Chapter 1

INTRODUCTION OF COMPRESSION AND ENCRYPTION TECHNIQUE

1.1 Introduction

The Bioinformatics research is mainly based on storage and doing something with great amounts of facts. The effective DNA "the code of life" compression-encryption is a challenging question in the area.

The DNA ordering apparatus is capable of producing a great quantity of genomic facts. Those very great volumes of facts needs place for storing, fast transmission, equipping with quick acquiring of any record and higher practicality. In fundamental research, genomic facts give a better comprehension of the cellular activity and evolution of organism[1]. In the field of biomedicine, it is utilized for making observations about the smallest units and diseases based on gene order and their differences. The modern diagnostic tools are made stronger by the use of this technology [2]. In farming and studies related to food, it is made to reproduce the best type of plants and animals. It is also used for making observations about virus effects on one another [3]. The knowledge of DNA, RNA and amino acid orders of proteins of molecular biology are stored in databases. It is experienced that the DNA database sizes are quickly increasing. As an outcome of that it is needed to store and transmit data effectively with small amount of supporting facts. The storage costs have a measurable size of total price in the work of arts and observation of DNA orders. The huge increase in DNA order is manageable because of the great increase in the storing capacity of the disk.

Standard compression technique failed to compress these orders. Nowadays, new algorithms have been introduced specially for this purpose.

The size of DNA orders varying from terabytes to petabytes in size[4-6]. The size of the DNA database becomes bigger twice or thrice in a year [6]. The DNA have within some logical organization[7], for this reason data structure to store, to obtain and process this data with small amount of support is a difficult and very hard work[8]. So it needs an effective compression algorithm to store these great masses of DNA facts. The DNA order has some special structures [9] researchers are kept in mind and offered several DNA compression. The growth rate [10] of genomic facts have exceeded the rate of data growth as predicted by Moore's law.

The relation between compression and encryption is alike. The methods are used to reduce the Redundancy in the original message. In harmony with Shannon[11], for an errorless technique of lossless compression, the mean bit rate is same as the original entropy. The

purpose of minimizing the redundancy of compression algorithm is to conserve the place for storing or bandwidth in communication. If compression and encryption processes are merged, we can get other characteristics of compression-encryption algorithm shown in fig. 1.1. In selective encryption, a part of the bit stream is encrypted, this part of the message plays a more important role during decompression and the other part of the bit stream are not affected. In a perfect compression technique, a part of a plain text message which is unencrypted, this part of the message is statistically not dependent, remaining part of the plain text is encrypted. So by having knowledge of the not encrypted text message, a cryptanalyst cannot deduce anything from the encrypted text message.

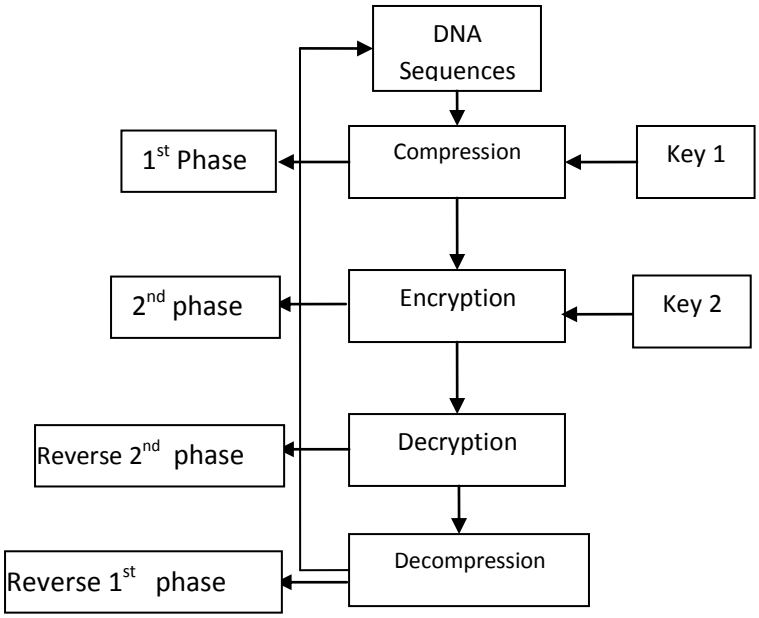

Fig. 1.1 Compression & encryption process diagram

If compression and encryption are mixed we can realize three benefits are : conservation of place for storing space, encryption price is reduced and stop attacks on source bit stream which is statistical property.

