COMPARATIVE STUDIES ON GENOMIC SEQUENCES COMPRESSIBILITY OF DIFFERENT ORGANISMS

A DISSERTATION

Submitted in the partial fulfillment of the requirements for the award of the degree of

MASTER OF SCIENCE

IN

BIOTECHNOLOGY

Under the supervision of:

Dr. Vikas Handa

Assistant Professor

Department of Biotechnology

Submitted by:

Japleen Kaur

Registration no.

301501004

DEPARTMENT OF BIOTECHNOLOGY

THAPAR UNIVERSITY,

PATIALA -147004

JULY 2017

CANDIDATE'S DECLARATION

I hereby declare that the project work entitled "**COMPARATIVE STUDIES ON GENOMIC SEQUENCES COMPRESSIBILITY OF DIFFERENT ORGANISMS**" in the partial fulfilment of the requirement for the award of the degree of **Master in Science** in Biotechnology, Department of Biotechnology, Thapar University, Patiala, is an authentic record of my work during the period of one year from July 2016 to July 2017, under the guidance of **Dr. Vikas Handa**, Assistant Professor , Thapar University, Patiala. The matter embodied in this thesis has not been submitted in any part or full to any other University or Institute for the award of any degree in India or abroad.

Japleen Kaur

JAPLEEN KAUR

(301501004)

CERTIFICATE

This is to certify that the thesis entitled "**COMPARATIVE STUDIES ON GENOMIC SEQUENCES COMPRESSIBILITY OF DIFFERENT ORGANISMS**" submitted by **Ms. Japleen kaur** in partial fulfilment of requirement for the award of degree of **Master in Science** in the Department of Biotechnology, Thapar University, Patiala, is the record of the candidate's own independent original work carried out by her, under my supervision and guidance. The matter embodied in this thesis has not been submitted in part of full to any other University or Institute for the award of any degree.

Dr. Vikas Handa

Assistant Professor

Department of Biotechnology

Thapar University

Patiala, Punjab

Japleen Kaur

M.Sc Biotechnology

(301501004)

Thapar University

Patiala, Punjab

ACKNOWLEDGEMENT

I thank the Almighty for showering his blessings throughout the preparation of my thesis.

First and foremost, I would like to express my sincere and profound gratitude to all those people who have made this dissertation possible. Firstly I would like to acknowledge the Head of Department **Dr. Moushumi Ghosh** for believing in me and giving me a chance to prove myself and to my thesis supervisor, **Dr. Vikas Handa**, Assistant Professor, for his valuable guidance, undaunted motivation, encouragement, constant support and sound advice. His rare academic and professional insight, commitment and admirable dedication to the subject have always been a source of motivation for me. I would also like to thank her for providing the best laboratory facilities for conducting my research work. I would also like to thank him for providing the best laboratory facilities for conducting my research work.

I owe a heartfelt thanks to **Dr. Rana** for helping and guiding me in my project.

I am extremely thankful to **Ms. Gurpreet Kaur**, for her immense help, valuable suggestions and necessary guidance. I would also like to thank my lab-mates, **Ms. Ravinder Kaur** and **Ms. Meera Sharma** for their support and help.

I shall retain my thankful indebtedness to my mother **Upkar Kaur** and my father **Swaran Singh Sohi** for giving me freedom and opportunity to pursue my own interest and for believing in me and enduring with me during difficult times.

Japleen Kaur

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

A	Adenine
C	Cytosine
chl.	Chloroplast
CDS	Coding DNA Sequence
CV	Measure of Dispersion
DNA	Deoxyribose Nucleic Acid
G	Guanine
mt.	Mitochondria
NCBI	National Center for Biotechtecnology Information
PCR	Percentage Compression Ratio
RLE	Run Length Encoding
RNA	Ribonucleic Acid
T	Thymine
VNTR	Variable Number Tandem Repeats

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

The eukaryotic DNA is highly complex depending upon the organisms. The genome complexity can be analysed by reassociation kinetics which in turn is related to the genomic contents such as coding sequences and repeat sequences. The coding sequences are usually unique i.e., they do not contain repetitive sequences whereas non coding sequences usually consist of repetitive DNA sequences. In this study genome complexity has been studied by DNA sequence compression. Lossless sequence compressibility depends upon the repetition of sequences. It has been found that in comparison with RLE method, LZ algorithm is more efficient in sequence compression. DNA sequence compression by either of the two methods could not show much difference among various genomes with varying evolutionary lineages. However, Percentage Compression Ratio (PCR) exhibited significant correlations with G+C content and sequence heterogeneity of different genomes studied.

Keywords: DNA sequence, RLE, LZ algorithm, sequence compression, repetitive sequences.

CHAPTER 1 INTRODUCTION

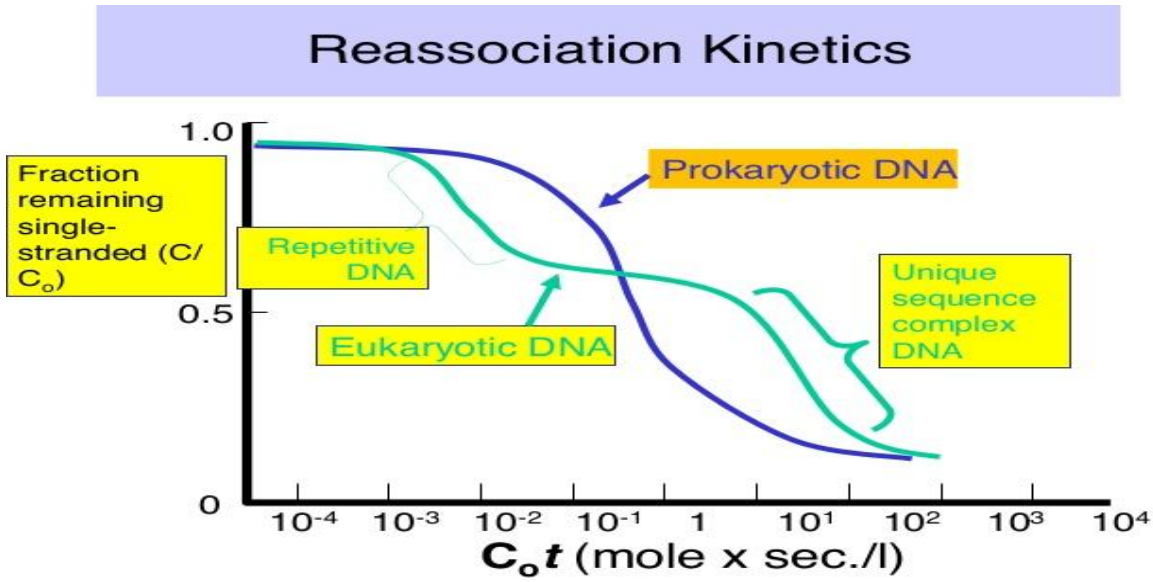
DNA consists of genetic and biological information. It is a bioploymer that consists of four different types of deoxyribonucleotides i.e., Guanine, Adenine, Thymine and Cytosine. These are categorized under Purines and Pyrimidines. These four nucleotides are joined to each other forming a chain, by covalent bonds between the sugar of nucleotides and the phosphate of the next, resulting in sugar phosphate backbone. The backbone of DNA is resistant to cleavage, and both the strands of the double-stranded DNA store exactly same biological information. This biological information is replicated as such.DNA is a genetic material in most of the living organisms and it carries genetic information in the form of sequence of four different bases from parent cell to daughter cells. Central dogma states that the genetic information flows from DNA–>RNA–>Protein. DNA sequence of genome can be transcribed into RNA and most of them translated into proteins. DNA as the genetic material has the following features: (i) can store genetic information (ii) Able to generate its replica through the process of replication (iii) should have its own mechanism to decode the genetic information into functional molecules such as RNAs and their translation into proteins. The large part of the DNA i.e., more than 98% for humans is non-coding, means that these sections of sequences do not serve as patterns for the protein sequences. Similarly, there is coding regions which is that portion of DNA that codes for some protein. Also called as Coding DNA Sequences (or CDS). The region of coding sequences is bounded nearer the 5' end by start codon and nearer the 3'end with stop codon. Understanding genomic sequences has wide applications, from synthesis of medicines to genetic screening and engineering. The knowledge of structure of genomic sequence is important for its cognition.DNA sequence analysis becomes important part in modern molecular biology. As DNA sequence is composed of four nucleotide bases—adenine (abbreviated as A), cytosine (C), guanine (G), and thymine (T) in any order. With four different nucleotides, 2 nucleotides could only code for maximum 4^2 of amino acids, but 3 nucleotides could only code for a maximum of 4^3 amino acids. Every three bases can translate to a single amino acid, called a codon. A short DNA sequence can contain less genetic information, while lots of bases may contain much more genetic information, and any two nucleotides switch place may change the meaning of genetic messages.

In simpler organisms almost the entire DNA consists of unique sequences. In higher organisms there can be large amounts of repetitive DNA. There are two types of repetitive

DNA: Tandem repeats these sequences are present in at least 10^5 copies per genome and they are typically short and are present in clusters in which the given sequence repeats itself, over and over again without interruption and Dispersed repeats which are moderately repeated fraction of the genomes of plants and animals i.e. it varies from about 20% to more than 80% of the total DNA depending upon organism. The length of the non-repetitive DNA component tends to increase as the complexity of organisms increase. Large amount of DNA present in the plants and animals indicates the presence of repetitive DNA. Most genes are present in non- repetitive DNA. This indicates that genetic complexity is proportional to the amount of non- repetitive DNA. **(Primrose, 2013)**

Genome complexity can be analysed by reassociation kinetics. The kinetics of DNA reassociation reveal DNA classes differing in repetition frequency. Double stranded DNA in solution is heated and denatured forming single stranded DNA. The DNA solution is cooled slowly as when it is cooled quickly the strands won't leave the single stranded state. The optimum temperature for reassociation is 25°C below the melting temperature (T_m). At this temperature 50% of duplex is dissociated. Various concentrations of DNA are incubated at varying times at temperature allowing reassociation. The larger and more complex an organism‘s genome is, the longer it will take for complimentary strands to bump into one another and hybridize. **(Primrose, 2013)**

Fig1.1 Genome complexity analysed by reassociation kinetics ($C_0t_{1/2}$ =1/K).



(Source: http://image.slidesharecdn.com/genomestructure-medics)

C_o is the concentration of single stranded DNA at the beginning of the reassociation reactions. It is measured in moles of nucleotide units per litre.

t is the time in seconds of the reassociation kinetics.

Hence more will be the sequence complexity more will be number of unique sequences. Moreover the reassociation kinetics will be more in this case as it has more unique sequences.

Three components of Reassociation Kinetics

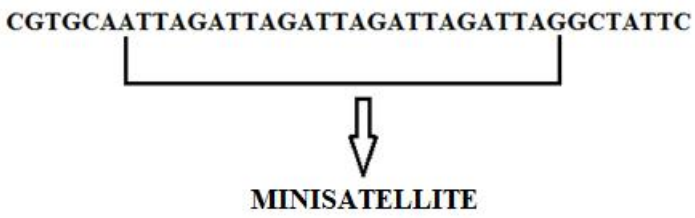
A) Highly repetitive DNA sequences

a) These sequences are present in at least 10^5 copies per genome.
b) They are typically short and are present in clusters in which the given sequence repeats itself, over and over again without interruption.
c) This type of sequences is said to be tandem repeated sequence.

1) Minisatellite DNAs

I. These sequences range from about 12 to 100 base pairs in length and are found in clusters.
II. They occupy considerably shorter stretches of the genome than the satellite sequences (form very large clusters each containing upto several million base pairs of DNA).
III. The length of a particular minisatellite locus is highly variable in the population even among the same members of the same family.
IV. It is used to identify individuals in criminal or paternity cases through the technique of DNA fingerprinting.

Fig1.2:-Depicts the repeated clusters of nucleotides called Minisatellite.



