



Cairo University

EFFICIENT COMPRESSION FOR DNA SEQUENCES USING TREE OF FILES

BY

Eng. Nour Saeed Ibrahim Bakr

A Thesis Submitted to the Faculty of Engineering at Cairo University in
Partial Fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY
In
Biomedical Engineering and Systems

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2018

EFFICIENT COMPRESSION FOR DNA SEQUENCES USING TREE OF FILES

BY

Eng. Nour Saeed Ibrahim Bakr

A Thesis Submitted to the Faculty of Engineering at Cairo University in
Partial Fulfillment of the Requirements for the Degree of

**DOCTOR OF PHILOSOPHY
In
Biomedical Engineering and Systems**

Under Supervision of

Prof. Dr. Amr Abdel Rahman Sharawi

Associate Professor, Biomedical Engineering and Systems Department,
Faculty of Engineering, Cairo University, Egypt.

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2018

EFFICIENT COMPRESSION FOR DNA SEQUENCES USING TREE OF FILES

BY

Eng. Nour Saeed Ibrahim Bakr

A Thesis Submitted to the Faculty of Engineering at Cairo University in
Partial Fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY
In
Biomedical Engineering and Systems

Approved By the Examining Committee:

Prof. Dr.: Amr Abdel Rahman Sharawi , **Thesis Main Advisor**
Associate Professor, Biomedical Engineering and Systems Department,
Faculty of Engineering, Cairo University, Egypt.

Prof. Dr.: Manal Abdel Wahed Abdel Fattah Abdel Wahed , **Internal Examiner**
Professor, Biomedical Engineering and Systems Department,
Faculty of Engineering, Cairo University, Egypt.

Prof. Dr.: Ahmed Farag Ali Mohamed Seddik , **External Examiner**
Professor, Biomedical Engineering Department,
Faculty of Engineering, Helwan University, Egypt.

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2018

Engineer: Nour Saeed Ibrahim Bakr.
Date of Birth: 19 / 2 / 1981
Nationality: Egyptian.
E-mail: bioeng_nour@ieee.org
Phone. 01000-98-1103
Address: Sharqia Governorate, 10th of Ramadan City,
Neighboring 27, Plot 278.
Registration Date: 1 / 3 / 2011
Awarding Date: / / 2018
Degree: Doctor of Philosophy.
Department: Biomedical Engineering and Systems.



Supervisor: Ass. Prof. Amr Abdel Rahman Sharawi.

Examiners: Ass. Prof. Amr Abdel Rahman Sharawi. , Thesis Main Advisor
Prof. Manal Abdel Wahed Abdel Fattah Abdel Wahed. , Internal Examiner
Prof. Ahmed Farag Ali Seddik, Helwan University. , External Examiner

Title of Thesis: Efficient Compression for DNA Sequences using Tree of Files

Key Words: DNA - compression – gzip – Bzip2 - 7z

Summary:

Huge amounts of genomic data generated by Modern DNA sequencing instruments present a difficult challenge to effective storage and fast transmission. This data is important for the visions it permits into the health of individuals and whole populations, and will continue to be of advantage into the future as medical knowledge grows. General purpose compression tools fail regarding this challenge. Special algorithms have been proposed for the compression of genomics data but at the cost of time.

We present a novel, fast, loss-less and reference-free compression method to improve the compression ratio and compression time of DNA sequence compression using both a tree of files and some general purpose compression tools namely gzip, Bzip2 and 7z. It consists of three pre-processing steps then applying these tools on the input sequences. The first step is to generate tree of binary files and then reduce the size of a largest binary file and finally convert all these binary files to symbolic files.

Validation results show successful compression at a reasonable compression ratio while providing the lowest compression time when compared with popular top compression algorithms. The compression ratio of gzip, Bzip2 and 7z on DNA sequences after pre-processing steps was also improved by 6.59%, 7.12% and 5.69% respectively on average and at a high speed equivalent to ten times the best available methods.

