Studies on Lossless Data Compression-Based Analysis of Different Types of DNA Sequence.

A dissertation

Submitted in partial fulfilment of the requirement

For the award of degree of

Masters of Technology

In

Biotechnology

(June,2022)

Under the guidance of

Dr. Vikas Handa

Submitted by

Hardik Bhan

Roll No. : 602004008

DEPARTMENT OF BIOTECHNOLOGY

THAPAR INSTITUTE OF ENGINEERING AND TECHNOLOGY,

PATIALA

ACKNOWLEDGEMENT

I am indebted to Dr. Prakash Gopalan, Director, Thapar Institute of Engineering and Technology (TIET) for providing the opportunity and facilities of institute to give me a chance to carry out dissertation work as part of M.Tech. curricular requirement. With great reverence, I express my warmest feeling with deep sense of gratitude to Dr. Vikas Handa, Assistant Professor, TIET, who agreed to take upon and guided for this dissertation and training. I have no word to express my heartfelt thanks to him for his illuminating guidance, unfailing encouragement, supervision and keen interest during the course of dissertation.

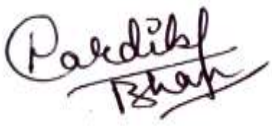
I would like to express my heartfelt respect to HOD, Dr. M. S. Reddy, Department of Biotechnology, Thapar Institute of Engineering and Technology, Patiala for his kind suggestions and foresightedness. At the end, thanks to my family and friends for giving me strength to keep going.

Name: Hardik Bhan

Signature:

DECLARATION

I hereby declare that the project work entitled (**"Studies on Lossless Data Compression-Based Analysis of Different Types of DNA Sequence"**) is an authentic record of my own work carried out at, **Thapar Institute of Engineering and Technology** as requirements of 12 months dissertation for the award of degree of Masters of Technology in Biotechnology, under the guidance of Dr. Vikas Handa, during June, 2021 to, June, 2022.



(Signature of student)

Hardik Bhan

602004008

Date: 30-06-2022

Certified that the above statement made by the student is correct to the best of our knowledge and the matter embodied in this dissertation has not been submitted to award of any Degree or certificate in any other university/institute.

Dr. Vikas Handa

(Assistant Professor)

CONTENTS

List of Figures

List of Tables

List of Abbreviations

Abbreviations	Extension
DNA	Deoxyribonucleic Acid
A	Adenine
T	Thymine
C	Cytosine
G	Guanine
tRNA	transfer ribonucleic acid
rRNA	ribosomal ribonucleic acid
mRNA	messenger ribonucleic acid
UTR	Un-translated Region
LZ77	Lempel Ziv 77
LZ78	Lempel Ziv 78
NCBI	National Center for Biotechnology Information
ANOVA	Analysis of Variance
WW Runs Test	Wald–Wolfowitz Runs Test
RS	Randomized Sequence

ABSTRACT

Genomic sequence data are produced in enormous quantities by modern biology. This is increasing the demand for effective sequence compression and analysis techniques. Data communication and storage are frequently thought to benefit from data compression and the related information theory-derived approaches. In computational biology research, data compression and related information-theoretic methods are frequently employed. The current work is based in the direction of understanding the complexity and compressibility of different types of genome sequences. As the sequences become more and more complex, the compressibility reduces.

Keywords: DNA, NCBI, WW Runs Test, ANOVA, tRNA, mRNA.

Chapter 1

INTRODUCTION

All genetic information encoded in DNA is found in an organism's genome. As a result, sequencing the genome, which dictates how organisms survive, develop, and multiply, is critical. Due to tremendous efforts in DNA sequencing over the last three decades, the whole genome sequence of many creatures, including humans, is now known, and genomic databases are growing exponentially with time. While specific genomes, like viruses and bacteria, are exceedingly small, some genomes, like some plants, can be nearly incomprehensibly enormous. Why there does not seem to be a consistent relationship between biological complexity and genome size is still somewhat perplexing.

Genome information is stored in DNA in the form of a base sequence. The exception is only RNA viruses. DNA is composed of four bases adenine (A), cytosine(C), guanine(G), and thymine(T) which are covalently bonded to the phosphate backbone.

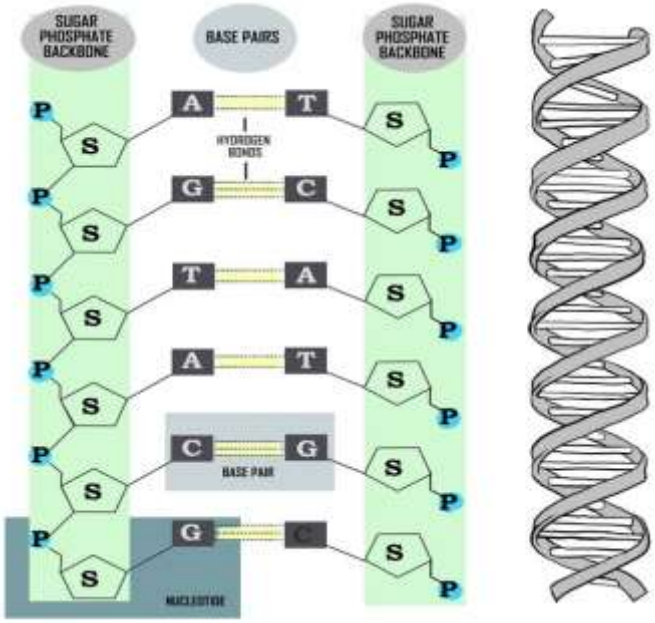


Fig. 1: Detailed Structure of DNA

(https://www.bing.com/images/search?q=dna+structure&qpvt=dna+structure&form=IGRE&first=1&tsc=ImageHoverTitle)

The structure of genes varies greatly among biological organisms. Bacterial genomes, for example, are virtually entirely made up of genes, but genes in higher eukaryotes can be little islands in a sea of non-coding DNA. As one progresses up the evolutionary tree, even the genes might become more complex architecturally. The genomes of bacteria, viruses, and organelles, on the one hand, and the nuclear genomes of eukaryotes, on the other, differ significantly at the whole-genome level. There are significant variances in the sorts of sequences observed, the quantity of DNA, and the number of chromosomes among eukaryotes [1].

Organisms	Size of genome (bp)
Bacteria	**10^6 to 10^7**
Plant	**10^7 to 10^{11}**
Insect	**10^8 to 10^9**
Mammal	**10^9 to 10^{10}**

Table 1: Variation in genome size of various organisms.

The Central Dogma of Molecular Biology states that biological information is transferred from DNA-to-DNA during replication. DNA information can be copied to RNA (mRNA) called transcription, and then proteins can be synthesized using this information in RNA as a template referred to as translation. Protein synthesis represents the basic biological process by which the cells build their specific proteins. These proteins are informational macromolecules. Proteins perform a vast array of functions within the living cell [1].

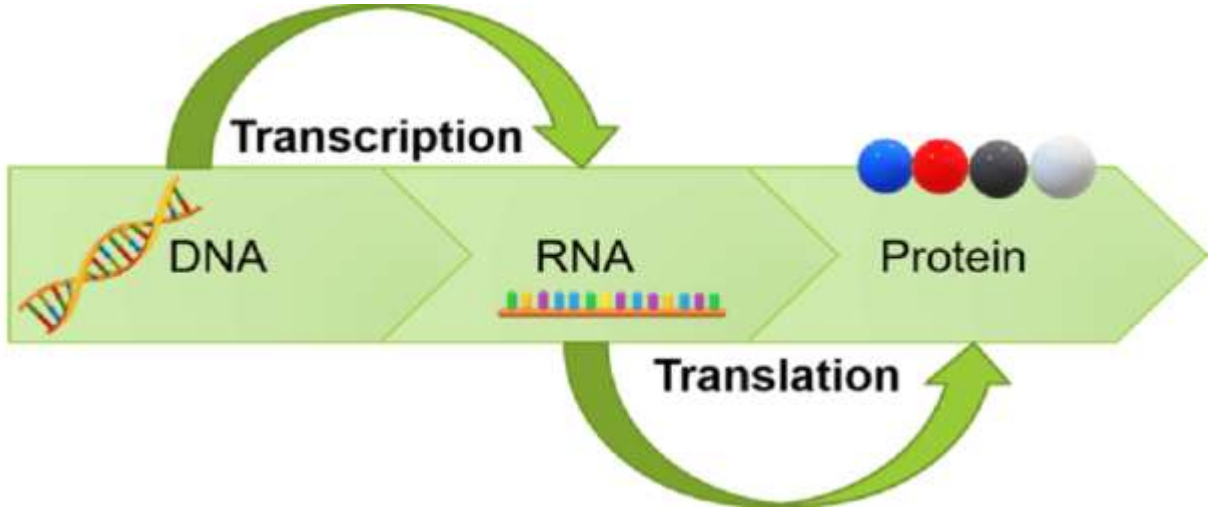


Fig.2: Central Dogma of Molecular Biology.