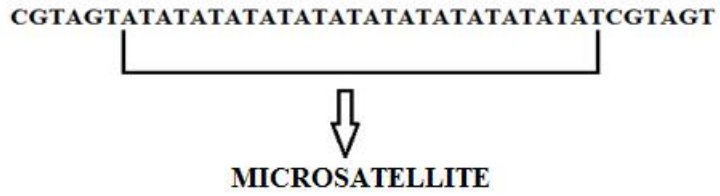
V. It is also termed as VNTRs (Variable Number Tandem Repeats).

2) Microsatellite DNAs

I. These are shortest sequences (1 to 3 nucleotide repeats) and are typically present in small clusters of about 10 to 40 base pairs in length.

II. These are scattered quite evenly through the DNA - more than 100,000 different loci are present in the human genome.

Fig1.3:-Depicts the repeated clusters of nucleotides called Microsatellite.



B) Moderately repeated DNA sequence

a) Moderately repeated fraction of the genomes of plants and animals i.e. it varies from about 20% to more than 80% of the total DNA depending upon organism.
b) This type is said to be dispersed repeats.
c) Dispersed repeats are segments of DNA that occur multiple times at more or less random positions in the genome. They are typically transposable elements, large segments that encode a protein responsible for the moving of the segment from one site to another.

C) Nonrepeated DNA Sequence

a) These are present in a single copy of genome which includes genes that exhibit Mendalian patterns of inheritance.
b) They always localise to a particular site on a particular chromosome.

Kolmogorov complexity is the measure of computational resources that specifies the particular object, and is also called as descriptive complexity. Moreover, it can also be measured in the form of compressibility of sequence. It is based on Algorithmic Information Theory considering objects as individual symbol strings and also it has the property to remain unchanged if the size of the system changes.**(Emmert-Streib F., 2010)**

DNA sequence contains ORF's, transposons, tandem repeats, duplicate genes and pseudo genes. Non coding regions contain tandem repeats, dispersed repeats, transposons and pseudo genes whereas Coding regions contain ORF's. Coding regions and Exons are more complex than Non coding and Introns as they contain more repetitive sequences. So, the compressibility will be less in case of Coding regions. The DNA contains direct repeats which show the type of regularities i.e., the repeated segment in a DNA sequence. Compressibility can be done by replacing such regularities with certain codes thereby decreasing the file size and reducing the storage efficiency. **(Delahaye J. P. et al., 1996)**

Compression can be also used to find out the relatedness of coding and non-coding sequences. The percent compression ratio (PCR) of coding and non-coding regions of DNA must be different. Hence, the compression technique can be used to test the non-randomness of different type of DNA sequence. **(Tungadri Bose et al., 2012)**

1.1 COMPRESSION

DNA is a genetic material in most of the eukaryotic organisms. The amount of DNA extracted from these organisms is increasing exponentially which yields major problem like storage and transfer of data efficiently. Compression is the type of technique which can decrease the storage requirements and thereby increase the transmission speed. It can also be used as measure of information content and for making inferences such as relatedness between two sequences. Lossless compression is one of the method in which DNA sequence can be subjected for compression just like other data used in computer files.**(Afify H., 2011)**

1.2 DATA COMPRESSION

Various types of compression algorithms are present which can be used to compress different types of data. When this compression algorithm compress the text document then it is called as Data compression. On the basis of type of data, it can be either lossy or lossless type of data compression. Lossless compression is used for most of the DNA sequences as in this type of compression, the original data is not lost.

1.2.1 LOSSY COMPRESSION

Compression is lossy when the original data cannot be brought back again or cannot be recovered once the file or data is compressed for example Image file (.jpg etc). As Image file once compressed cannot be regained its original size after decompression however the meaning of certain images does not get affected if the decompression losses some bytes.

1.2.2 LOSSLESS COMPRESSION

Compression is lossless when the data or file that has been compressed can be decoded back or after decompression process recovers exact and original form without any loss of information. It is in usually in case of text compression as here losing a character in the text may change its meaning therefore we consider lossless text compression methods. This point is very crucial as decompressed version of the sequence must enclose all the information contained in the genuine sequence.

Sequence - GGGGGGATATATATATGGCGGCGGC

Compression- [G6[AT5[GGC3

Decompression- GGGGGGATATATATATGGCGGCGGC

Run Length Encoding (RLE) is the simplest form of lossless data compression.

Today, there are many DNA sequences which are becoming available and their information for DNA sequences is stored in molecular biology databases. Moreover, their size and importance will be getting bigger in future therefore this information need to stored and communicated efficiently. Furthermore, sequence compression can be used to define the similarities among biological sequences. There exist 2 characteristic structures of DNA sequences. One is called palindrome or reverse compliments and other one is approximate repeats. Several specific algorithm for DNA sequences are present which uses these structures and can compress them in less than two bits per symbol. **(Matsumoto, T., et. al)**

CHAPTER 2 REVIEW OF LITERATURE

2.1 RELATED WORK

Compression algorithms designed for purpose of compression are many available but here some have been used like:-

2.1.1 Run Length Encoding:

Run Length Encoding (RLE) is a simple form of lossless data compression in which runs* of data is stored as single data value and count, rather than as original run*.

**sequences in which the same data value occurs in many consecutive data elements are runs of data while other are called as non runs.*

RLE can be used to compress the homopolymeric nucleotide repeat, di-nucleotide repeat, tri-nucleotide repeat and tandem repeats and tandem repeats.

Fig.2.1:- Compression of DNA sequence by RLE Algorithm

Length

for example:- Sequence ⟶ GGGGGGATATATATATGGCGGCGGC

Compression ⟶ [G6[AT5[GGC3

The RLE algorithm uses these runs to compress the original data while keeping all the non-runs out of reach of this compression process.**(Kodituwakku S.R. et. al).**

2.1.2 LZ algorithm:

LZ algorithm replaces strings of characters with single type of codes. It do not any analyse the incoming text. Instead, it adds every new string of characters to a table of strings. When a single code is an output instead of a string of characters this leads to compression. LZ algorithm compression provides a better compression ratio in most of the applications, that is

why it became the first widely used general-purpose method for compression**(Parihar B. et al., July 2013).**

The LZ algorithm is considered to be a mixture of random characters & repeated substrings.

For example., AT**ACGT**GC**ACGT**TA

↓

Can be coded as ATACGTGC(3,4)TA where (3,4)- denotes repeats 4 characters starting from position 3**(J. Ziv and A. Lempel, 1977).**

2.1.3 DNA Compress:

DNA Compress uses LZW compression scheme. The basic idea of LZW scheme is to replace repeat regions by references to a Dictionary. Compression is done in two phases: (i) Find all approximate repeats including complementary palindromes. (ii) Encode approximate repeat regions and non-repeat regions. Each repeat region is checked to see whether it reduces the size, otherwise that repeat is discarded **(Chen X. et al., 2002).**

2.1.4 LZ77:

It is based on LZ algorithm. It adds every new string of characters to a table of strings. This was the first compression algorithm described by Ziv and Lempel and is commonly referred to as LZ77. The compression is done by replacing repeated occurrence of data with reference to a dictionary. It is the modified form of LZ algorithm, which is used to compress the biological sequences such as DNA and RNA. It can compress the text well by using small amount of memory and fast speed. Memory requirement is very low for both encoding and decoding. However the encoding of the data can be time consuming**(Parihar, B 2013 and Sheng Bao et al., 2005).**

2.1.5 Percentage Compression Ratio:

The Compression amount applied to raw sequence is expressed as ratio. The PCR is the Ratio between the size of the Compressed Data and the size of the Uncompressed data. **(Kodituwakku S.R. et. al).**

$$\text{Percentage Compression Ratio} = \frac{\text{Compressed sequence}}{\text{uncompressed sequence}} \text{ X } 100$$

Genome complexity:

Genome complexity can be analysed by using reassociation kinetics. An organism's DNA solution can be heated in solution until it melts, and then cooled to allow DNA strands to reassociate forming double stranded DNA. This is typically done after shearing the DNA to form any fragments a few hundred bases in length. The larger and more complex an organism's genome is, the longer it will take for complimentary strands to bump into one another and hybridize. **(Primrose, 2013).**

Sequence complexity:

Kolmogorov Complexity:

Complexity of sequence can be measured in the form of compressibility of sequence known as Kolmogorov complexity. Kolmogorov complexity is based on algorithmic information theory considering objects as individual symbol strings, whereas the measures effective measure complexity, excess entropy, predictive information or thermodynamic depth relate objects to random variables and are ensemble based. **(Emmert-Strieb F. 2010).**

CHAPTER 3 SCOPE OF STUDY

Genome sequence may be categorized into two broad classes, coding and non-coding regions. Coding sequences usually consist of unique sequences while non-coding sequences may consist of large amounts of repeat sequences. Unlike unique sequences, repeat sequences can be compressed more effectively in lossless manner. Thus quality of a sequence to get compressed with varying efficiency may be used to assess the genome complexity. Genomes with different extent of repetitive sequences may be compared based on percentage compression ratio (ratio of compressed sequence length and uncompressed sequence length). RLE method is aimed to assess the tandem repeat sequence content while LZ algorithm is expected compress all types of repeat sequences. This study was also aimed to investigate relation between base composition of genome and its compressibility.

CHAPTER 4 OBJECTIVES

- Development of lossless compression algorithm by recording macros for efficient compression of large DNA sequence.
- Comparison of genome complexity based on Genome Sequence Compressibility.

CHAPTER 5 DATA SOURCE

Nucleotide sequence of various organisms were searched from the given NCBI website.

(www.ncbi.nlm.nih.gov)

The Genomes of different organisms have been arranged according to their complexity and were searched through NCBI by their following Accession number :-

Table5.1:- Showing 24 different genomic sequences with their Accession number.

SEQUENCE	ACCESSION NO.
Bacteriophage lambda	NC_001416.1
M13 phage	NC_003287.2
T4 phage	NC_000866.4
T7 phage	NC_001604.1
Haemophilis influenzae chr.	NC_000907.1
Escherichia coli K12 str.	HG738867.1
Sacchromyces cerevisiae mt.	KR260477.1
Caenorhabditis elegans mt.	NC_001328.1
Drosophila melanogaster mt.	NC_024511.2
Danio rerio mt.	NC_002333.2
Mus musculus mt.	NC_005089.1
Homo sapiens mt	NC_012920.1
Hepatitis B virus	AF384371.1
Adeno virus	NC_002077.1
Herpes Simplex Virus 1 Herpes Simplex Virus 2	NC_001806.2 NC_001798.2
Sacchromyces cerevisiae chr.4	NC_001147.6
Caenorhabditis elegans chr.3	NC_003281.10
Drosophila melanogaster chr.3L	NT_037436.4
Danio rerio chr.24	NC_007135.7
Mus musculus chr.19	NC_000085.6
Homo sapiens chr.21	NT_011512.12