ACKNOWLEDGMENTS

Every thing I have has been given to me from **God** and I am eternally grateful that His blessings and guidance that supported me in all my endeavors.

I would especially like to thank my supervisors **Prof. Amr A. Sharawi** for all of the faith, support, encouragement and feedback he has provided to me in this research, till its final stages.

Special thanks are sent to all my leaders and colleagues at **Higher Technological Institute**, especially **Dr. Khaled A. Shafei** and **Dr. Amal S. Eldesouky** who supported me throughout the work, and was kind to help at times of crisis.

Also, I give special thanks to **my family**, especially **my Parents** and **my Wife**. I could not have done it without their continuous encouragement and support.

Finally, **to all those who helped me along the way**, I thank you.

Nour S. Bakr

TABLE OF CONTENTS

Acknowledgements.	i
Table of contents.	ii
List of Figures.	v
List of Tables.	vi
List of Abbreviations.	viii
Abstract.	xi
 Chapter 1: Introduction.	 1
1.1 Thesis overview.	1
1.1.1 Problem Definition.	1
1.1.2 Thesis Objectives.	1
1.1.3 Test samples.	1
1.1.4 Methods.	2
1.1.5 Results.	2
1.2 Thesis Organization.	2
 Chapter 2: Background and Related Work.	 3
2.1 Biological Overview.	3
2.1.1 Bioinformatics.	3
2.1.2 Live organisms.	3
2.1.3 The cell.	4
2.1.4 DNA.	6
2.1.5 Human genome.	8
2.2 Genetic data.	9
2.2.1 DNA sequencing process.	9
2.2.2 Genome Projects.	10
2.2.3 Biological database.	11
2.2.3.1 Overview.	11
2.2.3.2 GenBank Data Base.	11
2.2.3.3 Biological data in GenBank.	12
2.3 Data compression.	14
2.3.1 Introduction.	14
2.3.1.1 Lossless and Lossy Compression.	14
2.3.1.2 Compression ratio.	14
2.3.1.3 Most famous general purpose approaches.	14
2.3.2 Compression of biological sequences.	15
2.3.2.1 Important of Genetic data compression.	15

2.3.2.2 Compression modes of Genetic data.	16
2.3.2.2.1 Horizontal (reference free) Mode.	16
2.3.2.2.2 Vertical (reference based) Mode.	17
2.3.2.3 Limitations of reference based Methods.	18
2.3.3 Compression of DNA using general purpose approaches.	18
2.3.4 Special DNA compression approaches.	19
2.3.5 The importance of General Purpose Approaches.	25
Chapter 3: The Proposed Compression Method.	27
3.1 Introduction.	27
3.2 DNA test sequences.	28
3.3 Methodology.	30
3.3.1 Overview.	30
3.3.2 Compression Method.	30
3.3.2.1 Generate file tree and built three binary files.	30
3.3.2.2 Minimize the size of a largest binary file.	32
3.3.2.3 Convert all binary files to symbolic files.	32
3.3.2.4 Apply the general purpose compression algorithm.	33
3.3.3 Compression algorithm.	34
3.3.4 Decompression Method.	36
3.3.4.1 Apply the general purpose Decompression algorithm.	36
3.3.4.2 Generate file tree and convert all symbolic files to binary files.	36
3.3.4.3 Restore the original size of the main binary file.	37
3.3.4.4 Generate DNA sequence file.	38
3.3.5 Decompression algorithm.	39
Chapter 4: Results and Discussions.	41
4.1 Compression of complete genomes of five Mitochondria.	41
4.1.1 Results.	41
4.1.2 Discussions.	43
4.2 Compression of complete genomes of five Chloroplasts.	44
4.2.1 Results.	44
4.2.2 Discussions.	45
4.3 Compression of complete genomes of ten Bacteria.	46
4.3.1 Results.	46
4.3.2 Discussions.	48
4.4 Compression of all genomes.	49
4.4.1 Results.	49
4.4.2 Discussions.	50