The main purpose is 1) to minimize the storing space 2) ensuring safety of data 3) difference in between formed of small units DNA and not natural DNA. Another aim is to lessen the running time of the program and preserve the safety of facts. This undertaking is to simplify the design, so as to make it understandable and minimize the complexity. Gaining high speed and best compression rate & ratio at the same is a hard task.

Every compression algorithm aims at the high compression rate and low decompression time and faster data transmission rate. The high information security and lesser time in decompression is the demand of every user, for that purpose developed two tier method for DNA sequence compression after encryption and decryption before decompression, unlike four pass process, there is not necessary to read the input text. This compression process is necessarily a special sign exchanged compression pattern that encodes with ASCII characters by replacement of nucleotide arranged consecutively. This process is to discover the best answer of decompression processes at the client end.

If standard security techniques are applied on small units of cellular DNA sequence directly observed low level security because the DNA order have only four nucleotide bases. Any authorized person can hack the sequence by trial and error process. After compression our algorithm produces two separate text files. First is compressed file containing ASCII code and non match nucleotide bases which give strong security and the second is dynamic library file of very small size for keeping records in comparison with compress file, which also has more than four nucleotide bases. If two files are sent separately the unauthorized person cannot track the file. Thus, ensuring safety of facts.

The algorithm is developed on the base of quickly moving and sensitive homology search[12] as our exact repeat, reverse, genetic palindromes & Palindrome and their combination search engine. This repeat, reverse, genetic palindromes& Palindrome and their combination substring produce online library file and put ASCII symbols in match position in the DNA sequence.

Nowadays many algorithms are offered for DNA orders compression on the idea of a special structure of the DNA order [13]. From literature, it is found that so many works are done on videos, images, speech by using selective encryption process. But few works using selective encryption on compressed DNA[14] orders have been done. But making a comparison of DNA computing, the operation of making observations of biological cryptology attracted less attention [15].

In many applications pattern matching string is an important operation and different pattern matching solution is presented. In this method sequence is scanned left-to-right using character shift rule [16]. These ideas give a careful way which representatively runs in sub-linear time for short pattern and small alphabet series.

An interesting application of string matching is pattern matching in DNA orders made up of four characters a, c, g, t. Using this technique compressed DNA order to change their size and I/O over head considerably and reduce time to look for pattern, exchange time from sender to receiver end directly in compressed DNA orders with information safety. This compression method for DNA orders is based on a search method.

Its features are low computational complexity, high safety and no distortion. On the other hand, secrecy and access key (encryption key) are controlled only by the given authority parties having the decryption key.

Therefore the compression of data is needed to minimize place for storing, transmission cost and network congestion [17]. The time demands algorithms that match designs in time and space in relation to make shorter size i.e without decompressing. This undertaking is mainly based on genomics and bioinformatics. It is intended for men of science, engineers, knowledge processing machine, computer programmers, or anybody having strong interest in science. On the other hand, we have attempted to consider the content to make some sense of bioinformatics. To develop the bioinformatics tools for genome compression, safety, these are needed, in case of the future of genome order usability.

1.2 Motivation

Some qualities of DNA orders reveal that they are organized. If the DNA sequence is totally random, the two bits encoding can be successfully applied in the DNA sequence, then compression and store easily. Every living organisms DNA sequence is the source of protein, so they must be logically organized[18]. A given DNA order contains many repetitions. For example, in a DNA sequence of TTTACGTTT, TTT is repeated in the given sequence. In DNA T and A are complement to each other. So are C and G. The complement of DNA order TTTACG would be AAATGC. If reversed TTTACG would get GCATTT. The CGTAAA is reverse complement of TTTACG. In a similar species the DNA order is greatly sized and contain many repetitions. It is known that only approximately 0.1% of the 3 GB human genome [19] is special, the remaining human genome is common. A DNA sequence is made-up of four symbols like G,C,A & T. Each three letter substring within the DNA orders is called codons. We have seen that there are 64 well known codons, which produces different 20 number amino acids. It is also observed that similar amino acid is produced from different codons [20]. Many unknown nucleotides are present in a human genome structure which is defined by N [21].

