Effect of Base Composition on DNA Sequence Traits

DISSERTATION

Submitted in the partial fulfilment for the award of the degree of

MASTER OF SCIENCE

IN

BIOTECHNOLOGY

Submitted by

Kirti Pal

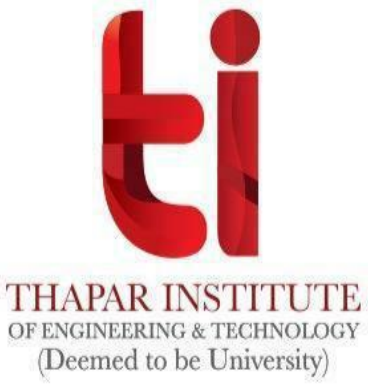
301601014

Under the Guidance of

Dr. Vikas Handa

Assistant Professor

Department of Biotechnology



THAPAR INSTITUE OF ENGINEERING AND TECHNOLOGY

PATIALA

JULY, 2018

CERTIFICATE

This is to certify that the thesis entitled **"Effect of Base Composition on DNA Sequence Traits"** submitted by **Ms. Kirti Pal** in partial fulfillment of requirement for the award of degree **Master in Science in Biotechnology,** Department of Biotechnology, Thapar University Patiala is the record of the candidate's own independent original work carried out by her. The report has not been submitted for the award of any other degree or certificates in this or any other university or institute.

Dr. Vikas Handa
Assistant Professor
Department of Biotechnology
Thapar University
Patiala, Punjab.

Kirti Pal
M.Sc. Biotechnology
(301601014)
Thapar University
Patiala, Punjab.

CANDIDATE'S DECLARATION

I hereby declare that the project work entitled "**EFFECT OF BASE COMPOSITION ON DNA SEQUENCE TRAITS**" in the partial fulfillment of the requirement for the awarded of the degree of master in science Biotechnology, Department of Biotechnology, and Thapar University, Patiala, is an authentic record of my work during the period of 6 month from Jan 2018 to 15 June 2018, under the guidance of **Dr. Vikas Handa,** Assistant professor, Thapar university, patiala. The matter embodies in thesis has not been submitted in any part or full to any other university or institute for the awarded any degree in India or abroad.

Kirti Pal

301601014

Date: 28-June-2018

ACKNOWLEDEGEMENT

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 has 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 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, **Mr. Gurparkash Singh Thind** and **Mr. Vardhan Chabra** for their support and help.

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

Kirti Pal

KIRTI PAL

Date: 28-June-2018

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVATIONS

A	Adenine
C	Cytosine
DNA	Deoxyribonucleic Acid
G	Guanine
NCBI	National Centre of Biotechnology Information
PCR	Percent Compression Ratio
T	Thymine

ABSTRACT

The eukaryotic DNA is highly complex depending upon the organisms. The genome complexity can be analyzed by re-association 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 Run Length Encoding (RLE) method 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- Genome complexity, Percentage compression ratio, Run length encoding, Base Composition

CHAPTER-1

INTRODUCTION

DNA is a polymer of deoxyribonucleotides. The length of DNA is defined as number of nucleotides present in it. A nucleotide has three components - a nitrogenous base, a pentose sugar (deoxyribose) and a phosphate group. There are two types of nitrogenous bases- Purines (Adenine and Guanine) and Pyrimidines (Cytosine and Thymine). Four bases of DNA i.e. A, T, G and C can be considered as four alphabets of DNA molecule. A nitrogenous base is linked to the pentose sugar through a N-glycosidic linkage to form a nucleoside. When a phosphate group is linked to 5‟-OH of a nucleoside through phosphoester linkage, a corresponding nucleotide is formed. Two nucleotides are linked through 3‟-5‟ phosphodiester linkage to form dinucleotide. The backbone in a polynucleotide chain is formed due to sugar and phosphate group. DNA as a genetic material carries information from cell to cell and from generation to generation.

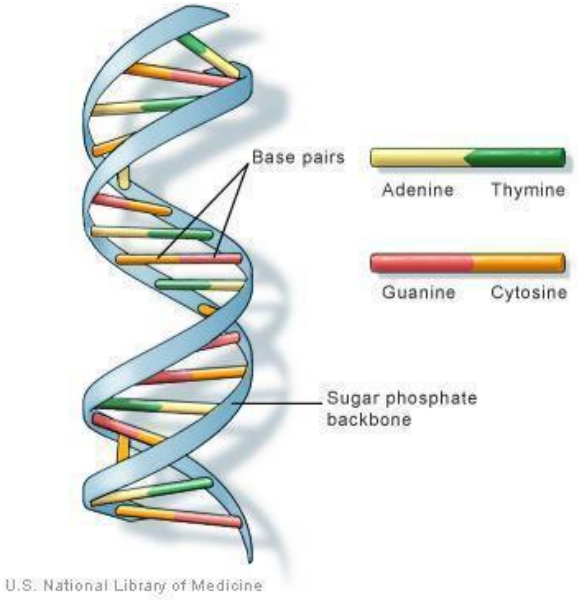


Fig1.1 DNA is a double helix formed by base pairs attached to a sugar-phosphate Backbone
(*Source (https://ghr.nlm.nih.gov/primer/basics/dna)*

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 (Fig.1.2).

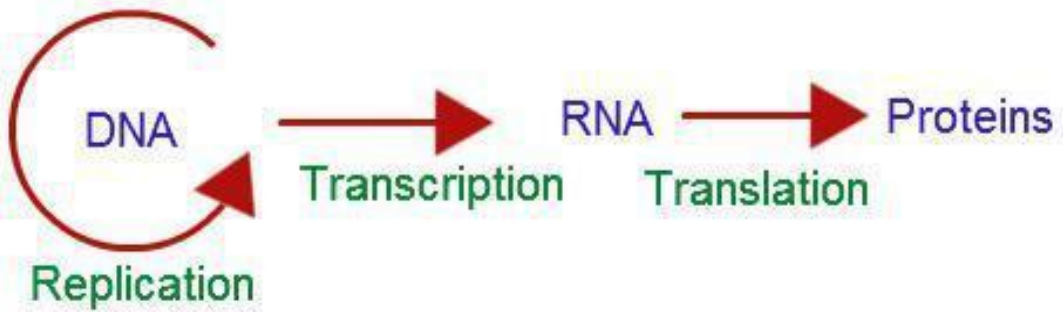


Fig1.2 Central dogma

Source (https://www.google.co.in/search?q=central+dogma)

The backbone of DNA is resistant to degradation, 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.

In higher eukaryotes, only ~ 2% of DNA is made up of protein-coding genes, the other 99% is non-coding. Non-coding DNA does not provide instructions for making proteins. Scientists once thought non-coding DNA was junk, with no known purpose. However, it is becoming clear that at least some of it is integral to the functions of cells, particularly the control of gene activity.

In prokaryotes, most of the DNA sequence is coding. Prokaryotic DNA rarely contains any intron or non-coding sequence. Lower eukaryotic DNA contains very few genes and the number of introns per gene is less. They contain introns with small length. In higher eukaryotes e.g. mammals have ~2% coding sequence. Most of the genes contain introns and the number of intron per gene is more. Length of the introns in higher eukaryotes is ~10 fold large than lower eukaryotes.

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 sequences: tandem repeats and dispersed repeats. 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 and Twyman, 2013).

Genome complexity can also be analysed using **reassociation kinetics**. An organism's DNA can be heated in solution until it melts, and then cooled to allow DNA strands to reassociate forming double-stranded DNA (Fig.1.3). This is typically done after shearing the DNA to form many 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.