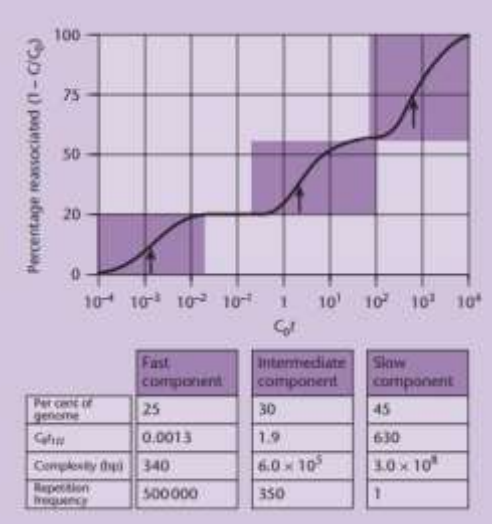
(https://biologydictionary.net/central-dogma/)

Genome complexity is based mainly on three components: non-repetitive, moderately repetitive and highly repetitive DNA. Tandem repeats are notable repeats in the genome, which makes them intriguing. They differ in the repeated sequence and the number of times it occurs in a particular genome. Many tandem repetitions can be found in DNA sequences, notably in higher eukaryotes and regions that produce non-coding RNA molecules like tRNA and rRNA. Dispersed repeats, such as transposons, are another form of the sequence found in the genome. Multiple copies of the same gene may exist in the genome. As DNA sequences are expected to be non-random, it is possible to remove redundancy, resulting in compression [2].

The non-repetitive DNA component typically increases as we move up the evolutionary tree. The presence of a lot of repetitive DNAs is indicated by the fact that many plants and animals have a substantially higher C-value. Most mRNA that hybridizes to DNA anneals to non-repetitive DNA, demonstrating that non-repetitive DNA contains most genes. In some genes, the coding sequence is interrupted by non-coding (untranslated)

sequences known as introns. Such genes are known as split genes, and the parts of these translated genes are known as exons. Split genes are rare in prokaryotes, although they are commoner in archaebacteria than eubacteria. Split genes are much commoner in eukaryotes, but the number of such genes, and the number and size of introns per gene, increase with genome complexity [2].

When dealing with complete genomes, one has to deal with millions or billions of base pairs. As a result, new and more effective strategies for compressing biological sequence data, particularly DNA sequences, are required. One of the ways to learn about DNA sequence complexity is to assess its compressibility.



	Fast component	Intermediate component	Slow component
Per cent of genome	25	30	45
$C_0t_{1/2}$	0.0013	1.9	630
Complexity (bp)	340	6.0×10^5	3.0×10^8
Repetition frequency	500000	350	1

Fig.3: The reassociation kinetics of a eukaryotic DNA sample showing genome complexity.

(Principles of Gene Manipulation, 6th Edition, 2001)

There are several challenges in compressing genomic sequences, and as genomic data is extensively used in the healthcare and medical industry, biological and technical implications will be required. Data compression and related information-theoretic techniques are widely used in computational biology research. The data structure, modelling, and speed are the main advantages of data compression in biological research.

Understanding genome sequences has many uses, from drug synthesis to gene screening and engineering. Knowing the structure of a line is essential to understanding it. Suppose a set of sequences has common structural properties shared by another sequence. In this case, they may be related in some way. Alternatively, the knowledge that applies to one can be helpful to the other. Compression helps to represent the information present in the genome [3].

Chapter 2

REVIEW OF LITERATURE

The composition of eukaryotic DNA is very complex and includes introns, exons, 3'UTRs, 5'UTRs, promoters, and intergenic sequences. The occurrence of lengthy DNA segments that don't seem to have any significance is one of the human genome's most striking features. This is what biologists refer to as non-coding DNA, and it comprises the majority of the genomic DNA of higher eukaryotes. These genes are sparse in the human genome and other higher eukaryotes. Most of them are interrupted by sequences, the introns, which are non-coding; i.e., they do not carry information for protein synthesis. During transcription, introns are therefore removed from the messenger RNA (mRNA), assembled only from the expressed parts of the gene, the exons. In the human genome, introns are ten times longer than exons and thus constitute the majority of the gene. Prokaryotes (such as bacteria) instead have a very compact genome without introns [4].

Coding sequences are unique and do not contain repetitive sequences, while non-coding sequences primarily contain repetitive DNA sequences. Therefore, exons' compressibility is expected to be less compared to other non-coding sequences. The compression ratios were calculated to analyze the compression ratios of the various sequences. The low compression ratio of coding DNA compared to non-coding DNA is just the beginning to show the application of information theory in genomics.

Over the last 20 years, genomic sequence compression can be divided into two categories. A specific compression algorithm was developed to efficiently compress sequence data to reduce resource consumption and study the usefulness of compressibility as a measure of information content for making inferences about sequences. Currently, for general use, only Lossless compression methods such as gzip and bzip2 are performed [5].

DNACompress, GenCompress, BioCompress etc., are some algorithms developed using DNA sequences' characteristics. It gains approximately a 22% compression ratio. Almost all algorithms only use the presence of only four bases, A, T, G, and C and the repetitive nature of DNA sequence. General compression algorithms exist based on Context Tree Weighting

(CTW) and Arithmetic Coding. Also, algorithms like LZ77, and LZ78 are used for this purpose, but they are not much efficient as far as DNA compression is concerned [6].

Data compression is essentially the process in which data is encoded in a smaller number of bits that is used by unencrypted data.
Data compression is a contentious topic in computational biology, bioinformatics and computer science. By 'compressing data', we specifically mean deriving techniques or, designing efficient algorithms to:

- represent data in a less redundant fashion
- remove the redundancy in data
- implement coding, including both encoding and decoding
- improve storage efficiency

Compression and Decompression goes hand-in-hand in any compression algorithm due to the nature of data compression. Decompression is equally considered as important as the compression because the compressed data needs to be restored back to standard state of usage.

2.1 Lossless compression vs Lossy Compression

In lossless compression, the original data is obtained without any loss on decompression. Our study and experiments have focused on the lossless compression of the DNA sequences as we cannot afford to lose any data after compression needed for further analysis and observations.

In lossy compression, the original data is lost during the compression process, i.e., retrieval of original data is impossible.

Various parameters can measure the performance of a compression algorithm. It depends on what is our priority concern. In our thesis, we have considered the easiest way to measure the effect of compression, i.e., compression ratio. The aim is to measure the impact of compression by the shrinkage of the size of the source in comparison with the size of the compressed version.

Compression Ratio: This is simply the ratio of size "after compression" to size "before compression."

$$\text{Compression ratio} = \frac{\text{size.after.compression}}{\text{size.before.compression}}$$

2.1.1 General Compression Algorithm

Many different compression algorithms were proposed by various scientists and researchers. The majority of lossless compression programmes employ two different types of algorithms: one that creates a statistical model for the input data and another that uses this model to map the input data to bit strings so that "probable" (frequently occurring) data will result in a shorter output than "improbable" data..

Statistical modeling algorithms for text include:

- Burrows-Wheeler transform
- LZ77
- LZW

Encoding algorithms to produce bit sequences are:

- Huffman coding
- Arithmetic coding

Block-sorting compression, often known as the **Burrows-Wheeler Transform (BWT)**, is an algorithm used in data compression methods like bzip2.. It was invented by Michael Burrows and David Wheeler.

Arithmetic coding is a type of lossless data compression method. It is a type of entropy encoding, however unlike other entropy encoding methods, which break up the input message into its individual symbols and then replace each one with a code word, this method simply reduces the message to a single number, a fraction n, where ($0.0 \leq n < 1.0$). Any of the several lossless data compression methods that function by looking for matches between the text to be compressed and a collection of strings included in a data structure (referred to as the "dictionary") maintained by the encoder are known as dictionary coders. When the encoder discovers a match of this sort, it replaces the reference with a pointer to the string's

location in the data structure. This idea underlies the LZ77 and LZ78 algorithms [6].

2.1.2 Related Existing Compression Algorithms

Tahi and Grumbach introduced two general groups for DNA sequence compression: intra-sequence similarities ‘or’ horizontal method and inter-sequence similarities ‘or’ vertical method. Tahi and Grumbach introduced the **BioCompress** algorithm. It was based on horizontal method. It searched for exact and reverse complement repetitions in DNA strings and then encoded them based on their position, previous repeat and repeat length. It encoded some unrepeated substrings by 2-bit method [7].

Chen et al. presented the **GenCompress** algorithm in which approximate repetitions were used for substrings [12].

Among statistical methods, **CDNA** was first in the field which was introduced by Loewenstren and Yianilos.

Run Length Encoding is a classic example of lossless compression. A run-length algorithm assigns codewords to consecutive recurrent symbols (called runs) instead of coding individual symbols. Run Length encoding follows a straightforward logic; it just picks the next unique character and appends the character and its count of subsequent occurrences in the encoded string. The idea is to reduce the total physical size of repeating characters in the data [15].

Example: Consider a string of DNA sequence AAAAAAATTTTTCCCCCGGGG (21 characters). This sequence can be compressed as $\mathbf{A_7T_5C_5G_4}$ (8 characters).

AIM & OBJECTIVE

In this thesis our AIM was to learn about the studies on lossless data compression-based analysis of different type of DNA sequences. For this study following objectives were performed:

1. To determine the compressibility of various types of biological sequences and to infer their complexity.

2. To compare complexity of intron and exon sequences using various tools related to information theory.

Chapter 3

MATERIALS AND METHODS

3.1 Data Extraction

The data used in our work consists of different types of genomic sequences which includes;

- Lambda genome (NC_001416.1)
- *A. thaliana* Chloroplast DNA (NC_000932.1)
- Hepatitis B virus isolate G376-A6 (AF384372.1)
- *Homo sapiens* Mitochondrial DNA (NC_012920.1)
- *Saccharomyces cerevisiae* S288C chromosome I (NC_001133.9)

All of the above-mentioned genomic sequences were obtained from the NCBI site (www.ncbi.nlm.nih.gov) [23].