When the DNA coding order is taken as input, the DNA to protein translation is done by sensing open reading frame. This order is then changed into amino acids taking 3 nucleotides at a time. Amino acid details are known from codon. Each codon is taken into account by changing single position getting another codon. The age of personalized genomics is giving help to individual genome being ordered for prevention, analyze and treatment of disease[22]. Genomic data are also made up of subsequences like promoters, functional motive and so on [23]. The analysis of the data gives knowledge [24] into individual being healthy and benefit from future medical research. In higher eukaryotes genomic data has in it many copies of repeat and essential genes. The analysis of repeat is vital as it has an effect on gene control when present in this field, with transcription factor [25].

1.3 Background of proposed research work

Life represents order. It is not without order or random[26]. The DNA orders that encode life is not random. Alternately we can say that the cellular DNA order must be able to compress. To support this fact, there are so many biological proofs to support the fact. It is well known that DNA orders, have in it many repetitions as specially seen in higher eukaryotes. All these evidences give more support that the DNA orders should be fairly compressible. It is very hard job to compress DNA sequences [27].

This was proved as for example in the forming of whole genome phylogenies [28]. There are many repeats within a given DNA order, which may come to mind more than once in a given DNA order [8]. It is accepted that a DNA sequence has many duplicates. The DNA sequences produce so many proteins. It is also assumed that at times gene copies itself for on-going development or for self interest purpose.

1.4 Problem domain

The DNA order is forming of only four nucleotide bases A, T/U, G & C, each base required eight bits for storing. However, if use quality example of software based on the theory of compression like the UNIX complex, pkzip, arj and compact or the MS-DOS archive programs the each software enlarge the file size with greater than eight bits/base, because the regularities in DNA orders are much more delicate, cannot compress the genome orders well because the regularities in DNA orders are ambiguous [29]. This software [30] is made to compress text file. It is observed that no marketable file-compression program does not achieve good results on benchmark DNA orders.

An organism's complete set of DNA/RNA is called its genome. Each DNA strand is made of four chemical units called nucleotide bases, which is composed of genetic "the letters". The bases are adenine(A), thymine(T), guanine(G), cytosine(C). Two bit representation using special sign is applicable if the DNA sequence is completely random. However, only a part of DNA order are produced in an applicable organisms, therefore the orders which appear in a living organism are looked on as non random and have some limitation. The two bits encoding is efficient if the bases are indiscriminately issued in the order[29].

The Huffman's static and adaptive model is unsuccessful in DNA sequence, because there are only four types of special signs in DNA orders and the chances of taking the place of the special signs are not extremely dissimilar. Huffman's lossless compression technique both of static and adaptive model are not well because of, in relation to DNA orders containing a very less number of different characters[29,8].

A dictionary based compression plan [31] reads in input data and look for groups of special signs that come into views as in a dictionary. If a string match is discovered, then a pointer or index into the dictionary can be output instead of the code for the special sign [32]. The longer the match, the better the compression ratio. In dictionary base compression the genomic identity is not fully preserved.

The substitution algorithm first search exact repeat like repeat, reverse, complement, reverse complement and palindrome then encode the match position, it will take more time for compression.

For DNA sequence compression use Lempel-Ziv compression technique as a reasonable default choice.

Jacob Ziv and Abraham Lempel in 1977 [33] and 1978 has published two landmark paper on compression based on adaptive dictionary based techniques.

The DNA sequence consists of only four bases, so, if selective encryption applied directly on that sequence, low level security is achieved. As a result, by using trial and error method any person can hack the DNA sequences easily. If the DNA sequence is large the symmetric key traditional block ciphers as Advanced Encryption Standard (AES), Data Encryption Standard (DES), and Escrowed Encryption Standard (RSA) are inapplicable. Directly applying standard encryption methods like DES, AES [14] , becomes prohibitive because of in relation to the processing time.