4.5 Comparing with other popular compression algorithms from point of view of compression ratio and compression time.	51
4.5.1 Results.	51
4.5.2 Discussions.	52
4.6 Over all discussions.	52
Chapter 5: Conclusions and Future Work.	55
5.1 Conclusions.	55
5.2 Future Work.	56
List of Publications.	57
References	58
Appendix A: The 64 symbols and their corresponding binary numbers.	66
Appendix B: Summary of some DNA reference-free algorithms.	67
Appendix C: Genbank and WGS statistics.	69
Appendix D: Benefits of genome research.	72
Appendix E: Lossless compression algorithms.	75
ملخص البحث	

LIST OF FIGURES

Figure		Page
2.1	Animal cell components.	5
2.2	Cell nucleus contains the chromosomes.	5
2.3	The structures of the four nucleotides in DNA.	6
2.4	Structure of genetic information which is arranged in a two stranded double helix and this DNA forms the chromosomes.	7
2.5	Successive enlargements of an organism to focus on the genetic material.	8
2.6	Automatic DNA Sequencer (illumine Hiseq2000).	9
2.7	Output of the DNA Sequencer.	10
2.8	Growth of biological data in GenBank and Whole Genome Shotgun projects.	12
2.9	Historic cost of sequencing a human genome.	13
3.1	Overview of the proposed method.	27
3.2	Voss mapping for the DNA sequence.	30
3.3	Generated tree with ten files for compression process.	31
3.4	Generated tree with six files from the input five files in the rectangular for decompression process.	36
4.1	The Average compression ratios using (gzip, Bzip2 and 7z) algorithms for all twenty complete genomes directly or after three preprocess steps.	49

LIST OF TABLES

Table		page
2.1	Growth of biological data in GenBank and WGS.	12
3.1	Complete genomes of five Mitochondria.	28
3.2	Complete genes of five Chloroplasts.	29
3.3	Complete genomes of ten Bacteria.	29
4.1	Compression ratio in bits/base (bpb) for the complete genomes of five Mitochondria before (C.R1) and after (C.R2) and percentage reduction achieved using the three preprocessing steps with gzip compression algorithm.	41
4.2	Compression ratio in bits/base (bpb) for the complete genomes of five Mitochondria before (C.R1) and after (C.R2) and percentage reduction achieved using the three preprocessing steps with Bzip2 compression algorithm.	42
4.3	Compression ratio in bits/base (bpb) for the complete genomes of five Mitochondria before (C.R1) and after (C.R2) and percentage reduction achieved using the three preprocessing steps with 7z compression algorithm.	42
4.4	Average compression ratio in bits/base (bpb) for the complete genomes of five Mitochondria before (C.R1) and after (C.R2) and percentage reduction achieved using the three preprocessing steps with all compression algorithms.	43
4.5	Compression ratio in bits/base (bpb) for the complete genomes of five Chloroplasts before (C.R1) and after (C.R2) and percentage reduction achieved using the three preprocessing steps with gzip compression algorithm.	44
4.6	Compression ratio in bits/base (bpb) for the complete genomes of five Chloroplasts before (C.R1) and after (C.R2) and percentage reduction achieved using the three preprocessing steps with Bzip2 compression algorithm.	44
4.7	Compression ratio in bits/base (bpb) for the complete genomes of five Chloroplasts before (C.R1) and after (C.R2) and percentage reduction achieved using the three preprocessing steps with 7z compression algorithm.	45