Fig5.1:- Screenshot of FASTA sequence of *Homo sapiens chromosome 21.*

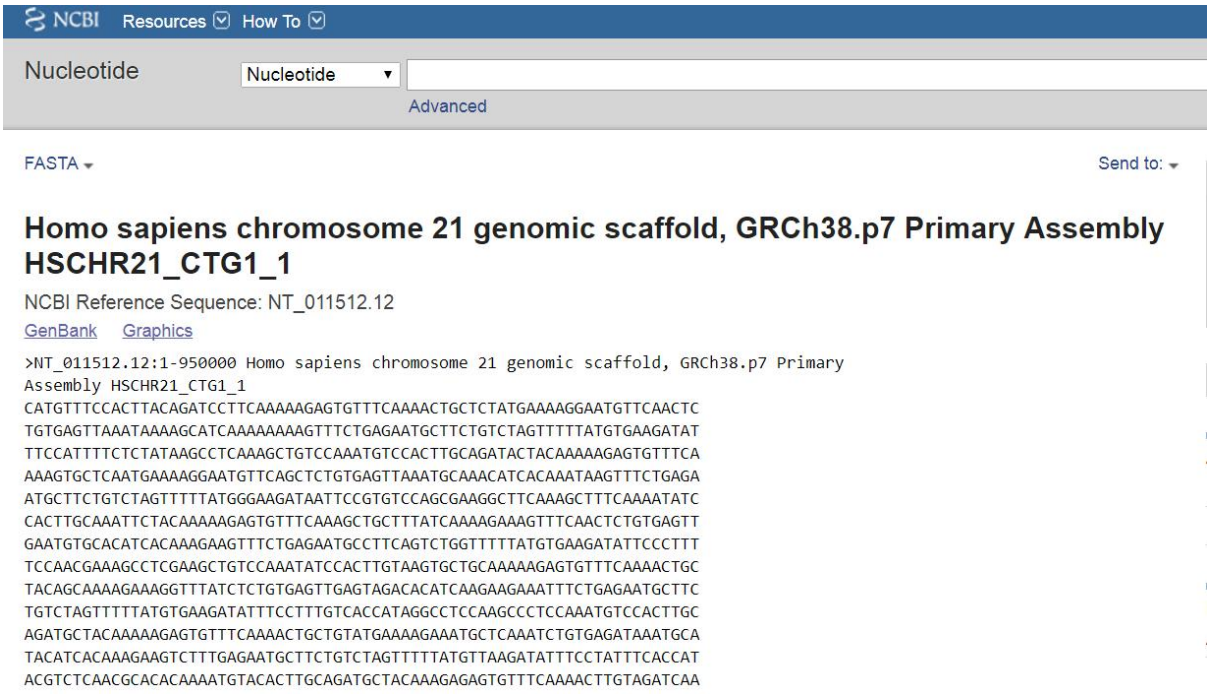
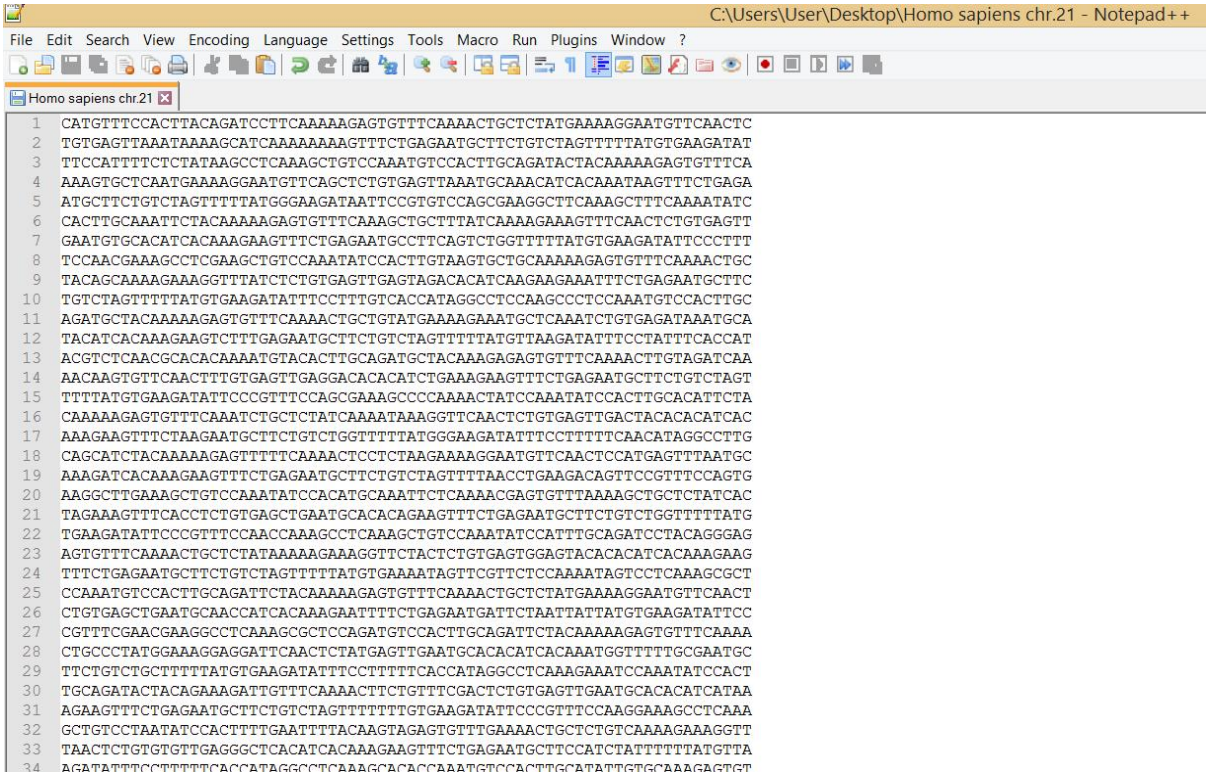


Fig5.2:-Screenshot of *Homo sapiens chromosome 21* sequence in notepad ++.



CHAPTER 6 MATERIAL & METHOD

6.1Sequence Analysis Tools:

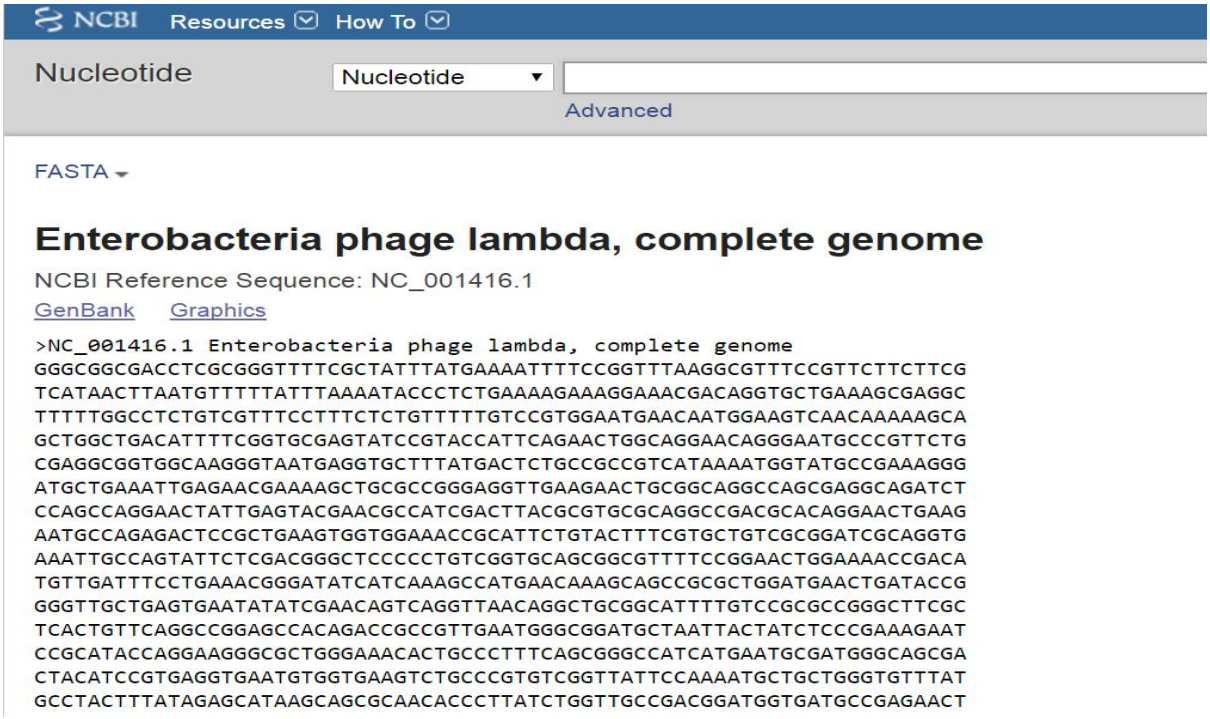
6.1.1Microsoft Excel: Microsoft Excel was used for analysing the statistical data of compressed sequence. Microsoft excel was mainly used for forming graphs and analysing graphical representation.

6.1.2 Notepad ++: Notepad ++ was used to design an algorithm for compressing and decompressing of DNA sequence. Macros were recorded for this purpose for compression and decompression.

6.1.3 FCGR: It is referred to as Computation of Chaos Game Representation of frequencies, where the Base composition was calculated of all possible nucleotide repeats. It was also used to investigate the relation between base composition of genome and its compressibility.

6.2 Methods: Nucleotide sequences of 22 different genomes in the FASTA format were downloaded from the Nucleotide database of NCBI website (www.ncbi.nlm.nih.gov/) through their accession number. Firstly Compressed *Bacteriophage lambda* sequence, accession no. NC_001416.1, 48502 bp long by RLE algorithm.

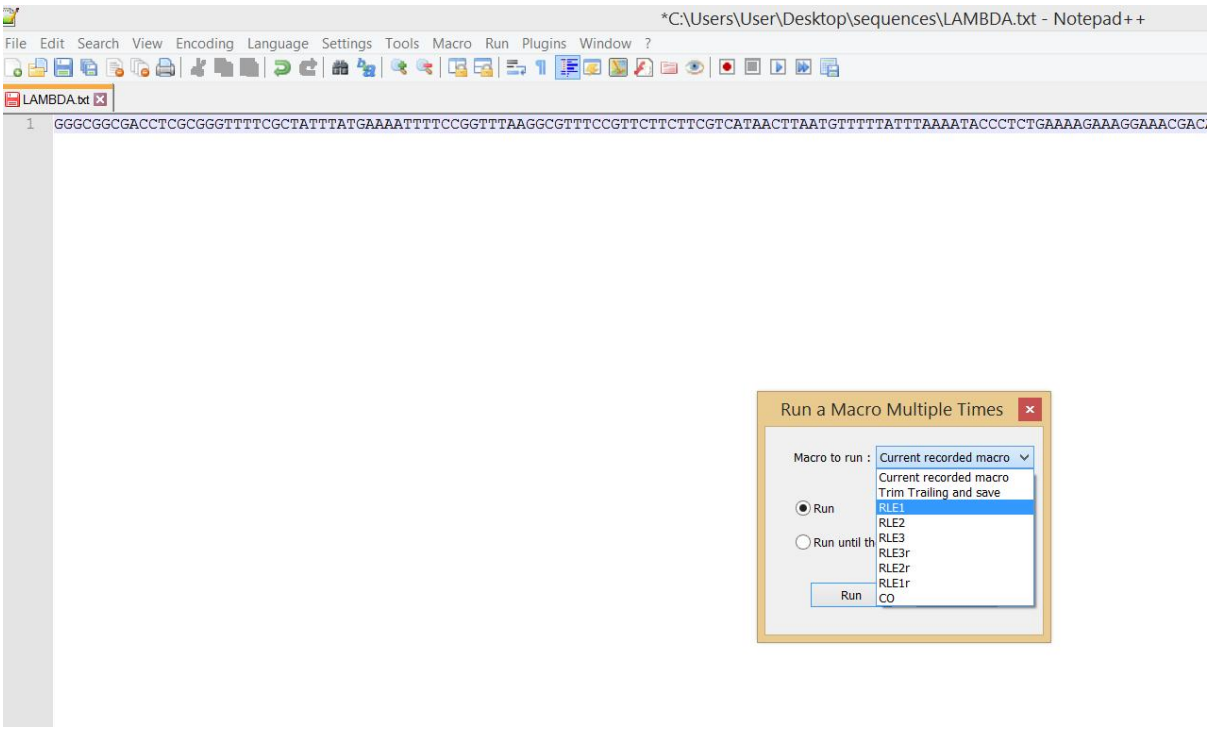
Fig6.1:- Screenshot showing *Bacteriophage lambda* sequence.



The FASTA sequence was opened in Notepad++. Macros were recorded to remove all spaces. Macros were recorded to design the RLE algorithm for compression of sequences. It was again recorded to decompress the compressed sequence.

