Dogma (الخطوة الأولى) والذي يسمى بعملية الاستنساخ، فان الرسالة النصية سوف تكون على شكل سلاسل من الاحماض الامينية تدعى بالأحماض الامينية الناقلة.

يتم تقسيم السلاسل الامينية الناقلة الى عدة مقاطع بناءا على تمثيل رياضي يتم انشاءه. ان التمثيل الرياضي الذي يقوم بتقسيم السلاسل الامينية الناقلة الى عدة مقاطع يعتمد اعتمادا كليا على طول الرسالة النصية المشفرة.

كذلك تم توظيف عملية الارساء البيولوجي في تضمين مقاطع الرسالة المشفرة داخل قواعد البيانات من خلال توليد مفتاحين (مفتاح الاخفاء و مفتاح المسافة). ان كل مقطع من مقاطع الرسالة المشفرة سوف يمر بمراحل من المعالجات الخاصة لضمان عملية التضمين داخل قواعد البيانات ، حيث يتم تحديد طول المقطع و المسافة بين كل مقطع و اخر ومن ثم امتلاك كل مقطع على كودونات محددة (كودون شفرة البدء و كودون شفرة النهاية) قبل ان تتم عملية التضمين داخل قاعدة البيانات.

تم اختبار طريقتنا للحفاظ على هوية تسلسل الحمض النووي واستخراج مقدار التشابه مع تسلسل الاحماض النووية الأخرى باستخدام برنامج بلاست (المركز الوطني لمعلومات التكنولوجيا الحيوية). حيث كانت هوية نتائج الاختبار ٩٦٪ بقاعدة بيانات فيتامين (د).

يبين هذا الاختبار عدم وجود تطابق بنسبة ١٠٠٪ بين قاعدة البيانات المختارة وباقي قواعد البيانات الأخرى
والتي اثبتت الحفاظ والأمنية على هوية تسلسل الحامض النووي والقدرة على NCBIالموجودة على الموقع
عدم قابلية الكشف و المقاومة. كذلك اثبت معيار التتبع

وزارة التعليم العالي والبحث العلمي
جامعة بابل
كلية العلوم للبنات
قسم علوم الحاسوب

حماية البيانات بالاعتماد على العلامة المائية للحامض النووي باستخدام تقنية الارساء

رسالة

مقدمة الى مجلس كلية العلوم للبنات / جامعة بابل استيفاء جزئي لشروط الحصول على درجة الماجستير في العلوم / علوم الحاسوب

بواسطة

الطالب / حيدر عبد الخالق عبد الرحيم

اشراف

أ.م.د.سحر عادل الباوي **أ.د. علي حسين المرزوكي**

1443 A.H **2022 A.D.**