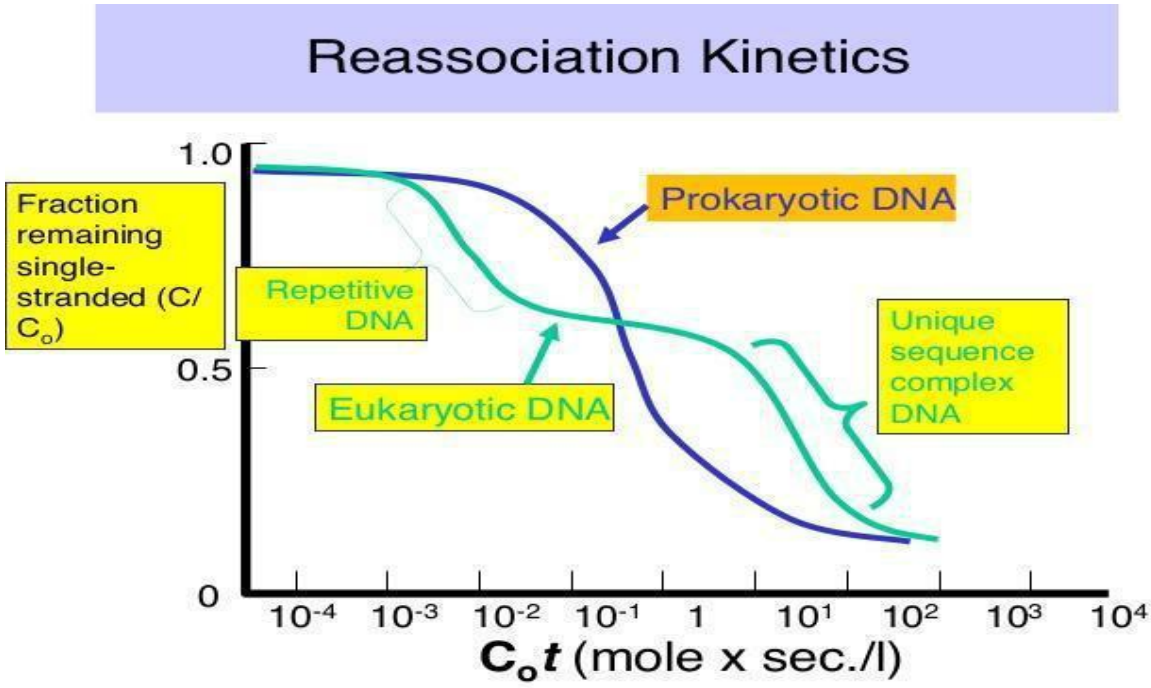


Fig. 1.3 Genome Complexity analysed by Reassociation Kinetics

Source (https://www.google.co.in/search?q=reassociation+kinetics+image)

Complexity can be measured in the form of compressibility of sequence known as Kolmogorov complexity. Kolmogorov complexity is related to algorithm information theory applied to strings of symbols. The normalized compression distance has been used as measure of Kolmogorov complexity *(Emmert- Streib, 2010)*.

Nature of DNA sequence is not random because DNA sequence contains tandem repeats, ORF, Transposons, Pseudo-genes and duplicate genes. Coding Region contains ORF and Non- Coding Region contains tandem repeats, dispersed repeats, transposes and Pseudo-genes. Therefore, the entire genome of the organism can be compressed using compression tools.

Compression can be used to find out the relatedness between different organisms. Therefore, the compression technique can be used to test the relatedness of DNA sequence of different organisms.

1.2 Data Compression

There are various types of compression algorithms which can be used to compress the different types of document. The compression algorithm used to compress the text document is called Data Compression. Data compression is of two types depending upon the type of data i.e. (I) Lossless compression and (ii) lossy compression. For DNA sequence compression Lossless compression technique can be used, because it is necessary that after decompression the decompressed data could be identical data.

1.2.1 Lossy Compression

It is a type of data compression where the data is compressed in such a way that after decompression the original data cannot be revived. Example of such type of data is compression of Image File (.jpg etc.). Image file once compressed cannot be regained its original size after decompression. The compression algorithm example is gzip compression algorithm.

1.2.2 Lossless Compression

Lossless compression is a type of data compression algorithms by which the original data can be perfectly reconstructed after decompression of the compressed data.

Lossless compression is used in cases where there is important that the original and the decompressed data be identical, or where deviations from the original data could be creating inaccuracy or error. Example of this type of document is text document.

Sequence→GGGGATATATGGCGGC

Compression→ [G4[AT3[GGC2

Decompression→GGGGATATATGGCGGC

Fig. 1.4 Example of lossless data compression

Examples of lossless data compression are: Run-Length Encoding, Huffman coding, Ziv and Lempel Algorithm and Burrow Wheeler transforms. Most simple algorithm for DNA sequence compression is Run -Length Encoding Algorithm.

1.2.3 Run Length Encoding

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 homo polymeric nucleotide repeat, di-nucleotide repeat, tri-nucleotide repeat and tandem repeats. It is not useful with data strings that don't have many runs as it could greatly increase the file size.

	Length
Sequence →GGGGATATATGGCGGC	= 16
Compression →[G4[AT3[GGC2	= 12

Fig. 1.5 Compression of DNA sequence by RLE Algorithm

Compression percentage of introns, exons, intergenic region, 3‟UTR, 5‟UTR and promoter region may vary according to the presence of repetitive DNA sequence. Therefore they all can be expressed in terms of compression percentage.

1.2.4 Base composition

Bases are not present in equal proportions in DNA sequence. Variation in base composition can be used to understand properties of DNA sequences. The ratio of

Guanine + cytosine (GC) to adenine + thymine (AT) nucleotides in a nucleic acid Base composition is usually expressed as a GC% value, for eg: 60% G+C. The GC content is of interest given those GC rich regions as associated with coding regions in genome.

CHAPTER-2

REVIEW OF LITREATURE

2.1 Related Work: Compression algorithms designed for purpose of compression Are many available but here some have been used like:-

Run length encoding - Run-length encoding (RLE) is a very simple form of lossless data compression in which runs of data are stored as a single data value and count, rather than as the original run.

It checks whether there are any repeating symbols or not, and is based on those redundancies and their lengths. Consecutive recurrent symbols are identified as runs and all the other sequences are considered as non-runs. For an example, the text "ABABBBBC" is considered as a source to compress, then the first 3 letters are considered as a non-run with length 3, and the next 4 letters are considered as a run with length 4 since there is a repetition of symbol B. The major task of this algorithm is to identify the runs of the source file, and to record the symbol and the length of each run. The Run Length Encoding algorithm uses those runs to compress the original source file while keeping all the non-runs without using for the compression process *(Kodituwakku et al. 2010).*

RLE can be used to compress the homo-polymeric nucleotide repeat, di-nucleotide repeat, tri-nucleotide repeats.

2.1.2: LZ Algorithm: For sequential data compression the first algorithm designed was LZ algorithm. It is based on dictionary coders and relies on exact repeats. It is the universal algorithm for text data compression. Therefore it can also be used to compress the DNA sequence. English alphabets from A-Z assigned Binary Codes such as A - 0001 (1), B – 00010 (2) and Z - 11010 (26). For DNA sequence binary codes are assigned as:

T- 10100

G- 111

C- 11

A-1

The DNA sequence is converted to binary numbers and then compressed (*Ziv and Lempel, 1997).*

2.1.3 LZ77: It is based on LZ algorithm. The compression is achieved 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 *(Sheng Bao et al., 2005).*

2.1.4 DNA Compress

Uncompress uses LZW compression scheme. The basic idea of LZW scheme is to replace repeat regions by references to a Dictionary. This algorithm is much faster than Decompress. Compression is done in two phases:

Find all approximate repeats including complementary palindromes.
Encode approximate repeat regions and non-repeat regions.

Each repeat region is check to see whether it reduces the size, otherwise that repeat is discarded (*Chen et al., 2002*).

2.1.5 Percentage Compression Ratio: PCR is the compression ratio is defined as the ratio between the uncompressed size and compressed size the compression amount applied raw sequence is expressed as ratio *(Kodituwakku et al. 2010).*

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

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 and Twyman, 2013*).

Sequence complexity:

Kolmogorov Complexity: Kolmogorov complexity is related to algorithm information theory applied to strings of symbols. The normalized compression distance has been used as measure of Kolmogorov complexity *(Emmert- Streib, 2010)*

Base composition: With a given DNA, base composition may vary and has effect on genetic information. One of the well studied example is variation in the GC content of DNA. It has been reported that majority of genes are found in GC rich regions which appear as light bands in chromosomes after Giemsa staining. Mammalian genomes have ~40% GC content however in different regions of genome it may locally vary from 38% to 55%. The variation of GC content at genome level is analyzed by studying isochores. H3 class of isochore which is has high GC content constitutes only 3% of genome but harbours >25% genes (*Gardiner, 1996).*

CpG islands are another important example of base composition variation found in higher eukaryotes. CpG islands are few to several hundred base pair long sequences often found in the 5' regions of housekeeping genes and are usually found to be unmethylated. These sequences are GC rich and have high CpG observed/expected ratio. CpG islands play important role in maintaining transcriptionally active state of the gene in the vicinity (*Gardiner-Garden and Frommer, 1987 & Takai and Jones, 2002*).