Run Length Encoding (RLE) compression algorithm was used for compression of the nucleotide sequence. It was used to compress homopolymeric nucleotide repeats, dinucleotide repeats and tri-nucleotide repeats. Lower limit for compression of the DNA sequence for homopolymeric nucleotide repeats, dinucleotide repeats and trinucleotide repeats is 4, 3 and 2 respectively.

Fig6.2:-Screenshot of running macros for Compression and decompression.



RLE1 ⟶ Compression for Homopolymeric repeats

RLE2 ⟶ Compression for Di-nucleotide repeats

RLE3 ⟶ Compression for Tri-nucleotide repeats

6.3 Calculation of upper limits of repeat units

To set the upper limit of repeat units, the probabilities of occurring homopolymeric nucleotide repeats, di-nucleotide repeats and tri-nucleotide repeats in a DNA sequence of 10^7 nucleotide bases were calculated using the following formulae;

for homopolymeric nucleotide repeats

$$\mathbf{[P(A)^n] \times N = 1}$$

$$\mathbf{[P(A)^n] = 1/N}$$

Apply natural log on both sides

$$\mathbf{n \ln P(A) = \ln(1/N)}$$

$$\mathbf{n_A = [\ln (1/N) / \ln P(A)]}$$

Where $N=10^7$ and $P(A) =1/4$

$$\mathbf{n_A = (\ln 1/10^7) / (\ln 1/4)}$$

$$\mathbf{n_A = 11.62}$$

$$\mathbf{n_A \sim 12}$$

For di-nucleotide repeat:

$$\mathbf{[P(AA)^n] \times N=1}$$

$$\mathbf{[P(AA)^n] = 1/ N}$$

Apply natural log on both sides:

$$\mathbf{n \ln P (AA) = \ln(1/N)}$$

$$\mathbf{n_{AA} = [\ln (1/N)]/ \ln P(AA)}$$

Where $N= 10^7$, $P (AA) = 1/16$

$$\mathbf{n_{AA} = (\ln 1/10^7) / (\ln 1/16)}$$

$$\mathbf{n_{AA} = 5.81}$$

$\mathbf{n_{AA} \sim 6}$

For tri-nucleotide repeat:

$\mathbf{[P\ (AAA)^n] \times N=1}$

$\mathbf{[P\ (AAA)^n] = 1/\ N}$

Apply natural log on both sides:

$\mathbf{n\ ln\ P\ (AAA) = ln\ (1/N)}$

$\mathbf{n_{AAA} = [\ ln\ (1/N)\ ]/\ ln\ P(AAA)}$

Where N= 10^7, P (AAA) = 1/64

$\mathbf{n_{AAA} = (ln\ 1/10^7)\ /\ (ln\ 1/64)}$

$\mathbf{n_{AAA} = 3.87}$

$\mathbf{n_{AAA} \sim 4}$

Where A- homopolymeric repeat, AA- dinucleotide repeat, AAA- trinucleotide repeat, P- probability

Therefore the upper limit to compress the DNA sequence for homopolymeric nucleotide repeat, dinucleotide repeat and tri-nucleotide repeat is 12, 6 and 4 respectively. However, addition of two more repeats will signify the rarer length of repeats. So, the upper limit to compress the DNA sequence for homopolymeric sequence is 14 for dinucleotide repeat is 8 and for trinucleotide repeat is 6.

Here, the upper limit is based on probilistic model of length of repeats. There may be longer repeats which may reduce the efficiency of compression in such cases.

Homopolymeric nucleotide repeat, dinucleotide repeat and tri-nucleotide repeat considered for compression were:

Homopolymeric repeats
As
Gs
Ts
Cs
Di-nucleotide repeats
GA or AG
GT or TG
GC or CG
AT or TA
AC or CA
TC or CT
Tri-nucleotide repeats
GGA or GAG or AGG
GGC or GCG or CGG
GGT or GTG or TGG
GAA or AAG or GAA
GAC or ACG or CGA
GAT or ATG or GAT
GCA or CAG or AGC
GCC or CCG or CGC
GCT or CTG or TGC
GTA or TAG or AGT
GTC or TCG or CGT
GTT or TTG or TGT
AAC or ACA or CAA
AAT or ATA or TAA
ACC or CCA or CAC
ACT or CTA or TAC
ATC or TCA or CAT

ATT or TTA or TAT
CCT or CTC or TCC
CTT or TTC or TCT

The macros were recorded separately for homopolymeric nucleotide repeat as N, for dinucleotide repeat as NN and for tri-nucleotide repeat as NNN.

Table6.1:-Homopolymeric Repeats were replaced as follows:

S.No	Repeats	Replacement
FOR HOMOPOLYMERIC REPEATS		
i.	AAAAAAAAAAAAAA	[A14
ii.	AAAAAAAAAAAAA	[A13
iii.	AAAAAAAAAAAA	[A12
iv.	AAAAAAAAAAA	[A11
v.	AAAAAAAAAA	[A10
vi.	AAAAAAAAA	[A9
vii.	AAAAAAAA	[A8
viii.	AAAAAAA	[A7
ix.	AAAAAA	[A6
x.	AAAAA	[A5
xi.	AAAA	[A4
i.	GGGGGGGGGGGGGG	[G14
ii.	GGGGGGGGGGGGG	[G13
iii.	GGGGGGGGGGGG	[G12
iv.	GGGGGGGGGGG	[G11
v.	GGGGGGGGGG	[G10
vi.	GGGGGGGGG	[G9
vii.	GGGGGGGG	[G8
viii.	GGGGGGG	[G7
ix.	GGGGGG	[G6

x.	GGGGG	[G5
xi.	GGGG	[G4
i.	CCCCCCCCCCCCCC	[C14
ii.	CCCCCCCCCCCCC	[C13
iii.	CCCCCCCCCCCC	[C12
iv.	CCCCCCCCCCC	[C11
v.	CCCCCCCCCC	[C10
vi.	CCCCCCCCC	[C9
vii.	CCCCCCCC	[C8
viii.	CCCCCCC	[C7
ix.	CCCCCC	[C6
x.	CCCCC	[C5
xi.	CCCC	[C4
i.	TTTTTTTTTTTTTT	[T14
ii.	TTTTTTTTTTTTT	[T13
iii.	TTTTTTTTTTTT	[T12
iv.	TTTTTTTTTTT	[T11
v.	TTTTTTTTTT	[T10
vi.	TTTTTTTTT	[T9
vii.	TTTTTTTT	[T8
viii.	TTTTTTT	[T7
ix.	TTTTTT	[T6
x.	TTTTT	[T5
xi.	TTTT	[T4

Accordingly, the algorithm for Homopolymeric repeats was obtained. Similarly it was made for Dinucleotide Repeats and Trinucleotide Repeats.

Table6.2:- Dinucleotide and Trinucleotide Repeats were replaced as follows:

FOR DI-NUCLEOTIDE SEQUENCES		
i.	GAGAGAGAGAGAGAGA	[GA8
ii.	GAGAGAGAGAGAGA	[GA7
iii.	GAGAGAGAGAGA	[GA6
iv.	GAGAGAGAGA	[GA5
v.	GAGAGAGA	[GA4
vi.	GAGAGA	[GA3
i.	GTGTGTGTGTGTGTGT	[GT8
ii.	GTGTGTGTGTGTGT	[GT7
iii.	GTGTGTGTGTGT	[GT6
iv.	GTGTGTGTGT	[GT5
v.	GTGTGTGT	[GT4
vi.	GTGTGT	[GT3
i.	GCGCGCGCGCGCGCGC	[GC8
ii.	GCGCGCGCGCGCGC	[GC7
iii.	GCGCGCGCGCGC	[GC6
iv.	GCGCGCGCGC	[GC5
v.	GCGCGCGC	[GC4
vi.	GCGCGC	[GC3
i.	ATATATATATATATAT	[AT8
ii.	ATATATATATATAT	[AT7
iii.	ATATATATATAT	[AT6
iv.	ATATATATAT	[AT5
v.	ATATATAT	[AT4
vi.	ATATAT	[AT3

i.	ACACACACACACACAC	[AC8
ii.	ACACACACACACAC	[AC7
iii.	ACACACACACAC	[AC6
iv.	ACACACACAC	[AC5
v.	ACACACAC	[AC4
vi.	ACACAC	[AC3
i.	TCTCTCTCTCTCTCTC	[TC8
ii.	TCTCTCTCTCTCTC	[TC7
iii.	TCTCTCTCTCTC	[TC6
iv.	TCTCTCTCTC	[TC5
v.	TCTCTCTC	[TC4
vi.	TCTCTC	[TC3
FOR TRI-NUCLEOTIDE SEQUENCES		
i.	GGAGGAGGAGGAGGAGGA	[GGA6
ii.	GGAGGAGGAGGAGGA	[GGA5
iii.	GGAGGAGGAGGA	[GGA4
iv.	GGAGGAGGA	[GGA3
v.	GGAGGA	[GGA2
i.	GGCGGCGGCGGCGGCGGC	[GGC6
ii.	GGCGGCGGCGGCGGC	[GGC5
iii.	GGCGGCGGCGGC	[GGC4
iv.	GGCGGCGGC	[GGC3
v.	GGCGGC	[GGC2

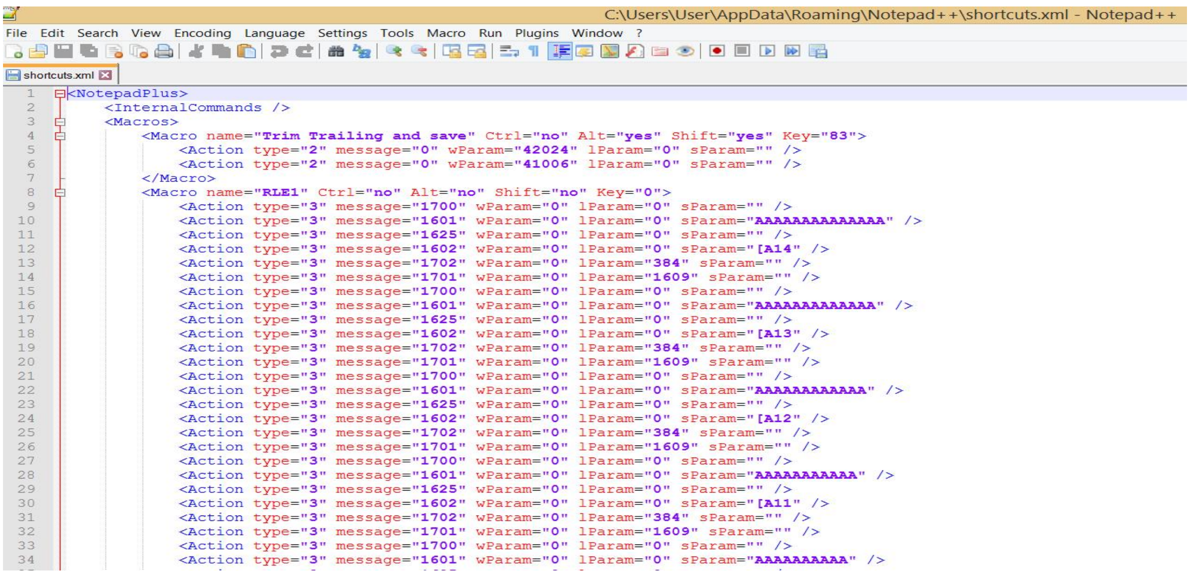
i.	GGTGGTGGTGGTGGTGGT	[GGT6
ii.	GGTGGTGGTGGTGGT	[GGT5
iii.	GGTGGTGGTGGT	[GGT4
iv.	GGTGGTGGT	[GGT3
v.	GGTGGT	[GGT2
i.	GAAGAAGAAGAAGAAGAA	[GAA6
ii.	GAAGAAGAAGAAGAA	[GAA5
iii.	GAAGAAGAAGAA	[GAA4
iv.	GAAGAAGAA	[GAA3
v.	GAAGAA	[GAA2
i.	GACGACGACGACGACGAC	[GAC6
ii.	GACGACGACGACGAC	[GAC5
iii.	GACGACGACGAC	[GAC4
iv.	GACGACGAC	[GAC3
v.	GACGAC	[GAC2
i.	GATGATGATGATGATGAT	[GAT6
ii.	GATGATGATGATGAT	[GAT5
iii.	GATGATGATGAT	[GAT4
iv.	GATGATGAT	[GAT3
v.	GATGAT	[GAT2
i.	GCAGCAGCAGCAGCAGCA	[GCA6
ii.	GCAGCAGCAGCAGCA	[GCA5
iii.	GCAGCAGCAGCA	[GCA4
iv.	GCAGCAGCA	[GCA3
v.	GCAGCA	[GCA2