22 Randomly selected different Human genes:

A1BG	**CBR3**	**HCST**	**LYAR**	**TNF**
ADM2	**CDK1**	**HKDC1**	**NOL7**	**UBL3**
APOE	**EMP2**	**IL6**	**OXTR**	
APOF	**F5**	**ING3**	**PKIA**	
BMP2	**GADD45A**	**INS**	**SPZ1**	

Table 2: Human genes

Ensembl database was used to download the exonic and intronic sequences of a sample set of randomly selected 22 human genes (https://asia.ensembl.org/index.html [24].

3.2 Experimental Design

- Three open online sources were used to convert text data to lossless data which included: gzip, bzip2 & deflate.

  Free online text compression tools - gzip, bzip2 and deflate
  https://www.txtwizard.net/compression

- For the lossless compression of exons and introns, the below mentioned open online source was used.

  Text to Deflate Compress using gzip, deflate and Brotli algorithms
  https://www.multiutil.com/

- Shuffle DNA (bioinformatics.org) - Shuffle DNA randomly shuffles a DNA sequence. It was used to randomize the DNA sequences i.e., its re-shuffles the DNA sequence without altering the base composition.
  https://www.bioinformatics.org/sms2/shuffle_dna.html

3.3 Compression of Extracted Genomic Sequences

The collected genomic sequences along with exon and intron sequences of 22 randomly selected human genes as mentioned in materials, were downloaded in FASTA format through NCBI and Ensembl respectively.

Lossless compression was done of the above-mentioned genomes. The randomization of sequences was achieved through Shuffle DNA. It was used to re-shuffle the DNA sequence without altering the base composition. These randomized sequences were used as controls.

The original and their respective shuffled sequences were converted into a single continuous string data with the help of “Macro recording” in **Notepad++**.For example if an exon of a gene is pasted as 12 strings of 60 bp each, all of them should be joined together to make it one single continuous string. It was required for lossless compression analysis.

The biological sequences and randomized sequences were then compressed through lossless compression method – gz, bzip2 and deflate and Brotli. The respective compressed sequences were also converted into single string data with the same procedure as mentioned in above step.

After lossless compression, compression ratios were calculated and analysed to determine the compressibility of various types of biological sequences and to optimize the compression tool for compression of various biological sequences.

3.4 Determining the randomness of exon & intron sequences by WW Runs Test

All the exon and intron sequences of 22 randomly selected human genes were subjected to WW Runs test in which the length of the sequences was calculated through LEN function in excel. For all the sequences, the consecutive repetitive bases (G,A,T & C) occurring in a sequence were replaced by their corresponding base, for example; if we consider a particular sequence of exon or intron – GGGGAAATTCCCCGAAGGAATTCC. This sequence can be re-written as GATCGAGATC. This is done with the help of Replace & Find tool of excel. Then lengths of these retrieved sequences were also calculated by the LEN function tool. Runs of the bases i.e., G,A,T&C were calculated for all the sequences. Further various parameters like mean value, standard deviation value and finally the Z-value were calculated. All the data was stored in excel file.

Chapter 4

RESULTS

4.1 Compression of genomes of diverse evolutionary lineages

In this experiment, we collected the genomic sequences of Lambda genome, *A. thaliana* Chloroplast DNA, Hepatitis B virus isolate G376-A6, *Homo sapiens* Mitochondrial DNA and *Saccharomyces cerevisiae* S288C chromosome I in FASTA format through NCBI. The sequences were subjected to lossless compression.

Each of the genomic sequences were re-shuffled 5 times to undergo randomization process. Each of the original and their respective shuffled sequences were converted into a single continuous string data with the help of “Macro recording” in Notepad++. Each of the original genomic sequences as well as their respective randomized sequences(RS) were then compressed through lossless compression method – gz, bzip2 and deflate.

The respective compressed sequences were also converted into single string data with the same procedure as mentioned in above step. After lossless compression, compression ratios were calculated and analysed to determine the compressibility of various types of biological sequences.

By performing the lossless compression for the above genomic sequences, it can be clearly observed that highest amount of compression is achieved in case of **bzip2 compression**. Not much effect is seen in the compressions of respective randomized DNA sequences (obtained through re-shuffling) of their individual original counterparts.

Biological sequences have lower compression ratio than the randomized sequences with same base composition. It may be inferred that biological sequences have some majorable hidden orders.

λ DNA	Compression Ratios		
	gz compression	bzip2 compression	deflate compression
Original Sequence	0.294	0.272	0.294
RS1	0.300	0.275	0.300
RS2	0.300	0.274	0.300
RS3	0.300	0.275	0.300
RS4	0.299	0.275	0.299
RS5	0.300	0.275	0.300

Table 3: Lambda genome compression data.

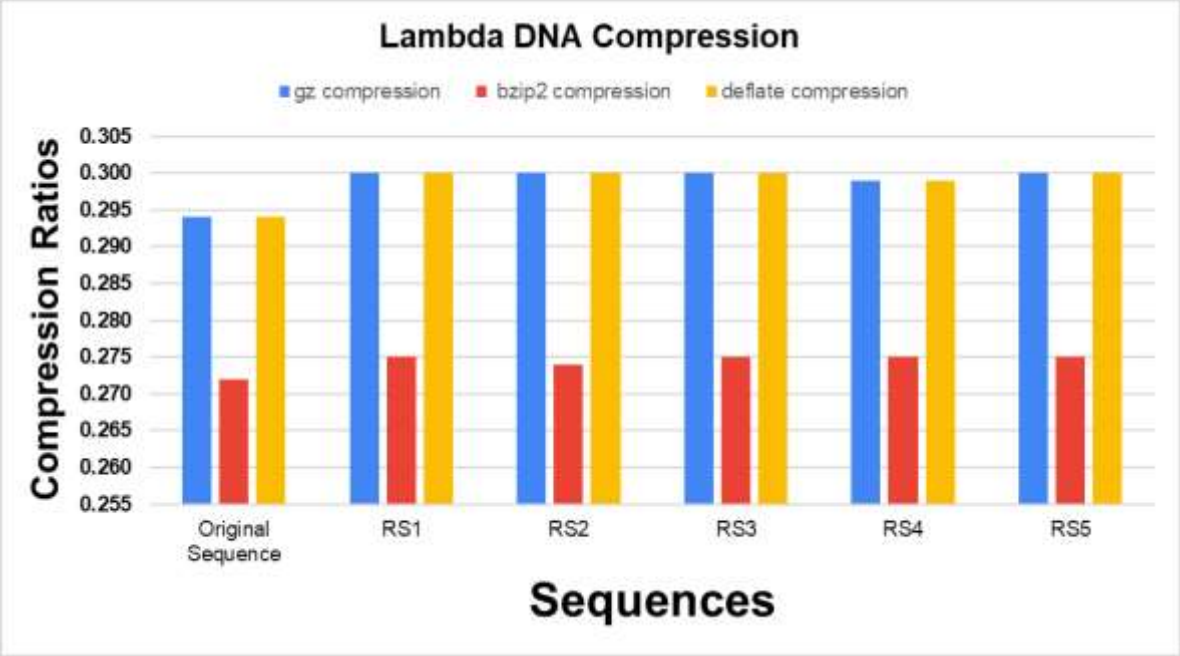


Fig. 4: Lambda Genome Lossless Compression

A. thaliana Chl. DNA	Compression Ratios		
	gz compression	bzip2 compression	deflate compression
Original Sequence	0.291	0.272	0.291
RS1	0.295	0.273	0.294
RS2	0.295	0.271	0.295
RS3	0.294	0.272	0.294
RS4	0.295	0.272	0.295
RS5	0.295	0.272	0.295

Table 4: *A.thaliana* Chloroplast DNA genome compression data.