4.8	Average compression ratio in bits/base (bpb) for the complete genomes of five Chloroplasts before (C.R1) and after (C.R2) and percentage reduction achieved using the three preprocessing steps with all compression algorithms.	45
4.9	Average compression ratio in bits/base (bpb) for the complete genomes of ten Bacteria before (C.R1) and after (C.R2) and percentage reduction achieved using the three preprocessing steps with all compression algorithms.	46
4.10	Compression ratio in bits/base (bpb) for the complete genomes of ten Bacteria before (C.R1) and after (C.R2) and percentage reduction achieved using the three preprocessing steps with gzip compression algorithm.	47
4.11	Compression ratio in bits/base (bpb) for the complete genomes of ten Bacteria before (C.R1) and after (C.R2) and percentage reduction achieved using the three preprocessing steps with Bzip2 compression algorithm.	47
4.12	Compression ratio in bits/base (bpb) for the complete genomes of ten Bacteria before (C.R1) and after (C.R2) and percentage reduction achieved using the three preprocessing steps with 7z compression algorithm.	48
4.13	Average compression ratio in bits/base (bpb) for all genomes sequences before (C.R1) and after (C.R2) and percentage reduction achieved using the three preprocessing steps with all compression algorithms.	49
4.14	The increasing and decreasing percentage compression ratios using (gzip, Bzip2 and 7z) for all twenty complete genomes directly or after three preprocess steps.	50
4.15	Comparing the compression ratio in bits/base (bpb) of the largest genomes size between our proposed method and others popular compression algorithms.	51
4.16	Comparing the compression time in minutes (sec) of the largest genomes size between our proposed method and others popular compression algorithms.	51
A.1	The 64 symbols and their corresponding binary numbers.	66
B.1	Compression ratio in bits/base (bpb) for some DNA reference-free algorithms on the standard eleven sequences.	67
B.2	Comparison for some DNA reference-free algorithms.	68
C.1	Data growth of GenBank and WGS from Dec. 1982 to Oct. 2017.	69

LIST OF ABBREVIATIONS

A	Adenine- DNA Base.
Arith-2	Order-2 Arithmetic coding, a lossless compression algorithm.
ASCII	American Standard Code for Information Interchange.
AVR	Approximate Repeat Vector.
AdpISPO	Self-adaptive Intelligent Single Particle Optimizer.
B1 : B10	Complete genomes of ten Bacteria.
bpb	bit per base.
BWT	Burrows–Wheeler Transform.
Bio-compress	DNA Compression Algorithm, reference-free, 1993.
Bio-compress2	DNA Compression Algorithm, reference-free, 1994.
BIND	DNA Compression Algorithm, reference-free, 2012.
BEETL	DNA Compression Algorithm, reference-free, 2014.
C	Cytosine- DNA Base.
CPU	Central Processing Unit.
C++	A general-purpose programming language.
C-fact	DNA Compression Algorithm, reference-free, 1996.
CR	Compression Ratio.
CLPSO	Comprehensive Learning Particle Swarm Optimizer.
CRAM	DNA Compression Algorithm, reference-based.
CTW	Context Tree Weighting, a lossless compression algorithm.
C1 : C5	Complete genomes of five Chloroplasts.
CTW+LZ	DNA Compression Algorithm, reference-free, 2000.
COMRAD	DNA Compression Algorithm, reference-free, 2009.
DNA	Deoxyribonucleic acid.
DDBJ	DNA Data Bank of Japan.
DNA compress	DNA Compression Algorithm, reference-free, 2002.
DNAC	DNA Compression Algorithm, reference-free, 2004.
DNASequitur	DNA Compression Algorithm, reference-free, 2004.
DNA Pack	DNA Compression Algorithm, reference-free, 2005.