i.	GCCGCCGCCGCCGCCGCC	[GCC6
ii.	GCCGCCGCCGCCGCC	[GCC5
iii.	GCCGCCGCCGCC	[GCC4
iv.	GCCGCCGCC	[GCC3
v.	GCCGCC	[GCC2
i.	GCTGCTGCTGCTGCTGCT	[GCT6
ii.	GCTGCTGCTGCTGCT	[GCT5
iii.	GCTGCTGCTGCT	[GCT4
iv.	GCTGCTGCT	[GCT3
v.	GCTGCT	[GCT2
i.	GTAGTAGTAGTAGTAGTA	[GTA6
ii.	GTAGTAGTAGTAGTA	[GTA5
iii.	GTAGTAGTAGTA	[GTA4
iv.	GTAGTAGTA	[GTA3
v.	GTAGTA	[GTA2
i.	GTCGTCGTCGTCGTCGTC	[GTC6
ii.	GTCGTCGTCGTCGTC	[GTC5
iii.	GTCGTCGTCGTC	[GTC4
iv.	GTCGTCGTC	[GTC3
v.	GTCGTC	[GTC2
i.	GTTGTTGTTGTTGTTGTT	[GTT6
ii.	GTTGTTGTTGTTGTT	[GTT5
iii.	GTTGTTGTTGTT	[GTT4
iv.	GTTGTTGTT	[GTT3
v.	GTTGTT	[GTT2

i.	AACAACAACAACAACAAC	[AAC6
ii.	AACAACAACAACAAC	[AAC5
iii.	AACAACAACAAC	[AAC4
iv.	AACAACAAC	[AAC3
v.	AACAAC	[AAC2
i.	AATAATAATAATAATAAT	[AAT6
ii.	AATAATAATAATAAT	[AAT5
iii.	AATAATAATAAT	[AAT4
iv.	AATAATAAT	[AAT3
v.	AATAAT	[AAT2
i.	ACCACCACCACCACCACC	[ACC6
ii.	ACCACCACCACCACC	[ACC5
iii.	ACCACCACCACC	[ACC4
iv.	ACCACCACC	[ACC3
v.	ACCACC	[ACC2
i.	ACTACTACTACTACTACT	[ACT6
ii.	ACTACTACTACTACT	[ACT5
iii.	ACTACTACTACT	[ACT4
iv.	ACTACTACT	[ACT3
v.	ACTACT	[ACT2
i.	ATCATCATCATCATCATC	[ATC6
ii.	ATCATCATCATCATC	[ATC5
iii.	ATCATCATCATC	[ATC4
iv.	ATCATCATC	[ATC3
v.	ATCATC	[ATC2

i.	ATTATTATTATTATTATT	[ATT6
ii.	ATTATTATTATTATT	[ATT5
iii.	ATTATTATTATT	[ATT4
iv.	ATTATTATT	[ATT3
v.	ATTATT	[ATT2
i.	CCTCCTCCTCCTCCTCCT	[CCT6
ii.	CCTCCTCCTCCTCCT	[CCT5
iii.	CCTCCTCCTCCT	[CCT4
iv.	CCTCCTCCT	[CCT3
v.	CCTCCT	[CCT2
i.	CTTCTTCTTCTTCTTCTT	[CTT6
ii.	CTTCTTCTTCTTCTT	[CTT5
iii.	CTTCTTCTTCTT	[CTT4
iv.	CTTCTTCTT	[CTT3
v.	CTTCTT	[CTT2

Fig6.3:- Screenshot of xml file for shortcuts of homopolymeric repeats.



```
<NotepadPlus>
    <InternalCommands />
    <Macros>
        <Macro name="Trim Trailing and save" Ctrl="no" Alt="yes" Shift="yes" Key="83">
            <Action type="2" message="0" wParam="42024" lParam="0" sParam="" />
            <Action type="2" message="0" wParam="41006" lParam="0" sParam="" />
        </Macro>
        <Macro name="RLE1" Ctrl="no" Alt="no" Shift="no" Key="0">
            <Action type="3" message="1700" wParam="0" lParam="0" sParam="" />
            <Action type="3" message="1601" wParam="0" lParam="0" sParam="AAAAAAAAAAAAAA" />
            <Action type="3" message="1625" wParam="0" lParam="0" sParam="" />
            <Action type="3" message="1602" wParam="0" lParam="0" sParam="[A14" />
            <Action type="3" message="1702" wParam="0" lParam="384" sParam="" />
            <Action type="3" message="1701" wParam="0" lParam="1609" sParam="" />
            <Action type="3" message="1700" wParam="0" lParam="0" sParam="" />
            <Action type="3" message="1601" wParam="0" lParam="0" sParam="AAAAAAAAAAAAA" />
            <Action type="3" message="1625" wParam="0" lParam="0" sParam="" />
            <Action type="3" message="1602" wParam="0" lParam="0" sParam="[A13" />
            <Action type="3" message="1702" wParam="0" lParam="384" sParam="" />
            <Action type="3" message="1701" wParam="0" lParam="1609" sParam="" />
            <Action type="3" message="1700" wParam="0" lParam="0" sParam="" />
            <Action type="3" message="1601" wParam="0" lParam="0" sParam="AAAAAAAAAAAA" />
            <Action type="3" message="1625" wParam="0" lParam="0" sParam="" />
            <Action type="3" message="1602" wParam="0" lParam="0" sParam="[A12" />
            <Action type="3" message="1702" wParam="0" lParam="384" sParam="" />
            <Action type="3" message="1701" wParam="0" lParam="1609" sParam="" />
            <Action type="3" message="1700" wParam="0" lParam="0" sParam="" />
            <Action type="3" message="1601" wParam="0" lParam="0" sParam="AAAAAAAAAAA" />
            <Action type="3" message="1625" wParam="0" lParam="0" sParam="" />
            <Action type="3" message="1602" wParam="0" lParam="0" sParam="[A11" />
            <Action type="3" message="1702" wParam="0" lParam="384" sParam="" />
            <Action type="3" message="1701" wParam="0" lParam="1609" sParam="" />
            <Action type="3" message="1700" wParam="0" lParam="0" sParam="" />
            <Action type="3" message="1601" wParam="0" lParam="0" sParam="AAAAAAAAAA" />
```

Fig6.4:-Screenshot of xml file for shortcuts of di-nucleotide repeats.

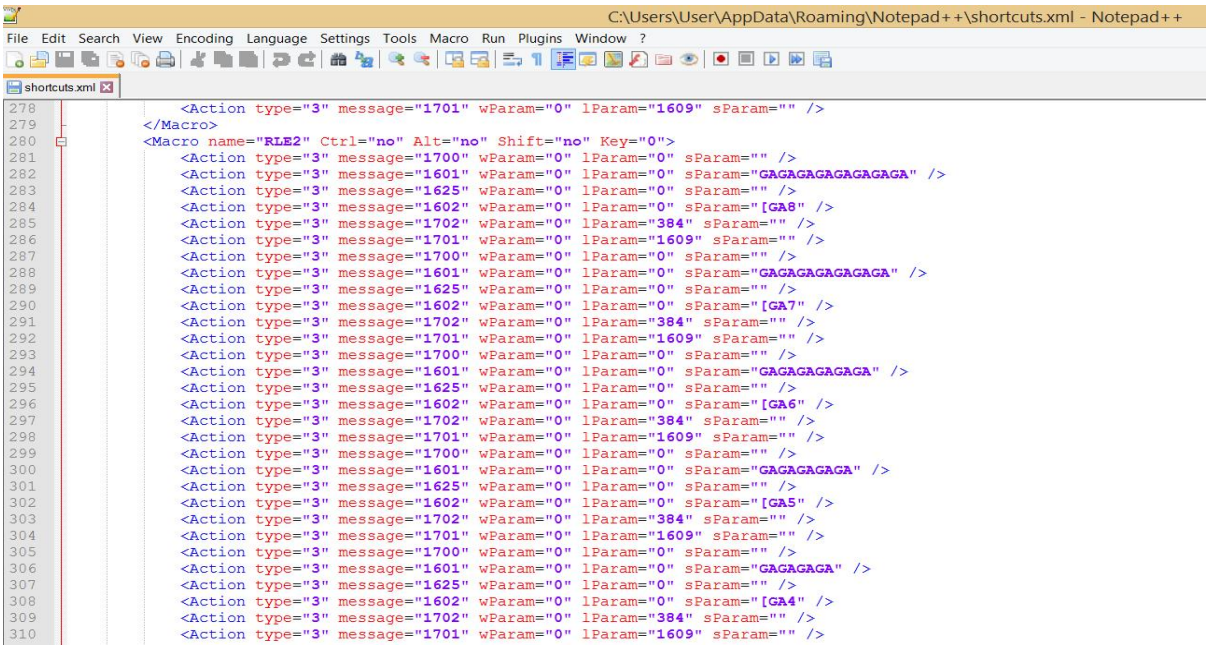
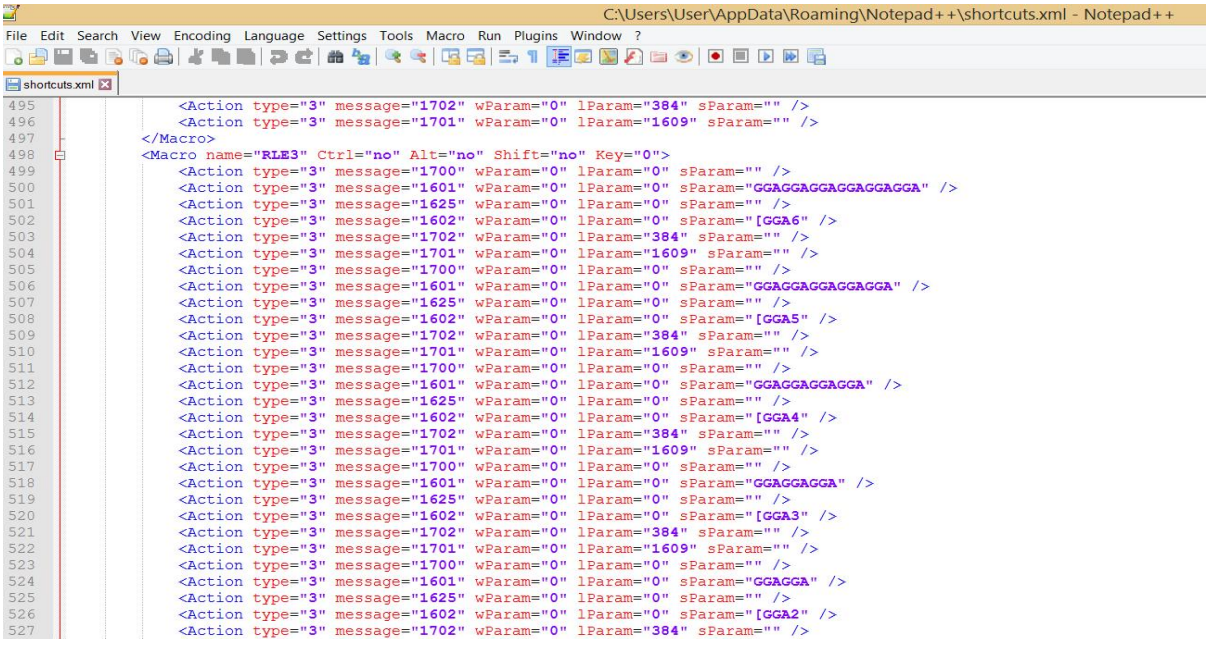


fig6.5:-Screenshot of xml file for shortcuts of tri-nucleotide repeats.