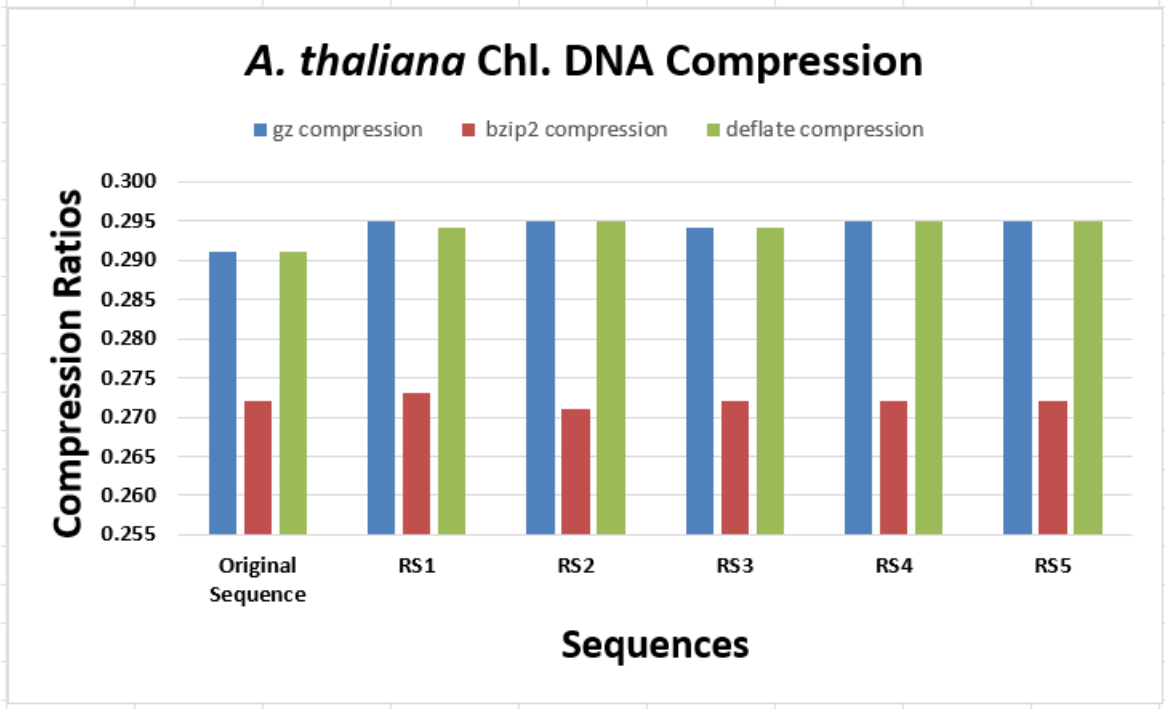


Fig. 5: *A. thaliana* Chloroplast DNA Lossless Compression

Hepatitis B virus	Compression Ratios		
	gz compression	bzip2 compression	deflate compression
Original Sequence	0.330	0.296	0.326
RS1	0.334	0.294	0.330
RS2	0.334	0.298	0.330
RS3	0.338	0.298	0.334
RS4	0.334	0.299	0.330
RS5	0.336	0.296	0.332

Table 5: Hepatitis B virus isolate G376-A6 genome compression data.

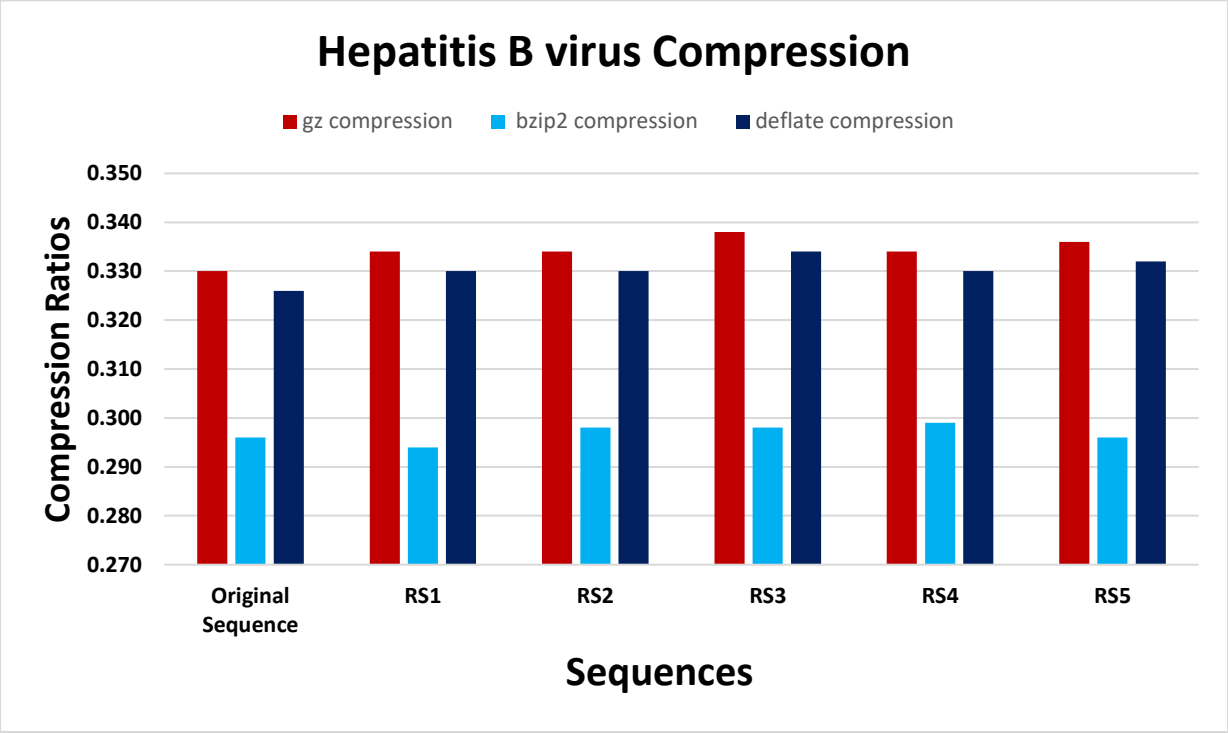


Fig. 6: Hepatitis B virus isolate G376-A6 Lossless Compression

Human Mitochondrial DNA	Compression Ratios		
	gz compression	bzip2 compression	deflate compression
Original Sequence	0.302	0.276	0.301
RS1	0.306	0.278	0.305
RS2	0.307	0.278	0.306
RS3	0.304	0.276	0.303
RS4	0.305	0.278	0.304
RS5	0.305	0.277	0.304

Table 6: Human Mitochondrial DNA compression data.

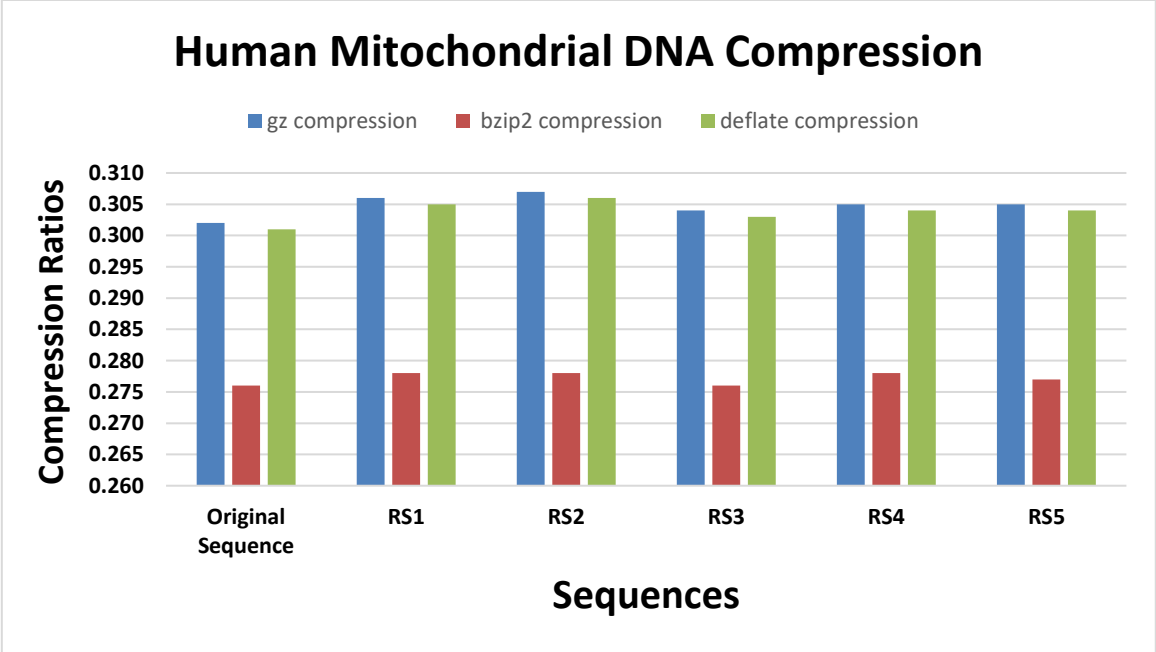


Fig. 7: Human Mitochondrial DNA Lossless Compression

S. cerevisae S288C chromosome I	Compression Ratios		
	gz compression	bzip2 compression	deflate compression
Original Sequence	0.287	0.270	0.287
RS1	0.295	0.273	0.295
RS2	0.295	0.273	0.295
RS3	0.294	0.273	0.294
RS4	0.295	0.273	0.295
RS5	0.294	0.273	0.294

Table 7: *Saccharomyces cerevisiae* S288C chromosome I compression data.