Base composition of genomes has also been studied related to Chargaff's second parity rule which states that in an individual DNA strand number of adenines are comparable to thymines and numbers of guanines are comparable to cytosines. It has been reported that difference in base composition at the dynamic part of genomes (single nucleotide polymorphic sites) indicates that mutational bias might be shaping the overall scenario of variation in base compositions (*Li et al., 2015*).

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 for efficiently compressing the large DNA sequence
- To measure compressibility of DNA sequence of different origin and correlate them with attribute based on base composition of the sequence

CHAPTER- 5

DATA SOURCE

Nucleotide sequence of various organisms was 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:

ORGANISM NAME	ACCESSION NO.	Sequence length
Bacteriophage lambda	NC_001416.1	1-48502
Haemophilis influenza	NC_000907.1	500001 to 700000
Homo sapiens chromosome 21	NC_000021.9	30000001-30200000
Homo sapiens mitochondrial DNA	NC_012920.1	1-16569
Escherichia coli K12	NC_000913.3	1-200000
Saccharomyces cerevisiae S288C chromosome IV	NC_001136.10	100001-300000
Drosophila melanogaster chromosome 2L	NT_033779.5	10000001-10200000
Saccharomyces cerevisiae S288C chromosome IV	NC_001136.10	500001 to 700000
Danio rerio strain Tuebingen chromosome 25	NC_007136.7	3850001 to 4050000
Xenopustropicalis strain Nigerian chromosome 1 Xenopus_tropicalis_v9.1	NC_030677.1	1650001 to 1850000
Oryzasativa Japonica Group cultivar Nipponbare chromosome 12, IRGSP-1.0	NC_029267.1	1000001 to 1200000
Arabidopsis thaliana chromosome 2 sequence	NC_003071.7	1000001 to 1200000
Caenorhabditis elegans chromosome 1	NC_003279.8	1000001 to 1200000
Gallus gallus chromosome 1	NC_006088.5	2000001 to 2200000
Ostreococcuslucimarinus CCE9901 chromosome 1	NC_009355.1	500001 to 700000

Table 5.1 Accession Number and Sequence Length of Genome sequence of 15 Different organisms.

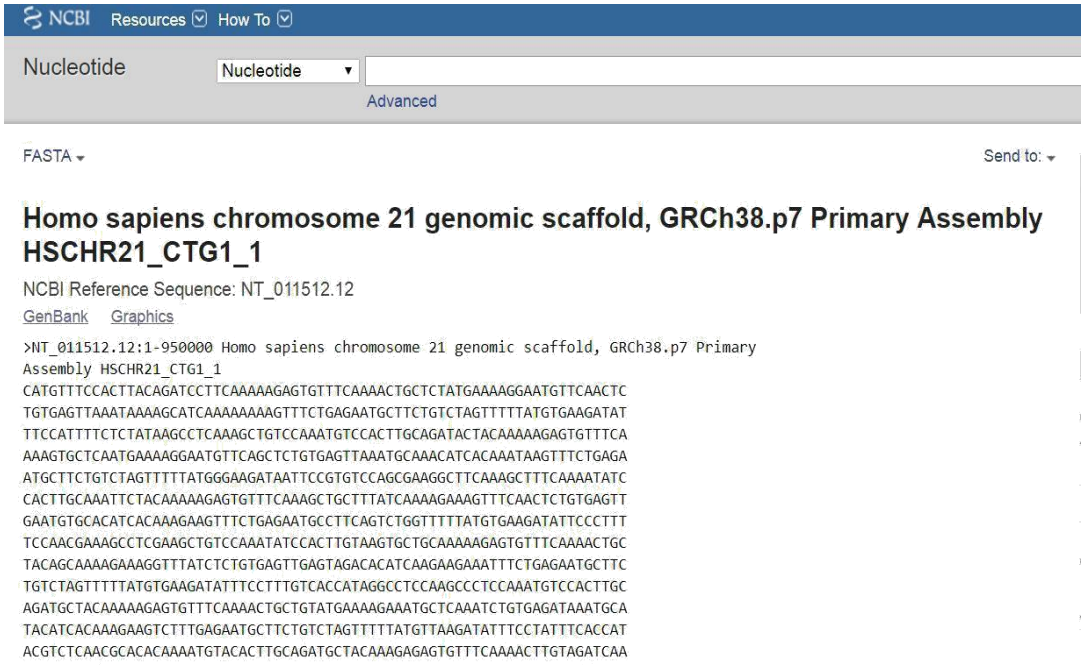


Fig5.1:- Screenshot of FASTA sequence of *Homo sapiens chromosome 21.*

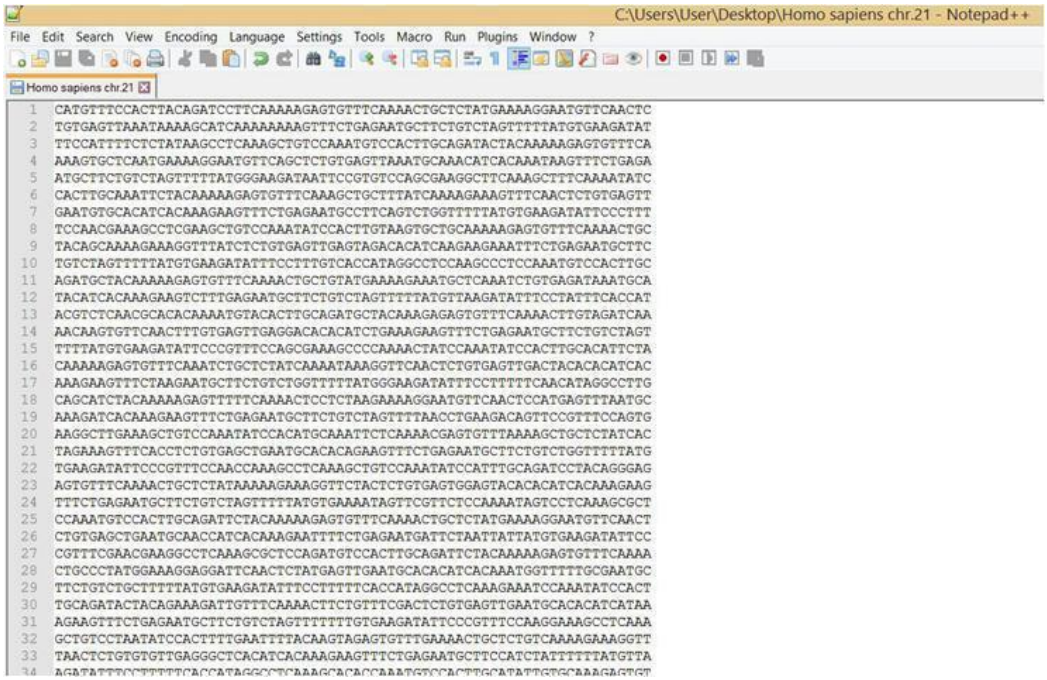


Fig5.2:-Screenshot of *Homo sapiens chromosome 21* sequence in notepad ++.

6.1Sequence Analysis Tools:

6.1.1Microsoft Excel: Microsoft Excel feature calculation graphing tools and macro programming language called visual basic for application. Microsoft excel was mainly used for forming graphs and analyzing 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: Full form of FCGR is Chaos Game Representation of frequencies as described by Deschavanne *et al.* (1999) represents frequencies of oligo-nucleotides. It was also used to investigate the relation between base composition of genome and its compressibility.