(xml files helps in viewing the recorded Macro in Notepad ++.)

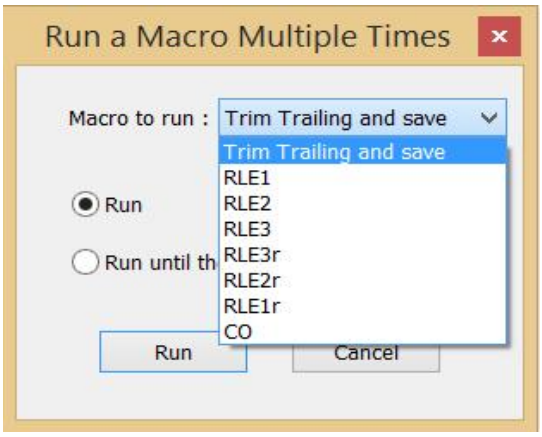
CHAPTER 7 RESULT

Genome contains genetic material i.e., DNA or RNA. It includes coding and noncoding DNA and the genetic material of the Mitochondria and Chloroplast. Genomic sequence contains unique sequences and repeated sequences. More will be the sequence complexity if there is more number of unique sequences. Hence the reassociation kinetics will be more in this case as it has more unique sequences. Unique sequences mostly occur in Coding regions of DNA whereas Repetitive sequences are more likely to occur in Non coding regions of DNA.

Repetitive sequences of DNA consist of Tandem repeats (at least 10^5 copies per genome) and Dispersed repeats (it varies from about 20% to more than 80% of the total DNA depending upon organism).

7.1 Genomic sequences with low complexity data can be compressed. Run Length Encoding (RLE) algorithm was developed using Notepad ++. 24 different Genomic sequences were taken from NCBI website (www.ncbi.nlm.nih.gov). RLE algorithm was obtained by recording Macros in Notepad ++. Macro was recorded for homopolymeric nucleotide repeats, di-nucleotide repeats and tri nucleotide repeats and saved. Similarly decompression algorithm was also developed using this method to check RLE compression.

fig7.1: Screenshot of saved RLE algorithms.



7.2 After performing RLE, the Percentage Compression Ratio (PCR) was calculated using formula as follows:

$$\text{Percentage Compression Ratio} = \frac{\text{Compressed sequence}}{\text{uncompressed sequence}} \text{ X } 100$$

7.3 Confirmation of lossless nature of the compression algorithm.

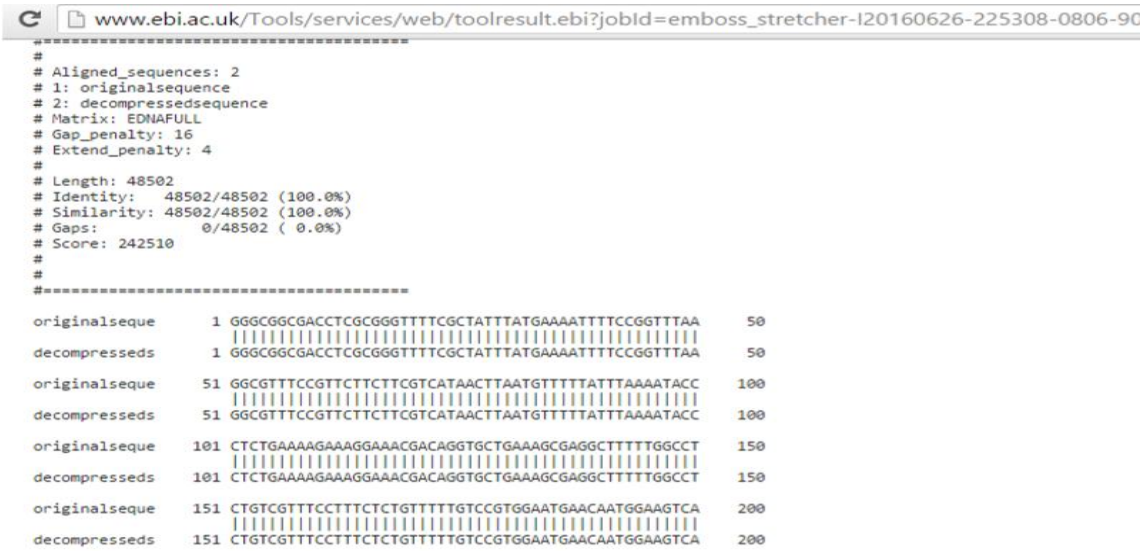
Decompression and alignment: Lossless compression was proved by the macro recording by 2 methods. To validate the lossless nature of compression algorithm:

1) Pairwise alignment was done of Uncompressed sequence with Compressed followed by Decompressed DNA sequence in MEGA.

Fig7.2:- Screenshot of Aligned sequence in MEGA 7

2) Second method to check whether the compression is lossy or lossless was done by aligning the sequences using pairwise sequence alignment.

Fig7.3:-Screenshot of Pairwise sequence alignment.



A 100% sequence identity between the uncompressed and compressed followed by uncompressed sequence verified that RLE method generated for DNA sequence compression was a lossless compression method.

Table7.1:- Consisting of PCR values by RLE method

S.No	SEQUENCE	SEQUENCE LENGTH	COMPRESSION BY RLE	PCR VALUES
1	*Bacteriophage lambda*	48502	47002	96.91
2	T7 phage	168903	39432	98.74
3	T4 phage	6407	162811	96.39
4	M13 phage	39937	6169	96.29
5	Adeno virus	4718	4582	97.12
6	Hepatitis B virus	3215	3109	96.7
7	Herpes Simplex Virus 1	152222	142861	93.85
8	Herpes Simplex Virus 2	154653	144534	93.46
9	*Sacchromyces cerevisiae* mt	76596	67451	88.06
10	*Caenorhabditis elegans* mt	13794	12678	91.91
11	*Drosophila melanogaster* mt	19524	17393	89.09
12	*Danio rerio* mt	16596	15921	95.93
13	*Mus musculus* mt	16299	15707	96.37
14	*Homo sapiens* mt	16569	15899	95.96
15	*Haemophilis influenza*	1830138	1750176	95.63
16	*Escherichia coli* K12 str.	4527246	4397658	97.14
17	*Sacchromyces cerevisiae* chr.4	1091291	1043107	95.58
18	*Caenorhabditis elegans* chr.3	5699999	5199688	91.22
19	*Drosophila melanogaster* chr.3L	5699973	5442043	95.47
20	*Danio rerio* chr.24	5699980	5364930	94.12
21	*Mus musculus* chr.19	5700005	5443523	95.5
22	*Homo sapiens* chr.21	5700005	5413796	94.98

7.4 Genomic sequences can also be compressed by LZ algorithm as it compress both Tandem and Dispersed repeats whereas RLE algorithm compress only Tandem repeats. Basically LZ algorithm is considered to be a mixture of random characters and repeated substrings.

Table7.2:- Consisting of PCR values by LZ method

S.No	SEQUENCE	SEQUENCE LENGTH	COMPRESSION BY LZ	PCR VALUES
1	*Bacteriophage lambda*	48502	14668	30.24
2	T7 phage	168903	46395	28.52
3	T4 phage	6407	2068	27.47
4	M13 phage	39937	11392	32.28
5	Adeno virus	4718	1494	31.67
6	Hepatitis B virus	3215	1055	32.81
7	Herpes Simplex Virus 1	152222	44101	28.97
8	Herpes Simplex Virus 2	154653	44183	28.57
9	*Sacchromyces cerevisiae* mt	76596	18346	23.95
10	*Caenorhabditis elegans* mt	13794	3934	28.52
11	*Drosophila melanogaster* mt	19524	5151	26.38
12	*Danio rerio* mt	16596	4789	28.86
13	*Mus musculus* mt	16299	4619	28.34
14	*Homo sapiens* mt	16569	4732	28.56
15	*Haemophilis influenzae*	1830138	538285	29.41
16	*Escherichia coli* K12 str.	4527246	1267485	28
17	*Sacchromyces cerevisiae* chr.4	1091291	303902	27.85
18	*Caenorhabditis elegans* chr.3	5699999	1510455	26.5
19	*Drosophila melanogaster* chr.3L	5699973	1592780	27.94
20	*Danio rerio* chr.24	5699980	1553501	27.25
21	*Mus musculus* chr.19	5700005	1568676	27.45
22	*Homo sapiens* chr.21	5700005	1551774	27.22

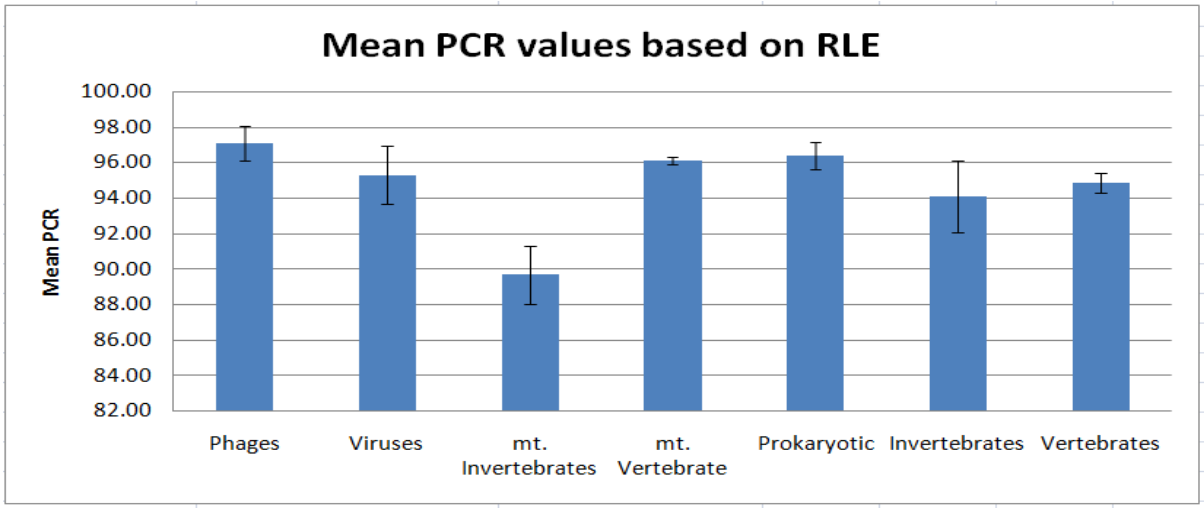
$$\frac{\text{RLE algorithm}}{\text{LZ algorithm}} \Rightarrow \frac{\text{Compresses Tandem Repeats}}{\text{Compresses Tandem and Dispersed Repeats}}$$

Table7.3:- Consisting of Ratios obtained by RLE/LZ.