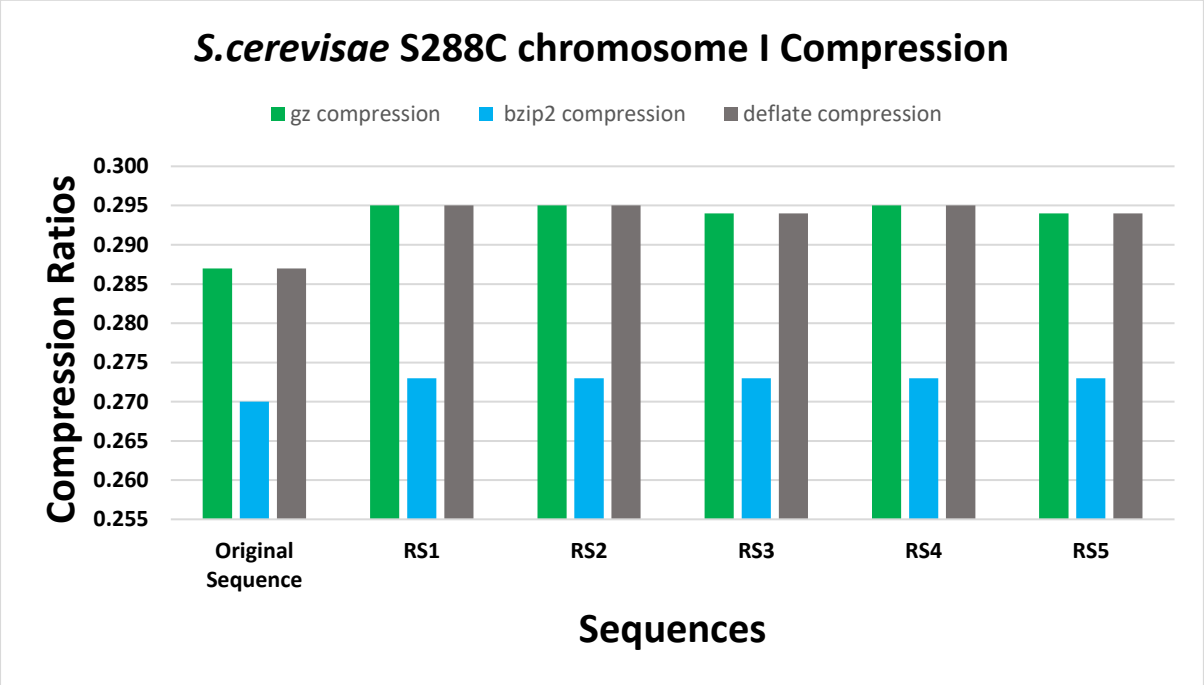


Fig. 8: *Saccharomyces cerevisiae* S288C chromosome I Lossless Compression

Biological Sequence	bzip2 compression Ratios
Hepatitis B virus	0.296
λ DNA	0.272
A. thaliana Chl. DNA	0.272
Human Mitochondrial DNA	0.276
S. cerevisae S288C chromosome I	0.270

Table 8: bzip2 compression ratio values for diverse genomic sequences.

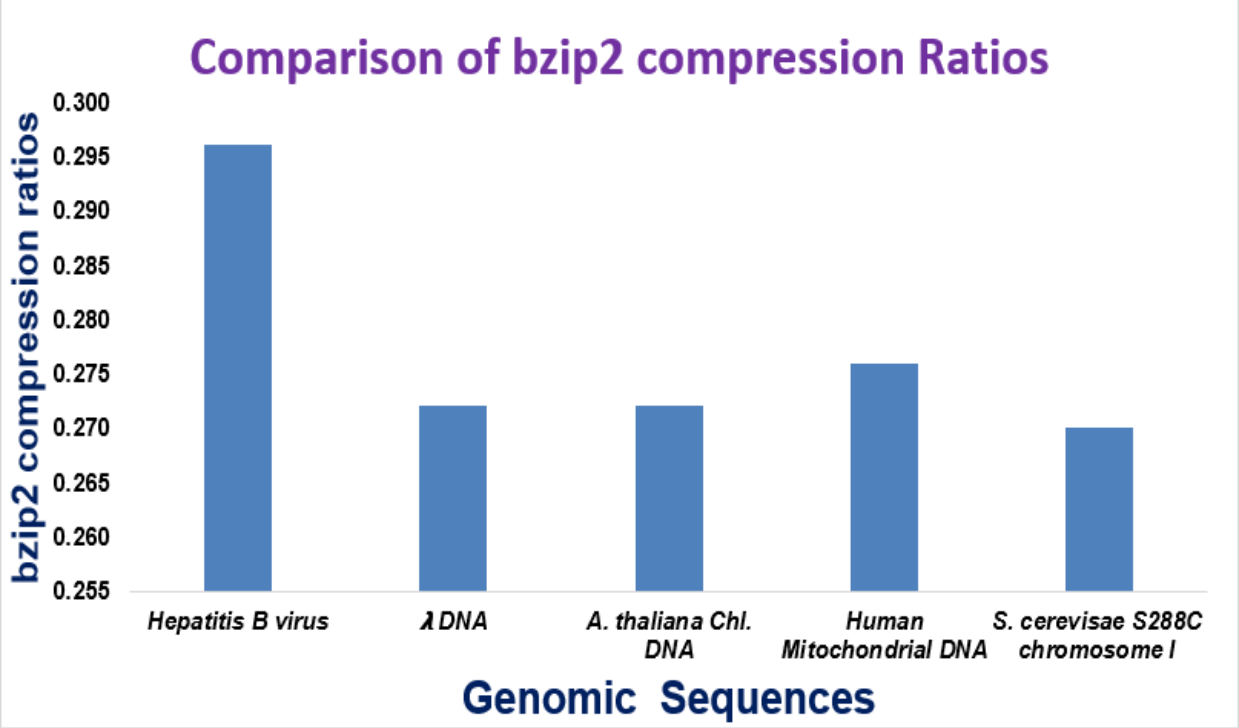


Fig. 9: Bar Graph depicting bzip2 compression values for diverse genomic sequences.

4.2 Compression of DNA sequence comparison for exons and introns

The experiment was performed by collecting the exon and intron sequences of 22 different randomly selected human genes from Ensembl genome browser. For each of the individual human gene, the sequence of exons and introns from single transcript were downloaded in FASTA format. All the exon and intron sequences of individual human gene were converted into the single string data by Macro recording in Notepad++ i.e., one string per exon or intron sequence.

All the exons and introns of each human gene were then subjected to lossless compression method- deflate, Gzip and Brotli respectively. After compression, their respective compression ratios were calculated and the whole set of data was demonstrated in Excel format file. The randomization of each of the original biological exon and intron sequences of individual human genes was done by re-shuffling through Shuffle DNA. The re-shuffling was done one time for each exon and intron sequences of every human.

Further, the lossless compression was done for each respective shuffled sequence. After lossless compression, the compression ratios were calculated accordingly. Determining the tool giving highest compression and enlisting the compression ratios obtained from the tool giving highest compression, for exons and introns in two separate columns. Different graphs were plotted to observe the comparison in Brotli compression ratios among exon and intron sequences of all the human genes.

After calculating the compression ratios, it was observed that the highest amount of compression in both exons and introns were obtained through **Brotli Lossless Compression**. Compression was observed more in introns than exons.

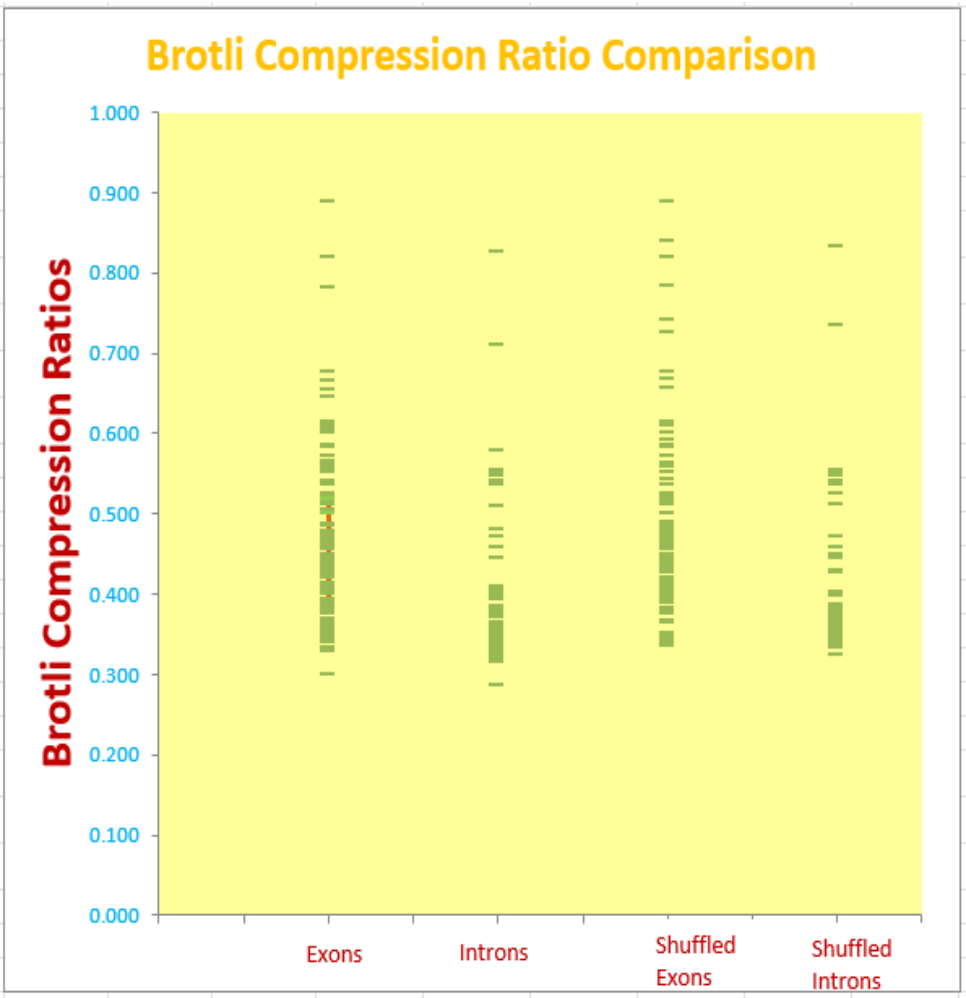


Fig. 10: X-Y Scatter Plot Representation

Based on scatter plot, it appears that exons have high compression ratio values with comparison to introns & their respective value mostly ranges between 0.3 to 0.7. Thus, it can be concluded that the compression in introns is higher than in exons. Exons have higher compression ratios than introns, however a similar result was also seen with randomized sequences.