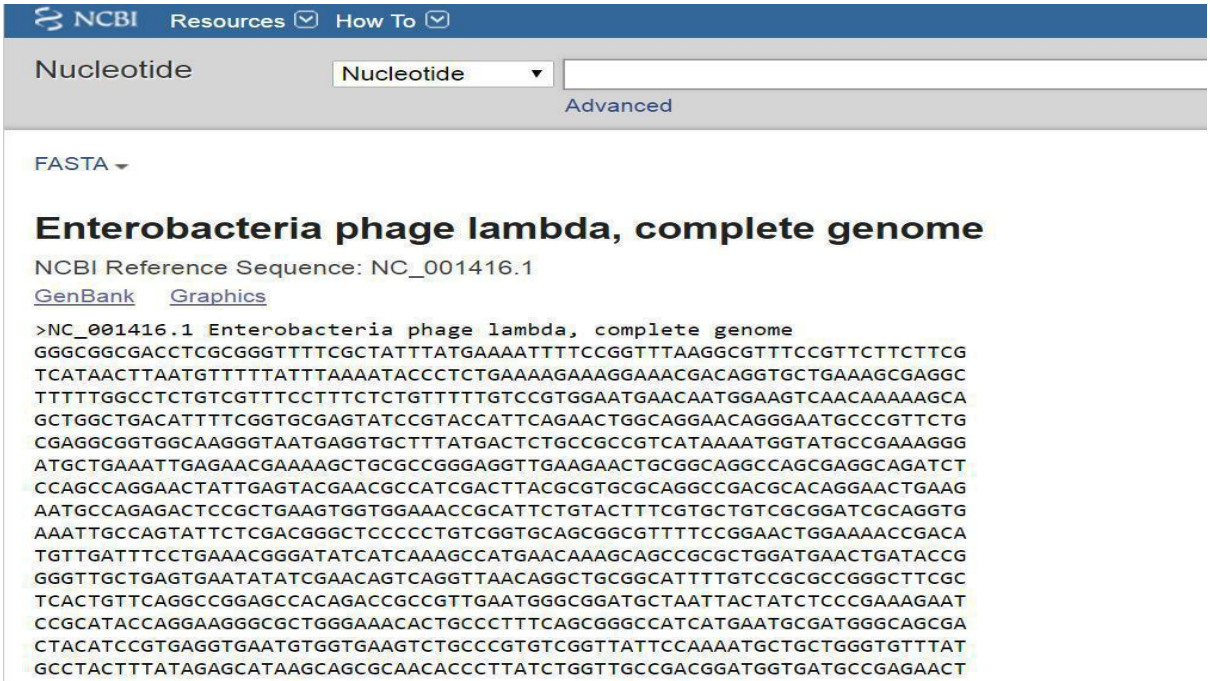


Fig6.1:- Sequence of *Bacteriophage lambda* in FASTA format.

6.2 Methods: Nucleotide sequences of 15 different genome sequences were downloaded from the Nucleotide database of NCBI website (www.ncbi.nlm.nih.gov/) in FASTA format. For example, *Bacteriophage lambda* sequence (accession number NC_001416.1, 48502 bp long). The length of the genome sequence was picked randomly was downloaded from NCBI and compressed by RLE algorithm using above mentioned macros recorded in Notepad++. Similarly other sequences were opened in Notepad++ and pre recorded macro was run to remove all spaces. Then the sequence as one continuous string was compressed (based on RLE algorithm) using the pre recorded macros. The lambda phage sequence was also decompressed by running another macro to check if the algorithm was perfectly lossless or not. Entire sequence of 48502 nucleotides was obtained with any change in the sequence when it was compressed and subsequently decompressed.

Run Length Encoding mainly used for compression of nucleotide sequence. It was used to compress homo-polymeric nucleotide repeats, di-nucleotide repeats and tri-nucleotide repeats. Lower limit for compression of DNA sequence for homo polymeric nucleotide repeats di-nucleotide repeats and tri-nucleotide repeats is 4, 3 and 2 respectively.

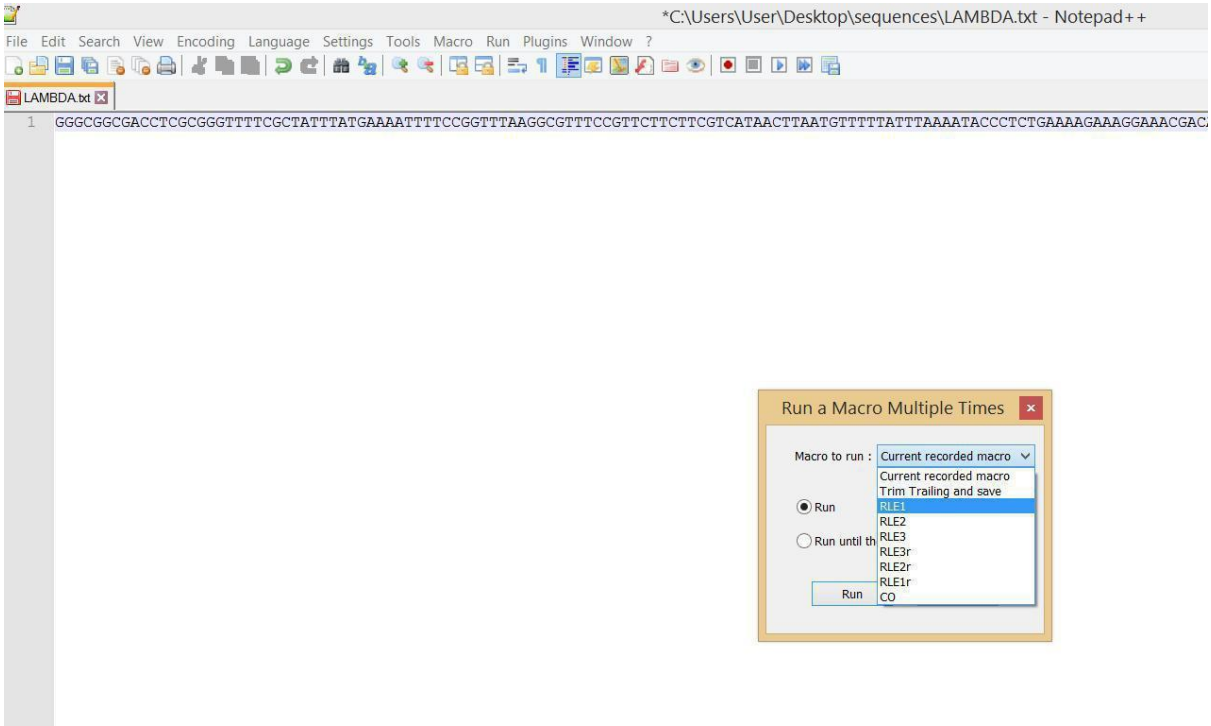


Fig6.2:-Screenshot of running macros for Compression and decompression

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 (*based on work of Gurpreet Kaur, Master's thesis, 2016)*

To set the upper limit of repeat units, the probabilities of occurring homo-polymer 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 repeat:

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

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

Apply natural log on both sides:

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

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

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

$$\mathbf{nA= (\ln 1/10^7) / (\ln ¼)}$$

$$\mathbf{nA=11.62}$$

$$\mathbf{nA\sim 12}$$

For di-nucleotide repeat:

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

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

Apply natural log on both sides:

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

nAA = [ln (1/N)]/ ln P(AA)

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

nAA= (ln 1/10^7) / (ln 1/16)

nAA=5.81

nAA~ 6

For tri-nucleotide repeat:

[P (AAA)] ^{n} × N=1

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

Apply natural log on both sides:

n ln P (AAA) = ln (1/N)

nAAA = [ln (1/N)]/ ln P(AAA)

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

nAAA= (ln 1/10^7) / (ln 1/64)

nAAA=3.87

nAAA~ 4

Therefore the upper limit to compress the DNA sequence for homo polymeric nucleotide repeat, di-nucleotide repeat and tri-nucleotide repeat is 12, 6 and 4 respectively.

Homo polymeric nucleotide repeat, di-nucleotide repeat and tri-nucleotide repeat considered for compression were:

Table 6.1 Nucleotide Repeats to be considered for compression

Homo polymeric repeats
As
Gs
Ts
Cs

Di-nucleotide repeats
GA or AG
GT or TG
GC or CG
Di-nucleotide repeats
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 homo-polymeric nucleotide repeat as N, for di nucleotide repeat as NN and for tri-nucleotide repeat as NNN.