S.No	SEQUENCE	SEQUENCE LENGTH	PCR VALUES BASED RLE	PCR VALUES BASED LZ	RATIO RLE/LZ
1	*Bacteriophage lambda*	48502	96.91	30.24	3.205
2	T7 phage	168903	98.74	28.52	3.462
3	T4 phage	6407	96.39	27.47	3.509
4	M13 phage	39937	96.29	32.28	2.983
5	Adeno virus	4718	97.12	31.67	3.067
6	Hepatitis B virus	3215	96.7	32.81	2.947
7	Herpes Simplex Virus 1	152222	93.85	28.97	3.240
8	Herpes Simplex Virus 2	154653	93.46	28.57	3.271
9	*Sacchromyces cerevisiae* mt	76596	88.06	23.95	3.677
10	*Caenorhabditis elegans* mt	13794	91.91	28.52	3.223
11	*Drosophila melanogaster* mt	19524	89.09	26.38	3.377
12	*Danio rerio* mt	16596	95.93	28.86	3.324
13	*Mus musculus* mt	16299	96.37	28.34	3.400
14	*Homo sapiens* mt	16569	95.96	28.56	3.360
15	*Haemophilis influenzae*	1830138	95.63	29.41	3.252
16	*Escherichia coli* K12 str.	4527246	97.14	28	3.469
17	*Sacchromyces cerevisiae* chr.4	1091291	95.58	27.85	3.432
18	*Caenorhabditis elegans* chr.3	5699999	91.22	26.5	3.442
19	*Drosophila melanogaster* chr.3L	5699973	95.47	27.94	3.417
20	*Danio rerio* chr.24	5699980	94.12	27.25	3.454
21	*Mus musculus* chr.19	5700005	95.5	27.45	3.479
22	*Homo sapiens* chr.21	5700005	94.98	27.22	3.489

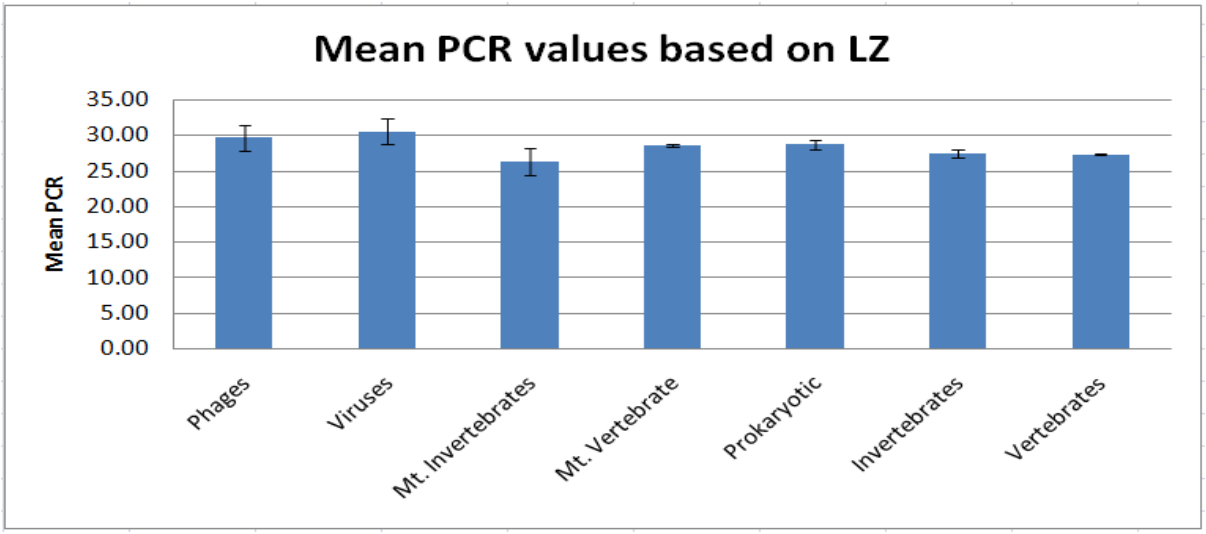
The first graph shows that the highest mean PCR value based on RLE compression was of Phages as this shows that it contained least number of Tandem repeats. Hence it shows lowest compressibility. Similarly lowest Mean PCR value came out to be of mitochondrial Vertebrates containing very high content of tandem repeats.

Fig7.4:- Graph showing MEAN values based on PCR_{RLE}.



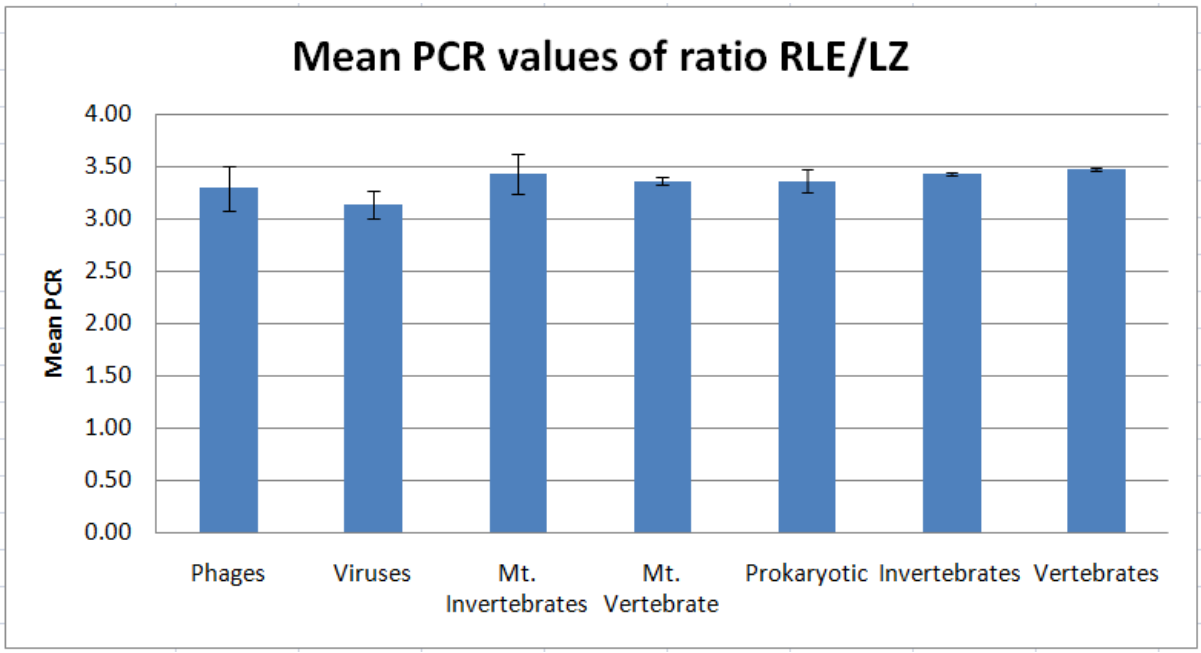
As LZ alg. is more efficient than RLE alg. as it compress the sequences containing Tandem as well as Dispersed repeats. So in second graph the least compression came out to be by LZ is of viruses. This shows it contains very low number of Dispersed repeats than other categorized genomic sequences.

Fig7.5:- Graph showing MEAN values based on PCR_{LZ}.



As LZ compression is more effective on non-coding regions. By interpreting the third graph the highest ratio value was of Vertebrates which signifies that LZ gives higher compression. However Viruses shows that LZ gives lowest compression among 7 categories.

Fig7.6:- Graph showing ratio of MEAN values based on PCR_{RLE}/PCRLZ.



Further it was attempted to study any relation between DNA sequence complexity and its base composition. PCR value was considered as measure of complexity and it was correlated with coefficient of variation (CV*) as a measure of heterogeneity of the base composition. Correlation coefficients were calculated between PCRs of various genomic sequences and their respective CVs. The results indicate that if CV is low the Heterogeneity measure will be high which leads to high compression and low complexity. Thus the Percentage Compression Ratio value will become low. Hence CV is inversely proportion to Percentage Compression Ratio (PCR). This was evident from the fact that correlation coefficients were less than zero.

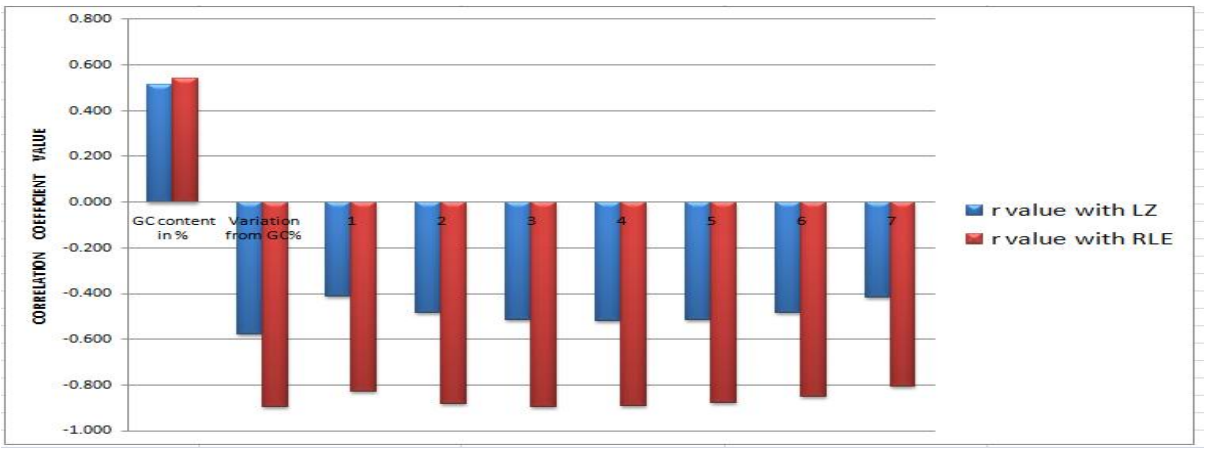
**It's an attribute which can be used to compare the data dispersion in different data sets.*

Table7.4:- Correlation coefficient (r) values and their respective P values with PCR_{RLE} and PCR_{LZ}.

	PCR_{RLE}		PCR_{LZ}	
	r	p-value	r	p-value
Compression by RLE	1	1	0.625	0.001091713
Compression by LZ	0.625	0.001091713	1	1
Sequence length	-0.001	0.997292353	-0.306	0.146516282
GC%	0.541	0.006304984	0.514	0.010178043
Variation from GC%	-0.898	2.60122E-09	-0.518	0.002923629
1	-0.828	5.8081E-07	-0.412	0.04558826
2	-0.881	1.28232E-08	-0.486	0.015953402
3	-0.895	3.57915E-09	-0.516	0.009916405
4	-0.892	4.98273E-09	-0.521	0.008982996
5	-0.88	1.50105E-08	-0.516	0.009857266
6	-0.854	1.09383E-07	-0.486	0.016100474
7	-0.808	1.83498E-06	-0.419	0.041411702

The negative correlation was stronger with LZ based PCR when compared to RLE based PCR. It showed that heterogeneity of the sequence affects PCR and as a result, sequence complexity when not only tandem repeats but also scattered repeats were taken into account. The correlation values were determined between PCR based on RLE or LZ method and CV of mono-, di-, tri-, tetra-, penta-, hexa-, and hepta-nucleotides. All the correlation coefficients were negative however the strongest effect was shown by tetra-nucleotide CV values. It may be concluded that tetra-nucleotide is the optimal word size for assessing sequence heterogeneity.

Fig7.7:- Graph showing Correlation Coefficient values.



CHAPTER 8 DISCUSSION

DNA as the genetic material contains many features such as, It can store genetic information, Able to generate its replica through the process of replication, It should have its own mechanism to decode the genetic information into functional molecules such as Proteins. Genome contains entire information about genetic material i.e., DNA or RNA as well as it includes coding and non coding DNA and the genetic material of the Mitochondria and Chloroplast. The DNA sequence contains Exons as well as Introns. Where exons are coding sequences and rest all are non coding sequences. The genetic sequence contains genomic information which leads to complexity of genome. In simpler organisms almost the whole of the DNA consists of unique sequences. In higher organisms there can be large amounts of repetitive DNA which decreases the complexity. Here, our analysis was on 24 different genomic sequences containing Phages, Viruses, Mitochondrial DNA, Chloroplast DNA, Prokaryotic DNA, Eukaryotic DNA. There are two types of repetitive DNA: First is Tandem repeats these sequences are present in at least 10^5 copies per genome and they are typically short and are present in clusters in which the given sequence repeats itself, over and over again without interruption and second is Dispersed repeats which are moderately repeated fraction of the genomes of plants and animals i.e. it varies from about 20% to more than 80% of the total DNA depending upon organism. The length of the non-repetitive DNA component tends to increase as the complexity of organisms increase. Large amount of DNA present in the plants and animals indicates the presence of repetitive DNA. Most genes are present in non- repetitive DNA. This indicates that genetic complexity is proportional to the amount of non- repetitive DNA. Also, the $C_0t_{1/2}$ value tends to increase with the complexity of genomic sequence.