Nowadays, so many computing tools are available for making or put right things and exchange networks at low costs relatively, the encryption standard is not as fast as it is expected by a user. In high speed network required high throughput encryption and decryption are gaining importance. Fast encryption algorithm is required nowadays for greater speed and safe communication of DNA order. It has been made clear that public key algorithm is slow, whereas symmetric key algorithms are much faster. In addition, public key systems are open to attack to selected plaintext. Symmetric key cryptography is widely used to find an answer to the old and wise hard question of communication over an unsafe channel. Nowadays the communication technology has increased in recent times and as an outcome, transmission over networks has been quicker desiring fast cryptographic changes for high speed safe connections. This is the guiding reasons behind the operation of making observations that lead to the development of an effective compression encryption method as described in this thesis.

1.5 Problem solving as a search task

Persons making observations down the ages have used general moves near on the natural pictures of DNA order, as a cord of characters. The compressibility of DNA order can take more chances of certain biological qualities. Other search tasks are as below

- Claude Shannon first pointed out the strong relationship between data compression and encryption.
- The phenomenal characteristic of genomic data contains so many repeats (e.g. ATGC) within a given DNA sequence
- In selective encryption process, a part of information is encrypted while the rest of the information remains unchanged.
- Create dynamic key Lookup table technique for security over static key
- Increase space for selection encryption by using ASCII code over nucleotide bases a,t,g & c

1.6 Proposed work & methodology

The main aim of any compression-encryption is undoubtedly to compress and provide safety. Gaining remarkable development in computational cost, time and compression, rate, speed &

selection encryption must not compromise with the safety. This being the cost, our proposed work offers to undertake the following

1. Developing new and modified Algorithm for executing lossless compression & selective encryption in DNA.

2. Compare the proposed algorithm with the existing algorithms in terms of compression rate, ratio, encryption speed and complexity in computation.

3. Our proposed algorithm enhanced the security level of selective encryption so as to reach at and above that of complete encryption and study the developed model with analysis.