Repeats were replaced in the following manner

Table 6.2 Replacement of nucleotide repeats using RLE

S. No.	Repeats	Replacement
1. Homo polymeric nucleotide repeat (N)		
(i)	AAAAAAAAAAAA	12A]
(ii)	AAAAAAAAAA	11A]
(iii)	AAAAAAAAA	10A]
(iv)	AAAAAAAA	9A]
(v)	AAAAAAA	8A]
(vi)	AAAAAA	7A]
(vii)	AAAAA	6A]
(viii)	AAAA	5A]
(ix)	AAA	4A]
(i)	GGGGGGGGGGG	12G]
(ii)	GGGGGGGGGG	11G]
(iii)	GGGGGGGGG	10G]
(iv)	GGGGGGGG	9G]
(v)	GGGGGGG	8G]
(vi)	GGGGGG	7G]
(vii)	GGGGG	6G]
(viii)	GGGG	5G]
(ix)	GGG	4G]
(i)	CCCCCCCCCCC	12C]
(ii)	CCCCCCCCCC	11C]
(iii)	CCCCCCCCC	10C]
(iv)	CCCCCCCC	9C]
(v)	CCCCCCC	8C]
(vi)	CCCCCC	7C]
(vii)	CCCCC	6C]

(viii)	CCCCC	5C]
(ix)	CCCC	4C]
(i)	TTTTTTTTTTT	12T]
(ii)	TTTTTTTTTT	11T]
(iii)	TTTTTTTTT	10T]
(iv)	TTTTTTTT	9T]
(v)	TTTTTTT	8T]
(vi)	TTTTTT	7T]
(vii)	TTTTT	6T]
(viii)	TTTT	5T]
(ix)	TTT	4T]
2. For di-nucleotide repeats (NN)		
(i)	GAGAGAGAGAGA	6GA]
(ii)	GAGAGAGAGA	5GA]
(iii)	GAGAGAGA	4GA]
(iv)	GAGAGA	3GA]
(i)	ATATATATATAT	6AT]
(ii)	ATATATATAT	5AT]
(iii)	ATATATAT	4AT]
(iv)	ATATAT	3AT]
(i)	ACACACACACAC	6AC]
(ii)	ACACACACAC	5AC]
(iii)	ACACACAC	4AC]
(iv)	ACACAC	3AC]
(i)	GTGTGTGTGTGT	6GT]
(ii)	GTGTGTGTGT	5GT]
(iii)	GTGTGTGT	4GT]
(iv)	GTGTGT	3GT]

(i)	GCGCGCGCGCGC	6GC]
(ii)	GCGCGCGCGC	5GC]
(iii)	GCGCGCGC	4GC]
(iv)	GCGCGC	3GC]
(i)	TCTCTCTCTCTC	6TC]
(ii)	TCTCTCTCTC	5TC]
(iii)	TCTCTCTC	4TC]
(iv)	TCTCTC	3TC]
3. For tri - nucleotide repeats (NNN)		
(i)	GGAGGAGGAGGA	4GGA]
(ii)	GGAGGAGGA	3GGA]
(iii)	GGAGGA	2GGA]
(i)	GGCGGCGGCGGC	4GGC]
(ii)	GGCGGCGGC	3GGC]
(iii)	GGCGGC	2GGC]
(i)	GGTGGTGGTGGT	4GGT]
(ii)	GGTGGTGGT	3GGT]
(iii)	GGTGGT	2GGT]
(i)	GAAGAAGAAGAA	4GAA]
(ii)	GAAGAAGAA	3GAA]
(iii)	GAAGAA	2GAA]
(i)	GACGACGACGAC	4GAC]
(ii)	GACGACGAC	3GAC]
(iii)	GACGAC	2GAC]
(i)	GATGATGATGAT	4GAT]

(ii)	GATGATGAT	3GAT]
(iii)	GATGAT	2GAT]
(i)	GCAGCAGCAGCA	4GCA]
(ii)	GCAGCAGCA	3GCA]
(iii)	GCAGCA	2GCA]
(i)	GCCGCCGCCGCC	4GCC]
(ii)	GCCGCCGCC	3GCC]
(iii)	GCCGCC	2GCC]
(i)	GCTGCTGCTGCT	4GCT]
(ii)	GCTGCTGCT	3GCT]
(iii)	GCTGCT	2GCT]
(i)	GTAGTAGTAGTA	4GTA]
(ii)	GTAGTAGTA	3GTA]
(iii)	GTAGTA	2GTA]
(i)	GTCGTCGTCGTC	4GTC]
(ii)	GTCGTCGTC	3GTC]
(iii)	GTCGTC	2GTC]
(i)	GTTGTTGTTGTT	4GTT]
(ii)	GTTGTTGTT	3GTT]
(iii)	GTTGTT	2GTT]
(i)	AACAACAACAAC	4AAC]
(ii)	AACAACAAC	3AAC]
(iii)	AACAAC	2AAC]
(i)	AATAATAATAAT	4AAT]

(ii)	AATAATAAT	3AAT]
(iii)	AATAAT	2AAT]
(i)	ACCACCACCACC	4ACC]
(ii)	ACCACCACC	3ACC]
(iii)	ACCACC	2ACC]
(i)	ACTACTACTACT	4ACT]
(ii)	ACTACTACT	3ACT]
(iii)	ACTACT	2ACT]
(i)	ATCATCATCATC	4ATC]
(ii)	ATCATCATC	3ATC]
(iii)	ATCATC	2ATC]
(i)	ATTATTATTATT	4ATT]
(ii)	ATTATTATT	3ATT]
(iii)	ATTATT	2ATT]
(i)	CCTCCTCCTCCT	4CCT]
(ii)	CCTCCTCCT	3CCT]
(iii)	CCTCCT	2CCT]
(i)	CTTCTTCTTCTT	4CTT]
(ii)	CTTCTTCTT	3CTT]
(iii)	CTTCTT	2CTT]

Compression scheme can be done in the following steps:

Table 6.3 Compression Scheme

S. No.	Steps
1	N+NN+NNN
2	N+NNN+NN
3	NN+NNN+N
4	NN+N+NNN
5	NNN+NN+N
6	NNN+N+NN

The Scheme with the Lowest PCR was selected for further analysis. Lowest PCR means the highest compressibility. We got highest compression with NN+NNN+N i.e. di nucleotide repeat compression followed by tri-nucleotide repeat compression followed by homo polymeric nucleotide repeat compression. All the 6 above mentioned permutations of compression algorithms were used to compress the sequences. Thus 6 different compressed lengths were obtained for each sequence. The lowest compressed length was selected using MS Excel function “Min” to calculate the PCR values.

CHAPTER -7

RESULT

Genome contains genetic material i.e., DNA or RNA. It includes coding and non-coding DNA. 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 cases 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++.15 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 homo polymeric nucleotide repeats, di-nucleotide repeats and tri nucleotide repeats and saved. Similarly decompression algorithm was also developed using this method to check RLE compression.

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

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

Fig7.1: Screenshot of RLE algorithms.