Kolmogorov complexity is the measure of computational resources that specifies the particular object, and is also called as descriptive complexity. Moreover, it can also be measured in the form of compressibility of sequence. It is based on Algorithmic Information Theory considering objects as individual symbol strings and also it has the property to remain unchanged if the size of the system changes.

In humans, more than 98% portion of the DNA is non-coding. Therefore these sequences can be compressed. Compression can be used to find out the relatedness of coding and non-coding sequences. The Percent Compression Ratio of coding and non-coding regions of DNA

is different. Therefore, the compression technique can be used to test the non- randomness of different type of DNA sequence.

Compression is the type of technique which can decrease the storage requirements and thereby increase the transmission speed. When the compression algorithm compress the text document then it is called as Data compression. On the basis of type of data, it can be either lossy or lossless type of data compression. Run-length Encoding (RLE) is a form of lossless data compression in which runs of data are stored as a single data value and count. This is most useful on data that contains many such runs. For DNA sequence, RLE can be used to compress the homopolymeric nucleotide repeat, di-nucleotide repeat, tri-nucleotide repeat and tandem repeats.

We designed RLE compression algorithm for homopolymeric, dinucleotide and tri-nucleotide repeats in the Notepad++. It is a type of lossless compression algorithm and to prove this we designed a decompression algorithm using Notepad++ which was the reverse of compression algorithm.

We took out the Mean Percentage Compression Ratio (PCR) with RLE compressed sequences. Among 7 categories i.e., Phages, Viruses, Mitochondrial Invertebrates, Mitochondrial Vertebrates, Prokaryotes, Invertebrates and Vertebrates the Phages had maximum values of PCR based on RLE while lowest was of mitochondrial Invertebrates. There wasn't any major difference among them.

LZ algorithm replaces strings of characters with single type of codes. It do not any analyse the incoming text. In this approach, the dictionary is simply a portion of the previously encoded sequence. Similarly Mean PCR values based on LZ was calculated the Viruses obtained highest value while mitochondrial Invertebrates was lowest and again no major difference was seen among categories.

Mean ratio values were calculated by using formula PCR_{RLE}/PCR_{LZ} which showed appropriate results that among 7 classes i.e., Phages, Viruses, Mitochondrial Invertebrates, Mitochondrial Vertebrates, Prokaryotes, Invertebrates and Vertebrates the lowest value was of Viruses whereas Vertebrates showed highest mean values. Here also not much difference was observed within other categories.

We also took Correlation coefficient values (r) between different attributes i.e., Genome size, GC% and Variation from GC%, Base Composition Values with PCR values of RLE and LZ.

This signifies that the Measure of Heterogeneity is proportional to CV and inversely proportional to Percentage Compression Ratio (PCR).

One of the reason that no big difference could be detected among sequence with very high coding content (viruses), high coding content (organelles and prokaryotic) and low coding content (higher Eukaryotes) could be attributed to failure of either method to compare dispersed repeats such as Transposons and Retrotransposons. As higher Eukaryotes consist of very high content of Transposons and Retrotransposons.

Sequence complexity might also depends on Base composition. The unequal distribution of the four bases may lead to lower complexity. Protein sequences composed of 20 amino acids is more complex than DNA sequence of same length consisting of 4 different bases. This may be extrapolated that if any one or more base are highly under or over represented then it will lead to lower complexity. For example GC or AT rich regions are less complex then sequences with comparable base composition of all four bases.

To investigate the above hypothesis , the base composition of all genomes was determined and its CV (measure of dispersion) was calculated to assess its unequal distribution. The results indicates that if CV is low the Heterogeneity measure will be high which leads to high compression and low complexity and thus the Percentage Compression Ratio value will be low. Hence, this signifies that CV is inversely proportional to Percentage Compression Ratio (PCR).

Further frequency of exhausting sets of di-nucleotides, tri-nucleotides, tetra-nucleotides, penta-nucleotides, hexa-nucleotides and hepta-nucleotides were also determined for each genome. Their CV's were also correlated. The PCR of LZ values showed strong correlation in comparison with the PCR of RLE values. As RLE is missing many repeat values which are assessed by LZ. No clear pattern could be obtained with either of the two compression tools examined in this study, owing to their inability to detect statistically very long repeat sequences. However a significant moderate (RLE) to strong (LZ) negative correlation between PCR and CV values were obtained. This shows that unequal distribution of bases strongly affects the genome complexity.

CHAPTER 9 CONCLUSION

Considering DNA sequences as string data, it may be compressed in lossless manner to relate its compressibility with its other properties such as genome complexity. There wasn't major difference among sequences related to content can be method to compress Dispersed repeats like Transposons. Both negative and positive correlation coefficient values were seen between different attributes. The LZ algorithm is more effective on Non coding regions as it contains more number of Dispersed repeats whereas RLE algorithm is more effective on Tandem repeats. To conclude the value of measure of Dispersion (CV) is inversely proportional to Percentage Compression Value (PCR). If the ratio of PCR_{RLE}/PCR_{LZ} value decreases then the Compression in RLE increases. This shows that PCR_{RLE}/PCR_{LZ} is inversely proportional to Compression in RLE. Moreover, the compression in LZ is directly proportional to PCR_{RLE}/PCR_{LZ}.

To get better picture of DNA sequence compression based assessment of genome complexity, more genomes should be analysed. Further a tool should be developed to handle compression of longer sequences also. With better tools and larger sample size, robust statistical analysis can make the final picture cleared.

REFRENCES

1. Primrose, S.B., & Twyman, R.(2013). *Principles of gene manipulation and genomics*. John Wiley & Sons.

2. Nalbantoglu, Ö. U., Russell, D.J.,& K. Sayood. Data Compression Concepts and Algorithms and Their Applications to Bioinformatics. Entropy 12 (1): 34-52 (2010).

3. Nelson, M., & Gailly, J.L. (1996). *The data compression book*(vol.2). New York: M&t Books.

4. Rajeswari, P.R., & Apparao, A. (2010). Genbit compress-algorithm for repetitive and non repetitive DNA sequences. *J Theor Appl Inf Technol,11(1*),25-29.

5. Emmert-Streib, F. (2010). Statistic Complexity: Combining Kolmogorov Complexity with an Ensemble Approach. *PLoS One*, 5(8), e12256.

6. Rajeswari, P.R., & Apparao, A. (2010). Genbit compress-algorithm for repetitive and non repetitive DNA sequences. *International Journal of Computer Science and Information Technology, 2, 25-29.*

7. Afify, H., Islam, M., & Wahed, M. A. (2011). DNA lossless differential compression algorithm based on similarity of genomic sequence database. *arXiv preprint arXiv*: 1109.0094.

8. Ziv J., Lempel A (1977). *A universal algorithm for sequential data compression* . IEEE Transactions on Information Theory, 23, 337-343.

9. Ziv J., Lempel A (1978). *Compression of individual sequence via variable length coding*. IEEE Transactions on lnformation Theory, 24, 530-536

10. Menconi, G., & Marangoni, R. (2006). A compression-based approach for coding sequences identification. I. Application to prokaryotic genomes. Journal of Computational Biology, 13(8), 1477-1488.

11. Rivals E, Dauchet M, Delahaye JP, Delgrange O. *Compression and genetic sequence analysis*.(1996). 315-322.

12. Cao, M. D., Dix, T. I., Allison, L., & Mears, C. (2007, March). *A simple statistical algorithm for biological sequence compression*. In Data Compression Conference, 2007. DCC'07 (pp. 4352). IEEE

13. Cox, A. J., Bauer, M. J., Jakobi, T., & Rosone, G. (2012). Large-scale compression of genomic sequence databases with the Burrows–Wheeler transform. Bioinformatics, 28(11), 1415-1419.

14. Elhai, J., *The Origin of Repeated Sequences in Genomes*. Center for the study of biological Complexity.

15. Hosseini, M., Pratas, D., & Pinho, AJ. (2016,October). *A Survey on Data Compression Methods for Biological Sequences*. Institute of Electronics and Informatics Engineering of Aveiro/Department of Electronics, Telecommunications and Informatics (IEETA/DETI), University of Aveiro, 3810-193 Aveiro, Portugal,2016.

16. Pratas, D., Pinho, AJ., & Paulo J. S. G. Ferreira. *Efficient Compression of Genomic Sequences*. IEETA - Institute of Electronics and Informatics Engineering of Aveiro DETI - Department of Electronics, Telecommunications and Informatics University of Aveiro, 3810-193 Aveiro, Portugal. In 2016 Data Compression Conference.

17. Rai, D.S., Bharti R.K., & Parihar B. *Survey of Compression of DNA Sequence.* International Journal of Computer Applications (0975 – 8887) Volume 73– No.6, July 2013.

18. Chen, X., Kwong, S., & Li, M. (2000, April). *A compression algorithm for DNA sequences and its applications in genome comparison*. In Proceedings of the fourth annual international conference on Computational molecular biology (p. 107). ACM.

19. Xin Chen, Ming Li, Bin Ma & Tromp J. (2002). *DNA Compress: fast and effective DNA sequence compression.* Bioinformatics, 18, 1696–1698.

20. Prokopenko M., Boschetti F., & Ryan A. (2009). *An information-theoretic primer on complexity, self-organization, and emergence*. Complexity 15: 11–28.

21. Grumbach S., Thai F. (1993). *Compression of DNA Sequences*. In Data Compression Conference Snowbird, Utah, USA. IEEE Computer Society Press.

22. Bell, T. C., Cleary, J. G., AND Witten, I. H. *Text Compression*. Prentice Hall, Upper Sadle River, NJ, 1990.

23. Sayood, K. *Introduction to Data Compression*. Academic Press, San Diego, CA, 1996, 2000.

24. Kuruppu, S., Puglisi, A. J. & Zobel, J. *Optimized relative Lempel-Ziv compression of genomes*. In: Proceedings of the ACSC Australasian Computer Science Conference (ed. Reynolds, M.). Australian Computer Society, Inc., Sydney, Australia, 91–98 (2011).

25. Salomon, D. & Motta, G. *Handbook of data compression*. Springer, London (2010).

26. Matsumoto, T., Sadakane, K., Imai, H. *Biological sequence compression algorithms*. Genome Inform. 2000, 11, 43–52.

27. Lanctot, J.K., Li, M., and Yang, E. *Estimating DNA sequence entropy*. Proceedings of the 11th Annual ACM-SIAM Symposium on Discrete Algorithms, 409-418, 2000.

28. Kodituwakku, S.R. & Amarasinghe U. S., *Comparison of lossless data compression algorithms for text data*. Indian Journal of Computer Science and Engineering Vol 1 No 4 416-425.

Compression

ORIGINALITY REPORT

%14	%11	%8	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	research.ijcaonline.org Internet Source	%2
2	www.coursehero.com Internet Source	%2
3	Liou, Cheng-Yuan, Shen-Han Tseng, Wei-Chen Cheng, and Huai-Ying Tsai. "Structural Complexity of DNA Sequence", Computational and Mathematical Methods in Medicine, 2013. Publication	%1
4	www.jsbi.org Internet Source	%1
5	en.wikipedia.org Internet Source	%1
6	www.bunnyclews.com Internet Source	%1
7	Frank Emmert-Streib. "Statistic Complexity: Combining Kolmogorov Complexity with an Ensemble Approach", PLoS ONE, 08/26/2010 Publication	%1