4. Our developed algorithm i,e RHUFF model is applied on Huffman's and RSA techniques.

1.7 Data flow diagram of DNA sequence compression & encryption

Process diagram of DNA sequence compression & encryption

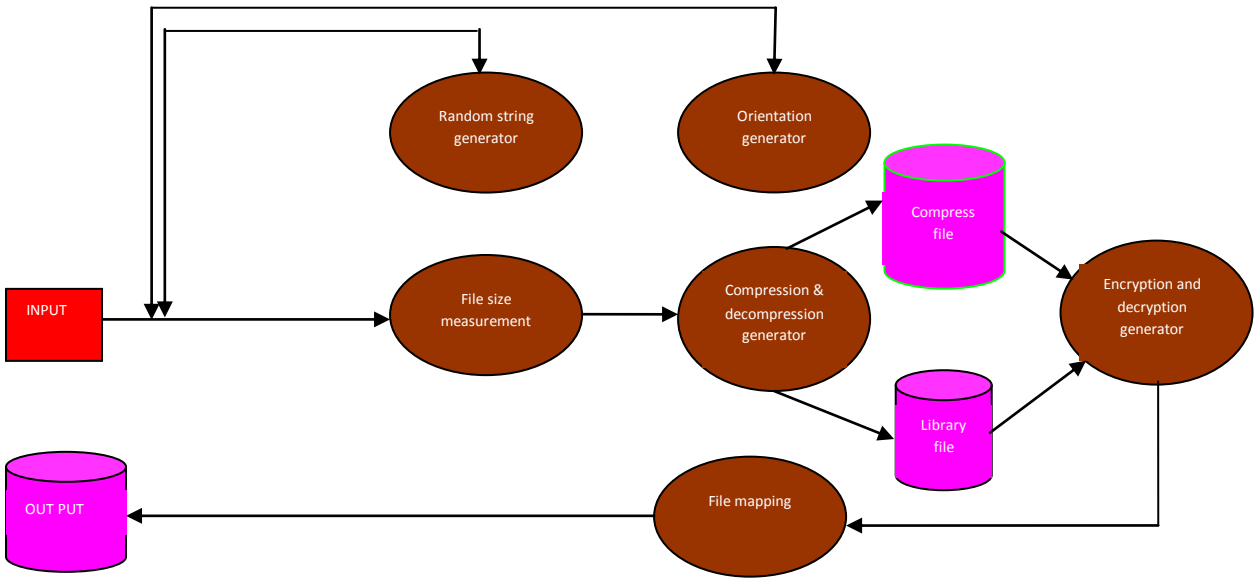

Fig.1.2 Process diagram of proposed work

Flow chart of DNA sequence compression & encryption

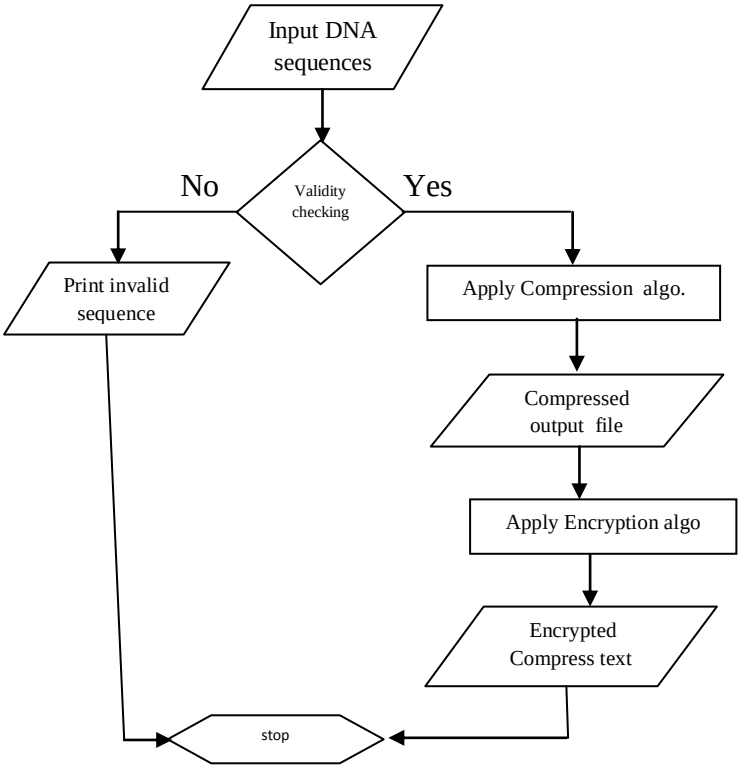

Fig.1.3 How to get encrypt DNA sequence from compress data

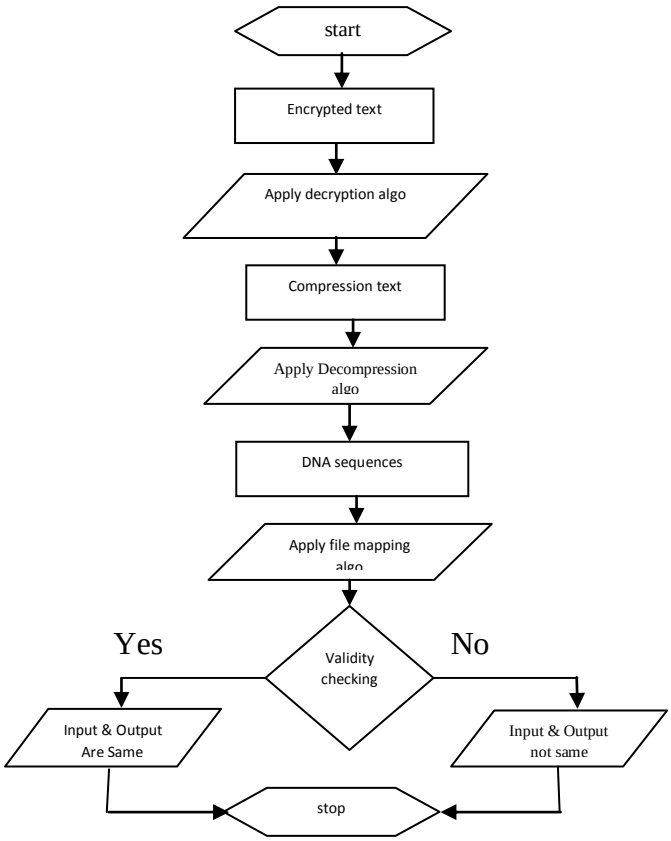

Fig.1.4 The reverse method to get original DNA sequence

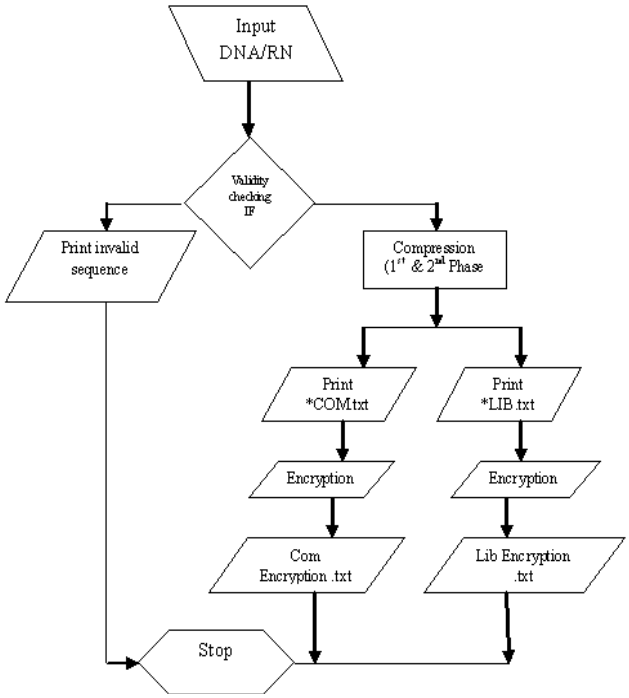

Fig. 1.5 DNA sequence compression process