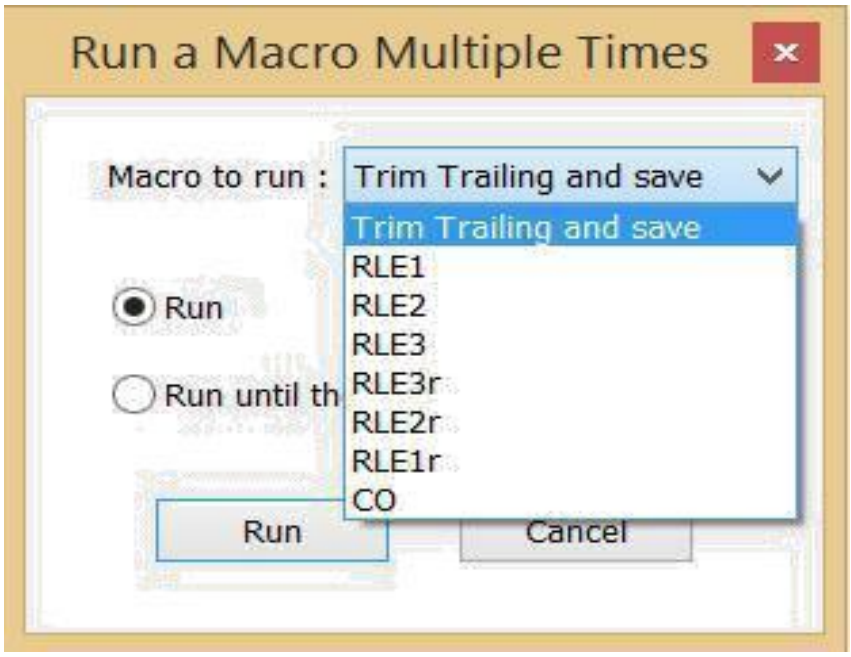


Table7.1: PCR values of 15 sequences from genomes of different organisms

S.No	SEQUENCE	SEQUENCE LENGTH	PCR VALUES	$(G+A+T+C)^{0.5}$
1	*Bacteriophage lambda*	48502	96.97	220.23
2	*Haemophilus influenzae* .	500001 to 700000	95.2135	447.21
3	*Homo sapiens* chromosome 21	30000001-30200000	95.2135	447.21
4	*Homo sapience Mitrocondrial DNA*	16569	95.9382	128.72
5	*Escherichia coli* K12	1-200000	97.4195	447.21
6	*Saccharomyces cerevisiae S288C chromosome IV*	100001-300000	95.794	447.21
7	*Drosophila melanogaster chromosome 2L*	10000001-10200000	95.262	447.21

8	*Saccharomyces cerevisiae S288C chromosome IV*	500001 to 700000	95.4745	447.21
9	*Daniorerio strain Tuebingen chromosome 25, GRCz11*	3850001 to 4050000	93.6785	447.21
10	*Xenopustropicalis strain Nigerian chromosome 1 Xenopus_tropicalis_v9.1*	1650001 to 1850000	95.128	447.21
11	*Oryzasativa Japonica Group cultivar Nipponbare chromosome 12, IRGSP-1.0*	1000001 to 1200000	95.4215	447.21
12	*Arabidopsis thaliana chromosome 2 sequence*	1000001 to 1200000	94.3535	447.21
13	*Caenorhabditis elegans chromosome I*	1000001 to 1200000	94.098	447.21
14	*Gallus gallus breed Red Jungle Fowl isolate RJF #256*	2000001 to 2200000	96.1015	447.21
15	*Ostreococcuslucimarinus CCE9901 chromosome 1*	500001 to 700000	94.404	447.21

DNA sequence compression may be related to the complexity of genomic information. To investigate if DNA sequence compressibility is related to base composition of DNA sequence, the measure of compression PCR was related to attributes related to base composition based attributes. The genomic sequences for which RLE based PCR was calculated, were analysed for calculating heterogeneity of sequence. Counts of the four bases and frequencies of all the possible di-nucleotides, and higher motifs up to hepta nucleotides were determined by FCGR tool for each sequence. Coefficient of variation was calculated for bases, di-nucleotides and the motifs for each sequence. A strong negative correlation was obtained for bases and all the oligo-nucleotide motifs however the strongest correlation was -0.757 between the PCR and CV values of di-nucleotides.

Another measure of distribution of bases in DNA sequences was based on the difference between counts of bases. It can be measured in three different manners based on difference between different bases:

a) G - A & T - C
b) G - T & A - C
c) G - C & A - T

If we consider a 2 dimensional random walk based on a DNA sequence where one step is moved in up, down, right or left direction depending on different bases, the final displacement can be measured as

$$((G-A)^2 + (T-C)^2)^{0.5}$$

For a situation a random walk displacement was calculated for set of base counts of each sequence based on *a*, *b* and *c* scenario. The random walk displacement was normalized by dividing it with √N where N is the total length of the sequence. Correlation coefficients were calculated between PCR and random walk displacement of all the sequences. A strong negative correlation was observed in *b* scenario which is in agreement with the correlation between PCR and CV values, however a weak positive correlation of r = 0.459 was observed in scenario *c*. This may be attributed to Chargaff's second parity rule which states that A% ~ T% and G% ~ C% in the same strand of DNA.

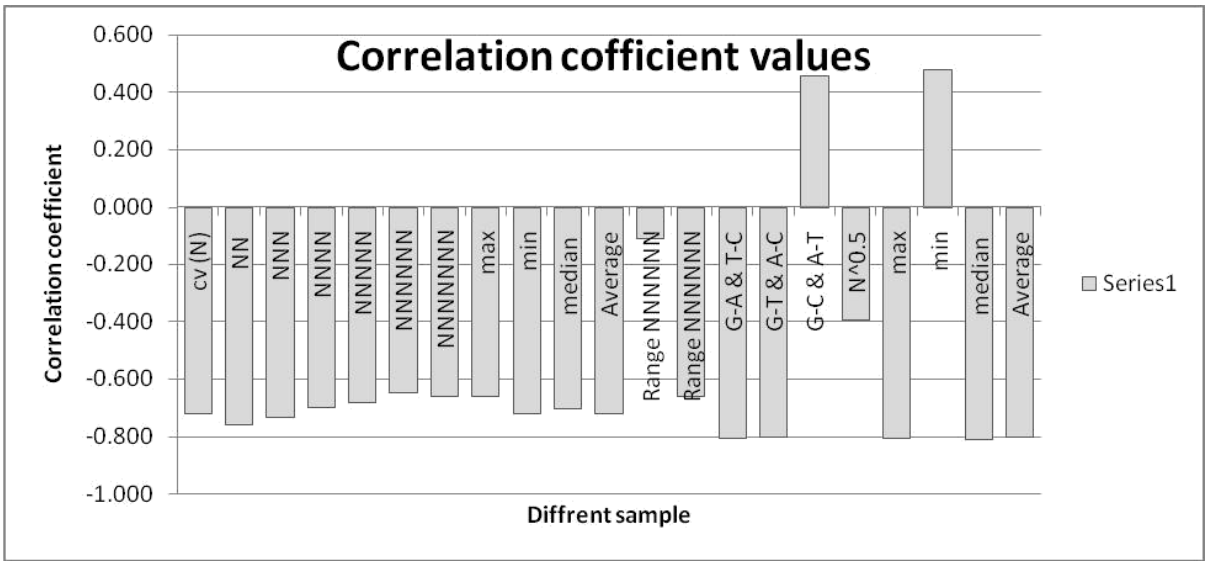


Fig 7.2: Graph showing correlation coefficient values

The correlation values were determined between PCR based on RLE 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 maybe concluded that tetra-nucleotide is the optimal word size for assessing sequence heterogeneity.