The major reason for difference in intron and exon compression ratios may largely be attributed to base composition rather than the type of sequences.

	Compression Ratio		
		AM	S.D.
Exons		0.475	0.106
Exons Shuffled		0.478	0.113
Introns		0.379	0.079
Introns Shuffled		0.376	0.078

Table 9: Bar Graph Statistical Data.

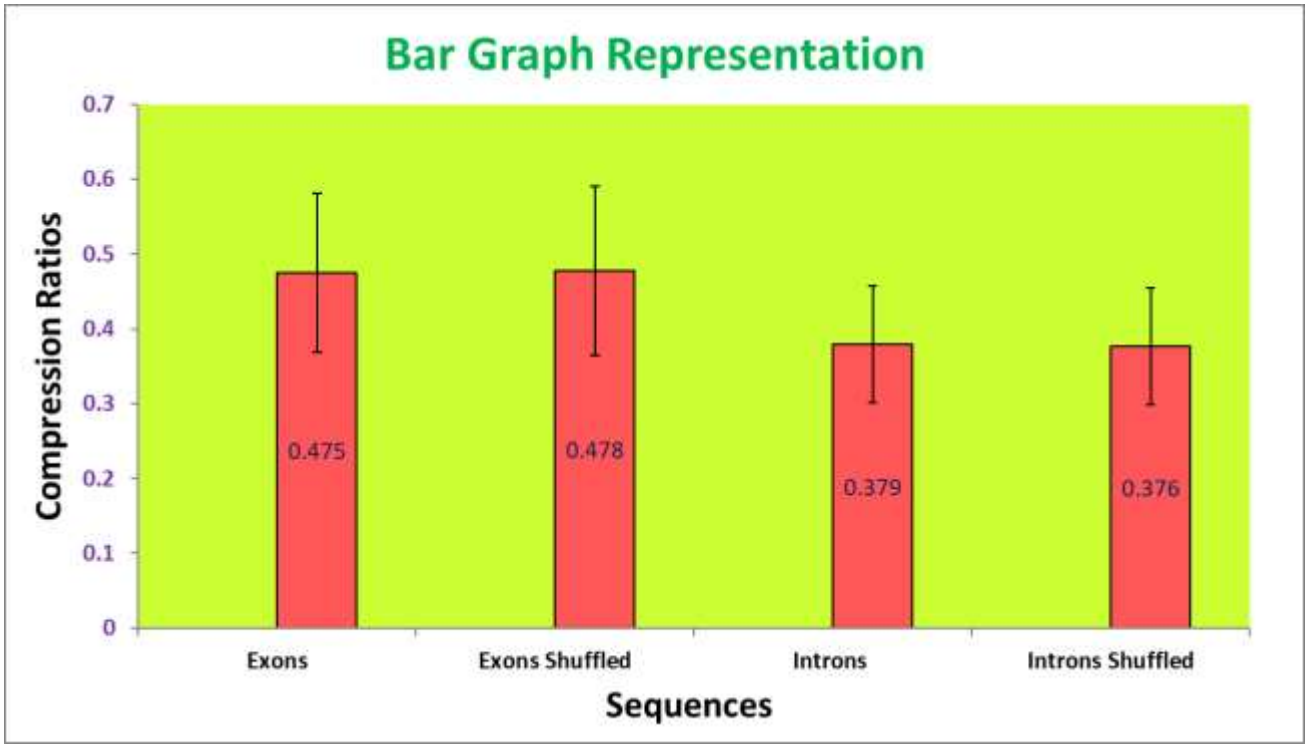


Fig. 11: Bar Graph Representation

The Bar Graph portrays the values of Average/Arithmetic Mean (A..M) and Standard Deviation (S.D.) for the **Brotli compression ratios** of exons and introns. The S.D. values are represented through error bars.

Exons		Difference
Min	0.300603	0.3006026
Q1	0.394558	0.0939552
Q2	0.457143	0.062585
Q3	0.526316	0.0691729
Max	0.888889	0.3625731

Introns		Difference
Min	0.28589	0.28589
Q1	0.341542	0.055652
Q2	0.352564	0.011022
Q3	0.378486	0.025922
Max	0.825544	0.447058

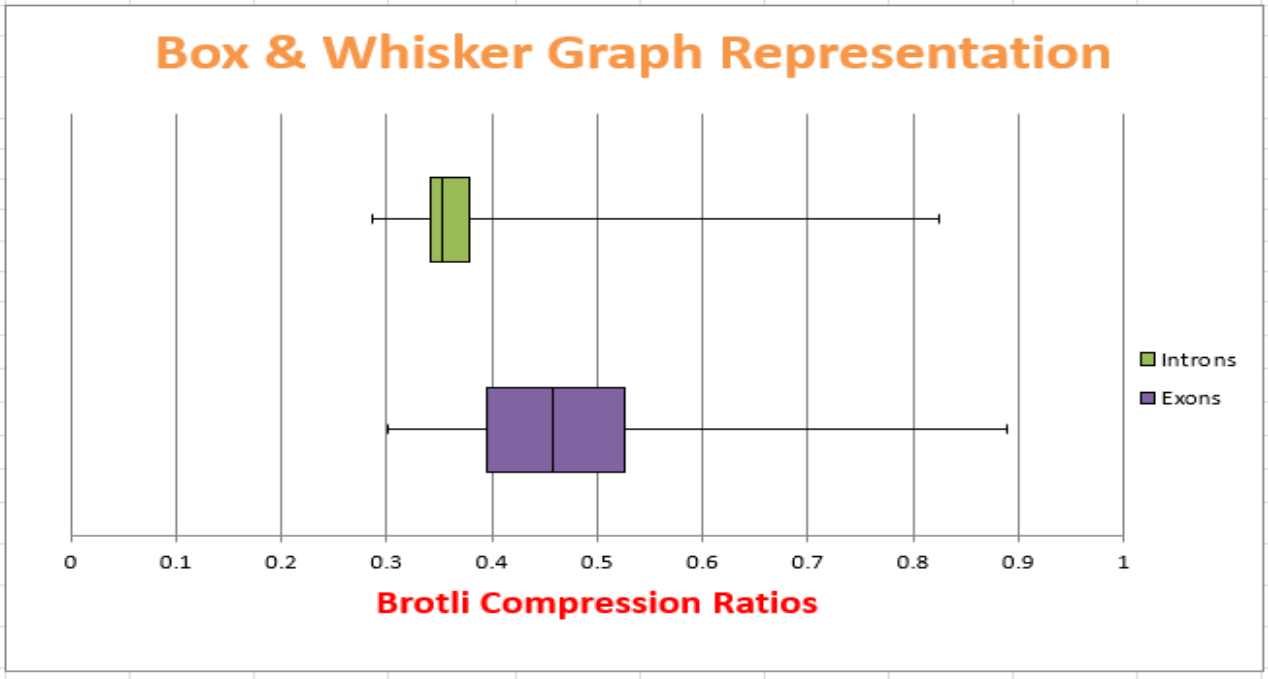


Fig. 12: Box & Whisker Graph Representation of Original Exonic and Intronic Biological Sequences

Exons		Difference
Min	0.335	0.335
Q1	0.397	0.062
Q2	0.459	0.062
Q3	0.520	0.061
Max	0.889	0.369

Introns		Difference
Min	0.324	0.324
Q1	0.340	0.016
Q2	0.349	0.009
Q3	0.3645	0.0155
Max	0.834	0.4695