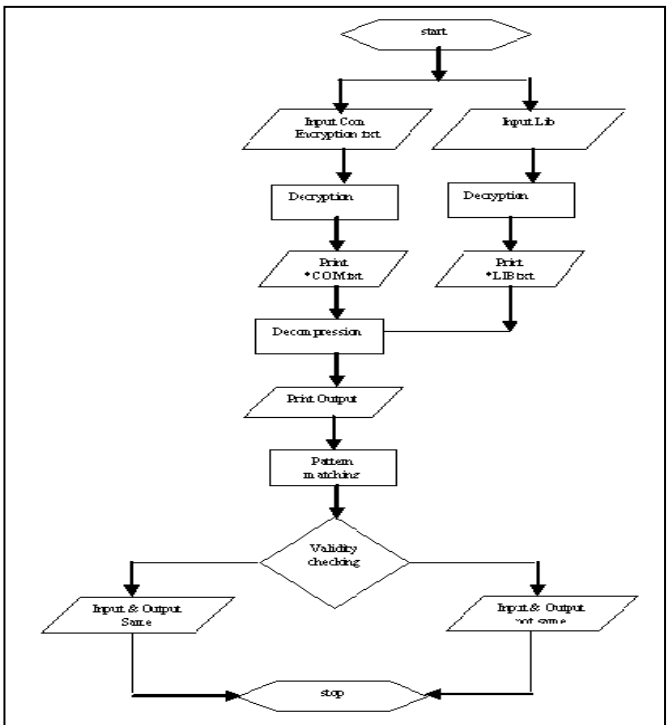

Fig. 1.6 DNA sequence decompression

1.8 Proposed research approach

Consider a DNA Sequence : atgcagtcatgtgttatgagtctgtacgactgatagatgtgtctcatatgccgtaatgcatg

acgagatg…….n. Where n is the no. of nucleotide base. Optimally shown that DNA sequence contain so many subsequences of repeat, reverse, palindrome and genetic palindrome/complement i,e repeat subsequence atg is repeated in many places in the same sequences, reverse subsequence gtc is repeated in many places in the same sequence, palindrome subsequence tgt repeat in many places of the same sequence and genetic palindrome tct repeat in many places of the same sequence.