Organism name	bacteriophage lambda	Hemeophilus influnzae	Homosapince mitrocondrial	Homo sapiens chromosome 21,	Escherichia coli str. K-12 substr. MG1655	Saccharomyces cerevisiae S288C chromosome IV	Drosophila melanogaster chromosome 2L	Saccharomyces cerevisiae S288C chromosome IV	Danio rerio strain Tuebingen chromosome 25, GRCz11	Xenopus tropicalis strain Nigerian chromosome 1, Xenopus_tropicalis_v9.1,	Oryza sativa Japonica Group cultivar Nipponbare chromosome 12, IRGSP-1.0	Arabidopsis thaliana chromosome 2 sequence	Caenorhabditis elegans chromosome I	Gallus gallus breed Red Jungle Fowl isolate RJF #256	Ostreococcus lucimarinus CCE9901 chromosome 1,
Compressed sequence	47030	190427	15896	190427	194839	191588	190524	190949	187357	190256	190843	188707	188196	192203	188808
Uncompressed sequence	48502	200000	16569	200000	200000	200000	200000	200000	200000	200000	200000	200000	200000	200000	200000
PCR	96.97	95.21	95.94	95.21	97.42	95.79	95.26	95.47	93.68	95.13	95.42	94.35	94.10	96.10	94.40
Cofficient of variation (N)	4.38	22.57	29.42	26.24	5.21	23.15	14.52	25.03	27.71	16.99	12.46	32.53	28.99	16.98	19.63
NN	15.28	38.11	44.50	41.65	17.74	35.56	28.12	38.23	42.18	30.34	20.76	49.22	57.72	33.28	44.83
NNN	26.72	50.31	57.65	55.32	29.50	45.58	39.32	49.18	57.31	42.54	28.00	63.72	88.39	45.67	63.83
NNNN	37.41	61.67	71.65	67.93	40.43	55.30	49.46	59.83	71.86	53.53	37.42	77.73	121.46	56.79	84.53
NNNNN	48.62	72.85	87.09	80.85	51.22	64.73	59.77	70.24	89.13	65.76	47.27	92.08	157.20	67.67	108.32
NNNNNN	56.82	83.51	71.05	94.33	62.92	57.29	71.60	81.53	112.55	82.92	58.80	107.99	196.32	121.25	136.17
NNNNNNN	74.49	97.22	101.63	111.53	77.39	70.90	88.53	96.43	151.58	115.18	75.28	128.05	242.99	147.39	174.10
Maximum value	74.49	97.22	101.63	111.53	77.39	70.90	88.53	96.43	151.58	115.18	75.28	128.05	242.99	147.39	174.10
Minimum value	4.38	22.57	29.42	26.24	5.21	23.15	14.52	25.03	27.71	16.99	12.46	32.53	28.99	16.98	19.63
Median	37.41	61.67	71.05	67.93	40.43	55.30	49.46	59.83	71.86	53.53	37.42	77.73	121.46	56.79	84.53
Average	37.67	60.89	66.14	68.26	40.63	50.36	50.19	60.07	78.90	58.18	40.00	78.76	127.58	69.86	90.20
Range NNNNNN	0.96	93.01	0.36	25.87	6.14	5.12	11.61	13.09	19.37	10.55	9.38	16.14	24.01	58.66	0.46
Range NNNNNNN	6.08	9.96	8.41	31.34	10.90	9.09	29.74	28.84	67.42	20.48	20.97	34.57	102.48	62.42	81.19
G-A & T-C	3.59	71.39	24.46	82.94	15.66	73.21	45.92	79.15	87.61	53.62	39.37	102.86	91.68	53.45	62.08
G-T & A-C	5.82	71.38	14.96	82.97	13.84	73.21	45.80	79.13	87.49	53.73	39.41	102.86	91.67	52.04	62.09
G-C & A-T	6.81	1.22	24.74	3.44	10.26	0.35	3.58	2.30	4.62	3.35	2.23	1.58	1.81	14.09	1.07
N^0.5	220.23	447.21	128.72	447.21	447.21	447.21	447.21	447.21	447.21	447.21	447.21	447.21	447.21	447.21	447.21
Maximum value	6.81	71.39	24.74	82.97	15.66	73.21	45.92	79.15	87.61	53.73	39.41	102.86	91.68	53.45	62.09
Minimum value	3.59	1.22	14.96	3.44	10.26	0.35	3.58	2.30	4.62	3.35	2.23	1.58	1.81	14.09	1.07
Median	5.82	71.38	24.46	82.94	13.84	73.21	45.80	79.13	87.49	53.62	39.37	102.86	91.67	52.04	62.08
Average	5.40	48.00	21.39	56.45	13.25	48.92	31.77	53.52	59.91	36.90	27.00	69.10	61.72	39.86	41.74