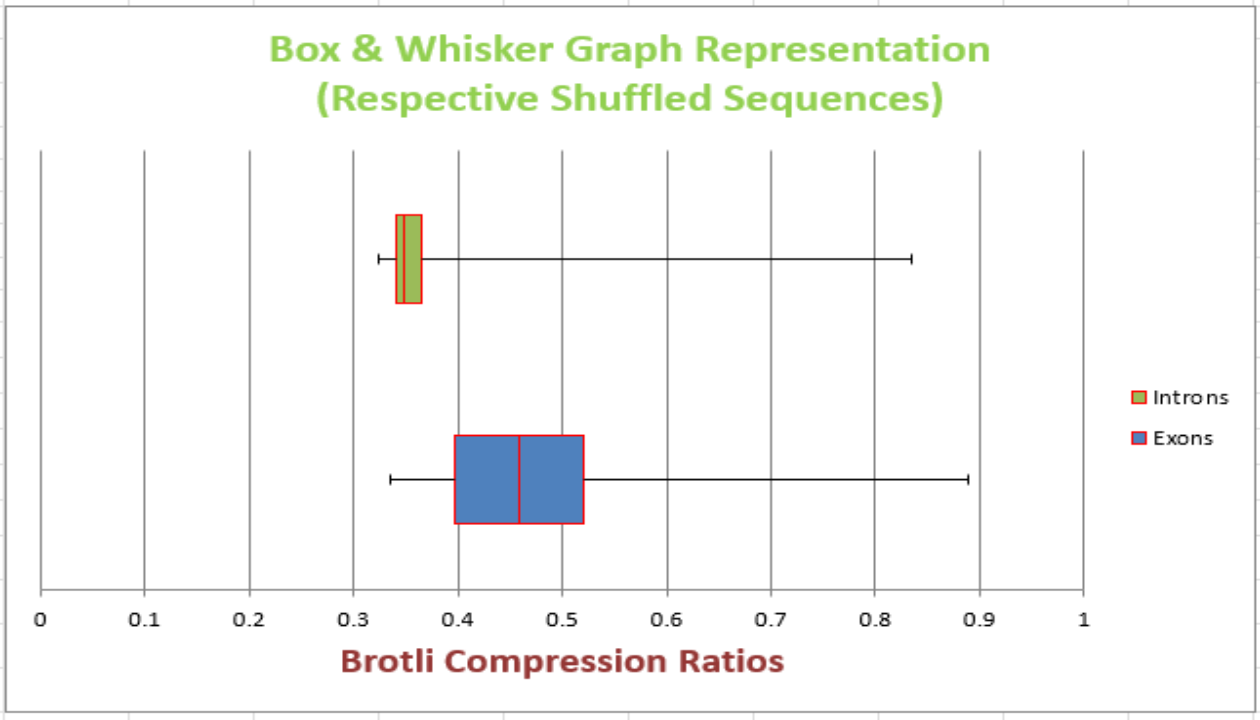


Fig. 13: Box & Whisker Graph Representation of Respective Shuffled Exonic and Intronic Biological Sequences

The Box & Whisker Plot determines:

1. **Minimum value:** Smallest value in data set.
2. **Second quartile:** A value that contains the bottom 25% of the data.
3. **Median value:** The middle number in a range of numbers.
4. **Third quartile:** A value that contains the upper 25% of the data.
5. **Maximum value:** Largest value in the data set.

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Exons	133	63.179	0.475	0.011
Exons Shuffled	133	63.508	0.478	0.013
Introns	111	42.103	0.379	0.006
Introns Shuffled	111	41.778	0.376	0.006

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	1.1731	3	0.391	41.620	6.08E-24	2.623
Within Groups	4.5473	484	0.009			
Total	5.7203	487				

Fig. 14: ANOVA Test Results

In order to compare the compression ratios of exons, introns, exons shuffled and introns shuffled sequences, if the sample belong to identical population, ANOVA was performed.

By performing ANOVA single factor, the p-value calculated was 6.07×10^{-24}.This p-value was less than the significance value $\alpha=0.05$ (estimated at 95% level of significance). Hence, we can conclude that the compression is more introns than exons.

Pairwise Comparisons		$HSD_{.05} = 0.0321$ $HSD_{.01} = 0.0390$	$Q_{.05} = 3.6458$ $Q_{.01} = 4.4259$
T_1:T_2	$M_1 = 0.48$ $M_2 = 0.48$	0.00	Q = 0.27 (*p* = .99747)
T_1:T_3	$M_1 = 0.48$ $M_3 = 0.38$	0.10	Q = 10.87 (*p* = .00000)
T_1:T_4	$M_1 = 0.48$ $M_4 = 0.38$	0.10	Q = 11.20 (*p* = .00000)
T_2:T_3	$M_2 = 0.48$ $M_3 = 0.38$	0.10	Q = 11.14 (*p* = .00000)
T_2:T_4	$M_2 = 0.48$ $M_4 = 0.38$	0.10	Q = 11.47 (*p* = .00000)
T_3:T_4	$M_3 = 0.38$ $M_4 = 0.38$	0.00	Q = 0.33 (*p* = .99556)

Fig. 15: Tukey HSD Test Results

NOTE: T_1 denotes the Exons, T_2 denotes Exons Shuffled, T_3 denotes the Introns and T_4 denotes Introns Shuffled.

The Tukey's HSD (honestly significant difference) procedure facilitates pairwise comparisons within your ANOVA data. The F-statistic in ANOVA tells you whether there is an overall difference between your sample means. Tukey's HSD test allows you to determine between which of the various pairs of means - if any of them - there is a significant difference.

A couple of things to note. First, a blue value indicates a significant result. Second, it's worth bearing in mind that there is some disagreement about whether Tukey's HSD is appropriate if the F-ratio score has not reached significance.

4.3 WW Runs Test

Sequence of characters can be ordered or complex. One of the ways of assessing the magnitude of disorder or randomness can be accomplished by Wald-Wolfowitz Runs Test.
WW Runs test includes calculation of various parameters, as in our case a biological sequence consists of four bases i.e., G, A, T&C the runs test can be extended to data with more than two categories, for in general: Mean is calculated as;

$$\mu_u = \frac{N(N+1) - \sum n_i^2}{N}$$

here, u stand for no. of runs.

The standard deviation is calculated by the formula:

$$\sigma_u = \frac{\sqrt{\sum n_i^2 \left[\sum n_i^2 + N(N+1)\right] - 2N \sum n_i^3 - N^3}}{\sqrt{N^2(N-1)}}$$

Where n_i is the number of items in category i, N is the total number of items (i.e., $N = \sum n_i$), and the summations are over all categories.

Finally, the Z-score is calculated as:

$$Z_c = \frac{|u - \mu_u| - 0.5}{\sigma_u}$$

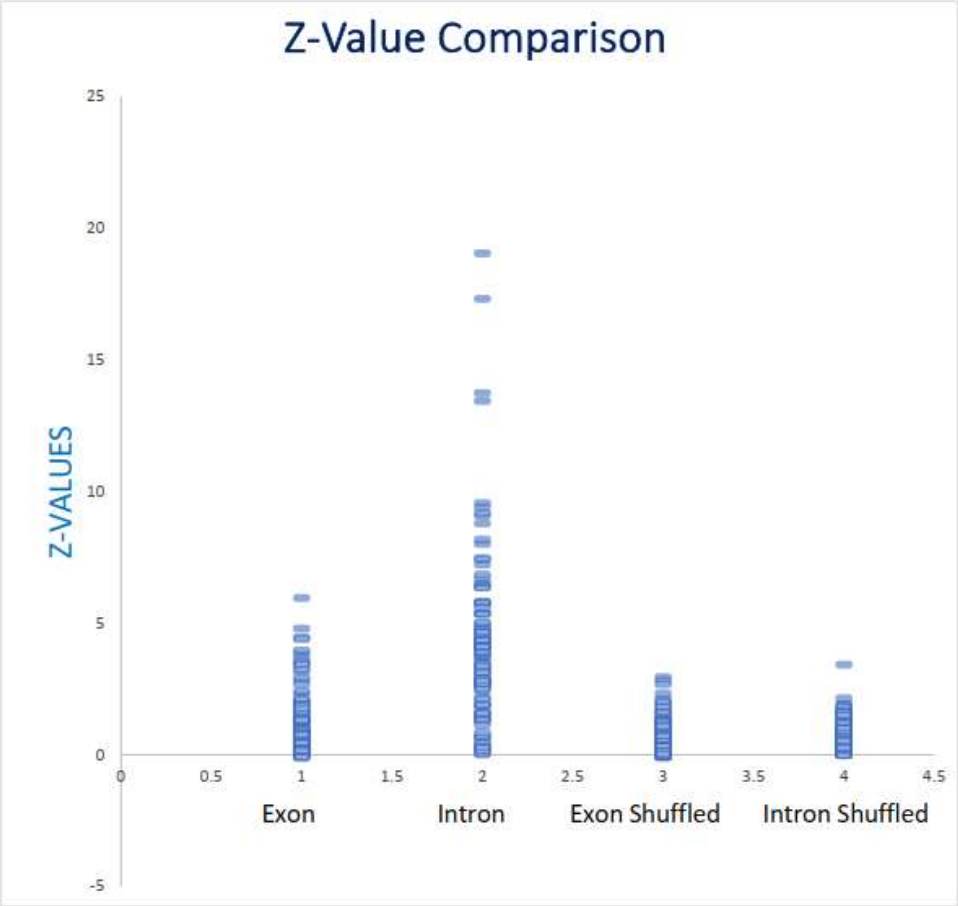


Fig. 16: Scatter Plot Representation showing Z-score comparison

Based on the scatter plot representing Z-score value comparison, we can infer that the exons have more complexity than introns. The complexity even increases in their respective shuffled sequences because as the sequences are shuffled their randomness increases as they become more complex.