So, four types of substring are available in a DNA sequence as Repeat(R), Reverse(R), Palindrome(P) and Genetic Palindrome/Complement (GP/C).

[A] Single substring substitution method

i) Repeat method

[B] Bi-combination of substring substitution method

(i) Reverse with Palindrome

(ii) Palindrome with Genetic palindrome

[C] Tri-combination of substring substitution method

(i) Repeat, Reverse and Genetic Palindrome

[F] Different Encryption process

a) Modified Huffman's technique

b) Modified RSA technique

1.9 The selective algorithm improvement on the standard approach

The selective algorithm deals with text file. The text file has many repeats that is why it is possible to encrypt. Here suppose it has need of approximate time T for every symbol in a given group of keys. The linear operations such as write, read and compare, takes time >>T, so, there is no need to consider the time in case of other computation. Consider, number of characters = 291 confidential information given is the word file (4 character string) number to

be kept secret characters = 4*8= 48 times savings standard approach = ((291-48)/291*100%=83.50%. My approach is=((291-4)/291)*100%=98.62%. Whereas our outcome show after compression it has come into existence four separate text file first two are compressed data have within 256 different characters, so it gives strong safety second text file is two separate library file of small size with comparison to compressed file which also has in it more than four symbols.

1.10 Thesis contribution

An effective tool for compression-encryption of DNA sequences by using exact repeat, reverse, complement & palindrome and their combination, modified Huffman's & RSA techniques. A clear account of our encoding of compression followed by encryption and decryption followed by decompression is briefly described. The biological DNA sequence contained so many repetitions. This important feature has not been observed by all available algorithms. All possible combinations of exact repeat and encryption from very small to possible maximum size and uniformity of a particular size are also not considered in the previous published works. All these drawbacks are successfully overcome by our algorithms.

1.11 Thesis outline

The thesis is organized in seven chapters.

Chapter-1 introduction, motivation, background of the proposed research work, problem domain, problem solving as a search task, proposed work & methodology, different data flow diagrams, the proposed research approach and the selectivity algorithm improvement on the standard approach

Chapter 2 preliminary background of proposed work. Theoretical background and discussed the and previous developmental work on compression & encryption methods.

Chapter 3 proposed pattern matching & substitution techniques of lossless DNA sequence compression using RP/GP^2 method with information storage and security. This method is applied on benchmark DNA order & equivalent artificial data and results is statistically analyzed including bar chart, line graph.

Chapter 4 proposed pattern matching & substitution techniques of lossless DNA Sequences Compression using Repeat technique and selective encryption using modified Huffman's technique. This algorithm is tested on benchmark DNA & equivalent artificial data and

results are statistically analyzed, including bar chart, line graph. Our results are compared with other existing techniques.

Chapter 5 proposed pattern matching & substitution techniques of the DNA sequence compression using GP^2R and selective encryption using modified RSA technique. This algorithm is tested on benchmark DNA & equivalent artificial data and results is statistically analyzed, including bar chart, line graph.

Chapter 6 proposed pattern matching & substitution techniques of lossless Compression algorithm for DNA Sequences based on R^2G techniques with security. This algorithm is tested on benchmark DNA & equivalent artificial data and results are statistically analyzed, including bar chart, line graph.

Chapter 7 proposed pattern matching & substitution techniques of lossless DNA compression & security based on Reverse technique. This algorithm is tested on benchmark DNA & equivalent artificial data and results is statistically analyzed, including bar chart, line graph.

Chapter 8 proposed pattern matching & substitution techniques of DNA compression & security using complement technique. This algorithm is tested on benchmark DNA & equivalent artificial data and results is statistically analyzed, including bar chart, line graph.

Chapter 9 finally described the conclusion and scope of further research work.