DSRC	DNA Compression Algorithm, reference-free, 2011.
DNAEnc3	DNA Compression Algorithm, reference-free, 2011.
DELIMINATE	DNA Compression Algorithm, reference-free, 2012.
DNA-compress	DNA COMpression based on a Pattern-Aware Contextual modeling Technique algorithm, reference-free, 2013.
DSCR2	DNA Compression Algorithm, reference-free, 2014.
EOF	End of File.
EMBL	European Bioinformatics Institute database.
DEFLATE	A lossless data compression algorithm.
FASTQZ	DNA Compression Algorithm, reference-based.
FCM-MX	Finite-Context Models, with orders (X) from two to sixteen.
Fqzcomp	DNA Compression Algorithm, reference-free, 2013.
Fastqz	DNA Compression Algorithm, reference-free, 2013.
FASTQ	A text-based format for storing both a nucleotide sequence and its corresponding quality scores.
FASTA	A text-based format for representing a nucleotide sequences.
G	Guanine- DNA Base.
Gen-compress	DNA Compression Algorithm, reference-free, 1999.
GeNML	DNA Compression Algorithm, reference-free, 2005.
GPCAs	General Purpose Compression Approaches.
GHz	Gigahertz, a unit of frequency.
GB	Gigabytes, a unit of memory size.
GSQZ	DNA Compression Algorithm, reference-free, 2010.
GeCo	DNA Compression Algorithm, reference-free, 2016.
GEO	Gene Expression Omnibus.
HTS	High Throughput Sequencing.
INSDC	International Nucleotide Sequence Database Collaboration.
ICGC	International Cancer Genome Consortium Project – 2010.
JGI	Joint Genome Institute.
K-mer	Typically refers to all the possible substrings of length k that are contained in a string.
LZ	Lempel–Ziv, a lossless data compression algorithm.

M1 : M5	Complete genomes of five Mitochondria.
MINCE	DNA Compression Algorithm, reference-free, 2015.
MST	Maximum Spanning Tree.
NML compress	DNA Compression Algorithm, reference-free, 2003.
NCBI	National Center for Biotechnology Information.
NGS	Next-Generation Sequencing.
ORCOM	DNA Compression Algorithm, reference-free, 2014.
PC	Personal Computer.
POMA	DNA Compression Algorithm, reference-free, 2011.
PATHENC	DNA Compression Algorithm, reference-based.
QUIP	DNA Compression Algorithm, reference-free, 2012.
ReCoil	DNA Compression Algorithm, reference-free, 2011.
RQS	DNA Compression Algorithm, reference-free, 2014.
RAM	Random Access Memory.
RNA	Ribonucleic acid.
SCAs	Specialized Compression Approaches.
SNP	Single nucleotide polymorphism.
SCALCE	DNA Compression Algorithm, reference-free, 2012.
SeqDB	DNA Compression Algorithm, reference-free, 2013.
SOLEXA	Illumina Solexa sequencing technology.
SOLiD	Sequencing by Oligonucleotide Ligation and Detection, a next-generation DNA sequencing technology.
SRA	Short Read Archive.
T	Thymine- DNA Base.
U	Uracil- RNA Base.
US	United States of America.
WGS	Whole Genome Shotgun projects.
XFCMs	Expanded Finite-Context Models.
XM	Expert Model, DNA Compression Algorithm, reference-free, 2007.
ZIP	An archive file format that supports lossless data compression.

ABSTRACT

Huge amounts of genomic data generated by Modern DNA sequencing instruments present a difficult challenge to effective storage and fast transmission. This data is important for the visions it permits into the health of individuals and whole populations, and will continue to be of advantage into the future as medical knowledge grows. General purpose compression tools fail regarding this challenge. Special algorithms have been proposed for the compression of genomics data but at the cost of time.

We present a novel, fast, loss-less and reference-free compression method to improve the compression ratio and compression time of DNA sequence compression using both a tree of files and some general purpose compression tools namely gzip, Bzip2 and 7z. It consists of three pre-processing steps then applying these tools on the input sequences. The first step is to generate tree of binary files and then reduce the size of a largest binary file and finally convert all these binary files to symbolic files.

Validation results show successful compression at a reasonable compression ratio while providing the lowest compression time when compared with popular top compression algorithms. The compression ratio of gzip, Bzip2 and 7z on DNA sequences after pre-processing steps was also improved by 6.59%, 7.12% and 5.69% respectively on average and at a high speed equivalent to ten times the best available methods.