Table 7.3: table showing CV Values

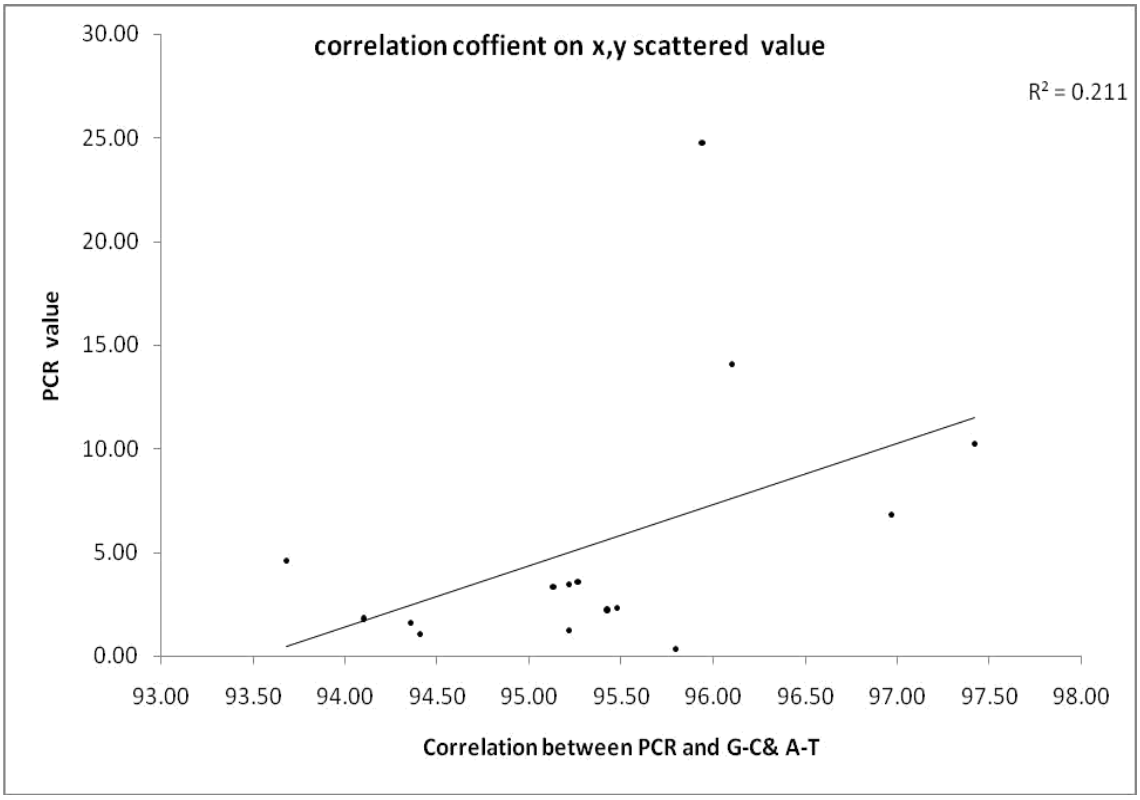


Fig7.3 Graph is showing correlation between PCR Values and G-C & A-T values

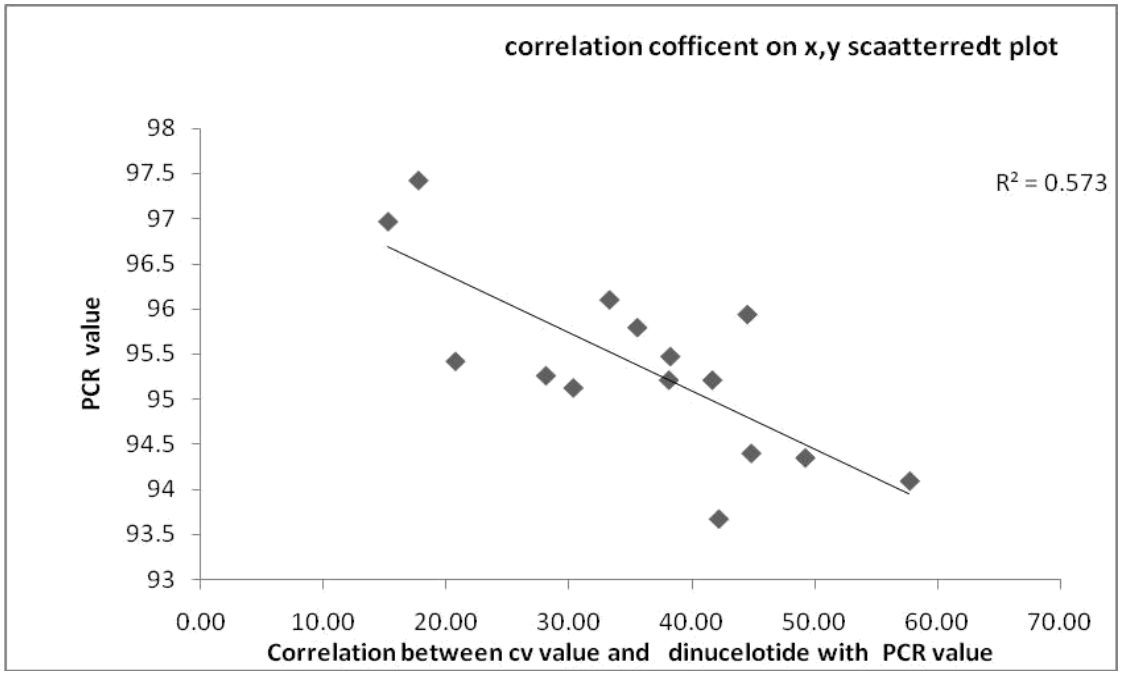


Fig 7.4 Graph is showing correlation between CV of di nucleotide with PCR value

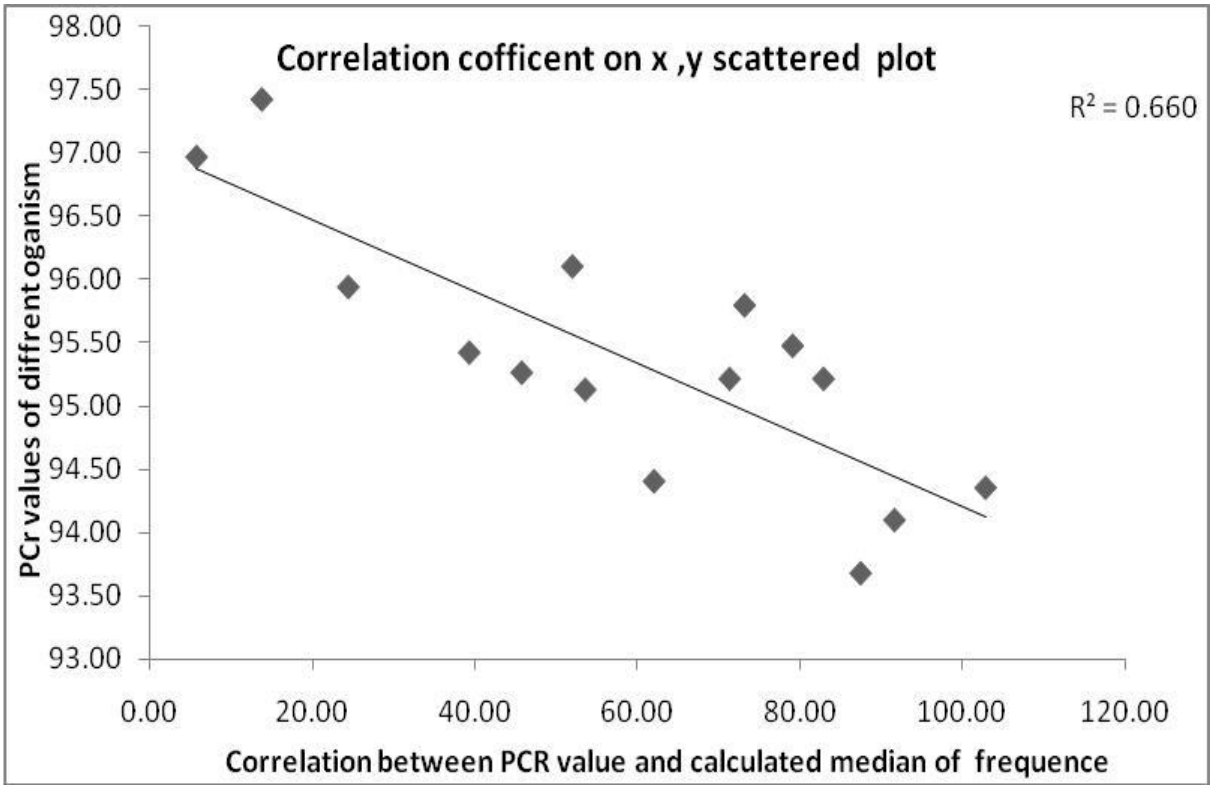


Fig 7.5 Graph is showing correlation between PCR values and median

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 15 different genomic sequences containing Phages, Viruses, Mitochondrial DNA, Chloroplast DNA, Prokaryotic DNA, and Eukaryotic DNA. There are two types of repetitive DNA: First is Tandem repeats these sequences are present in at least 10^5 copies per mammalian 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.

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

Sequence complexity might also depend 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 indicate that if CV is low the Heterogeneity measure will be low which leads to low compression. It is shown by negative correlation between PCR and CV. Higher the compression, lower is the PCR. Hence, this signifies that CV inversely proportional to Percentage Compression Ratio (PCR) can be used as a measure of DNA sequence complexity. So higher the CV value (heterogeneity) lower is the sequence complexity (a higher measure of compression). In the present study the maximum effect was found with CV of dinucleotides. This is also evident from the chaos game representation of frequencies of all the dinucleotides (Fig.7.4,7.5,7.6)

A similar relationship was found between compression and another measure of sequence heterogeneity at single nucleotide level. A strong negative correlation was found between median values of the three possible random walk displacements and PCR. In fact a stronger negative correlation value ($r = -0.813$) was obtained for the random walk displacement which again shows that heterogeneity of the DNA sequence can be best measured as random walk displacement and has inverse relationship with the DNA sequence complexity.

This work is based on compression of DNA sequence based on RLE which take into consideration on the contribution of tandem repeats. More accurate picture may be obtained is similar work is done with DNA compression measurements based on LZ algorithm which includes dispersed repeats into data compression.

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. To conclude the value of measure of Dispersion (CV) is inversely proportional to Percentage Compression Value (PCR).

The present study shows that DNA sequence heterogeneity can be measured as CV of di-nucleotide frequencies as well as random walk displacement and it can give a fair idea about the complexity of the genomic sequences without compressing them.

Chapater-10

References

- Bose, T., Mohammed, M. H., Dutta, A., &Mande, S. S. (2012). BIND–An algorithm for loss-less compression of nucleotide sequence data. *Journal of biosciences*, *37*(4), 785-789.

- Cao, M. D., Dix, T. I., Allison, L., & Mears, C. (2007) A simple statistical algorithm for biological sequence compression. In *Data Compression Conference, 2007. DCC'07* (pp. 43-52). IEEE.

- Chen, X., Kwong, S., & Li, M. (2000) 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.
- 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.

- Emmert-Streib, F. (2010). Statistic complexity: combining Kolmogorov complexity with an ensemble approach. *PLoS One*, *5*(8), e12256.

- Kodituwakku, S. R., & Amarasinghe, U. S. (2010). Comparison of lossless data compression algorithms for text data. *Indian journal of computer science and engineering*, *1*(4), 416-425. Kaur, M., Garg, E. U., & Sabo, T. (2015). A Review of Various Data Compression Techniques to form a New Technique for Text Data Compression. *Reason*, *1*(5).

- Li, P., Wang, S., Kim, J., Xiong, H., Ohno-Machado, L., & Jiang, X. (2013). DNA-COMPACT: DNA COM pressionBased on a P attern-A ware C on textual Modeling Technique. *PloS one*, *8*(11), e80377.

- 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.

- Mohammed, M. H., Dutta, A., Bose, T., Chadaram, S., &Mande, S. S. (2012). DELIMINATE—a fast and efficient method for loss-less compression of genomic sequences Sequence analysis. *Bioinformatics*, *28*(19), 2527-2529.

- Rajarajeswari, P., &Apparao, A. (2011). Dnabit compress–genome compression algorithm. *Bioinformation*, *5*(8), 350.

- Rajeswari, P. R., &Apparao, A. (2010). GenBit compress—algorithm for repetitive and non-repetitive DNA sequences. *J TheorApplInf Technol*,*11*(1), 25-29.

- Nalbantoglu, Ö. U., Russell, D. J., & Sayood, K. (2009). Data compression concepts and algorithms and their applications to bioinformatics. *Entropy*, *12*(1), 34-52.

- Takai, D., & Jones, P. A. (2002). Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proceedings of the national academy of sciences*, *99*(6), 3740-3745.

- Gardiner-Garden, M., & Frommer, M. (1987). CpG islands in vertebrate genomes. *Journal of molecular biology*, *196*(2), 261-282.

- Li, X., Scanlon, M. J., & Yu, J. (2015). Evolutionary patterns of DNA base composition and correlation to polymorphisms in DNA repair systems. *Nucleic acids research*, *43*(7), 3614-3625.

- Gardiner, K. (1996). Base composition and gene distribution: critical patterns in mammalian genome organization. *Trends in Genetics*, *12*(12), 519-524.