	Exons	Introns	Exons Shu	Introns Shuffled
Min	-0.110	0.009	**-0.097**	-0.015
Q1	0.340	2.143	**0.192**	0.262
Q2 (Median)	0.765	3.702	**0.531**	0.575
Q3	1.469	5.169	**1.047**	1.103
Max	5.955	19.044	**2.946**	3.425
Mean	1.118	4.143	0.677	0.734
Range	6.065	19.035	3.043	3.440

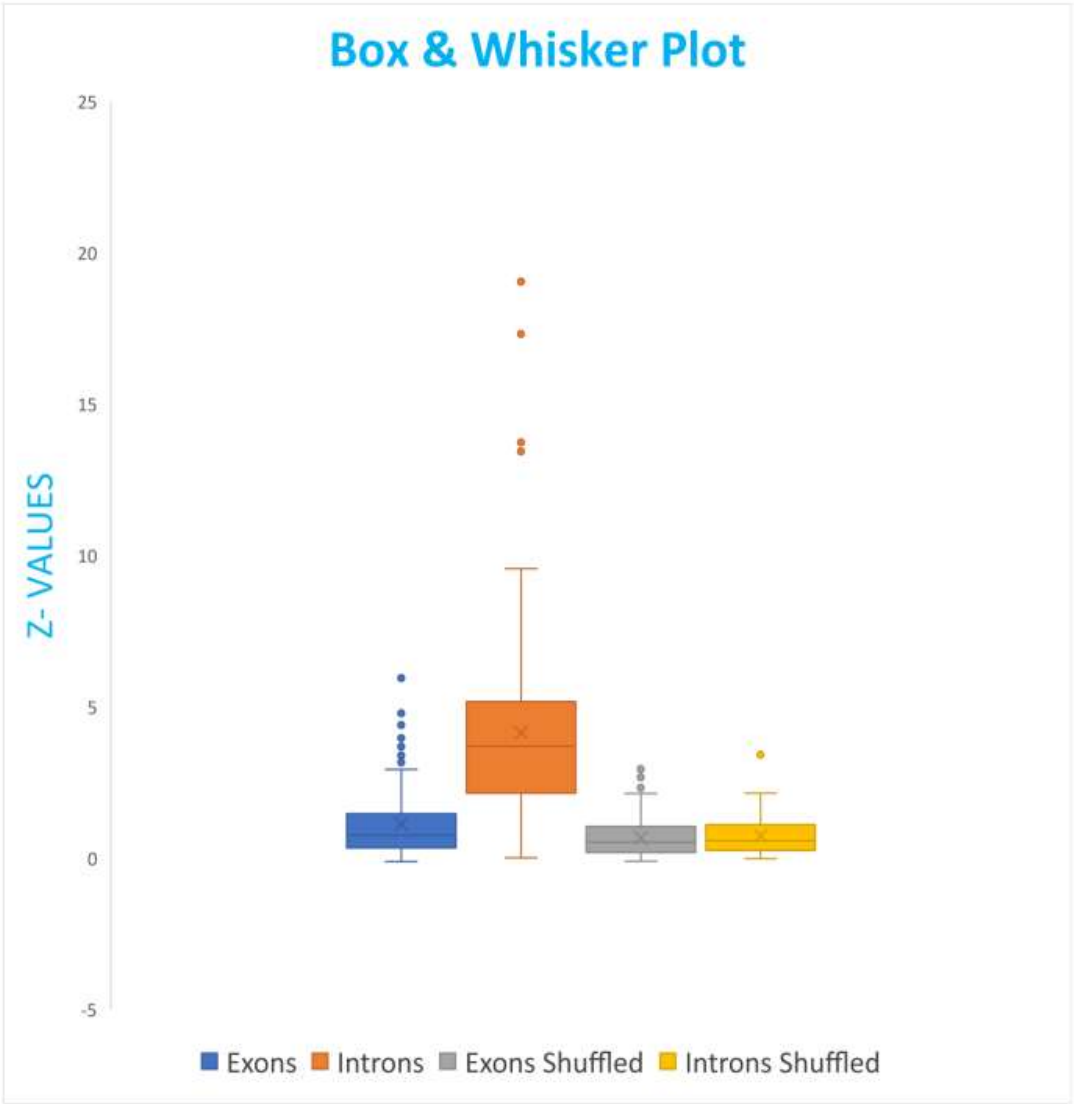


Fig. 17: Box-Whisker Graph Representation

Chapter 5

DISCUSSION

Complete genomic sequences from diverse range of organisms show a significant increase in genomic complexity. Changes include tapering of gene numbers due to retention of duplicate genes and a more rapid increase in the frequency of spliceosome introns and mobile genetic elements. The genome collects and stores information about the environment and is passed down from generation to generation.

Information theory shows that random sequences contain the least amount of information content, and well-preserved sequences contain the largest amount of information content. In other words, the actual information content ranges from zero in a completely random sequence to two bits in a stored sequence.

In our study, we have exploited the ways to assess the genome complexity of diverse range of organisms. One of the ways to confirm this is to use a regular archiving program to compress the DNA sequence. If the sequence is random, the compression is minimal. If the sequence is perfectly regular, the compression will be much higher.

The repetitive sequences are seen having less information content than the unique sequences. The more compressible the sequence , less will be the information content and vice-versa.

The exon (largely coding sequences) and intron (non-coding sequences) were studied by compression of their sequences. The exons have high compression ratio values with comparison to introns & their respective value mostly ranges between 0.3 to 0.7, however a similar result was also seen with randomized sequences. Thus, it can be concluded that the compression in introns is higher than in exons. The main reasons for difference in intron and exon sequences may be largely due to base composition parameter than the type of sequence. One major reason of not getting the expected results could be the fact that introns are usually several fold larger than exons in higher eukaryotes.

The randomness of exon and intron sequences was also checked to determine the sequence's complexity by performing the WW runs test. The scatter plot and box-whisker plot representing Z-score values give a reasonable inference that the exons have more complexity than introns, making introns more compressible. The complexity further increases in their respective shuffled sequences. As the sequences are shuffled, they become more and more complex.

Chapter 6

CONCLUSION

In the present study, we executed one of the ways to study the complexity of various genomic sequences by assessing their compressibility. Lossless compression methods were used to compress the genomic sequences. The comparison of randomness and complexity of intron and exon sequences were also inferred by using tools related to information theory.

Chapter 7

REFERENCES

1. Crick, F., 1970, Central Dogma of Molecular Biology. Nature 227, 561-563.
2. Edgell *et al.* 2000, Martinez-Arbaca & Toro 2000.
3. L. Rowen, G. Mahairas and L. Hood, “Sequencing the Human Genome,” Science, vol. 278, pp. 605-607, 1997.
4. Alberts *et al.* 2002. Molecular Biology of the Cell. 4th ed.
5. R. Giancarlo, D. Scaturro and F. Utro, “Textual Data Compression in Computational Biology: a synopsis,” Bioinformatics, vol. 25, no. 13, pp. 1575–1586, 2009.
6. X. Chen, M. Li, B. Ma and J. Tromp, “DNACompress: Fast and Effective DNA Sequence Compression,” Bioinformatics, vol. 18, no. 12, pp. 1696-1698, 2002.
7. Grumbach, S. and Tahi, F., A new challenge for compression algorithms: genetic sequences, J. Information.
8. S. Grumbach and F. Tahi, “Compression of DNA Sequences,” Proc. IEEE Symp. Data Compression, Snowbird, UT, pp. 340-350, 1993.
9. S. Grumbach and F. Tahi, “A New Challenge for Compression Algorithms: Genetic Sequences,” Information Processing Management, vol. 30, no. 6, pp. 875-886, 1994.
10. X. Chen, S. Kwong and M. Li, “A Compression Algorithm for DNA Sequences and its Applications in Genome Comparison,” Proc. 4th Ann. International Conf. Computational Molecular Biology, pp. 107, 2000.
11. X. Chen, M. Li, B. Ma and J. Tromp, “DNACompress: Fast and Effective DNA Sequence Compression,” Bioinformatics, vol. 18, no. 12, pp. 1696-1698, 2002.
12. Xin Chen *et al.*, DNACompress: fast and effective DNA sequence Compression, Bioinformatics Applications Note Vol. 18 no. 12 2002 Pages 1696–1698.
13. S Grumbach, F Tahi, LC INRIA, Compression of DNA sequences, Data Compression Conference, 1993. DCC'93., 1993.
14. Rivals, _E., Delgrange, O., Delahaye, J.-P., Dauchet, M., Delorme, M.-O., H_enaut, A., and Ollivier, E., Detection of signi_cant patterns by compression algorithms: the case of

approximate tandem repeats in DNA sequences, CABIOS, 13(2):131{136, 1997.
15. ttps://en.wikipedia.org/wiki/Run-length_encoding. www.datacompression.com
16. Salomon, David A Guide to Data Compression Methods. (London: Springer, 2001) [ISBN 0-387-95260-8].
17. Wayner, Peter Compression Algorithms for Real Programmers. (London: Morgan Kaufmann, 2000) [ISBN 0-12-788774-1].
18. Chapman, Nigel and Chapman, Jenny Digital Multimedia. (Chichester: John Wiley & Sons, 2000) [ISBN 0-471-98386-1].
19. Sayood, Khalid Introduction to Data Compression. 2nd edition (San Diego: Morgan Kaufmann, 2000) [ISBN 1-55860-558-4]
20. D. M. Loewenstern, H. Hirsh, P. Yianilos and M. Noordewier, “DNA Sequence Classification using Compression-based Induction,” Technical Report 95-04, DIMACS, 1995.
21. D. M. Loewenstern and P. N. Yianilos, “Significantly Lower Entropy Estimates for Natural DNA Sequences,” Proc. IEEE Data Compression Conference (DCC '97), pp. 151-160, 1997.
22. Edgell *et al*. 2000, Martinez-Arbaca & Toro 2000.
23. www.ncbi.nlm.nih.gov
24. https://asia.ensembl.org/index.html
25. Free online text compression tools - gzip, bzip2 and deflate https://www.txtwizard.net/compression
26. Text to Deflate Compress using gzip, deflate and Brotli algorithms. https://www.multiutil.com/.
27. https://www.bioinformatics.org/sms2/shuffle_dna.html
28. Principles of Gene Manipulation, 6th edition (2001).